Machine learning methods for efficient orthogonal functional basis selection to represent stochastic processes and functional data

Krzysztof Podgórski
Department of Statistics
Lund University

October 1, 2019

Credits

The target is develop an R-package for data driven basis selection for stochastic analysis and fda that combines machine learning methods and uses efficient treatment of spline orthogonalization. The work is jointly with **Xijia Liu, Umeå University** and **Hiba Nassar, Lund University**

Outline

Highly dimensional data through functional data

- In many recent application, the problem of high dimension relatively to sample size appears.

Highly dimensional data through functional data

- In many recent application, the problem of high dimension relatively to sample size appears.
- One of the approaches to eliminate this ‘overparameterization’ is through functional data.

Highly dimensional data through functional data

- In many recent application, the problem of high dimension relatively to sample size appears.
- One of the approaches to eliminate this ‘overparameterization’ is through functional data.
- In it multidimensional data are conceptually treated as functions.

Highly dimensional data through functional data

- In many recent application, the problem of high dimension relatively to sample size appears.
- One of the approaches to eliminate this ‘overparameterization’ is through functional data.
- In it multidimensional data are conceptually treated as functions.
- There are many issues with such approach that arises from the fact that functions are from infinite dimensional spaces and data always are finite dimensional.

Highly dimensional data through functional data

- In many recent application, the problem of high dimension relatively to sample size appears.
- One of the approaches to eliminate this ‘overparameterization’ is through functional data.
- In it multidimensional data are conceptually treated as functions.
- There are many issues with such approach that arises from the fact that functions are from infinite dimensional spaces and data always are finite dimensional.
- One of the key problem is a selection of convenient orthonormal basis.

Functional data

- **Random functions** of, let say, $\mathbf{t} \in \mathbb{R}^r$ observed 'repetitively' are frequently referred to as functional data.

Functional data

- **Random functions** of, let say, $\mathbf{t} \in \mathbb{R}^r$ observed 'repetitively' are frequently referred to as functional data.
- For simplicity, from now on functional data are

$$(x_i)_{i=1}^n\,, \tag{1}$$

  where $x_i = x_i(t)$'s, $t \in [0,1]$ produced by some **random 'mechanism'**.
- Reduce random 'mechanism' to independent sampling – **iid fd sample**
- Additionally we assume for each fixed t and s:

$$\sigma(t,s) = Cov(X(t), X(s))$$

  is well-defined.

How 'complicated' can be such data?

Not very complicated – Karhunen-Loéve Theorem

Theorem (Karhunen-Loéve Theorem – Part I)

If $\sigma(t,s)$ is continuous, then there exist a square summable sequence of non-negative numbers λ_k, an orthonormal basis e_k, $k \in \mathbb{N}_0$ in $L_2[0,1]$ and a sequence of zero-mean variance-one random variables Z_i such that

$$X(t) = \sum_{k=0}^{\infty} \sqrt{\lambda_k} Z_k \, e_k(t),$$

where the convergence is in the mean squared value and is uniform in t.

Random variables Z_k are 'observable'

$$Z_k = \int_0^1 X(t) e_k(t) \, dt / \sqrt{\lambda_i}.$$

Convenience of the representation

Theorem (Karhunen-Loéve Theorem – Part II)

The covariance function of the process is represented in the uniform convergence over $[0,1]^2$ *as*

$$\sigma(s,t) = \sum_{k=0}^{\infty} \lambda_k e_k(s) \bar{e}_k(t). \tag{2}$$

If we assume that λ_k *is also summable, then the average error in the norm of* $L_2[0,1]$ *can be bounded as follows*

$$E\|X - \sum_{k=0}^{n} \sqrt{\lambda_k} Z_k e_k\|_2^2 = \sum_{k=n+1}^{\infty} \lambda_k E Z_k^2 = \sum_{k=n+1}^{\infty} \lambda_k.$$

When would ‘things’ be easy for a statistician?

When would ‘things’ be easy for a statistician?

- If orthonormal basis e_k, $k \in \mathbb{N}_0$ were known, then we would observe independently for $i = 1, \ldots, n$

$$V_{k,i} = \sqrt{\lambda_k} Z_{k,i} = \int_0^1 X_i(t) e_k(t)\, dt, \ k = 0, 1, \ldots$$

and we can ‘easily’ **estimate** λ_k’s.

When would 'things' be easy for a statistician?

- If orthonormal basis e_k, $k \in \mathbb{N}_0$ were known, then we would observe independently for $i = 1, \ldots, n$

$$V_{k,i} = \sqrt{\lambda_k} Z_{k,i} = \int_0^1 X_i(t) e_k(t)\, dt, \ k = 0, 1, \ldots$$

and we can 'easily' **estimate** λ_k's.

The main problem is that the orthogonal basis of eigenfunctions is not known.

When would 'things' be easy for a statistician?

- If orthonormal basis e_k, $k \in \mathbb{N}_0$ were known, then we would observe independently for $i = 1, \ldots, n$

$$V_{k,i} = \sqrt{\lambda_k} Z_{k,i} = \int_0^1 X_i(t) e_k(t)\, dt, \ k = 0, 1, \ldots$$

and we can 'easily' **estimate** λ_k's.

The main problem is that the orthogonal basis of eigenfunctions is not known.

What can be done about it?

From sample of functions to sample of vectors and back

The most common approach in basic steps

From sample of functions to sample of vectors and back

The most common approach in basic steps

1. Pick-up a basis of functions $\{f_k\}_{k=0}^{\infty}$ that spans the space of all possibly observed functions –

From sample of functions to sample of vectors and back

The most common approach in basic steps

1. Pick-up a basis of functions $\{f_k\}_{k=0}^{\infty}$ that spans the space of all possibly observed functions – **somewhat arbitrarily**.

From sample of functions to sample of vectors and back

The most common approach in basic steps

1. Pick-up a basis of functions $\{f_k\}_{k=0}^{\infty}$ that spans the space of all possibly observed functions – **somewhat arbitrarily**.
2. Choose the number of basis elements N –

From sample of functions to sample of vectors and back

The most common approach in basic steps

1. Pick-up a basis of functions $\{f_k\}_{k=0}^{\infty}$ that spans the space of all possibly observed functions – **somewhat arbitrarily**.
2. Choose the number of basis elements N – **somewhat arbitrarily**.

From sample of functions to sample of vectors and back

The most common approach in basic steps

1. Pick-up a basis of functions $\{f_k\}_{k=0}^{\infty}$ that spans the space of all possibly observed functions – **somewhat arbitrarily**.
2. Choose the number of basis elements N – **somewhat arbitrarily**.
3. Decompose the observed sample x_i in the chosen basis, i.e. evaluate $\tilde{x}_{i,k} = \langle x_i, f_k \rangle$, take multivariate vectors $\tilde{\mathbf{x}}_i \in \mathbb{R}^N$ do some kind analysis on these vectors. For example find principal components, etc.

From sample of functions to sample of vectors and back

The most common approach in basic steps

1. Pick-up a basis of functions $\{f_k\}_{k=0}^{\infty}$ that spans the space of all possibly observed functions – **somewhat arbitrarily**.
2. Choose the number of basis elements N – **somewhat arbitrarily**.
3. Decompose the observed sample x_i in the chosen basis, i.e. evaluate $\tilde{x}_{i,k} = \langle x_i, f_k \rangle$, take multivariate vectors $\tilde{\mathbf{x}}_i \in \mathbb{R}^N$ do some kind analysis on these vectors. For example find principal components, etc.
4. 'Translate' the obtained conclusions on the functional models by returning to functions through the originally selected basis $\{f_k\}_{k=0}^{\infty}$. For example approximating the eigenfunctions $\{e_k\}_{k=0}^{\infty}$ and corresponding eigenvalues from the Karhune-Loéve decomposition.

From sample of functions to sample of vectors and back

The most common approach in basic steps

1. Pick-up a basis of functions $\{f_k\}_{k=0}^{\infty}$ that spans the space of all possibly observed functions – **somewhat arbitrarily**.
2. Choose the number of basis elements N – **somewhat arbitrarily**.
3. Decompose the observed sample x_i in the chosen basis, i.e. evaluate $\tilde{x}_{i,k} = \langle x_i, f_k \rangle$, take multivariate vectors $\tilde{\mathbf{x}}_i \in \mathbb{R}^N$ do some kind analysis on these vectors. For example find principal components, etc.
4. 'Translate' the obtained conclusions on the functional models by returning to functions through the originally selected basis $\{f_k\}_{k=0}^{\infty}$. For example approximating the eigenfunctions $\{e_k\}_{k=0}^{\infty}$ and corresponding eigenvalues from the Karhune-Loéve decomposition.

Functional or multivariate?

Practically by a **'non-statistical' decision** of the initial basis selection the functional problem is reduced in the statistical sense to a multivariate analysis problem. The rest are mathematical 'ornaments' to make things looking 'functional'.

Criticality of the initial choice of the basis

The following example while purely academic illustrates the point:

Criticality of the initial choice of the basis

The following example while purely academic illustrates the point:

Criticality of the initial choice of the basis

The following example while purely academic illustrates the point:

- Suppose that the **basis in K-L** $\{e_k\}_{k=0}^{\infty}$ is orthogonal to $\{f_k\}_{k=0}^{N}$. (For example, $\{e_k\}_{k=0}^{\infty} = \{f_k\}_{k=N+K}^{\infty}$, for some $K > 0$.)

Criticality of the initial choice of the basis

The following example while purely academic illustrates the point:

- Suppose that the **basis in K-L** $\{e_k\}_{k=0}^{\infty}$ is orthogonal to $\{f_k\}_{k=0}^{N}$. (For example, $\{e_k\}_{k=0}^{\infty} = \{f_k\}_{k=N+K}^{\infty}$, for some $K > 0$.)
- Then all $\tilde{\mathbf{x}}_i \in \mathbb{R}^N$ become equal to zero – **we end up with no data**.

Criticality of the initial choice of the basis

The following example while purely academic illustrates the point:

- Suppose that the **basis in K-L** $\{e_k\}_{k=0}^{\infty}$ is orthogonal to $\{f_k\}_{k=0}^{N}$. (For example, $\{e_k\}_{k=0}^{\infty} = \{f_k\}_{k=N+K}^{\infty}$, for some $K > 0$.)
- Then all $\tilde{\mathbf{x}}_i \in \mathbb{R}^N$ become equal to zero – **we end up with no data**.
- Even if $\{e_k\}_{k=0}^{\infty}$ is not quite orthogonal to $\{f_k\}_{k=0}^{N}$, for example,

$$\{e_k\}_{k=0}^{\infty} = \{f_k\}_{k=N-K}^{\infty}$$

but the **most influential eigenvalues** λ_k's correspond to e_k's that are orthogonal to $\{f_k\}_{k=0}^{N}$ Then the transformation to vectors **loses critical information** about the original data.

Criticality of the initial choice of the basis

Criticality of the initial choice of the basis

- Sometimes, a remedy is proposed by doing **cross-validation data mining** on the model for the **'optimal' selection** of N.

Criticality of the initial choice of the basis

- Sometimes, a remedy is proposed by doing **cross-validation data mining** on the model for the **'optimal' selection** of N.
- While it ties statistically (to some degree) the choice of the initial basis with the data, this generally may lead to vectors with large number of coordinates that are close to zero (irrelevant).

Criticality of the initial choice of the basis

- Sometimes, a remedy is proposed by doing **cross-validation data mining** on the model for the **'optimal' selection** of N.
- While it ties statistically (to some degree) the choice of the initial basis with the data, this generally may lead to vectors with large number of coordinates that are close to zero (irrelevant).
- For example, $\{e_k\}_{k=0}^{\infty} = \{f_k\}_{k=N+K}^{\infty}$, for some very large $K > 0$, will result with the initial large number of basis elements f_k, $k = 1, \ldots, N + K - 1$ completely useless.

Criticality of the initial choice of the basis

- Sometimes, a remedy is proposed by doing **cross-validation data mining** on the model for the **'optimal' selection** of N.
- While it ties statistically (to some degree) the choice of the initial basis with the data, this generally may lead to vectors with large number of coordinates that are close to zero (irrelevant).
- For example, $\{e_k\}_{k=0}^{\infty} = \{f_k\}_{k=N+K}^{\infty}$, for some very large $K > 0$, will result with the initial large number of basis elements f_k, $k = 1, \ldots, N + K - 1$ completely useless.
- Somewhat better, it would be to subsample from $\{f_k\}_{k=0}^{\infty}$ and decide for the choice that produces the most efficient model (through cross-validation?).

Criticality of the initial choice of the basis

- Sometimes, a remedy is proposed by doing **cross-validation data mining** on the model for the **'optimal' selection** of N.
- While it ties statistically (to some degree) the choice of the initial basis with the data, this generally may lead to vectors with large number of coordinates that are close to zero (irrelevant).
- For example, $\{e_k\}_{k=0}^{\infty} = \{f_k\}_{k=N+K}^{\infty}$, for some very large $K > 0$, will result with the initial large number of basis elements f_k, $k = 1, \ldots, N + K - 1$ completely useless.
- Somewhat better, it would be to subsample from $\{f_k\}_{k=0}^{\infty}$ and decide for the choice that produces the most efficient model (through cross-validation?).
- This would tie the base selection to the data, however it is conceptually unclear how to sample from an infinite set and even if defined it can be computationally prohibitive to do any cross-validation based 'optimal' selection.

Criticality of the initial choice of the basis

- Sometimes, a remedy is proposed by doing **cross-validation data mining** on the model for the **'optimal' selection** of N.
- While it ties statistically (to some degree) the choice of the initial basis with the data, this generally may lead to vectors with large number of coordinates that are close to zero (irrelevant).
- For example, $\{e_k\}_{k=0}^{\infty} = \{f_k\}_{k=N+K}^{\infty}$, for some very large $K > 0$, will result with the initial large number of basis elements f_k, $k = 1, \ldots, N + K - 1$ completely useless.
- Somewhat better, it would be to subsample from $\{f_k\}_{k=0}^{\infty}$ and decide for the choice that produces the most efficient model (through cross-validation?).
- This would tie the base selection to the data, however it is conceptually unclear how to sample from an infinite set and even if defined it can be computationally prohibitive to do any cross-validation based 'optimal' selection.
- Most importantly the important eigenvectors and eigenvalues can be spread over a large range of basis elements if the latter are chosen arbitrarily.

Outline

Popular bases orthogonal (O) or not (NO)

- the Fourier basis (O):

$$\{\sqrt{2}\sin(2\pi nx);\ n \in \mathbb{N}\} \cup \{\sqrt{2}\cos(2\pi nx);\ n \in \mathbb{N}\} \cup \{1\}.$$

- the Legendre polynomials for $[-1, 1]$ (O):

$$L_n(x) = \sum_{k=0}^{n} (-1)^k \binom{n}{k}^2 \left(\frac{1+x}{2}\right)^{n-k} \left(\frac{1-x}{2}\right)^k.$$

- the Hermite polynomials on $(-\infty, \infty)$, with Gaussian measure (O):

$$P_n(x) = n! \sum_{k=0}^{[n/2]} \frac{(-1)^k}{2^{2k}} \frac{x^{n-2k}}{(n-2k)!k!}.$$

- Wavelets basis (O),
- 'Haar' basis functions (O) for a knot-sequence τ:

$$B_{i,1}(x) = \begin{cases} 1 & ;\tau_i \leq x < \tau_{i+1}, \\ 0; & otherwise, \end{cases}$$

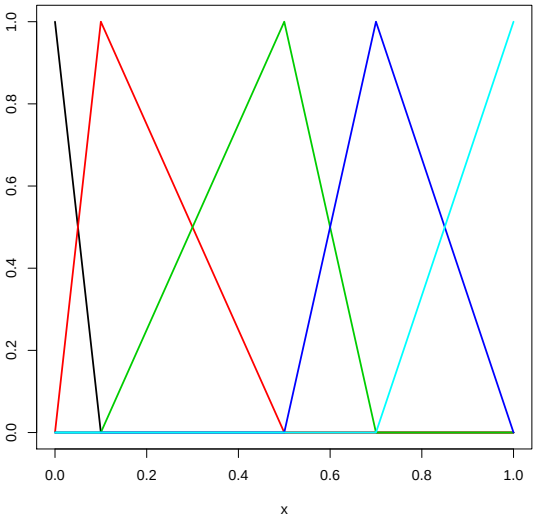
for $i = 1, ..., K + 1$.

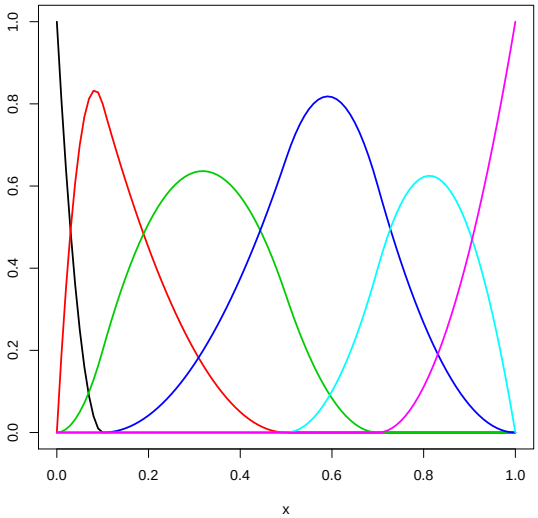
B-spline base

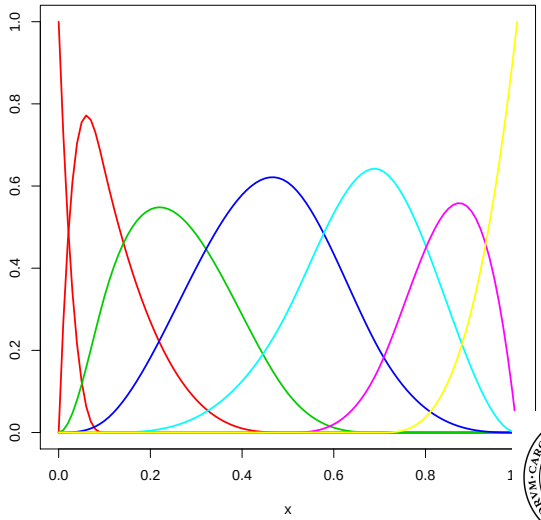
- B-splines (NO):

$$B_{i,m}(x) = \frac{x - \tau_i}{\tau_{i+m-1} - \tau_i} B_{i,m-1}(x) + \frac{\tau_{i+m} - x}{\tau_{i+m} - \tau_{i+1}} B_{i+1,m-1}(x),$$

for $i = 1, ..., K + 2M - m$. Thus with $M = 4$, $B_{i,4}$, $i = 1, ..., K + 4$ are the $K + 4$ cubic B-spline basis functions for the knot sequence τ. This recursion can be continued and will generate the B-spline basis for any order spline.







Which the base to choose?

- Typically, for a given set functional data one of the above bases is chosen based on the intuition on potential suitability for the data at hand.

Which the base to choose?

- Typically, for a given set functional data one of the above bases is chosen based on the intuition on potential suitability for the data at hand.
- The number of elements of the basis to be used is often tested in more or less ad hoc manner.

Which the base to choose?

- Typically, for a given set functional data one of the above bases is chosen based on the intuition on potential suitability for the data at hand.
- The number of elements of the basis to be used is often tested in more or less ad hoc manner.
- If we want to make our choice more data driven which of (infinitely many) bases to choose given the data.

Which the base to choose?

- Typically, for a given set functional data one of the above bases is chosen based on the intuition on potential suitability for the data at hand.
- The number of elements of the basis to be used is often tested in more or less ad hoc manner.
- If we want to make our choice more data driven which of (infinitely many) bases to choose given the data.
- We want explore efficiently the functional data for a way to represent them properly.

Which the base to choose?

- Typically, for a given set functional data one of the above bases is chosen based on the intuition on potential suitability for the data at hand.
- The number of elements of the basis to be used is often tested in more or less ad hoc manner.
- If we want to make our choice more data driven which of (infinitely many) bases to choose given the data.
- We want explore efficiently the functional data for a way to represent them properly.

Splines with properly selected knots seem to be very suitable for the problem.

Outline

Recap

- On common feature of different order of splines is that their are built on knots.
- Additional feature of the zero-order splines (Haar ‘splines’) is orthogonality and ‘locality’ with respect to the addition of the knots.
- For smoothed data one would wish for smooth basis.
- The B-splines are local but not orthogonal.
- They can be orthonormalized.
- There is a need for **good orthonormalization**.
- Strangely enough the problem did not receive much attention in the past.

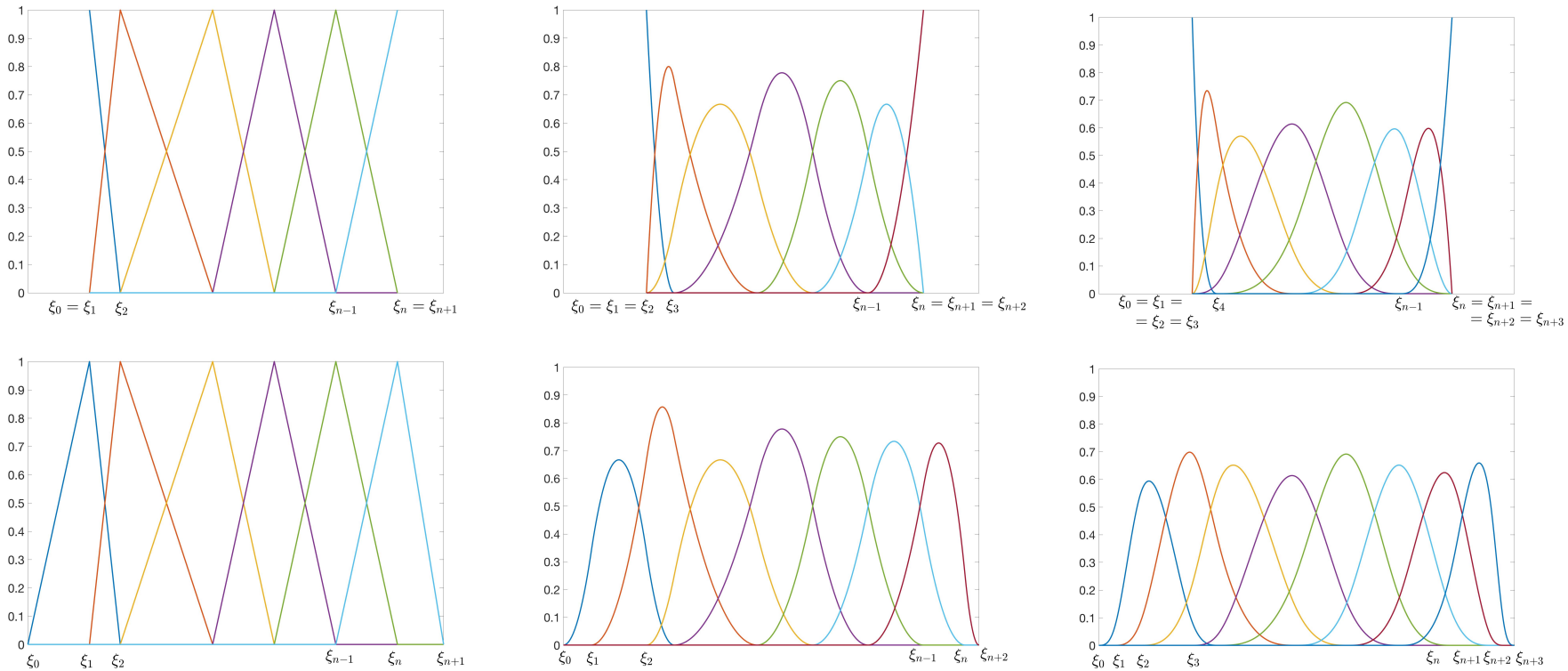
Primer on B-splines - knots and boundaries

- The set of knots is always ordered and represented as a vector ξ of ordered values.
- Typically, there are two alternative and essentially equivalent approaches to discussing splines at the end points of their range:
  1. **No boundary conditions** are imposed and ξ needs some initial superfluous knots located at zero in order to handle efficiently recurrent formulas.
  2. The second approach does not introduce any superficial knots. It rather imposes on **a spline and all its derivatives** of the order smaller than the order of splines the value of zero **at both the endpoints of the domain**.
- The two approaches are in a certain sense equivalent, for example, the first follows from the second by passing to the limit

$$\xi_0, \ldots, \xi_{K-1} \to \xi_K, \quad \xi_{K+n+1}, \ldots, \xi_{2K+n} \to \xi_{2K+n+1}$$

LUND UNIVERSITY

Primer on Splines – comparison of the two approaches



The two approaches to the endpoints for the first, second , and third order splines in the case of *B*-splines: superfluous knots at endpoints *(left)*, imposed zeros as the initial conditions at the endpoints *(right)*.

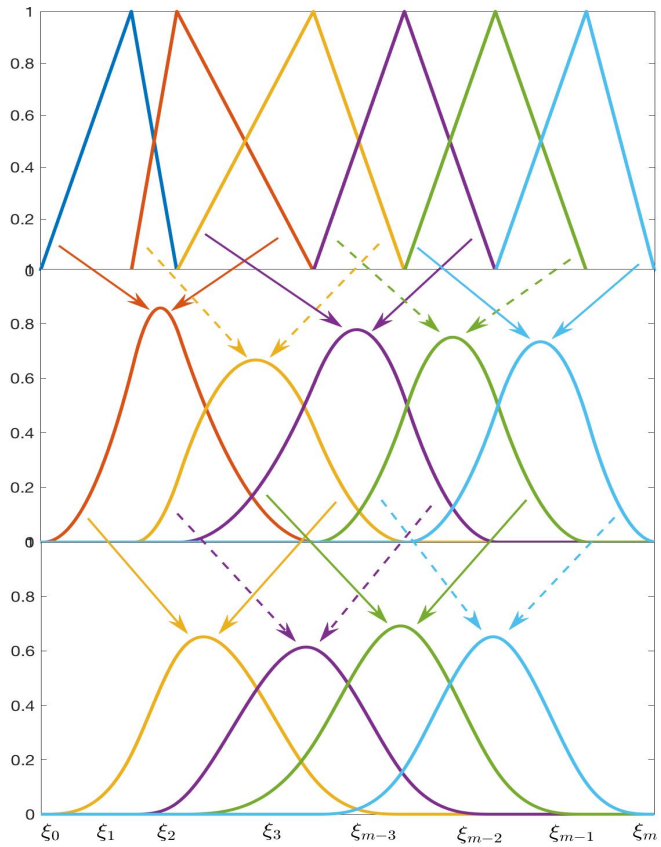
Primer on Splines – recurrent definition of B-splines

The following recursion relation leads to the definition of the splines of arbitrary order $k < m$. Suppose now that we have defined $B^{\xi}_{k-1,l}$, $l = 1, \ldots, m-k+1$. The B-splines of order k are defined, for $l = 1, \ldots, m-k$, by

$$B^{\xi}_{k,l}(x) = \frac{x - \xi_l}{\xi_{l+1} - \xi_l} B^{\xi}_{k-1,l}(x) + \frac{\xi_{l+2} - x}{\xi_{l+2} - \xi_{l+1}} B^{\xi}_{k-1,l+1}(x).$$

Primer on Splines – recurrence in thousand words

Primer on Splines – recurrence in thousand words



The recursion in definition of the *B*-splines, the first order splines *(top)*, the second order spline *(middle)*, and the third order spline *(bottom)*.

Orthogonalizing B - splines

Orthogonalizing B - splines

- B-splines seems to be nice because of their local supports.
- However despite their local supports they are not trully local –

Orthogonalizing B - splines

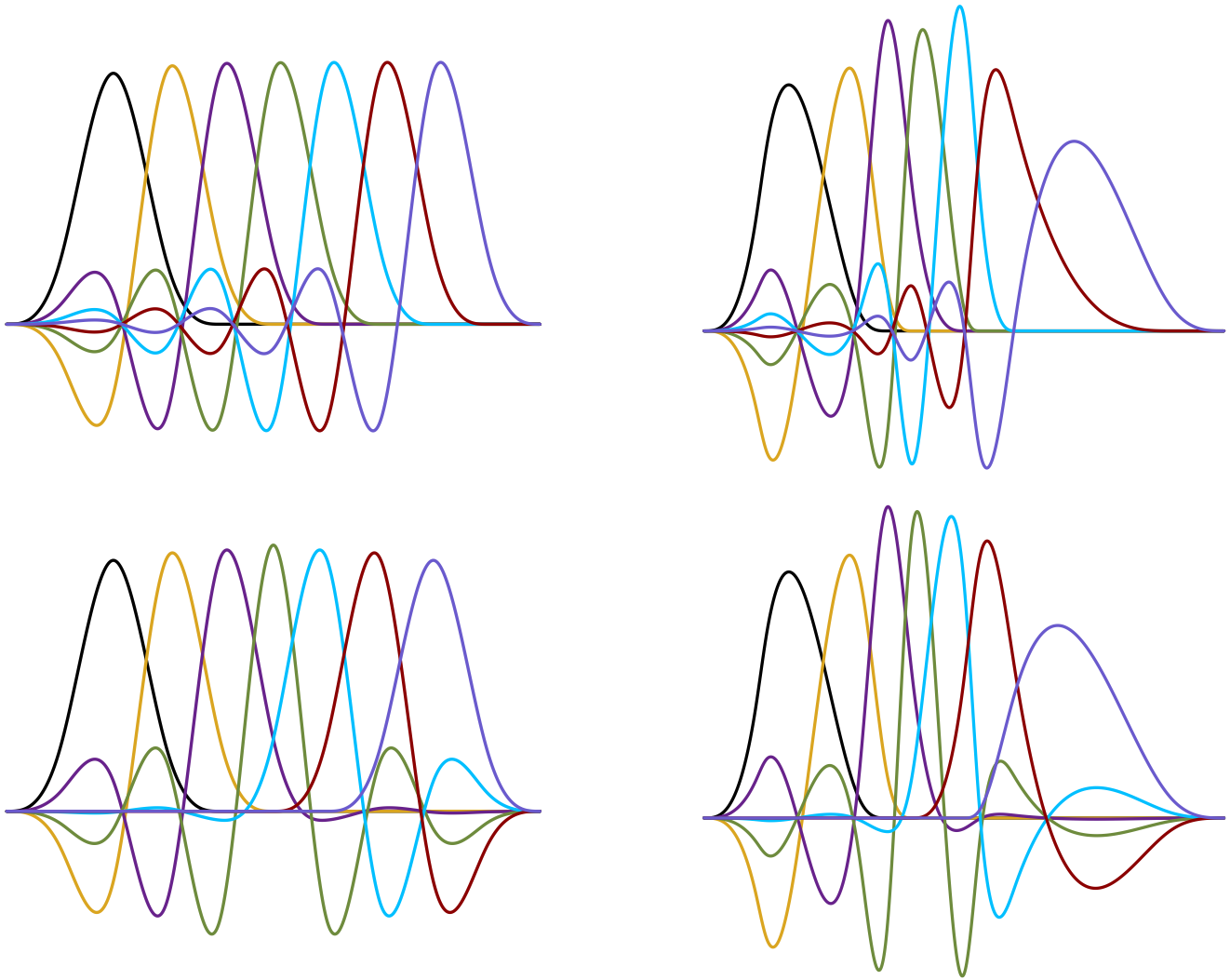
- B-splines seems to be nice because of their local supports.
- However despite their local supports they are not trully local – **try add one knot and perform Gram-Schmidt orthogonalization: 'everything' changes**.
- Having explicit orthonormalization of B splines is of interest.
- How to do it? There are many ways of orthogonalizing vectors.

Three orthogonalizations

Three orthogonalizations

- **Gramm-Schmidt** - everyone knows it.
- Löwdin Symmetric Orthogonalization - produces two-sided symmetry (with respect to the center), it has smaller total support for the elements than GS.
- Structured orthogonalization – **splinets**.

One- and two-sided orthogonalization



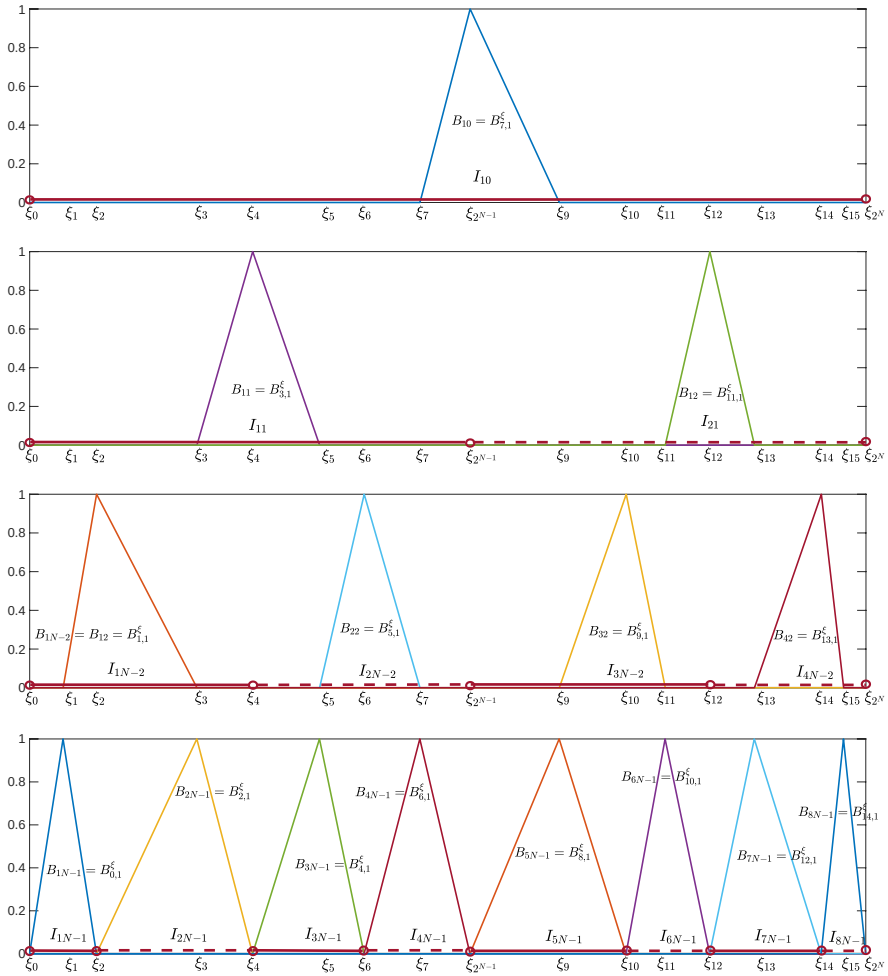
The third order *O*-splines. *Top*: one-sided left-to-right, ten knots; *Bottom*: two-sided, eleven knots; in the left column equally spaced knots, in the right column irregularly spaced knots.

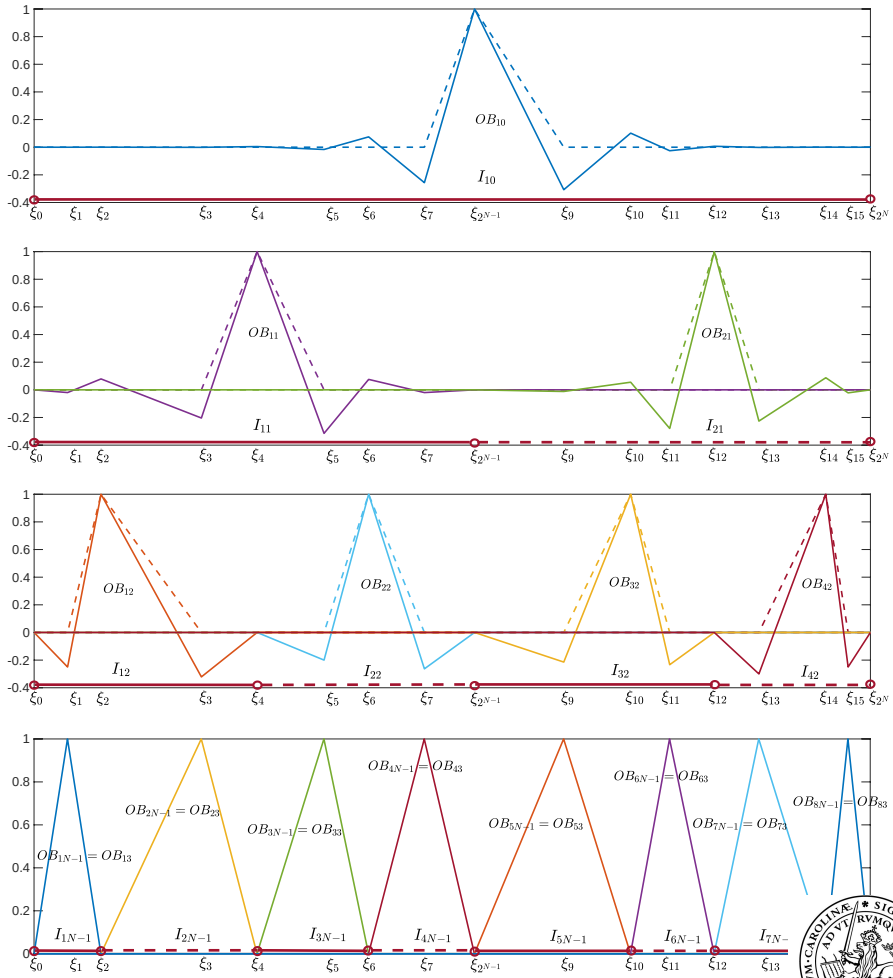
The graphs obtained using `splinets` package.

Splinets – highlights

- The term *splinets* is proposed to a unique ON family both preserving the locality in the supports and at the same time inducing the orthonormality.
- **The total size of the support** is at the order of $\log n$ while in the previous approaches it is at the order n.
- It uses the fact that for the B splines there are elements that have mutually disjoint supports and thus naturally orthogonal. We create disjoint support groups of subsequent splines. The number of element in a group is equal to the order of the spline.
- We perform two sided orthogonalization within the group. They are the initial vectors in our orthogonalization. We call them zero level O-spline in the splinet.
- Between subsequent zero order elements there are elements (their number depending on the order of the considered splines) that have support that is contained in the union of the support of these two zero order elements. We orthogonalize them with respect to the zero order O-splines.

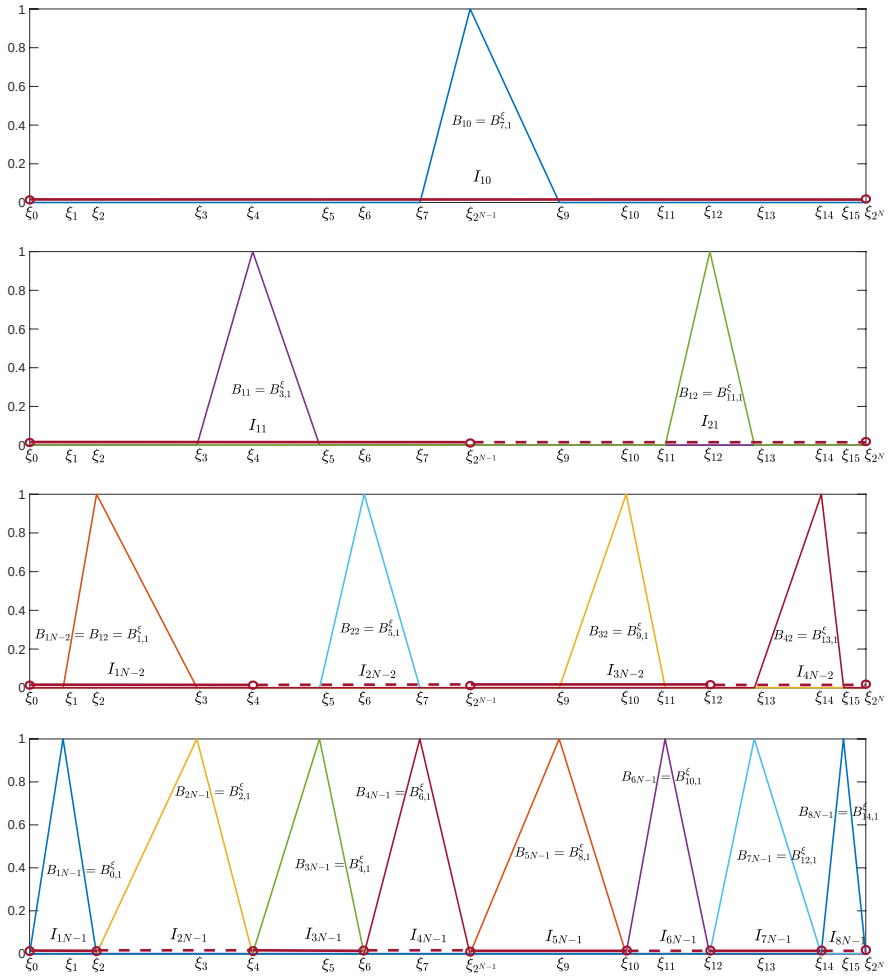
The first order dyadic case

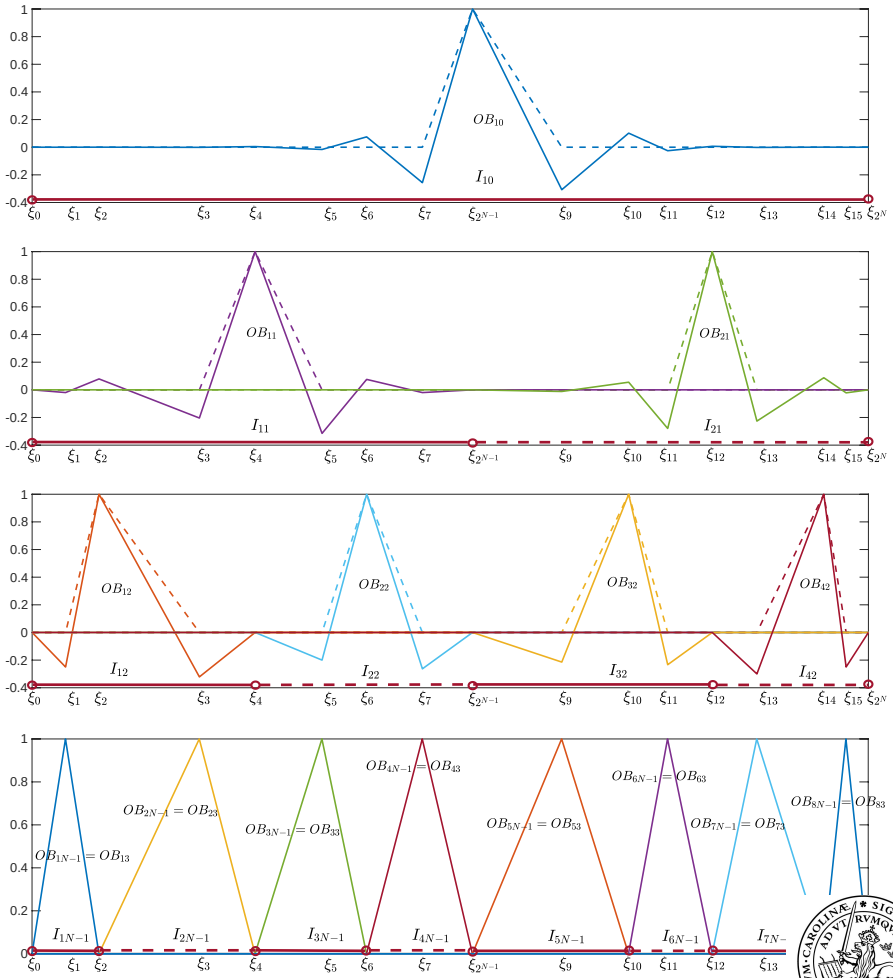




LUND UNIVERSITY

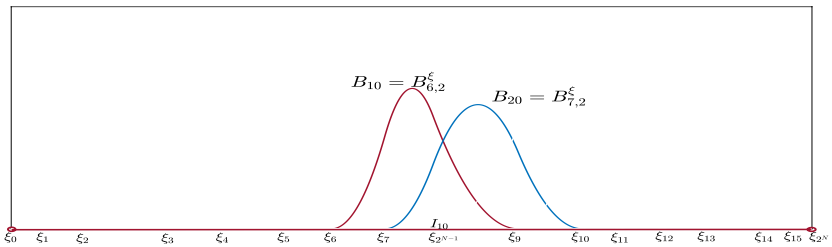
The first order dyadic case

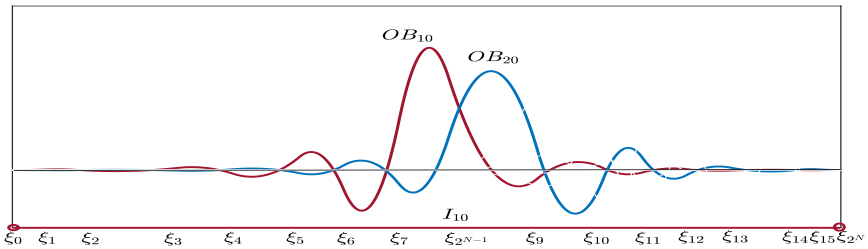


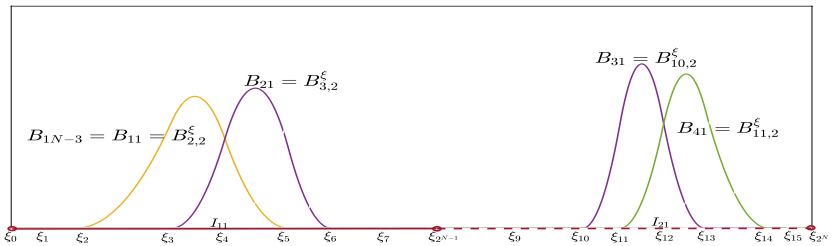


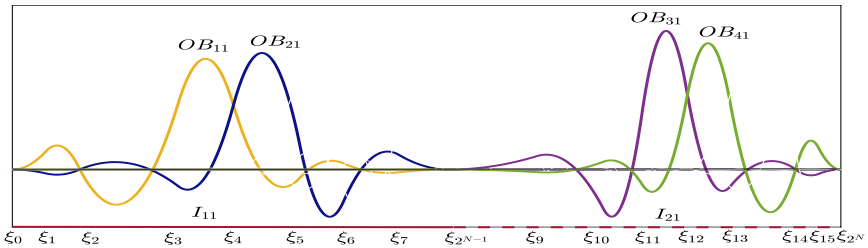
LUND UNIVERSITY

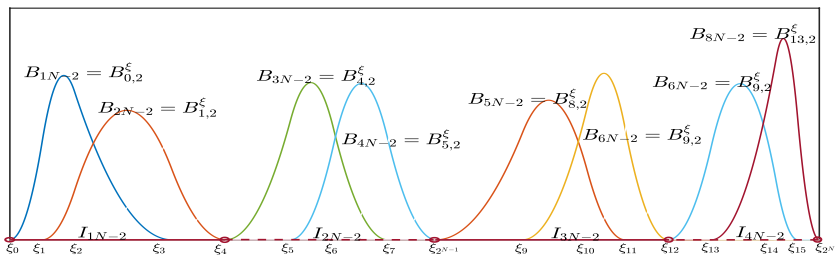
The second order dyadic case

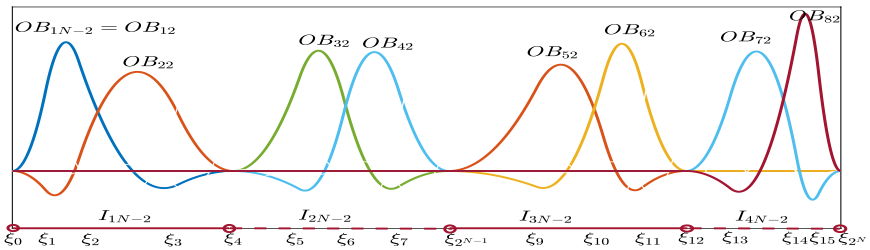












`splinets` – dedicated R-package

Features:

- General non-dyadic case covered.
- Spline object

$$\mathcal{S}(S) = \{k, \boldsymbol{\xi}, j, m, \mathbf{s_0}, \mathbf{s_1}, \ldots, \mathbf{s_k}\},$$

  where $[\xi_j, \xi_{j+m}]$ is the support of S and $\mathbf{s_0}, \mathbf{s_1}, \ldots, \mathbf{s_k}$ are $m+1$ dimensional vectors of values of the i-derivative of S at the knots given in $\left(\xi_j, \ldots, \xi_{j+m}\right)$, $i = 0, \ldots, k$.
- Algebra of spline objects.
- Computational efficiency and stability of the representation that follows from Taylor's representation, i.e. if $x \in [\xi_{j+r}, \xi_{j+r+1}]$, then

$$S(x) = \sum_{l=0}^{k} \frac{(x-\xi_{j+r})^l}{l!} s_{rl} = \sum_{l=0}^{k} \frac{(x-\xi_{j+r+1})^l}{l!} s_{r+1l}.$$

Outline

Setting the problem

- A method of adding knots based on the mean square error effectiveness of approximating the functional data $y_i = y_i(t)$, $t \in [0, 1]$, $i = 1, 2, \ldots n$.

Setting the problem

- A method of adding knots based on the mean square error effectiveness of approximating the functional data $y_i = y_i(t)$, $t \in [0, 1]$, $i = 1, 2, \ldots n$.
- The method is iterative and resembles **regression tree building** was inspired Regression Trees methods.

Setting the problem

- A method of adding knots based on the mean square error effectiveness of approximating the functional data $y_i = y_i(t)$, $t \in [0,1]$, $i = 1, 2, \ldots n$.
- The method is iterative and resembles **regression tree building** was inspired Regression Trees methods.
- For any functional data set $\mathcal{Y} = \{y_i \in L^2, i = 1, \ldots n\}$, the set of best least square **constant** predictors is a set of functions

$$y_i^{(0)} = \langle y_i, \mathbf{1} \rangle \mathbf{1}.$$

- The constant functions over $[0,1]$ can be viewed as 0-order splines with no internal knot points, and its one dimensional basis is given by the constant function $\mathbf{1}$.
- Thus we can set the initial set of knots to an empty set, i.e. $\mathcal{K}^{(0)} = \emptyset$, the initial basis $\mathcal{B}^{(0)} = \{\mathbf{1}\}$, and the projection to the space spanned by $\mathcal{B}^{(0)}$ is given by $\mathbf{P}^{(0)} y = \langle y, \mathbf{1} \rangle \mathbf{1}$.
- The average mean square error (AMSE) per function of approximations of y_i's by the optimal constant functions is given by

$$AMSE(\mathcal{Y}, \mathcal{B}^{(0)}) = \frac{1}{n} \sum_{i=1}^{n} \| y_i - \mathbf{P}^{(0)} y_i \|^2 = \frac{1}{n} \sum_{i=1}^{n} \| y_i - \langle y_i, \mathbf{1} \rangle \mathbf{1} \|^2.$$

LUND UNIVERSITY

Placing knots - do you remember regression trees?

- At the first step, $s = 1$, find a knot $\xi \in [0, 1]$ such that the optimal approximation of y by a linear combination of the 0-order splines with the set of knots $\mathcal{K}^{(1)} = \mathcal{K}^{(0)} \cup \{\xi\}$ yields the smallest AMSE, i.e. denote by $\mathcal{B}^{(1)}(\xi)$ the orthonormal base of piecewise constant functions over the intervals given by the knots in $\mathcal{K}^{(1)}(\xi)$. The new knot ξ_{new} is chosen as

$$\xi_{new} = \underset{\xi \in (0,1]}{\operatorname{argmin}} \, AMSE(\mathcal{Y}, \mathcal{B}^{(1)}(\xi)).$$

- Then the new, enlarged by one function, basis $\mathcal{B}^{(1)} = \mathcal{B}^{(1)}(\xi_{new})$ is uniquely defined the new (with one knot added at ξ_{new}) set of knots $\mathcal{K}^{(1)} = \mathcal{K}^{(1)}(\xi_{new})$.
- In the recurrent process, at the step s, we start with a sequence of knots $\mathcal{K}^{(s-1)}$ and search for a new knot ξ_{new} by considering $\mathcal{K}^{(s)}(\xi) = \mathcal{K}^{(s-1)} \cup \{\xi\}$ and the corresponding $\mathcal{B}^{(s)}(\xi)$ the orthonormal base of piecewise constant functions.

$$\xi_{new} = \underset{\xi \in (0,1]}{\operatorname{argmin}} \, AMSE(\mathcal{Y}, \mathcal{B}^{(s)}(\xi)).$$

LUND UNIVERSITY

Features of the knot selection process

- The 0-order splines constitute *O*-**basis**.

Features of the knot selection process

- The 0-order splines constitute *O*-**basis**.
- This basis is **local**, i.e. adding a new knot requires only a change in direct vicinity of the knot.

Features of the knot selection process

- The 0-order splines constitute *O*-**basis**.
- This basis is **local**, i.e. adding a new knot requires only a change in direct vicinity of the knot.
- At each new step most of the values from the previous steps can be utilized.
- Computations are fast and simple.

Features of the knot selection process

- The 0-order splines constitute *O*-**basis**.
- This basis is **local**, i.e. adding a new knot requires only a change in direct vicinity of the knot.
- At each new step most of the values from the previous steps can be utilized.
- Computations are fast and simple.
- The importance of particular knots can be measured by reduction of AMSE.
- Different measures of the stop criterion based on AMSE can be used.
- Modification of the method based on random selection of the knots, or bootstrapping from the functional data, or from knots using their importance can be utilized in the spirit of machine learning methods.
- The distributional properties of knots can be used as feature used for **aligning the functions – registering problem**.
- **Smooth B-splines** can be build on the final optimal choice of the basis.
- Smooth O-splines(?) can be even better as the final choice of the basis.

The final choice of the basis - splinet

- Given a choice of the knots a smooth bases can be utilized to be used in further functional data analysis.
- The disadvantage of *B*-splines is that although local they are not orthogonal except for the 0-order case.
- Search for the orthogonal version of *B*-splines with possibly maximal ‘locality’.
- Splinets are ideal for the purpose.

Outline

Simple generator of non-linear functional data

- functional data

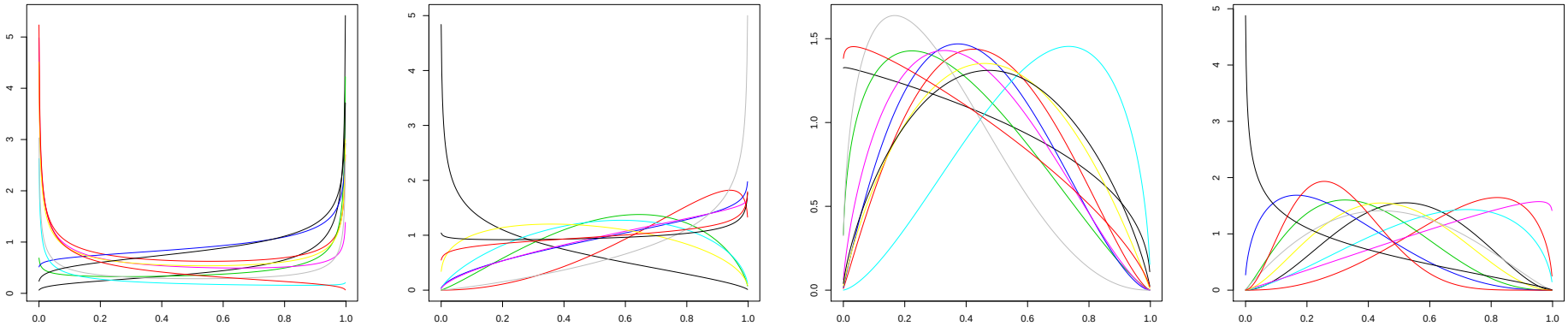
$$y_i = y_i(x),\ x \in (0,1),$$

we consider the **square roots of beta densities** that are proportional to $x^{\alpha-1}(1-x)^{\beta-1}$, $x \in (0,1)$, with positive α and β. Consequently all our functional data belong to the space of square integrable functions over $(0,1)$, are proportional to $x^{(\alpha-1)/2}(1-x)^{(\beta-1)/2}$ and they have the L_2 norm equal to one.

- For a given functional data size n, **pairs** (α_i, β_i), $i = 1, \ldots, n$ of parameters are **simulated independently** from two dimensional density proportional to

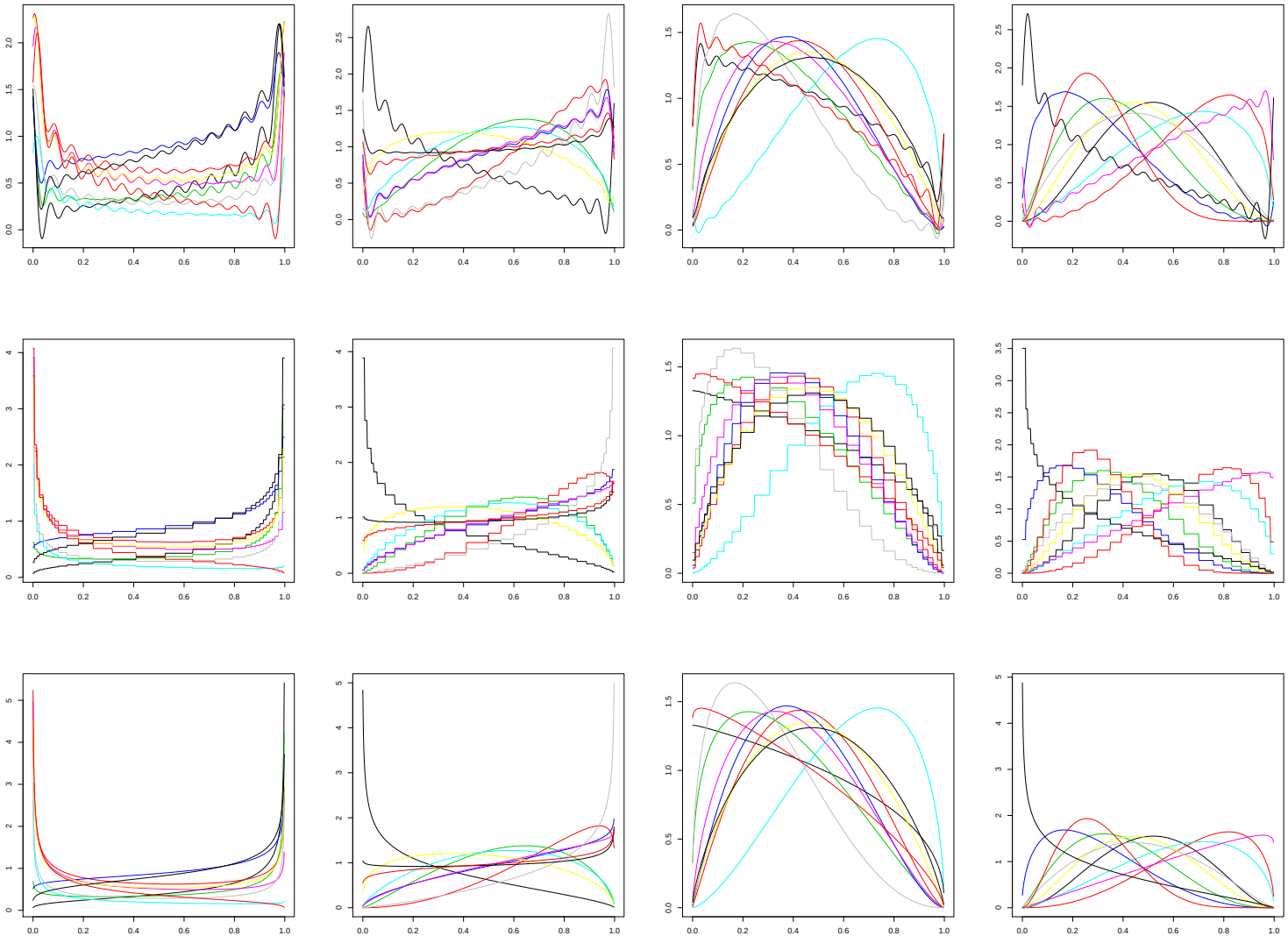
$$\alpha^{\tau_\alpha - 1}\beta^{\tau_\beta - 1}\exp\left(-2\left(\tau_\alpha\alpha + \tau_\beta\beta\right)\right).$$

Four data sets



Four samples of 10 functional data obtained from the square root beta density simulator for α and β sampled from gamma distribution with the shape τ equal from the right to the left to: $\tau_\alpha = \tau_\beta = 1$, $\tau_\alpha = \tau_\beta = 2$, $\tau_\alpha = \tau_\beta = 2.4$, $\tau_\alpha = \tau_\beta = 3$.

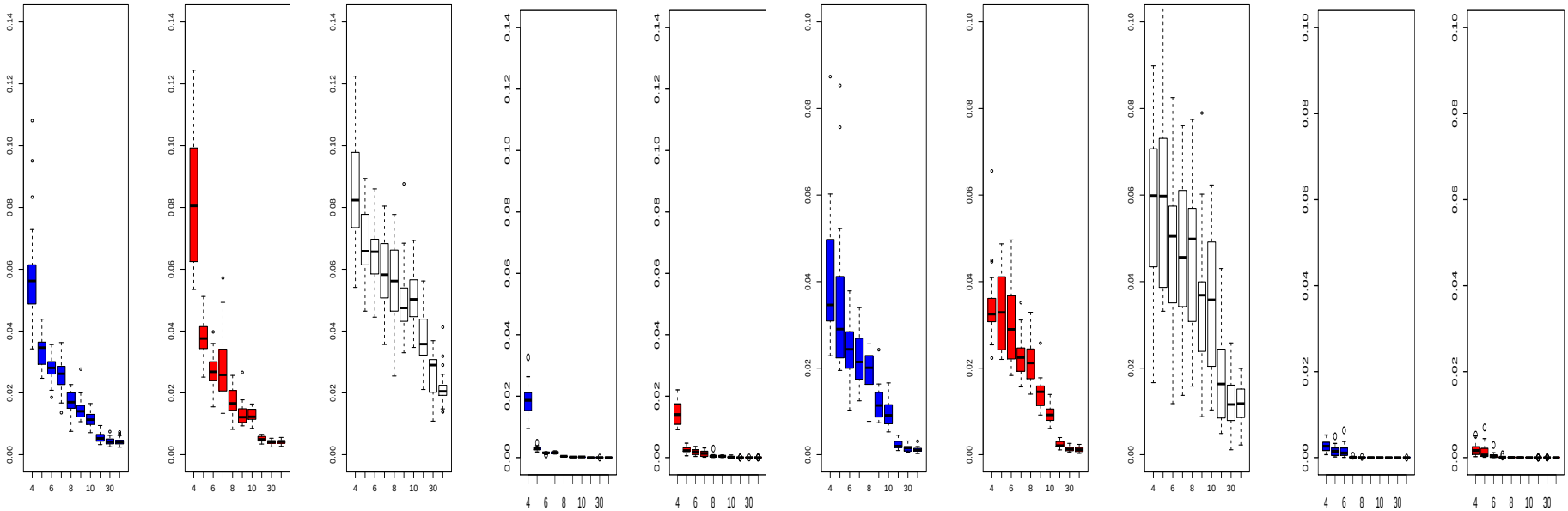
Fourier vs. piecewise constant vs. smooth splines



Monte Carlo study

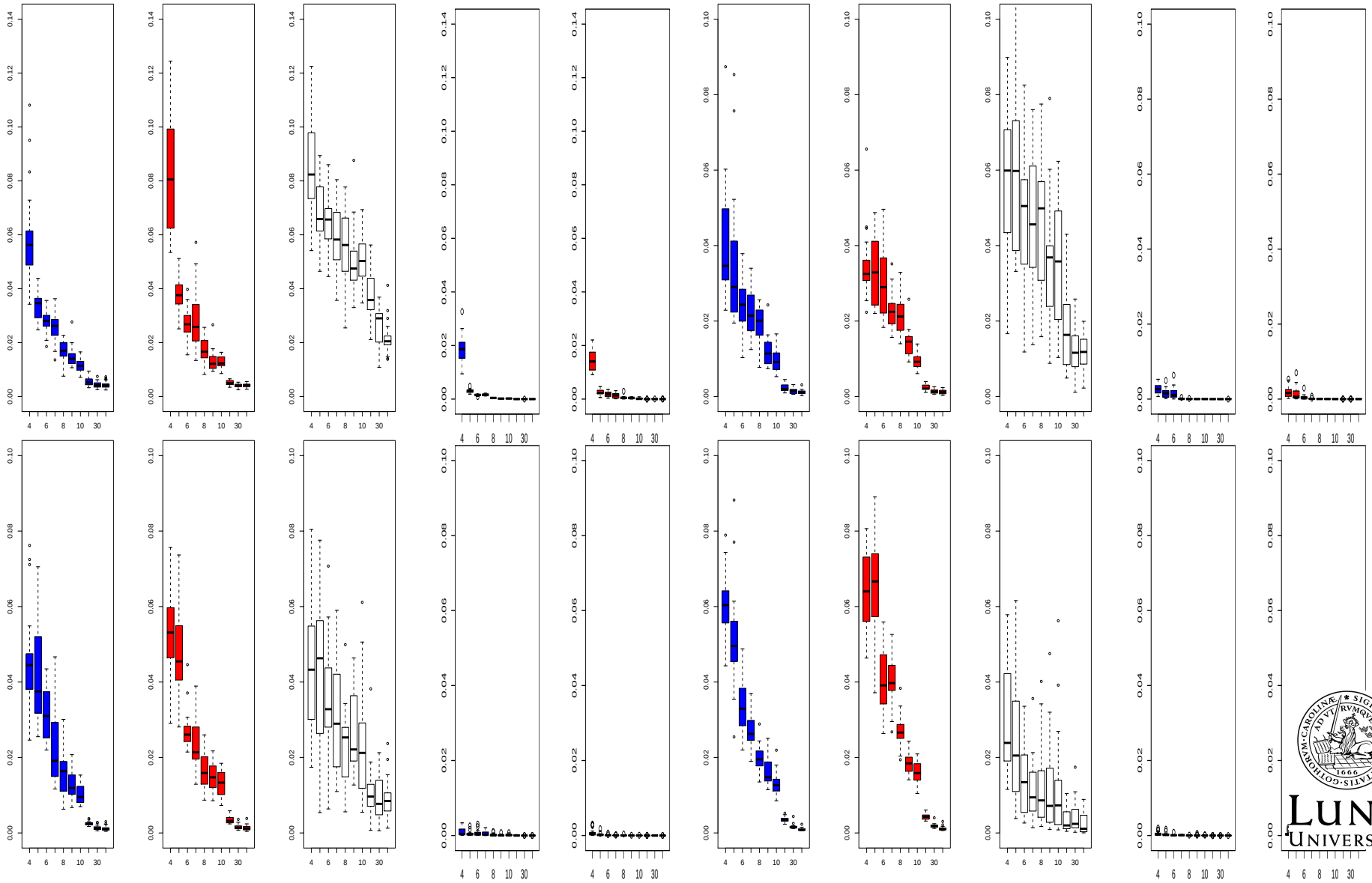
- A Monte Carlo study of the dependence of the **average mean square error (AMSE)** on the number basis elements.
- Monte Carlo samples of size 10 were drawn from the four specifications of the model based on beta densities.
- For each of this sample the Fourier approximations with the number of basis elements used increasing from 4 to 40 were evaluated and their AMSE over all 10 elements of the data evaluated.
- This procedure has been repeated independently 20 times resulting in 20 AMSE's for each size of the Fourier base used.
- The same was repeated for the piecewise constant data driven basis,
- Two cases: first, a new basis is selected anew for each MC sample, second, a piecewise constant basis is selected for the original sample and then used for every new MC sample.

Box-plots for MC study



The boxplots of AMSE's obtained from 20 Monte Carlo simulations as a function the base size, which ranges from 4 to 40. For each model 10 functional data were simulated and the orthonormal basis decomposition was run through these functional data with increasing number of basis elements. Four cases are presented in the groups of five graphs. In each of the four cases on the left (blue) and (red) corresponds to the piecewise constant data driven basis, in the left (blue) case a new basis is selected anew for each MC sample, in the middle (red) case a piecewise constant basis is selected for an original sample and then used for every new MC sample. The central picture (black and white) corresponds to the Fourier basis applied to the MC data. Finally on the right, the smoothed orthonormal splines are used. In the graphs we can see clearly benefits using data driven basis resulting in smaller variance and, except of few cases of small number of basis elements, in generally smaller AMSE for the piecewise constant bases. Clearly, the smoothed splines outperform other methods fo particular smooth data.

Box-plots for MC study



Conclusion

- The proposed data driven basis selection method searches through the functional data to find optimal allocation of the knots to construct initial orthonormal basis for subsequent FDA.

Conclusion

- The proposed data driven basis selection method searches through the functional data to find optimal allocation of the knots to construct initial orthonormal basis for subsequent FDA.
- Orthogonality of the zero order splines is exploited to allocate knots.

Conclusion

- The proposed data driven basis selection method searches through the functional data to find optimal allocation of the knots to construct initial orthonormal basis for subsequent FDA.
- Orthogonality of the zero order splines is exploited to allocate knots.
- Higher order orthogonal splinets can be used based on the obtained knots and assuring computational efficiency.

Conclusion

- The proposed data driven basis selection method searches through the functional data to find optimal allocation of the knots to construct initial orthonormal basis for subsequent FDA.
- Orthogonality of the zero order splines is exploited to allocate knots.
- Higher order orthogonal splinets can be used based on the obtained knots and assuring computational efficiency.
- Initial MC studies suggest that the approach more efficiently represents functional data.

Conclusion

- The proposed data driven basis selection method searches through the functional data to find optimal allocation of the knots to construct initial orthonormal basis for subsequent FDA.
- Orthogonality of the zero order splines is exploited to allocate knots.
- Higher order orthogonal splinets can be used based on the obtained knots and assuring computational efficiency.
- Initial MC studies suggest that the approach more efficiently represents functional data.
- Actual studies on FDA models and methods, for example, functional principle component analysis is needed.

Conclusion

- The proposed data driven basis selection method searches through the functional data to find optimal allocation of the knots to construct initial orthonormal basis for subsequent FDA.
- Orthogonality of the zero order splines is exploited to allocate knots.
- Higher order orthogonal splinets can be used based on the obtained knots and assuring computational efficiency.
- Initial MC studies suggest that the approach more efficiently represents functional data.
- Actual studies on FDA models and methods, for example, functional principle component analysis is needed.
- There is a potential to improve the method through machine learning style data mining through functional data (bootstrapping, boosting style).
- Multidimensional in argument approaches are possible through tensor spaces of splines.

Thank you!