M7777 Applied Functional Data Analysis
2. From Data to Functions – basis systems

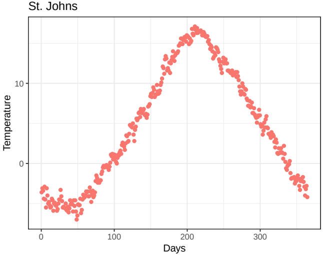
Jan Koláček (kolacek@math.muni.cz)

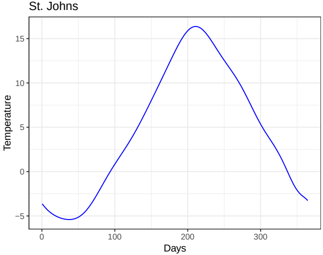
Dept. of Mathematics and Statistics, Faculty of Science, Masaryk University, Brno

From Data to Functions

How do we go from

data | to | functions?





Basis Expansions

We consider

$$y_i = x(t_i) + \varepsilon_i, \qquad \varepsilon_i \sim i.i.d$$

and

$$x(t_i) = \sum_{j=1}^{K} c_j \Phi_j(t_i).$$

Let us denote

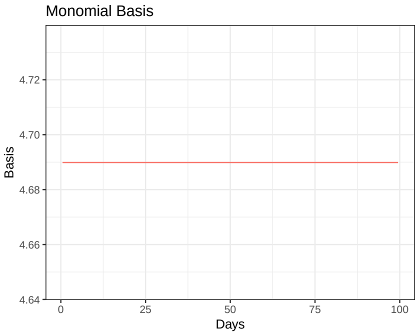
- $\mathbf{\Phi}^*(t) = (\Phi_1(t), \dots, \Phi_K(t))$ ... a **basis system** for $x(t)$
- $\mathbf{c} = (c_1, \dots, c_K)'$ ... **basis coefficients**

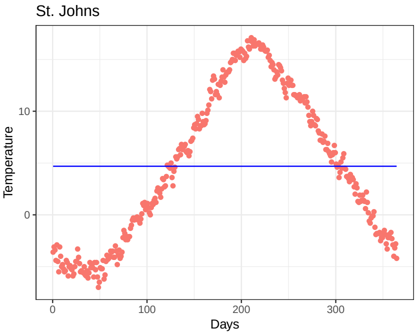
We write

$$x(t) = \mathbf{\Phi}^*(t)\mathbf{c}.$$

The Monomial Basis

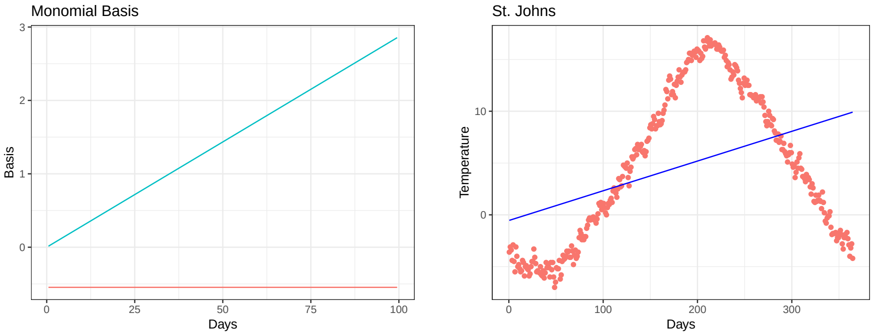
$$\mathbf{\Phi}^*(t) = (1)$$





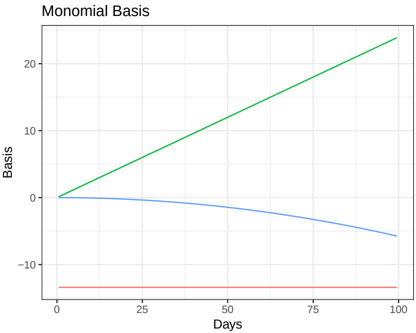
The Monomial Basis

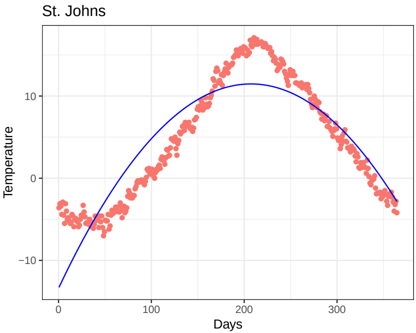
$$\mathbf{\Phi}^*(t) = (1, t)$$



The Monomial Basis

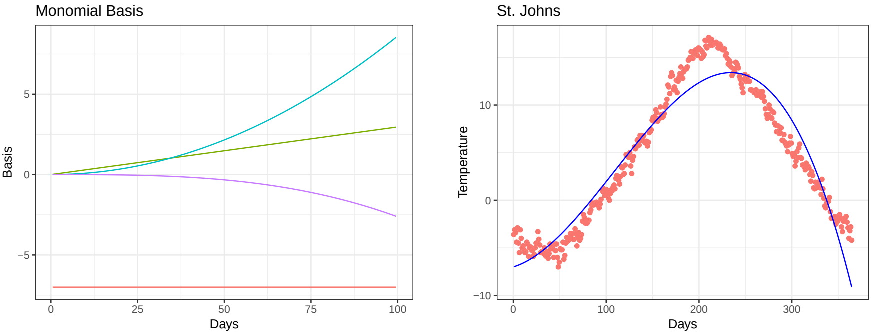
$$\mathbf{\Phi}^*(t) = (1, t, t^2)$$





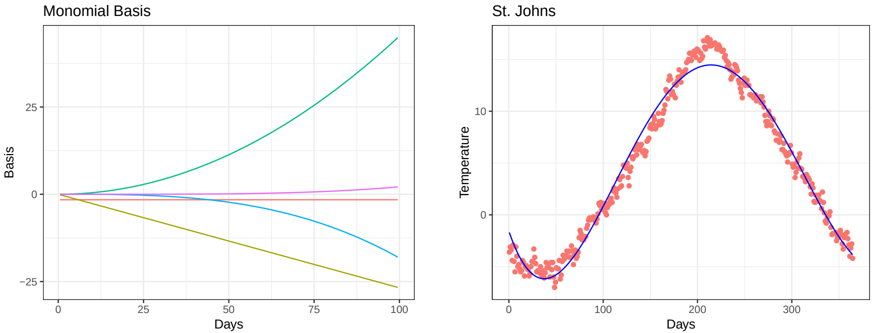
The Monomial Basis

$$\mathbf{\Phi}^*(t) = (1, t, t^2, t^3)$$



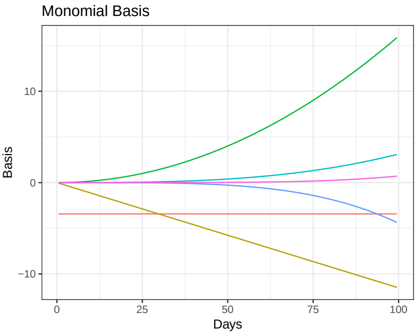
The Monomial Basis

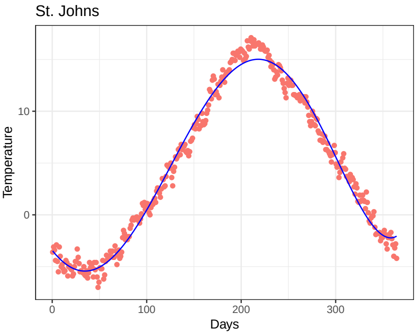
$$\mathbf{\Phi}^*(t) = (1, t, t^2, t^3, t^4)$$



The Monomial Basis

$$\mathbf{\Phi}^*(t) = (1, t, t^2, t^3, t^4, t^5)$$





The Monomial Basis

Summary

- Formula

$$x(t) = \sum_{j=0}^{K} c_j t^j$$

- numerically difficult for more than six terms
- problem with derivatives estimation

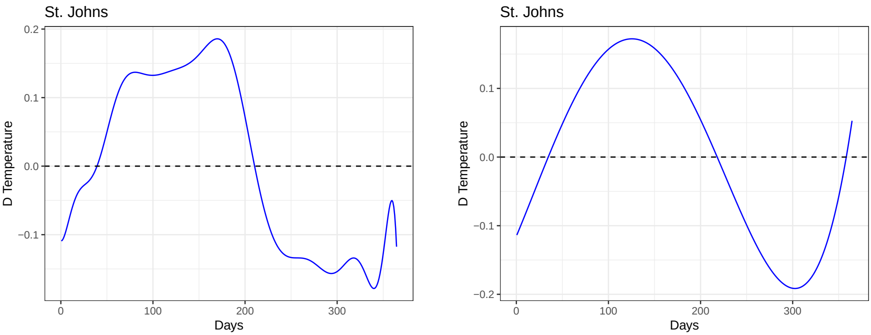
$$Dx(t) = \sum_{j=1}^{K} c_j j t^{j-1}$$

monomial derivatives get simpler, whereas the opposite happens in most real-world data (oversmoothing)

The Monomial Basis

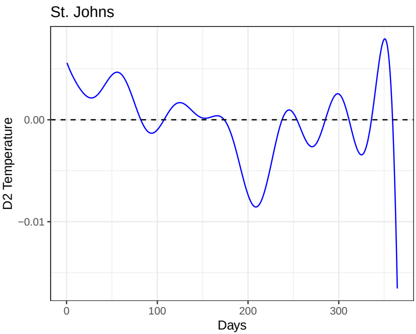
First derivative

Estimate

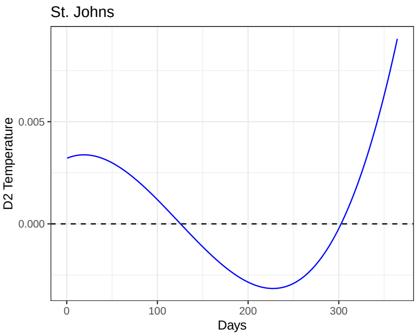


The Monomial Basis

Second derivative



Estimate



The Fourier Basis

The Fourier Basis

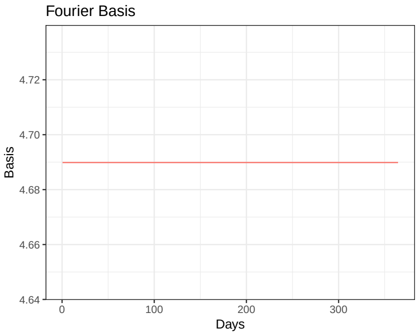
- Basis functions

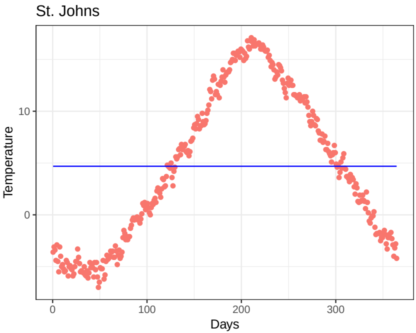
$$\mathbf{\Phi}^*(t) = (1, \sin(\omega t), \cos(\omega t), \sin(\omega 2t), \cos(\omega 2t), \ldots$$
$$\sin(\omega M t), \cos(\omega M t))$$

- ω ... defines the period of oscillation, i.e. $\omega = 2\pi/P$, P is the period
- $K = 2M + 1$ where M is the largest number of oscillations required in a period of length P

The Fourier Basis

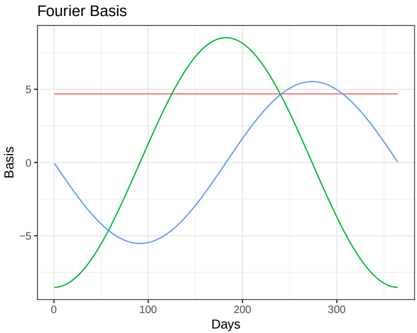
$$\mathbf{\Phi}^*(t) = (1)$$

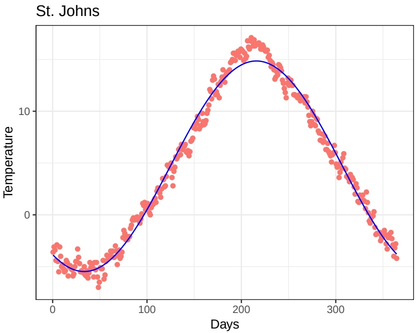




The Fourier Basis

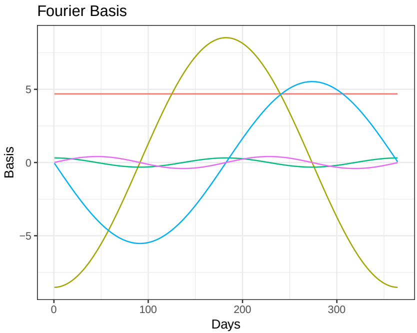
$$\mathbf{\Phi}^*(t) = (1, \sin(\omega t), \cos(\omega t))$$

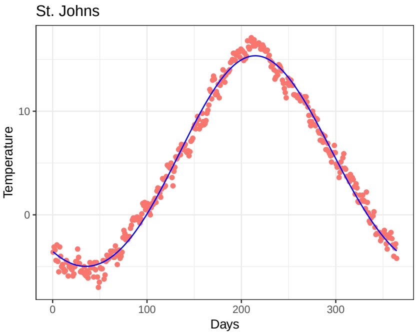




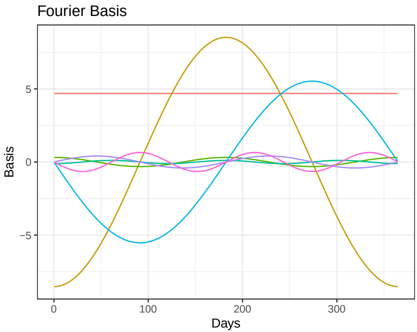
The Fourier Basis

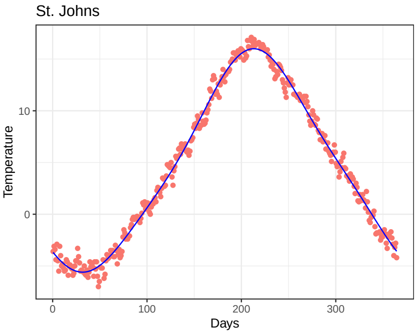
$$\mathbf{\Phi}^*(t) = (1, \sin(\omega t), \cos(\omega t), \sin(\omega 2t), \cos(\omega 2t))$$





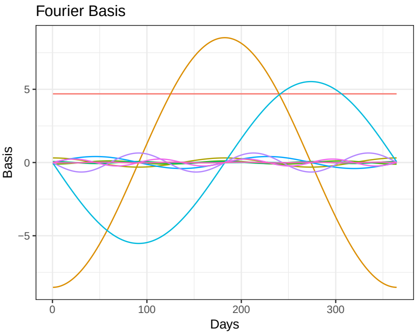
$$\mathbf{\Phi}^*(t) = (1, \sin(\omega t), \cos(\omega t), \sin(\omega 2t), \cos(\omega 2t), \sin(\omega 3t), \cos(\omega 3t))$$

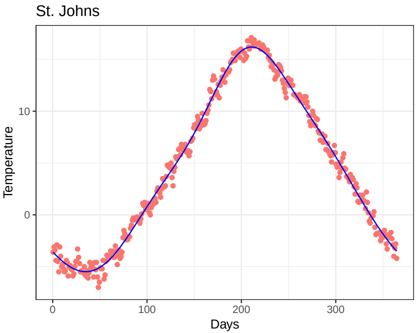




The Fourier Basis

$$\mathbf{\Phi}^*(t) = (1, \sin(\omega t), \cos(\omega t), \ldots, \sin(\omega 5t), \cos(\omega 5t))$$





The Fourier Basis

Summary

- Formula

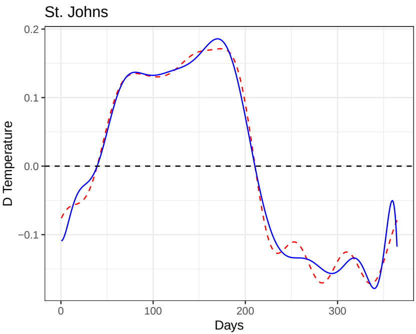
$$x(t) = c_1 + \sum_{j=1}^{M} c_{2j} \sin(\omega jt) + \sum_{j=1}^{M} c_{2j+1} \cos(\omega jt)$$

- Excellent computational properties, especially if the observations are equally spaced
- Natural for describing periodic data $\times$ inappropriate for special types of data (e.g. growth curves)
- Derivatives retain complexity, easy to compute

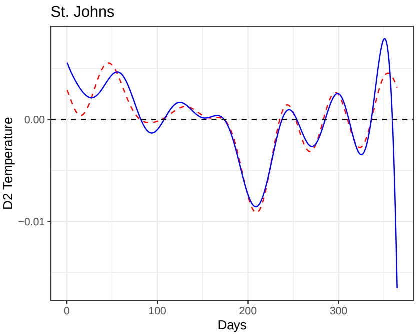
$$D \sin(\omega t) = \omega \cos(\omega t),\ D \cos(\omega t) = -\omega \sin(\omega t)$$

The Fourier Basis

First derivative & Estimate



Second derivative & Estimate



The Bsplines Basis

Splines

- Splines are polynomial segments joined end-to-end.
- Segments are constrained to be smooth at the joins.
- The points at which the segments join are called **knots**.
- System is defined by
  - The order m (order = degree+1) of the polynomial,
  - the location of the knots.

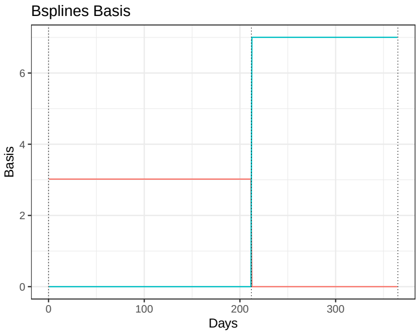
  Thus $K = \#interior\ knots + m$
- **Bsplines** are a particularly useful means of incorporating the constraints.

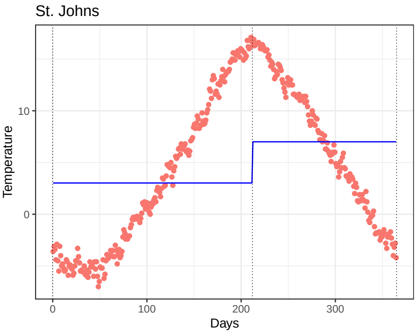
See de Boor, 2001, "A Practical Guide to Splines", Springer.

The Bsplines Basis

3 knots, local constants $(m = 1) \Rightarrow K = 2$

$$\mathbf{\Phi}^*(t) = (Bspl1.1(t), Bspl1.2(t))$$

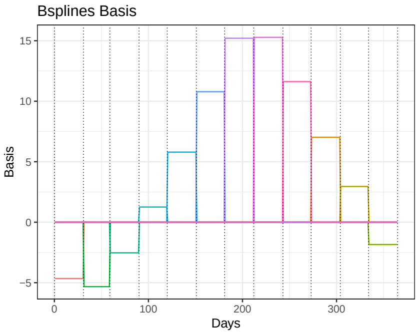


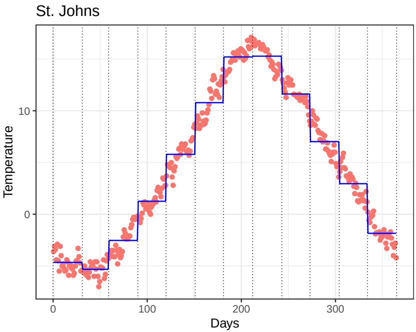


The Bsplines Basis

Knots monthly ($nknots = 13$), local constants ($m = 1$) $\Rightarrow K = 12$

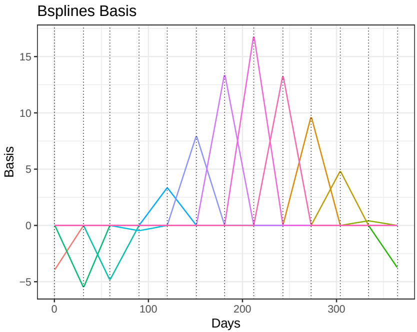
$$\mathbf{\Phi}^*(t) = (Bspl1.1(t), Bspl1.2(t), \ldots, Bspl1.12(t))$$

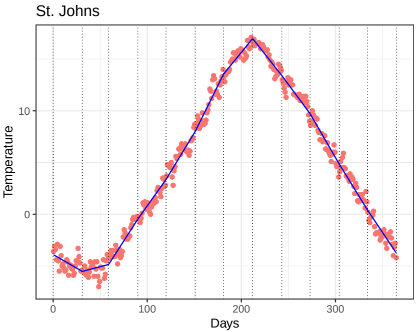




Knots monthly ($nknots = 13$), local linear ($m = 2$) $\Rightarrow K = 13$

$$\mathbf{\Phi}^*(t) = (Bspl2.1(t), Bspl2.2(t), \ldots, Bspl2.13(t))$$

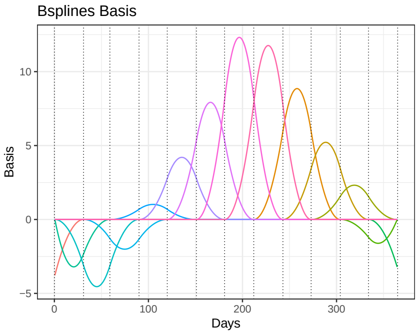


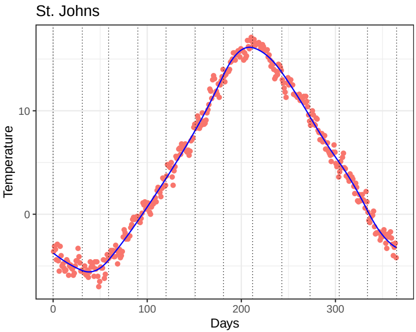


The Bsplines Basis

Knots monthly ($nknots = 13$), local quadratic ($m = 3$) $\Rightarrow K = 14$

$$\mathbf{\Phi}^*(t) = (Bspl3.1(t), Bspl3.2(t), \ldots, Bspl3.14(t))$$

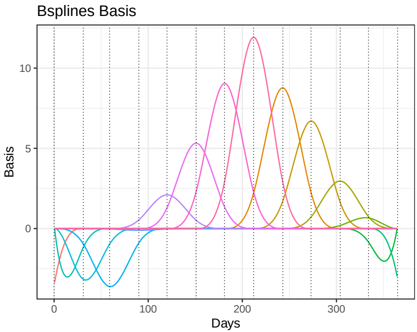


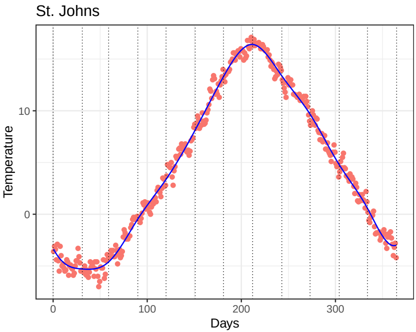


The Bsplines Basis

Knots monthly ($nknots = 13$), local cubic ($m = 4$) $\Rightarrow K = 15$

$$\mathbf{\Phi}^*(t) = (Bspl4.1(t), Bspl4.2(t), \ldots, Bspl4.15(t))$$

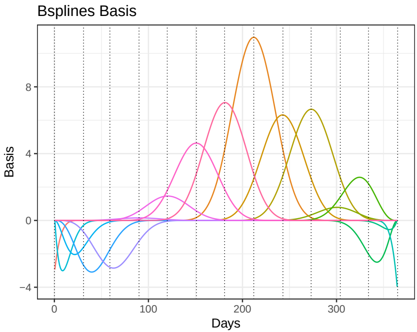


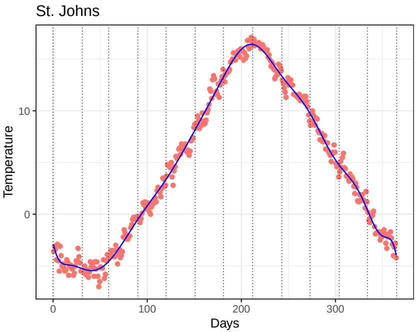


The Bsplines Basis

Knots monthly ($nknots = 13$), splines of order $m = 6 \Rightarrow K = 17$

$$\mathbf{\Phi}^*(t) = (Bspl6.1(t), Bspl6.2(t), \ldots, Bspl6.17(t))$$





The Bsplines Basis

Summary

- Number of basis functions:

$$\#interior\ \ knots + order$$

- Derivatives up to $m - 2$ are continuous.
- B-spline basis functions are positive over at most m adjacent intervals fast computation for even thousands of basis functions.
- Sum of all B-splines in a basis is always 1; can fit any polynomial of order m.
- Most popular choice is order 4, implying continuous second derivatives. Second derivatives have straight-line segments.

The Bsplines Basis

Choosing Knots and Order

- The order of the spline should be at least $k + 2$ if you are interested in k derivatives.
- Knots are often equally spaced (a useful default)
- But there are two important rules:
  - Place more knots where you know there is strong curvature, and fewer where the function changes slowly.
  - Be sure there is at least one data point in every interval.
- Later, we'll discuss placing a knot at each point of observation.
- Co-incident knots reduce the number of continuous derivatives at each point. This can be useful (more later).

The Bsplines Basis

Other

The `fda` library in `R` also allows the following bases:

Constant $\mathbf{\Phi}^*(t) = 1$, the simplest of all.

Power $\mathbf{\Phi}^*(t) = (t^{\lambda_1}, t^{\lambda_2}, t^{\lambda_3}, \dots, t^{\lambda_K})$, powers are distinct but not necessarily integers or positive.

Exponential $\mathbf{\Phi}^*(t) = (e^{\lambda_1 t}, e^{\lambda_2 t}, e^{\lambda_3 t}, \dots, e^{\lambda_K t})$

Other possible bases include

Wavelets especially for sharp, local features (not in `fda`)

Empirical functional Principal Components (special topics)

Problems to solve

1. Canadian Weather Data
   - Load the variable $\mathrm{CanadianWeather}$ from the fda package.
   - Set the time of each measurement to the half of the day, i.e. $t = (0.5, 1.5, 2.5, \ldots, 364.5)$
   - Create a cubic B-spline basis with knots at each point of season change and plot it (see Figure 1).
   - Create a cubic B-spline basis with knots at each fifth day and plot it.
   - Smooth data observed in Edmonton, Halifax, Montreal and Ottawa with using the created basis and plot the results (see Figure 2)
   - Do previous step with Fourier basis (see Figure 3). How many basis functions would be appropriate?
   - Plot the first derivatives of the Fourier-basis-smoothed functions (see Figure 4).
2. Refinery Data
   - Load the variable $\mathrm{refinery}$ from the fda package and plot it.
   - Create a cubic B-spline basis with knots $(0, 33, 67, 98, 130, 162, 193)$, smooth the data and plot the result. Then double (triple) the value 67 in knots and do the same (see all in the Figure 5).

Problems to solve

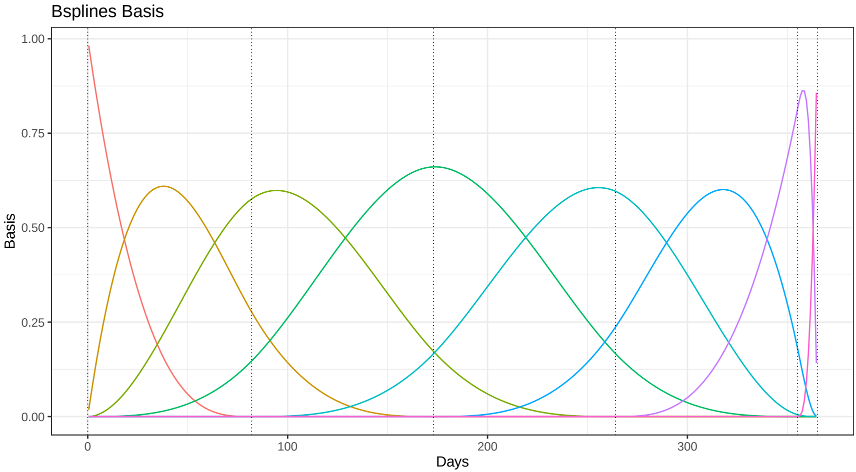


Figure 1.

Problems to solve

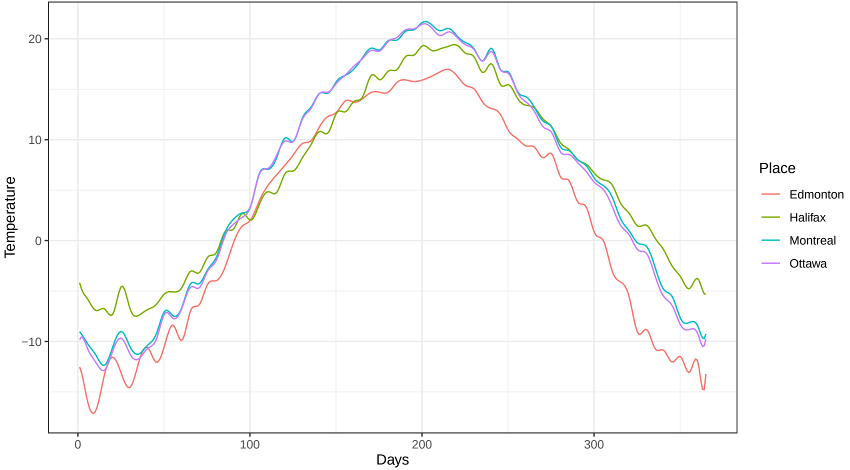


Figure 2.

Problems to solve

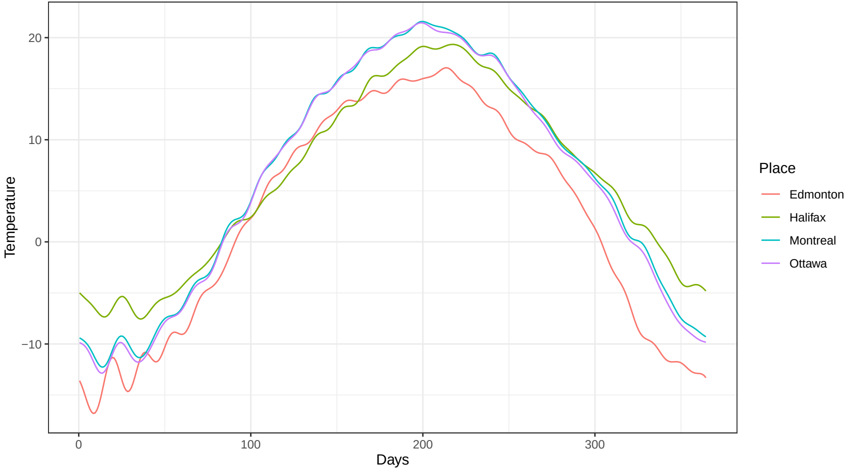


Figure 3.

Problems to solve

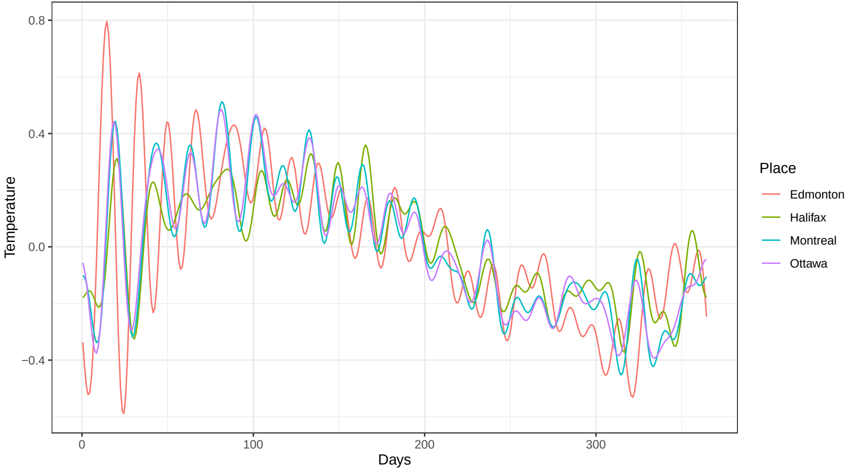


Figure 4.

Problems to solve

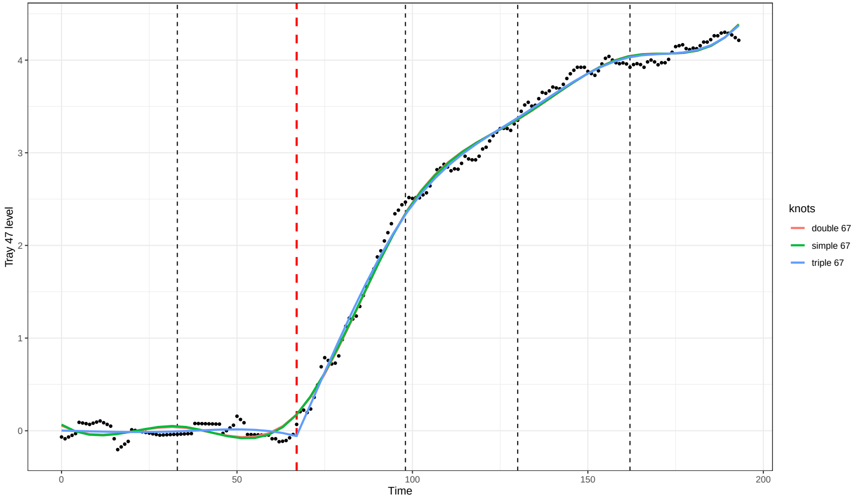


Figure 5.