Splines and basis expansion (week 3?)

6 Feb 2023

Table of contents

linear basis expansion

- transformations of various kinds

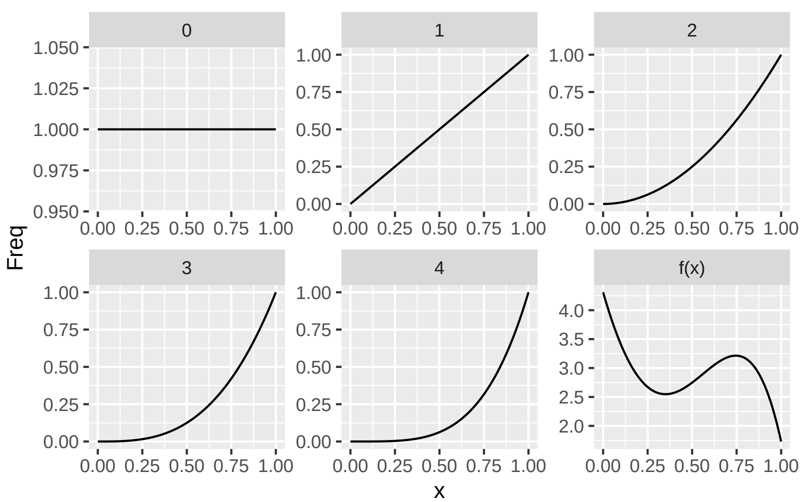
- quadratic expansion
- nonlinear transformations
- indicator variables

Select or regularize from the expanded set.

polynomial basis

- polynomial basis: $y_i = \sum_{j=0}^{n} \beta_j x_i^j$

```
## replicate figure 3.2 from Wood
library(ggplot2)
x <- seq(0, 1, length = 101)
n <- 4
y <- sapply(0:n, \(j) x^j)
beta <- c(4.31, -10.72, 16.8, 2.22, -10.88)
y <- cbind(y, fx = y %*% beta)
dimnames(y) <- list(x = x, j = c(0:n, "f(x)"))
yy <- as.data.frame(as.table(y))
yy$x <- as.numeric(as.character(yy$x))
ggplot(yy, aes(x, Freq)) + geom_line() + facet_wrap(~j, scale = "free")
```



piecewise polynomial bases

- constant, linear, continuous
- basis functions
- translate from x_i to columns of $\mathbf{X}$

splines

- **piecewise** polynomials with continuity/smoothness constraints
- very useful for function approximation
- convert a single numeric predictor into a flexible basis
- efficient
- with multiple predictors, consider **additive models**
- handle interactions (multidim smooth surfaces) *if reasonably low-dimensional*: tensor products etc.

spline terminology

- **knots**: breakpoints (boundary, interior)
- order-M (ESL): continuous derivatives up to order $M-2$ (cubic, $M=4$)
- typically $M=1$, 2, 4
- number of knots = df (degrees of freedom) -1 -intercept

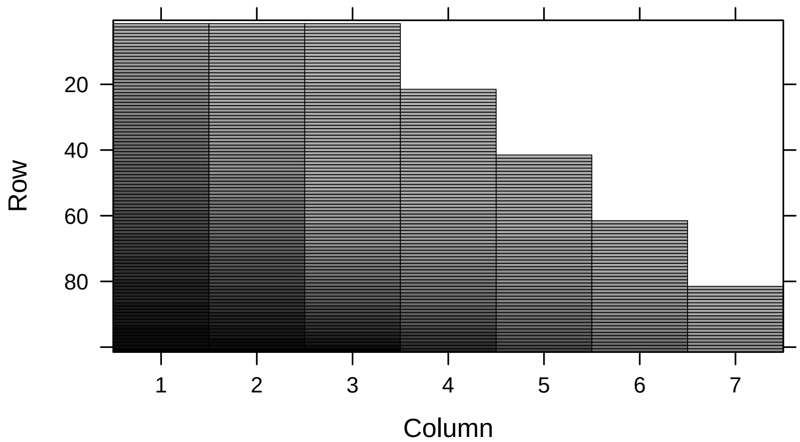
truncated power basis

- $X^0 \dots X^n$
- remaining columns are $(x-\xi_\ell)+^{M-1}$ where ℓ are the *interior knots*

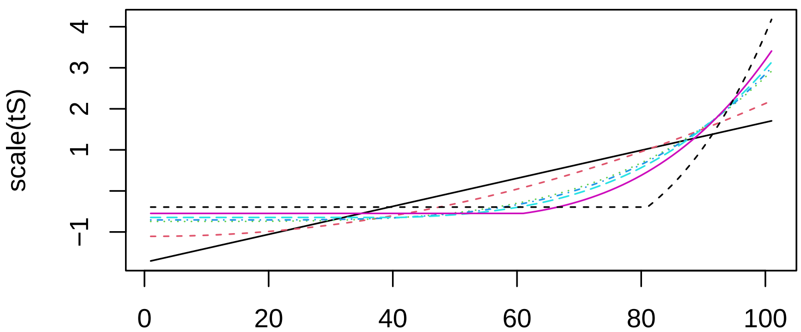
truncated power basis

- **Kronecker product**: blockwise multiplication ($\mathbf{A} \bigotimes \mathbf{B}$ multiplies $\mathbf{B}$ by each a_{ij})
- **Khatri-Rao product**: columnwise Kronecker product
    - super-handy for combining indicator variables with

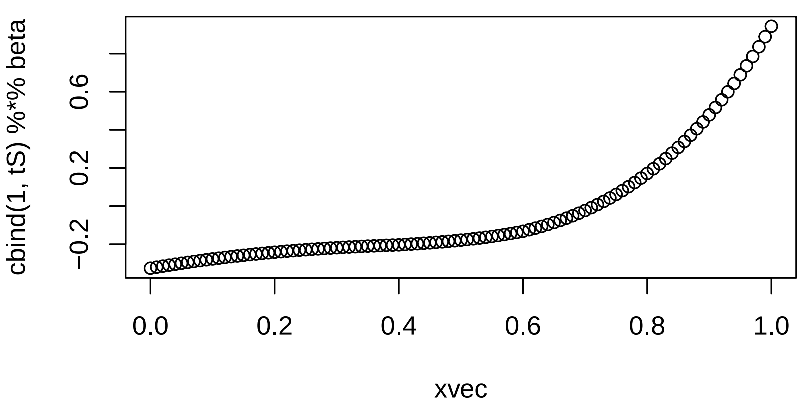
```
Loading required package: Matrix
```



Dimensions: 101 x 7




```r
set.seed(101)
beta <- rnorm(8)
plot(xvec, cbind(1, tS) %*% beta)
```



B-spline basis

- splines of a given order with *minimal support* (i.e., local)
- basis functions defined by recursion (not pretty)
- convenient for regression splines (see below)

natural cubic splines

- linear constraints beyond boundary knots (so 2d and 3d derivatives are 0 at the boundaries)

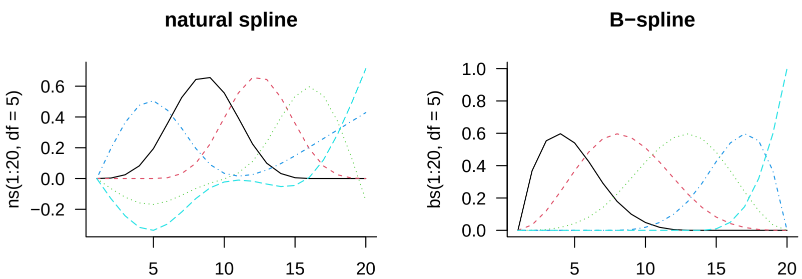
```
library(splines)
bb <- bs(1:20, df = 5)
attributes(bb)[c("degree", "knots", "Boundary.knots")]
```

```
$degree
[1] 3

$knots
33.33333% 66.66667%
 7.333333 13.666667

$Boundary.knots
[1]  1 20
```

```
nn <- ns(1:20, df = 7)
attributes(nn)[c("degree", "knots", "Boundary.knots")]
```

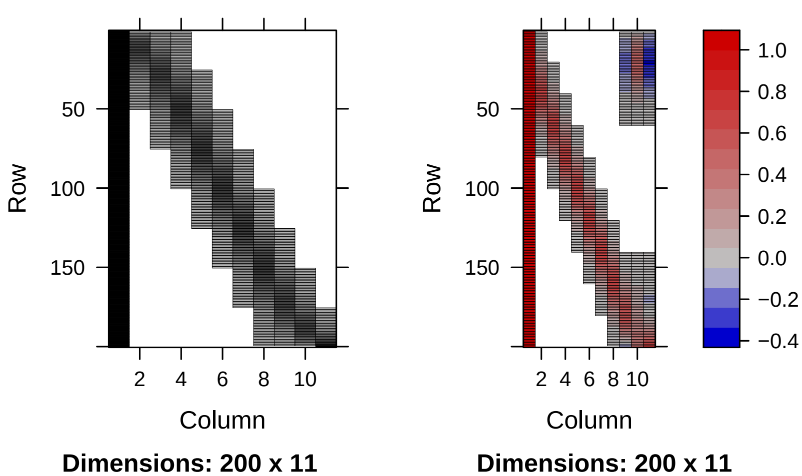
```
$degree
[1] 3

$knots
14.28571% 28.57143% 42.85714% 57.14286% 71.42857% 85.71429%
 3.714286  6.428571  9.142857 11.857143 14.571429 17.285714

$Boundary.knots
[1]  1 20
```

```
par(mfrow = c(1,2),las =1, bty ="l")
matplot(ns(1:20, df = 5), type = "l", main = "natural spline")
matplot(bs(1:20, df = 5), type = "l", main = "B-spline")
```



sparsity patterns

```
library(Matrix)
dd <- data.frame(x=1:200)
Xb <- model.matrix(~splines::bs(x, df = 10), data = dd)
Xn <- model.matrix(~splines::ns(x, df = 10), data = dd)
gridExtra::grid.arrange(
                   ncol = 2,
                   image(Matrix(Xb), aspect = "fill"),
                   image(Matrix(Xn), aspect = "fill")
                   )
```



smoothing splines

- as many knots as data points
- plus squared-second-derivative ("wiggliness") penalty

$$\text{RSS} + \lambda \int (f''(t))^2 \, dt$$

* defined on an infinite-dimensional space * minimizer is a natural cubic spline with knots at x_i

$$(\mathbf{y} - \mathbf{N}\theta)^\top (\mathbf{y} - \mathbf{N}\theta) + \lambda \theta^\top \Omega_N \theta$$

with $\{\Omega_N\}_{jk} = \int N_j''(t) N_k''(t) \, dt$ $$ **generalized** ridge regression: penalize by $\lambda \Omega_N$ rather than λI * same data augmentation methods as before except that now we use $\sqrt{\lambda} C$ where C is a matrix square root of Ω_N

See Wood (2017), Perperoglou et al. (2019)

Wood, Simon N. 2017. *Generalized Additive Models: An Introduction with R*. CRC Texts in Statistical Science. Chapman & Hall. https://www.amazon.com/Generalized-Additive-Models-Introduction-Statistical-ebook/dp/B071Z9L5D5/ref=sr_1_1?ie=UTF8&qid=1511887995&sr=8-1&keywords=wood+additive+models.

Perperoglou, Aris, Willi Sauerbrei, Michal Abrahamowicz, and Matthias Schmid. 2019. "A Review of Spline Function Procedures in R." *BMC Medical Research Methodology* 19 (1): 46. https://doi.org/10.1186/s12874-019-0666-3.

examples: South African heart disease

- use splines in a GLM with no additional effort
- fit splines to all continuous variables
- ESL says "use four natural spline bases" (… elements??)
- i.e. `df = 4` (no intercept)
- stepwise deletion via AIC
- why??
- showing p-values (why???)
- `stepAIC(..., direction = "backward")`

phoneme example

- combination of feature transformation (time to Fourier domain) and regularization
- smooth first, regress afterwards

degrees of freedom and smoother matrix

- The equivalent of the hat matrix is

$$\mathbf{N}(\mathbf{N}^\top\mathbf{N} + \lambda\Omega_N)^{-1}\mathbf{N}^\top$$

- "Smoother matrix"
- hat matrix is *idempotent* (why?), smoother matrix is *shrinking*
- smoother matrix has lower rank
- effective degrees of freedom = trace of hat matrix (again)

reduced-rank splines

- We *can* use as many knots as observations, but do we really need to?
- From `?mgcv::s`:

> … exact choice of 'k' is not generally critical: it should be chosen to be large enough that you are reasonably sure of having enough degrees of freedom to represent the underlying `truth' reasonably well, but small enough to maintain reasonable computational efficiency. Clearly`large' and 'small' are dependent on the particular problem being addressed.
>
> The default basis dimension, 'k', is the larger of 10 and 'm[1]' (spline order)

generalized cross-validation

Larsen (2015), Golub, Heath, and Wahba (1979)

- minimize $\mathrm{RSS}/(\mathrm{Tr}(\mathbf{I} - \S(\lambda)))^2$
- "a rotation-invariant version of PRESS" $(\sum(e_i/(1-h_i i))^2)$

Larsen, Kim. 2015. "GAM: The Predictive Modeling Silver Bullet | Stitch Fix Technology – Multithreaded." *MultiThreaded (StitchFix)*. https://multithreaded.stitchfix.com/blog/2015/07/30/g

Golub, Gene H., Michael Heath, and Grace Wahba. 1979. "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter." *Technometrics* 21 (2): 215–23. https://doi.org/10.1080/00401706.1979.10489751.

multidimensional splines

tensor product

thin-plate splines

null space