Basis expansions and Generalized additive models

Datasets

- HydroGeology.

Problems

Select any function from the table below. Generate a sample $(x_1, \ldots, x_N)$ from the uniform distribution on the interval defined by $\mathrm{dom}(f)$. Next, generate $y_n = f(x_n) + \varepsilon_n$, where $\varepsilon_n \sim \mathcal{N}(0, \sigma^2), n = 1, \ldots, N$. We assume that y is a dependent variable, and x is a predictor. Then do the following:

1. Construct a polynomial ridge regression with the degree of the polynomial and regularization parameter selected according to cross-validation. Calculate the GCV.
2. Approximate y with a cubic spline with fixed knots without smoothing. The number and placement of knots is arbitrary.
3. Approximate y with a cubic spline with fixed nodes and with smoothing. The number and placement of knots is arbitrary. Set the smoothing parameter by minimizing the GCV.
4. Display the basis functions for your spline.
5. Build a smoothing spline. Select a smoothing parameter using GCV.
6. Draw on the single plot the sample, the function $f(x)$ and the function estimates obtained by each of the methods above. Compare the results in terms of the number of degrees of freedom and the accuracy of the GCV.

For the HydroGeology dataset, fit the additive model to predict the value of variable q. Try to choose the parameters so as to minimize the cross validation error (or GCV).

Recommendations

Least squares fit for polynomial regression can be applied by calling

```
lm(y ~ poly(x, K)),
```

where `K` is a polynomial degree. To use regularization, the function `poly` can calculate the values of the basis functions, so that the regression plan matrix can be compiled.

The widest possibilities for constructing splines and generalized additive models are available through the `gam` function from the `mgcv` package. Splines are defined through the `s` function of this package, and are specified in a form similar to the use of `poly`. Read the man pages for `gam, s, gamObject, smooth.terms, cubic.regression.spline, smooth.construct, predict.gam`.

In particular, the basis for the spline is specified by the argument `bs` of the function `s`, the number of basis functions is determined by the argument `k` for `s`, the use of the penalty term is determined by the argument `fx` for `s`, the position of knots for splines is determined by the argument `knots` of the function `gam`. The value of the basis functions can be obtained using the `predict.gam` function with the argument `type = "lpmatrix"`.

$f(x)$	$\mathrm{dom}(f)$	σ	N
$\frac{\sin(x)}{x}$	$[0; 4\pi]$	0.2	50
$\cos x + \frac{x}{4}$	$[0; 4\pi]$	0.75	50
$\exp(-x^2) - \exp(-\frac{(x-2)^2}{8})$	$[-2; 4]$	0.2	50
$\exp(-x) + \frac{1}{4}\sin(4x)$	$[0; 4]$	0.1	50
$x \cos x$	$[0; 2\pi]$	1	50
$\ln\left(\frac{x}{\pi - x}\right) - \sin(4x)$	$(0; \pi)$	0.75	50
$\ln\left(\sin(x^2) + 1.2\right)$	$[0; 4\pi]$	0.25	50
$x^2 + \sin(2\pi x) I(x > 0)$	$[-2; 2]$	0.5	50
$\sqrt{\lvert x \rvert} - x \cos(\pi x) I(x > 0)$	$[-2; 2]$	0.4	50
$\ln(x) \sin(\pi x)$	$[0; 2]$	0.15	50
$\sin(x) \ln(1 + x^2)$	$[-\pi; \pi]$	0.5	50