Distributional Modelling in R

02-Smoothing - Exercises

1. In this example we analyze motor cycle accident data of the **MASS** package.

   ```
   R> data("mcycle", package = "MASS")
   ```

   (a) Generate a scatter plot to visualize the `mcycle` data. What insights can you derive from the plot?

   (b) From the `functions.R` script, use the implementation for B-spline basis functions and the penalty matrix. Set up a function `find.lambda()` that searches for the optimal smoothing parameter using either the AIC or BIC. The function should have the following arguments:

   ```
   find.lambda(y, B, K)
   ```

   where `y` is the response vector, `B` is the B-spline design matrix, and `K` is the penalty matrix.

   (c) For optimization, employ the general-purpose optimizer function `optimize()`. Find the optimal smoothing parameter in the `mcycle` example data using degree-3 P-splines with 80 knots and a second-order difference penalty.

   (d) The fitted regression curve can be expressed as $f(x) = B\boldsymbol{\beta}$. Confidence intervals can be computed with
   $$\hat{f}(x) \pm z_{1-\alpha/2}\hat{\sigma}\sqrt{\text{diag}(\mathbf{S}_\lambda \mathbf{S}_\lambda^\top)},$$
   where $\mathbf{S}_\lambda = \mathbf{B}(\mathbf{B}^\top\mathbf{B} + \lambda\mathbf{K})\mathbf{B}^\top$ is the smoother matrix and $\hat{\sigma}$ the estimated residual standard deviation
   $$\hat{\sigma} = \sqrt{\frac{1}{n-p}\hat{\varepsilon}^\top\hat{\varepsilon}}.$$
   Now, visualize the fitted curve along with 95% confidence intervals. To achieve this, calculate the number of parameters p utilized to fit the model by the trace of the smoother matrix. What observations and insights can you derive from this analysis?

   (e) Now, utilize the **mgcv** package for estimation. Explore the impact of adjusting the number of basis functions and/or modifying the type of basis function. Investigate how these alterations influence the model's performance or behavior.

2. In this example we analyze the Zambia nutrition data of the **R2BayesX** package.

   ```
   R> data("ZambiaNutrition", package = "R2BayesX")
   ```

   The primary interest is to model the dependence of stunting of newborn children, with an age ranging from 0 to 5 years, on covariates such as the body mass index of the mother, the age of the child and others. The map of Zambia is provided in

   ```
   R> data("ZambiaBnd", package = "R2BayesX")
   ```

   (a) In order to estimate a spatial model employing a Markov random field smoother, it is necessary to compute the penalty matrix based on neighboring regions. The `neighbormatrix()` function within the **bamlss** package facilitates this process. Carefully assess the outcomes.

   (b) Once the penalty matrix is established, proceed to estimate a spatial Generalized Additive Model (GAM) utilizing the **mgcv** package. Incorporate all relevant covariates along with a random effect component for the districts in Zambia. Subsequently, provide an interpretation of the obtained results.

   (c) Utilize the `predict()` method to generate predictions for both the structured and unstructured spatial effects. For effective visualization, consider employing the `plotmap()` function from the **bamlss** package or leverage the capabilities of the **sf** package. This will help provide a clear and insightful representation of the spatial effects.