Practical Four

February 2018

Overview

This practical is designed to compare the results when using a range of smoother-based approaches to an unevenly smooth function. In particular, we will see the value of spatially adaptive smoothing using underfitting and overfitting scores as evaluation metrics. In particular, the objectives are to understand:

1. the tension between fitting models which generate predictions which are close to the data and fitting models which are 'overfitted' to the data;
2. the relationship between variance and bias

Methods

To meet these objectives we will consider the following methods:

1. polynomials (with $d = 6$)
2. penalised regression splines (with complexity chosen using MGCV)
3. spatially adapative regression splines (with automated model selection using SALSA)

We will make these comparisons by examining how different methods perform when exposed to simulated data (so we know what the underlying function is and how each method reproduces 'truth').

Specifically, we will measure fit to the data using the residual sums of squares (RSS) and fit to the underlying function using the mean average squared error (ASE; the mean of the average squared difference between the fitted values and the underlying function evaluated at the observed x-values).

Comparing performance across methods

1. Read in the 'simulatedData.csv' file (from Moodle) which contains 100 sets of simulated data from:

$$y_i = \mu_i + N(0, \sigma = 0.2)$$

where

$$\mu_i = \sqrt{x_i(1 - x_i)}\sin\frac{2\pi(1 + 2^{(9-4*6)/5})}{x_i + 2^{(9-4*6)/5}}$$

where **x** is a sequence of 400 numbers from 0 to 1 (`x<- seq(0,1,length=400)`).

2. Fit a polynomial model to each simulated set and evaluate the RSS and ASE in each case using:

```r
dat<- read.csv("simulatedData.csv", header=T)
```

Create a place to store the results:

```r
RSS_poly<- c()
ASE_poly<- c()
```

plot the data and underlying function (`mu`) to examine how the fitted curves compare with the underlying function:

```
plot(dat$x, dat$response, pch=16, col="grey")
lines(dat$x,dat$mu, lwd=2)
```

Fit a polynomial model to each of the 100 simulated sets:

```
plot(dat$x, dat$response, pch=16, col="grey")
lines(dat$x,dat$mu, lwd=2)
for(i in 1:100){
  #print(i)
  #subset the data so it only uses one at a time:
  datasub<- dat[dat$sim==i,]
  polymod<- lm ( response ~ poly(x,6), data=datasub)
  lines(datasub$x,fitted(polymod), col=2)
   RSS_poly[i]<- sum(residuals(polymod)**2)
  ASE_poly[i]<- mean((datasub$mu-fitted(polymod))**2)
}
```

```
#find the mean of the RSS and ASE scores
mean(RSS_poly)
mean(ASE_poly)
```

3. Fit a penalised regression spline GAM to each set and evaluate the RSS and ASE in each case using:

```
RSS_gam<- c()
ASE_gam<- c()
```

Plot the data and underlying function (mu) to examine how the fitted curves compare with the underlying function.

Fit a PRS GAM to each of the 100 simulated sets:

```
plot(dat$x, dat$response, pch=16, col="grey")
lines(dat$x,dat$mu, lwd=2)
require(mgcv)

for(i in 1:100){
  #print(i)
  #subset the data so it only uses one at a time:
  datasub<- dat[dat$sim==i,]
  fitgam<- gam(response ~ s(x), data=datasub)
  lines(datasub$x,fitted(fitgam), col=2)
  RSS_gam[i]<- sum(residuals(fitgam)**2)
  ASE_gam[i]<- mean((datasub$mu-fitted(fitgam))**2)
}
```

```
mean(RSS_gam)
mean(ASE_gam)
```

4. Fit a spatially adaptive spline based model using SALSA to each simulated set and evaluate the RSS and ASE in each case using:

```
require(MRSea)
ASE_SALSA<- c()
knots<- c()
RSS_SALSA<- c()
```

```
plot(dat$x, dat$response, pch=16, col="grey")
lines(dat$x,dat$mu, lwd=2)
```

```r
for(i in 1:100){
  #print(i)

  datasub<- dat[dat$sim==i,]

  initialModel<- glm(response ~ 1, data=datasub)
  salsa1dlist<-list(fitnessMeasure = 'BIC',
                    minKnots_1d=c(2),
                    maxKnots_1d = c(40),
                    startKnots_1d = c(10),
                    degree=c(2),
                    maxIterations = 10,
                    gaps=c(0))

  # run SALSA
  salsa1dOutput<-runSALSA1D(initialModel, salsa1dlist,
                            varlist=c('x'),factorlist=NULL, datasub,
                            splineParams=NULL, suppress.printout=TRUE,
                            datain=datasub)

  ASE_SALSA[i]<- mean((datasub$mu-fitted(salsa1dOutput$bestModel))**2)
  knots[i]<-length(salsa1dOutput$splineParams[[2]]$knots)
  RSS_SALSA[i]<- sum(residuals(salsa1dOutput$bestModel)**2)
  lines(datasub$x,fitted(salsa1dOutput$bestModel), lwd=2, col=2)

}
```

```r
mean(RSS_SALSA)
mean(ASE_SALSA)
```

5. Examine the distribution of the number of knots obtained using SALSA across the 100 sets using:

```r
barplot(table(knots))
```

6. For the SALSA based models, examine the relationship between model complexity (using knots) and both RSS and ASE using:

```r
par(mfrow=c(1,2))
plot(knots, ASE_SALSA)
lines(lowess(x=knots, y=ASE_SALSA))
plot(knots, RSS_SALSA)
lines(lowess(x=knots, y=RSS_SALSA))
```

```r
summary(lm(ASE_SALSA ~knots))
```

```r
summary(lm(RSS_SALSA ~knots))
```

Questions

1. Which of the following about the PRS-based curves in this case is FALSE?
   - The PRS-based curves have higher variance and lower bias than the polynomial based curves.
   - The PRS-based curves have lower variance and higher bias than the curves based on the spatially adaptive method (via SALSA).
   - The PRS-based curves have a better average fit to the data than the polynomial curves.
   - The PRS-based curves have worse average fit to the data than the curves based on the spatially adaptive method (via SALSA).
   - The PRS-based curves have higher variance and higher bias than the curves based on the spatially adaptive method (via SALSA).

2. Which of the following about the spatially adaptive GAM-based curves (via SALSA) in this case is FALSE?
   - The spatially adaptive GAM-based curves (via SALSA) have lower variance and lower bias than the polynomial-based curves.
   - The spatially adaptive GAM-based curves (via SALSA) have higher variance and lower bias than the PRS-based curves.
   - The spatially adaptive GAM-based curves (via SALSA) return a better average fit to the data than the polynomial-based curves.
   - The spatially adaptive GAM-based curves (via SALSA) return a better average fit to the data than the PRS-based curves.

3. TRUE or FALSE? The polynomial curves are underfitting relative to the PRS-based curves.

4. TRUE or FALSE? The mean RSS score for the polynomial curves is about 1.4 times the size of the corresponding value for the PRS-based curves.

5. TRUE or FALSE? The mean ASE score for the polynomial curves is about 1.8 times the size of the corresponding value for the PRS-based curves.

6. TRUE or FALSE? The polynomial curves are overfitting relative to the spatially adaptive SALSA-based curves.

7. TRUE or FALSE? The mean RSS score for the polynomial curves is approximately 2.3 times the size of the corresponding value for the spatially adaptive SALSA-based curves.

8. TRUE of FALSE? The mean ASE score for the polynomial curves is approximately 8.4 times the size of the corresponding value for the spatially adaptive SALSA-based curves. 8.58153787

9. TRUE or FALSE? The PRS-based GAMs are returning overfitted models relative to the spatially adaptive SALSA-based curves.

10. TRUE or FALSE? The mean RSS score for the PRS-based curves is about 1.7 times the size of the corresponding value for the spatially adaptive SALSA-based curves.

11. TRUE or FALSE? The mean ASE score for the PRS-based GAMs is about 4.6 times the size of the corresponding value for the spatially adaptive SALSA-based curves.

12. TRUE or FALSE? The spatially adaptive SALSA-based GAMs are slower to fit to the data than PRS-based GAMs but the SALSA-based curves are closer to the underlying function compared to the PRS-based GAMs (which has a single smoothing parameter for the entire covariate range).

13. As model complexity increases for the SALSA based models, the ASE generally decreases until about 12 knots. Beyond 12 knots the ASE appears to be relatively similar even if the number of knots increases. This is the reason that the ASE is a good fit measure to use in practice when fitting models to observed data.

14. As model complexity increases for the SALSA based models, the RSS decreases. The RSS continues to decrease as the number of knots increases. This is the reason that overfitting can arise when the RSS is used to choose model complexity when fitting models to observed data.