# STAT 824 hw 04

Orthogonal series estimator, backfitting, sparse backfitting, bootstrap

1. Suppose $\{\varphi_j\}_{j=1}^{\infty}$ is a basis for all functions $f : [0,1] \to \mathbb{R}$ such that $\int_0^1 |f(x)|^2 dx < \infty$ which satisfies

$$\int_0^1 \varphi_j(x)\varphi_{j'}(x)dx = \begin{cases} 1, & j = j' \\ 0, & j \neq j'. \end{cases} \tag{1}$$

A basis with the above property is called an *orthonormal basis*. Assume we can represent $f$ as

$$f(x) = \sum_{i=1}^{\infty} \theta_j \varphi_j(x), \quad \text{where} \quad \theta_j = \int_0^1 f(x)\varphi_j(x)dx, \quad j = 1, 2, \ldots$$

We will consider estimating the approximation $f_n^N(x) = \sum_{i=1}^N \theta_j \varphi_j(x)$ for some finite $N$ in the context of nonparametric regression.

(a) Consider the *trigonometric basis*, which is given by $\varphi_1(x) = 1$, $\varphi_{2k}(x) = \sqrt{2}\cos(2\pi kx)$, and $\varphi_{2k+1}(x) = \sqrt{2}\sin(2\pi kx)$ for $k = 1, 2, \ldots$ for $x \in [0,1]$. Show that this basis is orthonormal, i.e. that it satisfies 1.

It's a little laborious, but you have to show

$$\int_0^1 \varphi_1(x)\varphi_1(x) = 1$$
$$\int_0^1 \varphi_1(x)\varphi_{2k}(x) = 0$$
$$\int_0^1 \varphi_1(x)\varphi_{2k+1}(x) = 0$$
$$\int_0^1 \varphi_{2k}(x)\varphi_{2k'+1}(x) = 0$$
$$\int_0^1 \varphi_{2k}(x)\varphi_{2k'}(x) = \begin{cases} 1, & k = k' \\ 0, & k \neq k' \end{cases}$$
$$\int_0^1 \varphi_{2k+1}(x)\varphi_{2k'+1}(x) = \begin{cases} 1, & k = k' \\ 0, & k \neq k', \end{cases}$$

which can be done with the aid of a calculus book giving antiderivatives for products of sines and cosines.

(b) Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be data pairs such that $Y_i = f(X_i) + \varepsilon_i$, where $X_i = i/n$, $i = 1, \ldots, n$ and $\varepsilon_1, \ldots, \varepsilon_n$ are independent with mean zero and variance $\sigma^2 < \infty$. Consider the estimator $\hat{f}_n^N$ of $f$ given by

$$\hat{f}_n^N(x) = \sum_{j=1}^{N} \hat{\theta}_j \varphi_j(x), \quad \text{where} \quad \hat{\theta}_j = n^{-1} \sum_{i=1}^{n} Y_i \varphi_j(X_i), \quad j = 1, \ldots, N. \tag{2}$$

This type of estimator is called an *orthogonal series estimator*. See [2] for more details.

i. For $x \in [0,1]$, find weights $W_{n1}(x), \ldots, W_{nn}(x)$ such that $\hat{f}_n^N(x) = \sum_{i=1}^n W_{ni}(x) Y_i$.

We have

$$W_{ni}(x) = \frac{1}{n} \sum_{j=1}^{N} \varphi_j(X_i)\varphi_j(x).$$

ii. Give the entries of the matrix $\mathbf{S}$ such that $\hat{\mathbf{f}}_n^N = \mathbf{S}\mathbf{Y}$, where $\hat{\mathbf{f}}_n^N = (\hat{f}_n^N(X_1), \ldots, \hat{f}_n^N(X_n))^T$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$.

The smoother matrix $\mathbf{S}$ is given by

$$\mathbf{S} = \begin{bmatrix} n^{-1}\sum_{j=1}^N \varphi_j(X_1)\varphi_j(X_1) & \ldots & n^{-1}\sum_{j=1}^N \varphi_j(X_n)\varphi_j(X_1) \\ \vdots & \ddots & \vdots \\ n^{-1}\sum_{j=1}^N \varphi_j(X_1)\varphi_j(X_n) & \ldots & n^{-1}\sum_{j=1}^N \varphi_j(X_n)\varphi_j(X_n) \end{bmatrix}$$

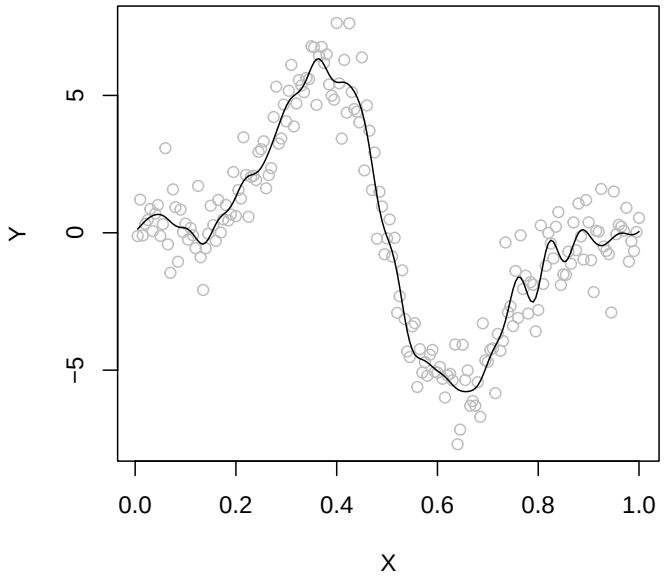
iii. Give the matrix $\mathbf{B}$ such that $\mathbf{S} = (1/n)\mathbf{B}\mathbf{B}^T$.

We can set $\mathbf{B} = (\varphi_j(X_i))_{1 \le i \le n, 1 \le j \le N}$.

iv. Generate data with the R code

```
m <- function(x){ - 25 * 4 * (2*x - 1) * dnorm(4*(2*x - 1))}
n <- 200
X <- c(1:n)/n
Y <- m(X) + rnorm(n,0,1)
```

Then make a scatterplot of the data with a curve overlaid which traces the fitted values $\hat{f}_n^N(X_1), \ldots, \hat{f}_n^N(X_n)$ of the estimator in (2) based on the trigonometric basis with functions for $k = 1, \ldots, 20$, such that $N = 41$. My plot looks like this:




Here is my code:

```
K <- 20 # makes N = 2 * K + 1 basis functions
B <- matrix(1,n,2*K + 1)
for( k in 1:K){

  B[,2*k] <- sqrt(2) * cos( 2*pi*k*X)
  B[,2*k + 1] <- sqrt(2) * sin( 2*pi*k*X)

}
S <- (1/n) * B %*% t(B)
m.hat <- S %*% Y

plot(Y ~ X, col = "gray")
lines(m.hat ~ X)
```

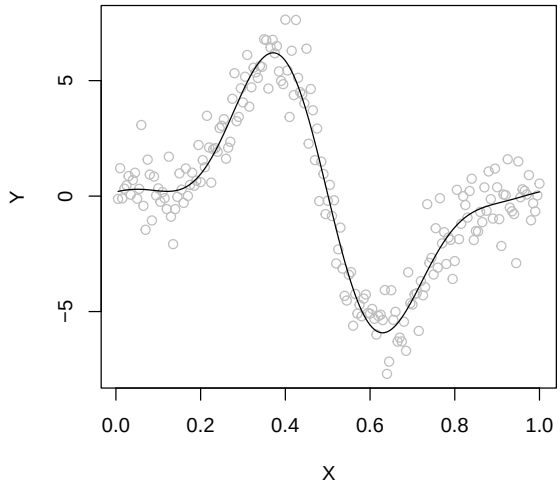
v. What do you notice about the quantities $n^{-1}\sum_{i=1}^{n}\varphi_j(i/n)\varphi_{j'}(i/n)$, $1 \le j, j' \le N$, in relation to the property in (1)? *Hint: These are the entries of the matrix* $(1/n)\mathbf{B}^T\mathbf{B}$*, which you can compute in R.*

We find that $(1/n)\mathbf{B}^T\mathbf{B} = \mathbf{I}_n$, so that when $X_i = i/n$ for $i = 1, \ldots, n$, the trigonometric basis is orthonormal with respect to the empirical distribution of $X_1, \ldots, X_n$.

vi. Now consider using the trigonometric basis with functions for $k = 1, \ldots, K$, giving $N = 2K+1$ total basis functions: Choose $K$ via leave-one-out crossvalidation (note that you can use the special trick for linear estimators to save computation time). Report the chosen value of $K$

and the corresponding number of basis functions $N$. Also make a scatterplot of the data with the curve tracing the fitted values overlaid.

For my data, leave-one-out crossvalidation selected $K = 3$, giving $N = 2(3) + 1 = 7$ total basis functions. Here is the plot:




2. Import into R the data in this `.Rdata` file and fit the additive model

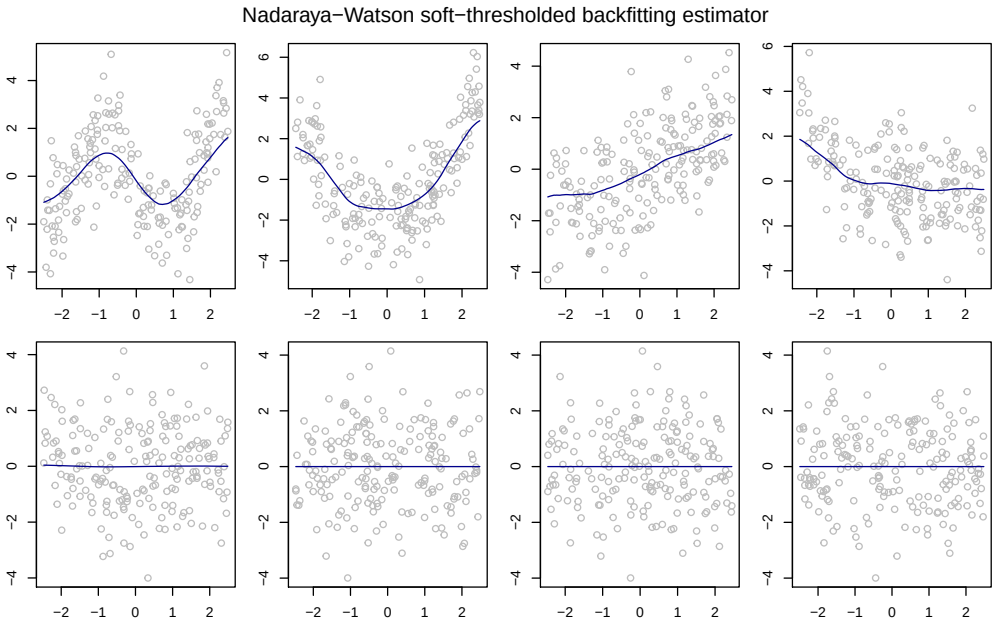
$$Y = \mu + m_1(X_1) + \cdots + m_8(X_8) + \varepsilon$$

with a soft-thresholded (sparse) Nadaraya-Watson backfitting estimator, enforcing the usual identifiability condition on the additive components.

(a) Give $\hat{\mu}$.

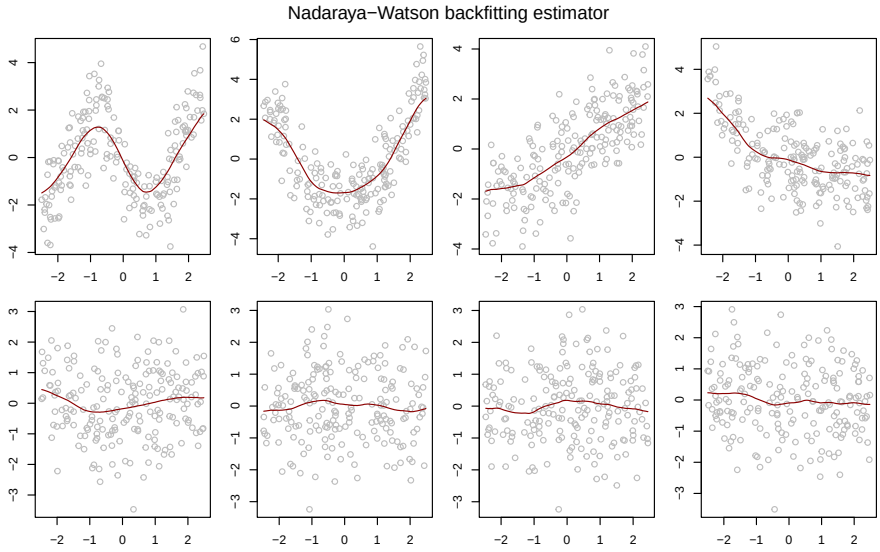
We estimate $\mu$ with $\hat{\mu} = \bar{Y}_n = 4$.

(b) Make a plot like the one pictured below (choose a bandwidth $h$ and a soft-thresholding parameter just by eyeballing the plot), where in panel $j$, the points $(Y_i - \sum_{k \neq j} \hat{m}_k(X_{ik}), X_{kj})$, $i = 1, \ldots, n$, are plotted along with a line tracing the fitted values $\hat{m}_j(X_{ij})$, $i = 1, \ldots, n$.




(c) Now fit Nadaraya-Watson backfitting estimator *without* soft-thresholding; make a similar plot.

Mine looks like this




3. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be iid realizations of $(X, Y)$. Let $\rho = \text{corr}(X, Y)$ and $\hat{\rho}$ be the sample correlation. If $(X, Y)$ are bivariate Normal then $\sqrt{n}(\zeta(\hat{\rho}) - \zeta(\rho)) \xrightarrow{\text{D}} \text{Normal}(0, 1)$ as $n \to \infty$ where

$$\zeta(\rho) = \frac{1}{2} \log\left(\frac{1+\rho}{1-\rho}\right).$$

(a) Let $Y|X \sim \text{Normal}(\rho X, 1 - \rho^2)$, $X \sim \text{Normal}(0, 1)$ so that $(X, Y)$ are bivariate standard Normal with correlation $\rho$. For $\alpha = 0.05$, $n = 50$, $\rho = 1/2$, and $B = 500$, run a simulation with 500

simulated data sets to compare the coverage of $\rho$ and the average width of the three intervals

$$\begin{aligned}\mathcal{A}_n &= [\zeta^{-1}(\zeta(\hat{\rho}) - n^{-1/2} z_{\alpha/2}), \zeta^{-1}(\zeta(\hat{\rho}) + n^{-1/2} z_{\alpha/2})] \\ \mathcal{B}_n^{\text{pctl}} &= [\hat{\rho}_n^{*((\alpha/2)B)}, \hat{\rho}_n^{*((1-\alpha/2)B)}] \\ \mathcal{B}_n^{\text{piv}} &= [\zeta^{-1}(2\hat{\zeta}_n - \hat{\zeta}_n^{*((1-\alpha/2)B)}), \zeta^{-1}(2\hat{\zeta}_n - \hat{\zeta}_n^{*((\alpha/2)B)})],\end{aligned}$$

where $\zeta^{-1}(z) = \frac{e^{2z}-1}{e^{2z}+1}$, $\hat{\rho}_n^{*(1)} \leq \cdots \leq \hat{\rho}_n^{*(B)}$ are sorted bootstrap realizations of $\hat{\rho}$ from samples drawn with replacement from $(X_1, Y_1), \ldots, (X_n, Y_n)$, and $\hat{\zeta}_n^{*(b)} = \zeta(\hat{\rho}_n^{*(b)})$ for $b = 1, \ldots, B$ with $\hat{\zeta}_n = \zeta(\hat{\rho}_n)$.

I obtained

| | asymp | boot pctl | boot piv |
|---|---|---|---|
| coverage | 0.950 | 0.946 | 0.952 |
| avg width | 0.409 | 0.416 | 0.420 |

(b) Now let $Y|X \sim \text{Normal}(X, \sigma^2)$, $X \sim \text{Exponential}(\lambda)$ with $\lambda = 1$ and $\sigma^2 = 3$. Find $\rho = \text{corr}(X, Y)$ and compare the coverage of $\rho$ and the width of the intervals for $\alpha = 0.05$, $n = 50$, and $B = 500$ as before.

I obtained

| | asymp | boot pctl | boot piv |
|---|---|---|---|
| coverage | 0.916 | 0.924 | 0.930 |
| avg width | 0.409 | 0.454 | 0.446 |

(c) Why does the asymptotic interval $\mathcal{A}_n$ perform poorly under the settings in part (b)?

The result $\sqrt{n}(\zeta(\hat{\rho}) - \zeta(\rho)) \xrightarrow{\text{D}} \text{Normal}(0, 1)$ depends on $(X, Y)$ having the bivariate Normal distribution. Since this is not the case, the asympototic variance may be incorrect. It seems that the assumption of Normality is very important to the reliability of this interval.

(d) Which interval performed best in parts (a) and (b)?

For me the very basic percentile interval performed the best. Very cool!

4. (Optional) Let $\mathbf{X} \in \mathbb{R}^{n \times p}$, $p < n$, be a full-rank matrix and let $\mathbf{Y} \in \mathbb{R}^n$ and partition the columns of $\mathbf{X}$ such that $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_{-1}]$. Let $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ be the vector such that $(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}$ and let $\hat{\boldsymbol{\beta}}$ be partitioned in the same way as $\mathbf{X}$ into

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_{-1} \end{bmatrix}.$$

Define $\mathbf{P}_1 = \mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T$ and $\mathbf{P}_{-1} = \mathbf{X}_{-1}(\mathbf{X}_{-1}^T\mathbf{X}_{-1})^{-1}\mathbf{X}_{-1}^T$, and let $\mathbf{X}_{1\setminus -1} = (\mathbf{I} - \mathbf{P}_{-1})\mathbf{X}_1$ be the residuals from regressions of the columns of $\mathbf{X}_1$ onto the columns of $\mathbf{X}_{-1}$.

(a) Let $\hat{\mathbf{Y}}_1 = \mathbf{X}_1\hat{\boldsymbol{\beta}}_1$ and let $\hat{\mathbf{Y}}_{-1} = \mathbf{X}_{-1}\hat{\boldsymbol{\beta}}_{-1}$.

i. Show that the normal equations $(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}$ are equivalent to

$$\hat{\mathbf{Y}}_1 = \mathbf{P}_1(\mathbf{Y} - \hat{\mathbf{Y}}_{-1})$$
$$\hat{\mathbf{Y}}_{-1} = \mathbf{P}_{-1}(\mathbf{Y} - \hat{\mathbf{Y}}_1).$$

The normal equations

$$\begin{bmatrix} \mathbf{X}_1^T\mathbf{X}_1 & \mathbf{X}_1^T\mathbf{X}_{-1} \\ \mathbf{X}_{-1}^T\mathbf{X}_1 & \mathbf{X}_{-1}^T\mathbf{X}_{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_{-1} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^T\mathbf{Y}_1 \\ \mathbf{X}_{-1}^T\mathbf{Y}_{-1} \end{bmatrix}$$

are equivalent to

$$(\mathbf{X}_1^T\mathbf{X}_1)\hat{\boldsymbol{\beta}}_1 + (\mathbf{X}_1^T\mathbf{X}_{-1})\hat{\boldsymbol{\beta}}_{-1} = \mathbf{X}_1^T\mathbf{Y}$$
$$(\mathbf{X}_{-1}^T\mathbf{X}_1)\hat{\boldsymbol{\beta}}_1 + (\mathbf{X}_{-1}^T\mathbf{X}_{-1})\hat{\boldsymbol{\beta}}_{-1} = \mathbf{X}_{-1}^T\mathbf{Y}.$$

These equations are in turn equivalent to

$$\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1^T\mathbf{X}_1)^{-1}[\mathbf{X}_1^T\mathbf{Y} - \mathbf{X}_1^T\mathbf{X}_{-1}\hat{\boldsymbol{\beta}}_{-1}]$$
$$\hat{\boldsymbol{\beta}}_{-1} = (\mathbf{X}_{-1}^T\mathbf{X}_{-1})^{-1}[\mathbf{X}_{-1}^T\mathbf{Y} - \mathbf{X}_{-1}^T\mathbf{X}_1\hat{\boldsymbol{\beta}}_1].$$

From here we may write

$$\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 = \mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T[\mathbf{Y} - \mathbf{X}_{-1}\hat{\boldsymbol{\beta}}_{-1}]$$
$$\mathbf{X}_{-1}\hat{\boldsymbol{\beta}}_{-1} = \mathbf{X}_{-1}(\mathbf{X}_{-1}^T\mathbf{X}_{-1})^{-1}\mathbf{X}_{-1}^T[\mathbf{Y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1],$$

which is what we wished to show.

ii. Show that

$$\begin{pmatrix} \mathbf{I} & \mathbf{P}_1 \\ \mathbf{P}_{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{Y}}_1 \\ \hat{\mathbf{Y}}_{-1} \end{pmatrix} = \begin{pmatrix} \mathbf{P}_1\mathbf{Y} \\ \mathbf{P}_{-1}\mathbf{Y} \end{pmatrix}.$$

From what we proved in the previous part, we may write

$$\hat{\mathbf{Y}}_1 = \mathbf{P}_1\mathbf{Y} - \mathbf{P}_1\hat{\mathbf{Y}}_{-1}$$
$$\hat{\mathbf{Y}}_{-1} = \mathbf{P}_{-1}\mathbf{Y} - \mathbf{P}_{-1}\hat{\mathbf{Y}}_1,$$

giving

$$\hat{\mathbf{Y}}_1 + \mathbf{P}_1\hat{\mathbf{Y}}_{-1} = \mathbf{P}_1\mathbf{Y}$$
$$\hat{\mathbf{Y}}_{-1} + \mathbf{P}_{-1}\hat{\mathbf{Y}}_1 = \mathbf{P}_{-1}\mathbf{Y}.$$

Writing the left side as a matrix multiplication gives the claim.

(b) Show that $\hat{\mathbf{Y}}_1 = (\mathbf{I} - \mathbf{P}_1\mathbf{P}_{-1})^{-1}\mathbf{P}_1(\mathbf{I} - \mathbf{P}_{-1})\mathbf{Y}$.

Using the block-matrix inversion formula, what we proved in the previous part gives

$$\begin{aligned}\hat{\mathbf{Y}}_1 &= (\mathbf{I} - \mathbf{P}_1\mathbf{P}_{-1})^{-1}\mathbf{P}_1\mathbf{Y} - (\mathbf{I} - \mathbf{P}_1\mathbf{P}_{-1})^{-1}\mathbf{P}_1\mathbf{P}_{-1}\mathbf{Y} \\ &= (\mathbf{I} - \mathbf{P}_1\mathbf{P}_{-1})^{-1}\mathbf{P}_1(\mathbf{I} - \mathbf{P}_{-1})\mathbf{Y}.\end{aligned}$$

(c) The Gauss–Seidel or backfitting algorithm for finding $\hat{\mathbf{Y}}_1$ and $\hat{\mathbf{Y}}_{-1}$ is the following:

Initialize $\hat{\mathbf{Y}}_1 \leftarrow \mathbf{0}$ and $\hat{\mathbf{Y}}_{-1} \leftarrow \mathbf{0}$. Then repeat the steps

i. $\hat{\mathbf{Y}}_1 \leftarrow \mathbf{P}_1(\mathbf{Y} - \hat{\mathbf{Y}}_{-1})$
ii. $\hat{\mathbf{Y}}_{-1} \leftarrow \mathbf{P}_{-1}(\mathbf{Y} - \hat{\mathbf{Y}}_1)$

until $\hat{\mathbf{Y}}_1$ and $\hat{\mathbf{Y}}_{-1}$ do not change.

Show that in the $k$th iteration of the backfitting algorithm, we have

$$\hat{\mathbf{Y}}_1^{(k)} \leftarrow \left[\mathbf{I} - \textstyle\sum_{l=0}^{k-1}(\mathbf{P}_1\mathbf{P}_{-1})^l(\mathbf{I} - \mathbf{P}_1)\right]\mathbf{Y}.$$

We have

$$\begin{aligned}\hat{\mathbf{Y}}_1^{(1)} &\leftarrow \mathbf{P}_1(\mathbf{Y} - \mathbf{0}) \\ \hat{\mathbf{Y}}_{-1}^{(1)} &\leftarrow \mathbf{P}_1(\mathbf{Y} - \hat{\mathbf{Y}}_1^{(1)}) = \mathbf{P}_{-1}\mathbf{Y} - \mathbf{P}_{-1}\mathbf{P}_1\mathbf{Y}\end{aligned}$$

Then

$$\begin{aligned}\hat{\mathbf{Y}}_1^{(2)} &\leftarrow \mathbf{P}_1(\mathbf{Y} - \hat{\mathbf{Y}}^{(1)}) = \mathbf{P}_1\mathbf{Y} - \mathbf{P}_1\mathbf{P}_{-1}\mathbf{Y} + \mathbf{P}_1\mathbf{P}_{-1}\mathbf{P}_1\mathbf{Y} \\ \hat{\mathbf{Y}}_{-1}^{(2)} &\leftarrow \mathbf{P}_1(\mathbf{Y} - \hat{\mathbf{Y}}_1) = \mathbf{P}_{-1}\mathbf{Y} - \mathbf{P}_{-1}\mathbf{P}_1\mathbf{Y} + \mathbf{P}_{-1}\mathbf{P}_1\mathbf{P}_{-1}\mathbf{Y} - \mathbf{P}_{-1}\mathbf{P}_1\mathbf{P}_{-1}\mathbf{P}_1\mathbf{Y}\end{aligned}$$

The pattern continues, and we have

$$\begin{aligned}\hat{\mathbf{Y}}_1^{(3)} &= \mathbf{P}_1\mathbf{Y} - \mathbf{P}_1\mathbf{P}_{-1}\mathbf{P}_1\mathbf{Y} + \mathbf{P}_1\mathbf{P}_{-1}\mathbf{P}_1\mathbf{P}_{-1}\mathbf{Y} - \mathbf{P}_1\mathbf{P}_{-1}\mathbf{P}_1\mathbf{P}_{-1}\mathbf{P}_1\mathbf{Y} \\ &= \left[\mathbf{P}_1 - \sum_{l=1}^{2}(\mathbf{P}_1\mathbf{P}_{-1})^l(\mathbf{I} - \mathbf{P}_1)\right]\mathbf{Y} \\ &= \left[\mathbf{I} - \sum_{l=0}^{2}(\mathbf{P}_1\mathbf{P}_{-1})^l(\mathbf{I} - \mathbf{P}_1)\right]\mathbf{Y}.\end{aligned}$$

From this we see that we will have

$$\hat{\mathbf{Y}}_1^{(k)} = \left[\mathbf{I} - \sum_{l=0}^{k-1}(\mathbf{P}_1\mathbf{P}_{-1})^l(\mathbf{I} - \mathbf{P}_1)\right]\mathbf{Y}.$$

(d) Show that

$$\mathbf{I} - \sum_{l=0}^{\infty} (\mathbf{P}_1 \mathbf{P}_{-1})^l (\mathbf{I} - \mathbf{P}_1) = (\mathbf{I} - \mathbf{P}_1 \mathbf{P}_{-1})^{-1} \mathbf{P}_1 (\mathbf{I} - \mathbf{P}_{-1}),$$

in consequence of which $\hat{\mathbf{Y}}_1^{(k)} \to \hat{\mathbf{Y}}_1$ as $k \to \infty$. You will make use of the fact that for any real-valued square matrix $\mathbf{A}$, $\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \cdots = (\mathbf{I} - \mathbf{A})^{-1}$, provided $\lambda_{\max}(\mathbf{A}^T \mathbf{A}) < 1$, and you may assume $\lambda_{\max}(\mathbf{P}_1 \mathbf{P}_{-1} \mathbf{P}_1) < 1$.

We have

$$\begin{aligned}
\mathbf{I} - \sum_{l=0}^{\infty} (\mathbf{P}_1 \mathbf{P}_{-1})^l (\mathbf{I} - \mathbf{P}_1) &= \mathbf{I} - \sum_{l=0}^{\infty} (\mathbf{P}_1 \mathbf{P}_{-1})^l + \sum_{l=0}^{\infty} (\mathbf{P}_1 \mathbf{P}_{-1})^l \mathbf{P}_1 \\
&= -\sum_{l=1}^{\infty} (\mathbf{P}_1 \mathbf{P}_{-1})^l + \sum_{l=0}^{\infty} (\mathbf{P}_1 \mathbf{P}_{-1})^l \mathbf{P}_1 \\
&= -\sum_{l=0}^{\infty} (\mathbf{P}_1 \mathbf{P}_{-1})^l \mathbf{P}_1 \mathbf{P}_{-1} + \sum_{l=0}^{\infty} (\mathbf{P}_1 \mathbf{P}_{-1})^l \mathbf{P}_1 \\
&= \sum_{l=0}^{\infty} (\mathbf{P}_1 \mathbf{P}_{-1})^l (\mathbf{P}_1 - \mathbf{P}_1 \mathbf{P}_{-1}) \\
&= (\mathbf{I} - \mathbf{P}_1 \mathbf{P}_{-1})^{-1} \mathbf{P}_1 (\mathbf{I} - \mathbf{P}_{-1}),
\end{aligned}$$

since $\lambda_{\max}(\mathbf{P}_{-1} \mathbf{P}_1 \mathbf{P}_{-1}) = \lambda_{\max}(\mathbf{P}_1 \mathbf{P}_{-1} \mathbf{P}_1) < 1$. This is using the fact that $\mathbf{A}^T \mathbf{A}$ and $\mathbf{A} \mathbf{A}^T$ have the same nonzero eigenvalues (pg. 266 of Monahan [1]). This shows that the Gauss–Seidel, or backfitting algorithm, works.

It is worthwhile to consider the condition $\lambda_{\max}(\mathbf{P}_1 \mathbf{P}_{-1} \mathbf{P}_1) < 1$. This condition, we find, is satisfied if the matrix $\mathbf{X}$ has full-column rank. To see why, define

$$\rho = \sup \left\{ \frac{\mathbf{h}_1^T \mathbf{h}_2}{\|\mathbf{h}_1\|_2 \|\mathbf{h}_2\|_2}, 0 \neq \mathbf{h}_1 \in \mathcal{C}(\mathbf{X}_1), 0 \neq \mathbf{h}_2 \in \mathcal{C}(\mathbf{X}_{-1}) \right\}.$$

If $\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_{-1}]$ has full-column rank, then its columns are linearly independent, meaning that, given a linear combination $\mathbf{h}_1$ of the columns of $\mathbf{X}_1$, we cannot find a linear combination $\mathbf{h}_2$ of the columns of $\mathbf{X}_{-1}$ that is equal to $\mathbf{h}_1$. This gives $\rho < 1$. Using this fact, we write

$$\sup_{x \in \mathbb{R}^n} \frac{\mathbf{x}^T \mathbf{P}_1 \mathbf{P}_{-1} \mathbf{P}_1 \mathbf{x}}{\|\mathbf{P}_1 \mathbf{x}\|_2 \|\mathbf{P}_{-1} \mathbf{P}_1 \mathbf{x}\|_2} < 1 \implies \mathbf{x}^T \mathbf{P}_1 \mathbf{P}_{-1} \mathbf{P}_1 \mathbf{x} < \|\mathbf{P}_1 \mathbf{x}\|_2 \|\mathbf{P}_{-1} \mathbf{P}_1 \mathbf{x}\|_2 \quad \forall \mathbf{x} \in \mathbb{R}^n. \tag{3}$$

Since $\mathbf{P}_1$ is a projection matrix, $\|\mathbf{P}_1 \mathbf{x}\|_2 \leq \|\mathbf{x}\|_2$ for all $\mathbf{x}$, since

$$\sup_{\mathbf{x} \in \mathbb{R}^n} \frac{\|\mathbf{P}_1 \mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sup_{\mathbf{x} \in \mathbb{R}^n} \sqrt{\frac{\mathbf{x}^T \mathbf{P}_1 \mathbf{x}}{\|\mathbf{x}\|_2^2}} = \sqrt{\lambda_{\max}(\mathbf{P}_1)},$$

and the eigenvalues of a projection matrix are all in $\{0, 1\}$. Applying this also to $\mathbf{P}_{-1}$, we have

$$\|\mathbf{P}_1 \mathbf{x}\|_2 \|\mathbf{P}_{-1} \mathbf{P}_1 \mathbf{x}\|_2 \leq \|\mathbf{x}\|_2^2,$$

so that (3) gives

$$\lambda_{\max}(\mathbf{P}_1\mathbf{P}_{-1}\mathbf{P}_1) = \sup_{x\in\mathbb{R}^n} \frac{\mathbf{x}^T\mathbf{P}_1\mathbf{P}_{-1}\mathbf{P}_1\mathbf{x}}{\|\mathbf{x}\|_2^2} < 1.$$

# References


[1] John F Monahan. *A primer on linear models*. CRC Press, 2008.

[2] Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.