Sparse Estimation for Functional Semiparametric Additive Models

Authors: Peijun Sang, Richard A. Lockhart and Jiguo Cao
Presenter: Tianyu Guan

Simon Fraser University

February 27, 2019

Outline

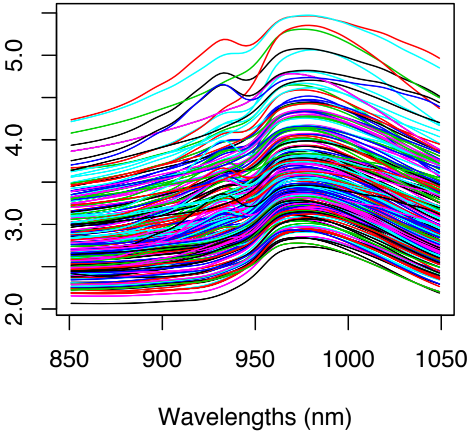
Tecator Data

- 240 meat samples
- 100-channel spectrum of absorbance with wavelength ranging from $850 - 1050$nm (851nm, 853nm, ..., 1049nm) - function
- the contents of moisture (water) - scalar
- the contents of fat - scalar
- the contents of protein- scalar

Objective

- Predict the content of protein

Tecator Data - spectrum of absorbance



Previous Studies

- Regress the content of protein on the functional spectral trajectories

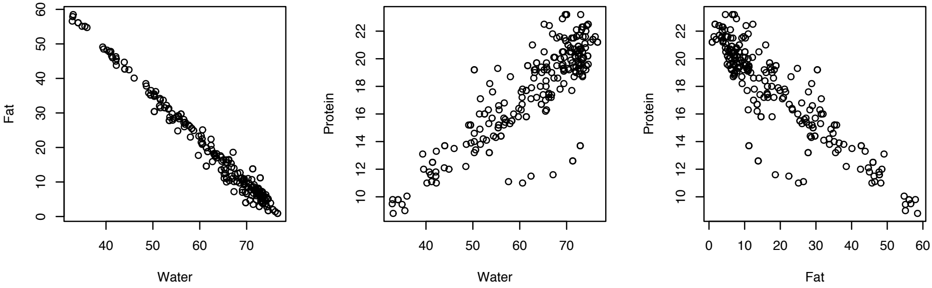
$$Y = b + \int X(t)\beta(t)\,\mathrm{d}\,t + \varepsilon$$

- Regress the content of protein on the scaled FPC scores of the functional spectral trajectories(a regularized functional additive model proposed by Zhu et al. (2014))

$$Y = b + \sum_{k=1}^{s} f_k(\zeta_k) + \varepsilon$$

- Y is the response (the content of protein)
- X is the functional predictor (the functional spectral trajectory)
- β is the unknown coefficient function
- $f_k(\cdot)$ are unknown smooth functions
- s is a sufficiently large number such that $f_k \equiv 0$ when $k > s$
- ζ_k is the scaled FPC score
- ε is the error

Tecator Data - pairewise scatter plots



- Each content is highly correlated with the other two contents
- Add the content of water(scalar) and fat(scalar) into the regression model

Functional Semiparametric Additive Model (FSAM)

- Predict a scalar response variable using both scalar and functional predictors
- Functional covariate is represented by its scaled leading FPC scores (non-parametric additive components)
- Scalar covariates are modeled linearly (parametric)

$$Y = b + \sum_{k=1}^{s} f_k(\zeta_k) + \mathbf{z}^T \boldsymbol{\alpha} + \varepsilon$$

- $\mathbf{z} = (z_1, \ldots, z_p)^T$ is a p-dimensinal scalar covariate (e.g. $p = 2$ in the Tector Data example. z_1 is the content of water and z_2 is the content of fat.)
- $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^T$ is the coefficient vector
- The authors develop a method (COSSO[1] penalty) for estimating the FSAM by smoothing and slecting non-vanishing components for the functional covariate

1. COSSO: component selection and smoothing operator

Models with effect of scalar predictors

$$Y = b_0 + \int X(t)\beta(t)\,\mathrm{d}\,t + \mathbf{z}^T\boldsymbol{\alpha} + \varepsilon \tag{1}$$

FPCS on X

- The covariance function of X is $G(s,t) = \sum_{k=1}^{\infty} \lambda_k \psi_k(s)\psi_k(t)$
- $\lambda_1, \lambda_2, \ldots$ are eigenvalues of G, $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$
- $\psi_1(t), \psi_2(t), \ldots$ are the corresponding orthonormal eigenfunctions
- $\int \psi_j \psi_k \,\mathrm{d}\,t = 1$ if $j = k$
- $\int \psi_j \psi_k \,\mathrm{d}\,t = 0$ if $j \neq k$
- $X(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_k \psi_k(t)$
- $\mu(t)$ is the mean function of X
- $\xi_k = \int (X(t) - \mu(t))\psi_k(t)\,\mathrm{d}\,t$ is the uncorrelated FPC score

Model (1) can be written as

$$Y = b + \sum_{k=1}^{\infty} \xi_k b_k + \mathbf{z}^T \boldsymbol{\alpha} + \varepsilon$$

- $b = b_0 + \int \mu(t)\beta(t)\,\mathrm{d}\,t$
- $b_k = \int \psi_k(t)\beta(t)\,\mathrm{d}\,t$

To allow for greater flexibility

$$Y = b + \sum_{k=1}^{\infty} f_k(\xi_k) + \mathbf{z}^T \boldsymbol{\alpha} + \varepsilon \tag{2}$$

- $f_k(\cdot)$ are unknown smooth functions

Standardization of ξ_k

$$\zeta_k = \Phi(\lambda_k^{-1/2}\xi_k)$$

- $Var(\xi_k) = \lambda_k$
- $\Phi(\cdot)$ is a continuously differentiable map from R to $[0, 1]$
- A wide range of cumulative distribution functions (CDFs) can be used as $\Phi(\cdot)$
- $\Phi(\cdot)$ is the CDF of $N(0, 1)$ in the paper

Advantages

- ζ_k have similar or the same variations
- $\zeta_k \in [0, 1]$ is convenient to do the model regularization
- When the distribution of ξ is close to Gaussian, ζ is approximately uniform in $[0, 1]$, which is convenient for nonparametric modeling on the effect of ζ

Model (2) can be written as

$$Y = b + \sum_{k=1}^{\infty} f_k(\zeta_k) + \mathbf{z}^T \boldsymbol{\alpha} + \varepsilon$$

The truncated model is

$$Y = b + \sum_{k=1}^{s} f_k(\zeta_k) + \mathbf{z}^T \boldsymbol{\alpha} + \varepsilon \tag{3}$$

- s is large enough that $f_k \equiv 0$ when $k > s$

For elements f, g, h

Operation of addition

- $f + g = g + f$
- $(f + g) + h = f + (g + h)$
- for any two elements f and g, there exists an element h such that $f + h = g$

Operation of scalar multiplication

- $\alpha(f + g) = \alpha f + \alpha g$
- $(\alpha + \beta)f = \alpha f + \beta f$
- $1f = f$ and $0f = 0$
- α and β are real numbers

Linear space

A set $\mathcal{L}$ of such elements form a linear space

- If f, $g \in \mathcal{L}$ implies that $f + g \in \mathcal{L}$ and $\alpha f \in \mathcal{L}$

Functional

A functional in a linear space $\mathcal{L}$ operates on an element $f \in \mathcal{L}$ and returns a real number as its value

Linear functional

A linear functional $L \in \mathcal{L}$ satisfies

- $L(f + g) = Lf + Lg$
- $L(\alpha f) = \alpha Lf, f, g \in \mathcal{L}$, α is real

Bilinear form

A bilinear form $J(f, g)$ in a linear space $\mathcal{L}$ and takes $f, g \in \mathcal{L}$ as arguments and returns a real value and satisfies

- $J(\alpha f + \beta g, h) = \alpha J(f, h) + \beta J(g, h)$
- $J(f, \alpha g + \beta h) = \alpha J(f, g) + \beta J(f, h)$
- $J(\cdot, \cdot)$ is symmetric if $J(f, g) = J(g, f)$
- A symmetric bilinear form is non-negative definite if $J(f, f) \geq 0$ $\forall f \in \mathcal{L}$
- A symmetric bilinear form is positive definite if $J(f, f) > 0$ $\forall f \in \mathcal{L}$

Inner product

- A linear space is often equipped with an inner product
- An inner product is a positive definite bilinear form with a notation $(\cdot,\cdot)$
- An inner product defines a norm in the linear space, $\|f\| = \sqrt{(f,f)}$
- A norm measures the distance between elements in the space $\|f - g\|$
- The Cauchy-Schwarz inequality $|(f,g)| \leq \|f\|\|g\|$ hold in such a linear space
- The triangle inequality $\|f + g\| \leq \|f\| + \|g\|$ hold in such a linear space
- A sequence satisfying $\lim_{n,m\to\infty}\|f_n - f_m\| = 0$ is called a Cauchy sequence
- A linear space is $\mathcal{L}$ is complete if every Cauchy sequence in $\mathcal{L}$ converges to an element in $\mathcal{L}$

Hilbert space

A Hilbert space $\mathcal{H}$ is a complete inner product linear space

- For every g in a Hilbert space $\mathcal{H}$, $L_g f = (g, f)$ defines a continuous linear functional L_g
- Conversely, every continuous linear functional L in $\mathcal{H}$ has a representation $Lf = (g_L, f)$ for some $g_L \in \mathcal{H}$, called the representer of L

Riesz representation

Theorem: For every continuous linear functional L in a Hilbert space $\mathcal{H}$, there exists a unique $g_L \in \mathcal{H}$ such that $Lf = (g_L, f), \forall f \in \mathcal{H}$

- A linear functional L is continuous if $\lim_{n\to\infty} Lf_n = Lf$ whenever $\lim_{n\to\infty} f_n = f$

Evaluation functional

- Define the evaluation functional as $[x](\cdot)$
- $[x]f = f(x)$

Reproducing kernel Hilbert space (RKHS)

Consider a Hilbert space $\mathcal{H}$ of functions on domain $\mathcal{X}$. If the evaluation functional $[x]f = f(x)$ is continuous in $\mathcal{H}$, $\forall x \in \mathcal{X}$, then $\mathcal{H}$ is called a reproducing kernel Hilbert space

- By the Riesz representation theorem, for the evaluation functional $[x]f = f(x)$, there exists $R_x \in \mathcal{H}$, the representer, such that $[x]f = f(x) = (R_x, f)$, $\forall f \in \mathcal{H}$

Reproducing kernel Hilbert space (RKHS)

Consider a Hilbert space $\mathcal{H}$ of functions on domain $\mathcal{X}$. If the evaluation functional $[x]f = f(x)$ is continuous in $\mathcal{H}$, $\forall x \in \mathcal{X}$, then $\mathcal{H}$ is called a reproducing kernel Hilbert space

- By the Riesz representation theorem, for the evaluation functional $[x]f = f(x)$, there exists $R_x \in \mathcal{H}$, the representer, such that $[x]f = f(x) = (R_x, f)$, $\forall f \in \mathcal{H}$
- $(R_x, R_y) = (R_y, R_x) = R_x(y) = R_y(x)$
- The function $R(x,y) = R_x(y) = (R_x, R_y)$ is symmetric bivariate
- $R(x,y)$ has the reproducing property $(R(x,\cdot), f(\cdot)) = f(x)$ and is called the reproducing kernel
- the reproducing kernel is unique when it exists

Example: lth-order Soblev space on [0,1] $\mathcal{H}$

Definition: $\mathcal{H}$ is a collection of functioins on $[0,1]$ whose first $(l-1)$th derivatives are absolutely continuous and the lth derivative belongs to $L^2[0,1]$

- $L^2[0,1]$ is a Hilbert space which is a collection of all square integrable functioins on $[0,1]$
- $\mathcal{H}$ is a reproducing kernel Hilbert space equipped with

$$(f,g) = \sum_{\nu=0}^{l-1}\left(\int_0^1 f^{(\nu)}\,\mathrm{d}x\right)\left(\int_0^1 g^{(\nu)}\,\mathrm{d}x\right) + \int_0^1 f^{(l)} g^{(l)}\,\mathrm{d}x$$

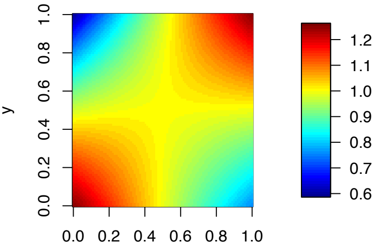
Example: lth-order Soblev space on [0,1] $\mathcal{H}$

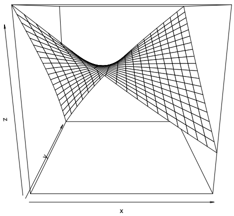
For $l = 2$

$$R(x, y) = 1 + k_1(x)k_1(y) + k_2(x)k_2(y) - k_4(x - y)$$

- $k_1(x) = x - 0.5$
- $k_2(x) = \frac{1}{2}\left(k_1^2(x) - \frac{1}{12}\right)$
- $k_4(x) = \frac{1}{24}\left(k_1^4(x) - \frac{k_1^2(x)}{2} + \frac{7}{240}\right)$

The paper focuses on the second order Soblev space with $l = 2$





Left: contour plot of $R(x, y)$. Right: 3d plot of $R(x, y)$

The model is

$$Y = b + \sum_{k=1}^{s} f_k(\zeta_k) + \mathbf{z}^T \boldsymbol{\alpha} + \varepsilon \tag{4}$$

- The regression function $f(\zeta) = b + \sum_{k=1}^{s} f_k(\zeta_k)$
- $f_k \in \bar{H}$, $k = 1, \ldots, s$
- $f(\zeta)$ lies in the truncated subspace $\mathcal{F}^s = 1 \bigoplus \sum_{k=1}^{s} \bar{H}$
- $\mathcal{F}^s = 1 \bigoplus \sum_{k=1}^{s} \bar{H}$ is direct sum of the space of constant and s copies of $\bar{H}$

Consider the model

$$Y = b + \sum_{k=1}^{s} f_k(\zeta_k) + \varepsilon$$

for now. The loss function with the Component Selection and Smoothing Operator (COSSO) is defined as

$$Q(f) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - f(\zeta_i)\}^2 + \tau^2 J(f) \tag{5}$$

- $J(f) = \sum_{k=1}^{s} \|\mathcal{P}^k f\|$ is the COSSO penalty
- $\mathcal{P}^k f$ denotes the projection of f onto $\bar{H}$ with the argument being the kth component of ζ, i.e. f_k

An equivalent reformulation

Minimizing

$$Q(f) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - f(\zeta_i)\}^2 + \tau^2 J(f) \qquad (6)$$

is equivalent to minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i - f(\zeta_i)\}^2 + \lambda_0 \sum_{k=1}^{s} \theta_k^{-1}\|\mathcal{P}^k f\|^2 + \lambda \sum_{k=1}^{s} \theta_k \qquad (7)$$

- with respect to f and $\boldsymbol{\theta}$
- subject to $\theta_k > 0$, $k = 1, \ldots, s$

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i - f(\zeta_i)\}^2 + \lambda_0 \sum_{k=1}^{s}\theta_k^{-1}\|\mathcal{P}^k f\|^2 + \lambda \sum_{k=1}^{s}\theta_k$$

- From the spline literature[2] the minimizer of the above loss function has the form $f(\zeta) = b + \sum_{k=1}^{s}\theta_k \sum_{i=1}^{n} c_i R(\hat{\zeta}_{ik}, \zeta_k)$
- $R(\cdot,\cdot)$ is the reproducing kernel of $\bar{H}$
- $f_k(\zeta_k) = \sum_{i=1}^{n} c_i \theta_k R(\hat{\zeta}_{ik}, \zeta_k)$, $\hat{\zeta}_{ik}$ is the estimated scaled FPC scores
- $\mathbf{c} = (c_1, \ldots, c_n)^T$ is a vector of unknown parameters
- $\sum_{k=1}^{s}\theta_k^{-1}\|\mathcal{P}^k f\|^2 = \sum_{k=1}^{s}\theta_k \mathbf{c}^T R_k \mathbf{c} = \mathbf{c}^T R_\theta \mathbf{c}$
- R_k denote the $n \times n$ matrix with the (j,l) entry $R(\hat{\zeta}_{jk}, \hat{\zeta}_{lk})$
- $R_\theta = \sum_{k=1}^{s}\theta_k R_k$

2. e.g. Chapter 10 (Additive and Interaction Splines) of Wahba, Grace. Spline models for observational data written by Grace Wahba in 1990

- $f_k(\zeta_k) = \sum_{i=1}^{n} c_i \theta_k R(\hat{\zeta}_{ik}, \zeta_k)$

$$\begin{aligned}
\|\mathcal{P}^k f\|^2 = \|f_k\|^2 &= \left(\sum_{i=1}^{n} c_i \theta_k R(\hat{\zeta}_{ik}, \zeta_k), \sum_{j=1}^{n} c_j \theta_k R(\hat{\zeta}_{jk}, \zeta_k) \right) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \theta_k^2 c_i c_j \left(R(\zeta_k, \hat{\zeta}_{ik}), R(\zeta_k, \hat{\zeta}_{jk}) \right) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \theta_k^2 c_i c_j R(\hat{\zeta}_{ik}, \hat{\zeta}_{jk}) \\
&= \theta_k^2 \mathbf{c}^T R_k \mathbf{c}
\end{aligned}$$

$$\sum_{k=1}^{s} \theta_k^{-1} \|\mathcal{P}^k f\|^2 = \mathbf{c}^T \left(\sum_{k=1}^{s} \theta_k R_k \right) \mathbf{c} = \mathbf{c}^T R_\theta \mathbf{c}$$

Using the estimated scaled FPC scores, the loss function is

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i - f(\hat{\zeta}_i)\}^2 + \lambda_0 \sum_{k=1}^{s} \theta_k^{-1} \|\mathcal{P}^k f\|^2 + \lambda \sum_{k=1}^{s} \theta_k$$

- $f_k(\zeta_k) = \sum_{j=1}^{n} c_j \theta_k R(\hat{\zeta}_{jk}, \zeta_k)$ and $f_k(\hat{\zeta}_{ik}) = \sum_{j=1}^{n} c_j \theta_k R(\hat{\zeta}_{jk}, \hat{\zeta}_{ik})$
- $f(\hat{\zeta}_{ik}) = b + \sum_{k=1}^{s}\sum_{j=1}^{n} c_j \theta_k R(\hat{\zeta}_{jk}, \hat{\zeta}_{ik})$

It can be written as

$$\|\mathbf{Y} - \mathbf{1}_n b - R_\theta \mathbf{c}\|_E^2 + \lambda_0 \mathbf{c}^T R_\theta \mathbf{c} + \lambda \mathbf{1}_n^T \boldsymbol{\theta}$$

- $\|\cdot\|_E$ represents the Euclidean norm
- $\mathbf{1}_n$ is the vector of ones of length n
- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s)$

To minimize

$$\|\mathbf{Y} - \mathbf{1}_n b - R_\theta \mathbf{c}\|_E^2 + \lambda_0 \mathbf{c}^T R_\theta \mathbf{c} + \lambda \mathbf{1}_n^T \boldsymbol{\theta} \tag{8}$$

We alternatively solve for $(b, \mathbf{c})$ with $\boldsymbol{\theta}$ fixed and then solve for $\boldsymbol{\theta}$ with $(b, \mathbf{c})$ fixed

- When θ is fixed, solving (8) is equivalent to solving

$$\min_{b,\mathbf{c}} \|\mathbf{Y} - \mathbf{1}_n b - R_\theta \mathbf{c}\|_E^2 + \lambda_0 \mathbf{c}^T R_\theta \mathbf{c} \tag{9}$$

- When $(b, \mathbf{c})$ is fixed (8) becomes

$$\min_{\boldsymbol{\theta} \geq \mathbf{0}} (\mathbf{v} - G\boldsymbol{\theta})^T (\mathbf{v} - G\boldsymbol{\theta}) + n\lambda \mathbf{1}_s \boldsymbol{\theta} \tag{10}$$

  - $\mathbf{v} = \mathbf{y} - (1/2) n\lambda_0 \mathbf{c} - \mathbf{1}_n b$
  - G is $n \times s$ matrix with the kth column being $R_k \mathbf{c}$

- We consider an equivalent optimization problem: for some $M \geq 0$, find

$$\min_{\boldsymbol{\theta}} (\mathbf{v} - G\boldsymbol{\theta})^T (\mathbf{v} - G\boldsymbol{\theta}) \quad \text{subject to } \mathbf{1}_s^T \boldsymbol{\theta} \leq M \text{ and } \boldsymbol{\theta} \geq \mathbf{0}_s \tag{11}$$

Algorithm 1 Iterative updating for regularized functional semiparametric additive model

Step 1: Start with an initial value of $\boldsymbol{\alpha}$, say $\hat{\boldsymbol{\alpha}}^{(0)}$, and an initial value of $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}}^{(0)}$.

Step 2: Use the current estimate $\hat{\boldsymbol{\alpha}}^{(m)}$ and $\hat{\boldsymbol{\theta}}^{(m)}$ to obtain estimates $\hat{b}^{(m+1)}$ and $\hat{\mathbf{c}}^{(m+1)}$ by solving (9), in which $\boldsymbol{y}$ is replaced by $\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\alpha}}^{(m)}$.

Step 3: Use the current estimate $\hat{\boldsymbol{\alpha}}^{(m)}$, $\hat{b}^{(m+1)}$ and $\hat{\mathbf{c}}^{(m+1)}$ to obtain an updated estimate $\hat{\boldsymbol{\theta}}^{(m+1)}$ by solving (11), in which $\boldsymbol{v}$ is replaced by $\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\alpha}}^{(m)} - (1/2)n\lambda_0\hat{\mathbf{c}}^{(m+1)} - \mathbf{1}_n\hat{b}^{(m+1)}$.

Step 4: Use the estimate $\hat{b}^{(m+1)}$, $\hat{\mathbf{c}}^{(m+1)}$ and $\hat{\boldsymbol{\theta}}^{(m+1)}$ to obtain an updated estimate $\hat{\boldsymbol{\alpha}}^{(m+1)}$ by solving a least squares problem.

Step 5: Repeat Steps 2, 3 and 4 until $||\hat{\boldsymbol{\alpha}}^{(m+1)} - \hat{\boldsymbol{\alpha}}^{(m)}|| < \epsilon$, where ϵ is a pre-determined tolerance value.

Tuning parameter selection

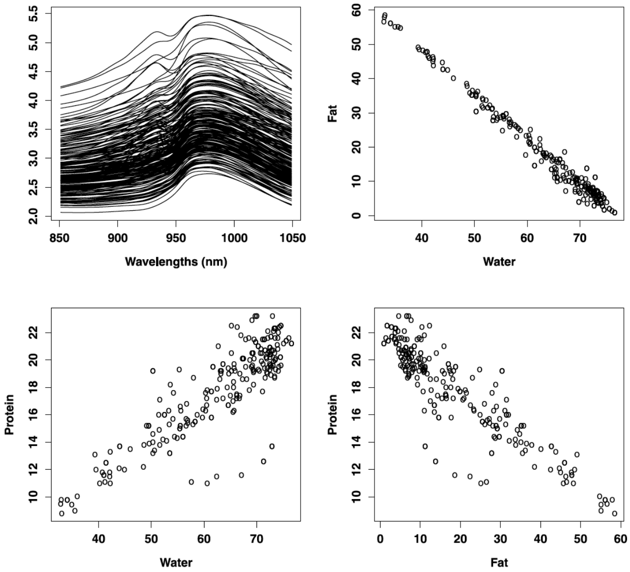
- In step 1 in Algorithm 1, the initial value of $\boldsymbol{\theta}$ is chosen as $\mathbf{1}_s$
- Cross validation (or GCV) is used to choose tuning parameter λ_0 when solving for $(b, \mathbf{c})$ with $\boldsymbol{\theta}$ fixed
- When solving for $\boldsymbol{\theta}$ with $(b, \mathbf{c})$ fixed, we use the chosen value of λ_0 and use CV to tune M

Tecator Data

- 240 meat samples
- 100-channel spectrum of absorbance with wavelength ranging from $850-1050$nm (851nm, 853nm, ..., 1049nm) - function
- the contents of moisture (water) - scalar
- the contents of fat - scalar
- the contents of protein- scalar

Objective

- Predict the content of protein using spectral trajectories, the content of water and the content of fat



2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
850
900
950
1000
1050
Wavelengths (nm)
Fat
0
10
20
30
40
50
60
40
50
60
70
Water
Protein
10
12
14
16
18
20
22
40
50
60
70
Water
Protein
10
12
14
16
18
20
22
0
10
20
30
40
50
60
Fat

Table 4
Summary of prediction error and proportion of variance explained on the test set of each model. FAM represents the functional additive model [24] where only the $\hat{\zeta}_i$s are considered as explanatory variables. MARS_0 denotes the MARS model considering only the $\hat{\zeta}_i$s as explanatory variables while neglecting the effect of the fat content. $d = 10$ and $d = 20$ indicate that 10 and 20 leading FPCs are initially retained, respectively.

	$d = 20$					
	FSAM-COSSO	CSEFAM	FSAM-GAMS	FAM	MARS	MARS_0
MSPE	0.52	0.71	0.84	0.73	0.83	1.18
R^2	0.97	0.96	0.95	0.96	0.95	0.93
	$d = 10$					
	FSAM-COSSO	CSEFAM	FSAM-GAMS	FAM	MARS	MARS_0
MSPE	0.92	1.99	1.35	1.42	0.97	1.01
R^2	0.95	0.88	0.92	0.92	0.94	0.94

- MSPE is the mean squared prediction error
- R^2 is the quasi-R^2 defined by

$$R^2 = 1 - \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 / \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

- Sang, Peijun, Richard A. Lockhart, and Jiguo Cao. Sparse estimation for functional semiparametric additive models. *Journal of Multivariate Analysis 168 (2018): 105-118*.
- Wong, Raymond KW, Yehua Li, and Zhengyuan Zhu. Partially linear functional additive models for multivariate functional data. *Journal of the American Statistical Association (2018): 1-13*.
- Zhu, Hongxiao, Fang Yao, and Hao Helen Zhang. Structured functional additive regression in reproducing kernel Hilbert spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76, no. 3 (2014): 581-603*.
- Wahba, Grace. Spline models for observational data. Vol. 59. Siam, 1990.

Thank You !