

Modeling Aggregated Functional Data

Ronaldo Dias¹

¹IMECC/UNICAMP

Conference on Nonparametric Statistics for Big Data, and
Celebration to Honor Professor

Grace Wahba

Type of Consumers:



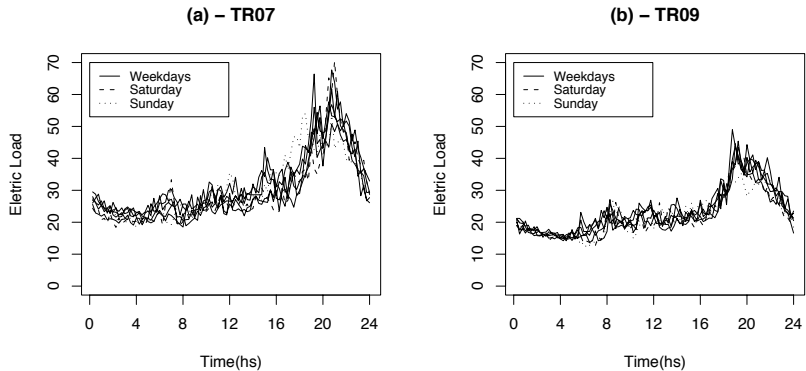


Figure: Observed curves for TR07 and TR09.

Aggregated data

- Sample J days of the electric load of M transformers (trafos).
- Although the load for each trafo is a continuous curve, it is observed only at T points ($t_1 < t_2 < \dots < t_T$).
- The electric load of trafo i is composed by the sum of $N_i = N_{1,i} + \dots + N_{C,i}$ curves where $N_{i,c}$ is the number of consumers type c .
- The market (the number of consumers of each type) of each trafo is known.

Market of trafo i : $(N_{1,i}, N_{2,i}, \dots, N_{C,i})$

Total load $Y_{i,j}(t)$ for the i th trafo at time t of the j th day

$$Y_{i,j}(t) = \sum_{c=1}^C \sum_{n_c=1}^{N_{c,i}} W_{c,j,n_c,i}(t),$$

$$t \in [0, 24], \quad i = 1, \dots, M, \quad j = 1, \dots, J$$

$W_{c,j,n_c,i}(t)$: load of the n_c th consumer of type c , at day j for trafo i .

Nonparametric regression model

typology (mean curve): $\alpha_c(t)$ for consumer type c

$$W_{c,j,n_c,i}(t) = \alpha_c(t) + \varepsilon_{c,j,n_c,i}(t),$$

$\varepsilon_{c,j,n_c,i}(t)$: mean zero Gaussian random process independent and identically distributed for fixed c .

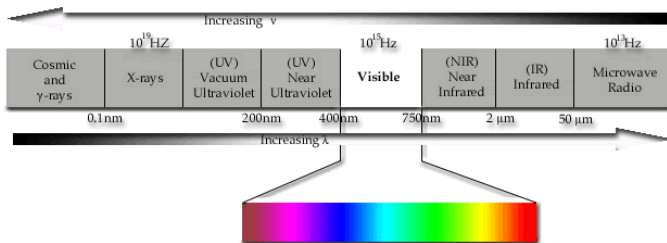
Therefore,

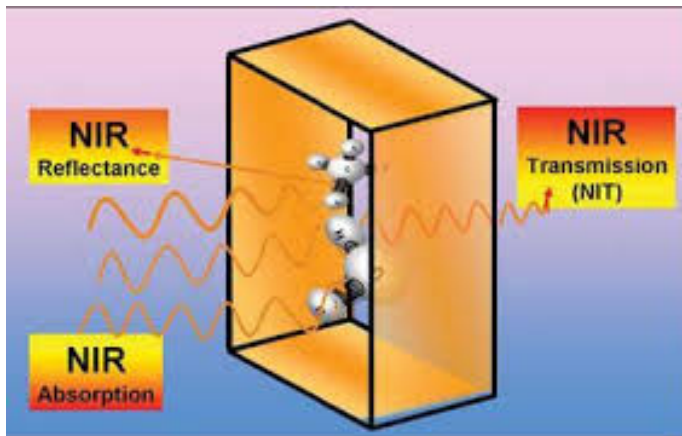
$$Y_{i,j}(t) = \sum_{c=1}^C N_{c,i} \alpha_c(t) + \varepsilon_{i,j}(t),$$

where

$$\varepsilon_{i,j}(t) = \sum_{c=1}^C \sum_{n_c=1}^{N_{c,i}} \varepsilon_{c,j,n_c,i}(t).$$

Wavelength in nanometers:





- When a substance is submitted to different wavelengths, the overtones and combinations in the NIR band will produce very complex patterns that characterize the constituents of the sample.
- The different characteristics of the constituents of the sample overlap and give rise to a curve which is a sum of several curves depending on the **concentration of each constituent**.

Calibration problem

- There are several techniques to measure the constituents of a sample.
- Some of them are very precise and expensive (analytical techniques).
- Some of them are not so precise but inexpensive (e.g. NIR spectroscopy).
- To use inexpensive techniques it is necessary to **calibrate the instrument**.
- For practical reasons it is only possible to measure the spectral data at a finite number of wavelengths $t_1 < t_2 < \dots < t_T$. Often T is in the range 100–200 or even more.

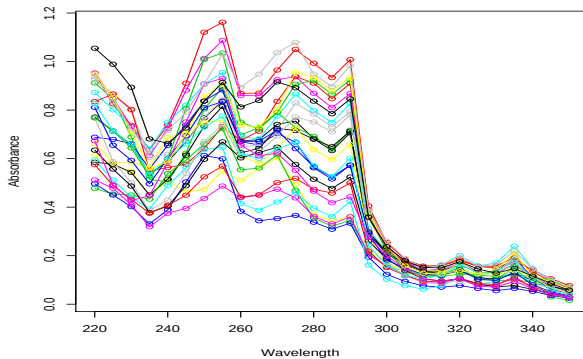
Prediction problem

- After calibrating the inexpensive procedure
- Take new measurements
- Estimate the concentrations of the constituents of the new sample.

Polycyclic aromatic hydrocarbons(PAH)

- EAS (Electronic Absorption Spectroscopy)
- 25 chemical samples
- Each sample with 10 different chemical elements (constituents)
- 27 wavelengths (220nm–350nm)

PAH Dataset



- A chemical sample is a compound of several constituents.
- Each constituent of interest is called an analyte. The amount of the analyte ℓ is called y_ℓ , $\ell = 1, \dots, m$.
- A sample is called closed if $y_1 + \dots + y_m = 1$, i.e. when all constituents in the compound are analyzed.
- The absorbance of a given analyte ℓ at a particular wavelength t will be denoted by $a_\ell(t)$.
- We assume that there are n samples of varying compositions.
- Y dataset: concentrations measured by a reference method ($y_{\ell,i}$ for $i = 1, \dots, n$ and $\ell = 1, \dots, m$).
- **Spectral data** refers to the absorbances $W(t_1), \dots, W(t_T)$ measured by NIR instrument at T wavelengths $t_1 < t_2 < \dots < t_T$.

The model: Beer-Lambert law

- For the i th closed sample composed of m constituents and T wavelengths is given by

$$W_i(t_j) = \sum_{\ell=1}^K y_{i,\ell} a_{\ell}(t_j) + \epsilon_i(t_j), \text{ for } j = 1, \dots, T$$

- Usually only a subset of constituents is considered. In this case, we do not have the restriction $y_{i,1} + \dots + y_{i,m} = 1$ but a modification of the Beer-Lambert law is used:

$$W_i(t_j) = \alpha_j + \sum_{\ell=1}^K y_{i,\ell} a_{\ell}(t_j) + \epsilon_i(t_j), \text{ for } j = 1, \dots, T$$

Calibration Sample

$$W_i(t_j) = \alpha_j + \sum_{\ell=1}^K y_{i,\ell} \alpha_{\ell}(t_j) + \epsilon_i(t_j), \text{ for } j = 1, \dots, T$$

Prediction sample

$$W_i^*(t_j) = \alpha_j + \sum_{\ell=1}^K y_{i,\ell}^* \alpha_{\ell}(t_j) + \epsilon_i(t_j), \text{ for } j = 1, \dots, T$$

Calibration problem Given $y_{i,\ell}$ and $W_i(t_j)$ estimate $\alpha_{\ell}(t_j)$.

Prediction problem Given a new dataset $W_i^*(t_j)$, estimate $y_{i,\ell}^*$.

Functional data

To consider the Beer-Lambert law to be a discretization of the true model.

$$W(t) = \alpha(t) + \sum_{\ell=1}^K y_{\ell} a_{\ell}(t) + \epsilon(t),$$

for all $t \in [A, B]$ where $\epsilon(t)$ is a Gaussian process with covariance function given by $\sigma(s, t) = \text{Cov}(\epsilon(s), \epsilon(t))$.

Covariance structure

- We assume that the correlation between points $Y_{ij}(t)$ and $Y_{ij}(s)$ decays exponentially, as $|t - s|$ increases.
- For each analyte c , we assign an exponential correlation function with decay parameter $\phi > 0$ for the Gaussian process associated to each $\epsilon_{ic}(\cdot)$.
- For the i -th curve, let

$$\begin{aligned} Z_i(t, s) &= \text{Cov}(\epsilon_i(t), \epsilon_i(s)) \\ &= \sum_{c=1}^C y_{ic} \sigma^2 \exp(-\phi |t - s|) \end{aligned}$$

Prior specification for $\Theta = (\beta, \sigma^2, \phi)$

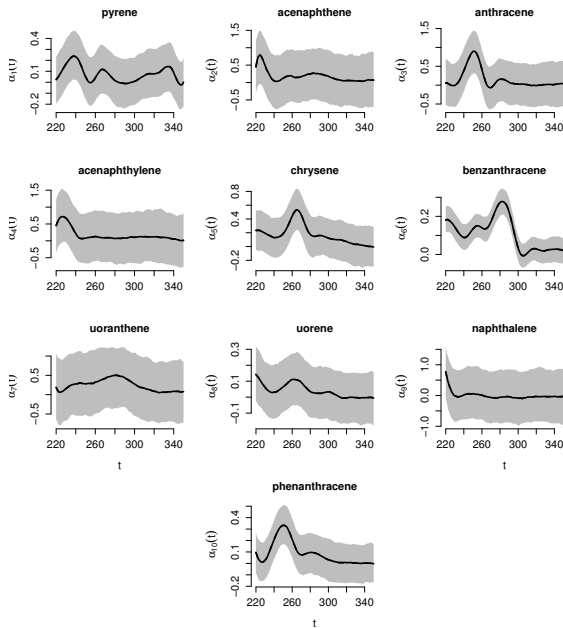
- Prior independence among the components of Θ
- For β_c : K -dimensional multivariate normal distributions with known mean vector $\mathbf{b}_c$ and diagonal covariance matrices Ω_c , $c = 1, 2, \dots, C$, with elements in the diagonal fixed at some large value.
- For σ^2 we assume an inverse gamma prior distribution with shape parameter d and rate parameter l .
- For ϕ , we assign a gamma prior distribution with shape parameter p and rate parameter q fixed at some reasonable value. For example, we can use the idea of practical range. The mean of the prior, p/q can be fixed such that at a reasonable distance, the correlation is close to zero, say 0.05. More specifically, we fix the mean at the value that solves $0.05 = \exp(-\phi^* \text{dist})$, where ϕ^* is the prior mean guess we need, and dist is a fixed distance.

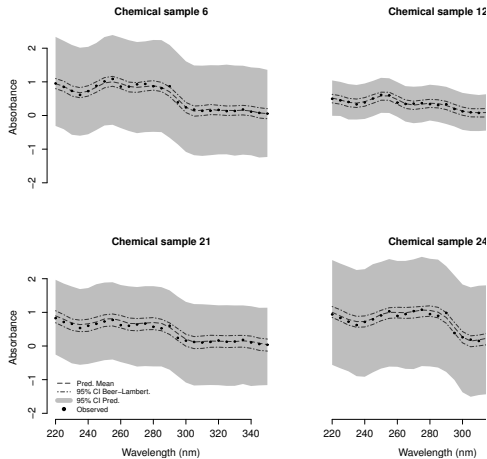
Posterior distribution for $\Theta = (\beta, \sigma^2, \phi)$

- The full conditional posterior distributions of β_c are normal distributions, which can be easily sampled.
- For σ^2 and ϕ we make use of the Metropolis-Hastings algorithm with log-normal proposals based on the current value of the chain, and some fixed variance, tuned to give reasonable acceptance rates.
- The MCMC algorithm was implemented in `ox`.

- 25 chemical samples
- 10 different constituents (pyrene, acenaphthene, anthracene, acenaphthylene, chrysene, benzantracene, uoranthene, uorene, naphthalene, phenanthracene)
- 27 wavelengths (220nm–350nm)
- 14 B-spline basis with 10 interior knots equally spaced in the interval (220,350) to estimate the latent absorbance curves of each constituent, $\alpha_c(\cdot)$.

- gamma prior for ϕ
- For σ^2 we assigned an inverse gamma prior distribution with parameters 2 and 0.2.
- We let the MCMC algorithm run for 100,000 iterations, considered the first 5,000 as burn-in and kept every 100-th sample to avoid auto-correlation between the sampled values.





- Calibration sample

$$W_i(t_j) = \alpha_j + \sum_{\ell=1}^K y_{i,\ell} \alpha_{\ell}(t_j) + \epsilon_i(t_j), \text{ for } j = 1, \dots, T$$

- Prediction sample

$$W_i^*(t_j) = \alpha_j + \sum_{\ell=1}^K y_{i,\ell}^* \alpha_{\ell}(t_j) + \epsilon_i(t_j), \text{ for } j = 1, \dots, T$$

- Consider as data:

- 1 $W_i(t_j)$ for $i = 1, \dots, I, j = 1, \dots, T$

- 2 $y_{i,\ell}$ for $i = 1, \dots, I$, and $\ell = 1, \dots, C$

- 3 $W_i^*(t_j)$ for $i = 1, \dots, I^*, j = 1, \dots, T$

- Consider as parameters:

- 1 The β such that $\alpha_c(t) = \sum_{k=1}^K \beta_{c,k} B_k(t)$,

- 2 σ^2 and ϕ parameters of the covariance

- 3 $y_{i,\ell}^*$ for $i = 1, \dots, I^*$, and $\ell = 1, \dots, C$

Calibration Procedure

Smoothing Splines Let $\mathbf{X} = \mathbf{Y} \otimes \mathbf{B}$. Minimize w.r.t. β

$$(\mathbf{W} - \mathbf{X}\beta)' (\mathbf{W} - \mathbf{X}\beta) + \lambda\beta' (\mathbf{I}_{m \times m} \otimes \mathbf{R}) \beta,$$

therefore

$$\hat{\beta} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_{m \times m} \otimes \mathbf{R})^{-1} \mathbf{X}'\mathbf{W},$$

where $\mathbf{R}$ is a matrix with entries

$$\mathbf{R}_{i,j} = \int_{-\infty}^{\infty} D^2 B_i(t) D^2 B_j(t) dt,$$

and D^2 is the second order differential operator $\partial^2/\partial t^2$. The degree of smoothness will be controlled either by inspection of the curves plots or by minimization of the generalized cross-validation criteria (Wahba, 1990).

Prediction

- Once we have the estimated analyte absorbance curves $\hat{\theta}_\ell(t)$, $\ell = 0, \dots, m$, we can proceed to perform *prediction*.
- a new set of J spectra measured by the same NIR instrument, $W_j^*(t)$, $j = 1, \dots, J$, all of them observed at the wavelengths $t = t_1, \dots, t_T$ (this assumption can be relaxed as long as the new wavelengths are within $[t_1, t_T]$).
- Using this new set of data, we want to predict now the new concentrations $y_{j,\ell}^*$ using the Beer-Lambert relationship

$$W_j^*(t) = \sum_{k=1}^K \left[\beta_{0,k} + \sum_{\ell=1}^m \beta_{\ell,k} y_{j,\ell}^* \right] B_k(t) + \epsilon_j(t).$$

Prediction

- A naïve way to obtain estimates for $y_{j,\ell}^*$, $\ell = 1, \dots, m$, $j = 1, \dots, J$ is to use the estimated spectra $\hat{\theta}_\ell$ from the calibration sample, and find which set of $\mathbf{y}_j^*$ minimizes the square error of prediction

$$\hat{\mathbf{y}}_j^* = \arg \min_{\mathbf{y}_j^*} \sum_{n=1}^T \sum_{\ell=1}^m \left(w_j^*(t_n) - y_{j,\ell}^* \hat{\theta}_\ell(t_n) \right)^2.$$

Prediction

- To estimate the variance of the $\hat{\mathbf{y}}_j^*$, we use a leave-one-out jackknife technique.
 - 1 For $i = 1, \dots, I$, leave out all data related to $W_i(t)$ and find the estimate $\theta_\ell^{(-i)}(t)$. Then,
 - 2 Use $\hat{\theta}_\ell^{(-i)}(t)$ to “estimate” $y_{i,\ell}$ as $\hat{y}_{i,\ell}^{(-i)}$.
 - 3 Compare $y_{i,\ell}$ with $\hat{y}_{i,\ell}^{(-i)}$ using

$$S_\ell^2 = \frac{1}{I} \sum_{i=1}^I \left(y_{i,\ell} - \hat{y}_{i,\ell}^{(-i)} \right)^2$$

to estimate the variance of $\hat{y}_{i,\ell}^{(-i)}$.

- 4 For any new curve $W_j^*(t)$ consider the confidence interval to be $\hat{y}_{j,\ell}^* \pm cS_\ell$ (estimators are normally distributed conditionally on the calibration sample).

Prediction

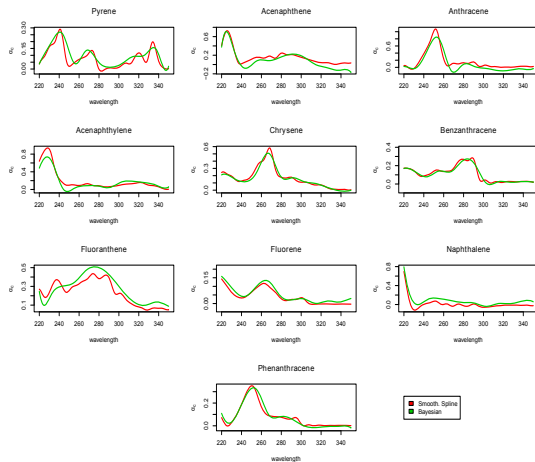
- In order to compare the prediction performance of different approaches we are going to use the Standard Error of Prediction (SEP), which is the root mean square of the difference between the true and the predicted content. The contribution of component ℓ is given by

$$\text{SEP}_\ell = \left(\frac{1}{J-1} \sum_{j=1}^J (y_{j,\ell}^* - \hat{y}_{j,\ell}^*) \right)^{1/2}$$

and the overall SEP is given by

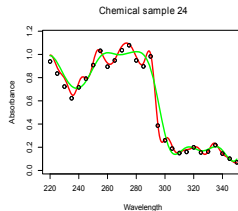
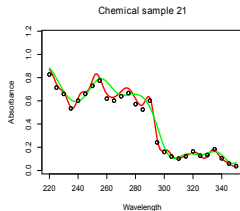
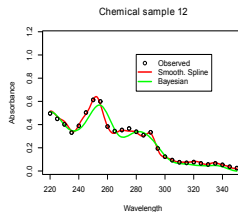
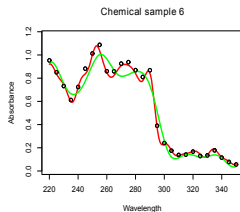
$$\text{SEP} = \left(\frac{1}{mJ-1} \sum_{\ell=1}^m \sum_{j=1}^J (y_{j,\ell}^* - \hat{y}_{j,\ell}^*) \right)^{1/2}.$$

PAH data



Estimated absorbance spectra $\hat{\theta}_\ell$, $\ell = 1, \dots, 10$ for PAH dataset

PAH data



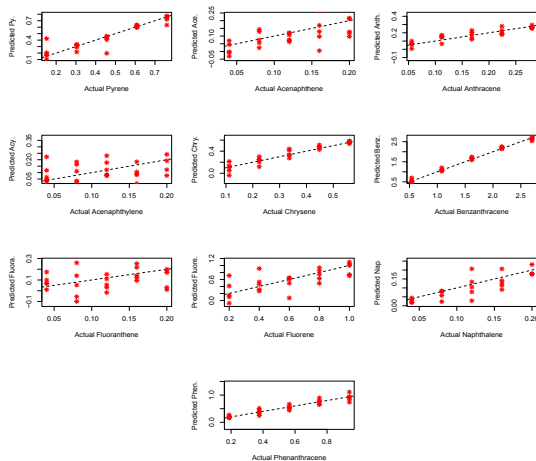
Estimated aggregated curves $\sum_{\ell=1}^m y_{i\ell} \hat{\theta}_{\ell}(t)$ for chemical samples $i=6, 12, 21$ and 24 .

PAH data

Constituent	Smoothing Splines
Pyrene	0.05
Acenaphthene	0.07
Anthracene	0.04
Acenaphthylene	0.42
Chrysene	0.27
Benzantracene	1.64
Fluoranthene	0.41
Fluorene	0.56
Naphthalene	0.11
Phenanthracene	0.36

Table: Leave-one-out Standard Deviation estimates (mg/l) for PAH data.

PAH data



The predicted values $\hat{y}$ for the Independent Test Set against their true values.

PAH data

Constituent	Smoothing Sp	MLR	PCR	PLS
Pyrene	0.09	0.09	0.09	0.09
Acenaphthene	0.06	0.06	0.03	0.03
Anthracene	0.04	0.03	0.03	0.03
Acenaphthylene	0.09	0.09	0.06	0.06
Chrysene	0.06	0.06	0.05	0.04
Benzanthracene	0.07	0.07	0.07	0.06
Fluoranthene	0.09	0.09	0.08	0.08
Fluorene	0.23	0.24	0.20	0.15
Naphthalene	0.04	0.04	0.03	0.04
Phenanthracene	0.09	0.08	0.09	0.08
Overall	0.10	0.11	0.09	0.08

Table: SEP (mg/l) for PAH data. Smoothing spline (λ minimizing GCV); MLR: multiple linear regression; PCR: Principal Components Regression with 10 components (Brereton, 2003); PLS: Partial Least Squares, same number of components as PCR.

Concluding Remarks

- In general, data-driven smoothness capture more of the local variation of the data and produce less smooth estimates than Bayesian smoothness. It also does perform better in the prediction problem.
- Variability of prediction estimates can be estimated using a leave-one-out jackknife technique (unnecessary in Bayesian setting).
- The functional approach has two main advantages over multivariate methods: (i) it can be applied if the spectra are measured at distinct wavelengths for distinct curves; (ii) it provides estimates for the individual absorption curves for each constituent.

Work in progress

- Calibration is satisfactory if assumption of independent errors is violated, but simulated prediction does not perform as well as partial least squares methods.
- Smoothness of absorbance curves is assumed to be the homogeneous, but it might be unrealistic.

References I

Richard Brereton. Chemometrics: Data Analysis For the Laboratory and Chemical Plant. Wiley, Chichester, 2003.

Grace Wahba. Spline Models for Observational Data. SIAM:PA, 1990.