Modeling Infrared Spectra : an Algorithm for an Automatic and Simultaneous Analysis

ESREL 2021

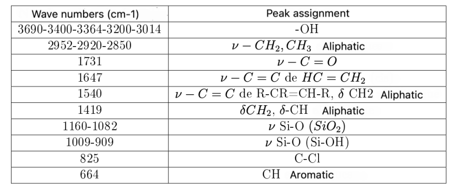
C. Butucea (ENSAE), J.-F. Delmas (Ecole des Ponts), A. Dutfoy (EDF R&D), C. Hardy (EDF R&D, Ecole des Ponts)

Context of the study

- Infrared spectra measure the interaction of infrared radiations with the matter and reveal the presence of chemical substances in a material.
- Infrared spectroscopy has become widespread in the industry for nondestructive testing.

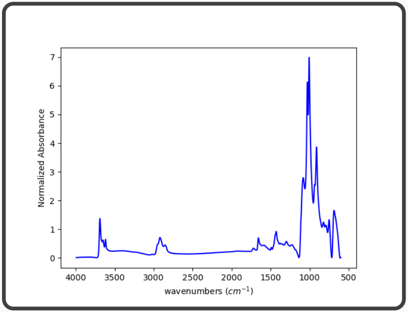
Context of the study

Infrared spectroscopy for nondestructive testing



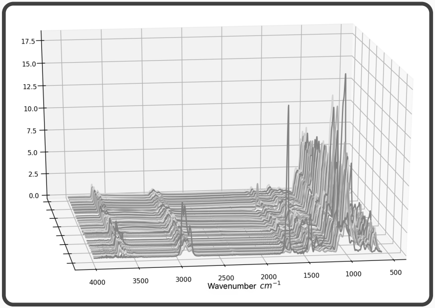
Wave numbers (cm-1)	Peak assignment
3690-3400-3364-3200-3014	-OH
2952-2920-2850	$\nu - CH_2, CH_3$ Aliphatic
1731	$\nu - C = O$
1647	$\nu - C = C$ de $HC = CH_2$
1540	$\nu - C = C$ de R-CR=CH-R, δ CH2 Aliphatic
1419	δCH_2, δ-CH Aliphatic
1160-1082	ν Si-O (SiO_2)
1009-909	ν Si-O (Si-OH)
825	C-Cl
664	CH Aromatic

Location of peaks and corresponding bonds for polychloroprene ([Tchalla, 2017]).



Context of the study

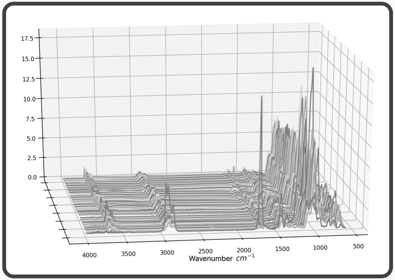
Dealing with a large dataset of infrared spectra



Infrared spectra from 72 polychloroprene rubbers used in a marine environment at different aging levels.

Context of the study

- Analyzing infrared spectra usually requires an expertise (guess on the locations of peaks, numbers of peaks...).
- Principal component analysis or partial least square analysis produce results for which it is difficult to give a physical interpretation.



Our goal is to recover in an automatic and simultaneous way the peaks (locations and amplitudes) associated with the chemical compounds of a material.

Context of the study

→ Recover the locations of the peaks to identify chemical substances

→ Recover the amplitudes of the peaks to determine concentrations

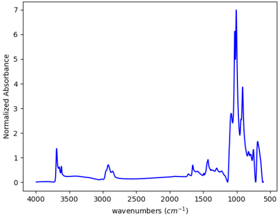
The Model

The Model

The spectra are modeled by linear combinations of peaks whose shape and position are parametrized with an additive noise.

$$y(\sigma_j) = \sum_{k=1}^{K} \beta_k^\star \, \phi(\theta_k^\star, \sigma_j) + w_j,$$

$$1 \leq j \leq T.$$



- $w_j \sim \mathcal{N}(0, s^2), i.i.d,$
- $\phi(\theta_k^\star, \sigma_j) = \frac{\varphi(\theta_k^\star, \sigma_j)}{\sqrt{\Delta_T} \left\| \varphi(\theta_k^\star, \cdot) \right\|_{\ell_2}},$
- $\theta_k^\star \in \Theta \subset \mathbb{R}^d,$

where Δ_T is the step in the discretization scheme.

The model

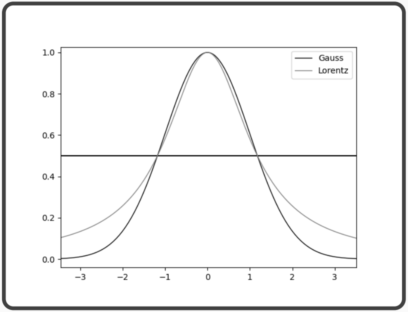
Shape of the parametric family :

$$\varphi_{Gauss} : \Theta \times \mathcal{D} \to \mathbb{R}$$

$$((\mu,\nu),\sigma) \mapsto \exp\left(-\frac{(\sigma-\mu)^2}{2\nu^2}\right)$$

$$\varphi_{Lorentz} : \Theta \times \mathcal{D} \to \mathbb{R}$$

$$((\mu,\nu),\sigma) \mapsto \frac{1}{1+\frac{(\sigma-\mu)^2}{2\nu^2}}$$



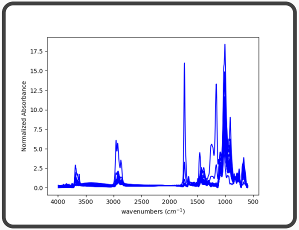
Gaussian and Lorentz functions with the same half-width.

The Model

We observe n spectra $(y_i)_{1\leq i\leq n}$ discretized on T wavenumbers $(\sigma_j)_{1\leq j\leq T}$.

$$y_i(\sigma_j) = \sum_{k=1}^{K} B^\star_{ik}\,\phi(\theta^\star_k, \sigma_j) + w_{ij},$$

$$1 \leq j \leq T, \quad 1 \leq i \leq n.$$



- $w_{ij} \sim \mathcal{N}(0, s^2), i.i.d,$
- $S^\star = \{k, ||B^\star_{.,k}|| \neq 0\}.$

The matrix $B^\star$ is sparse columnwise, i.e Card$(S^\star) < K$.

The peaks are shared by all the spectra in the dataset but their amplitudes are specific to each spectrum.

The Model

A matrix form for the model:

$$\boxed{Y = B^\star \Phi(\vartheta^\star) + W}$$

- $Y_{i,j} = y_i(\sigma_j), \quad Y \in \mathbb{R}^{n \times T}$
- $\vartheta^\star = \left(\theta_1^\star, \cdots, \theta_K^\star\right), \quad \vartheta^\star \in \mathbb{R}^{d \times K}$
- $\Phi(\vartheta)_{k,j} = \phi(\theta_k, \sigma_j), \quad \Phi(\vartheta) \in \mathbb{R}^{K \times T}$
- $W_{ij} \sim \mathcal{N}(0, s^2), i.i.d, \quad W \in \mathbb{R}^{n \times T}$

The Model

A matrix form for the model:

$$\boxed{Y = B^\star \Phi(\vartheta^\star) + W, \quad Y \in \mathbb{R}^{n \times T}}$$

- K is an upper bound of the number of peaks in the mixture (arbitrarily large).
- $B^\star$ is sparse columnwise.

The model

→ Recover the locations of the peaks $\vartheta^\star_{S^\star}$

→ Recover the amplitudes of the peaks $B^\star$

up to a joint permutation on the columns of $B^\star$ and $\vartheta^\star_{S^\star}$.

Optimization Problem

Optimization Problem

We formulate a non-linear least square problem with a Group-Lasso penalization term weighted by a real parameter $\lambda > 0$:

$$\boxed{(\hat{B}, \hat{\vartheta}) \in \underset{B \in \mathbb{R}^{n \times K}, \vartheta \in \Theta_{K,T}(h)}{\operatorname{argmin}} \quad \frac{1}{2nT} ||Y - B\Phi(\vartheta)||_{\ell_2}^2 + \lambda ||B||_{1,2}}$$

- $\|B\|_{1,2} = \sum_{k=1}^{K} \|B_{\cdot,k}\|_{\ell_2}$

- $\Theta_{K,T}(h) \subset \Theta^K$ with $h > 0$, is the set of parameters $\vartheta = (\theta_1, \cdots, \theta_K) \in \Theta^K$ such that for all $1 \leq \ell, k \leq K, \ell \neq k$:

$$\Delta_T \left| \langle \phi(\theta_\ell), \phi(\theta_k) \rangle \right| < h.$$

Optimization Problem

We formulate a non-linear least square problem with a Group-Lasso penalization term weighted by a real parameter $\lambda > 0$:

$$\boxed{(\hat{B}, \hat{\vartheta}) \in \underset{B \in \mathbb{R}^{n \times K}, \vartheta \in \Theta_{K,T}(h)}{\operatorname{argmin}} \quad \frac{1}{2nT} ||Y - B\Phi(\vartheta)||_{\ell_2}^2 + \lambda ||B||_{1,2}}$$

The set $\hat{S}$ gathers the indices of the active peaks used to fit the spectra :

$$\hat{S} = \{k : \text{ there exists } 1 \leq i \leq n, \hat{B}_{ik} \neq 0\}.$$

→ $\hat{B}$ et $\hat{\vartheta}_{\hat{S}}$ are estimators of $B^\star$ and $\vartheta^\star_{S^\star}$ (up to a joint permutation on the columns of $B^\star$ and $\vartheta^\star_{S^\star}$).

Algorithm

Algorithm

One would like to solve the problem:

$$(\hat{B}, \hat{\vartheta}) \in \underset{B \in \mathbb{R}^{n \times K}, \vartheta \in \Theta_{K,T}(h)}{\operatorname{argmin}} \quad \frac{1}{2nT} ||Y - B\Phi(\vartheta)||_{\ell_2}^2 + \lambda ||B||_{1,2}$$

→ Non convex problem!

Algorithm 1:

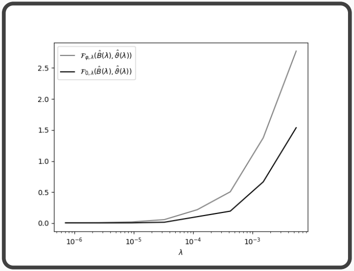
```
Data: $Y$                                Output: $\vartheta, B$
Input: $\varphi, \lambda, h$             Initialize: $i := 0,\ R^{(0)} := Y\ ,\ \vartheta^{(0)} := \emptyset$
while $i < K$ do
    $\theta^{(i+\frac{1}{2})} \in \operatorname{argmax}_{\theta \in \Theta} \left\| R^{(i)} \phi(\theta)^\top \right\|_{\ell_2}^2$
    $\vartheta^{(i+\frac{1}{2})} = \left(\vartheta^{(i)}, \theta^{(i+\frac{1}{2})}\right)$
    $B^{(i+\frac{1}{2})} \in \operatorname{argmin}_{B \in \mathbb{R}_+^{n \times (i+1)}} \mathcal{F}_{\lambda,\varphi}(B, \vartheta^{(i+\frac{1}{2})})$
    $\vartheta^{(i+1)} \in \operatorname{argmin}_{\vartheta \in \Theta^{i+1}} \mathcal{F}_{\lambda,\varphi}(B^{(i)}, \vartheta)$ initialized in $\vartheta^{(i+\frac{1}{2})}$
    Merging routine $(\vartheta^{(i+1)},\ h)$
    $B^{(i+1)} \in \operatorname{argmin}_{B \in \mathbb{R}_+^{n \times (i+1)}} \mathcal{F}_{\lambda,\varphi}(B, \vartheta^{(i+1)})$
    $R^{(i+1)} = Y - B^{(i+1)} \Phi(\vartheta^{(i+1)})$
    $i = i + 1$
end
```

We note for a matrix $B \in \mathbb{R}^{n \times m}$ et $\vartheta \in \Theta^m$,

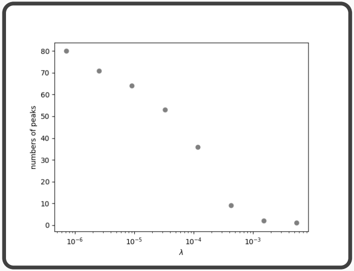
$$\mathcal{F}_{\lambda,\varphi}(B, \vartheta) = \frac{1}{2nT} \left\| Y - B\Phi(\vartheta) \right\|_{\ell_2}^2 + \lambda \|B\|_{1,2}.$$

An application to group polychloroprene samples with respect to aging

Resolution for spectra from polychloroprene rubbers



Mean square error $\mathcal{F}_{0,\varphi}(\hat{B}(\lambda),\hat{\vartheta}(\lambda))$ and penalized mean square error $\mathcal{F}_{\lambda,\varphi}(\hat{B}(\lambda),\hat{\vartheta}(\lambda))$ seen as functions of λ.

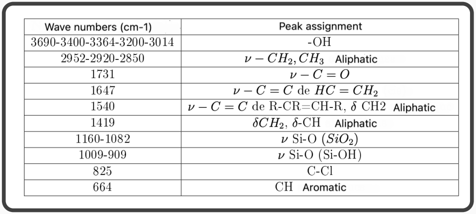


Number of peaks found by the algorithm to fit the spectra of polychloroprene samples as a function of the tuning parameter λ.

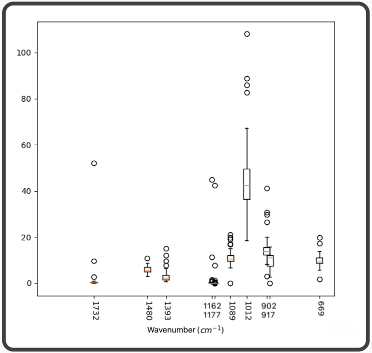
We note for a matrix $B \in \mathbb{R}^{n\times m}$ et $\vartheta \in \Theta^m$,

$$\mathcal{F}_{\lambda,\varphi}(B,\vartheta) = \frac{1}{2nT}\,\|Y - B\Phi(\vartheta)\|_{\ell_2}^2 + \lambda\|B\|_{1,2}.$$

Resolution for spectra from polychloroprene rubbers



Wave numbers (cm-1)	Peak assignment
3690-3400-3364-3200-3014	-OH
2952-2920-2850	$\nu - CH_2, CH_3$ Aliphatic
1731	$\nu - C = O$
1647	$\nu - C = C$ de $HC = CH_2$
1540	$\nu - C = C$ de R-CR=CH-R, δ CH2 Aliphatic
1419	δCH_2, δ-CH Aliphatic
1160-1082	ν Si-O (SiO_2)
1009-909	ν Si-O (Si-OH)
825	C-Cl
664	CH Aromatic

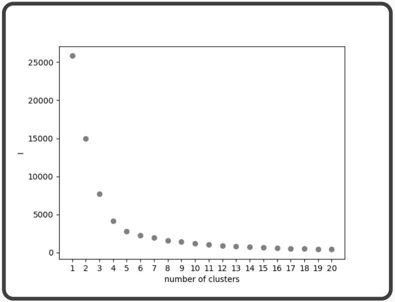


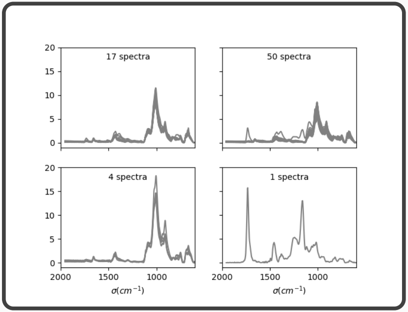
Boxplot for the amplitudes of the 10 most significant peaks for the 72 polychloroprene spectra in the dataset.

The locations of the peaks found by the algorithm are consistent with those established by previous work in the field of chemistry.

Clustering for spectra from polychloroprene rubbers

We run a k-means on the n vectors of amplitude $\hat{B}_{i,.} \in \mathbb{R}^K$





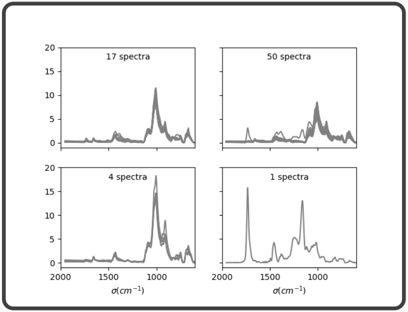
The k-means algorithm aims to partition the n vectors into M sets $\mathcal{A} = \{\mathcal{A}_1, \cdots, \mathcal{A}_M\}$ with barycenters $\{\beta_1, \cdots, \beta_M\}$ so as to minimize the within-cluster sum of squares :

$$\min_{\mathcal{A}} \underbrace{\sum_{\ell=1}^{M} \sum_{i \in A_\ell} \left\| \hat{B}_{i,.} - \beta_\ell \right\|_{\ell_2}^2}_{:=I(M)}.$$

Aging process of polychloroprene rubbers

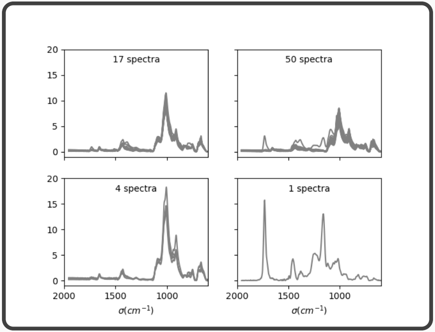
- Running a k-means on the amplitude vectors $\hat{B}_{i,.} \in \mathbb{R}^K, \quad 1 \leq i \leq n$
- Running a k-means on the amplitude vectors restricted to silica, silanol and carbonyl peaks

$\rightarrow$ It yields the same results.



The main differences between the spectra are due to the peaks of carbonyl, silanol and silica.

Aging process of polychloroprene rubbers



Previous work in chemistry has also shown that the amplitudes of peaks at 1731 cm^{-1} (Carbonyls), $1160 - 1082$ (Silice) cm^{-1} and $1009 - 909$ cm^{-1}(Silanol) evolve with age (hydrolysis of silica and oxidation reaction) in a marine environment ([Le Gac et al., 2012]).

Conclusion

- The spectra are modeled under the physical constraints by linear combinations of peaks.
- A numerical method is proposed with an off-the-grid scheme to estimate the parameters of the model.
- The parameters have a physical interpretation.
- This general framework can be applied to many other branches of spectroscopy.