Non-linear regression model in therapeutic monitoring of an anticancer molecule by Surface Raman Enhanced Spectroscopy

–Séminaire du CMAP–

Tom Rohmer, Laetitia Le, Marc Lavielle

Centre de Mathématiques Appliquées, Xpop Team

April 9, 2019

Summary

Summary

Chemotherapy

At the European Georges Pompidou Hospital, AP-HP, Paris

- Around 34 000 antineoplastic preparations produced per year
- Manual production by staff pharmacy
- 80% of the chemotherapy production analytcially controlled to ensure right drug at the right dose
- 0.3% of non-compliant preparations (molecule or dose)

Chemotherapy

Chemotherapy in France (Inca 2015):

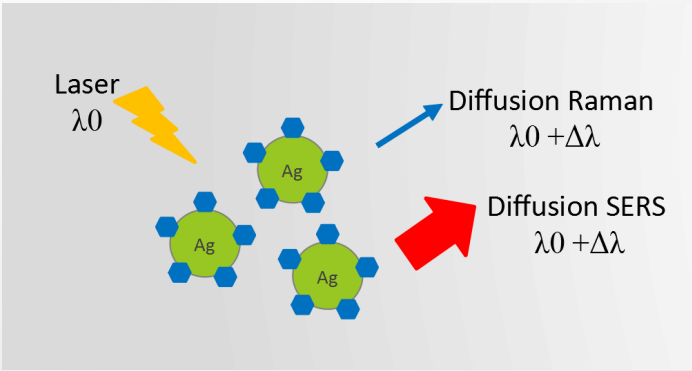
- 2 405 252 chemotherapy sessions (308 634 patients)
- 792 healthcare establishments
- Risk of medication errors

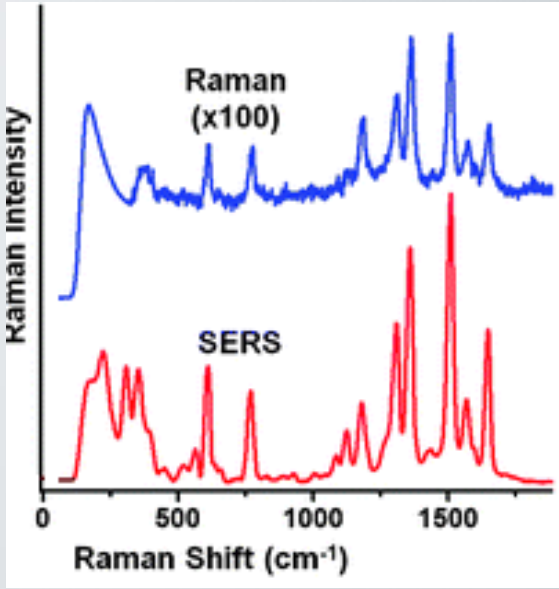
Objective:

Development of a simple, rapid and handled analytcial method to analyze in real time antineoplastic drug

Surface-enhanced Raman spectroscopy (SERS)

- A surface-sensitive technique that enhances Raman scattering by molecules adsorbed on rough metal surfaces or by nanostructures
- Enhancement factor up to 10^{11}
- Two mechanisms of the enhancement effect described:
  1. Electromagnetic effect with an excitation of localized surface plasmons
  2. Chemical effect with the formation of chemical bonds with the surface

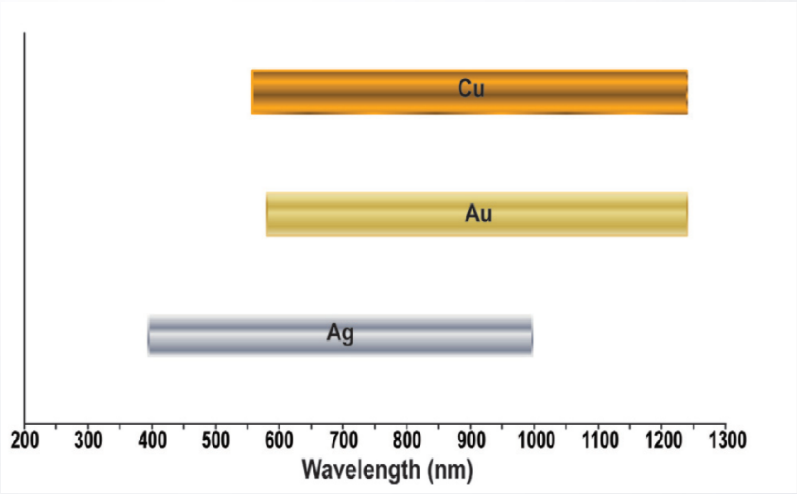




Surface-enhanced Raman spectroscopy (SERS)

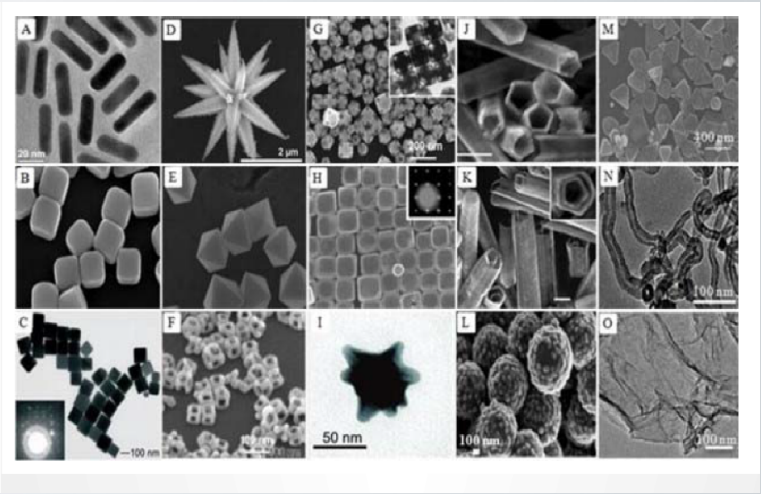
SERS substrates

- Solid substrate ou colloidal suspension
- Nature (gold, silver, copper,...)
- Form (sphere, stick, star...)
- Size



Parameters influencing SERS signal

- Analyte
- Substrate
- Acquisition parameters



Sample preparation

Molecule : 5-fluorouracile (5FU)

- Diluted in ultrapure water at various concentration
- From 0.5 to 12 mg/mL of 5FU

nanoparticules preparations

- Spherical silver Nps in water suspension
- Lee and Meisel technique
- By chemical reduction of AgNO3 by citrate
- Size $\sim$ 50 nm

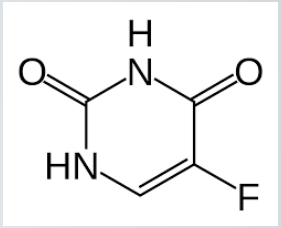


Figure: 5-Fluorouracile

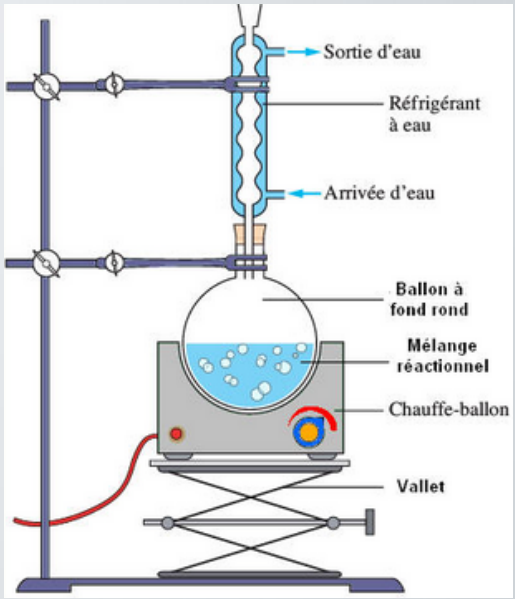


Figure: Silver Nps syntheses

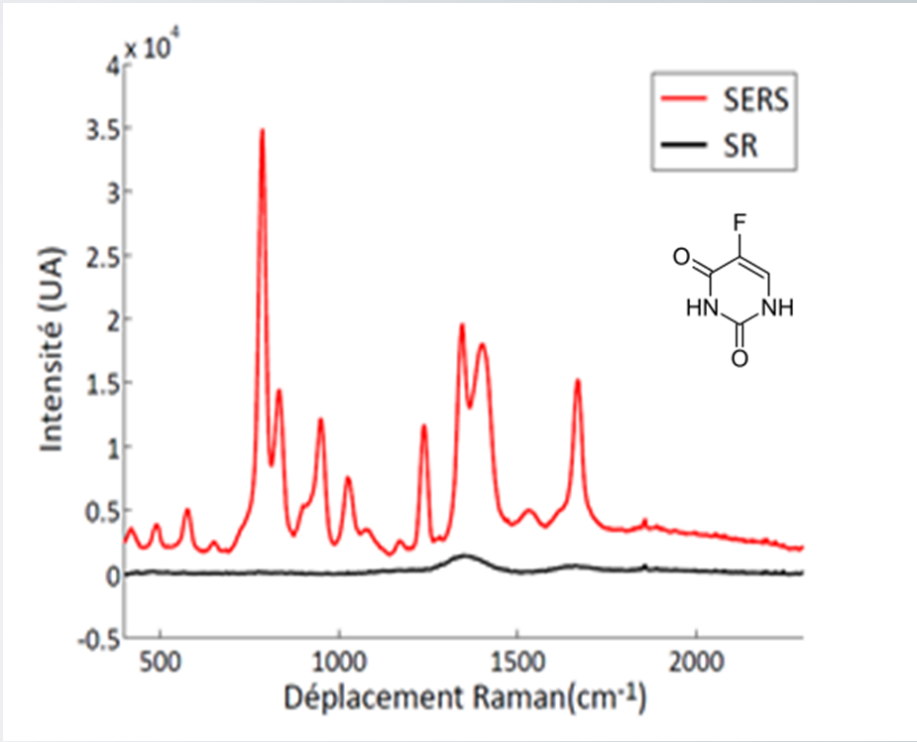
SERS analysis

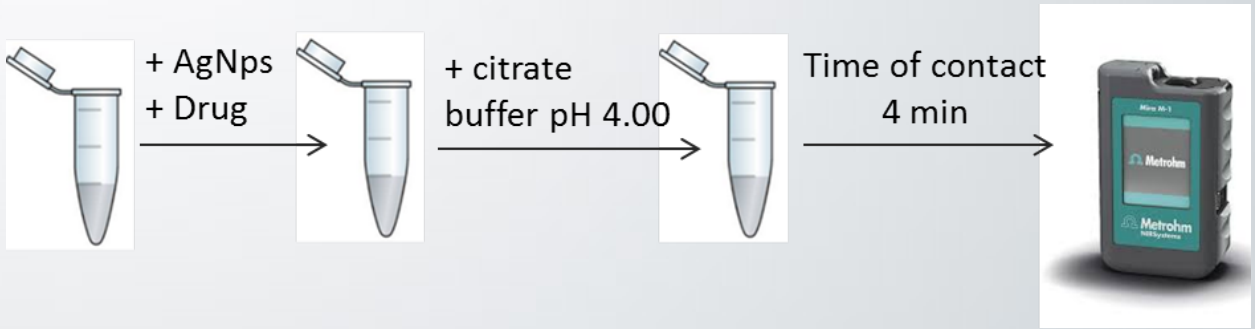
Preparation

- 400 μL of AgNps suspension
- 100 μL of 5FU solution
- 20 μL of citrate buffer pH 4.00

SERS acquisition

- MIRA spectrometer (Methrom)
- Spectral resolution : 12 to 14 cm^{-1}
- Spectral range : from 400 to 2300 cm^{-1}
- Time of acquisition : 2 s





Protocol

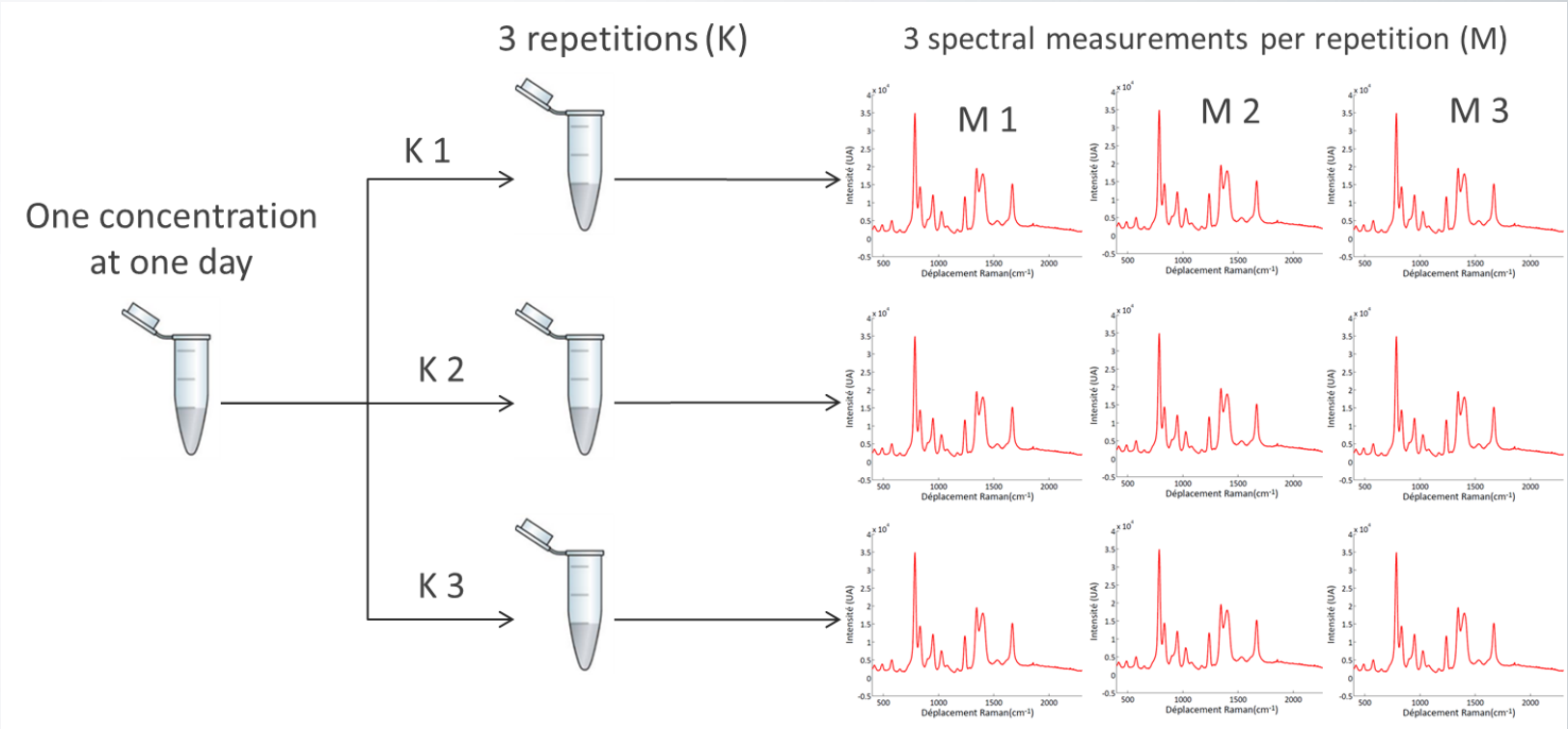


Figure: Experiment Protocol for each of the 9 nominal concentrations and each of the 7 series

Summary

Pretreatment I: Outliers detection

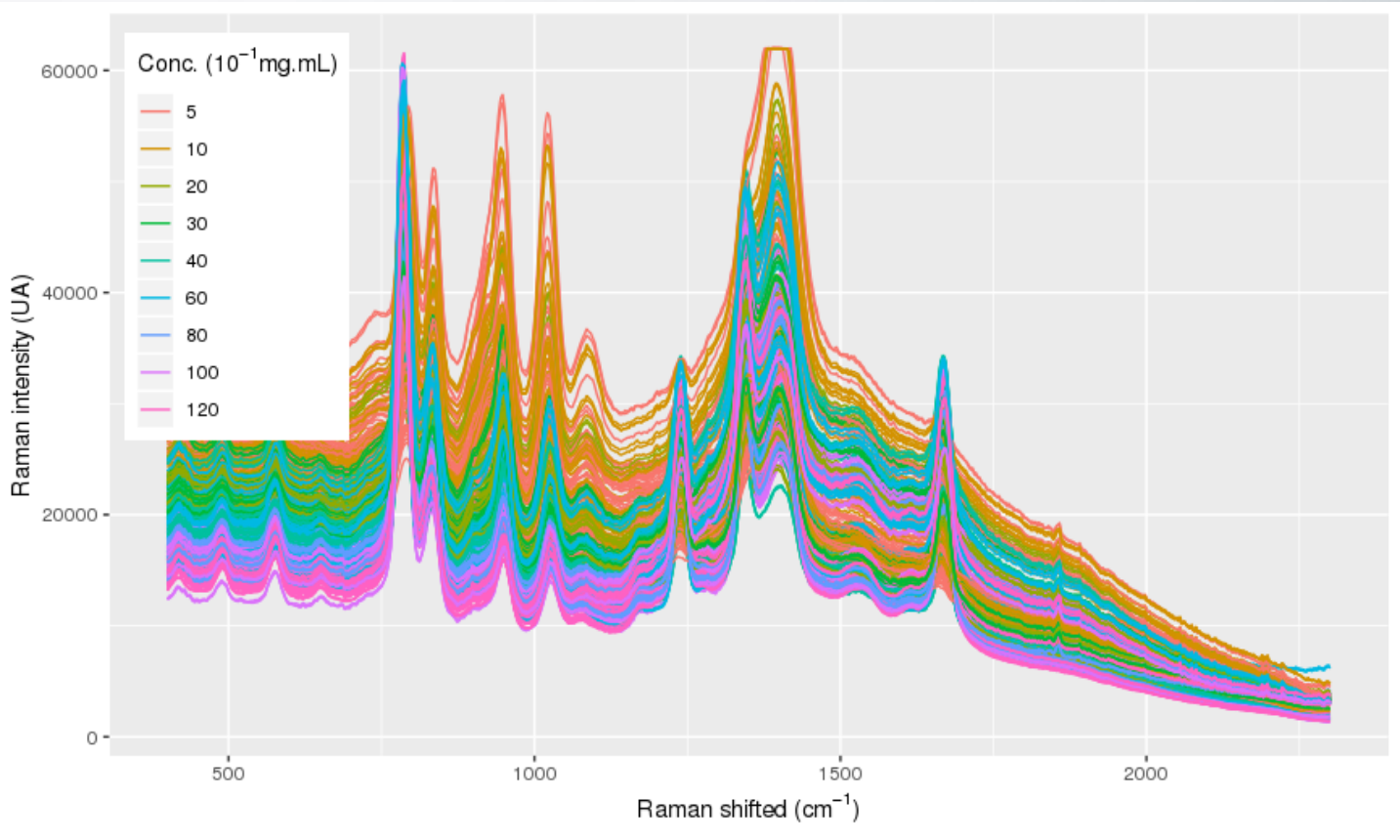


Figure: Spectra by nominal concentration: Signal saturation, exclusions

Pretreatment II: measurements aggregation

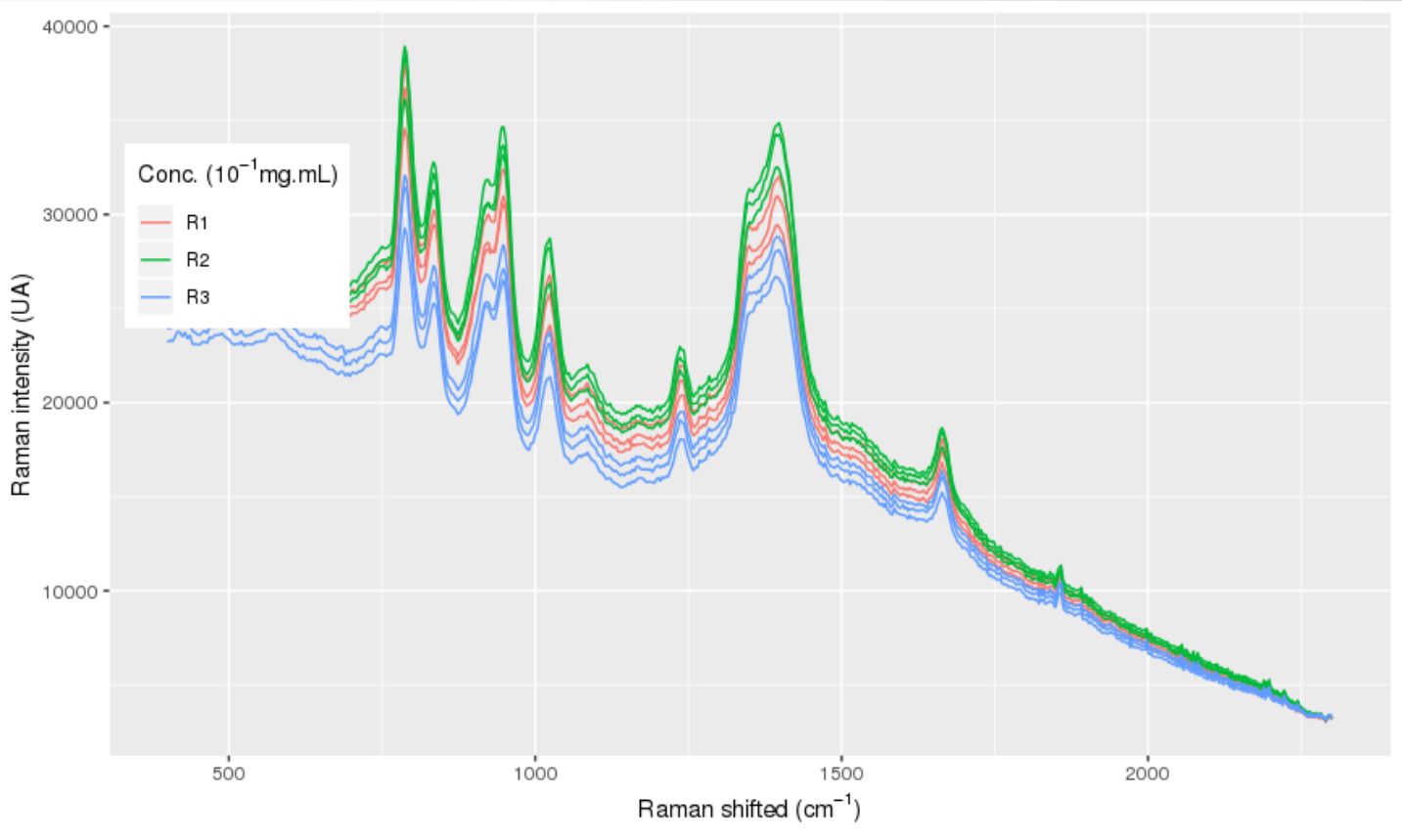


Figure: aggregate the 3 measurements: average spectrum

Pretreatment III: renormalization by the citrate band ($1022 cm^{-1}$)

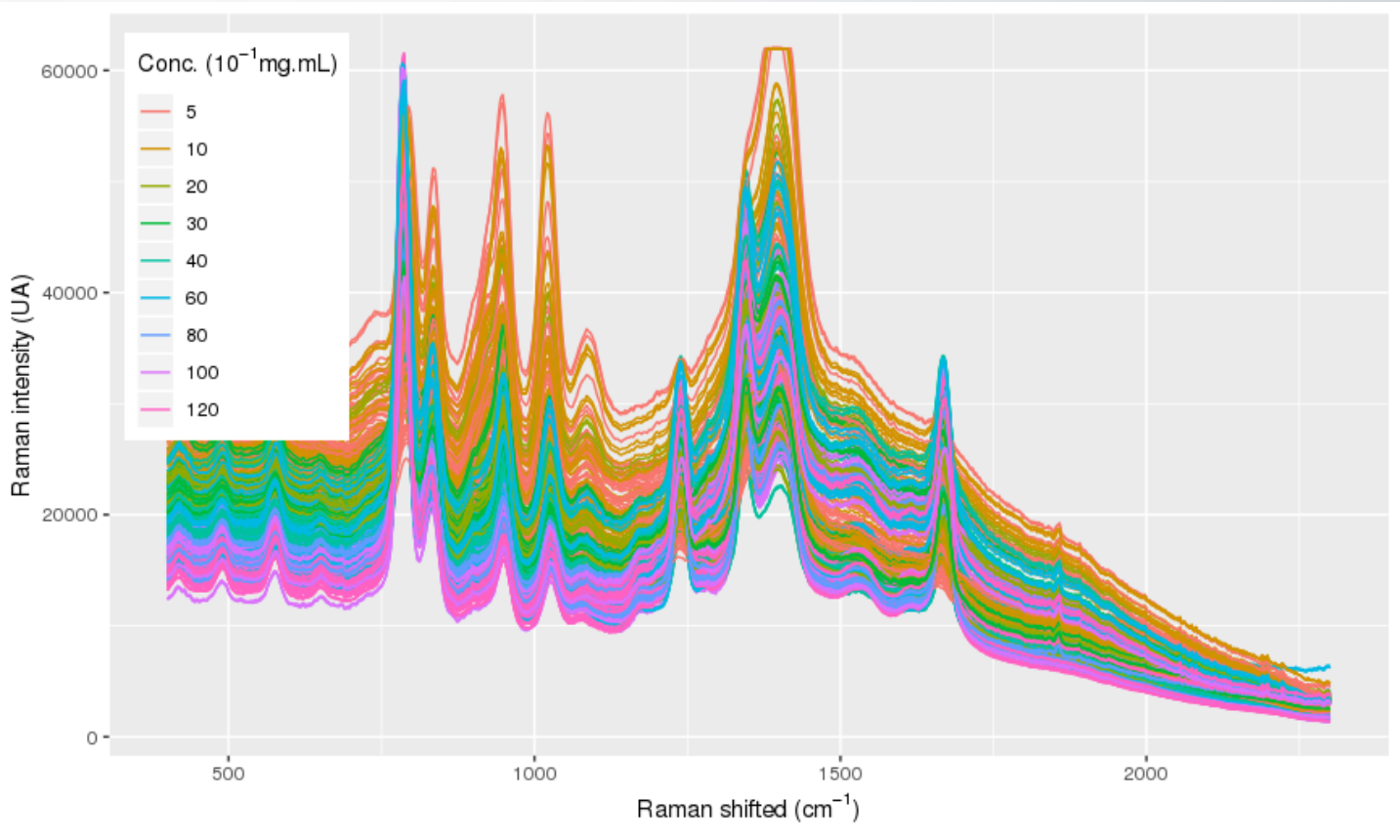
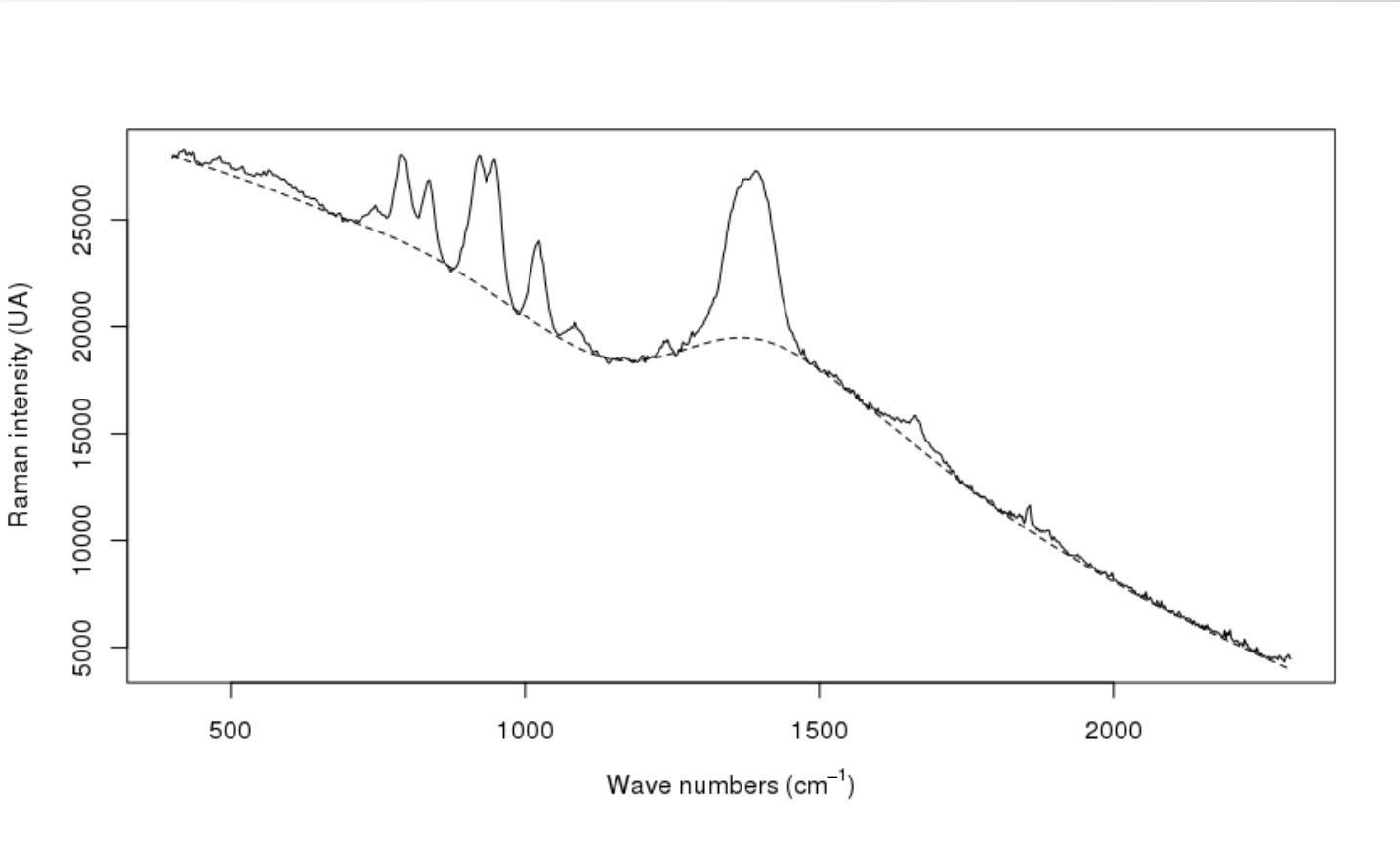


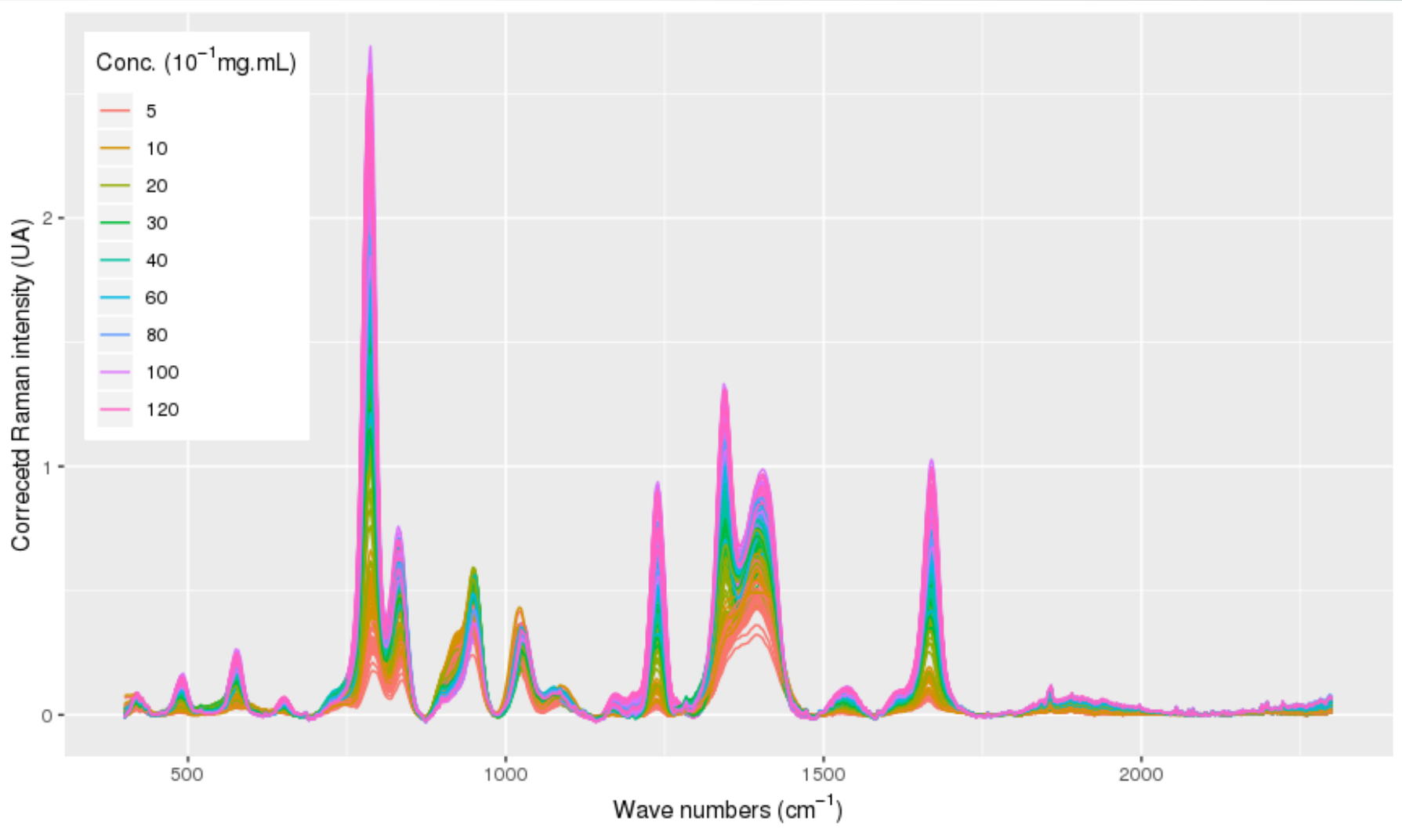
Figure: Spectra by nominal concentration: re-normalization

Pretreatment IV: baseline correction



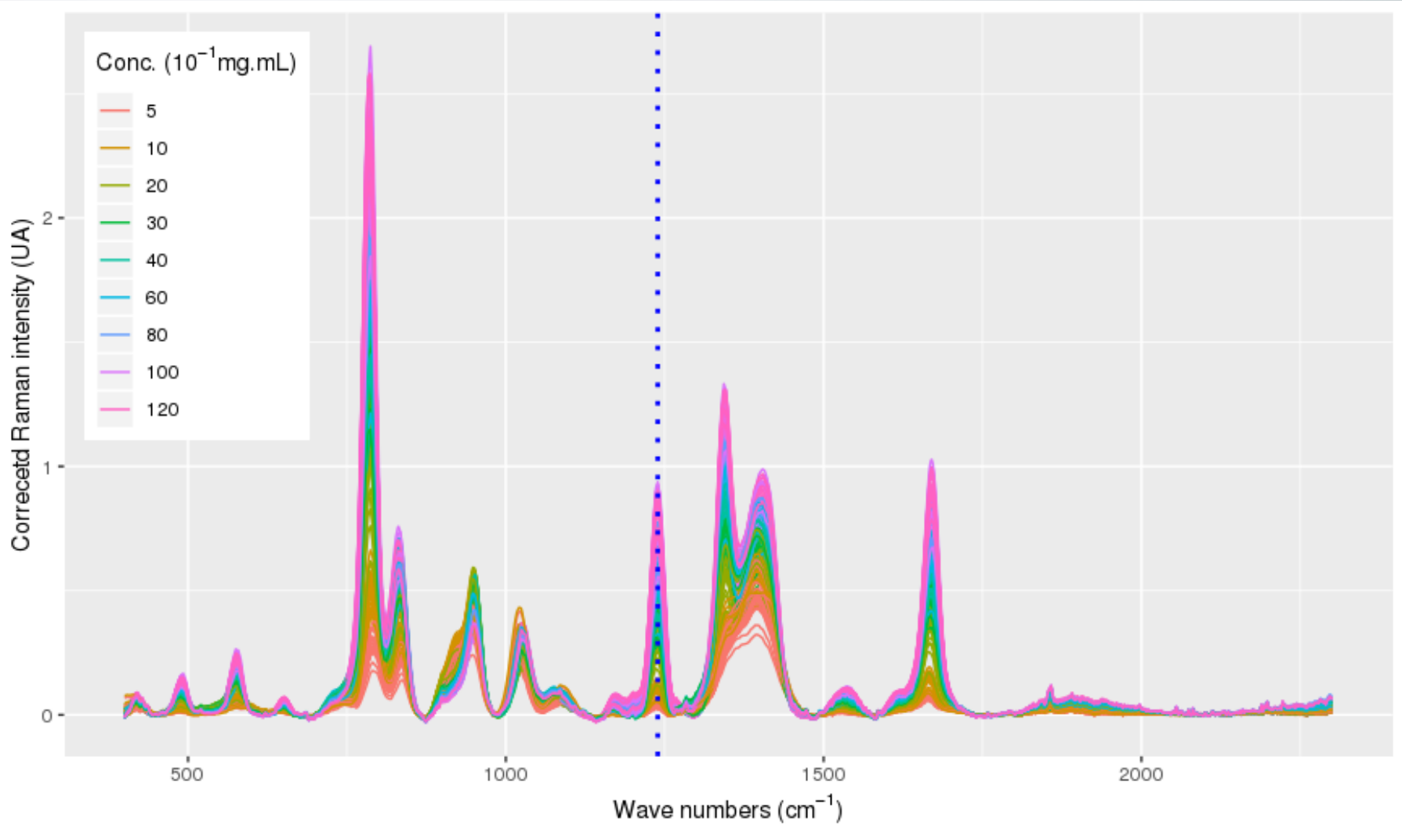
- Baseline correcion with asymmetric least squares smoothing
- Taking into account the flexibility of the baseline

Corrected spectra

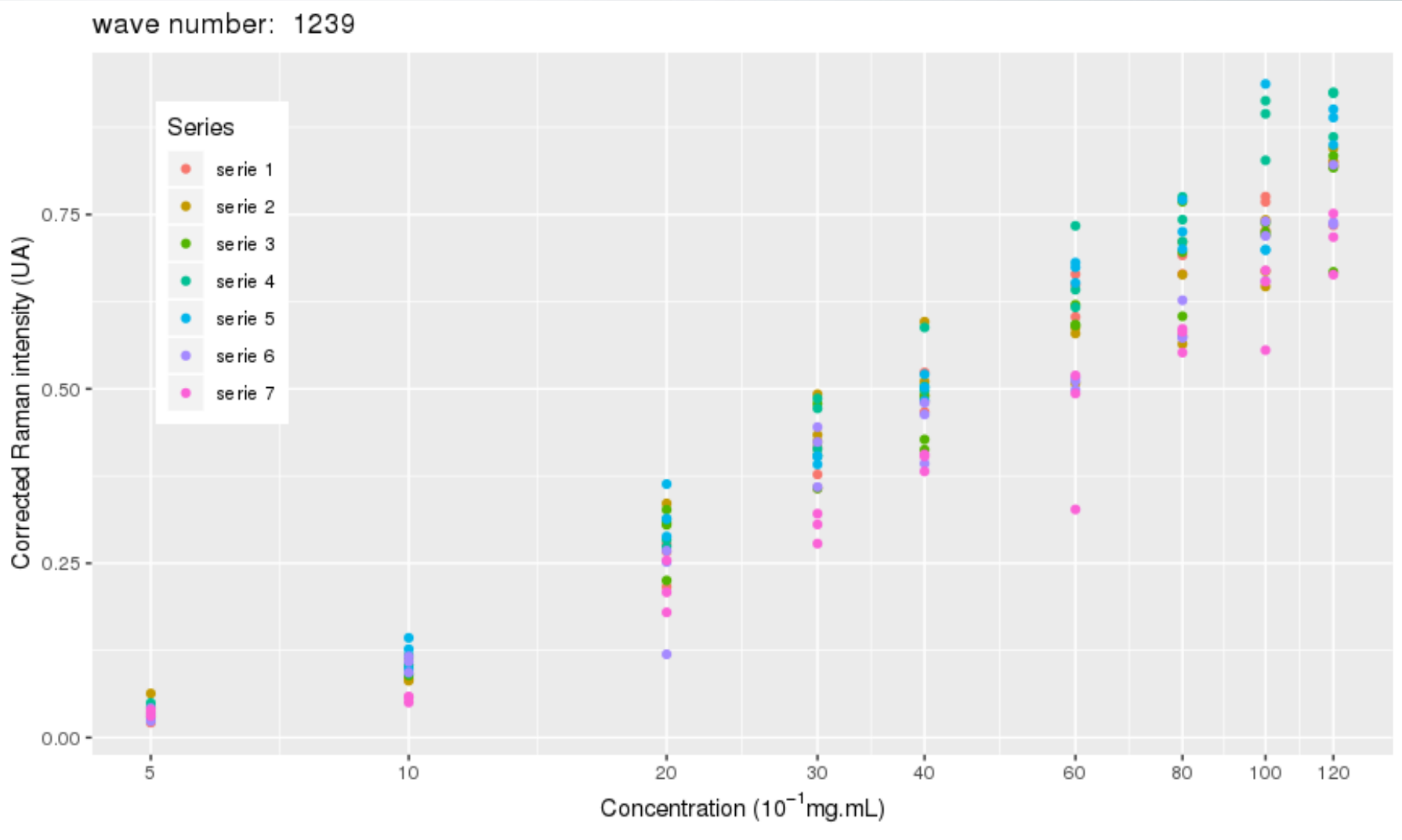


Summary

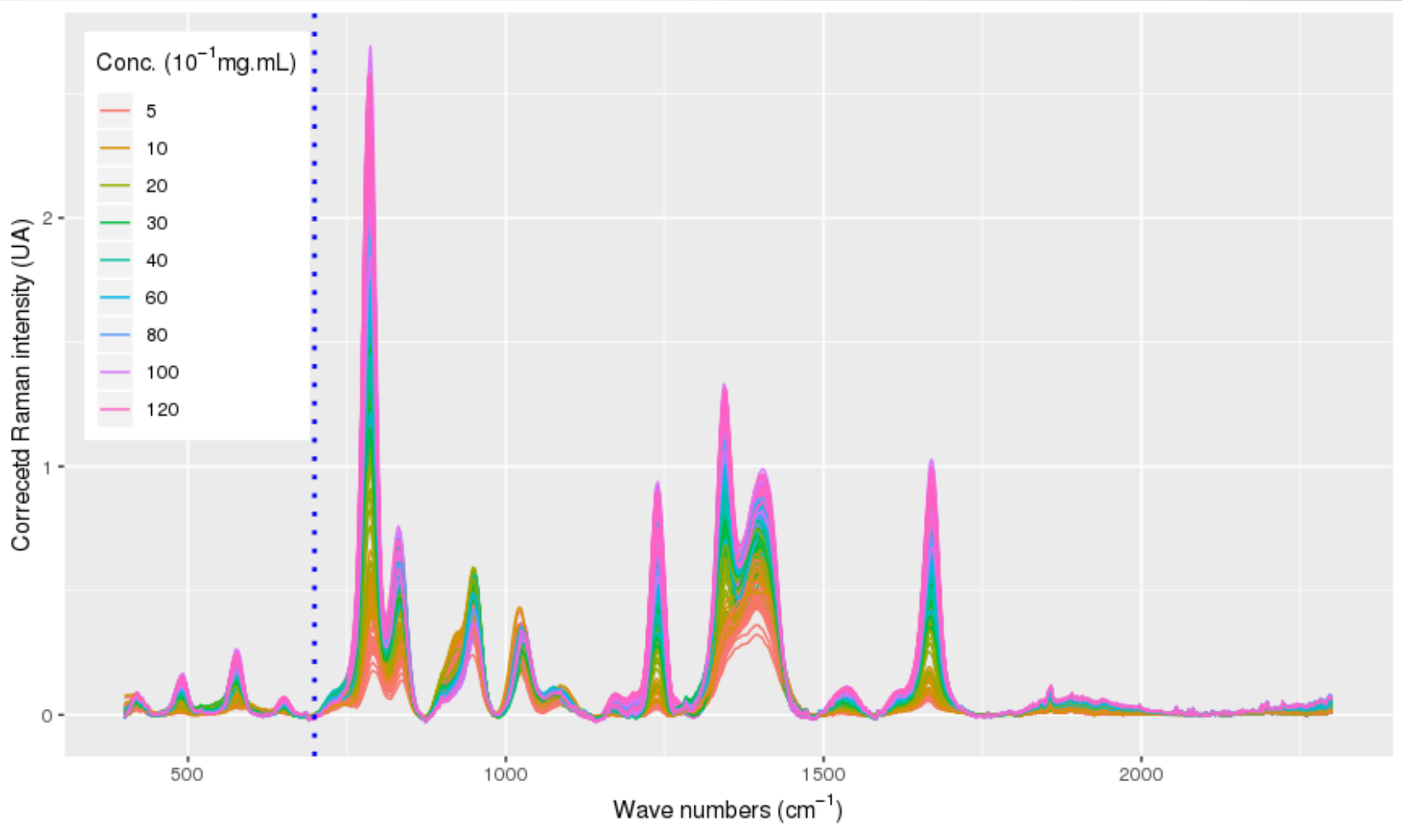
Wave number 1239



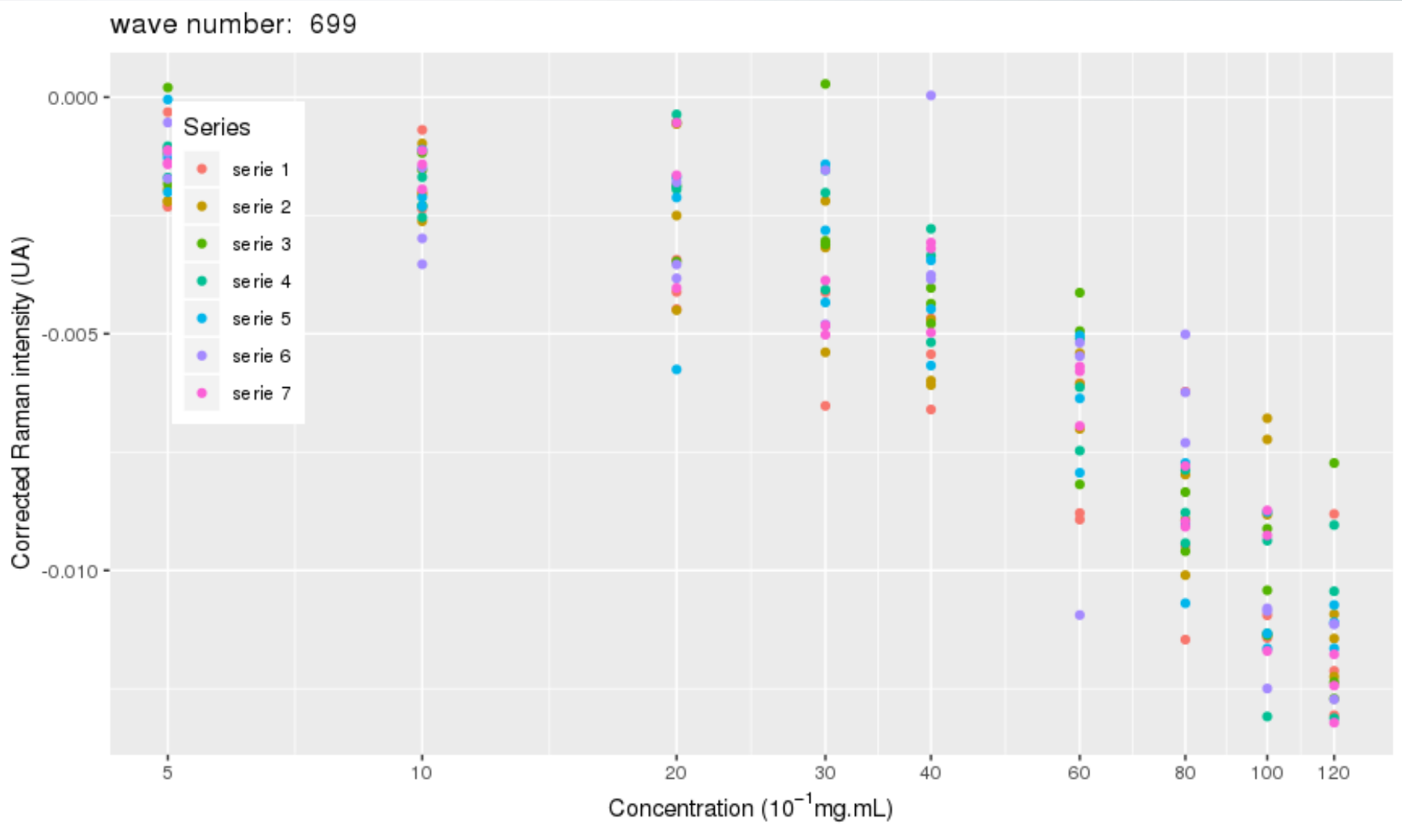
Linear correlation: 0.926



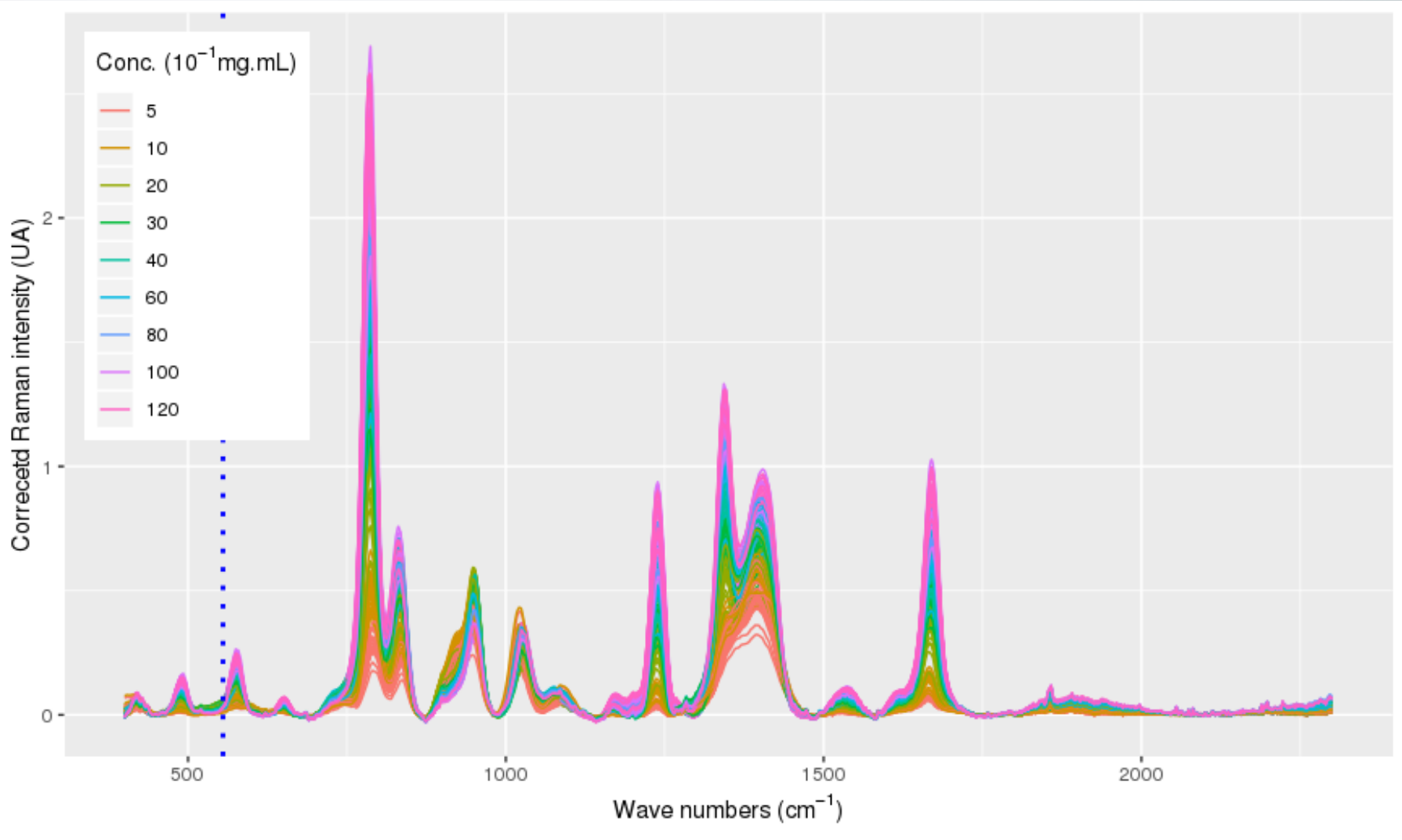
Wave number 699



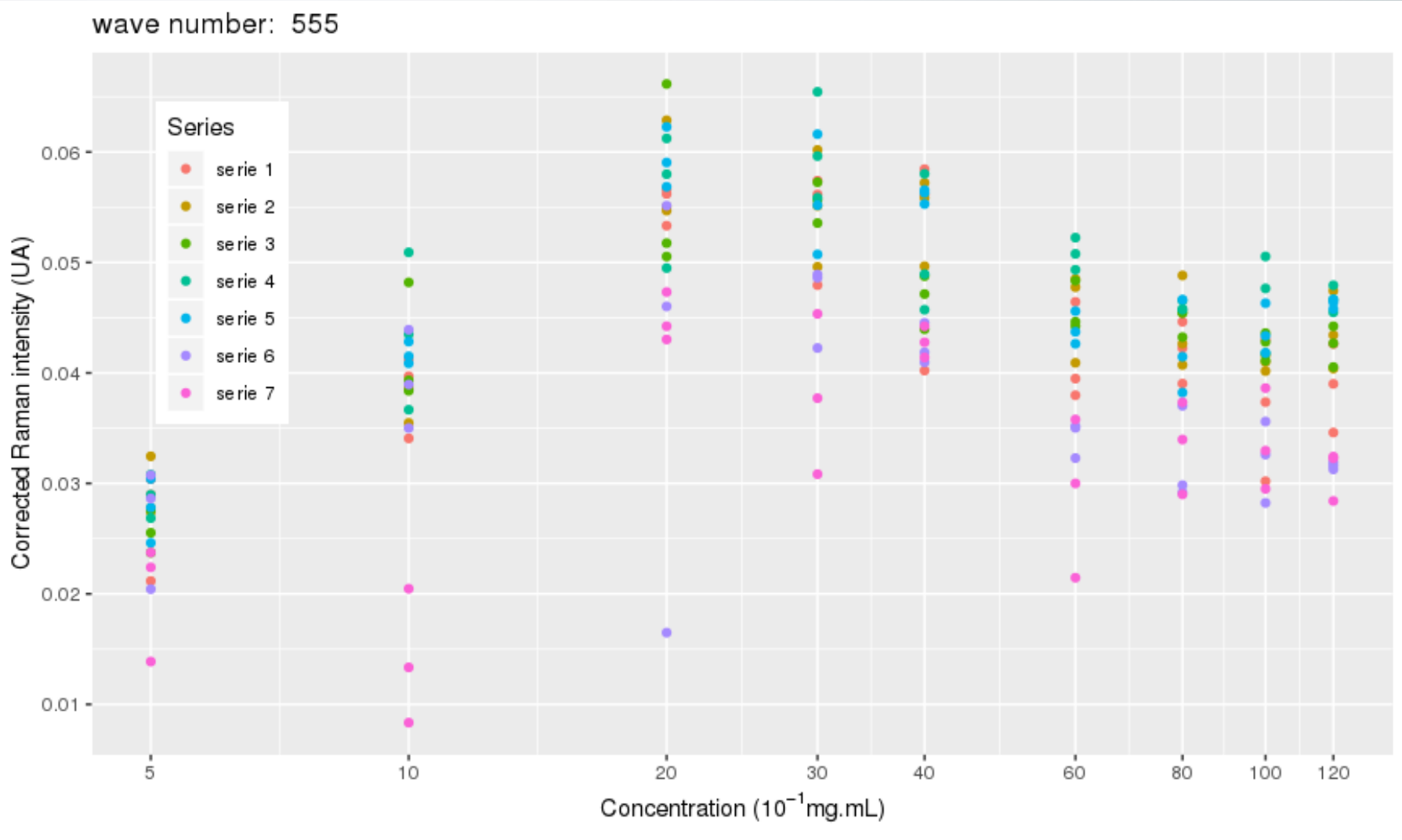
Linear Correlation: -0.929



Wave number 555



Linear correlation: 0.004



Regression models

Parametric regression model

$$y_{ijk\ell} = f(c_j, \theta_{i\ell}) + e_{ijk\ell},$$

with

- $i = 1, \ldots, I$ series of measures
- $c_1, \ldots, c_J$ are the J nominal concentrations
- $k = 1, \ldots, K$ repetitions
- $\ell = 1, \ldots, L$ Wave numbers
- $\theta_{i\ell}$ unknown parameters vector (to estimate)
- f is the structural model to define
- $e_{ijk\ell} \sim_{indpt.} \mathcal{N}(0, \sigma^2_{ij\ell})$

structural model

$$y_{ijk\ell} = f(c_j, \theta_{i\ell}) + e_{ijk\ell},$$

Structural model: In view of the data, we propose a sigmoid-type regression, that is for $\theta_{i\ell} = (S_{i\ell}, A_{i\ell}, \gamma_{i\ell}, \tau_{i\ell}) \in \mathbb{R}^4$:

non-linear regression type

$$f(c_j, \theta_{i\ell}) = S_{i\ell} + \frac{(A_{i\ell} - S_{i\ell})}{1 + \exp(-\gamma_{i\ell}(\log(c_j) - \tau_{i\ell}))}.$$

residual error model

$e_{ijk\ell} \sim \mathcal{N}(0, \sigma^2_{ij\ell})$ with

$$\sigma^2_{ij\ell} = g(c_j, \theta_{i\ell}, \xi_{i\ell}),$$

- Constant error model

$$\xi_{i\ell} = a_{i\ell}, \quad g(c_j, \theta_{i\ell}, \xi_{i\ell}) = a^2_{i\ell}$$

residual error model

$e_{ijk\ell} \sim \mathcal{N}(0, \sigma^2_{ij\ell})$ with

$$\sigma^2_{ij\ell} = g(c_j, \theta_{i\ell}, \xi_{i\ell}),$$

- Constant error model

$$\xi_{i\ell} = a_{i\ell}, \quad g(c_j, \theta_{i\ell}, \xi_{i\ell}) = a^2_{i\ell}$$

- Proportional error model

$$\xi_{i\ell} = b_{i\ell}, \quad g(c_j, \theta_{i\ell}, \xi_{i\ell}) = b^2_{i\ell} f(c_j, \theta_{i\ell})^2$$

residual error model

$e_{ijk\ell} \sim \mathcal{N}(0, \sigma^2_{ij\ell})$ with

$$\sigma^2_{ij\ell} = g(c_j, \theta_{i\ell}, \xi_{i\ell}),$$

- Constant error model

$$\xi_{i\ell} = a_{i\ell}, \quad g(c_j, \theta_{i\ell}, \xi_{i\ell}) = a^2_{i\ell}$$

- Proportional error model

$$\xi_{i\ell} = b_{i\ell}, \quad g(c_j, \theta_{i\ell}, \xi_{i\ell}) = b^2_{i\ell} f(c_j, \theta_{i\ell})^2$$

- Combined error model

$$\xi_{i\ell} = (a_{i\ell}, b_{i\ell}), \quad g(c_j, \theta_{i\ell}, \xi_{i\ell}) = a^2_{i\ell} + b^2_{i\ell} f(c_j, \theta_{i\ell})^2$$

Estimation problem, Maximum Likelihood Estimator (MLE)

for $\ell = 1, \dots L$ and $i = 1, \dots, I$, the vector parameters $(S_{i\ell}, A_{i\ell}, \gamma_{i\ell}, \tau_{i\ell})$ and $\xi_{i\ell} = (a_{i\ell}, b_{i\ell})$ are unknown and must be conjointly estimated by maximizing the log-likelihood of the statistical models, that are:

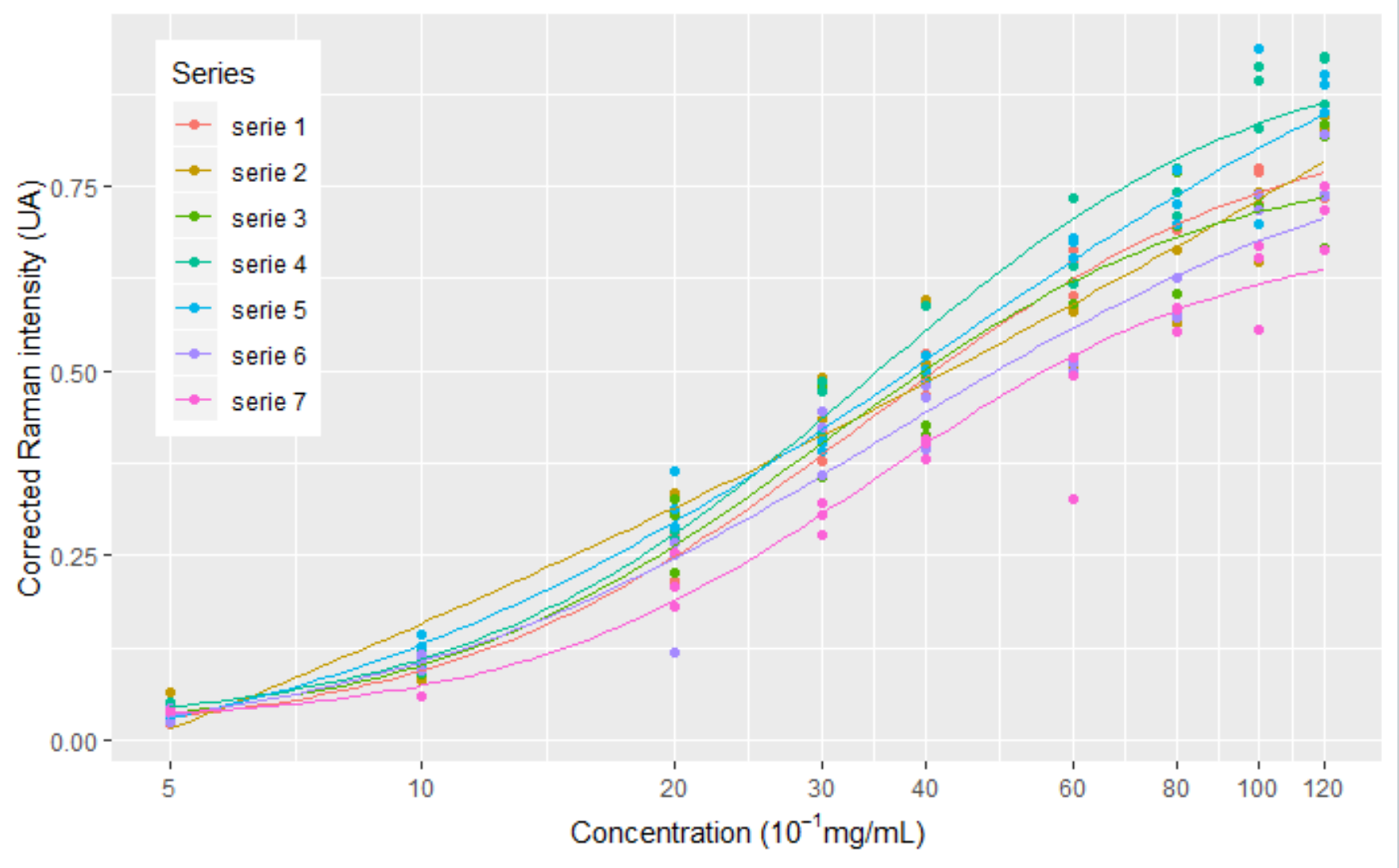
$$\mathcal{L}(\theta_{i\ell}, \xi_{i\ell} | \boldsymbol{y}, \boldsymbol{c}) = -\frac{1}{2} \sum_{j=1}^{J} \sum_{k=1}^{K} \left(\frac{(y_{ijk\ell} - f(c_j, \theta_{i\ell}))^2}{g(c_j, \theta_{i\ell}, \xi_{i\ell})} + \log(2\pi g(c_j, \theta_{i\ell}, \xi_{i\ell})) \right)$$

That is to say in the combinated residual errors model:

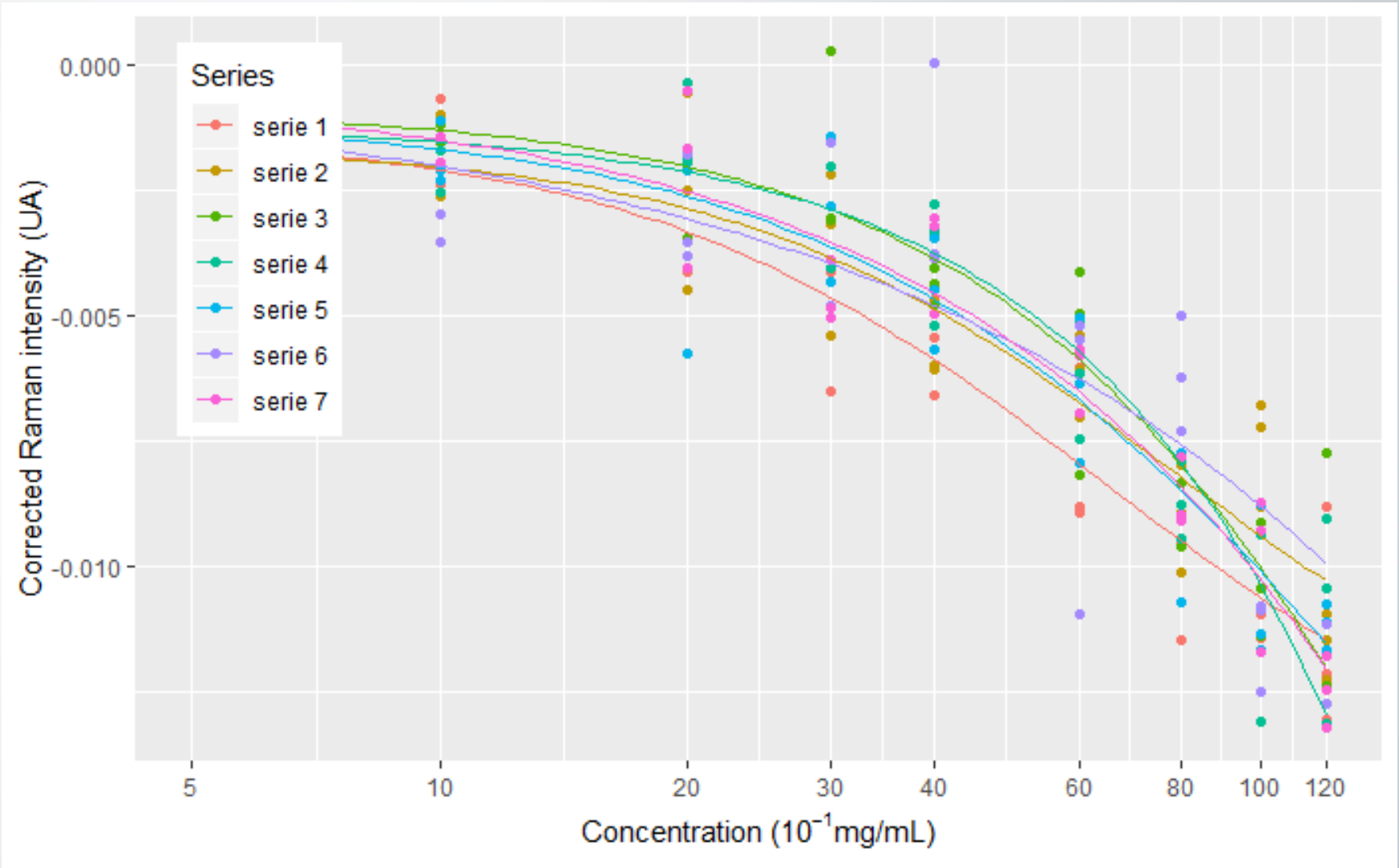
$$(\hat{\theta}_{i\ell}^{MLE}, \hat{a}_{i\ell}^{MLE}, \hat{b}_{i\ell}^{MLE}) = \arg \min_{(\theta,a,b) \in \mathbb{R}^6} \left(\sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(y_{ijk\ell} - f(c_j, \theta))^2}{a_{i\ell}^2 + b_{i\ell}^2 f(c_j, \theta_{i\ell})^2} + K \sum_{j=1}^{J} \log(a_{i\ell}^2 + b_{i\ell}^2 f(c_j, \theta_{i\ell})^2) \right).$$

▷ Optimization problem

Fit, wave number 1239



Fit, wave number 699



Predicted spectra

Based on the regression model, predicted intensities can be obtained by $\widehat{y}_{ijk\ell} = f(c_j, \widehat{\theta}_{i\ell})$, but an inverse problem is needing to estimate the 5FU concentration of new spectra.

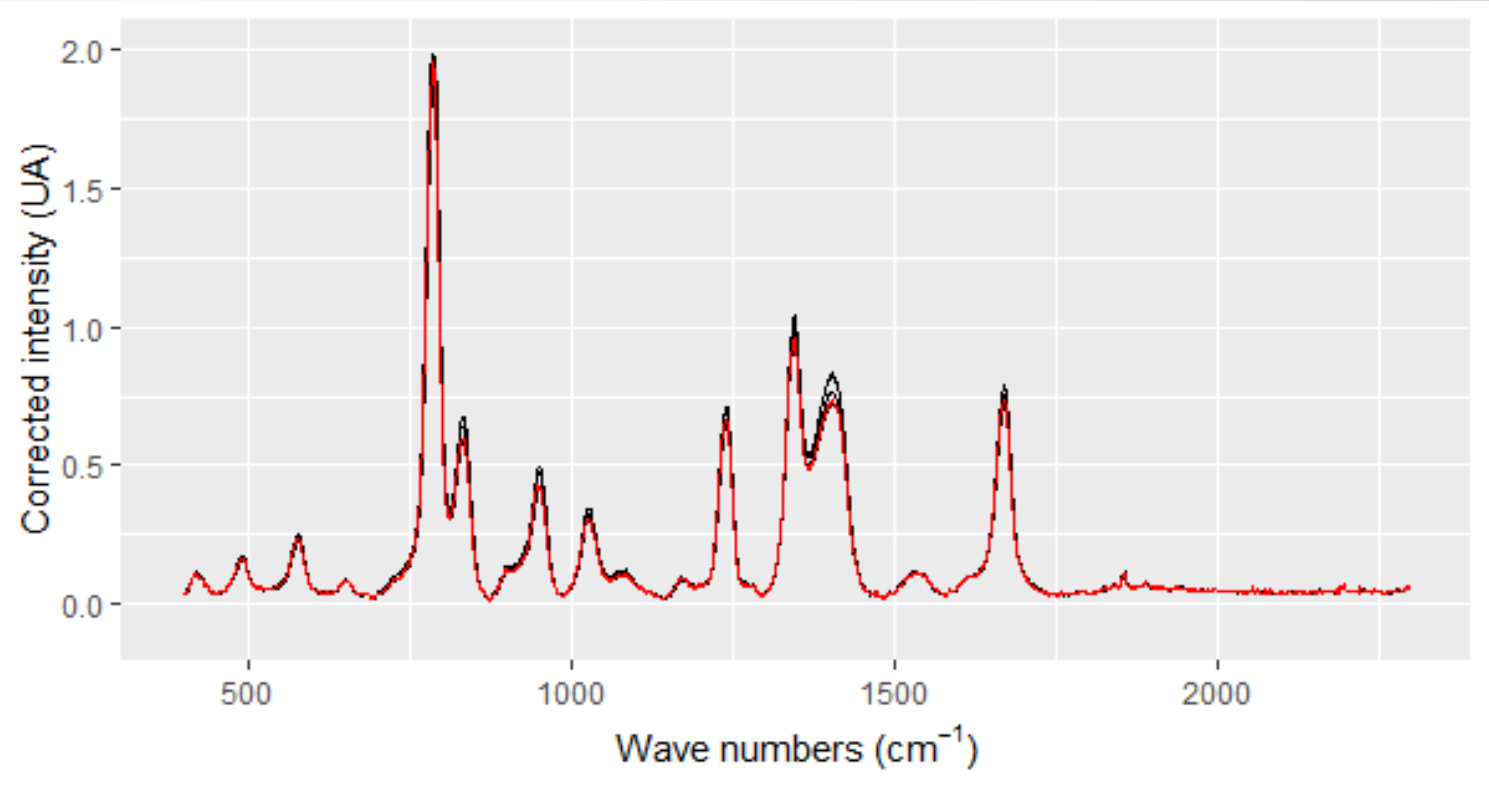


Figure: predicted SERS intensity (red) and SERS intensity (black) for the first series and concentration $0.6 mg.mL^{-1}$

Predicted spectra

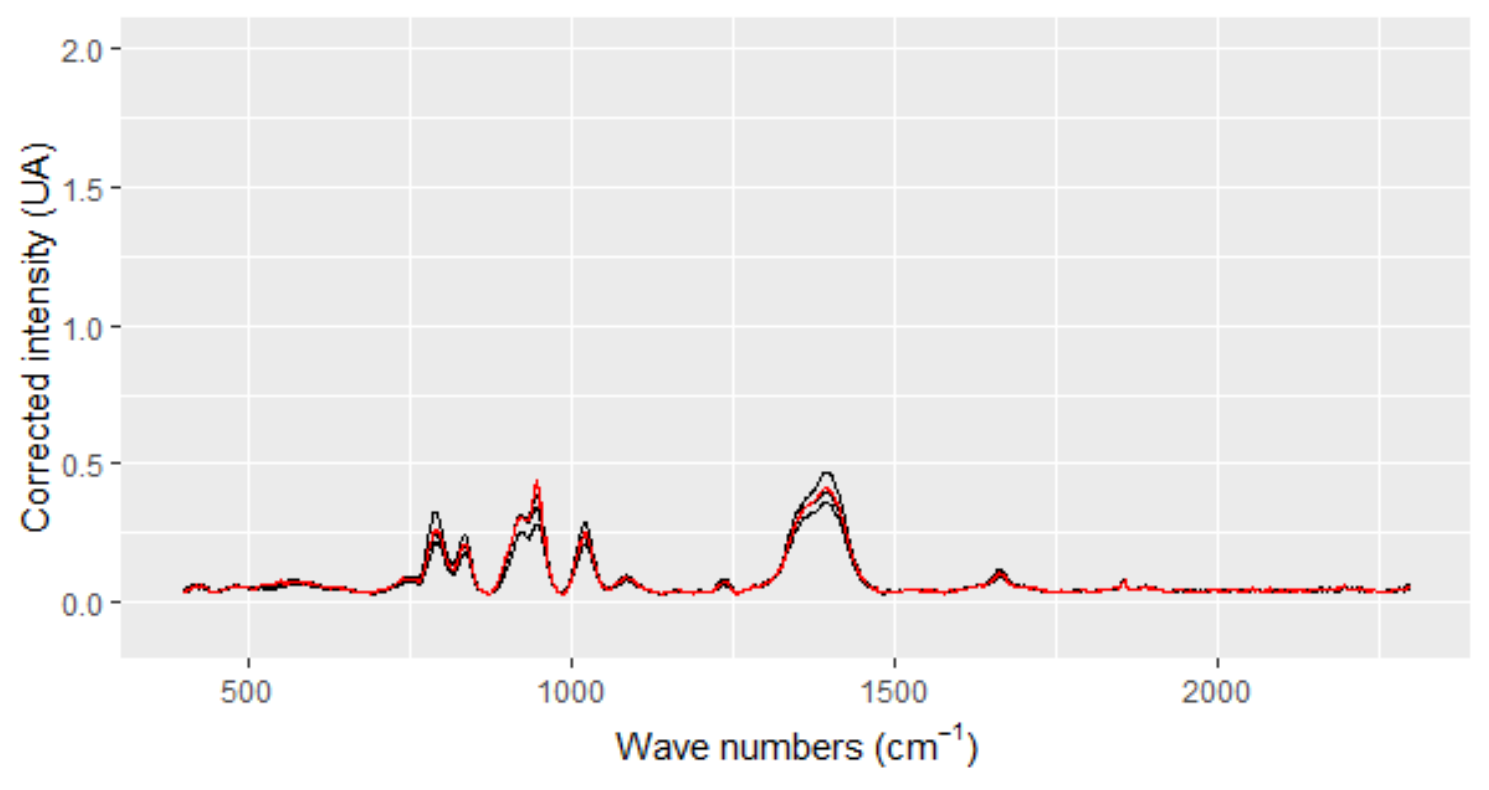


Figure: predicted SERS intensity (red) and SERS intensity (black) for the first series and concentration $0.05 mg.mL^{-1}$

Inverse problem I

A training dataset constituted with $(I-1)$ series of measurements will be used to estimate the model. The last set of measurements constituting the test dataset will be used to assess the predictive performance of the model.

Denote for $k = 1, \ldots, K$ the K repetitions of a new spectra $(y_{k\ell}^{new})_{\ell=1,\ldots,L}$ in the test dataset. A strategy to estimate the concentration of the new spectrum can be to maximize the log-likelihood of the model:

$$\mathcal{LL}_i(c|\boldsymbol{y}^{new}, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\xi}}) = -\frac{1}{2}\sum_{\ell=1}^{L}\sum_{k=1}^{K}\left(\frac{\left(y_{k\ell}^{new} - f(c, \widehat{\theta}_{i\ell})\right)^2}{g(c, \widehat{\theta}_{i\ell}, \widehat{\xi}_{i\ell})} + \log(2\pi g(c, \widehat{\theta}_{i\ell}, \widehat{\xi}_{i\ell}))\right)$$

What about i?

Inverse problem II

For measurements whose series would be known, the concentration could be estimated by

$$\hat{c}_i^{MLE} = \arg\max_{c \in \mathbb{R}+} \mathcal{LL}_i(c|\boldsymbol{y}^{new}, \widehat{\boldsymbol{\theta}}_i, \widehat{\boldsymbol{\xi}}_i).$$

In a general setting, the series is unknown. A strategy of nearest neighbors is thus considered to estimate the concentration:

$$\hat{c}^{MLE} = \arg\max_{c \in \mathbb{R}+} \max_{i \in \{1,\dots,I-1\}} \mathcal{LL}_i(c|\boldsymbol{y}^{new}, \widehat{\boldsymbol{\theta}}_i, \widehat{\boldsymbol{\xi}}_i).$$

Evaluate the errors of the model by Cross-Validation

Mean absolute relative error (MARE): $\varepsilon_i = \frac{1}{J}\sum_{j=1}^{J}\frac{|\hat{c}_{j,i} - c_j|}{c_j}$ with

- J Number of concentration to predict for the base ith
- $\hat{c}_{j,i}$ Predicted concentrations relatively to the base ith
- c_j Nominal concentrations

test dataset	cst		prop		comb	
	train	test	train	test	train	test
I1	0.083	0.103	0.068	0.051	0.056	0.090
I2	0.072	0.137	0.064	0.127	0.047	0.115
I3	0.084	0.074	0.068	0.053	0.052	0.067
I4	0.083	0.083	0.069	0.060	0.055	0.069
I5	0.083	0.069	0.068	0.086	0.055	0.087
I6	0.079	0.113	0.062	0.166	0.053	0.118
I7	0.072	0.130	0.052	0.193	0.052	0.147
Mean $\bar{\varepsilon}$	0.079	0.102	0.064	0.105	0.053	0.099

Table: Full dataset. Mean relative error (top) in the training dataset (bottom) in the test dataset

Wavenumbers selection, stepwise algorithm

In order to optimize the model, one of the strategy is to select pertinent wavenumbers. We use classical stepwise algorithm.

For M the band size and n the number of wavenumbers removed after optimization.
For a sufficiently small positive integer M,

1. Divide spectra in several bands with length M;
2. Remove successfully each bands and recalculate the mean relative errors $\bar{\varepsilon}$ for the test datasets;
3. Remove the band for which the average of the mean relative errors calculated in step 2 is the lowest;
4. While the mean relative error decreases, reiterate step 2 and 3.

Wavenumbers selection, stepwise algorithm

	l_1	l_2	l_3	l_4	l_5	l_6	l_7	mean	n
M=50	0.050	0.086	0.052	0.047	0.059	0.062	0.129	0.069	1151
M=20	0.031	0.092	0.060	0.052	0.060	0.053	0.110	0.065	1180

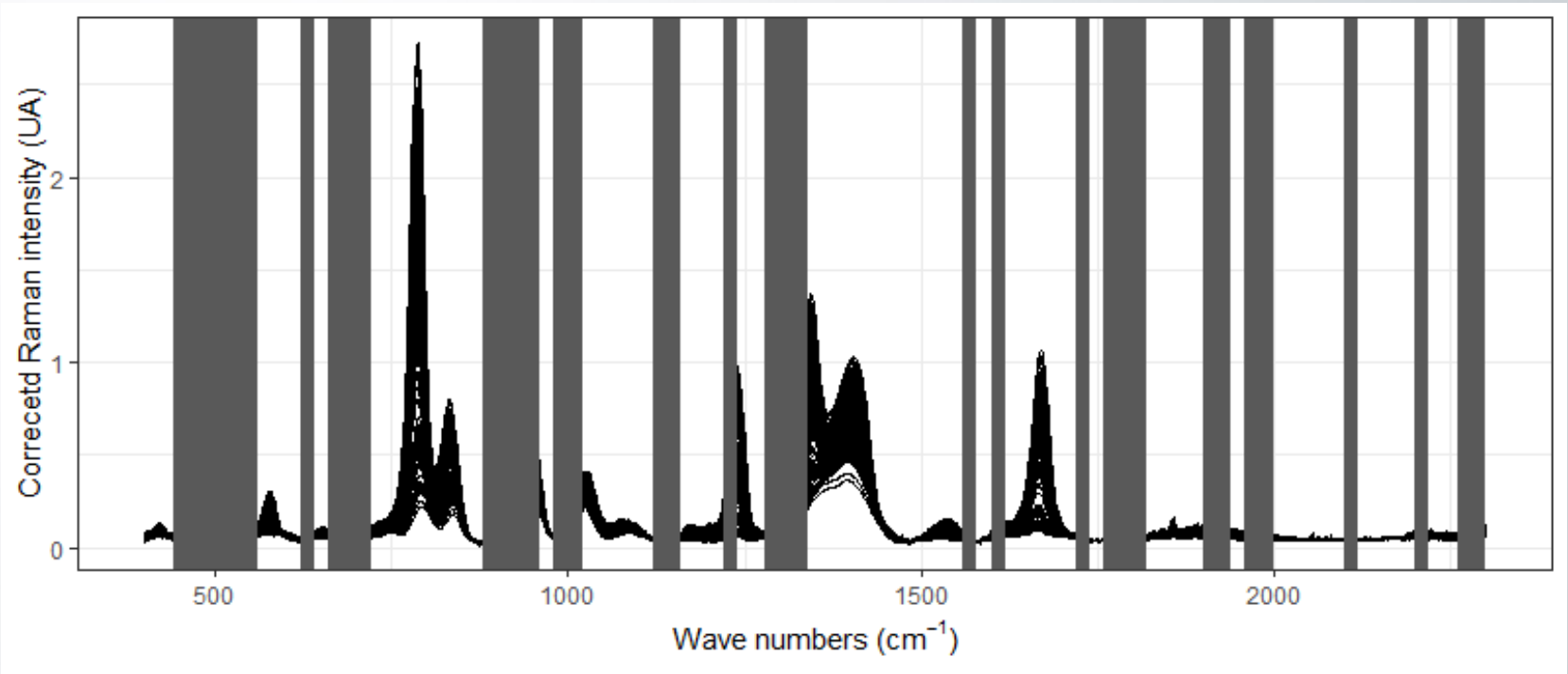


Figure: Stepwise for $M = 20$

Selection by stepwise algorithm

Table: Concentrations estimated by the model and corresponding mean absolute relative error (MARE) after wavenumbers selection (for $M = 20$)

Conc.	I_1	I_2	I_3	I_4	I_5	I_6	I_7	MARE
5	5.13	5.16	5.72	5.62	4.76	5.08	5.18	6.07
10	9.97	9.19	9.03	8.82	9.80	10.00	5.98	10.28
20	19.07	23.48	20.22	19.98	23.74	15.82	23.11	11.20
30	29.38	31.49	30.55	32.51	27.67	33.33	32.68	6.43
40	40.05	37.24	39.98	40.50	37.96	42.15	38.04	3.39
60	62.63	52.45	67.82	59.05	59.33	59.84	50.65	6.94
80	79.80	63.72	81.65	77.34	75.51	74.26	80.20	5.58
100	112.04	92.06	94.83	107.49	92.32	99.46	93.59	6.75
120	121.96	121.63	112.30	119.73	118.93	120.57	115.30	2.13
MARE	3.11	9.20	5.96	5.17	5.96	5.26	11.03	6.53

Comparison with the usual method

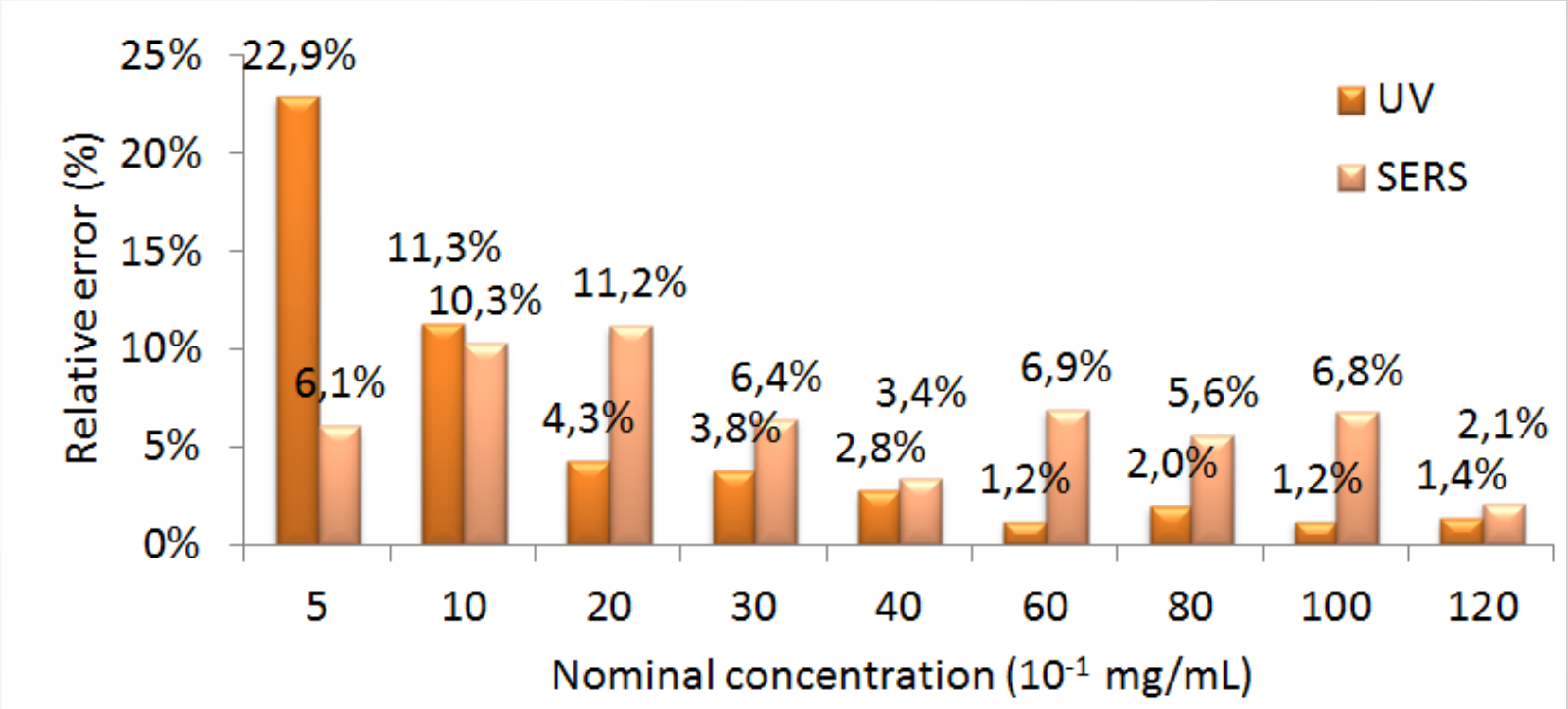


Figure: Mean absolute relative error (MARE) of concentration values predicted by UV method and the SERS model

Application to Raman spectroscopy

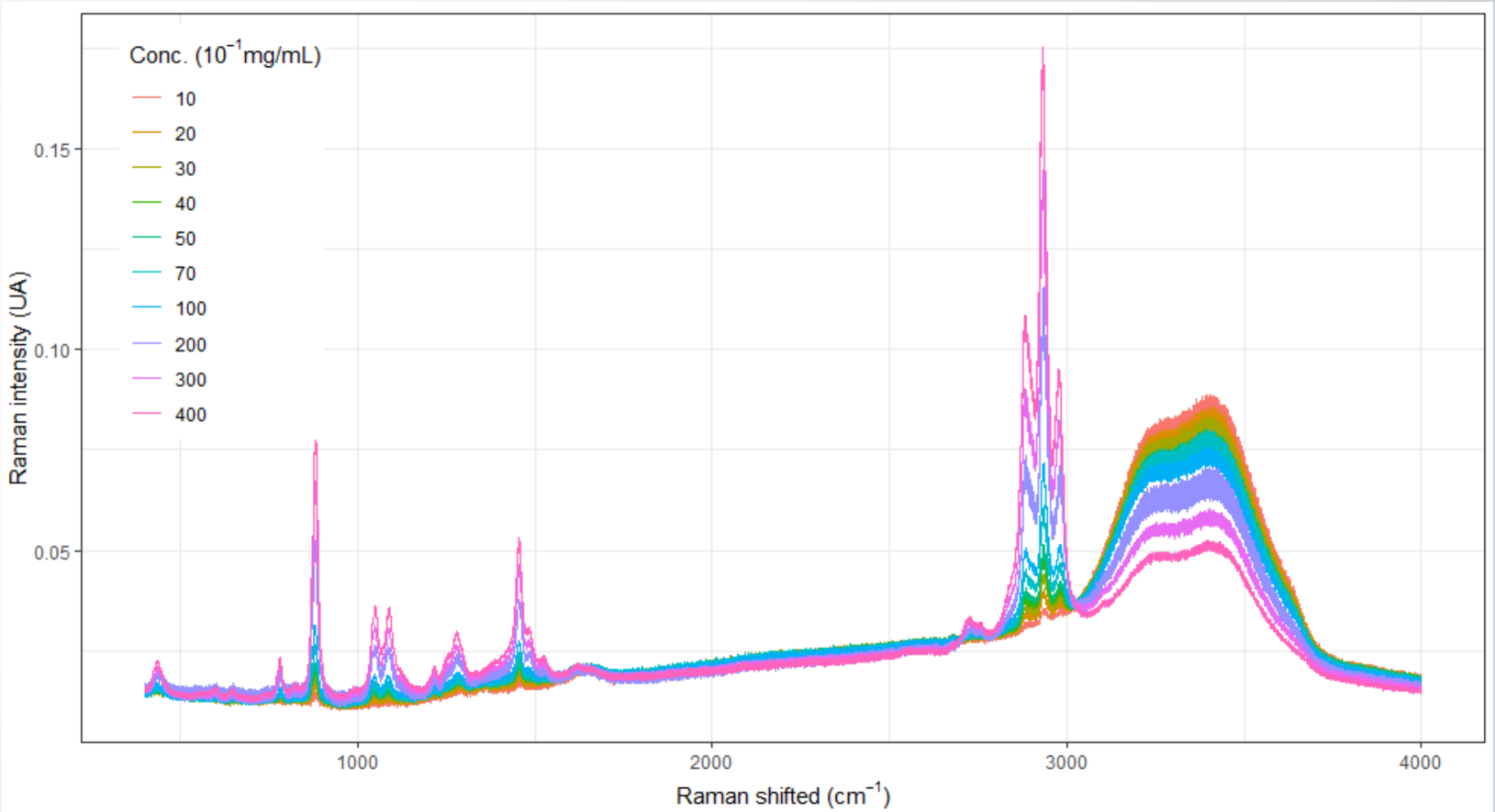


Figure: Raman spectra of Gemcitabine for 10 to 400 10^{-1} mg/mL (n=90 spectra)

Application to Raman spectroscopy

Table: Gemcitabine concentrations predicted by the model (10^{-1} mg/mL)

Conc.	I_1	I_2	I_3
10	8.57	8.37	7.23
20	19.62	20.25	21.29
30	30.53	30.23	32.18
40	40.51	41.08	42.29
50	50.72	49.84	48.07
70	70.64	72.13	70.65
100	99.21	98.68	97.75
200	199.76	199.23	197.69
300	297.66	298.16	298.03
400	402.56	390.94	402.79
MARE	2.4	2.9	5.7

Perspectives

1. Other approaches to optimize the model
   - Wavenumbers selection
   - Experiment planning (Nominal concentrations, Repetitions, series, ...
2. Transpose the model to predict concentration value in biological matrix
   - Analysis in blood, serum, plasma
   - Adapt the chemotherapy dose according to its biological concentration
   - Develop the Therapeutic drug monitoring (TDM) in oncology
   - Reduce adverse effects of chemotherapy