Sequential Monte Carlo for Bayesian Analysis of Raman Spectroscopy

Matt Moores
`mmoores@uow.edu.au`

Lecturer in Statistical Science
NIASRA, University of Wollongong

Associate Investigator
Australian Research Council Centre of Excellence
in Mathematical and Statistical Frontiers

Acknowledgements

University of Strathclyde:

- Karen Faulds
- Duncan Graham
- Kirsten Gracie
- Ivan Ramos Sasselli

Los Alamos National Laboratory:

- Kary Myers

University of Warwick:

- David Firth
- Jianyin Peng
- Jake Carson

Alan Turing Institute for Data Science:

- Mark Girolami

Raman Spectroscopy

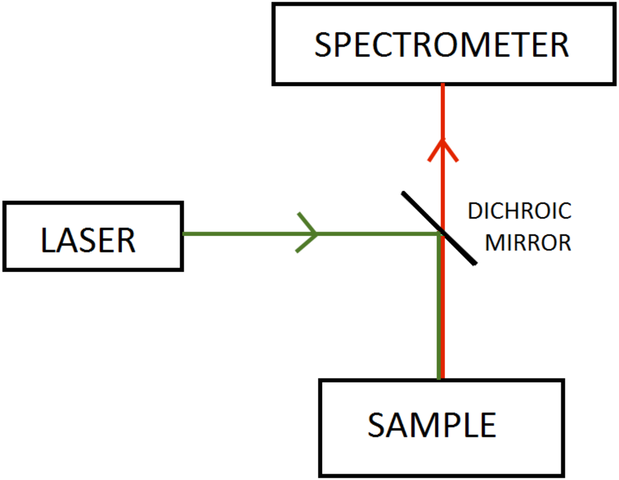
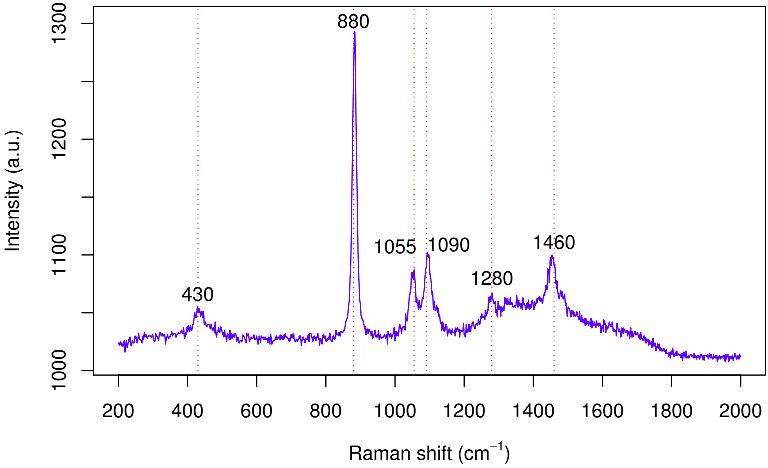


Illustration courtesy Jake Carson (U. Warwick)

Observed Spectrum



Raman scattering

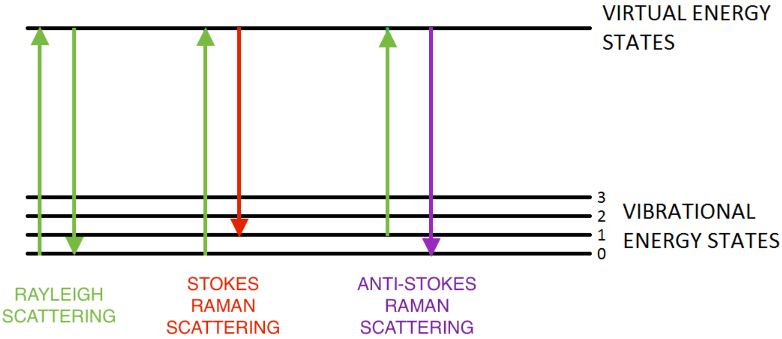


Illustration courtesy Jake Carson (U. Warwick)

Surface-enhanced Raman scattering (SERS)

- Raman signal enhanced by proximity to nanoparticles
- Functionalisation using antibodies

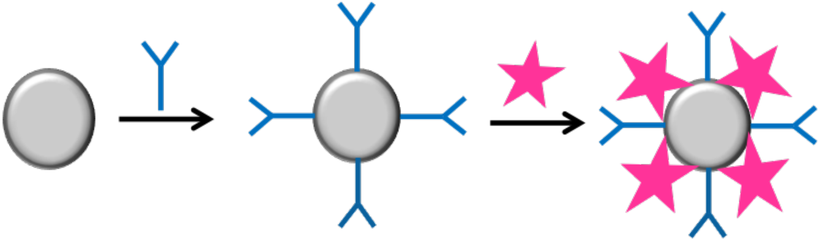
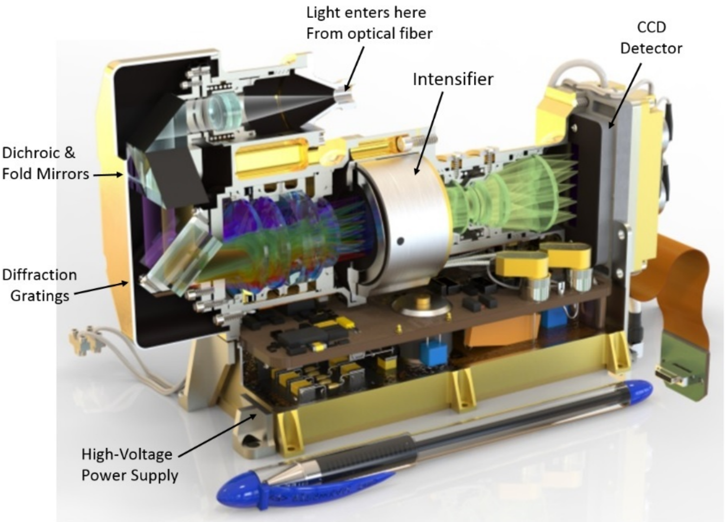


Illustration courtesy Kirsten Gracie (U. Strathclyde)

Mars 2020

SuperCam

- 532nm Raman spectroscopy at up to 12 m distance



Functional Model

Separate the hyperspectral signal into 3 components:

$$\mathbf{y}_i = \xi_i(\tilde{\nu}) + s_i(\tilde{\nu}) + \boldsymbol{\epsilon}_i \qquad (1)$$

where:

- $\mathbf{y}_i$ is a an observed spectrum, discretised at multiple wavenumbers $\nu_j \in \widetilde{\mathcal{V}}$ (cm^{-1})
- $\xi_i(\tilde{\nu})$ is a smooth baseline function
- $s_i(\tilde{\nu})$ is the spectral signature of the molecule
- $\epsilon_{i,j}$ $\sim \mathcal{N}\left(0, \sigma_\epsilon^2\right)$ is additive, zero mean white noise

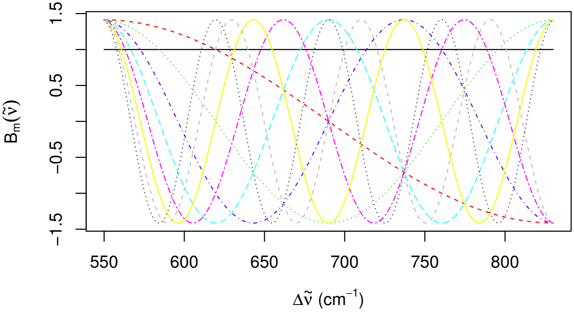
Baseline

Penalised spline:

$$\xi_i(\tilde{\nu}) = \sum_{m=1}^{M} B_m(\tilde{\nu})\alpha_{i,m} \quad (2)$$

$$\pi(\alpha_{i,\cdot}) \sim \mathcal{N}_M(0, \Sigma_\lambda) \quad (3)$$

where $B_m(\tilde{\nu})$ are Demmler-Reinsch or B-spline basis functions



Spectral Signature

An additive mixture of radial basis functions:

$$s_i(\tilde{\nu}) = \sum_{p=1}^{P} A_{i,p} f\left(\tilde{\nu} \mid \ell_p, \varphi_p\right) \tag{4}$$

where:

ℓ_p is the location of peak p

$A_{i,p}$ is the amplitude

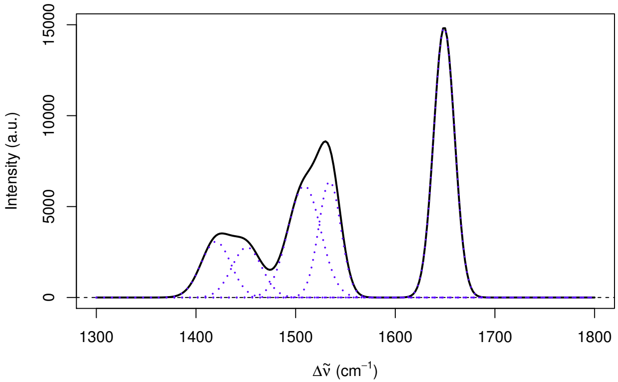
φ_p is the scale (broadening)

Squared exponential

Peak broadening function is an unnormalised Gaussian density:

$$f\left(\nu_j \mid \ell_p, \varphi_p\right) = \exp\left\{\frac{(\nu_j - \ell_p)^2}{2\varphi_p^2}\right\} \tag{5}$$

$$FWHM = 2\sqrt{2\ln 2}\,\varphi_p \tag{6}$$

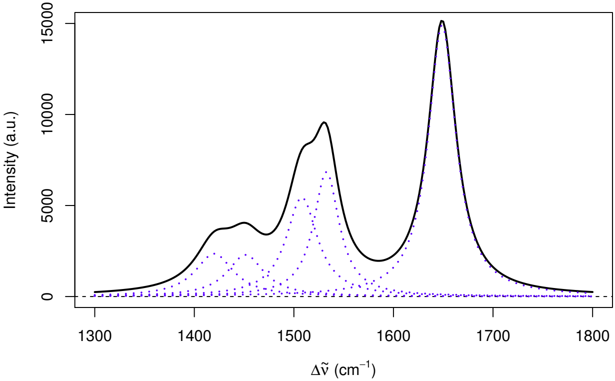


Lorentzian Peaks

Long-range dependence between peaks can be modelled using an unnormalised Cauchy density:

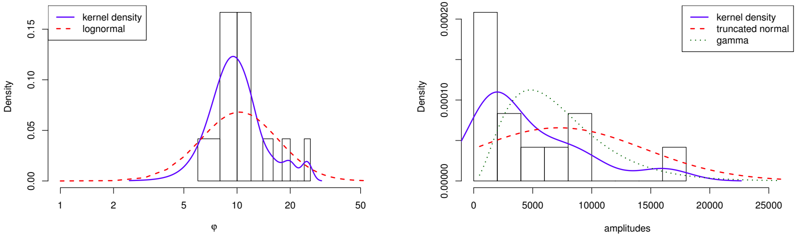
$$f\left(\nu_j \mid \ell_p, \varphi_p\right) = \frac{\varphi_p^2}{(\nu_j - \ell_p)^2 + \varphi_p^2} \tag{7}$$

$$FWHM = 2\varphi_p \tag{8}$$



Informative priors

Obtained from manual peak fitting of independent data:



(a) Scale parameters, φ (cm^{-1})

(b) Amplitudes, **A** (arbitrary units)

RRUFF Project

Anorthite R040059

Browse Search Results

<< Previous | Back to Search Results | Next >>

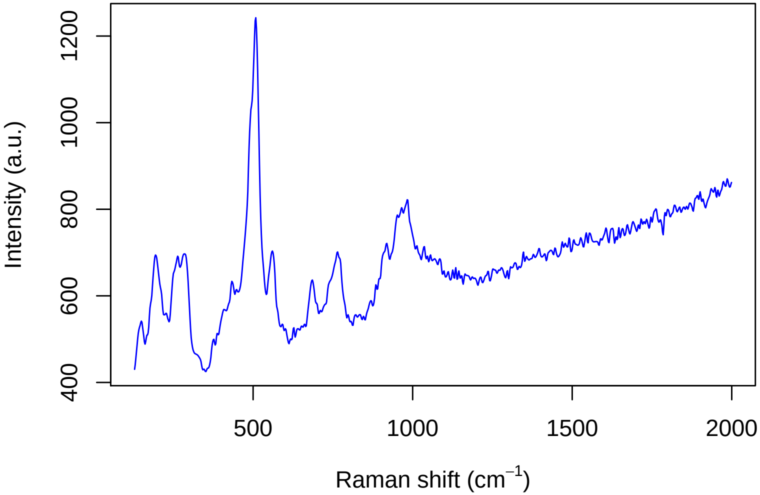
Record 11 of 282

Name: Anorthite
RRUFF ID: R040059
Ideal Chemistry: $Ca(Al_2Si_2O_8)$
Locality: Miyakejima, Japan
Source: University of Arizona Mineral Museum 4079 [view label]
Owner: RRUFF
Description: Colorless fragments of crystals
Status: The identification of this mineral has been confirmed by X-ray diffraction and chemical analysis

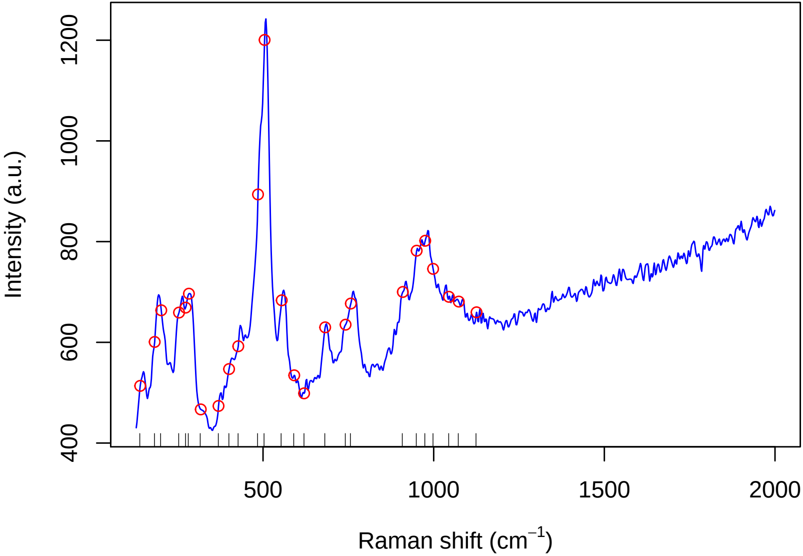
Mineral Group: [feldspar (58)]
Quick search: [All Anorthite samples (12)]

Lafuente, Downs, Yang & Stone (2015) In: *Highlights in Mineralogical Crystallography*, pp 1–30.

Raman Spectrum



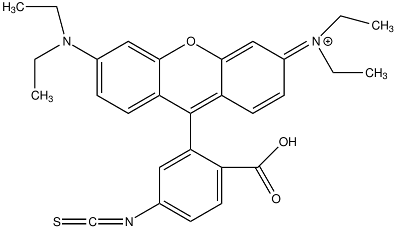
Peak Locations



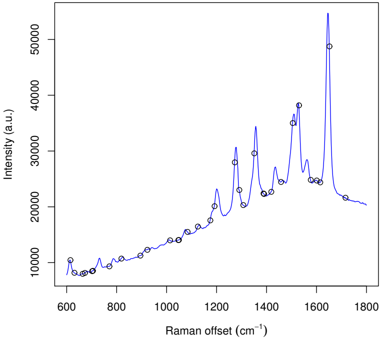
Sharma, Simons & Yoder Jr. (1983) *American Mineralogist*, **68**: 1113–1125.

Time-dependent density functional theory

Locations of Raman peaks can be predicted from chemical structure using a quantum mechanical model (TD-DFT):



(a) Rhodamine B



(b) Raman spectrum

Watanabe, Hayazawa, Inouye & Kawata (2005) *J. Phys. Chem. B* **109**(11)

TD-DFT, continued

Off-the-shelf implementations:

- Gaussian 09, Amsterdam Density Functional (ADF), Quantum Espresso (QE), etc.
- B3LYP functional with the basis set of $6\text{-}311{+}{+}G(d,p)$

Rhodamine B: $C_{28}H_{31}ClN_2O_3$ (PubChem ID 6694)

1. Use the crystal structure of the molecule as an initial geometry
2. Optimise to obtain the resting state (energy minimum)
3. Calculate potential energy distributions for the vibrational modes (C–H out-of-plane bend, C=C symmetric stretch, etc.)
4. Apply selection rules to determine Raman scattering frequencies and infrared absorption frequencies

Becke (1993) *J Chem. Phys.* **98**: 5648.
Lee, Yang & Parr (1988) *Phys. Rev. B* **37**: 785.

Markov chain Monte Carlo

MCMC targeting the joint posterior $\pi\left(\mathbf{A}, \boldsymbol{\varphi}, \boldsymbol{\ell} \mid y_i(\tilde{\nu})\right)$

Marginal likelihood is available in closed form:

$$\begin{aligned} p(y_i(\tilde{\nu}) \mid \mathbf{A}, \boldsymbol{\varphi}, \boldsymbol{\ell}) &= \int\int p(y_i(\tilde{\nu}) \mid \boldsymbol{\Theta})\, \pi(\boldsymbol{\alpha})\, \pi(\sigma_\epsilon^2)\; d\boldsymbol{\alpha}\, d\sigma_\epsilon \\ &= \frac{p(y_i(\tilde{\nu}) \mid \boldsymbol{\Theta})\, \pi(\boldsymbol{\alpha})\, \pi(\sigma_\epsilon^2)}{p(\boldsymbol{\alpha}, \sigma_\epsilon^2 \mid y_i(\tilde{\nu}), \mathbf{A}, \boldsymbol{\varphi}, \boldsymbol{\ell})} \end{aligned}$$

Given random walk proposals for $\mathbf{A}', \boldsymbol{\varphi}', \boldsymbol{\ell}'$,
accept with probability $\min(1, \rho_t)$ where:

$$\rho_t = \frac{p(y_i(\tilde{\nu}) \mid \mathbf{A}', \boldsymbol{\varphi}', \boldsymbol{\ell}')\pi(\mathbf{A}')\pi(\boldsymbol{\varphi}')\pi(\boldsymbol{\ell}')}{p\left(y_i(\tilde{\nu}) \mid \mathbf{A}^{(t-1)}, \boldsymbol{\varphi}^{(t-1)}, \boldsymbol{\ell}^{(t-1)}\right) \pi(\mathbf{A}^{(t-1)})\pi(\boldsymbol{\varphi}^{(t-1)})\pi(\boldsymbol{\ell}^{(t-1)})}$$

Chib (*JASA*, 1995) “Marginal Likelihood from the Gibbs Output.”

Rao-Blackwellized Particle Filter

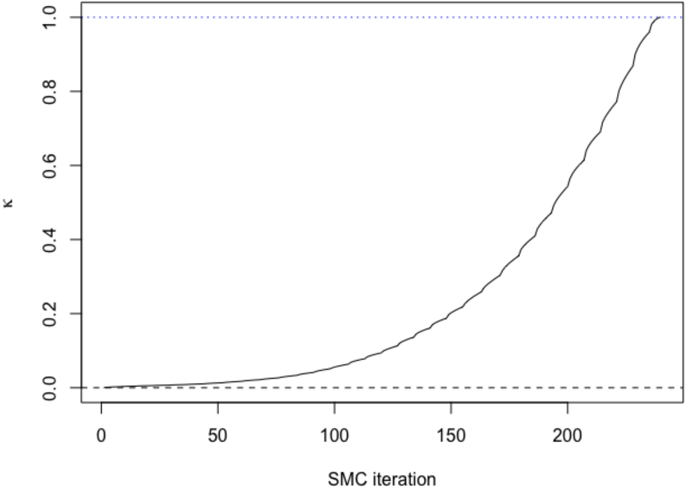
Particle-based method targeting a sequence of partial posteriors $\pi_t\left(\mathbf{A}, \varphi, \ell \mid y_i(\tilde{\nu})\right)$

Algorithm 1 SMC

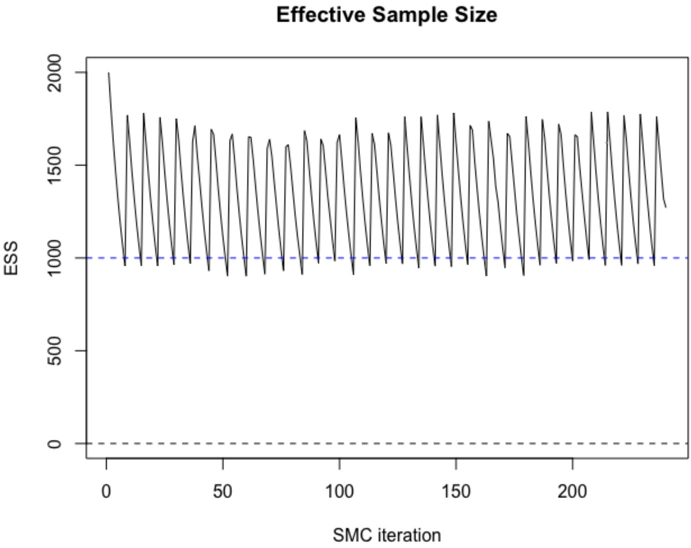
```
Initialise $\varphi^{(q)}$, $\mathbf{A}^{(q)}$, $\ell^{(q)}$ $\forall\, q \in \{1, \ldots, Q\}$
Initialise importance weights, $w_0^{(q)} = \frac{1}{Q}$
for all iterations $t = 1, \ldots, T$ do
    Update importance weights:
        $$w_t^{(q)} \propto w_{t-1}^{(q)} \frac{p\left(y_i(\tilde{\nu}) \mid \ell, \mathbf{A}, \varphi\right)^{\kappa_t}}{p\left(y_i(\tilde{\nu}) \mid \ell, \mathbf{A}, \varphi\right)^{\kappa_{t-1}}} \qquad (9)$$
    Resample particles if $ESS_t$ is below threshold
    for all particles $q \in \{1, \ldots, Q\}$ do
        Update $\varphi^{(q)}$, $\mathbf{A}^{(q)}$, $\ell^{(q)}$ using MCMC steps
    end for
end for
```

Del Moral, Doucet & Jasra (2006)

Likelihood Tempering



ESS


Effective Sample Size
ESS
2000
1500
1000
500
0
0
50
100
150
200
SMC iteration

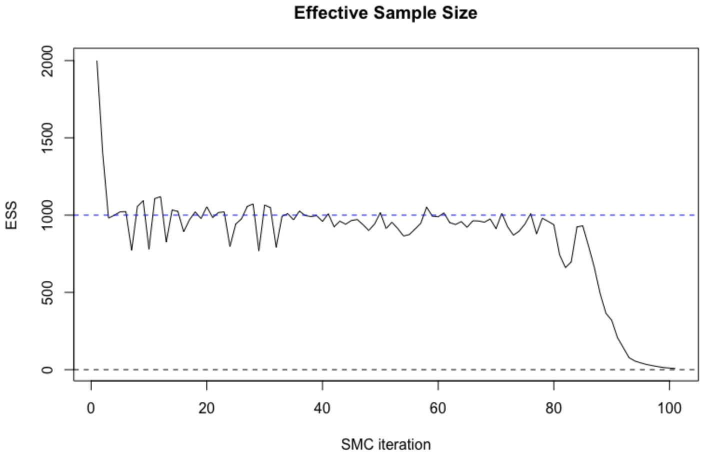
Resampling

- Multinomial *(bootstrap particle filter)*
- Systematic
- Stratified
- Residual

Can parallelize by ordering the ancestry vector

Douc, Cappé & Moulines (*Proc. 4th IEEE ISPA*, 2005) "Comparison of resampling schemes for particle filtering."
Murray, Lee & Jacob (*JCGS*, 2016) "Parallel resampling in the particle filter."

SMC collapse



serrsBayes

An R package for Bayesian modelling and quantification of Raman spectroscopy using sequential Monte Carlo (SMC) algorithms:

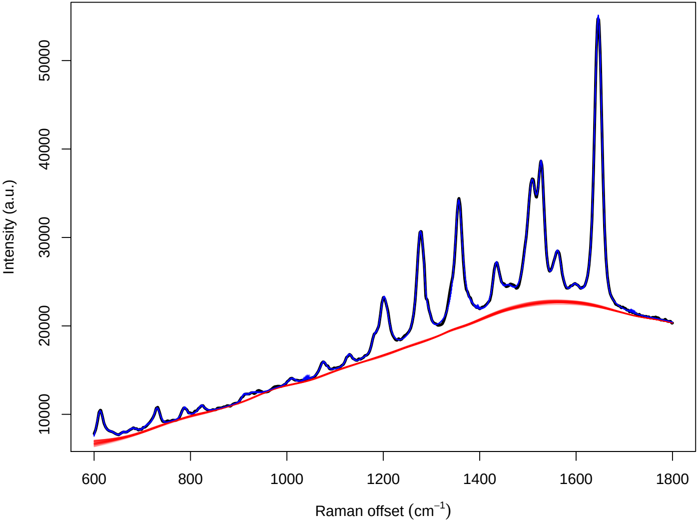
- RcppEigen for fast linear algebra in C++
- OpenMP for parallelism

```
library(serrsBayes)
library(hyperSpec)

spec ← read.spc("spectrum.spc")
lPriors ← list(..)
result ← fitSpectraSMC(spec$wl, spec$spc, lPriors)
```

Bates & Eddelbuettel (2013) Fast and Elegant Numerical Linear Algebra Using the `RcppEigen` Package. *J. Stat. Soft.* **52**(5): 1–24.
Beleites & Sergo (2014) `hyperSpec`: a package to handle hyperspectral data sets in R.

Posterior distribution



Summary

serrsBayes provides an open-source approach to analysis of spectroscopy:

- Joint estimation of baseline and peaks
- 95% CI for peak locations, amplitudes, and FWHM

Ongoing and future work:

- Scalable computation using a divide-and-conquer algorithm
- T-optimum experimental design for multiplex Raman
- Spatial and temporal modelling of Raman maps
- Other types of spectroscopy (RF, X-ray, LIBS)

For Further Reading I

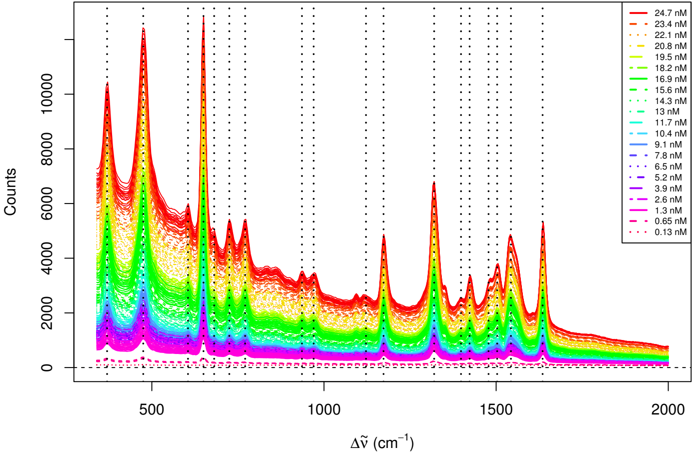
- Moores, Gracie, Carson, Faulds, Graham & Girolami
  Bayesian modelling and quantification of Raman spectroscopy.
  arXiv preprint `arXiv:1604.07299` [stat.AP]
- Noonan, Asiala, Grassia, MacRitchie, Gracie, Carson, Moores, et al.
  In vivo multiplex molecular imaging of vascular inflammation using surface-enhanced Raman spectroscopy.
  To appear in *Theranostics*.
- Gracie, Moores, Smith, Harding, Girolami, Graham, & Faulds
  Preferential attachment of specific fluorescent dyes and dye labelled DNA sequences in a SERS multiplex.
  Anal. Chem., 88(2): 1147–1153, 2016.
- Zhong, Girolami, Faulds & Graham
  Bayesian methods to detect dye-labelled DNA oligonucleotides in multiplexed Raman spectra.
  J. R. Stat. Soc. Ser. C, 60(2): 187–206, 2011.

For Further Reading II

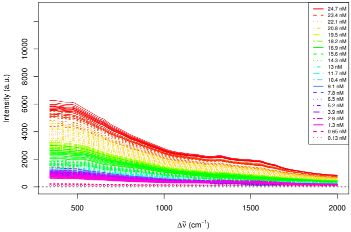
- Särkkä, Vehtari & Lampinen
  Rao-Blackwellized particle filter for multiple target tracking
  Information Fusion 8(1): 2–15, 2007.
- R. Douc, O. Cappé & E. Moulines
  Comparison of resampling schemes for particle filtering
  In *Proc. 4th IEEE Int. Symp. Image and Signal Processing and Analysis*, 2005.
- L.M. Murray, A. Lee & P.E. Jacob
  Parallel resampling in the particle filter
  J. Comput. Graph. Stat. **25**(3): 789–805, 2016.
- F. Lindsten, A.M. Johansen, C.A. Naesseth, B. Kirkpatrick, T.B. Schön, J.A.D. Aston & A. Bouchard-Côté
  Divide-and-conquer with sequential Monte Carlo
  J. Comput. Graph. Stat. **26**(2): 445–458, 2017.

Dilution study for FAM

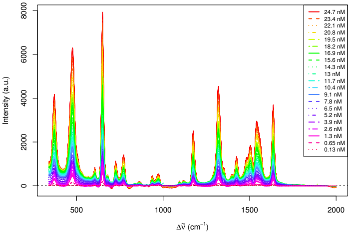
315 spectra at 21 different concentrations, from 0.13 to 24.7 nM



Results: Baseline correction



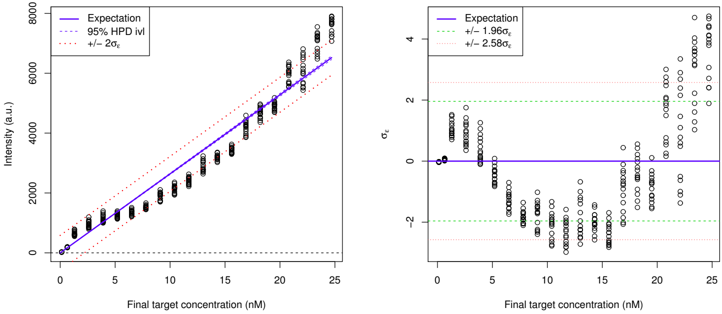
(a) Posterior means of the baselines



(b) Baseline-corrected spectra

Results: Quantification

SERS peak intensities at 650cm^{-1}: 95% CI [257.7; 262.5] $\times c_i$



(a) Linear regression for β_p

(b) Standardised residuals