Probabilistic signal estimation for vibrational spectroscopy with a flexible non-stationary Gaussian process baseline model

David Frich Hansen[a,*], Tommy Sonne Alstrøm[a], Mikkel N. Schmidt[a]

[a] *DTU Compute, Technical University of Denmark, Richard Petersens Plads 324, 2800, Kgs. Lyngby, Denmark*

Abstract

Vibrational spectroscopy techniques enable accurate chemical detection and quantification, but the extraction of spectral peak parameters is frequently hampered by an underlying baseline. Because the signal and baseline are additive, it is difficult to distinguish between signal peaks and baseline effects when the baseline is not smooth. Using surface enhanced Raman spectroscopy (SERS) and near-infrared (NIR) spectroscopy as examples, we show how to estimate the signal and the baseline jointly while imposing a high-capacity non-stationary Gaussian process on the baseline. This allows us to both obtain accurate estimation and meaningful uncertainty estimates on interpretable peak parameters. We demonstrate this on artificially generated SERS maps, a challenging real-world SERS case, and a benchmark NIR dataset.

Keywords: Surface enhanced Raman spectroscopy, SERS, Near-infrared spectroscopy, NIR, Gaussian processes, baseline correction, vibrational spectroscopy, Markov chain Monte Carlo, MCMC

1. Introduction

A critical task in chemical science is determining the composition of unknown compounds, a process known as fingerprinting. Vibrational spectroscopy is a class of such fingerprinting techniques that works by measuring changes in the vibrational state of molecules in response to light interaction. Near-infrared spectroscopy (NIR), infrared spectroscopy (IR), and Raman spectroscopy (Siesler, 2016) are examples of vibrational techniques that can be used to detect, identify, or quantify chemical compounds in a substance. Irradiating the sample and measuring the energy absorbed, transmitted, or inelastically scattered (de-

*Corresponding author

Email addresses: dfha@dtu.dk (David Frich Hansen), tsal@dtu.dk (Tommy Sonne Alstrøm), mnsc@dtu.dk (Mikkel N. Schmidt)

pending on the technique) at a range of wavelengths yields a spectrum with a pattern of spectral lines that characterize the molecular structure.

The detection procedure then entails identifying a distinctive set of spectral features associated with the compound. The measured spectrum can be compared to a library of known spectra to identify a compound among a set of possibilities.

Typically, quantification is accomplished by obtaining a measure of peak intensity, such as the magnitude of a specific mode or the area under one or more characteristic spectral lines. A calibration curve can then be used to map these numbers to a concentration. The requirement to precisely and reliably determine the location, magnitude, and width of the spectral lines is shared by all of these tasks.

Spectral imaging is a popular acquisition method in which the sample is measured in a spatial grid rather than at a single point. A spectral map has several advantages over a single spectrum, including the ability to reduce noise and unmix spectral information using multivariate statistical techniques (Tauler et al., 1993). Furthermore, if the sample being analyzed is spatially inhomogeneous, having multiple spectra is extremely beneficial because each spectrum can convey vastly different information. While multivariate statistical techniques provide significant analysis benefits, they typically require additional post-processing to extract important spectral characteristics such as peak locations.

This work is centered on spectral imaging, specifically surface enhanced Raman spectroscopy (SERS), but the methods presented are general and easily extendable to other types of vibrational spectroscopy. While the infrequent occurrence of inelastic scattering is a significant drawback of Raman scattering, in SERS (Fleischmann et al., 1974; Jeanmaire and Van Duyne, 1977; Albrecht and Creighton, 1977) the material of interest is adsorbed on a nano-structured surface (hereafter referred to as a substrate) which locally amplifies the electrical field and greatly increases the signal strength. Since its discovery, SERS has been a popular analysis method at very low concentration levels.

Typical statistical analysis methods include multivariate curve resolution, non-negative matrix factorization, and peak parameter fitting; however, these methods commonly ignore uncertainty in measurements reflected in estimated model parameters and can have difficulty identifying variation between samples of the same material. Furthermore, SERS measurements are often contaminated with an unknown baseline, making it difficult to extract the pure underlying Raman signal. Many of the common analysis methods necessitate the removal of such baseline, and because the signal and baseline are additive, inaccurate baseline estimation introduces errors in the estimation of the SERS signal. When the baseline is complex and not spatially uniform due to random variation across the SERS map, it is particularly challenging to distinguish the signal from the background, and suitable scattering correction is necessary.

The choice of baseline removal methods for spectroscopy are many, and numerous work has been carried out in the last decades. A good overview of the typical baseline removal procedure is given by Yang et al., where the pipeline

is described as a three-step procedure, 1) smoothing, 2) baseline correction, 3) peak picking (Yang et al., 2009). This pipeline shows there are significant choices that need to be made. One can design the processing as three individual units, or have an algorithm that estimates baseline and peak-picking jointly. Methods that simultaneously estimate baselines and peaks started to receive attention in the last ten years and can work with either methods that can learn to ignore the baseline (Schmidt et al., 2019), or methods that explicitly model both the baseline and peak identification (Picaud et al., 2018).

Turning the attention to more classical methods, where the baseline is considered without taking occurrences of peaks into account, a popular approach is to fit a low-order polynomial model to the spectrum and use that as the baseline. Here the experimenter needs to choose the order of the polynomial which can have a high impact on the quality of the baseline estimation. (Gornushkin et al., 2003). These methods are rudimentary and prone to estimating crucial parts of the spectrum as part of the baseline. Other approaches that are more robust towards tuning of model parameters are based on asymmetric least squares which ensures the baseline estimation matches the trend of the spectrum and ignore peaks (Eilers, 2004). An improved method that innovates this idea is adaptive iteratively reweighted penalized least squares (airPLS) (Zhang et al., 2010), which can be used without any user input. The method works by iteratively changing the weights of the errors between the estimated baseline and the original spectrum. Further reading on baseline removal methods is found in (Yang et al., 2009; Li et al., 2020; Pan et al., 2022; Chen et al., 2023).

We propose an algorithm that performs simultaneous estimation of baseline and parameters of the spectral characteristics while performing additive and multiplicative scattering correction. The main contribution is a novel approach to modeling the baseline using a non-stationary Gaussian process. This enables the model to separately adapt the baseline model's flexibility within and outside of spectral regions with the target SERS signal, which significantly improves the signal estimation compared with existing approaches. Because our method is written in a Bayesian framework, posterior uncertainty is naturally provided for all estimated signal parameters.

2. Methods and theory

Let $X \in \mathbb{R}^{N \times W}$ be the observed spectral map, where $N = N_x \cdot N_y$ is the number of observed spectra and W is the number of observed wavenumbers such that $X_{n,w}$ is the measured intensity at wavenumber index w in spectrum n (see Figure 1). We model the signal and baseline as additive, such that the observed spectral map is given by

$$X_{n,w} = I_{n,w} + L_{n,w} + \varepsilon_{n,w} \tag{1}$$

where I is our model for the underlying signal, L is our baseline model, and ε is additive noise.

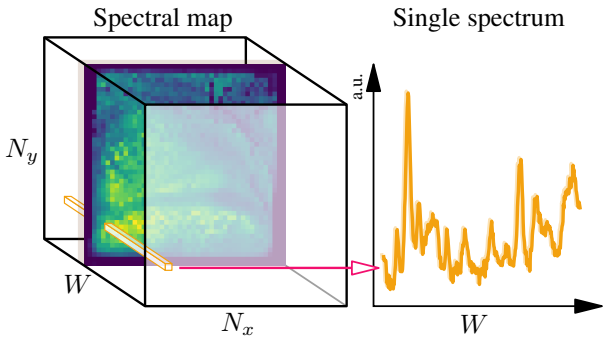

Figure 1: A spectral map containing $N = N_x \cdot N_y$ spectra measured at W wavenumbers, $\omega_1, \ldots \omega_W$. A "slice" of the map is visualized as an image, and a "fibre" corresponding to single spectrum is visualized as a line graph.

This additive model highlights a fundamental challenge with modeling spectral maps: components of the observed data can equivalently be attributed to the three model parts: signal, baseline, or noise. To separate and identify the three parts, further constraints are necessary. In a Bayesian setting, such constraints are in the form of functional and distributional assumptions. Typically, the spectral signal is modeled using a parametric peak model, the baseline follows a smooth and not too flexible model, for example a spline, and the noise is assumed independent and Gaussian.

In this paper, we address a problem that arises frequently in practice when the observed baseline is not sufficiently smooth. In that case, a more expressive baseline model might be chosen; however, due to additivity, a more flexible baseline will compete with the signal model to explain the observed data, resulting in poorer estimates of the peak parameters. To address this problem, we propose using a very flexible non-parametric baseline model based on a Gaussian process (GP), but limiting its flexibility by increasing smoothness around the peak locations where the signal model is active.

To highlight the benefits of our locally smooth (non-stationary) GP baseline model, we contrast it with the commonly used smooth B-spline background (Han and Ram, 2020). In addition, we compare with a manually selected, locally linear baseline (Göksel et al., 2021) as well as no baseline correction, while keeping all other modeling choices the same.

2.1. Pseudo-Voigt spectral model

We model the spectral peaks using a pseudo-Voigt curve (Alstrøm et al., 2017; Demtröder, 2014; Li and Dai, 2012), defined as a linear interpolation between a Lorentzian and a Gaussian lineshape,

$$V(\omega; c, \gamma, \eta) = \eta \underbrace{\frac{1}{1 + \left(\frac{\omega - c}{\gamma}\right)^2}}_{\mathcal{L}} + (1 - \eta) \underbrace{\exp\left[-\frac{1}{2}\left(\frac{\omega - c}{\gamma}\right)^2\right]}_{\mathcal{G}} \tag{2}$$

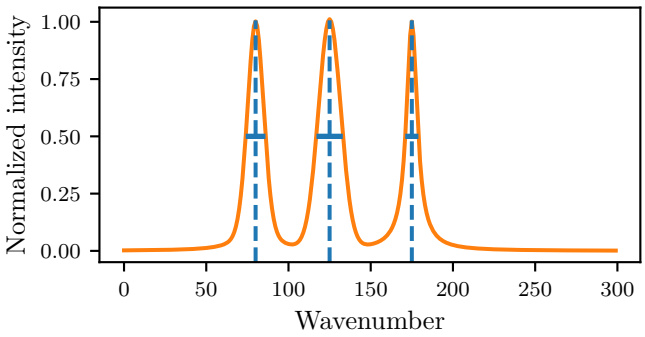

Figure 2: Pseudo-Voigt curve. The vertical lines shows the locations, c_k, and horizontal lines indicate the full width-at-half-maximums, $2\gamma_k$.

where $\mathcal{L}$ and $\mathcal{G}$ are the Lorentzian and Gaussian lineshapes with location c and scale γ, and evaluated at wavenumber ω. $\eta \in [0, 1]$ denotes the *Lorentzianity* of the curve, and informally determines its sharpness. This formulation of the pseudo-Voigt curve is height-normalized to unity. Examples of such a curves can be seen in Figure 2. In our setting, c denotes the position on the wavenumber axis, and 2γ is the full width at half maximum (FWHM).

Since many analytes have multiple Raman peaks, we gather K pseudo-Voigt curves in a matrix $V \in \mathbb{R}^{K \times W}$ with components

$$V_{k,w} = V(\omega_w; c_k, \gamma_k, \eta_k). \tag{3}$$

As mentioned previously, the enhancement of the SERS substrate is non-uniform, and the analyte may be distributed unevenly across the substrate. To model this, we introduce a set of spectrum and peak specific amplitude parameters, $\alpha_{n,k}$, which gives us our full pseudo-Voigt spectral peak model,

$$I_{n,w} = \sum_{k=1}^{K} \alpha_{n,k} V_{k,w}. \tag{4}$$

Under this model, each spectrum in the map share peak locations and shapes, but are allowed different peak amplitudes.

2.2. Baseline model

We assume that the shape of the baseline is shared across the SERS map, and varies spatially only with a multiplicative scale, β_n, and an additive bias, κ_n. This gives a rank-1 baseline model,

$$L_{n,w} = \beta_n B_w + \kappa_n. \tag{5}$$

In case this assumption is too crude, it is straightforward to generalize to a rank-P baseline model, where P baseline shapes are shared across the map with individual scales and biases,

$$L_{n,w} = \sum_{p=1}^{P} \beta_{n,p} B_{w,p} + \kappa_{n,p}. \tag{6}$$

This is relevant, for instance, when numerous separate contaminants contribute to the baseline; nevertheless, we limit our exposition to $P = 1$, which we found to be adequate in our practical setting.

2.3. Noise model

We assume that the observation noise is independent and identically Gaussian,

$$\varepsilon_{n,w} \sim \mathcal{N}(0, \tau^{-1}), \tag{7}$$

with mean zero and precision τ. While this assumption works well in most practical situations, it is occasionally useful to extend it to handle e.g. shot noise caused by external disturbances.

2.4. Full model

Combining the signal, baseline, and noise models yields full SERS model,

$$X_{n,w} = \underbrace{\sum_{k=1}^{K} \alpha_{n,k} V_{k,w}}_{I_{n,w}} + \underbrace{\beta_n B_w + \kappa_n}_{L_{n,w}} + \varepsilon_{n,w}. \tag{8}$$

Left is to define the baseline, B, specify prior distributions for the model parameters, and device a suitable inference procedure, which we cover in the sections that follow.

2.5. Non-stationary Gaussian process baseline

A Gaussian process (GP) is a stochastic process $\{Y_t : t \in \mathcal{T}\}$ where every finite realization over a subset of the index set $\mathcal{T}$ have consistent multivariate Gaussian distributions. The process is completely specified by its mean function, $m(t)$ and its covariance function $c(t, t')$. While not many restrictions are put on the mean function, it is a requirement that the covariance function is positive definite. This means that the kernel matrix, $K \in \mathbb{R}^{n \times n}$, with components

$$K_{i,j} = c(t_i, t_j), \tag{9}$$

is positive definite for any $t_1, \ldots, t_n \subset \mathcal{T}$. While it is not trivial to construct a positive definite function, it is relatively easy to make new constructions by composing known positive definite functions using a set of combination rules (Bishop, 2006). For a thorough introduction on GP's for machine learning, we refer to the text by Rasmussen and Williams (2005).

In regression, it is common to use a stationary GP, where the covariance only depends on the distance between inputs. A commonly used kernel, which we will use as a starting point, is the squared exponential,

$$c(t, t') = \nu \cdot \exp\left(-\frac{\|t - t'\|^2}{2\ell^2}\right). \tag{10}$$

Here, $t, t' \in \mathbb{R}^D$ are two (vector) inputs, $\|\cdot\|$ denotes any norm (usually the Euclidean), ℓ^2 is a length scale parameter which controls flexibility of the GP (how quickly it fluctuates), and ν determines the function variance.

The problem with using a stationary GP to model the spectral baseline is that it has the same level of flexibility everywhere. With too much flexibility, the baseline model might capture parts of the Raman signal of interest, and with too little flexibility a complex baseline is not modelled in sufficient detail: In both cases this can lead to inaccurate estimation of the signal parameters.

One idea could be to let the length scale vary with t in a squared exponential kernel; however, this does not lead to a valid covariance function without significant changes to the kernel (Gibbs, 1997; Heinonen et al., 2015). Instead, we construct a non-stationary GP using input-warping (Vinokur and Tolpin, 2021): If $c(t, t')$ is a valid covariance, the input-warped covariance $c(\phi(t), \phi(t'))$ is also positive definite for any function $\phi(t)$ (Bishop, 2006).

With an appropriate choice of warping function $\phi(t)$, we can increase the smoothness of the GP in peak regions. To gain some intuition, consider a scalar index $t \in \mathbb{R}$ and a linear warping function, $\phi(t) = a \cdot t$: As a decreases, the smoothness of the GP will increase as points t and t' will be closer together and thus more correlated. Thus, we design the function $\phi(t)$ to increase linearly at unit rate outside peak regions and at a higher rate r in peak regions.

To interpolate smoothly between the regions we employ the soft window function,

$$\Xi(\omega; \gamma) = \frac{1}{2}\left[\tanh\left(\frac{\omega + \frac{\gamma}{2})}{s}\right) - \tanh\left(\frac{\omega - \frac{\gamma}{2})}{s}\right)\right], \tag{11}$$

centered at zero, of width 2γ, and with s controlling its softness.

To govern the slope of the input warping we construct a function that has value 1 outside peak regions and has value a under each peak with smooth interpolation between the regions,

$$\xi(\omega) = 1 - (1 - a) \cdot \max_{k \in 1 \ldots k} \Xi(\omega - c_k; \gamma_k). \tag{12}$$

Here we use the maximum of the soft window functions across peaks to correctly handle the situation with multiple overlapping peaks. The width of each peak region is taken as the peak half-width.

Finally we construct the warping function as

$$\phi(\omega) = \int_0^\omega \xi(\omega')\mathrm{d}\omega', \tag{13}$$

which we compute in practice by numerical integration with the rectangle rule,

$$\phi(w) = \Delta\omega \sum_{i=1}^{w} \xi(\omega_i), \tag{14}$$

where $\Delta\omega = \omega_{k+1} - \omega_k$ is the resolution of the wavenumber axis.

An illustrative example of the procedure, including example draws from the non-stationary Gaussian process are given in Figure 3.

2.6. B-spline baseline

We contrast our method with the joint signal and baseline estimation technique proposed by Han and Ram (2020), where the baseline is modeled using a B-spline: A piecewise polynomial of fixed degree, joining at a predefined set of knots. For a fixed set of knots, the B-spline is uniquely characterized by its coefficients. For a general introduction to B-splines, we refer to Hastie et al. (2009).

Following Han and Ram (2020), we use cubic splines, $d = 3$, as a good trade-off between flexibility and smoothness, and use equally spaced knots on the wavenumber axis (with multiplicity $d+1$ at the endpoints). We precompute the associated basis matrix, such that the baseline may be expressed as,

$$B_w = \sum_{m=1}^{M_b} \Psi_{w,m} \lambda_m, \tag{15}$$

where Ψ is the B-spline basis matrix, and λ are the coefficients to be estimated. The number of spline basis functions is $M_b = M_k - d - 1$ where M_k is the number of knots and d is the degree of the spline.

2.7. Probabilistic model and inference

A major advantage of estimating the signal and baseline jointly is the ability to obtain uncertainty estimates on parameters and predictions, even when parameters are strongly correlated. To do this, we formulate the model as a joint distribution of the data and parameters and use Bayes' theorem to obtain the posterior distribution of parameters given the observed data.

Based on the noise assumption, the likelihood is Gaussian,

$$p(X_{n,w}|\theta) = \mathcal{N}\left(\sum_{k=1}^{K} \alpha_{n,k} V_{k,w} + \beta_n B_w + \kappa_n, \tau^{-1}\right), \tag{16}$$

where $\theta = \{\boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\beta}, B, \boldsymbol{\kappa}, \tau\}$ denotes all model parameters. We now introduce the following priors,

$$\begin{aligned}
c_k &\overset{\text{i.i.d}}{\sim} \mathcal{U}(c_\text{a}, c_\text{b}), & \alpha_{n,k} &\overset{\text{i.i.d}}{\sim} \mathcal{N}^+(\mu_\alpha, \tau_\alpha^{-1}),\\
\gamma_k &\overset{\text{i.i.d}}{\sim} \mathcal{N}^+(\mu_\gamma, \tau_\gamma^{-1}), & \eta_k &\overset{\text{i.i.d}}{\sim} \mathcal{U}(0,1),\\
\beta_n &\overset{\text{i.i.d}}{\sim} \mathcal{N}^+(\mu_\beta, \tau_\beta^{-1}), & B|\boldsymbol{c},\boldsymbol{\gamma},\boldsymbol{\eta} &\sim GP\left(0, c\left(\phi(w), \phi(w')\right)\right),\\
\kappa_n &\overset{\text{i.i.d}}{\sim} \mathcal{N}^+(\mu_\kappa, \tau_\kappa^{-1}), & \tau &\sim \mathcal{G}(a_\tau, b_\tau).
\end{aligned} \tag{17}$$

Here, $\mathcal{N}^+(\mu, \sigma^2)$ is a normal distribution truncated to the non-negative half-axis, with parameters μ and σ^2, $\mathcal{G}(a, b)$ is a Gamma distribution, and $\mathcal{U}(a, b)$ is a continuous uniform distribution on the interval $[a, b]$. All hyperparameters can be chosen to yield fairly uninformative priors, but if further knowledge e.g. about the analyte's peak positions is available, one can switch to more informative priors to aid the inference process. Guidelines for selecting appropriate priors are included in Appendix C.

This gives us our full joint probabilistic model,

$$p(X, \theta) = \left(\prod_{n,w} p(X_{n,w}|\theta)\right)\left(\prod_k p(c_k)p(\gamma_k)p(\eta_k)\prod_n p(\alpha_{n,k})\right) \times \left(\prod_n p(\beta_n)p(\kappa_n)\right) p(B|c, \gamma, \eta)p(\tau), \quad (18)$$

where we readily get the posterior distribution using Bayes' theorem.

For the experimental comparison to a B-spline baseline, we replace the GP prior by a Gaussian prior on the B-spline coefficients,

$$\lambda_m \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \tau_\lambda^{-1}), \quad (19)$$

and keep the rest of the model specification identical.

2.7.1. Inference

We propose a Metropolis-within-Gibbs Markov chain Monte Carlo (MCMC) sampling procedure. Because the peak amplitudes and all parameters of the baseline models are linear in the Gaussian likelihood and have (truncated) Gaussian priors, their conditional distributions are tractable. The specifics of how this allows us to construct standard Gibbs updates for these parameters is given in the supplementary material. However, because of the strong coupling between the baseline and signal model, using only separate Gibbs updates does not work. Conditioned on a given peak model, the Gibbs update for the baseline will be very constrained (it must fit the residual), and the MCMC sampler will thus tend to get locked in suboptimal solutions. To solve this, we use a joint signal and baseline update in the form of a Metropolis-Hastings step as outlined below, which leads to much faster mixing. To simplify the presentation, we omit some conditionals in the distributions, as they do not change when a new peak is proposed:

1. For each peak, $k = 1, \ldots, K$:
   (a) Propose new peak parameters $(c_k', \gamma_k', \eta_k')$ from a random walk proposal with some stepsize, which may be varied over the run or randomly drawn from a small set of values.
   (b) Conditioned on the proposed peak, propose a new baseline, B' from its Gibbs distribution.
   (c) For each spectrum n, propose an amplitude $\alpha_{n,k}'$ from its Gibbs distribution.

(d) Accept or reject this set of parameters jointly using a Metropolis-Hastings correction.

2. Sample α using its Gibbs distribution as in step 1c).
3. Sample B using its Gibbs distribution as in step 1b).
4. Sample β from its Gibbs distibution.
5. Sample κ from its Gibbs distribution.
6. Sample τ from its Gibbs distribution.

The process above is repeated until the desired number of samples has been drawn. Details of the derivation of the updates are given in Appendix A. We note that the computation time required to estimate the NSGP-baseline model is higher than the B-spline baseline model introduced by Han and Ram (2020) but is generally on the order of minutes depending on the size of the map (our inference approach scales cubically in the number of wavenumbers) and on the number of samples to be drawn.

While it would be possible to extend the inferences procedure to also learn additional hyperparameters, such as ℓ and a which control the GP lenghtscale and s which controls the smoothness of the transition regions, in this work we choose these manually.

There are multiple strategies to estimate the number of peaks K from the data: Han and Ram (2020) introduce reversible jump MCMC steps to the sampler, which allows the parameters to change dimensionality. This approach can seamlessly be integrated with our method. An alternative proposed by Li et al. (2020) is to assume that there is a peak located at every wavenumber, and enforce sparsity on the amplitudes of the peaks: While this circumvents estimating the peak locations, in our experience it makes it much more difficult to estimate the peak amplitudes. In the following, we assume that the number of peaks is known, and focus on the estimation of the peak parameters.

3. Results and discussion

We first evaluate the proposed methods on synthetic data and compare with the B-spline baseline model. Next, we examine the performance of the model on a difficult Methotrexate SERS map dataset. Finally we compare with established datadriven baseline estimation methods on a standard NIR data set of corn.

3.1. Synthetic data experiments

In all of the synthetic data experiments we use the following settings:

- Simulated maps consist of $N = 5$ spectra with shared parameters except for independent peak amplitudes, background scale and bias, as well as observation noise.
- The wavenumber axis is defined as $\omega_w = w$ for $w = 1, \ldots, 300$.

- We simulate $K = 3$ spectral peaks with
  - locations $c_k \sim \mathcal{U}(25, 275)$,
  - peak widths $\gamma_k \sim \mathcal{N}^+(7, 1)$,
  - lorentzianities $\eta_k \sim \mathcal{U}(0, 1)$, and
  - amplitudes $\alpha_{n,k} \sim \mathcal{U}(\frac{1}{2}, 1)$,

  with the constraint that the distance between peak locations is at least 25 to avoid overlap. Thus each simulated spectral curve contains peaks with different locations, amplitudes, and shapes. As detailed in the individual experiments, the amplitudes are further multiplied by a scale factor to control the signal strength.

- Baselines are generated according to one of the following procedures:
  1. AR(2): A zero mean unit variance second-order autoregressive process
  $$B_w = -0.9505 B_{w-2} + 1.95 B_{w-1} + 7.04 \cdot 10^{-3} \varepsilon_w \tag{20}$$
  with $\varepsilon_w \sim \mathcal{N}(0, 1)$ and steady state initial conditions.
  2. B-spline: A realization of a B-spline with random coefficients drawn from a zero mean Gaussian distribution
  $$\lambda_m \sim \mathcal{N}(0, 1). \tag{21}$$
  The number of basis functions is chosen as $M_b = 8$ as suggested by Han and Ram (2020) since this achieves smooth baseline that looks realistic. In all experiments on synthetic data, the B-spline baseline model has the correct number of basis functions.
  3. NSGP: A realization of our proposed non-stationary GP with length-scale $\lambda = 1 \cdot 10^3$ outside peak regions and relative lengths scale $a = 0.01$ inside peak regions, transition smoothness $s = 3$, and variance $\nu = 1$.

  These procedures generate baselines with fairly similar variance and autocorrelation (see Figure 4). Finally, the baselines are
  - scaled by $\beta_n \sim \mathcal{N}^+(1, 0.1)$, and
  - shifted by $\kappa_n \sim \mathcal{N}^+(3, 0.1)$.

- The noise variance is fixed at $\tau^{-1} = 0.01$.

Based on these choices, SERS maps are simulated according to Equation 8.

Table 1: Mean absolute error in estimating the peak location and amplitude with different simulated baselines.

Parameter	Model	Simulated baseline		
		AR(2)	B-spline	NSGP
c_k (location)	NSGP	**0.802**	0.816	**0.276**
	B-spline	0.840	**0.764**	0.991
α_k (amplitude)	NSGP	**0.270**	**0.239**	**0.195**
	B-spline	0.510	0.440	0.500

3.1.1. Different simulated baselines

In the first experiment we compare the NSGP and B-spline models on the three different simulated baselines. We expect the best performance when the model matches the simulated baseline, while the AR(2) serves as an additional test where the modeling assumptions are not matched.

We focus our attention on the estimation of the location and amplitude parameters, and present the results in terms of the mean absolute error on these parameters, averaged over 100 simulated maps. All maps were generated at an intermediat signal-to-baseline (SBR) ratio, with a unit amplitude scale factor, which means that most signal peaks are clearly visible in the spectra. An example spectrum is illustrated in Figure 5 (Ampl. scale 1).

Using the MCMC procedure, we draw 10 000 samples, discard the first 5 000 samples for burn-in, and use the mean value of the posterior marginals to estimate the parameters of interest.

The results of the experiment are given in Table 1. Regarding the location parameter, the NSGP outperforms the B-spline on the AR(2) background. The B-spline model performs slightly better than the NSGP on the B-spline baseline, whereas the NSGP model performs substantially better than the B-spline on the on the NSGP baseline. Regarding the amplitude parameter, the NSGP model performs better than the B-spline model in all cases, including when the baseline is generated from a B-spline model.

3.1.2. Amplitude estimation for small amplitude peaks

An especially important aspect of certain types of chemometrics for spectroscopy is the ability to accurately detect and quantify signals at weak signal strengths, where the signal-to-baseline ratio is low.

To test our model's ability to detect signals at low SBR, we created a synthetic data set of 1 000 maps with the peak amplitude scale factor varied logarithmically from 0.1 to 10, and employing an AR(2) baseline. Examples of simulated spectra at different SBR can be seen in Figure 5.

For the inference, we again draw 10 000 samples and discard 5 000 as burn-in. We evaluate our NSGP baseline correction method against the previously described B-spline baseline correction method and compare the results of the extracted amplitudes against the true underlying amplitudes. The results are

binned according the individual peak magnitudes. We show the results in Figure 6 where the true underlying amplitudes are shown against the estimated amplitudes for the two models and in Figure 7 where the relative errors are shown.

We see that when the underlying peak amplitudes are sufficiently small, our model outperforms the B-spline baseline model, with lower errors. This is due to the our model better handling the trade-off between modeling the signal with baseline and with the spectral model.

3.2. Detecting and quantifying methotrexate in human serum

Methotrexate (MTX) is a widely used anti-cancer drug often used in patients with eg. leukemia, breast cancer and more, and has high toxicity in high-dosis cases (HD-MTX), especially if the usage is long Howard et al. (2016); Paci et al. (2014). Thus, to manage the dosis and toxicity of MTX in HD-MTX patients, an expensive rescue-drug, leucovorin (LV) is administered along with MTX. This inconvenience can be alleviated by employing so-called therapeutic drug monitoring of MTX, and recently it has been shown by Göksel et al. (2021) that MTX can be detected in human serum with SERS using minimal sample preparation.

In the work of Göksel et.al. a simple linear baseline correction is applied since the characteristic Raman peak of MTX (at 687 cm^{-1}) is of known position, and thus such a baseline is easily applicable, that is, they fit a linear model to points around the peak locations, excluding the wavenumbers under the peak. We call this approach a manual local linear fit. This approach to baseline correction has the advantage that it is simple, computationally cheap compared to our flexible approach and is easily interpretable. However, it requires extensive prior information regarding the analyte and quite substantial manual labour, especially if the assumption of a common baseline across the SERS map is not relaxed. The raw data is shown in Appendix B.

We analyze the same dataset of MTX in filtered (artificial) human serum that is used by Göksel et al. (2021). It consists of 7 different concentration levels (0, 2.5, 5, 10, 15, 20 and 30 μM MTX respectively) and there are 3 maps for each concentration level, corresponding to 21 SERS maps in total. Each of the maps are of (roughly) size 45×45 and consists of 1686 wavenumbers from 50 cm^{-1} to 3300 cm^{-1} at a resolution of approximately 2 cm^{-1}. For the experiments below, we consider only wavenumbers from 648 cm^{-1} to 735 cm^{-1} for simplicity, and we height-normalize each map to have a maximum intensity of 2. We compare our non-stationary GP baseline model to the B-spline approach, a manual local linear fit with the shared baseline assumption relaxed (ie. each spectrum has an individual baseline) and simply reading off the amplitude of the raw spectra with no baseline correction. Since no signal is present along the edges of the maps, we remove a few of the spectra corresponding to the edges of the substrate.

Using the MCMC procedure described previously, we draw 60.000 samples, and discard 30.000 as burn-in. The priors used for this problem are (for the

NSGP baseline model),

$$
\begin{aligned}
c_k &\overset{\text{i.i.d}}{\sim} \mathcal{U}(683, 691), & \alpha_{n,k} &\overset{\text{i.i.d}}{\sim} \mathcal{N}^+(0.1, 0.1),\\
\gamma_k &\overset{\text{i.i.d}}{\sim} \mathcal{N}^+(2, 0.05), & \eta_k &\overset{\text{i.i.d}}{\sim} \mathcal{U}(0, 1),\\
\beta_n &\overset{\text{i.i.d}}{\sim} \mathcal{N}^+(0.1, 0.001), & \kappa_n &\overset{\text{i.i.d}}{\sim} \mathcal{N}^+(3, 1),\\
\tau &\sim \mathcal{G}(1, 1).
\end{aligned}
$$

The NSGP background model is tuned to appropriate hyperparameters. The B-spline is set with 15 basis functions, yielding 19 knots in total, and the prior for β_n is changed to a $\mathcal{N}^+(1, 0.01)$ due to convergence issues. This was selected by tuning the parameters such that the baseline achieved a good trade-off between flexibility of the baseline and accurate signal extraction. All other priors are identical for the two baseline models.

We compare the results by using calibration curves, which are computed by fitting a linear regression model to map the estimated peak amplitudes to estimated concentrations. We select the top 20% spectra as measured on estimated peak amplitudes to compute these calibration curves.

The results are seen in Figure 8 where the recalibrated curves for the NSGP baseline model, the B-spline baseline model, a manual local linear correction and a case where the signal intensity is read off with no correction or peak fitting is present. Both the B-spline baseline model and the direct peak amplitude readings essentially fail to correlate with the concentration, whereas the manual local linear fit and our proposed method display a clear relation with the concentration. Some examples of the model fit for the NSGP model and B-spline model can be seen in Figure 9.

From here, it is clear, that appropriate baseline correction is indeed needed for accurate calibration. Our model outperforms the B-spline baseline model due to the complicated nature of the problem (see eg. Figure 9), and performs similar to the manual linear baseline correction, particularly in the challenging low concentration domain. While our method does not outperform the manual local linear baseline estimation in this case, an advantage of our approach is that it reduces the requirement for manual labor and makes it possible to automate the task to a higher degree.

3.3. Baseline correction of corn NIR spectra

To compare our method to established datadriven baseline estimation techniques, we examine its performance on a NIR dataset of corn samples (`http://eigenvector.com`). The corn dataset consists of 80 NIR spectra of measured at 700 wavenumbers from 1100 nm to 2498 nm. For each sample, four target measurements of moisture, oil, protein, and starch are given.

Following Li et al. (2020), the comparison is carried out by fitting a partial least squares (PLS) regression model on the baseline subtracted data to predict the target variables. We compare with two established methods, namely adaptive iteratively reweighted penalized least squares (airPLS, Zhang et al., 2010),

and sparse Bayesian learning for baseline correction (SBL-BC, Li et al., 2020). In addition, we include results on PLS regression on the raw data, i.e. no baseline subtraction (NO).

For the proposed method, the priors were set as fairly uninformative,

$$
\begin{aligned}
c_k &\overset{\text{i.i.d}}{\sim} \mathcal{U}(1100, 2498), & \alpha_{n,k} &\overset{\text{i.i.d}}{\sim} \mathcal{N}^+(0.1, 0.25),\\
\gamma_k &\overset{\text{i.i.d}}{\sim} \mathcal{N}^+(50, 400), & \eta_k &\overset{\text{i.i.d}}{\sim} \mathcal{U}(0, 1),\\
\beta_n &\overset{\text{i.i.d}}{\sim} \mathcal{N}^+(1, 1), & \kappa_n &\overset{\text{i.i.d}}{\sim} \mathcal{N}^+(1, 1),\\
\tau &\sim \mathcal{G}(1, 1).
\end{aligned}
$$

with the prior peak amplitude and width priors set to approximately match the peaks in the data as assessed by visual inspection.

To provide a suitable model order and a rough initial fit, we used $K = 13$ peaks initialized at wavenumbers 1200, 1364, 1464, 1570, 1688, 1754, 1776, 1932, 2106, 2286, 2314, 2342, and 2492, chosen by manual inspection of the spectra. Using the described MCMC procedure, we generated 500 samples and used only the last sample as a point prediction. We disregarded the estimated peak parameters and used only the estimated baseline. For airPLS we used a fairly strong smoothness of $\lambda = 5 \cdot 10^7$ which we found to give good results across all targets.

The dataset and fitted baselines are illustrated in Figure 11. Both SBL-BC and airPLS smoothly track the baseline. With the high smoothness we found optimal for airPLS the estimated baselines are nearly linear. The baselines estimated by NSGP are significantly less smooth but flat in the peak regions, and do not track the spectra as closely because of the additive signal model used in the fitting procedure.

We then estimated the four targets using PLS. We randomly selected 65 of the 80 observations for training and used 5-fold crossvalidation within the 65 training observations to select the number of PLS components. We then fitted the selected PLS model on the 65 training observations and tested on the 15 held-out observations. The procedure was averaged over 1000 random splits.

Results are shown in Table 2. Finally, to examine how sensitive the methods are to the size of the training data, we repeated the entire procedure for training set sizes of $60, 61, \ldots, 70$. These results are shown in Figure 10. In line with previously published results, no baseline correction performed best for predicting moisture. For predicting oil, all methods were near identical, perhaps with a slight advantage to airPLS. Predicting protein, SBL-BC was substantially better than the other methods. Finally, for predicting starch, NSGP performed best with both airPLS and SBL-BC performing substantially worse than NSGP and NO.

4. Conclusion

We have developed and described a novel approach to joint estimation of signal and baseline for vibrational spectroscopy in the case of an additive base-

	Moisture	Oil	Protein	Starch
NO	**0.007**	0.025	0.080	0.117
SBL-BC	0.049	0.025	**0.064**	0.183
airPLS	0.037	**0.024**	0.086	0.169
NSGP	0.023	0.025	0.083	**0.111**

Table 2: Root mean squared error of prediction on the Corn NIR dataset. Comparison of no baseline correction (NO), adaptive iteratively reweighted penalized least squares (airPLS), sparse Bayesian learning for baseline correction (SBL-BC) and non-stationary Gaussian process (NSGP) baseline estimation methods.

line using non-stationary Gaussian processes for the baseline. We have shown that our model outperforms a similar approach using B-splines for the baseline, exemplified with surface enhanced Raman spectroscopy, both in the case of synthetic and real-life SERS data.
Our approach, while requiring some tuning of hyperparameters and inference process, provides a very flexible approach where very little user-input is required, and achieves similar performance to a manual baseline correction, which is both labor intensive and requires prior knowledge of the analyte to be effective. Furthermore, our approach has the benefit of giving natural uncertainty quantification, as it is formulated in a fully Bayesian framework.

Our approach is generally applicable to any model of the form,

$$X_w = f(X_w; \theta) + B_w + \varepsilon,$$

where f is a known function of parametes θ and the baseline B is has local differences in flexibility, ie. where the signal model and baseline model is competing to explain observations.

Acknowledgements

Funding: This work was supported by the Independent Research Fund Denmark [grant number 9131-00039B].

Appendix A. Derivation of sampling distributions

In this section we derive the relevant conditional distributions that are used in the Metropolis-within-Gibbs sampling scheme. We will provide some detail, but since some of these results are standard in the context of Bayesian statistics, they will only be summarized.

For convenience, we restate the factorized model, (18), here

$$p(\boldsymbol{X}, \boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\beta}, B, \boldsymbol{\kappa}, \tau) = \prod_{n,w} p(X_{n,w}|\alpha_{n,:}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \beta_n, B_w, \kappa_n, \tau) \times \prod_k p(c_k) \prod_k p(\gamma_k) \prod_k p(\eta_k) \prod_{n,k} p(\alpha_{n,k}) \times \prod_n p(\beta_n) \prod_n p(\kappa_n) \times p(B|\boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\eta}) \times p(\tau).$$

and the priors (17),

$$\begin{aligned} c_k &\overset{\text{i.i.d}}{\sim} \mathcal{N}^+(\mu_c, \tau_c^{-1}) & \alpha_{nk} &\overset{\text{i.i.d}}{\sim} \mathcal{N}^+(\mu_\alpha, \tau_\alpha^{-1}) \\ \gamma_k &\overset{\text{i.i.d}}{\sim} \mathcal{N}^+(\mu_\gamma, \tau_\gamma^{-1}) & \eta_k &\overset{\text{i.i.d}}{\sim} \mathcal{U}(0,1) \\ \beta_n &\overset{\text{i.i.d}}{\sim} \mathcal{N}^+(\mu_\beta, \tau_\beta^{-1}) & B|c,\gamma,\eta &\sim GP(0, C(w,w')) \\ \kappa_n &\overset{\text{i.i.d}}{\sim} \mathcal{N}^+(\mu_\kappa, \tau_\kappa^{-1}) & \tau &\sim \mathcal{G}(a_\tau, b_\tau) \end{aligned}$$

Finally, recall that the likelihood is Gaussian,

$$X_{n,w} \sim \mathcal{N}\left(\sum_{k=1}^{K} \alpha_{n,k} V_w(c_k, \eta_k, \gamma_k) + \beta_n B_w + \kappa_n, \tau^{-1}\right).$$

Appendix A.1. Gibbs update for τ

Since the prior for the likelihood precision is a gamma distribution, $\tau \sim \mathcal{G}(a_\tau, b_\tau)$ and the likelihood is Gaussian, we obtain a standard result in Bayesian statistics, namely that the relevant Gibbs distribution is,

$$\tau|\boldsymbol{X}, \boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{B}, \boldsymbol{\kappa} \sim \mathcal{G}\left(a_\tau + \tfrac{NW}{2}, b_\tau + \tfrac{1}{2}\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_F^2\right), \tag{A.1}$$

where N is the number of spectra in the map, and W is the number of measured wavenumbers, $\hat{\boldsymbol{X}} = \boldsymbol{\alpha}\boldsymbol{V}(\boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\eta}) + \boldsymbol{\beta}\boldsymbol{B}^\top + \boldsymbol{\kappa}\mathbf{1}_W^\top$ is our model reconstruction, and $\|\cdot\|_F$ is the Frobenius norm of a matrix.

Appendix A.2. Gibbs update for B

We seek to compute,

$$p(B|\boldsymbol{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\kappa}, \tau) \propto p(B, \boldsymbol{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\kappa}, \tau).$$

Considering the joint distribution as a function of B, we see that up to proportionality, we have,

$$p(B|\boldsymbol{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\kappa}, \tau) \propto p(\boldsymbol{X}|B, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\kappa}, \tau) p(B|0, C)$$

where C is the non-stationary kernel for the Gaussian process described in subsection 2.5. In the following, we assume that the length-scale parameter ℓ^2 as well as s in the transformation are deterministic hyperparameters.
Inserting the Gaussian likelihood and the GP-prior we obtain,

$$p(B|\boldsymbol{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\kappa}, \tau) \propto \prod_n \prod_w \mathcal{N}(X_{n,w}|\hat{X}_{n,w}, \tau^{-1}) \mathcal{N}(B|0, C),$$

yielding (to proportionality),

$$p(B|\boldsymbol{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\kappa}, \tau) \propto \exp\left[-\frac{1}{2}\tau \|\boldsymbol{X} - \boldsymbol{\alpha}\boldsymbol{V} - \boldsymbol{\beta}B^\top - \boldsymbol{\kappa}\mathbf{1}_W^\top\|_F^2 - \frac{1}{2}B^\top C^{-1} B\right],$$

where $\|\cdot\|_F$ is the Frobenius norm.
Defining $\boldsymbol{\phi} = \boldsymbol{\alpha}\boldsymbol{V} + \boldsymbol{\kappa}\mathbf{1}^\top$, this is rewritten,

$$\begin{aligned}\log p(B|\boldsymbol{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\kappa}, \tau) \propto \\ \frac{1}{2}\Big[\tau\|\boldsymbol{X}\|_F^2 + \tau\|\boldsymbol{\phi}\|_F^2 + \tau B^\top \boldsymbol{\beta}^\top \boldsymbol{\beta} I_W B - \\ 2\tau\langle\boldsymbol{\phi}, \boldsymbol{X}\rangle_F - 2\tau B^\top \boldsymbol{X}\boldsymbol{\beta} + 2\tau B^\top \boldsymbol{\phi}\boldsymbol{\beta} + B^\top C^{-1} B\Big],\end{aligned}$$

where the proportionality is up to an additive constant not dependent on B and $\langle\cdot,\cdot\rangle_F$ denotes the Frobenius inner product.
Since we are interested in a distribution over B, we leave out all terms not dependent on B, getting to proportionality,

$$\begin{aligned}\log p(B|\boldsymbol{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\kappa}, \tau) \propto \\ -\frac{1}{2}[\tau B^\top \boldsymbol{\beta}^\top \boldsymbol{\beta} I_W B - 2\tau B^\top \boldsymbol{X}\boldsymbol{\beta} + 2\tau B^\top \boldsymbol{\phi}\boldsymbol{\beta} + B^\top C^{-1} B],\end{aligned}$$

where I_W is the identity matrix of size $W \times W$.
Reordering, we get,

$$\begin{aligned}\log p(B|\boldsymbol{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\kappa}, \tau) \propto \\ -\frac{1}{2}B^\top(\tau\boldsymbol{\beta}^\top\boldsymbol{\beta} I_W + C^{-1})B + \tau B^\top(\boldsymbol{X} - \boldsymbol{\phi})\boldsymbol{\beta}.\end{aligned}$$

Define now $A^{-1} = \tau\boldsymbol{\beta}^\top\boldsymbol{\beta} I_W + C^{-1}$ and $\mu = \tau A(\boldsymbol{X} - \boldsymbol{\phi})\boldsymbol{\beta}$. If we insert, we get,

$$\log p(B|\boldsymbol{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\kappa}, \tau) \propto -\frac{1}{2}B^\top A^{-1} B + B^\top A^{-1}\mu,$$

which is the log-density of a multivariate Gaussian in B with covariance A and mean μ, meaning that,

$$B|\boldsymbol{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\kappa}, \tau \sim \mathcal{N}\left(\tau\left[\tau\boldsymbol{\beta}^\top\boldsymbol{\beta} I_W + K^{-1}\right]^{-1}(\boldsymbol{X} - \boldsymbol{\phi})\boldsymbol{\beta}\,, \left[\tau\boldsymbol{\beta}^\top\boldsymbol{\beta} I_W + K^{-1}\right]^{-1}\right).$$

Appendix A.3. Gibbs update for $\boldsymbol{\kappa}$

Recall that $\boldsymbol{\kappa}$ is the translational bias for the model and that a truncated normal prior is imposed.

The derivation of the Gibbs sampling distribution is very similar in all but algebra to the derivation for B, and so we shall skip over some details in the derivation.

The derivation starts by noting that the truncated normal distribution can be written simply as,

$$p(\kappa_n|\mu_\kappa, \tau_\kappa^{-1}) \propto \mathcal{N}(\kappa_n|\mu_\kappa, \tau_\kappa^{-1})\mathbb{I}(\kappa_n \geq 0),$$

where $\mathbb{I}(A)$ is the indicator function for boolean variable A, ie.

$$\mathbb{I}(\kappa_n \geq 0) = \begin{cases} 1, & \kappa_n \geq 0 \\ 0, & \text{otherwise} \end{cases}.$$

This is extensible to the log-domain, where we simply define,

$$\log \mathbb{I}(\kappa_n \geq 0) = \begin{cases} 0, & \kappa_n \geq 0 \\ -\infty, & \text{otherwise} \end{cases}.$$

It is then straightforward to show that (by simple algebra),

$$p(\kappa_n|\boldsymbol{X}, \boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\beta}, B, \tau) \propto p(\boldsymbol{X}|\boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\beta}, B, \boldsymbol{\kappa}, \tau)p(\kappa_n|\mu_\kappa, \tau_\kappa^{-1}),$$

is in fact a truncated normal distribution with parameters

$$\begin{aligned} \mu &= \tau \frac{(\boldsymbol{X} - \boldsymbol{\alpha}\boldsymbol{V} - \boldsymbol{\beta}B^\top)\mathbf{1}_W + \mu_\kappa \tau_\kappa}{W\tau + \tau_\kappa} \\ \sigma^2 &= (W\tau + \tau_\kappa)^{-1} \end{aligned}$$

such that,

$$\kappa_n|\boldsymbol{X}, \boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\beta}, B, \tau \overset{\text{i.i.d}}{\sim} \mathcal{N}^+\left(\tau \frac{(\boldsymbol{X} - \boldsymbol{\alpha}\boldsymbol{V} - \boldsymbol{\beta}B^\top)\mathbf{1}_W + \mu_\kappa \tau_\kappa}{W\tau + \tau_\kappa}, (W\tau + \tau_\kappa)^{-1}\right)$$

Appendix A.4. Gibbs update for $\boldsymbol{\alpha}$

Recall that we impose an i.i.d truncated Gaussian prior on the spectrum specific peak amplitudes, $\alpha_{k,n}$, ie.

$$p(\alpha_{k,n}) \sim \mathcal{N}^+(\mu_\alpha, \tau_\alpha^{-1}).$$

With the Gaussian likelihood we have employed, the derivation of the update is analogous to that of κ_n, and we get

$$\alpha_{n,k}|\boldsymbol{X}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\beta}, B, \boldsymbol{\kappa}, \tau \overset{\text{i.i.d}}{\sim} \mathcal{N}^+(\mu, \sigma^2),$$

where

$$\begin{aligned} \sigma^2 &= \tau_\alpha + \tau V(c_k, \gamma_k, \eta_k)^\top V(c_k, \gamma_k, \eta_k) \\ \mu &= \sigma^2 \left(\mu_\alpha \tau_\alpha + \tau \sum_w V_{w,k} \left[X_{n,w} - \beta_n B_w - \kappa_n - \sum_{k' \neq k} \alpha_{n,k'} V_{k',w} \right] \right) \end{aligned}$$

Appendix A.5. Gibbs update for $\boldsymbol{\beta}$

Since we have the prior $\beta_n \overset{\text{i.i.d}}{\sim} \mathcal{N}^+(\mu_\beta, \tau_\beta^{-1})$, and the likelihood is Gaussian, we are again in an analogous situation to that of the Gibbs update of κ_n. This gives us,

$$p(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\eta}, B, \boldsymbol{\kappa}, \tau) = \mathcal{N}^+(\boldsymbol{\mu}, \Sigma),$$

where,

$$\begin{aligned}\Sigma &= \frac{1}{\tau_\beta + \tau B^\top B} I_N \\ \boldsymbol{\mu} &= \Sigma \left(\tau \left[\boldsymbol{X} - \boldsymbol{\alpha}\boldsymbol{V}(\boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\eta}) - \boldsymbol{\kappa}\mathbf{1}_W^\top\right] B + \mu_\beta \tau_\beta \mathbf{1}_N\right)\end{aligned}$$

where I_N is the identity matrix of size $N \times N$ and $\mathbf{1}_M$ is a $M \times 1$ vector of ones.

Appendix A.6. Gibbs update for spline coefficients

For the B-spline baseline model, we precompute the B-spline Vandermonde matrix evaluated at all wavenumbers $\boldsymbol{w}$. Denote this matrix $\boldsymbol{\Psi}$. Denote the coefficients $\boldsymbol{\lambda}$. Then, the baseline is expressed as,

$$B = \boldsymbol{\Psi}\boldsymbol{\lambda}.$$

We impose a zero-mean Gaussian prior on the spline coefficients,

$$\boldsymbol{\lambda} \sim \mathcal{N}(0, \tau_\lambda^{-1} I). \tag{A.2}$$

Since $\boldsymbol{\Psi}$ is deterministic, the derivation is very similar to the one for $\boldsymbol{\kappa}$, as we simply compute,

$$p(\boldsymbol{\lambda}|\boldsymbol{X}, \boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\kappa}, \tau,) \propto p(\boldsymbol{X}|\boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\kappa}, \tau, \boldsymbol{\lambda}) p(\boldsymbol{\lambda}|\tau_\lambda), \tag{A.3}$$

which by simple algebra can be shown to be a multivariate Gaussian with parameters,

$$\begin{aligned}\Sigma &= \left(N\tau \boldsymbol{\Psi}^\top \boldsymbol{\Psi} + \tau_\lambda I\right)^{-1} \\ \mu &= \Sigma \left[\tau \left(\boldsymbol{X} - \boldsymbol{\alpha}\boldsymbol{V} - \boldsymbol{\kappa}^\top \mathbf{1}_W\right) \boldsymbol{\Psi}\boldsymbol{\beta}\right]\end{aligned}$$

Appendix B. Raw MTX spectra

Figure B.12 shows the raw MTX data. Each color represents experimental repetitions of the experiment at the corresponding concentration. The bold colored line denotes the mean spectrum, obtained by averaging over each spectrum. The vertical red, dotted line is the location of the peak of interest at 687 cm^{-1}. The spectra have been preprocessed only such that they are on the same scale for visualization purposes (ie. $\kappa_n \approx 0$ for all n). The figure shows why the problem is challenging. This is namely because the amplitude of the peak at 687 cm^{-1} is not increasing much, even at drastically higher concentrations. This is in correspondence with our findings, see eg. Figure 8.

Appendix C. Guidelines for setting hyperparameters

The proposed method for joint peak and baseline estimation has a number of hyperparameters which must be set. In the following we give general guidelines for reasonable prior and parameters settings which are not strongly informative.

$c_\mathbf{a}$, $c_\mathbf{b}$ The upper and lower limit for peak positions is typically set to cover the entire spectral range, such that c_a is set to the lowest wavenumber and c_b is set to the highest wavenumber.

μ_α, τ_α^{-1} The mean and variance of peak amplitudes should be set to cover the expected range of peak amplitudes. For example, if peak amplitudes are typically in the range 0–1 units, reasonable settings would be $\mu_\alpha = 0.5$, $\tau_\alpha^{-1} = 0.5^2$.

μ_γ, τ_γ^{-1} The mean and variance of the peak widths should be set to cover the expected range of peak widths. For example, if a typical peak widts are in the range 20–40 units, reasonable settings would be $\mu_\gamma = 30$ and $\tau_\gamma^{-1} = 10^2$

μ_β, τ_β^{-1} The mean and variance for the multiplicative baseline scale should match the typical basline level. For example, if the baseline is typically in the range 2–4 units across the different spectra, reasonable settings would be $\mu_\beta = 3$ and $\tau_\beta = 1^2$.

μ_κ, τ_κ^{-1} The additive bias for the baseline should typically have zero mean prior with a high variance.

α_τ, β_τ The prior for the noise variance should typically set to $\alpha_\tau = \beta_\tau = 1$ or smaller to be relatively non-informative.

a The relative lengthscale (flatness) in peak regions should typically be around 0.01.

s The smoothness of the interpolation between peak and non-peak regions should typically be set between 10–20 % of the expected typical peak width.

References

Albrecht, M.G., Creighton, J.A., 1977. Anomalously intense raman spectra of pyridine at a silver electrode. Journal of the american chemical society 99, 5215–5217.

Alstrøm, T.S., Schmidt, M.N., Rindzevicius, T., Boisen, A., Larsen, J., 2017. A pseudo-Voigt component model for high-resolution recovery of constituent spectra in Raman spectroscopy. Proceedings of the 42nd Ieee International Conference on Acoustics, Speech and Signal Processing , 2317–2321doi:10.1109/ICASSP.2017.7952570.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. volume 4. Springer. URL: http://www.library.wisc.edu/selectedtocs/bg0137.pdf, doi:10.1117/1.2819119.

Chen, H., Shi, X., He, Y., Zhang, W., 2023. Automatic background correction method for laser-induced breakdown spectroscopy. Spectrochimica Acta Part B: Atomic Spectroscopy , 106763.

Demtröder, W., 2014. Laser Spectroscopy 1. Springer Berlin Heidelberg, Berlin, Heidelberg. URL: http://link.springer.com/10.1007/978-3-642-53859-9, doi:10.1007/978-3-642-53859-9.

Eilers, P.H., 2004. Parametric time warping. Analytical chemistry 76, 404–411.

Fleischmann, M., Hendra, P.J., McQuillan, A.J., 1974. Raman spectra of pyridine adsorbed at a silver electrode. Chemical physics letters 26, 163–166.

Gibbs, M.N., 1997. Bayesian Gaussian Processes for Regression and Classification. Ph.D. thesis.

Göksel, Y., Zor, K., Rindzevicius, T., Thorhauge Als-Nielsen, B.E., Schmiegelow, K., Boisen, A., 2021. Quantification of Methotrexate in Human Serum Using Surface-Enhanced Raman ScatteringToward Therapeutic Drug Monitoring. ACS Sensors 6, 2664–2673. doi:10.1021/acssensors.1c00643.

Gornushkin, I., Eagan, P., Novikov, A., Smith, B., Winefordner, J., 2003. Automatic correction of continuum background in laser-induced breakdown and raman spectrometry. Applied spectroscopy 57, 197–207.

Han, N., Ram, R.J., 2020. Bayesian modeling and computation for analyte quantification in complex mixtures using Raman spectroscopy. Computational Statistics and Data Analysis 143, 106846. URL: https://doi.org/10.1016/j.csda.2019.106846, doi:10.1016/j.csda.2019.106846.

Hastie, T., Friedman, J.H., Tibshirani, R., 2009. The Elements of Statistical Learning : Data Mining, Inference, and Prediction. Springer New York.

Heinonen, M., Mannerström, H., Rousu, J., Kaski, S., Lähdesmäki, H., 2015. Non-Stationary Gaussian Process Regression with Hamiltonian Monte Carlo URL: http://arxiv.org/abs/1508.04319.

Howard, S.C., McCormick, J., Pui, C.h., Buddington, R.K., Harvey, R.D., 2016. Preventing and Managing Toxicities of High-Dose Methotrexate. The Oncologist 21, 1471–1482. URL: http://dx.doi.org/10.1634/theoncologist.2015-0164, doi:10.1634/theoncologist.2015-0164.

Jeanmaire, D.L., Van Duyne, R.P., 1977. Surface raman spectroelectrochemistry: Part i. heterocyclic, aromatic, and aliphatic amines adsorbed on the anodized silver electrode. Journal of electroanalytical chemistry and interfacial electrochemistry 84, 1–20.

Li, H., Dai, J., Pan, T., Chang, C., So, H.C., 2020. Sparse Bayesian learning approach for baseline correction. Chemometrics and Intelligent Laboratory Systems 204, 104088. doi:10.1016/j.chemolab.2020.104088.

Li, J., Dai, L., 2012. A hard modeling approach to determine methanol concentration in methanol gasoline by Raman spectroscopy. Sensors and Actuators, B: Chemical 173, 385–390. URL: http://dx.doi.org/10.1016/j.snb.2012.07.012, doi:10.1016/j.snb.2012.07.012.

Paci, A., Veal, G., Bardin, C., Levêque, D., Widmer, N., Beijnen, J., Astier, A., Chatelut, E., 2014. Review of therapeutic drug monitoring of anticancer drugs part 1 - Cytotoxics. doi:10.1016/j.ejca.2014.04.014.

Pan, L., Zhang, P., Daengngam, C., Peng, S., Chongcheawchamnan, M., 2022. A review of artificial intelligence methods combined with raman spectroscopy to identify the composition of substances. Journal of Raman Spectroscopy 53, 6–19.

Picaud, V., Giovannelli, J.F., Truntzer, C., Charrier, J.P., Giremus, A., Grangeat, P., Mercier, C., 2018. Linear maldi-tof simultaneous spectrum deconvolution and baseline removal. BMC bioinformatics 19, 1–20.

Rasmussen, C.E., Williams, C.K.I., 2005. Gaussian Processes for Machine Learning. The MIT Press. URL: https://direct.mit.edu/books/book/2320/gaussian-processes-for-machine-learning, doi:10.7551/mitpress/3206.001.0001.

Schmidt, M.N., Alstrøm, T.S., Svendstorp, M., Larsen, J., 2019. Peak detection and baseline correction using a convolutional neural network, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 2757–2761.

Siesler, H., 2016. Vibrational Spectroscopy. Reference Module in Materials Science and Materials Engineering URL: https://linkinghub.elsevier.com/retrieve/pii/B9780128035818013187, doi:10.1016/B978-0-12-803581-8.01318-7.

Tauler, R., Kowalski, B., Fleming, S., 1993. Multivariate curve resolution applied to spectral data from multiple runs of an industrial process. Analytical Chemistry 65, 2040–2047. URL: https://pubs.acs.org/doi/abs/10.1021/ac00063a019, doi:10.1021/ac00063a019.

Vinokur, I., Tolpin, D., 2021. Warped Input Gaussian Processes for Time Series Forecasting. Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12716, 205–220. doi:10.1007/978-3-030-78086-9{_}16.

Yang, C., He, Z., Yu, W., 2009. Comparison of public peak detection algorithms for maldi mass spectrometry data analysis. BMC bioinformatics 10, 1–13.

Zhang, Z.M., Chen, S., Liang, Y.Z., 2010. Baseline correction using adaptive iteratively reweighted penalized least squares. Analyst 135. doi:10.1039/b922045c.

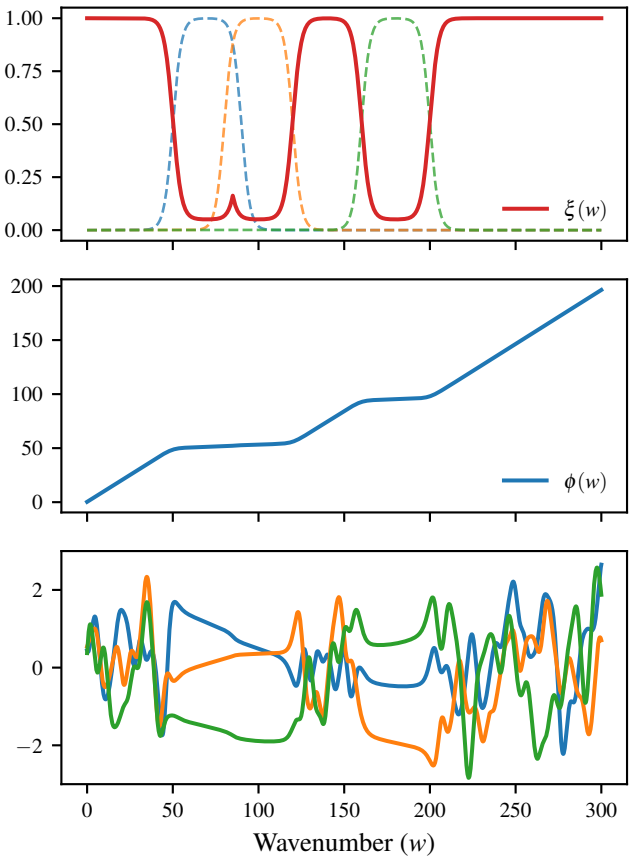

Figure 3: The transformation used for input-warping in the GP. Top: Soft window functions (dotted lines) are placed on each spectral peak (located at $w = 70, 100, 180$) and combined to form the warping function slope, $\xi(w)$. Middle: Integration of the slope yields the warping function, $\phi(w)$. Bottom: Three samples from a GP with an input-warped squared exponential kernel show that random functions are more smooth in the peak regions.

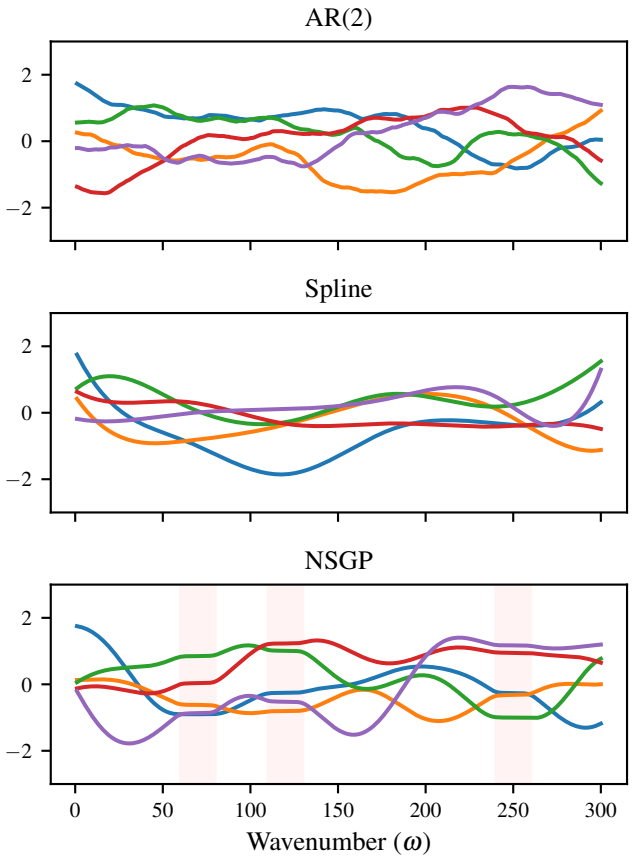

Figure 4: Examples of simulated baselines. The NSGP baseline is generated conditioned on peak positions $c_1 = 70, c_2 = 120, c_3 = 250$.

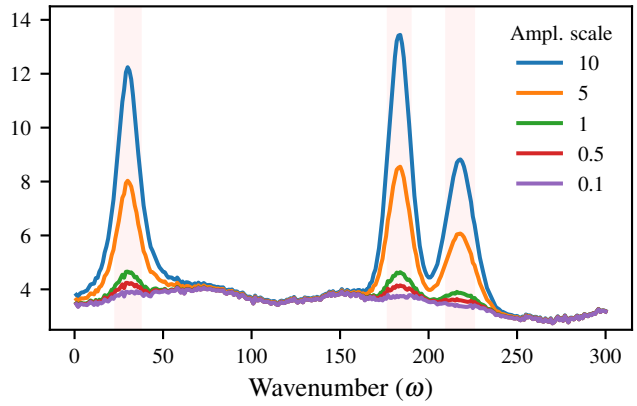

Figure 5: Examples of generated data with amplitude scales in the range 0.1–10, here visualized with the same baseline.

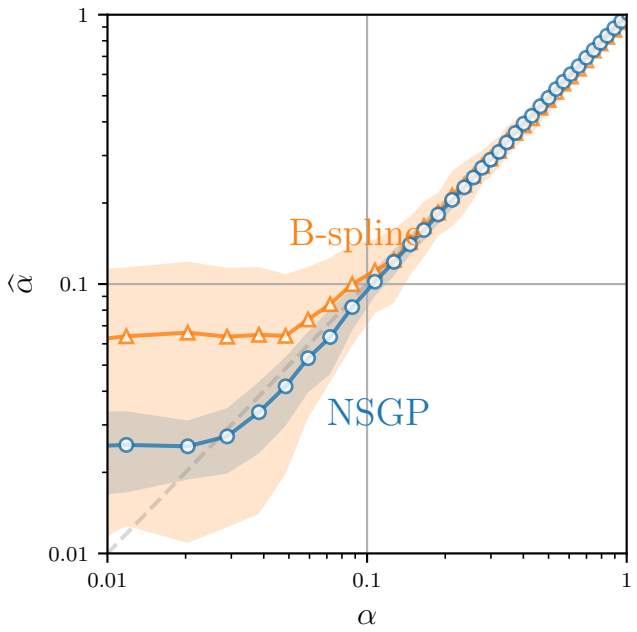

Figure 6: Raw estimation of amplitudes in synthetic spectra, binned on true amplitudes. Perfect estimation would mean that the points would lie on the dotted gray line. Colored in regions denote standard deviations within the bins. Note that both axes are logarithmic.

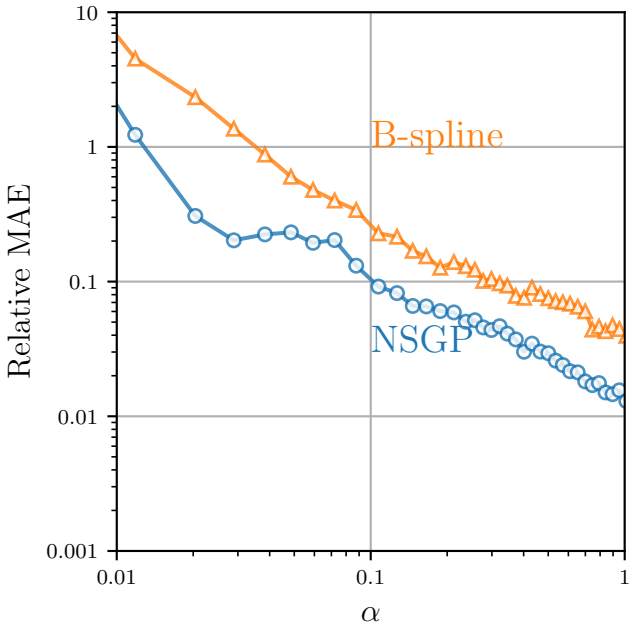

Figure 7: Relative mean absolute error of estimated amplitudes as a function of true signal amplitude. Note that both axes are logarithmic. Note also, that in this plot, we look at each peak independently, thus ignoring the SERS map connection between some peaks.

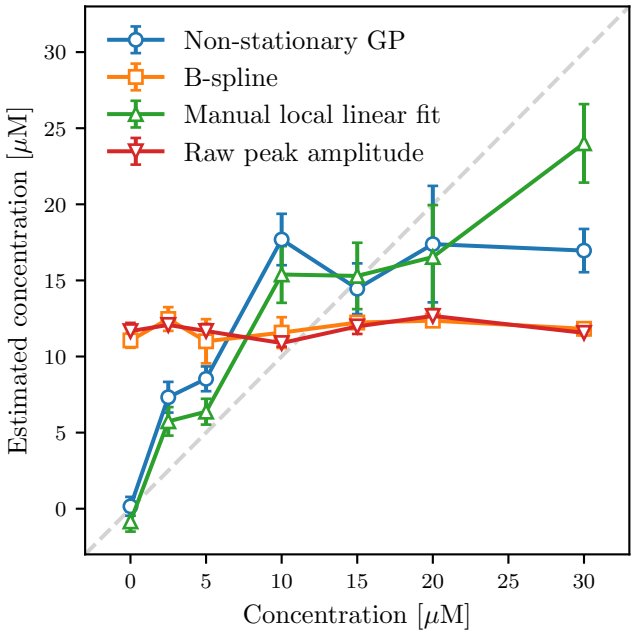

Figure 8: Calibration curves for MTX concentration estimation, computed by simple linear regression on the estimated peak amplitude.

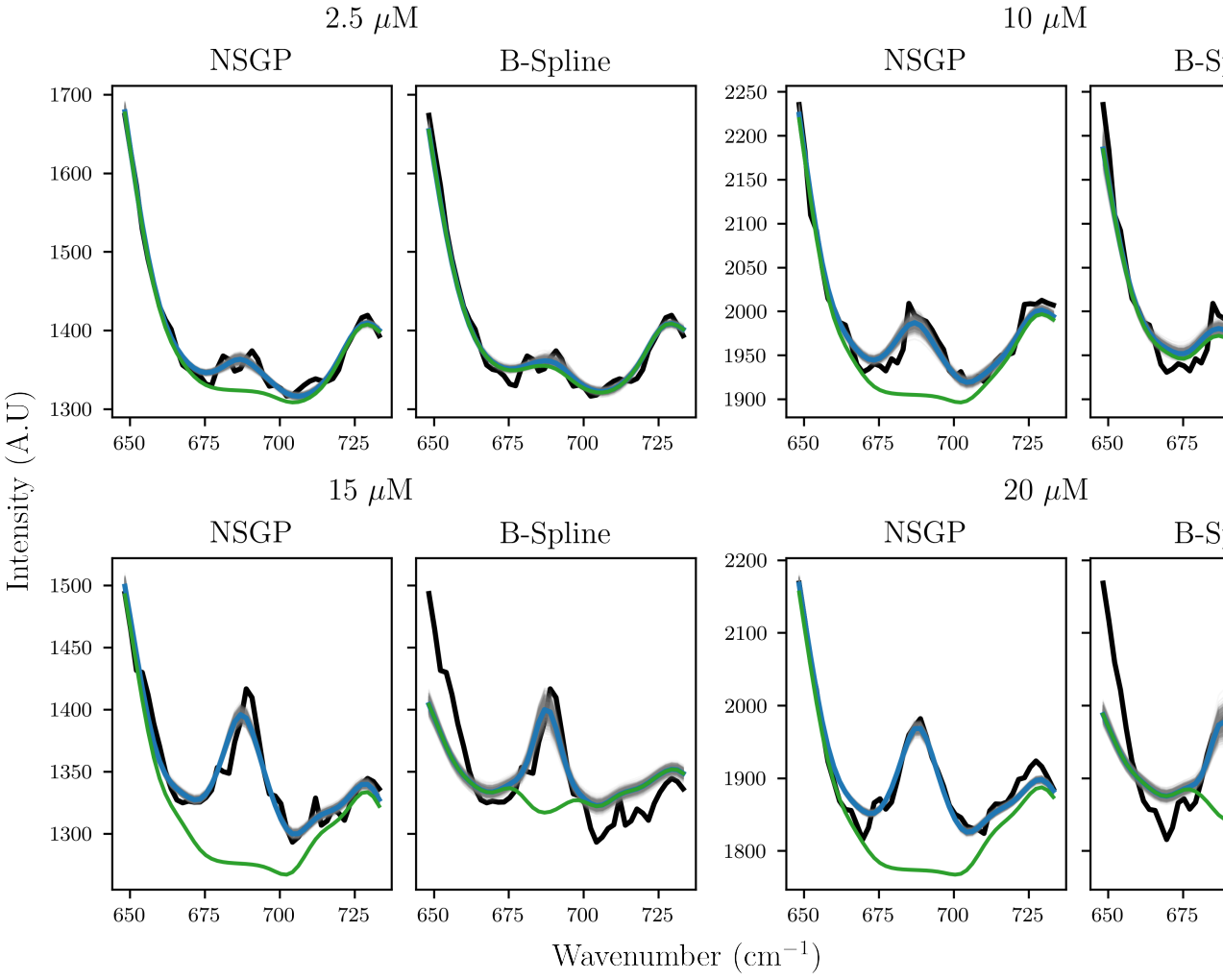

Figure 9: Examples of model reconstructions for the MTX data for different increasing concentrations. Note that these are merely single spectra in select SERS maps. Blue line is the estimated spectrum of the posterior, green line is the expected baseline of the posterior. We see that for especially low concentrations, the NSGP baseline still allows for some of the signal to be explained by the peak model.

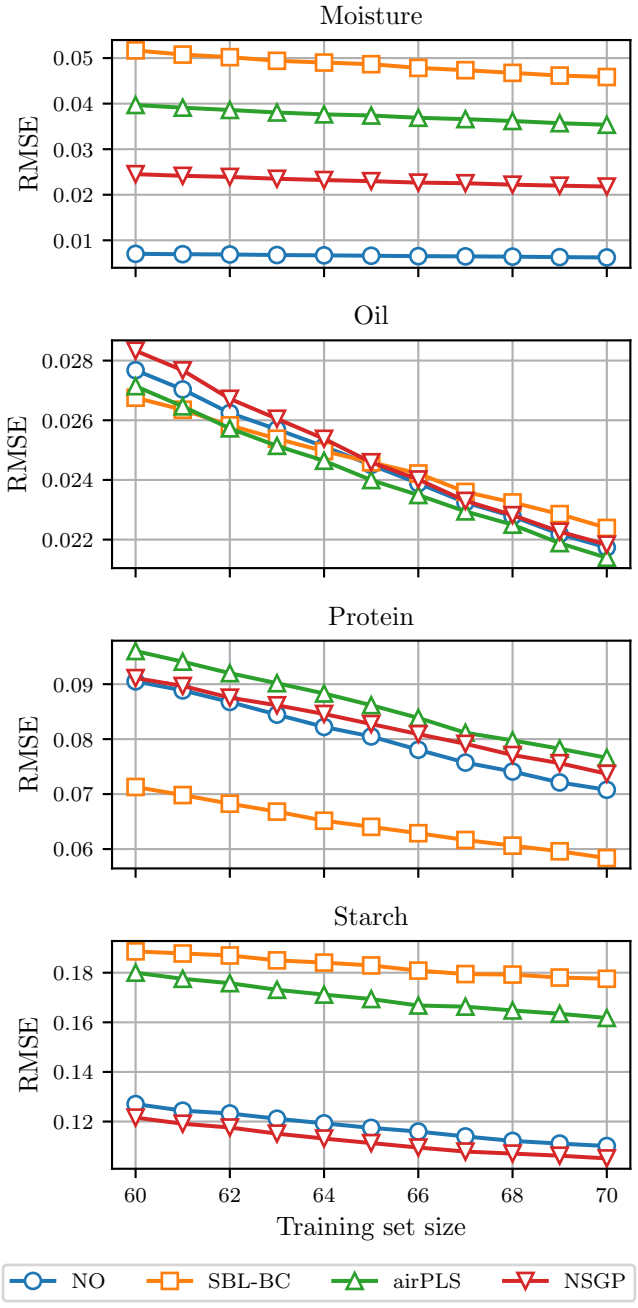

Figure 10: Root mean squared error of prediction on the Corn NIR dataset. Comparison of no baseline correction (NO), adaptive iteratively reweighted penalized least squares (airPLS), sparse Bayesian learning for baseline correction (SBL-BC) and non-stationairy Gaussian process (NSGP) baseline estimation methods for varying sizes of the training data set.

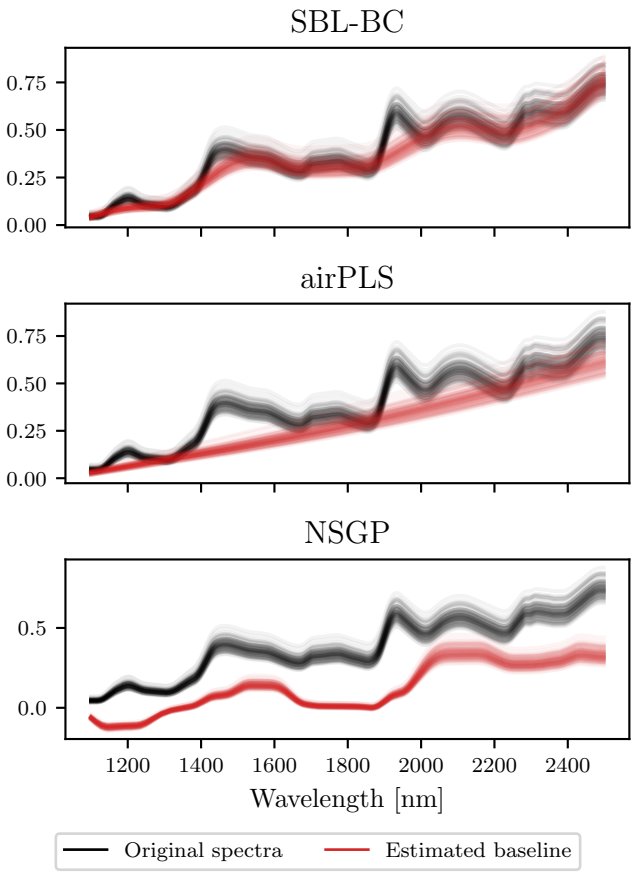

Figure 11: Original sepctra and estimated baselines on the Corn NIR dataset.

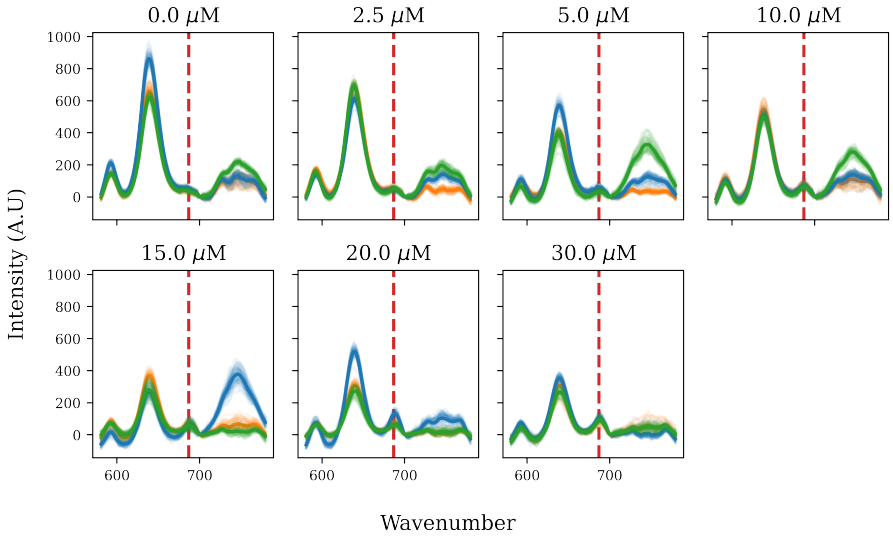

Figure B.12: Raw MTX data translated to be on the same scale (i.e. ensuring $\kappa_n \approx 0$ for all spectra). Note that the signal strength does not increase much as concentration increases.