Bayesian Temporal Signal Extraction on Sparse Sensor Data: Applications in LC-MS

Jacob Lamadrid[1]

[1] *Georgia Institute of Technology*

Peak and noise estimation has been a common challenge in scientific signal analysis due to the necessity to confidently understand and define the points of interest in highly complex signals, particularly in the uncertainty quantification of sensor data. In this project, I propose an application of a Bayesian method for peak and noise estimation with temporal considerations representing comprehensive aggregate analysis and sensor validation. The framework for this applied approach follows techniques provided by Tokuda et al. which focuses on estimating noise variance and the number of peaks via Bayesian model selection and calculating the posterior density of each peak identified. However, for highly complex and massive spectra, complete spectra analysis and MCMC methods are computationally infeasible, instead requiring the exploration of strongly informed priors and Variational Bayes for this task. I apply these methods to real LC-MS recordings in this paper, but applications in neuroscience, cosmology, engineering, etc. exist as well.

I. INTRODUCTION

Signal estimation in spectroscopic data has focused on the complex deconvolution of overlapping peaks, in particular how to estimate the number of these peaks present in addition to noise. In this project, I modify this approach in applying these methods to sparse sensor data for computationally efficient estimation and built-in analysis across recordings. The primary utility of this is in feature extraction and uncertainty quantification for a more comprehensive data analysis pipeline, in which significant signal features and sensor error are simultaneously characterized. This approach improves confidence in signal quality and serves as an effective dimensionality reduction tool for high-dimensional datasets

The primary application of this project will be in signal extraction from large liquid chromatography–mass spectrometry (LC-MS) recordings. Due to factors such as biological noise, chemical variability, and sensor error, basic frequentist methods are unreliable for properly analyzing these spectra across time points. While complex signal processing methods may be capable of signal extraction in many of these spectra, reliability is still a significant concern and additional information about the underlying chemical features and sensor dynamics are left out entirely. This is where Bayesian methods may be utilized in order to provide certainty to these recordings across spectra and provide significant quantitative information for downstream data analysis, such as biomarker discovery and abundance comparison.

In applying this method to highly complex spectra, the utilization of informed priors based on a standard peak detection method, segmenting peak regions representing relevant m/z values, is necessary for efficient peak and noise estimation to occur. Additionally, the use of Variational Inference (VI) is employed to avoid the computational cost and restrictions of the standard Markov Chain Monte Carlo (MCMC). By treating detected peaks as prior parameters, this framework adapts the generative model proposed by Tokuda et al. to model sparse spectra as a probabilistic sum of single peaks and noise.

II. METHODS

A. Data Processing

The dataset used for this LC-MS application is from the Metabolomics Workbench Study ST003061 in which three tea cultivars of *Camellia sinensis* were treated and analyzed by UPLC-Quadrupole Time-of-Flight Mass Spectrometry using an ACQUITY UPLC system in tandem to a SYNAPT G2-Si QTOF mass spectrometer. These recordings are represented by 2,212 spectra over the course of 16 minutes with each spectra containing around 38,000 points. Without modification, this would imply the analysis of around 8.4×10^7 points. For this reason, significant data manipulation and adjustment to the implementation of the provided model is required.

Firstly, gathered tdf data files are converted to mzML files via MSConvert, then read into R via MSnbase. With this, it is possible to extract intensity and m/z (mass-to-charge ratio) arrays, at which point data is binned and the top n peaks in all individual spectra are identified and gathered for use as priors. For the results and visualizations presented in this paper, $n = 5$. This was chosen based on the approximate number of visually detectable peaks per spectrum, but the threshold for pracma's 'findpeaks' function for peak identification was set as

$$\text{threshold} = \beta + 3\left(\frac{\sum_{i=1}^{N} y_i - \text{med}(y)}{N}\right) \tag{2.1}$$

resulting in still many spectra containing less than 5 regions for estimation, in which case only those chosen were analyzed. This is a general estimate of the number of useful peaks based on outlier detection chosen in favor of computational efficiency, but it has also been suggested by Shao et al. that a linear regression model may be trained to predict this n. From these rough estimates, each surrounding window is treated as an individual probabilistic estimation.

This fracturing of the data into smaller subsets allows the RStan models to efficiently run, but still sampling methods are not feasible due to the complexity and high dimensionality of the dataset. With these modifications, the total number of points to be analyzed is reduced roughly to 2×10^5, with this number varying by input due to concentrations of low vs high mass ions.

B. Probabilistic Models

In modeling the data, the generative framework presented by Tokuda et al. is adapted. Each spectral segment consisting of N data points $\{x_i, y_i\}_{i=1}^N$ is modeled as a sum of latent Gaussian peaks, a constant baseline, and additive noise. Here, $y_i \in \mathbb{R}$ represents the **intensity** and $x_i \in \mathbb{R}$ represents the **mass-to-charge ratio (m/z)** as

$$y_i = f(x_i; \mathbf{w}) + \epsilon_i \tag{2.2}$$

$$f(x_i; \mathbf{w}) = \beta + \sum_{k=1}^{K} a_k \phi_k(x_i; \mu_k, \rho_k) \tag{2.3}$$

$$\phi_k(x_i; \mu_k, \rho_k) = \exp\left[-\frac{\rho_k}{2}(x_i - \mu_k)^2\right] \tag{2.4}$$

$$p(y_i \mid x_i, \mathbf{w}, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}[y_i - f(x_i; \mathbf{w})]^2\right) \tag{2.5}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ represents sensor noise. The parameter set $\mathbf{w} = \{\beta, \{a_k, \mu_k, \rho_k\}_{k=1}^K\}$ includes the baseline offset β, the peak amplitude a_k, the m/z position μ_k, and the peak precision ρ_k (inversely related to peak width). While this framework allows for K peaks, applications in sparse signals function at $K = 1$ within each isolated window to ensure computational feasibility.

Additionally, a second model was created for a robust quantification of signal confidence beyond an arbitrary R^2 value, which only represents correlation. The first model is effectively used as a peak model (H_1), with a separate model representing a noise model (H_0) in which predictive capacity is evaluated only with access to the baseline offset, β, and sensor noise, ϵ parameters

$$y_i = \beta + \epsilon_i \tag{2.6}$$

$$p(y_i \mid \beta, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \beta)^2\right) \tag{2.7}$$

As seen in figure 1, these models are evaluated using the Bayesian Information Criterion (BIC) which penalizes model complexity (k) against the maximized log-likelihood ($\hat{L}$). The Peak Model ($k = 5$) is formatted as

$$\mathrm{BIC}_{\mathrm{peak}} = 5\ln(N) - 2\ln(\hat{L}_{\mathrm{peak}}) \tag{2.8}$$

$$\ln(\hat{L}_{\mathrm{peak}}) = \sum_{i=1}^{N} \ln p(y_i \mid x_i, \mathbf{w}, \sigma) \tag{2.9}$$

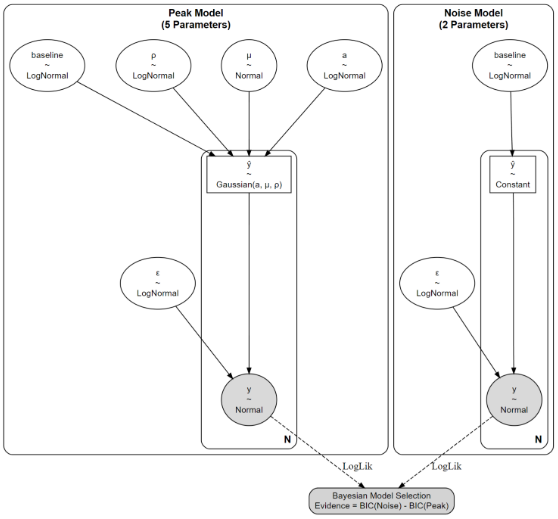

FIG. 1. RStan Model Architecture

and the Noise Model ($k = 2$) as

$$\mathrm{BIC}_{\mathrm{noise}} = 2\ln(N) - 2\ln(\hat{L}_{\mathrm{noise}}) \tag{2.10}$$

$$\ln(\hat{L}_{\mathrm{noise}}) = \sum_{i=1}^{N} \ln p(y_i \mid \beta, \sigma) \tag{2.11}$$

Model selection is performed by calculating the evidence difference $\Delta\mathrm{BIC} = \mathrm{BIC}_{\mathrm{noise}} - \mathrm{BIC}_{\mathrm{peak}}$, with a positive difference indicating evidence for the presence of a signal peak over noise. The evidence required is for quality purposes is 2, with some scores reaching > 50. This BIC value is an approximation used for generic filtering, not strict Bayesian evidence, as well as the effective number of parameters under variational Bayes is ambiguous due to the use of priors.

C. Fitting

Properly fitting this method on complex spectra requires extensive fine tuning depending on the use case. These parameters depend on the specific representation of the signal and the expectation of the underlying mechanisms that produce the data. For these LC-MS spectra, extreme granularity is required with window regions measured in Daltons (Da), optimally looking at windows ±0.4 from the defined position prior. This window allows for proper distribution representation in amplitude, position, and precision (LogNormal, Normal, LogNormal). Significant reductions in precision accuracy occur with greater windows, while position accuracy reduces with smaller windows, both tending toward uniform distributions in their respective cases, resulting from posterior diffusion. The general joint posterior distribution of these parameters with a window of 0.8 Da is shown in figure 2.

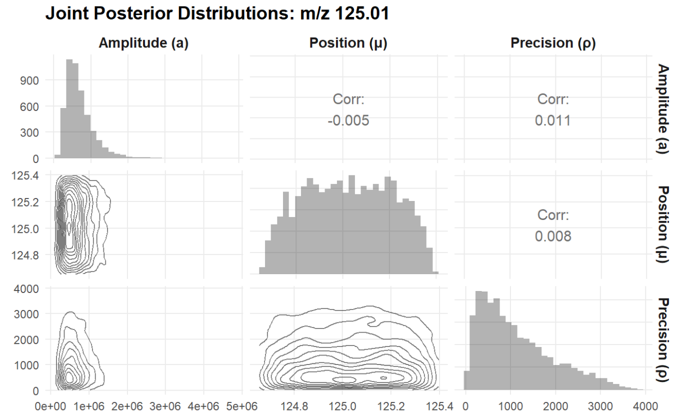

FIG. 2. Posterior Distributions with 0.8 Da Window

As mentioned, informed priors are utilized in fitting, with a set as the height of the identified peak, μ as the location of the max height on the x-axis, β as the median value of the window region, and ϵ as the standard deviation of the region. ρ is the most weakly informed prior in which domain knowledge is utilized over statistical estimation. For the purpose of mass spectrometry, this value is set to 1000.0, representing an expected standard deviation of 0.03 Da or around a 0.12 Da base width of the peak.

These priors are utilized in the proper RStan models, which are then fit via Variational Bayes with the convergence tolerance on the relative norm of the objective function set as 0.005. Drawbacks in accuracy exist when using Variational Inference when compared to traditional MCMC methods due to its reliance on Mean-field approximation. This approximation may underestimate the variance contained in the data, producing predictions more suitable in ideal scenarios. This concern however is reduced in the case of temporal or repeated measures datasets such as LC-MS. These approximations become more stable as the sample size grows, at which point aggregation of regions is possible and quality may be better ensured.

D. Evaluation

Model evaluation is largely focused on sanity of the final spectrum reconstruction in a chemical identity context, posterior predictive distributions, and distributions of metrics in sensor calibration, position, and abundance. Confidence metrics are largely based on the generated "evidence" scores (ΔBIC) as well as the R^2 value. As a posterior predictive check, an example of a confident vs. uncertain peak estimation is provided in figure 3, in which a significant difference in posterior predictive power is presented, largely explained by mass-to-charge ratio, reflective of the velocity of the ion across the sensor, and in some other cases, extreme precision of the peak. This velocity and mass difference is noted in the analysis of the results. This displays the models capabilities in deciphering signal from noise when sufficient data is provided while also providing a reasonable Highest Density Interval (HDI) for uncertainty quantification, a key point of this Bayesian analysis.

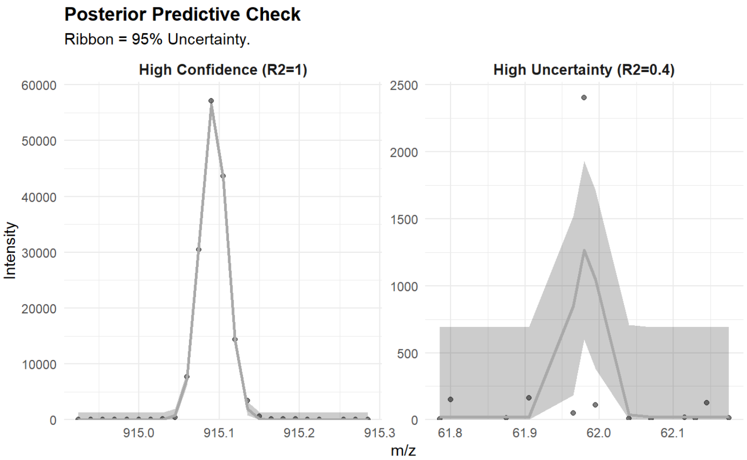

FIG. 3. Posterior Predictive Confidence Examples

III. OUTCOMES

A. Results

The primary result of these methods is the final collapsed reconstruction across recordings which are provided in figure 4 with a single mzML file per final spectrum in which each individual cultivar represented. This reconstruction provides a comprehensive visual aggregate of chemical features at the designated m/z values, indicating specific compounds in combination with other identified peaks. This effectively reduces all information contained in the 2,212 spectra across recordings into a single spectrum. From this we can easily quantify the compositions of different cultivar and subsequently analyze the differences with an estimate of error throughout the process.

From this initial analysis, we see consistencies in the reconstruction as well as clear biomarker and recording differences. Firstly, intensity ranges vary drastically, but relative abundance remains generally consistent, possibly suggesting differences in treatment or settings between sample measurements. We also see similar groups across cultivars as well as specific m/z identifiers, such as 125.01, 289.05, 305.4, and some others. This is expected with these values likely displaying the models ability to accurately identify Catechin and Gallocatechin as well as "a second ion containing important structural information...at m/z 125 in the MS/MS product ion spectra of all of the reference compounds" (Miketova et al.). This par-

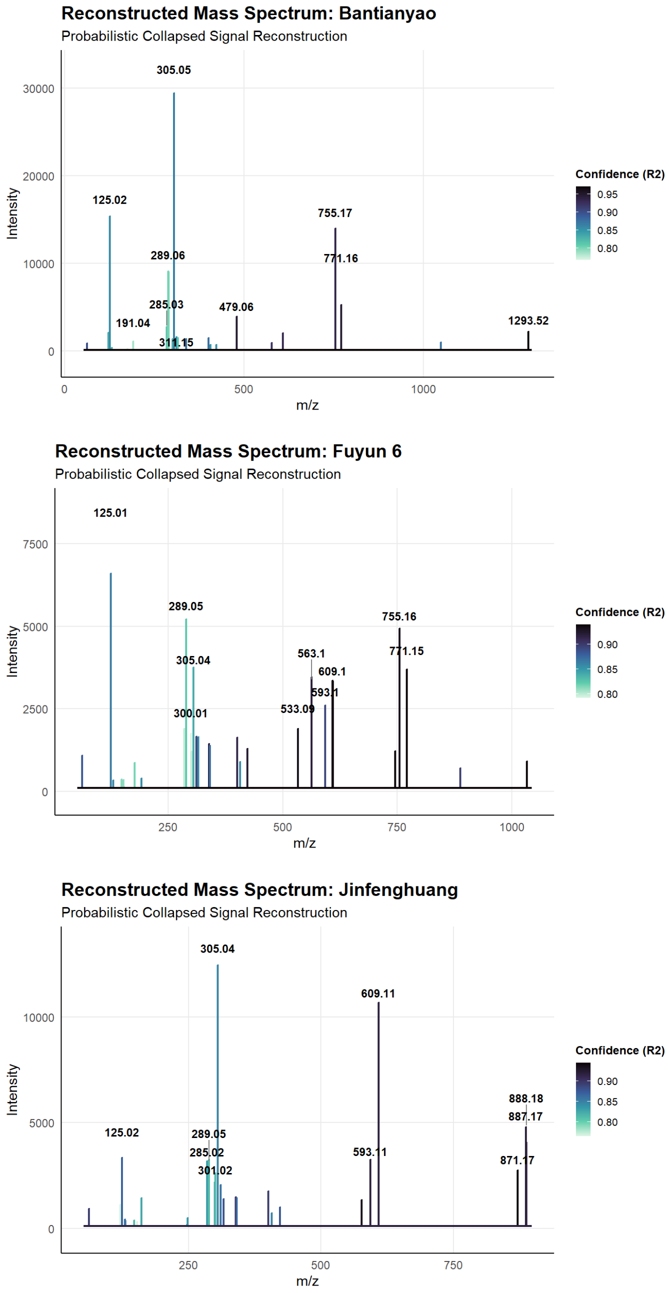

FIG. 4. Probabilistic Reconstructions. Top: Bantianyao. Middle: Fuyun 6. Bottom: Jinfenghuang. Transparency is scaled by R^2.

tially validates the models ability to identify true peaks, and in this case, chemical signatures. It is also worth noting that the criteria for final peaks shown on this graph has not been rigorously optimized, but all necessary data for personal filtering and answering specific scientific questions are available in the output RDS file. By example a large peak at m/z 169 is present in all samples, but at a low confidence and has thus been removed for visibility.

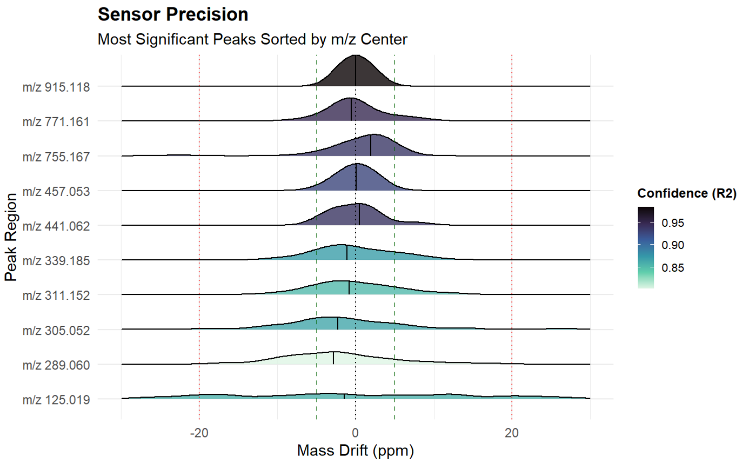

FIG. 5. Sensor Drift Measured in PPM

With this initial validation, much more information can reliably be extracted from this data, particularly in regards to sensor calibration and how this relates to the physical properties of the samples at hand. The most central aspects of this project are in the accurate probabilistic modeling of both intensity and position, both of which are impacted by stochastic error in sensor and biological noise. With these data, we can quantify the error, or drift, of the sensor from its target position as seen in figure 5. We can now quantify uncertainty in the utilization of specific m/z peaks extracted in measuring the expected drift from the mean point in the region and the overall posterior distribution of points.

As seen in the figures in this section, there is a direct correlation with the mass of an ion and the confidence we have in detecting its presence in the sample. This correlation is significant in the final aggregated and filtered data in which a Pearson correlation of 0.649 exists between the R^2 score and the final m/z position. The correlation is even stronger when looking at ΔBIC at 0.787. This significant relationship between position and quality is explained by the underlying physical properties of the samples and the sensor as ions are accelerated by an electric field, with lighter ions travel at higher velocities, while heavier ions travel slower. The lower velocity of high-mass ions allows for greater perceptional power by the detector, resulting in improved peak shape definition and reduced posterior variance compared to the fast-moving low-mass ions. Generally, as speed increases, sensor abilities diminish, as evidenced by the resulting mass drift findings. Overall predictive capabilities drop across all parameters as mass decreases, but this uncertainty is reported by the model and can be modeled in itself.

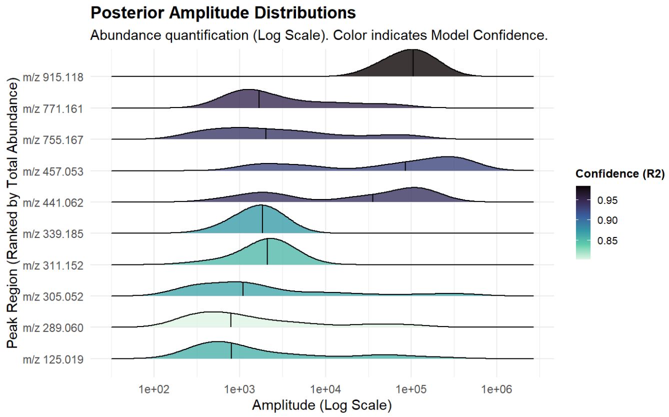

FIG. 6. Amplitude Distributions Across Positions

B. Conclusion

Probabilistic modeling of complex noisy mass spectrometry spectra for composition estimation in terms of both compound type and abundance is possible by many methods, but through Bayesian modeling, the necessary and accurate uncertainty quantification of this data is possible. Through the temporal modeling of spectra across cultivars, we can estimate specific cultivar compositions using a relatively low number of spectra at which point general conclusions of the samples or even of the sensor's abilities may be made. The primary utilization of this method is not in comprehensive biological or chemical analysis, but rather in sensor uncertainty and accurate point identification across samples. As mentioned, this methodology can be easily modified and generalized to other research areas. For example, NASA's lightkurve dataset provides luminosity signals from sensors pointed at other solar systems in which peaks such as solar flares or inversely orbital patterns can be gathered. This is one of many possible applications in sparse signal extraction which this probabilistic model may be used. Future work for this project may include hierarchical modeling across groups, improve peak number estimation as previously mentioned, or improvements in fitting methodology for higher confidence posterior predictions.

IV. REFERENCES

S. Tokuda, *et al.*, "Simultaneous Estimation of Noise Variance and Number of Peaks in Bayesian Spectral Deconvolution," *Journal of the Physical Society of Japan*, vol. 86, no. 2, p. 024001, 2016. DOI: 10.7566/jpsj.86.024001

Metabolomics Workbench, Project ID PR001907. Available at `https://www.metabolomicsworkbench.org`. DOI: 10.21228/M8714X.

V. Mazet, *et al.*, "Unsupervised Joint Decomposition of a Spectroscopic Signal Sequence," *Signal Processing*, vol. 109, pp. 193–205, 2014. DOI: 10.1016/j.sigpro.2014.10.032

G. C. Allen and R. F. McMeeking, "Deconvolution of Spectra by Least-Squares Fitting," *Analytica Chimica Acta*, vol. 103, no. 1, pp. 73–108, 1978.

H. Sayama, "Mean-Field Approximation," in *Introduction to the Modeling and Analysis of Complex Systems*. [Online]. Available: `https://math.libretexts.org/`

P. Miketova, *et al.*, "Tandem Mass Spectrometry Studies of Green Tea Catechins," *Journal of Mass Spectrometry*, vol. 35, no. 7, pp. 860–869, 2000.

W. Shao and H. Lam, "Denoising Peptide Tandem Mass Spectra for Spectral Libraries: A Bayesian Approach," *Journal of Proteome Research*, vol. 12, no. 7, pp. 3223–3232, 2013. DOI: 10.1021/pr400080b