Bayesian–MAP Method for Spectral Decomposition

Louis Bouchard
(Dated: January 27, 2025)

I propose that we test a Bayesian–MAP method for the decomposition of NMR spectra into known and unknown constituents. The main application would be metabolomics, where a complicated NMR spectrum is "fitted" to a set of known metabolite spectra. The approach allows for low-pass baseline modeling of unknown spectral contributions and an optional sparse (L1-penalized) prior on the metabolite coefficients. The method iterates between updating the known metabolite concentrations via least-squares or Lasso and refining the baseline through a smooth filter of the residual. This yields a flexible, computationally simple method for accurate metabolite quantification in complex mixtures. In tests I have done so far, on fake data, I was able to get improved performance compared to a simple least-squares fit. This approach makes it clear how unmodeled spectral components may be captured either probabilistically or via baseline filtering. We show this hybrid approach mitigates some important limitations of pure Gaussian assumptions and diminishes unphysical oscillations in unmodeled terms while giving better recovery of coefficients when there are unknown signals or partial overlaps.

INTRODUCTION

NMR spectroscopy is useful tool in metabolomics, where measurements contain signals from countless chemical species present in complex mixtures such as biological tissues or fluids. Despite its utility, analyzing NMR data becomes difficult when the measured spectrum involves both well-characterized reference components and a vast number of unknown molecules (including metabolites not in a spectral library). Our goal is to decompose the observed spectrum into contributions from known references, typically collected as columns in a dictionary matrix $X_{\text{int}} \in \mathbb{R}^{m\times n}$, and from an unmodeled portion that subsumes any baseline drifts, partial overlaps, or genuinely unknown metabolites. Let $\boldsymbol{\beta} \in \mathbb{R}^n$ denote the concentration (or amplitude) vector for the known dictionary spectra, and let $Z \in \mathbb{R}^m$ encapsulate these unmodeled signals plus baseline structure.

Traditional linear least-squares or purely Gaussian Bayesian treatments can experience difficulties when the unknown portion in Z becomes significant or exhibits slowly varying and partially overlapping peaks. Previous Bayesian approaches [2, 3] manage noise well through Gaussian priors on $\boldsymbol{\beta}$, but an overly flexible baseline model can absorb large segments of the measured data and inadvertently suppress or inflate the dictionary coefficients. In contrast, we propose an iterative maximum *a posteriori* (MAP) method in which Z is explicitly filtered or smoothly constrained, rather than merely modeled as a Gaussian random vector with high covariance. We also introduce an optional ℓ_1-type penalty on $\boldsymbol{\beta}$ (a Lasso prior) that encourages sparsity in the dictionary coefficients if only a few known metabolites are believed to be present. Although the resulting scheme is not purely conjugate, it remains computationally tractable by alternating between a baseline-filtering step and either a least-squares or Lasso update for $\boldsymbol{\beta}$. The reader should be warned that this writeup mainly presents a measure-theoretic grounding for the framework; we follow the notation of Karatzas and Shreve [1], and then develop the numerical procedure that handles NMR spectral decomposition.

PROBLEM FORMULATION

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space, and let $Y\colon \Omega \to \mathbb{R}^m$ denote the random variable representing a measured NMR spectrum of a complex mixture. Suppose a dictionary of n reference spectra is available, arranged into a matrix $X_{\text{int}} \in \mathbb{R}^{m\times n}$, where each column encodes one well-characterized metabolite or known component. Let $\boldsymbol{\beta}_{\text{int}}\colon \Omega \to \mathbb{R}^n$ be a random variable whose realization $\boldsymbol{\beta}_{\text{int}}(\omega)$, $\omega \in \Omega$ corresponds to the amplitude or concentration of each known reference in the observed spectrum $Y(\omega)$. Define $Z\colon \Omega \to \mathbb{R}^m$ to be a random variable capturing unmodeled (unknown) spectral contributions, which may include baseline distortions, overlapping peaks not accounted for in X_{int}, or other systematic effects.

We represent measurement noise by an additional random variable $\epsilon\colon \Omega \to \mathbb{R}^m$, and assume $\epsilon(\omega)$ follows an isotropic Gaussian law $\mathcal{N}(0, \sigma^2 I)$. The essential observation model for a single realization $\omega \in \Omega$ is

$$Y(\omega) \;=\; X_{\text{int}}\,\boldsymbol{\beta}_{\text{int}}(\omega) \;+\; Z(\omega) \;+\; \epsilon(\omega). \tag{1}$$

In standard Gaussian Bayesian treatments, $Z(\omega)$ might also be assigned a broad Gaussian prior. Here, we adopt a different principle. Rather than letting $Z(\omega)$ fluctuate freely with large covariance, we impose a low-pass or slowly varying constraint on Z via a regularizing functional Ψ, consistent with the view that Z should capture slowly varying baseline contributions. We may specify a prior on Z by

$$\mathbb{P}\big(Z \in \mathrm{d}z\big) \;\propto\; \exp\!\big(-\Psi(z)\big)\,\mathrm{d}z, \tag{2}$$

where $\Psi\colon \mathbb{R}^m \to \mathbb{R}_+$ penalizes rapid oscillations or high-frequency content. In practice, one might realize $\Psi(z)$ through a low-pass filter (e.g. Butterworth) or a total variation (TV) norm.

We also allow $\boldsymbol{\beta}_{\text{int}}$ to admit a sparsity-inducing prior. Let $\boldsymbol{\beta}_{\text{int}}(\omega)$ have density proportional to

$$\exp\Big(-\tfrac{1}{2}\,(\boldsymbol{\beta}_{\text{int}}-\mu_{\text{int}})^T\,\Sigma_{\text{int}}^{-1}\,(\boldsymbol{\beta}_{\text{int}}-\mu_{\text{int}})\;-\;\lambda\,\|\boldsymbol{\beta}_{\text{int}}\|_1\Big),$$

where $\mu_{\text{int}} \in \mathbb{R}^n$ and $\Sigma_{\text{int}} \in \mathbb{R}^{n\times n}$ define a baseline Gaussian prior, and the term $\lambda\,\|\boldsymbol{\beta}_{\text{int}}\|_1$ enforces an ℓ_1 penalty for Lasso-type sparsity. Setting $\lambda = 0$ recovers a purely Gaussian prior.

From a measure-theoretic viewpoint, we combine these priors and the likelihood implied by (1) to form a posterior distribution on $(\boldsymbol{\beta}_{\text{int}}, Z)$, typically expressed as

$$\mathrm{d}\mathbb{P}\big(\boldsymbol{\beta}_{\text{int}}, Z \mid Y\big)\;\propto\;\exp\Big(-\tfrac{1}{2}\,\|Y-X_{\text{int}}\,\boldsymbol{\beta}_{\text{int}}-Z\|^2_{\sigma^{-2}I}\Big)\,\exp\big(-\Psi(Z)\big)\,\exp\Big(-\tfrac{1}{2}\,(\boldsymbol{\beta}_{\text{int}}-\mu_{\text{int}})^T\,\Sigma_{\text{int}}^{-1}\,(\boldsymbol{\beta}_{\text{int}}-\mu_{\text{int}})\;-\;\lambda\,\|\boldsymbol{\beta}_{\text{int}}\|_1\Big). \tag{3}$$

A maximum *a posteriori* (MAP) estimate for $(\boldsymbol{\beta}_{\text{int}}, Z)$ is then obtained by maximizing (3) or, equivalently, minimizing the sum of the squared residual, the low-pass penalty $\Psi(Z)$, and the Gaussian-plus-sparsity prior on $\boldsymbol{\beta}_{\text{int}}$. Because the low-pass constraint on Z and the ℓ_1 penalty on $\boldsymbol{\beta}_{\text{int}}$ are both non-Gaussian, there is no closed-form solution; the resulting MAP problem is solved by iterative, alternating updates of $\boldsymbol{\beta}_{\text{int}}$ and Z. Although this partial loss of conjugacy complicates the analysis, it grants a more realistic treatment of baseline drifts and the potential for sparsity in the metabolite coefficients (and hopefully, enabling a more accurate spectral decomposition of Y).

ITERATIVE ALGORITHM

We partition the unknowns into (β_{int}, Z). Although each alone is simpler to solve for, their coupling in (1) precludes a closed-form joint solution. Hence, we proceed with the following steps:

1. **(Beta-step)**: Fix Z and solve for β_{int}.
   - If $\lambda > 0$ (sparsity), we solve a Lasso problem:

   $$\hat{\beta} = \arg\min_{\beta}\;\tfrac{1}{2}\|(Y-Z)-X_{\text{int}}\beta\|^2 + \lambda\|\beta\|_1.$$

   - If $\lambda = 0$ (pure Gaussian prior), we solve ordinary least-squares or a ridge-regularized version if needed.
2. **(Baseline-step)**: Fix β_{int} and update Z as the low-pass or otherwise-penalized solution of

   $$\hat{Z} = \arg\min_{Z}\;\tfrac{1}{2}\|Z-(Y-X_{\text{int}}\,\beta_{\text{int}})\|^2 + \Psi(Z).$$

   For a simple *low-pass filter*, we set $Z = \text{Filter}_{\text{lowpass}}\big(Y - X_{\text{int}}\,\beta\big)$.

We iterate these two steps until convergence. Although each step may be cast in measure-theoretic or variational form, the resulting procedure is straightforward in practice.

If we attempted a fully Bayesian treatment of Z via (2) with $\Psi(Z)$ corresponding to a complicated non-Gaussian prior (like an indicator of low-frequency content), we would sample from or maximize the posterior (3). The alternating approach is a common deterministic approximation to the resulting MAP system.

IMPLEMENTATION DETAILS

In practice, X_{int} is built from known metabolite spectra. In the code I sent you, I have simulated NMR spectra (fake data). However, in practice, you would load them from real data files:

$$X_{\text{int}} = \big[\;\underline{x_1}\;\;\underline{x_2}\;\;\ldots\;\;\underline{x_n}\big],$$

where each $\underline{x_i} \in \mathbb{R}^m$ is a (possibly normalized) spectrum.

The baseline Z is computed from the residual $R = Y - X_{\text{int}}\beta$ via a low-pass filter. If b, a are the Butterworth filter coefficients of user-selected cutoff ω_{cut}, we solve

$$Z = \text{filtfilt}(b, a,\; R).$$

This imposes that Z has minimal high-frequency content. Alternately, one might use total variation or wavelet-based priors [4].

For the *Beta-step*, we either proceed according to no sparseness or L1-sparseness. For no sparseness, we solve the classical least-squares $\hat{\beta} = (X_{\text{int}}^T X_{\text{int}})^{-1} X_{\text{int}}^T (Y - Z)$ or a ridge-regularized problem. For L1-sparseness we solve $\min_\beta\;\tfrac{1}{2}\|(Y-Z)-X_{\text{int}}\beta\|^2 + \lambda\|\beta\|_1$, via Lasso routines (coordinate descent, proximal gradient, etc.).

Algorithmic Steps

1. **Initialize** $\beta^{(0)} = 0$, $Z^{(0)} = 0$.
2. For $k = 1, 2, \ldots$
   (a) **(Beta-step)**:

$$\beta^{(k)} \leftarrow \begin{cases} \arg\min_\beta \|Y - Z^{(k-1)} - X_{\text{int}}\beta\|^2, & \text{(no L1)} \\ \arg\min_\beta \frac{1}{2}\|Y - Z^{(k-1)} - X_{\text{int}}\beta\|^2 + \lambda\|\beta\|_1, & \text{(L1-sparse)} \end{cases}$$

   (b) **(Z-step)**:

$$R = Y - X_{\text{int}}\beta^{(k)}, \quad Z^{(k)} = \text{filtfilt}(b, a,\ R),$$

   or an equivalent baseline penalty approach.
3. Stop when $\|\beta^{(k)} - \beta^{(k-1)}\| + \|Z^{(k)} - Z^{(k-1)}\| < \varepsilon$.

NOTES

This method generalizes earlier Gaussian Bayesian methods by explicitly modeling the unmodeled portion Z as a low-frequency baseline rather than a broad Gaussian with large covariance. We also allow L1-sparse priors on β to handle large dictionaries of possible reference spectra. Despite the loss of conjugacy, the approach remains computationally efficient: each iteration solves a straightforward linear or Lasso subproblem for β, plus a single filtering step for Z. Numerical experiments, which you are welcome to experiment with using the attached code, demonstrate substantial improvement in matching *true* metabolite concentrations if unmodeled signals are truly slow-varying or if only a few references are active. The method also eliminates large oscillations and overfitting in Z, since the baseline is constrained to remain smooth. Finally, it offers greater stability compared to unconstrained Bayesian baselines that could absorb the entire signal, forcing β to zero. In many real NMR datasets, baseline drift is the dominant unknown effect. Low-pass filtering thus effectively isolates the low-frequency baseline, letting the dictionary handle distinct spectral peaks.

POSSIBLE EXTENSIONS

This Bayesian–MAP approach to NMR spectral decomposition integrates a smooth, low-pass baseline model and an optional sparse prior on the dictionary coefficients. Our measure-theoretic rationale treats the unknown baseline Z as having a non-Gaussian prior $\exp(-\Psi(Z))$ and the coefficients β as possibly L1-penalized. An alternating iterative scheme yields solutions reflecting the correct division between known references and slowly varying unknown signals. Further refinements could include replacing simple low-pass with total variation or wavelet-based priors on Z. We could also include positivity constraints on β for physically meaningful metabolite concentrations. We could also use hierarchical priors for large-scale metabolomics applications, where multiple spectra share hyperparameters.

[1] I. Karatzas and S. Shreve, *Brownian Motion and Stochastic Calculus*. Springer, 1991.

[2] G. O. Roberts and J. Rosenthal, "Bayesian spectral analysis and model selection," *Ann. Statist.*, vol. 32, no. 3, pp. 561–594, 2004.

[3] A. Gelman *et al.*, *Bayesian Data Analysis*. CRC Press, 2013.

[4] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D*, vol. 60, pp. 259–268, 1992.