

Bayesian Metrology in Metabolomics

Duke University

Andrew Cron

January 27, 2011

Introduction

The Model

Informative Priors

Chemistry/Physics

Biological Knowledge and Machine Information

Inference and Example

Future Work

The Big Picture

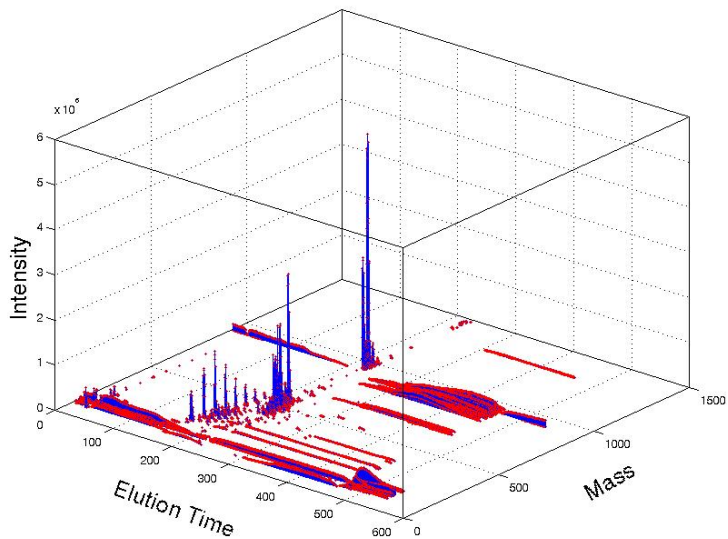
Metabolites are light-weight molecules that are intermediates or products in metabolism. Recent technology has made it possible to estimate the abundance of different metabolites in tissue samples from plants and animals. This has launched the new science of **metabolomics**.

The Big Picture

Metabolites are light-weight molecules that are intermediates or products in metabolism. Recent technology has made it possible to estimate the abundance of different metabolites in tissue samples from plants and animals. This has launched the new science of **metabolomics**.

Metabolomics is often compared with genomics and proteomics, but there are reasons to hope that statistical inference with metabolite data will be easier than for genes or proteins due to our vast prior knowledge.

Mass Spec Data



Non-Linear Regression for 2D Peaks

Let Y_i be the i^{th} intensity and X_i be the corresponding Elution Time and Mass. We will define a non-linear regression model:

$$Y_i = \sum_{j=1}^J \beta_j f_j(X_i; \mu_j, \sigma_j) + \epsilon_i \text{ where } f_j : \mathbb{R}^2 \rightarrow \mathbb{R}.$$

Furthermore, we will want f_j to have its only mode at μ_j . For our current example, we will let f_j be a bivariate gaussian kernel with mode μ_j and variance σ^2 . This can (and should) be generalize to a more flexible shape.

While it may not be completely reasonable, we will also assume that ϵ_i are independent normal errors.

Isotopes, Adducts, and Dimers

- ▶ **Isotopes:** Given the mass, chemical formulation, and intensity of any peak in an isotopic series, the mass and intensities of the other peaks are known (up to machine error). Even if the chemical formulation is not known, the masses are known.

Isotopes, Adducts, and Dimers

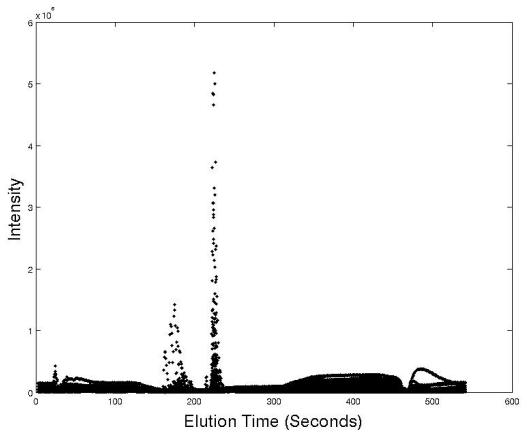
- ▶ **Isotopes:** Given the mass, chemical formulation, and intensity of any peak in an isotopic series, the mass and intensities of the other peaks are known (up to machine error). Even if the chemical formulation is not known, the masses are known.
- ▶ **Adducts:** The masses are known if they exist, but the chemical abundance can vary across experiments.
- ▶ **Dimers:** Similar to adducts.

Isotopes, Adducts, and Dimers

- ▶ **Isotopes:** Given the mass, chemical formulation, and intensity of any peak in an isotopic series, the mass and intensities of the other peaks are known (up to machine error). Even if the chemical formulation is not known, the masses are known.
- ▶ **Adducts:** The masses are known if they exist, but the chemical abundance can vary across experiments.
- ▶ **Dimers:** Similar to adducts.

- ▶ **Punchline!** These translate trivially to a hierarchical prior on μ_j .

Elution Time



We know a compound's peaks will all have the same elution time. However, the elution time is measured with more uncertainty and compounds often overlap.

Biological Knowledge and Machine Information

Biologists have a good idea of what metabolites are in certain tissue samples. The less known factor is their abundance. Furthermore, certain well understood compounds are in every experiment for machine calibration.

Finally, the machines are well understood, so there is plenty of prior knowledge on the variances.

Bayesian Inference

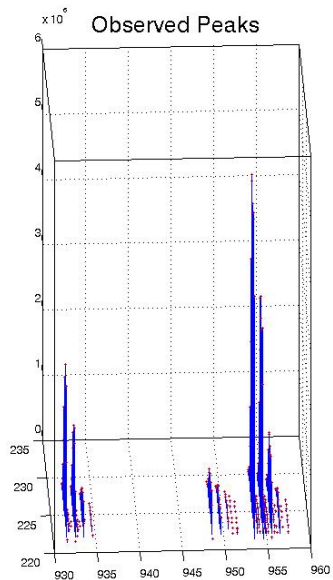
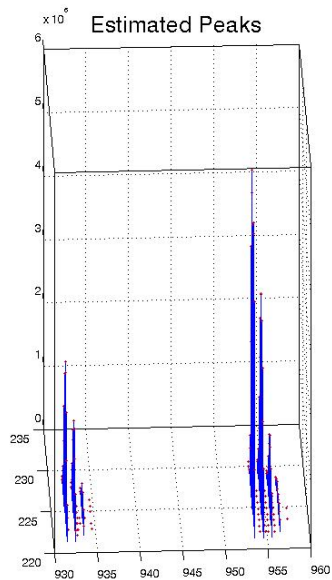
All of this information can be easily translate into a set of informative and dependent priors on μ_j , β_j , and σ_j . For instance, for the isotopic series, let $\mu_{1,1} \sim N(m, s)$. Then define $\mu_{j,1} | \mu_{1,1} \sim N(\mu_{1,1} + (j - 1), s_j)$.

Inference: We can use MCMC for posterior inference. However, due to the non-linearity of the model, Metropolis Hastings can be used updating one peak at a time.

Example with Glycocholate

Consider the case when only one known compound is analyzed in the mass spectrometer. Our goal will be to find that compound and determine how much of it is in the sample.

Results



Future Work

- ▶ This can be trivially extended to locate many known peak patterns by letting

$$Y_i = \sum_{k=1}^K w_k \sum_{j=1}^J \beta_{j,k} f_{j,k}(X_i; \mu_{j,k}, \sigma_{j,k}) \text{ where } \sum_{k=1}^K w_k = 1.$$

The w_k essentially summarize the tissue sample's metabolic state.

- ▶ A semi-parametric form of the peaks should be considered, so that the decay around the peak is not assumed to look like a normal, etc.
- ▶ There are known issues with the intensity detectors (saturation) that needs to be accounted for in the model to get an accurate metabolic picture.

Thank You!

Questions?