Modeling Background Noise for Denoising in Chemical Spectroscopy

June 29, 2009

Richard Barnard
Department of Mathematics Louisiana State University

Talk Outline

MALDI-TOF Mass Spectrometer

We will consider data sets obtained via **M**atrix **A**ssisted **L**aser **Desorption**/**I**onization **T**ime **O**f **F**light Mass Spectrometer.

- Analyte sample is placed in a matrix solution.

MALDI-TOF Mass Spectrometer

We will consider data sets obtained via **M**atrix **A**ssisted **L**aser **Desorption**/**I**onization **T**ime **O**f **F**light Mass Spectrometer.

- Analyte sample is placed in a matrix solution.
- Pulsed laser fired at mixture, ionizing analyte.

MALDI-TOF Mass Spectrometer

We will consider data sets obtained via **M**atrix **A**ssisted **L**aser **Desorption**/**I**onization **T**ime **O**f **F**light Mass Spectrometer.

- Analyte sample is placed in a matrix solution.
- Pulsed laser fired at mixture, ionizing analyte.
- Analyte ions travel along a path of known length, striking a detector.

MALDI-TOF Mass Spectrometer

We will consider data sets obtained via **M**atrix **A**ssisted **L**aser **Desorption**/**I**onization **T**ime **O**f **F**light Mass Spectrometer.

- Analyte sample is placed in a matrix solution.
- Pulsed laser fired at mixture, ionizing analyte.
- Analyte ions travel along a path of known length, striking a detector.
- Time of flight can be used to determine mass to charge ratio.

Mass Spectrum

Resulting data is a set of 50,000-100,000 data pairs (time/mass-to-charge ratio and intensity). Our spectra will be from SRM 2881, a polystyrene, obtained from NIST. Noise from various sources can lead to uncertainty (see Guttman, Flynn, Wallace, and Kearsley 2009).

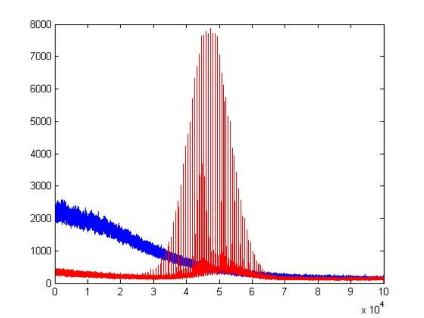


Figure: Analyte(red) and corresponding background(blue), low noise

Mass Spectrum

Resulting data is a set of 50,000-100,000 data pairs (time/mass-to-charge ratio and intensity). Our spectra will be from SRM 2881, a polystyrene, obtained from NIST. Noise from various sources can lead to uncertainty (see Guttman, Flynn, Wallace, and Kearsley 2009).

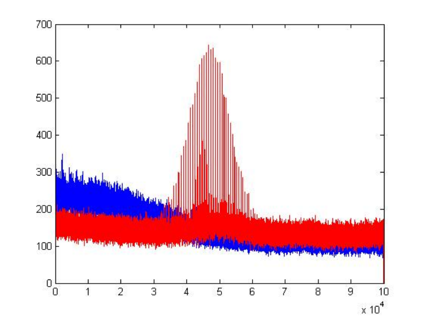


Figure: Analyte(red) and corresponding background(blue), with noise

Mass Spectrum

Resulting data is a set of 50,000-100,000 data pairs (time/mass-to-charge ratio and intensity). Our spectra will be from SRM 2881, a polystyrene, obtained from NIST. Noise from various sources can lead to uncertainty (see Guttman, Flynn, Wallace, and Kearsley 2009).

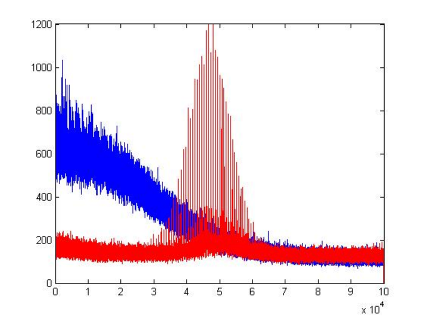


Figure: Analyte(red) and corresponding background(blue),with noise

Overview

- Fit background spectrum to stochastic differential model

Overview

- Fit background spectrum to stochastic differential model
- Determine the mean and variance of noise

Overview

- Fit background spectrum to stochastic differential model
- Determine the mean and variance of noise
- Segment spectrum

Overview

- Fit background spectrum to stochastic differential model
- Determine the mean and variance of noise
- Segment spectrum
- Use Tikhonov regularization on each segment

Background Model

We fit the analyte-free spectrum to a Stochastic Differential Equation with time dependent coefficients

$$dX_t = (a_0(t) + a_1(t)X_t)dt + b_0(t)X_t(t)dW_t$$

$\{W_t\}$ is a Wiener Process, $W_t - W_s \sim N(0, t-s), s < t$

Discretization

Given the background data $\{X(i)\}$ at discrete points, we use Euler-Maruyama discretization:

$$\Delta X(i) = (a_0(i) + a_1(i)X(i)\delta + b_0(i)X(i)\Delta W_i \qquad (1)$$

Discretization

Given the background data $\{X(i)\}$ at discrete points, we use Euler-Maruyama discretization:

$$\Delta X(i) = (a_0(i) + a_1(i)X(i)\delta + b_0(i)X(i)\Delta W_i \qquad (1)$$

Given a window size for regression h, we use the *Epanechnikov Kernel*

$$K_h(z) = \frac{3}{4h}(1 - z^2)$$

for $z \in (-1, 0)$ and $K_h \equiv 0$ off $(-1, 0)$.

Estimating a_0, a_1

In order to estimate a_0, a_1 at each i, we look to minimize

$$\min_{a_0,a_1} \sum_{j=1}^{N} \left(\frac{X(j+1)-X(j)}{\delta} - a_0(i) - a_1(i)X(j)\right)^2 K_h\left(\frac{\delta(j-i)}{h}\right). \tag{2}$$

Estimating a_0, a_1

In order to estimate a_0, a_1 at each i, we look to minimize

$$\min_{a_0, a_1} \sum_{j=1}^{N} \left(\frac{X(j+1) - X(j)}{\delta} - a_0(i) - a_1(i)X(j)\right)^2 K_h\left(\frac{\delta(j-i)}{h}\right). \tag{2}$$

For $Y(j) = X(j+1) - X(j),\ \tau_{ij} = \frac{\delta(j-i)}{h}$

$$\tilde{a}_0(i) = \frac{\sum Y(j)K_h(\tau_{ij}) - \delta a_1(i)K_h(\tau_{ij})}{\delta K_h(\tau_{ij})}$$

Estimating a_0, a_1

In order to estimate a_0, a_1 at each i, we look to minimize

$$\min_{a_0,a_1} \sum_{j=1}^{N} \left(\frac{X(j+1)-X(j)}{\delta} - a_0(i) - a_1(i)X(j)\right)^2 K_h\left(\frac{\delta(j-i)}{h}\right). \tag{2}$$

For $Y(j) = X(j+1) - X(j),\ \tau_{ij} = \frac{\delta(j-i)}{h}$

$$\begin{aligned} \tilde{a}_1(i) &= \frac{1}{\delta(\sum K_h(\tau_{ij}) \sum K_h(\tau_{ij}) X(j)^2 - (\sum K_h(\tau_{ij}) X(j))^2)} \\ &* \quad (\sum K_h(\tau_{ij}) \sum Y(j) X(j) K_h(\tau_{ij}) \\ &- \quad \sum Y(j) K_h(\tau_i j) \sum X(j) K_h(\tau_{ij})) \end{aligned}$$

Estimating b_0

Therefore $\Delta X(i) - (\tilde{a}_0(i) + \tilde{a}_1(i)X(i))\delta \approx b_0(i)X(i)\Delta W_i$,
We set

$$\tilde{E}_i = \frac{\Delta X(i) - (\tilde{a}_0(i) + \tilde{a}_1(i)X(i))\delta}{\delta}$$

Then we find $\tilde{b}_0(i)$ by maximizing at each i

$$-\frac{1}{2}\sum_{j=1}^{N} K_h(\tau_{ij})(\log(b^2X^2(i)) + \frac{\tilde{E}_i^2}{b^2X^2(i)}. \tag{3}$$

$$\tilde{b}_0(i) = \frac{\sum_{j=1}^{N} K_h(\tau_{ij})\tilde{E}_i^2|X(i)|^{-2}}{\sum_{j=1}^{N} K_h(\tau_{ij})} \tag{4}$$

Mean and Variance

$E[X(t)]$ solves the initial value problem

$$y'(t) = a_0(t) + a_1(t)y(t), y(0) = X(0)$$

which we solve using a first order forward Euler scheme. The variance of the noise is given by

$$\delta(b_0(t)X(t))^2$$

Segmentaton

- We want to use denoising algorithms that take advantage of knowledge about the noise.

Segmentaton

- We want to use denoising algorithms that take advantage of knowledge about the noise.
- Many assume constant variance of the noise.

Segmentaton

- We want to use denoising algorithms that take advantage of knowledge about the noise.
- Many assume constant variance of the noise.
- We partition the data and take an approximation of the variance on each segment.

Segmentation

Given a number L we partition the background spectrum into L intervals, I_ℓ, such that

$$||\sigma(t)|_{I_\ell}||_1 = \frac{1}{L}||\sigma(t)||_1 \qquad (5)$$

where σ is the variance of the background spectrum.

Tikhonov Regularization

We look to minimize

$$f_{\lambda,L}(x_{est}) = ||x_{est} - x_{obs}||_2^2 + \lambda||Lx_{est}||_2^2$$

Parameter Selection and Segmentation

UPRE is an unbiased estimator of the mean squared error of predictive error P_λ of,

$$\frac{1}{N}||P_\lambda||^2 = \frac{1}{N}||x_\lambda - x_{true}||^2,$$

UPRE Cont'd

We use the following UPRE functional,

$$\begin{aligned} U(\lambda) &= E(\frac{1}{N}||P_\lambda||^2) \\ &= \frac{1}{N}||r_\lambda||^2 + \frac{2\sigma^2}{N} trace(A_\lambda) - \sigma^2, \end{aligned}$$

where r_λ is the residual and $A_\lambda = (I + \lambda I)^{-1}$. We can take the mean of $\sigma(t)$ for the above σ

UPRE Cont'd

We use the following UPRE functional,

$$\begin{aligned} U(\lambda) &= E(\frac{1}{N}||P_\lambda||^2) \\ &= \frac{1}{N}||r_\lambda||^2 + \frac{2\sigma^2}{N} trace(A_\lambda) - \sigma^2, \end{aligned}$$

where r_λ is the residual and $A_\lambda = (I + \lambda I)^{-1}$. We can take the mean of $\sigma(t)$ for the above σ The optimal λ is defined to be,

$$\lambda_{opt} = \min_{\lambda} \{U(\lambda)\}.$$

Algorithm Summary

Given h, ϵ, L, background spectrum, and analyte spectrum:

1. Fit background spectrum to discreted stochastic model, using h for regression
2. Partition time/mass-per-charge interval into segments
3. Use UPRE to establish on each corresponding segment of analyte spectrum an optimal λ and use Tikhonov regularization
4. Repeat (2) and (3) with increased number of segments until improvement in normalized L^1 is less than ϵ or number of segments is equal to L.

Noise Model

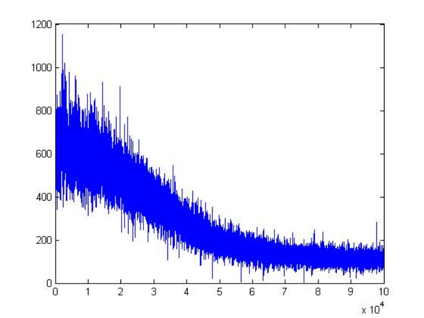


Figure: Simulated Background Spectrum from Noise model for 2nd Noisy Spectrum

Denoising Results

With $h = 10$, tolerance at .001, and max iterations 20,

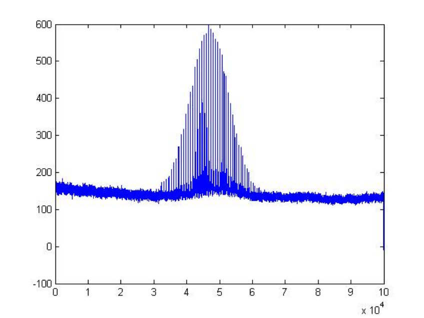


Figure: 1st Noisy Set

Denoising Results

With $h = 10$, tolerance at .001, and max iterations 20,

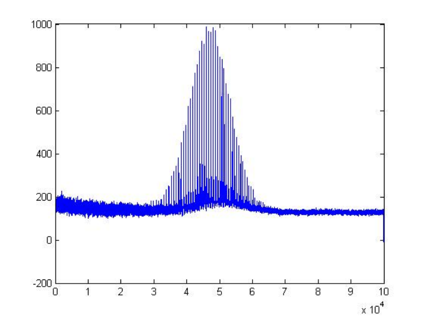


Figure: 2nd Noisy set

Normalized Denoised Results

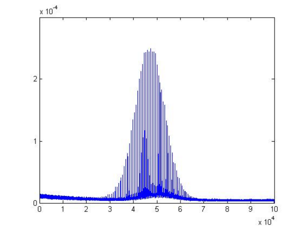


Figure: Low Noise Spectrum divided by its L^1

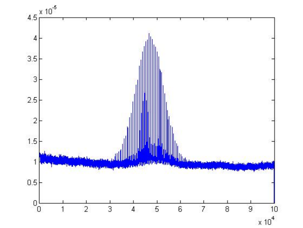


Figure: 1st Noisy Spectrum, Denoised, similarly normalized

Normalized Denoised Results

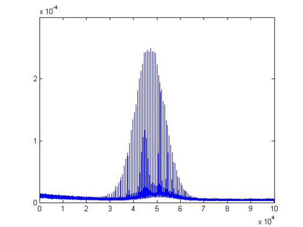


Figure: Low Noise Spectrum divided by its L^1

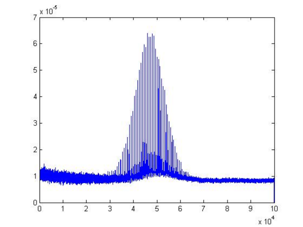


Figure: 2nd Noisy Spectrum, Denoised, similarly normalized

Normalized Denoised Results

Normalized L^1 distance from Best Set and Noisy Spectrum

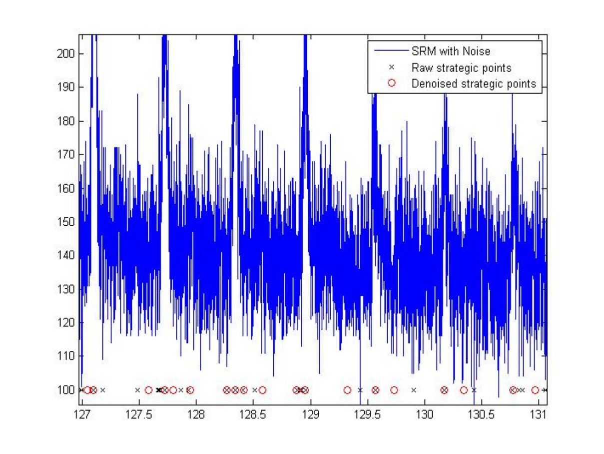
	Noisy	Denoised
1st Set	.5720	.5682
2nd Set	.4950	.4912

Strategic Points

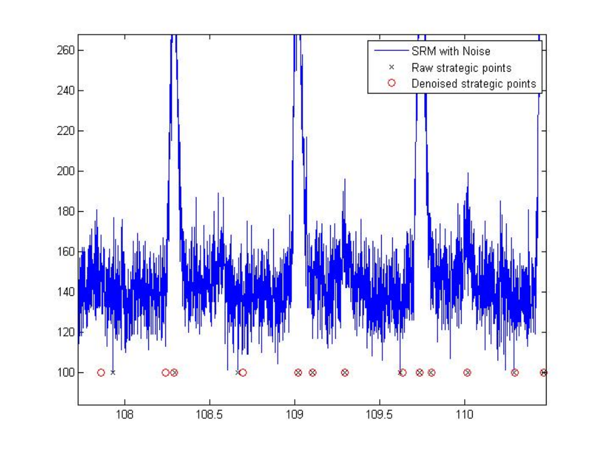
We create a set of strategic points using the following algorithm

1. Set first and last data points as strategic points
2. Find data point with maximum orthogonal distance from line segment connecting two consecutive strategic points
3. This point becomes a new strategic point
4. Repeat until maximal orthogonal distance is below prescribed tolerance

Strategic Points



Strategic Points



Conclusions

- Modeled noise by SDE
- Created an algorithm to denoise spectrum by segmentation
- Smoothes without moving peak locations

Future Work

- Peak height is reduced, possibly fit strategic point height to pre-denoised level
- Investigate other regularization techniques
- Filter strategic points to remove insignificant peaks for better estimation of oligomer peaks

Thank You!