MCMC methods for NEMS Mass Spectrometry

PERENON Rémi¹, Ali Mohammad-Djafari², Pierre Grangeat¹

remi.perenon@cea.fr

15 November 2011
1 CEA Leti – DTBS – STD – LE2S – PROTIS Team
2 UMR 8506 (CNRS – Université Paris Sud – Supélec) LSS – GPI

DRT/LETI/DTBS/STD/LE2S 11-254
Outline

- Introduction
  - Mass spectrometry problem
  - NEMS opportunity

- Direct problem
  - Physics
  - Bayesian modeling

- Inversion
  - Estimation : MCMC method
  - Sampling the “list”

- Results & conclusion
Outline

- Introduction
  - Mass spectrometry problem
  - NEMS opportunity

- Direct problem
  - Physics
  - Bayesian modeling

- Inversion
  - Estimation : MCMC method
  - Sampling the “list”

- Results & conclusion
Mass spectrometry problem

- **Proteomic**
  - Proteome: whole proteins. Proteomic: study of proteome
  - Biomarker discovery / quantification
  - Clinical applications
    - Bacteria recognition
    - Early cancer detection (CEA Leti PROTIS)
  - Device used: mass spectrometer

- **Mass Spectrometer goal**
  - Mass spectra estimation
  - Quantification
  - Mass estimation

- **Performance criteria**
  - Sensitivity: How many molecules needed?
  - Mass resolution?
  - Mass range...
NEMS opportunity 1/2

- Current MS
  - Many devices
    - Ion traps
    - Quadrupoles
    - ToF
    - ...
  - Based on a FLOW mode

![Diagram of separation and detection in MS](image)

State-of-the-art MS

- Ions flow
- Electrons flow
- quantification
NEMS opportunity 2/2

- Flow mode => Counting mode

- Potentially sensitive to a single molecule!
Outline

- Introduction
  - Mass spectrometry problem
  - NEMS opportunity

- Direct problem
  - Physics
  - Bayesian modeling

- Inversion
  - Estimation : MCMC method
  - Sampling the “list”

- Results & conclusion
Physics 1/2

\[ RF \propto \sqrt{\frac{k}{M}} \]

\[ RF_{M+\Delta m} = RF_M - \alpha \cdot \Delta m \]

\[ y = -\sum_{i=1}^{N} m_i \cdot h(t_i) \]

\( h \): infinite impulse response

Incident molecules

Adsorption

Resonant frequency reading

PLL / SO

Observed signal

Impulse response

Time
Naik et al., Towards single molecule nanomechanical mass spectrometry, Nature Nanotechnology, 2009, DOI: 10.1038/NNANO.2009.152
Bayesian modeling 1/5

- Why Bayesian modeling?
  - Stronger estimation due to *a priori*
  - Stronger estimation if EAP used
  - Problem ill-posed without *a priori* information:

Both give almost the same signal!
Bayesian modeling 2/5

- Incident molecules
- Prior information
- Likelihood
- Inversion based on *a posteriori* law

**Observed signal**

- Mass spectra estimation
- Quantity vs Mass

- Time: $t_1$, $t_2$, $t_3$
Bayesian modeling 3/5

- Molecules => amplitude-modulated Dirac comb $m$

- Marked point process => list representation

\[
\begin{align*}
\text{mass} & \quad 100 \quad 200 \quad 200 \quad 50 \\
\text{time} & \quad 10 \quad 35 \quad 68 \quad 97
\end{align*}
\]

\[
\begin{align*}
10 & \mid 100 \\
35 & \mid 200 \\
68 & \mid 200 \\
97 & \mid 50 \\
t_i & \mid m_i
\end{align*}
\]

N unknown
Bayesian modeling 4/5

- Introducing parameters
  - \( \{t_i, m_i\} \) : list
  - \( \sigma^2 \) : noise variance
  - \( P \) : “density” parameter \( (t_i) \)
  - \( m_i \) : Two cases “Free masses” or “Mixture-modeled masses”
  - Mass mixture : Mixture of gaussian/gamma laws, mixture of Dirac distributions ...

- Hierarchical modeling
Bayesian modeling 5/5

- Likelihood
- A priori laws
  - $P$ : beta law
  - $\sigma^2$ : inverse-Gamma law
  - Model : mixture of Dirac each mass $M_k$ follow a gamma law
  - List : Poisson process (parameter $P$) ...
  - ... and Dirac mixture
  - $m$ is marginalized

- Able to write total \textit{a posteriori} law ... and conditional \textit{a posteriori} law for all unknown parameters

\[
N\left(y - \sum_{i=1}^{N} m_i \cdot h(t_i), \sigma^2 \cdot I_d \right) \\
\times \beta(P|a_p, b_p) \\
\times \Gamma^{-1}(\sigma^2|k_\sigma, \beta_\sigma) \\
\times \prod_{k=1}^{K} \Gamma(M_k|k_M, \theta_M) \\
\times N^P \cdot (T - N)^{1-P} \\
\times \prod_{i=1}^{N} \sum_{k=1}^{K} \frac{1}{I} \cdot \delta(m_i - M_k)
\]
Outline

- Introduction
  - Mass spectrometry problem
  - NEMS opportunity

- Direct problem
  - Physics
  - Bayesian modeling

- Inversion
  - Estimation: MCMC method
  - Sampling the “list”

- Results & conclusion
Estimation : MCMC method

- **MCMC**
  - is a way to explore space according to *a posteriori* law
  - offers estimators as expectation/maximum *a posteriori*

- **Gibbs sampler**
  - Every unknown parameter is sampled according to its conditional *a posteriori* law
  - $\sigma^2$ : inverse Gamma law
  - $P$ : beta law
  - Sampling list (and model)
    - $N$ is unknown : no information on $\{t_i\}$ (difference with classical MS)
    - Infinite Impulse Response : Not-independent mass sampling if “once at a time”...
    - Mixture of Dirac model : how to change model to fit data ?
Sampling the “list” 1/3

- Three or four steps
  - 1. Defining neighbor lists
  - 2. Estimating a posteriori law for each neighbor list and choose one
  - **CASE FREE-MASSES** 3. Sample every $m_i$
  - **CASE MIXTURE MODEL** 3. Affecting every events to one class
  - **CASE MIXTURE MODEL** 4. Sample mass $M_k$ for each class

- Neighbor lists
  - We base our work on Single Most Likely Replacement (SMLR) idea: birth and death process on list entries
  - For a given list $L$, a neighbor list $L_n$ is a list where:
    - One event is created
    - One event is removed
    - One event is moved (change $t_i$)
    - Two events are merged
    - One event is created AND its “mass” is taken in each/near event
    - One event is removed AND its “mass” is distributed on each/near event
    - ...
Sampling the “list” 2/3
Sampling the “list” 3/3

- Choosing a list among the neighborhood
  - “Metropolis-Hastings like”
  - Estimate a posteriori probability for each neighbor list
  - Choose a neighbor list according to normalized a posteriori probability
  - Accept this sample with probability \( \frac{p_{AP}(\text{new \_list})}{p_{AP}(\text{current \_list})} \)

- Affecting an event \( t_i \) to a class
  - ML computed between \( t_i \) and \( t_{i+1} \)

- Sample \( m_i \) or mass for each class \( M_k \)
  - Numerical estimation of a posteriori law
  - Inverse repartition method

- Create an event
  - Very long if a large number of sample
  - Idea: find an easy way to calculate “regions of interest” => create new events in this region of interest
Outline

- Introduction
  - Mass spectrometry problem
  - NEMS opportunity

- Direct problem
  - Physics
  - Bayesian modeling

- Inversion
  - Estimation: MCMC method
  - Sampling the “list”

- Results & conclusion
Results and conclusion 1/5
Results and conclusion 2/5

σ = 1000

Free-mass

Estimated list histogram

True list histogram

Noisy signal
Noise-free signal
Estimated signal
Results and conclusion 3/5

\[ \sigma = 1000 \]

Two classes

Estimated list histogram

True list histogram
Results and conclusion 4/5

Mass EAP = 804.6 on 100 samples

σ = 1000

Estimated list histogram

True list histogram

Noisy signal
Noise-free signal
Estimated signal

One class
Results and conclusion 5/5

- New technology ...
- ... so new processings for this field
- Bayesian hierarchical modeling
- List mode problem => marked point process
- MCMC
  - Estimation of instrument parameters and laws hyperparameters
  - Robust estimation
- Sample the list
  - « Metropolis-Hasting like » defining a neighborhood for a list
  - Use class information to robust mass estimation
- Application to other marked point process problems?
Some references

Thank You