Pics, ligne de base, bruit : séparation ternaire de sources assistée (BEADS : positivité, parcimonie), spectres chimiques & miscellanées

L. DUVAL, A. PIRAYRE
IFP Energies nouvelles

X. NING, I. W. SELESNICK
Polytechnic School of Engineering, New York University

23 mars 2018
Old peaks cast long shadows

Chromatography: the traditional 2D way.
Old peaks cast long shadows

Chromatography: individual 1D peaks for single compounds
Old peaks cast long shadows

Chromatography: ternary sources separated
Old peaks cast long shadows

Chromatography: observed signal
Old peaks cast long shadows

Chromatography: wrapping it up
The quick version

- **Issue**: how to accurately & repeatably quantize peaks?
  - avoiding separate baseline and noise removal
- **Question**: where is the string behind the bead?
  - without too accurate models for: peak, noise, baseline

- **Answer**: use main measurement properties + optimization
  - sparsity+symmetry, stationarity, smoothness
- **BEADS**: Baseline Estimation And Denoising w/ Sparsity
  - other properties + optimization for further processing (BARCHAN)
Outline

INTRODUCTION
   FOREWORD
   OUTLINE*
   BACKGROUND

BEADS MODEL AND ALGORITHM
   NOTATIONS
   COMPOUND SPARSE DERIVATIVE MODELING
   MAJORIZE-MINIMIZE TYPE OPTIMIZATION

EVALUATION AND RESULTS
   GC: SIMULATED BASELINE AND GAUSSIAN NOISE
   GC: SIMULATED POISSON NOISE
   GC: REAL DATA
   GC×GC: REAL DATA

ONGOING, EXTENSIONS, CONCLUSION
Background on background

Image processing: varying illumination

- Background affects quantitative evaluation/comparison
- In other domains: (instrumental) bias, (seasonal) trend
- In analytical chemistry: drift, continuum, wander, baseline
- Very rare cases of parametric modeling (piecewise linear, polynomial, spline)
Background on background

Econometrics: trends and seasonality

- Background affects quantitative evaluation/comparison
- In other domains: (instrumental) bias, (seasonal) trend
- In analytical chemistry: drift, continuum, wander, *baseline*
- Very rare cases of parametric modeling (piecewise linear, polynomial, spline)
Background on background

Biomedical: ECG isoelectric line or baseline wander

- Background affects quantitative evaluation/comparison
- In other domains: (instrumental) bias, (seasonal) trend
- In analytical chemistry: drift, continuum, wander, baseline
- Very rare cases of parametric modeling (piecewise linear, polynomial, spline)
Background on background

- Background affects quantitative evaluation/comparison
- In other domains: (instrumental) bias, (seasonal) trend
- In analytical chemistry: drift, continuum, wander, baseline
- Very rare cases of parametric modeling (piecewise linear, polynomial, spline)
Background on background

Analytical chemistry, biological data

- Signal separation into three main morphological components
Notations and assumptions

*Morphological decomposition*: $y = x + f + w$, signals in $\mathbb{R}^N$

- $y$: observation (spectrum, analytical data)
- $x$: clean series of peaks (no baseline, no noise)
- $f$: baseline
- $w$: noise

*Assumption*: without peaks, the baseline can be (approx.) recovered from noise-corrupted data by low-pass filtering

- $\hat{f} = L(y - \hat{x})$: $L$: low-pass filter; $H = I - L$: high-pass filter
- formulated as $\|y - \hat{x} - \hat{f}\|_2^2 = \|H(y - \hat{x})\|_2^2$
- Going further with $D_i$: differentiation operators
Compound sparse derivative modeling

An estimate \( \hat{x} \) can be obtained via:

\[
\hat{x} = \arg \min_x \left\{ F(x) = \frac{1}{2} \| H(y - x) \|_2^2 + \sum_{i=0}^{M} \lambda_i R_i(D_i x) \right\}.
\]
Compound sparse derivative modeling

Examples of (smooth) sparsity promoting functions for $R_i$

- $\phi_i^A = |x|
- \phi_i^B = \sqrt{|x|^2 + \epsilon}
- \phi_i^C = |x| - \epsilon \log (|x| + \epsilon)$
Compound sparse derivative modeling
Take the positivity of chromatogram peaks into account:

$$\hat{x} = \arg \min_x \left\{ F(x) = \frac{1}{2} \| H(y - x) \|_2^2 \right. \right.$$

$$+ \lambda_0 \sum_{n=0}^{N-1} \theta_\varepsilon(x_n; r) + \sum_{i=1}^{M} \lambda_i \sum_{n=0}^{N_i-1} \phi ([D_i x]_n) \left\}. \right.$$

Start from:

$$\theta(x; r) = \begin{cases} x, & x \geq 0 \\ -rx, & x < 0 \end{cases}$$
Compound sparse derivative modeling
Take the positivity of chromatogram peaks into account:

\[
\hat{x} = \arg \min_x \left\{ F(x) = \frac{1}{2} \| H(y - x) \|_2^2 \right. \\
\left. + \lambda_0 \sum_{n=0}^{N-1} \theta_\epsilon(x_n; r) + \sum_{i=1}^M \lambda_i \sum_{n=0}^{N_i-1} \phi(D_i x_n) \right\}.
\]

and majorize it

![The majorizer \( g(x, v) \) for the penalty function \( \theta(x; r) \), \( r = 3 \)](image-url)
Compound sparse derivative modeling

Take the positivity of chromatogram peaks into account:

\[
\hat{x} = \arg \min_x \left\{ F(x) = \frac{1}{2} \| H(y - x) \|_2^2 + \lambda_0 \sum_{n=0}^{N-1} \theta_\epsilon(x_n; r) + \sum_{i=1}^M \sum_{n=0}^{N_i-1} \lambda_i \phi([D_i x]_n) \right\}
\]

then smooth it:

![The smoothed asymmetric penalty function θ_ε (x; r), r = 3](image-url)
Compound sparse derivative modeling

Take the positivity of chromatogram peaks into account:

\[ \hat{x} = \arg \min_x \left\{ F(x) = \frac{1}{2} \| H(y - x) \|_2^2 + \sum_{n=0}^{N-1} \theta \epsilon(x_n; r) + \sum_{i=1}^{M} \sum_{n=0}^{N_i-1} \phi([D_i x]_n) \right\}. \]

then majorize it:

\[ g_0(x, v) = \begin{cases} 
\frac{1+r}{4|v|} x^2 + \frac{1-r}{2} x + |v| \frac{1+r}{4}, & |v| > \epsilon \\
\frac{1+r}{4\epsilon} x^2 + \frac{1-r}{2} x + \epsilon \frac{1+r}{4}, & |v| \leq \epsilon.
\end{cases} \]
Overall principle for Majoration-Minimization

\[ G(x, x_k) \]

\[ G(x, x_{k+1}) \]

MM principles.
BEADS Algorithm (short)

Input: \( y, \ A, \ B, \lambda_i, \ i = 0, \ldots, M \)

1. \( b = B^TBA^{-1}y \)

2. \( x = y \)  \hspace{1cm} (Initialization)

Repeat

3. \( [\Lambda_i]_{n,n} = \frac{\phi'([D_i x]_n)}{[D_i x]_n}, \quad i = 0, \ldots, M, \)

4. \( M = \sum_{i=0}^{M} \lambda_i D_i^T \Lambda_i D_i \)

5. \( Q = B^T B + A^T M A \)

6. \( x = A Q^{-1} b \)

Until converged

8. \( f = y - x - BA^{-1}(y - x) \)

Output: \( x, f \)
Evaluation 1

Simulated chromatograms w/ polynomial+sine baseline
Evaluation 1 with Gaussian noise
Evaluation 2

Simulated chromatograms w/ limited power spectrum noise
Evaluation 2 with Gaussian noise

![Graph showing mean output SNR vs SNR in dB]

<table>
<thead>
<tr>
<th></th>
<th>0 dB</th>
<th>10 dB</th>
<th>20 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std</td>
<td>Mean</td>
</tr>
<tr>
<td>BEADS</td>
<td>18.75</td>
<td>3.71</td>
<td>19.99</td>
</tr>
<tr>
<td>backcor</td>
<td>17.20</td>
<td>4.57</td>
<td>18.93</td>
</tr>
<tr>
<td>airPLS</td>
<td>16.71</td>
<td>4.80</td>
<td>17.52</td>
</tr>
</tbody>
</table>
Evaluation 3 with Poisson noise

Simulated chromatograms w/ Poisson noise
Results: mono-dimensional chromatography (data 1)

Original, superimposed, clean, noise
Results: two-dimensional chromatography (data 2)

Original data
Results: two-dimensional chromatography (data 2)

2D background (estimated)
Results: two-dimensional chromatography (data 2)

Noise (estimated)
Results: two-dimensional chromatography (data 2)
Results: two-dimensional chromatography (data 2)

Original data (again!)
Results: two-dimensional chromatography (data 3)
Results: two-dimensional chromatography (data 3)

2D background (estimated)
Results: two-dimensional chromatography (data 3)

Noise (estimated)
Results: two-dimensional chromatography (data 3)

BEADS corrected data
Results: two-dimensional chromatography (data 3)

Original data (again!)
Results: computing scalability

Linear cost per sample (almost)
Ongoing work

- Tests on analytical chemistry data: NIR, NMR, XPS
- Novel filtering: improved Savitzky-Golay filters
- Novel deconvolution: sparse & positive with norm ratios

SOOT: Non-convex $\ell_0$ count index approximation

- Novel metrics: errors related to peak quantities
- Baseline and noise use: uncertainty, trace products
- 2D chromatography comparisons: BARCHAN warping
- Improved usability: parameter estimation
BARCHAN: 2D chromatography warping

Semi-rigid morphing of two different 2D chromatograms.
BARCHAN: 2D chromatography warping

Ingredients of a GMM plus EM optimization:

- Point sets \(X = \{X_1, \ldots, X_N\}\) and \(Y = \{Y_1, \ldots, Y_M\}\)
- \(p(X_n) = \frac{w}{N} + \sum_{m=1}^{M} \frac{1-w}{2M\pi\sigma^2} \exp \left( -\frac{\|X_n - T(Y_m)\|^2}{2\sigma^2} \right)\)
- \(\min_{\sigma, W, s, t} E = E_1(\sigma, W, s, t) + \frac{\lambda}{2} \text{Tr}(W^\top GW)\)

Calculated deformation of a 2D chromatogram with BARCHAN.
BARCHAN: 2D chromatography warping

Ingredients of a GMM plus EM optimization:

- Point sets $X = \{X_1, \ldots, X_N\}$ and $Y = \{Y_1, \ldots, Y_M\}$
- $p(X_n) = \frac{w}{N} + \sum_{m=1}^{M} \frac{1-w}{2M\pi\sigma^2} \exp\left(-\frac{\|X_n - T(Y_m)\|^2}{2\sigma^2}\right)$
- $\min_{\sigma, W, s, t} E = E_1(\sigma, W, s, t) + \frac{\lambda}{2} \text{Tr}(W^\top GW)$

Calculated deformation of a 2D chromatogram with BARCHAN.
BARCHAN: 2D chromatography warping

Ingredients of a GMM plus EM optimization:

- Point sets $X = \{X_1, \ldots, X_N\}$ and $Y = \{Y_1, \ldots, Y_M\}$
- $p(X_n) = \frac{w}{N} + \sum_{m=1}^{M} \frac{1-w}{2M\pi\sigma^2} \exp\left(-\frac{||X_n-T(Y_m)||^2}{2\sigma^2}\right)$
- $\min_{\sigma, W, s, t} E = E_1(\sigma, W, s, t) + \frac{\lambda}{2} \text{Tr}(W^\top GW)$

Calculated deformation of a 2D chromatogram with BARCHAN.
Improved usability: parameter estimation

- Cut-off frequency estimation
Improved usability: parameter estimation

- Noise, asymmetry ($r$) and regularization ($\lambda$)
Extended applications

- Lidar application
Extended applications

- Engine knocking application
Other known uses

- A fairly generic model (sparsity, positivity/negativity), reused by other authors
  - gas chromatography: mono-dimensional and comprehensive/two-dimensional
  - Raman spectra: biological and biomedical
  - MUSE (Multi Unit Spectroscopic Explorer): astronomical hyperspectral galaxy spectrum
  - X-ray absorption spectroscopy (XAS), X-ray diffraction (XRD), and combined XAS/XRD
  - high-resolution mass spectrometry
  - postprandial Plasma Glucose (PPG), multichannel electroencephalogram (EEG) and single-channel electrocardiogram (ECG)
  - arabic characters
Other known uses

- A fairly generic model (sparsity, positivity/negativity), reused by other authors
  - gas chromatography: mono-dimensional and comprehensive/two-dimensional
  - Raman spectra: biological and biomedical
  - MUSE (Multi Unit Spectroscopic Explorer): astronomical hyperspectral galaxy spectrum
  - X-ray absorption spectroscopy (XAS), X-ray diffraction (XRD), and combined XAS/XRD
  - high-resolution mass spectrometry
  - postprandial Plasma Glucose (PPG), multichannel electroencephalogram (EEG) and single-channel electrocardiogram (ECG)
  - arabic characters
Conclusions

▶ Joint baseline/background and noise estimation
  ▶ Interaction between “separative science” and “source separation”
  ▶ Little “hard” modeling
  ▶ Easy to tune, scalable
  ▶ Codes in Matlab, R and C++

▶ A wide range of applications to unveil

A little more: additional references

Comparison of conventional gas chromatography and comprehensive two-dimensional gas chromatography for the detailed analysis of petrochemical samples.
*J. Chrom. A*, 2004, [http://dx.doi.org/10.1016/j.chroma.2004.05.071](http://dx.doi.org/10.1016/j.chroma.2004.05.071)

Characterisation of middle-distillates by comprehensive two-dimensional gas chromatography (GC × GC): A powerful alternative for performing various standard analysis of middle-distillates.
*J. Chrom. A*, 2005, [http://dx.doi.org/10.1016/j.chroma.2005.05.106](http://dx.doi.org/10.1016/j.chroma.2005.05.106)

C. Vendeuvre, R. Ruiz-Guerrero, F. Bertoncini, L. Duval, and D. Thiébaut.
Comprehensive two-dimensional gas chromatography for detailed characterisation of petroleum products.

X. Ning, I. W. Selesnick, and L. Duval.
Chromatogram baseline estimation and denoising using sparsity (BEADS).

A. Repetti, M. Q. Pham, L. Duval, E. Chouzenoux, and J.-C. Pesquet.
Euclid in a taxicab: Sparse blind deconvolution with smoothed $\ell_1/\ell_2$ regularization.
*IEEE Signal Process. Lett.*, 2015, [http://dx.doi.org/10.1109/LSP.2014.2362861](http://dx.doi.org/10.1109/LSP.2014.2362861)

BARCHAN: Blob Alignment for Robust CHromatographic ANalysis.
*Journal of Chromatography A.*, 2017, [http://dx.doi.org/10.1016/j.chroma.2017.01.003](http://dx.doi.org/10.1016/j.chroma.2017.01.003)

Peaks, baseline and noise separation.
BEADS Algorithm

We now have a majorizer for $F$

\[
G(x, v) = \frac{1}{2} \|H(y - x)\|_2^2 + \lambda_0 x^T [\Gamma(v)] x \\
+ \lambda_0 b^T x + \sum_{i=1}^{M} \left[ \frac{\lambda_i}{2} (D_i x)^T [\Lambda(D_i v)] (D_i x) \right] + c(v).
\]

Minimizing $G(x, v)$ with respect to $x$ yields

\[
x = \left[ H^T H + 2\lambda_0 \Gamma(v) + \sum_{i=1}^{M} \lambda_i D_i^T [\Lambda(D_i v)] D_i \right]^{-1} \left( H^T H y - \lambda_0 b \right)
\]

with notations

\[
c(v) = \sum_n \left[ \phi(v_n) - \frac{v_n}{2} \phi'(v_n) \right].
\]
BEADS Algorithm

We now have a majorizer for $F$

$$G(x, v) = \frac{1}{2} \|H(y - x)\|_2^2 + \lambda_0 x^T [\Gamma(v)] x$$

$$+ \lambda_0 b^T x + \sum_{i=1}^{M} \left[ \frac{\lambda_i}{2} (D_i x)^T [\Lambda(D_i v)] (D_i x) \right] + c(v).$$

Minimizing $G(x, v)$ with respect to $x$ yields

$$x = \left[ H^T H + 2\lambda_0 \Gamma(v) + \sum_{i=1}^{M} \lambda_i D_i^T [\Lambda(D_i v)] D_i \right]^{-1} \left( H^T H y - \lambda_0 b \right).$$

with notations

$$[\Gamma(v)]_{n,n} = \begin{cases} 
\frac{1+r}{4|v_n|}, & |v_n| \geq \epsilon \\
\frac{1+r}{4\epsilon}, & |v_n| \leq \epsilon 
\end{cases}$$
BEADS Algorithm

We now have a majorizer for $F$

$$G(x, v) = \frac{1}{2} \|H(y - x)\|_2^2 + \lambda_0 x^T [\Gamma(v)] x$$

$$+ \lambda_0 b^T x + \sum_{i=1}^{M} \left[ \frac{\lambda_i}{2} (D_i x)^T [\Lambda(D_i v)] (D_i x) \right] + c(v).$$

Minimizing $G(x, v)$ with respect to $x$ yields

$$x = \left[ H^T H + 2 \lambda_0 \Gamma(v) + \sum_{i=1}^{M} \lambda_i D_i^T [\Lambda(D_i v)] D_i \right]^{-1} \left( H^T H y - \lambda_0 b \right).$$

with notations

$$[\Lambda(v)]_{n,n} = \frac{\phi'(v_n)}{v_n}$$
BEADS Algorithm

We now have a majorizer for $F$

$$G(x, v) = \frac{1}{2} \|H(y - x)\|^2_2 + \lambda_0 x^T [\Gamma(v)] x$$

$$+ \lambda_0 b^T x + \sum_{i=1}^{M} \left[ \frac{\lambda_i}{2} (D_i x)^T \Lambda(D_i v) (D_i x) \right] + c(v).$$

Minimizing $G(x, v)$ with respect to $x$ yields

$$x = \left[ H^T H + 2\lambda_0 \Gamma(v) + \sum_{i=1}^{M} \lambda_i D_i^T \Lambda(D_i v) D_i \right]^{-1} \left( H^T H y - \lambda_0 b \right).$$

with notations

$$[b]_n = \frac{1 - r}{2}$$
BEADS Algorithm

Writing filter $H = A^{-1}B \approx BA^{-1}$ (banded matrices) we have

$$x = AQ^{-1} \left( B^T BA^{-1} y - \lambda_0 A^T b \right)$$

where $Q$ is the banded matrix,

$$Q = B^T B + A^T M A,$$

and $M$ is the banded matrix,

$$M = 2\lambda_0 \Gamma(v) + \sum_{i=1}^{M} \lambda_i D_i^T [\Lambda(D_i v)] D_i.$$
BEADS Algorithm

Using previous equations, the MM iteration takes the form:

\[
M^{(k)} = 2\lambda_0 \Gamma(x^{(k)}) + \sum_{i=1}^{M} \lambda_i D_i^T [\Lambda(D_i x^{(k)})] D_i.
\]

\[
Q^{(k)} = B^T B + A^T M^{(k)} A
\]

\[
x^{(k+1)} = A [Q^{(k)}]^{-1} \left( B^T B A^{-1} y - \lambda_0 A^T b \right)
\]