Statistical proximal point methods for image reconstruction

Stéphane Chrétien, Alfred Hero and Robinson Piramuthu

Institute of Statistics and Operations Research

Free University of Brussels

and

Electrical Engineering Computer Science Dpt.

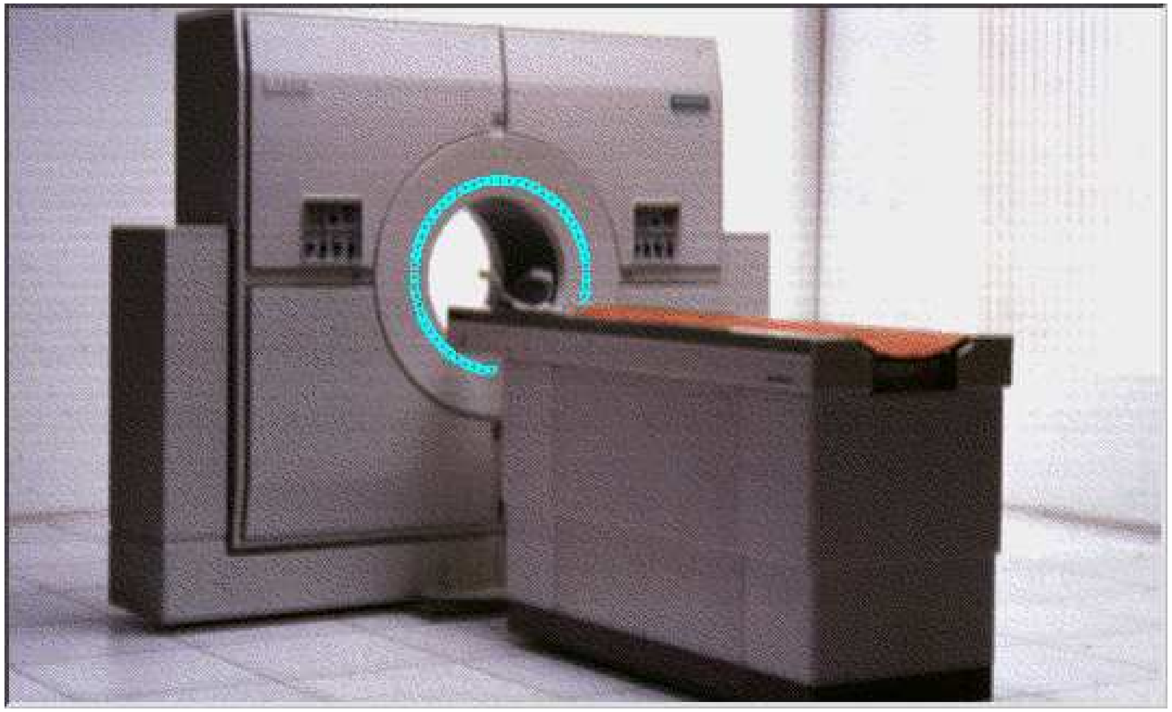
The University of Michigan, Ann Arbor

PLAN OF THE PRESENTATION

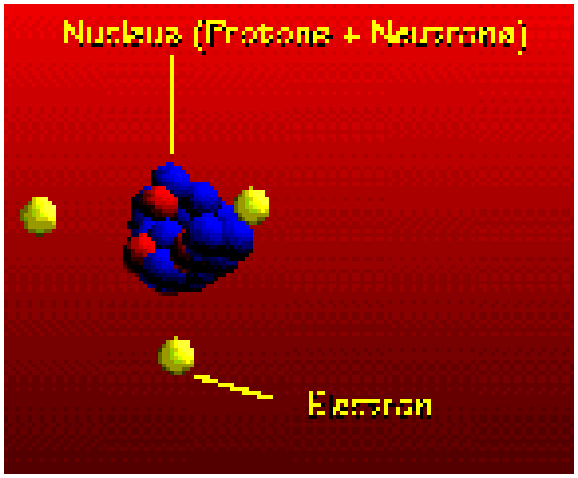
0. Positon Emission Tomography
1. The EM algorithm
2. Acceleration via Kullback proximal generalizations of EM
3. The SAGE method
4. Kullback proximal acceleration of SAGE
5. Numerical results

POSITON EMISSION TOMOGRAPHY (PET)

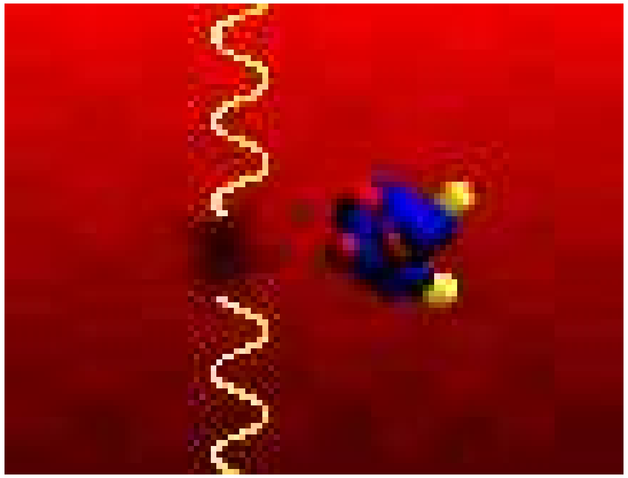
Physical description



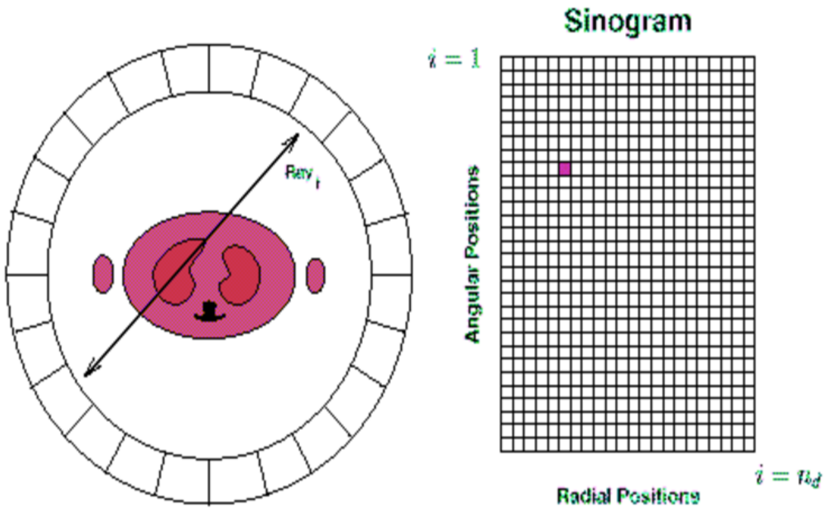
Emission of a gamma ray



$\Rightarrow$

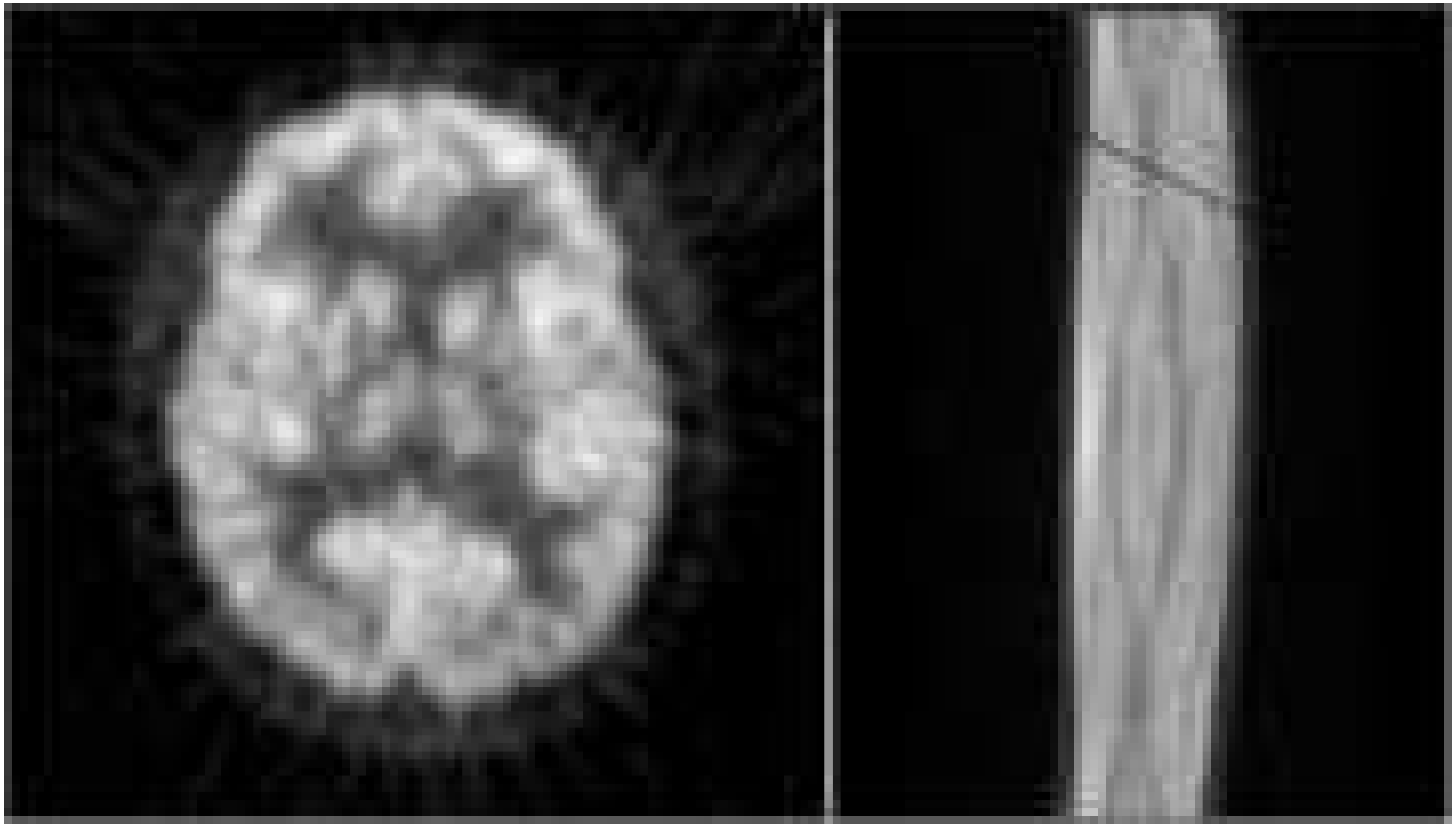


PET Example



$n_d \approx (n_{\text{crystals}})^2$

Example of observed samples



Statistical model

- radionucleide distribution discretized into p pixels with <u>emission rates</u> $\theta = [\theta_1, \ldots, \theta_p]'$
- N detectors. Let $N_{i,j}$ denote the <u>number of emissions</u> from the j^{th} pixel and detected by the i^{th} detector.
- assume

$$N_{i,j} \sim \text{ Poisson } \left\{ a_{i,j} \lambda_j \right\}, \;\; a_{i,j} \geq 0 \tag{1}$$

- observations:

$$Y_i = \sum_{j=1}^{p} N_{i,j} + R_i \tag{2}$$

where $R_i =$"noise"

Reconstruction strategy: maximum likelihood estimation

Y is Poisson$\left\{ \sum_{j=1}^{N} a_{i,j}\theta_j + r_i \right\}$

Log-likelihood:

$$\log g(y;\theta) = \sum_{j=1}^{N} a_{i,j}\theta_j + r_i + y_i \log(\sum_{j=1}^{N} a_{i,j}\theta_j + r_i) \tag{3}$$

Maximum-Likelihood estimator

$$\theta^* \in \text{argmax}_{\theta \geq 0} \log g(y;\theta)$$

$\Rightarrow$ **NEED FOR AN ITERATIVE METHOD !!**

1. THE EM ALGORITHM

The general maximum likelihood estimation problem:

$$\text{Find } \theta_{ML} = \text{argmax}_{\theta \in R^p} l_y(\theta)$$

with

$$l_y(\theta) = \log g(y;\theta)$$

where

- y observed sample of a random variable Y,
- $g(y;\theta)$ density of Y at y,
- θ deterministic parameter.

Framework

We postulate a "more informative" $\mathcal{X}$ in which the log-likelihood can be "easily" maximized

- X random variable on $\mathcal{X}$,
- $f(x;\theta)$ density of X parametrized by θ,
- there exists a many to one mapping

  $$x \mapsto y = h(x)$$

  such that

  $$g(y;\theta) = \int_{h^{-1}(\{y\})} f(x;\theta) d\mu(x),$$

  for an adequate measure $\mu(x)$.

Equivalence between "complete" and "incomplete" problems

If

$$y = h(x)$$

$$\boxed{\begin{aligned} \text{then } \theta_{ML} = \text{argmax}_{\theta \in R^p} \log g(y;\theta) \Leftrightarrow \\ \theta_{ML} = \text{argmax}_{\theta \in R^p} \log f(x;\theta) \end{aligned}}$$

PROBLEM

Find x satisfying $y = h(x)$ is hard !

SOLUTION

Remplace $\log f(x;\theta)$ by its MMSE

$$\mathsf{E}[\log f(x;\theta) \mid y]$$

$\boxed{\text{EM algorithm}}$

Define

$$k(x \mid y; \theta) = \frac{f(x; \theta)}{g(y; \theta)}$$

and

$$E_\theta[\phi(x) \mid y] = \int_{h^{-1}(\{y\})} \phi(x) k(x \mid y; \theta) d\mu(x)$$

for all integrable ϕ.

Thus, the EM algorithm is defined by

$$\boxed{\begin{aligned} &\text{Calculate } Q(\theta \mid \theta^k) = \mathsf{E}_{\theta^k}[\log f(X;\theta)|y], \\ &\theta^{k+1} = \text{argmax}_{\theta \in R^p} Q(\theta \mid \theta^k) \end{aligned}}$$

$\boxed{\text{IMPORTANT PROPERTIES OF EM}}$

- monotonicity of $l_y(\theta^k)$!

$$l_y(\theta^{k+1}) \geq l_y(\theta^k), \quad k = 1, 2, \ldots$$

- convergence towards a stationary point of $l_y(\theta)$
- if local convexity $\Rightarrow$ EM converges linearly towards a local maximizer

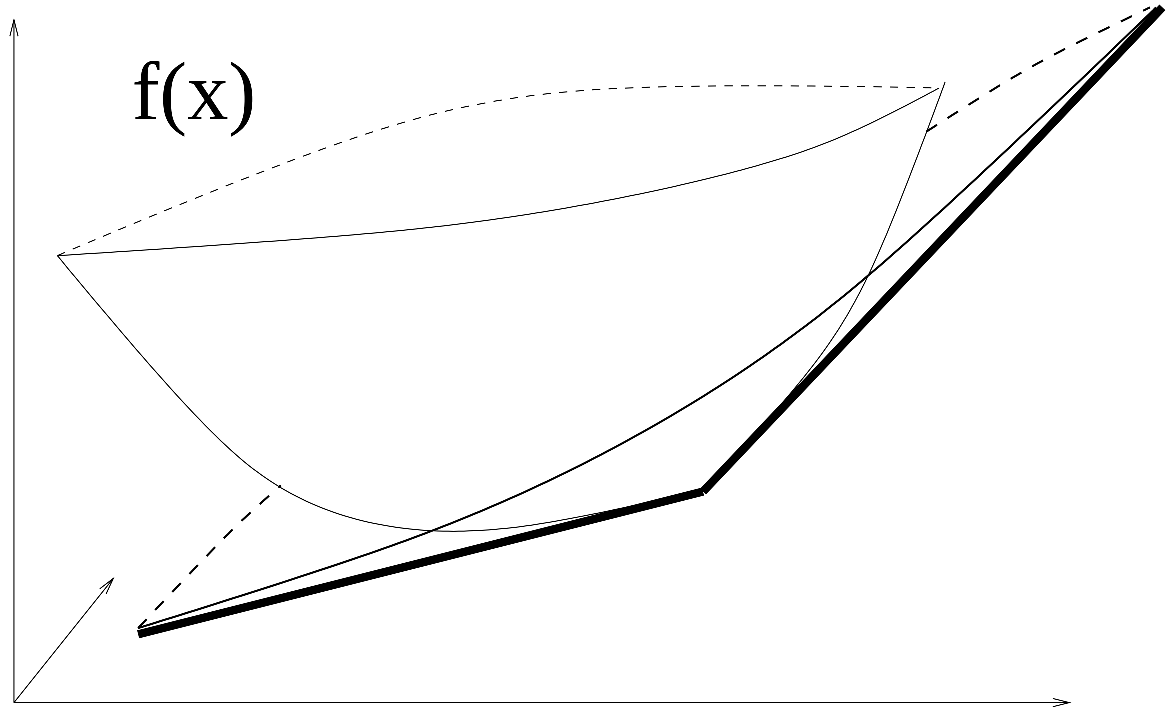
$\boxed{\text{MAIN DRAWBACK}}$

SLOW CONVERGENCE IN LATE ITERATIONS !!

2. THE PROXIMAL APPROACH

Proximal point methods

Originally introduced to adapt **steepest descent** to **non-differentiable** optimization



Proximal point algorithm introduced by Rockafellar in 1976, after the work of Moreau and Minty

$$\min_{\theta} f(\theta)$$

$$\Downarrow$$

$$\theta^{k+1} = \mathrm{argmin}_{\theta} f(\theta) + \frac{1}{2}\|\theta - \theta^k\|^2$$

Why proximal methods ?

$$\theta^{k+1} - \theta^k = -\nabla F(\theta^k)$$

where $F(\theta) = \min_{\theta'} f(\theta') + \frac{1}{2}\|\theta' - \theta^k\|^2$

F is the **Moreau-Yosida regularization** of f

- Equivalent to **Augmented Lagrangien** in the dual space of an optimization problem
- **Regularizing effect**: f convex $\Rightarrow$ F convex and $\mathcal{C}^1$

Generalized proximal methods

Introduced by Teboulle and others around 1990, after the work of Eckstein and Eggermont

$$\min_{\theta} f(\theta) \quad \textbf{subject to } \theta \geq 0 \text{ !!}$$

$$\Downarrow$$

$$\theta^{k+1} = \text{argmin}_{\theta} f(\theta) + \beta_k d_{\phi}(\theta, \theta^k)$$

where

$$d_{\phi}(\theta, \theta^k) = \sum_i \theta_i^k \phi(\theta_i/\theta_i^k) \tag{4}$$

Particularities of generalized proximal methods

- d_ϕ is a ϕ divergence !!
- $\phi(t) \to +\infty$ when $t \to 0 \Rightarrow$ **preserves nonnegativity** !
- useful in inequality constrained optimization

<u>example :</u>

$\phi = t \log t - t + 1$, dom(ϕ)= $[0, +\infty]$

$\Rightarrow d_\phi(\theta', \theta) = \sum_i \theta_i \log \frac{\theta_i}{\theta'_i} + \theta'_i - \theta_i$

Kullback-Leibler !!

PROXIMAL INTERPRETATION OF EM

The maximization step is defined by

$$\theta^{k+1} = \mathrm{argmax}_{\theta\in R^p} \mathsf{E}_{\theta^k}[\log f(X;\theta) \mid y] \tag{5}$$

$$\Rightarrow \theta^{k+1} = \mathrm{argmax}_{\theta\in R^p} \Big\{ \log g(y;\theta) + \mathsf{E}_{\theta^k}[\log \tfrac{f(X;\theta)}{g(y;\theta)} \mid y] \Big\}$$

$$\begin{aligned}\Rightarrow \theta^{k+1} = \mathrm{argmax}_{\theta\in R^p} \{&\log g(y;\theta) + \mathsf{E}_{\theta^k}[\log k(X;\theta) \mid y] \\ &-\mathsf{E}_{\theta^k}[\log k(X;\theta^k) \mid y]\},\end{aligned}$$

$$\Rightarrow \theta^{k+1} = \mathrm{argmax}_{\theta\in R^p} \log g(y;\theta) - \mathsf{E}_{\theta^k}[\log \frac{k(x;\theta^k)}{k(x;\theta)} \mid y],$$

Define

$$I(\theta, \theta^k) = \mathsf{E}_{\theta^k}[\log \frac{k(X \mid y; \theta^k)}{k(X \mid y; \theta)} \mid y],$$

Kullback divergence ! Therefore, EM is equivalent to

$$\boxed{\theta^{k+1} = \operatorname{argmax}_{\theta \in R^p} \log g(y; \theta) - I(\theta, \theta^k) \qquad (6)}$$

$$\Downarrow$$

The EM algorithm is a PROXIMAL algorithm using a KULLBACK divergence with UNIT RELAXATION

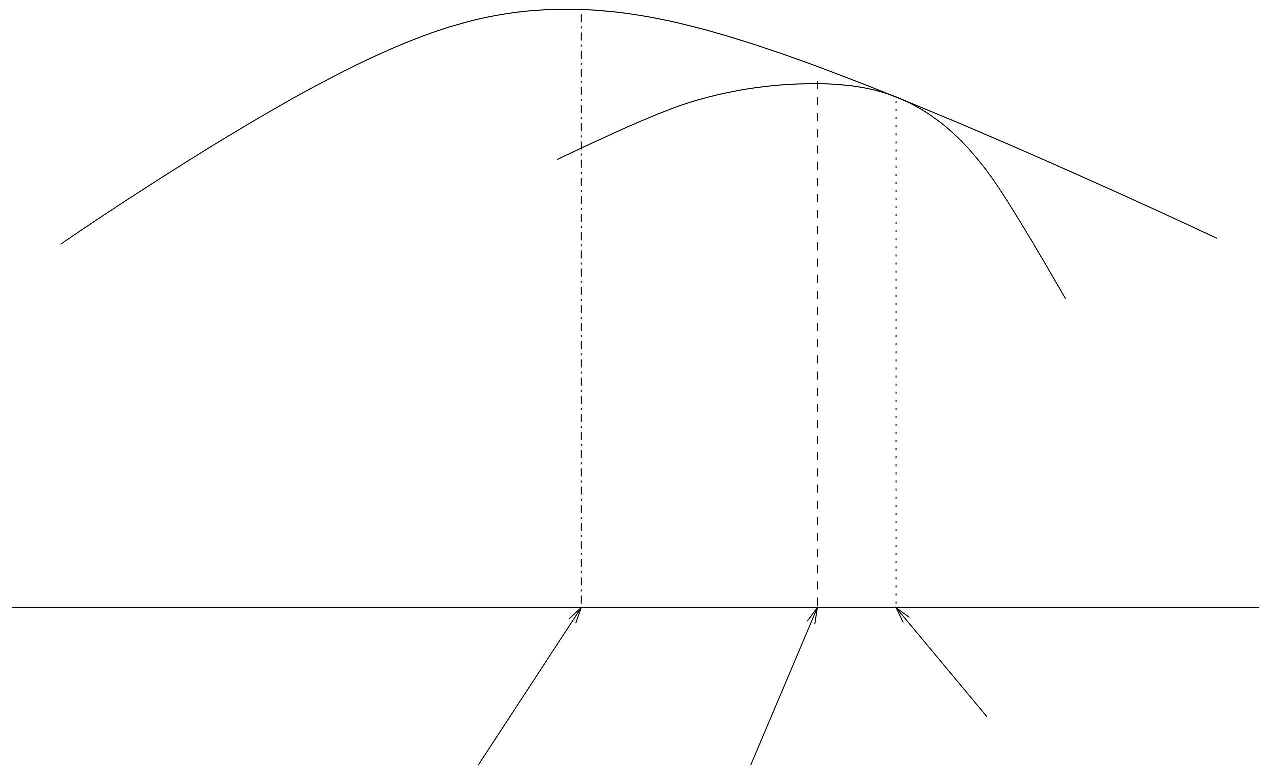


Figure 1: Effect of the regularization

EM and Kullback-Prox for PET

Complete data: $(\{N_{i,j}\}_{j=1}^{p}, \{R_i\})_{i=1}^{N}$

EM algorithm:

$$\theta_i^{k+1} = \frac{\theta_i^k}{\sum_j A_{j,i}} \sum_j \frac{y_j A_{j,i}}{\sum_i A_{j,i}\theta_i^k} \tag{7}$$

Proximal representation:

$$\theta^{k+1} = \mathrm{argmax}_\theta l_y(\theta) - I(\theta, \theta^k) \tag{8}$$

where

$$I(\theta, \theta') = \sum_j \frac{1}{a_j^\mathsf{T}\theta'} \Big\{ \sum_i y_j A_{j,i} \theta_i' \log \frac{\theta_i'}{\theta_i} + y_j a_j^\mathsf{T}\theta' \log\Big(\frac{a_j^\mathsf{T}\theta}{a_j^\mathsf{T}\theta'}\Big)\Big\}. \tag{9}$$

3. MAIN CONVERGENCE RESULTS

THE KULLBACK PROXIMAL POINT ALGORITHM

$$\theta^{k+1} = \mathrm{argmax}_{\theta \in R^p} \log g(y;\theta) - \lambda_k I(\theta, \theta^k),$$

- $\left\{k(x \mid y;\theta)\right\}_{\theta \in R^p}$ regular family
- $\left\{\lambda_k\right\}_{k \in N} \in R_+$

Advantages

- EM is a particular case

$$\lambda_k = 1 \;\; \forall k \in N$$

- One can use $\{\lambda_k\}_{k \in N}$ in order to **accelerate the convergence** !!

$\boxed{\text{THEOREMS}}$

Theorem 1

Assume that I is differentiable at (θ, θ) for all $\theta \in \mathbb{R}_+^p$. Assume that

$$\sum \frac{1}{\beta_k} = +\infty. \tag{10}$$

Then, the Kullback-proximal algorithm converges monotonically towards a maximizer of l_y

One can obtain faster convergence !

Theorem 2

Assume

- **unique** maximizer in **positive orthant**
- λ_k **converges towards zero**

Then, θ^k converges **superlinearly** towards a **global maximizer** of l_y

$$\lim_{k\to} \frac{\|\theta^{k+1} - \theta^*\|}{\|\theta^k - \theta^*\|} = 0 \tag{11}$$

4. THE SAGE METHODOLOGY

MAIN IDEAS

One can get better fit between $Q(\theta, \theta^k)$ and $l_y(\theta)$ along a single coordinate axis than over the whole parameter space

$\Rightarrow$ at iteration k, maximize over one coordinate (e.g. $\theta_{k \mod p}$) with **"adapted" complete** variable called **hidden data** (Hero & Fessler 1994) ...

... while fixing θ_j for all $j \neq (k \mod p)$

- SAGE maximizes a **tighter lower bound** on $l_y(\theta)$
- **monotone convergence** garanteed
- speed of convergence can be **improved** by **choice of hidden data**

SAGE FOR PET

For iteration k, the "adapted" complete variable is $(N_{i,k \mod p}, R_i)_{i=1}^N$

$$Q_k(\theta_j \mid \theta^k) = \sum_{i=1}^{N} \Big(-a_{ij}\lambda_j + \theta_j^k a_{i,j} y_i / (\sum_{j=1}^{p} a_{ij}\theta_j + r_i) \log(a_{i,j}\theta_j) \Big)$$

Maximizing analytically,

$$\begin{aligned} \theta_j^{k+1} &= \theta_j^k \frac{\sum_{i=1}^N a_{i,j} y_j}{(\sum_{j=1}^p a_{ij}\theta_j + r_i)(\sum_{i=1}^N a_{i,j})} && \text{for } j = k \mod p \\ \theta_j^{k+1} &= \theta_j^k && \text{for } j \neq k \mod p \end{aligned} \tag{12}$$

5. THE PROXIMAL SAGE ALGORITHM

Apply the same proximal strategy as for EM ...

$$\theta_j^{k+1} = \operatorname{argmax}_{\substack{\theta_j \geq 0 \\ \theta_i = \theta_i^k \;\; i \neq j}} \log g(y;\theta) - \lambda_k I_k(\theta, \theta^k), \qquad \text{for } j = k \mod p$$

$$\theta_j^{k+1} = \theta_j^k \qquad \text{for } j \neq k \mod p$$

with

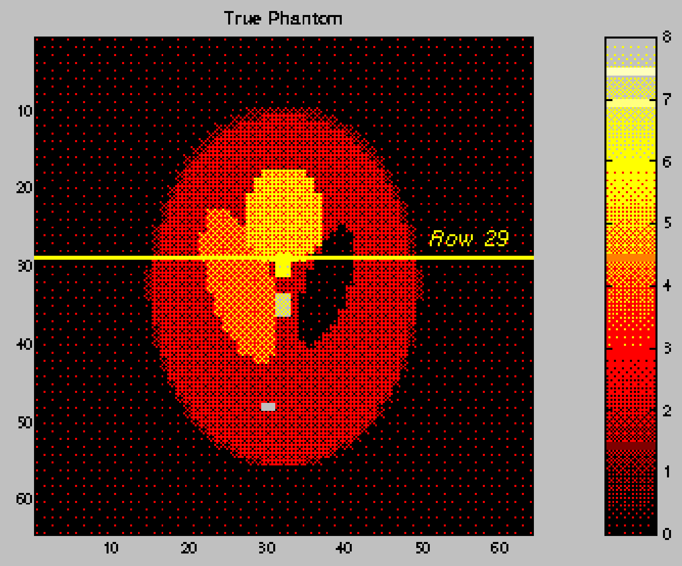
$$I_k(\theta, \theta^k) = \mathsf{E}_{\theta^k}[\log \frac{k(X^k \mid y;\theta^k)}{k(X^k \mid y;\theta)} \mid y],$$

where X^k is the **adapted** complete data (or "hidden data")

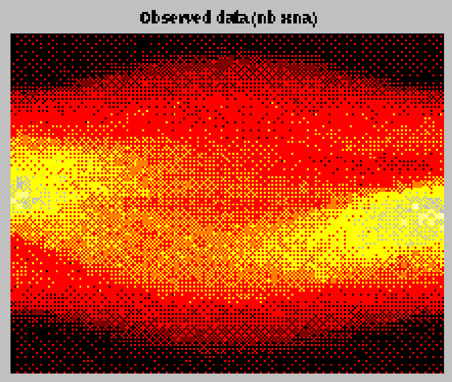
NB: iterations have to be approximated !

6. NUMERICAL EXPERIMENTS

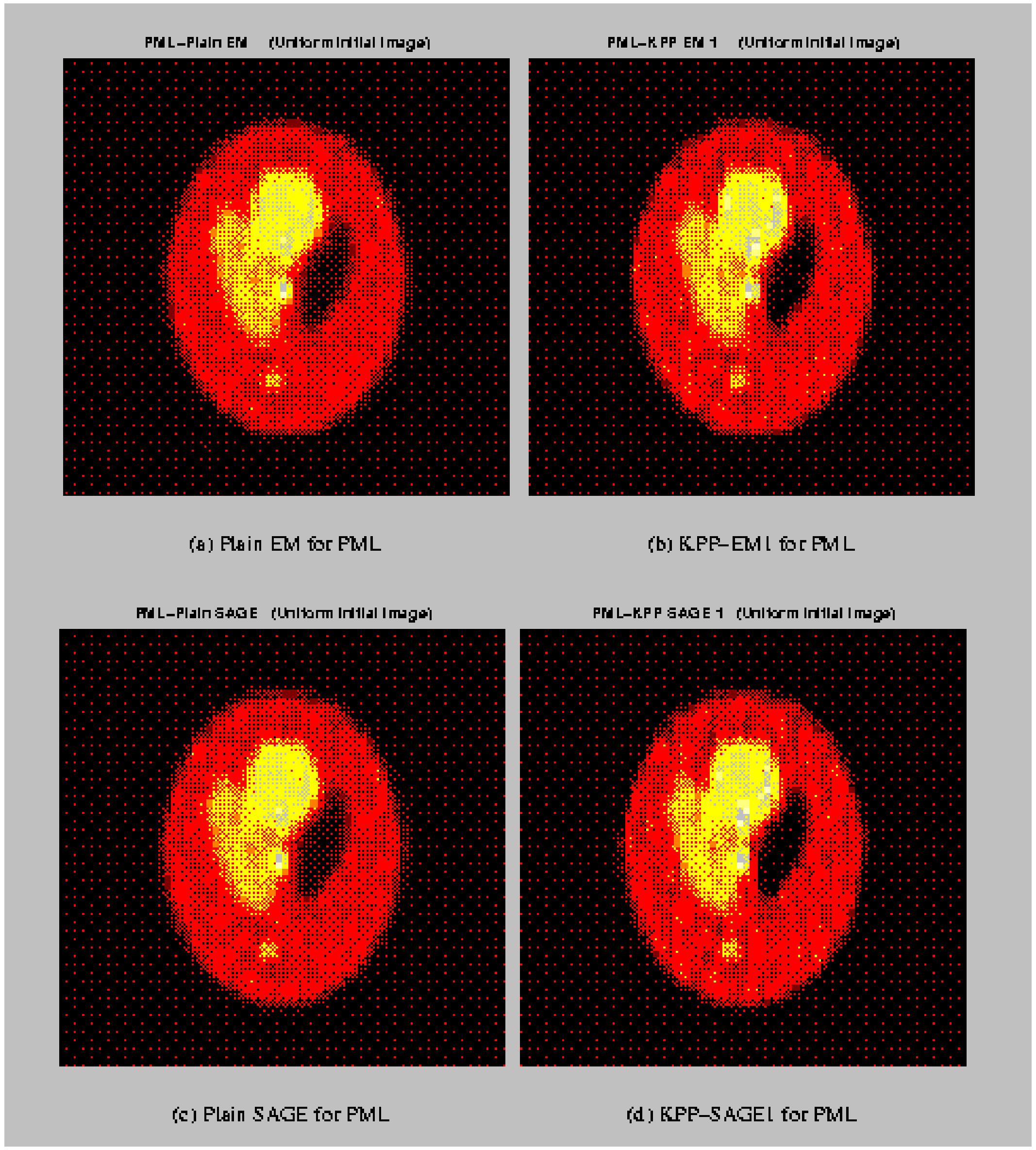
ML–Plain EM: 7.3875	ML–KPP EM: 7.3967
ML–Plain SAGE: 6.0849	ML–KPP SAGE: 7.8280
PML–Plain EM: 7.3966	PML–KPP EM–1: 7.4408 PML–KPP EM–2: 7.4428
PML–Plain SAGE: 6.1043	PML–KPP SAGE–1: 7.8270 PML–KPP SAGE–2: 7.8348

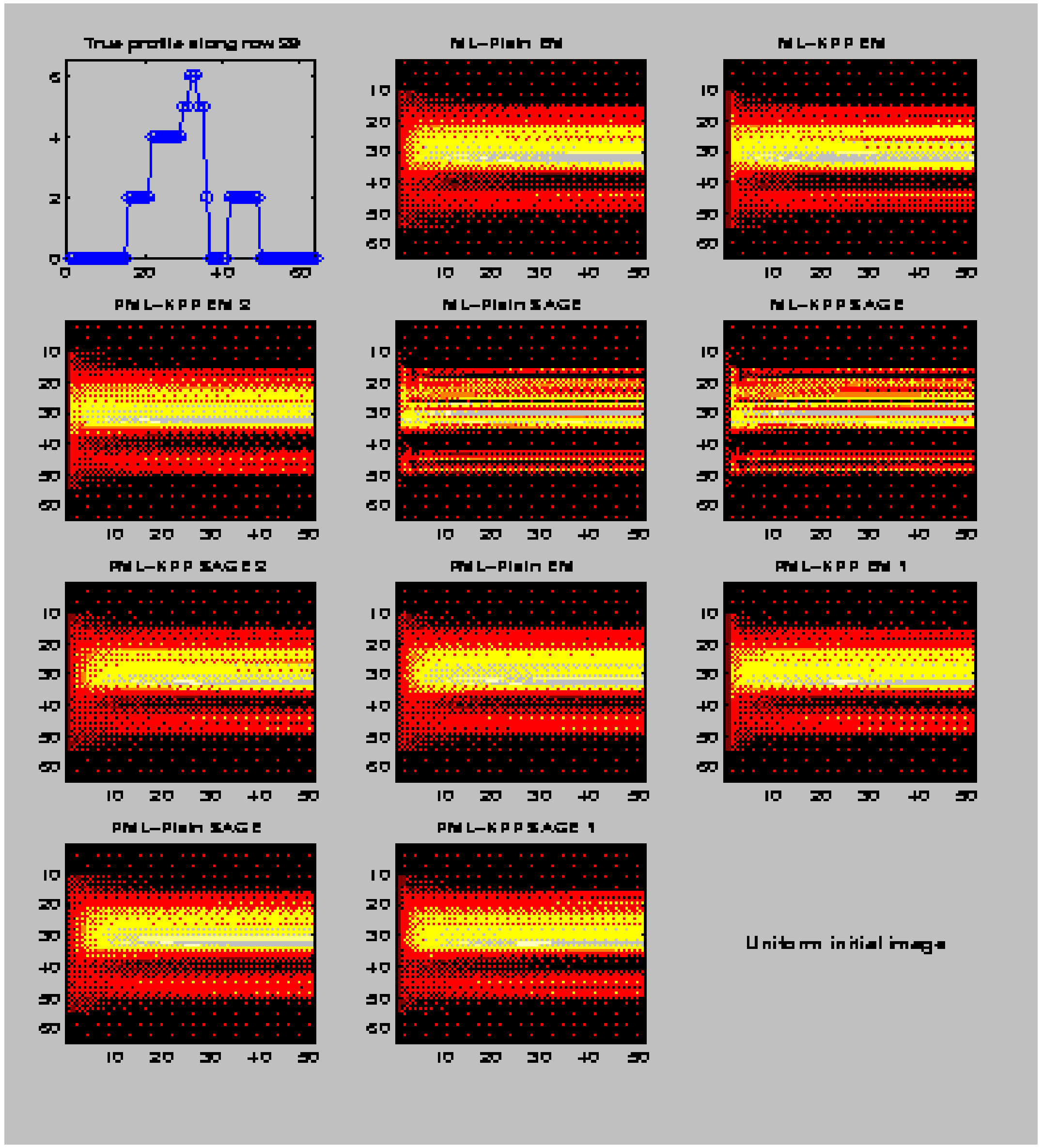


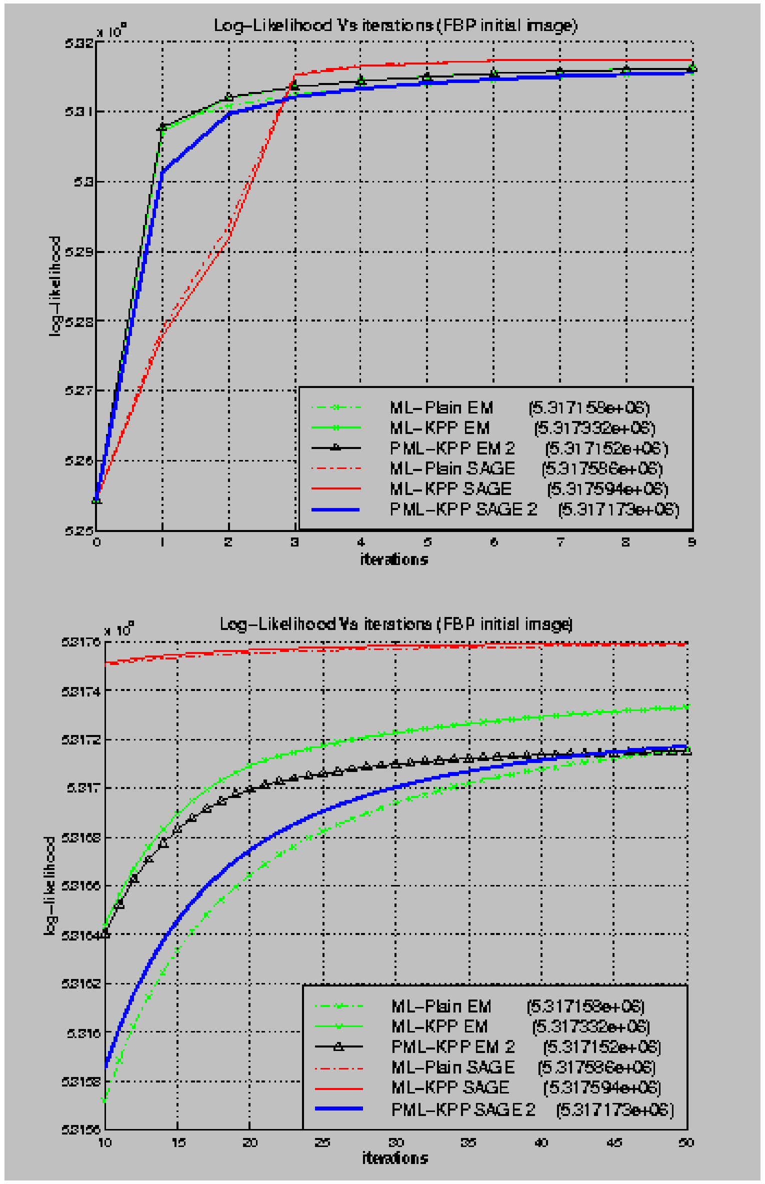
(a) True



(b) Observed


PML−Plain EM (Uniform initial image)
PML−KPP EM 1 (Uniform initial image)
(a) Plain EM for PML
(b) KPP−EM1 for PML
PML−Plain SAGE (Uniform initial image)
PML−KPP SAGE 1 (Uniform initial image)
(c) Plain SAGE for PML
(d) KPP−SAGE1 for PML


True profile along row
ML-Plain EM
ML-KPP EM
PML-KPP EM 2
ML-Plain SAGE
ML-KPP SAGE
PML-KPP SAGE 2
PML-Plain EM
PML-KPP EM 1
PML-Plain SAGE
PML-KPP SAGE 1
Uniform initial image


Log−Likelihood Vs iterations (FBP initial image)
log−likelihood
iterations
ML−Plain EM (5.317158e+06)
ML−KPP EM (5.317332e+06)
PML−KPP EM 2 (5.317152e+06)
ML−Plain SAGE (5.317586e+06)
ML−KPP SAGE (5.317594e+06)
PML−KPP SAGE 2 (5.317173e+06)
Log−Likelihood Vs iterations (FBP initial image)
log−likelihood
iterations
ML−Plain EM (5.317158e+06)
ML−KPP EM (5.317332e+06)
PML−KPP EM 2 (5.317152e+06)
ML−Plain SAGE (5.317586e+06)
ML−KPP SAGE (5.317594e+06)
PML−KPP SAGE 2 (5.317173e+06)