THE EM ALGORITHM AND BEYOND

Alfred O. Hero
Dept. of EECS
Univ. of Michigan
Ann Arbor, MI 48109-2122

Outline

1. **Some difficult ML and PML inverse problems**
   EM algorithm: the basics
   Focus applications
   Convergence properties
   Acceleration of EM: SAGE
   Concluding remarks

1 Some difficult ML and PML inverse problems

1. Gaussian inverse problem:

– Estimating amplitudes of superimposed signals

2. Poisson inverse problem:

– PET image reconstruction

3. Filtered Poisson inverse problem: estimating shot noise parameters

– estimating shot noise parameters

1.1 Gaussian inverse problem

Measure a linear comb. of n samples of p signals $\underline{a}_1, \ldots, \underline{a}_p$

$$\begin{aligned} \underline{Y} &= \underline{a}_1\, \theta_1 + \ldots + \underline{a}_p\, \theta_p + \underline{W}_y \\ &= \mathbf{A}\underline{\theta} + \underline{W}_y \end{aligned}$$

* $\mathbf{A} = [\underline{a}_1, \ldots, \underline{a}_p]$: known signal vectors
* $\underline{W}_y \sim \mathcal{N}_n(0,\ \mathbf{R}_y)$: Gaussian noise
* $\underline{\theta}$: vector of unknown amplitudes

This is linearized version of general superimposed signal model

$$\underline{Y} = \underline{a}_1(\theta_1) + \ldots + \underline{a}_p(\theta_p) + \underline{W}_y$$

Penalized Log-Likelihood function:

$$
\begin{aligned}
l(\underline{\theta}) &= \ln f(\underline{Y};\underline{\theta}) - \underline{\theta}^T \mathbf{P} \underline{\theta} = \\
&= \underbrace{-\tfrac{1}{2}[\underline{Y} - \mathbf{A}\underline{\theta}]^T \mathbf{R}_y^{-1} [\underline{Y} - \mathbf{A}\underline{\theta}] - \tfrac{1}{2}\ln(2\pi|\mathbf{R}_y|)}_{=\ln f} - \eta \underline{\theta}^T \mathbf{P} \underline{\theta}
\end{aligned}
$$

where

* $\mathbf{P} > 0$ is a penalty (smoothing) matrix

* $\eta > 0$ is a regularization (smoothing) parameter

Penalized Maximum Likelihood (PML) estimator

$$
\begin{aligned}
\hat{\underline{\theta}} &= \text{argmax}_{\underline{\theta}}\, l(\underline{\theta}) \\
&= [\mathbf{A}^T \mathbf{R}_y^{-1} \mathbf{A} + \eta \mathbf{P}]^{-1} \mathbf{A}^T \mathbf{R}_y^{-1} \underline{Y}
\end{aligned}
$$

Online computational complexity: $O(np)$

1.2 Poisson inverse problem

Measure a set $\underline{Y}$ of n projections of p pixels $\underline{\theta}$ onto n detectors

$$\underline{Y} = \mathbf{A}\underline{\theta} + \underline{n}_y$$

* $\mathbf{A}$: known matrix of (projection) transition probabilities

* $\underline{n}_y$: Poisson noise with covariance

$$\mathrm{cov}_{\underline{\theta}}\left(\mathrm{diag}[\mathbf{A}\underline{\theta} + \underline{r}]\right)$$

* $\underline{\theta}$: unknown vector (image)

Penalized Log-Likelihood function:

$$l(\underline{\theta}) = \sum_{d=1}^{n} \left[-\mu_d(\underline{\theta}) + y_k \log \mu_k(\underline{\theta})\right] - \eta \underline{\theta}^T \mathbf{P} \underline{\theta}$$

where $\mu_d(\underline{\theta}) = E_{\underline{\theta}}[y_d]$

$$\mu_d(\underline{\theta}) = \sum_{i=1}^{p} a_{di}\theta_i + r_d$$

ML and PML are not of analytic form

1.3 Filtered Poisson process

Measure a waveform $y(t)$ consisting of shot noise and Gaussian noise

$$y(t) = \sum_{i=1}^{N_t} a_i g(t - t_i) + w_y(t), \qquad t \in [0, T]$$

* $g(t)$: known function of time
* $a_i \sim \mathcal{N}(\mu_a, \sigma_a^2)$: i.i.d. Gaussian amplitudes
* $\{N_t : t \in [0, T]\}$: inhomogeneous Poisson process with intensity $\lambda(t; \underline{\theta})$
* $t_1, \ldots, t_n$ arrival times of points
* $\underline{\theta}$: unknown parameters to be estimated

Likelihood function is not closed form:

$$
\begin{aligned}
L(\underline{\theta}) &= \sum_{n=0}^{\infty} e^{-\Lambda - \frac{n\mu_a^2}{2\sigma_a^2}} \\
&= \int_{t_1,\ldots,t_n} \phi(t_1,\ldots,t_n)\, dt_1 \ldots\ dt_n \quad .
\end{aligned}
$$

where

$$
\phi(t_1,\ldots,t_n) = \lambda(t_i,\underline{\theta}) \exp\left[\frac{\mu_a^2}{2\sigma_a^2}[\alpha\underline{X}_f + \underline{1}]^T[I + R_p\alpha]^{-1}[\alpha\underline{X}_f + \underline{1}]\right]
$$

and $\mathbf{R}_p = ((< g(t-t_i), g(t-t_j) >))_{i,j=1}^n$

2 EM algorithm: the basics

Standard iterative algorithms one might consider:

* **Exhaustive search**: impractical in many cases

$$\hat{\underline{\theta}} = \text{argmax}_{\underline{\theta}} l(\underline{\theta})$$

* **Steepest ascent**: no monotonicity, diffble required

$$\hat{\underline{\theta}}^{k+1} = \hat{\underline{\theta}}^{k} + \mu \, \nabla l(\hat{\underline{\theta}}^{k})$$

* **Newton-Raphson**: instability problems, diffble required

$$\underline{\hat{\theta}}^{k+1} = \underline{\hat{\theta}}^{k} - \mu \left[\nabla^2 l(\underline{\hat{\theta}}^{k})\right]^{-1} \nabla l(\underline{\hat{\theta}}^{k})$$

* **Fisher scoring**: matrix inversion, diffble required

$$\underline{\hat{\theta}}^{k+1} = \underline{\hat{\theta}}^{k} + \mu\, F^{-1}(\underline{\hat{\theta}}^{k})\, \nabla l(\underline{\hat{\theta}}^{k})$$

where $\mathbf{F}(\underline{\theta}) = E_{\underline{\theta}}[-\nabla^2 l(\underline{\theta})]$

2.1 Alternative: the EM algorithm

Basic idea:

1. Postulate a more informative data set $\underline{X}$ whose log-likel is easily maximized

$$l_x(\underline{\theta}) = \ln f(\underline{X}; \underline{\theta})$$

2. Implement following heuristic:

* assume have computed $\hat{\underline{\theta}}^k$

Since $l_x(\underline{\theta})$ is unknown, consider maximizing instead its MMSEE

$$Q(\underline{\theta}, \hat{\underline{\theta}}^k) = E_{\hat{\underline{\theta}}^k}[\ln f(\underline{X}; \underline{\theta})|\underline{Y}]$$

Terminology:

- Complete Data: $\underline{X}$
- Incomplete Data: $\underline{Y}$

Complete data admissibility condition:

$$f(\underline{X}|\underline{Y};\underline{\theta}) = f(\underline{X}|\underline{Y})$$

EM ALGORITHM

Initialize: $\underline{\theta}^0$, $k = 0$

Do for $k = 1, \ldots$

[E Step] Estimate

$$Q(\underline{\theta}, \hat{\underline{\theta}}^k) = E_{\hat{\underline{\theta}}^k}[\ln f(\underline{X}; \underline{\theta}) | \underline{Y}]$$

[M Step] Maximize

$$\underline{\theta}^{k+1} = \mathrm{argmax}_{\underline{\theta}}\, Q(\underline{\theta}, \underline{\theta}^k)$$

Important properties of EM algorithm:

- Unlike off-the-shelf packages EM iteration is adapted to the statistical model
- EM algorithm guarantees monotone increase in $l(\underline{\theta}^k)$

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

- EM algorithm will converge to MLE for convex $l(\underline{\theta})$
- Speed of convergence depends on "distance" between $l_x(\underline{\theta})$ and $l_y(\underline{\theta})$
- At each iteration EM maximizes a lower bound on $l(\underline{\theta})$

2.2 Applications

2.2.1 Application 1: Superimposed signals

$$
\begin{aligned}
\underline{Y} &= \mathbf{A}\underline{\theta} + \underline{W}_y \\
&= \mathbf{BC}\underline{\theta} + \underline{W}_y \\
&= \begin{bmatrix} b_{11}, & \cdots & b_{1m} \\ & & \\ \vdots & \vdots & \\ b_{n1} & \cdots & b_{nm} \end{bmatrix} \begin{bmatrix} c_{11}, & c_{1p} \\ \vdots & \vdots \\ \vdots & \vdots \\ c_{m1} & c_{mp} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_p \end{bmatrix} + \begin{bmatrix} w_1 \\ \vdots, \\ w_n \end{bmatrix}
\end{aligned}
$$

Now hypothesize complete data

$$\underline{X} = \mathbf{C}\underline{\theta} + \underline{w}_x$$

where $\underline{w}_x \sim \mathcal{N}(0, \mathbf{R}_x)$

Find $Q(\underline{\theta}, \underline{\theta}^k)$ is quadratic and EM alg. is

$$\underline{\theta}^{k+1} = [I - \mathbf{F}_x^{-1}\mathbf{F}_y]\underline{\theta}^k + \mathbf{F}_x \mathbf{A}^T R_y^{-1} \underline{Y}$$

where $\mathbf{F}_x$ and $\mathbf{F}_y$ are curvature matrices (FIM)

$$\mathbf{F}_x = \mathbf{C}^T \mathbf{R}_x^{-1} \mathbf{C}, \qquad \mathbf{F}_y = \mathbf{A}^T \mathbf{R}_y^{-1} \mathbf{A} + \eta \mathbf{P}$$

Convergence:

As EM is linear it is easy to verify that

$$\Delta\underline{\theta}^{k+1} = \left[I - \mathbf{F}_x^{-1}\mathbf{F}_y\right]\Delta\underline{\theta}^k$$

where

* $\Delta\underline{\theta}^k = \underline{\theta}^k - \underline{\theta}^*$

* $\underline{\theta}^* =$ PML estimator

Remarks

- When $\mathbf{F}_x$ is diagonal EM is equivalent to Jacobi iterations
- $\mathbf{F}_x$ is (-) curvature of $l_x(\underline{\theta})$
- $\mathbf{F}_y = \mathbf{A}^T\mathbf{R}_y^{-1}\mathbf{A}$ is (-) curvature of $l_y(\underline{\theta})$

- By **data processing thm**: $[I - \mathbf{F}_x^{-1}\mathbf{F}_y]$ is non-expansive and EM converges
- Root convergence factor is:

$$\rho = \rho([I - \mathbf{F}_x{-}1\mathbf{F}_y]) = \max_k |\lambda_k^{I-F_x^{-1}F_y}| \in [0,1)$$

Common choice for $\underline{X}$ (Feder&Weinstein):

* $\underline{X}$ direct observation of individual θ_i's

$$\mathbf{C} = \begin{bmatrix} \underline{a}_1 & 0 & \cdots & 0 \\ 0 & \underline{a}_2 & \vdots & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & \ddots & 0 & \underline{a}_p \end{bmatrix} \qquad (np \times p)$$

For admissibility require

* $\mathbf{B} = [\mathbf{I}_p | \cdots | \mathbf{I}_p]^T$

* $R_y = \mathbf{B}\mathbf{R}_x\mathbf{B}^T$ or

$$\mathbf{R}_x = \frac{1}{p} I_p \bigotimes \mathbf{R}_y$$

Obtain EM algorithm (diagonal $\mathbf{P}$):

$$\theta_i^{k+1} = \frac{1}{p\underline{a}_i^T \mathbf{R}_y^{-1} \underline{a}_i + P_{ii}} \underline{\theta}_i^T \mathbf{R}_y^{-1} (\underline{Y} - \mathbf{A}\underline{\theta}_i^k + p\underline{a}_i \theta_i^k)$$

Remarks

- **Online computional complexity:** $O(n)$ per iteration
- This EM algorithm is equivalent to Jacobi iterations
- Complete data FIM: $\mathbf{F}_x = p\mathrm{diag}(\underline{a}_k^T \mathbf{R}_y^{-1} \underline{a}_k)$
- $\rho([I - \mathbf{F}_x{-1}\mathbf{F}_y]) = \rho([1 - 1/p\mathbf{K}]) \Rightarrow$ convergence is VERY SLOW for large p

2.2.2 Poisson inverse problem

$$\underline{Y} \quad = \quad \mathbf{A}\underline{\theta} + \underline{n}_y$$

Common complete data (Shepp&Vardi): directly observe each pixel

$$\underline{X} = \underline{\theta} + \underline{n}_x, \qquad (p \text{ data points})$$

where: $\underline{n}_x$ is Poisson with rate $\mathrm{diag}(\theta_i)$

Find that $Q(\underline{\theta}, \underline{\theta}^k)$ is convex in $\underline{\theta}$

$\Rightarrow \nabla^{10} Q(\underline{\theta}, \underline{\theta}^k) = 0$ gives

$$\underline{\theta}_i^{k+1} = \frac{\underline{\theta}^k}{a_{*i}} \sum_{d=1}^{n} \frac{y_d a_{di}}{\mu_d(\underline{\theta}^k)}$$

2.2.3 Shot noise

$$\begin{aligned} y(t) &= \sum_{i=1}^{N_t} a_i g(t-t_i) + w_y(t), \quad t \in [0,T] \\ &= \int_0^t a(t) g(t-\tau)\, dN_\tau + w_y(t), \quad t \in [0,T] \end{aligned}$$

* $dN\tau$ is Poisson impulse train

Hypothesize complete data: $\underline{X} = dN_t$

Find:

$$Q(\underline{\theta}, \underline{\theta}^k) = \int_0^T d\tilde{N}_t(\underline{\theta}^k) \ln \lambda(t; \underline{\theta}) dt$$

where $d\tilde{N}_t(\underline{\theta}^k)$ is MMSEE of dN_t

$$d\tilde{N}_t(\underline{\theta}^k) := E[dN_t|y;\underline{\theta}^k]$$

Difficulty: finding MMSEE is difficult

A Solution (Antoniadis & Hero, IEEE SP 94): truncated Volterra series approximation

$$\begin{aligned} d\tilde{N}_t(\underline{\theta}^k) &= \underbrace{E[dN_t] + \int_0^t h(t,\tau)y(\tau)d\tau}_{\text{LMSEE}} \\ &\quad + \int_0^t d\tau_1 \int_0^{\tau_1} d\tau_2 h(t,\tau_1,\tau_2)y(\tau_1)y(\tau_2)\ldots \end{aligned}$$

Limiting form of data-linearized E step

$$\tilde{Q}(\underline{\theta}, \underline{\theta}^k) = \frac{\sigma_a^2}{\mu_a^2 + \sigma_a^2} \int_0^T \lambda\left(t; \tau^{(k)}\right)\, dt + \frac{\mu_a^2}{\mu_a^2 + \sigma_a^2} \int_0^T \lambda\left(t; \tau_o\right)\, dt$$

3 EM algorithm: convergence properties

Start from relation between Q, l and conditional entropy H

$$l(\underline{\theta}) = Q(\underline{\theta}, \underline{\theta}^k) + H(\underline{\theta}|\underline{\theta}^k)$$

where

$$H(\underline{\theta}|\underline{\theta}^k) = E[-\ln f(\underline{X} \mid \underline{Y}; \underline{\theta}) \mid \underline{Y}; \underline{\theta}^k]$$

Important observation:

$$H(\underline{\theta}|\underline{\theta}^k) \geq H(\underline{\theta}^k|\underline{\theta}^k) = \text{cond. entropy} \geq 0$$

This implies: $l(\underline{\theta}^{k+1}) \geq l(\underline{\theta}^k)$

Proof: use $Q(\underline{\theta}^{k+1}, \underline{\theta}^k) \geq Q(\underline{\theta}^k, \underline{\theta}^k)$ and

$$l(\underline{\theta}^{k+1}) - Q(\underline{\theta}^{k+1}, \underline{\theta}^k) = H(\underline{\theta}^{k+1}|\underline{\theta}^k) \geq H(\underline{\theta}^k|\underline{\theta}^k) = l(\underline{\theta}^k) - Q(\underline{\theta}^k, \underline{\theta}^k)$$

3.1 Qualitative convergence analysis of EM

Implications

* $Q(\underline{\theta}, \underline{\theta}^k)$ is a lower bound on $l(\underline{\theta})$ for each k
* $l(\underline{\theta}) - Q(\underline{\theta}, \underline{\theta}^k) \geq H(\underline{\theta}^k | \underline{\theta}^k)$
* $l(\underline{\theta}^k) - Q(\underline{\theta}^k, \underline{\theta}^k) = H(\underline{\theta}^k | \underline{\theta}^k)$
* $H(\underline{\theta}^k | \underline{\theta}^k) = 0$ when $\underline{Y}$ "determines" $\underline{X}$.

Qualitative insight:

Asymptotic convergence speed of EM depends on two things:

1. difference between curvature of $Q(\underline{\theta}, \underline{\theta}^k)$ and curvature of $l(\underline{\theta})$

$\Rightarrow \nabla^{20} Q(\underline{\theta}, \underline{\theta}^k) - \nabla^{20} l(\underline{\theta}) \geq 0$

2. The distance $H(\underline{\theta}^k, \underline{\theta}^k)$ between them at the point $\underline{\theta}^k$

$\Rightarrow$ "Osculating parabola" with low curvature is best Q!

3.2 Quantitative convergence analysis of EM

Can show (Hero&Fessler, Statistica Sinica, 95) for general “A algorithm”:

$$\Delta\underline{\theta}^{k+1} = \mathbf{M}(\underline{\theta}^{k+1}, \underline{\theta}^{k})\Delta\underline{\theta}^{k}, \quad \text{qusi-linear representation}$$

where:

$$\mathbf{M}(\underline{\theta}^{k+1}, \underline{\theta}^{k}) = [-\nabla^{20}Q(\underline{\theta}^{k+1}, \underline{\theta}^{k})]^{-1}\ \nabla^{11}Q(\underline{\theta}^{k+1}, \underline{\theta}^{k})$$

Sufficient condition for convergence:

For some norm $\|\underline{u}\|_T := \underline{u}^T \mathbf{T} \underline{u}$

$$\|\mathbf{M}\|_T = \alpha < 1$$

in which case we would have monotone convergence in norm

$$\|\Delta\underline{\theta}^{k+1}\|_T \leq \alpha \, \|\Delta\underline{\theta}^{k}\|_T, \qquad k = 0, 1, \ldots$$

Region of convergence

By investigating conditions on $\underline{\theta}^0$ for which $\|\mathbf{M}\|_T = \alpha < 1$ we obtain a **Region of monotone convergence**

$$\mathcal{R}^+ = \{\underline{\theta}^0 : \|\Delta\underline{\theta}^{k+1}\|_T \leq \alpha \, \|\Delta\underline{\theta}^k\|_T, \; k > 0\}$$

* α controls the **rate of convergence**

3.2.1 Application 1: Superimposed signals

EM algorithm

$$\underline{\theta}^{k+1} = [I - \mathbf{F}_x^{-1}\mathbf{F}_y]\underline{\theta}^k + \mathbf{F}_x \mathbf{A}^T R_y^{-1} \underline{Y}$$

is equivalent to:

$$\Delta\underline{\theta}^{k+1} = \underbrace{[I - \mathbf{F}_x^{-1}\mathbf{F}_y]}_{\mathbf{M}} \Delta\underline{\theta}^k$$

or

$$\mathbf{F}_x^{\frac{1}{2}} \Delta\underline{\theta}^{k+1} = \underbrace{[\mathbf{F}_x^{-\frac{1}{2}}(\mathbf{F}_x - \mathbf{F}_y)\mathbf{F}_x^{-\frac{1}{2}}]}_{\mathbf{M}} \mathbf{F}_x^{\frac{1}{2}} \Delta\underline{\theta}^k$$

and

$$\|\Delta\underline{\theta}^{k+1}\|_{F_x^{\frac{1}{2}}} \leq \rho(I - \mathbf{F}_x^{-1}\mathbf{F}_y) \; \|\Delta\underline{\theta}^{k}\|_{F_x^{\frac{1}{2}}}, \qquad k = 0, 1, \ldots$$

$\Rightarrow$ EM iterates converge monotonically in norm $\| \cdot \|_{F_x^{\frac{1}{2}}}$

As $\mathbf{M}$ does not depend on $\underline{\theta}^k, \underline{\theta}^{k+1}$ the EM converge globally.

3.2.2 Poisson inverse problem

For complete data: $\underline{X} = \underline{\theta} + \underline{n}_x$

Can show:

* $\underline{\theta}^k$ does not converge monotonically to $\underline{\theta}^*$ in any norm

* $\ln \underline{\theta}^k$ converges monotonically to $\ln \underline{\theta}^*$ in the norm

$$\sum_{b=1}^{p} a_{*b}\theta_b^*(\ln \theta_b^{k+1} - \ln \theta_b^*) \leq \alpha \sum_{b=1}^{p} a_{*b}\theta_b^*(\ln \theta_b^{k} - \ln \theta_b^*)$$

$$\alpha = \rho(I - \mathbf{F}_x^{-1}(\underline{t}^*)\mathbf{F}_y(\underline{t}^*))$$

* FIM's $(\mathbf{P} = 0)$ are

$$
\begin{aligned}
\mathbf{F}_y &= \sum_{d=1}^{n} \frac{\underline{a}_d \, \underline{a}_d^T}{\mu_d(\underline{\theta})} \\
\mathbf{F}_x &= \mathrm{diag}_i \left(\frac{1}{\theta_i} \right)
\end{aligned}
$$

* $1 - \frac{1}{p} \ \leq \ \alpha < 1$

4 Acceleration of EM: SAGE

Main Idea: can get better fit between $Q(\underline{\theta}, \underline{\theta}^k)$ and $l(\underline{\theta})$ along a single coordinate axis θ_i

$\Rightarrow$ at iteration k maximize $l(\theta_k)$ while fixing θ_j, $j \neq k$.

Define: subsets of indices

$$S^k = \{kmodp\}, \qquad \overline{S}^k = \{1, \ldots, k-1, k+1, \ldots, p\}$$

For $S = S^k$ we alternate between updating

1. $\underline{\theta}_S$: the working parameters

and

2. $\underline{\theta}_{\overline{S}}$: the rest of the parameters

Define Q^k as the function

$$Q^k(\underline{\theta}_{S^k}, \underline{\theta}^k) = E[\ln f(\underline{X}_{S^k} | \underline{\theta}_{S^k}, \underline{\theta}_{\overline{S}^k}) \mid \underline{Y}, \underline{\theta}_{S^k}, \underline{\theta}_{\overline{S}^k})$$

where $\underline{X}_S$ is a "hidden data set" satisfying

$$f(X_S | \underline{Y}, \underline{\theta}_S, \underline{\theta}_{\overline{S}}) = f(X_S | \underline{Y}, \underline{\theta}_{\overline{S}})$$

This gives the Space Alternating Generalized EM (SAGE) algorithm

SAGE ALGORITHM

Initialize: S^0, $\overline{S}^0$, $k=0$

Do for $k = 1, \ldots$

[E Step] Estimate

$$Q(\underline{\theta}_{S^k}, \underline{\theta}^k) = E[\ln f(\underline{X} | \underline{\theta}_{S^k}, \underline{\theta}_{\overline{S}^k}) | \underline{Y}, \underline{\theta}_{S^k}, \underline{\theta}_{\overline{S}^k}]$$

[M Step] Maximize

$$\underline{\theta}^{k+1} = \text{argmax}_{\underline{\theta}_{S^k}} \; Q(\underline{\theta}_{S^k}, \underline{\theta}^k)$$

-

$$\underline{\theta}_{\overline{S}^{k+1}} = \underline{\theta}_{\overline{S}^k}$$

Important properties of SAGE algorithm:

- SAGE offers potential for much faster convergence than EM
- SAGE algorithm guarantees monotone increase in $l(\underline{\theta}^k)$
- Cyclic SAGE algorithm will converge to PML for convex $l(\underline{\theta})$
- At each iteration SAGE maximizes a tighter lower bound on $l(\underline{\theta})$
- Speed of convergence depends on choice of hidden data

$$\Delta\underline{\theta}^{k+1} = [M^{S^k} \cdots M^{S^1}]\, \Delta\underline{\theta}^k$$

4.0.3 Application 1: Superimposed signals

For comparison: recall EM complete data (np data samples)

$$\underline{X} = \mathbf{C}\underline{\theta} + \mathcal{N}\left(0,\ \frac{1}{p}\,\mathbf{I}_p \bigotimes R_y\right)$$

$$\mathbf{C} = \begin{bmatrix} \underline{a}_1 & 0 & \cdots & 0 \\ 0 & \underline{a}_2 & \vdots & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & \ddots & 0 & \underline{a}_p \end{bmatrix} \qquad (np \times p)$$

with associated FIM: $\mathbf{F}_x = p\mathrm{diag}(\underline{a}_k^T \mathbf{R}_y^{-1} \underline{a}_k)$

For SAGE we choose less informative "hidden data sets" (p data samples)

$$\underline{X}^k = \underline{a}_k \underline{\theta} + \mathcal{N}\left(0,\ \mathbf{R}_y\right), \qquad (k = k\ modp)$$

with associated FIM: $\mathbf{F}_x^{S^k} = \underline{a}_k^T \mathbf{R}_y \underline{a}$

Special case of 2 signals

Define complementary angle:

$$\alpha = \cos(\phi) := \frac{< \underline{a}_1 \underline{a}_2 >_{R_y}}{\|\underline{a}_1\|_{R_y}\ \underline{a}_2\|_{R_y}}$$

Find:

$$\rho_{\mathsf{SAGE}} = \alpha^2, \qquad \rho_{\mathsf{EM}} = 1/2 + \alpha/2$$

4.0.4 Application 2: Poisson inverse problem

For comparison, recall old choice of complete data (p data samples):

$$\underline{X} = \underline{\theta} + \underline{n}_x$$

* $\underline{n}_x$ is Poisson noise with $\mathrm{var}_{\underline{\theta}}(\underline{n}_x) = \mathrm{diag}(\theta_i)$

with associated FIM

$$\mathbf{F}_x = \mathrm{diag}\left(\frac{1}{\theta_i}\right)$$

Here we choose hidden data as (1 data sample)

$$\underline{X}^{S^k} = \theta_k + n_k$$

* n_k is Poisson noise with $\mathrm{var}_{\underline{\theta}}(n_k) = \mu_d = \sum_{i=1}^{p} a_{ki}\theta_i$

with associated FIM

$$\mathbf{F}_x^{S^k} = \mathrm{diag}\left(\frac{1}{\mu_i}\right)$$

5 Concluding remarks

- EM has been largely rendered obsolete by coordinatewise update methods (SAGE, ECM, AECM)
- EM/SAGE algorithms can also be applied to iterative CR bound computations (Hero&Fessler IT94)
- EM/SAGE can be viewed as special case of proximal point iterations with Kullback-Liebler penalty (Chretien&Hero SIAM)
- SAGE has been applied to multi-user communications (Nelson&Poor COM96, Sharfer&Hero SP)
- Hybrid EM-MCMC algorithms are available for accomplishing M-step (Laviel AnnStat)

•