Space-Alternating Generalized Expectation-Maximization Algorithm

Jeffrey A. Fessler, *Member, IEEE,* and Alfred O. Hero, *Member, IEEE*

Abstract— **The expectation-maximization (EM) method can facilitate maximizing likelihood functions that arise in statistical estimation problems. In the classical EM paradigm, one iteratively maximizes the conditional log-likelihood of a single unobservable *complete data space*, rather than maximizing the intractable likelihood function for the measured or *incomplete* data. EM algorithms update all parameters *simultaneously*, which has two drawbacks: 1) slow convergence, and 2) difficult maximization steps due to coupling when smoothness penalties are used.**

This paper describes the space-alternating generalized EM (SAGE) method, which updates the parameters *sequentially* by alternating between several small *hidden-data spaces* defined by the algorithm designer. We prove that the sequence of estimates monotonically increases the penalized-likelihood objective, we derive asymptotic convergence rates, and we provide sufficient conditions for monotone convergence in norm. Two signal processing applications illustrate the method: estimation of superimposed signals in Gaussian noise, and image reconstruction from Poisson measurements. In both applications, our SAGE algorithms easily accommodate smoothness penalties and converge faster than the EM algorithms.

I. Introduction

In a variety of signal processing applications, direct calculations of maximum-likelihood (ML), maximum *a posteriori* (MAP), or maximum penalized-likelihood parameter estimates are intractable due to the complexity of the likelihood functions or to the coupling introduced by smoothness penalties or priors. EM algorithms and generalized EM (GEM) algorithms [1] have proven to be useful for iterative parameter estimation in many such contexts, e.g., [2] and [3]. In the classical formulation of an EM algorithm, one supplements the observed measurements, or *incomplete data*, with a single *complete-data space* whose relationship to the parameter space facilitates estimation. An EM algorithm iteratively alternates between an E-step, calculating the conditional expectation of the complete-data log-likelihood, and an M-step, *simultaneously* maximizing that expectation with respect to all of the unknown parameters. EM algorithms are most useful when the M-step is easier than maximizing the original likelihood. The simultaneous update used by a classical EM algorithm necessitates overly informative complete-data spaces, which in turn lead to slow convergence. In this paper we show improved convergence rates by updating the parameters *sequentially* in small groups.

The convergence rate of an EM algorithm is inversely related to the Fisher information of its complete-data space [1], and we have previously shown that less-informative complete-data spaces lead to improved asymptotic convergence rates [4]–[6]. Less informative complete-data spaces can also lead to larger step sizes and greater likelihood increases in the early iterations [5]-[7]. Since the relationship between complete-data space information and convergence is therefore more than just an asymptotic phenomenon, we believe that one should strive to minimize the information of the complete-data space. However, in the classical EM formulation a less informative complete data space can lead to an intractable maximization step [1], [5], due to the simultaneous update employed by EM algorithms. (As an example, the least-informative admissible "complete" data space would be the measurement space itself!)

To circumvent this tradeoff between convergence rate and complexity, in this paper we extend the concepts in [4] and [6] by proposing a new space-alternating generalized EM (SAGE) method. This method is suited to problems where one can sequentially update small groups of the elements of the parameter vector. Rather than using one large complete-data space, we associate with each group of parameters a *hidden-data space* (Definition 2 in Section II), which would be a complete-data space in the sense of [1] if the other parameters were known. We define a flexible admissibility criterion that ensures that the algorithm monotonically increases the penalized-likelihood objective. In the examples we describe here and in [8], one can design the hidden-data space for each parameter subset to be considerably less informative than the natural single complete-data space. This reduction leads to faster convergence.

Convergence rate is one of two motivations for the SAGE method. In applications such as tomographic imaging and image restoration, where the parameter space is very large, it is often necessary or desirable to regularize using smoothness penalties. Such penalties usually introduce couplings that render intractable the maximization steps of classical EM methods [9]. Several approaches to this problem have been proposed, many motivated by emission tomography, including GEM algorithms [10]-[12], linearizations of the penalty function [9], line searches [13], applying *ad hoc* smoothing in lieu of a smoothness penalty [14], red-black orderings [15], and majorization of the penalty functional [16], [17]. These methods

Manuscript received May 28, 1993, revised February 4, 1994. This work was supported by a DOE Alexander Hollaender Postdoctoral Fellowship, DOE Grant DE-FG02-87ER60561; NSF Grant BCS-9024370; and NCI Grants CA-54362-02 and CA-60711-01. The associate editor coordinating the review of this paper and approving it for publication was Prof. Stanley J. Reeves.

J. Fessler is with the Division of Nuclear Medicine, University of Michigan, Ann Arbor, MI 48109–0552, USA.

A. O. Hero is with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA.

IEEE Log Number 9403729.

are all rooted in the classical EM method, and often they share its slow convergence. In contrast, by using a separate hidden-data space for each parameter, a SAGE algorithm intrinsically decouples the parameter updates. Surprisingly, not only is the maximization simplified, but the convergence rate is improved as well. Two related approaches that also decouple the update are the hybrid ICM-EM algorithm of Abdalla and Kay [18] and the coordinate-wise Newton–Raphson method of Bouman and Sauer [19], [20].

A variety of methods have been proposed for accelerating EM algorithms, most of which are based on standard *numerical* tools such as Aitken's acceleration [21], over-relaxation [22], line-searches [23], Newton methods [24], [19], and conjugate gradients [23], [25]. These methods, although often effective, do not guarantee monotone increases in the objective unless one explicitly computes the objective function. The SAGE method is based fundamentally on *statistical* considerations, and monotonicity is guaranteed. The relative importance of monotonicity and convergence rate will of course be application-dependent.

When the EM algorithm was first introduced, discussants questioned the term "algorithm" since the general method does not prescribe specific computational steps for particular applications [1]. The SAGE method is similarly general, if not more so! Therefore, we devote much of this paper to a detailed comparison of SAGE and classical EM for two signal processing applications: estimation of superimposed signals in Gaussian noise, and image reconstruction from Poisson measurements. We have simplified the examples for the purposes of illustration, while hopefully retaining sufficient complexity that the reader will gain insight into how to apply SAGE to other problems.

The organization of this paper is as follows. Section II defines the generalized concept of "hidden data space," describes the general form of the SAGE algorithm, and establishes monotonicity in objective. Sections III and IV describe the applications. Appendix A discusses convergence of the algorithm and a region of monotone convergence in norm. Appendix B establishes that the region of monotone convergence is nonempty for suitably regular problems. Appendix C examines the relationship between convergence rate and Fisher information of the hidden-data spaces.

II. The SAGE Algorithm

A. Problem

Let the observation Y have the probability density[1] function $f(y;\theta_{\text{true}})$, where θ_{true} is a parameter vector residing in subset Θ of the p-dimensional space $\mathbb{R}^p$. Given a measurement realization $Y = y$, our goal is to compute the maximum penalized-likelihood estimate $\hat{\theta}$ of θ_{true}, defined by

$$\hat{\theta} \triangleq \arg\max_{\theta\in\Theta} \Phi(\theta)$$

where

$$\Phi(\theta) \triangleq \log f(y;\theta) - P(\theta). \tag{1}$$

[1] For simplicity, we restrict our description to continuous random variables. The method is easily extended to general distributions [4].

Unfortunately, direct maximization of Φ is often intractable due to the complexity of f, the coupling in P, or both. Thus, one must resort to iterative methods, and in many problems it is natural to consider subsets of the elements of the parameter vector θ. (Updating in subsets also often leads to remarkably fast convergence, e.g., [26].) The following definition formalizes this idea.

Definition 1: A set S is defined to be an *index set* if it i) is nonempty, ii) is a subset of the set $\{1,\ldots,p\}$, and iii) has no repeated entries. The set $\tilde{S}$ denotes the complement of S intersected with $\{1,\ldots,p\}$.

Let the cardinality of S be m. Then, we use θ_S to denote the m dimensional vector consisting of the m elements of θ indexed by the members of S. Similarly, define $\theta_{\tilde{S}}$ to be the $p-m$ dimensional vector consisting of the remaining elements of θ. For example, if $p = 5$ and $S = \{1,3,4\}$, then $\tilde{S} = \{2,5\}$, $\theta_S = [\theta_1 \quad \theta_3 \quad \theta_4]'$, and $\theta_{\tilde{S}} = [\theta_2 \quad \theta_5]'$, where $'$ denotes matrix transpose. Note that when we use S as a superscript, as in ϕ^S defined below, it serves as a reminder that the function or matrix depends on the choice of S.

One more notational convention will be used hereafter. Functions like $\Phi(\theta)$ expect a p-dimensional vector argument, but it is often convenient to split the argument θ into two vectors: θ_S and $\theta_{\tilde{S}}$, as defined above. Therefore, we define expressions such as the following to be equivalent: $\Phi(\theta_S,\theta_{\tilde{S}}) = \Phi(\theta)$.

In a "grouped coordinate-ascent" method, one sequences through different index sets $S = S^i$ and updates only the elements θ_S of θ while holding the other parameters $\theta_{\tilde{S}}$ fixed [27]. At the ith iteration one would usually like to assign θ_S^{i+1} to the argument that maximizes $\Phi(\theta_S,\theta_{\tilde{S}}^i)$ over θ_S. However, in applications such as the imaging problem described in Section III, there is *no analytical form* for the maximum of $\Phi(\theta_S,\theta_{\tilde{S}})$ over θ_S, *even if the index set S contains only one element.* One could apply numerical line-search methods, but these can be computationally demanding if evaluating $\Phi(\theta_S,\theta_{\tilde{S}}^i) - \Phi(\theta^i)$ for several values of θ_S is expensive.

The basic idea behind the SAGE method is borrowed directly from the EM method. By introducing a "hidden-data" space for θ_S based on the statistical structure of the likelihood, we replace the maximization of $\Phi(\theta_S,\theta_{\tilde{S}}^i)$ over θ_S with the maximization of another functional $\phi^S(\theta_S;\theta^i)$. If the hidden-data space is chosen wisely, then one can maximize the function $\phi^S(\cdot;\theta^i)$ *analytically*, obviating the need for line searches. Even if one cannot maximize ϕ^S analytically, one can often choose hidden-data spaces such that it is easier to evaluate $\phi^S(\cdot;\theta^i) - \phi^S(\theta_S^i;\theta^i)$ than $\Phi(\cdot,\theta_{\tilde{S}}^i) - \Phi(\theta_S^i,\theta_{\tilde{S}}^i)$, so line searches for maximizing $\phi^S(\cdot;\theta^i)$ would be cheaper than line searches for maximizing $\Phi(\cdot;\theta_{\tilde{S}}^i)$. Just as for an EM algorithm, the functionals ϕ^S are constructed to ensure that increases in ϕ^S yield increases in Φ. Furthermore, we have found empirically for tomography that by using a new hidden-data space whose Fisher information is small, the analytical maximum of $\phi^S(\cdot;\theta^i)$ increases $\Phi(\cdot,\theta_{\tilde{S}}^i)$, nearly as much as maximizing $\Phi(\cdot,\theta_{\tilde{S}}^i)$ itself. This is formalized in Appendix C, where we prove that less informative hidden-data spaces lead to faster asymptotic convergence rates. In summary, the SAGE

method uses the underlying statistical structure of the problem to replace cumbersome or expensive numerical maximizations with analytical or simpler maximizations.

B. Hidden-Data Space

To generate the functions ϕ^S for each index set S of interest, we must identify an admissible hidden-data space defined in the following sense:

Definition 2: A random vector X^S with probability density function $f(x;\theta)$ is an *admissible hidden-data space* with respect to θ_S for $f(y;\theta)$ if the joint density of X^S and Y satisfies

$$f(y,x;\theta) = f(y \mid x;\theta_{\bar{S}})f(x;\theta) \tag{2}$$

i.e., the conditional distribution $f(y \mid x;\theta_{\bar{S}})$ must be independent of θ_S. In other words, X^S must be a complete-data space (in the sense of [1]) for θ_S given that $\theta_{\bar{S}}$ is known.

A few remarks may clarify this definition's relationship to related methods.

- The complete-data space for the classical EM algorithm *et al.* [1] is contained as a special case of Definition 2 by choosing $S = \{1,\ldots,p\}$ and requiring Y to be a deterministic function of X^S [4].
- Under the decomposition (2), one can think of Y as the output of a noisy channel that may depend on $\theta_{\bar{S}}$ but not on θ_S, as illustrated in Fig. 1.
- We use the term "hidden" rather than "complete" to describe X^S, since in general X^S will not be complete for θ in the original sense of Dempster *et al.* [1]. Even the aggregate of X^S over all of S will not in general be an admissible complete-data space for θ.
- The most significant generalization over the EM complete-data that is embodied by (2) is that the conditional distribution of Y on X^S *is* allowed to depend on all of the other parameters $\theta_{\bar{S}}$ (Fig. 1). In the superimposed signal application described in Section IV, it is precisely this dependency that leads to improved convergencerates. It also allows significantly more flexibility in the design of the distribution of X^S.
- The cascade EM algorithm [28] is an alternative generalization based on a hierarchy of nested complete-data spaces. In principle, one could further generalize the SAGE method by allowing hierarchies for each X^S.

C. Algorithm

An essential ingredient of any SAGE algorithm is the following conditional expectation of the log-likelihood of X^S:

$$\begin{aligned} Q^S(\theta_S;\overline{\theta}) &= Q^S(\theta_S;\overline{\theta}_S,\overline{\theta}_{\bar{S}}) \\ &\triangleq E\{\log f(X^S;\theta_S,\overline{\theta}_{\bar{S}}) \mid Y = y;\overline{\theta}\} \\ &= \int f(x \mid Y = y;\overline{\theta}) \log f(x;\theta_S,\overline{\theta}_{\bar{S}})dx. \end{aligned} \tag{3}$$

We combine this expectation with the penalty function:

$$\phi^S(\theta_S;\overline{\theta}) \triangleq Q^S(\theta_S;\overline{\theta}) - P(\theta_S,\overline{\theta}_{\bar{S}}). \tag{4}$$

Let $\theta^0 \in \Theta$ be an initial parameter estimate. A generic SAGE algorithm produces a sequence of estimates $\{\theta^i\}_{i=0}^{\infty}$ via the following recursion:

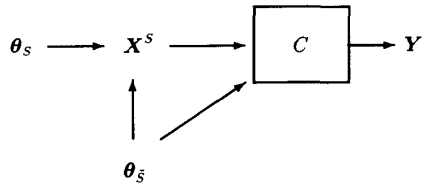

Fig. 1. Representing the observed data Yas the output of a possibly noisy channel C whose input is the hidden-data X^S.

SAGE Algorithm

For $i = 0, 1, \ldots$ {

1) Choose an index set $S = S^i$.
2) Choose an admissible hidden-data space X^{S^i} for θ_{S^i}.
3) E-step: Compute $\phi^{S^i}(\theta_{S^i};\theta^i)$ using (4).
4) M-step:

$$\theta_{S^i}^{i+1} = \arg\max_{\theta_{Si}} \phi^{S^i}(\theta_{S^i};\theta^i), \tag{5}$$

$$\theta_{\bar{S}^i}^{i+1} = \theta_{\bar{S}^i}^{i}. \tag{6}$$

5) Optional:[2] Repeat steps 3 and 4.

},

where the maximization in (5) is over the set

$$\Theta^S(\theta^i) = \{\theta_{S^i} : (\theta_{S^i},\theta_{\bar{S}^i}^i) \in \Theta\}. \tag{7}$$

If one chooses the index sets and hidden data spaces appropriately, then typically one can combine the E-step and M-step via an analytical maximization into a recursion of the form $\theta_{S^i}^{i+1} = g^{S^i}(\theta^i)$. The examples in later sections illustrate this important aspect of the SAGE method.

Note that if for some index set S one chooses $X^S = Y$, then for that S one sees from (3) and (4) that $\phi^S(\theta_S;\theta^i) = \Phi(\theta_S,\theta_{\bar{S}}^i)$. Thus, grouped coordinate-ascent [27] is a special case of the SAGE method, which one can use with index sets S for which $\Phi(\theta_S,\theta_{\bar{S}}^i)$ is easily maximized.

Rather than requiring a strict maximization in (5), one could settle simply for local maxima [4], or for mere increases in ϕ^S, in analogy with GEM algorithms [1]. These generalizations provide the opportunity to further refine the tradeoff between convergence rate and computation per iteration.

D. Choosing Index Sets

To implement a SAGE algorithm, one must choose a sequence of index sets $S^i, i = 0, 1, \ldots$. This choice is as much art as science, and will depend on the structure and relative complexities of the E- and M-steps for the problem. To illustrate the tradeoffs, we focus on imaging problems, for which there are at least four natural choices for the index sets: 1) the entire image, 2) individual pixels, i.e.

$$S^i = \{1 + (i \text{ modulo } p)\} \tag{8}$$

[2] Including the optional subiterations of the E- and M-steps yields a "greedier" algorithm. In the few examples we have tried in image reconstruction, the additional greediness was not beneficial. (This is consistent with the benefits of *under*-relaxation for coordinate-ascent analyzed in [29].) In other applications however, such subiterations may improve the convergence rate, and may be computationally advantageous over line-search methods that require analogous subiterations applied directly to Φ.

(this was used in the ICM-EM algorithm of [18]), 3) grouping by rows or by columns, and 4) "red-black" type orderings. These four choices lead to different tradeoffs between convergence rate and ability to parallelize. A "red-black" grouping was used in a modified EM algorithm in [15] to address the M-step coupling introduced by the smoothness penalties. However, those authors recently concluded [16] that a new simultaneous-update algorithm by De Pierro [17] is preferable. Those methods use the same complete-data space as in the conventional EM algorithm for image reconstruction [3], so the convergence rate is still slow. Since the E-step for image reconstruction naturally decomposes into p separate calculations (one for each element of θ), it is natural to update individual pixels (8). By using the less informative hidden-data spaces described in Section III, we show in [8] and [30] that the SAGE algorithm converges faster than the GEM algorithm of Hebert and Leahy [10], which in turn is faster than the new method of De Pierro [17]. Thus, for image reconstruction, it appears that (8) is best for serial computers.

As noted by the reviewers, for image restoration problems with spatially-invariant systems, one can compute the E-step of the conventional EM algorithm using fast Fourier transforms (FFT's). A SAGE algorithm with single-element index sets (8) would require direct convolutions. Depending on the width and spectrum of the point-spread function, the improved convergence rate of SAGE using (8) may be offset by the use of direct convolution. A compromise would be to group the pixels alternately by rows and by columns. This would allow the use of 1-D FFT's for the E-step, yet could still retain some of the improved convergence rate. Nevertheless, the SAGE method may be most beneficial in applications with spatially-variant responses.

Regardless of how one chooses the index sets, we have constructed ϕ^S to ensure that increases in ϕ^S lead to monotone increases in Φ, as shown next.

E. Monotonicity

Let S and X^S respectively denote an index set and hidden data space used in a SAGE algorithm. Under mild regularity conditions [1], [4], one can apply Bayes' theorem to (3) to see that

$$\begin{aligned} Q^S(\theta_S;\overline{\theta}) &= \int f(x \mid Y = y;\overline{\theta}) \log f(x;\theta_S,\overline{\theta}_{\bar{S}})dx \\ &= L(\theta_S,\overline{\theta}_{\bar{S}}) + H^S(\theta_S;\overline{\theta}) - W^S(\overline{\theta}) \end{aligned} \tag{9}$$

where

$$L(\theta_S,\overline{\theta}_{\bar{S}}) \triangleq \log f(y;\theta_S,\overline{\theta}_{\bar{S}}),$$

$$H^S(\theta_S;\overline{\theta}) \triangleq E\{\log f(X^S \mid Y = y;\theta_S,\overline{\theta}_{\bar{S}}) \mid Y = y;\overline{\theta}\} \tag{10}$$

and due to (2)

$$W^S(\overline{\theta}) \triangleq \int f(x \mid Y = y;\overline{\theta}) \log f(y \mid x;\overline{\theta}_{\bar{S}})dx.$$

Note that W^S is independent of θ_S, so it does not affect the maximization (5). Using these definitions and Jensen's inequality [1], one can easily show that

$$H^S(\theta_S;\overline{\theta}) \le H^S(\overline{\theta}_S;\overline{\theta}), \quad \forall\theta_S, \quad \forall\overline{\theta} \tag{11}$$

from which the following theorem follows directly.

Theorem 1: Let θ^i denote the sequence of estimates generated by a SAGE algorithm (5). Then 1) $\Phi(\theta^i)$ is monotonically nondecreasing, 2) if $\hat{\theta}$ maximizes Φ, then $\hat{\theta}$ is a fixed point of the SAGE algorithm, and 3)

$$\Phi(\theta^{i+1}) - \Phi(\theta^i) \ge \phi^S(\theta_S^{i+1};\theta^i) - \phi^S(\theta_S^i;\theta^i).$$

Proof: From (4) and (9) it follows that

$$\begin{aligned} &\Phi(\theta_S,\overline{\theta}_{\bar{S}}) - \Phi(\overline{\theta}) \\ &\quad = \phi^S(\theta_S;\overline{\theta}) - H^S(\theta_S;\overline{\theta}) - (\phi^S(\overline{\theta}_S;\overline{\theta}) - H^S(\overline{\theta}_S;\overline{\theta})). \end{aligned}$$

Thus, if $\phi^S(\theta_S;\overline{\theta}) \ge \phi^S(\overline{\theta}_S;\overline{\theta})$, then $\Phi(\theta_S,\overline{\theta}_{\bar{S}}) \ge \Phi(\overline{\theta})$ using (11). The results then follow from the definition of the SAGE algorithm. □

Standard numerical methods require evaluation of $\Phi(\theta^{i+1}) - \Phi(\theta^i)$ to ensure monotonicity. That requirement is obviated for SAGE methods by the monotonicity theorem above.

F. Convergence

For a well-behaved objective Φ, the monotonicity property ensures that the sequence $\{\theta^i\}$ will not diverge, but it does not guarantee convergence even to a local maximum of Φ. (Some EM algorithms have fixed points that are not local maxima [1], [31].) Therefore, in the appendices we provide additional theorems that give sufficient conditions for convergence in norm, and that characterize the asymptotic convergence rate. To summarize briefly, these theorems show under suitable regularity conditions that:

- If a SAGE algorithm is initialized in a region suitably close to a local maximum in the interior of Θ, then the sequence of estimates will converge monotonically in *norm* to it. (This may not apply when the local maximum lies on the boundary of Θ, as often happens in the example in Section III.)
- For suitably regular objectives, the region of monotone convergence in norm is guaranteed to be nonempty [43].
- The asymptotic convergence rate of a SAGE algorithm will be improved if one chooses a less informative hidden-data space.

This last point is subtle, but is perhaps one of the most important conclusions of our analyses since it emphasizes the need for careful design of the hidden-data spaces. Less informative hidden-data spaces yield faster convergence, but more informative hidden-data spaces may yield easier M-steps [5], [8], [30].

III. Example 1
Linear Poisson Measurements

The EM method has been used for over a decade to compute ML estimates of radionuclide distributions from tomographic data, such as that measured by positron emission tomography (PET) [3], [32]. In this section we present a brief review of the

classical EM algorithm for this problem, and then introduce two SAGE algorithms. The second SAGE algorithm is based on a new hidden-data space, and converges faster than even an accelerated EM algorithm. For simplicity we focus in this paper on ML estimation; the penalized version is described in [8] and [30].

Assume that a radionuclide distribution can be discretized into p pixels with emission rates $\lambda = [\lambda_1, \ldots, \lambda_p]'$. Assume that the emission source is viewed by N detectors, and let N_{nk} denote the number of emissions from the kth pixel that are detected by the nth detector. Assume the variates N_{nk} have independent Poisson distributions:

$$N_{nk} \sim \text{Poisson}\{a_{nk}\lambda_k\} \tag{12}$$

where the a_{nk} are nonnegative constants that characterize the system [3]. The detectors record emissions from several source locations, so at best one would observe only the sums $\sum_{k=1}^{p} N_{nk}$, rather than each N_{nk}. Background emissions, random coincidences, and scatter contaminate the measurements, so we observe

$$Y_n = \sum_{k=1}^{p} N_{nk} + R_n$$

where $\{R_n\}$ are independent Poisson variates

$$R_n \sim \text{Poisson}\{r_n\} \tag{13}$$

with means $\{r_n\}$ assumed known for simplicity. Thus

$$Y_n \sim \text{Poisson}\left\{\sum_{k=1}^{p} a_{nk}\lambda_k + r_n\right\}. \tag{14}$$

Given realizations $\{y_n\}$ of $\{Y_n\}$, the log-likelihood for this problem is given by [3]:

$$\log f(y;\lambda) = \sum_{n=1}^{N} \left(-\overline{y}_n(\lambda) + y_n \log \overline{y}_n(\lambda)\right)$$

where

$$\overline{y}_n(\lambda) = \sum_{k=1}^{p} a_{nk}\lambda_k + r_n.$$

We would like to compute the ML estimate $\hat{\lambda}$ from y.

To apply coordinate ascent directly to this likelihood, one might try to update λ_k by equating the derivative of the likelihood to zero

$$0 = -a_{\cdot k} + \sum_{n=1}^{N} a_{nk} \frac{y_n}{a_{nk}(\lambda_k - \lambda_k^i) + \overline{y}_n(\lambda^i)} \tag{15}$$

where $a_{\cdot k} = \sum_{n=1}^{N} a_{nk}$. Unfortunately, this equation has no analytical solution. A line-search method would require multiple evaluations of (15), which would be expensive—hence the popularity of EM-type algorithms [3] that require no line searches.

The complete-data space for the classical EM algorithm [3] for this problem is the set of unobservable random variates

$$X^1 = \{\{N_{nk}\}_{k=1}^{p}, \{R_n\}\}_{n=1}^{N}. \tag{16}$$

For this complete-data space, the Q function (3) becomes (see (4) of [3])

$$Q^1(\lambda;\lambda^i) = \sum_{n=1}^{N}\sum_{k=1}^{p} \left(-a_{nk}\lambda_k + \overline{N}_{nk}\log(a_{nk}\lambda_k)\right)$$

where [3]

$$\overline{N}_{nk} = E\{N_{nk} \mid Y = y; \lambda^i\} = \lambda_k^i a_{nk} y_n / \overline{y}_n(\lambda^i).$$

Maximizing $Q^1(\cdot;\lambda^i)$ analytically leads to the algorithm which follows.

ML-EM Algorithm for Poisson Data

for $i = 0, 1, \ldots$ {

$$\overline{y}_n := \sum_{k=1}^{p} a_{nk}\lambda_k^i + r_n, \quad n = 1, \ldots, N$$

for $k = 1, \ldots, p$ {

$$e_k = \sum_{n=1}^{N} a_{nk} y_n / \overline{y}_n$$
$$\lambda_k^{i+1} = \lambda_k^i e_k / a_{\cdot k} \tag{17}$$

}

}.

In words, the previous parameter estimate is used to compute predicted measurements, those predictions are divided into the measurements and backprojected to form multiplicative correction factors, and the estimates are *simultaneously* updated using those correction factors. This EM algorithm converges globally [3], [5], but slowly. The root-convergence factor is very close to 1 (even if $p = 1$ [5]), since the complete-data space is considerably more informative than the measurements [5], [8], [30].

We now derive two SAGE algorithms for this problem, both of which use individual pixels for the index sets: $S^i = \{k\}$, where $k = 1 + (i \text{ modulo } p)$. The most obvious hidden-data for λ_k is just

$$X^{S^k} = \{N_{nk}, R_n\}_{n=1}^{N}$$

which is a subset of the classical complete-data space (16). The Q^{S^k} function for the kth parameter is:

$$Q^{S^k}(\lambda_k;\lambda^i) = \sum_{n=1}^{N} \left(-a_{nk}\lambda_k + \overline{N}_{nk}\log(a_{nk}\lambda_k)\right).$$

Maximizing $Q^{S^k}(\cdot;\lambda^i)$ analytically yields the following algorithm:

ML-SAGE-1 Algorithm for Poisson Data

Initialize: $\overline{y}_n = \sum_{k=1}^{p} a_{nk}\lambda_k^0 + r_n, \quad n = 1, \ldots, N.$

for $i = 0, 1, \ldots$ {

for $k = 1, \ldots, p$ {

$$\begin{aligned} e_k &= \sum_{n=1}^{N} a_{nk} y_n / \overline{y}_n \\ \lambda_k^{i+1} &= \lambda_k^i e_k / a_{\cdot k} \\ \lambda_j^{i+1} &= \lambda_j^i, j \neq k \\ \overline{y}_n &:= \overline{y}_n + (\lambda_k^{i+1} - \lambda_k^i) a_{nk}, \quad \forall n{:}a_{nk} \neq 0 \end{aligned} \tag{18}$$

}

}.

This SAGE algorithm updates the parameters *sequentially*, and immediately updates the predicted measurements $\overline{y}_n$ within the inner loop, whereas the ML-EM algorithm waits until all parameters have been updated. ML-SAGE-1 is the unregularized special case of the ICM-EM algorithm of [18]; a local convergence result for ICM-EM was mentioned in [18].

We found that ML-SAGE-1 converges somewhat faster than ML-EM for well-conditioned problems, but the difference is minimal for poorly-conditioned problems. The reason is that X^{S^k} is still overly informative since the background events are isolated from the parameter being updated (*cf.* (12) and (13)) [8], [30]. Therefore, we now introduce a new, less informative hidden-data space that associates some of the uncertainty of the background events R_n with the particular parameter λ_k as it is updated [8], [30]. Whereas the ordinary complete-data space has some intuitive relationship with the underlying image formation physics, this new hidden-data space was developed from a statistical perspective on the problem and its Fisher information. First, define

$$z_k = \min_{n:a_{nk} \neq 0} \{r_n / a_{nk}\}$$

and define unobservable independent Poisson variates

$$\begin{aligned} Z_{nk} &\sim \text{Poisson}\{a_{nk}(\lambda_k + z_k)\} \\ B_{nk} &\sim \text{Poisson}\left\{ r_n - a_{nk} z_k + \sum_{j \neq k} a_{nj} \lambda_j \right\} \end{aligned} \tag{19}$$

and let the hidden-data space for λ_k *only* be

$$X^{S^k} = \{Z_{nk}, B_{nk}\}_{n=1}^{N}.$$

Then, clearly

$$Y_n = Z_{nk} + B_{nk}$$

has the appropriate distribution (14) for any particular k. We have absorbed all of the background events into the terms Z_{nk} and B_{nk} which are associated with λ_k. Thus, the aggregate of all p of the hidden-data spaces is *not* an admissible hidden-data space for the entire parameter vector λ. Using a similar derivation as in [3] (see [8] and [30] for details), one can show

$$Q^{S^k}(\lambda_k; \lambda^i) = \sum_{n=1}^{N} (-a_{nk}(\lambda_k + z_k) + \overline{Z}_{nk} \log(a_{nk}(\lambda_k + z_k))$$

where

$$\overline{Z}_{nk} = E\{Z_{nk} \mid Y = y; \lambda^i\} = (\lambda_k^i + z_k) a_{nk} y_n / \overline{y}_n(\lambda^i).$$

Maximizing $Q^{S^k}(\cdot; \lambda^i)$ analytically (subject to the nonnegativity constraint) yields the ML-SAGE-2 algorithm, which has the same sequential structure as ML-SAGE-1, except that (18) is replaced by:

$$\lambda_k^{i+1} := \max\left\{ (\lambda_k^i + z_k) e_k / a_{\cdot k} - z_k, 0 \right\}.$$

Provided $z_k \neq 0$, which is always the case in PET since random coincidences are pervasive, this remarkably small modification yields significant improvements in convergence rate.

The Fisher information for the classical complete-data space with respect to λ is diagonal with entries

$$a_{\cdot k} / \hat{\lambda}_k$$

provided the ML estimate $\hat{\lambda}$ is positive. In contrast, the Fisher information for the new hidden-data space is diagonal with entries

$$a_{\cdot k} / (\hat{\lambda}_k + z_k)$$

which is clearly smaller since $z_k > 0$. The improved convergence rate of ML-SAGE-2 is closely related to this difference.

To illustrate, Fig. 2 displays the likelihood $\Phi(\theta^i)$ versus iteration for the ML-EM algorithm and for ML-SAGE-2 applied to a simulation of PET data. The image was an 80 × 110 discretization of a central slice of the digital 3-D Hoffman brain phantom (2 mm pixel size). The sinogram size was 70

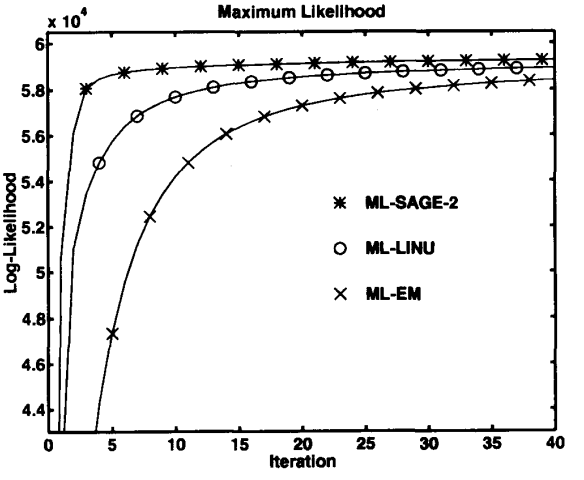

Fig. 2. Comparison of log-likelihood increase $\log f(y; \theta^i) - \log f(y; \theta^0)$ versus iteration i for ML-EM, ML-LINU, and ML-SAGE-2 algorithms, for image reconstruction from PET measurements with 9% random coincidences. ML-SAGE-2 clearly reaches the asymptote sooner.

radial bins (3 mm wide) by 100 angles. A 900 000-count noisy projection was generated using (6-mm-wide) strip-integrals for $\{a_{nk}\}$ [29], including the effects of nonuniform head attenuation and nonuniform detector efficiency. A uniform field of random coincidences was added, reflecting a scan with 9% of the total counts due to randoms (i.e., $\sum_{n=1}^{N} r_n \approx 0.1 \sum_{n=1}^{N} \overline{y}_n(\lambda)$), a typical fraction for a PET study. Further details can be found in [8] and [30], including comparisons over a large range of r_n's. Also shown in Fig. 2 is the LINU unbounded line-search acceleration algorithm described by Kaufman [23]. The ML-SAGE-2 likelihood clearly increases faster and reaches its asymptote sooner than both the ML-EM and ML-LINU algorithms.[3] (ML-SAGE-2 was also considerably easier to implement than the bent-line LINU method.)

The convergence in norm given by Theorem 3 of Appendix A is inapplicable to this Poisson example when the ML estimate has components that are zero, i.e., when the ML estimate lies on the boundary of the nonnegative orthant [33]. See [30] for a global convergence proof for ML-SAGE-1 and ML-SAGE-2 similar to the proofs in [3] and [17].

The reader may wonder whether one can also find a better complete-data space for the classical EM algorithm. Because the EM update is simultaneous, one must distribute the background events among *all* pixels; therefore, the terms z_k are reduced by a factor of roughly $\sqrt{p}$ [8], [30]. Since $\sqrt{p}$ is in the hundreds, the change in convergence rate is insignificant, which is consistent with the small reduction in Fisher information [8], [30]. Other simultaneous updates [17] similarly do not improve much [30]. Apparently one benefits most from this less informative hidden-data space by using a SAGE method with the parameters grouped into many small index sets.

An alternative to SAGE is the coordinate-wise sequential Newton–Raphson updates recently proposed by Bouman and Sauer [19]. That method is not guaranteed to be monotonic, but when it converges it might do so somewhat faster than SAGE since it is even greedier. One can obtain similar (but monotonic) greediness by using multiple subiterations of the E- and M-steps in the SAGE algorithm, as indicated by Step 5 of the generic SAGE algorithm. However, for the few cases we have tested, we have not observed any improvement in convergence rates using multiple subiterations. Although further investigation of the tradeoffs available is needed, including comparisons with possibly superlinear methods such as preconditioned conjugate gradient [23], [34], it appears that the statistical perspective inherent to the SAGE method is a useful addition to conventional numerical tools.

IV. Example 2
Linear Gaussian Measurements

The Poisson problem has important practical applications, but the nonlinearity of the algorithms complicates a formal analysis of the convergence rates. In this section, we analyze the problem of estimating superimposed linear signals in Gaussian noise [2], [9]

$$Y = a_1\theta_1 + \cdots + a_p\theta_p + \epsilon = \mathbf{A}\theta + \epsilon \tag{20}$$

where $\mathbf{A} = [a_1 \ \ldots \ a_p]$, and ϵ is additive zero-mean Gaussian noise with covariance Π, i.e., $\epsilon \sim \mathcal{N}(\mathbf{0}, \Pi)$. For simplicity we consider a quadratic penalty $P(\theta) = \frac{1}{2}\theta'\mathbf{P}\theta$, so the penalized-likelihood objective function is

$$-\Phi(\theta) = \frac{1}{2}(y - \mathbf{A}\theta)'\Pi^{-1}(y - \mathbf{A}\theta) + \frac{1}{2}\theta'\mathbf{P}\theta.$$

Such objective functions arise in many inverse problems [9]. We assume $\mathbf{A}$ has full column rank, $\mathbf{P}$ is symmetric nonnegative definite, and the intersection of the null spaces of $\mathbf{P}$ and $\mathbf{A}$ is empty, in which case the (unique) penalized-likelihood estimate is

$$\hat{\theta} = (\mathbf{A}'\Pi^{-1}\mathbf{A} + \mathbf{P})^{-1}\mathbf{A}'\Pi^{-1}y. \tag{21}$$

If $\mathbf{A}$ is large, or if positivity constraints on θ are appropriate, then (21) is impractical and iterative methods may be useful. (One can also think of (20) as a linearization of the more interesting nonlinear problem [2].) We present the linear version here since we can derive exact expressions for the convergence rates. We first present admissible hidden-data spaces for this problem, derive EM and SAGE algorithms, and then prove that the SAGE algorithm converges faster.

Since the mean of Y is linear in θ, the conventional complete-data [2], [9] for the EM algorithm for this problem is also linear in θ. Here, we restrict our attention to hidden-data spaces X^S whose means are also linear in θ, and for which the conditional mean of Y given X^S is linear in X^S and $\theta_{\bar{S}}$. Considering a general index set S, the natural hidden-data space for θ_S is

$$X^S \sim \mathcal{N}(\mathbf{B}\theta_S + \tilde{\mathbf{B}}\theta_{\bar{S}}, \mathbf{C})$$
$$Y \mid X^S = x \sim \mathcal{N}(\mathbf{G}x + \tilde{\mathbf{G}}\theta_{\bar{S}}, \mathbf{W})$$

which is admissible provided the two normal distributions are independent and consistent with (20), i.e., $\mathbf{A}_S = \mathbf{GB}$, $\mathbf{A}_{\bar{S}} = \mathbf{G}\tilde{\mathbf{B}} + \tilde{\mathbf{G}}$, and $\Pi = \mathbf{W} + \mathbf{GCG}'$. The log-likelihood for X^S is given by

$$\begin{aligned}\log f(X^S; \theta_S, \theta^i_{\bar{S}}) &= c_1 \\ &\quad - \frac{1}{2}(X^S - \mathbf{B}\theta_S - \tilde{\mathbf{B}}\theta^i_{\bar{S}})'\mathbf{C}^{-1}(X^S - \mathbf{B}\theta_S - \tilde{\mathbf{B}}\theta^i_{\bar{S}}) \\ &= c_2 + (\mathbf{B}\theta_S + \tilde{\mathbf{B}}\theta^i_{\bar{S}})'\mathbf{C}^{-1}\left(X^S - \frac{1}{2}(\mathbf{B}\theta_S + \tilde{\mathbf{B}}\theta^i_{\bar{S}})\right)\end{aligned}$$

where c_1 and c_2 are independent of θ_S. By standard properties of joint normal distributions

$$\begin{aligned}\overline{X}^S &= E\{X^S \mid Y = y; \theta^i\} \\ &= \mathbf{B}\theta^i_S + \tilde{\mathbf{B}}\theta^i_{\bar{S}} + \mathbf{CG}'\Pi^{-1}(y - \mathbf{A}\theta^i).\end{aligned}$$

[3] Fast convergence is clearly desirable for regularized objective functions, but we advise caution when using "stopping rules" in conjunction with coordinate-based algorithms for the unregularized case, since for such algorithms the *high* spatial frequencies converge faster than the low frequencies [26].

The ϕ^S function of (4) is thus

$$\phi^S(\theta_S;\theta^i) = (\mathbf{B}\theta_S + \tilde{\mathbf{B}}\theta^i_{\tilde{S}})'\mathbf{C}^{-1}\left(\overline{X}^S - \frac{1}{2}(\mathbf{B}\theta_S + \tilde{\mathbf{B}}\theta^i_{\tilde{S}})\right), - \frac{1}{2}\begin{bmatrix}\theta_S\\ \theta^i_{\tilde{S}}\end{bmatrix}'\begin{bmatrix}\mathbf{P}_1\mathbf{P}_2\\ \mathbf{P}'_2\mathbf{P}_3\end{bmatrix}\begin{bmatrix}\theta_S\\ \theta^i_{\tilde{S}}\end{bmatrix} + c_2$$

which, maximized over θ_S, yields the generic combined E- and M-step

$$\begin{aligned}\theta^{i+1}_S &= (\mathbf{F}_{X^S} + \mathbf{P}_1)^{-1}[\mathbf{B}'\mathbf{C}^{-1}(\overline{X}^S - \tilde{\mathbf{B}}\theta^i_{\tilde{S}}) - \mathbf{P}_2\theta^i_{\tilde{S}}]\\ &= \theta^i_S + (\mathbf{F}_{X^S} + \mathbf{P}_1)^{-1}\mathbf{A}'_S\Pi^{-1}[y - \mathbf{A}\theta^i]\\ &\quad - (\mathbf{F}_{X^S} + \mathbf{P}_1)^{-1}[\mathbf{P}_1\mathbf{P}_2]\theta^i \qquad (22)\end{aligned}$$

where $\mathbf{F}_{X^S} = \mathbf{B}'\mathbf{C}^{-1}\mathbf{B}$ is the Fisher information of X^S for θ_S.

A. EM Algorithm

The ordinary EM algorithm [2], [9] is based on the following choices for the complete-data space: $S = \{1,\dots,p\}$, $\mathbf{B} = \text{diag}\{a_k\}$, $\mathbf{C} = \frac{1}{p}\mathbf{I}_p \otimes \Pi$, $\mathbf{G} = (\mathbf{1}'_p \otimes \mathbf{I}_p)$, and $\tilde{\mathbf{B}} = \tilde{\mathbf{G}} = \mathbf{W} = \mathbf{0}$, where diag$\{\cdot\}$ denotes a diagonal or block-diagonal matrix appropriately formed, $\mathbf{1}_p$ denotes the p vector of ones, and $\otimes$ is the Kronecker matrix product. Note that these choices distribute a fraction $\frac{1}{p}$ of the noise covariance Π to each signal vector a_k. Thus, $\mathbf{F}_X = p\,\text{diag}\{a'_k\Pi^{-1}a_k\}$, which being a diagonal matrix is easily inverted. However, since $S = \{1,\dots,p\}$, the penalized EM algorithm (22) requires inversion of $\mathbf{F}_X + \mathbf{P}$, which could be just as difficult as inverting $\mathbf{A}'\Pi^{-1}\mathbf{A} + \mathbf{P}$ for a general $\mathbf{P}$. Therefore, we consider the case where $\mathbf{P} = \text{diag}\{P_{kk}\}$, for which the EM algorithm *simultaneously* updates all parameters via

$$\begin{aligned}\theta^{i+1}_k &= \frac{1}{p}(a'_k\Pi^{-1}a_k + P_{kk}/p)^{-1}a'_k\Pi^{-1}(y - \mathbf{A}\theta^i)\\ &\quad + (a'_k\Pi^{-1}a_k + P_{kk}/p)^{-1}a'_k\Pi^{-1}a_k\theta^i_k \qquad (23)\end{aligned}$$

for $k = 1,\dots,p$.

B. SAGE Algorithm

Because of the additive form of (20), it is natural for the SAGE algorithm to update each parameter θ_k individually, i.e., $S^i = \{k\}$ where $k = 1 + (i \text{ modulo } p)$. In light of the discussion in Appendix C, we would like the Fisher information of the hidden-data space for θ_k to be small, so we associate *all* of the noise covariance with the signal vector a_k

$$\begin{aligned}X^{S^k} &\sim \mathcal{N}(a_k\theta_k, \Pi)\\ Y &= X^{S^k} + \sum_{j\neq k} a_j\theta_j.\end{aligned}$$

Thus, $\mathbf{F}_{X^{S^k}} = a'_k\Pi^{-1}a_k$, which is p times less informative than the EM case, which associates only a fraction $1/p$ of the noise covariance with each signal. (This provides a statistical interpretation of the modified EM algorithms in [35] and [36].)

The above choice for the hidden-data space corresponds to $\mathbf{B} = a_k$, $\mathbf{C} = \Pi$, $\tilde{\mathbf{B}} = \mathbf{W} = \mathbf{0}$, $\mathbf{G} = \mathbf{I}$, and $\tilde{\mathbf{G}} = [a_1 \dots a_{k-1}a_{k+1} \dots a_p]$, which, substituted into (22), yields the following algorithm.

SAGE Algorithm for Superimposed Signals

```
Initialize:        ε̂ = y − Aθ^0
for  i = 0, 1, ...
        k = 1 + (i modulo p),   S = {k},
        θ_k^{i+1} := θ_k^i − (a_k'Π^{-1}a_k + P_kk)^{-1} P_k θ^i
                   + (a_k'Π^{-1}a_k + P_kk)^{-1} a_k'Π ε̂
        ε̂ := ε̂ + (θ_k^{i+1} − θ_k^i) a_k,
        θ_j^{i+1} := θ_j^i,   j = 1, ..., k−1, k+1, ..., p,
}
```

where P_{kk} is the kth diagonal entry of $\mathbf{P}$, and $\mathbf{P}_k$ is the kth row of $\mathbf{P}$. Note that unlike the EM algorithm, the SAGE algorithm circumvents the need to invert $\mathbf{P}$ by performing a *sequential* update, so a nondiagonal smoothness penalty $\mathbf{P}$ is entirely feasible.

C. Convergence

To establish convergence of the EM and SAGE algorithms, we use Definition 3 and Theorem 3 of Appendix A. A few definitions are needed. Let $\mathbf{H} = \mathbf{A}'\Pi^{-1}\mathbf{A} + \mathbf{P}$ be the Hessian for this problem, and decompose it by

$$\mathbf{H} = \mathbf{L}_H + \mathbf{D}_H + \mathbf{L}'_H \qquad (24)$$

where $\mathbf{D}_H$ is a diagonal matrix with the diagonal entries of $\mathbf{H}$, and $\mathbf{L}_H$ is a strictly lower triangular matrix. Similarly, let

$$\begin{aligned}\mathbf{F} = \mathbf{A}'\Pi^{-1}\mathbf{A} &= \mathbf{L}_F + \mathbf{D}_F + \mathbf{L}'_F,\\ \mathbf{D}_H &= \mathbf{D}_F + \mathbf{D}_P\end{aligned}$$

where $\mathbf{D}_P = \text{diag}\{P_{kk}\}$ and $\mathbf{F}$ is the Fisher information for Y with respect to θ.

Let $\|x\|$ denote the standard Euclidian norm of a vector x, and for a nonsingular matrix $\mathbf{T}$ define $\|x\|_{\mathbf{T}} = \|\mathbf{T}x\|$, which induces the matrix norm

$$\|\mathbf{A}\|_{\mathbf{T}} = \max_x \frac{\|\mathbf{A}x\|_{\mathbf{T}}}{\|x\|_{\mathbf{T}}} = \|\mathbf{T}\mathbf{A}\mathbf{T}^{-1}\|.$$

In addition, let $\rho(\mathbf{A})$ denote matrix spectral radius, the maximum magnitude eigenvalue of $\mathbf{A}$.

SAGE Algorithm: From the SAGE algorithm given above, one can show (cf. proof of Theorem 3) that

$$\theta^{(i+1)p} - \hat{\theta} = \mathbf{M}_p \cdot \cdots \cdot \mathbf{M}_1 \cdot (\theta^{ip} - \hat{\theta}) \qquad (25)$$

where

$$\begin{aligned}\mathbf{M}_k &= \mathbf{I} - \mathbf{e_k}H^{-1}_{kk}\mathbf{e_k}'\mathbf{H},\\ &= \mathbf{H}^{-1/2}\left(\mathbf{I} - \mathbf{H}^{1/2}\mathbf{e_k}(H_{kk})^{-1}\mathbf{e_k}'\mathbf{H}^{1/2}\right)\mathbf{H}^{1/2},\\ &= \mathbf{T}^{-1}\left(\mathbf{I} - \mathbf{t}_k(\mathbf{t}'_k\mathbf{t}_k)^{-1}\mathbf{t}'_k\right)\mathbf{T},\\ &= \mathbf{T}^{-1}\mathcal{P}^{\perp}_k\mathbf{T}\end{aligned}$$

and where $\mathbf{T} = \mathbf{H}^{1/2}$, the kth column of $\mathbf{T}$ is $\mathbf{t}_k$, $\mathcal{P}^{\perp}_k$ is the orthogonal projection onto $\mathbf{t}_k$, and $\mathbf{e}_k$ is the kth unit vector of length p. Since an orthogonal projection is nonexpansive, $\|\mathbf{M}_k\|_{\mathbf{T}} \leq 1$, which confirms condition 2 of Definition 3. To confirm condition 3, rewrite the SAGE algorithm using (24) as

$$\theta^{(i+1)p} - \hat{\theta} = [\mathbf{I} - (\mathbf{D}_H + \mathbf{L}_H)^{-1}\mathbf{H}](\theta^{ip} - \hat{\theta})$$

which is the Gauss–Siedel iteration (see p. 72 of [37]). Condition 3 follows from p. 109 of [37] since

$$\|\mathbf{I} - (\mathbf{D}_H + \mathbf{L}_H)^{-1}\mathbf{H}\|_{\mathbf{T}} = \|\mathbf{M}_p \cdots \mathbf{M}_1\|_{\mathbf{T}} < 1.$$

EM Algorithm: One can use (21) and (23) to show that

$$\theta^{i+1} - \hat{\theta} = \mathbf{M} \cdot (\theta^i - \hat{\theta})$$

for the EM algorithm, where (cf. (37))

$$\mathbf{M} = \mathbf{I} - (p\mathbf{D}_F + \mathbf{P})^{-1}\mathbf{H}.$$

Thus, the EM algorithm is closely related to the simultaneous overrelaxation (JOR) iteration (p. 72 of [37]). To establish that $\|\mathbf{M}\|_{\mathbf{T}} < 1$ for $\mathbf{T} = \mathbf{H}^{1/2}$ using Theorem 4, we must show that $\mathbf{S} + \mathbf{S}' > \mathbf{H}$, where in this case $\mathbf{S} = p\mathbf{D}_F + \mathbf{P}$. Since $\mathbf{H} = \mathbf{L}_F + \mathbf{D}_F + \mathbf{L}'_F + \mathbf{P}$ and $\mathbf{P} \geq \mathbf{0}$, it suffices to show that $p\mathbf{D}_F > \mathbf{L}_F + \mathbf{D}_F + \mathbf{L}'_F$, or equivalently that $p\mathbf{I} > \overline{\mathbf{L}} + \mathbf{I} + \overline{\mathbf{L}}$, where $\overline{\mathbf{L}} = \mathbf{D}_F^{-1/2}\mathbf{L}_F\mathbf{D}_F^{-1/2}$. Since $\mathbf{A}'\Pi^{-1}\mathbf{A}$ is positive definite by assumption, $x'(\overline{\mathbf{L}} + \mathbf{I} + \overline{\mathbf{L}}')x > 0$ for any nonzero x; therefore, using $x = \mathbf{e}_j \pm \mathbf{e}_k$, we see that $\overline{L}_{ij} \in (-1, 1)$. Thus, for any nonzero x, $x'(\overline{\mathbf{L}} + \mathbf{I} + \overline{\mathbf{L}}')x < (\sum_k \mid x_k \mid)^2 \leq p\|x\|^2$, where the second inequality is a special case of Hölder's inequality. The result then follows.

We have thus established that both the EM and SAGE algorithm converge globally. The convergence is globally monotonic in norm with respect to the norm $\mathbf{T} = \mathbf{H}^{1/2}$, i.e., $\mathcal{R}_+$ is all of $\mathbb{R}^p$.

D. Convergence Rates

To compare the root-convergence factors of EM and SAGE, we focus on the case where $\mathbf{P}$ is diagonal, since otherwise the EM algorithm is in general impractical. Therefore, from the results above

$$\begin{aligned}\rho_{\mathrm{EM}} &= \rho\big(\mathbf{I} - (p\mathbf{D}_F + \mathbf{P})^{-1}\mathbf{H}\big) \\ &= \rho\big(\mathbf{I} - ((p-1)\mathbf{D}_F + \mathbf{D}_H)^{-1}\mathbf{H}\big) \qquad (26)\\ \rho_{\mathrm{SAGE}} &= \rho\big(\mathbf{I} - (\mathbf{D}_H + \mathbf{L}_H)^{-1}\mathbf{H}\big) \qquad (27)\end{aligned}$$

since $\mathbf{P} = \mathbf{D}_P$ for diagonal $\mathbf{P}$.

Theorem 2: For linear superimposed signals in Gaussian noise with a diagonal penalty matrix, the SAGE algorithm asymptotically converges faster than the EM algorithm, i.e.

$$\rho_{\mathrm{SAGE}} < \rho_{\mathrm{EM}} < 1.$$

Proof: The right inequality follows from $\rho_{\mathrm{EM}} \leq \|\mathbf{M}\|_{\mathbf{T}} < 1$. From (24), $\mathbf{I} = \overline{\mathbf{L}}_H + \overline{\mathbf{D}}_H + \overline{\mathbf{L}}'_H$ where $\overline{\mathbf{L}}_H = \mathbf{H}^{-1/2}\mathbf{L}_H\mathbf{H}^{-1/2}$ and $\overline{\mathbf{D}}_H = \mathbf{H}^{-1/2}\mathbf{D}_H\mathbf{H}^{-1/2}$. Thus, for any vector x

$$x'\overline{\mathbf{L}}_H x = x'\overline{\mathbf{L}}'_H x = (\|x\|^2 - x'\overline{\mathbf{D}}_H x)/2. \qquad (28)$$

Since $\mathbf{I} - \mathbf{G}^{-1}\mathbf{H}$ is similar to the real symmetric matrix $\mathbf{I} - \mathbf{G}^{-1/2}\mathbf{H}\mathbf{G}^{-1/2}$, the eigenvalues of $\mathbf{I} - (\mathbf{D}_H + \mathbf{L}_H)^{-1}\mathbf{H}$ are real. For $\nu = \rho_{\mathrm{SAGE}} \in [0, 1)$ there exists $v \neq \mathbf{0}$ such that

$$[\mathbf{I} - (\mathbf{D}_H + \mathbf{L}_H)^{-1}\mathbf{H}]v = \nu v$$

thus

$$[\mathbf{I} - (\overline{\mathbf{D}}_H + \overline{\mathbf{L}}_H)^{-1}]x = \nu x$$

where $x = \mathbf{H}^{1/2}v$. Rearranging and multiplying both sides by x'

$$\|x\|^2 = (1-\nu)x'(\overline{\mathbf{L}}_H + \overline{\mathbf{D}}_H)x.$$

Combining with (28)

$$x'\overline{\mathbf{D}}_H x = \frac{1+\nu}{1-\nu}\|x\|^2.$$

By the invariance of eigenvalues under similarity transforms

$$\begin{aligned}\rho_{\mathrm{EM}} &= \rho(\mathbf{I} - ((p-1)\mathbf{D}_F + \mathbf{D}_H)^{-1}\mathbf{H}) \\ &= \rho(\mathbf{I} - ((p-1)\mathbf{D}_F + \mathbf{D}_H)^{-1/2}\mathbf{H} \\ &\quad \times ((p-1)\mathbf{D}_F + \mathbf{D}_H)^{-1/2}) \\ &\geq 1 - \frac{\|\mathbf{H}^{1/2}((p-1)\mathbf{D}_F + \mathbf{D}_H)^{-1/2}z\|^2}{\|z\|^2}\end{aligned}$$

for any z (by definition of spectral radius). In particular, for $z = ((p-1)\mathbf{D}_F + \mathbf{D}_H)^{1/2}\mathbf{H}^{-1/2}x$:

$$\begin{aligned}\rho_{\mathrm{EM}} &\geq 1 - \frac{\|x\|^2}{\|((p-1)\mathbf{D}_F + \mathbf{D}_H)^{1/2}\mathbf{H}^{-1/2}x\|^2} \\ &= 1 - \frac{\|x\|^2}{x'[(p-1)\mathbf{H}^{-1/2}\mathbf{D}_F\mathbf{H}^{-1/2} + \overline{\mathbf{D}}_H]x} \\ &\geq 1 - \frac{\|x\|^2}{x'\overline{\mathbf{D}}_H x} \\ &= 1 - \frac{1}{(1+\nu)/(1-\nu)} = \left(\frac{2}{1+\nu}\right)\nu \\ &> \nu = \rho_{\mathrm{SAGE}}\end{aligned}$$

where the last inequality follows from $\nu \in [0, 1)$. □

The inequalities in this proof are rather loose, and often the difference in convergence rate between EM and SAGE is more dramatic than the proof might suggest. To illustrate, consider the case where $\mathbf{P} = \mathbf{0}$. Then returning to (25), for the EM algorithm we have

$$\mathbf{M} = \frac{1}{p}\sum_{k=1}^{p}\mathbf{M}_k = \mathbf{T}^{-1}\left(\frac{1}{p}\sum_{k=1}^{p}\mathcal{P}_k^{\perp}\right)\mathbf{T}.$$

Since eigenvalues are invariant to similarity transforms, it follows that root-convergence factors for the two algorithms are given by the spectral radii

$$\begin{aligned}\rho_{\mathrm{EM}} &= \rho\left(\frac{1}{p}\sum_{k=1}^{p}\mathcal{P}_k^{\perp}\right), \\ \rho_{\mathrm{SAGE}} &= \rho\left(\prod_{k=1}^{p}\mathcal{P}_k^{\perp}\right)\end{aligned}$$

i.e., for the EM algorithm we have a convex combination of orthogonal projections and for the SAGE algorithm we have the product of those projections. Thus, this SAGE algorithm is closely related to the method of alternating projections [38], [39]. In particular, if $\mathbf{P} = \mathbf{0}$ and the columns of $\mathbf{A}$ are orthogonal, then $\rho_{\mathrm{SAGE}} = 0$ whereas $\rho_{\mathrm{EM}} \geq 1 - 1/p$, i.e.,

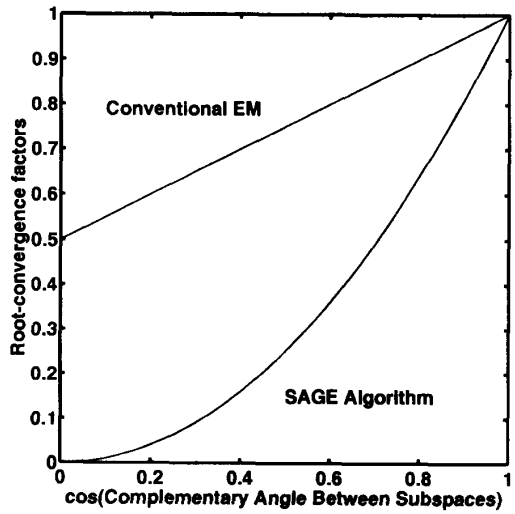

Fig. 3. Comparison of root-convergence factors for conventional EM algorithm and proposed SAGE algorithm versus complementary angle between subspaces of superimposed signals. The SAGE algorithm has a significantly improved convergence rate.

the SAGE algorithm converges in one iteration, while EM converges very slowly.

When $p = 2$, using a Gram–Schmidt argument one can show that $\mathbf{t}_1 = [1\ 0]'$ and $\mathbf{t}_2 = [\alpha\sqrt{1-\alpha^2}]'$ where $\alpha = |a_1'\Pi^{-1}a_2|/(\|a_1\|\|a_2\|)$ is the cosine of the complementary angle between a_1 and a_2. Thus

$$\rho_{\mathrm{SAGE}} = \rho\left(\begin{bmatrix}\alpha^2 & 0\\ 0 & 0\end{bmatrix}\right) = \alpha^2$$

$$\rho_{\mathrm{EM}} = \rho\left(\frac{1}{2}\begin{bmatrix}1-\alpha^2 & -\alpha\sqrt{1-\alpha^2}\\ -\alpha\sqrt{1-\alpha^2} & 1+\alpha^2\end{bmatrix}\right) = \frac{1}{2}+\frac{\alpha}{2}.$$

Fig. 3 illustrates that the root-convergence factor of SAGE is significantly smaller than that of EM, which substantially reduces the number of iterations required.

Not only is $\rho_{\mathrm{SAGE}} < \rho_{\mathrm{EM}}$, but also $\rho_{\mathrm{SAGE}} < \rho_{\mathrm{EM}}^2$, so one SAGE iteration is better than two EM iterations, at least when $p = 2$. Thus, even though the EM algorithm appears to have the advantage that one can parallelize the M-step using p processors that simultaneously update all parameters, in this case the convergence rate of the parallel algorithm is so much slower that a sequential update may be preferable. This depends, of course, on how difficult the M-step is; in the nonlinear case discussed in [2], the M-step is presumably fairly difficult, so parallelization may be advantageous. Equations (26) and (27) help one examine these types of tradeoffs.

V. Discussion

We have described a generalization of the classical EM algorithm in which one alternates between several hidden-data spaces rather than using just one, and updates only a subset of the elements of the parameter vector each iteration. By updating the parameters sequentially, rather than simultaneously, we demonstrated that SAGE algorithms yield faster convergence than EM algorithms in two signal processing applications.

The particular SAGE algorithms that we presented in this paper sacrifice one important characteristic of the EM algorithm: they are less amenable to a parallel implementation since they are coordinate-wise methods. However, the general SAGE method is very flexible, and work is in progress on more parallelizable algorithms using index sets S consisting of several elements of θ [30]. The benefits of parallelization must be weighed against the convergence rates for each application.

It is probably no coincidence that the applications we put forth are ones in which the terminology "incomplete-data" and "complete-data" as introduced in [1] are somewhat unnatural. In most of the statistical applications discussed in [1], there is a clearly identifiable portion of the data that is "missing," and hence one natural complete-data space. In contrast, there is nothing really "incomplete" about tomographic measurements; the problem is simply that the log-likelihood is difficult to maximize. The EM algorithm is thus just a computational tool. (To further illustrate this point, note that in classical missing data problems the estimates of the missing data may be of some intrinsic interest, whereas the "complete-data" for tomography is never explicitly computed and would be of little use anyway.) SAGE algorithms may be most useful in such contexts.

We have emphasized that the SAGE algorithm improves the asymptotic convergence rate. The actual convergence rate will certainly depend on how close the initial estimate is to a fixed-point. In tomography and image restoration, fast linear algorithms can provide good initializers for penalized likelihood estimation. A greedy algorithm like SAGE is likely to be most beneficial in applications where such initializers are available.

Appendix A
Monotone Convergence in Norm

Because the SAGE "algorithm" is so general, a single convergence theorem statement/proof cannot possibly cover all cases of interest (see, for example, the variety of special cases considered for the classical EM algorithm in [40].) Here we adopt the Taylor expansion approach of [4] since it directly illuminates the convergence rate properties and prescribes a region of monotone convergence in norm. However, this general approach has the drawback that it assumes the fixed point lies in the interior of Θ. This restriction is often not a necessary condition, and at least for some applications one can often find specific convergence results without the restriction, e.g., [3] and [30]. Readers who are satisfied with the assurance of monotonicity of the objective $\Phi(\theta^i)$, as provided by Theorem 1, may wish to simply skim this Appendix.

For simplicity, we discuss only the case where the index sets S^i are chosen cyclically with period K, i.e., $S^i = S^k$ where $k = 1 + (i \bmod K)$. We also assume that $\bigcup_{k=1}^{K} S^k = \{1,\ldots,p\}$ so that each parameter is updated at least once per cycle.

Before stating the convergence theorem, several definitions are needed. Consider an index set S, and let m denote its

cardinality. Bearing in mind our notational convention that $\phi^S(\theta_S;\overline{\theta}) = \phi^S(\theta_S;\overline{\theta}_S,\overline{\theta}_{\bar{S}})$, we define the $m \times m$ matrices

$$(\nabla^{200}\phi^S)(\theta_S;\overline{\theta}) = (\nabla_{\theta_S}\nabla'_{\theta_S}\phi^S)(\theta_S;\overline{\theta}_S,\overline{\theta}_{\bar{S}})$$

and

$$(\nabla^{110}\phi^S)(\theta_S;\overline{\theta}) = (\nabla_{\overline{\theta}_S}\nabla'_{\theta_S}\phi^S)(\theta_S;\overline{\theta}_S,\overline{\theta}_{\bar{S}})$$

and the $m \times (p-m)$ matrix

$$(\nabla^{101}\phi^S)(\theta_S;\overline{\theta}) = (\nabla_{\overline{\theta}_{\bar{S}}}\nabla'_{\theta_S}\phi^S)(\theta_S;\overline{\theta}_S,\overline{\theta}_{\bar{S}})$$

where ∇ denotes the (row) gradient operator and ∇' its transpose. Let $\hat{\theta}$ be a fixed point of the SAGE algorithm, and define

$$\mathbf{U}^S(\theta_S;\overline{\theta}) = -\int_0^1 (\nabla^{200}\phi^S)(t\theta_S + (1-t)\hat{\theta}_S; t\overline{\theta} + (1-t)\hat{\theta})dt \quad (29)$$

$$\mathbf{V}^S(\theta_S;\overline{\theta}) = \int_0^1 (\nabla^{110}\phi^S)(t\theta_S + (1-t)\hat{\theta}_S; t\overline{\theta} + (1-t)\hat{\theta})dt \quad (30)$$

and

$$\mathbf{W}^S(\theta_S;\overline{\theta}) = \int_0^1 (\nabla^{101}\phi^S)(t\theta_S + (1-t)\hat{\theta}_S; t\overline{\theta} + (1-t)\hat{\theta})dt. \quad (31)$$

Let $\mathbf{R}^S$ denote the $p \times p$ permutation matrix that reorders the elements of $\{S, \bar{S}\}$ into $\{1,\ldots,p\}$. Then define the $p \times p$ composite matrix

$$\mathbf{M}^S(\theta_S;\overline{\theta}) = \mathbf{R}^S \begin{bmatrix} \mathbf{U}^S(\theta_S;\overline{\theta})^{-1}[\mathbf{V}^S(\theta_S;\overline{\theta})\mathbf{W}^S(\theta_S;\overline{\theta})] \\ \mathbf{0}_{(p-m)\times m} \qquad\qquad \mathbf{I}_{p-m} \end{bmatrix} (\mathbf{R}^S)' \quad (32)$$

where $\mathbf{I}_n$ denotes the $n \times n$ identity matrix.

In addition, define

$$\widetilde{\Theta}^S(\overline{\theta}) = \{\theta_S \in \Theta^S(\overline{\Theta}) : \phi^S(\theta_S;\overline{\theta}) \geq \phi^S(\overline{\theta}_S;\overline{\theta})\}.$$

With the above definitions, we can define the following region of monotone convergence in norm to $\hat{\theta}$.

Definition 3: $\mathcal{R}_+ \subset \Theta$ is a region of monotone convergence in norm if there exists a nonsingular $p \times p$ matrix $\mathbf{T}$ such that $\mathcal{R}_+$ is an open ball with respect to the norm $\|\cdot\|_{\mathbf{T}}$ and

1. For $k = 1,\ldots,K$, $\mathbf{U}^{S^k}(\theta_{S^k};\overline{\theta})$ is invertible for all $\overline{\theta} \in \mathcal{R}_+$ and for all $\theta_{S^k} \in \widetilde{\Theta}^{S^k}(\overline{\theta})$ (see (7)).
2. For $k = 1,\ldots,K$, $\|\mathbf{M}^{S^k}(\theta_{S^k};\overline{\theta})\|_{\mathbf{T}} \leq 1$ for all $\overline{\theta} \in \mathcal{R}_+$ and for all $\theta_{S^k} \in \widetilde{\Theta}^{S^k}(\overline{\theta})$,
3. There exists $\alpha < 1$ such that for any $\overline{\theta}^1,\ldots,\overline{\theta}^K \in \mathcal{R}_+$ and $\theta_{S^k} \in \widetilde{\Theta}^{S^k}(\overline{\theta}^k)$, $\quad k = 1,\ldots,K$

$$\|\mathbf{M}^{S^K}(\theta_{S^K};\overline{\theta}^K)\cdots\mathbf{M}^{S^1}(\theta_{S^1};\overline{\theta}^1)\|_T \leq \alpha. \quad (33)$$

In general, $\mathbf{T}$ may depend on $\hat{\theta}$, but we do not allow $\mathbf{T}$ to vary with iteration. The hard work is to verify condition 3 (see Appendix B), but if one can do so, the reward is the following theorem.

Theorem 3: Assume i) $S^i = S^k$ where $k = 1+(i \text{ modulo } K)$ and $\bigcup_{k=1}^K S^k = \{1,\ldots,p\}$, ii) $\hat{\theta}$ is a fixed point of the SAGE algorithm (5) in the interior of Θ, iii) for all $\overline{\theta} \in \mathcal{R}_+$ the maximum over θ_{S^k} of $\phi^{S^k}(\theta_{S^k};\overline{\theta})$ is in the interior of $\Theta^{S^k}(\overline{\theta})$, iv) $\phi^{S^k}(\theta_{S^k};\overline{\theta})$ is twice differentiable in both arguments $\forall\overline{\theta} \in \Theta$ and $\forall\theta_{S^k} \in \Theta^{S^k}(\overline{\theta})$, and v) the region of monotone convergence $\mathcal{R}_+$ for a norm $\|\cdot\|_{\mathbf{T}}$ is nonempty.

1. If $\theta^0 \in \mathcal{R}_+$ then
$$\|\theta^{i+1} - \hat{\theta}\|_{\mathbf{T}} \leq \|\theta^i - \hat{\theta}\|_{\mathbf{T}} \ \forall i$$
and
$$\|\theta^{(i+1)K} - \hat{\theta}\|_{\mathbf{T}} \leq \alpha\|\theta^{iK} - \hat{\theta}\|_{\mathbf{T}} \quad (34)$$
where $\alpha < 1$ is defined by (33). Therefore, $\|\theta^{iK} - \hat{\theta}\|_{\mathbf{T}}$ converges monotonically to zero with at least linear rate.
2. The root-convergence factor [41] of the subsequence $\{\theta^{iK}\}_{i=0}^{\infty}$ is given by the spectral radius
$$\rho\Big(\mathbf{M}^{S^K}(\hat{\theta}_{S^K};\hat{\theta})\cdots\mathbf{M}^{S^1}(\hat{\theta}_{S^1};\hat{\theta})\Big) \quad (35)$$
which is bounded above by $\alpha < 1$.

Note that by the equivalence of matrix norms (p. 29 of [37]), monotone convergence with respect to the norm $\|\cdot\|_T$ implies convergence with respect to any other norm, although probably nonmonotonically. Since the index sets are chosen cyclically, a "full iteration" consists of K updates; therefore, (34) bounds the convergence rate of the subsequence $\{\theta^{iK}\}_{i=0}^{\infty}$.

Proof: Consider the ith iteration and let $S = S^k$ where $k = 1 + (i \text{ modulo } K)$. Define

$$z = \begin{bmatrix} \theta_S \\ \overline{\theta}_S \\ \overline{\theta}_{\bar{S}} \end{bmatrix}, \quad \hat{z} = \begin{bmatrix} \hat{\theta}_S \\ \hat{\theta}_S \\ \hat{\theta}_{\bar{S}} \end{bmatrix}$$

and let $\phi^S(z) = \phi^S(\theta_S;\overline{\theta})$. Let

$$d(z) = d(\theta_S;\overline{\theta}) = (\nabla'_{\theta_S}\phi^S)(z)$$

then by assumption iv) we can apply the Taylor formula with remainder [42] to expand $d(z)$ about $\hat{z}$

$$d(z) = d(\hat{z}) + \int_0^1 (\nabla d)(tz + (1-t)\hat{z})dt\,(z - \hat{z}).$$

Since $\hat{\theta}$ is a fixed point of the SAGE algorithm, by assumption iii) and iv) $d(\hat{z}) = \mathbf{0}$. Observe that by the definitions (29)–(31):

$$(\nabla d)(z) = [-\mathbf{U}^S(\theta_S;\overline{\theta})\ \mathbf{V}^S(\theta_S;\overline{\theta})\ \mathbf{W}^S(\theta_S;\overline{\theta})].$$

By assumptions iii) and iv) $d(\theta_{S^k}^{i+1};\theta^i) = \mathbf{0}$ for the SAGE algorithm (5), so

$$\mathbf{U}^{S^k}(\theta_{S^k}^{i+1};\theta^i)(\theta_{S^k}^{i+1} - \hat{\theta}_{S^k}) = \mathbf{V}^{S^k}(\theta_{S^k}^{i+1};\theta^i)(\theta_{S^k}^i - \hat{\theta}_{S^k}) + \mathbf{W}^{S^k}(\theta_{S^k}^{i+1};\theta^i)(\theta_{\bar{S}^k}^i - \hat{\theta}_{\bar{S}^k}). \quad (36)$$

By property 1 (invertibility) of Definition 3

$$\theta_{S^k}^{i+1} - \hat{\theta}_{S^k} = \mathbf{U}^{S^k}(\theta_{S^k}^{i+1};\theta^i)^{-1}\mathbf{V}^{S^k}(\theta_{S^k}^{i+1};\theta^i)(\theta_{S^k}^i - \hat{\theta}_{S^k}) + \mathbf{U}^{S^k}(\theta_{S^k}^{i+1};\theta^i)^{-1}\mathbf{W}^{S^k}(\theta_{S^k}^{i+1};\theta^i)(\theta_{\bar{S}^k}^i - \hat{\theta}_{\bar{S}^k}).$$

From (6) the components of $\theta^i_{\tilde{S}^k}$ are just copied, so after permuting using $\mathbf{R}^S$ (32)

$$\theta^{i+1} - \hat{\theta} = \mathbf{M}^{S^k}(\theta^{i+1}_{S^k}; \theta^i)(\theta^i - \hat{\theta}) \tag{37}$$

where $\mathbf{M}^{S^k}$ was defined by (32). Therefore, since $\theta^0 \in \mathcal{R}_+$ by property 2 of Definition 3

$$\|\theta^{i+1} - \hat{\theta}\|_{\mathbf{T}} \leq \|\theta^i - \hat{\theta}\|_{\mathbf{T}}$$

and therefore, it follows by induction that $\theta^i \in \mathcal{R}_+$. A full cycle consists of one update over each of the K index sets; therefore, applying (37) K times

$$(\theta^{(i+1)K} - \hat{\theta}) = \mathbf{M}^{S^k}(\theta^{(i+1)K}_{S^K}; \theta^{(i+1)K-1}) \times \cdots \mathbf{M}^{S^1}(\theta^{iK+1}_{S^1}; \theta^{iK})(\theta^{iK} - \hat{\theta}).$$

Thus, by property 3 of Definition 3

$$\|\theta^{(i+1)K} - \hat{\theta}\|_{\mathbf{T}} \leq \alpha \|\theta^{iK} - \hat{\theta}\|_{\mathbf{T}}$$

and therefore, the subsequence $\{\theta^{iK}\}_{i=0}^{\infty}$ converges monotonically in norm to $\hat{\theta}$ as $i \to \infty$ with linear rate at most α.

By continuity of the derivatives of Φ_{S^k} one can show [4] that the root convergence factor of the subsequence θ^{iK} is governed by the spectral radius

$$\rho(\mathbf{M}^{S^K}(\hat{\theta}_{S^K}; \hat{\theta}) \cdots \mathbf{M}^{S^1}(\hat{\theta}_{S^1}; \hat{\theta})).$$

Since the spectral radius is bounded above by any matrix norm, the root convergence factor is bounded above by α. □

Appendix B
$\mathcal{R}_+$ Is Nonempty

In this appendix, we show that the region of monotone convergence in norm $\mathcal{R}_+$ is nonempty for suitably regular problems. Thus, the conditions for Theorem 3 are reasonable, and the superimposed signals example in Section IV is a concrete example.

First note that from (10) one can show that

$$(\nabla^{110} H^S)(\hat{\theta}_S; \hat{\theta}) = -(\nabla^{200} H^S)(\hat{\theta}_S; \hat{\theta}),$$
$$(\nabla^{101} H^S)(\hat{\theta}_S; \hat{\theta}) = \mathbf{0}$$

(cf. (3.16) of [1]). For an index set S, define

$$\mathbf{F}_{\mathbf{X}^S|\mathbf{y}} = -(\nabla^{200} H^S)(\hat{\theta}_S; \hat{\theta})$$

then from (10), one can see that the matrix $\mathbf{F}_{\mathbf{X}^S|\mathbf{y}}$ is the conditional Fisher information of X^S for θ_S, given $Y = y$ and given all of the other parameters $\theta_{\tilde{S}}$.

Define the Hessian of the objective at $\hat{\theta}$ by

$$\mathbf{H} = -\nabla^2 \Phi(\hat{\theta})$$

and the following submatrices of the Hessian

$$\mathbf{H}^S = -(\nabla_{\theta_S} \nabla'_{\theta_S} \Phi)(\hat{\theta})$$
$$\mathbf{H}^{S,\tilde{S}} = -(\nabla_{\theta_S} \nabla'_{\theta_{\tilde{S}}} \Phi)(\hat{\theta}).$$

(To simplify notation we leave implicit the dependence on $\hat{\theta}$.) Note that $\mathbf{H}^S$ is the curvature of the objective Φ with respect to θ_S, and $\mathbf{H}^{S,\tilde{S}}$ is the coupling between θ_S and $\theta_{\tilde{S}}$ induced by Φ. Combining the above definitions with (29)–(31)

$$\mathbf{U}^S = \mathbf{H}^S + \mathbf{F}_{\mathbf{X}^S|\mathbf{y}}$$
$$\mathbf{V}^S = \mathbf{F}_{\mathbf{X}^S|\mathbf{y}}$$
$$\mathbf{W}^S = -\mathbf{H}^{S,\tilde{S}}. \tag{38}$$

If $\mathbf{H}$ is positive definite, then $\mathbf{U}^S$ will be invertible; therefore, by (32)

$$\mathbf{M}^S = \mathbf{R}^S \begin{bmatrix} (\mathbf{U}^S)^{-1}\mathbf{V}^S & (\mathbf{U}^S)^{-1}\mathbf{W}^S \\ 0 & \mathbf{I} \end{bmatrix} (\mathbf{R}^S)'.$$

Substituting in (38)

$$\begin{aligned}
&(\mathbf{R}^S)'\mathbf{M}^S\mathbf{R}^S \\
&= \begin{bmatrix} (\mathbf{H}^S + \mathbf{F}_{\mathbf{X}^S|\mathbf{y}})^{-1}\mathbf{F}_{\mathbf{X}^S|\mathbf{y}} & -(\mathbf{H}^S + \mathbf{F}_{\mathbf{X}^S|\mathbf{y}})^{-1}\mathbf{H}^{S,\tilde{S}} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \\
&= \mathbf{I} - \begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix} (\mathbf{H}^S + \mathbf{F}_{\mathbf{X}^S|\mathbf{y}})^{-1} \begin{bmatrix} \mathbf{H}^S \mathbf{H}^{S,\tilde{S}} \end{bmatrix} \\
&= \mathbf{I} - \begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix} (\mathbf{H}^S + \mathbf{F}_{\mathbf{x}^S|\mathbf{y}})^{-1} [\mathbf{I}\ \mathbf{0}](\mathbf{R}^S)'\mathbf{H}\mathbf{R}^S.
\end{aligned}$$

Thus

$$\mathbf{H}^{\frac{1}{2}}\mathbf{M}^S\mathbf{H}^{-\frac{1}{2}} = \mathbf{I} - \mathbf{H}^{\frac{1}{2}}\mathbf{R}^S \begin{bmatrix} (\mathbf{H}^S + \mathbf{F}_{\mathbf{X}^S|\mathbf{y}})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} (\mathbf{R}^S)'\mathbf{H}^{\frac{1}{2}}. \tag{39}$$

For simplicity, we now consider the case the index sets are disjoint and are chosen cyclically in the natural order, i.e., $S^i = S^k$, where $k = (1 + i \text{ modulo } K)$ and $\{S^1, \ldots, S^K\} = \{1, \ldots, p\}$. In that case, it follows from (39) that

$$\mathbf{M}^{S^K} \cdots \mathbf{M}^{S^1} = \mathbf{I} - (\tilde{\mathbf{D}}_H + \tilde{\mathbf{D}}_F + \tilde{\mathbf{L}}_H)^{-1}\mathbf{H} \tag{40}$$

where $\tilde{\mathbf{D}}_F$ is block-diagonal with $\mathbf{F}_{\mathbf{x}^S|\mathbf{y}}$ in the kth block, and

$$\mathbf{H} = \tilde{\mathbf{L}}_H + \tilde{\mathbf{D}}_H + \tilde{\mathbf{L}}'_H \tag{41}$$

where $\tilde{\mathbf{D}}_H$ is a block diagonal matrix containing the diagonal blocks of $\mathbf{H}$ that correspond to the subsets S^k, and $\tilde{\mathbf{L}}_H$ is the corresponding strictly lower block triangular matrix. We can thus establish that $\|\mathbf{M}^{S^K} \cdots \mathbf{M}^{S^1}\|_{\mathbf{T}} < 1$ by using the following "splitting matrix" theorem (p. 79 of [37]).

Theorem 4: If $\mathbf{H}$ is positive definite and $\mathbf{S}$ is invertible, then

$$\|\mathbf{I} - \mathbf{S}^{-1}\mathbf{H}\|_{\mathbf{T}} < 1$$

for $\mathbf{T} = \mathbf{H}^{\frac{1}{2}}$ if

$$\mathbf{S} + \mathbf{S}' > \mathbf{H}. \tag{42}$$

From (40), for a SAGE algorithm $\mathbf{S} = \tilde{\mathbf{D}}_H + \tilde{\mathbf{D}}_F + \tilde{\mathbf{L}}_H$, so in light of (41), condition (42) of Theorem 4 is satisfied. Thus, $\|\mathbf{M}^{S^K} \cdots \mathbf{M}^{S^1}\| < 1$.

Using the relationships derived above, one can establish the following result [43].

Theorem 5: Let $\hat{\theta}$ be a fixed point of a SAGE algorithm, and assume that Φ is strictly concave on an open set local to $\hat{\theta}$ (so that $\mathbf{H}$ is positive definite). Then if Φ and the functions ϕ^S are all twice continuously differentiable near $\hat{\theta}$, there exists a nonempty region of monotone convergence in norm $\mathcal{R}_+$ satisfying the conditions of Definition 3 for the norm induced by $\mathbf{T} = \mathbf{H}^{\frac{1}{2}}$.

APPENDIX C
FISHER INFORMATION

From (35) we see that the root-convergence factor of a SAGE algorithm is given by the spectral radius of a product of matrices $\mathbf{M}^S(\hat{\theta}_S; \hat{\theta})$ of the form (32). For an EM algorithm, this spectral radius increases towards 1 as the complete-data becomes more informative, i.e., as its Fisher information increases [1], [4], [5]. In this section we demonstrate that a similar relationship holds for the convergence rate of a SAGE algorithm.

Defining

$$\delta^i = \mathbf{H}^{\frac{1}{2}} \cdot (\theta^i - \hat{\theta})$$

we see from (39) that for δ^i small

$$\delta^{i+1} \approx$$

$$\left(\mathbf{I} - \mathbf{H}^{\frac{1}{2}}\mathbf{R}^S \begin{bmatrix} (\mathbf{H}_S + \mathbf{F}_{\mathbf{X}^S|\mathbf{y}})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} (\mathbf{R}^S)' \mathbf{H}^{\frac{1}{2}}\right)\delta^i. \quad (43)$$

This last equation suggests that minimizing $\mathbf{F}_{\mathbf{X}^S|\mathbf{y}}$ will improve the rate of convergence of $\|\delta^i\|$ to 0. To demonstrate this more formally, let $\tilde{\mathbf{D}}_{F_1}$ and $\tilde{\mathbf{D}}_{F_2}$ be the block diagonal aggregate of the Fisher information matrices for two SAGE algorithms with $\tilde{\mathbf{D}}_{F_1} < \tilde{\mathbf{D}}_{F_2}$. Then one can use (40) with an argument similar to the proof of Theorem 2 to show that

$$\rho(\mathbf{I} - (\tilde{\mathbf{D}}_H + \tilde{\mathbf{D}}_{F_1} + \tilde{\mathbf{L}}_H)^{-1}\mathbf{H}) < \rho(\mathbf{I} - (\tilde{\mathbf{D}}_H + \tilde{\mathbf{D}}_{F_2} + \tilde{\mathbf{L}}_H)^{-1}\mathbf{H}).$$

Thus, less informative hidden-data spaces lead to smaller root-convergence factors and hence faster converging SAGE algorithms. In particular, once one has chosen the index sets S^k the optimal hidden-data space from the point of view of *asymptotic* convergence rate would simply be $X^S = Y$, since then $\mathbf{F}_{\mathbf{X}^S|\mathbf{y}} = 0$. But that choice will often lead to an intractable M-step. The SAGE algorithm allows one to choose hidden-data spaces whose Fisher information matrices are much smaller than that of the usual complete data of an EM algorithm.

Finally, note that from (43) we see that since $\mathbf{H}$ is determined by Φ, once the index sets are chosen, the only design issue left is to choose the hidden-data X^S. This choice should be made by considering the tradeoff between making $\mathbf{F}_{\mathbf{X}^S|\mathbf{y}}$ small but yet making the M-step tractable.

ACKNOWLEDGMENT

The first author gratefully acknowledges helpful discussions on the superimposed signals application with Y. Bresler and S.-F. Yau. The authors thank the reviewers for helpful suggestions, including references to [18] and [15].

REFERENCES

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc. Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[2] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the EM algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 4, pp. 477–489, Apr. 1988.

[3] K. Lange and R. Carson, "EM reconstruction algorithms for emission and transmission tomography," *J. Comp. Assisted Tomography*, vol. 8, no. 2, pp. 306–316, Apr. 1984.

[4] A. O. Hero and J. A. Fessler, "Asymptotic convergence properties of EM-type algorithms," Communications and Signal Processing Laboratory, Dept. of EECS, Univ. of Michigan, Ann Arbor, MI, Technical Report 282, Apr. 1993.

[5] J. A. Fessler, N. H. Clinthorne, and W. Leslie Rogers, "On complete data spaces for PET reconstruction algorithms," *IEEE Trans. Nucl. Sci.*, vol. 40, no. 4, pp 1055–1061, Aug. 1993.

[6] J. A. Fessler and A. O. Hero, "Complete-data spaces and generalized EM algorithms," in *Proc. IEEE Conf. ASSP*, 1993, vol. 4, pp. 1–4.

[7] D. G. Politte and D. L. Snyder, "Corrections for accidental coincidences and attenuation in maximum-likelihood image reconstruction for positron-emission tomography," *IEEE Trans. Med. Imag.*, vol. 10, no. 1, pp. 82–89, Mar. 1991.

[8] J. A. Fessler and A. O. Hero, "New complete-data spaces and faster algorithms for penalized-likelihood emission tomography," in *Conf. Rec. IEEE Nucl. Sci. Symp. Med. Imag.*, 1993, pp. 1897–1901.

[9] P. J. Green, "On use of the EM algorithm for penalized likelihood estimation," *J. Royal Stat. Soc. Series B*, vol. 52, no. 3, pp. 443–452, 1990.

[10] T. Hebert and R. Leahy, "A Bayesian reconstruction algorithm for emission tomography using a Markov random field prior," in *Proc. SPIE 1092, Med. Imag. III: Image Processing*, 1989, pp. 458–466.

[11] ——, "A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors," *IEEE Trans. Med. Imag.*, vol. 8, no. 2, pp. 194–202, June 1989.

[12] J. A. Fessler, N. H. Clinthorne, and W. L. Rogers, "Regularized emission image reconstruction using imperfect side information," *IEEE Trans. Nucl. Sci.*, vol. 39, no. 5, pp. 1464–1471, Oct. 1992.

[13] K. Lange, "Convergence of EM image reconstruction algorithms with Gibbs smoothing," *IEEE Trans. Med. Imag.*, vol. 9, no. 4, pp. 439–446, Dec. 1990 (Corrections, June 1991).

[14] B. W. Silverman *et al.* "A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography," *J. Royal Stat. Soc. Series B*, vol. 52, no 2, pp. 271–324, 1990.

[15] A. R. De Pierro, "A generalization of the EM algorithm for maximum likelihood estimates from incomplete data," Med. Imag. Processing Group, Dept. of Radiology, Univ. of Pennsylvania, Technical Report MIPG119, 1987.

[16] G. T. Herman, A. R. De Pierro, and N. Gai, "On methods for maximum a posteriori image reconstruction with a normal prior," *J. Visual Commun. Image. Represent.*, vol. 3, no. 4, pp. 316–324, Dec. 1992.

[17] A. R. De Pierro, "A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography," to be published in *IEEE Trans. Med. Imag.*

[18] M. Abdalla and J. W. Kay, "Edge-preserving image restoration," in *Stochastic Models, Statistical Methods, and Algorithms in Image Analysis* (P. Borne, A. Frigessi, and M. Piccioni, Eds.)., vol. 74 of *Lecture Notes in Statistics*. New York: Springer, 1992, pp. 1–13.

[19] C. Bouman and K. Sauer, "Fast numerical methods for emission and transmission tomographic reconstruction," in *Proc. Conf. Inform. Sci. Syst. Johns Hopkins*, 1993.

[20] ——, "A unified approach to statistical tomography using coordinate descent optimization," 1993, submitted to *IEEE Trans. Med. Imag.*

[21] T. A. Louis, "Finding the observed information matrix when using the EM algorithm," *J. Royal Stat. Soc. Series B*, vol. 44, no. 2, pp. 226–233, 1992.

[22] R. M. Lewitt and G. Muehllehner, "Accelerated iterative reconstruction for positron emission tomography based on the EM algorithm for maximum likelihood estimation," *IEEE Trans. Med. Imag.*, vol. MI–5, no. 1, pp. 16–22, Mar. 1986.

[23] L. Kaufman, "Implementing and accelerating the EM algorithm for positron emission tomography," *IEEE Trans. Med. Imag.*, vol. MI–6, no. 1, pp. 37–51, Mar. 1987.

[24] I. Meilijson, "A fast improvement to the EM algorithm on its own terms," *J. Royal Stat. Soc. Series B*, vol. 5, no. 1, pp. 127–138, 1989.

[25] M. Jamshidian and R. I. Jennrich, "Conjugate gradient acceleration of the EM algorithm," *J. Amer. Stat. Assoc.*, vol. 88, no. 421, pp. 221–228, 1993.

[26] K. Sauer and C. Bouman, "A local update strategy for iterative reconstruction from projections," *IEEE Trans. Signal Processing*, vol. 41, no. 2, pp. 534–548, Feb. 1993.

[27] W. H. Press *et al.*, *Numerical Recipes in C*. Cambridge, UK: Cambridge Univ. Press, 1988.

[28] M. Segal and W. Weinstein, "The cascade EM algorithm," *Proc. IEEE*,

vol 76, no. 10, pp. 1388–1390, Oct. 1988.
[29] J. A. Fessler, "Penalized weighted least-squares image reconstruction for positron emission tomography," *IEEE Trans. Med. Imag.*, vol. 13, no 2, pp. 290–300, June 1994.
[30] J. A. Fessler and A. O. Hero, "Space-alternating generalized EM algorithms for penalized maximum-likelihood image reconstruction," Commun. and Signal Processing Lab., Dept. of Elec. Eng. and Comp. Sci., Univ. of Michigan, Ann Arbor, MI, Tech. Rep. 286, Feb. 1994.
[31] R. Boyles, "On the convergence of the EM algorithm," *J. Royal Stat. Soc. Series B*, vol. 45, no. 1, pp. 47–50, 1993.
[32] L. A. Shepp and Y. Verdi, "Maximum likelihood reconstruction for emission tomography," *IEEE Trans. Med. Imag.*, vol. MI–1, no. 2, pp. 113–122, Oct. 1982.
[33] C. L. Byrne, "Iterative image reconstruction reconstruction algorithms based on cross-entropy minimization," *IEEE Trans. Imag. Processing*, vol. 2, no. 1, pp. 96–103, Jan. 1993.
[34] L. Kaufman, "Maximum likelihood, least squares, and penalized least squares for PET," *IEEE Trans. Med. Imag.*, vol. 12, no. 2, pp. 200–214, June 1993.
[35] J. A. Fessler, "Object-based 3-D reconstuction of arterial trees from a few projections," Ph.D. thesis, Stanford Univ., Stanford, CA, Aug. 1990.
[36] D. Chazan, Y. Stettiner, and D. Malah, "Optimal multi-path estimation using the EM algorithm for co-channel speech separation," in *Proc. IEEE Conf. Acoust., Speech, Signal Processing*, 1993, vol. 2, pp. 728–731.
[37] D. M. Young. *Iterative Solution of Large Linear Systems*. New York: Academic, 1971.
[38] I. Ziskind and M. Wax, "Maximum liklihood localization of multiple sources by alternating projection," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 26, no. 10, pp. 1553–1560, Oct. 1988.
[39] S. Kayalar and H. L. Weinert, "Error bounds for the method of alternating projections," *Math Contr. Signals Syst.*, vol. 1, pp. 43–59, 1988.
[40] C. F. J. Wu, "On the convergence properties of the EM algorithm," *Ann. Stat.*, vol. 11, no. 1, pp. 95–103, 1983.
[41] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*. New York: Academic, 1970.
[42] E. Polak, *Computational Methods in Optimization: a Unified Approach*. Orlando, FL: Academic, 1971.
[43] A. O. Hero and J. A. Fessler, "Convergence in norm for alternating expectation-maximization (EM)-type algorithms," to be published in *Statistica Sinica*.

Jeffrey A. Fessler (S'83–M'90) received the B.S.E.E. degree from Purdue University in 1985, the M.S.E.E. degree from Stanford University in 1986, and the M.S. degree in statistics from Stanford University in 1989. From 1985 to 1988, he was a National Science Foundation Graduate Fellow at Stanford, where he earned the Ph.D. degree in electrical engineering in 1990. From 1991 to 1992, he was a Department of Energy Alexander Hollaender Post-Doctoral Fellow in the Division of Nuclear Medicine at the University of Michigan.

Since 1993, he has been an Assistant Professor in Nuclear Medicine and the Bioengineering Program of the University of Michigan. His research interests are in statistical aspects of medical imaging.

Alfred O. Hero (S'79–M'84) was born in Boston, MA on December 5, 1955. He received the B.S. degree *summa cum laude* from Boston University in 1980 and the Ph.D. degree from Princeton University in 1984, both in electrical engineering. He held the G.V.N. Lothrop Fellowship in Engineering at Princeton University

He is presently Associate Professor of Electrical Engineering and Computer Science at the University of Michigan, Ann Arbor. From 1987 to 1989, he was Visiting Scientist at the M.I.T. Lincoln Laboratory, Lexington, MA. In 1991, he was Visiting Professor at the Ecole National de Techniques Avancees in Paris, France. In 1993, he was a Whitney Clay Ford Fellow at the Ford Motor Company. His research interests are in the areas of detection and estimation theory applied to statistical signal and image processing.

Dr. Hero is a member of Tau Beta Pi, the New York Academy of Sciences, and Commission C of the International Union of Radio Science. He was Chairman for Publicity for the 1986 IEEE International Symposium on Information Theory. He is General Chairman for the 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing.