Lecture Notes on the EM Algorithm

Mário A. T. Figueiredo
Instituto de Telecomunicações,
Instituto Superior Técnico
1049-001 Lisboa,
Portugal
mtf@lx.it.pt

June 4, 2008

Abstract

This is a tutorial on the EM algorithm, including modern proofs of monotonicity, and several examples focusing on the use of EM to handle heavy-tailed models (Laplace, Student) and on finite mixture estimation.

1 The Algorithm

Consider a general scenario in which we have observed data $\mathbf{x}$, and a set of unknown parameters $\boldsymbol{\theta}$. Let us also assume some prior $p(\boldsymbol{\theta})$ for the parameters, which could well be a flat prior. The *a posteriori* probability function $p(\boldsymbol{\theta}|\mathbf{x})$ is proportional to $p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Now, suppose that finding the MAP estimate of $\boldsymbol{\theta}$ would be easier if we had access to some other data $\mathbf{y}$, that is, it would be easy to maximize $p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, where $p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta})$ is related to $p(\mathbf{x}|\boldsymbol{\theta})$ via marginalization

$$p(\mathbf{x}|\boldsymbol{\theta}) = \int p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}) \, d\mathbf{y}. \tag{1}$$

The *expectation-maximization* (EM) algorithm is an iterative procedure which can be shown to converge to a (local) maximum of the marginal *a posteriori* probability function $p(\boldsymbol{\theta}|\mathbf{x}) = p(\mathbf{x}|\boldsymbol{\theta}) \, p(\boldsymbol{\theta})$, without the need to explicitly manipulate the marginal likelihood $p(\mathbf{x}|\boldsymbol{\theta})$. EM was first proposed in [9], and since then it has attracted a great deal of interest and stimulated a considerable amount of research. For example, the EM algorithm has been often used

in image restoration/reconstruction problems (see, *e.g.*, [18], [19], [25], [26], [30], [41], [45], [47], [50], [56]). Instances of the use of EM in statistics, computer vision, signal processing, machine learning, and pattern recognition are too numerous to be listed here. Several books have been fully devoted to the EM algorithm, while many others contain large portions covering this technique [34], [38], [51].

In its original formulation, EM is presented as an algorithm to perform ML parameter estimation with missing data [9], [34], [38], [51]; *i.e.* the unobserved $\mathbf{y}$ is said to be *missing*, and the goal is to find the ML estimate of $\boldsymbol{\theta}$. Usually, $\mathbf{z} = (\mathbf{y}, \mathbf{x})$ is called the *complete* data, while $p(\mathbf{z}|\boldsymbol{\theta})$ is termed the *complete* likelihood function. The complete likelihood is supposed to be relatively easy to maximize with respect to $\boldsymbol{\theta}$. In many cases, the missing data is artificially inserted as a means of allowing the use of the EM algorithm to find a difficult ML estimate. Specifically, when $p(\mathbf{x}|\boldsymbol{\theta})$ is difficult to maximize with respect to $\boldsymbol{\theta}$, but there is an alternative model $p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta})$ which is easy to maximize with respect to $\boldsymbol{\theta}$, and such that $p(\mathbf{x}|\boldsymbol{\theta})$ is related to $p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta})$ via marginalization (Eq. (1)).

The concept of observed/complete data may also be generalized; the observed data does not have to be a "portion" of the complete data (as in the previous paragraph), but any non-invertible transformation of the complete data, *i.e.*, $\mathbf{x} = h(\mathbf{z})$, where $\mathbf{z}$ is the complete data. In this case, the marginal likelihood is obtained by computing

$$p(\mathbf{x}|\boldsymbol{\theta}) = \int_{h^{-1}(\mathbf{x})} p(\mathbf{z}|\theta)\, d\mathbf{z},$$

where $h^{-1}(\mathbf{x}) = \{\mathbf{z} : h(\mathbf{z}) = \mathbf{x}\}$ (actually, the integral should be replaced by a sum if the the set $h^{-1}(\mathbf{x})$ is composed of isolated points; a uniform notation could be obtained with Stieltjes integrals [6]). For example, the complete data may be a vector $\mathbf{z} \in I\!R^d$ and the observations may be only the absolute values of the components, that is $\mathbf{x} = |\mathbf{z}| \in (I\!R_o^+)^d$, where $|\cdot|$ denotes the component-wise absolute value [for example, $|(-1, 2, -3)| = (1, 2, 3)$]. In this case, $h^{-1}(\mathbf{x}) = \{\mathbf{z} : |\mathbf{z}| = \mathbf{x}\}$ contains 2^d elements. In these notes, we will not pursue this concept of complete/observewd data.

The EM algorithm works iteratively by alternatingly applying two steps: the E-Step (*expectation*) and the M-Step (*maximization*). Formally, let $\widehat{\boldsymbol{\theta}}^{(t)}$, for $t = 0, 1, 2, ...$, denote the successive parameter estimates; the E and M steps are defined as:

E-Step: Compute the conditional expectation (with respect to the missing $\mathbf{y}$) of the logarithm of the complete *a posteriori* probability function,

$\log p(\mathbf{y}, \boldsymbol{\theta}|\mathbf{x})$, given the observed data $\mathbf{x}$ and the current parameter estimate $\widehat{\boldsymbol{\theta}}^{(n)}$ (usually called the *Q-function*):

$$\begin{aligned} Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(t)}) &\equiv E[\log p(\mathbf{y}, \boldsymbol{\theta}|\mathbf{x})|\mathbf{x}, \widehat{\boldsymbol{\theta}}^{(t)}] \\ &\propto \log p(\boldsymbol{\theta}) + E[\log p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta})|\mathbf{x}, \widehat{\boldsymbol{\theta}}^{(t)}] \\ &= \log p(\boldsymbol{\theta}) + \int p(\mathbf{y}|\mathbf{x}, \widehat{\boldsymbol{\theta}}^{(t)}) \log p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta})\, d\mathbf{y}. \end{aligned} \tag{2}$$

M-Step: Update the parameter estimate according to

$$\widehat{\boldsymbol{\theta}}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(t)}). \tag{3}$$

The process continues until some stopping criterion is met.

Several generalizations, particularizations, and accelerated versions of the EM algorithm have been proposed; see [24], [34], [38], [40], [51] and references therein.

2 Monotonicity

Consider the function $\mathcal{E}(\boldsymbol{\theta}) : I\!R^p \rightarrow I\!R$, whose maximum with respect to $\boldsymbol{\theta}$ is sought; this could be $\log p(\mathbf{x}|\boldsymbol{\theta})$, in the maximum likelihood case, $\log p(\mathbf{x}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$, for MAP estimation. The EM algorithm can be shown to monotonically increase $\mathcal{E}(\boldsymbol{\theta})$, *i.e.*, the sequence of estimates verifies $\mathcal{E}(\widehat{\boldsymbol{\theta}}^{(t+1)}) \geq \mathcal{E}(\widehat{\boldsymbol{\theta}}^{(t)})$. This is a well-known result which was first proved in [9]. The monotonicity conditions for generalized and modified versions of EM were further studied in [55]. A simple and elegant proof recently proposed in [6] (which we will review below) sees EM under a new light that opens the door to extensions and generalizations: EM belongs to a class of iterative methods called *proximal point algorithms* (PPA). A related earlier result, although without identifying EM as a PPA, can be found in [43].

A (generalized) PPA is defined by the iteration

$$\widehat{\boldsymbol{\theta}}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \left\{ \mathcal{E}(\boldsymbol{\theta}) - \beta_t\, d(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(t)}) \right\}, \tag{4}$$

where β_t is a sequence of positive numbers and $d(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(t)})$ is a penalty function verifying $d(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(t)}) \geq 0$ and $d(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(t)}) = 0$ if and only if $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(t)}$). The original PPA, with $d(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(t)}) = \| \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}^{(t)} \|^2$, was proposed and studied in [36]

and [48]. Generalized versions, with penalty functions other than $\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}^{(t)}\|^2$, have been considered by several authors. For an introduction to PPA which includes a comprehensive set of pointers to the literature, see [1] (Chapter 5).

Monotonicity of PPA iterations is a trivial consequence of the monotonicity of the penalty function. From the definition of the iteration in Eq. (4), we have

$$\mathcal{E}(\widehat{\boldsymbol{\theta}}^{(t+1)}) - \beta_t\, d[\widehat{\boldsymbol{\theta}}^{(t+1)}, \widehat{\boldsymbol{\theta}}^{(t)}] \geq \mathcal{E}(\widehat{\boldsymbol{\theta}}^{(t)}), \tag{5}$$

because, by definition, the maximum of $\mathcal{E}(\boldsymbol{\theta}) - \beta_t\, d[\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(t)}]$ is at $\widehat{\boldsymbol{\theta}}^{(t+1)}$, and $d[\widehat{\boldsymbol{\theta}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}] = 0$. Consequently,

$$\mathcal{E}(\widehat{\boldsymbol{\theta}}^{(t+1)}) - \mathcal{E}(\widehat{\boldsymbol{\theta}}^{(t)}) \geq \beta_t\, d[\widehat{\boldsymbol{\theta}}^{(t+1)}, \widehat{\boldsymbol{\theta}}^{(t)}] \geq 0,$$

since $d(\widehat{\boldsymbol{\theta}}^{(t+1)}, \widehat{\boldsymbol{\theta}}^{(t)}) \geq 0$, which establishes that $\{\mathcal{E}(\widehat{\boldsymbol{\theta}}^{(t)}),\ t = 0, 1, 2, ...\}$ is a non-decreasing sequence.

Another (equivalent) view of EM, sees it as a so-called *bound optimization algorithm* (BOA) [27]. Behind the monotonicity of EM is the following fundamental property of the Q-function: the difference $\mathcal{E}(\boldsymbol{\theta}) - Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(t)})$ attains its minimum for $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(t)}$. This can be seen from

$$\begin{aligned}
\mathcal{E}(\widehat{\boldsymbol{\theta}}^{(t+1)}) &= \mathcal{E}(\widehat{\boldsymbol{\theta}}^{(t+1)}) - Q(\widehat{\boldsymbol{\theta}}^{(t+1)}|\widehat{\boldsymbol{\theta}}^{(t)}) + Q(\widehat{\boldsymbol{\theta}}^{(t+1)}|\widehat{\boldsymbol{\theta}}^{(t)}) \\
&\geq \mathcal{E}(\widehat{\boldsymbol{\theta}}^{(t)}) - Q(\widehat{\boldsymbol{\theta}}^{(t)}|\widehat{\boldsymbol{\theta}}^{(t)}) + Q(\widehat{\boldsymbol{\theta}}^{(t+1)}|\widehat{\boldsymbol{\theta}}^{(t)}) \\
&\geq \mathcal{E}(\widehat{\boldsymbol{\theta}}^{(t)}) - Q(\widehat{\boldsymbol{\theta}}^{(t)}|\widehat{\boldsymbol{\theta}}^{(t)}) + Q(\widehat{\boldsymbol{\theta}}^{(t)}|\widehat{\boldsymbol{\theta}}^{(t)}) \\
&= \mathcal{E}(\widehat{\boldsymbol{\theta}}^{(t)}),
\end{aligned} \tag{6}$$

where the first inequality is due to the fact that $\mathcal{E}(\boldsymbol{\theta}) - Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(t)})$ attains its minimum for $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$, and the second inequality results from the fact that $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(t)})$ attains its maximum for $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(t+1)}$. It is clear that this view is equivalent to the PPA interpretation; since $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(t)}) = \mathcal{E}(\boldsymbol{\theta}) - \beta_t\, d[\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(t)}]$, we have that

$$\mathcal{E}(\boldsymbol{\theta}) - Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(t)}) = \beta_t\, d[\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(t)}]$$

which, by construction, attains its minimum for $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(t)}$.

It turns out that EM is a PPA (as shown in the next paragraph) with

$\mathcal{E}(\boldsymbol{\theta}) = \log p(\mathbf{x}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\mathbf{x})$, $\beta_t = 1$, and

$$\begin{aligned} d(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(t)}) &= \mathcal{D}_{\mathrm{KL}}\left[p(\mathbf{y}|\mathbf{x}, \widehat{\boldsymbol{\theta}}^{(t)}) \parallel p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\right] && (7) \\ &= \int p(\mathbf{y}|\mathbf{x}, \widehat{\boldsymbol{\theta}}^{(t)}) \log \frac{p(\mathbf{y}|\mathbf{x}, \widehat{\boldsymbol{\theta}}^{(t)})}{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})}\, d\mathbf{y} && (8) \\ &= E\left[\log \frac{p(\mathbf{y}|\mathbf{x}, \widehat{\boldsymbol{\theta}}^{(t)})}{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})} \middle| \mathbf{x}, \widehat{\boldsymbol{\theta}}^{(t)}\right] && (9) \end{aligned}$$

is the Kullback-Leibler divergence between $p(\mathbf{y}|\mathbf{x}, \widehat{\boldsymbol{\theta}}^{(t)})$ and $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ [4], [6]. Since the Kullback-Leibler divergence satisfies the conditions $d(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(t)}) \geq 0$ and $d(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(t)}) = 0$ if and only if $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(t)}$ (see, for example, [7]), monotonicity of EM results immediately.

Let us now confirm that EM is a PPA with the choices expressed in the previous paragraph. This can be done by writing Eq. (4) with the choices mentioned in the previous paragraph (assuming a flat prior, for simplicity) and dropping all terms that do not depend on $\boldsymbol{\theta}$:

$$\begin{aligned} \widehat{\boldsymbol{\theta}}^{(t+1)} &= \arg\max_{\boldsymbol{\theta}} \left\{ \log p(\boldsymbol{\theta}|\mathbf{x}) - E\left[\log \frac{p(\mathbf{y}|\mathbf{x}, \widehat{\boldsymbol{\theta}}^{(t)})}{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})} \middle| \mathbf{x}, \widehat{\boldsymbol{\theta}}^{(t)}\right] \right\} \\ &= \arg\max_{\boldsymbol{\theta}} \left\{ \log p(\boldsymbol{\theta}) + \log p(\mathbf{x}|\boldsymbol{\theta}) + E\left[\log p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) | \, \mathbf{x}, \widehat{\boldsymbol{\theta}}^{(t)}\right] \right\} \\ &= \arg\max_{\boldsymbol{\theta}} \{ \underbrace{\log p(\boldsymbol{\theta}) + E\left[\log p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}) | \, \mathbf{x}, \widehat{\boldsymbol{\theta}}^{(t)}\right]}_{Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(t)})} \} \qquad (10) \end{aligned}$$

(compare Eq. (10) with Eq. (2)).

The monotonicity of EM makes it dependent on initialization. In other words, if the objective function $p(\boldsymbol{\theta}|\mathbf{x})$ is not concave (*i.e.*, has several local maxima) EM converges to a local maximum which depends on its starting point. In problems with multimodal objective functions, strategies have to be devised to address this characteristic of EM.

In the so-called *generalized* EM algorithm (GEM), $\widehat{\boldsymbol{\theta}}^{(t+1)}$ is chosen not necessarily as a maximum of $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(t)})$, but simply as verifying

$$Q(\widehat{\boldsymbol{\theta}}^{(t+1)}|\widehat{\boldsymbol{\theta}}^{(t)}) \geq Q(\widehat{\boldsymbol{\theta}}^{(t)}|\widehat{\boldsymbol{\theta}}^{(t)}).$$

Clearly, the proof of monotonicity of EM can be extended to GEM because the inequality in Eq. (5) also applies to GEM iterations.

3 Convergence to a Stationary Point

Of course monotonicity is not a sufficient condition for convergence to a (maybe local) maximum of $p(\boldsymbol{\theta}|\mathbf{x})$. Proving this convergence requires further smoothness conditions on $\log p(\boldsymbol{\theta}|\mathbf{x})$ and $\mathcal{D}_{\mathrm{KL}}\left[p(\mathbf{y}|\mathbf{x},\widehat{\boldsymbol{\theta}}^{(t)}) \parallel p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})\right]$ with respect to $\boldsymbol{\theta}$ (see [6], [51]). At a limit fixed point of the EM/PPA algorithm, *i.e.*, when $t \to \infty$, we have

$$\widehat{\boldsymbol{\theta}}^{(\infty)} = \arg\max_{\boldsymbol{\theta}} \left\{ \log p(\boldsymbol{\theta}|\mathbf{x}) - \mathcal{D}_{\mathrm{KL}}\left[p(\mathbf{y}|\mathbf{x},\widehat{\boldsymbol{\theta}}^{(\infty)}) \parallel p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})\right] \right\};$$

since both terms are smooth with respect to $\boldsymbol{\theta}$, it turns out $\widehat{\boldsymbol{\theta}}^{(\infty)}$ must be a stationary point, that is[1]

$$\left.\frac{\partial \log p(\boldsymbol{\theta}|\mathbf{x})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(\infty)}} - \left.\frac{\partial \mathcal{D}_{\mathrm{KL}}\left[p(\mathbf{y}|\mathbf{x},\widehat{\boldsymbol{\theta}}^{(\infty)}) \parallel p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})\right]}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(\infty)}} = 0.$$

Since the Kullback-Leibler divergence $\mathcal{D}_{\mathrm{KL}}[p(\mathbf{y}|\mathbf{x},\widehat{\boldsymbol{\theta}}^{(\infty)}) \parallel p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})]$ has a minimum (which is zero) for $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(\infty)}$, its partial derivative with respect to $\boldsymbol{\theta}$, at $\widehat{\boldsymbol{\theta}}^{(\infty)}$, is zero. We can can then conclude that the fixed points of the EM/PPA algorithm are in fact stationary points of the marginal log-likelihood function $\log p(\boldsymbol{\theta}|\mathbf{x})$.

4 The EM Algorithm for Exponential Families

The EM algorithm becomes particularly simple in the case where the complete data probability density function belongs to the exponential family [3]. When that is the case, we can write

$$\begin{aligned} p(\mathbf{y},\mathbf{x}|\boldsymbol{\theta}) &= \phi(\mathbf{y},\mathbf{x})\,\psi(\boldsymbol{\xi}(\boldsymbol{\theta}))\,\exp\{\boldsymbol{\xi}(\theta)^T\mathbf{t}(\mathbf{y},\mathbf{x})\} \\ &= \phi(\mathbf{z})\,\psi(\boldsymbol{\xi}(\boldsymbol{\theta}))\,\exp\left\{\sum_{j=1}^{k}\xi_j(\theta)t_j(\mathbf{z})\right\}, \end{aligned} \tag{11}$$

where $\mathbf{z} = (\mathbf{y},\mathbf{x})$ denotes the complete data, $\mathbf{t}(\mathbf{y},\mathbf{x}) = \mathbf{t}(\mathbf{z}) = [t_1(\mathbf{z}),...,t_k(\mathbf{z})]^T$ is the (k-dimensional) vector of sufficient statistics and $\boldsymbol{\xi}(\boldsymbol{\theta})$ the (also k-dimensional) *natural* (or *canonical*) parameter. Let us write the E-step in

[1] Here, we are using $\frac{\partial}{\partial\boldsymbol{\theta}}$ to denote gradient with respect to $\boldsymbol{\theta}$, in case it is multidimensional.

Eq. (2) for this case, with respect to the canonical parameter (for which we are assuming a flat prior; generalization to any prior $p(\boldsymbol{\xi})$ is trivial)

$$\begin{aligned} Q(\boldsymbol{\xi}|\widehat{\boldsymbol{\xi}}^{(t)}) &= \int p(\mathbf{y}|\mathbf{x},\widehat{\boldsymbol{\xi}}^{(t)}) \left[\log \phi(\mathbf{z}) + \log \psi(\boldsymbol{\xi}) + \boldsymbol{\xi}^T \mathbf{t}(\mathbf{z})\right] d\mathbf{y} \\ &= \underbrace{\int p(\mathbf{y}|\mathbf{x},\widehat{\boldsymbol{\xi}}^{(t)}) \log \phi(\mathbf{z})\, d\mathbf{y}}_{\text{independent of } \boldsymbol{\xi}} + \underbrace{\log \psi(\boldsymbol{\xi})}_{\substack{\text{Gibbs free} \\ \text{energy)}}} + \boldsymbol{\xi}^T E\left[\mathbf{t}(\mathbf{z})|\,\mathbf{x},\widehat{\boldsymbol{\xi}}^{(t)}\right]. \end{aligned}$$

That is, the E-step consists in computing the conditional expectation of the sufficient statistics, given the observed data and the current estimate of the parameter. To find $\widehat{\boldsymbol{\xi}}^{(t+1)}$ we have to set the gradient of $Q(\boldsymbol{\xi}|\widehat{\boldsymbol{\xi}}^{(t)})$ with respect to $\boldsymbol{\xi}$ to zero, which leads to the following update equation:

$$\widehat{\boldsymbol{\xi}}^{(t+1)} = \text{Solution w.r.t. } \boldsymbol{\xi} \text{ of} \left(-\frac{\partial \log \psi(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} = E\left[\mathbf{t}(\mathbf{y},\mathbf{x})|\,\mathbf{x},\widehat{\boldsymbol{\xi}}^{(t)}\right]\right). \quad (12)$$

Now, recall that (see [3])

$$-\frac{\partial \log \psi(\boldsymbol{\xi})}{\partial \xi_i} = \frac{\partial \log Z(\boldsymbol{\xi})}{\partial \xi_i} = E\left[t_i(\mathbf{y},\mathbf{x})|\boldsymbol{\xi}\right],$$

which means that the M-step is equivalent to solving, with respect to $\boldsymbol{\xi}$, the following set of equations

$$E\left[t_i(\mathbf{y},\mathbf{x})|\boldsymbol{\xi}\right] = E\left[t_i(\mathbf{y},\mathbf{x})|\,\mathbf{x},\widehat{\boldsymbol{\xi}}^{(t)}\right], \quad \text{for } i = 1,...,k.$$

In words, the new estimate of $\boldsymbol{\xi}$ is such that the (unconditional) expected value of the sufficient statistics coincides with the conditional expected value of the sufficient statistics, given the observations and the current parameter estimate $\widehat{\boldsymbol{\xi}}^{(t)}$.

5 Examples

5.1 Gaussian variables with Gaussian noise

Let us consider a set of n real-valued previous observations $\mathbf{g} = \{g_1,...,g_n\}$, which are noisy versions of (unobserved) $\mathbf{y} = \{f_1,...,f_n\}$; the noise has zero mean and (known) variance σ^2. Assume prior knowledge about the f_i's is expressed via a common Gaussian probability density function $\mathcal{N}(f_i|\mu,\tau^2)$,

with μ and τ^2 both unknown. The goal is to perform inference about τ^2 and μ.

To implement the marginal maximum likelihood (MML) criterion we have to compute the marginal

$$\begin{aligned} p(\mathbf{g}|\mu,\tau^2) &= \prod_{i=1}^{n} \int p\left(f_i, g_i|\mu,\tau^2\right) df_i \\ &\propto \prod_{i=1}^{n} \int p\left(g_i|f_i,\sigma^2\right) p(f_i|\mu,\tau^2) \\ &\propto \prod_{i=1}^{n} \mathcal{N}(g_i|\mu,\tau^2+\sigma^2) \end{aligned}$$

which is a natural result since the mean of the g_i is of course μ, and its variance $\tau^2+\sigma^2$. In this case, MML estimates are simply

$$\widehat{\mu} = \frac{1}{n}\left(\sum_{j=1}^{n} g_j\right) \tag{13}$$

$$\widehat{\tau^2} = \left(\frac{1}{n}\left(\sum_{j=1}^{n}(g_j-\widehat{\mu})^2\right)-\sigma^2\right)_+, \tag{14}$$

where $(\cdot)_+$ is the positive part operator defined as $x_+ = x$, if $x \geq 0$, and $x_+ = 0$, if $x < 0$.

We chose this as the first example of EM because it involves a complete likelihood function in exponential form, and because we have simple exact expressions for the MML estimates of the parameters. Let us now see how the EM algorithm can be used to obtain these estimates. Of course, in practice, there would be no advantage in using EM instead of the simple closed form expressions. Only in cases where such expressions can not be obtained is the use of EM necessary.

The complete likelihood function is

$$p(\mathbf{y},\mathbf{g}|\mu,\tau^2) = \prod_{i=1}^{n} \mathcal{N}(g_i|f_i,\sigma^2)\,\mathcal{N}(f_i|\mu,\tau^2) \tag{15}$$

Since we are considering σ^2 as known, we can omit it from the parameteri-

zation and write this likelihood function in exponential form as

$$
\begin{aligned}
p(\mathbf{y},\mathbf{g}|\boldsymbol{\xi}) &= \underbrace{\frac{1}{(2\pi\sigma)^n}\exp\left\{-\sum_{i=1}^{n}\frac{(f_i-g_i)^2}{2\sigma^2}\right\}}_{\phi(\mathbf{y},\mathbf{g})} \\
&\times \underbrace{\left(\frac{1}{\tau^2}\right)^{\frac{n}{2}}\exp\left\{-\frac{n\mu^2}{2\tau^2}\right\}}_{\psi(\boldsymbol{\xi})} \\
&\times \underbrace{\exp\left\{\frac{\mu}{\tau^2}\sum_{i=1}^{n}f_i-\frac{1}{\tau^2}\sum_{i=1}^{n}\frac{f_i^2}{2}\right\}}_{\exp\{\boldsymbol{\xi}^T\mathbf{t}(\mathbf{y},\mathbf{g})\}} \qquad (16)
\end{aligned}
$$

where $\boldsymbol{\xi} = [\xi_1, \xi_2]^T = [\mu/\tau^2, 1/\tau^2]^T$ is the vector of canonical parameters, and

$$
\mathbf{t}(\mathbf{y},\mathbf{g}) = \begin{bmatrix} t_1(\mathbf{y},\mathbf{g}) \\ t_2(\mathbf{y},\mathbf{g}) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} f_i \\ -\sum_{i=1}^{n}\frac{f_i^2}{2} \end{bmatrix}
$$

is the vector of sufficient statistics.

As seen above, the E-step involves computing the conditional expectation of the sufficient statistics, given the observations and the current parameter estimate $\widehat{\boldsymbol{\xi}}^{(t)}$. Due to the independence expressed in Eq. (15), it is clear that

$$
p(\mathbf{y}|\mathbf{g},\sigma^2,\widehat{\mu}^{(t)},\widehat{\tau^2}^{(t)}) = \prod_{i=1}^{n} p(f_i|g_i,\sigma^2,\widehat{\mu}^{(t)},\widehat{\tau^2}^{(t)});
$$

It is a simple task to show that

$$
p(f_i|g_i,\sigma^2,\widehat{\mu}^{(t)},\widehat{\tau^2}^{(t)}) = \mathcal{N}\left(f_i \,\middle|\, \frac{\widehat{\mu}^{(t)}\sigma^2 + g_i\widehat{\tau^2}^{(t)}}{\sigma^2+\widehat{\tau^2}^{(t)}},\ \frac{\sigma^2\widehat{\tau^2}^{(t)}}{\sigma^2+\widehat{\tau^2}^{(t)}}\right).
$$

Using this result, we can write the conditional expectations of the sufficient

statistics, which is all that is required in the E-step:

$$
\begin{aligned}
E\left[t_1(\mathbf{y},\mathbf{g})\,\middle|\,\mathbf{g},\widehat{\boldsymbol{\xi}}^{(t)}\right] &= \frac{n}{\sigma^2+\widehat{\tau^2}^{(t)}}\left(\widehat{\mu}^{(t)}\sigma^2+\frac{\widehat{\tau^2}^{(t)}}{n}\sum_{i=1}^{n}g_i\right)\\
E\left[t_2(\mathbf{y},\mathbf{g})\,\middle|\,\mathbf{g},\widehat{\boldsymbol{\xi}}^{(t)}\right] &= -\frac{n\sigma^2\widehat{\tau^2}^{(t)}}{2(\sigma^2+\widehat{\tau^2}^{(t)})}-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{\widehat{\mu}^{(t)}\sigma^2+g_i\widehat{\tau^2}^{(t)}}{\sigma^2+\widehat{\tau^2}^{(t)}}\right)^2.
\end{aligned}
$$

Finally, to obtain the updated parameters, we have to solve Eq. (12). From Eq. (16),

$$
\begin{aligned}
-\frac{\partial\log\psi(\boldsymbol{\xi})}{\partial\xi_1} &= \frac{n\,\xi_1}{\xi_2} = n\mu\\
-\frac{\partial\log\psi(\boldsymbol{\xi})}{\partial\xi_2} &= -\frac{n}{2}\left(\frac{1}{\xi_2}+\frac{\xi_1^2}{\xi_2^2}\right)=-\frac{n}{2}\left(\tau^2+\mu^2\right), \qquad (17)
\end{aligned}
$$

leading to the following updated estimates

$$
\begin{aligned}
\widehat{\mu}^{(t+1)} &= \frac{1}{\sigma^2+\widehat{\tau^2}^{(t)}}\left(\widehat{\mu}^{(t)}\sigma^2+\frac{\widehat{\tau^2}^{(t)}}{n}\sum_{i=1}^{n}g_i\right)\\
\widehat{\tau^2}^{(t+1)} &= \left(\frac{\sigma^2\widehat{\tau^2}^{(t)}}{\sigma^2+\widehat{\tau^2}^{(t)}}-\left(\widehat{\mu}^{(t+1)}\right)^2+\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\widehat{\mu}^{(t)}\sigma^2+g_i\widehat{\tau^2}^{(t)}}{\sigma^2+\widehat{\tau^2}^{(t)}}\right)^2\right)_+.
\end{aligned}
$$

where the positive part operator $(\cdot)_+$ appears because we are under the constraint $\tau^2 \geq 0$. Notice that the update equation for $\widehat{\mu}^{(t+1)}$ has a simple meaning: it is a weighted average of the previous estimate $\widehat{\mu}^{(t)}$ and the mean of the observations.

In Figure 1 we plot the successive estimates $\widehat{\mu}^{(t)}$ and $\widehat{\tau}^{(t)}$ obtained with a size $n = 20$ sample with $\mu = 1$, $\tau = 1$, and $\sigma = 0.5$. The horizontal dashed line represents the exact MML estimate given by Eqs. (13) and (14). Observe how the EM estimates converge fast to these exact estimates. The EM algorithm was initialized with $\widehat{\mu}^{(0)} = 0$ and $\widehat{\tau^2}^{(0)}$ set to the sample variance of the observations. In Figure 2 we repeat the experiment, now with $\sigma = 2.5$; observe how the convergence to the MML solutions is slower (particularly for the estimate of τ).

5.2 Laplacian priors

Let us consider a simple Gaussian likelihood for a scalar observation g,

$$
p(g|f) = \mathcal{N}(g|f,\sigma^2) \qquad (18)
$$

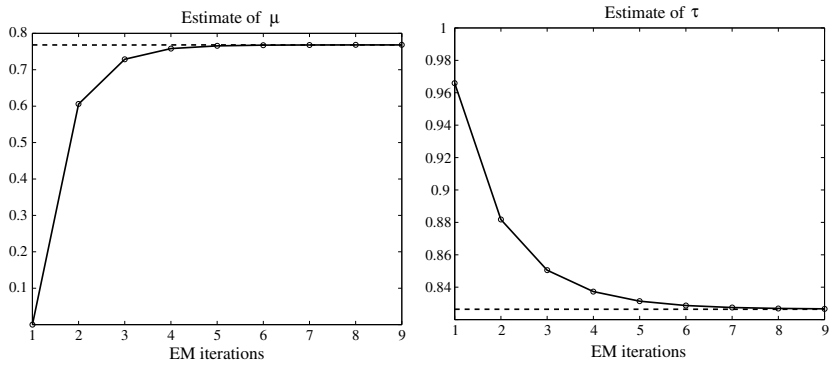

Figure 1: Evolution of the EM estimates of μ and τ (for $\mu = 1$, $\tau = 1$, and $\sigma = 0.5$); the horizontal dashed line represents the exact MML estimate given by Eqs. (13) and (14).

(where σ^2 is known) and the goal of estimating f. Now consider a zero-mean Gaussian prior on f, but with unknown variance τ^2, *i.e.*, $p(f|\tau^2) = \mathcal{N}(f|0, \tau^2)$. To complete the hierarchical Bayes model, we need a hyper-prior; let us consider an exponential density

$$p(\tau^2) = \frac{1}{\eta} \exp\{-\frac{\tau^2}{\eta}\}, \tag{19}$$

for $\tau^2 \geq 0$, where η is a (known) hyper-parameter. This choice can be justified by invoking that it is the maximum entropy prior for a positive parameter with a given mean (the mean is η).

The set of unknowns is (f, τ^2), and the a posteriori probability density function is

$$p(f, \tau^2|g) \propto p(g|f)\, p(f|\tau^2)\, p(\tau^2) \tag{20}$$

(as usual, we omit known parameters). The marginal posterior on f is

$$p(f|g) = \int_0^\infty p(f, \tau^2|g)\, d\tau^2 \tag{21}$$

$$\propto p(g|f) \underbrace{\int_0^\infty p(f|\tau^2)\, p(\tau^2)\, d\tau^2}_{\text{equivalent prior } p(f)}. \tag{22}$$

The equivalent prior $p(f)$, obtained by integrating with respect to τ^2, turns out to be

$$p(f) = \frac{1}{\sqrt{2\eta}} \exp\{-\sqrt{\frac{2}{\eta}}\, |f|\}, \tag{23}$$

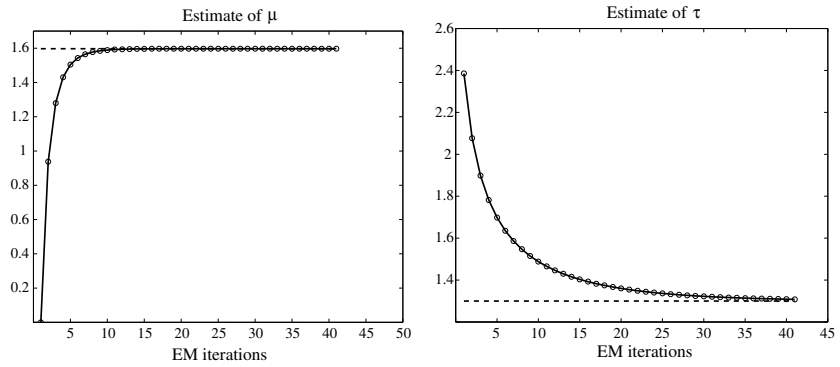

Figure 2: Evolution of the EM estimates of μ and τ (for $\mu = 1$, $\tau = 1$, and $\sigma = 2.5$); the horizontal dashed line represents the exact MML estimate given by Eqs. (13) and (14). Observe the slower convergence rate, when compared with the plots in Figure 1.

which is called a Laplacian density and plotted in Fig. 3 for several values of the parameter η. Since the likelihood $p(g|f)$ is Gaussian, the MAP estimate is simply

$$\widehat{f}_{\text{MAP}} = \arg\min_{f} \left\{ \frac{1}{2\sigma^2}(f-g)^2 + \sqrt{\frac{2}{\eta}}\,|f| \right\} \tag{24}$$

the solution of which is

$$\widehat{f}_{\text{MAP}} = \text{sign}(g) \left(|g| - \sigma^2 \sqrt{\frac{2}{\eta}} \right)_+ . \tag{25}$$

In the previous equation, $(\cdot)_+$ denotes the "the positive part" operator which is defined as $x_+ = x$, if $x \geq 0$, and $x_+ = 0$, if $x < 0$. The notation $\text{sign}(g)$ stands for the sign of g, *i.e.*, $\text{sign}(g) = 1$, if $g > 0$, while $\text{sign}(g) = -1$ when $g < 0$. The function in Eq. (25), which known as the soft-threshold, was proposed in [11] and [12] to perform wavelet-based signal denoising/estimation (see also [44]). For obvious reasons, the quantity $t_{\text{MAP}} \equiv \sigma^2\sqrt{\frac{2}{\eta}}$ is called the threshold.

In the multivariate case, we have the likelihood function for the observed vector $\mathbf{g}$

$$p(\mathbf{g}|\mathbf{y}) = \mathcal{N}(\mathbf{g}|\mathbf{H}\mathbf{y}, \sigma^2\mathbf{I}) \tag{26}$$

where σ^2 is known, $\mathbf{H}$ is a matrix, and the goal is to estimate $\mathbf{y}$ (which is d-dimensional) from $\mathbf{g}$ (which is n-dimensional). As above, we take a Gaussian

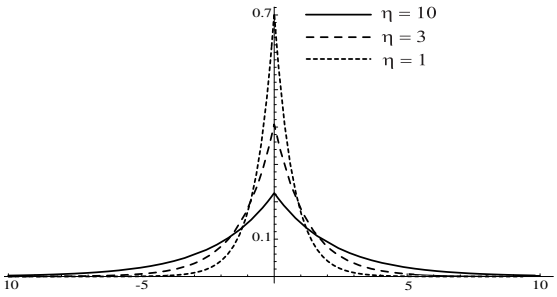

Figure 3: Laplacian probability density function (defined in Eq. (23))for several values of the parameter η.

prior for $\mathbf{y}$ under which the f_i's are assumed independent, zero-mean, and with (unknown) variances τ_i^2. That is,

$$p(\mathbf{y}|\boldsymbol{\tau}) = \mathcal{N}(\mathbf{y}|0, \boldsymbol{\tau}) \tag{27}$$

where $\boldsymbol{\tau} = \text{diag}\{\tau_1^2, ..., \tau_d^2\}$. The hierarchical Bayes model is completed with a set of independent exponential hyper-priors

$$p(\boldsymbol{\tau}) = \prod_{i=1}^{d} \frac{1}{\eta} \exp\{-\frac{\tau_i^2}{\eta}\} = \frac{1}{\eta^d} \exp\{-\frac{1}{\eta} \sum_{i=1}^{d} \tau_i^2\}, \tag{28}$$

for $\tau_i^2 \geq 0$, where η is a (known) hyper-parameter. The idea of placing independent Gaussian priors, with independent variances, for each f_i is related to the so-called *automatic relevance determination* (ARD) approach described in [35] and [42]; however, in ARD, there is no hyper-prior for these variances (or in other words, there is a flat hyper-prior).

The MAP estimate of $\mathbf{y}$ is given by

$$\widehat{\mathbf{y}}_{\text{MAP}} = \arg\max_{\mathbf{y}} \int p(\mathbf{g}|\mathbf{y})\, p(\mathbf{y}|\boldsymbol{\tau})\, p(\boldsymbol{\tau})\, d\boldsymbol{\tau}. \tag{29}$$

Since both the likelihoods and the prior $p(\mathbf{y}|\boldsymbol{\tau})$ are factorized, this integration can be performed separately with respect to each τ_i^2 leading to

$$\widehat{\mathbf{y}}_{\text{MAP}} = \arg\min_{\mathbf{y}} \left\{ \parallel \mathbf{Hf} - \mathbf{g} \parallel_2 + \sqrt{\frac{8\,\sigma^4}{\eta}} \parallel \mathbf{y} \parallel_1, \right\} \tag{30}$$

In the previous equation, $\| \cdot \|_2$ denotes the standard Euclidean norm, while $\| \cdot \|_1$ stands for the L_1 norm, *i.e.*, $\| \mathbf{x} \|_1 = \sum_i |x_i|$; $(\mathbf{Hf})_i$ stands for the

i-th component of vector $\mathbf{Hf}$. Notice that Eq. (30) is a multidimensional version of Eq. (24).

This type of mixed L_2/L_1 criterion has been suggested as a way of promoting *sparseness* of the estimate, *i.e.* to look for an estimate $\widehat{\mathbf{y}}$ such only a few of its components are different from zero (see [5], [13], [31], [46], [52], [54] and references therein). In some applications, the columns of matrix $\mathbf{H}$ are seen as an over-complete set of basis functions. The goal is then to obtain a representation of the observed vector $\mathbf{g}$ as a linear combination of those basis functions, $\mathbf{g} = \mathbf{Hf}$, under a sparseness constraint, that is, such that only a few elements of $\mathbf{y}$ are non-zero [5].

In the statistical regression literature, this criterion is known as the *lasso* (standing for *least absolute shrinkage and selection operator*). The *Lasso* approach was proposed in [52] (see also [15]) as a compromise solution between standard ridge regression (which results from a Gaussian prior on the coefficients) and subset selection (where the best set of regressors is chosen).

We will see how to use the EM algorithm to solve Eq. (30). Clearly, the unknown quantity is $\mathbf{y}$, the observed data is $\mathbf{g}$, and the missing data is $\boldsymbol{\tau}$. The complete log-posterior from which we could easily estimate $\mathbf{y}$ if $\boldsymbol{\tau}$ was observed is

$$\log p(\mathbf{y}|\boldsymbol{\tau}, \mathbf{g}) = \log p(\mathbf{y}, \boldsymbol{\tau}, \mathbf{g}) - \underbrace{\log p(\boldsymbol{\tau}, \mathbf{g})}_{\text{indep. of } \mathbf{y}} = \log \left(p(\mathbf{g}|\mathbf{y}) p(\mathbf{y}|\boldsymbol{\tau}) p(\boldsymbol{\tau}) \right) + K,$$

where K is some constant independent of $\mathbf{y}$. Inserting Eqs. (26), (27), and (28), and absorbing all terms that do not depend on $\mathbf{y}$ in constant K, we obtain

$$\log p(\mathbf{y}|\boldsymbol{\tau}, \mathbf{g}) = -\frac{1}{2\sigma^2} \parallel \mathbf{Hy} - \mathbf{g} \parallel_2^2 - \frac{1}{2} \sum_{i=1}^{d} \frac{f_i^2}{\tau_i^2} + K. \tag{31}$$

We can now write the E-step as the computation of the expected value (with respect to the missing data $\boldsymbol{\tau}$) of the right hand side of Eq. (31)

$$Q(\mathbf{y}|\widehat{\mathbf{y}}^{(t)}) = -\frac{1}{2\sigma^2} \parallel \mathbf{H}\widehat{\mathbf{y}}^{(t)} - \mathbf{g} \parallel_2^2 - \frac{1}{2} \sum_{i=1}^{d} f_i^2 E\left[\frac{1}{\tau_i^2} | \widehat{\mathbf{y}}^{(t)} \right]. \tag{32}$$

Again using the models in Eqs. (26), (27), and (28), it is easy to show (by applying Bayes law) that

$$p(\boldsymbol{\tau}|\widehat{\mathbf{y}}^{(t)}, \mathbf{g}) = \prod_{i=1}^{d} \left(\pi\eta\tau_i^2 \right)^{-\frac{1}{2}} \exp\left\{ \widehat{f}_i^{(t)} \left(\sqrt{\frac{2}{\eta}} - \frac{1}{\tau_i^2} \right) - \frac{\tau_i^2}{\eta} \right\}$$

and finally,

$$E\left[\frac{1}{\tau_i^2}|\widehat{\mathbf{y}}^{(t)}\right] = E\left[\frac{1}{\tau_i^2}|\widehat{f}_i^{(t)}\right] = \left|\widehat{f}_i^{(t)}\right|^{-1}\sqrt{\frac{2}{\eta}} \equiv w_i^{(t)}. \tag{33}$$

Letting $\mathbf{W}^{(t)} \equiv \text{diag}(w_1^{(t)}, w_2^{(t)}, ..., w_d^{(t)})$, we can write the M-step (*i.e.*, the maximization of Eq. (32) with respect to $\mathbf{y}$) as

$$\widehat{\mathbf{y}}^{(t+1)} = \arg\max_{\mathbf{y}} \left\{ \| \mathbf{H}\widehat{\mathbf{y}}^{(t)} - \mathbf{g} \|_2^2 + \sigma^2 \mathbf{y}^T \mathbf{W}^{(t)} \mathbf{y} \right\},$$

which has a simple analytical solution

$$\widehat{\mathbf{y}}^{(t+1)} = \left[\sigma^2 \mathbf{W}^{(t)} + \mathbf{H}^T\mathbf{H}\right]^{-1} \mathbf{H}^T\mathbf{g}. \tag{34}$$

Summarizing, the E-step consists in computing the diagonal matrix $\mathbf{W}^{(t)}$ whose elements are given by Eq. (33), and the M-step updates the estimate of $\mathbf{y}$ following Eq. (34).

Let us conclude by showing a simple regression example, using radial basis functions (RBF). Consider a set of d (non-orthogonal) Gaussian-shaped functions

$$\phi_i(x) = \exp\left\{ -\left(\frac{x - c_i}{h}\right)^2 \right\}, \quad \text{for } i = 1, 2, ..., d;$$

the goal is to approximate an unknown function $g = \psi(x)$, of which we are given a set of noisy observations $\{(x_1, g_1), ..., (x_n, g_n)\}$, as a linear combination of those basis functions:

$$\widehat{\psi}(x) = \sum_{j=1}^{d} \widehat{f}_j \, \phi_j(x).$$

Since we do not have a generative model for the $\{x_i, \ i = 1, ..., n\}$, the observation model can be written as in Eq. (26) where $\mathbf{y} = [f_1, ..., f_d]^T$, $\mathbf{g} = [g_1, ..., g_n]^T$, and the element (i, j) of matrix $\mathbf{H}$ is given by $H_{i,j} = \phi_j(x_i)$. In this example, we consider a true function that can actually be written as $\psi(x) = \sum_{j=1}^{d} f_j \, \phi_i(x)$, with $d = 40$, $f_2 = 2$, $f_{12} = -2$, $f_{24} = 3$, $f_{34} = -4$, and all the 36 remaining f_i's equal to zero. The centers of the basis function are fixed at $c_i = 5\,i$, for $i = 1, ..., 40$, and the width parameter h is set to 10. The resulting function is plotted in Figure 4 (a), together with 200 noisy observations placed at the integers $1, 2, ...200$, with noise standard deviation equal to 0.8. After running the EM algorithm from these noisy observations,

with $\eta = 0.15$, the function estimate is the one plotted in Figure 4 (b). For comparison, we also show a ridge estimate (also called *weight decay* in the neural networks literature) with a zero-mean unit-variance Gaussian prior on $\mathbf{y}$, and maximum likelihood estimate, both in Figure 4 (c); notice how these estimates are more affected by noise, specially in the flat regions of the true function. Finally, Figure 4 (d) shows the elements of $\widehat{\mathbf{y}}$ obtained under the lasso (circles) and ridge (stars) criteria; observe how the lasso estimate is dominated by a few large components, while the ridge estimate does not exhibit this sparse structure.

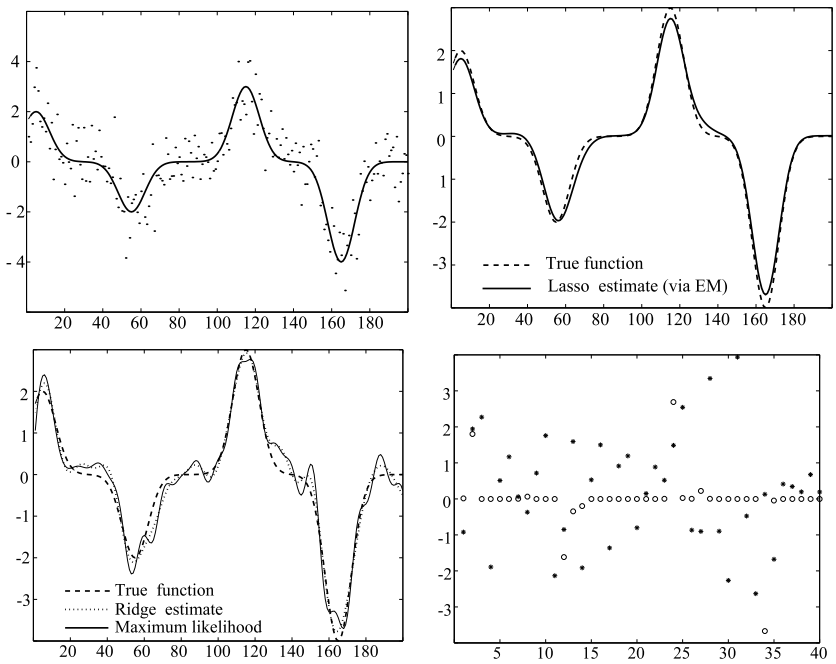

Figure 4: (a) True function (solid line) and observed noisy data (dots), for $\sigma = 0.8$. (b) Function estimate from the lasso criterion (solid line) and the original function. (c) Ridge and maximum likelihood estimates. (d) Elements of $\widehat{\mathbf{y}}$ for the lasso estimate (circles) and the ridge estimate (stars).

Finally, we mention that this approach can easily be extended to consider σ^2 as an unknown to be estimated from the data.

5.3 Laplacian noise: robust estimation

Let us consider the same type of hierarchical Bayes modelling but now focusing on the noise variance. Consider the observation model

$$g_i = f + u_i, \quad i = 1, ..., n,$$

where we have n noisy observations of the same unknown quantity f for which we have a flat prior $p(f) =$"constant". If the noise samples are independent, Gaussian, have zero mean, and common variance σ^2, the MAP/ML estimate of f is

$$\widehat{f}_{\text{LS}} = \arg\min_f \sum_{i=1}^{n} (g_i - f)^2 = \frac{1}{n} \sum_{i=1}^{n} g_i, \tag{35}$$

which is the well known *least squares* (LS) criterion. The main criticism on the LS approach is its lack of *robustness*, meaning that it is too sensitive to *outliers*. Outliers are observations that may not have been generated according to the assumed model (the typical example being a variance much larger than the assumed σ^2). Estimators that are designed to overcome this limitation are called *robust* (see, for example, [21]). One of the best known robust criteria is obtained by replacing the squared error in the LS formulation by the absolute error, leading to what is usually called the *least absolute deviation* (LAD) criterion

$$\widehat{f}_{\text{LAD}} = \arg\min_f \sum_{i=1}^{n} |g_i - f|; \tag{36}$$

see [2] for a textbook devoted to the LAD criterion. The solution to this optimization problem is well known. Let the set of sorted observations be denoted as $\{f_{s_1}, f_{s_2}, ..., f_{s_n}\}$, where $\{s_1, s_2, ..., s_n\}$ is a permutation of $\{1, 2, ..., n\}$ such that $f_{s_1} \leq f_{s_2} \leq ... \leq f_{s_n}$. Then, the solution of Eq. (36) is

$$\widehat{f}_{\text{LAD}} = \begin{cases} f_{s_{\frac{n+1}{2}}} & \Leftarrow \quad n \text{ is odd} \\ \text{any value in } \left[f_{s_{\frac{n}{2}}}, f_{s_{\frac{n}{2}+1}}\right] & \Leftarrow \quad n \text{ is even,} \end{cases} \tag{37}$$

or, in other words, $\widehat{f}_{\text{LAD}}$ is the *median* of the observation, that is, a value that has an equal number of larger and smaller observations than it. We now illustrate with a very simple example why this is called a robust estimator. Let the set of observations (after sorting) be $1, 2, 3, 4, 5, 6, 7, 8, 9$. Then, the mean and the median coincide, and both $\widehat{f}_{\text{LAD}}$ and $\widehat{f}_{\text{LS}}$ are equal

to 5. Now suppose we replace one of the observations with a strong outlier, $1, 2, 3, 4, 5, 6, 7, 9, 80$; in this case, $\widehat{f}_{\mathrm{LAD}}$ is still 5, while $\widehat{f}_{\mathrm{LS}}$ is now 13, highly affected by the outlier. Of course the outlier can affect the median; for example if the data is $-80, 1, 2, 3, 4, 5, 6, 7, 9$, $\widehat{f}_{\mathrm{LAD}}$ is now 4, but $\widehat{f}_{\mathrm{LS}}$ is (approximately) -4.7, thus much more sensitive to the presence of the outlier.

The very notion of outlier seems to point the way to a hierarchical Bayes formulation of the problem. Suppose that it is known that the noise is zero-mean Gaussian, but the specific variance of each noise sample u_i, denoted σ_i^2, is not known in advance. Then it is natural to include these variances in the set of unknowns (now $(f, \sigma_1^2, ..., \sigma_n^2)$) and provide a prior, which we assume to have the form $p(\sigma_1^2, ..., \sigma_n^2) = p(\sigma_1^2) \cdots p(\sigma_n^2)$. Since we are actually not interested in estimating the variances, the relevant posterior is

$$\begin{aligned} p(f|g_1, ..., g_n) &\propto \int_0^\infty \cdots \int_0^\infty p(f, \sigma_1^2, ..., \sigma_n^2|g_1, ..., g_n)\, d\sigma_1^2 \cdots d\sigma_n^2 \\ &\propto \int_0^\infty \cdots \int_0^\infty \left(\prod_{i=1}^n p(g_i|f, \sigma_i^2)\, p(\sigma_i^2) \right) d\sigma_1^2 \cdots d\sigma_n^2 \\ &= \prod_{i=1}^n \underbrace{\int p(g_i|f, \sigma_i^2)\, p(\sigma_i^2)\, d\sigma_i^2}_{\text{effective likelihood } p(g_i|f)} . \end{aligned} \tag{38}$$

If we adopt an exponential hyper-prior for the variance

$$p(\sigma_i^2) = \frac{1}{\eta} \exp\{-\frac{\sigma_i^2}{\eta}\}, \tag{39}$$

the resulting effective likelihood is

$$p(g_i|f) = \frac{1}{\sqrt{2\eta}} \exp\left\{ -\sqrt{\frac{2}{\eta}}\, |g_i - f| \right\}. \tag{40}$$

Then, the MAP/ML estimate of f is given by Eq. (36), regardless of the value of the hyper-parameter η. In conclusion, the LAD criterion can be interpreted as the MMAP/MML criterion resulting from a hierarchical Bayes formulation where the noise samples are assumed to have unknown independent variances with exponential (hyper)priors.

Of course the same line of thought can be followed in regression problems, where rather than estimating a scalar quantity f, the goal is to fit some function to observed data. Results along the same lines for other noise models can be found in [16] and [29].

The EM algorithm can also be used to do LAD regression by resorting to the hierarchical interpretation of the Laplacian density. In this case, the observation model is again linear

$$\mathbf{g} = \mathbf{H}\mathbf{f} + \mathbf{u} \tag{41}$$

(where matrix $\mathbf{H}$ is known), but now each element of the noise vector $\mathbf{u} = [u_1, ..., n_n]^T$ has its own variance, *i.e.*, $p(u_i|\sigma_i^2) = \mathcal{N}(u_i|0, \sigma_i^2)$, where $\boldsymbol{\sigma} \equiv [\sigma_1^2, ..., \sigma_n^2]^T$, and we have independent exponential hyper-priors for these variances,

$$p(\boldsymbol{\sigma}) = \frac{1}{\eta^n} \prod_{i=1}^{n} \exp\{-\frac{\sigma_i^2}{\eta}\}.$$

The model is completed with a flat prior for $\mathbf{f}$, *i.e.*, $p(\mathbf{f}) \propto$ "constant".

To adopt an EM approach, we interpret $\boldsymbol{\sigma}$ as missing data, $\mathbf{g}$ is of course the observed data, and the goal is to estimate $\mathbf{f}$. The complete log-likelihood from which we could easily estimate $\mathbf{f}$ if $\boldsymbol{\sigma}$ was observed is

$$\log p(\mathbf{g}, \boldsymbol{\sigma}|\mathbf{y}) = \log p(\mathbf{g}|\boldsymbol{\sigma}, \mathbf{g}) + \underbrace{\log p(\boldsymbol{\sigma})}_{\text{indep. of } \mathbf{y}} = -\frac{((\mathbf{H}\,\mathbf{f})_i - g_i)^2}{2\sigma_i^2} + K,$$

where $(\mathbf{H}\,\mathbf{f})_i$ denotes the i-th component of $\mathbf{H}\,\mathbf{f}$; constant K includes all terms that do not depend on $\mathbf{f}$. We can now write the E-step as

$$Q(\mathbf{f}|\widehat{\mathbf{f}}^{(t)}) = -\frac{1}{2}(\mathbf{H}\,\mathbf{f} - \mathbf{g})^T \mathbf{W}^{(t)} (\mathbf{H}\,\mathbf{f} - \mathbf{g}) + K \tag{42}$$

where $\mathbf{W}^{(t)} \equiv \text{diag}(w_1^{(t)}, w_2^{(t)}, ..., w_d^{(t)})$ and

$$w_i^{(t)} = E\left[\frac{1}{\sigma_i^2}|\mathbf{f}^{(t)}\right] = E\left[\frac{1}{\sigma_i^2}|f_i^{(t)}\right] = |(\mathbf{H}\,\mathbf{f})_i - g_i|^{-1}\sqrt{\frac{2}{\eta}}; \tag{43}$$

observe the similarity with Eq. (33). Maximizing Eq. (42) with respect to $\mathbf{y}$ leads to

$$\widehat{\mathbf{f}}^{(t+1)} = \left[\mathbf{H}^T \mathbf{W}^{(t)} \mathbf{H}\right]^{-1} \mathbf{H}^T \mathbf{W}^{(t)} \mathbf{g}, \tag{44}$$

which is a *weighted least squares* (WLS) estimate. Summarizing, the E-step consists in computing the diagonal matrix $\mathbf{W}^{(t)}$ whose elements are given by Eq. (43), and the M-step updates the estimate of $\mathbf{f}$ following Eq. (44). Since each M-step implements a WLS estimate, and the weighting is updated at each iteration, this EM algorithm can be considered an *iteratively*

reweighted least squares (IRLS) procedure [49], with a particular choice of the reweighting function.

We conclude with a simple illustration. Consider that the goal is to fit a straight line to the noisy data in Figure 5 (a); this data clearly has eight outliers. Still in Figure 5 (a) we show the standard least squares regression, clearly affected by the anomalous data. Figures 5 (b), (c), and (d) then show the results obtained with the EM algorithm just described, after 1, 4, and 10 iterations. Observe how the outliers are progressively ignored.

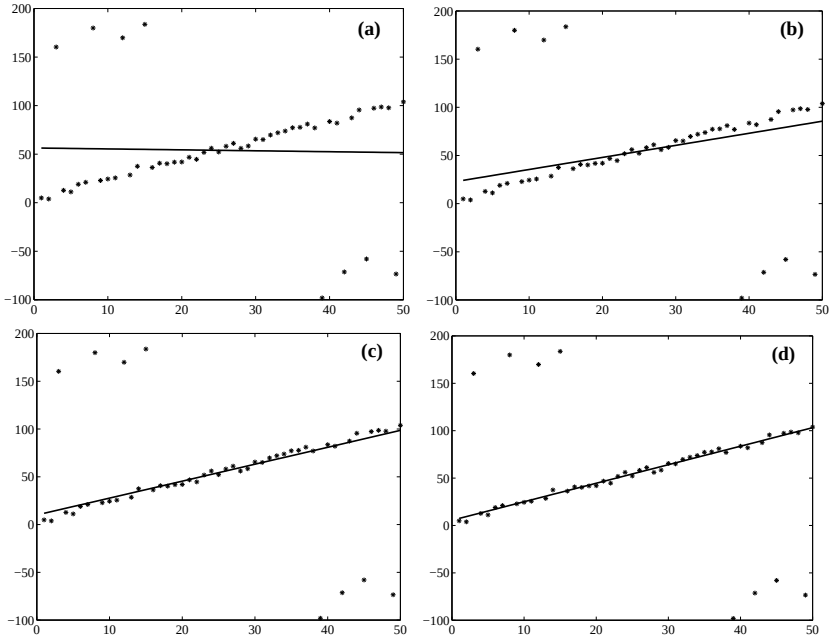

Figure 5: (a) Observed noisy data and standard least squares linear regression. (b), (c), and (d): LAD estimate via EM after 1, 4, and 10 iterations, respectively.

As in the previous example, it is possible to extend this approach to estimate also the hyper-parameter η from the data.

5.4 More on robust estimation: Student t distribution

Student t distributions (univariate or multivariate) are a common replacement for the Gaussian density when robustness (with respect to outliers) is sought [28]. The robust nature of the t distribution is related to the fact that it exhibits heavier tails than the Gaussian. However, unlike the Gaussian

density, ML (or MAP) estimates of its parameters can not be obtained in closed form. In this example, we will show how the EM algorithm can be used to circumvent this difficulty. To keep the notation simple we will focus on the univariate case.

A variable X is said to have a Student t distribution if its density has the form

$$p(x) = t_\nu(\mu, \sigma^2) \equiv \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi\sigma^2}} \left(1 + \frac{1}{\nu}\frac{(x-\mu)^2}{\sigma^2}\right)^{-\frac{\nu+1}{2}}$$

where μ is the mean, σ is a scale parameter, and ν is called the *number of degrees of freedom*. The variance, which is only finite if $\nu > 2$, is equal to $\sigma^2(\nu/(\nu-2))$. When $\nu \to \infty$, this density approaches a $\mathcal{N}(\mu, \sigma^2)$. Figure 6 plots a $t_1(0,1)$, a $t_4(0,1)$, and a $\mathcal{N}(0,1)$ density; observe how the t densities decay slower than the Gaussian (have heavier tails).

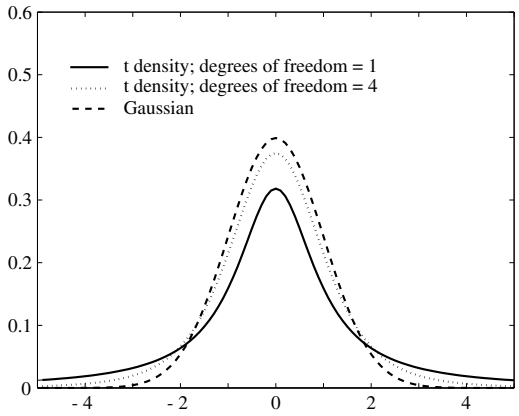

Figure 6: Three densities: $t_1(0,1)$, $t_4(0,1)$, and $\mathcal{N}(0,1)$. Observe the heavier tails of the t densities and how the $t_4(0,1)$ is closer to the Gaussian than the $t_1(0,1)$.

Now suppose that given a set of n observations $\mathbf{x} = \{x_1, ..., x_n\}$, the goal is to find the ML estimates of μ and σ (assuming ν is known), for which there is no closed form solution. The door to the use of the EM algorithm is opened by the observation that a t-distributed variable can be obtained by a two-step procedure. Specifically, let Z be a random variable following a $\mathcal{N}(0,1)$ distribution. Let $Y = Y'/\nu$ be another random variable, independent from Z, where Y' obeys a *chi-square*[2] distribution with ν degrees of freedom

[2]The *chi-square* distribution is a particular case of a Gamma distribution with $\alpha = \nu/2$

(notice that $p(y) = \nu\chi^2_\nu(\nu y)$). Then

$$X = \mu + \frac{\sigma\, Z}{\sqrt{Y}}$$

follows a Student-t distribution $t_\nu(\mu, \sigma)$.

This decomposition of the Student-t distribution suggests the "creation" of a missing data set $\mathbf{y} = \{y_1, ..., y_n\}$ such that x_i is a sample of $X_i = \mu + \frac{\sigma\, Z}{\sqrt{y_i}}$. In fact, if $\mathbf{y}$ was observed, each x_i would simply be a noisy version of the constant μ contaminated with a zero-mean Gaussian perturbation of variance σ^2/y_i (because Z is $\mathcal{N}(0,1)$). The complete log-likelihood function is then

$$\begin{aligned} \log p(\mathbf{x}, \mathbf{y}|\mu, \sigma^2) &= \sum_{i=1}^{n} \left(\log \mathcal{N}(x_i|\mu, \sigma^2/y_i) + \log \nu + \log \chi^2_\nu(\nu y_i)\right) \\ &\propto -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i\,(x_i - \mu)^2 \end{aligned} \tag{45}$$

where we have dropped all additive terms that do not depend on μ or σ^2. Notice that Eq. (45) is linear in the missing variables $\mathbf{y}$; accordingly, the Q-function can be obtained by plugging their conditional expectations into the complete log-likelihood function, *i.e.*,

$$\begin{aligned} Q\left(\mu, \sigma^2|\widehat{\mu}^{(t)}, \widehat{\sigma}^{(t)}\right) &= E\left[\log p(\mathbf{x}, \mathbf{y}|\mu, \sigma^2)\,\middle|\,\mathbf{x}, \widehat{\mu}^{(t)}, \widehat{\sigma}^{(t)}\right] \\ &\propto \log p\left(\mathbf{x}, E\left[\mathbf{y}|\mathbf{x}, \widehat{\mu}^{(t)}, \widehat{\sigma}^{(t)}\right]\middle|\,\mu, \sigma^2\right). \end{aligned}$$

To obtain the conditional expectations $E[y_i|x_i, \widehat{\mu}^{(t)}, \widehat{\sigma}^{(t)}]$, notice that we can see $p(x_i|y_i, \mu, \sigma_2) = \mathcal{N}(x_i|\mu, \sigma^2/y_i)$ as a likelihood function and $p(y_i) = \nu\chi^2_\nu(\nu y_i)$ as conjugate prior. In fact, as stated in footnote 2 in the previous page, $\chi^2_\nu(y') = \mathrm{Ga}(y'|\nu/2, 1/2)$, and so $p(y_i) = \nu\chi^2_\nu(\nu y_i) = \mathrm{Ga}(y_i|\nu/2, \nu/2)$. The corresponding posterior is then

$$\mathrm{Ga}\left(y_i\,\middle|\,\frac{\nu}{2} + \frac{1}{2}, \frac{\nu}{2} + \frac{1}{2(\widehat{\sigma}^{(t)})^2}\sum_{i=1}^{n}(x_i - \widehat{\mu}^{(t)})^2\right)$$

---

and $\beta = 1/2$. Its probability density function is

$$\chi^2_\nu(x) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)} x^{\frac{\nu-2}{2}} \exp\{-\frac{x}{2}\}.$$

This density is probably best know by the fact that it characterizes the sum of the squares of ν independent and identically distributed zero-mean unit-variance Gaussian variables.

whose mean is

$$w_i^{(t)} \equiv E[y_i|x_i, \widehat{\mu}^{(t)}, \widehat{\sigma}^{(t)}] = \frac{\nu+1}{\nu + \frac{(x_i - \widehat{\mu}^{(t)})^2}{2(\widehat{\sigma}^{(t)})^2}}.$$

Finally, plugging $w_i^{(t)}$ in the complete log-likelihood function, and maximizing it with respect to μ and σ leads to the following update equations:

$$\widehat{\mu}^{(t+1)} = \frac{\sum_{i=1}^n x_i\, w_i^{(t)}}{\sum_{i=1}^n w_i^{(t)}} \tag{46}$$

$$\widehat{\sigma}^{(t+1)} = \sqrt{\frac{\sum_{i=1}^n \left(x_i - \widehat{\mu}^{(t+1)}\right)^2 w_i^{(t)}}{n}}. \tag{47}$$

To illustrate the use of Student's t distribution in robust estimation, consider the data in Figure 7 (a); it consists of 100 noisy observations of the underlying, supposedly unknown, constant μ, which is equal to 1. Most of the observations were obtained by adding i.i.d. Gaussian noise, of standard deviation equal to 0.8; however, 10 of them are outliers, *i.e.* they are abnormally large observations which are not well explained by the Gaussian noise assumption. If we ignore the presence of the outliers and simply compute the mean of the observations (which is the ML estimate of μ under the i.i.d. Gaussian noise assumption) we obtain $\widehat{\mu} = 2.01$, clearly affected by the presence of the outliers; the corresponding estimate of σ^2 is 3.0, also unduly affected by the outliers. If we could remove the outliers and compute the mean of the remaining observations, the results would be $\widehat{\mu} = 1.06$ and $\widehat{\sigma} = 0.85$, much closer to the true values. The results of using a t model for the noise (with $\nu = 2$) are also depicted in Figure 7 (b-d). Figure 7 (b) and (d) show the evolution of the estimates $\widehat{\mu}^{(t)}$ and $\widehat{\sigma}^{(t)}$; the final values are $\widehat{\mu} = 1.07$ and $\widehat{\sigma} = 0.81$, very close to the true values. Finally, 7 (c) plots the final values of w_i showing that, at the positions of the outliers, these values are small, thus down-weighting these observations in the mean and variance estimates (Eqs. (47) and (47)).

6 Unsupervised Classification and Mixture Models

6.1 The EM algorithm for mixtures

Let us now consider a classification problem over K classes, which we label $\{1, 2, ..., K\}$. Assume that the number of classes K is known. The class-

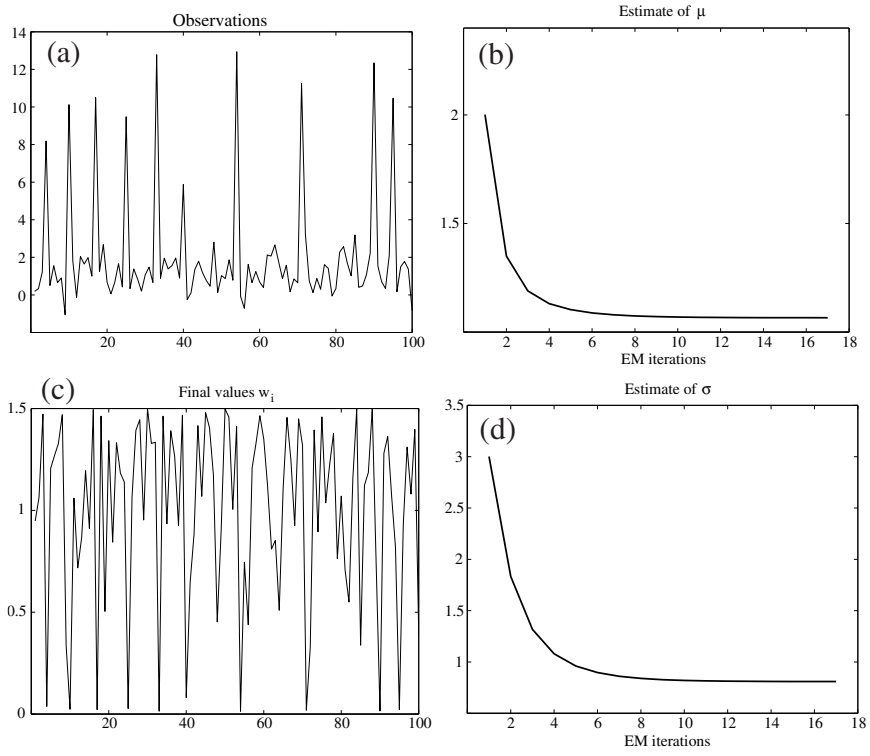

Figure 7: (a) Data with outliers. (b) Evolution of the estimate of the mean produced by the EM algorithm, under a Student t noise model. (c) Final values w_i (see text). (d) Evolution of the estimate of the scale parameter σ.

conditional probability functions,

$$p(\mathbf{g}|\boldsymbol{\theta}_s, \boldsymbol{\theta}_0), \quad \text{for } s \in \{1, 2, ..., K\},$$

have unknown parameters which are collected in $\boldsymbol{\theta} = \{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K\}$; $\boldsymbol{\theta}_0$ contains the parameters common to all classes. Let us stress again that there may exist known parameters, but we simply omit them from the notation. Also, we assume that the functional form of each class-conditional probability function is known; for example, they can all be Gaussian with different means and covariance matrices, *i.e.*, $p(\mathbf{g}|\boldsymbol{\theta}_s) = \mathcal{N}(\mathbf{g}|\boldsymbol{\mu}_s, \mathbf{C}_s)$ in which case $\boldsymbol{\theta}_s = \{\boldsymbol{\mu}_s, \mathbf{C}_s\}$ and $\boldsymbol{\theta}_0$ does not exist, or Gaussian with different means but a common covariance matrix, *i.e.*, $p(\mathbf{g}|\boldsymbol{\theta}_s) = \mathcal{N}(\mathbf{g}|\boldsymbol{\mu}_s, \mathbf{C})$ in which case $\boldsymbol{\theta}_s = \boldsymbol{\mu}_s$ and $\boldsymbol{\theta}_0 = \mathbf{C}$. If the *a priori* probabilities of each class are also considered unknown, they are collected in a vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_K)$.

In an unsupervised learning scenario, we are given a data set containing observations $\mathcal{G} = \{\mathbf{g}_1, ..., \mathbf{g}_N\}$ whose true classes are unknown. The usual goal is to obtain estimates of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\alpha}}$ from the data. Since the class of each observation is unknown, the likelihood function is

$$p(\mathbf{g}_i|\boldsymbol{\alpha}, \boldsymbol{\theta}) = \sum_{s=1}^{K} \alpha_s f(\mathbf{g}_i|\boldsymbol{\theta}_s, \boldsymbol{\theta}_0), \quad \text{for } i = 1, 2, ..., N; \tag{48}$$

this type of probability density function (a linear combination of a set of, usually simpler, probability density functions) is called a *finite mixture* [37], [39], [53].

Before proceeding, let us point out that the use of finite mixture models is not limited to unsupervised learning scenarios [22], [23], [37]. In fact, mixtures constitute a flexible and powerful probabilistic modeling tool for univariate and multivariate data. The usefulness of mixture models in any area which involves statistical data modelling is currently widely acknowledged [33], [37], [39], [53]. A fundamental fact is that finite mixture models are able to represent arbitrarily complex probability density functions. This fact makes them well suited for representing complex likelihood functions (see [17] and [20]), or priors (see [8] and [10]) for Bayesian inference. Recent theoretical results concerning the approximation of arbitrary probability density functions by finite mixtures can be found in [32].

Let us consider some prior $p(\boldsymbol{\alpha}, \boldsymbol{\theta})$; the MAP estimate (ML, if this prior is flat) is given by

$$\left(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\theta}}\right)_{\text{MAP}} = \arg\max_{\boldsymbol{\alpha}, \boldsymbol{\theta}} \{\log p(\boldsymbol{\alpha}, \boldsymbol{\theta}) + \underbrace{\sum_{i=1}^{N} \log \sum_{s=1}^{K} \alpha_s p(\mathbf{g}_i|\boldsymbol{\theta}_s)}_{\text{log-likelihood: } L(\boldsymbol{\alpha}, \boldsymbol{\theta}, \mathcal{G})} \}$$

which does not have a closed form solution (even if the prior is flat). However, estimating the mixture parameters is clearly a missing data problem where the class labels of each observation are missing and the EM algorithm can be adopted. In fact, EM (and some variants of it) is the standard choice for this task [37], [38], [39]. To derive the EM algorithm for mixtures, it is convenient to formalized the missing labels as follows: associated to each observation $\mathbf{g}_i$, there is a (missing) label $\mathbf{z}_i$ that indicates the class of that observation. Specifically each label is a binary vector $\mathbf{z}_i = [z_i^{(1)}, ..., z_i^{(K)}]^T$ such that $z_i^{(s)} = 1$ and $z_i^{(j)} = 0$, for $j \neq s$, indicates that $\mathbf{g}_i$ was generated by class s. If the missing data $\mathcal{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_N\}$ was observed together with

the actually observed $\mathcal{G}$, we could write the *complete* loglikelihood function

$$L_c(\boldsymbol{\alpha}, \boldsymbol{\theta}, \mathcal{G}, \mathcal{Z}) \equiv \log \prod_{i=1}^{N} p(\mathbf{g}_i, \mathbf{z}_i | \boldsymbol{\alpha}, \boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{s=1}^{K} z_i^{(s)} \log \left(\alpha_s p(\mathbf{g}_i | \boldsymbol{\theta}_s) \right). \quad (49)$$

Maximization of $L_c(\boldsymbol{\alpha}, \boldsymbol{\theta}, \mathcal{G}, \mathcal{Z})$ with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ is simple and leads to the following ML estimates:

$$\widehat{\alpha_s} = \frac{1}{N} \sum_{i=1}^{N} z_i^{(s)}, \quad \text{for } s = 1, 2, ..., K, \quad (50)$$

$$\widehat{\boldsymbol{\theta}}_s = \arg\max_{\boldsymbol{\theta}_s} \sum_{i=1}^{N} z_i^{(s)} \log p(\mathbf{g}_i | \boldsymbol{\theta}_s), \quad \text{for } s = 1, 2, ..., K. \quad (51)$$

That is, the presence of the missing data would turn the problem into a supervised learning one, where the parameters of each class-conditional density are estimated separately. It is of course easy to extend this conclusion to cases where we have non-flat priors $p(\boldsymbol{\alpha}, \boldsymbol{\theta})$.

Recall that the fundamental step of the EM algorithm consists in computing the conditional expectation of the complete loglikelihood function, given the observed data and the current parameter estimate (the E-step, Eq. (2)). Notice that the complete loglikelihood function in Eq. (49) is linear in the missing observations; this allows writing

$$\begin{aligned} Q(\boldsymbol{\alpha}, \boldsymbol{\theta} | \widehat{\boldsymbol{\alpha}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}) &= E\left[L_c(\boldsymbol{\alpha}, \boldsymbol{\theta}, \mathcal{G}, \mathcal{Z}) | \mathcal{G}, \widehat{\boldsymbol{\alpha}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)} \right] \\ &= L_c\left(\boldsymbol{\alpha}, \boldsymbol{\theta}, \mathcal{G}, E[\mathcal{Z} | \mathcal{G}, \widehat{\boldsymbol{\alpha}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}] \right) \end{aligned}$$

because expectations commute with linear functions. In other words, the E-step reduces to the computation of the expected value of the missing data, which is then plugged into the complete log-likelihood function. Since the elements of $\mathcal{Z}$ are all binary variables, we have

$$w_i^{(s)} \equiv E[z_i^{(s)} | \mathcal{G}, \widehat{\boldsymbol{\alpha}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}] = \Pr[z_i^{(s)} = 1 | \mathcal{G}, \widehat{\boldsymbol{\alpha}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}]$$

which is the probability that $\mathbf{g}_i$ was generated by class s, given the observations $\mathcal{G}$ and the current parameter estimates. This probability can easily be obtained from Bayes law,

$$w_i^{(s)} = \Pr[z_i^{(s)} = 1 | \mathcal{G}, \widehat{\boldsymbol{\alpha}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}] = \frac{\widehat{\alpha_s}^{(t)} p(\mathbf{g}_i | \widehat{\boldsymbol{\theta}_s}^{(t)})}{\sum_{r=1}^{K} \widehat{\alpha_r}^{(t)} p(\mathbf{g}_i | \widehat{\boldsymbol{\theta}_r}^{(t)})}, \quad (52)$$

because $w_i^{(s)}$ and $\widehat{\alpha_s}^{(t)}$ can be seen, respectively, as the *a posteriori* and *a priori* probabilities that $\mathbf{g}_i$ belongs to class s (given the current parameter estimates). The M-step then consists in solving Eqs. (50) and (51) with $w_i^{(s)}$ replacing $z_i^{(s)}$.

6.2 Particular case: Gaussian mixtures

Gaussian mixtures, *i.e.*, those in which each component is Gaussian, $p(\mathbf{g}|\boldsymbol{\theta}_s) = \mathcal{N}(\mathbf{g}|\boldsymbol{\mu}_s, \mathbf{C}_s)$, are by far the most common type. From an unsupervised classification perspective, this corresponds to assuming Gaussian class-conditional densities. For simplicity, let us assume that there are no constraints on the unknown parameters of each component (mean and covariance matrix) and that the prior is flat, *i.e.*, we will be looking for ML estimates of the mixture parameters. In this case, the parameters to be estimated are $\boldsymbol{\alpha}$ and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K\}$, where $\boldsymbol{\theta}_s = \{\boldsymbol{\mu}_s, \mathbf{C}_s\}$ (the mean and covariance of component s). The E-step (Eq. (52)), in the particular case of Gaussian components becomes

$$w_i^{(s)} = \frac{\widehat{\alpha_s}^{(t)} \mathcal{N}(\mathbf{g}_i|\widehat{\boldsymbol{\mu}_s}^{(t)}, \widehat{\mathbf{C}_s}^{(t)})}{\sum_{r=1}^{K} \widehat{\alpha_r}^{(t)} \mathcal{N}(\mathbf{g}_i|\widehat{\boldsymbol{\mu}_r}^{(t)}, \widehat{\mathbf{C}_r}^{(t)})}.$$

Concerning the M-step, for the Gaussian case we have

$$\widehat{\alpha_s}^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} w_i^{(s)}, \tag{53}$$

$$\widehat{\boldsymbol{\mu}_s}^{(t+1)} = \frac{\sum_{i=1}^{N} \mathbf{g}_i w_i^{(s)}}{\sum_{i=1}^{N} w_i^{(s)}}, \tag{54}$$

$$\widehat{\mathbf{C}_s}^{(t+1)} = \frac{\sum_{i=1}^{N} (\mathbf{g}_i - \widehat{\boldsymbol{\mu}_s}^{(t+1)})(\mathbf{g}_i - \widehat{\boldsymbol{\mu}_s}^{(t+1)})^T w_i^{(s)}}{\sum_{i=1}^{N} w_i^{(s)}}, \tag{55}$$

all for $s = 1, 2, ..., K$. Notice that the update equations for the mean vectors and the covariance matrices are weighted versions of the standard ML estimates (sample mean and sample covariance).

We will now illustrate the application of the EM algorithm in learning a mixture of four bivariate equiprobable Gaussian classes. The parameters of the class-conditional Gaussian densities are as follows:

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 0 \\ -3 \end{bmatrix} \quad \boldsymbol{\mu}_3 = \begin{bmatrix} 3 \\ 2.5 \end{bmatrix} \quad \boldsymbol{\mu}_4 = \begin{bmatrix} -3 \\ 2.5 \end{bmatrix},$$

and

$$\mathbf{C}_1 = \begin{bmatrix} 0.5 & 0 \\ 0 & 1 \end{bmatrix} \qquad \mathbf{C}_2 = \begin{bmatrix} 1.5 & 0 \\ 0 & 0.1 \end{bmatrix}$$
$$\mathbf{C}_3 = \begin{bmatrix} 1 & -0.3 \\ -0.3 & 0.2 \end{bmatrix} \quad \mathbf{C}_4 = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 0.2 \end{bmatrix}.$$

Figure 8 shows 1000 samples (approximately 250 per class) of this densities and the evolution of the component estimates obtained by the EM algorithm (the ellipses represent level curves of each Gaussian density). We also show the initial condition from which EM was started.

The estimates of the parameters obtained were

$$\widehat{\alpha_1} = 0.2453 \quad \widehat{\alpha_2} = 0.2537 \quad \widehat{\alpha_3} = 0.2511 \quad \widehat{\alpha_4} = 0.2499,$$

(recall that the classes are equiprobable, thus the true values are $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.25$),

$$\widehat{\boldsymbol{\mu}_1} = \begin{bmatrix} 0.036 \\ 0.057 \end{bmatrix} \quad \widehat{\boldsymbol{\mu}_2} = \begin{bmatrix} 0.091 \\ -3.022 \end{bmatrix} \quad \widehat{\boldsymbol{\mu}_3} = \begin{bmatrix} 3.017 \\ 2.475 \end{bmatrix} \quad \widehat{\boldsymbol{\mu}_4} = \begin{bmatrix} -2.971 \\ 2.543 \end{bmatrix},$$

and

$$\widehat{\mathbf{C}_1} = \begin{bmatrix} 0.448 & 0 \\ 0 & 0.930 \end{bmatrix} \qquad \widehat{\mathbf{C}_2} = \begin{bmatrix} 1.318 & 0 \\ 0 & 0.098 \end{bmatrix}$$
$$\widehat{\mathbf{C}_3} = \begin{bmatrix} 0.968 & -0.2785 \\ -0.2785 & 0.209 \end{bmatrix} \quad \widehat{\mathbf{C}_4} = \begin{bmatrix} 1.049 & 0.302 \\ 0.302 & 0.197 \end{bmatrix}.$$

One of the main features of EM is its greedy (local) nature which makes proper initialization an important issue. This is specially true with mixture models, whose likelihood function has many local maxima. Local maxima of the likelihood arise when there are too many components in one region of the space, and too few in another, because EM is unable to move components from one region to the other, crossing low likelihood regions. This fact is illustrated in Figures 9 and 10, where we show an example of a successful and an unsuccessful initialization.

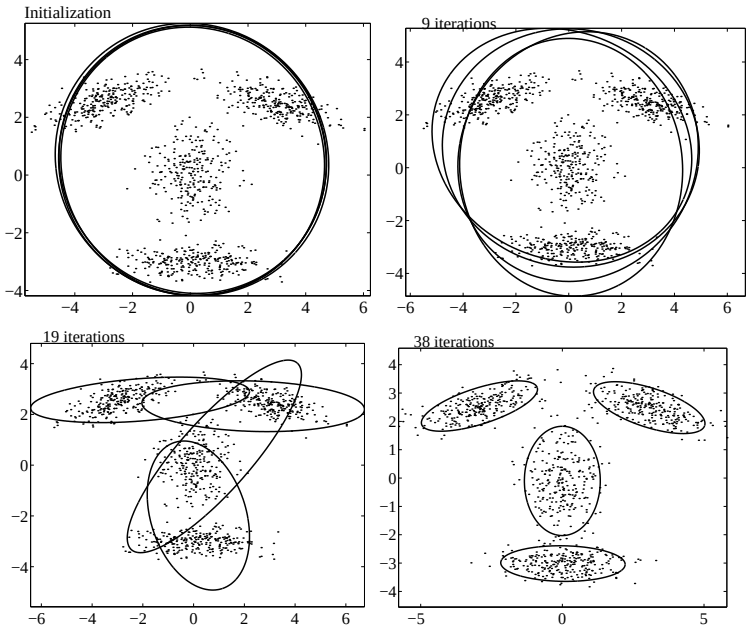

Figure 8: The EM algorithm applied to 1000 (unclassified) samples of the four Gaussian classes described in the text. Convergence is declared when the relative increase in the objective function (the likelihood in this case) falls below a threshold; with the threshold set to 10^{-5} convergence happened after 38 iterations. The ellipses represent level curves of each Gaussian density.

It should be pointed out that looking for the ML estimate of the parameters of a Gaussian mixture is a peculiar type of optimization problem. On one hand, we have a multi-modal likelihood function with several poor local maxima which make pure local/greedy methods (like EM) dependent on initialization. But on the other hand, we do not want a global maximum, because the likelihood function is unbounded. As a simple example of the unbounded nature of the likelihood function, consider n real observations $x_1, ...x_n$ to which a two-component univariate Gaussian mixture is to be fitted. The logarithm of the likelihood function is

$$\sum_{i=1}^{n} \log \left(\frac{\alpha \, e^{-\frac{(x_i-\mu_1)}{2\sigma_1^2}}}{\sqrt{2\pi\sigma_1^2}} + \frac{(1-\alpha) \, e^{-\frac{(x_i-\mu_2)}{2\sigma_2^2}}}{\sqrt{2\pi\sigma_2^2}} \right)$$

With $\mu_1 = x_1$, the first term can be made arbitrarily large by letting σ_1^2

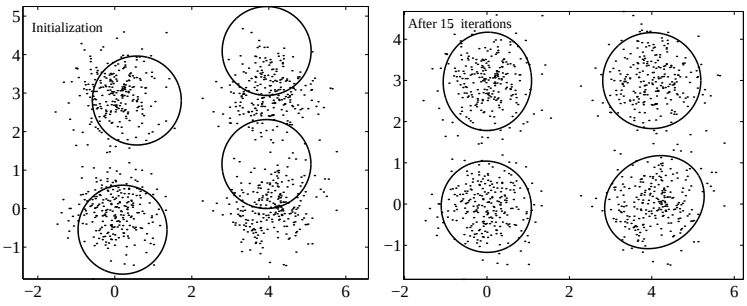

Figure 9: Example of a successful run of EM.

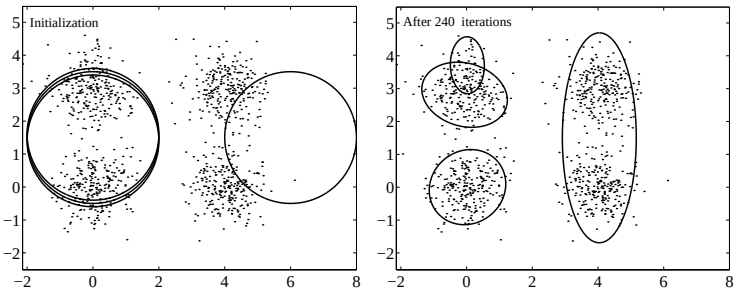

Figure 10: Example of a unsuccessful run of EM due to poor initialization.

approach zero, while the other terms of the sum remain bounded. In conclusion, what is sought for is a "good" local maximum, not a global maximum. For several references on the EM initialization problem for finite mixtures, and a particular approach, see [14], [39].

References

[1] D. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, MA, 1999.

[2] P. Bloomfield and W. Steiger. *Least Absolute Deviations: Theory, Applications, and Algorithms.* Birkhäuser, Boston, 1983.

[3] L. Brown. *Foundations of Exponential Families.* Institute of Mathematical Statistics, Hayward, CA, 1986.

[4] G. Celeux, S. Chrétien, F. Forbes, and A. Mkhadri. A component-wise EM algorithm for mixtures. Technical Report 3746, INRIA Rhône-Alpes, France, 1999. Available at `http://www.inria.fr/RRRT/RR-3746.html`.

[5] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal of Scientific Computation*, 20(1):33–61, 1998.

[6] S. Chrétien and A. Hero III. Kullback proximal algorithms for maximum likelihood estimation. *IEEE Transactions on Information Theory*, 46:1800–1810, 2000.

[7] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.

[8] S. Dalal and W. Hall. Approximating priors by mixtures of natural conjugate priors. *Journal of the Royal Statistical Society (B)*, 45, 1983.

[9] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

[10] P. Diaconis and D. Ylvisaker. Conjugate priors for exponential families. *Annals of Statistics*, 7:269–281, 1979.

[11] D. Donoho and I. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[12] D. Donoho and I. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995.

[13] M. Figueiredo. Adaptive sparseness for supervised learning. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 25:1150–1159, 2003.

[14] M. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396, 2001.

[15] W. J. Fu. Penalized regression: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998.

[16] F. Girosi. Models of noise and robust estimates. Massachusetts Institute of Technology. Artificial Intelligence Laboratory (Memo 1287) and Center for Biological and Computational Learning (Paper 66), 1991.

[17] T. Hastie and R. Tibshirani. Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society (B)*, 58:155–176, 1996.

[18] T. Hebert and R. Leahy. A generalized EM algorithm for 3D Bayesian reconstruction from Poisson data using Gibbs priors. *IEEE Transactions on Medical Imaging*, MI-8:194–202, 1989.

[19] T.J. Hebert and Y. Leahy. Statistic-based MAP image reconstruction from Poisson data using Gibbs priors. *IEEE Transactions on Signal Processing*, 40(9):2290–2303, 1992.

[20] G. Hinton, P. Dayan, and M. Revow. Modeling the manifolds of images of handwritten digits. *IEEE Transactions on Neural Networks*, 8:65–74, 1997.

[21] P. Huber. *Robust statistics*. John Wiley, New York, 1981.

[22] A. K. Jain and R. Dubes. *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, N. J., 1988.

[23] A. K. Jain, R. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–38, 2000.

[24] M. Jamshidian and R. Jennrich. Acceleration of the EM algorithm by using wuasi-Newton methods. *Journal of the Royal Statistical Society (B)*, 59(3):569–587, 1997.

[25] R. Lagendijk, J. Biemond, and D. Boekee. Identification and restoration of noisy blurred images using the expectation-maximization algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(7):1180–1191, July 1990.

[26] K. Lange. Convergence of EM image reconstruction algorithms with Gibbs smoothing. *IEEE Med. Im.*, 9(4):439–446, 1991.

[27] K. Lange, D. Hunter, and I. Yang. Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9:1–59, 2000.

[28] K. Lange, R. Little, and J. Taylor. Robust statistical modeling using the t-distribution. *Journal of the American Statistical Association*, 84:881–896, 1989.

[29] K. Lange and J. Sinsheimer. Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics*, 2:175–198, 1993.

[30] K. Lay and A. Katsaggelos. Blur identification and image restoration based on the EM algorithm. *Optical Engineering*, 29(5):436–445, May 1990.

[31] M. Lewicki and T. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12:337–365, 2000.

[32] J. Li and A. Barron. Mixture density estimation. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems 12*. MIT Press, 2000.

[33] B. Lindsay. *Mixture Models: Theory, Geometry, and Applications*. Institute of Mathematical Statistics and American Statistical Association, Hayward, CA, 1995.

[34] R. Little and D. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, New York, 1987.

[35] D. MacKay. Bayesian non-linear modelling for the 1993 energy prediction competition. In G. Heidbreder, editor, *Maximum Entropy and Bayesian Methods*, pages 221–234. Kluwer, Dordrecht, 1996.

[36] B. Martinet. Regularisation d'inéquations variationnelles par approximations successives. *Revue Française d'Informatique et de Recherche Operationnelle*, 3:154–179, 1970.

[37] G. McLachlan and K. Basford. *Mixture Models: Inference and Application to Clustering*. Marcel Dekker, New York, 1988.

[38] G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. John Wiley & Sons, New York, 1997.

[39] G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, New York, 2000.

[40] X. Meng and D. van Dyk. The EM algorithm – an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society (B)*, 59(3):511–567, 1997.

[41] M. I. Miller and C. S. Butler. 3-D maximum likelihood a posteriori estimation for single-photon emission computed tomography on massively parallel computers. *IEEE Transactions on Medical Imaging*, 12(3):560–565, 1993.

[42] R. Neal. *Bayesian Learning for Neural Networks*. Springer Verlag, New York, 1996.

[43] R. Neal and G. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. Kluwer Academic Publishers, 1998.

[44] R. T. Ogden. *Essential Wavelets for Statistical Applications and Data Analysis*. Birkhäuser, Boston, MA, 1997.

[45] J. Ollinger and D. Snyder. A preliminary evaluation of the use of the EM algorithm for estimating parameters in dynamic tracer studies. *IEEE Transactions on Nuclear Scince*, 32:3575–3583, 1985.

[46] B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[47] W. Qian and D. M. Titterington. Bayesian image restoration – an application to edge-preserving surface recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(7):748–752, 1993.

[48] R. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14:877–898, 1976.

[49] P. Rowsseeuw and A. Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, New York, 1987.

[50] L. Shepp and Y. Vardi. Maximum likelihood reconstruction for emission tomography. *IEEE Transactions on Medical Imaging*, MI-1(2):113–122, October 1982.

[51] M. Tanner. *Tools for Statistical Inference*. Springer-Verlag, New York, 1993.

[52] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (B)*, 58:267–288, 1996.

[53] D. Titterington, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, Chichester (U.K.), 1985.

[54] P. Williams. Bayesian regularization and pruning using a Laplace prior. *Neural Computation*, 7:117–143, 1995.

[55] C. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

[56] J. Zhang. The mean field theory in EM procedures for blind Markov random field image restoration. *IEEE Transactions on Image Processing*, 2(1):27–40, January 1993.