1 EM Algorithm and Mixtures.

1.1 Introduction

The Expectation-Maximization (EM) iterative algorithm is a broadly applicable statistical technique for maximizing complex likelihoods and handling the incomplete data problem. At each iteration step of the algorithm, two steps are performed: (i) E-Step consisting of projecting an appropriate functional containing the augmented data on the space of the original, incomplete data, and (ii) M-Step consisting of maximizing the functional. The name EM algorithm was coined by Dempster, Laird, and Rubin in their fundamental paper [1], often referred to as DLR paper. But if one comes up with smart idea, one may be sure that other smart guys in history thought about it.

The EM algorithm relates to MCMC as a forerunner by its data augmentation step that replaces simulation by maximization. Newcomb [7] was interested in estimating the mixtures of normals in 1886. McKendrick [5] and Healy and Westmacott [3] proposed iterative methods that, in fact, are examples of the EM algorithm. Dozens of papers proposing various applications of EM appeared before the DLR paper in 1997. However, the DLR paper was the first to unify and organize the approach.

1.2 What is EM?

Let Y be a random vector corresponding to the observed data y and having a postulated pdf as $f(y, \psi)$, where $\psi = (\psi_1, \ldots, \psi_d)$ is a vector of unknown parameters. Let x be a vector of augmented (so called complete) data, and let z be the additional data, $x = [y, z]$.

Denote by $g_c(x, \psi)$ the pdf of the random vector corresponding to the complete data set x. The log-likelihood for ψ, if x were fully observed, would be

$$\log L_c(\psi) = \log g_c(x, \psi).$$

The incomplete data vector y comes from the "incomplete" sample space $\mathcal{Y}$. There is a 1-1 correspondence between the complete sample space $\mathcal{X}$ and the incomplete sample space $\mathcal{Y}$. Thus, for $x \in \mathcal{X}$, one can uniquely find the "incomplete" $y = y(x) \in \mathcal{Y}$. Also, the incomplete pdf could be found by properly integrating out the complete pdf,

$$g(y, \psi) = \int_{\mathcal{X}(y)} g_c(x, \psi) dx,$$

where $\mathcal{X}(y)$ is the subset of $\mathcal{X}$ constrained by the relation $y = y(x)$.

Let $\psi^{(0)}$ be some initial value for ψ. At the k-th step the **EM** algorithm one performs the following two steps:

E-Step. Calculate

$$Q(\psi, \psi^{(k)}) = \mathbb{E}_{\psi^{(k)}}\{\log L_c(\psi)|y\}.$$

M-Step. Choose any value $\psi^{(k+1)}$ that maximizes $Q(\psi, \psi^{(k)})$, i.e.,

$$(\forall \psi) Q(\psi^{(k+1)}, \psi^{(k)}) \geq Q(\psi, \psi^{(k)}).$$

The **E** and **M** steps are alternated until the difference

$$L(\psi^{(k+1)}) - L(\psi^{(k)})$$

becomes small in absolute value.

Next we illustrate the EM algorithm on a famous example first considered by Fisher and Balmukand [2]. It is also discussed in Rao's monograph [8] and Mclachlan and Krishnan [6].

1.2.1 Fisher's Example

Here is the background. This description follows a superb 2002 lecture by Terry Speed of UC at Berkeley. In modern terminology, one has two linked bi-allelic loci, A and B say, with alleles A and a, and B and b, respectively, where A is dominant over a and B is dominant over b. A double heterozygote $AaBb$ will produce gametes of four types: AB, Ab, aB and ab. Since the loci are linked, the types AB and ab will appear with a frequency different from that of Ab and aB, say $1-r$ and r, respectively, in males, and $1-r'$ and r' respectively in females. Here we suppose that the parental origin of these heterozygotes is from the mating $AABB \times aabb$, so that r and r' are the male and female recombination rates between the two loci. The problem is to estimate r and r', if possible, from the offspring of selfed double heterozygotes. Since gametes AB, Ab,aB and ab are produced in proportions $(1-r)/2, r/2, r/2$ and $(1-r)/2$ respectively by the male parent, and $(1-r')/2, r'/2, r'/2$ and $(1-r')/2$ respectively by the female parent, zygotes with genotypes $AABB, AaBB, \ldots$ etc, are produced with frequencies $(1-r)(1-r')/4, (1-r)r'/4$, etc.

The problem here is this: although there are 16 distinct offspring genotypes, taking parental origin into account, the dominance relations imply that we only observe 4 distinct phenotypes, which we denote by A^*B^*, A^*b^*, a^*B^* and a^*b^*. Here A^*(respectively B^*) denotes the dominant, while a^* (respectively b^*) denotes the recessive phenotype determined by the alleles at A (respectively B).

Figure 1: Sir R. A. Fisher (1890-1962)

Thus individuals with genotypes $AABB, AaBB, AABb$ or $AaBb$, which account for 9/16 gametic combinations (check!), all exhibit the phenotype A^*B^*, i.e. the dominant alternative in both characters, while those with genotypes $AAbb$ or $Aabb$ (3/16) exhibit the phenotype A^*b^*, those with genotypes $aaBB$ and $aaBb$ (3/16) exhibit the phenotype a^*B^*, and finally the double recessives $aabb$ (1/16) exhibit the phenotype a^*b^*. It is a slightly surprising fact that the probabilities of the four phenotypic classes are definable

in terms of the parameter $\psi = (1-r)(1-r')$, as follows: a^*b^* has probability $\psi/4$ (easy to see), a^*B^* and A^*b^* both have probabilities $(1-\psi)/4$, while A^*B^* has probability 1 minus the sum of the preceding, which is $(2+\psi)/4$. Now suppose we have a random sample of n offspring from the selfing of our double heterozygote. Thus the 4 phenotypic classes will be represented roughly in proportion to their theoretical probabilities, their joint distribution being multinomial,

$$\textit{Multinomial}\left[n; \frac{2+\psi}{4}, \frac{1-\psi}{4}, \frac{1-\psi}{4}, \frac{\psi}{4}\right]. \tag{1}$$

Note that here neither r nor r' will be separately estimable from these data, but only the product $(1-r)(1-r')$. Note that since we know that $r \leq 1/2$ and $r' \leq 1/2$, it follows that $\psi \geq 1/4$.

How do we estimate ψ? Fisher and Balmukand [2] discuss a variety of methods that were in the literature at the time they wrote, and compare them with maximum likelihood, which is the method of choice in problems like this. We describe a variant of their approach to illustrate the **EM** algorithm.

Let $y = (125, 18, 20, 34)$ be a realization of vector $y = (y_1, y_2, y_3, y_4)$ believed to be coming from the multinomial distribution given in (1).

The probability mass function, given the data, is

$$g(y,\psi) = \frac{n!}{y_1!y_2!y_3!y_4!}(1/2+\psi/4)^{y_1}\ (1/4-\psi/4)^{y_2+y_3}\ (\psi/4)^{y_4}.$$

The log likelihood, after omitting an additive term not containing ψ is:

$$\log L(\psi) = y_1\log(2+\psi) + (y_2+y_3)\log(1-\psi) + y_4\log(\psi).$$

By differentiating with respect to ψ one gets

$$\partial \log L(\psi)/\partial\psi = \frac{y_1}{2+\psi} - \frac{y_2+y_3}{1-\psi} + \frac{y_4}{\psi}.$$

The equation $\partial \log L(\psi)/\partial\psi = 0$ can be solved and solution (produced by MATHEMATICA©) is $\psi = \frac{15+\sqrt{53809}}{394} \approx 0.626821$.

Assume that instead of original value y_1 the counts y_{11} and y_{12}, such that $y_{11} + y_{12} = y_1$, could be observed, and that their probabilities are $1/2$ and $\psi/4$, respectively. The "complete data" can be defined as $x = (y_{11}, y_{12}, y_2, y_3, y_4)$.

The probability mass function of incomplete data y is $g(y,\psi) = \sum g_c(x,\psi)$, where

$$g_c(x,\psi) = c(x)(1/2)^{y_{11}}(\psi/4)^{y_{12}}(1/4-\psi/4)^{y_2+y_3}(\psi/4)^{y_4},$$

$c(x)$ is free of ψ, and the summation is taken over all values of x for which $y_{11} + y_{12} = y_1$.

The "complete" log likelihood is

$$\log L_c(\psi) = (y_{12}+y_4)\log(\psi) + (y_2+y_3)\log(1-\psi). \tag{2}$$

Our goal is to find the conditional expectation of $\log L_c(\psi)$ given y, using the starting point for $\psi^{(0)}$,

$$Q(\psi,\psi^{(0)}) = \mathbb{E}_{\psi^{(0)}}\{\log L_c(\psi)|y\}.$$

As $\log L_c$ is a linear function in y_{11} and y_{12}, the **E-Step** is done by simply by replacing y_{11} and y_{12} by their conditional expectations, given y.

Considering Y_{11} to be a random variable corresponding to y_{11}, it is easy to see that

$$Y_{11} \sim Bin\left(y_1, \frac{1/2}{1/2 + \psi^{(0)}/4}\right).$$

Thus, the conditional expectation of Y_{11} given y_1 is

$$\mathbb{E}_{\psi^{(0)}}(Y_{11}|y_1) = \frac{\frac{y_1}{2}}{\frac{1}{2} + \frac{\psi^{(0)}}{4}} = y_{11}^{(0)}.$$

Of course, $y_{12}^{(0)} = y_1 - y_{11}^{(0)}$. This completes the **E-Step** part.

In the **M-Step** one chooses $\psi^{(1)}$ so that $Q(\psi, \psi^{(0)})$ is maximized. After replacing y_{11} and y_{12} by their conditional expectations $y_{11}^{(0)}$ and $y_{12}^{(0)}$ in the Q-function, the maximum is obtained at

$$\psi^{(1)} = \frac{y_{12}^{(0)} + y_4}{y_{12}^{(0)} + y_2 + y_3 + y_4} = \frac{y_{12}^{(0)} + y_4}{n - y_{11}^{(0)}}.$$

Now, the **E-** and **M-**Steps are alternating. At the iteration k we have

$$\psi^{(k+1)} = \frac{y_{12}^{(k)} + y_4}{n - y_{11}^{(k)}},$$

where $y_{11}^{(k)} = \frac{1}{2} y_1/(1/2 + \psi^{(k)}/4)$ and $y_{12}^{(k)} = y_1 - y_{11}^{(k)}$.

The insert from the matlab program `emexample.m` illustrates approximation of the MLE of ψ by the EM algorithm.

```
function psi = emexample(y1, y2, y3, y4, tol, start)
%--------------
 psi_last = 0;
 n = y1 + y2 + y3 + y4;
 psi_current = start;
 psi = psi_current;
 while (abs(psi_last-psi) > tol )
   [y11, y12] = estep(psi_current, y1);
   psi = mstep(y12, y11, y4, n);
   psi_last = psi_current;
   psi_current = psi
end %while
%--------------------------
function psi_new = mstep(y12, y11, y4, n)
 psi_new = (y12+y4)/(n-y11);
%--------------------------
function [y11, y12] = estep(psi_current, y1)
 y11 = (1/2*y1)/(1/2+psi_current/4);
 y12 = y1 - y11;
%--------------------------
```

When applied on the data $y = (125, 18, 20, 34)$ with tolerance of 10^{-6}, with $\psi^{(0)} = 0.5$ the folowing script is obtained

```
>> format long
>> emexample(125, 18, 20, 34, 10^-6, 0.5)
psi_current = 0.60824742268041 psi_current = 0.62432105036927
psi_current = 0.62648887907967 psi_current = 0.62677732234731
psi_current = 0.62681563211004 psi_current = 0.62682071901931
psi_current = 0.62682139445598

ans = 0.62682139445598

>> ans - (15+sqrt(53809))/394

ans = -1.034149983425436e-007
```

1.3 MAP Principle

All information in Bayesian inference is contained in the posterior and posterior location measures (mean, median, mode) are standard Bayes rules for the location parameters. Typically, it is more difficult to exhibit the mean or median of a posterior, than the value at which the posterior is maximized, a posterior mode. This is because for the mean or median, an exact expression for the posterior is needed. MAP rules that maximize the posterior maximize, at the same time, the product of the likelihood and prior, and are typically shrinkage rules.

To illustrate some MAP rules, assume that the likelihood is normal (density ϕ) and that the parameter of the interest, θ is the location. Given an observation z, the posterior distribution of θ is proportional to

$$\pi(\theta|z) \propto \phi(z-\theta) \cdot \pi(\theta). \tag{3}$$

Let $s(\theta) = -\log \pi(\theta)$ be the score of the prior. Notice that the posterior is maximized at the same argument at which

$$s(\theta) - \log \phi(z-\theta) = \frac{1}{2\sigma^2}(z-\theta)^2 + s(\theta) \tag{4}$$

is minimized. If $s(\theta)$ is strictly convex and differentiable, the minimizer of (4) is a solution $\hat{\theta}$ of

$$s'(\theta) + \frac{1}{\sigma^2}(\theta - z) = 0.$$

One finds,

$$\hat{\theta} = h^{-1}(z), \quad h(u) = u + \sigma^2 s'(u). \tag{5}$$

Generally, the inversion in (5) may not be analytically feasible, but a solution may be achieved via an approximate sequence of invertible functions. Several examples of prior distributions on θ for which an analytical maximization is possible and authors provide some examples. Some additional solvable cases can be found in Fan (1997), Hyvärinen (1998), and Wang (1999).

For example, if $\pi(\theta) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|\theta|}$, then $s'(\theta) = \sqrt{2}\,\text{sign}(\theta)$, and $\hat{\theta}(z) = \text{sign}(z)\ \max(0, |z| - \sqrt{2}\sigma^2)$.

For

$$\pi(\theta) \propto e^{-a\theta^2/2 - b|\theta|},\ a, b > 0,$$

i.e., if $s'(\theta) = a\,\theta + b\,\text{sign}(\theta)$, the MAP rule is

$$\hat{\theta}(z) = \frac{1}{1+\sigma^2 a}\text{sign}(z)\ \max(0, |z| - b\,\sigma^2).$$

If π is a "supergaussian" probability density,

$$\pi(\theta) \propto \left[\sqrt{\alpha(\alpha+1)} + \left|\frac{\theta}{b}\right|\ \right]^{\alpha+3},$$

the corresponding MAP rule is

$$\hat{\theta}(z) = \text{sign}(z)\max\left(0, \frac{|z|-ab}{2} + \frac{1}{2}\sqrt{(|z|+ab)^2 - 4\sigma^2(\alpha+3)}\right), \tag{6}$$

where $a = \sqrt{\alpha(\alpha+1)/2}$, and $\hat{\theta}(z)$ is set to 0 if the square root in (6) is imaginary.

Leporini and Pesquet (1998) explore cases for which the prior is an exponential power distribution [$\mathcal{EPD}(\alpha, \beta)$]. If the noise also has an $\mathcal{EPD}(a, b)$ distribution with $0 < \beta < b \leq 1$, this MAP solution is a hard-thresholding rule. If $0 < \beta \leq 1 < b$ then the resulting MAP rule is

$$\hat{\theta}(z) = z - \left(\frac{\beta a^b}{b\alpha^\beta}\right)^{1/(b-1)} |z|^{(\beta-1)/(b-1)} + o(|z|^{(\beta-1)/(b-1)}).$$

The same authors consider also the Cauchy noise and explore properties of the resulting rules. When the priors are hierarchical (mixtures) Leporini, Pesquet, and Krim (1999) demonstrated that the MAP solution can be degenerated and suggested Maximum Generalized Marginal Likelihood method. Some related derivations can be found in Chambolle et al. (1998) and Leporini and Pesquet (1999).

The EM algorithm can be readily adapted to Bayesian context to maximize the posterior distributions. The benefit of MAPs over other posterior location measures is that the maximum of the posterior $\pi(\psi|y)$, if it exists, coincides with the maximum of the product of the likelihood and prior $f(y|\psi)\pi(\psi)$, the unnormalized posterior. By taking the logarithm one can see that finding the MAP estimator amounts to maximizing

$$\log \pi(\psi|y) = \log L(\psi) + \log \pi(\psi).$$

The EM algorithm can be readily implemented as follows:
E-Step. At $(k+1)$st iteration calculate

$$\mathbb{E}_{\psi^{(k)}}\{\log \pi(\psi|x)|y\} = Q(\psi, \psi^{(k)}) + \log \pi(\psi).$$

The E-Step coincides with the traditional EM algorithm, only calculation is in finding $Q(\psi, \psi^{(k)})$.
M-Step. Choose $\psi^{(k+1)}$ to maximize $Q(\psi, \psi^{(k)}) + \log \pi(\psi)$. The M-Step differs from the EM, since the objective function to be maximized over all ψ's contains an additive factor, from the prior. However, the

presence of this additional term will make the objective function more concave and thus increase the speed of convergence.

Fisher's Genomic Example Revisited, with a Bayesian Hat. Assume that we elicit Beta, $Beta(\nu_1, \nu_2)$, prior on unknown ψ,

$$\pi(\psi) = \frac{1}{B(\nu_1, \nu_2)} \psi^{\nu_1 - 1}(1-\psi)^{\nu_2 - 1},$$

where $B(\nu_1, \nu_2) = \int_0^1 \psi^{\nu_1 - 1}(1-\psi)^{\nu_2 - 1} d\psi$ is the standard Beta function. The Beta distribution is a natural conjugate for the missing data distribution, say for y_{12} which is $Bin(y_1, \frac{\psi/4}{1/2+\psi/4})$.

Thus, the log-posterior (additive constants ignored) is

$$\begin{aligned} \log \pi(\psi|x) &= \log L(\psi) + \log p(\psi) \\ &= (y_{12} + y_4 + \nu_1 - 1)\log\psi + (y_2 + y_3 + \nu_2 - 1)\log(1-\psi). \end{aligned}$$

The **E-Step** is done by replacing y_{12} by its conditional expectation $y_1 \frac{\psi^{(k)}/4}{1/2+\psi^{(k)}/4}$. This step coincides with the standard EM algorithm.

The **M-Step**, resulting in the $(k+1)$st iteration of the MAP, is

$$\psi^{(k+1)} = \frac{y_{12}^{(k)} + y_4 + \nu_1 - 1}{y_{12}^{(k)} + y_2 + y_3 + y_4 + \nu_1 + \nu_2 - 2}.$$

When the Beta distribution is uniform (i.e., for $\nu_1 = \nu_2 = 1$), the MAP solution coincides with the standard ML solution.

Exercise. Adapt the matlab program `emexmple.m` to accommodate Beta $\mathcal{B}e(\alpha, \beta)$ prior on ψ.

1.4 EM via Monte Carlo

One of the difficulties of general EM is in finding the expected log-likelihood when the augmented data enter the complete likelihood in a non-linear fashion (unlike the Fisher's example). Wei and Tanner (1990) proposed Monte Carlo approach in finding expected log-likelihood $Q(\psi, \psi^{(k)})$. Their proposal consists of generating the variates $Z_1, Z_2, \ldots, Z_m$ from the conditional distribution $h(z|y,\psi) = f(x|\psi)/g(y|\psi)$ (recall $x = [yz]$, where y was observed and z is the completion). Then in the M-step, one maximizes an approximation to $Q(\psi, \psi^{(k)})$,

$$\hat{Q}(\psi, \psi^{(k)}) = \frac{1}{m}\sum_{i=1}^{m} \log L^c(\psi|y, Z_i).$$

Since

$$\frac{1}{m}\sum_{i=1}^{m} \log L^c(\psi|y, Z_i) \to Q(\psi, \psi^{(k)}),$$

the Monte Carlo EM converges to regular EM when $m \to \infty$.

Example. In the Fisher's Example, simulate $Y_{11,i},\ i = 1, \ldots, m$, from $\mathcal{B}in\left(y_1, \frac{1/2}{1/2+\psi^{(n-1)}/4}\right)$, producing $\bar{Y}_{11} = \frac{1}{m}\sum_{i=1}^{m} Y_{11,i}$ and $\bar{Y}_{12} = y_1 - \bar{Y}_{11}$. Thus, the subsequent value of the parameter is:

$$\psi^{(n)} = \frac{\bar{Y}_{12} + y_4}{\bar{Y}_{12} + y_2 + y_3 + y_4} = \frac{\bar{Y}_{12} + y_4}{n - \bar{Y}_{11}}.$$

In many situations of interest the model with missing data is weakly identifiable, i.e. the observed data likelihood function exhibits several local maxima of comparable magnitude (up to sampling fluctuations) even for comparatively large sample sizes n. In such situations Monte Carlo EM could lead to inefficient procedures.

Stochastic EM (S-EM) Celeux and Diebolt (1985, 1988) incorporates an S-step which simulates a realization Z^* of the missing data set from the posterior density $\pi(Z|y, \psi^{(k)})$ based on the current estimate, which is then updated by maximizing the likelihood function of the restored data set (y, Z^*) and there is no need to compute $Q(\psi, \psi^{(k)})$. The S-step can be completed through Gibbs sampling or Hastings-Metropolis. See also Diebolt and Ip (1996).

2 Mixtures

Estimating mixtures of distributions is an important task is statistics. Pattern recognition, data mining and other modern statistical tasks often call for mixture estimation. The mixture problem has several levels of difficulty: (i) estimate weights of a fixed number of fully known distributions, (ii) estimate weights of a fixed number of partially specified distributions, and (iii) estimate weights of an unknown number of partially specified distributions. In this handout we will overview tasks (i) and (ii) by applying EM and Gibbs sampling schemes respectively. We will address task (iii) as an illustration of reversible jump MCMC later.

2.1 EM treatment

Suppose we want to estimate weights of a fixed number of fully known distributions. We illustrate the EM approach which introduces unobserved indicators with the goal of simplifying the likelihood. The weights are estimated by the maximum likelihood method. Assume that a sample $X_1, X_2, \ldots, X_n$ comes from the mixture

$$f(x, \omega) = \sum_{j=1}^{k} \omega_j f_j(x),$$

where the weights $0 \le \omega_j \le 1$ are unknown and constitute $(k-1)$-dimensional vector $\omega = (\omega_1, \ldots, \omega_{k-1})$ and $\omega_k = 1 - \omega_1 - \cdots - \omega_{k-1}$. The class-densities $f_j(x)$ are fully specified.

Even in this simplest case when the only parameters are weights ω, the log-likelihood assumes quite complicated form,

$$\sum_{i=1}^{n} \log f(x_i, \omega) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} \omega_j f_j(x_i) \right).$$

The derivatives with respect to ω_j lead to the system of equations, not solvable in a closed form.

But here it comes the **EM**. Augment the data $x = (x_1, \ldots, x_n)$ by an unobservable matrix $z = (z_{ij}, i = 1, \ldots, n; j = 1, \ldots, k)$. The values z_{ij} are indicators, defined as

$$z_{ij} = \begin{cases} 1, & \text{observation } x_i \text{ comes from the distribution } f_j \\ 0, & \text{else} \end{cases}$$

The unobservable matrix z tells us (in an oracular fashion) where the ith observation x_i comes from. Note that each row of z contains only one 1 and $k-1$ 0's. With the augmented data, $x = (y, z)$ the (complete) likelihood takes quite a simple form,

$$\prod_{i=1}^{n} \prod_{j=1}^{k} (\omega_j f_j(x_i))^{z_{ij}} .$$

The complete log-likelihood is

$$\log L_c(\omega) = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} \log \omega_j + C,$$

where $C = \sum_i \sum_j z_{ij} \log f_j(x_i)$ is free of ω.

Assume that the mth iteration of weights is $\omega^{(m)}$ is already obtained. The mth **E-Step** is

$$\mathbb{E}_{\omega^{(m)}}(z_{ij}|x) = \mathbb{P}_{\omega^{(m)}}(z_{ij} = 1|x) = z_{ij}^{(k)},$$

where $z_{ij}^{(m)}$ is the posterior probability of the ith observation coming from the jth mixture-component, f_j, in the iterative step m.

$$z_{ij}^{(m)} = \frac{\omega_j^{(m)} f_j(x_i)}{f(x_i, \omega^{(m)})}.$$

Because the $\log L_c(\omega)$ is linear in the z_{ij}'s, $Q(\omega, \omega^{(m)})$ is simply $\sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij}^{(m)} \log \omega_j + C$. The subsequent **M-Step** is simple; $Q(\omega, \omega^{(m)})$ is maximized by

$$\omega_j^{(m+1)} = \frac{\sum_{i=1}^{n} z_{ij}^{(m)}}{n}.$$

The following m-script (`mixture_cla.m`) illustrates the above. A sample of size 150 is generated from the mixture $f(x) = 0.5\mathcal{N}(-5, 2^2) + 0.3\mathcal{N}(0, 0.5^2) + 0.2\mathcal{N}(2, 1)$, where $\mathcal{N}$ denotes the normal distribution. The mixing weights are estimated by the **EM** algorithm. $M = 20$ iterations of the EM algorithm yielded $\hat{\omega} = (0.4977, 0.2732, 0.2290)$. Figure 2.1 gives the histogram of data, theoretical mixture and the EM estimate.

```
%Mixture by EM algorithm
 clear all
 close all
%-----------------figure defaults
disp('Mixture by EM algorithm')
lw = 2;
set(0, 'DefaultAxesFontSize', 16);
```

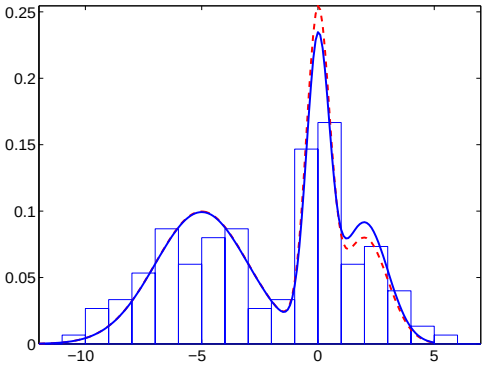

Figure 2: Observations from the $0.5\mathcal{N}(-5, 2^2) + 0.3\mathcal{N}(0, 0.5^2) + 0.2\mathcal{N}(2, 1)$ mixture (histogram), the mixture (dotted line) and EM estimated mixture (solid line).

```
fs = 14;
msize = 5;
%----------------
randn('state',3) %set the seeds (state) to have
rand ('state',3) %the constancy of results
%--------------generate the data ----------------
x=[]; n=150;
for i=1:n
    ra=rand(1,1);
    add = 2 * randn - 5; %N(-5,4) 50% comp 1
    if ra < 0.2
       add =  randn + 2; %N(2,1) 20% comp 2
    elseif ra > 0.7
       add = 0.5*randn; %N(0,0.25) 30% comp 3
    end
   x = [x add];
end
% we got 150 obsercations from
% 0.5  N(-5, 2^2) + 0.3 N(0, 0.5^2) + 0.2 N(2,1).
%------- forget the weights now -----------------------------
k = 3; %number of distributions
omega_current = 1/3 * ones(1, 3); %------ignorance start------
%---------
mus    = [-5   0   2 ];   %-------- we know the distributions
sig2s  = [4  0.25   1 ];  %---- exactly, so means, vars are known
%---------------EM starts!-------------------------------------
for  M = 1 : 20            %---- # of E-M cycles. 20 is plenty!
szs=[]; z=zeros(n,k);
%-------------E-Step, conditional expectations of z's
for i=1:n
    for j=1:k
        z(i,j) = omega_current(j) * 1./(sqrt(2*pi*sig2s(j))) * ...
            exp( -(x(i)-mus(j))^2/(2*sig2s(j)) );
    end
end
%  z(i,j)  are probabilities that observation i is coming
%  from the component j.    But the probs are not normalized.
```

```
%  The following cycle will compute the normalizing constant.
for i=1:n
    sz = 0;
    for j=1:k
        sz = sz + z(i,j);
    end
    szs =[szs, sz];
end
norm=[];
for j=1:k
norm = [norm szs'];
end                          % norm needed to normalize z_{ij}'s
zm = z ./ norm;              % now the z's are normalized
%---------------- M-Step, just plug in zk's
omega_new = sum(zm)/n;       % and repeat...
omega_current = omega_new;
end
omega_current

figure(1)
 histo(x,25,0,1)
 hold on
 xx=-12:0.1:5; omega=[0.5 0.3 0.2]; mixt=0; mixe=0;
 for j=1:k
 mixt = mixt+omega(j) .* 1./(sqrt(2*pi*sig2s(j))) .* ...
            exp( -(xx-mus(j)).^2./(2*sig2s(j)) );
    end
 plot(xx, mixt, 'r--','linewidth',lw)
 for j=1:k
 mixe = mixe+omega_current(j) .* 1./(sqrt(2*pi*sig2s(j))) .* ...
            exp( -(xx-mus(j)).^2./(2*sig2s(j)) );
 end
 plot(xx, mixe, 'b-','linewidth',lw)
 axis tight
 print -depsc 'C:\Brani\Courses\Bayes\Handouts\Working12\Figs\mix_em.eps'
```

2.2 Gibbs Sampling Treatment

As an example of the MCMC approach to mixture estimation, consider the model in which observations $x_1, \ldots, x_n$ follow the model

$$f(x|\psi) = \sum_{j=1}^{k} \omega_j f_j(x|\mu_j, \sigma^2),$$

where $f_j(x|\mu, \sigma^2)$ are normal densities with unknown means μ_j and unknown common variance σ^2. The parameter vector is $\psi = (\omega, \mu, \sigma^2) = (\omega_1, \ldots, \omega_k, \mu_1, \ldots, \mu_k, \sigma^2)$.

Let the prior on the weights be Dirichlet,

$$[\omega] \sim \mathcal{D}ir(\alpha, \alpha, \ldots, \alpha).$$

The prior locations and variance have independent priors,

$$\begin{aligned} [\mu_j] &\sim \mathcal{N}(0, \tau^2), \ \textit{i.i.d.}; \ j = 1, \ldots, k \\ [\sigma^2] &\sim \mathcal{IG}amma(\delta, \gamma). \end{aligned}$$

The joint posterior is proportional to the joint distribution,

$$\pi(\omega, \mu, \sigma^2|x) \propto f(x, \omega, \mu, \sigma^2),$$

which is equal to

$$f(x|\omega, \mu, \sigma^2)\pi(\omega, \mu, \sigma^2) = f(x|\omega, \mu, \sigma^2)\pi(\mu)\pi(\omega)\pi(\sigma^2).$$

Now, by plugging the component models in the previous expression we obtain the unnormalized posterior,

$$\begin{aligned}\pi(\omega, \mu, \sigma^2|x) \quad \propto \quad & \prod_{i=1}^{n}\left[\sum_{j=1}^{k}\omega_j \frac{1}{\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu_j)^2\right\}\right] \quad \times \quad \prod_{j=1}^{k}\exp\left\{-\frac{\mu_j^2}{2\tau^2}\right\} \\ & \times \quad \prod_{j=1}^{k}{\omega_j}^{\alpha} \quad \times \quad \frac{1}{\sigma^{2(\gamma+1)}}\exp\left\{-\frac{\delta}{\sigma^2}\right\}\end{aligned}$$

Instead of dealing with this complex posterior, we introduce model indicators. Let $s_i \in \{1, 2, \ldots, k\}$ be the indicator of model from which the observation x_i is coming. Then the likelihood

$$f(x|\psi) = \sum_{j=1}^{k}\omega_j f_j(x|\mu_j, \sigma^2),$$

can be replaced by

$$\begin{aligned}[x|s = j, \mu_j, \sigma^2] \quad &\sim \quad f_j(x|\mu_j, \sigma^2) \\ P(s = j) &= \omega_j.\end{aligned}$$

Here s is the indicator for x. Since s_i are independent and identically distributed,

$$\pi(s_1, s_2, \ldots, s_n|\omega_1, \ldots, \omega_k) = \prod_{i=1}^{n} P(s_i|\omega) = \prod_{j=1}^{k}\omega_j^{n_j},$$

where n_j is the number of observations coming from jth density, i.e., $n_j = \#s_i$ such that $s_i = j$, $i = 1, \ldots, n$. The posterior is proportional to

$$\begin{aligned}f(x, s, \omega, \mu, \sigma^2) \quad \propto \quad & f(x|s, \omega, \mu, \sigma^2)\pi(s|\omega)\pi(\omega)\pi(\mu)\pi(\sigma^2) \\ = \quad & \frac{1}{\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_{s_i})^2\right\} \times \prod_{i=1}^{n}\omega_{s_i} \times \prod_{j=1}^{k}{\omega_j}^{\alpha} \times \\ & \times \exp\left\{-\frac{1}{2\tau^2}\sum_{j=1}^{k}\mu_j^2\right\} \times \frac{1}{\sigma^{2(\gamma+1)}}\exp\left\{-\frac{\delta}{\sigma^2}\right\}.\end{aligned}$$

To apply the Gibbs sampler we need to find full conditionals. Since

$$\pi(\omega|\mu, \sigma^2, s, x) \propto f(x, s, \omega, \mu, \sigma^2) \propto \pi(s|\omega)\pi(\omega) = \prod_{i=1}^{n}\omega_{s_i} \prod_{j=1}^{k}{\omega_j}^{\alpha},$$

the full conditional on ω is Dirichlet,

$$[\omega|\mu,\sigma^2,s,x] \sim \mathcal{D}ir(\alpha+n_1,\ldots,\alpha+n_k).$$

The full conditional on μ is multivariate normal, with independent components. From $f(x,s,\omega,\mu,\sigma^2)$ we choose factors containing μ and normalize,

$$\begin{aligned}
\pi(\mu|\omega,\sigma^2,s,x) &\propto \exp\left\{-\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu_{s_i})^2\right\}\times\exp\left\{-\frac{1}{2\tau^2}\sum_{j=1}^{k}\mu_j^2\right\}\\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{k}\left(-2\mu_j\sum_{i|s_i=j}x_i+n_j\mu_j^2\right)-\frac{1}{2\tau^2}\sum_{j=1}^{k}\mu_j^2\right\}\\
&\propto \exp\left\{\sum_{j=1}^{k}\left[\left(-\frac{n_j}{2\sigma^2}-\frac{1}{2\tau^2}\right)\mu_j^2+\frac{1}{\sigma^2}\mu_j\sum_{i|s_i=j}x_i\right]\right\}\\
&\propto \prod_{j=1}^{k}\exp\left\{-\frac{1}{2}\left(\frac{n_j}{\sigma^2}+\frac{1}{\tau^2}\right)\left[\mu_j^2-2\frac{\frac{1}{\sigma^2}\sum_{i|s_i=j}x_i}{\frac{n_j}{\sigma^2}+\frac{1}{\tau^2}}\mu_j+\left(\frac{\frac{1}{\sigma^2}\sum_{i|s_i=j}x_i}{\frac{n_j}{\sigma^2}+\frac{1}{\tau^2}}\right)^2\right]\right\}.
\end{aligned}$$

Thus,

$$[\mu_j|\omega,\sigma^2,s,x] \sim \mathcal{N}\left(\frac{\tau^2\sum_{i|s_i=j}x_i}{n_j\tau^2+\sigma^2},\left(\frac{n_j}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\right),\ \textit{i.i.d}\ \ j=1,\ldots,k.$$

The common variance has two factors in the joint distribution, $f(x|s,\omega,\mu,\sigma^2)\pi(\sigma^2)$,

$$\pi(\sigma^2|\omega,\mu,s,x) \propto \frac{1}{\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_{s_i})^2\right\}\ \times\ \frac{1}{\sigma^{2(\gamma+1)}}\exp\left\{-\frac{\delta}{\sigma^2}\right\}.$$

This is the inverse gamma distribution,

$$[\sigma^2|\omega,\mu,s,x] \sim \mathcal{IG}amma\left(\gamma+\frac{n}{2},\delta+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_{s_i})^2\right).$$

Finally, the full conditional for s is proportional to $f(x|s,\mu,\sigma^2)\pi(s|\omega)$,

$$\begin{aligned}
\pi(s|\omega,\mu,\sigma^2,x) &\propto \frac{1}{\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_{s_i})^2\right\}\ \times\ \prod_{j=1}^{k}{\omega_j}^{n_j}\\
&\propto \prod_{i=1}^{n}\omega_{s_i}\exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu_{s_i})^2\right\}.
\end{aligned}$$

By independence,

$$\mathbb{P}(s_i=j) \propto \omega_j\exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu_j)^2\right\}.$$

All conditionals are found and the Gibbs sampler can be started. As an illustration we generated $n = 40$ observations from the mixture $0.1\mathcal{N}(1, 2^2) + 0.7\mathcal{N}(9, 2^2) + 0.2\mathcal{N}(20, 2^2)$. Figure 2.2 shows posterior histograms for all parameters (panels (a-c)) and performance of the algorithm on simulated data. Quite amazing that only 40 observations produced good estimators for 7 parameters! The matlab program implementing the above Gibbs sampler is provided next. You can find it on our web page as `gibbs.m`.

```
clear all
close all
%--------------------------------
  lw = 2.5;
  set(0, 'DefaultAxesFontSize', 16);
  fs = 16;
  msize = 10;
%---------------------------------------
disp('Gibbs for Mixture')
rand('seed', 1)   %initialize
randn('seed',1)   %random number generators
% simulate data--------------------------
n=40; us = rand(1,n);
for i=1:n
    if us(i) < 0.1
          x(i) = 2 * randn +  1; %0.1 of N(1,2^2)
    elseif us(i) > 0.8
          x(i) = 2 * randn + 20; %0.2 of N(20,2^2)
    else  x(i) = 2 * randn +  9; %0.7 of N(9,2^2)
    end
end
x = sort(x);
%fixed (hyper)parameters
k=3;
tau =     12;
alpha =   5;
delta =   30;
gamma =   10;
iterations = 5000;
burn = 1000;
%-------------------------
% initial parameters
sig2 = 5;
s0=repeat((1:k)', ceil(n/k));
s1 = s0'; s2 = s1(:); s = s2(1:n); %s is latent variable
w  = repeat(1/k, k);  %equal weight
mu = repeat(10,k);
%
%---------for iter = 1:iterations
mus=[]; sig2s=[]; ws=[];
  h=waitbar(0,'Simulation in progress');
for iter=1: burn + iterations
for i=1:n
   for j=1:k
    sij(i,j)= (s(i)==j);
   end
end
nj=sum(sij);
w = (rand_dirichlet(alpha + nj, 1))';
```

```
%---------------------------
for j = 1:k
    sj(j) = sum(x(s==j));
end
mu = sqrt( 1./( nj/sig2 + 1/tau^2)) .* randn(1,k) + (tau^2.*sj)./(nj*tau^2+sig2);
%---------------------------
for j = 1:k
    sm2j(j) = sum ( (x(s==j) - mu(j)).^2 );
end
sig2 = 1./rand_gamma( gamma + n/2,   delta + 1/2 * sum(sm2j), 1, 1);
%------------------------------------------------------------

pr = zeros(n,k);
for i = 1:n
    for j=1:k
        pr(i,j) = w(j)*exp( - (x(i) - mu(j))^2 /(2 * sig2));
    end
end
tot= repeat((sum(pr'))',3);
pr=pr./tot;
for i=1:n
 [aa,bb]=rand_multinomial(1,pr(i,:));
  s(i) = find(bb==1);
end
%------------------------------------------------------------
mus=[mus; mu];
sig2s=[sig2s, sig2];
ws = [ws; w];
  waitbar(iter/(burn+iterations))
end
close(h)
muss=mus(burn+1:end, :);
mmu= mean(muss)
sig2ss = sig2s(burn+1: end);
msig2= mean(sig2ss)
wss = ws(burn+1:end, :);
mw = mean(wss)

figure(1)
subplot(3,1,1)
hist(muss(:,1),50)
subplot(3,1,2)
hist(muss(:,2),50)
subplot(3,1,3)
hist(muss(:,3),50)
print -depsc 'C:\Brani\Courses\Bayes\Handouts\Working12\Figs\MixGibbs1.eps'

figure(2)
hist(sig2ss, 50)
print -depsc 'C:\Brani\Courses\Bayes\Handouts\Working12\Figs\MixGibbs2.eps'
figure(3)
subplot(3,1,1)
hist(wss(:,1),50)
subplot(3,1,2)
hist(wss(:,2),50)
subplot(3,1,3)
```

```
hist(wss(:,3),50)
print -depsc 'C:\Brani\Courses\Bayes\Handouts\Working12\Figs\MixGibbs3.eps'

figure(4)
histo(x,30,0,1)
hold on
cee=linspace(-5,25,300);
est=mw(1).*1./sqrt(2 * pi * msig2).*exp(-1/(2*msig2) * (cee - mmu(1)).^2)+...
    mw(2).*1./sqrt(2 * pi * msig2).*exp(-1/(2*msig2) * (cee - mmu(2)).^2)+...
    mw(3).*1./sqrt(2 * pi * msig2).*exp(-1/(2*msig2) * (cee - mmu(3)).^2);
plot(cee, est,'b-')
 theo = 0.1 * 1./sqrt(2 * pi * 4).*exp(-1/(2*4) * (cee - 1).^2)+...
        0.2 * 1./sqrt(2 * pi * 4).*exp(-1/(2*4) * (cee - 20).^2)+...
        0.7 * 1./sqrt(2 * pi * 4).*exp(-1/(2*4) * (cee - 9).^2);
plot(cee, theo,'r--')
print -depsc 'C:\Brani\Courses\Bayes\Handouts\Working12\Figs\MixGibbs4.eps'
```

3 Exercises

1. (Example from Arslan et al. (1993)) The observed data $y = (-20, 1, 2, 3)$ is assumed to follow a Students t-distribution with 0.05 degrees of freedom and unknown location parameter μ. The log-likelihood given by

$$\log L(\theta|y) = -0.525 \sum_{i=1}^{4} \log\{0.05 + (y_i - \theta)^2\},$$

does not admit a closed form solution for the MLE of μ. In a complete data formulation with $x = [y, z]$, the missing variables $z = (z_1, \ldots, z_4)$ are defined so that $Y_i|z_i \sim \mathcal{N}(\mu, 1/z_i)$, independently for $i = 1, \ldots, 4$ and $Z_i \sim \mathcal{G}amma(0.025, 0, 025)$. The complete log-likelihood can be written as

$$\log f(x|\theta) = C - 0.475 \sum_{i=1}^{4} \log z_i - 0.025 \sum_{i=1}^{4} z_i - 0.5 \sum_{i=1}^{4} z_i (y_i - \theta)^2.$$

The conditional distribution of Z_i given y_i is gamma,

$$Z_i|y_i \sim \mathcal{G}amma\left(0.525, 0.025 + \frac{(y_i - \theta)^2}{2}\right).$$

(a) Plot the log-likelihood function. Show (numerically) that there are 4 local maxima at

$$\theta_1 = -19.993, \theta_2 = 1.086, \theta_3 = 1.997, \text{ and } \theta_4 = 2.906.$$

(b) Show that the recursions induced by the EM algorithm are defined by

$$\theta^{(k)} = \frac{\sum_{i=1}^{4} y_i \varphi_i(\theta^{(k-1)})}{\sum_{i=1}^{4} \varphi_i(\theta^{(k-1)})},$$

where $\varphi_i(\theta) = 1.05 \left(0.05 + (y_i - \theta)^2\right)^{-1}$.

(c) Apply the recursions from (b) to starting values $(-30, -18, 1.5, 2.5, 30)$. Are any of the maxima from (a) missed?

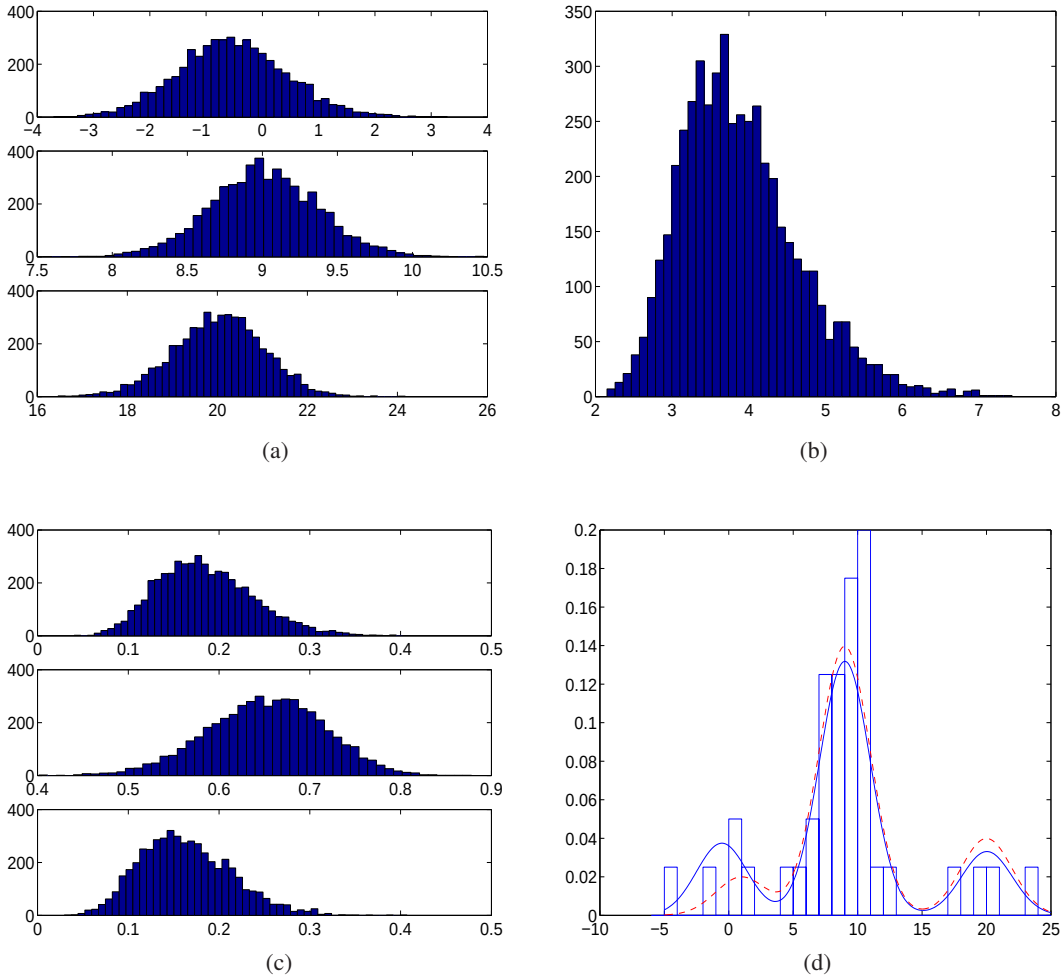

Figure 3: Histograms of traces for (a) μ_1, μ_2, μ_3; (b) σ^2; (c) $\omega_1, \omega_2, \omega_3$; (d) Sample of size 40 (normalized histogram), true mixture (dashed), MCMC estimator (solid).

References

[1] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion). *Journal of the Royal Statistical Society, B,* **39**, 1–38.

[2] Fisher, R.A. and Balmukand, B. (1928). The estimation of linkage from the offspring of selfed heterozygotes. *Journal of Genetics,* **20**, 79–92.

[3] Healy M.J.R., Westmacott M.H. (1956). Missing values in experiments analysed on automatic computers. *Appl. Statistics*,**5**, 203–306.

[4] Arslan, O., Constable, P.D.L. and Kent, J.T. (1993). Domains of convergence for the EM algorithm: A cautionary tale in a location estimation problem. *Statistical Computing*, **3**, 103–108.

[5] McKendrick, A. G. (1926). Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society,* **44**, 98–130.

[6] McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions,* Wiley, NY.

[7] Newcomb, S. (1980). A Generalized Theory of the Combination of Observations so as to Obtain the Best Result. In Stigler, M., Editor, *American Contributions to Mathematical Statistics in the Nineteenth Century,* Volume 2, 343-366. Arno Press, New York.

[8] Rao, C. R. (1973) *Linear Statistical Inference and its Applications,* Second Edition, Wiley, 368–369.

[9] Abramovich, F., Sapatinas, T., and Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *J. Roy. Statist. Soc. Ser. B*, **60**, 725–749.

[10] Berger, J. and Pericchi, L. (1996). The Intrinsic Bayes Factor for Model Selection and Prediction. *J. Amer. Statist. Assoc.*,**91**, 109–122.

[11] Chambolle, A., DeVore, R. A., N-Y. Lee N-Y., and Lucier, B. J. (1998). Nonlinear Wavelet Image Processing: Variational Problems, Compression, and Noise Removal through Wavelet Shrinkage, *IEEE Trans. Image Process.*,**7**, 319–355.

[12] Chipman, H., Kolaczyk, E., and McCulloch, R. (1997). Adaptive Bayesian Wavelet Shrinkage. *J. Amer. Statist. Assoc.*,**92**, 1413–1421.

[13] Clyde, M. A., Parmigiani, G., and Vidakovic, B. (1998). Multiple Shrinkage and Subset Selection in Wavelets. *Biometrika,* **85**, 391–402.

[14] Celeux, G. and Diebolt, J. (1985). The SEM algorithm: A probabilisitic teacher algorithm derived from the EM algorithm for the mixture problem. *Comput. Statist. Quater.*,**2**, 73–82.

[15] Celeux, G. and Diebolt, J. (1988). A probabilistic teacher algorithm for iterative maximum likelihood estimation. *Classification and Related Methods of Data Analysis*, (H. H. Bock, Editor.), North-Holland, 617–623.

[16] Diebolt, J. and Ip, E. H. S. (1996). A stochastic EM algorithm for approximating the maximum likelihood estimate. In: *Markov Chain Monte Carlo in Practice,* (W.R. Gilks, S.T. Richardson and D.J. Spiegelhalter, Eds.). Chapman and Hall.

[17] Fan, J. (1997), Comment on "Wavelets in Statistics: A Review" by A. Antoniadis, *Italian Jour. Statist.*,**6**, 97–144.

[18] Hyvärinen, A. (1998). Sparse Code Shrinkage: Denoising of Nongaussian Data by Maximum Likelihood Estimation, Technical Report A51, Helsinki University of Technology, Finland.

[19] Leporini, D., and Pesquet, J.-C. (1998). Wavelet Thresholding for a Wide Class of Noise Distributions, EUSIPCO'98, Rhodes, Greece, September 1998, 993–996.

[20] Leporini, D., and Pesquet, J.-C. (1999). Bayesian wavelet denoising: Besov priors and non-gaussian noises, *Signal Processing,* **81**, 55–67.

[21] Leporini, D., Pesquet, J.-C., Krim, H. (1999). Best Basis Representations with Prior Statistical Models, In: *Bayesian Inference In Wavelet Based Models*, Editors P. Müller and B. Vidakovic, Lecture Notes in Statistics,**141**, 109–113, Springer Verlag, New York.

[22] Pesquet, J.-C., Krim, H., Leporini, D., and Hamman, E. (1996). Bayesian approach to best basis selection, Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 7-10 May, Atlanta, GA. **5**, 2634–2637.

[23] Vidakovic, B. (1998a). Nonlinear Wavelet Shrinkage With Bayes Rules and Bayes Factors, *Journal of the American Statistical Association,* **93**, 441, 173–179.

[24] Vidakovic, B. (1998b). Wavelet-based nonparametric Bayes methods. In: Practical Nonparametric and Semiparametric Bayesian Statistics. Editors D. Dey, P. Müller and D. Sinha, Lecture Notes in Statistics **133** , 133–155, Springer-Verlag, New York.

[25] Wang, Y. (1999). An Overview of Wavelet Regularization, In: *Bayesian Inference In Wavelet Based Models*, Editors P. Müller and B. Vidakovic, Lecture Notes in Statistics,**141**, 109–113, Springer Verlag, New York.

[26] Wei, G.C.G. & Tanner, M. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation. *J. Amer. Statist. Assoc.*, **85**, 699–704.