Improvements to Expectation-Maximization approach for unsupervised classification of remote sensing data

Thales Sehn Korting[1]
Luciano Vieira Dutra[1], **Leila Maria Garcia Fonseca**[1]
Guaraci Erthal[1], **Felipe Castro da Silva**[1]

[1]Image Processing Division
National Institute for Space Research – INPE
São José dos Campos – SP, Brazil

tkorting, dutra, leila, gaia, felipe@dpi.inpe.br

Abstract. *In statistical pattern recognition, mixture models allow a formal approach to unsupervised learning. This work aims to present a modification of the Expectation-Maximization clustering method applied to remote sensing images. The stability of its convergence has been increased by supplying the results of the well-known K-Means algorithm, as seed points. Hence, the accuracy has been improved by applying cluster validity measures to each configuration, varying the initial number of clusters. High-resolution urban scenes has been tested, and we show a comparison to supervised classification results. Performance tests were also realized, showing the improvements of our proposal, in comparison to the original one.*

1. Introduction

Generally, a color composition of some remote sensing image behaves as a mixture of several colors, which changes gradually according x and y pixel positions. If a specialist performs a manual classification in a certain image, and after views its scatter plot, the classes will appear together, in such a way that linear classification algorithms will not have success when classifying it. Figure 1 shows one example of this idea.

In this Figure, we used 6 classes, namely *Streets, Pools, Roofs, Shadows, Greens*, and *Others*. By visualizing the scatter plots, which draws the pixel occurrence and also pixel class for bands RG, RB and GB, it seams clear that classes named roofs and swimming pools are linearly separable from the rest, as shown in the second scatter plot (Figure 1c). However, the other 4 classes remain together, and it's a challenging task to discover their statistical distributions. Each class can be thought as an independent variable; as they are a fraction of a total (the entire image), it characterizes a mixture model.

One way to estimate mixture models is to assume that data points have "membership" in one of the distributions present in the data. At first, such membership is unknown. The objective is to estimate suitable parameters for the model, where the connection to the data points is represented as their membership in the individual model distributions.

In statistical pattern recognition, such mixture models allow a formal approach to unsupervised learning (*i.e.* clustering) [Figueiredo and Jain 2002]. A standard method to fit finite mixture models to observed data is the *Expectation-Maximization* (EM) algorithm, first proposed by [Dempster et al. 1977]. EM is an iterative procedure which

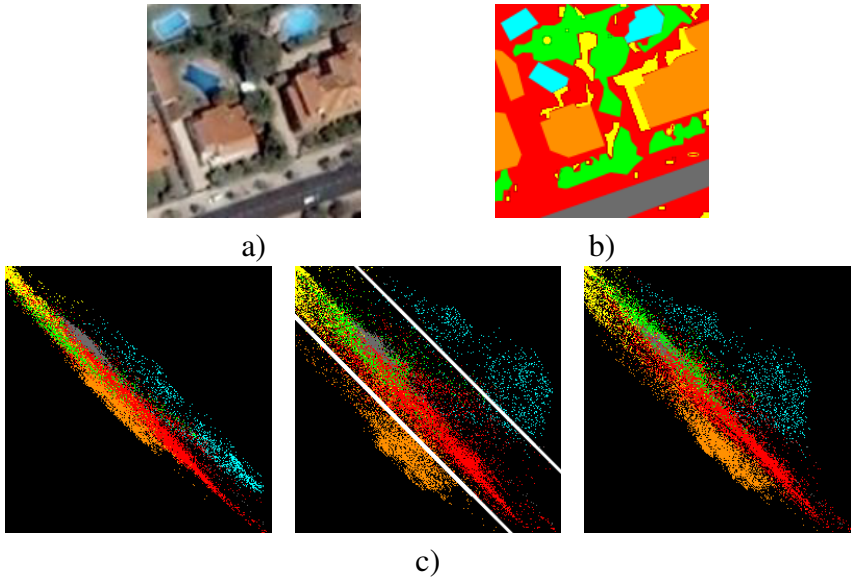


Figure 1. a) Example remote sensing image. b) Manual Classification. c) Scatter Plots of bands RG, RB and GB considering manual classification.

converges to a (local) maximum of the marginal *a posteriori* probability function without manipulating the marginal likelihood $p(\theta|\mathbf{x})$ [Figueiredo 2004]:

$$p(\theta|\mathbf{x}) = p(\mathbf{x}|\theta)p(\theta) \tag{1}$$

where θ is a set of unknown parameters from $\mathbf{x}$. Therefore, EM estimates the components probabilities present in a certain cluster. In our case, the input is composed by the image pixels, and the parameters are mean and variance.

In other words, EM is a general method of estimating the features of a given data set, when the data are incomplete or have missing values [Bilmes 1998]. This algorithm has been used in several areas, such as image reconstruction [Lay and Katsaggelos 1990, Qian and Titterington 1993, Shepp and Vardi 1982], signal processing, and machine learning [Beal and Ghahramani 2003, Guo and Rodriguez 1992, Lawrence and Reilly 1990].

The finite mixture models are able to represent arbitrarily complex probability density functions [Figueiredo 2004]. This fact makes EM approach proper for representing complex likelihood functions, considering Bayesian inference. Being an iterative procedure, the EM method can present high computational cost. So, in this article we present a variation of the EM algorithm, increasing stability and capability, by providing the first set of parameters from K-Means algorithm and performing clustering validation.

The paper is organized as follows. Section 2 starts explaining the EM approach and its application to mixture models, followed by how to estimate the parameters using such method. After, in Section 3 we show our main contribution describing the "improved EM" approach. We discuss the implemented system, divided by modules on the whole process. Section 4 presents some results when applying the method to urban remote sensing images, and a discussion over the performance achieved using the suggested

improvements. In Section 5 we conclude with some remarks about the results and future works.

2. The standard EM algorithm

An image pixel might behave differently if it comes from an edge rather than a smooth region. Therefore, the global behavior is likely to be a mixture of the two distinctive behaviors [Bouman 1995]. The objective of the mixture distributions is to produce a probabilistic model composed of a subclasses set. In our approach, each class is characterized by a set of parameters describing the mean and variance of the spectral components.

EM algorithm is based on the Bayesian theory. We assume the algorithm will estimate M clusters (or classes) $C_j, j = 1, \ldots M$. For each of the N input vectors $\mathbf{x}_k, k = 1, \ldots N$, the algorithm calculates its probability $P(C_j|\mathbf{x}_k)$ to belong to a certain class [Theodoridis and Koutroumbas 2003]. The highest probability will point to the vector's class.

Being an unsupervised classification method, there is no training stage. The image and the number of clusters to be estimated form the input. The attributes-vector is composed of the pixel-value for each band. So, an image with three bands produces a 3D-space for the whole set, and so on.

2.1. Computing EM

The EM algorithm works iteratively by applying two steps: the E-step (*Expectation*) and the M-step (*Maximization*). Formally, $\hat{\theta}(t) = \{\mu_j(t), \Sigma_j(t)\}, j = 1, \ldots M$ stands for successive parameter estimates. The method aims to approximate $\hat{\theta}(t)$ to real data distribution when $t = 0, 1, \ldots$

E-step: This step calculates the conditional expectation of the complete *a posteriori* probability function;

M-step: This step updates the parameter estimation $\hat{\theta}(t)$.

Each cluster probability, given a certain attribute-vector, is estimated as following:

$$P(C_j|\mathbf{x}) = \frac{|\Sigma_j(t)|^{-GB} e^{\eta_j} P_j(t)}{\sum_{k=1}^{M} |\Sigma_k(t)|^{-GB} e^{\eta_k} P_k(t)} \tag{2}$$

where

$$\eta_i = -\frac{1}{2}(\mathbf{x} - \mu_i(t))^T \Sigma_x^{-1}(t)(\mathbf{x} - \mu_i(t))$$

With such probabilities, one can now estimate the mean, covariance, and the *a priori* probability for each cluster, at time $t + 1$, according to Equations 3, 4, and 5:

$$\mu_j(t+1) = \frac{\sum_{k=1}^{N} P(C_j|\mathbf{x}_k)\mathbf{x}_k}{\sum_{k=1}^{N} P(C_j|\mathbf{x}_k)} \tag{3}$$

$$\Sigma_j(t+1) = \frac{\sum_{k=1}^{N} P(C_j|\mathbf{x}_k)(\mathbf{x}_k - \mu_j(t))(\mathbf{x}_k - \mu_j(t))^T}{\sum_{k=1}^{N} P(C_j|\mathbf{x}_k)} \tag{4}$$

$$P_j(t+1) = \frac{1}{N} \sum_{k=1}^{N} P(C_j|\mathbf{x}_k) \tag{5}$$

These steps are performed until reaching the convergence, according the following equation [Theodoridis and Koutroumbas 2003]:

$$\| \theta(t+1) - \theta(t) \| < \varepsilon \tag{6}$$

where $\| \ . \ \|$, in this implementation, is the Euclidean distance between the vectors $\mu(t+1)$ and $\mu(t)$, and ε is a threshold chosen by the user. After the calculations, Equation 2 is used to classify the image. The next section explains the classification in detail.

3. The "improved EM" approach

Figure 2 shows a diagram composed of four modules, presenting our method, according equations presented in the previous section, and with the contributions presented by this paper.

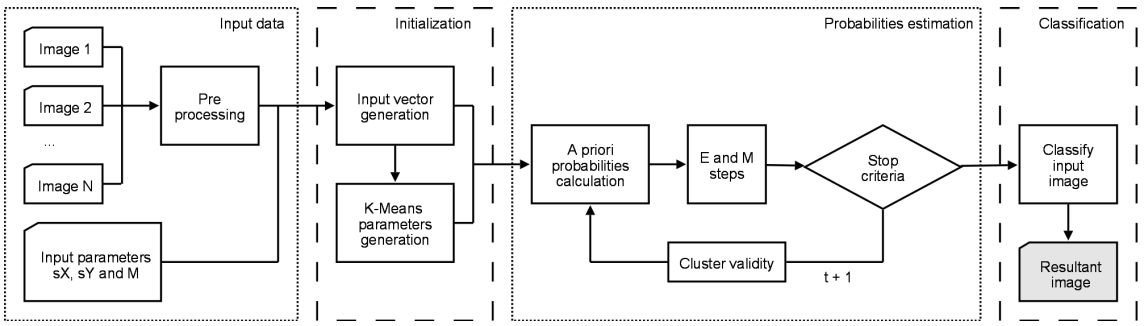


Figure 2. System's diagram.

The implementation follows this script:

Input data: this module deals with N images and the input parameters called sampling rate (sX and sY), on directions x and y. This rate aims to reduce the input data, building the input vector as a fraction of the image pixels. M stands for the number of clusters the algorithm has to estimate. Here we propose a preprocessing stage, removing, from the input data, pixels close to the image border because, because of sensor features, sometimes they are not trustworthy;

Initialization: using the sampling rate, we build the instance set $\mathbf{x}$, and create the θ set, with seed points provided by the K-Means algorithm. On the standard EM implementation, the first set of parameters are randomized, and this is one of the main causes of the high computational cost of this algorithm, and the risk of converging to local minimums;

Probabilities estimation: this module performs the iterative procedure of successive parameter estimation and cluster validity, described below. Such technique aims to certify the number of classes provided by the user, and guarantee that all clusters are distant from each other. While t increases, a test is performed to check if the algorithm has already converged, or a maximum number of iterations have been reached;

Classification: here the final classification is performed. For each of the N pixels $\mathbf{x}_k$ is associated the class with higher probability, that is, find $P(C_j|\mathbf{x}_k) > P(C_i|\mathbf{x}_k), j \neq i$ and classify $\mathbf{x}_k$ as C_j.

The "Initialization" and "Probabilities estimation" modules were adjusted to carry out more stability and capability to the results. We introduced the solution to use K-Means for producing the first set of unknown parameters θ, *i.e.* when $t = 0$. Applying this to the EM approach, we reduce the number of iterations, thus reducing computational time.

Sometimes, the algorithm is not able to converge, during the "Probabilities estimation" module, to the entire set of classes, because of the mixture models natural behavior. On our approach, we modified each iteration of this module by validating the current clustering arrangement. During convergence, if a cluster center is approaching another one, then one of them is randomly modified for the next iteration. This aims to "shake" the values, so that cluster C_j may converge to another class, far from C_i in the attribute space.

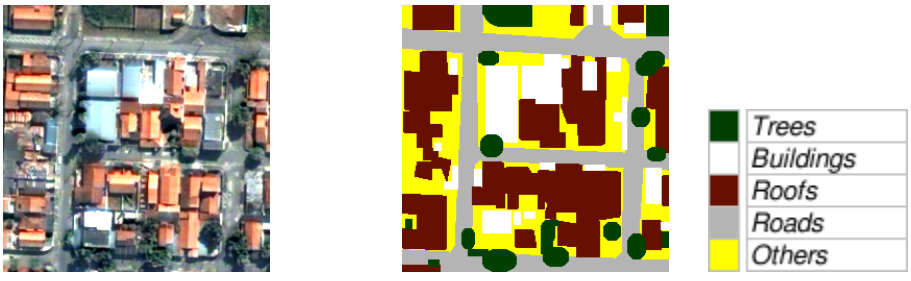
Considering clustering validation, we also perform cluster exclusion when some of them have a low probability. It was implemented because sometimes the user-supplied parameters can have a mistaken number of parameters, or the attributes distribution doesn't allow detecting a certain number of clusters. Through a threshold η, the cluster exclusion is implemented according the following equation:

$$\text{if } P_i(t) < \eta \text{ then exclude cluster } C_i \tag{7}$$

4. Results

This section presents some results, applying the EM algorithm to classify remote sensing images.

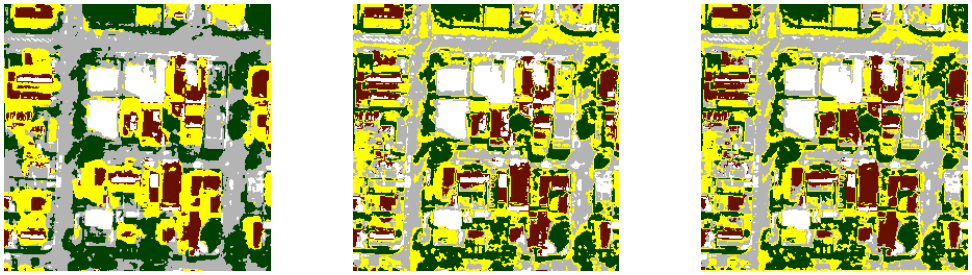
Firstly we have a color composition of an urban area from São José dos Campos – Brazil. Such image was taken in January 2004, from QuickBird, and the composition is R3G2B1. Figure 3 shows the original image and its manual classification, with respect to 5 classes, namely *Trees, Buildings, Roofs, Roads*, and *Others*. In order to analyze the results and compare it with another known methods, we performed the classification using three algorithms: improved EM, KMeans and Maximum Likelihood (ML). The classification results are shown in Figure 4.



a) b)

Figure 3. a) Color composition R3G2B1 of QuickBird scene from São José dos Campos – Brazil. b) Manual classification.

To prove the enhancement on the results, by the use of our improved EM approach, we show on Table 1 the agreement matrices for each of the tested algorithms.



a) b) c)

Figure 4. Classification results using: a) Improved EM, b) KMeans, and c) ML.

By observing the tables we can note that the only algorithm which achieved more than 70% of correct matches was the improved EM method. We should already expect better results than KMeans, since it provides the first set of parameters, and improved EM adjust them in a better way. However, the ML algorithm is supervised, and the training stage was performed with enough samples. Another point that must be taken into consideration is the low matches of the last class, named *Others*. Even being a bad result, this class stands for the less important set of objects in the scene, that even was not classified by the speciallist.

Since the improved EM algorithm got good results for "urban classes", like *Trees* and *Roads*, also better than the other algorithms, we must point out that our approach can be used successfully for such kind of image classification.

Table 1. Agreement Matrices for: a) Improved EM, b) KMeans, and c) ML. Classes are: 1) Trees, 2) Buildings, 3) Roofs, 4) Roads, and 5) Others.

	1	2	3	4	5
1	0,87	0,05	0,05	0,25	0,40
2	0,00	0,57	0,07	0,03	0,03
3	0,00	0,02	0,31	0,00	0,01
4	0,10	0,28	0,05	0,70	0,39
5	0,03	0,08	0,53	0,02	0,17

a)

	1	2	3	4	5
1	0,66	0,07	0,15	0,19	0,30
2	0,01	0,59	0,10	0,06	0,04
3	0,00	0,01	0,35	0,01	0,00
4	0,03	0,23	0,09	0,49	0,25
5	0,30	0,09	0,31	0,26	0,40

b)

	1	2	3	4	5
1	0,67	0,07	0,16	0,18	0,31
2	0,00	0,56	0,06	0,05	0,03
3	0,00	0,01	0,36	0,00	0,01
4	0,03	0,27	0,13	0,48	0,26
5	0,29	0,09	0,29	0,30	0,39

c)

Figure 5a shows a CBERS-2 color composition of bands 2, 3, and 4. Three classes are identified on this image, namely *Urban, Vegetation*, and *Water*. Figures 5b and 5c show, respectively, the scatter plot and the classified image for different classification methods[1]: EM, ML, and Euclidean Distance (ED). We show the scatter plots, so that the reader is able to perceive the mixture model present in such image, and also to draw the classification result, since the classes are exposed on each combination of bands RG, RB and GB. And, in comparison to the other approaches, EM got the smoothest thematic

[1]Software SPRING was used to perform such classifications [Câmara et al. 1996]

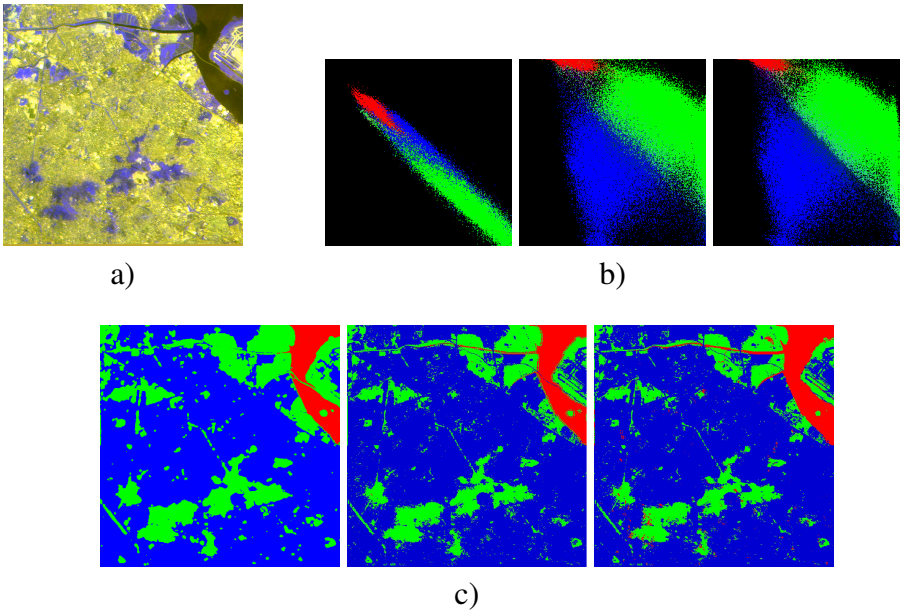


Figure 5. a) CBERS-2 color composition R2G3B4. b) Scatter plots. c) Classification (from left to right) using EM, ML, and ED methods.

map. It is important to point out that both methods (ML and ED) are supervised; however, visually the EM result is more satisfactory, as we can observe comparing results on Figure 5c.

4.1. Discussion

EM algorithm presents some drawbacks. Being a local method, it is sensitive to initialization because the likelihood function of a mixture model is not unimodal [Figueiredo and Jain 2002]. This was the main cause for using K-Means as first set of parameters. For certain mixture types, it may converge to the parameter space boundary, leading to meaningless estimates.

However, to test the performance increase we have performed several tests, using the original EM proposal, and the modified approach. The tests considered processing time until convergence, for both approaches. We used 5 different images and input parameters, so the final increasing in performance is unbiased. Table 2 shows the results, considering image size, number of classes for each one, and computational time until convergence.

Table 2. Comparison between original and improved approaches.

	Image1	Image2	Image3	Image4	Image5
Image size	512×512	512×512	200×200	512×384	264×377
# of classes	4	4	5	6	5
Δt_1 original EM	$467s$	$467s$	$103s$	$402s$	$202s$
Δt_2 improved EM	$140s$	$148s$	$29s$	$105s$	$70s$
$\Delta t_1/\Delta t_2$	3.335	3.155	3.551	3.828	2.885

Calculating the average values for time decrease, showed in Table 2 at the line $\Delta t_1/\Delta t_2$, we reach the value 3.35. This means that our improved approach is around

$3\times$ faster than the original, and considering the showed results, its also more robust to outliers.

Images classified by pixel-based methods (not on region), generally, present a noisy appearance because of some isolated pixels that are misclassified [Guo and Moore 1991]. To fix such problem, we can use some post-classification method. One expects some degree of spatial correlation among neighborhood pixels, so we can remove isolated misclassification, resulting in a smoothed map.

Even becoming faster than the original approach, the EM algorithm is still more expensive than the other methods. It performs calculations of inverse matrix and determinant at each iteration, for the whole set of data. One approach, to reduce the computational demand, is to assume that all covariance matrices are diagonal or that they are equal to each other [Theodoridis and Koutroumbas 2003]. In this case, only one inversion matrix is needed at each iteration step, however, the system loses in generalization.

5. Conclusion

This work has presented an improvement to the EM Clustering Method, by using K-Means results as input, and some cluster validity techniques. Estimating mixture parameters is clearly a missing data problem, where the cluster labels of each observation are unknown [Figueiredo 2004]. The EM algorithm can be adopted, as we have proposed in this work, as a standard choice for this task.

One advantage of the EM algorithm is that its convergence is smooth and not vulnerable to instabilities. However, we have shown that wrong initial parameters might result in meaningless classification. Therefore the proposed approach, which estimates the first parameters using K-Means, increases the resultant accuracy.

In [Starck et al. 1998] they present the recovery of Gaussian-like clusters, applying the *à trous* wavelet. Future works include tests not only with K-Means approach but with this one as well. We also intend to perform another preprocessing stage, searching for outliers and removing them from the whole data set.

We have shown how to implement an EM algorithm and how to apply it to unsupervised image classification. As well, some classification results obtained with the proposed method and others were shown to compare their accuracy. We have implemented the algorithm using TerraLib library [Câmara et al. 2000], which is available for free download at `http://www.terralib.org/`. We also developed a software for unsupervised image classification, available at `http://www.dpi.inpe.br/~tkorting/`.

References

Beal, M. and Ghahramani, Z. (2003). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian Statistics*, 7:453–464.

Bilmes, J. (1998). A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. *manuscript, International Computer Science Institute*.

Bouman, C. (1995). Cluster: An unsupervised algorithm for modeling gaussian mixtures. *preprint available at http://www. ece. purdue. edu/ bouman/ softare/cluster/manual. pdf*.

Câmara, G., Souza, R., Freitas, U., and Garrido, J. (1996). SPRING: integrating remote sensing and GIS by object-oriented data modelling. *Computers & Graphics*, 20(3):395–403.

Câmara, G., Souza, R., Pedrosa, B., Vinhas, L., Monteiro, A., Paiva, J., Carvalho, M., and Gatass, M. (2000). TerraLib: Technology in Support of GIS Innovation. *II Workshop Brasileiro de Geoinformática, GeoInfo2000*.

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

Figueiredo (2004). Lecture Notes on the EM Algorithm. Technical report, Institute of Tele-communication.

Figueiredo, M. and Jain, A. (2002). Unsupervised learning of finite mixture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(3):381–396.

Guo, G. and Rodriguez, G. (1992). Estimating a Multivariate Proportional Hazards Model for Clustered Data Using the EM Algorithm, with an Application to Child Survival in Guatemala. *Journal of the American Statistical Association*, 87(420):969–976.

Guo, L. and Moore, J. (1991). Post-classification Processing For Thematic Mapping Based On Remotely Sensed Image Data. *Geoscience and Remote Sensing Symposium, 1991. IGARSS'91.'Remote Sensing: Global Monitoring for Earth Management'., International*, 4.

Lawrence, C. and Reilly, A. (1990). An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins: Structure, Function and Genetics*, 7:41–51.

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

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

Shepp, L. and Vardi, Y. (1982). Maximum likelihood reconstruction for emission tomography. *IEEE Trans. Med. Imag*, 1(2):113–122.

Starck, J., Murtagh, F., and Bijaoui, A. (1998). *Image Processing and Data Analysis: The Multiscale Approach*. Cambridge University Press.

Theodoridis, S. and Koutroumbas, K. (2003). *Pattern Recognition*. Academic Press.