Stochastic Modeling Techniques and Data Analysis international conference (SMTDA '10), Chania, Greece, June 8-11, 2010.

EM and ICE in Hidden and Triplet Markov Models

Wojciech Pieczynski

CITI Department, Institut Telecom, Telecom Sudparis
Evry, France
Email: wojciech.pieczynski@it-sudparis.eu

Abstract: This paper addresses the problem of parameter estimation in the case of hidden data. The aim is to discuss two general iterative parameter estimation methods "Expectation-Maximization" (EM) and "Iterative Conditional Estimation" (ICE) in the context of the classical Hidden Markov Models (HMMs) and in the context of the recent Triplet Markov Models (TMMs). A very general method of TMMs identification based on ICE and copulas is also specified.
Keywords: Hidden data, parameter estimation, Expectation-Maximization, Iterative Conditional Estimation, hidden Markov models, triplet Markov models, copulas.

1 Introduction

Let $Y = (Y_1,...,Y_n)$ be observed data and $X = (X_1,...,X_n)$ hidden ones. In the whole paper, each Y_i takes its values from the set of real numbers R, and each X_i takes its values from a finite set of classes $\Omega = \{\omega_1,...,\omega_K\}$. Let $p_\theta(x,y)$ be the probability distribution depending on a parameter $\theta \in R^m$, and let $l_\theta(x,y) = \log[p_\theta(x,y)]$ be the log-likelihood. Besides, let $\hat{\theta}(X,Y)$ be an estimator of $\theta \in R^m$ defined from complete data (X,Y). Both "Expectation-Maximization" (EM) and "Iterative Conditional Estimation" (ICE) define a sequence of parameters from the observation y. After having chosen an initial value θ^0, the EM sequence is defined by

$$\theta^{q+1}(y) = \arg\max_\theta E[l_\theta(X,Y)\big|Y = y, \theta^q], \quad (1.1)$$

while the ICE sequence is defined by

$$\theta^{q+1}(y) = E[\hat{\theta}(X,Y)\big|Y = y, \theta^q]. \quad (1.2)$$

The EM method (McLachlan and Krishnan (1997)) is well known and widely used, while ICE is less popular. However, ICE has been successfully used in different problems of unsupervised image processing; let us mention (Cao et al. (2005), Carincotte et al. (2006), Derrode and Pieczynski (2004), Destrempes and Mignotte

(2004), Provost et al. (2004), and Salzenstein et al. (2007)), among recent references. Concerning general considerations to compare EM and ICE, let us underline the following points:

(i) ICE is more general than EM because the estimator $\hat{\theta}(X,Y)$ can be of any form; in particular, it can be the "maximum likelihood" (ML) estimator or not. It is also often easier to perform because the maximization step does not exist in ICE ;

(ii) as stated in Delmas (1997), in the case of an exponential family of distributions EM and ICE can produce the same sequence (θ^q) ;

(iii) many comparisons between EM and ICE have been performed in classical contexts with Gaussian noise, like adaptive estimations (Peng and Pieczynski (1995)), hidden Markov chains (Benmiloud and Pieczynski (1995)), or hidden Markov trees (Monfrini (2002)). In all these situations the EM formulae are computable and it turns out that both EM and ICE methods are of quite a comparable efficiency ;

(iv) the use of EM is justified by the theoretical results concerning the optimal asymptotic behavior of the ML estimator, and by the fact that EM produces a sequence (θ^q) such that the sequence $p(y|\theta^q)$, being increasing, often converges to a local maximum. We have to notice that this does not imply the convergence of (θ^q) to the real parameter θ; however, if the initial value θ^0 is close enough to the real value θ, the convergence can be shown under some mild hypotheses. The idea behind ICE is different and is based on the following. Assuming that $\hat{\theta}(X,Y)$ has interesting quadratic error - or is even optimal, being, for example, an ML estimator in an exponential model - one wishes to approximate it by a function of the only observed variables y. The "best" - with regard to the same "quadratic error" criterion - approximation is the conditional expectation. As this expectation depends on the parameter, we arrive at (1.2). Concerning the convergence of ICE, let us mention a recent theoretical result obtained in the case of independent data (Pieczynski (2008)). As in the case of EM, convergence can be obtained under some reasonable hypotheses if the initial value θ^0 is close enough to the real value θ ;

(v) EM encounters more difficulties in hidden Markov field models, where the maximization step cannot be calculated and one is obliged to simplify the model, for example by introducing the "mean field" as indicated in (Celeux et al. (2006)). ICE can be used without model modification, even in more complex situations, as in the context of recent triplet Markov fields (Benboudjema and Pieczynski (2007)).

The aim of the paper is to discuss and compare the difficulties when applying these two methods in the context of the classical Hidden Markov Models (HMMs (Cappe et al. (2005), Ephraim (2002), (Koski (2001)) and the recent Triplet Markov Models (TMMs (Pieczynski and Desbouvries (2005), (Pieczynski (2007),

Pieczynski (2010)). A very general method of TMM identification based on ICE and copulas is also briefly described.

2 Pairwise and Hidden Markov Models

Let us consider the couple of stochastic sequences $(X,Y)=(X_1,Y_1,...,X_n,Y_n)$ and let us set $Z=(X,Y)=(Z_1,...,Z_n)$, with $Z_1=(X_1,Y_1)$, …, $Z_n=(X_n,Y_n)$. The couple $Z=(X,Y)$ is a "Pairwise Markov Model" (PMM) if its distribution is given by

$$p(z)=p(z_1)p(z_2|z_1)...p(z_n|z_{n-1}). \quad (2.1)$$

We will say that a PMM $Z=(X,Y)$ is "stationary" if the distributions $p(z_i,z_{i+1})$ do not depend on $i=1$, …, $n-1$. Thus the distribution of a stationary PMM (SPMM) is given by $p(z_1,z_2)$, which can be written:

$$p(z_1,z_2)=p(x_1,x_2)p(y_1,y_2|x_1,x_2). \quad (2.2)$$

There are then two kinds of SPMM $Z=(X,Y)$. Either X is a Markov chain or it is not. If it is, the SPMM $Z=(X,Y)$ will be called a "stationary hidden Markov model" (SHMM), which is consistent with the fact that the hidden model is a Markov one. One can then show that a "reversible" (which means that $p(z_i,z_{i+1})=p(z_{i+1},z_i)$) SPMM is an SHMM if, and only if, $p(y_1,y_2|x_1,x_2)$ in (2.2) verifies

$$p(y_1|x_1,x_2)=p(y_1|x_1). \quad (2.3)$$

In fact, a reversible SPMM $Z=(X,Y)$ is an SHMM if, and only if, the two equivalent conditions: (i) for each $2\le i\le n$, $p(y_i|x_i,x_{i-1})=p(y_i|x_i)$; (ii) for each $1\le i\le n$, $p(y_i|x)=p(y_i|x_i)$, are verified (Pieczynski (2007)).

Let us remark that the very classical SHMM, whose distribution is defined by

$$p(x,y)=p(x_1)p(x_2|x_1)...p(x_n|x_{n-1})\,p(y_1|x_1)...p(y_n|x_n), \quad (2.4)$$

is obtained when $p(y_1,y_2|x_1,x_2)$ in (2.2) verifies

$$p(y_1,y_2|x_1,x_2)=p(y_1|x_1)\,p(y_2|x_2), \quad (2.5)$$

which is stronger than (2.3).

In a similar way to the classical HMMs, the transitions $p(x_{i+1}|x_i, y)$ and the marginal distributions $p(x_i| y)$ can be computed in the following way. Let us consider the following "forward" $\alpha(x_i) = p(y_1,...,y_{i-1}, z_i)$ and "backward" $\beta(x_i) = p(y_{i+1},..., y_n|z_i)$ probabilities, which again give the classical probabilities when the PMM considered is an HMM. Then we have

$$\alpha_1(x_1) = p(z_1), \text{ and } \alpha_{i+1}(x_{i+1}) = \sum_{x_i \in \Omega} \alpha_i(x_i) p(z_{i+1}|z_i) \text{ for } 2 \le i \le n; \quad (2.6)$$

$$\beta_n(x_n) = 1, \text{ and } \beta_i(x_i) = \sum_{x_{i+1} \in \Omega} \beta_{i+1}(x_{i+1}) p(z_{i+1}|z_i) \text{ for } 1 \le i \le n-1; \quad (2.7)$$

$$p(x_{i+1}|x_i, y) = \frac{p(z_{i+1}|z_i)\beta_{i+1}(x_{i+1})}{\beta_i(x_i)}; \quad (2.8)$$

$$p(x_i|y) = \frac{\alpha_i(x_i)\beta_i(x_i)}{\sum_{x_i' \in \Omega} \alpha_i(x_i')\beta_i(x_i')}; \quad (2.9)$$

$$p(x_i, x_{i+1}|y) = p(x_i|y) p(x_{i+1}|x_i, y). \quad (2.10)$$

The formulae (2.6)-(2.10) are extensions of the well known HMM formulae, which are obtained by taking $p(z_1) = p(x_1)p(y_1|x_1)$ and $p(z_{i+1}|z_i) = p(x_{i+1}|x_i)p(y_{i+1}|x_{i+1})$.

Let us underline the fact that considering SPMMs which are not SHMMs (in which (2.3) does not hold) can be of real interest in the unsupervised segmentation of real or simulated data: see different results presented in (Derrode and Pieczynski (2004)).

3. EM and ICE in SPMM

Let us consider an SPMM whose distribution given by (2.2) is such that $p(y_1, y_2|x_1, x_2)$ are Gaussian. The parameters to be estimated are $p_{jk} = p(x_1 = \omega_j, x_2 = \omega_k)$ and the mean vectors M_{jk} and variance-covariance matrices Γ_{jk} of the Gaussian distributions $p(y_1, y_2|x_1 = \omega_j, x_2 = \omega_k)$. In both EM and ICE methods (p_{jk}) are re-estimated by

$$p_{jk}^{q+1} = \frac{1}{n-1}\sum_{i=1}^{n-1} p^q(x_i = \omega_j, x_{i+1} = \omega_k \big| y), \quad (3.1)$$

where $p^q(x_i = \omega_j, x_{i+1} = \omega_k \big| y)$ are computed with (2.6)-(2.10).
The parameters M_{jk} and Γ_{jk} are re-estimated in EM by

$$M_{jk}^{q+1} = \frac{\sum_{i=1}^{n-1} \begin{bmatrix} y_i \\ y_{i+1} \end{bmatrix} p^q(x_i = \omega_j, x_{i+1} = \omega_k \big| y)}{\sum_{i=1}^{n-1} p^q(x_i = \omega_j, x_{i+1} = \omega_k \big| y)}; \quad (3.2)$$

$$\Gamma_{jk}^{q+1} = \frac{\sum_{i=1}^{n-1} (\begin{bmatrix} y_i \\ y_{i+1} \end{bmatrix} - M_{jk}^{q+1})(\begin{bmatrix} y_i \\ y_{i+1} \end{bmatrix} - M_{jk}^{q+1})^T p^q(x_i = \omega_j, x_{i+1} = \omega_k \big| y)}{\sum_{i=1}^{n-1} p^q(x_i = \omega_j, x_{i+1} = \omega_k \big| y)}, \quad (3.3)$$

while in ICE they are re-estimated by

$$M_{jk}^{q+1} = \frac{\sum_{i=1}^{n-1} \begin{bmatrix} y_i \\ y_{i+1} \end{bmatrix} 1_{[x_i^q = \omega_j, x_{i+1}^q = \omega_k]}}{\sum_{i=1}^{n-1} 1_{[x_i^q = \omega_j, x_{i+1}^q = \omega_k]}}; \quad (3.4)$$

$$\Gamma_{jk}^{q+1} = \frac{\sum_{i=1}^{n-1} (\begin{bmatrix} y_i \\ y_{i+1} \end{bmatrix} - M_{jk}^{q+1})(\begin{bmatrix} y_i \\ y_{i+1} \end{bmatrix} - M_{jk}^{q+1})^T 1_{[x_i^q = \omega_j, x_{i+1}^q = \omega_k]}}{\sum_{i=1}^{n-1} 1_{[x_i^q = \omega_j, x_{i+1}^q = \omega_k]}}, \quad (3.5)$$

where $x^q = (x_1^q, ..., x_n^q)$ is sampled according to $p(x|y)$ using the current values of the parameters.
Dealing with the Gaussian case under consideration here with either EM or ICE would probably provide similar results, as they do in the classical HMMs. However, when one leaves the Gaussian case and deals with the "generalized" mixture estimation, ICE is much easier to apply. In the SHMM context one is faced with the "generalized" mixture estimation problem when the forms of the noise distributions $p(y_1 \big| x_1 = \omega_j)$ are not known and can vary with the class ω_j. However, for each ω_j one knows that $p(y_1 \big| x_1 = \omega_j)$ belongs to a given set of

forms. For example, one knows that $p(y_1|x_1=\omega_1)$ is either Gaussian or gamma, $p(y_1|x_1=\omega_2)$ can be Gaussian, exponential, or Rayleigh, … and so on. Such situations are of interest and they can occur, in particular, in radar images models (Delignon and Pieczynski (2002), Nadarajah and Kotz (2008)). Estimating such a mixture therefore contains two problems: (i) finding the right form for each class; and (ii) estimating the related parameters. ICE has been extended to a "generalized" ICE (GICE) to deal with such problems in SHMMs in (Giordana and Pieczynski (1997)) and different experiments have shown its efficiency. Afterwards, the extension of ICE to "generalized" reversible SPMMs has been suggested in (Pieczynski (2010)). Let us briefly recall its principle.

For K classes there are $(K-1)K/2$ distributions $p(y_1,y_2|x_1,x_2)$ on R^2. Besides, let $H(y_1,y_2)$ be a cumulative distribution function (cdf) over R^2, and $H_1(y_1)$, $H_2(y_2)$ the related marginal cdfs. Then, according to the Sklar theorem (Brunel and Pieczynski (2005), Nelsen (1998)), there is a unique cdf C on $[0,1]^2$ with uniform marginal distributions (called a "copula") such that

$$H(y_1,y_2)=C(H_1(y_1),H_2(y_2)) \qquad (3.6)$$

Thus each of the $(K-1)K/2$ distributions $p^{ij}(y_1,y_2)=p(y_1,y_2|x_1=\omega_i,x_2=\omega_j)$ on R^2 is defined by $(K-1)K/2$ marginal distributions $p^{ij}(y_1)$ and $(K-1)K/2$ copulas C^{ij}. Assuming that for each (i,j) the form of the marginal distribution $p^{ij}(y_1)$ belongs to a given set $\Phi^{ij}=\{F_1^{ij},...,F_{r(i,j)}^{ij}\}$ of admissible forms and the form of the copula C^{ij} belongs to a given set $\mathrm{X}^{ij}=\{\mathrm{C}_1^{ij},...,\mathrm{C}_{m(i,j)}^{ij}\}$ of admissible forms, one is faced with the following problem : for each (i,j) select from Φ^{ij} and X^{ij} the correct forms and estimate the related parameters. At each iteration of ICE these two problems are then dealt with using $x^q=(x_1^q,...,x_n^q)$ sampled according to $p(x|y,\theta^q)$.

4. Generalized ICE in Stationary Triplet Markov Models

Let us consider the couple (X,Y) as above. Let $U=(U_1,...,U_n,)$ be a third random chain, each U_i taking its values from $\Lambda=\{\lambda_1,...,\lambda_M\}$. The triplet $T=(X,U,Y)$ is called a "Triplet Markov Model" (TMM) if its distribution is a Markovian one. Setting $V=(X,U)$ one sees that a TMM can also be seen as a PMM (V,Y); in fact, $V=(V_1,...,V_n)$ with each V_i taking its values from a finite set $\Omega\times\Lambda$. Thus both X and U can be estimated by some Bayesian method, and the parameters can be estimated with EM or ICE as discussed above.

The choice of the interpretation of the third chain U and the choice of the Markovian distribution for $T=(X,U,Y)$ lead to a very rich family of possible distributions for (X,Y). One possible choice is a Markov distribution for $V=(X,U)$ such that X is a semi-Markov chain (Pieczynski and Desbouvries (2005)); $T=(X,U,Y)$ is then a classical hidden semi-Markov chain. Other choices lead to a non-stationary distribution for (X,Y), where the switches among the different stationarities are modeled by U (Lanchantin and Pieczynski (2004)). Let us also mention the use of TMM to perform the Dempster-Shafer fusion in a Markovian context (Pieczynski (2007)). It is also possible to consider multivariate U, to model different properties simultaneously. For example, one can take $U=(U^1,U^2)$, where U^1 models the semi-Markovianity of X and U^2 models its non-stationarity (Lapuyade-Lahorgue and Pieczynski (2006)). In each of these situations, one can then apply the "generalized" ICE described above to the related PMM $T=(X,U,Y)=(V,Y)$.

References

1. Benboudjema D., and Pieczynski, W., Unsupervised statistical segmentation of non stationary images using triplet Markov fields, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **29**, 8, 1367-1378 (2007).
2. Benmiloud, B. and Pieczynski, W., Estimation des paramètres dans les chaînes de Markov cachées et segmentation d'images*, Traitement du Signal*, **12**, 5, 433-454 (1995).
3. Brunel, N., and Pieczynski, W., Unsupervised signal restoration using hidden Markov chains with copulas, *Signal Processing*, **85**, 12, 2304-2315 (2005).
4. Cao, Y. F., Sun, H., and Xu, X., An unsupervised segmentation method based on MPM for SAR images, *IEEE Geoscience and Remote Sensing Letters*, 2, 1, 55-58 (2005).
5. Cappé, O., Moulines, E., and Ryden, T., *Inference in hidden Markov models*, Springer, Series in Statistics, Springer (2005).
6. Carincotte, C., Derrode, S., and Bourennane, S., Unsupervised change detection on SAR images using fuzzy hidden Markov chains, *IEEE Trans. on Geoscience and Remote Sensing*, **44**, 2, 432-441 (2006).
7. Celeux, G., Forbes, F., and Peyrard, N., EM procedures using mean field-like approximations for Markov model-based segmentation, *Pattern Recognition*, **36**, 131-144 (2006).
8. Delignon, Y., and Pieczynski, W., Modeling non Rayleigh speckle distribution in SAR images, *IEEE Trans. on Geoscience and Remote Sensing*, **40**, 6, 1430-1435 (2002).

9. Delmas, J.-P., An equivalence of the EM and ICE algorithm for exponential family, *IEEE Trans. on Signal Processing*, **45**, 10, 2613-2615 (1997).
10. Derrode, S., and Pieczynski, W., Signal and image segmentation using pairwise Markov chains, *IEEE Trans. on Signal Processing*, **52**, 9, 2477-2489 (2004).
11. Destrempes, F., and Mignotte, M., A statistical model for contours in images, *IEEE Trans. on Pattern Analysis and Machine Intelligence,* 26, 5, 626-638 (2004).
12. Ephraim, Y., Hidden Markov processes, *IEEE Trans. on Information Theory*, **48**, 6, 1518-1569 (2002).
13. Giordana, N., and Pieczynski, W., Estimation of generalized multisensor hidden Markov chains and unsupervised image segmentation, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **19**, 5, 465-475 (1997).
14. Koski, T., *Hidden Markov models for bioinformatics*, Kluwer Academic Publishers, Netherlands (2001).
15. Lanchantin, P., and Pieczynski, W., Unsupervised non stationary image segmentation using triplet Markov chains, ACVIS 04, Aug. 31-Sept. 3, Brussels, Belgium, 2004.
16. Lapuyade-Lahorgue, J., and Pieczynski, W., Unsupervised segmentation of hidden semi-Markov non stationary chains, MaxEnt 2006, Paris, France, July 8-13, 2006.
17. McLachlan, G. J. and Krishnan, T., *EM Algorithm and Extensions*, Wiley (1997).
18. Monfrini, E., Identifiabilité et méthode des moments dans les mélanges généralisés de distributions du système de Pearson, Thèse de l'Université Paris VI, soutenue le 4 janvier 2002.
19. Nadarajah, S., and Kotz, S., Intensity models for non-Rayleigh speckle distributions, *International Journal of Remote Sensing*, **29**, 2, 529-541 (2008).
20. Nelsen, R. B., *An introduction to Copulas*. Number 139 in Lecture Notes in Statistics. Springer-Verlag (1998).
21. Peng, A. and Pieczynski, W., Adaptive mixture estimation and unsupervised local Bayesian image segmentation, *Graphical Models and Image Processing*, **57**, 5, 389-399 (1995).
22. Pieczynski, W., and Desbouvries, F., On triplet Markov chains, (ASMDA 2005), Brest, France, May 2005.
23. Pieczynski, W., Multisensor triplet Markov chains and theory of evidence, *International Journal of Approximate Reasoning*, **45**, 1, 1-16 (2007).
24. Pieczynski, W., Sur la convergence de l'estimation conditionnelle itérative, *Comptes Rendus*, **346**, 7-8, 457-460 (2008).

25. Pieczynski, W., *Triplet Markov chains and image segmentation*, chapter 4 in Inverse Problems in Vision and 3D Tomography, A. Mohammed-Djafari ed., Wiley (2010).
26. Provost, J.-N., Collet, C., Rostaing, P., Pérez, P.,and Bouthemy, P., Hierarchical Markovian segmentation of multispectral images for reconstruction of water depth maps, *Computer Vision and Image Understanding*, **93**, 2, 155-174 (2004).
27. Salzenstein F., Collet, C., Le Cam, S. and Hatt, M., Non stationary fuzzy Markov chains, *Pattern Recognition Letters*, **28**, 16, 2201-2208 (2007).