ESTIMATING THE OFFSET PARAMETERS OF A MIXTURE IN THE FOURIER DOMAIN

S. Vaton and T. Chonavel

ENST Bretagne, BP 832,29285 Brest Cedex, France.
e-mail: Sandrine.Vaton, Thierry.Chonavel@enst-bretagne.fr

ABSTRACT

In this contribution we present an algorithm for estimating some parameters of offset in the case of incomplete data. This estimation cannot be performed directly with an EM or SEM method because the density of local extrema in the likelihood map grows exponentially with the number of observations and because the SEM method provides a monotonic sequence of estimates so that bad initialization cannot be recovered. We perform the estimation in the Fourier domain. The offsets in time domain are transformed into pulsations in the Fourier domain. We minimize a quadratic distance between the parametric and empirical sampled Fourier transform with an EM method. Contrary to the problems encountered in the time domain the asymptotic loglikelihood of the sampled empirical Fourier transform is continuous w.r.t. the parameters of offset. We discuss the influence of the frequencies at which the Fourier transform is sampled and we present a numerical study of convergence of the proposed algorithms.

1. INTRODUCTION

Some previous studies have proved that Fourier transform based techniques enable estimation of the proportions of a mixture (see [1] and the references therein) . In this paper we discuss the possibility of using weighted distances between Fourier transforms and their empirical counterparts for estimating some parameters of offset in the case of incomplete data, a problem that cannot be solved by classical approaches. Semi-Markov processes (HMM, MRP...) with shifted exponential conditional laws $f_{\theta,\lambda}(x) = \lambda \exp(-\lambda(x-\theta))\mathbb{1}_{[\theta,+\infty[}(x)$ are generalizations of standard models of teletraffic such as the Poisson process or the MMPP. These models with offset fit the marginal distribution of the inter-event times better than standard models [2]. Modeling and estimating teletraffic is an important issue for dimensioning telecommunication networks.

The rest of the paper is organised as follows. In Section 2 we explain why MLE in the time domain is not viable in the case of incomplete data. In Section 3 we recall some techniques for estimating a single offset and we note that these techniques do not generalize to the case of incomplete data. In order to overcome these problems we present two Fourier transform based algorithms. The influence of the points at which the Fourier transform is sampled and the performances of the algorithm is discussed.

This work was supported by France Telecom research center under contract number PE95-7633.

2. FAILURE OF MLE IN THE TIME DOMAIN

2.1. Failure of the EM algorithm

Denote by $L(x;\theta)$ the loglikelihood of the observations $x = x_{1:T}$ when the parameters of shift are equal to $\theta = (\theta_i)_{1\leq i\leq K}$. When $(X_t)_t$ is an i.i.d. sequence distributed as a mixture $L(x;\theta) = \sum_t \log(\sum_i \pi_i f_{0,\lambda_i}(x_t - \theta_i))$ and $L(x;\theta)$ is discontinuous at any point $\theta = (\theta_i)_{1\leq i\leq K}$ such that one of the x_t is equal to one of the θ_i. Consequently the likelihood map $L(x;\theta)$ has infinitely many local extrema that attract the EM algorithm [3]. When $(X_t)_t$ is a HMM same discontinuities are encountered.

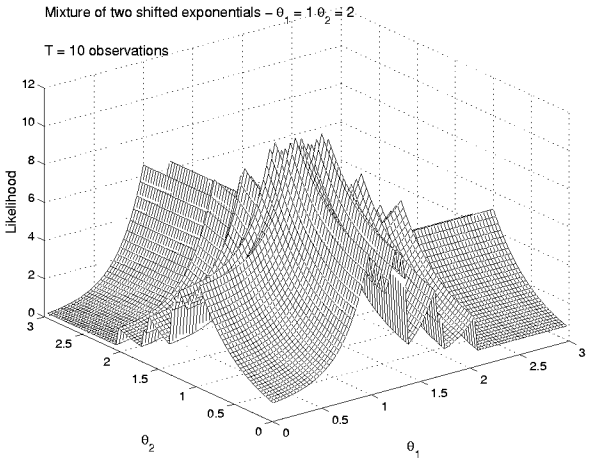

Figure 1: Discontinuities of the likelihood in time domain. $\pi_1 = \pi_2 = 0.5, \theta_1 = 1, \theta_2 = 2, \lambda_1 = 2, \lambda_2 = 4$.

2.2. Failure of the SEM algorithm

In the presence of local extrema one often uses the SEM algorithm [4]. Each iteration of the SEM algorithm can be decomposed into two steps:

1. Stochastic Expectation step

Compute by means of the forward backward algorithm the distribution of the unobserved $S_{1:T}$ conditionally to the observed $X_{1:T} = x_{1:T}$ for the value θ^k of the set of parameters. Simulate $S_{1:T}$ under this distribution. The result is denoted by $s^k = s^k_{1:T}$.

2. Maximization step

Maximize the complete loglikelihood $L(x_{1:T}, s^k_{1:T}; \theta)$:

$$L(x_{1:T}, s^k_{1:T}; \theta) \quad = \quad L(s^k_{1:T}) + \sum_i \sum_t \mathbb{1}_{s^k_t = i} \log f_{\theta_i, \lambda_i}(x_t)$$

If $x_t < \theta^k_i$ then $\mathbb{P}(s_t = i \mid x_t; \theta^k) = 0$ and a.s. $s^k_t \neq i$ so that $\theta^k_i \leq \min(x_t; s^k_t = i)$. If $x_t < \theta_i$ then $\sum_t \mathbb{1}_{s^k_t = i} \log f_{\theta_i, \lambda_i}(x_t) = -\infty$ so that $\theta^{k+1}_i \leq \min_{1 \leq t \leq T; s^k_t = i}(x_t)$ and this function strictly increases over $]-\infty, \min(x_t; s^k_t = i)[$ yielding $\theta^{k+1}_i = \min(x_t; s^k_t = i) \geq \theta^k_i$.

The sequence $(\theta^k_i)_{k \in \mathbb{N}}$ increases so that a bad initialization cannot be corrected.

2.3. Lack of Cramer-Rao Lower Bound (CRLB)

Maximum likelihood techniques are justified by the fact that the MLE is asymptotically unbiased and that its variance converges to $CRLB = \mathbb{E}_\theta(\frac{d^2\theta}{d\theta^2} \log f(X; \theta))$. For the problem under study the loglikelihood is not derivable w.r.t. the parameters of shift and the conditions under which the CRLB is derived are not fulfilled. However, in the Fourier domain the normalized sampled Fourier transform converges to a gaussian distribution which belongs to the exponential family $f(x; \theta) = c(\theta)h(x)\exp(< \alpha(\theta), T(x) >)$ and the MLE is known in this case to be efficient.

3. ESTIMATION OF A SINGLE SHIFT PARAMETER

In this Section we recall some techniques for estimating a parameter of offset when the observations are i.i.d. with p.d.f. $f_\theta(x) = f_0(x - \theta)$ and $f_0(x) = 0$ if $x < 0$.

3.1. Maximum Likelihood Estimation

The loglikelihood of $x_{1:T}$ is $L(x_{1:T}; \theta) = \sum_{t=1}^T \log f_0(x_t - \theta)$ so that $L(x_{1:T}; \theta) = -\infty$ if $x_t < \theta$ for some t. In the case when $f_0(\bullet)$ strictly decreases over $\mathbb{R}^+$ and in particular in the case of the shifted exponential distribution $\hat{\theta}_{ML} = X_{\min} \triangleq \min_{1 \leq t \leq T} X_t$.

3.2. Bayesian estimation techniques

The choice of a prior distribution is usually the controversial point in Bayesian estimation. Denote by $g(\theta)$ the prior distribution of θ. Then $f(x_{1:T}) = (f_0 * g)(x_{1:T})$ where $*$ denotes convolution. $f(x_{1:T})$ is maximum for $g(\theta) = \frac{f_0(x_{1:T} - \theta)}{\int f_0(x_{1:T} - \theta)\, d\theta}$ which acts as a matched filter on $f_0(x_{1:T})$. For this choice of prior distribution the Maximum A Posteriori is $\hat{\theta}_{MAP} = Arg\max_\theta f(\theta \mid x_{1:T}) = X_{\min}$ when f_0 strictly decreases over $\mathbb{R}^+$ and the mean squared error estimator is $\hat{\theta}_{EQM} = \mathbb{E}(\theta \mid X_{1:T}) = X_{\min} - (2\lambda T)^{-1}$ in the case of the shifted exponential distribution.

3.3. Pitman estimator

Consider the set of all equivariant estimators of θ, $S(x_{1:T} + \mu) = \mu + S(x_{1:T})$. It is natural to look for an estimator satisfying this property when one is concerned with the estimation of a parameter of shift. The equivariant estimator with minimum quadratic risk function is the Pitman estimator [5]. In the case of the shifted exponential distribution $\hat{\theta}_{PIT} = X_{\min} - (\lambda T)^{-1}$.

3.4. Barankin estimator

The Barankin bound [6] is the greatest lower bound among minimum variance bounds for unbiased estimators. As a byproduct the calculation of this bound supplies a locally, and possibly globally, minimum variance unbiased estimate. In particular, for a shifted exponential distribution, we can obtain the minimum variance unbiased estimator $\hat{\theta}_B(X) = X_{\min} - (\lambda T)^{-1}$ [6].

The above methods do not permit estimation of the offset parameters of a mixture. In order to overcome the limitations of these methods we propose solving this problem in the Fourier transform domain.

4. ESTIMATION IN THE FOURIER DOMAIN

4.1. The offsets in the time domain are transposed into pulsations in the Fourier domain

Denote by $\Phi(\omega) = \int e^{i\omega x} f(x) dx$ the characteristic function of f(x). Note that a shift in the time domain is equivalent to a modulation by a complex exponential in the Fourier domain the offset parameters $(\theta_i)_{1 \leq i \leq K}$ being the frequencies of the modulating complex exponentials : $\Phi(\omega) = \sum_i \pi_i e^{i\omega\theta_i} \Phi_{0,\lambda_i}(\omega)$. This note will permit the construction of an estimator of $\theta = (\theta_i)_{1 \leq i \leq K}$ from a sampled empirical estimate $(\hat{\Phi}(\omega_1), \cdots, \hat{\Phi}(\omega_L))$ where $\hat{\Phi}(\omega) = \frac{1}{T} \sum_{1 \leq t \leq T} e^{i\omega X_t}$.

4.2. Central Limit Theorem [7]

Denote by $\omega_1, \omega_2, \cdots, \omega_L$ the pulsations at which the Fourier transform is sampled. Denote by $m(\theta) = (\Phi(\omega_1), \Phi(\omega_2), \cdots, \Phi(\omega_L))^T$ the sampled Fourier transform of $f(\bullet)$ and denote the empirical sampled Fourier transform by $Z_T = T^{-1} \sum Y_t$ where Y_t is the vector with entry $1 \leq l \leq L$ is $e^{i\omega_l X_t}$.

Theorem 1 *Suppose that $(X_t)_t$ is i.i.d. or a finite state irreducible Markov chain; then it holds that $\sqrt{T}(Z_T - m(\theta)) \sim \mathcal{AN}(0, \Gamma(\theta))$ where $\Gamma(\theta) = \mathbb{E}_\theta((Y_t - m)(Y_t - m)^H)$ in the i.i.d. case and $\Gamma_\theta = \sum_{\tau \in \mathbb{Z}} \mathbb{E}_\theta((Y_{t+\tau} - m)(Y_{t+\tau} - m)^H)$ in the HMM case.*

4.3. Maximum Likelihood Estimation

Our procedure consists in minimizing the asymptotic loglikelihood of the normalized sampled Fourier transform :

$$J(\theta) = \log|\Gamma(\theta)| + \frac{1}{2} T (Z_T - m(\theta))^H \Gamma^{-1}(\theta)(Z_T - m(\theta))$$

Standard optimization techniques (gradient method, Newton method, conjugate gradient method ...) can be used to maximize $J(\theta)$. Let us consider the asymptotic covariance matrix $\Gamma(\theta)$. When $(X_t)_t$ is i.i.d. then $\Gamma_{ij}(\theta) = \Phi_\theta(\omega_i - \omega_j)$. When $(X_t)_t$ is a HMM it is simpler to replace $\Gamma(\theta)$ with a consistent estimate $\hat{\Gamma}_T$ and to optimize $K(\theta) = (Z_T - m(\theta))^H \hat{\Gamma}_T^{-1} (Z_T - m(\theta))$.

4.3.1. Problems related to sampling the Fourier transform

Contraction of the loglikelihood map

Suppose that $\omega_1 < \omega_2 < \cdots < \omega_L$. If $\omega_1 \leq 2\pi / X_{\max}$ where $X_{\max} = \max_{1 \leq t \leq T} X_t$ there exists a one to one mapping

between $(X_t)_{1\leq t\leq T}$ and $(Y_t)_{1\leq t\leq T}$. In what follows we discuss the influence of the choice of the points $(\omega_1,\cdots,\omega_L)$ at which the Fourier transform is sampled.

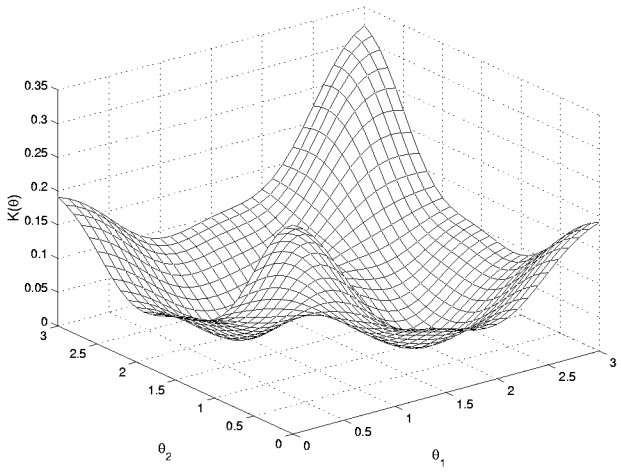

Figure 2: $\omega = 2\pi/X_{\max}$, $K = 2$,$\pi_1 = \pi_2 = 0.5$,$\lambda_1 = \lambda_2 = 1$,$\theta_1 = 1$,$\theta_2 = 2$,$T = 5000$

Note that $(\omega_1,\cdots,\omega_L) \rightarrow \alpha(\omega_1,\cdots,\omega_L)$results in a contraction of $\theta \rightarrow K(\theta)$ when $\alpha > 1$ (Fig. 2) and in a dilatation of $\theta \rightarrow K(\theta)$ when $\alpha < 1$ (Fig. 3). For the sake of simplicity we suppose in what follows that the sampling is regular $\omega_k = k\omega_1$. Let us study the effects of some variations on ω_1. First of all there are clearly more local extrema in $\theta \rightarrow K(\theta)$ as ω_1 increases because of the contraction effect mentioned above. The optimization algorithms consequently get trapped in the local extrema if ω_1 is too large. For example let us choose $T = 1000, L = 1, K = 2, \pi_1 = \pi_2 = 0.5, \lambda_1 = \lambda_2 = 2, \theta_1 = 1, \theta_2 = 2$. The initialization of θ is random with uniform distribution on $[0,3]\times[0,3]$. If the condition $\omega_1 < \frac{2\pi}{X_{\max}}$ is not fulfilled, the estimate $\hat{\theta}$ gets trapped in the local extrema of $K(\theta)$: for $\omega_1 = \frac{2\pi}{X_{\max}}\ \sigma_\theta^2 \simeq 10^{-3}$ and for $\omega_1 = 10\frac{2\pi}{X_{\max}}\ \sigma_\theta^2 \simeq 1$ where σ_θ^2 denotes the sum of the variances of the estimate of the offsets. It results from this discussion that ω_1 should be chosen as small as possible in order to avoid local extrema. But another effect of the choice of frequencies at which the Fourier transform is sampled is that the covariance matrix Γ has too high a condition number for optimization to be possible when ω_1 is very small. This point is discussed in what follows.

Condition number of the covariance matrix

Recall first of all that if the $(X_t)_t$ are i.i.d. then the matrix Γ has entry (i,j) $\Gamma(i,j) = \Phi(\omega_i - \omega_j)$. Since $\omega_k = k\omega_1$ the covariance matrix Γ tends to the rank one matrix ee^T where $e = [1,1,\cdots,1]^T$ when ω_1 tends to zero. The condition number of Γ is thus very high if ω_1 is too small. Table 1 provides the condition number of Γ for different values of $\omega_1 = \alpha 2\pi/X_{\max}$ and a growing number L of sampling points.

As one can see in Table 1 the condition number of Γ grows very quickly with the number L of frequencies at which the Fourier transform is sampled. A contradictory effect of L is that the variance of the estimate $\hat{\theta}$ is lowest if the number L of sampling points is high; the variance of the estimate decreases as $O(1/L)$ until $L = T$ and then the variance stabilizes. Table 2 shows this result.

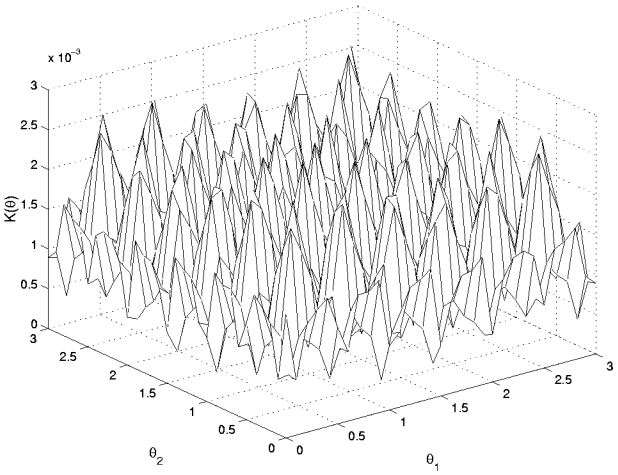

Figure 3: $\omega = 20 \times 2\pi/X_{\max}$, $K = 2$,$\pi_1 = \pi_2 = 0.5$,$\lambda_1 = \lambda_2 = 1$,$\theta_1 = 1$,$\theta_2 = 2$,$T = 5000$

L α	0.001	0.01	0.1	1	10	100	1000
2	10^5	10^4	100	1	1	1	1
5	10^{13}	10^9	10^5	10	1	1	1
10	10^{18}	10^{17}	10^9	100	10	1	1

Table 1: Condition number of Γ with $T = 1000$; $K = 2$; $\pi(1) = \pi(2) = 0.5$; $\lambda_1 = \lambda_2 = 2$; $\theta_1 = 1$; $\theta_2 = 2$.

In order to deal with the contradictory effects of the choice of the frequencies at which the Fourier transform is sampled we suggest the following approach. We successively estimate θ for simultaneously increasing values of α and L so that the condition number of $\hat{\Gamma}$ remains reasonable and we use the previous estimate of θ as an initial guess while changing α and L. The use of this previous estimate avoids converging to a local minimum of $K(\theta)$ when α increases. A possible alternative approach consists in replacing the consistent estimate $\hat{\Gamma}$ with the identity matrix in the criterion $K(\theta)$.

4.4. Least mean square estimation

In this Section we propose to undertake a least mean square estimation i.e. to minimize $M(\theta) = \|Z_T - m(\theta)\|^2$.

4.4.1. Expectation Maximization method

The above problem is equivalent to maximizing the likelihood of $\hat{Z}_T$ w.r.t. θ for the following distribution :

$$\left\{\begin{array}{lcl} Z & = & W_1 + W_2 + \cdots + W_K \\ (W_k)_k & & independent \\ W_k & \sim & \mathcal{N}(m_k(\theta_k), \Gamma_k) \end{array}\right.$$

where $m_k(\theta_k) = \pi_k[e^{j\omega_1\theta_k}\Phi_{0,\lambda_k}(\omega_1),\cdots,e^{j\omega_L\theta_k}\Phi_{0,\lambda_k}(\omega_L)]^T$. An EM algorithm makes it possible to split the optimization of the likelihood into simpler one dimensional optimizations.

1. Expectation step

Compute $Q(\theta,\theta_k) = \mathbb{E}(L(W,Z;\theta) \mid Z=z;\theta_k) = C - \frac{1}{2}\sum_i \Gamma_i^{-1/2}\|m_i(\theta_i^k) - m_i(\theta_i) + \Gamma_i\Gamma^{-1}(z - \sum_{j=1}^K m_j(\theta_j^k)\|^2\Gamma_i^{-H/2}$.

2. Maximization step

If one supposes that the $W_i s$ all have the same covariance matrix $\Gamma_i = K^{-1}\Gamma$ the criterion to minimize w.r.t. θ_i reduces to $\|m_i(\theta_i^k) - m_i(\theta_i)) + K^{-1}(z - \sum_{j=1}^K m_j(\theta_j^k)\|^2$ and $\theta_i^{k+1} = Arg\min_{\theta_i} \|m_i(\theta_i^k) - m_i(\theta_i)) + K^{-1}(z - \sum_{j=1}^K m_j(\theta_j^k)\|^2$.

4.4.2. Simplified maximization step

The criterion $\|m_i(\theta_i^k) - m_i(\theta_i)) + K^{-1}(z - \sum_{j=1}^K m_j(\theta_j^k)\|^2$ can be reduced to a quadratic form in θ_i though yielding analytical reestimation for θ_i.

Denote by $\alpha_i^l = m_i(\theta_i) - K^{-1}(z - m(\theta^k))$ and by $\psi_l = Arg(\alpha_i^l)$ then $\alpha_i^l \simeq \pi_i \left|\Phi_{0,\lambda_i}(\omega_l)\right| e^{j\psi_l}$ and a second order development yields $\|m_i(\theta_i) - \alpha_i\|^2 \simeq \sum_l \pi_i^2 \left|\Phi_{0,\lambda_i}(\omega_l)\right|^2 (\omega_l\theta_i - \psi_l)^2$ and

$$\theta_i^{k+1} = \frac{\sum_l \left|\Phi_{0,\lambda_i}(\omega_l)\right|^2 \omega_l\psi_l}{\sum_l \left|\Phi_{0,\lambda_i}(\omega_l)\right|^2 \omega_l^2}$$

4.5. Example

The algorithm is used to estimate the offsets of a four component mixture of shifted exponential distributions. The parameters of offset are $\theta_1 = 1, \theta_2 = 2, \theta_3 = 3$ and $\theta_4 = 5$ and the intensities of the exponentials are all equal to $\lambda = 2$. The Fourier transform of the mixture is considered at only one ($L = 1$) pulsation $\omega = \frac{2\pi}{X_{\max}}$. The EM method 4.4.1 as well as the optimization of the simplified criterion 4.4.2 are considered. The Table below lists the mean and variance of the estimates of the offsets for $T = 1000$ independent realisations of the four component mixture. The performance of the algorithms in the case of small samples has also been investigated. For $T = 100$ independent realisations of the mixture the obtained bias is of the order of 0.1 times the true parameters of offset and the variance of the estimates is between 0.02 and 0.10 for both the EM and the simplified algorithm. The estimates at successive iterations of the EM algorithm are plotted in Figure 4 for small samples $T = 100$.

L	1	10	100	1000	10000	100000
σ_θ	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-6}	10^{-6}

Table 2: Variance of $\hat{\theta}$. $T = 1000; K = 2; \pi(1) = \pi(2) = 0.5; \lambda_1 = \lambda_2 = 2; \theta_1 = 1; \theta_2 = 2; \omega = 2\pi/X_{\max}$

	Algorithm 4.4.1		Algorithm 4.4.2	
	$\mathbb{E}(\hat{\theta}_k)$	$var(\hat{\theta}_k)$	$\mathbb{E}(\hat{\theta}_k)$	$var(\hat{\theta}_k)$
θ_1	0.99	1.5×10^{-3}	1.06	1.9×10^{-3}
θ_2	2.21	1.7×10^{-3}	2.20	1.9×10^{-3}
θ_3	2.72	2.1×10^{-3}	2.70	2.5×10^{-3}
θ_4	5.06	6×10^{-3}	5.07	7.6×10^{-3}

Table 3: Mean and variance of the estimators.

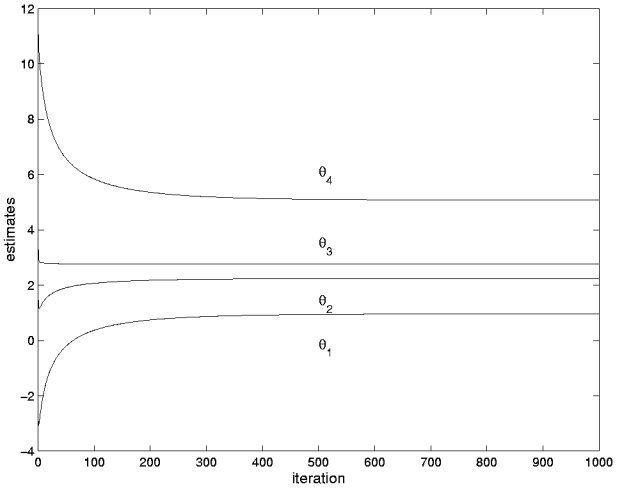

Figure 4: $T = 100$. $L = 1$. $\omega = 2\pi/X_{\max}$. $\theta_1 = 1, \theta_2 = 2, \theta_3 = 3, \theta_4 = 5$. $\hat{\theta_1} = 1.10, \hat{\theta_2} = 2.29, \hat{\theta_3} = 2.78, \hat{\theta_4} = 5.10$.

5. CONCLUSION

Time domain approaches do not enable the estimation of a mixture in some cases when offset parameters are unknown. We have proposed some Fourier transform based algorithms for estimating those parameters. When all but the offset parameters are known a good choice of sampling for the characteristic function results in no local extrema. In this context analytical reestimation formulae have been derived thus permiting estimation with extremely low computational burden and fast convergence.

6. REFERENCES

[1] J.L. Bryant and A.S. Paulson, "Some comments on characteristic-function based estimations," *Sankhya A*, vol. 41, pp. 109–116, 1979.

[2] S.Vaton, H.Korezlioglu, and T.Chonavel, "Modelling lan traffic data as a locally stationary semi-markov process," in *Performance Modelling and Evaluation of ATM Networks Vol.3*, D.D.Kouvatsos, Ed. 1997, Chapmann and Hall.

[3] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood estimation from incomplete data via the em algorithm," *JRSS-B*, vol. 39, pp. 1–38, 1977.

[4] G. Celeux and J. Diebolt, "The sem algorithm: a probabilistic teacher algorithm derived from the em algorithm for the mixture problem," *Comp. Stat. Quat.*, vol. 2, pp. 73–82, 1985.

[5] E.Pitman, "The estimation of location and scale parameters of a continuous population of any given form," *Biometrika*, , no. 30, pp. 391–421, 1939.

[6] T.L. Marzetta, "Computing the barankin bound by solving an unconstrained quadratic optimization problem," in *ICASSP'97*, Munich, Germany, april 1997, pp. 3829–3832.

[7] P. Doukhan, *Mixing: properties and examples*, Lecture Notes in Statistics. Springer-Verlag, 1994.