

Parameter Estimation: Known Vector Signals in Unknown Gaussian Noise¹

G. R. Dattatreya² and Xiaori (Frank) Fang

Abstract

This paper develops recursive, convergent estimators for the parameters of finite Gaussian mixtures with a common covariance matrix. The mean vectors (signals) of the component densities are assumed to be known. The motivation for the study stems from digital communication. The basic approach is first illustrated for the case of an independent identically distributed sequence of samples from a univariate mixture of M classes (symbols). This is accomplished through the development of a convergent stochastic approximation form of estimator for the common variance value. The asymptotic variance of the estimated variance is derived. A batch processing alternative that possesses a sufficient statistic is developed for the case of a fixed size sample set. Three generalizations are studied. The first extends from the case of the univariate data to multivariate data. The second generalization allows for the statistical dependence of successive vector signals. Finally, the case of dependent successive vector signals along with dependent successive additive noise vectors is treated. In each case, convergent estimators for all unknown parameters are developed. Many cases are illustrated with simulation experiments. Results presented are applicable to communication engineering, pattern recognition, and some special image processing problems.

Keywords: Finite Gaussian Mixtures, Parameter Estimation, Stochastic Approximation, Sufficient Statistics, and Blind Channel Estimation.

¹A condensed version of portions from this manuscript forms part of "Parameter Estimation and Application of a class of Gaussian Image Models," Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation, Dallas, Texas, April 1994, pp. 18 - 23.

²Corresponding author. Authors' Address:

Department of Computer Science
University of Texas at Dallas
MS EC 3.1, P. O. Box 830688
Richardson, TX 75083-0688
fax: 972-883-2349
email: datta@utdallas.edu

1 Introduction

Finite mixture density models have found applications in many areas such as biological classification, medicine, marketing, speech recognition, and digital communication. Gaussian mixtures predominate the literature and applications. The general problem of correctly estimating all the parameters is very complicated. Many published approaches concentrate on obtaining locally optimum solutions, accelerating the rate of convergence, and special cases when some of the parameters of the mixture density are known. This paper studies estimation of parameters of a mixture random variable (or vector) with M Gaussian components whose means values (or vectors) are known and whose variances (or covariance matrices) are equal and unknown. The prior or mixing probabilities are also not known.

A Motivation

In digital transmission with Automatic Gain Control (AGC) at the receiver, the received signal is amplified based on the level of the received carrier and this brings the equivalent noise-free signal levels for the different symbols to predetermined values. This technique is used in many communication digital systems [1], for example, in Cellular Telephony with CDMA. In Coherent Demodulation, the pilot carrier provides the reference phase so that the noise-free phase values for the received symbols are predetermined. The received phase in such systems are evaluated from the amplitudes in quadrature. The amplitude noises are generally modeled as Gaussian with symbol-independent variance [1]. Combinations of the above amplitude and phase modulation schemes are also commonly employed in wireless digital communication systems including satellite communication systems. In simple cases, the symbols are considered to be equiprobable and the value of the noise variance does not affect the detection scheme, although it affects the probability of correct detection. Detection accuracy can be improved by using the correct prior symbol probabilities and the noise variance in a maximum *a posteriori* probability classification scheme. Due to AGC and/or fading-compensation, the noise variance is not known and can also be slowly varying. Thus, a receiver that can recursively estimate the required prior probabilities and the noise variance on-line, as it operates with unclassified data, is very useful in the above systems.

Multipath fading is common in wide band CDMA systems. The different faded components are called fingers. Coherent systems with “selection combining (SC)” combine a preselected number of the strongest fingers for detection [2]. Noise variances in all the fingers are required in order to rank them for SC. These noise variances should be estimated from the mixture data. This problem further motivates our study here.

The simplest case of the problem is the estimation of the prior probabilities P_1, \dots, P_M and the common variance σ^2 of the M components of a univariate Gaussian mixture density with known mean values μ_1, \dots, μ_M , from an i.i.d. sequence of mixture samples $\{x_n\}$. A digital communication application of this case is the pulse amplitude modulated transmission of a statistically independent and identically distributed (i.i.d.) sequence of symbols from a multi-symbol set, added by an i.i.d. noise sequence, with AGC at the receiver. The case of coherent demodulation with the phase evaluated from amplitudes in quadrature requires the processing of the vector of received amplitudes. In narrow band signal processing, noise values affecting these amplitudes are anticipated to be correlated (Pawula, Rice, and Roberts [3]). The estimation of covariance between these noise components (in addition to the estimation of other parameters) corresponds to the case of known *vector* signals in unknown Gaussian noise. The sequence of symbols (at the data source) given to the transmitter may be naturally correlated in some applications. Even if they are not, the communication system may encode them into convolutional (or other forms of) codes (Proakis [1]) for higher detection accuracy. This case requires the parameter estimator to deal with a Markov sequence of symbols (pattern classes). In the case of coherent demodulation, again, signal processing can be in a narrow band which results in successive noise vectors to form a statistically dependent sequence. All these cases constitute extensions of the basic univariate case with an i.i.d. sequence of mixture samples. These extensions are also studied in this paper and convergent estimators are developed. Many cases are demonstrated with simulation results.

B Literature survey

Beginning with the century old Pearson’s [4] work on the method of moments to estimate the five parameters of a mixture of two univariate normal densities, a large number of results on Gaussian mixture parameter estimation have been reported. The book by Titterton *et al.* [5] covers all topics of mixture densities and comparatively discusses techniques for the estimations

of their parameters. Redner and Walker [6] is a comprehensive survey paper on the estimation of parameters of the mixtures. Other key references on the topic of mixture densities are Charlier [7], Pearson and Lee [8], Burrau [9], Preston [10], Dick and Boeden [11], and Rao [12]. Rao [12] assumed equal variances of the two component densities in Pearson's problem, used four sample moments, and constructed a cubic equation. Some authors extended Pearson's method of moments to more general mixtures of normal densities and to mixtures of other continuous densities. Pollard [13] studied the mixture of three univariate normal densities. Cooper [14] and Day [15] have extended Rao's method to the mixture of multivariate normals with a common covariance matrix. The mixture of two exponential densities is studied by Gumbel [16] and Rider [17], and the mixture of two gamma distributions, by John [18].

Kazakos [19] reports results on maximum likelihood and recursive estimation of the mixing probabilities only, in multiclass finite mixtures. Dattatreya and Kanal [20] and [21] approach the same problem from a different direction. They use sample mean vectors of nonlinear transformations of the data and derive a convergent estimator for any estimable problem, in [20]. They also develop an asymptotically efficient estimator for the same parameters, in [21].

The recent book *Finite Mixture Models* [22] by McLachlan and Peel gives extensive coverage to the use of EM algorithm for fitting finite mixture models, especially normal mixture models. The choice of the root (from the multiple maxima of the likelihood function), the problem of unbounded likelihood function in the case of unequal variances, and the choice of the number of components are studied in depth. Figueiredo and Jain [23] propose a single unsupervised algorithm that is capable of selecting the number of components and without requiring careful initialization, unlike the standard EM algorithm. They seamlessly integrate estimation and model selection by annihilating unnecessary components of the mixture in the maximization step of the algorithm. Their experiments involving Gaussian mixtures testify for the good performance of the approach.

Dattatreya [24] develops a joint estimator for the prior probabilities and the (possibly) different class conditional variances, when the class-mean values of the data are known. The estimator converges if the following conditions are satisfied: (a) The estimator equations possess a unique vector root in the region of the unknown parameters and (b) the class dependent variances are small and bounded by a specified quantity. Condition (b) is usually satisfied in digital communication where the noise variances (which may be symbol-dependent due to nonlinear signal transformations) are

usually small in comparison with the differences in the mean values of transmitted signals corresponding to the different symbols. The uniqueness of the vector root of the estimation equations is hard to examine. A special case of this problem occurs when the class-conditional variance is independent of the class label. This special case also has important applications as mentioned in the above subsection on Motivation. We study this special case in this paper and present a solution that overcomes the above restrictive conditions (a) and (b).

C Organization of the paper

We introduce the problem and develop a recursive estimator for the common variance in Section 2. Convergence of the estimator and the expression for the asymptotic variance are proved in Section 3. Section 4 presents a technique to deal with a fixed size data set and a simulation result. Section 5 extends the estimation techniques and proofs to dependent cases of class label and noise sequence. Section 6 concludes the paper. Some cumbersome derivations are dealt with in Appendix I and Appendix II.

2 Development of the Estimator

A sequence of outcomes of i.i.d. random variables, $x^n = \{x_1, \dots, x_n\}$, arrives. Each sample is an outcome of the random variable X with the density function

$$p(x) = \sum_{i=1}^M P_i p(x|\omega_i). \quad (1)$$

The above mixture density is characterized by the following.

$$p(x|\omega_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu_i)^2}{2\sigma^2}\right), \quad i = 1, \dots, M. \quad (2)$$

$p(x|\omega_i)$, $i = 1, \dots, M$ are the class conditional probability densities of X with known μ_i . The row vector composed of $\mu_j^2, j = 1, \dots, M$ is denoted by \mathcal{M} . The common variance of the Gaussian components is σ^2 which is unknown.

M is the known finite number of classes.

$0 < P_i < 1$, $i = 1, \dots, M$ are the unknown prior or mixing probabilities. \mathbf{P} is the column vector of these prior probabilities.

An estimator is required for the prior probabilities in (1) as well as the variance of the class conditional density functions in (2). The assumption $P_i > 0$ is very realistic and avoids complications of generating finite samples from a class, in a run of unbounded number of mixture samples. Let $\hat{P}_{i(n)}$ and $\hat{\sigma}_{(n)}^2$ denote the estimates of P_i and σ^2 , respectively, using n samples; $\hat{\mathbf{P}}_{(n)} = [\hat{P}_{1(n)}, \dots, \hat{P}_{M(n)}]^T$ is the vector estimate of prior probabilities. Random variables are denoted by upper case letters; their outcomes, by lower case letters. Bold letters are used to indicate vectors. Matrices are denoted by bold upper case letters, their inverses and transposes with superscripts -1 and T , respectively.

To develop an estimator for σ^2 , define

$$h_i(x) = \frac{1}{\sqrt{2\pi\rho^2}} \exp\left(-\frac{(x - \mu_i)^2}{2\rho^2}\right), i = 1, \dots, M \quad (3)$$

as M functions of the data x with a control parameter ρ . Let $\mathbf{H}(\sigma^2)$ be an $M \times M$ matrix with elements

$$\begin{aligned} h_{ij} &= E[h_i(x)|\omega_j] \\ &= \frac{1}{\sqrt{2\pi(\sigma^2 + \rho^2)}} \exp\left(-\frac{(\mu_i - \mu_j)^2}{2(\sigma^2 + \rho^2)}\right) \quad i, j = 1, \dots, M. \end{aligned} \quad (4)$$

Note that h_{ij} are continuous functions of σ^2 and $\mathbf{H}(\sigma^2)$ is invertible for any finite σ^2 , as shown in [20]. Also, $h_i(X)$ have finite variances. As in [20],

$$\mathbf{P} = \mathbf{H}^{-1}(\sigma^2) \begin{bmatrix} E[h_1(X)] \\ \vdots \\ E[h_M(X)] \end{bmatrix} = \mathbf{H}^{-1}(\sigma^2) E[\mathbf{h}(X)], \quad (5)$$

where $E[\mathbf{h}(X)]$ is the column vector of $E[h_i(X)]$. However, the corresponding estimator

$$\hat{\mathbf{P}}_{(n)} = \mathbf{H}^{-1}(\sigma^2) \frac{1}{n} \sum_{i=1}^n \begin{bmatrix} h_1(x_i) \\ \vdots \\ h_M(x_i) \end{bmatrix} \quad (6)$$

is unrealizable since σ^2 is unknown. But (6) contains σ^2 only in the matrix \mathbf{H} ; the data transformations $h_i(x_k)$ are realizable. It is possible to reorder (6) to express it in a recursive form. Now,

consider the second moment of X ; we have

$$\begin{aligned} E[X^2] &= \int x^2 p(x) dx \\ &= \sigma^2 + \mathbf{M}P. \end{aligned} \quad (7)$$

Combining (5) and (7),

$$\sigma^2 = E[X^2] - \mathbf{M}\mathbf{H}^{-1}(\sigma^2)E[\mathbf{h}(X)]. \quad (8)$$

We note here that (8) is valid for every unlabeled data sample, even if successive data samples form a dependent sequence. This property is used in Section 5. A recursive estimator for σ^2 is now obtained by combining (6) and (8), as

$$\hat{\sigma}_{(0)}^2 = \text{a positive number}, \quad (9)$$

$$\hat{\sigma}_{(n)}^2 = \frac{n-1}{n}\hat{\sigma}_{(n-1)}^2 + \frac{1}{n}[x_n^2 - \mathbf{M}\mathbf{H}^{-1}(\hat{\sigma}_{(n-1)}^2)\mathbf{h}(x_n)]. \quad (10)$$

In (10), n represents the numbers of samples used to compute the current estimate. To ensure that the argument of \mathbf{H} is a meaningful variance value, we use a clipped version of $\hat{\sigma}_{(n-1)}^2$ defined by

$$\tilde{\sigma}_{(n-1)}^2 = \begin{cases} \hat{\sigma}_{(n-1)}^2 & \text{if } 0 \leq \hat{\sigma}_{(n-1)}^2 \leq \sigma_{\max}^2, \\ \sigma_{\max}^2 & \text{if } \hat{\sigma}_{(n-1)}^2 > \sigma_{\max}^2, \\ 0 & \text{if } \hat{\sigma}_{(n-1)}^2 < 0, \end{cases} \quad (11)$$

where $\sigma_{\max}^2 > \sigma^2$ is some known upper bound on σ^2 . Clipping at the lower limit of 0 is to ensure that the variance argument of $\mathbf{H}^{-1}(\tilde{\sigma}_{(n-1)}^2)$ in (12) is non-negative. Clipping at 0 does not preclude $\sigma^2 = 0$ in the mixture data model. If σ^2 is known to be less than σ_{\max}^2 , clipping at the upper bound helps in using a better approximation for $\mathbf{H}^{-1}(\sigma^2)$ in (12). A more general estimator with α as a control parameter (in addition to the control parameter ρ^2) is

$$\hat{\sigma}_{(n)}^2 = \hat{\sigma}_{(n-1)}^2 - \frac{\alpha}{n}[\hat{\sigma}_{(n-1)}^2 - x_n^2 + \mathbf{M}\mathbf{H}^{-1}(\tilde{\sigma}_{(n-1)}^2)\mathbf{h}(x_n)]. \quad (12)$$

For prior probabilities, we propose

$$\hat{P}_{(0)} = \left[\frac{1}{M}, \dots, \frac{1}{M}\right]^T, \quad (13a)$$

$$\hat{P}_{(n)} = \begin{cases} \mathbf{H}^{-1}(\tilde{\sigma}_{(n-1)}^2)\frac{1}{n}\sum_{i=1}^n \mathbf{h}(x_i) & \text{if this evaluates to a valid probability vector,} \\ \hat{P}_{(n-1)} & \text{otherwise; } n = 1, 2, \dots \end{cases} \quad (13b)$$

The value of $\frac{1}{n} \sum_{i=1}^n \mathbf{h}(x_i)$ can be obtained in a recursive manner. Alternatively, a scaled and normalized version of the first part of (13b) can be used to ensure that $\hat{\mathbf{P}}_{(n)}$ is always a valid probability vector.

3 Convergence Analysis

In the following, we analyze the asymptotic properties of (12) and (13). Replace the outcomes x_n and $h(x_n)$ in (12) by their mean values and compensate for the error by using y_n as the outcome of a corresponding zero mean random variable Y_n . The resulting random variable estimator is

$$\hat{\sigma}_{(n)}^2 = \hat{\sigma}_{(n-1)}^2 - \frac{\alpha}{n} \left[\hat{\sigma}_{(n-1)}^2 - E[X^2] + \mathcal{M}\mathbf{H}^{-1}(\tilde{\sigma}_{(n-1)}^2)E[\mathbf{h}(X)] \right] + \frac{\alpha}{n}Y_n, \quad n = 1, 2, \dots, \quad (14)$$

where

$$Y_n = X_n^2 - E[X^2] - \mathcal{M}\mathbf{H}^{-1}(\tilde{\sigma}_{(n-1)}^2)\{\mathbf{h}(X_n) - E[\mathbf{h}(X)]\}. \quad (15)$$

Similar to the comment following (8), Y_n is zero mean for every unlabeled data sample X_n , even if $\{X_i\}$ is a dependent sequence of random variables. Adding and subtracting

$$\frac{\alpha}{n}\mathcal{M}\mathbf{H}^{-1}(\sigma^2)E[\mathbf{h}(X)] = \frac{\alpha}{n}\mathcal{M}\mathbf{P} = \frac{\alpha}{n} \left[E[X^2] - \sigma^2 \right] \quad (16)$$

to the RHS of (14), we have

$$\begin{aligned} \hat{\sigma}_{(n)}^2 &= \hat{\sigma}_{(n-1)}^2 - \frac{\alpha}{n} \left[\hat{\sigma}_{(n-1)}^2 - E[X^2] + \mathcal{M}\mathbf{P} \right] \\ &\quad - \frac{\alpha}{n}\mathcal{M} \left[\mathbf{H}^{-1}(\tilde{\sigma}_{(n-1)}^2) - \mathbf{H}^{-1}(\sigma^2) \right] E[\mathbf{h}(X)] + \frac{\alpha}{n}Y_n \end{aligned} \quad (17)$$

$$= \hat{\sigma}_{(n-1)}^2 - \frac{\alpha}{n} \left[\hat{\sigma}_{(n-1)}^2 - \sigma^2 + \mathcal{M} \left\{ \mathbf{H}^{-1}(\tilde{\sigma}_{(n-1)}^2)\mathbf{H}^{-1}(\sigma^2) - \mathbf{I} \right\} \mathbf{P} \right] + \frac{\alpha}{n}Y_n \quad (18)$$

$$= \hat{\sigma}_{(n-1)}^2 - \frac{\alpha}{n}f(u, v) + \frac{\alpha}{n}Y_n \quad (19)$$

where $u = \hat{\sigma}_{(n-1)}^2 + \rho^2$, $v = \sigma^2 + \rho^2$, and $f(u, v)$, is the quantity within the square brackets in (19). The behavior of $f(u, v)$ influences the convergence properties of the estimator. The scalar function $f(u, v)$ can also be expressed as

$$f(u, v) = \left[(\hat{\sigma}_{(n-1)}^2 - \sigma^2)[1, \dots, 1] + \mathcal{M} \left\{ \mathbf{H}^{-1}(\tilde{\sigma}_{(n-1)}^2)\mathbf{H}^{-1}(\sigma^2) - \mathbf{I} \right\} \right] \mathbf{P} \quad (20)$$

$$= \mathbf{F}(u, v)\mathbf{P} \quad (21)$$

$$= (u - v)\mathbf{G}(u, v)\mathbf{P}. \quad (22)$$

The last two equations define the row vectors $\mathbf{F}(u, v)$ and $\mathbf{G}(u, v)$. The components of the row vector $\mathbf{G}(u, v)$ turn out to be continuous and differentiable in the region of interest of (u, v) . The minimum of the M components of the row vector $\mathbf{G}(u, v)$ is defined as the scalar $g(u, v)$ and

$$g^* = \min_{\forall v} g(u, v). \quad (23)$$

The minimum above is taken over any known possible region of v , determined by ρ^2 and σ_{\max}^2 . Appendix I conducts a brief study of some properties of the above functions. The conclusions there are as follows.

1. We can easily obtain a numerical plot of a Boolean function of $\text{sign}[g(u, v)]$ on a two-dimensional plane segment of (u, v) and identify the regions satisfying the required properties, for any specific problem specifications.
2. Most practical problems result in plots with a sufficient and a convenient region for

$$\rho^2 < u, v < \rho^2 + \sigma_{\max}^2. \quad (24)$$

3. If there is such a sufficient region as above, we can easily choose a value for the control parameter ρ^2 to satisfy (24), from the plot.

THEOREM 1: Assume that the $g(u, v) > 0$ in the region of the unknown σ^2 , for a chosen ρ^2 . Let a large α be chosen such that the inequality $2\alpha g^* > 1$ is satisfied. Then the sequence of estimates $\hat{\sigma}_{(n)}^2$ given by (12) converges [25] to σ^2 in the mean square sense (m.s.) and with probability 1 (w.p.1). Furthermore, the asymptotic variance of $\hat{\sigma}_{(n-1)}^2$ is given by

$$\lim_{n \rightarrow \infty} n^{\frac{1}{2}} \left\{ E[(\hat{\sigma}_{(n)}^2)] - \left(E[\hat{\sigma}_{(n)}^2] \right)^2 \right\} = \frac{\alpha \psi^2}{2\alpha \mathbf{G}(\rho^2 + \sigma^2, \rho^2 + \sigma^2) \mathbf{P} - 1} \quad (25)$$

where ψ^2 is the variance of Y_n evaluated at $\hat{\sigma}_{(n-1)}^2 = \sigma^2$. An expression for ψ^2 is derived in Appendix II.

PROOF: The random variable estimator in (19) has a form equivalent to the Robbins-Monro stochastic approximation procedure studied in Sacks ([26], Section 3, pp. 379 - 381). Other than our use of n in place of $n + 1$ in [26], they are identical. The following verifies the conditions in [26] sufficient for the convergence of $\hat{\sigma}_{(n)}^2$ in (19).

The domain of each of u, v is limited to $[\rho^2, \rho^2 + \sigma_{\max}^2]$. From (15), $E[Y_n] = 0$ for every possible $\hat{\sigma}_{(n)}^2$.

$$(u - v)f(u, v) > 0, \forall u \neq v \quad (26)$$

satisfying Assumption (A1) in Sacks ([26], p. 379). From (58) and Remark 4 in Appendix I,

$$\frac{f(u, v)}{u - v} = \frac{\mathbf{F}(u, v)\mathbf{P}}{u - v} = \mathbf{G}(u, v)\mathbf{P} \quad (27)$$

is positive and bounded from above and below by strictly positive finite values for all u including for $u = v$. This satisfies Assumption (A2) in Sacks ([26], p. 380). Remark 5 in Appendix I satisfies Assumption (A3) in Sacks ([26], p. 380). Invertibility of $\mathbf{H}(\hat{\sigma}_{(n-1)}^2)$ and bounds on $\mathbf{h}(x_n)$ in (15) implies that

$$\sup_{\hat{\sigma}_{(n-1)}^2} E[Y_n^2] < \infty. \quad (28)$$

From Appendix II,

$$\lim_{\hat{\sigma}_{(n)}^2 \rightarrow \sigma^2} E[Y_n^2] = \psi^2. \quad (29)$$

This satisfies Assumption (A4) in Sacks ([26], p. 380). From (15), Y_n is continuous as $\hat{\sigma}_{(n-1)}^2$ varies around σ^2 . Also,

$$y_n = O(x_n^2), \text{ as } |y_n| \rightarrow \infty. \quad (30)$$

Since X is a finite Gaussian mixture with finite mean components, we have

$$\lim_{R \rightarrow \infty} \lim_{\epsilon \rightarrow 0^+} \sup_{|u-v| < \epsilon} \int_{|y_n|^2 > R} y_n^2 p(y_n) dy_n = 0, \quad (31)$$

satisfying Assumption (A5) in Sacks ([26], p. 380). Therefore, following the Theorem in Sacks ([26], p. 381), we have that

$$\lim_{n \rightarrow \infty} n^{\frac{1}{2}} \left\{ E[(\hat{\sigma}_{(n)}^2)] - \left(E[\hat{\sigma}_{(n)}^2] \right)^2 \right\} = \frac{\alpha \psi^2}{2\alpha \mathbf{G}(\rho^2 + \sigma^2, \rho^2 + \sigma^2) \mathbf{P} - 1} \quad (32)$$

which of course implies convergence in the mean square sense, provided α is such that the denominator in the above expression is positive. Convergence with probability 1 follows under even slightly weaker condition, as noted by Sacks ([26], p. 380), following his statement of Assumption (A5). This concludes the proof. \square

It is very difficult to compare the asymptotic variance in (32) with the Cramer-Rao lower bound for the following reasons. Integrations to derive the components of the Fisher's information matrix

for the joint estimation of σ^2 and the $M - 1$ independent components of the M prior probabilities cannot be analytically carried out. Analytical inversion of the matrix also appears to be virtually infeasible.

THEOREM 2: Estimator (13) for the prior probabilities converges in m.s. and w.p.1, to true prior probabilities, under the same condition as in *THEOREM 1*.

PROOF: $\mathbf{H}^{-1}(\tilde{\sigma}_{(n-1)}^2)$ is continuous at $\tilde{\sigma}_{(n-1)}^2 = \sigma^2$, and $\tilde{\sigma}_{(n-1)}^2$ as a function of $\hat{\sigma}_{(n-1)}^2$ is continuous at $\hat{\sigma}_{(n-1)}^2 = \sigma^2$. *THEOREM 1* shows $\hat{\sigma}_{(n-1)}^2 \xrightarrow{wp1} \sigma^2$. Therefore, from Serfling [27, Theorem, p. 24] we have $\tilde{\sigma}_{(n-1)}^2 \xrightarrow{wp1} \sigma^2$, and $\mathbf{H}^{-1}(\tilde{\sigma}_{(n-1)}^2) \xrightarrow{wp1} \mathbf{H}^{-1}(\sigma^2)$. We know that $\frac{1}{n} \sum_{i=1}^n h(x_i) \xrightarrow{wp1} E[h(X)]$ even if successive pattern samples form a regular Markov chain [28]. Indeed, even the added Gaussian noise sequence does not need to be white for the above convergence³. $\hat{\mathbf{P}}_{(n)}$ is defined in (13b) to be $\hat{\mathbf{P}}_{(n-1)}$ only if $H^{-1}(\tilde{\sigma}_{(n-1)}^2) \frac{1}{n} \sum_{i=1}^n \mathbf{h}(x_i)$ is not valid probability vector. However, the assumption $P_i > 0$ ensures that around converging ordered pair point $(\sigma^2, E(\mathbf{h}(X)))$, the function $\hat{\mathbf{P}}_{(n)}$ is continuous. Hence the theorem in Serfling [27, p. 24] is applicable and shows the convergence of $\hat{\mathbf{P}}_{(n)}$, the function of $\tilde{\sigma}_{(n-1)}^2$ and $\frac{1}{n} \sum_{i=1}^n h(x_i)$, to the true prior probability vector, \mathbf{P} , w.p.1. (13b) also bounds the components of $\hat{\mathbf{P}}_{(n)}$ by a constant vector $[1, \dots, 1]^T$. It follows from Serfling [27, Theorem, p. 11], that

$$\lim_{n \rightarrow \infty} E(\hat{\mathbf{P}}_{(n)} - \mathbf{P})^2 = 0, \quad (33)$$

concluding the proof. □

4 Discussion and Simulation Experiment

A Batch processing with a finite number of samples

In digital transmission with on-line parameter estimation, the recursive estimator developed above is well suited. In some applications, the parameters are required to be estimated after collecting a fixed size data set. This requires a non-recursive estimator. To accomplish this, the approach developed above is easily modified as follows. Reproducing (8), we have

$$\sigma^2 = E[X^2] - \mathcal{M}\mathbf{H}^{-1}(\sigma^2)E[\mathbf{h}(X)]. \quad (34)$$

³The absolute value of the correlation coefficient between noise added to any two signal samples should merely be less than unity.

Under the conditions developed for the convergence of the estimator, the above equation in the one unknown σ^2 has exactly one root in $0 \leq \sigma^2 \leq \sigma_{\max}^2$ which can be very easily evaluated numerically. Therefore, the algorithm for estimation with finite sample set is to substitute $E[X^2]$ above by the sample average of the second moment moment of the data and substitute $E[\mathbf{h}(X)]$ by the corresponding sample average of the $\mathbf{h}(x_i), i = 1, \dots, n$ and numerically evaluate σ^2 . The uniqueness and the limited one-dimensional range for the root ensures a very efficient computation. The $M + 1$ sample averages used in this technique are the sufficient statistics, so that the individual data points are not required after the calculation of sufficient statistic. This method also gives the unique estimates for the prior probabilities through the use of (5) with the estimated variance. In contrast, the EM algorithm [22] requires repeated evaluation of densities for candidate parameter values as the parameters are iteratively updated. The EM algorithm is also known to converge to one of the many possible local maxima of the likelihood functions. Furthermore, after the initial rapid approach towards a maxima, the batch processing EM algorithm is known to be very slow in convergence.

B A simulation experiment

Several simulation experiments were conducted. For the problem of univariate mixture of four classes with means 0.0, 1.0, 2.0, and 3.0, respectively, we attempted with several different values for ρ . and obtained anticipated converging estimates. Fig. 1 shows the plot of these estimates for P_1, \dots, P_4 , and σ^2 upto 2,000 samples. The class probabilities P_1, \dots, P_4 and σ^2 are indicated in the simulation plots. The initial value for the estimator is chosen as $\sigma_{(0)}^2 = 0.25$, a quarter fraction of the square of the minimum difference between class means. For this and other cases, we also simulated with an adaptive value for ρ as $\rho_n^2 = \tilde{\sigma}_{(n-1)}^2$. Results showed good convergence.

5 Extensions

A Multivariate Gaussian mixture

The estimators (12) and (13) can be extended to the case of multivariate Gaussian class conditional densities. Each sample in the sequence $\mathbf{x}_1, \mathbf{x}_2, \dots$ is now a d -component column vector and is an

outcome of \mathbf{X} with density function

$$\begin{aligned} p(\mathbf{x}) &= \sum_{i=1}^M P_i p(\mathbf{x}|\omega_i) \\ &= \sum_{i=1}^M P_i \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)\right], \end{aligned} \quad (35)$$

where $\boldsymbol{\mu}_i$ is a d -component mean vector for class ω_i , $\boldsymbol{\Sigma}$ is the $d \times d$ covariance matrix, and $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$. The matrix of second moments is

$$\begin{aligned} E[\mathbf{X}\mathbf{X}^T] &= \sum_{i=1}^M P_i E[\mathbf{X}\mathbf{X}^T|\omega_i] \\ &= \boldsymbol{\Sigma} + \sum_{i=1}^M P_i \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T. \end{aligned} \quad (36)$$

Each of the diagonal elements of $\boldsymbol{\Sigma}$ can be estimated in same manner as in (12); for $i = 1, \dots, d$,

$$\hat{\sigma}_{ii(n)} = \hat{\sigma}_{ii(n-1)} - \frac{\alpha}{n} [\hat{\sigma}_{ii(n-1)} - x_{ni}^2 + \boldsymbol{\mathcal{M}}_i \mathbf{H}_i^{-1}(\tilde{\sigma}_{ii(n-1)}) \mathbf{h}(x_{ni}, \rho_i)], \quad (37)$$

where x_{ni} is the i th component of the vector \mathbf{x}_n and

$$\mathbf{h}(x_{ni}, \rho_i) = [h_1(x_{ni}, \rho_i), \dots, h_2(x_{ni}, \rho_i)]^T \quad (38)$$

with

$$h_j(x_{ni}, \rho_i) = \frac{1}{\sqrt{2\pi\rho_i^2}} \exp\left(-\frac{(x_{ni} - \mu_{ji})^2}{2\rho_i^2}\right). \quad (39)$$

The prior probability vector can be estimated by using the average of the estimates of $\hat{\mathbf{P}}(n)$ over $\tilde{\sigma}_{ii(n)}$ as follows.

$$\hat{\mathbf{P}}(n) = \begin{cases} \frac{1}{d} \sum_{i=1}^d [\mathbf{H}_i^{-1}(\tilde{\sigma}_{ii(n-1)}) \frac{1}{n} \sum_{k=1}^n \mathbf{h}(x_{ki}, \rho_i)] & \text{if RHS is a valid probability vector,} \\ \hat{\mathbf{P}}(n-1) & \text{otherwise.} \end{cases} \quad (40)$$

(40) is the average of convergent estimates of (13b), so it is convergent. For the off-diagonal elements of $\boldsymbol{\Sigma}$, $\sigma_{ij}; i \neq j; i, j = 1, \dots, d$, note that

$$\sigma_{ij}^2 < \sigma_{ii} \sigma_{jj}. \quad (41)$$

We estimate σ_{ij} by

$$\hat{\sigma}_{ij(n)} = \begin{cases} \frac{1}{n} \sum_{k=1}^n x_{ki} x_{kj} - \sum_{m=1}^M \hat{P}_{m(n)} \mu_{mi} \mu_{mj} & \text{if } 0 < \hat{\sigma}_{ij(n)}^2 < |\hat{\sigma}_{ii(n)} \hat{\sigma}_{jj(n)}|, \\ \hat{\sigma}_{ij(n-1)} \sqrt{|\hat{\sigma}_{ii(n)} \hat{\sigma}_{jj(n)}|} / \sqrt{|\hat{\sigma}_{ii(n-1)} \hat{\sigma}_{jj(n-1)}|} & \text{otherwise.} \end{cases} \quad (42)$$

The second part of (42) ensures that the estimate of Σ is a valid covariance matrix at every n .

THEOREM 3: The estimates of covariances in (42) converge to the true covariances in m.s. and w.p.1.

PROOF: The sample mean of the product $X_{ki}X_{kj}$ (viewed as a sequence of k) and $\hat{P}_{(n)}$ converge to finite values w.p.1 and in m.s. (even if $\{\mathbf{x}_n\}$ is a dependent sequence). Use of Serfling [27, Theorem, p. 24] establishes the convergence of $\hat{\sigma}_{ij(n)}$ to σ_{ij} w.p.1. $\hat{\sigma}_{ij(n)}$ is bounded by $\sqrt{|\hat{\sigma}_{ii(n)}\hat{\sigma}_{jj(n)}|}$ for all n . And $\hat{\sigma}_{ii(n)}\hat{\sigma}_{jj(n)}$ is known to converge in m.s. to constant. Hence by Serfling [27, Theorem, p. 11], $\hat{\sigma}_{ij(n)}$ converges to σ_{ij} in m.s. also. \square

Although (35) in the beginning of this section appears to imply an assumption that Σ is non-singular, we do not find Σ^{-1} or $(\hat{\Sigma}_{(n)})^{-1}$ in the estimators, and the convergence properties are valid even if Σ is singular. Hence, a user need not know in advance that the data vector is linearly independent. Fig. 2 and Fig. 3 show the simulation results for a two dimensional Gaussian mixture of four classes. The true values of the four class probabilities are in the vector $\mathbf{P} = [0.34 \ 0.5 \ 0.11 \ 0.05]^T$. The mean vectors of the four classes are $[0.5, 0.5]^T$, $[1.5, 1.5]^T$, $[2.5, 2.5]^T$, and $[3.5, 3.5]^T$ respectively. The common covariance matrix of the bivariate data for each class is

$$\Sigma = \begin{bmatrix} 0.125 & 0.1 \\ 0.1 & 0.125 \end{bmatrix}. \quad (43)$$

Fig. 2 shows the estimates of the parameters of the covariance matrix. Fig. 3 shows the plot of the class probabilities. Simulation is carried out upto 1,200 data samples.

B Markov sequence of class labels

We now study dependent sequences. The simplest extension is the case of the stationary finite Markov signal sequence to which stationary white Gaussian noise is added to generate unlabeled pattern samples. The parameters of the model are the mean vectors (known) of the M classes, the common covariance matrix Σ , and the $M \times M$ transition probability matrix \mathbf{Q} . This is also known as the hidden Markov model [22]. We assume that the Markov chain is regular [29], as is the case in most practical problems. The prior probability vector \mathbf{P} is a function of \mathbf{Q} . The most appealing application of such a model is in the reception of a sequence of Markov symbols transmitted over an additive white Gaussian noise channel. The receiver requires the data statistics, \mathbf{Q} , as well as noise statistics, Σ for a successful implementation of the Viterbi Algorithm [1]. The dependent sequence

of data random variables $\{X_i\}$ are identical and each is distributed as X . As long as we consider the statistics of one such random variable (without any condition of earlier random variables), (8) is valid and takes the form

$$\sigma^2 = E[X_i^2] - \mathbf{M}\mathbf{H}^{-1}(\sigma^2)E[\mathbf{h}(X_i)]. \quad (44)$$

Similarly, the derivation of the various forms of the estimator in (10) through (19) are also valid for a dependent sequence $\{X_i\}$. Conditions in the Sacks' SA theorem are not changed for such a dependent sequence. Hence, we have the following conclusion.

THEOREM 4: The estimator $\hat{\Sigma}_{(n)}$ defined by (37) and (42) converges to Σ w.p.1 and in m.s. even for the present case of Markov sequence of signals. \square

Indeed, by the same argument, the estimator converges for other statistical structures [30] such as in the patterns forming a 2-dimensional (non-degenerate) discrete random field lattice with additive white Gaussian noise. Our estimator for Σ is meaningful only if \mathbf{P} is unknown. So, in the case of the present hidden Markov model, simultaneous estimation of \mathbf{Q} is important. [28] develops a convergent estimator for \mathbf{Q} which requires the full knowledge of Σ . The following straightforward modification extends it to the case of unknown Σ : Estimate Σ using the procedures developed in earlier sections. Simultaneously, estimate \mathbf{Q} as in [28] with the only difference being the use of $\hat{\Sigma}_{(n)}$ and the \mathbf{h} functions used in Section 5 A. Convergence of \mathbf{Q} follows from the convergence of $\hat{\Sigma}_{(n)}$ and the use of converging $\hat{\Sigma}_{(n)}$ in estimating \mathbf{Q} . Similar conclusions hold for the multidimensional lattice of patterns studied in [30], provided the additive Gaussian field is white.

Simulation experiments with a univariate mixture data of four classes were conducted. The sequence of class labels of successive data samples form a regular Markov chain with transition probabilities θ_{ij} , i being the class of the present sample and j , the class of the next sample in the sequence of data samples. The transition probabilities uniquely determine the stationary class probabilities and the latter are not plotted. σ^2 is the variance of the additive white noise. Simulation is conducted upto 2,000 samples. Fig. 4 shows the plot of the estimates of σ^2 . Figs. 5, 6, and 7 plot the transition probabilities. The true values of the parameters are also plotted with horizontal lines and identified. The theoretical error rate for the Viterbi algorithm for the above data parameters is between 8% and 9%. The simulation of the Viterbi algorithm with estimated parameters (based on 2000 samples) resulted in an error rate very close to that with the true parameters.

C Dependent additive Gaussian noise sequence over independent signal sequence

Section 5 B studied the case of dependent class labels in a sequence. We can similarly relax the restriction on the additive noise. The simplest situation is when the dependence of the additive noise sequence is completely representable with the expectations of the products of the components of noise vector added to two successive pattern samples. Suppose this to be the model. Also assume that the absolute value of the corresponding correlation coefficients to be strictly less than unity; this assumption merely eliminates full correlation. Then the estimation of all other parameters (Σ , and P) is unaffected. Neither are their convergence properties, due to the irrelevance of the dependency of Y_n in the Sacks' SA theorem [26]. Therefore, we can concentrate on estimating the dependency of successive noise samples.

Consider the scalar random variable X as the signal U with additive noise V . That is

$$X = U + V, \quad (45)$$

where the sample space for U is $\{\mu_1, \dots, \mu_M\}$ and V is Gaussian with zero mean. The correlation between two successive samples can be written as

$$\begin{aligned} E[X_{n-1}X_n] &= \sum_{i=1}^M \sum_{j=1}^M E[X_{n-1}X_n | \Omega_{n-1} = \omega_i, \Omega_n = \omega_j] P[\omega_i] P[\omega_j] \\ &= \sum_{i=1}^M \sum_{j=1}^M E[(\mu_i + V_{n-1})(\mu_j + V_n)] P[\omega_i] P[\omega_j] \\ &= \sum_{i=1}^M \sum_{j=1}^M P[\omega_i] P[\omega_j] \mu_i \mu_j + E[V_{n-1}V_n]. \end{aligned} \quad (46)$$

Estimation of $P[\omega_i]$ and convergence properties follow the techniques in Section 2. So, $R = E[V_{n-1}V_n]$, the correlation between noises added to two successive signals can be estimated by

$$\hat{R}_{(n)} = \frac{1}{n-1} \sum_{i=1}^{n-1} x_i x_{i+1} - \sum_{i=1}^M \sum_{j=1}^M \hat{P}[\omega_i] \hat{P}[\omega_j] \mu_i \mu_j. \quad (47)$$

Convergence of $\hat{R}_{(n)}$ to R w.p.1 and in m.s. are straightforward to establish, as in earlier sections.

Extension to vector \mathbf{X} to estimate the covariance matrix of the noise vector, and simultaneously, the correlation matrix between successive noise vectors is also straightforward. The estimator for

the correlation matrix of the additive noise vector is

$$\hat{\mathbf{R}}_{(n)} = \frac{1}{n-1} \sum_{i=1}^{n-1} \mathbf{x}_i \mathbf{x}_{i+1}^T - \sum_{i=1}^M \sum_{j=1}^M \hat{P}[\omega_i] \hat{P}[\omega_j] \boldsymbol{\mu}_i \boldsymbol{\mu}_j^T. \quad (48)$$

D Dependent additive noise over dependent class label sequence

Let $\mathbf{x}_1, \mathbf{x}_2, \dots$ be received data vector sequence. Define

$$\mathbf{y}_i = \begin{bmatrix} \mathbf{x}_i \\ \mathbf{x}_{i+1} \end{bmatrix}, i = 1, 2, \dots \quad (49)$$

Now each \mathbf{y}_i is an outcome from one of M^2 classes. Convergent estimation of covariance matrix of \mathbf{Y} as well as the M^2 prior probabilities follow the procedure developed in Section 5 A. The sequence $\{\mathbf{y}_i\}$ and the corresponding class sequence $\{(\omega_i, \omega_{i+1})\}$ are dependent sequences. However, the irrelevance of such dependence follows along the lines argued in Section 3 B.

Dattatreya [28] develops computationally efficient solution to the linear equations for the case of fully known statistics of independent additive noise sequence over a regular Markov sequences of class labels. Such techniques are applicable here also. In particular, the solution of the M^2 linear equations can be obtained with the inversion of an $M \times M$ matrix. Finally, similar techniques and computational improvements are straightforward to work out, for compatible two dimensional image models.

6 Conclusion

In this paper, we have developed and studied a parameter estimator for finite Gaussian mixtures under the assumption of known class means. We used the method of “moments of data transformations” and the procedure of stochastic approximation. The approach is easily modified to a simple non-recursive (batch processing) technique to evaluate the unique vector estimate from a fixed size sample data set. Extensions of the estimation techniques to multivariate data with a common covariance matrix for all classes, Markov sequence of class probabilities, and dependent sequence of the noise sequence are developed. In applications of the approach to digital communication, the signal to noise ratio will be much higher than what is used in simulation experiments, and the speed of convergence is expected to be sufficient. In image processing applications, the number of available samples (pixels) will be large enough to provide estimates close to true values.

7 Summary

Estimation of parameters of a multiclass finite Gaussian mixture has many applications in pattern recognition, image processing, and communication engineering. This paper studies parameter estimators for such mixtures under the assumption of known class means. Example problems in digital communication are described to motivate our study. The development uses the method of “moments of data transformations” and the procedure of stochastic approximation.

The simplest case of the problem is the estimation of the prior probability vector \mathbf{P} and common variance σ^2 of the M components of a univariate Gaussian mixture density with known mean values μ_1, \dots, μ_M , from an i.i.d. sequence of mixture samples $\{x_n\}$. We solve this in Section 2 as follows. Estimate the noise statistics assuming knowledge of true signal probabilities; substitute the estimates of signal probabilities in place of true signal probabilities. This approach yields a recursive estimator expression for the noise statistic, without containing any explicit reference to signal probabilities or their estimates.

Section 3 deals with the convergence properties of the estimators, and their proofs. We demonstrate that the mean vectors for most practical problems yield feasible values for the control parameter ρ^2 which can be obtained from numerical plots of a specified Boolean function over a two dimensional plane segment. An application of a Stochastic Approximation Theorem proves the convergence and helps in deriving the asymptotic variance of the estimates of the unknown variance.

In Section 4, a batch processing alternative to the recursive estimator is derived. It uses transformations of all the available data once and computes sufficient statistics. These are used with numerical root finding of a well behaved function in a limited interval. Therefore, the numerical algorithm converges very efficiently.

Extensions to vector data, Markov signals, and dependent noise sequence are concisely worked out in Section 5. Convergent estimators are developed for all unknown quantities for each model. The most general case is the following: Class labels form a discrete random field over a k -dimensional lattice. Each class label corresponds to a known vector signal. A Gaussian noise vector is added to each signal vector. Noise vectors added to signal vectors at neighboring lattice points are not necessarily independent. However, both the statistical structures of the signal random field and the noise random field are completely representable by joint signal probabilities and joint noise

covariance values in a finite size neighborhood. The most appealing special case preserving key aspects of this generality occurs in the communication of first order Markov sequence of symbols over an additive first order autoregressive Gaussian noise channel. Many cases are demonstrated with simulation experiments.

Acknowledgment

The authors thank John Fonseka for helpful discussions on motivating examples in digital communications.

Appendix I Properties of $f(u, v)$

Let

$$\mathbf{F}(u, v) = (u - v)[1, \dots, 1] + \mathcal{M}\left\{\sqrt{\frac{u}{v}}(\mathbf{I} + \mathbf{B})^{-1}(\mathbf{I} + \mathbf{A}) - \mathbf{I}\right\} \quad (50)$$

be a row vector with M components where \mathbf{A} and \mathbf{B} are $M \times M$ matrices with components

$$a_{ii} = b_{ii} = 0 \quad (51)$$

$$a_{ij} = \left[\exp -\frac{(\mu_i - \mu_j)^2}{2v} \right], i \neq j \quad (52)$$

$$b_{ij} = \left[\exp -\frac{(\mu_i - \mu_j)^2}{2u} \right], i \neq j. \quad (53)$$

The motivation for the above definition is the following. If every component of $\mathbf{F}(u, v)$ has the same sign as $u - v$, then $f(u, v) = \mathbf{F}(u, v)\mathbf{P}$ will also have the same sign as $u - v$, for every possible prior probability vector \mathbf{P} . Differentiation of vectors and matrices is carried out element by element, with respect to a specified single variable.

Remark 1: As a consequence of the invertibility of $\mathbf{H}(\tilde{\sigma}_{(n-1)}^2)$, $(\mathbf{I} + \mathbf{B})$ and $(\mathbf{I} + \mathbf{A})$ are invertible for all finite $u, v > 0$.

Remark 2: We can differentiate $\mathbf{Z} = (\mathbf{I} + \mathbf{B})^{-1}$ wrt u by differentiating both sides of

$$\mathbf{Z}(\mathbf{I} + \mathbf{B}) = \mathbf{I}. \quad (54)$$

Using this approach, we arrive at

$$\frac{d}{du}\mathbf{F}(u, v) = [1, \dots, 1] + \frac{1}{2\sqrt{uv}}\mathcal{M}(\mathbf{I} + \mathbf{B})^{-1}\left[\mathbf{I} - \frac{1}{u}\mathbf{C}(u)(\mathbf{I} + \mathbf{B})^{-1}\right](\mathbf{I} + \mathbf{A}). \quad (55)$$

where the $M \times M$ matrix $\mathbf{C}(u)$ is composed of elements $c_{ij} = b_{ij}(\mu_i - \mu_j)^2$. Thus, $\frac{d}{du}\mathbf{F}(u, v)$ exists for all finite $u, v > 0$.

Remark 3:

$$\lim_{u \rightarrow v^+} \frac{d}{du}\mathbf{F}(u, v) = \lim_{u \rightarrow v^-} \frac{d}{du}\mathbf{F}(u, v) \quad (56)$$

$$= [1, \dots, 1] + \frac{1}{2v}\mathcal{M} - \frac{1}{2v^2}\mathcal{M}(\mathbf{I} + \mathbf{A})^{-1}\mathbf{C}(v) \quad (57)$$

Denote the above derivative by $\mathbf{F}'(v, v)$. Note that $f'(v, v) = \mathbf{F}'(v, v)\mathbf{P}$.

Remark 4: Let

$$\mathbf{G}(u, v) = \frac{\mathbf{F}(u, v)}{u - v} = [1, \dots, 1] + \mathcal{M} \frac{\sqrt{\frac{u}{v}}(\mathbf{I} + \mathbf{B})^{-1}(\mathbf{I} + \mathbf{A}) - \mathbf{I}}{u - v}. \quad (58)$$

Evaluation of $\mathbf{G}(v, v) = \lim_{u \rightarrow v} \mathbf{G}(u, v)$ using L'Hospital's rule results in $\mathbf{G}(v, v) = \mathbf{F}'(v, v)$

Remark 5: Let

$$\mathbf{F}(u, v) = (u - v)\mathbf{G}(v, v) + \boldsymbol{\delta}(u, v). \quad (59)$$

The quantity $\boldsymbol{\delta}(u, v)$ is a matrix. Then,

$$\lim_{u \rightarrow v^+} \frac{\boldsymbol{\delta}(u, v)}{u - v} = \lim_{u \rightarrow v^-} \frac{\boldsymbol{\delta}(u, v)}{u - v} \quad (60)$$

$$= \lim_{u \rightarrow v} \frac{\mathbf{F}(u, v) - \mathbf{F}(v, v)}{u - v} - \mathbf{G}(v, v) \quad (61)$$

$$= \mathbf{F}'(v, v) - \mathbf{G}(v, v) \quad (62)$$

$$= \mathbf{0}. \quad (63)$$

Hence, $\boldsymbol{\delta}(u, v) = o(|u - v|)$ as $u - v \rightarrow 0$.

Remark 6: From (59), for small values of $u - v$, the sign of $\mathbf{F}(u, v)$ is the same as the sign of $\mathbf{G}(v, v)$.

Indeed, evaluating the limit, we find that

$$\lim_{v \rightarrow 0} \mathbf{G}(v, v) = \infty. \quad (64)$$

This implies the existence of a bound λ such that

$$\frac{\mathbf{F}(u, v)}{u - v} > \mathbf{0}, \quad \forall 0 < u, v < \lambda. \quad (65)$$

Remark 7: It is important for $f(u, v)$ to have the same sign as $u - v$ for the range of u and v and for all possible prior probability vectors \mathbf{P} . Let

$$g(u, v) = \min\{\mathbf{G}(u, v)\} \quad (66)$$

where the minimum is taken over the M components of the row vector $\mathbf{G}(u, v)$. One approach to satisfy the required condition is to attempt to choose ρ^2 such that $g(u, v)$ is positive for u, v in the interval $[\rho^2, \rho^2 + \sigma_{\max}^2]$. The region of positive $g(u, v)$ can be plotted to attempt to choose ρ . We plotted this function for several mean vectors including $[0, 1, 2, 3]$, $[-2, -1, 0, 1]$, $[1, 2, 4, 8]$, and $[-4, -3, 0, 4]$. Many plots yielded no negative value for $g(u, v)$ at all. Some of the plots had a negative region for high v and very low u . The most illustrative negative components are found for the mean vector $[0, 1, 2, 3]$. The plot of $g(u, v)$ for this case is shown in Fig. 8. The curve separates the regions between positive and negative values of $g(u, v)$. In the plot, the region to the left of, above, and below the curve corresponds to positive $g(u, v)$ so that $\mathbf{G}(u, v)\mathbf{P}$ is positive for any possible \mathbf{P} . In the region to the right of the curve (the region enclosed by the curve and the boundary of the plot at $v = 2.7$), $g(u, v)$ is negative indicating that at least one of the components of the vector $\mathbf{G}(u, v)$ is negative so that the corresponding $\mathbf{G}(u, v)\mathbf{P}$ may become negative for certain values of \mathbf{P} .

Finally, we note that most practical cases of mean vectors appear to yield sufficient regions for u, v to ensure positive $g(u, v)$.

Appendix II Variance of Y

The random variable Y is defined as Y_n when $\hat{\sigma}_{(n-1)}^2 = \sigma^2$. In the expression for Y_n , the random variable X_n can be replaced by X . Also, in this Appendix, we use \mathbf{H} to denote $\mathbf{H}(\sigma^2)$, for simpler notation. Thus,

$$Y = X^2 - E[X^2] - \mathcal{M}\mathbf{H}^{-1}\mathbf{h}(X) + \mathcal{M}\mathbf{P} = X^2 - \mathcal{M}\mathbf{H}^{-1}\mathbf{h}(X) - \sigma^2, \quad (67)$$

a zero mean random variable. Therefore, $\text{Variance}[Y] = E[Y^2] = \psi^2$, the notation used in the body of the paper. From (67),

$$\psi^2 = E\left[\left\{X^2 - \mathcal{M}\mathbf{H}^{-1}\mathbf{h}(X)\right\}^2\right] - \sigma^4. \quad (68)$$

Consider

$$E \left[\left\{ X^2 - \mathbf{M}\mathbf{H}^{-1}\mathbf{h}(X) \right\}^2 \right] = E \left[X^4 + \mathbf{M}\mathbf{H}^{-1}\mathbf{h}(X)\mathbf{h}^T(X)\mathbf{H}^{-T}\mathbf{M}^T - 2\mathbf{M}\mathbf{H}^{-1}X^2\mathbf{h}(x) \right] \quad (69)$$

$$= E[X^4] + \mathbf{M}\mathbf{H}^{-1}E[\mathbf{h}(X)\mathbf{h}^T(X)]\mathbf{H}^{-T}\mathbf{M}^T - 2\mathbf{M}\mathbf{H}^{-1}E[X^2\mathbf{h}(X)]. \quad (70)$$

In (70), $E[X^4]$ can be obtained as a mixture of the moments of Gaussian random variables (Papoulis [25]). With this approach,

$$E[X^4|\omega_j] = 3\sigma^4 + 6\sigma^2\mu_j^2 + \mu_j^4 \quad (71)$$

so that

$$E[X^4] = 3\sigma^4 + 6\sigma^2 \sum_{j=1}^M P_j \mu_j^2 + \sum_{j=1}^M P_j \mu_j^4. \quad (72)$$

Next, consider $E[\mathbf{h}(X)\mathbf{h}^T(X)]$, which is the $M \times M$ matrix composed of elements

$$E[h_i(X)h_j(X)] \quad (73)$$

$$= \frac{1}{(2\pi)^{\frac{3}{2}}(\sigma^2 + \rho^2)\sqrt{\sigma^2}} \int_{-\infty}^{\infty} \sum_{k=1}^M P_k \exp \left[-\frac{(x - \mu_i)^2 + (x - \mu_j)^2}{2(\sigma^2 + \rho^2)} - \frac{(x - \mu_k)^2}{2\sigma^2} \right] dx. \quad (74)$$

By completing the square inside the exponential to express the above as a product of an expression and the integral of a perfect Gaussian function, and then integrating, we obtain

$$E[h_i(X)h_j(X)] \quad (75)$$

$$= \sum_{k=1}^M \frac{P_k}{2\pi\sqrt{(\sigma^2 + \rho^2)(3\sigma^2 + \rho^2)}} \quad (76)$$

$$\times \exp \left[-\frac{\sigma^2(\mu_i^2 + \mu_j^2 + \mu_k^2) + \rho^2\mu_k^2}{2\rho^2(\sigma^2 + \rho^2)} + \frac{[\sigma^2(\mu_i + \mu_j + \mu_k) + \rho^2\mu_k]^2}{2\rho^2(3\sigma^2 + \rho^2)(\sigma^2 + \rho^2)} \right]. \quad (77)$$

Next, consider $E[X^2\mathbf{h}(X)]$, a column vector of M components of the form

$$E[X^2\mathbf{h}_i(X)] = \sum_{j=1}^M P_j \int_{-\infty}^{\infty} x^2 \frac{1}{2\pi\sqrt{\sigma^2(\sigma^2 + \rho^2)}} \exp \left[-\frac{(x - \mu_i)^2}{2(\sigma^2 + \rho^2)} - \frac{(x - \mu_j)^2}{2\sigma^2} \right] dx. \quad (78)$$

By completing the square to obtain a perfect Gaussian function with a factor x^2 , and then integrating, we obtain

$$E[X^2\mathbf{h}_i(X)] = \sum_{j=1}^M P_j \frac{\frac{\sigma^2(\sigma^2 + \rho^2)}{2\sigma^2 + \rho^2} + \left(\frac{\sigma^2\mu_i + \sigma^2\mu_j + \rho^2\mu_j}{2\sigma^2 + \rho^2} \right)^2}{\sqrt{2\pi(2\sigma^2 + \rho^2)}} \quad (79)$$

$$\times \exp \left[-\frac{(\sigma^2 \mu_i + \sigma^2 \mu_j + \rho^2 \mu_j)^2}{2\sigma^2(\sigma^2 + \rho^2)(2\sigma^2 + \rho^2)} + \frac{\sigma^2 \mu_i^2 + \sigma^2 \mu_j^2 + \rho^2 \mu_j^2}{2\sigma^2(\sigma^2 + \rho^2)} \right]. \quad (80)$$

Substituting the RHS of (77) and (80) in (70) and the resulting equivalent of (70) in (68), we get an expression for the require ψ^2 . It is very lengthy and not particularly illustrative to reproduce it.

References

- [1] J. G. Proakis, Digital Communications. NY: McGraw-Hill (1995).
- [2] M. K. Simon and M. -S. Alouini, Digital Communication over Generalized Fading Channels: A unified approach to the Performance Analysis. NY: John Wiley (2000).
- [3] R. F. Pawula, S. O. Rice, and J. H. Roberts, Distribution of the phase angle between two vectors perturbed by Gaussian noise, IEEE Trans. Communications (1982) 1828-1840.
- [4] K. Pearson, Contributions to the mathematical theory of evolution, Phil. Trans. Royal Soc. 185A (1894) 71-110.
- [5] D. M. Titterington, A. F. M. Smith, and U. E. Makov, Statistical Analysis of Finite Mixture Distributions. New York: John Wiley and Sons, Inc. (1985).
- [6] R. A. Redner and H. F. Walker, Mixture densities, maximum likelihood, and the EM algorithm, SIAM Review (26) (1984) 195-239.
- [7] C. V. L. Charlier, Researches into the theory of probability, Acta Univ. Lund.(Neue Folge. Abt. 2) 1 (1906) 33-38.
- [8] K. Pearson and A. Lee, On the generalized probable error in multiple normal correlation, Biometrika 6 (1908-09) 59-68.
- [9] C. Burrau, The half-invariants of the sum of two typical laws of errors, with an application to the problem of dissecting a frequency curve into components, Skand. Aktuarietidskrift 17 (1934) 1-6.

- [10] E. J. Preston, A graphical method for the analysis of statistical distributions into two normal components, *Biometrika* 40 (1953) 460-464.
- [11] N. P. Dick and D. C. Boeden, Maximum likelihood estimation for mixture of two normal distributions, *Biometrics* 29, (1973) 781-791.
- [12] C. R. Rao, *Advanced Statistical Methods in Biometric Research*, New York: John Wiley and Sons, Inc. (1952).
- [13] H. S. Pollard, On the relative stability of the median and the arithmetic mean, with particular reference to certain frequency distributions which can be dissected into normal distributions, *Ann. Math. Statist.* 5 (1934) 227-262.
- [14] P. W. Cooper, Some topics on nonsupervised adaptive detection for multivariate normal distributions, in J. T. Tou, ed., *Computer and Information Sciences, II*, NY: Academic Press (1967) 143-146.
- [15] N. E. Day, Estimating the components of a mixture of normal distributions, *Biometrika* 56 (1969) 463-474.
- [16] E. J. Gumbel, La dissection d'une repartition, *Annales de l'Universite de Lyon* 3 A (1939) 39-51.
- [17] P. R. Rider, The method of moments applied to a mixture of two exponential distributions, *Ann. Math. Statist.*, 32 (1961) 143-147.
- [18] S. John, On identifying the population of origin of each observation in a mixture of observations from two gamma populations, *Technometrics* 12 (1970) 565-568.
- [19] D. Kazakos, Recursive estimation of prior probabilities using a mixture, *IEEE Trans. Inform. Theory* 23 (1977) 203-211.
- [20] G. R. Dattatreya and L. N. Kanal, Estimation of mixing probabilities in multiclass finite mixtures, *IEEE Trans. Syst., Man, and Cybern.* 20 (1990) 149-158.
- [21] G. R. Dattatreya, Asymptotically efficient estimation of prior probabilities in multiclass finite mixtures, *IEEE Trans. Information Theory*, 37 (1991) 482-489.

- [22] G. McLachlan and D. Peel, *Finite Mixtures Models*. NY: John Wiley (2000).
- [23] M. A. T. Figueiredo and A. K. Jain, Unsupervised learning of finite mixture models, *IEEE Trans. Pattern Anal. Machine Intell.* 24 (2002) 381 - 396.
- [24] G. R. Dattatreya, Gaussian parameter estimation with known means and unknown class-dependent variances, *Pattern Recognition* 35 (7) (2002) 1611-1616.
- [25] A. Papoulis, *Probability, Random Variables and Stochastic Processes*. NY: McGraw Hill (1984).
- [26] J. Sacks, Asymptotic distribution of stochastic approximation procedures, *Ann. Math Statist.*, 29 (1958) 373 - 405.
- [27] R. J. Serfling, *Approximation Theorems of Mathematical Statistics*, New York: John Wiley and Sons, Inc. (1980).
- [28] G. R. Dattatreya, Estimation of prior and transition probabilities in multiclass finite Markov mixtures, *IEEE Trans. Syst., Man, and Cybern.* 21, (1991) 419-426.
- [29] J. G. Kemeny and J. L. Snell, *Finite Markov Chains*. NY: Van Nostrand (1960)
- [30] G. R. Dattatreya, Unsupervised context estimation in a mesh of pattern classes for image recognition, *Pattern Recognition* 24 (7) (1991) 685-694.

About the Authors – G. R. DATTATREYA received the B. Tech. degree in Electrical Engineering from the Indian Institute of Technology, Madras, M.E. in Electrical Communication Engineering, and Ph.D. from the Department of Computer Science and Automation, Indian Institute of Science, Bangalore, India in 1975, 77, and 81, respectively. During 1981-82, he was a Senior Scientist at the Scientific Analysis Group, Delhi, India, and worked on Pattern Recognition and Speech Processing problems. During 1983-85, he was a Visiting Assistant Professor at the Machine Intelligence and Pattern Analysis Laboratory, Department of Computer Science, University of Maryland, College Park, where he taught and conducted research in Information Processing. He is currently an Associate Professor in the Department of Computer Science, University of Texas at Dallas. During June - Dec. 1996, he was a consultant on the Malaysia Polytechnic Project, Batu Pahat, Johor, Malaysia. During June 1999 - May 2000, he was a Visiting Professor at the Center for Artificial Intelligence, ITESM, Monterrey, Mexico. His current research interests are Stochastic Modeling, Parameter Estimation, and Adaptive Optimization in Communication, Signal Processing, and Computer Network Systems. He is working on Routing and load balancing in ad hoc networks, Tractable models for bursty traffic data packet generation, parameter estimation, and network operation simulation, Medium access control, Blind channel estimation.

XIAORI (FRANK) FANG received his M. S. degree in Computer Science from the University of Texas at Dallas. He has worked for US Data and NORTEL Telecommunication companies.

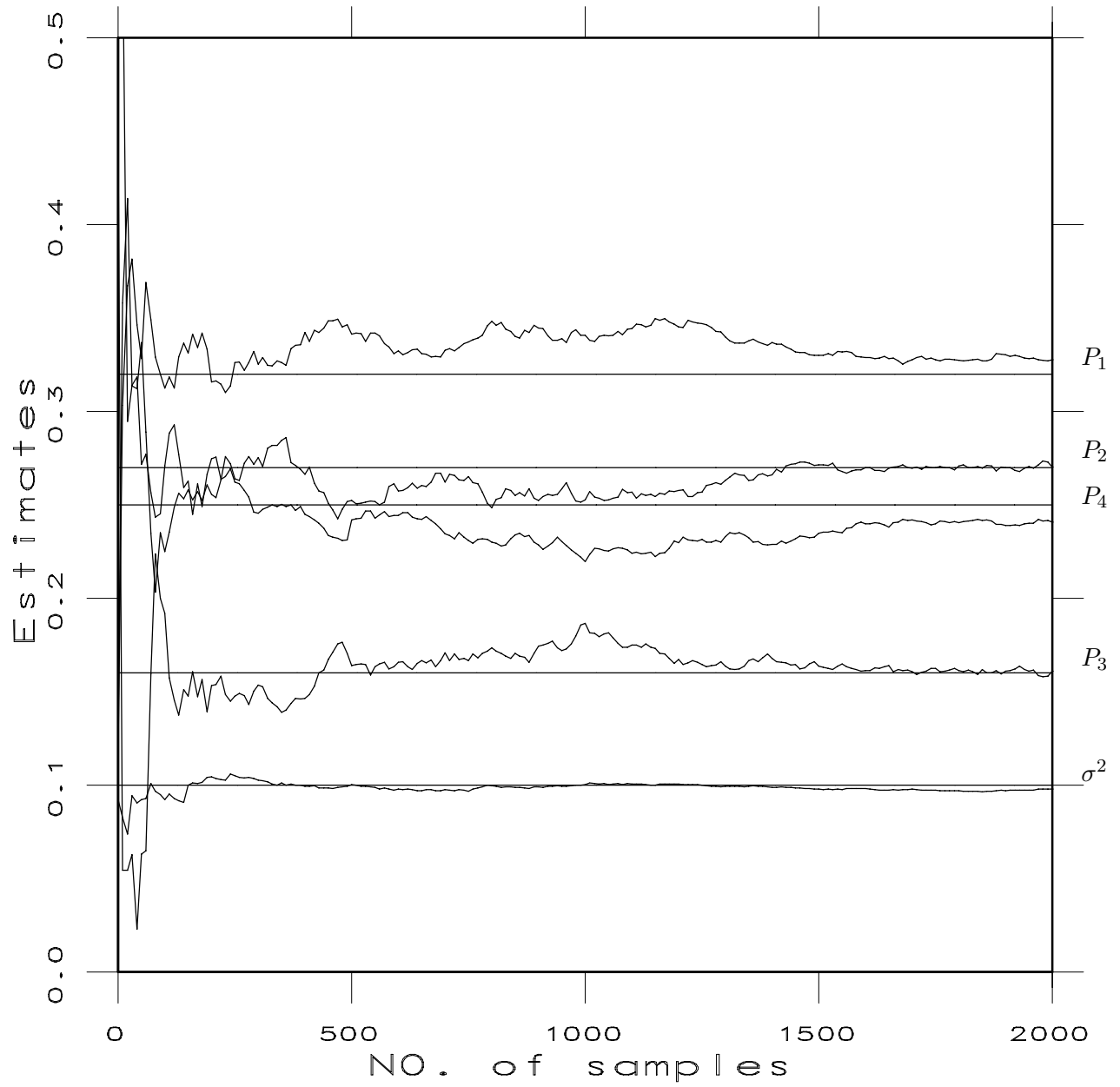


Figure 1: I. I. D. sequence of univariate mixture; $\rho^2 = 0.5$.

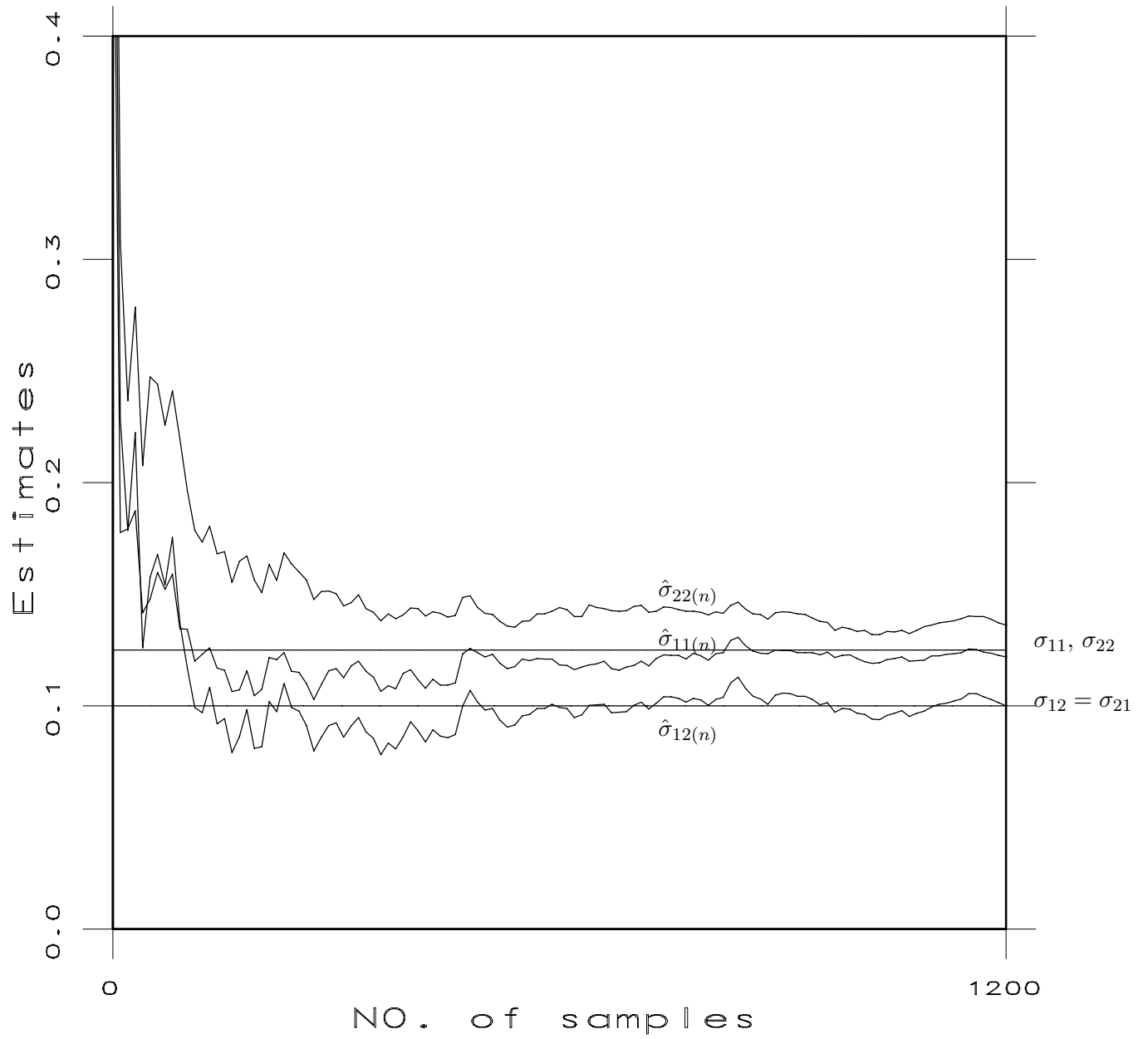


Figure 2: I. I. D. sequence of bivariate mixture; estimates of components of covariance matrix.

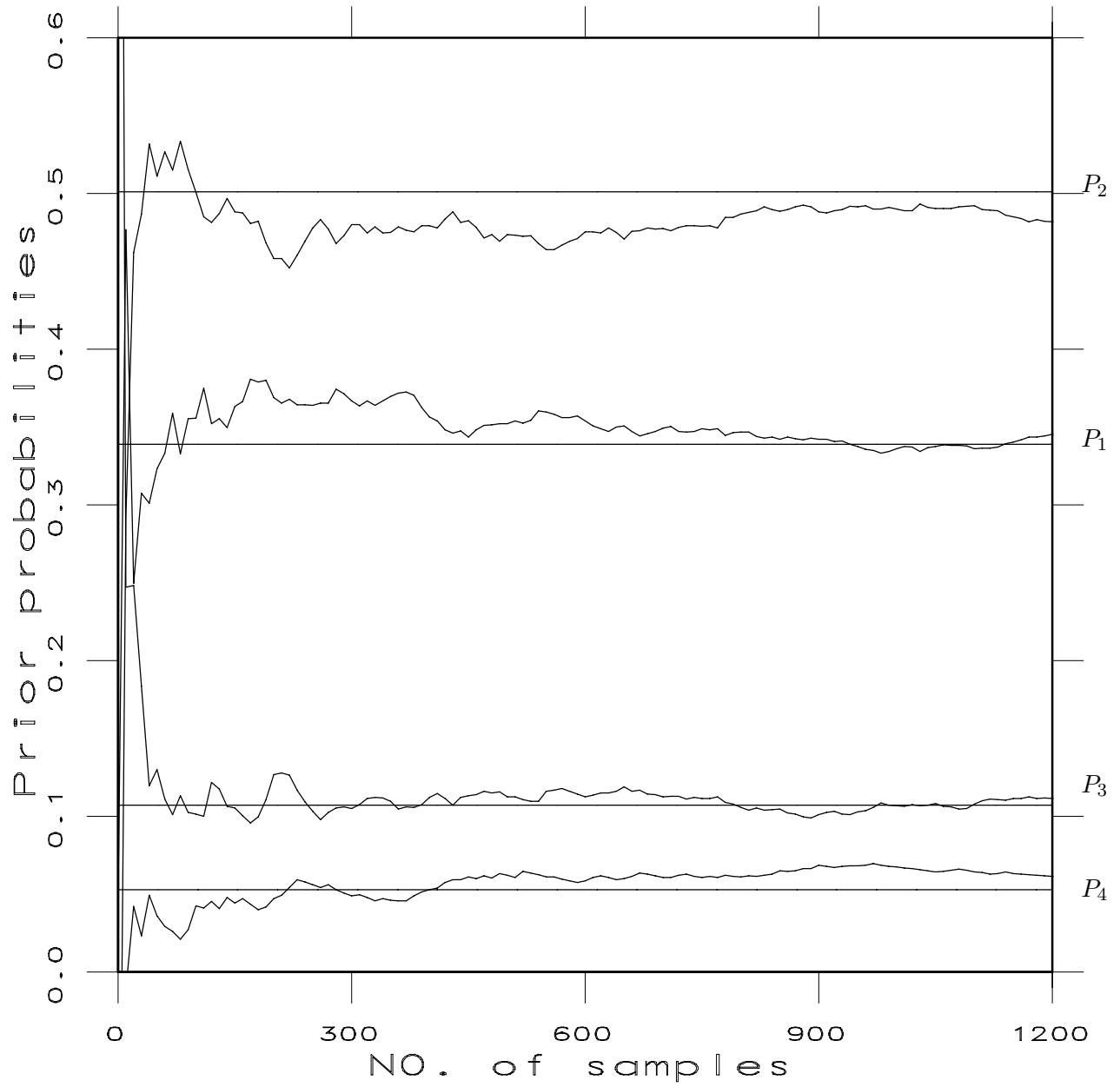


Figure 3: I. I. D. sequence of bivariate mixture; estimates of prior probabilities.

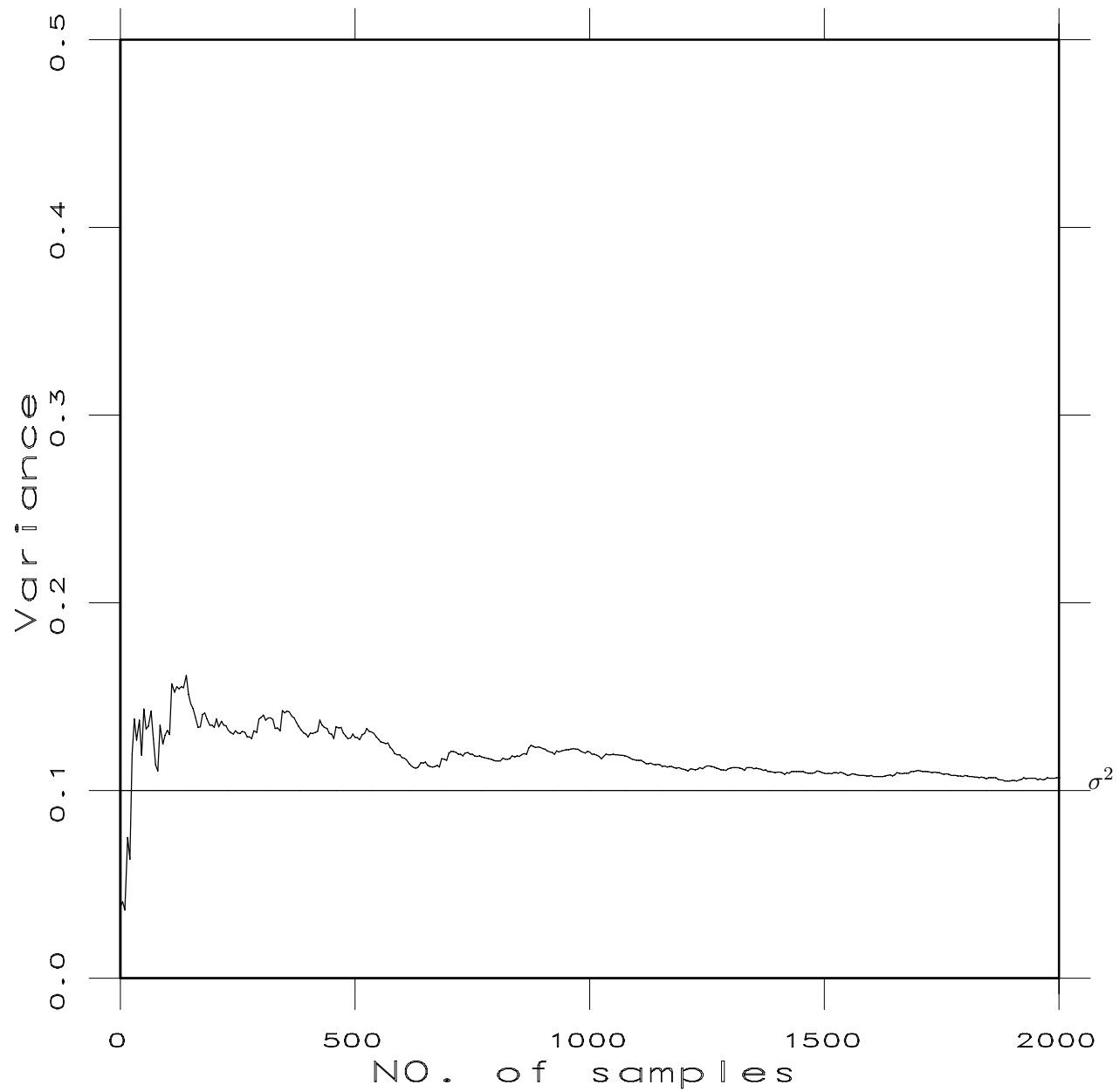


Figure 4: Estimation of noise variance; Markov sequence of signals.

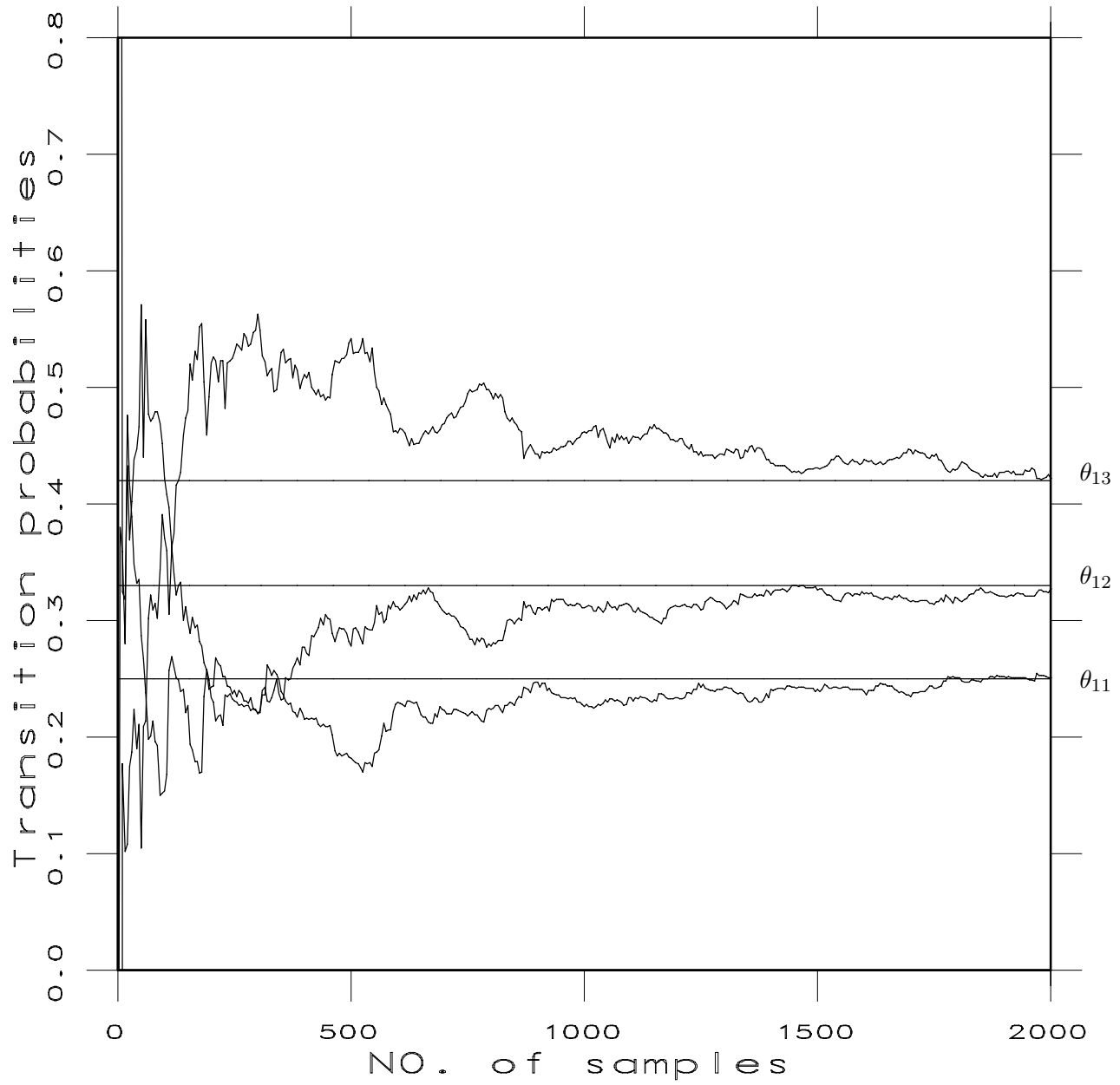


Figure 5: Estimation of transition probabilities of Markov sequence.

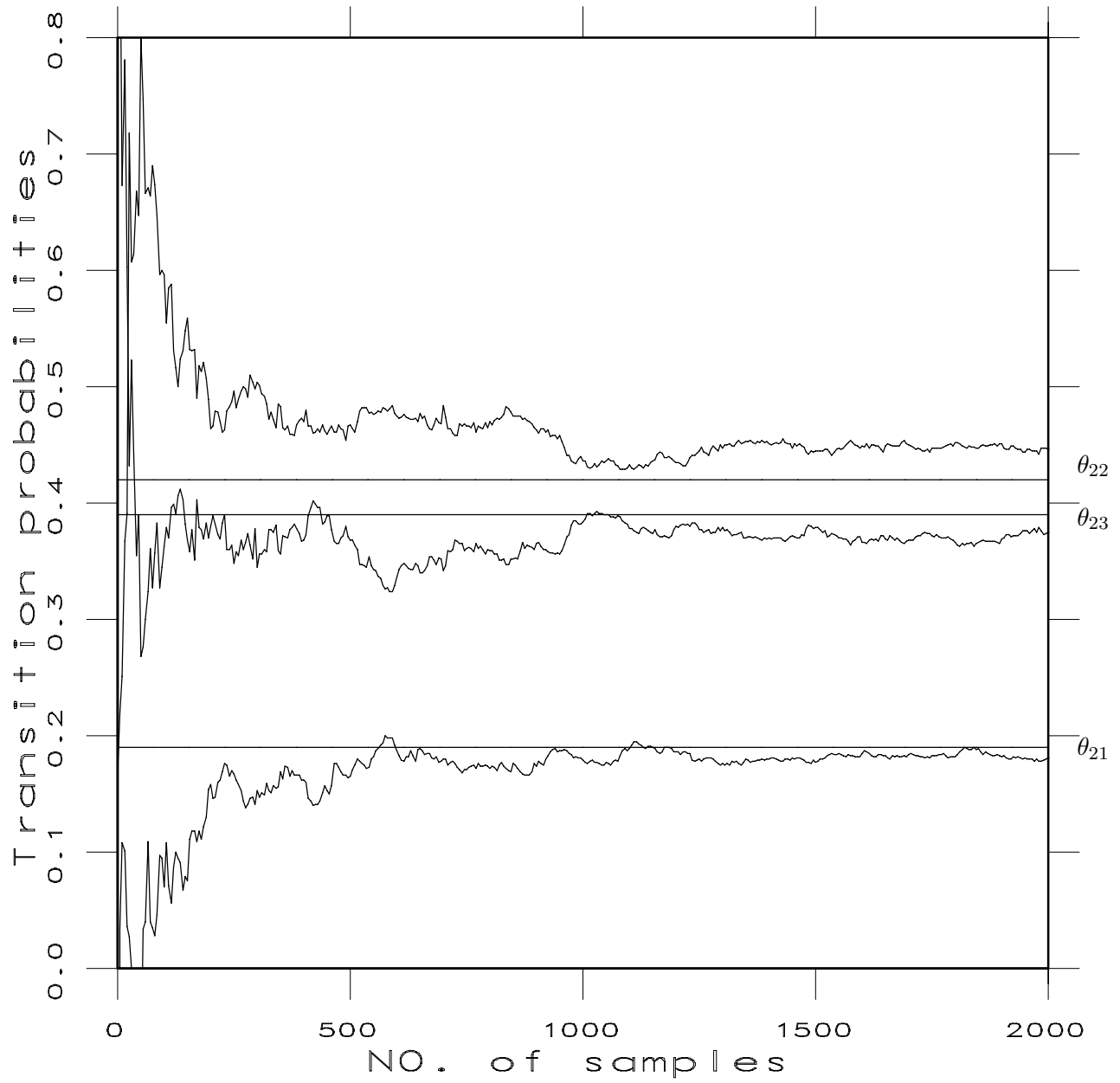


Figure 6: Estimation of transition probabilities of Markov sequence.

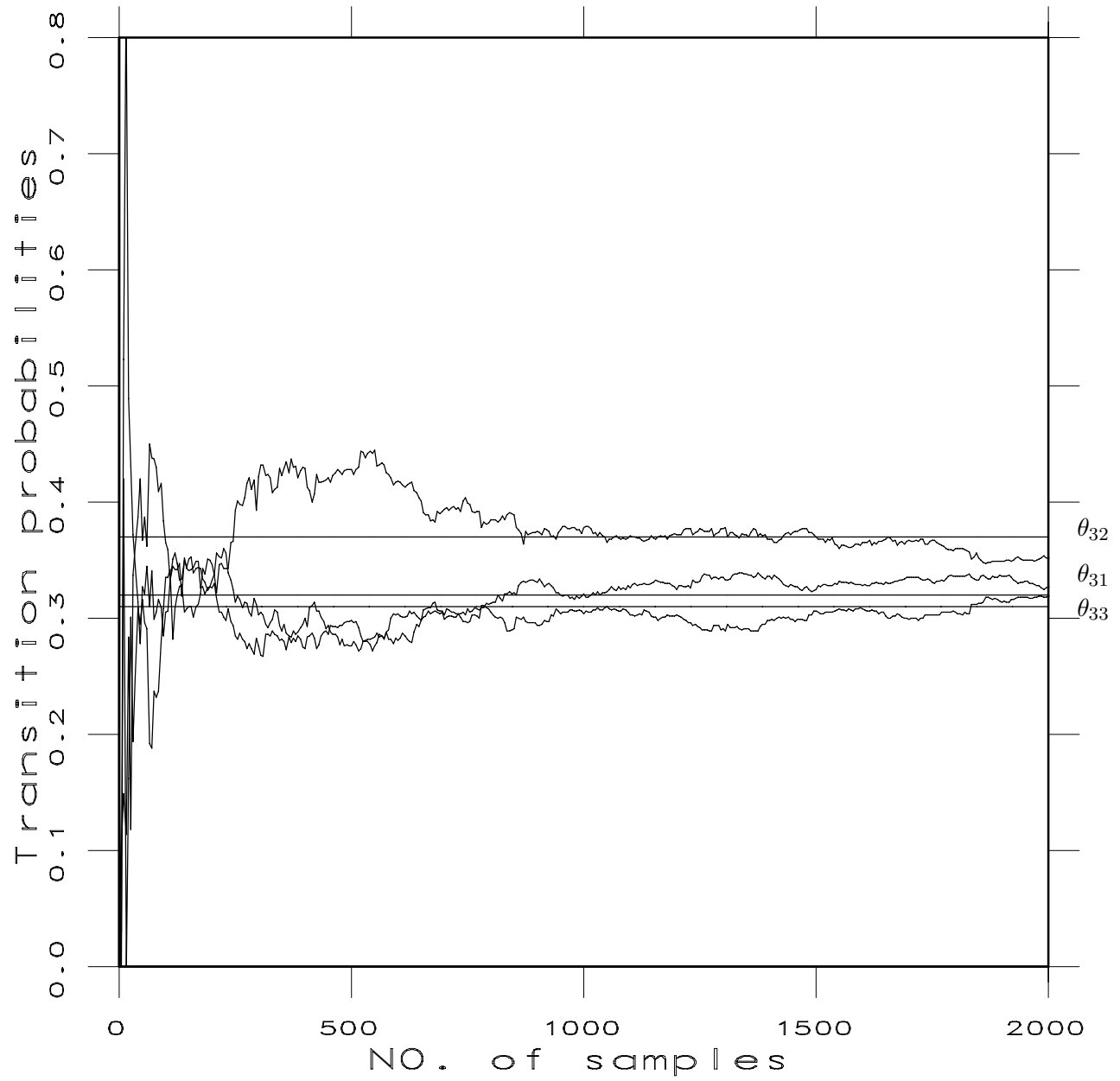


Figure 7: Estimation of transition probabilities of Markov sequence.

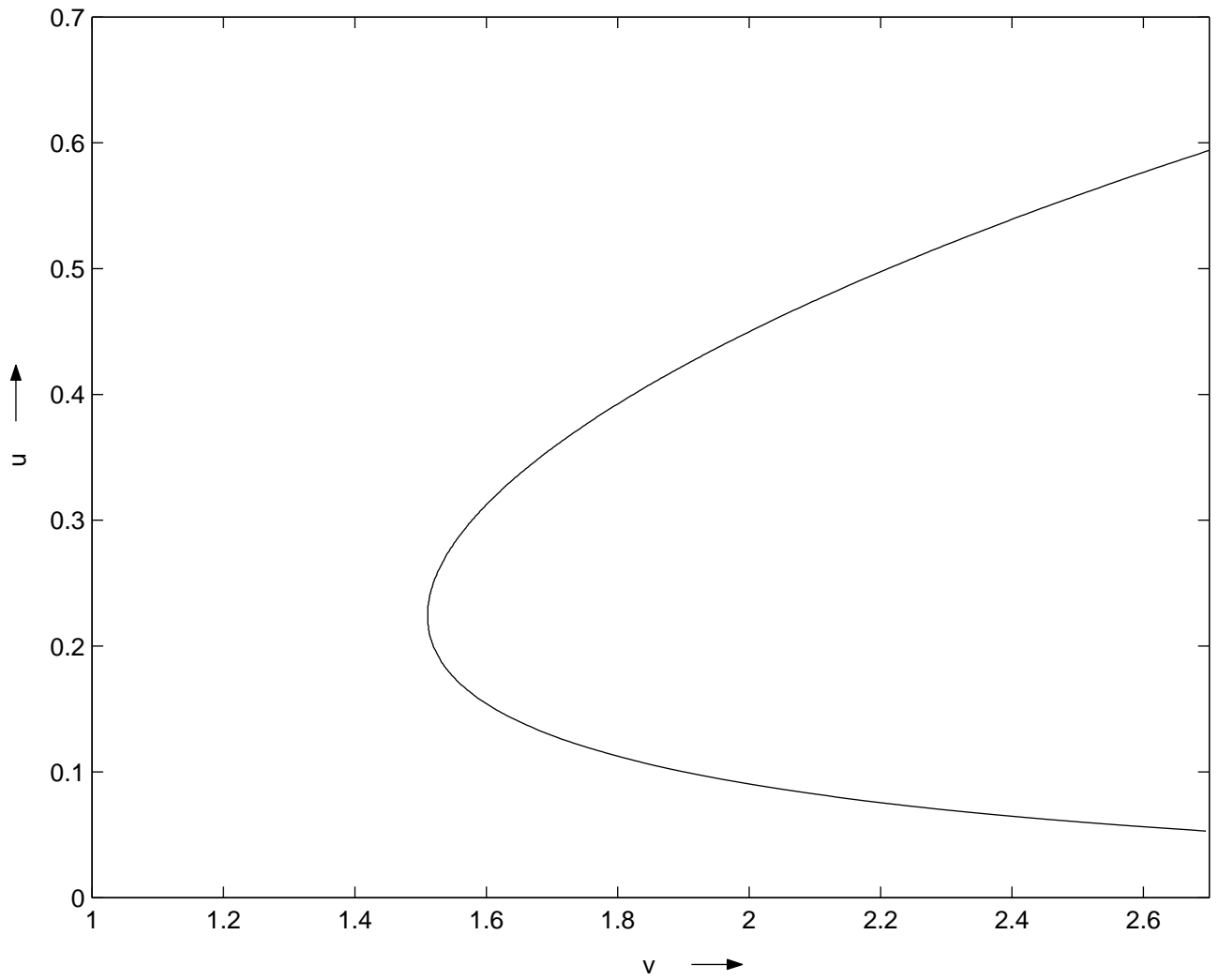


Figure 8: Plot of boundary between signs of $g(u, v)$ for mean vector $[0,1,2,3]$.

Figure Captions

1. I. I. D. sequence of univariate mixture; $\rho^2 = 0.5$.
2. I. I. D. sequence of bivariate mixture; estimates of components of covariance matrix.
3. I. I. D. sequence of bivariate mixture; estimates of prior probabilities.
4. Estimation of noise variance; Markov sequence of signals.
5. Estimation of transition probabilities of Markov sequence.
6. Estimation of transition probabilities of Markov sequence.
7. Estimation of transition probabilities of Markov sequence.
8. Plot of boundary between the signs of $g(u, v)$ for mean vector $[0,1,2,3]$.