Sequential inference in finite mixture models and time series

T.N. Sriram*
Department of Statistics
University of Georgia
Athens, GA, USA
tn@stat.uga.edu

Abstract

In this article, we will discuss two estimation scenarios, where estimation will follow the essential principles underlying sequential inference, but the sample size will be fixed in advance. More specifically, the first scenario is concerned with estimation of the unknown number of components in a finite mixture model, where our desired objective is to minimize a model selection criterion based on Hellinger distance. In the second scenario, we are concerned with estimation of number of sufficient dimensions in the context of time series, where the desired objective is to maximize Kullback-Leibler type information. Thus, in each of the two scenarios, the article takes a broader view of sequential inference, focusing on sequentially achieving the desired objectives.

Key words and Phrases: Finite mixtures; Hellinger Information Criterion; Nonlinear time series; Threshold; Time series central subspace.

* The work on finite mixtures is with Mi-Ja Woo and the work on time series is with Jin-Hong Park and Xiangrong Yin.

1 Introduction

Sequential estimation generally refers to parametric inference based on a random sample size determined using a sampling scheme. Such a scheme, defined usually via a stopping rule, collects observations either one-at-a-time or in batches until a pre-specified objective is achieved. Sequential sampling is not only inherently cost effective, but also leads to procedures that achieve a desired objective. These features make sequential inference a more attractive alternative over traditional parametric inference based on a fixed sample size.

In this article, we will discuss two estimation scenarios, where estimation will follow the essential principles underlying sequential inference, but the sample size will be fixed in advance. More specifically, the first scenario is concerned with estimation of the unknown number of components in a finite mixture model, where our desired objective is to minimize a model selection criterion based on Hellinger distance. In the second scenario, we are concerned with estimation of number of sufficient dimensions in the context of time series, where the desired objective is to maximize Kullback-Leibler information. Thus, in each of the two scenarios, the article takes a broader view of sequential inference, focusing on sequentially achieving the desired objectives.

2 Finite mixtures

Finite mixture models provide a natural way of modeling unobserved population heterogeneity, which is often encountered in data sets arising from biological, physical and social sciences. A complication in many applications is that there is not much *a priori* information about the number of mixture components, termed *mixture complexity.* Estimation of mixture complexity is a fundamental problem because correct identification of mixture complexity followed by efficient estimation of all parameters would lead to finding a mixture with fewest possible components.

Consider a parametric family of density functions $\mathcal{F}_m = \{f_{\boldsymbol{\theta}_m} : \boldsymbol{\theta}_m \in \Theta_m \subseteq R^p\}$ for each fixed integer $1 \le m < \infty$ such that $f_{\boldsymbol{\theta}_m}$ can be represented as a finite mixture of the form

$$f_{\boldsymbol{\theta}_m}(x) = \sum_{i=1}^{m} \pi_i f(x|\boldsymbol{\phi}_i), \quad x \in \mathcal{X} \subseteq \mathcal{R}, \tag{2.1}$$

where the component densities $f(x|\boldsymbol{\phi}_i) \ge 0$, $\int f(x|\boldsymbol{\phi}_i)dx = 1$, $\boldsymbol{\phi}_i \in \Phi \subseteq R^s$, the mixing proportions $\pi_i \ge 0$, $\sum\limits_{i=1}^{m} \pi_i = 1$ for $i = 1, \dots, m$ and $\boldsymbol{\theta}_m = (\pi_1, \dots, \pi_{m-1}, \boldsymbol{\phi}_1^T, \dots, \boldsymbol{\phi}_m^T)^T$. The class $\mathcal{F}_m \subseteq \mathcal{F}_{m+1}$ for all m and we denote $\mathcal{F} = \bigcup\limits_{m=1}^{\infty} \mathcal{F}_m$. Let $X_1, \dots, X_n$ be independent and identically distributed random variables with an unknown density function f_0. Define

$$m_0 = m(f_0) = \min\{m : f_0 \in \mathcal{F}_m\}. \tag{2.2}$$

If indeed f_0 is a finite mixture then $m_0 < \infty$ and it denotes the true mixture complexity; otherwise $m_0 = \infty$. Note that m_0 represents the most parsimonious mixture model representation for f_0. We now describe an estimation procedure which proceeds sequentially to estimate m_0.

Before we propose an estimator of m_0, we define an estimator of $\boldsymbol{\theta}_m$ for each fixed $m \geq 1$. To this end, define the Hellinger distance between two densities f and g by $H^2(f,g) = ||f^{1/2} - g^{1/2}||_2^2$, where $||\cdot||_2$ is the L_2 norm. Let $\hat{f}_n$ be a kernel density estimator of f_0 of the form

$$\hat{f}_n(x) = \frac{1}{nc_n}\sum_{i=1}^{n} K(\frac{x - X_i}{c_n}) \tag{2.3}$$

where K is a density on $\Omega \subseteq R$ and the bandwidth $c_n = c_n(X_1, \ldots, X_n)$ satisfy regularity conditions given in the Theorem stated below. When $m \geq 1$, the minimum Hellinger distance (MHD) estimator $\hat{\boldsymbol{\theta}}_{n,m}^{MHD}$ of $\boldsymbol{\theta}_m$ is that value for which $H(f_{\hat{\boldsymbol{\theta}}_{n,m}^{MHD}}, \hat{f}_n) = \min_{\boldsymbol{t}_m \in \Theta_m} H(f_{\boldsymbol{t}_m}, \hat{f}_n)$.

Note that it is possible to view the estimation of mixture complexity m_0 as a model selection problem by defining a criterion based on the Hellinger distance. To this end, first note that in the context of minimum Hellinger distance estimation, the statistic $H^2(f_{\hat{\boldsymbol{\theta}}_{n,m}^{MHD}}, \hat{f}_n)$ is particularly appropriate for measuring goodness-of-fit of mixture models. Motivated by the classical Akaike type of criterion for model selection, we may consider a model selection criterion of the form

$$HIC = H^2(f_{\hat{\boldsymbol{\theta}}_{n,m}^{MHD}}, \hat{f}_n) + n^{-1}b(n)\nu(m) \tag{2.4}$$

where $b(n)$ depends only on n and $\nu(m)$ is the number of parameters in the mixture model. Here, the value of m yielding the minimum HIC specifies the best model. Since $\mathcal{F}_m \subseteq \mathcal{F}_{m+1}$, we have $H^2(f_{\hat{\boldsymbol{\theta}}_{n,m}^{MHD}}, \hat{f}_n) \geq H^2(f_{\hat{\boldsymbol{\theta}}_{n,m+1}^{MHD}}, \hat{f}_n)$. Therefore, in (2.4) we penalize the goodness-of-fit statistic by a term proportional to the number of parameters in the mixture model. A simple heuristic to search for the best model from a sequence of nested models is to try successive models, starting with the smallest, and stop with model m when the HIC value for model m is lesser than that for model $(m+1)$. That is, this heuristic stops when

$$H^2(f_{\hat{\boldsymbol{\theta}}_{n,m}^{MHD}}, \hat{f}_n) + n^{-1}b(n)\nu(m) \leq H^2(f_{\hat{\boldsymbol{\theta}}_{n,m+1}^{MHD}}, \hat{f}_n) + n^{-1}b(n)\nu(m+1)$$

or, equivalently,

$$H^2(f_{\hat{\boldsymbol{\theta}}_{n,m}^{MHD}}, \hat{f}_n) - H^2(f_{\hat{\boldsymbol{\theta}}_{n,m+1}^{MHD}}, \hat{f}_n) \leq n^{-1}b(n)[\nu(m+1) - \nu(m)]. \tag{2.5}$$

Setting $\alpha_{n,m} = n^{-1}b(n)[\nu(m+1) - \nu(m)]$ in (2.5) naturally leads us to the following estimator of m_0 defined by

$$\hat{m}_n = \min\{m : H^2(f_{\hat{\boldsymbol{\theta}}_{n,m}^{MHD}}, \hat{f}_n) - H^2(f_{\hat{\boldsymbol{\theta}}_{n,m+1}^{MHD}}, \hat{f}_n) \leq \alpha_{n,m}\} \tag{2.6}$$

where $\{\alpha_{n,j}; j \geq 1\}$ are positive sequences of threshold values chosen in such a way that they converge to zero as $n \to \infty$. We define $\hat{m}_n = \infty$ if the minimum in (2.6) does not exist.

Given a data set, computation of $\hat{m}_n$ in (2.6) is clearly a sequential process. The procedure starts by assuming that the data comes from a mixture with a single component ($m = 1$) whose form is known except for the parameter values. After fitting a nonparametric density estimator $\hat{f}_n$, the MHD estimate of the $\boldsymbol{\theta}_1$ is computed, which yields the best parametric fit $f_{\hat{\boldsymbol{\theta}}_{n,1}^{MHD}}$ and a goodness-of-fit measure $H^2(f_{\hat{\boldsymbol{\theta}}_{n,1}^{MHD}}, \hat{f}_n)$. Next, another component density is added yielding a mixture of two components ($m = 2$). As in the first stage, the best parametric fit $f_{\hat{\boldsymbol{\theta}}_{n,2}^{MHD}}$ and a goodness-of-fit measure $H^2(f_{\hat{\boldsymbol{\theta}}_{n,2}^{MHD}}, \hat{f}_n)$ are computed using the MHD estimate of $\boldsymbol{\theta}_2$. The difference between the two goodness-of-fit measures is then compared with the threshold value $\alpha_{n,1}$. The above sequential procedure of adding one more component to the previous mixture is repeated until the first value $m = k$ for which the difference between goodness-of-fit measures computed at the k-th and the $(k+1)$-th stage falls below the corresponding threshold value $\alpha_{n,k}$. At this time, the sequential procedure terminates, declaring k as an estimate of the number of components in the mixture. Note that at this stage our sequential procedure automatically provides a best parametric fit determined by MHD estimates of mixture parameters in a k-component mixture. The following theorem shows that sequential procedure $\hat{m}_n$ is strongly consistent as $n \to \infty$.

Theorem. Suppose $X_1, \ldots, X_n$ are independent and identically distributed random variables with an unknown density function f_0. Suppose the bandwidth c_n in (2.3) satisfies $c_n + (nc_n)^{-1} \to 0$ a.s. as $n \to \infty$. If f_0 is a finite mixture with mixture complexity $m_0 < \infty$, then for any sequence $\alpha_{n,m} \to 0$ the estimator $\hat{m}_n$ defined in (2.6) is strongly consistent, i.e., as $n \to \infty$

$$\hat{m}_n \to m_0 \quad \text{a.s.}$$

If f_0 is *not* a finite mixture, then $\hat{m}_n \to \infty$ a.s.

For a proof of the Theorem; see Woo and Sriram (2006). In addition, Woo and Sriram (2006) also illustrate the performance of $\hat{m}_n$ via extensive simulations when the true mixture components are normal or when they are symmetric departures from postulated component normality. The latter simulations shows that the sequential procedure $\hat{m}_n$ is robust under model misspecification. Here we present only one simulation for the case when the true components are normal; for other simulations on robustness, see Woo and Sriram (2006).

The first simulation demonstrates the performance of (2.6) for the target density given by

$$f(x) = (1/2)\phi(x|(0,10)) + (1/4)\phi(x|(-0.3, 0.05)) + (1/4)\phi(x|(0.3, 0.05)), \tag{2.7}$$

where ϕ denotes the normal density with mean and variance identified inside the parentheses. We implemented our sequential procedure for a sample of size $n = 1000$ drawn from (2.7). We performed 100 Monte Carlo replications of our sequential algorithm, each yielding an

Table 1: Sequential Mixture Complexity Estimation [Mixture in (2.7) has $m_0 = 3$ components]

	Estimated number of components							
	1	2	3	4	5	6	7	8
$n = 1000$								
MHDE	0	26	*74					
MKE	0	18	*63	10	2	3	1	3
R&W	0	0	0	1	89	10		
Bootstrap	0	79	15	4	2			

estimate $\hat{m}_n$ of mixture complexity m_0. We then tallied the estimated number of components (out of 100 replications). These counts are reported in Table 1, where MHDE corresponds to our estimate $\hat{m}_n$ given by (2.6), which is compared to MKE, R&W, and Bootstrap methods proposed by James et al. (2001, see Table 1), Roeder and Wasserman (1997) and McLachlan (1987), respectively. In this case, the true mixture complexity $m_0 = 3$ and we denote only the highest percentage of correct identifications by an asterisk in Table 1. The numbers in Table 1 show that our sequential procedure detects the true mixture complexity correctly about 74% of times, which is more than the percentage of correct detections by the other procedures.

3 Time Series

Time series analysis has been an active area of research for decades. Over the years, the scientific community has witnessed development of many useful parametric and nonparametric methods for analyzing time series data. Nevertheless, there is a never-ending quest to build new and modern methodologies to analyze time series data.

Suppose $\{x_t; t \geq 1\}$ is a time series. The primary goal of time series analysis is forecasting, which requires inference about the conditional distribution of $x_t|X_{t-1}$, where $X_{t-1} = (x_{t-1}, ..., x_{t-p})^T$ for some suitable lag $p \geq 1$. Here, we will assume that such a lag value p exists and is known. Our goal is to find finitely many linear combinations, $\mathit{\Phi}_1^T X_{t-1}, \cdots, \mathit{\Phi}_q^T X_{t-1}$, with $q \leq p$ such that the conditional distribution of $x_t|X_{t-1}$ is same as the conditional distribution of $x_t|(\mathit{\Phi}_1^T X_{t-1}, \cdots, \mathit{\Phi}_q^T X_{t-1})$. This is equivalent to finding a $p \times q$ matrix $\Phi = (\mathit{\Phi}_1, ..., \mathit{\Phi}_q)$ such that

$$x_t \perp\!\!\!\perp X_{t-1}|\Phi^T X_{t-1}, \tag{3.1}$$

that is to say, x_t is independent of X_{t-1} given $\Phi^T X_{t-1}$. If such a Φ exists, then the $p \times 1$ vector X_{t-1} can be replaced by the $q \times 1$ vector $\Phi^T X_{t-1}$ (with q much smaller than p) without loss of information. This would represent a sufficient yet useful reduction in the dimension of X_{t-1}, where all the information in X_{t-1} about x_t is contained in the q-linear combinations.

We define a dimension reduction subspace for x_t on X_{t-1} as any subspace $\mathbf{S}(\Phi)$ of $\mathbf{R}^p$

for which (3.1) holds. Note that (3.1) holds trivially for $\Phi = I_p$ (Identity matrix), which implies that a dimension reduction space always exists. Since our primary aim is to reduce the dimension, we naturally seek a minimum dimension reduction space for x_t on X_{t-1}. To this end, we define the intersection of all dimension reduction spaces as a *Time Series Central Subspace* (TSCS), denoted by $\mathbf{S}_{x_t|X_{t-1}}(\Phi_d)$, if the intersection is itself a dimension reduction space, where $\dim(\mathbf{S}_{x_t|X_{t-1}}(\Phi_d)) = d$ and $\Phi_d = (\mathit{\Phi}_1, ..., \mathit{\Phi}_d)$. Clearly, TSCS is the minimum dimension reduction subspace.

Our definition of TSCS is general enough to include many well known linear and nonlinear time series. It is possible to identify the basis for TSCS corresponding to autoregressive model of order p, threshold autoregressive model of order p, and autoregressive conditionally heteroskedastic models; see Park et al.(2009) for details. In view of the large class of examples, we will henceforth assume that TSCS exists. We will focus on simultaneous estimation of minimum dimension d of $\mathbf{S}_{x_t|X_{t-1}}(\Phi_d)$ and Φ_d. From a time series data analysis point of view, a simultaneous estimation of d and Φ_d would provide minimum linear combinations of X_{t-1}, which would provide an initial phase when an adequate parsimoniously parameterized time series model is not yet available.

Suppose for a moment d and p are known. Park et al.(2009) motivated estimation of Φ_d by $\hat{\Phi}_n = \arg\max_{\mathbf{h}} \hat{\Psi}_n(\mathbf{h})$ based on a Kullback-Leibler type information, where

$$\hat{\Psi}_n(\mathbf{h}) = \frac{1}{n}\sum_{t=1}^{n} \log \frac{p_n(\mathbf{h}^T X_{t-1}, x_t)}{p_n(x_t)p_n(\mathbf{h}^T X_{t-1})},$$

p_n's are product or univariate Gaussian kernel density estimators and the maximization is over all $p \times d$ matrices $\mathbf{h}$ satisfying the constraint $\mathbf{h}^T\mathbf{h} = I_d$. They also show that, under certain regularity conditions, $\hat{\Phi}_n$ converges almost surely to Φ_d; see Park et al.(2009).

In practice, however, minimum dimension d is seldom known; hence has to be estimated from sample. Motivated by the sequential approach described in section 2, we now propose a sequential estimation procedure which would simultaneously yield an estimator of d and Φ_d, when p is known. Note that if $p = 1$, then $d = 1$, and therefore there is no need for dimension reduction. Thus, our estimation procedure starts by fixing a value of p (≥ 2) and determines

$$\hat{d} = \min\{k(\leq (p-1)) : \hat{\Psi}_n(\hat{\mathbf{h}}_{(k+1)}) - \hat{\Psi}_n(\hat{\mathbf{h}}_k) \leq \tau_{p,n}\}, \tag{3.2}$$

where $\hat{\mathbf{h}}_k = \arg\max_{\mathbf{h}_k} \hat{\Psi}_n(\mathbf{h}_k)$ and the maximization is over all $p \times k$ matrices $\mathbf{h}_k$, and $\{\tau_{p,n}; n \geq 1\}$ is a sequence of non-negative threshold values chosen in such a way that it converges to zero as $n \to \infty$.

As in section 2, for a data set, the computation of (3.2) is a sequential process. The procedure in (3.2) successively compares the difference $\hat{\mathbf{D}}_k = \hat{\Psi}_n(\hat{\mathbf{h}}_{(k+1)}) - \hat{\Psi}_n(\hat{\mathbf{h}}_k)$ (this is > 0 because of a result in Park et al., 2009) with the threshold value $\tau_{p,n}$ starting with $k = 1$, and stops at the first value of $k = l$ for which $\hat{\mathbf{D}}_l$ is at or below the threshold. This yields an estimate $\hat{d}$ of d for a given value of p, which in turn yields an estimate $\hat{\Phi}_{n,\hat{d}}$ of Φ_d. Obviously, if $\hat{\mathbf{D}}_k$ never falls below the threshold $\tau_{p,n}$, then $\hat{d} = p$. Park et al. (2009) also show that $\hat{d}$ is strongly consistent for d; see their article for details. We now present a small simulation study for a nonlinear time series model to illustrate the performance of our sequential procedure $\hat{d}$ in (3.2).

Model: Let $x_t = -1 - \cos((\pi/2)(x_{t-1})) - \cos((\pi/2)(1/\sqrt{5})(x_{t-3} + 2x_{t-6})) + 0.2\varepsilon_t$, where true $p = 6$ and $d = 2$, and $\{\varepsilon_t\}$ is a sequence of independent $N(0, 1)$ random variables. We assess the performance of our sequential estimate $\hat{d}$ in (3.2). The sample size for this study is $n = 300$. Table 2 reports f_i, the frequency of $\hat{d} = i$, based on 200 Monte Carlo replications using threshold values $\tau_{p,n} = \chi^2_p(0.05)/(2n)$ (0.05-threshold) and $\tau_{p,n} = \chi^2_p(0.01)/(2n)$ (0.01-threshold), where $\chi^2_p(\alpha)$ is the $100(1-\alpha)$ percentile of Chi-square distribution with p degrees of freedom. Here, f_{i+} denotes the frequency of $\hat{d} \geq i$. Table 2 shows that $\hat{d}$ with 0.05-threshold and 0.01-threshold correctly estimate the true dimension, $d = 2$, about 78% to 80% of the times when the lag p is 6.

Table 2: Frequency of estimated dimension for 0.05-threshold and 0.01-threshold, based on 200 Monte Carlo replications. The true dimension is $d = 2$.

n	lag p	0.05-threshold	0.01-threshold
300	6	f_1=21 f_2=156* f_{3+}=23	f_1=21 f_2=159* f_{3+}=20

References

[1] James, L. F., Priebe, C. E., and Marchette, D. J. (2001). Consistent Estimation of Mixture Complexity. *The Annals of Statistics*, **29**, 1281–1236.

[2] McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society*, Ser. C (Applied Statistics) **36**, 318–324.

[3] Park, J. H., Sriram, T. N. and Yin, X. (2009). Dimension reduction in time series. *Statistica Sinica*, To appear.

[4] Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, **92**, 894–902.

[5] Woo, Mi-Ja, and Sriram, T. N. (2006). Robust estimation of mixture complexity. *Journal of the American Statistical Association*, **101**, 1475–86.