Weighted Chinese restaurant processes and Bayesian mixture models

Albert Y. Lo[1]
Department of Information and Systems Management
The University of Science and Technology
Clear Water Bay, Hong Kong

Lawrence J. Brunner[2]
Department of Statistics
The University of Toronto
Toronto M5S 1A1, Canada

Anthony T. Chan[3]
Department of Information and Systems Management
The University of Science and Technology
Clear Water Bay, Hong Kong

Research Report
Version 1.0
December, 1996
(Typos corrected)

---

[1]Supported in part by Hong Kong Research Grants Council Competitive Earmarked Research Grant 674/95.

[2]Supported in part by Natural Sciences and Engineering Research Council of Canada Grant OGPIN 014

[3]Supported in part by (Hong Kong) RGC Direct Allocation Grant 92/93.BM13

Summary.

The weighted Chinese restaurant process is a sequential seating algorithm which generates a random partition of the data. Repeated simulation of the weighted Chinese restaurant process can be used to provide an iid Monte Carlo approximation to posterior quantities for Bayesian mixture models. Numerical examples illustrating the algorithm are given in Bayesian deconvolutions, i.e., estimation of the mixing distribution, and Bayesian density estimations. Numerical comparisons of the weighted Chinese restaurant process and a recently proposed missing data sequential imputation algorithm show that the weighted Chinese restaurant process performs consistently better. The weighted Chinese restaurant process also provides an iid Monte Carlo approximation to posterior quantities for Bayesian mixture hazard rate models. Comparison of sequential seating and sequential imputation algorithms are discussed using Blackwell's criterion of comparison of experiments.

1. Introduction.

Kuo (1986), studying Bayesian mixture models, argued that an evaluation of the mixture of Dirichlet process posteriors based on an iid simulation of the extended Polya urn sequence [Blackwell–MacQueen (1973)] would not work, and proposed an alternative iid Monte Carlo method based on the Chinese restaurant process of sampling partitions [Aldous (1985); see also Ferguson (1983)]. However, numerical results [Ji (1991)] indicate that Kuo's method fails to produce acceptable estimates if the true density under question is multi–modal, while its performance for a unimodal density model is mixed. Kuo's method has difficulties due to the fact that the Chinese restaurant process essentially samples from the prior distribution, and it captures little contribution from the likelihood part of the data.

The Chinese restaurant process is a sequential seating algorithm in which a partition of the set {1,...,n} is built up by seating the integers 1,...,n into cells of the partition sequentially, one at a time. Here we propose another sequential seating algorithm which, in the course of the simulation, accounts for both contributions from both the likelihood and the prior distribution. This is achieved by attaching to each integer a weight based on a Bayesian predictive argument. The resulting algorithm is called the weighted Chinese restaurant restaurant process (WCR). The WCR is essentially a new kind of clustering algorithm where the probability of assigning a unit to a cluster is proportional to the product of the cluster size and the Bayesian predictive density of the unit to be assigned (conditional on the cluster). Section 2 describes the WCR algorithm. Section 3 discusses its application in Bayesian mixture models. In this case the WCR produces a random partition which, up to a product factor, has a distribution explicitly represented in Theorem 2 in Lo (1984). The WCR algorithm is illustrated with numerical examples in Bayesian

deconvolution, kernel density estimation, and location models where the mixture distribution is a nuisance parameter. Effects of nested models and sequential orders are also investigated using a mixture of normal models.

While the present paper was in draft form, Liu (1996) published a paper studying mixture of binomial models from a Bayesian viewpoint. In this paper, he proposed a sequential imputation algorithm for missing data [Kong, Liu, and Wong (1994)] to evaluate posterior distributions which are mixtures of Dirichlet processes. Up to a product factor, Liu's algorithm is a simulation of a random vector (missing data), the distribution of which was explicitly represented by Theorem 1 in Lo (1984). Section 4 discusses an extension of Liu's simulation algorithm to mixtures of arbitrary kernels, called a weighted Polya process (WP). The WP algorithm dictates that the missing values are either drawn from the (sub−)posterior distribution given one observation, or that it be one of the previously drawn missing values. The posterior distribution, given one single observation, differs little from the prior distribution and, as a result, the imputed missing values are approximately observations from the prior distribution. As such, the WP is sensitive to the choice of prior parameters even for large sample sizes. While the WP improves significantly on Kuo's raw Chinese restaurant process algorithm by accounting for the peaky integrand in the course of the sequential sampling, the improvement is greatly retarded by this sensitivity. Numerical examples comparing the WCR and WP are given in connection with the Bayesian deconvolution and the Bayesian kernel density estimation. The tack data [Beckett and Diaconis (1994)] are examined using a mixture of binomial models.

Section 5 discusses the application of the WCR method in Bayesian mixture hazard rate models. Mixture hazard rates have been popular models for emission tomography; see Chapter 3 in Snyder and Miller (1991) for references. Here it is illustrated that the WCR and WP algorithms for computing posterior moments applies as well.

Section 6 discusses the comparison of the sequential imputation and sequential seating experiments using Blackwell's notion of comparing experiments [Blackwell (1951, 1953)]. An interpretation of Lemma 2 in Lo (1984) results in a conditional distribution of a sequentially imputed missing random vector given a sequential seating random partition. This in part explains the noisy feature of the WP in comparison with the WCR. Another consequence of this conditional distribution is in the application of Markov chain Monte Carlo methods [Hastings (1970)] to Bayesian mixture models, which has the merit of eliminating the normalization constant upon convergence to stationarity. The state−of−the−art Markov chain Monte Carlo method for Bayesian mixture models constructs a Markov chain based on sequentially imputing

the "missing data", and Liu (1996) provided an excellent summary and recent references for evaluating posterior distributions which are mixtures of Dirichlet processes; see also Diebolt and Robert (1994). In this approach, the state space is the collection of all possible missing values. The conditional distribution obtained reveals that the search for better Markov chain Monte Carlo method for Bayesian mixture models should be confined to the class of sequential seating Markov chains which have a finite state space being the collection of partitions of {1,...,n}.

Section 2. Sequential seating: a weighted Chinese restaurant process (WCR).

The (unweighted) Chinese restaurant process is a procedure for randomly partitioning the integers {1,...,n} into subgroups that are called tables (or cells). It takes its name from a seating process allegedly witnessed by Jim Pitman in a Bay area Chinese restaurant he frequented [see Aldous (1985)]. A Chinese restaurant process with parameter c>0 selects a random partition by sequentially assigning the integers to tables/cells as follows: Customers 1,...,n enter the restaurant in the order written and they are seated one after the other. Initially, all tables in the restaurant are folded up. When customer 1 comes in, a table is opened to seat him/her. After customers labelled 1,...,k−1 (k≥2) are seated, customer k with be seated on an empty table with probability c/[c+k−1]; otherwise, he/she sits on an occupied table with probability proportional to the number of occupants at that table. The seating process will continue until all customers are seated. (In this paper, we only consider restaurants with n or more tables.)

The Chinese restaurant process $\mathbf{p}$ with parameter c has the density

(2.1) $\quad q(\mathbf{p}|c)=[B_{n-1}(c)]^{-1}c^{n(\mathbf{p})}\times\Pi_{1\le i\le n(\mathbf{p})}(e_i-1)!,$

where $\quad B_{n-1}(c)=c(c+1)\times...\times(c+n-1),$

$\mathbf{p}=\{C_1,...,C_{n(\mathbf{p})}\}$ is a partition of {1,...,n} into $n(\mathbf{p})$ tables (i.e., disjoint subsets of {1,...,n}) and $e_1,...,e_{n(\mathbf{p})}$ are table sizes. The density (2.1) can be proved by letting g's to be the identity function in Lemma 2 in Lo (1984) and setting $\alpha(R)=c$.

The seating probabilities are parameters which define a sequential seating process. In the WCR case, these seating probabilities are defined in terms of a (prior) mixing measure $\alpha(du)$, the number of customers to be seated n, and a nonnegative and finite (likelihood) weighting function $w_j(u)$ for customer j, j=1,...n. Define the "marginal" weight for a table C by

(2.2) $\quad m(C)\equiv\int\Pi_{j\in C}w_j(u)\alpha(du).$

The marginal weights are assumed to satisfy $m(C)<\infty$. Since m(C)=0 implies that m(r,C)=0 for $r\notin C$, we define "predictive" weights by the ratio

(2.3) $\quad m(r|C)\equiv m(r,C)/m(C)$, for $r\notin C$; $\equiv 0$ if m(C)=0.

By (2.3), for any table C, the marginal weight m(C) can be written as a product of predictive weights, by adding customers one at a time, starting from an empty table. According to this definition, the order of seating customers is irrelevant. For example, suppose C has e elements, and $i_1,i_2,...,i_e$ is any ordering of them, the product rule states that

(2.4) $\quad m(C)=\Pi_{1\le j\le e}m(i_j|i_1,...,i_{j-1})$, $[m(j|i_0)\equiv m(j).]$

The essence of this argument is that the numerator of a term in the product cancels with the denominator of the next term. A similar cancellation reduces the Kaplan−Meier estimator to the

empirical distribution function in the absence of incomplete observations.

A "posterior distribution" of u given table C is defined by

$$\pi(du|C) \propto \Pi_{j\in C} w_j(u)\alpha(du).$$

The notion of posterior distribution is less significant in the development of the WCR, which is founded on a (Bayesian) predictive argument.

In the WCR, customer k is seated on an empty table with probability proportional to m(k); otherwise, he/she sits on an occupied table with probability proportional to the product $e_i m(k|C_i)$. More precisely, the WCR algorithm for seating customers 1,...,n is:

(2.5) Set $\lambda(0)=m(1)$.

Step 1: Assign 1 to the first table with probability $\lambda(0)/\lambda(0)=1$.

Step k (k=2,...,n):

Given $\mathbf{p}=\{C_1,C_2,...,C_{\#(\mathbf{p})}\}$ from step k−1, with table sizes $e_1,...,e_{\#(\mathbf{p})}$,

calculate $\lambda(k-1)=m(k)+\Sigma_{1\le i\le \#(\mathbf{p})} e_i m(k|C_i)$.

Assign k to a new table with probability $m(k)/\lambda(k-1)$;

otherwise, k sits on table C_i with probability $e_i m(k|C_i)/\lambda(k-1)$,

i=1,...,#(**p**).

The completion of Step n results in a WCR process $\mathbf{p}=\{C_1,C_2,...,C_{n(\mathbf{p})}\}$ with sizes $e_1,...,e_{n(\mathbf{p})}$, respectively.

Going through the n−step WCR algorithm with the product rule (2.4) operating at each step, we obtain the density of the WCR as

(2.6) $$q(\mathbf{p}|\alpha,n,\mathbf{w})=[\Pi_{1\le i\le n(\mathbf{p})}(e_i-1)!m(C_i)]\times(\Lambda_{n-1})^{-1},$$

where $\Lambda_{n-1}=\lambda(0)\times...\times\lambda(n-1)$.

The weighted Chinese restaurant process $q(\mathbf{p}|\alpha,n,\mathbf{w})$ reduces to the Chinese restaurant process $q(\mathbf{p}|c)$ if $w_i(u)\equiv 1$, and the measure $\alpha(.)$ is finite with total mass c. For the Chinese restaurant process, the predictive weights m(k|C) remain a constant one for all nonempty tables C. The following is a change of variables lemma connects the Chinese restaurant process and the WCR.

Lemma 1. $B_{n-1}(1)q(\mathbf{p}|1)\times\Pi_{1\le i\le n(\mathbf{p})}m(C_i)=\Lambda_{n-1}q(\mathbf{p}|\alpha,n,\mathbf{w})$.

The way incoming customers in the WCR are assigned to occupied tables deserves comments as it reveals a Bayesian way of performing cluster analysis for a set of data $\{x_i, i=1,...,n\}$ by means of predictive densities rather than the usual ones based on a distance function defined between (groups of) data [Duda and Hart (1973)]. Identify the observation x_i with customer i, i=1,...,n. The definitions of the marginals and predictive weights (2.3) state

that at the kth step, the kth customer is assigned to tabe C_i with seating probability proportional to $e_i m(k|C_i)$. Regarding C_i as a cluster of data, and the predictive weight $m(k|C_i)$ defined in (2.3) is really the value of a predictive density, conditional on C_i, evaluated at a future observation k. The predictive weight $m(k|C_i)$ is large if k (i.e., x_k) is close to (the x_j in) C_i; otherwise $m(k|C_i)$ is small. Hence if k is "close to" C_i, the seating probability that it will be grouped into C_i is also large. So (on average) eventually a partition $\mathbf{p}$ from a WCR is formed by tables C_i i=1,...,n($\mathbf{p}$), which yields clusters of the observed data. This phenomenon is made clear in the context of Bayesian mixture density and hazard rate modelss discussed in the next section and Section 5.

Section 3. Bayesian mixture model.

A mixture model is a family of densities defined by

(3.1) $$f(x|G) = \int k(x|u)G(du),\ G\in\Theta,$$

where the parameter space Θ is the collection of distributions; x and u are points in Euclidean spaces and G is a distribution of u. The kernel $k(.|.)$ is given, and for each u, $k(.|u)$ is a density of x. The kernel densities $\{k(.|u), \text{all } u\}$ are the extreme points of the model (3.1). The model densities have desirable smoothness properties which sometimes can be characterized via extreme point representations. On the other hand, the mixture model often arises as a result of missing information in the sense that a complete observation (x,u) is not available. Instead, one observes the variable x which is a randomization of u. Let $x_1,...,x_n|G$ be i.i.d. observations from the mixture density $f(x|G)$. The problem is to estimate G based on the sample $x_1,...,x_n$. Assuming a Dirichlet process prior on G [Ferguson (1973)] with shape measure $\alpha(.)$, Lo (1984) obtains the posterior distribution of G explicitly as a mixture of Dirichlet processes [Antoniak (1974)], and represents the posterior mean of $f(t|G)$ as an average over partitions of the set $\{1,...,n\}$ [see the following expression (3.2)]. The number of partitions of the set $\{1,...,n\}$ is called Bell's number, which increases roughly as the factorial of n. As a result, the exact evaluation of the posterior mean is formidable for sample sizes larger than twelve. This section uses a weighted Chinese restaurant process of simulating random partitions which may be averaged to yield an approximation of the stated sum over partitions. To describe it, one needs the notation of a sub–Bayesian system:

Suppose $\alpha(.)$ is a finite mixing measure with total mass $\alpha(R)$. u has a (sub–)prior distribution $\pi(du)=\alpha(du)/\alpha(R)$, and $y_1,...,y_n|u$ are iid $k(.|u)$. Given a table C, the marginal density of the y_j, $j\in C$ are then $m(y_j,j\in C)/\alpha(R)$. Denote the predictive density of y_r given $\{y_j,j\in C\}$ by $m(y_r|C)\equiv m(y_r,y_j,j\in C)/m(y_j,j\in C)$, where $y_r\notin C$. The (sub–) posterior distribution of $u|y_j,j\in C$ is denoted by $\pi(du|C)$.

The posterior mean of the mixture density, $\hat{f}(t)=E[f(t|G)|\mathbf{x}]$, has the representation [Theorem 2 in Lo (1984)]

(3.2) $$\hat{f}(t) \propto \Sigma_{\mathbf{p}}\{m(t)+\Sigma_{1\le i\le n(\mathbf{p})}e_i m(t|C_i)\}\times q(\mathbf{p}|1)\Pi_{1\le i\le n(\mathbf{p})}m(x_j,j\in C_i)$$

Here $m(t)=\int k(t|u)\alpha(du)$, i.e., $\alpha(R)$ times the predictive density based on no observation. It follows that the predictive density $\hat{f}(t)$ is a two–layer mixture of sub–predictive density $m(t|C_i)$'s which are kernel functions with variable bandwidths. This contrasts significantly with the classical kernel estimator [Rosenblatt (1956), Parzen (1962) and Cencov (1962)] which is a one–layer mixture of kernels with a fixed bandwidth.

Kuo (1986) proposed an iid Monte Carlo method to evaluate (3.2) based on sampling a Chinese restaurant process $\mathbf{p}$ from $q(\mathbf{p}|\alpha(R))= q(\mathbf{p}|1)\alpha(R)^{-n(\mathbf{P})}/B_{n-1}(\alpha(R))$. One feature of the Chinese restaurant process is that large tables have a higher probability of receiving newcomers and, as a result, it will grow larger still. According to Korwar and Hollander (1973), the number of occupied tables in a Chinese restaurant process is approximately $\alpha(R)\times\log(n)$. The presence of only a very few occupied tables in a random partition results in peaky integrands, the product of which is highly variable. This variability in effect drastically reduces the efficiency of Kuo's method based on sampling a Chinese restaurant process $q(\mathbf{p}|\alpha(R))$. On the other hand, the weighted Chinese restaurant process accounts for the peaky integrands in the course of simulation and the problem of highly variable peaky integrands diminishes. An inspection of the summand of (3.2) suggests that $k(x_i|u)=w_i(u)$, and identifying x_i as i, i=1,...,n, result in

$$\hat{f}(t) \propto \Sigma_{\mathbf{p}}\{m(t)+\Sigma_{1\le i\le n(\mathbf{p})}e_i m(t|C_i)\}\times\Lambda_{n-1}q(\mathbf{p}|\alpha,n,\mathbf{k})$$

which can be written as, since $\int m(t)dt=\alpha(R)$,

$$\hat{f}(t)=[\alpha(R)+n]^{-1}\times\Sigma_{\mathbf{p}}\Lambda_n q(\mathbf{p}|\alpha,n,\mathbf{k})/\Sigma_{\mathbf{p}}\Lambda_{n-1}q(\mathbf{p}|\alpha,n,\mathbf{k}) \tag{3.3}$$

where $\lambda(n)=m(t)+\Sigma_{1\le i\le n(\mathbf{p})}e_i m(t|C_i)$ is defined at the completion of the nth step of the algorithm, $\mathbf{p}$ is a partition of {1,....,n}. Run the WCR process M times independently to get M iid partitions and compute $\Lambda_{n-1}(m)$, $\Lambda_n(m)$, m=1,...,M. (Set $x_{n+1}=t$.) The WCR approximation to $\hat{f}(t)$ is

$$\hat{f}_M(t)=[\alpha(R)+n]^{-1}\times\Sigma_{1\le m\le M}\Lambda_n(m)/\Sigma_{1\le m\le M}\Lambda_{n-1}(m). \tag{3.4}$$

A WCR approximation to a higher posterior moment is essentially an extension of the algorithm to more steps. A higher posterior cross moment is a similar sum over partitions of the set {1,....,n,n+1,...,n+k−1} where k is the total order of the cross moments. As such, it can be written as an expectation with repect to $q(\mathbf{p}'|\alpha,n+k-1,\mathbf{k}')$ where $\mathbf{p}'$ is a partition of {1,...,n+k−1} and $\mathbf{k}'$ has n+k−1 components. A WCR, extended to n+k−1 steps, provides an appropriate approximation. For example, the posterior (j,k)th−cross moment is, j+k≥1,

$$\begin{aligned} &E[\{f(s|G)\}^j\times\{f(t|G)\}^k|x_1,...,x_n] \\ &=\{[\alpha(R)+n]\times...\times[\alpha(R)+n-1+j+k]\}^{-1} \\ &\qquad\times\textstyle\int\Lambda_{n-1+j+k}q(\mathbf{p}'|\alpha,n+j+k-1,\mathbf{k})/\int\Lambda_{n-1}q(\mathbf{p}|\alpha,n,\mathbf{k}) \end{aligned} \tag{3.5}$$

where $\mathbf{p}'$ is a partition of {1,...,n+j+k−1}, $w_i(u)=k(s|u)$ for i=n+1,...,n+j, $w_i(u)=k(t|u)$ for i=n+j+1,...,n+s+k−1, and $\lambda_{n-1+j+k}=m(t)+\Sigma_i e_i m(t|C_i)$; the index i runs over indices of currently available tables. Run the (n+j+k−1)−step WCR algorithm M times independently to get $\Lambda_{n-1}(m)$, and $\Lambda_{n-1+j+k}(m)$, m=1,...,M. The M−average

(3.6) $\{[\alpha(R)+n]\times\ldots\times[\alpha(R)+n-1+j+k]\}^{-1}\times\Sigma_{1\le m\le M}\Lambda_{n-1+j+k}(m)/\Sigma_{1\le m\le M}\Lambda_{n-1}(m)$

is a WCR approximation to (3.5). Joint posterior cross–product moments of $f(t|G)$, $t=t_1,\ldots,t_k$ can be handled analogously.

Numerical computations showed that the WCR algorithm performs well for mixture models that generate unimodal mixture densities. A list includes the scale–mixture of exponentials [Jewell (1982)], of uniforms [Brunner and Lo (1989)], and of normals. Finite mixture models [Everitt and Hand (1981); Titterington and Smith (1985); Diebolt and Robert (1994)] form a finite dimensional subset of the model considered in this paper. However, the posterior quantities derived by assuming a finite mixture model are also sums over partitions and are not any simpler. A WCR algorithm in this case can be defined by a properly chosen discrete $\alpha(.)$. A more interesting testing ground is provided by mixture models with multimodal mixture densities, and this will form the main body of the following numerical examples. Conjugate priors [see for example De Groot (1970)] exist for the sub–systems in these examples and we shall use them. To obtain a posterior density with respect to an arbitrary prior density, the posterior density with respect to conjugate priors is weighted with a weight being the ratio of the two prior densities.

Unless otherwise specified, the Monte Carlo sample size M is 1000 and $\alpha(R)$ is 2.5 in all figures.

Example 3.1 Bayesian deconvolution. Location or scale mixture models, i.e., $k(x|u)=k(x-u)$ or $k(x|u)=uk(xu)$, respectively, are in general identifiable in G and it is useful to calculate a point estimate of G. The posterior mean of $G(u)$, $\hat{G}(u)=E[G(u)|x_1,\ldots,x_n]$, has the same expression as (3.3) with the same $\lambda(1),\ldots,\lambda(n)$ and

(3.7) $\lambda(n)=\alpha(u)+\Sigma_{1\le i\le n(\mathbf{p})}e_i\pi(u|C_i)$.

That is,

(3.8) $\hat{G}(u) \propto \Sigma_{\mathbf{p}}[\alpha(u)+\Sigma_{1\le i\le n(\mathbf{p})}e_i\pi(u|C_i)]\Lambda_{n-1}q(\mathbf{p}|\alpha,n,\mathbf{k})$.

$\hat{G}(du)$ is $\alpha(du)/\alpha(R)$ with probability $\alpha(R)[\alpha(R)+n]^{-1}$; otherwise it is proportional to $e_i\pi(du|C_i)$]. Notice that $\pi(du|C_i)$ peaks for a large table C_i. Since $\hat{G}(u)$ is basically an average of $\lambda(n)=\alpha(u)+\Sigma_{1\le i\le n(\mathbf{p})}e_i\pi(u|C_i)$, it is approximately a mixture of step functions. In comparison, the maximum likelihood estimator of $G(u)$ is exactly discrete [Lindsay (1983)]. We illustrate the Bayesian deconvolution by

(i) location mixture of N(u,1) s. We use data from a three–peak location mixture of normals: 0.125N(–5,1)+ 0.375N(0,1) + 0.5N(5,1). The sub–prior $\alpha(.)/\alpha(R)$ is N(m,1/t). For table C_i, the sub–sample posterior $\pi(du|C_i)$ is $N(m_i,1/t_i)$, where $t_i=t+e_i$, $m_i= t_i^{-1}(tm+ e_i\bar{x}_i)$; $\bar{x}_i=e_i^{-1}\Sigma_{j\in C(i)}x_j$. The sub–sample predictive density $m(x|C_i)$ is $N(m_i,1+1/r_i)$. Figure 3.1a plots

$\hat{G}_M(u)$ for a nested sample of n=10, 50, 150, and 300.

(ii) location mixture of U(u−1/2,u+1/2) s. We use data from a three−peak location mixture of uniforms: 0.2U(−0.9,0.1)+0.4U(0,1) + 0.4U(0.9,1.9). Note that $\hat{G}(u)$ is differentiable if the shape $\alpha(.)/\alpha(R)$ is differentiable. Let $\alpha(.)/\alpha(R)$ be N(0,1). Figure 3.1b plots $\hat{g}_M(u)=(d/du)\hat{G}_M(u)$ for a nested sample of n=10, 50, 150, and 300; for large n (=300), $\hat{g}_M(u)$ is a mixture of δ−functions approaching a mixture of three δ−functions with mass 0.2, 0.4, and 0.4 at −0.5, 0.5, and 1.4, respectively. The deconvolution effect is clear. However, the deconvolution effect reduces if we move the two uniforms at both ends closer. Figure 2.1c plots three different $\hat{f}_M(t)$ for n=10, 50, and 150. The convergence of $\hat{f}_M(t)$ to the (dotted) mixture density f(x|G) is clear.

Example 3.2. Location and scale mixture models and a local band for a density. Since a density lies in the L^1−closure of the location and scale mixtures, it would be reasonable to use the location and scale mixture model to estimate an arbitrary density [Lo (1984)]. Here, k(x|u) is $\tau k(\tau(x-s)$ where $u=(\tau,s)$ is two dimensional. The idea is that by allowing G to carry mass at τ close to zero, the consistent behavior of a shrinking kernel in the frequentist kernel density estimator can be captured. Pick a standard normal k(.). Suppose $\alpha(dv)/\alpha(R)$ is a gamma−normal distribution defined by: τ is gamma (a,1/b) and μ|τ is normal (m,1/(τt)). [The mean of gamma (a,1/b) is a/b.] In this case, we say that (τ,μ) is gamma−normal (a,1/b; m,1/t). The sub−posteriors are gamma−normal (a_i,1/b_i; m_i,1/t_i) where $t_i=t+e_i$, $m_i=(mt+\bar{x}_i e_i)/t_i$, $a_i=a+e_i/2$, and $b_i=b+2^{-1}[\Sigma_{j\in C(i)}(x_j-\bar{x}_i)^2+(m-\bar{x}_i)^2/(t^{-1}+e_i^{-1})]$. The sub−sample predictive density $m(x|C_i)$ is a t−density with degrees of freedom $2a_i$, location m_i, and pecision $(a_i/b_i)t_i/(t_i+1)$.

The interplay between the prior parameters a, b, m and t is subtle. Inspection of the sub−predictive t densities suggests that t can not be very large, or else $m_i \approx m$ and $\hat{f}(t)$ will be approximately unimodal with mode close to m; on the other hand, a small t results in a predictive t−density centered at $\bar{x}_i$ and reveals the data structure better. We shall confine our choice to small t. To keep things simple, we assume a and |m| are moderate, say, bounded by 2. If b is large, the precision $(a_i/b_i)t_i/(t_i+1)$ for the predictive t−densities will be small, resulting in a flat $\hat{f}(t)$; so a large b is not recommended. The analysis can then be confined to small t and moderate size b. It turns out that for sample sizes less than 400, small t (t<0.1) and moderate b ($b\in[0.5, 3]$) are desirable; see Example 4.3 in the next section for detailed analysis of this issue. Figure 3.2a displays WCR approximations to the local band $E[f(t|G)|\mathbf{x}]\pm 2\sigma[f(t|G)|\mathbf{x}]$, for a grid of t and for nested data n=10, 50, 150, 300 from a three−peak location−scale mixture of

normals (a=0.5, b=0.5, m=1, and t=0.1; M=2000). The evaluation of $\sigma[f(t|G)|\mathbf{x}]$ requires a WCR approximation to the second posterior moment $E[f(t|G)^2|\mathbf{x}]$ given by (2.9) with j=0 and k=2. Figure 3.2b displays WCR approximations to the local band $E[f(t|G)|\mathbf{x}]\pm 2\sigma[f(t|G)|\mathbf{x}]$ for nested data n=10, 50, 150, and 300 from a log–normal density (a=1.5, b=1.5, m=0, and t=0.05; M=2000).

Example 3.3. Effect of nested mixture models. Here we investigate the effect of nested models. The two models we use for comparison are the location and scale mixture of normals (dashed), and its strict subset, the location mixture of normals (solid). Data are from a location mixture of normal density with four peaks (dotted; the small peak at 3 is barely detectable): $(1/8)N(-2.5,s^2)+(2/8)N(0,s^2)+(1/8)N(3,s^2)+(4/8)N(5,s^2)$ where s=0.7. Figure 3.3a displays $\hat{f}_M(t)$ for nested data. The location mixture model WCR reveals the peak more clearly. For the same large data set of size 400, we ran several WCR. It turns out that the $\hat{f}_M(t)$ based on the location model is quite stable and consistently reveal the small peak (1st column in Figure 3.3b). The performance of the $\hat{f}_M(t)$ based on the location–scale mixture model is less stable (2 nd column in Figure 3.3c), and it fails to reveal the small peak in 20% of the trials.

Example 3.4. WCR and the sequential order of the data. Technically, the WCR depends on the order in which the customers (data) arrive. Figure 3.4 displays $\hat{f}_M(t)$ (M=2000) based on sorted and unsorted data. We use data from a five–peak location mixture of normals. The $\hat{f}_M(t)$ based on the order statistics (sorted data) does not appear to be more accurate for samples of sizes larger than 150. The fact that $\hat{f}_M(t)$ perform better for positive t appears to be due to the prior assignment that m=1. More numerical studies based on the location–scale mixture of normal model and the three–peak data in Example 3.2 show that sorting the data in descending and ascending order, and Siegel–Tukey ranking the data, produce almost identical $\hat{f}_M(t)$ for nested sample sizes n=10, 50, 150, and 300. These numerical results suggest that the dependence of WCR $\hat{f}_M(.)$ on the ordering of the data is minor.

Example 3.5 A location problem. Here $k(x|u,\theta)=uk(u(x-\theta))$ where the parameter (θ,G) has a joint distribution: $\theta\sim p(d\theta)$ and $G|\theta\sim D(dG|\alpha_\theta)$, G is supported by $[0,\infty)$. Given the parameter (θ, G), the model density is $f(x|\theta,G)=\int uk(u(x-\theta))G(du)$ where θ is a location parameter and G is an unknown probability on the half line. The extreme point method states that $\{f(x|\theta,G)=\int uk(u(x-\theta))G(du)\}$ consists of all unimodal and symmetric density with mode θ. For the location problem, G is a nuisance parameter. Average out G to get the posterior distribution of θ given the data as $p(d\theta|\mathbf{x})$:

$$p(d\theta|\mathbf{x}) \propto \Sigma_{\mathbf{p}} q(\mathbf{p}|\alpha_\theta(R))\ \Pi_{1\le i\le n(\mathbf{p})}[\int \Pi_{j\in C_i} uk(u(x_j-\theta))\alpha_\theta(du)/\alpha_\theta(R)]\pi(d\theta).$$

Assume that α_θ is independent of θ (i.e., the shape G and the location θ are independent), and

that $p'(\theta)\equiv 1$ (i.e., a "flat" prior), the posterior density of θ is

(3.9) $$p(\theta|\mathbf{x}) \propto \Sigma_{\mathbf{p}} q(\mathbf{p}|\alpha(R))\ \Pi_{1\le i\le n(\mathbf{p})}[\int \Pi_{j\in C_i} uk(u(x_j-\theta))\alpha(du)/\alpha(R)$$
$$\propto \Sigma_{\mathbf{p}} \Lambda_{n-1}(\theta) q(\mathbf{p}|\alpha,n,\mathbf{w}_\theta)$$

where $q(\mathbf{p}|\alpha,n,\mathbf{w}_\theta)$ is the WCR density defined in (2.6) with $w_{\theta i}(u)=uk(u(x_i-\theta))$. [The same modification applies in the definition of $\lambda(j)$.] Λ_{n-1} depends on θ and is denoted by $\Lambda_{n-1}(\theta)$. According to the WCR, $p(\theta|\mathbf{x})$ is approximated by the density $\hat{p}_M(\theta)$ where

(3.10) $$\hat{p}_M(\theta) \propto \Sigma_{1\le m\le M} \Lambda_{n-1}(\theta;m);$$

$\Lambda_{n-1}(\theta;m)$, m=1,...,M are $\Lambda_{n-1}(\theta)$ obtained from running the WCR algorithm M times.

The location problem was discussed by Brunner and Lo (1989, 1994) where k(.) is essentially a $U(-1/2,1/2)$ density and $u>0$. Here we discuss a smooth subset of this model by assuming that k(.) is a N(0,1) density. Suppose then $\alpha(du)/\alpha(R)$ is gamma $(a,1/b)$. Then, the sub−sample posteriors are also gamma: $\pi(du|C_i)$ is gamma $(a_i,1/b_i)$, where $a_i=a+e_i/2$; $b_i=b+2^{-1}\Sigma_{j\in C(i)}x_j^2$. The sub−sample predictive density $m(x|C_i)$ is a t−density with degrees of freedom $2a_i$, location 0, precision (a_i/b_i). Figure 3.5 displays $\hat{p}_M(\theta)$ (M=2000) for a grip of θ, for nested samples from standard normal and Cauchy densities.

Remark 3.1. In another direction, the model (3.1) can be extended to the case that the kernel $k(.|u)$ depends on the i th observation through some extraneous regression variable z_i which may contain other parameters θ of interest (Example 3.5 is a case in point). But this extension to (3.1) does not present additional technical difficulties as was noted by Lo (1978): The likelihood function is $\Pi_i\int k_i(x_i|u)G(du)$, and one simply uses $k_i(x_i|u)$ to play the role of $k(x_i|u)$, resulting in a WCR algorithm [i.e., $w_i(u)=k_i(x_i|u)$], extended to cover this regression sampling plans. Regression problems using a Bayesian mixture model approach were also considered by Bunke (1985). Brunner (1995) assumed a uniform k(.) and discussed linear regression.

Section 4. Sequential imputation: a weighted Polya process (WP).

An extended Polya urn sequence $U_1,...,U_n$ with parameter a finite measure $\alpha(.)$ is defined as follows [Blackwell and MacQueen (1973)]: U_1 has distribution $\alpha(.)/\alpha(R)$. Given $U_1=u_1$, U_2 equals u_1 with probability $1/[\alpha(R)+1]$; otherwise, U_2 has distribution $\alpha(.)/\alpha(R)$. Given $U_1=u_1,...,U_{n-1}=u_{n-1}$, $U_n=u_j$ with probability $1/[\alpha(R)+n-1]$ for $j=1,...,n-1$; otherwise, U_n has distribution $\alpha(.)/\alpha(R)$. Assume the mixture model (3.1) and G has a Dirichlet process prior $D\ (dG|\alpha)$ [Ferguson (1973)] with shape measure $\alpha(.)$, the posterior distribution of G is an average with respect to the distribution of an extended Polya sequence as follows [Lo (1984)]:

(4.1) $$\pi(dG|x_1,...,x_n) \propto \int D\ (dG|\alpha+\Sigma_i\delta_{u_i})\Pi_i k(x_i|u_i)\mu(d\mathbf{u}|\alpha),$$

where the measure $\mu(d\mathbf{u}|\alpha)$ is defined by

(4.2) $$\mu(d\mathbf{u}|\alpha)=\Pi_{1\le i\le n}(\alpha+\Sigma_{1\le j\le i-1}\delta_{u_j})(du_i).$$

The normalized $\mu(d\mathbf{u}|\alpha)$, $\mu(d\mathbf{u}|\alpha)/B_{n-1}(\alpha(R))$, is the distribution of the extended Polya sequence.

One can define a sequential sampling scheme based on sampling the random variable U_i in the posterior representation (4.1) weighted by $\Pi_i k(x_i|u_i)$. This is the approach taken in Liu (1996) for mixtures of binomial kernels. The extension, called the weighted Polya process (WP), is described as follows. Given a finite shape measure α, a kernel $k(.|.)$ and data $x_1,...,x_n$: Set $\kappa(0)=m(x_1)$, and using the notation for the Bayesian sub-system in the last section.

Step 1: U_1 has distribution $\pi(du|x_1)$.

Step k (k=2,...,n): Given $u_1,u_2,...,u_{k-1}$,

calculate $\kappa(k-1)=m(x_k)+k(x_k|u_1)+...+k(x_k|u_{k-1})$.

U_k equals u_j with probability $k(x_k|u_j)/\kappa(k-1)$, $j=1,...k-1$;

otherwise, U_k has distribution $\pi(du|x_k)$.

The joint distribution of $U_1,...,U_n$ is denoted by $P(d\mathbf{u}|\alpha,n,\mathbf{k})$. Note that if the kernels $k(.|.)\equiv 1$, and the shape measure is finite with with total mass $\alpha(R)$ the WP reduces to the extended Polya process with shape measure $\alpha(.)$. An inspection of (4.1) results in the following change of variable result connecting the distribution $\mu(d\mathbf{u}|\alpha)$ and the WP distribution $P(d\mathbf{u}|\alpha,n,\mathbf{k})$.

Lemma 2. Let $K_{n-1}=\kappa(0)\times...\times\kappa(n-1)$. For all $s\ge 0$,

$$\int s(\mathbf{u})\Pi_{1\le i\le n}k(x_i|u_i)\mu(d\mathbf{u}|\alpha)=\int s(\mathbf{u})K_{n-1}\ P(d\mathbf{u}|\alpha,n,\mathbf{k}).$$

According to Lemma 2, (4.1) can be written as, for any $h \geq 0$,

(4.3) $$E[h(G)|x_1,...,x_n] = \int[\int h(G)\mathbb{D}\ (dG|\alpha+\Sigma_{1\leq i\leq n}\delta_{u_i})]K_{n-1}P(d\mathbf{u}|\alpha,n,\mathbf{k})/\int K_{n-1}P(d\mathbf{u}|\alpha,n,\mathbf{k}).$$

The iid WP method is defined by running the WP algorithm M times independently to get $u_1(m),...,u_n(m)$, and $K_{n-1}(m)$, $m=1,...,M$. The M–average

(4.4) $$\hat{h}_M = [\alpha(R)+n]^{-1}\times\Sigma_{1\leq m\leq M}[\int h(G)\mathbb{D}\ (dG|\alpha+\Sigma_{1\leq i\leq n}\delta_{u_i(m)}]K_{n-1}(m)/\Sigma_{1\leq k\leq M}K_{n-1}(m)$$

is an iid MC approximation to $E[h(G)|x_1,...,x_n]$.

The WP is easy to implement; it is a straightforward simulation of the "missing value" u_i based on $k(.|.)$ and $\alpha(.)/\alpha(R)$. It does not account for the data reduction part of the missing values, and the cost is heavy.

Example 4.1. Bayesian deconvolution (cont.) For location or scale models, $k(x|u)=k(x-u)$ or $k(x|u)=uk(xu)$, respectively. The WCR and WP are compared in the case of estimating G. Here

(4.5) $$\hat{G}(y) \propto \int[\alpha(y)+\Sigma_{1\leq i\leq n}I\{u_i\leq y\}K_{n-1}P(d\mathbf{u}|\alpha,n,\mathbf{k})$$

and the WP simulated $\hat{G}(y)$ (dotted) is essentially an average of $\alpha(y)+\Sigma_{1\leq i\leq n}I\{u_i\leq y\}$ and hence is discrete [up to the prior factor factor of $\alpha(y)$]. In comparison, the WCR $\hat{G}(y)$ (dashed) given by (3.8) is differentiable since it is essentially an average of the sub–posterior distribution function $\pi(du|C_i)$, which is differentiable if $\alpha(.)$ is differentiable. The sensitivity of the $\hat{G}_M(u)$ to the choice of prior parameters can also be detected. Data n=400 are from a three–peak location mixture of normals. Moving from the first column to the second column, the normal sub–prior mean increases from 0.1 to 4.9 (while the sub–prior precision remains a constant t=0.1), $\hat{G}_M$ right–shifts accordingly. Moving from the first row to the second row, the normal sub–prior precision increases from 0.1 to 0.5 (while the sub–prior mean m remains constant), the accuracy of $\hat{G}_M$ deteriorates.

Figure 4.1 suggests that the WP is more sensitive to the choice of prior parameters than the WCR. Examining the WP algorithm, one sees that at the kth step U_k would be one observation from $\pi(du|x_k)$, or that it is one of the previous $U_1,...,U_{i-1}$. Since $\pi(du|x_k)$ differs from $\pi(du)$ only by one observation, the U_i are close to a sample from the sub–prior $\pi(du)$. This could disrupt the WP algorithm if the data likelihood and the sub–prior $\pi(du)$ are approximately "orthogonal" (for a lack of a better word). The next example isolates this effect.

Example 4.2. Sensitivity of WP to the prior. We use the location–scale mixture of $N(u,0.7^2)$'s; $\alpha(.)/\alpha(R)$ is $N(m,1/t)$. Figure 3.2 displays $\hat{f}_M(t)$. Data are n=300 observations

from a unimodal $N(0,0.7^2)$. One expects a good fit due to the unimodal nature of the data. At m=3 and t=5, the sub–prior $\alpha(.)/\alpha(R)$ being $N(3,0.447^2)$ is almost "orthogonal" to a $N(0, 0.7^2)$ in the sense that they have almost no overlap. The resulting WP approximation is bad; in comparison, the WCR approximation is already indistinguishable from the target $N(0,0.7^2)$. Deflating the sub–prior $N(3,1/t)$ by reducing the precision t from t=5, 3.5, 2, 0.01 results in incresingly better WP approximations; at t=0.01, the WP approximation is indistinguishable from the target $N(0,0.7^2)$.

For a defence against such orthogonal property between the data likelihood and the prior density, one could choose a prior which is "flat". For example, in the last Example 4.2, t=0.01 corresponds to a flat $N(3,10^2)$ sub–prior, and results in an excellent WP approximation. However, there is a question of the degree of "flatness". The following example suggests that in some cases, too "flat" will result in noise artifact for WP and, to a lesser degree, for WCR.

Example 4.3. Location and scale mixture of normals (cont.). Data (n=300) are from the three–peak location and scale mixture of normals of Example 3.2, and the sub–prior used is the gamma–normal conjugate prior discussed in that example. $\hat{f}(t)$ is essentially a mixture of sub–predictive t densities, with degrees of freedom $2a_i$, location m_i, and pecision $(a_i/b_i)t_i/(t_i+1)$. In Example 3.2, we gave the reasons for considering only the case of small and moderate b and t. Set m=–2 and a=1.5 and observe $\hat{f}_M(t)$ as a function of b and t [WCR (dashed) and WP (solid)]. A moderate size t, say t=2, and small b (b=0.5, 0.05, 0.005) result in a $\hat{f}_M(t)$ with a sharp and narrow peak at m=–2 (first column of Figure 4.3a) which can be explained by the no data predictive t–density with location m (=–2) as the precision $(a/b)t/(t+1)$ increases. For a small t, say t=0.0005, decreasing b (b=10, 1.5, 0.05) produces noise artifact (second column of Figure 4.3a). The best results are obtained from moderate $b\in[0.5, 2]$ and a small t<0.01. Figure 4.3b exemplifies the effect of different $\alpha(R)$'s. The parameters are set at a moderate b (=1.5) and a small t (=0.0005). Increasing $\alpha(R)$ (20 or 100) shifts the mass of a WP $\hat{f}_M(t)$ to the tails; the WCR $\hat{f}_M(t)$ remains stable. In conclusion, the "window of opportunity" for WP is very narrow, at least for sample sizes less than 400.

Example 4.4. Mixture of binomial (m,u) kernels; 0≤u≤1. If $\alpha(.)/\alpha(R)$ is beta (a,b), the sub–posterior distribution of u given $x_j, j\in C_i$ is beta $(a+\Sigma_{j\in C_i}x_j,\ b+me_i-\Sigma_{j\in C_i}x_j)$. The predictive density for the next observation conditional on table C_i, $m(x|C_i)$, is beta–binomial. The sub–prior is a "flat" uniform (0,1) distribution, and the WCR and the WP are compared vs the changes of $\alpha(R)$. The data are 320 tack data from Beckett and Diaconis (1994); the data were analysed by Kong, Liu and Wong (1992) and Liu (1996). $\hat{f}_M(t)$ is a histogram, so are its WCR and WP approximations. Figure 4.4a plots the WCR $\hat{f}_M(t)$ (dashed), the WP $\hat{f}_M(t)$

(solid), and the histogram of the data (dotted). The mixture of binomial (m,u) density is determined by m+1 mixture probabilities which sum up to unity. The m probabilities determine the first m moments of G, which does not determine G in general. Thus, the model is not identifiable in G and estimating G is a problem. Nevertheless, Figure 4.4b plots a discrete WP $\hat{G}_M(u)$ (solid), a WCR $\hat{G}_M(u)$ (dashed). The derivative of a WCR $\hat{G}_M(u)$ is plotted in the graph imediately below it, which reveals two peaks at 0.5 and 0.8. The graphs for the WP $\hat{G}_M(u)$ fail to reveal two jumps at 0.5 and 0.8; repeated WP Monte Carlo simulations result in different $\hat{G}_M(u)$.

Example 4.5. An empirical WCR. The WCR density has two key factors: the Chinese restaurant process weight $\alpha(R)^{n(\mathbf{p})}\Pi_{1\le i\le n(\mathbf{p})}(e_i-1)!$ and the "predictive weight" $\Pi_{1\le i\le n(\mathbf{p})}\Pi_{1\le j\le e_i} m(x(i)_j|x(i)_1,,...,x(i)_{j-1})$. One suspects that the contribution of the predictive weight would be paramount, and an empirical WCR which is guided by the predictive weights would perform well. To define it, set $\alpha(R)\equiv 1$, $\lambda(0)=m(x_1)$. After customers 1,...,k−1 are seated (k=2,...,n), define

$$\lambda(k-1)=m(x_k)+\Sigma_i m(x_k|C_i),$$

the index i runs over indices of currently available tables. The seating probability for customer k is then $m(x_k|C_i)/\lambda(k-1)$ for table C_i; customer k sits at an empty table with probability $m(x_k)/\lambda(k-1)$. At the completion of the n−th step, define

$$\lambda(n)=m(t)+\Sigma_{1\le i\le n(\mathbf{p})} e_i m(t|C_i).$$

Denote the probability distribution of $\mathbf{p}$ induced by this sequential seating by $q_e(\mathbf{p}|\mathbf{x})$, and an empirical WCR density by

$$\hat{e}(t) =[\alpha(R)+n]^{-1}\Sigma_{\mathbf{p}}\Lambda_n q_e(\mathbf{p}|\mathbf{x})/\Sigma_{\mathbf{p}}\Lambda_{n-1} q_e(\mathbf{p}|\mathbf{x}).$$

Define $\hat{e}_M(t)$ accordingly. It turns out that in general the $\hat{e}_M(t)$ performs slightly better than the WP, yet is worse than the WCR. Figure 4.5 displays $\hat{e}_M(t)$, WCR, and WP for the location and scale mixture of normal model. Data are the three−peak data of Example 4.3. $\alpha(R)=1$.

We conclude this section with a discussion of the running time (of compiled C code on a SUN Sparc 20) of the WCR and WP algorithms. A scrutiny of the two algorithms reveals that they differ in the sense that the WP algorithm requires the simulation a missing value from a posterior distribution in each step of the loop, and the WCR requires the selection of a table out of $n(\mathbf{p})+1$ possible tables with prescribed seating probabilities. Additional tables mean more updating and calculations for the WCR. As a result, the WCR algorithm runs more slowly as the number of tables increases. The probability of opening a new table is proportional to $\alpha(R)$, hence, the WCR runs slower when $\alpha(R)$ is large. For comparison, the WP is essentially a

constant function of $\alpha(R)$. Table 4.1 illustrates the running time as a function of $\alpha(R)$ using the three–peak data from the location and scale mixture of normal model in Example 4.3 (calculating $\hat{f}_M(u)$ at a grip of 350 u on the u–axis; a=b=1.5, m=0.0, t=0.0005). The sample size is fixed at n=150.

Table 4.1 CPU time for sequential sampling algorithms vs changes in $\alpha(R)$

$\alpha(R)$	0.1	1	5	10
WCR	1'01"	1'38"	2'41"	3'29"
WP	11'12"	11'15"	11'14"	11'15"

The comparison of the WCR and the WP in terms of the sample size is more complicated. Table 4.2 summarizes a study of the running time vs the sample size where the missing value u has a dimension of one. The data are from the three–peak location mixture of normals in Example 4.1 ($\alpha(R)$=1.0; m=0.0, t=0.1). In each case, $\hat{f}_M(x)$ are evaluated for 350 grips on the x–axis.

Table 4.2 CPU time for sequential sampling algorithms vs changes in sample size (u is one dimensional)

n	50	150	300
WCR	0'16"	0'23"	0'31"
WP	2'53"	9'45"	23'49"

Table 4.3 summarizes a study of the running time vs the sample size n where the missing value u=(τ,μ) has a dimension of two. The data are from the three–peak location and scale mixture of normals in Example 4.3 ($\alpha(R)$=1.0; a=b=1.5, m=0.0, t=0.0005). In each case, $\hat{f}_M(x)$ are evaluated for 350 grips on the x–axis.

Table 4.3 CPU time for sequential sampling algorithms vs changes in sample size (u is two dimensional)

n	50	150	300
WCR	1'16"	1'41"	2'13"
WP	3'19"	11'14"	26'36"

(For M=2000 repetitions, the running time approximately doubles.) The two tables suggest that an increase in the dimension of the missing value affects the speed of the WCR more than the WP. That is, for high dimensional mixture models, the management and updating of the opened tables is less efficient than simulating a high–dimensional missing value, and the WP could have an edge in terms of a smaller running time. However, this possible effect should be investigated in a thorough study of high dimensional mixture problems.

Section 5. The mixture hazard rate model.

The likelihood function of a hazard rate point process model is proportional to

(5.1) $$[\Pi_{1\le i\le n}\, r(x_i)]\, \exp\{-\int_I Y(s)r(s)ds\}$$

where r(s) is a hazard rate, I is the interval in which the point process is being observed, x_i is the "failure time", and Y(s) is a left continuous integer valued function of the data. The likelihood function of the multiplicative point process models [Aalen (1981)] involves a product of these likelihood factors, each of which can be treated independently (by a Bayesian) and creates no additional complexities [Lo and Weng (1989)]. The hazard rate model (5.1) and its multiplicative extension include point process models such as life testing models with censored data, Poisson models and competing risk models, amomg other point process models. In a mixture hazard rate model, the hazard rate r(.) depends on a kernel k(.|.) and a mixing measure $\mu(du)$ on the "missing" variable u such that

(5.2) $$r(s|\mu)=\int k(s|u)\mu(du);$$

$k(.|.)\ge 0$ and satisfies integrability conditions in both variables. Often, it is convenient to assume that for each u, k(.|u) is a density [see Section 5 in Lo and Weng (1989), and Chapter 3 in Snyder and Miller (1991)]. In some cases, the kernel k(.|.) generates a hazard rate which has desirable smoothness properties. For example, the scale mixture of uniforms generates monotone hazard rates, and the scale mixture of exponentials generates "completely monotone" hazard rates [see page 20 in Feller (1971)]. On the other hand, the mixture hazard rate model often arises as a result of missing information. The most renowned example perhaps is the emission tomography models where the data are in fact from a Poisson point process with mixture hazard rates; see Chapter 3 in Snyder and Miller (1991) for references.

In their discussion of the Bayesian mixture hazard rate model, Lo and Weng (1989) argued that the likelihood function (5.1) [and (5.2)] looks like a gamma density in $\mu(.)$, and suggested a weighted gamma process prior for $\mu(.)$. A random measure $\nu(.)$ is a gamma process with shape measure $\alpha(.)$ if (i) $\nu(.)$ is "independent increment" and (ii) for each A, $\nu(A)$ is a gamma $(\alpha(A),1)$ random variable. [The theory of Dirichlet process can be understood via a gamma process in the sense that $\nu(.)/\nu(R)$ is a Dirichlet process with a finite shape measure $\alpha(.)$.] The random process $\mu(.)$ defined by $\mu(A)=\int_A \beta(u)\nu(du)$ is called a weighted gamma process with shape $\alpha(.)$ and multiplier $\beta(.)\ge 0$; its distribution is denoted by $G\ (d\mu|\alpha,\beta)$. See Lo (1982) for the calculus of weighted gamma processes. In this gamma process prior setting, hazard rates which are scale mixture of uniform kernels were considered by Dykstra and Laud (1971) in life testing models. Theorem 4.1 in Lo and Weng (1989) provides a representation

which states that the posterior distribution of μ for model (5.1) and (5.2) is an extended Polya mixture of weighted gamma processes given by

(5.3) $\pi(d\mu|data) \propto \int G\ (d\mu|\alpha+\Sigma_i\delta_{u_i},\beta^*)\Pi_i k^*(x_i|u_i)\mu(d\mathbf{u}|\alpha)$,

where $\beta^*(u)=\beta(u)/[1+\beta(u)\int_I Y(t)\times k(t|u)dt]$,

$k^*(t|u)=\beta^*(u)k(t|u)$,

and $\mu(d\mathbf{u}|\alpha)$ is defined in (4.2). This mixture of gamma processes posterior distribution is another example of (4.1) where $k^*(x_i|u)$ playing the role of the likelihood weight $k(x_i|u)$, and WCR and WP can be defined by a change to the *–notation. That is, putting $w_i(u)=k^*(x_i|u)$ in the definition of the WCR marginal and predictive weights [(2.2) and (2.3)] The definitions of Λ^*_{n-1}, Λ^*_n and $q(\mathbf{p}|\alpha,n,\mathbf{k^*})$ follow. The posterior mean of $r(t|\mu)=E[r(t|\mu)|x_1,...,x_n]$ is [Theorem 4.2 in Lo and Weng (1989)], $t=x_{n+1}$,

(5.5) $\hat{r}(t)=\Sigma_{\mathbf{p}}\Lambda^*_n q(\mathbf{p}|\alpha,n,\mathbf{k^*})/\Sigma_{\mathbf{p}}\Lambda^*_{n-1}q(\mathbf{p}|\alpha,n,\mathbf{k^*})$.

This expression is almost identical to the expression of $\hat{f}(t)$ in (3.3), except the factor $[\alpha(R)+n]^{-1}$ which now becomes a part of β^* in the definition of k^*. Sampling the WCR from $q(\mathbf{p}|\alpha,n,\mathbf{k^*})$ results in an iid Monte Carlo approximation to $\hat{r}(t)$ and to posterior moments of μ.

One could also incorporate a regression model here by letting $w_j(u)\equiv k(x_j|u)$ which depends on a regression variable z_j.

Section 6. Comparing simulation experiments.

The usual way of comparing the performance of different Monte Carlo methods is to compare the variances or the covariance matrices of the resulting Monte Carlo approximations. However, Blackwell's notion of comparison of experiments [Blackwell (1951, 1953); see also De Groot (1970) and Strassen (1965)] provides a desirable alternative. Suppose the experiment is to observe two random variables T and S. If there is a known nondegenerate conditional distribution of T given S, F(dt|s), then S is sufficient for T in the sense that if S is observed, then one should not observe T (without taking into account the observed S) as one can always produce an observation in T using the conditional distribution F(dt|S). In short, if you know S, T is still uncertain. This comparison implies that S is less variable and hence more informative than T. (If there is also a known nondegenerate distribution of S given T, then one can say that S and T are equally informative.)

Accordingly, the comparison of the Chinese restaurant process and Blackwell–MacQueen's extended Polya urn process is as follows. S is a random partition $\mathbf{p}$, and T is the urn (missing value) vector $\mathbf{u}$. According to discussions in Sections 2 and 3, $\mathbf{p}$ has a marginal density $q(\mathbf{p}|c)$ and $\mathbf{u}$ has a marginal distribution $\mu(d\mathbf{u}|\alpha)/B_{n-1}(c)$, where c=the total mass of $\alpha(.)$. Korwar and Hollander (1973) constructed a conditional distribution of $\mathbf{u}$ given $\mathbf{p}$ which can be described as follows: Given $\mathbf{p}=C_1,...,C_{n(\mathbf{p})}$, define

(6.1) (I) $U_i|\mathbf{p}$ has distribution $\alpha(du)/\alpha(R)$, for $i=1,...,n(\mathbf{p})$,

(II) $U_1,...,U_{n(\mathbf{p})}|\mathbf{p}$ are independent,

(III) for $i=1,...,n(\mathbf{p})$, duplicate U_i a total of e_i times, and denote them by U_j, $j\in C_i$; $i=1,...,n(\mathbf{p})$.

The existence of the conditional distribution (6.1) supports Kuo's (1986) initial proposal to use the Chinese restaurant process rather than the extended Polya urn to perform the iid MC approximation to posterior quantities in Bayesian mixture models. Unfortunately, the contribution from the peaky integrands dominates and blurs the otherwise observable improvements.

Conditional distributions of $\mathbf{u}|\mathbf{p}$ also exist in comparing a sequential imputation $\mathbf{u}$ which has a distribution proportional to $K_{n-1}P(d\mathbf{u}|\alpha,n,\mathbf{k})$, and a sequential seating $\mathbf{p}$ which has a distribution proportional to $\Lambda_{n-1}q(\mathbf{p}|\alpha,n,\mathbf{k})$. First, we extend Lemma 2 in Lo (1984) from a finite $\alpha(.)$ to a σ–finite $\alpha(.)$ so that both mixture density models and hazard rates models can be treated together. Recall definition (4.2) that $\mu(d\mathbf{u}|\alpha)=\Pi_{1\le i\le n}(\alpha+\Sigma_{1\le j\le i-1}\delta_{u_j})(du_i)$.

Lemma 3. For a σ–finite $\alpha(.)$ and all $g_i \geq 0$,

$$\int\{\Pi_{1\leq i\leq n} w_i(u_i)\}\mu(d\mathbf{u}|\alpha)=\Sigma_{\mathbf{p}}[\Pi_{1\leq i\leq n(\mathbf{p})}(e_i-1)!m(C_i)].$$

The above equality is true if $\alpha(.)$ is restricted to a set A_k such that $\alpha(A_k)$ is finite [Lemma 2 in Lo (1984)]. Lemma 3 follows from letting A_k increases to R such that $\alpha(A_k)$ increases to $\alpha(R)$. The existence of such a sequence $\{A_k\}$ is guaranteed by the σ–finiteness of $\alpha(.)$. Denote a WP distribution of $\mathbf{u}$ generated from the weights $w_i(u)$ rather than $k(x_i|u)$, i=1,...,n, by $P(d\mathbf{u}|\alpha,n,\mathbf{w})$. Lemma 3 and the product rule of probability (2.4) imply the following:

Lemma 4. For all $g_i \geq 0$,

$$\int\{\Pi_{1\leq i\leq n} g_i(u_i)\}\times K_{n-1}P(d\mathbf{u}|\alpha,n,\mathbf{w})$$
$$=\Sigma_{\mathbf{p}}\{\Pi_{1\leq i\leq n(\mathbf{p})}\int\Pi_{j\in C_i} g_j(u)\pi(du|C_i)\}\times\Lambda_{n-1}q(\mathbf{p}|\alpha,n,\mathbf{w});$$

in particular, $\int K_{n-1}P(d\mathbf{u}|\alpha,n,\mathbf{w})=\Sigma_{\mathbf{p}}\Lambda_{n-1}q(\mathbf{p}|\alpha,n,\mathbf{w})$.

The equality in Lemma 4 is valid for all nonnegative g_i and hence defines a conditional distribution as follows: Suppose $\mathbf{p}$ has a (marginal) density proportional to $\Lambda_{n-1}q(\mathbf{p}|\alpha,n,\mathbf{w})$, and the $\mathbf{u}$ has a (marginal) distribution proportional to $K_{n-1}P(d\mathbf{u}|\alpha,n,\mathbf{w})$. Given $\mathbf{p}=C_1,...,C_{n(\mathbf{p})}$, define $\mathbf{u}|\mathbf{p}$ as in (6.1) except that (I) is replaced by

(I') $U_i|\mathbf{p}$ has distribution $\pi(du|C_i)$ for i=1,...,n($\mathbf{p}$).

Use Lemma 4 and the double expectation formula to rewrite the n–fold integrals in (4.1) and (5.3) as sums of n($\mathbf{p}$)–fold integrals, and we arrive at

Theorem 1. Suppose the joint distribution of $\mathbf{p}$, $u_1,...,u_n$ is specified by (I'), (II), and (III).

(i) For the Bayesian mixture density model with posterior distribution given by (4.1),

$$\pi(dG|\mathbf{x}) \propto \Sigma_{\mathbf{p}}\int_{\mathbb{D}} (dG|\alpha+\Sigma_{1\leq i\leq n(\mathbf{p})}e_i\delta_{u_i})[\Pi_{1\leq i\leq n(\mathbf{p})}\pi(du_i|C_i)]\Lambda_{n-1}q(\mathbf{p}|\alpha,n,\mathbf{k});$$

(ii) for the Bayesian mixture hazard rate model with posterior distribution (5.3),

$$\pi(d\mu|\mathbf{x}) \propto \Sigma_{\mathbf{p}}\int_{\mathbb{G}} (d\mu|\alpha+\Sigma_{1\leq i\leq n(\mathbf{p})}e_i\delta_{u_i},\beta^*)[\Pi_{1\leq i\leq n(\mathbf{p})}\pi(du_i|C_i)]\Lambda_{n-1}q(\mathbf{p}|\alpha,n,\mathbf{k^*}).$$

Theorem 1 states that all posterior quantities are obtainable from sampling a partition, and then sampling from the transition probability given the observed partition. Discussions in Section 3 point out that for evaluating posterior moments, sampling from the transition probability can be replaced by running an extended WCR.

Wing Hung Wong recently brought to our attention of Liu, Chen, and Wong (1996) which discussed an improvement of the sequential imputation method [Kong, Liu, Wong (1994)] based on a rejection scheme. The resulting rejection control sequential imputation reduces the variability of the product WP K_{n-1} factor, resulting in an improved sequential

imputation simulation. In this regard, one anticipates that a WCR equiped with a rejection control strategy reduces the variability of the WCR product factor Λ_{n-1} and will fare even better. The ideal sequential imputation samples a missing value vector from a distribution proportional to $K_{n-1}P(d\mathbf{u}|\alpha,n,\mathbf{k})$; likewise, an ideal sequential seating experiment samples a random partition from a distribution proportional to $\Lambda_{n-1}q(\mathbf{p}|\alpha,n,\mathbf{k})$. Such ideal sequential sampling experiments reduce the variabilities in K_{n-1} and Λ_{n-1} to zero. According to (I'), (II), and (III), Theorem 1(i), and (4.1), the latter beats the former. Similar conclusions hold for the mixture hazard rate model using Theorem 1(ii).

Lemma 4 applies to the MCMC method for Bayesian mixture models yields the following: Upon convergence to stationarity, a sequential imputation MCMC approximates a distribution proportional to $K_{n-1}P(d\mathbf{u}|\alpha,n,\mathbf{k})$. A sequential seating MCMC approximates a distribution proportional to $\Lambda_{n-1}q(\mathbf{p}|\alpha,n,\mathbf{k})$. According to (I'), (II), and (III), Theorem 1(i), and (4.1), a sequential seating MCMC is better in the sense of comparison of experiments; furthermore, it houses a simpler convergence theory since its state space is finite. Similar conclusions holds for the mixture hazard rate model.

Acknowledgements. The authors are greatly indebted to Mr. Martin Law and Mr. Man–Wai Ho for help in various stages of the project. We would also like to thank Ms. Chi–Man Tsui and Ms. Amy C. Yip for their input to this project.

References

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics,* **6**, 701–726.

Aldous, D.J. (1985). Exchangeability and Related Topics. Lecture Notes in Mathematics. 1117 Springer–Verlag.

Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics,* **2**, 1152-1174.

Blackwell, D. (1951). Comparison of experiments. *Proc. 2nd Berkeley Symp. Math. Statist. Prob.* Univer. Calif. Press, Berkeley, Calif., 93–102.

Blackwell, D. (1953). Equivalent comparison of experiments. *Ann. Math. Statist.,* **24**, 265–272.

Blackwell, D. and MacQueen, J. (1973). Fergusion distribution via Polya urn schemes. *The Annals of Statistics,* **1** 353–355.

Brunner, L.J. (1995). Bayesian linear regression with error terms that have symmetric unimodal densities. *Journal of Nonparametric Statistics,* **4**, 335-348.

Brunner, L.J. and Lo, A.Y. (1994). A Bayes approach to directional data. *Canadian Journal of Statistics,* **22** 401-412.

Brunner, L.J. and Lo, A.Y. (1989). Bayes method for a symmetric and unimodal density and its mode. *The Annals of Statistics,* **17**, 1550-1566.

Bunke, O. (1985). Bayesian estimators in semiparametric problems. Preprint Nr. 102, Sektion Mathematik, Humboldt–Universitat Zu Berlin.

Censov, N.N. (1962). Evaluation of an unknown distribution density from observations. *Soviet Math.,* **3**, 1559–1562.

Diebolt, J. and Robert, C.P. (1994). Estimation of finite mixture distributions through Bayesian Sampling. *JRSS* B, **56**, 363–375.

De Groot, M. (1970). *Optimal Statistical Decisions* McGraw–Hill, Inc.

Duda, R.O and Hart, P.E. (1973). *Pattern Classification and Scene Analysis,* John Wiley and Sons.

Beckett, L. and Diaconis, P. (1994). Spectral Analysis for Discrete Longitudinal Data, *Adv. in Math.*, **103**, 107–128.

Doob, J. (1949). Application of the theory of martingales. *Coll. Int. du CNRS.* Paris, 23–27.

Dykstra, R.L. and Laud, P. (1981). A Bayesian nonparametric approach to reliability. *The Annals of Statistics,* **9**, 356–367.

Everitt, B.S. and Hand, D.J. (1981). *Finite Mixture Distributions.*Chapman and Hall, London

Feller, W. (1971). *An Introduction to Probability and its Applications V.II.* John Wiley and Sons

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics,* **1**, 209–230.

Ferguson, T.S. (1983). Bayesian density estimation by mixture of Normal distributions. *Recent Advances in Statistics* Academic Press, 278–302.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika,* **57**, 97–109.

Jewell, N.P. (1982). Mixtures of exponential distributions. *The Annals of Statistics,* **10**, 479–484.

Ji, Wenyun (1991). The evaluation of Bayes estimates for mixture models. Thesis, Master of Arts degree in Statistics. SUNY at Buffalo.

Kong, A., Liu, J. and Wong, W.H. (1994). Sequential imputations and Bayesian missing data problems. *J. Amer. Statist. Assoc.,* **91**, 11684–11688.

Korwar, R.M. and Hollander, M. (1973). Contributions to the theory of Dirichlet processes. *The Annals of Statistics,* **1**, 705–711.

Kuo, L. (1986). Computations of mixtures of Dirichlet processes. *SIAM J. Sci. Stat. Comput.,* **7**, 60–71.

Lindsay, B. (1983). The geometry of mixture likelihoods, I and II. *The Annals of Statistics,* **11**, 86–94 and 783–792.

Liu, J. (1996). Nonparametric Hierarchical Bayes via Sequential Imputations. *The Annals of Statistics,* **24**, 910–930.

Liu, J., R. Chen, and W.H. Wong (1996). Sequential importance sampling with rejection control. Preprint.

Lo, A.Y. (1978). Bayesian nonparametric density methods. Technical Report, Department of Statistics, University of California at Berkeley.

Lo, A.Y. (1982). Bayesian nonparametric statistical inference for Poisson point process. *Z. Wahr. verw. Gebiete.,* **59**, 55–66.

Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics,* **12**, 351–357.

Lo, A.Y. and Weng, C.S. (1989). On a class of Bayesian nonparametric estimates: I. Hazard rate estimates. *Ann. Instit. Statist. Math.,* **41**, 227–245.

Parzen, E. (1962). On the estimation of a probability density and mode. *Ann. Math. Statist.,* **33**, 1065–1076.

Robert, C.P. (1995). Convergence control methods for Markov chain Monte carlo algorithms. *Statistical Science,* **10**, 231–253.

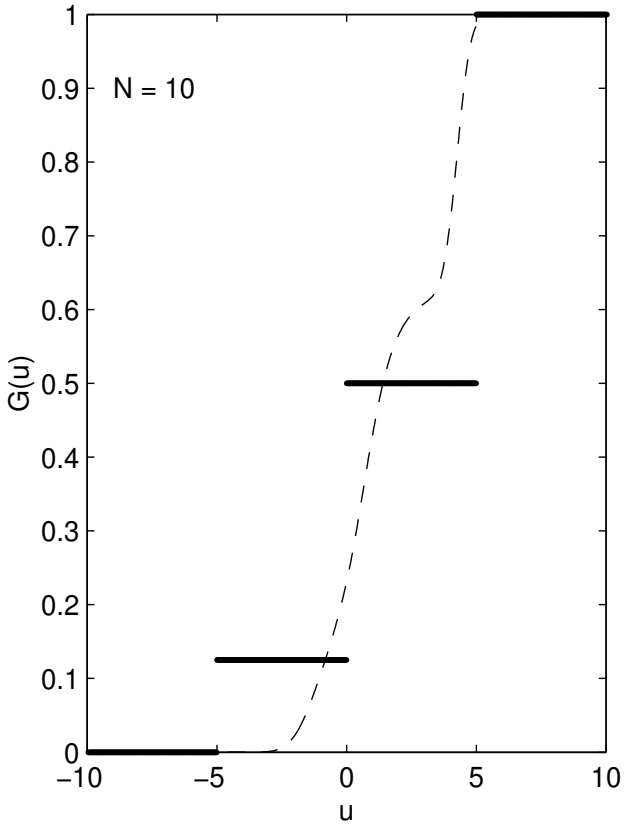
Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.,* **27**, 832–835.

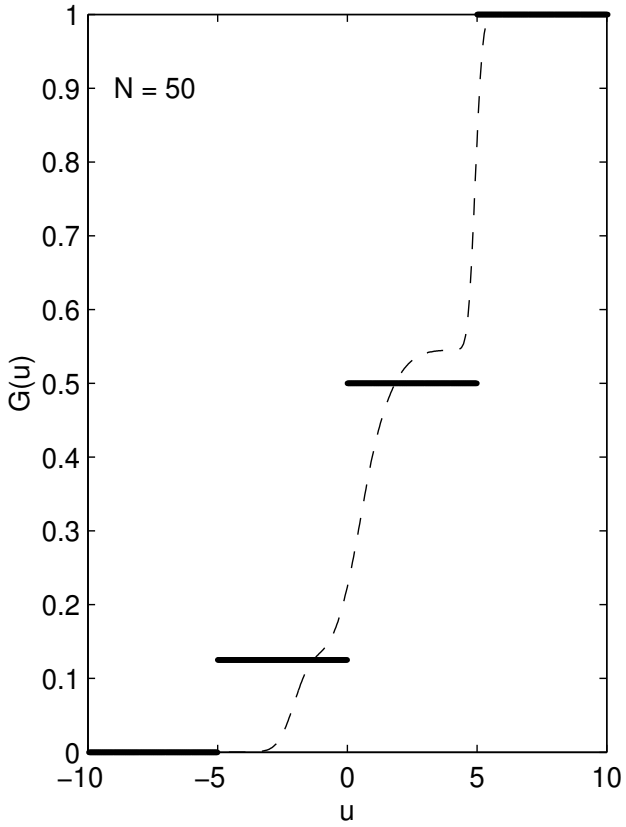
Snyder, D.L. and Miller, M.I. (1991). *Random Point Processes in Time and Space.* Springer–Verlag New York, Inc.

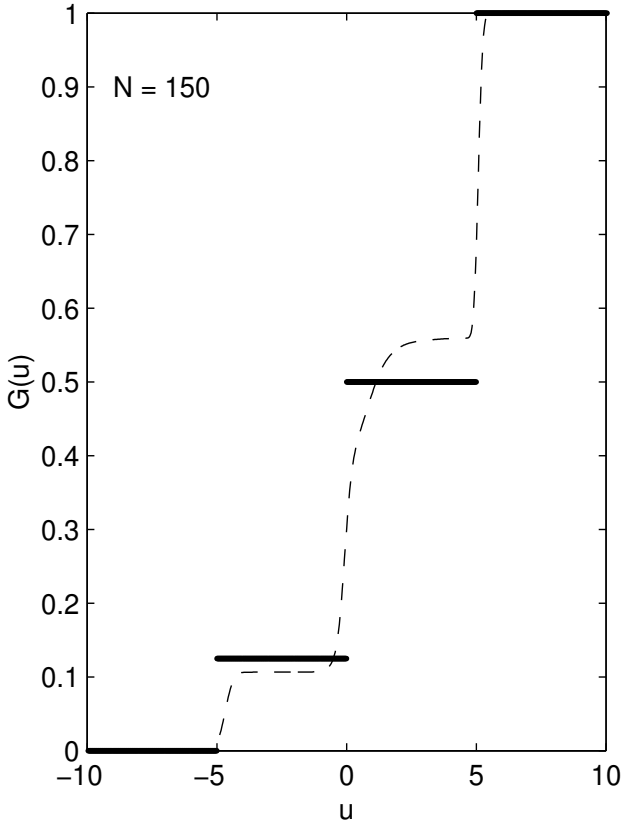
Strassen, V. (1965). The existence of probability measures with given marginals. *Ann. Math. Statist.,* **36**, 423–439.

Titterington, D.M., Smith, A.F.M., and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions* Wiley, New York.

Fig 3.1a Data: 0.125N(−5,1)+0.375N(0,1)+0.5N(5,1) [m=0.1 & t=0.1]







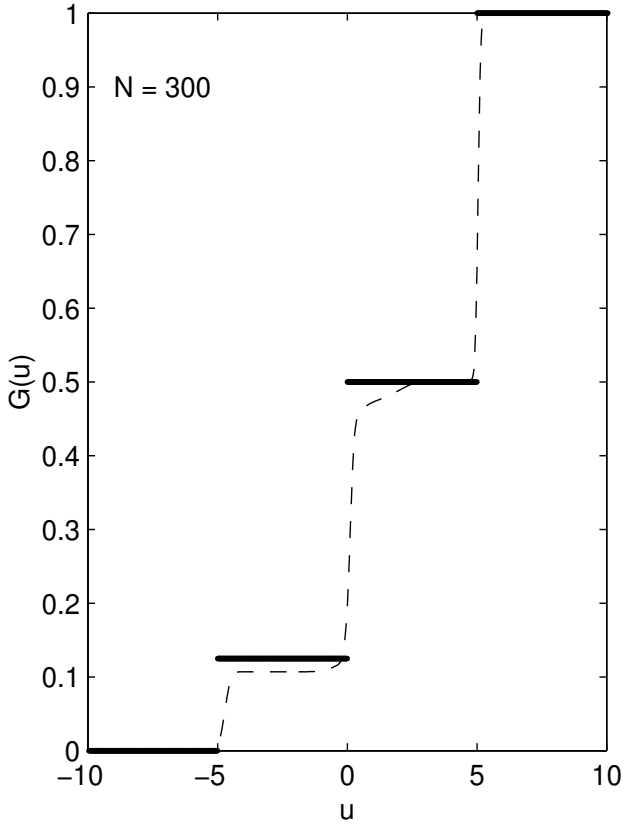


Fig 3.1b Uniform Kernel < M = 2000 >

Data: 0.2U(−0.9,0.1)+0.4U(0,1)+0.4U(0.9,1.9)

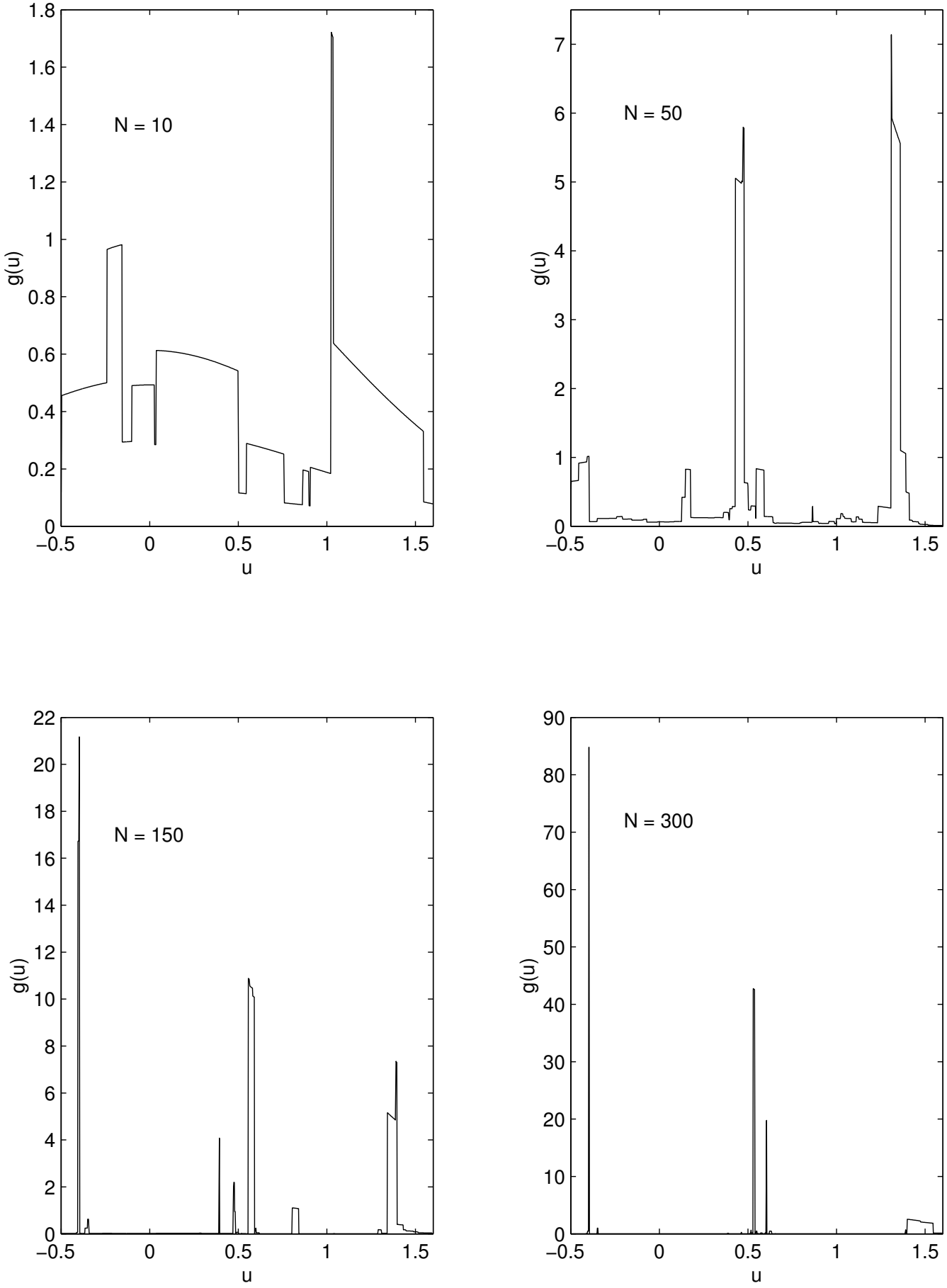


Fig 3.1c Uniform Kernel < M = 2000 >

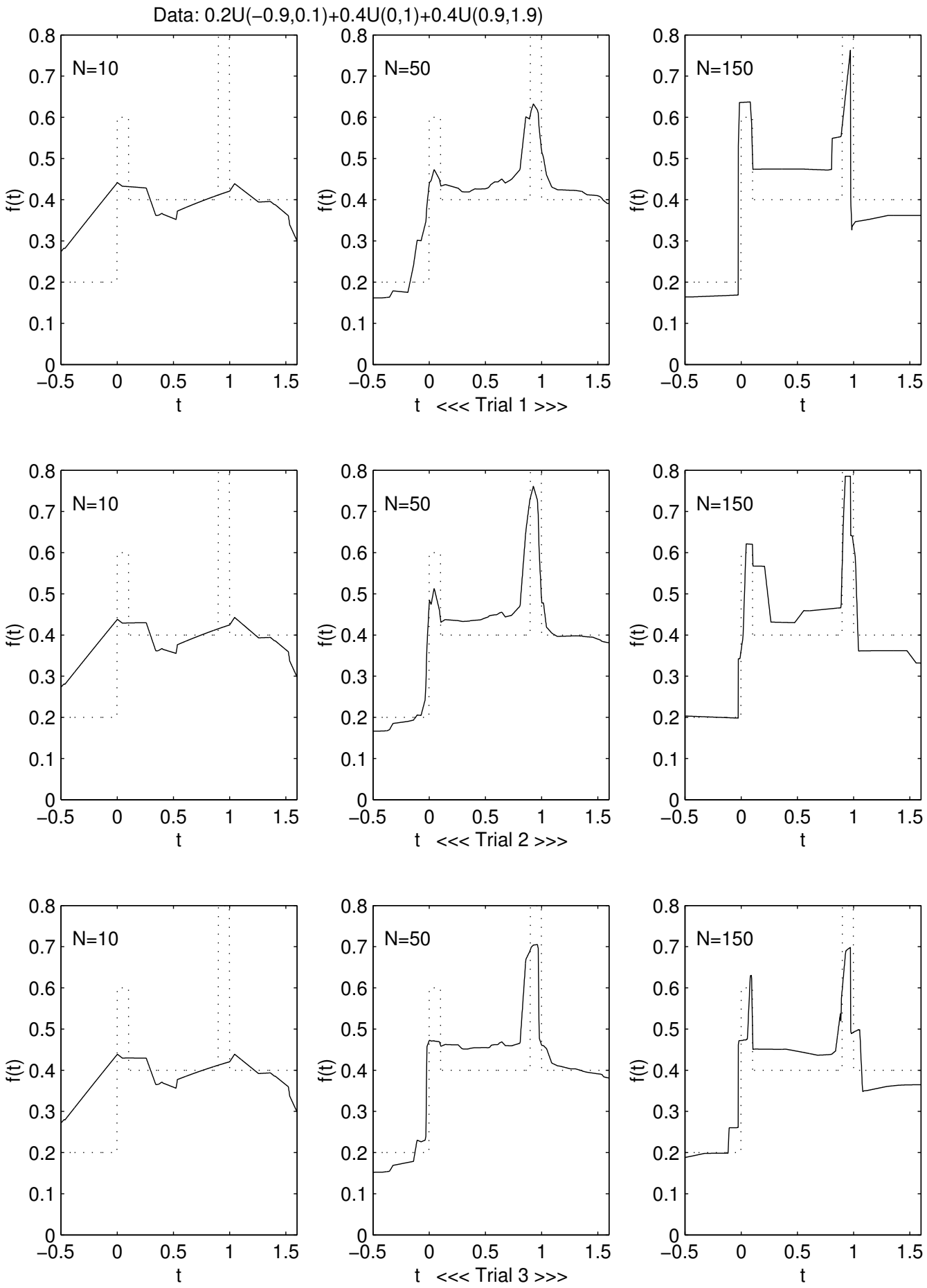


Fig 3.2a Data: 0.6N(−5,0.25)+0.25N(0,1)+0.15N(5,2.25) [a=b=0.5,m=1.0,t=0.1; M=2000]

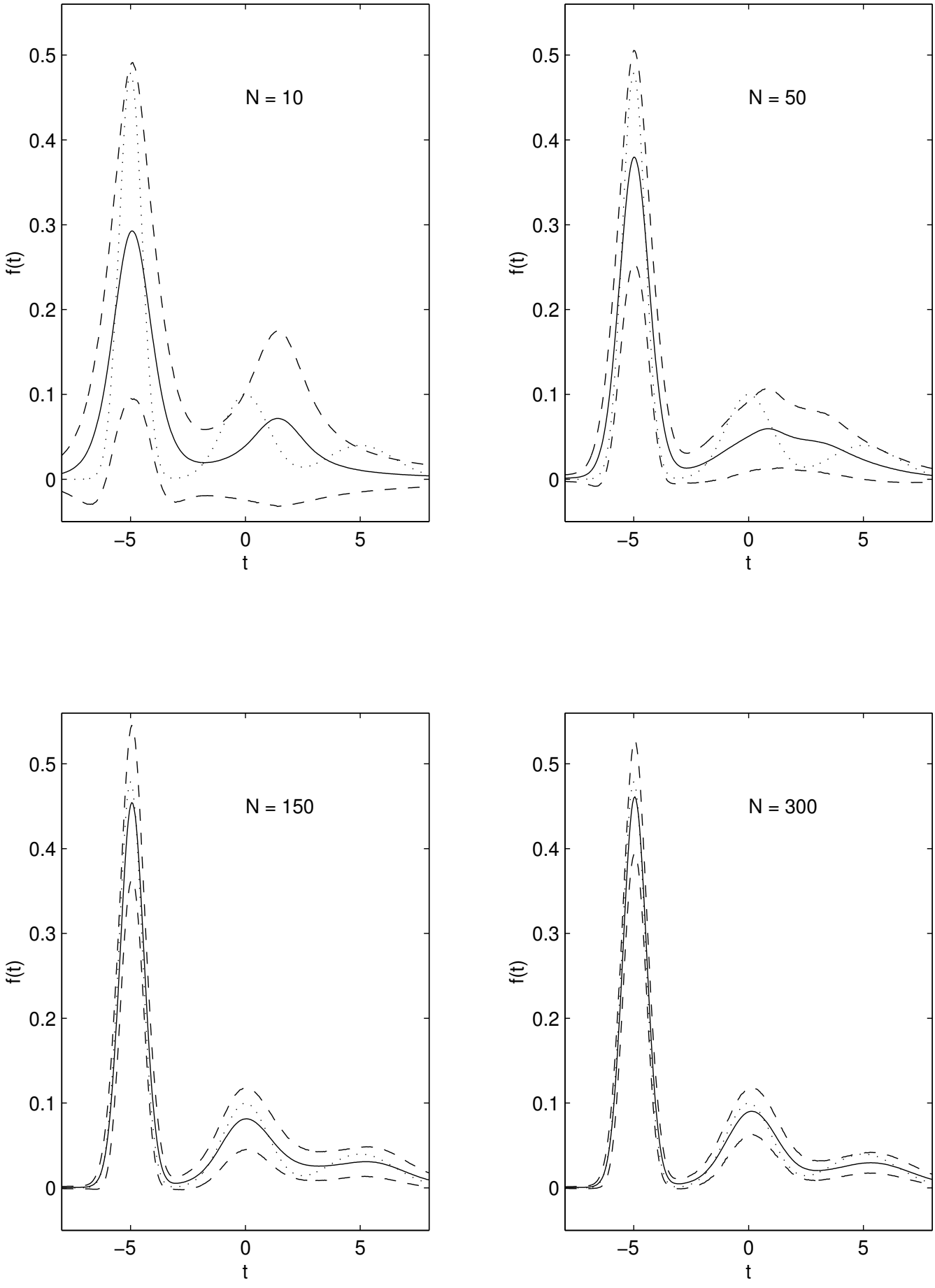


Fig 3.2b Data: exp(0.8N(0,1)+1.5) [a=1.5,b=1.5,m=0,t=0.05; M=2000]

Solid: Approximation By WCR ––– Dot: True Lognormal Density

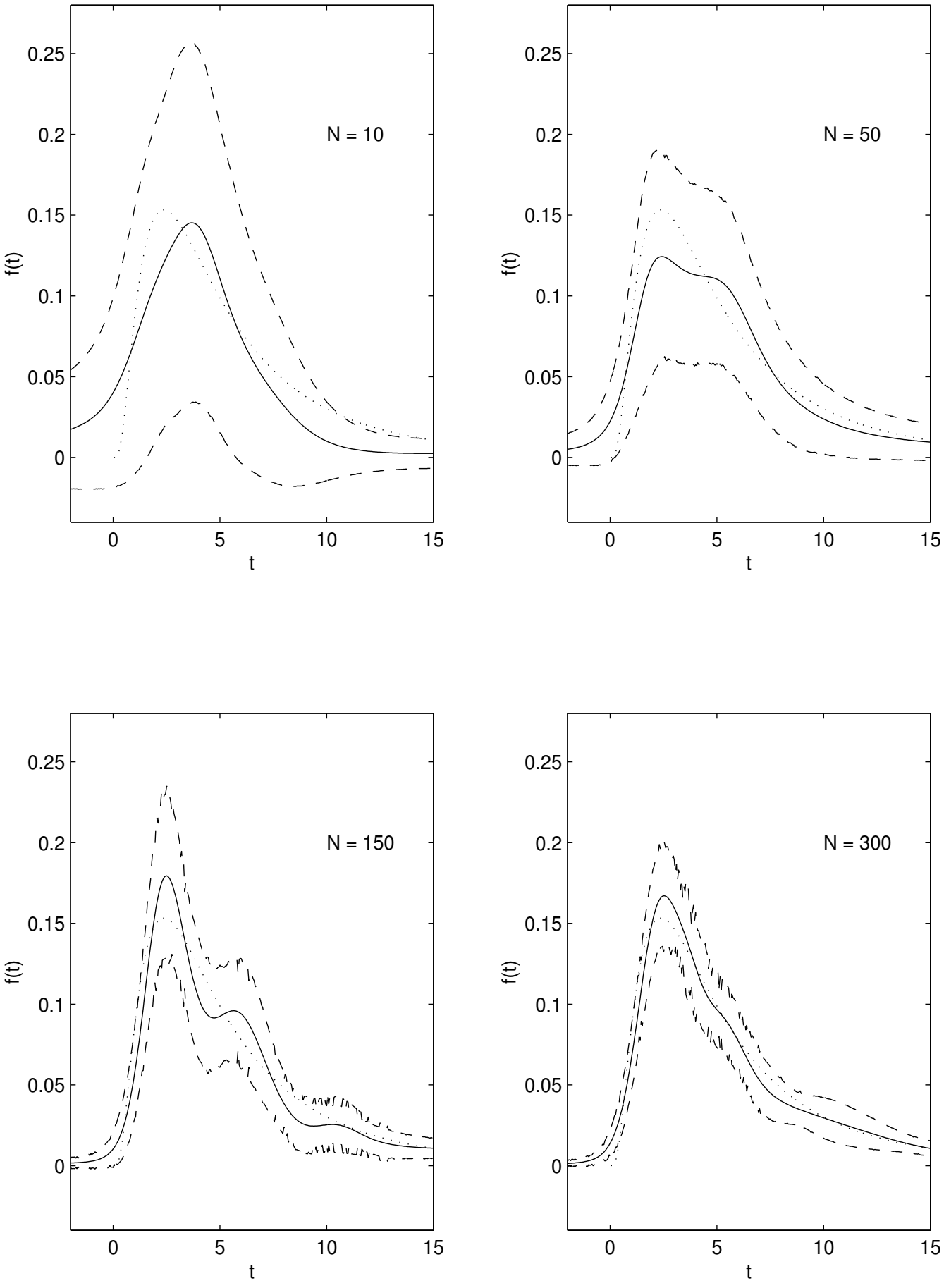


Fig 3.3a [a=b=1.5,m=0.0,t=0.0005]

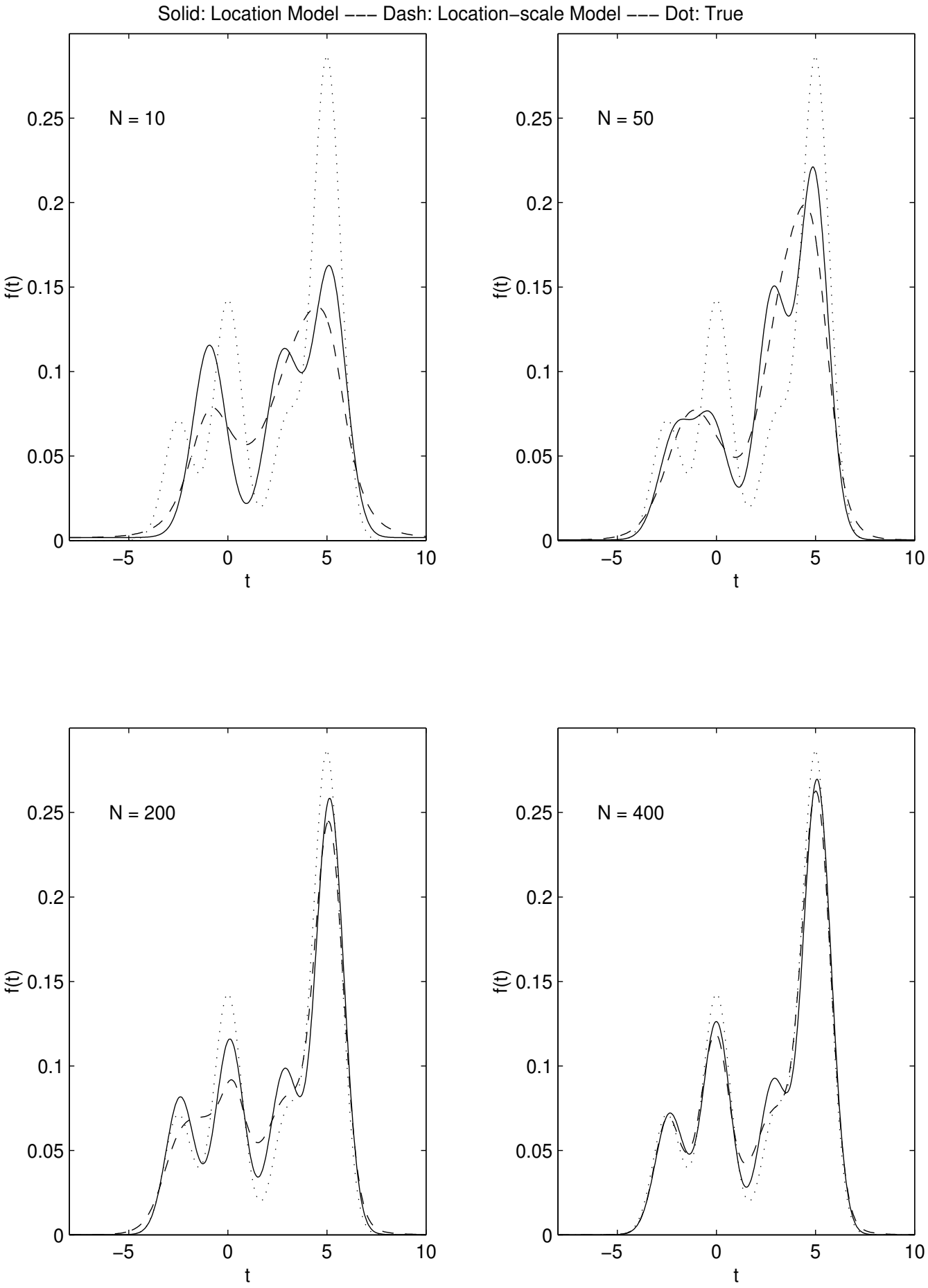


Fig 3.3b [a=b=1.5,m=0.0,t=0.0005]

Solid: Location Model --- Dash: Location–scale Model --- Dot: True

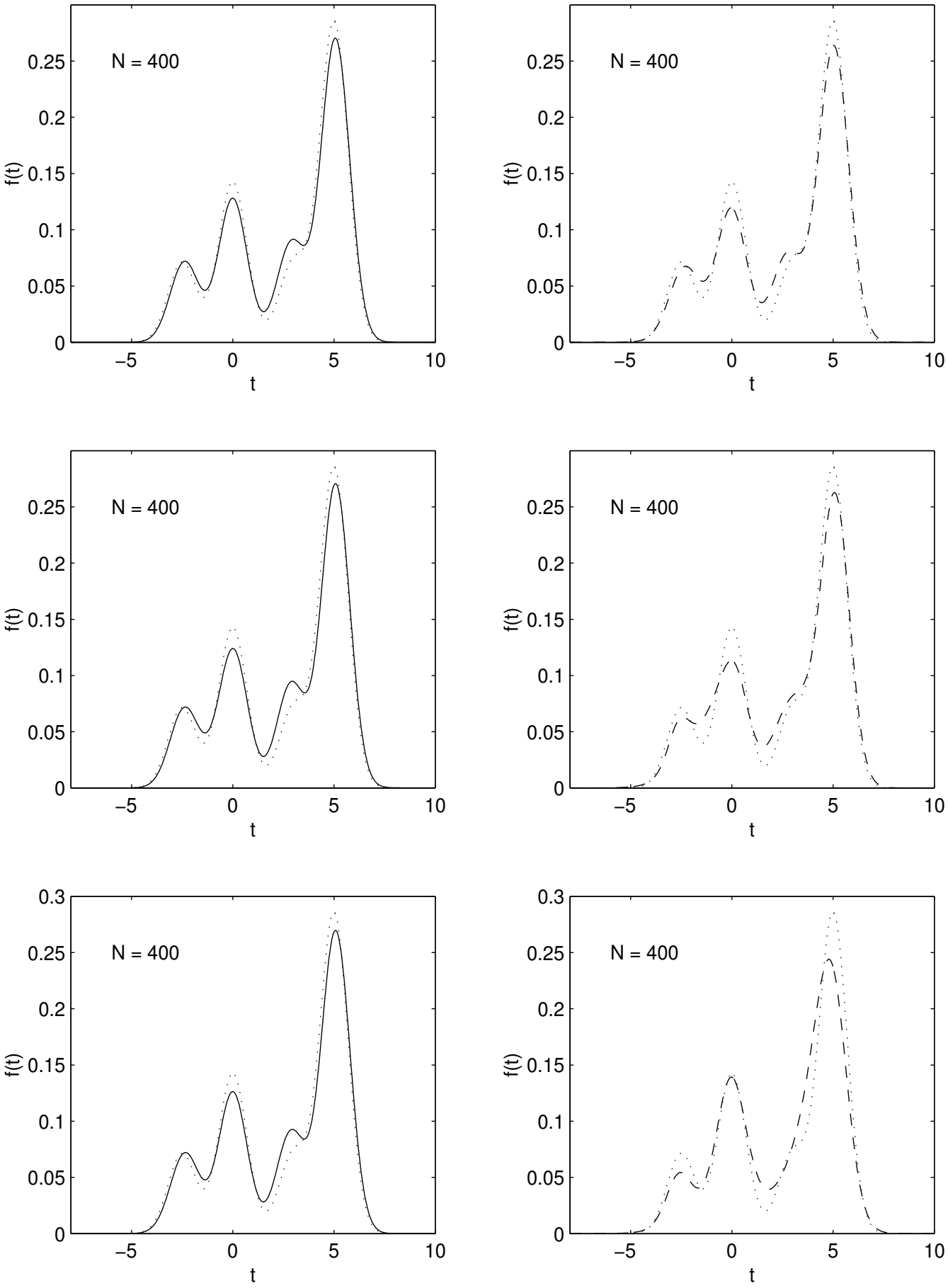
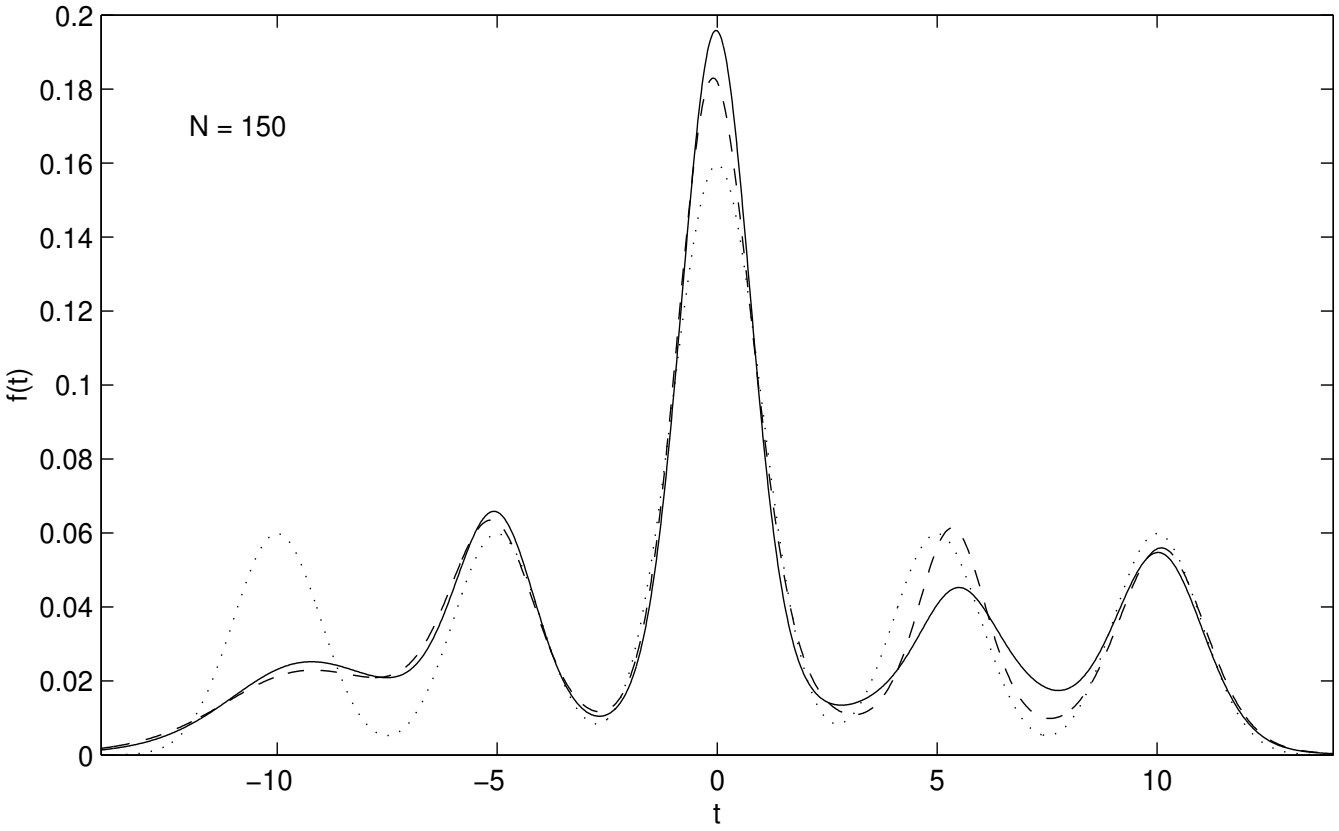


Fig 3.4 Data: 0.15N(−10,1)+0.15N(−5,1)+0.4N(0,1)+0.15N(5,1)+0.15N(10,1)



Solid: Sorted −−− Dash: Unsorted −−− Dot: True [a=0.5,b=0.5,m=1.0,t=1.0; M=2000]

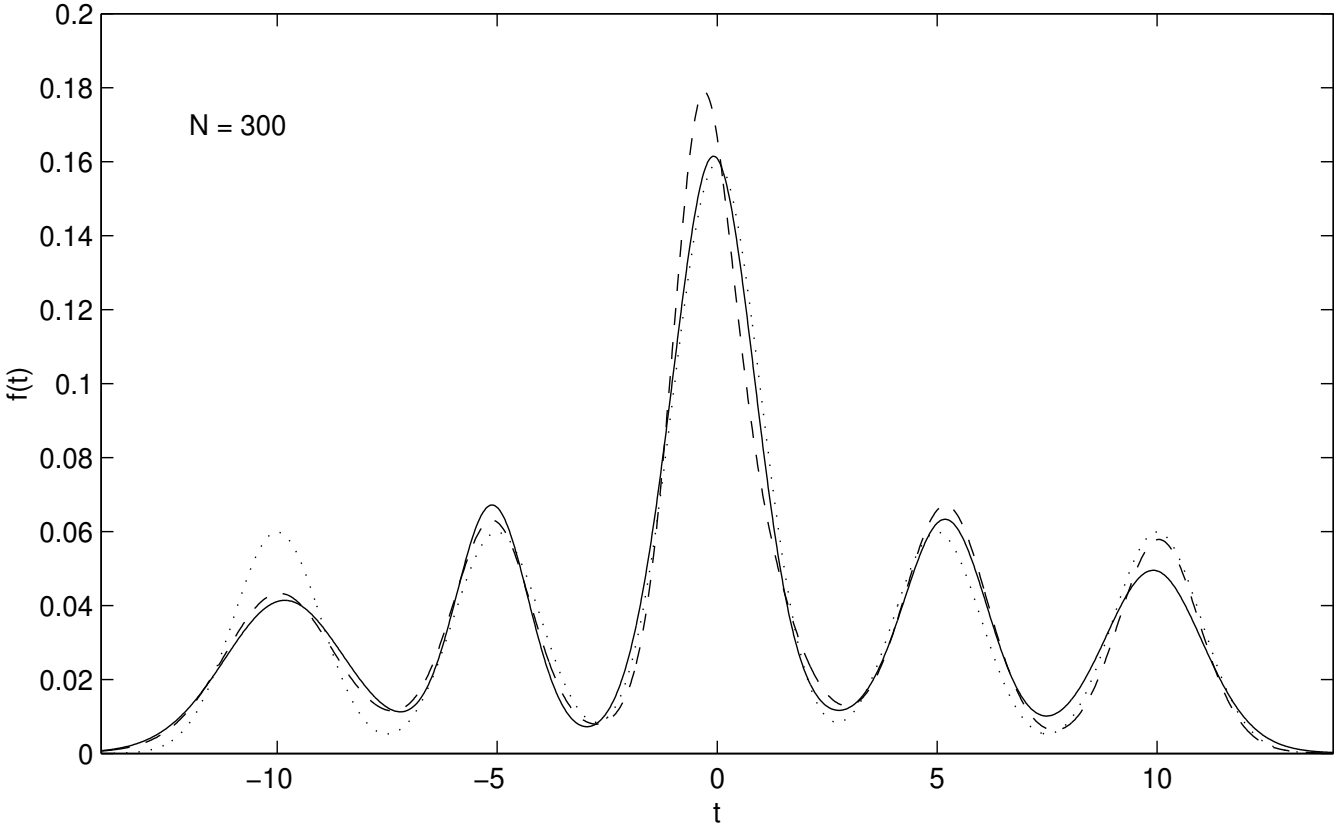


Fig 3.5 Data: Cauchy(0,1) & N(0,1) < ALPHA(R) = 1.0; M = 2000 >

Dash: N(0,1) −−− Solid: Cauchy(0,1)

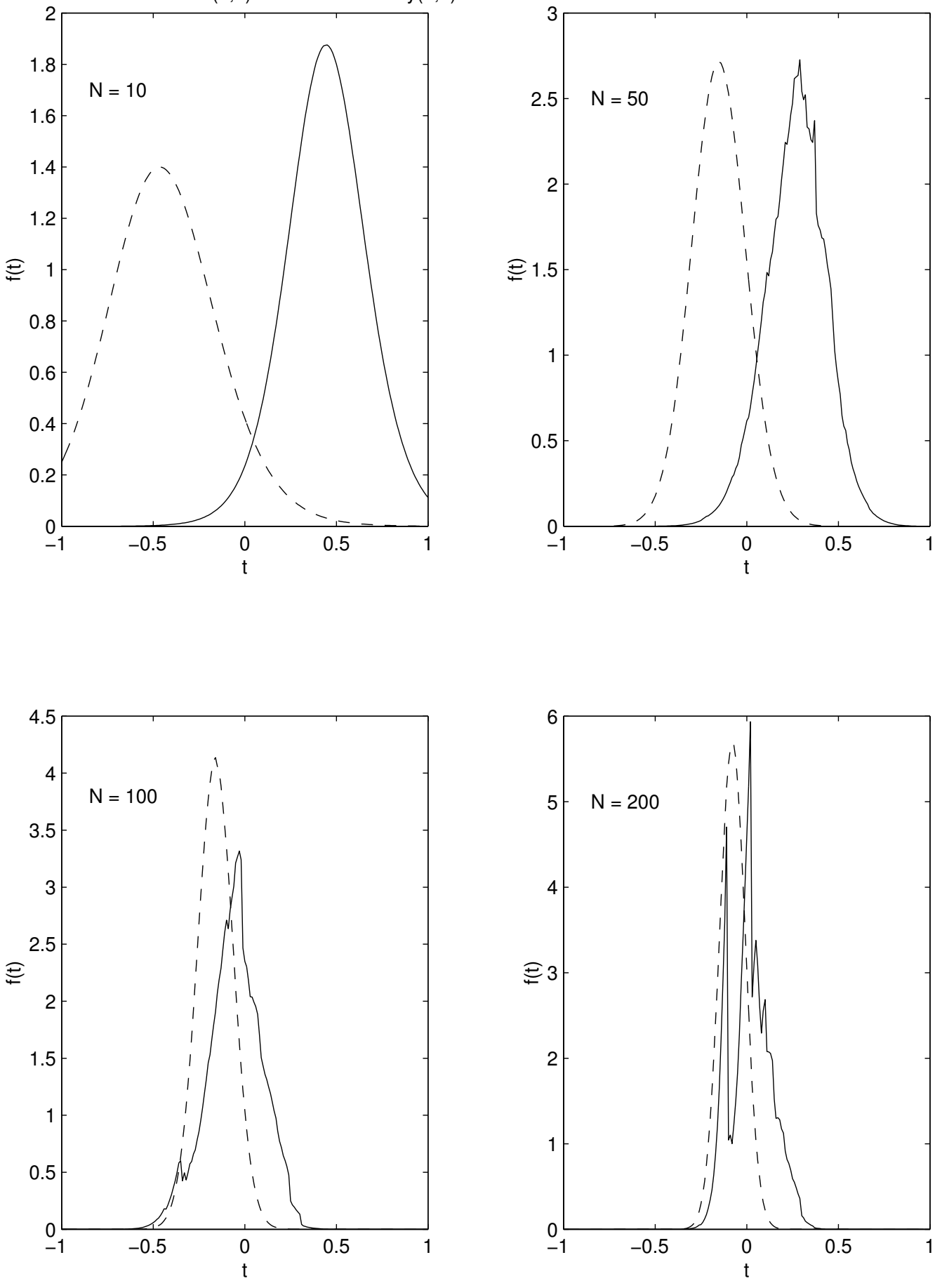
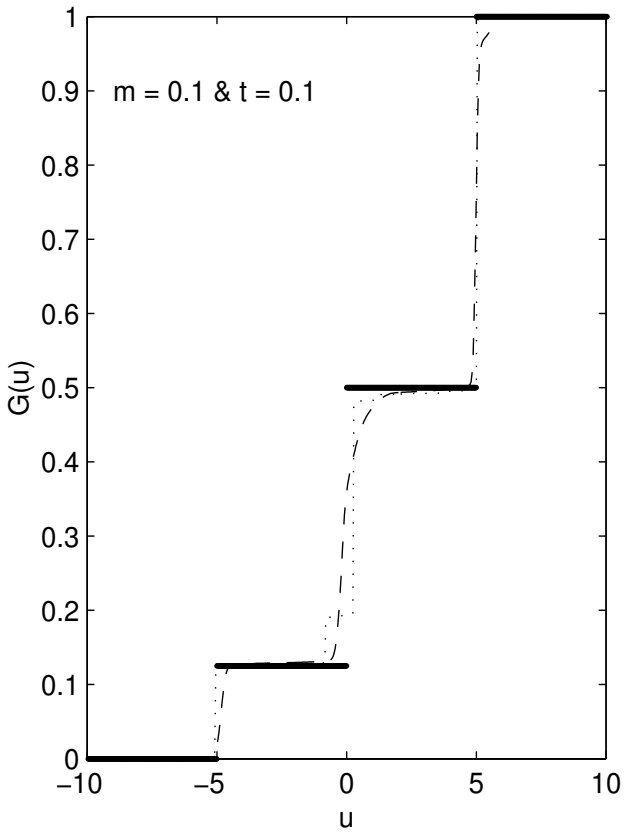
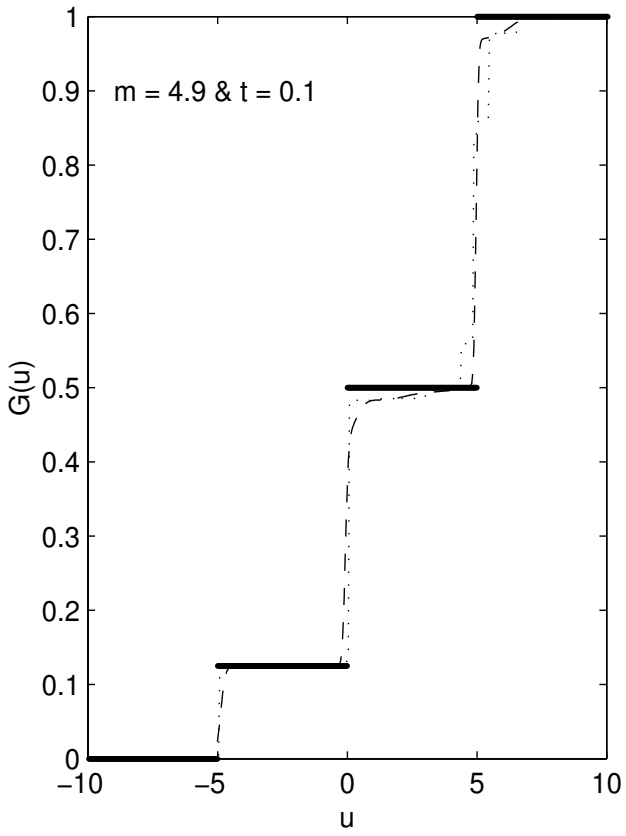
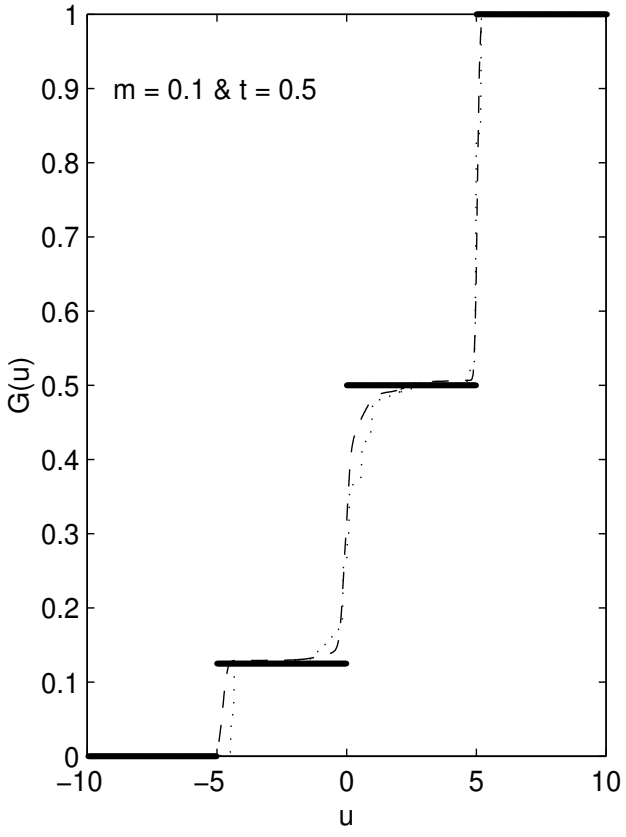


Fig 4.1 Data: 0.125N(−5,1)+0.375N(0,1)+0.5N(5,1) (N = 400)







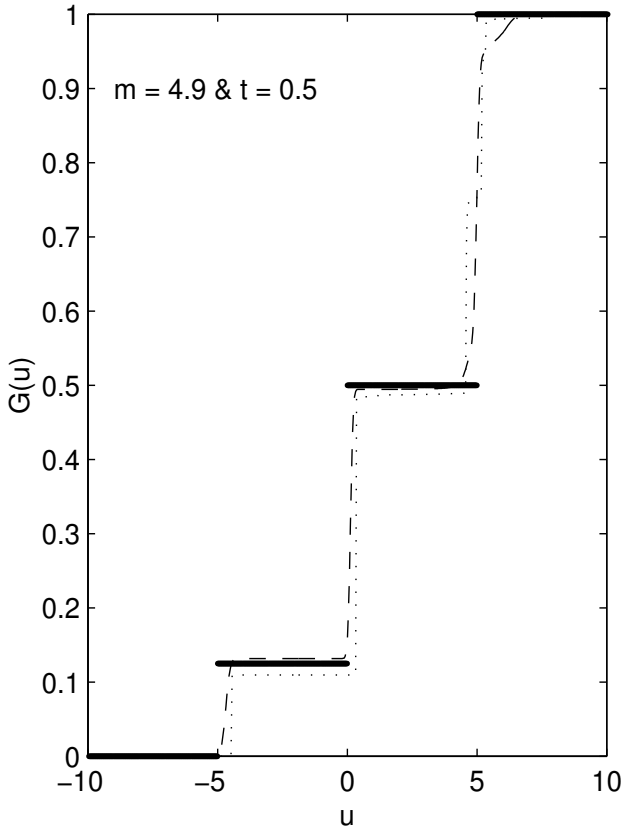


Fig 4.2 Location Model (N = 300)

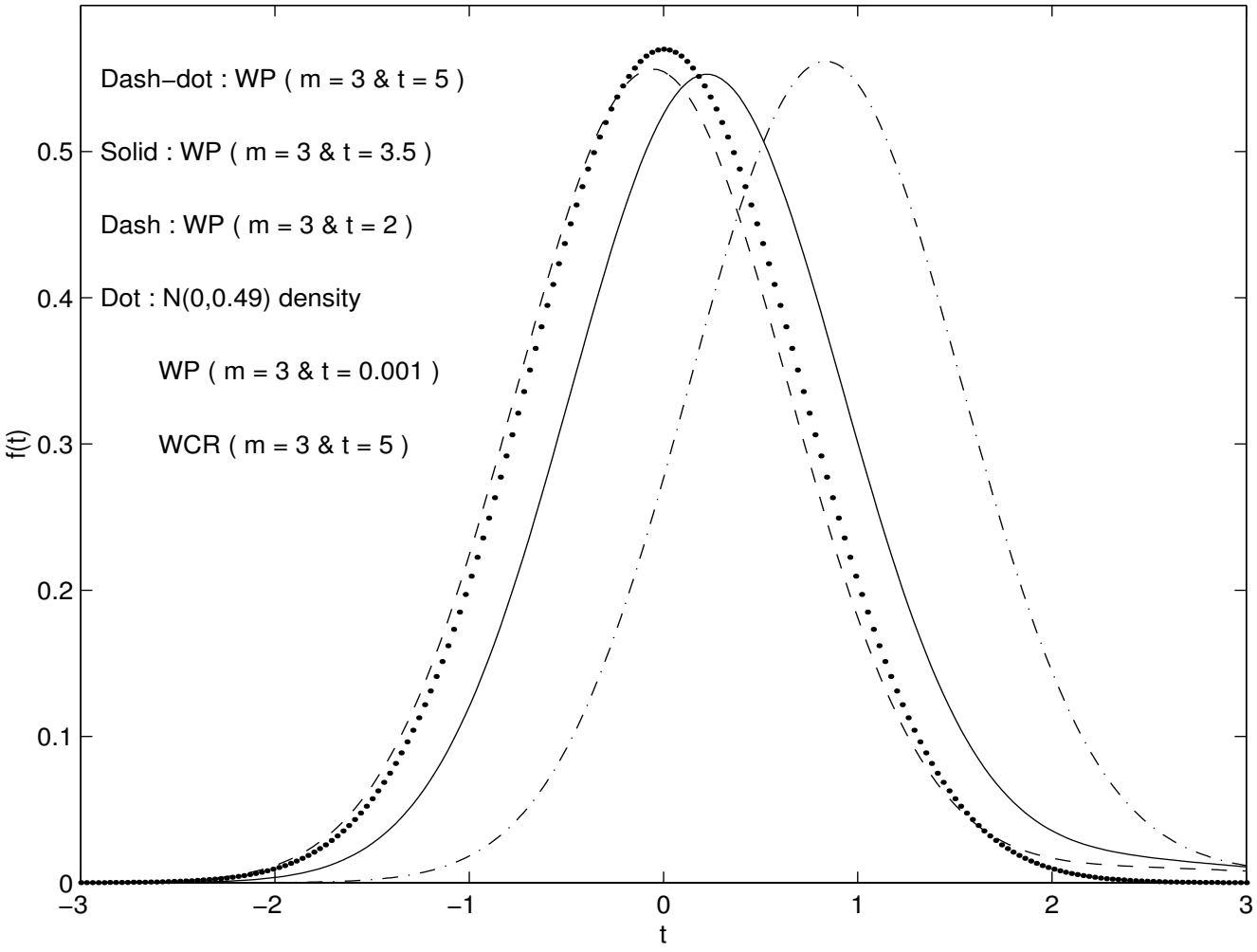


Fig 4.3a Location−scale Mixture Model (N = 300)

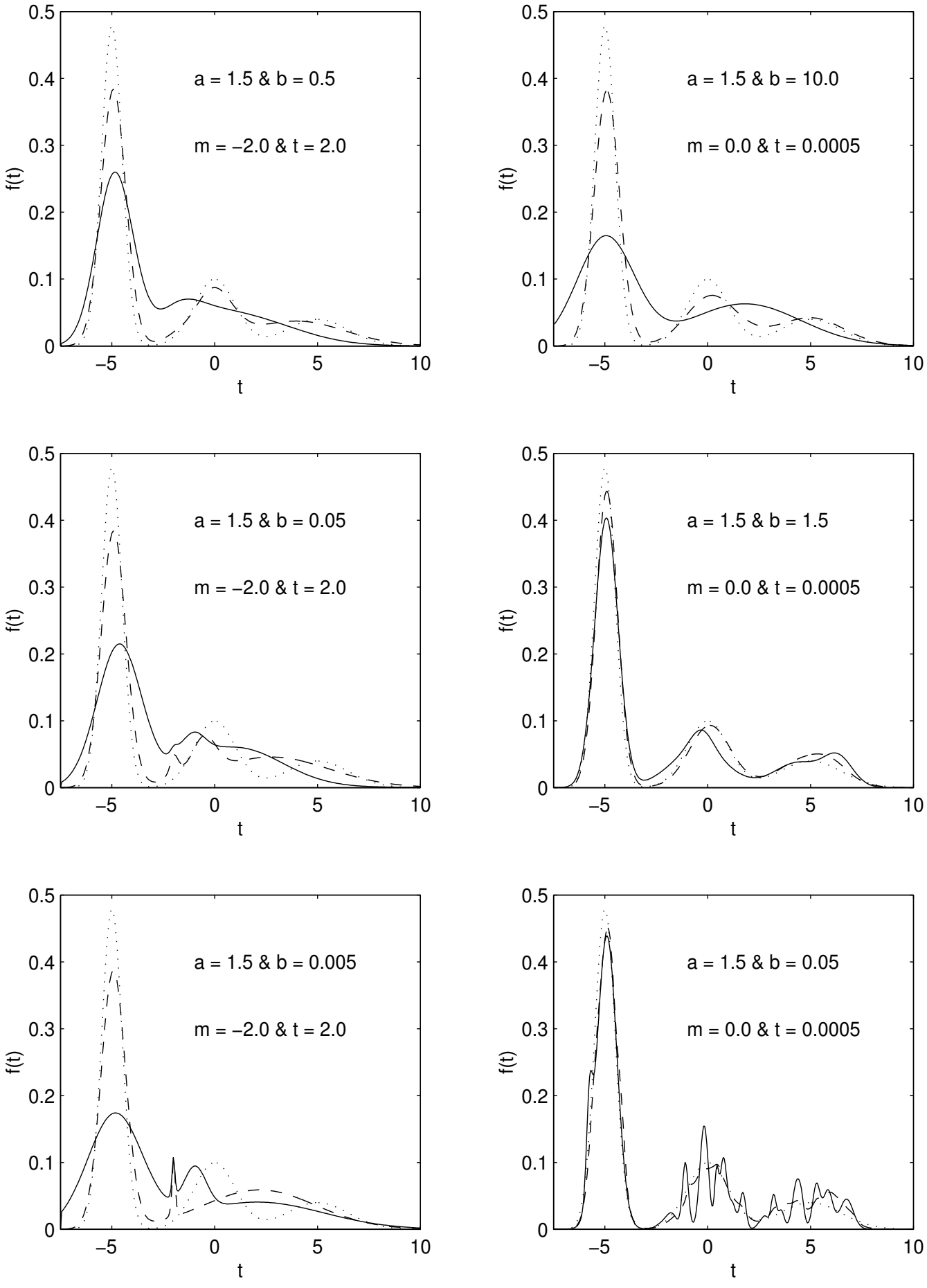


Fig 4.3b Location−scale Mixture Model (N = 200) [a=b=1.5,m=0,t=0.0005]

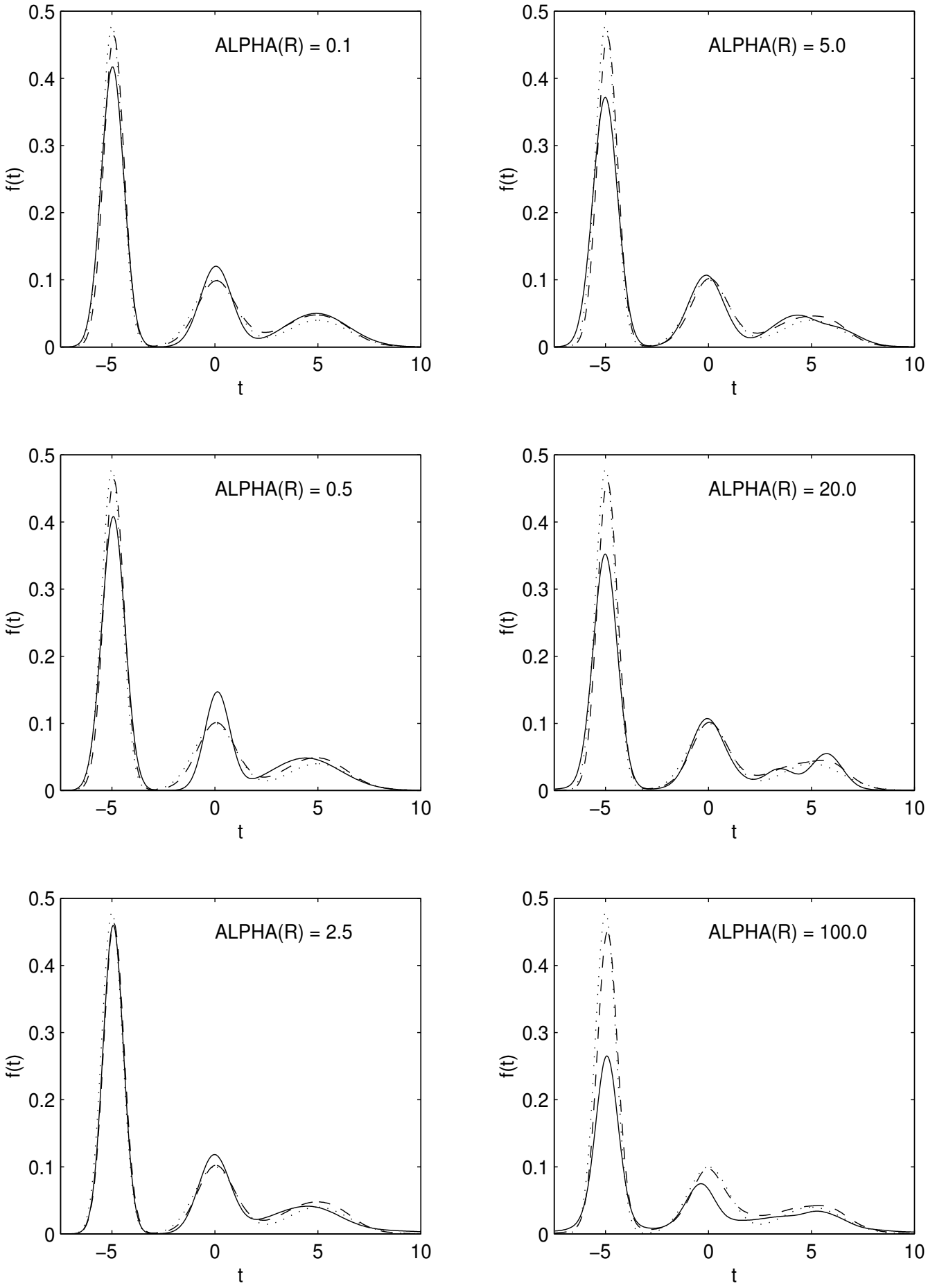
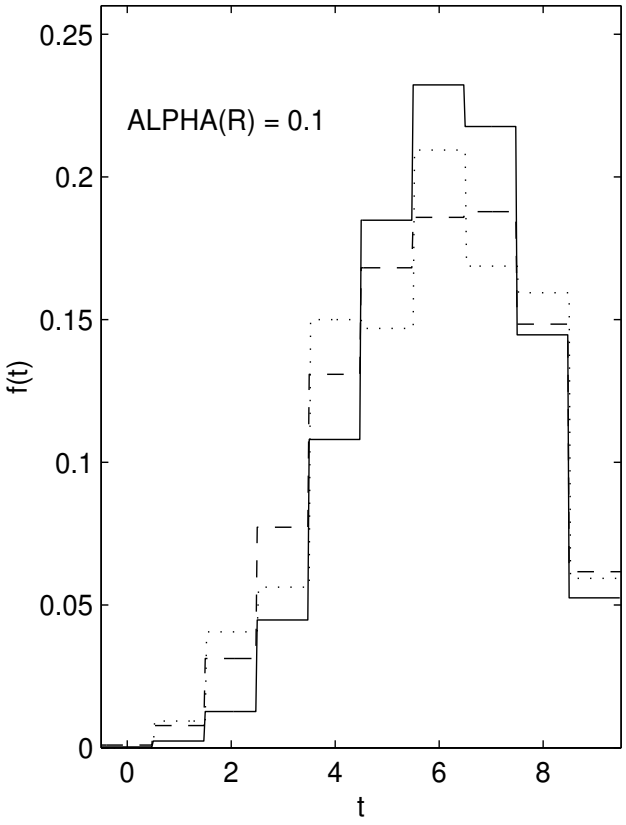
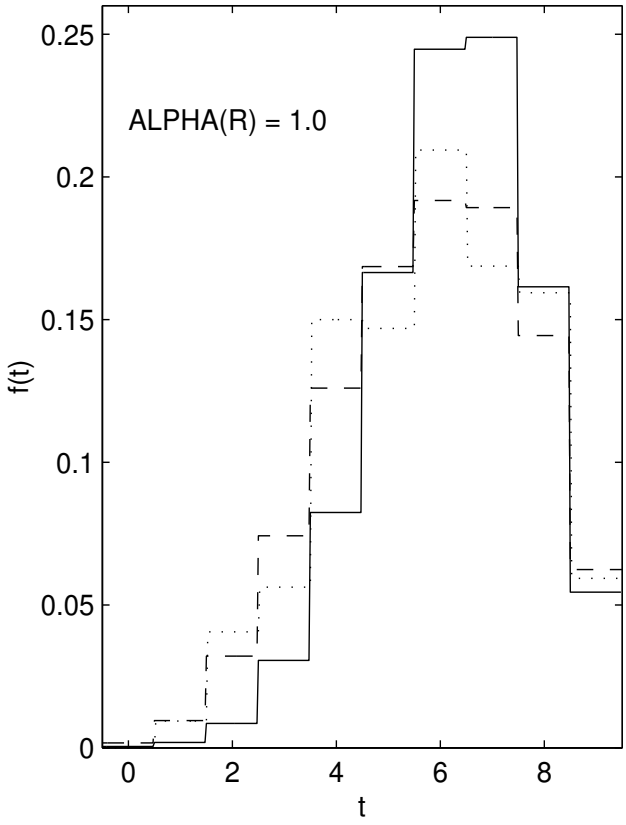
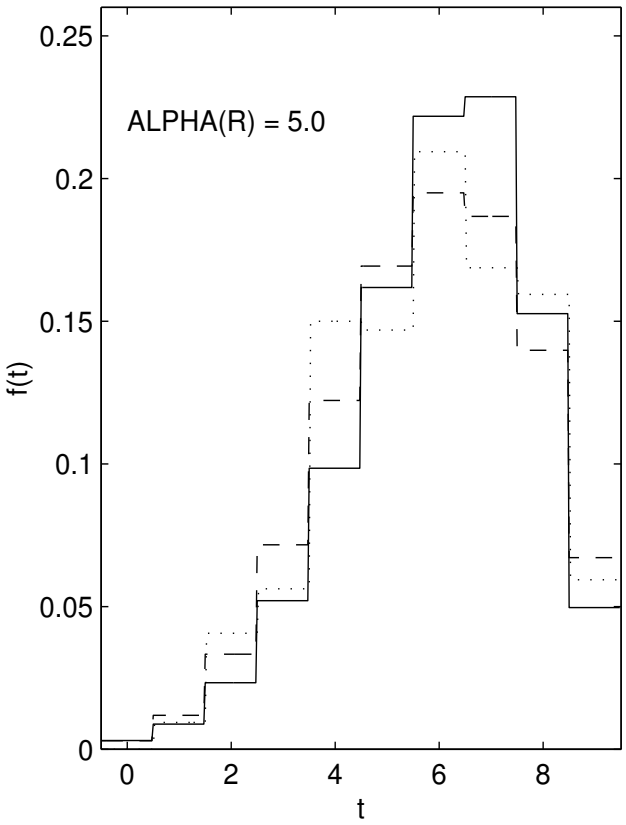


Fig 4.4a Binomial Kernel Bi(9,u) (N = 320)







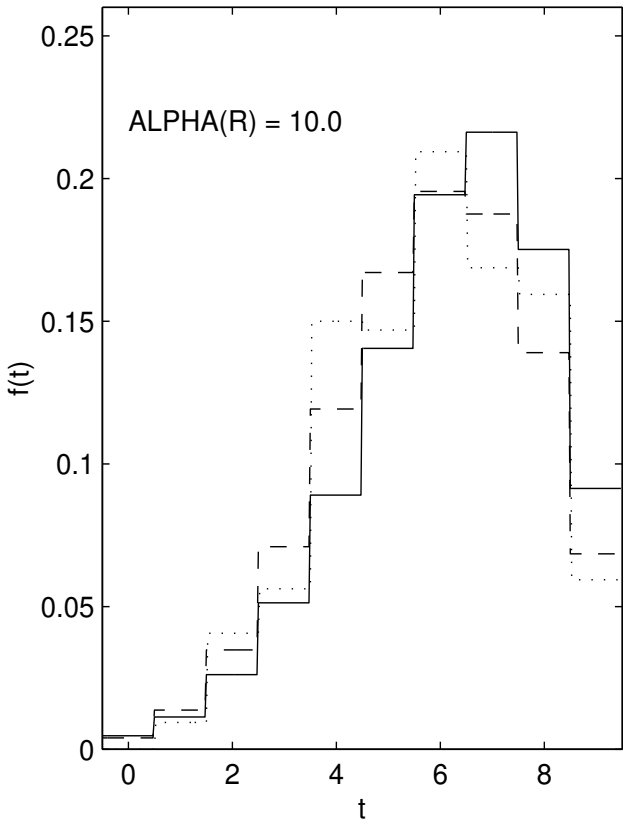
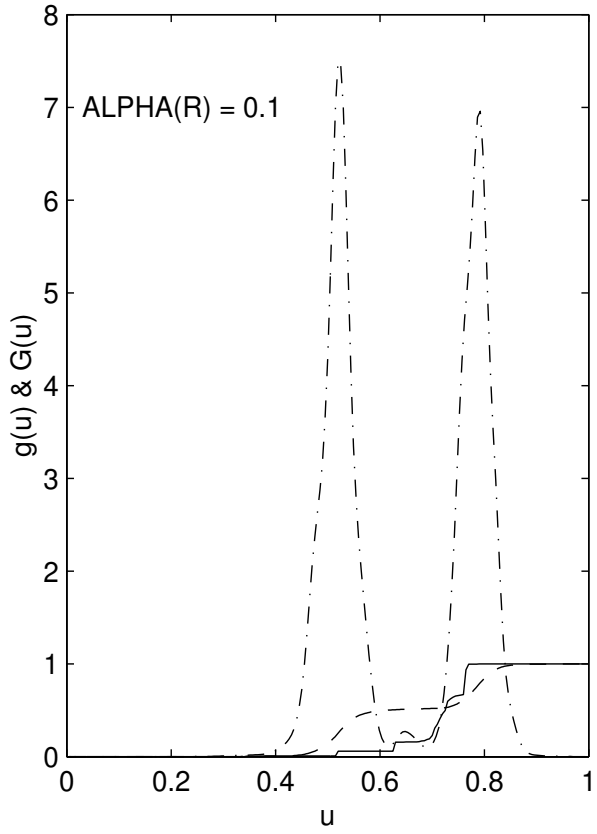
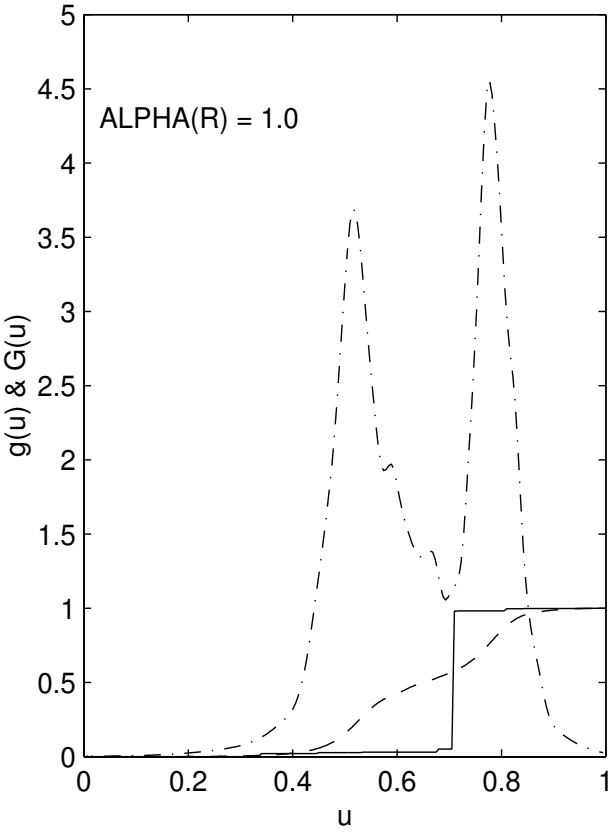
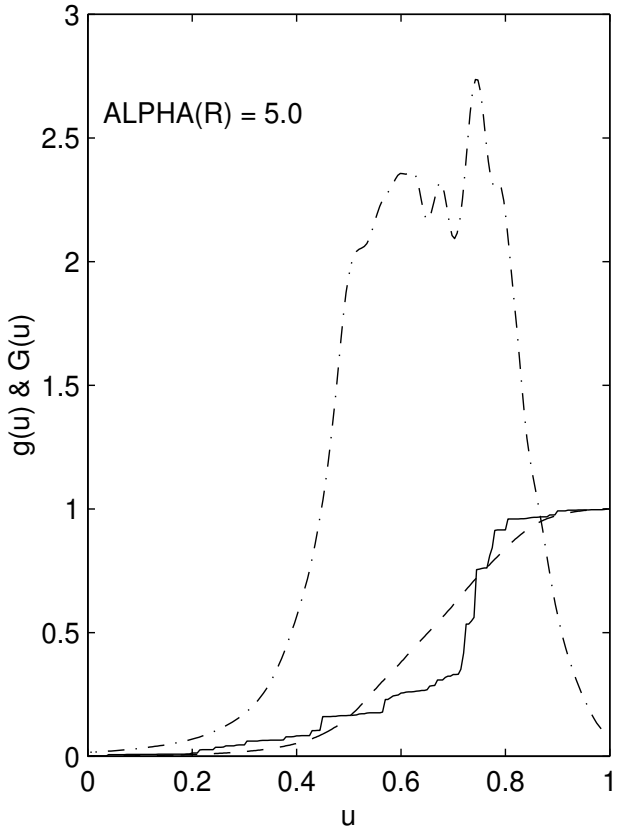


Fig 4.4b Binomial Kernel Bi(9,u) (N = 320)







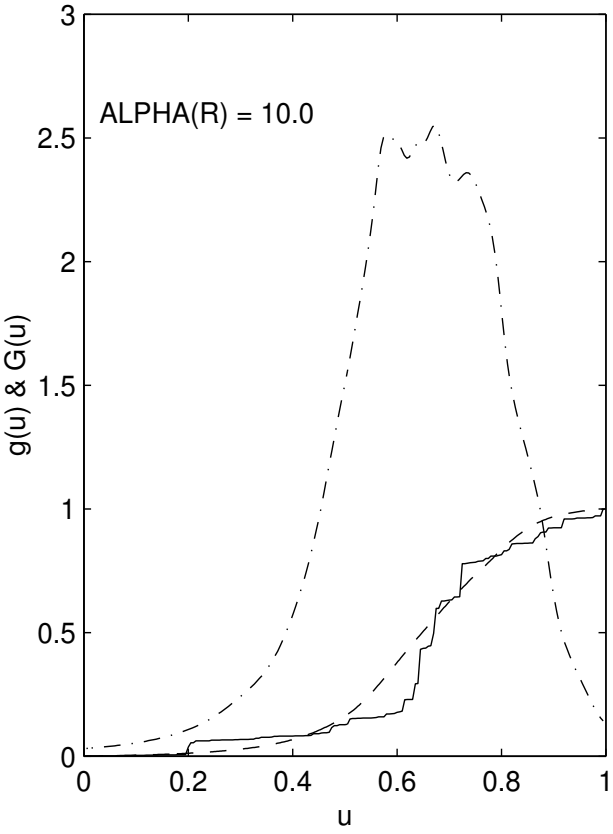


Fig 4.5 Location−scale Mixture Model (N = 300)