DRAWING A RANDOM SAMPLE FROM A DENSITY SELECTED AT RANDOM[1]

by

Jeesen Chen and Herman Rubin
University of Cincinnati and Purdue University

Technical Report #83-15

Department of Statistics
Purdue University

April 1983
(Revised April 1986)

[1]This research was partially supported by Army Research Office Contract #DAAG-29-80-K-0043.

ABSTRACT:

We provide a model to build up prior distributions over density functions, to select a density from that prior. We also indicate how to draw a random sample from density, which is arbitrary close to the randomly selected density, in a reasonable amount of time.

AMS 1980 subject classification. Primary 65U05, secondary 65C10

KEY WORDS: Prior distribution, orthonormal basis, random sample generation.

§1. Introduction.

One of the methods used in practice to compare density estimators is to generate random samples and compare the goodness of fit for those samples. Usually this is done for common distributions such as the normal, gamma, Cauchy, triangular, etc. However, if one believes the distribution to be of one of those forms, much better parametric procedures exist.

Since a motivation for "nonparametric" density estimation is that the density is not of a standard parametric form, it seems desirable to use a "typical" density of the type under consideration. Consequently we suggest that densities should be selected by a random process, and that estimators should be compared on random samples form those densities. If one density estimator performs consistently better than another for randomly selected densities, there is reason to believe it a better density estimator.

How can one choose a probability density at random? In this paper, we propose a method to represent a density function, which converts the above problem into choosing a point at random from the infinite-dimensional sphere. In Section 3, we show how this can be accomplished in practice for a specific model.

In Section 4, we give a general method to generate random numbers according to the density selected. We also give an example to illustrate how one can improve the efficiency of the general procedure by using local bounds for the selected density function.

§2. The Model.

Let $\{\phi_n(t) : n=1,2,\ldots\}$ be an orthonormal basis for the square integrable functions with respect to the measure μ, S be the set of sequences of real (or

complex) numbers with sum of absolute squares 1. For $A = (a_n : n=1,2,\ldots)$, we use $f_A(t)$ to denote $|\sum_n a_n\phi_n(t)|^2$. It is easy to see that

(I) $f_A(\cdot)$ is a probability density function with respect to μ if $A \in S$, and

(II) for any probability density function $f(\cdot)$, it is possible to find an $A \in S$ such that $f(t) = f_A(t)$ μ-almost everywhere.

In fact, a possible choice $A = (a_n : n=1,2,\ldots)$ in (II) is $a_n = \int\sqrt{f(t)}\,\phi_n(t)\mu(dt)$.

By the above scheme, we have transformed the problem of choosing a probability density at random (i.e. according to some distributions over the set of all probability density functions) to the problem of choosing a point from the infinite dimensional unit sphere S according to some distributions over S.

The following proposition shows that we are able to use a finite number of terms to represent a density function which is arbitrarily close to a preassigned density function. This is important for the purpose of simulation.

For $(a_n : n=1,2,\ldots) = A \in S$ and $0 < \varepsilon < 1$, let $T_\varepsilon(A)$ be the first j such that $\sum_{1\leq i\leq j} |a_i|^2 > 1-\varepsilon$.

Proposition 1. Suppose $g(t)$ is a given density function, and $A = (a_n : n=1,2,\ldots)$ where

$$a_n = \int\sqrt{g(t)}\,\overline{\phi}_n(t)\,\mu(dt)\ ,\ n = 1,2,\ldots\ .$$

Define $\tilde{A} = (\tilde{a}_n : n=1,2,\ldots)$ as

(1)
$$\tilde{a}_n = \begin{cases} a_n & \text{if } n \leq T_\varepsilon(A) \\ \alpha\{1-\sum_{i=1}^{n-1} |a_i|^2\}^{1/2} & \text{if } n = T_\varepsilon(A)+1 \\ 0 & \text{otherwise} \end{cases} .$$

where α is chosen so that the real part of $\tilde{a}_n \bar{a}_n$ is not greater than 0 and $|\alpha| = 1$. Then:

(a) the squared Kakutani distance between g and $f_{\tilde{A}}$,

(2)
$$K(g, f_{\tilde{A}}) = \int (|g(t)|^{1/2} - |f_{\tilde{A}}(t)|^{1/2})^2 \mu(dt) < 2\varepsilon ;$$

(b) the absolute deviation between g and $f_{\tilde{A}}$,

(3)
$$L_1(g, f_{\tilde{A}}) = \int |g(t) - f_{\tilde{A}}(t)| \mu(dt) < \sqrt{8\varepsilon} .$$

Proof.

From the definition of $\tilde{a}_n$, it is easy to see that $|a_n - \tilde{a}_n|^2 \leq |a_n|^2 + |\tilde{a}_n|^2$ for $n \geq T_\varepsilon(A) + 1$. Hence

$$\begin{aligned} K(g, f_{\tilde{A}}) &= \int (|\sum a_n \phi_n(t)| - |\sum \tilde{a}_n \phi_n(t)|)^2 \, dt \\ &\leq \int (|\sum (a_n - \tilde{a}_n)\phi_n(t)|)^2 \, dt \\ &= \sum_n |a_n - \tilde{a}_n|^2 \\ &= \sum_{n=T_\varepsilon(A)+1}^{\infty} |a_n - \tilde{a}_n|^2 \\ &\leq \sum_{n \geq T_\varepsilon(A)+1} |a_n|^2 + \sum_{n \geq T_\varepsilon(A)+1} |\tilde{a}_n|^2 \\ &< 2\varepsilon . \end{aligned}$$

For (b), we apply (a) and the following inequality:

$$(L_1(g,f_{\tilde{A}}))^2 \leq \{\int |g^{1/2}(t)-f_{\tilde{A}}^{1/2}(t)| \cdot |g^{1/2}(t)+f_{\tilde{A}}^{1/2}(t)| \mu(dt)\}^2$$

$$\leq \int |g^{1/2}(t)-f_{\tilde{A}}^{1/2}(t)|^2 \mu(dt) \cdot \int |g^{1/2}(t)+f_{\tilde{A}}^{1/2}(t)|^2 \mu(dt)$$

$$\leq K(g,f_{\tilde{A}}) \cdot \int 2(g(t)+f_{\tilde{A}}(t))\mu(dt)$$

$$= 4\, K(g,f_{\tilde{A}}) \ .$$

Q.E.D.

§3. An Example of a Method to Choose a Density Function at Random.

Suppose $R_1, R_2, \ldots$ is a sequence of random variables such that

$$1 = R_0 \geq R_1 \geq R_2 \geq \ldots \ , \tag{4}$$

and

$$R_n \to 0 \quad \text{almost surely} \ . \tag{5}$$

For any sequence of numbers with absolute value 1, $\{\theta_n\}$, if we define $A = (a_n : n=1,2,\ldots)$ as

$$a_n = \theta_n (R_{n-1}-R_n)^{1/2} \ , \tag{6}$$

then $A \in S$ almost surely. Therefore, $f_A(t) = |\sum a_n \phi_n(t)|^2$ is a density function randomly chosen according to a prior distribution.

One way to construct the R_n sequence is the following:

Let $U_1, U_2, \ldots$ denote a sequence of independent random variables having a nontrivial distribution on (0,1). (For example, we may chose U_i to be uniformly distributed over (0,1).) Define R_j as

(7) $$R_j = U_1 U_2 \cdots U_j .$$

It is easy to see that the sequence $\{R_j\}$ satisfies (4) and (5).

Notice that in order to fully determine a density function $f_A(\cdot)$, we need to get infinitely many random variables U_n. This is obviously impossible to achieve in a finite time. But the truncation modification procedure describe in the Proposition 1 give us a computable reasonable approximation to $f_A(\cdot)$. The number of U's we need to compute $f_A(\cdot)$ is $T_\varepsilon(A)$, which is usually not very large as shown in the following two propositions.

Proposition 2. If the U's are i.i.d., then

(8) $$\frac{\log \varepsilon}{E(\log U)} \leq E(T_\varepsilon(A)+1) \leq \frac{2 \log \varepsilon}{E(\max(\log U, \log \varepsilon))} .$$

Proof. Since $1-R_n = (1-R_1)+(R_1-R_2)+\ldots+(R_{n-1}-R_n) = \sum_{1\leq j\leq n} a_j^2$, we have

(9) $$\begin{aligned} \{T_\varepsilon(A)+1=n\} &= \{n \text{ is the first } j \text{ such that } R_j < \varepsilon\} \\ &= \{n \text{ is the first } j \text{ such that } \sum_{1\leq i\leq j} \log U_i \leq \log \varepsilon\}. \end{aligned}$$

The above equation implies that $T_\varepsilon(A)+1$ is a stopping time for the i.i.d. sequence $\{\log U_i, i=1,2,\ldots\}$ and the following inequality

(10) $$\sum_{1\leq i\leq T_\varepsilon(A)+1} \log U_i < \log \varepsilon \leq \sum_{1\leq i\leq T_\varepsilon(A)} \log U_i .$$

Applying the Wald equation to the first half of (10), we have

$$\log \varepsilon \geq E\Big(\sum_{1\leq i\leq T_\varepsilon(A)+1} \log U_i\Big) \tag{11}$$

$$= E(\log U)\, E(T_\varepsilon(A)+1) .$$

Since the second half of (10) implies

$$2 \log \varepsilon \leq \sum_{1\leq i\leq T_\varepsilon(A)} \log U_i + \log \varepsilon \tag{12}$$

$$\leq \sum_{1\leq i\leq T_\varepsilon(A)+1} \max(\log U_i, \log \varepsilon) ,$$

we have

$$2 \log \varepsilon \leq E \sum_{1\leq i\leq T_\varepsilon(A)+1} \max(\log U_i, \log \varepsilon) \tag{13}$$

$$= E(\max(\log U, \log \varepsilon))\, E(T_\varepsilon(A)+1).$$

Q.E.D.

Proposition 3. If the U's are i.i.d. Uniform (0,1) random variables, then $T_\varepsilon(A)-1$ has a Poisson distribution with mean $\lambda = -\log \varepsilon$.

Proof. This is a well-known result. See, for example, Karlin and Taylor (1981), page 128, problem 16.

For the above case, if we set $\varepsilon = 10^{-22}$, so the L_1 distance is less than 10^{-10}, then

$$E(T_\varepsilon(A)) = 51.66 , \tag{14}$$

$$\mathrm{var}(T_\varepsilon(A)) = 50.66 , \tag{15}$$

$$P(T_\varepsilon(A) \leq 26) < 0.488 \times 10^{-5} , \tag{16}$$

and

$$(17) \qquad P(T_\varepsilon(A) \geq 86) < 0.633 \times 10^{-5} .$$

§4. Generating a Random Sample According to a Randomly Selected Density: An Example.

Once we have density function f, we are able to generate random numbers according to this density function. A general method is the (von Neumann's) Acceptance-Rejection Procedure which is stated in Rubin (1976) as follows:

> To obtain a random variable whose distribution has density f (with respect to Lebesgue measure), one obtains a random variable Y whose density is b.g where $f \leq g$, and then set X = Y with probability f(Y)/g(Y).

(See also Ripley (1983), page 311.)

The acceptance rate of this procedure is b, i.e. on the average one need to generate 1/b number Y's to compute a X. If we want to generate a large sample, it is worthwhile to use better local bounds in the acceptance-rejection procedure. To illustrate this point, let us consider the following example:

Suppose we want to generate random sample according to density function

$$(18) \qquad f_{\tilde{A}}(t) = \{a(0) + \sum_{j=1}^{58} a(2j-1)\sqrt{2}\,\sin(2\pi jt) + a(2j)\sqrt{2}\,\cos(2\pi jt)\}^2 ,$$

where $\{a(i)\}$ are reported in Table 1. (The graph of $f_{\tilde{A}}(t)$ is in Figure 1. Although $f_{\tilde{A}}(t)$ has extremely small values around [0.375, 0.438], it has only two zeros.) This density function $f_{\tilde{A}}$ is a randomly selected density function according to the procedure we describe in Section 3, with orthonormal basis

$$\{\phi_n(t):n=1,2,\ldots\} = \{1,\sqrt{2}\sin(2\pi jt), \sqrt{2}\cos(2\pi jt) : j=1,2,\ldots\} ,$$

uniform (0,1) U's and randomly chosen sign $\{\theta_n\}$.

Since, for this orthonormal basis,

$$\{\sum \tilde{a}_n\phi_n(t)\}^2 \leq \{\sum|\tilde{a}_n|\}^2, \tag{19}$$

if we put $b = \{\sum|\tilde{a}_n|\}^{-2}$, we have

$$f_{\tilde{A}}(t) \leq \frac{1}{b}\, 1_{[0,1]}(t) = g(t) . \tag{20}$$

Hence, the following Acceptance-Rejection Procedure results X which has density function $f_{\tilde{A}}$ if the Y has density function $b\, g(t) = 1_{[0,1]}(t)$.

Acceptance-Rejection Procedure:

(AR1) Generate a $Y \sim U(0,1)$.

(AR2) Generate $U \sim U(0,1)$. If $\{\sum|\tilde{a}_n|\}^2 \cdot U > \{\sum\tilde{a}_n\phi_n(Y)\}^2$, GO TO (AR1).

(AR3) Return $X = Y$.

If we know the maximum value of $f_{\tilde{A}}(\cdot)$ is M, we could replace $\{\sum|\tilde{a}_n|\}^2$ in (AR2) by M. This modified Acceptance-Rejection Procedure has better acceptance rate, 1/M, than that of the Procedure (AR1) ~ (AR3).

Another method to increase the acceptance rate is to use better local bounds and to apply a variants of Acceptance-Rejection method, namely, the Squeeze Method. (See Marsaglia (1977) page 321 or Ripley (1983) page 312.)

Squeeze Method:

(S1) Break the unit interval into k subintervals of equal length $I_1, I_2, \ldots, I_k$. Compute the maximum, M_j, and the minimum, m_j, of $f_A^{\sim}(t)$ on each subinterval I_j.

(S2) Select the j-th interval I_j with probability $M_j / \sum_{i=1}^{k} M_i$.

(S3) Select X uniformly in I_j.

(S4) Generate $U \sim U(0,1)$. If $M_j U < m_j$, accept X and we can resue $M_j U/m_j$ as a uniform random variable independent of X. Else accept X exactly when $M_j U < f(X)$.

In the example, $f_A^{\sim}(t)$ defined in (18), we find that if we use k = 32 intervals, that to compute 1000 sample points, on the average 1087 X's (Step (S3)) will be generated and 173 densities evaluated. This compare with on the average 2767 Y's (Step (AR1)) will be generated in the modified Acceptance-Rejection Procedure.

§5. Conclusions and Remarks.

It is possible to build up prior distributions over density functions, to select a density according to that prior, and generate random numbers according to density function, which is arbitrary close to the density chosen at random, within reasonable amount of time.

There are many problems related to this simulation scheme. What properties do these randomly chosen densities $f_A(t)$ have? How to choose unimodal density function at random? How to randomly choose density function which have specific moments?

We are unable to answer all of the above questions, but the following facts are known to us:

(a) If the elements of the orthonormal basis are smooth functions and the sequence $(a_n : n=1,2,\ldots) \in S$ converges to zero sufficiently fast, the resulting density $(\sum a_n \phi_n(t))^2$ will be smooth.

(b) Given a density function $g(t)$ and a positive number δ, there is a positive number ε such that with positive probability that $L_1(g,f_A)$, $K(g,f_A)$, $L_1(g,f_{\tilde{A}})$ and $K(g,f_{\tilde{A}})$ are all less than δ. Here $A \in S$ is the random point described in Section 3 and $\tilde{A} \in S$ is the random point resulting from A and the truncation-modification procedure described in Proposition 1.

We shall report these results and another properties of $f_A(t)$ elsewhere.

Finally, we want to mention that Kraft (1964) has another method to randomly select a probability density function on the unit interval. However, these densities are discontinuous at all binary rationals almost surely.

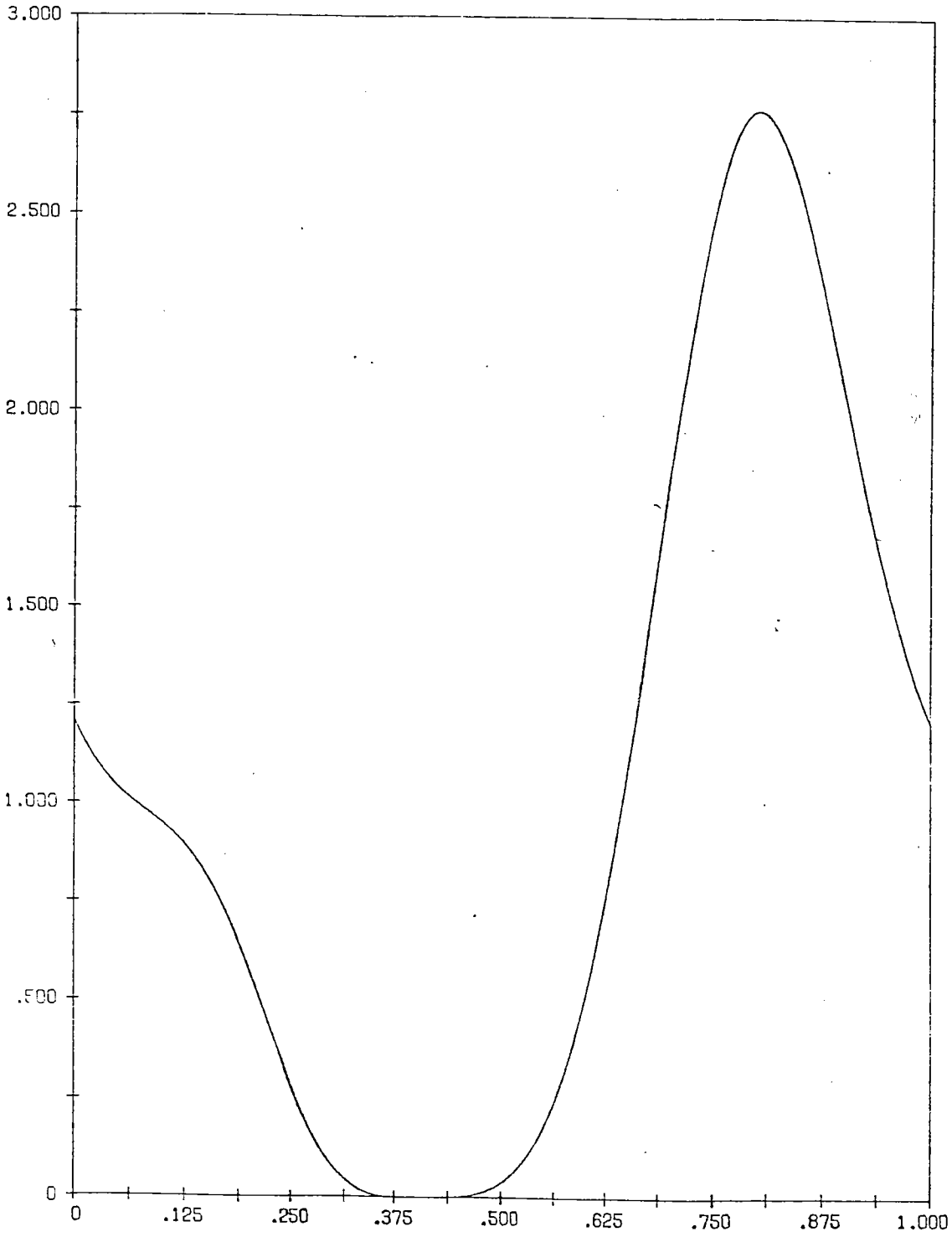
ACKNOWLEDGEMENT:

The authors would like to thank Dr. Paul Horn for his opinions about the presentation of this paper.

TABLE 1

i	a(i)	i	a(i)
0	8.4461134e-01		
1	-3.5837020e-01	2	3.6554103e-01
3	1.4423854e-01	4	-4.8704963e-02
5	3.6595802e-02	6	-7.0169992e-03
7	-3.3638465e-03	8	2.3260103e-03
9	1.7838685e-04	10	-2.0001775e-04
11	-7.2732561e-05	12	8.9707323e-05
13	-3.2776456e-05	14	2.0666291e-05
15	3.4747862e-05	16	-4.2741440e-06
17	-5.4365186e-06	18	6.5046789e-07
19	-4.7919963e-07	20	-1.0162950e-06
21	6.9635951e-07	22	1.1581500e-07
23	2.6266013e-07	24	4.3748485e-08
25	-1.1980375e-08	26	-2.4469099e-08
27	3.0901869e-09	28	5.2340562e-09
29	-1.6479048e-08	30	-1.3634036e-08
31	6.5788516e-09	32	-1.8500272e-09
33	1.1240625e-09	34	1.9715822e-09
35	-9.1657520e-10	36	-6.6356275e-10
37	1.8146884e-10	38	-1.6252322e-10
39	1.4994553e-10	40	-1.5135062e-10
41	1.4873387e-10	42	-5.9860760e-11
43	-7.1017821e-11	44	-1.5742200e-11
45	-2.5028370e-11	46	1.0976158e-11
47	-2.3798249e-11	48	-7.1806823e-12
49	-8.1359469e-13	50	-1.4062408e-12
51	2.5047494e-13	52	2.2632739e-13
53	2.5240195e-14	54	2.0613403e-14
55	-1.2514841e-14	56	8.1441427e-15
57	-3.9124416e-15	58	-1.1805710e-15
59	2.3080753e-15	60	-1.8432151e-15
61	-3.1181220e-16	62	-1.0039514e-15
63	-1.3554017e-15	64	-8.2923404e-16
65	3.4104557e-16	66	4.5893563e-16
67	-3.2739767e-16	68	-1.6183522e-16
69	-8.8326178e-17	70	2.2490170e-17
71	2.3272892e-17	72	5.8299995e-18
73	-5.1930393e-18	74	-1.5302493e-18
75	-4.5455540e-18	76	-6.5737401e-18
77	3.3693718e-18	78	2.4165115e-18
79	-1.9060753e-18	80	-1.1327650e-18
81	-8.6395760e-20	82	4.9009035e-19
83	2.6434637e-19	84	1.8145851e-19
85	4.7897024e-20	86	-1.8178549e-19
87	-4.6609308e-20	88	2.9336716e-20
89	1.8388663e-20	90	2.7382083e-20
91	-6.8909226e-20	92	1.0031774e-20
93	-8.8649134e-21	94	8.3491477e-21
95	-6.8955675e-22	96	-2.7963895e-21
97	1.7187625e-21	98	1.3820918e-21
99	-2.7333553e-22	100	-4.3345872e-22
101	5.7603077e-22	102	-1.3667905e-22
103	6.6515355e-23	104	-3.4873525e-23
105	-3.4498206e-23	106	-1.0584718e-23
107	-1.4207942e-23	108	-4.3143942e-24
109	-3.2761956e-24	110	1.6307769e-25
111	3.1245310e-25	112	-2.0749040e-25
113	5.8563529e-25	114	-1.4548586e-25
115	2.0857506e-25	116	-1.1968444e-25

FIGURE 1


3.000
2.500
2.000
1.500
1.000
.500
0
0
.125
.250
.375
.500
.625
.750
.875
1.000

BIBLIOGRAPHY

Karlin, S. and Taylor, H. M. (1981). A Second Course in Stochastic Processes, Academic Press, Inc., New York.

Kraft, Ch. H. (1964). A class of distribution function processes which has derivatives, J. Appl. Probability, 1, p. 385-388.

Marsaglia, G. (1977). The squeeze method for generating gamma variates. Comp. & Maths. with Appls., 3, 321-325.

Ripley, B. D. (1983). Computer generation of random variables: a tutorial. International Statistical Review, 51, 301-309.

Rubin, H. (1976). Some fast methods of generating random variables with preassigned distribution I. General acceptance - rejection procedures. Mimeograph Series #466, Purdue University.