The Philippine Statistician, 1998
Vol. 47, Nos. 1-4, pp. 1-10

A Stochastic Clustering Algorithm for Panel Data with Applications to Polling[1]

Daniel C. Bonzo[2]

ABSTRACT

Political parties commission tracking studies during election years to benchmark the 'winnability' of its slate. Data coming from such polls constitute a panel of index measurements which can be used as inputs in formulating campaign strategies. A non-hierarchical approach to cluster analysis can be used to segment candidates into safe, within striking distance and losing groups using a panel of index measurements. Such grouping information can be used in designing strategies and allocating resources.

In this paper, a stochastic version of the Kosmelj and Batagelj (1990) approach to clustering panel data is presented. The method makes use of a probability link function in defining cluster inertias with the aim of preserving the clusters' probabilistic structure. Given a fixed number of clusters, the optimal constitution of a set of clusters is determined through the application of a variant of the relocation-swapping algorithm. This variant makes use of adaptive simulated annealing in minimizing an objective function defined by the cluster inertias. Annealing ensures location of optimal points using a stochastic search that terminates with probability one.

Keywords and Phrases: non-hierarchical cluster analysis, polling, panel data, annealing, probability link function

1. Introduction

Cluster analysis is one of the most if not the most popularly used multivariate technique in applications. This popularity stems from its distribution free nature and the variety of problems that it treats. Recently, it has been utilized as a pattern recognition tool in projection pursuit problems and unsupervised classification problems and is used as a benchmark in neural network applications such as credit scoring. Aside from these popular applications, the technique has been found useful in polling (see Bonzo, 1995). The application consists of forming clusters of competing candidates with the purpose of forming strategic decisions such as campaign strategies and allocation of resources. Such information is of utmost importance especially in a highly centralized campaign strategy such as the ones being waged by big political parties, the objective of which is to position as many of its candidates in the winners' column.

In a typical election year, big political parties usually commission survey organizations to track the 'winnability' of their slate of candidates, especially if the position is national in scope such as in senatorial elections. Data coming from such tracking studies are of the cross-sectional and time series type, i.e, panel where each candidate's 'winnability' is measured through time. Hence, the usual clustering procedures have to be modified so as not to ignore the effect of time. 'Winnability' here is measured in terms of a performance indices which can be a function of the candidate's rating, i.e., the proportion of voters who are willing to vote for a candidate. These indices differ per voting region and can change per tracking period. Hence a candidate's data

---

[1] This paper was presented during the first quarter meeting of the Philippine Statistical Association, PSSC Bldg., Commonwealth Ave., Quezon City and the 25th Annual Conference of the Mathematical Society of the Philippines.

[2] Associate Professor of Statistics, UP School of Statistics

profile is multivariate in nature and indexed by time. Based on these data, one can draw a list of 'safe' candidates, i.e., assured of winning, and another set who are within striking distance. If a party's candidates are within striking distance then they can throw appropriate resources to increase the chance of the candidates to win come election time. Such will entail identifying and working in geographic areas where the candidates' performance indices are weak.

In his application of cluster analysis to poll data panels, the author used the classificaton approach of Kosmelj and Batagelj (1990). The approach is non-hierarchical and made use of a maximal relocation-swapping algorithm to optimize cluster inertias. The cluster analysis output was then used as a feed in determining an odds of winning scale for candidates using generalized linear models (GLIM) utilizing the logistic link function.

In this paper the clustering approach is modified further to incorporate stochastic features. Instead of using plain distance functions, probability link functions are utilized as indicators of object similarities. The advantage of such an approach is documented by Bacelar-Nicolau and Nicolau (1997), especially in cases where preservation of the probabilistic structures of clusters is important. The clustering problem is then presented as an optimization problem with the objective function recast to allow the application of stochastic optimization techniques such as adaptive simulated annealing (ASA). This technique has the advantage of ensuring that global optima will be achieved using some statistical measures, i.e., convergence is attained with probability one.

The plan of the paper is as follows. In section 2, the longitudinal approach to clustering is discussed with specific concentration on probability link functions. Section 3 presents a brief account of annealing methods and their algorithmic construction. Section 4 presents a modification of the maximal relocation-swapping using ASA. Finally, some issues and concerns regarding the application of the technique are discussed in section 5.

2. Longitudinal Approach to Classification

The following approach was first used by Kosmelj and Batagelj . The approach has the advantage of flexibility, specifically in the definition of the object weights and time weights and is applicable even for short observation times. However, instead of using distances to indicate similarity or dissimilarity, probability link functions are used.

Consider index measurements X_{jti} (in p dimension) of object j at time t_i, $j = 1, \ldots, n$, $i = 1, 2, \ldots, T$, where the ordering property

$$t_1 < t_2 < \ldots < t_T$$

holds. Denote by X the corresponding data matrix of the index measurements. Let C be a partition of X into c clusters, i.e.,

$$C = C_1 \cup C_2 \cup \ldots \cup Cc$$

and the collection of such clusters as C.

Corresponding to the index measurements, define auxiliary measurements Z_{jk} as

$$Z_{jk} = \begin{cases} 1, & j \in C_k \\ 0, & j \notin C_k \end{cases}, \text{k} = 1, \ldots, \text{c}$$

The number of units in cluster k is then denoted by $Y_k = \sum_{j=1}^{n} Z_{jk}$. Also denote by Z the data matrix of the auxiliary variables. Denote by $Z_{\cdot k}$ the columns of Z, representing cluster information and Z_j. The rows of Z represent object information. Thus, $Y_k = 1_n'Z_{\cdot k}$, $Y = 1_n'Z$ and $C_k = Z_{\cdot k}$. Hence, the collection C is defined by

$$C = \{ Z : \sum_{k=1}^{c} Z_{jk} = 1, \sum_{j=1}^{n}\sum_{k=1}^{c} Z_{jk} = n \}$$

Define the inertia of the kth cluster as

$$f_k(Y_k(Z_{\cdot k})) = \sum_{j1<j2} w_k(j_1) w_k(j_2) \delta_k(j_1, j_2) / w_k(Y_k) \qquad (1)$$

where $w_k(j)$ is the weight unit j and is defined as

$$w_k(j) = w_j Z_{jk}$$

and $w_k(Y_k) = \sum_{j=1}^{n} w_k(j)$. The term $\delta(j_1, j_2)$ is called the validation link between unit j_1 and j_2 in a probability scale. It is sometimes called the linkage validity coefficient and can be interpreted as a similarity probabilistic coefficient. Objects with equal pairwise link coefficients are said to belong to the same cluster. In elementary applications, this similarity coefficient can be used as an input to hierarchical procedures in lieu of distance coefficients.

The allocation problem looks at the optimization of the cluster inertia as given in (1). Analytically, one is interested in minimizing the Lagrangian equation

$$\varphi(Z,\lambda) = f(Y(Z)) - \sum_{j=1}^{n} \lambda_j (1' Z_{j\cdot} - 1) - \lambda_{n+1} (\sum_{j=1}^{n}\sum_{k=1}^{c} Z_{jk} - n) \qquad (2)$$

with respect to Z, $Z \in C$ and λ where f(Y(Z)) is the sum of the cluster inertias given in (1). The optimization can move along (Y(Z), λ) and then along Z as the chain rule implies to represent relocation (and therefore the number of objects in a cluster) and swapping (and therefore cluster membership), respectively. Details of the relocation and swapping algorithm are given in Bonzo(1995).

For panel data applications, the coefficient is taken as

$$\delta_k(j_1, j_2) = \sum_{i=1}^{T} \alpha_{ti} P(\Delta_k(X_{j1ti}, X_{j2ti}) \le \Delta_k(x_{j1ti}, x_{j2ti})) \qquad (3)$$

where the x_{jt}'s are the observed values. Δ_k is taken as the Mahalanobis distance with respect to the centroid and dispersion of cluster k, i.e.,

$$\Delta_k(X_{j1ti}, X_{j2ti}) = (X_{j1ti} - \mu_{kti})' \Sigma_{kti}^{-1} (X_{j1ti} - \mu_{kti}). \quad (4)$$

Given data observations, δ_k can be estimated by its sample version d_k where

$$d_k(j_1, j_2) = \sum_{i=1}^{T} \alpha_{ti} P(\hat{\Delta}_k(X_{j1ti}, X_{j2ti}) \le \hat{\Delta}_k(x_{j1ti}, x_{j2ti}))$$

where

$$\hat{\Delta}_k(X_{j1ti}, X_{j2ti}) = (X_{j1ti} - \bar{X}_{kti})' S_{kti}^{-1} (X_{j1ti} - \bar{X}_{kti}).$$

and the sample mean and covariances are computed from the objects in C_k. Given appropriate regularity conditions, $\hat{\Delta}_k \xrightarrow{d} T^2$, where Δ_k is distributed $\chi^2(p)$. Also, $|d_k(j_1, j_2) - \delta_k(j_1, j_2)| \xrightarrow{a.s.} 0$. Hence as the constitution of C_k increases, the probabilistic structure of clusters remains intact. This is concretized in the following theorems:

Theorem 1. Suppose that X_{1ti}, $X_{2ti} \in C_k$ and are jointly normal with identical mean μ and variance Σ, with $\text{cov}(X_{1ti}, X_{2ti}) = \Sigma$, then $\hat{\Delta}_k(X_{1ti}, X_{2ti}) \to_d \Delta_k(X_{1ti}, X_{2ti}) = \chi^2(p)$ as $Y_k \to \infty$.

Proof: For simplicity, the index t_i is dropped in the following proof. Suppose X_1 and X_2 are jointly normal with

$$E(X_1) = E(X_2) = \mu$$

$$V(X_1) = V(X_2) = \text{cov}(X_1, X_2) = \Sigma$$

Then if we define $Y_i = \Sigma^{-1/2}(X_i - \mu)$, i = 1, 2, we have $Y_1' Y_2 = \frac{1}{2} Y' A Y$

where

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \quad \text{and} \quad A = \begin{pmatrix} 0 & I_p \\ I_p & 0 \end{pmatrix}.$$

Note that $Y \sim N(\nu, B)$ where ν and B are defined as

$$\nu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ and } B = \begin{pmatrix} I_p & I_p \\ I_p & I_p \end{pmatrix}.$$

Note that $\frac{1}{2} AB$ is idempotent and $(X_1 - \mu)' \Sigma^{-1} (X_2 - \mu) \sim \chi^2(p)$.

Now, we estimate μ by $\bar{X}$ and Σ by S. Thus by WLLN, $\bar{X} \xrightarrow[P]{} \mu$ and $S \xrightarrow[P]{} \Sigma$. If we let t $= 1_P$ and $T = 1_P 1_P'$ we have

$$(X_1 - \bar{X})' S^{-1} (X_2 - \bar{X}) = (X_1 - \mu)' (\Sigma^{-1} + o_P(1) T)(X_2 - \mu) + (X_1 - \mu)' (\Sigma^{-1} + o_P(1) T) t\, o_P(1)$$
$$+ o_P(1)\, t' (\Sigma^{-1} + o_P(1) T)(X_2 - \mu) + o_P(1)\, t' (\Sigma^{-1} + o_P(1) T) t\, o_P(1)$$
$$= (X_1 - \mu)' \Sigma^{-1} (X_2 - \mu) + o_P(1)$$

Thus, $(X_1 - \bar{X})' S^{-1} (X_2 - \bar{X}) \rightarrow_P (X_1 - \mu)' \Sigma^{-1} (X_2 - \mu)$ which implies that $(X_1 - \bar{X})' S^{-1} (X_2 - \bar{X})$ $\rightarrow_d \chi^2(p)$. Hence, $\hat{\Delta}_k(X_{1ti}, X_{2ti}) \rightarrow_d \Delta_k(X_{1ti}, X_{2ti}) = \chi^2(p)$ as $Y_k \rightarrow \infty$.

In Theorem 1, the constraining assumption is the joint normality of the data. With large sample sizes this assumption can be dropped and the result will still hold, i.e., $\hat{\Delta}_k(X_{1ti}, X_{2ti}) \rightarrow_d \Delta_k(X_{1ti}, X_{2ti})$ as $Y_k \rightarrow \infty$.

Theorem 2. Let d_k be the 'estimator' of δ_k with $\hat{\Delta}_k(X_{1ti}, X_{2ti})$ in dk 'estimating' $\Delta_k(X_{1ti}, X_{2ti})$ in δ_k. Then $| d_k(j_1, j_2) - \delta_k(j_1, j_2) | \rightarrow_{a.s.} 0$. For $j_1, j_2, \in C_k$.

Proof: Consider

$$| d_k(j_1, j_2) - \delta_k(j_1, j_2) | \leq \sum_{i=1}^{T} \alpha_{ti} |F_{Yk} - F|$$

where

$$F_{Yk} = P(\hat{\Delta}_k(X_{j1ti}, X_{j2ti}) \leq \hat{\Delta}_k(x_{j1ti}, x_{j2ti}))$$
$$F = P(\Delta_k(X_{j1ti}, X_{j2ti}) \leq \Delta_k(x_{j1ti}, x_{j2ti}))$$

Let $\tilde{F}_{Yk}$ the empirical distribution function of $\Delta_k(X_{1ti}, X_{2ti})$ constructed from pairs (x_{j1ti}, x_{j2ti}) in cluster C_k. Thus,

$$\tilde{F}_{Yk} - F = o_P(1) \text{ and } \quad F_{Yk} - F = o_P(1)$$

together imply that $F_{Yk} - \tilde{F}_{Yk} = o_P(1)$. Thus,

$$\sup_{t \in R} |F_{Yk}(t) - F(t)| = \sup_{t \in R} |\tilde{F}_{Yk}(t) - F(t)| + o_P(1).$$

From the Glivenko-Cantelli theorem, $\sup_{t \in R} |\tilde{F}_{Yk}(t) - F(t)| \rightarrow_{a.s.} 0$. Hence, $\sup_{t \in R} |F_{Yk}(t) - F(t)|$ $\rightarrow_{a.s.} 0$ and thus the result.

3. Simulated Annealing and Its Variants

The optimization technique known as simulated annealing (SA) borrowed concepts from the physical process known as annealing. In order to avoid metastable states produced by quenching, metals are often cooled very slowly. This allows the metals to order themselves into stable, structurally strong, low energy configurations. Annealing then gives the opportunity to

jump out of local minima with a reasonable probability while the temperature is still relatively high.

Simulated annealing has been successfully applied to optimization problems involving integers and has been reformulated to suit general purpose optimization problems. The mathematical interpretation of annealing is a cooling schedule in a decreasing function of time. It has been shown that SA is a random walk with bias.

As described by Ingber (1993), SA consists of three functional relationships: g(y), the probability density of a state space of D parameters Y = { Y(i), i = 1, . . . , D}; h(ΔE), the probability for acceptance of a new cost function given in the previous value; T(l), the schedule of 'annealing' the 'temperature' T in annealing time steps l, i.e., changing the volatility or fluctuations of one or both of the two previous probability densities.
One variant of the SA is called the Boltzmann annealing (BA). The method utilizes the Gibbs distribution

$$G(y) = \frac{1}{Z}\exp\{-H(y)/T\}$$

where

$$Z = \sum_{y}\exp\{-H(y)\}.$$

H is a Hamiltonian operator whose derivative defines the acceptance of the new cost function in defining the probabilities of the state parameters Y. The Hamiltonian of a system specifies its total energy, i.e., the sum of the kinetic energy and its potential energy, in terms of the Lagrangian function derived in earlier studies of dynamics and of the position and momentum of each of the particles. The scheduling is made such that T is selected no faster than

$$T(l) = \frac{T_0}{\ln(l)} \qquad (5)$$

where T_0 is large enough.

In optimizing f (minima), a typical algorithm for a continuous nonlinear annealing with a finite termination is given as follows:

0. Initialize (T, y) and set l = 0.
1. Generate another vector w from y (in its neighborhood).
2. If f(w) - f(y) ≤ 0, accept w as the new state.
3. Else, if $e^{-(f(w)-f(y))/T(l)}$ > u (generated from U(0,1)), then accept w as the new state and set y = w.
4. Lower T(l) to T(l+1) and increment l.
5. If l ≤ L, go to step 1.
6. Report all points y for which the best function value is achieved.

In the above algorithm, step 5 guarantees termination. In the absence of 5, termination is assured with probability 1. Satisfaction of the condition in step 2 means that a better value for f is achieved. Step 3 (annealing step) specifies the acceptance criterion and note that the structure

is of the Gibbs type with h(Δy) = f(w) - f(y). In this case, the probability of acceptance is $e^{-h(\Delta y)/T(l)}$. Increasing T(l) in step 4 has the effect of increasing the acceptance probability in step 3 and thus guaranteeing convergence.

In some instances the convergence may be painfully slow and therefore a revision of the annealing step may be necessary. Such a revision involves the schedule T(l) to speed up the search of the optimum value. This procedure is known as quenching. The procedure in (5) can be modified to allow a logarithmic schedule, i.e.,

$$T(l) = T_0 \frac{\ln(l_0)}{\ln(l)}$$

where l_0 is a starting index, or an exponential schedule such as

$$T(l) = T_0 \exp\{(c-1)l\}, \; 0 < c < 1$$

for expediency sake.

Adaptive simulated annealing (ASA) considers the parameter $y_l^{(i)}$ in dimension i generated at annealing time l with range $y_l^{(i)} \in (A_i, B_i)$ calculated with the random variable $v^{(i)}$ as

$$y_{l+1}^{(i)} = y_l^{(i)} + v^{(i)} (B_i - A_i) \qquad (6)$$

where $v^{(i)} \in [-1, 1]$.

Define he generating function

$$g_T(v) = \prod_{i=1}^{D} g_T^i(v^{(i)})$$

where $gT^{(i)}$ is the marginal of $v^{(i)}$ and

$$G_T^{(i)}(v^{(i)}) = \int_{-1}^{vi} g_T^{(i)}(r)dr$$

$$= \frac{1}{2} + \frac{\text{sgn}(v^{(i)})}{2} \frac{\ln(1+|v^{(i)}|/T_i)}{\ln(1+1/T_i)}$$

Here, $v^{(i)}$ is generated from $u^{(i)}$ with U(0,1) disribution, i.e.,

$$v^{(i)} = \text{sgn}(u^{(i)} - 1/2) T_i [(1+1/T_i)^{|2u(i)-1|} - 1] \qquad (7)$$

Clearly, the annealing schedule can be shown to satisfy

$$T_i(l) = T_{0i} \exp(-c_i l^{1/D}) \tag{8}$$

One can choose c_i to be equal to $r_i \exp\{-s_i/D\}$ where r_i and s_i are 'free' parameters to help tune ASA for specific problems.

4. A Revised Algorithm

The alternating relocation-swapping algorithm used to find the optimal non-hierarchical clustering constitution can be modified using ASA concepts. Given the function $\varphi(Z,\lambda)$ in (4), consider $y = (y(Z), \lambda)$, the state parameters. Here, $D = k + n + 1$. Then the annealing algorithm can be modified as follows.

0. Given values r_i and s_i at annealing time l and bound L, choose the schedule according to (8) and iniialize y.
1. Generate a random vector w using (6) where v is given in (7).
2. If $\varphi(w) - \varphi(y) \leq 0$, accept w as a new state and set y= w.
3. Else, if $\exp\{ -(\varphi(w) - \varphi(y))/T(l) > U(0,1)$, then accept was the new state and set y=w.
4. Obtain $\tilde{Z}$ from Z by swapping elements of C_k's such that $y(\tilde{Z}) = y(Z)$. Retain $\tilde{Z}$ if $\varphi(y(\tilde{Z})) - \varphi(y(Z)) < 0$.
5. Do step 4 until it is not possible to improve $\varphi(y(\tilde{Z}))$ anymore.
6. Lower T(l) to T(l + 1) and increment l.
7. If $i \leq L$, go to step 1.
8. Report all points y for which φ is lowest.

Note that steps 4 and 5 incorporate the swapping idea in non-hiearchical clustering procedures and thus is a main point of variation in traditional ASA applications. Also, the steps can take a while, especially if the number of objects n is large. One can improve on these steps by retaining the best $\tilde{Z}$ structure in the previous step.

5, Some Concerns

The clustering algorithm developed in section 4 has the non-hierarchical case as the principal application. The technique is basically stochastic in nature and the method flexible and general enough. Its desirability also hinges on the preservation of the probabilistic structure of the clusters. However, there are numerous areas in which the approach can be reconfigured. These then would dictate future research directions.

First, it would be informative to know how small sample sizes can affect the clustering results. This evaluation can be done by assessing the bias and misspecifications that may arise in applications. Results of the evaluation can then be used for adjustment purposes in order to male the technique robust.

The second area of improvement concerns the determination of the object weights $w_k(j)$. Statistical ideas such as the introduction of minimization criterion in determining the weights can be used. The Bayesian approach in selecting $w_k(j)$ can also be explored. The effect of such choices (both object and time weights) and the distances, either singly or in combination on the

actual clustering can then be evaluated especially in confirmatory problems. The effect of the choice of weight can be evaluated in terms of its impact on the speed of the algorithm, as well as how it affects cluster structure. One interesting aspect of this evaluation is the investigation of classes of 'smooth' functions from which the weights can be chosen from to ensure robustness.

The effect of missing data can also be looked at especially in the case when X is unbalanced. If imputation is done, bias evaluation should show the extent to which the cluster preservation objective was affected so that appropriate adjustments can be developed.

The constant α_{ti} basically determines the degree of importance that one attaches to observations taken at time t_i. The usual approach is to adopt a geometric determination of the constants with the more recent observations receiving the higher weights. As in the case of the object weights, one can apply statistical ideas to optimally choose the weighting pattern. Of specific interest is the case when T is large. The question then is to determine at what time points are to be dropped, i.e., which α_{ti} will be set to zero.

In the definition of δ_k in (2), one can choose a robust alternative to Δ_k in (3) such as L_1 distances. Measures of data depth such as in Liu and Singh (1993) can be used. It is well known that the effect of outliers is a function of the dimensionality of the data.

The preservation of the probabilistic structure of the clusters can easily be guaranteed by an IID structure among the index measurements. However, in the polling application, especially in the case of well-coordinated campaigns, these measurements are highly dependent. Forms of regularity conditions that will guarantee probabilistic structure preservation are of interest.

Lastly, since the viability of the annealing algorithm is a function of the annealing schedule, proper choice of the free parameters r_i and s_i is necessary. In this regard, simulation studies can be conducted to come up with heuristics, more importantly in cases when the number of clusters c is large.

6. Acknowledgments

The author would like to thank Profs. Bacelar-Nicolau, Mohammed Goria, Nye John and Dr. Tom Anderberg for sharing some ideas regarding clustering and theory of classification in general and for the very interesting discussions during the ISI delegation visit to Beijing last October, 1997. He is also very grateful to the incisive comments given by a referee. The comments did not only help the author present the material better but has given him additional impetus to pursue the topic further.

7. References

BACELAR-NICOLAU, H. AND NICOLAU, F. 1997. *Classification of Variables vs. Classification of Subjects: A Probabilistic Approach for Binary Data, In the Proceedings of the International Statistical Institute (Istanbul),* Session 51, pp. 349-352.

BONZO, D.I.B.C. 1995. *Determination of an Odds of Winning Scale in a Contested Election when Panel Data are Available*, In the Proceedings of the 1995 Philippine National Convention on Statistics (DEVSTAT).

FAHRMEIR, L. AND TUTZ, G. 1994. *Multivariate Statistical Modeling Based on Generalized linear Models*. New York: Springer.

GEMAN, S. AND GEMAN, D. 1984. *Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration in Images*, IEEE Trans. Patt. Anal. Mac. Int., 6, pp. 721-741.

INGBER, L. 1993. *Adaptive Simulated Annealing (ASA).* (ftp.alumni.caltech.edu: /pub/ingber/ASA-shar, AS-shar.Z, ASA.tar.Z, ASA.tar.gz, ASA.zip)

INGBER, L. 1993. *Simulated Annealing: Practice versus Theory*, J. Math'l Comput. and Modeling, 18, pp. 29-57.

JOBSON, J.D. 1992. *Applied Multivariate Data Analysis, vol. 2 (Categorical and Multivariate Methods).* New York: Springer.

KIRKPATRICK, S., GELATT, C.D. AND VECCHI, M.P. 1983. *Optimization by Simulated Annealing, Science*, 220, pp. 671-680.

KOSMELJ, K. AND BATAGELJ, V. 1990. *Cross-sectional Approach for Clustering Time Varying Data, J. of Classification*, 7, pp. 99-109.

LIU, R.Y. AND SINGH, K. 1993. *A Quality Index Based on Data Depth and Multivariate Rank* Tests, J. American Statist. Soc., 88, pp. 252-260.

SERFLING, R.J. 1980. *Approximation Theorems of Mathematical Statistics*. New York: Wiley.