Biometrika (1994), **81**, 4, *pp.* 721–9
Printed in Great Britain

Exponential mixtures and quadratic exponential families

BY PETER MCCULLAGH

Department of Statistics, University of Chicago, 5734 University Avenue, Chicago, Illinois 60637, U.S.A.

SUMMARY

Conditions are derived under which quadratic and polynomial exponential models can be generated as mixtures of linear exponential models. The conditions are highly restrictive for continuous sample spaces, but less restrictive in the discrete case. Some properties of binary quadratic exponential models are explored with a view towards finding models that have properties suitable for epidemiological applications.

Some key words: Cluster sampling; Mixture model; Quadratic exponential model.

1. INTRODUCTION

Correlated responses are common in many fields of application such as time series, spatial statistics and longitudinal studies. In medical statistics and in epidemiological studies, correlation can arise because of cluster sampling. Individuals in a cluster have in common unobserved traits, either genetic or environmental, as a result of which their responses such as susceptibility to disease, attitude to education, political affiliation and so on, tend to be alike. Such a description carries the implicit assumption that, conditional on the value of the unobserved trait, individuals within a cluster respond independently. Linear exponential-family models have been widely and successfully used for the analysis of independent responses. Quadratic Gibbsian models such as the Ising model have a lengthy history as models for physical phenomena such as ferromagnetism. More recently, similar quadratic exponential models have been put forward as a way of accommodating correlations of the type that occur in longitudinal studies and in cluster sampling (Zhao & Prentice, 1990, 1991; Fitzmaurice & Laird, 1993). The main purpose of this note is to investigate the conditions under which a mixture of independent linear exponential family distributions gives rise to a quadratic exponential model, or more generally a polynomial exponential model. Section 5 takes a critical look at how properties of quadratic exponential models are related to the behaviour one would expect in epidemiological applications.

2. CLUSTER SAMPLING AND EXPONENTIAL MIXTURES

Let $Y^1, \ldots, Y^n$ denote the responses of the n individuals in a cluster. Suppose that, conditionally on the unobserved trait, the observations are independent, and that the density of Y^j is

$$f_{\theta_j}(y) = e^{y\theta_j - k(\theta_j)} f_0(y)$$

with respect to either counting measure in the discrete case or Lebesgue measure in the continuous case. Thus $k(\theta)$ is the cumulant generating function of the initial distribution

f_0. The conditional distribution is of the linear exponential-family type with canonical parameter θ_j and cumulant function $k(.)$. The joint conditional density is thus

$$\prod_j \exp\{y^j\theta_j - k(\theta_j)\} f_0(y^j) \tag{1}$$

with respect to the product measure. The canonical parameter θ is here considered to be a function of the unobserved trait.

If the canonical parameter vector $(\theta_1, \ldots, \theta_n)$ has joint distribution function $F(.)$ on $\Theta \subset R^n$, the unconditional distribution of the observations Y is

$$p_n(y) = \int_\Theta \exp\{y^j\theta_j - \sum k(\theta_j) + \log f_0(y)\}\, dF(\theta) \tag{2}$$

with respect to the product measure on the sample space. In this paper, we seek to identify the family of distributions $\mathscr{F}$ on Θ such that, for all $F \in \mathscr{F}$,

$$\log p_n(y) = P_m(y) + d(y), \tag{3}$$

where $P_m(y) \equiv P_m(y; \phi)$ is a polynomial in y of degree not more than m with coefficients ϕ, and $d = d(y_1, \ldots, y_n)$ is an arbitrary function. An important point here is that P_m is required to be a polynomial of fixed maximal degree m independent of the cluster size n. Also, (3) need be satisfied only for y in the sample space: in the binary case this means $y^j = 0, 1$ only. For $m = 2$, (3) is called a quadratic exponential model.

Since $d(y)$ is an arbitrary function, it might appear that P_m can be incorporated into d without loss of generality and without restriction on p_n. However, the term exponential family refers not to a particular distribution, but to a family of distributions. In the case of (3), the family is indexed by the coefficients in the polynomial P_m. Condition (3) must be satisfied by the same function d for each $F \in \mathscr{F}$. Only the coefficients in P_m depend on the particular choice of F. Thus, the condition places no restrictions on any particular F, but it does impose restrictions on the family $\mathscr{F}$.

3. TWO EXAMPLES

We consider two examples, one in which the observations are continuously distributed and one in which the observations are binary.

The simplest examples concern the normal distribution. Suppose, in the notation previously established, that

$$Y^j|\theta \sim N(\theta_j, 1), \quad \theta \sim N_n(\mu, \Sigma).$$

Then the unconditional distribution is $Y \sim N_n(\mu, \Sigma + I)$. The conditional distributions are of the linear exponential type, and the unconditional distributions are of the quadratic exponential type. The coefficients in the polynomial P_2 are $(\Sigma + I)^{-1}\mu$ and $-\frac{1}{2}(\Sigma + I)^{-1}$ respectively.

To take a second example, let the conditional distribution of Y^j be Bernoulli with parameter $\pi_j = e^{\theta_j}/(1 + e^{\theta_j})$. Suppose that the joint density of $\theta_1, \ldots, \theta_n$ is given by

$$f_n(\theta) = \phi_n(\theta; \mu, \Sigma) \prod_j (1 + e^{\theta_j})/M, \tag{4}$$

where $\phi_n(\ )$ denotes the normal density with mean vector μ_j and covariance matrix Σ_{ij}, and M is a normalisation constant depending on μ, Σ. Then the unconditional joint distribution of Y is

$$p_n(y) = \int_\Theta e^{\theta_j y^j} \phi_n(\theta; \mu, \Sigma)\, d\theta/M = \exp(\mu_j y^j + \tfrac{1}{2}\Sigma_{ij} y^i y^j)/M. \tag{5}$$

Evidently, $M = \sum_y \exp(\mu_j y^j + \frac{1}{2}\Sigma_{ij} y^i y^j)$ with summation over all n-component binary vectors y. This is a discrete quadratic exponential model whose canonical parameters μ, Σ are the mean-value parameters from the density ϕ_n. Because of the binary nature of the observations, there is some redundancy in the parameterisation, which may be eliminated most conveniently by taking either $\Sigma_{jj} = 0$ or $\mu_j = 0$. In other words, several distinct mixing distributions within the class (4) produce the same quadratic exponential distribution (5).

4. Conditions for existence

Let $S \subset R^n$ be the sample space. For convenience of notation we write $y\theta$ for the linear combination $y^j\theta_j$, and $k(\theta)$ for the joint cumulant generating function. We seek conditions on the class of distributions $\mathscr{F}$ such that for each F_0, F_1 in $\mathscr{F}$ the ratio of unconditional densities satisfies

$$\frac{\int \exp\{y\theta - k(\theta)\} f_0(y)\, dF_1(\theta)}{\int \exp\{y\theta - k(\theta)\} f_0(y)\, dF_0(\theta)} = \exp\{P_m(y)\}. \tag{6}$$

The common function $d(y)$ in (3) cancels out in the ratio. This is slightly more general than the formulation in § 2 because the components of Y in $f_0(y)$, and in the exponential family generated from f_0, need not be conditionally independent.

We first note that (3) is satisfied, in a sense trivially with $m = 1$, by any distribution degenerate at a point $\theta_0 \in \Theta$. With F_0 equal to any such point distribution, we have from (6)

$$\int_\Theta \exp\{y\theta - k(\theta)\}\, dF(\theta) = \exp\{P_m(y) + y\theta_0 - k(\theta_0)\} \tag{7}$$

for each $F \in \mathscr{F}$ and for each $y \in S$.

It is convenient in what follows to arrange matters so that $0 \in S$. This can be done without loss of generality by translation of S. For each $F \in \mathscr{F}$ we define a new probability distribution F^* on Θ by

$$dF^*(\theta) = M \exp\{-k(\theta)\}\, dF(\theta),$$

where M is a normalisation constant. The family thus generated is denoted by $\mathscr{F}^*$. Putting $y = 0$ in (7) gives $\log M = -P_m(0) + k(\theta_0)$. Thus, for each $F^* \in \mathscr{F}^*$, and for each $y \in S$ we have

$$\int_\Theta e^{y\theta}\, dF^*(\theta) = \exp\{P_m(y) - P_m(0)\}. \tag{8}$$

To make further progress, it is necessary to consider a number of cases separately, the distinctions having to do with the nature of the sample space. Three of the most important special cases are now considered.

Continuous case. In the continuous case where S contains an open neighbourhood of the origin, (8) implies that the cumulant generating function of F^* is a polynomial.

Marcinkiewicz's theorem (Marcinkiewicz, 1938; Lukacs, 1958; Moran, 1984, p. 277) then implies that $m=2$ and F^* is normal. It follows that F and F^* have the same support, namely $\Theta = R^n$ and $|k(\theta)| < \infty$ for all θ. Consequently $\mathscr{F}$ is equal to the set of distributions with densities

$$f(\theta) = \phi_n(\theta; \mu, \Sigma)e^{k(\theta)}/M(\mu, \Sigma) \tag{9}$$

for θ, $\mu \in R^n$, Σ positive semi-definite, and

$$M = \int \phi_n(\theta; \mu, \Sigma)e^{k(\theta)}\, d\theta < \infty.$$

No polynomial exponential family of degree $m > 2$ can be generated as a mixture of linear exponential models.

As a corollary, if Θ does not coincide with R^n, $\mathscr{F}$ contains only singletons or degenerate distributions. In such cases, no nontrivial polynomial exponential family can be generated by mixing on the canonical parameter. For example, if $n=1$ and $f_0(y) = e^{-|y|}/2$ or $f_0(y) = e^y/(1+e^y)^2$, Θ is the interval $(-1, 1)$ so $\mathscr{F}$ contains only singletons. In fact, the condition $M < \infty$ implies $\int \exp(y^T\Sigma y/2) f_0(y)\, dy < \infty$ for some nontrivial Σ, thereby excluding exponential families whose tails are heavier than normal.

Binary case. Let S be the set of 2^n n-component binary vectors y. Condition (8) need only be satisfied on this set, giving a total of 2^n constraints on the monomial moments of the variables $\{e^{\theta_j}\}$ in the distribution F^*. For $m=2$, density (9) with $\mu=0$ satisfies the required conditions, but, unlike the continuous case, it does not exhaust the possibilities. We now show that any polynomial P_m of any degree gives rise to a valid set of moments.

First observe that the canonical parameter space Θ coincides with R^n, and the mean-value space $\mathscr{M}$ is the unit cube with components $\pi_j = e^{\theta_j}/(1+e^{\theta_j})$. The extreme points in the closure of $\mathscr{M}$ are the 2^n n-component vectors in S. For $y \in S$ and $\pi \in \mathrm{cl}(\mathscr{M})$, the conditional distributions have the form

$$f(y|\pi) = \pi^y(1-\pi)^{1-y}, \tag{10}$$

where π^y is the product $\prod \pi_j^{y_j}$. If $\pi \in S$, that is an extreme point of $\mathrm{cl}(\mathscr{M})$, this conditional probability is zero at all points of S except $y = \pi$. Let $p(y)$ be an arbitrary distribution on S. Such a distribution can be realised as a mixture over $\mathscr{M}$ of independent distributions (10). In fact, the mixing distribution need only be supported on the extreme points of $\mathscr{M}$ as follows:

$$p(y) = \sum_{\pi \in S} f(y|\pi)p(\pi).$$

Any mixing distribution F on the canonical parameter space Θ induces a mixing distribution G on the mean-value space $\mathscr{M}$ by component-wise transformation. All that is required of G is that its 2^n monomial moments should match those of p. In symbols, for each $y \in S$,

$$\int_{\mathscr{M}} \pi^y\, dG(\pi) = \sum_{x \in S} x^y p(x).$$

The preceding argument shows that this is achievable in a trivial way for all distributions p on S by taking $G = p$ on the extreme points, but there are infinitely many distributions satisfying the moment condition. To conclude, polynomial exponential models of all orders, and without restriction on P_m, can be generated as mixtures of independent Bernoulli variables.

Lattice case $S = Z^n$. Let S be the set of integer-valued n-component vectors y. Condition (8) gives the joint moments of all orders of the random variables $\{e^{\theta_j}\}$ in the distribution F^*. It should be borne in mind that not all polynomials P_m give rise to valid moments. Further, even if P_m gives rise to a valid set of moments, these moments may not identify F^* uniquely.

Let m be a positive integer, and let $P_m(.)$ be a polynomial of degree m satisfying $P_m(0) = 0$. Suppose that for each integer r, positive or negative, the exponential moment

$$M(r) = \int_{-\infty}^{\infty} e^{rx}\, dF^*(x) = \exp\{P_m(r)\} \tag{11}$$

is finite. The $M(t)$ exists for all real t. In fact, $M(iz)$ is an entire characteristic function of order m. Consequently, by the Hadamard factorisation theorem

$$M(z) = G(z) \exp\{P_m(z)\},$$

where G, the canonical product of the zeros of M, takes the value unity on the integers. If $m = 1$, the moments determine F^* uniquely as a degenerate distribution. If $m = 2$, there is a range of solutions including $G \equiv 1$. Any distribution having the same exponential moments as the normal will satisfy (11). This determines the entire class of distributions F^* satisfying (11) with $m = 2$. For example, the distribution with density $\phi(x)\{1 + \varepsilon \sin(\omega x)\}$ for $|\varepsilon| \leqslant 1$ has moment generating function $\exp(t^2/2)\{1 + \varepsilon e^{-\omega^2/2} \sin(\omega t)\}$. If ω is an integer multiple of π, these moment generating functions coincide on the integers, and the entire family has the same exponential moments as the standard normal, $M(r) = \exp(r^2/2)$ for integer r.

For $m > 2$ convexity of M on the real line implies that r is even. In addition, Theorem A of Lukacs (1958) imposes severe constraints on G, such that the exponent of convergence of the canonical product of the roots must equal m. In other words, if $\{z_n\}$ are the roots of M arranged in increasing order of modulus, then

$$\rho = \limsup_{n \to \infty} \frac{\log n}{\log |z_n|} = m.$$

Consequently, the number of roots less than t in modulus is $O(t^{m+\varepsilon})$, so the roots are increasingly dense in the plane as $|z| \to \infty$. It follows that G cannot be periodic even though $G = 1$ on the real integers. While one could perhaps construct such a function G, it seems doubtful that it could satisfy other requirements for a characteristic function.

5. Limit distributions for exchangeable binary models

5·1. *Strong dependence*

Let Y be an n-component binary vector with joint distribution

$$p_n(y) = \exp(\gamma_1 y_. - \tfrac{1}{2}\gamma_2 y_.^2)/M, \tag{12}$$

where $y_.$ is the sum of the components. Then the marginal distribution of $Y_.$ is

$$\operatorname{pr}(Y_. = y) = \binom{n}{y} \exp(\gamma_1 y - \tfrac{1}{2}\gamma_2 y^2)/M. \tag{13}$$

If $\gamma_2 = 0$, $Y_.$ is binomial with parameter $\exp(\gamma_1)/\{1+\exp(\gamma_1)\}$, and the standardised sum has a limiting normal distribution. We now show that if $\gamma_2 > 0$ no limit distribution exists but the asymptotic behaviour as $n \to \infty$ is as follows:

$$Y_. = \{\log(n\gamma_2) - \log\log n + \gamma_1\}/\gamma_2 + O_p(1).$$

As a consequence, $\bar{Y}_n \to 0$ in probability. By contrast, if $\gamma_2 < 0$, $Y_. - n$ tends to zero in probability.

To prove these results, consider the function

$$g_n(y) = \gamma_1 y - \tfrac{1}{2}\gamma_2 y^2 - \log\Gamma(y+1) - \log\Gamma(n-y+1)$$

for $0 \leqslant y \leqslant n$. This is the exponent in the density (13). For $\gamma_2 \geqslant 0$, g_n is unimodal on $[0, n]$ with derivative

$$g_n'(y) = \gamma_1 - \gamma_2 y - \psi(y+1) + \psi(n-y+1). \tag{14}$$

Now, g_n' is monotone, decreasing from $\psi(n+1) - \psi(1) + \gamma_1$ at $y=0$ to $-n\gamma_2 - \psi(n+1) + \psi(1) + \gamma_1$ at $y = n$. For sufficiently large n, $g'(0) > 0$. Similarly, if $\gamma_2 \geqslant 0$, $g'(n) < 0$, so $g_n'(\hat{y}_n) = 0$ has a solution in the interval $(0, n)$. For $\gamma_2 > 0$, the approximate solution for large n is

$$\gamma_2 \hat{y}_n = \log(n\gamma_2) - \log\log n + \gamma_1 + o(1),$$

so $\hat{y}_n$ is $O(\log n)$. Finally, we observe that for any fixed δ

$$g_n(\hat{y}_n + \delta) - g_n(\hat{y}_n) = -\tfrac{1}{2}\gamma_2\delta^2 + o(1),$$

so g_n is approximately quadratic in any $O(1)$ neighbourhood of $\hat{y}_n$, and this region contains essentially all the probability. Were it not for the restriction to integers, the conclusion would be that $Y_. - \hat{y}_n$ has a limiting normal distribution with zero mean and variance $1/\gamma_2$. As it is, $Y_.$ is asymptotically 'discrete normal' with probability mass function proportional to

$$\text{pr}(Y_. = y) \propto \exp\{-\gamma_2(y - \hat{y}_n)^2/2\} \tag{15}$$

on the integers. The centred random variable $Y_. - \hat{y}_n$ is $O_p(1)$ with roughly zero mean and variance $1/\gamma_2$, but it does not have a limit distribution. On the subsequence for which the fractional part of $\hat{y}_n$ is approximately a constant, α, the centred random variable $Y_. - [\hat{y}_n]$ does have a limit distribution, the discrete normal with parameters α, $1/\gamma_2$. For purposes of approximation, however, it is best to use (15) directly with γ_2 replaced by $-g_n''(\hat{y}_n)$.

5·2. *Weak dependence*

Let Y have joint distribution (12) with γ_2 replaced by γ_2/n: this form is sometimes called the Curie–Weiss model for spontaneous magnetisation (Ellis, 1985, p. 98). Then the marginal distribution of $Y_.$ is

$$\text{pr}(Y_. = y) = \binom{n}{y} \exp(\gamma_1 y - \tfrac{1}{2}\gamma_2 y^2/n)/M.$$

Depending on the value of (γ_1, γ_2), the limiting distribution of $Y_.$ can take one of several forms.

The exponent

$$g_n(y) = \gamma_1 y - \tfrac{1}{2}\gamma_2 y^2/n - \log\Gamma(y+1) - \log\Gamma(n-y+1)$$

has derivative

$$g_n'(y) = \gamma_1 - \gamma_2 y/n - \psi(y+1) + \psi(n-y+1).$$

For large n, the equation $g_n'(y) = 0$ has either one or three roots: in the latter case some of the roots may be coincident. The main cases to be distinguished are as follows:

(i) for $\gamma_2 > -4$, g_n' is monotone decreasing, so there is a single root, $\hat{y} = n\pi$ with $0 < \pi < 1$;

(ii) for $\gamma_2 = 2\gamma_1 = -4$ there are three coincident roots at $\hat{y} = n/2$;

(iii) for $\gamma_2 = 2\gamma_1 < -4$ the three roots, $n\pi$, $n/2$, $n(1-\pi)$, are symmetrically located about $n/2$;

(iv) in all other cases there is either exactly one root, or, if there are three roots, one is dominant.

In case (i) the asymptotic behaviour of the exponent is such that

$$\sqrt{n}(\bar{Y} - \pi) \sim N\{0, \sigma^2(\pi)\}, \tag{16}$$

where the limiting proportion π is the root of the equation $\gamma_1 - \gamma_2\pi - \log\{\pi/(1-\pi)\} = 0$. The limiting variance is

$$\sigma^2(\pi) = \frac{\pi(1-\pi)}{1 + \gamma_2\pi(1-\pi)},$$

so there is excess asymptotic dispersion if $\gamma_2 < 0$.

In case (ii), $n^{1/4}(\bar{Y} - \frac{1}{2})$ has a nondegenerate, nonnormal limiting distribution in which the density at t is proportional to $\exp(-4t^4/3)$.

In (iii), the case of so-called spontaneous magnetisation, the asymptotic distribution of $\bar{Y}$ is an equally-weighted two-component normal mixture centred at π and $1-\pi$ with equal variances $\sigma^2(\pi)/n$. So the limit distribution of $\bar{Y}$ has probability $\frac{1}{2}$ at the points π and $1-\pi$.

In case (iv), if there is one root, the result is the same as (16). If there are three roots $\pi_1 \leqslant \pi_2 < \frac{1}{2} < \pi_3$, the third is dominant. Conversely, if $\pi_1 < \frac{1}{2} < \pi_2 \leqslant \pi_3$, the first root is dominant. The limit result is the same as (16) at the dominant root.

5·3. *Implications for applications*

The applications that I have in mind here are sociological or epidemiological in which the clusters are families, households, schools, villages or small communities in which there is a high degree of within-cluster contact or relatedness relative to that between clusters. The intention is to construct a model in which incidence rates depend on cluster-specific or individual-specific covariates, but the pairwise within-cluster association, however measured, is constant across clusters and accounts for the correlations among subjects. To keep the discussion focused, we consider here the simplest case in which observations within cluster are exchangeable, and only cluster-specific covariates are available.

A standard modelling procedure is to begin with a distribution such as (12) with parameters $(\gamma_1^{(i)}, \gamma_2^{(i)})$ for the ith cluster. Dependence on cluster-specific covariates is usually incorporated through a linear regression model for $\gamma_1^{(i)}$: the simplest model of association is to take $\gamma_2^{(i)}$ to be constant over clusters, although more complicated models for association can certainly be contemplated (Fitzmaurice & Laird, 1993). Although there is no fundamental reason to exclude it, cluster size is rarely included as an explicit covariate.

A quite different approach is to construct models for the mean-value parameters (Zhao

& Prentice, 1990), as opposed to the canonical parameters (γ_1, γ_2). The criticisms that follow refer mostly to models specified directly in terms of the canonical parameters, but they apply also to the so-called mixed parameterisation (Fitzmaurice & Laird, 1993) in which a regression model is set up for the association parameters of a strong-dependence distribution of the type (12).

Two objections to the use of the strong-dependence model are as follows. First, under the component-wise transformation $Y \mapsto 1 - Y$, the parameters in (12) transform to

$$\gamma_1 \mapsto n\gamma_2 - \gamma_1, \quad \gamma_2 \mapsto \gamma_2.$$

Thus, unless the cluster size is included as a covariate in the regression for $\gamma_1^{(i)}$, the model is not closed under re-labelling of outcomes. In a technical sense, family (12) is not closed because, if $\gamma_2 \neq 0$, the transformed γ_1 is $O(n)$. The weakly dependent Curie–Weiss model overcomes this objection because the induced transformation is independent of n:

$$\gamma_2 \mapsto \gamma_2 - \gamma_1, \quad \gamma_2 \mapsto \gamma_2,$$

and γ_2 is invariant. Whether cluster size is included as a covariate or not, the two models are generally different because in the first case pairwise interactions are assumed constant across clusters, whereas in the second case they are assumed to be inversely proportional to cluster size. The two models are equivalent only if all clusters are of equal size.

A second objection concerns the auto-logistic property

$$\log \text{odds}\,(Y_1 = 1 \mid \text{others}) = \begin{cases} \gamma_1 - \tfrac{1}{2}\gamma_2 - \gamma_2 y_{\cdot}^{(1)} & \text{(strong)}, \\ \gamma_1 - \tfrac{1}{2}\gamma_2/n - \gamma_2 y_{\cdot}^{(1)}/n & \text{(weak)}, \end{cases} \tag{17}$$

where $y_{\cdot}^{(1)}$ is the sum of the others in the cluster. The case of most interest is $\gamma_2 < 0$, which is assumed for convenience in the discussion that follows. In the strongly dependent case the log odds has a lower bound of $\gamma_1 - \frac{1}{2}\gamma_2$, and an upper bound of $\gamma_1 + \frac{1}{2}\gamma_2 - n\gamma_2$ tending to infinity with cluster size. While such asymmetry is usually undesirable in practice, a more compelling objection is that the conditional log odds in the first part of (17) depends on the remaining cluster total regardless of cluster size. Regardless of the parameter values, the first equation (17) gives the same probability for the two cases $(y_{\cdot}^{(1)}, n) = (5, 6)$ and $(y_{\cdot}^{(1)}, n) = (5, 101)$, even though $y_{\cdot}^{(1)} = 5$ constitutes a 100% success rate in the first case and only 5% in the second. In the weakly-dependent model the log odds in (17) has a lower limit of $\gamma_1 - \frac{1}{2}\gamma_2/n$ and an upper limit of $\gamma_1 - \gamma_2 + \frac{1}{2}\gamma_2/n$. In practice one might expect the upper limit to increase slowly with n because a 100% success rate among the neighbours is more convincing for larger n. The weakly-dependent model exhibits this effect in a weak form.

In many ways, the most plausible parameter values of epidemiological applications are those for which the cluster mean $\bar{Y}$ has a nondegenerate limiting distribution. The actual cluster average $\bar{y}^{(i)}$ is then considered to be indicative of the specific exposure or immunity of the ith cluster, both of which vary from cluster to cluster. Although there is no reason in practice to expect the limiting distribution to be discrete or symmetric about $\frac{1}{2}$, this argument favours the weakly dependent model with parameter values in the region $\gamma_1 = -2 - \varepsilon_1$, $\gamma_2 = -4 - \varepsilon_2$ for small positive values of $\varepsilon_1, \varepsilon_2$.

ACKNOWLEDGEMENT

I am grateful to Yali Amit for helpful discussions concerning limiting distributions for Gibbs models.

References

Ellis, R. S. (1985). *Entropy, Large Deviations, and Statistical Mechanics*. New York: Springer-Verlag.

Fitzmaurice, G. M. & Laird, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika* **80**, 141–51.

Lukacs, E. (1958). Some extensions of a theorem of Marcinkiewicz. *Pacific J. Math.* **8**, 487–501.

Marcinkiewicz, J. (1938). Sur une propriété de la loi de Gauss. *Math Zeitung* **44**, 612–8.

Moran, P. (1984). *An Introduction to Probability Theory*. Oxford University Press.

Zhao, L. P. & Prentice, R. L. (1990). Correlated binary data regression using a quadratic exponential model. *Biometrika* **77**, 642–8.

Zhao, L. P. & Prentice, R. L. (1991). Use of quadratic exponential model to generate estimating equations for means, variances, and covariances. In *Estimating Functions*, Ed. V. P. Godambe, pp. 103–17. Oxford: Clarendon Press.

[*Received January* 1994. *Revised May* 1994]