



UW Biostatistics Working Paper Series

6-11-2003

Asymptotics for Marginal Generalized Linear Models With Sparse Correlations

Thomas Lumley

University of Washington, tlumley@u.washington.edu

Nicole Mayer Hamblett

Children's Hospital and Regional Medical Center, mayerh@u.washington.edu

Suggested Citation

Lumley, Thomas and Mayer Hamblett, Nicole, "Asymptotics for Marginal Generalized Linear Models With Sparse Correlations" (June 2003). *UW Biostatistics Working Paper Series*. Working Paper 207.
<http://biostats.bepress.com/uwbiostat/paper207>

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

Copyright © 2011 by the authors

1 INTRODUCTION

In many applications, data are weakly dependent and the form of the dependence is not of primary interest. The systematic variation in the mean response may be modelled using standard regression methods, but inference regarding these mean contrasts requires estimation of covariances.

A semiparametric estimating equation approach has been used in similar situations where we are unable or unwilling to specify a fully parametric model for the data. Even when a full parametric model can be specified, simple estimating equations can still provide a reduction in computational effort and a gain in robustness of inference, with consistent parameter estimation and valid testing under weaker assumptions than are required for maximum likelihood methods. These benefits are illustrated by the fitting of generalized linear models using quaslikelihood (Wedderburn, 1974), later extended by Liang & Zeger (1986) to longitudinal data. Under this approach, a model is specified for only the first two moments of the data. The resulting parameter estimates are consistent if the mean is correctly specified and relatively efficient if the second moment assumptions are approximately correct.

If we have a scalar response Y_j and a p -vector of predictors X_j for observation j , the marginal generalized linear model specifies

$$g(E[Y_j|X_j]) \equiv g(\mu_j) = X_j'\beta \quad (1.1)$$

for a p -vector of parameters β , where g is a monotone, smooth known function called the link. We are interested in inference that is valid under this mean restriction, together with necessary moment restrictions, and that is still relatively efficient in the submodel where

$$\text{var}[Y_j|X_j] = \phi V(\mu_j)$$

for a known variance function $V(\cdot)$. We estimate β by solving the quasiscore equations

$$U_n(\beta) = \sum_{j=1}^n U_j(\beta) = \sum_{j=1}^n \frac{\partial \mu_j}{\partial \beta} \frac{(Y_j - \mu_j)}{V(\mu_j)} = 0.$$

These have mean zero at the true value of β for any distribution satisfying the mean model and are the exact score equations for the exponential family distribution contained in the submodel with this link and variance function (McCullagh & Nelder, 1989). We will refer to the loglikelihood for this exponential family distribution as the “independence working loglikelihood” or “quasilikelihood”.

Using information sandwich estimators (Huber, 1967; White, 1984; Royall, 1986; Lin & Wei, 1989), standard error estimates are consistent if the first moment is correctly specified. These standard error estimators are based on empirical variances computed from independent subsets of the data and so are not directly applicable to some important correlated data designs. For data measured over time or space, modifications of the sandwich estimators have been constructed (Newey & West, 1987; Lele, 1991; Lumley & Heagerty, 1999) substituting asymptotic independence for exact independence.

We deal here with another important case of correlated data which occurs when the correlation matrix is sparse but not block diagonal so that the data cannot be decomposed into independent blocks even though most pairs of observations are independent. The most common example of this is a crossed experimental design where observations are correlated if they share any one of the design factors. Perhaps the best known example of a crossed design with non-Gaussian response is the salamander mating experiment analyzed by McCullagh & Nelder (1989), Karim & Zeger (1992), Shun (1997), and others.

In this paper, we give conditions that allow marginal generalized linear models to be estimated using the quasiscore equations. In Section 2, we describe the two applications that motivated our research. The first is a method for modelling changing patterns of genetic variation of the human immunodeficiency virus (HIV) within an infected patient; the second is an investigation of the properties of quasiscore estimation for longitudinal

data where the number of observations on each individual is large.

Other examples for which our methods are potentially relevant occur in education (Rasbash & Goldstein, 1994), reproductive medicine (Clayton & Écochard, 1997), and medical diagnostics (Nelson, 1999). In many cases, sparsely correlated data have been analyzed by fitting a generalized linear mixed model (Breslow & Clayton, 1993) in which the correlation is modelled using latent Gaussian random variables. Estimation in these models is computationally difficult. More importantly, the parameters in a generalized linear mixed model have a different interpretation from those in a marginal generalized linear model and so it is useful to be able to fit either class of model as appropriate.

In the case of a complete crossed design where every level of each factor is crossed with every level of every other factor, an analysis using U -processes may be possible. This would allow weaker moment and smoothness assumptions (e.g. de la Peña & Giné, 1999; Nolan & Pollard, 1987) than we require. Our methods, however, also apply to incomplete crossed designs and other sources of sparsely correlated data that lack the special structure of U -statistics, as will be demonstrated in the genetic variation example.

The central limit theorem that we require to establish conditions for the estimation of generalized linear models using quasiscore equations is proven in Section 3 and in Section 4, we derive the conditions for consistency and asymptotic Normality of regression parameter estimates and consistency of an empirical variance estimator.

2 EXAMPLES

2.1 *Modelling Patterns of HIV Genetic Variation*

In studies of HIV genetic variation, the emphasis is often on describing patterns of evolutionary change within the context of specific evolutionary models. Inference for parameters in these models can therefore rely on underlying structural modelling assumptions (Hillis et al., 1996, Felsenstein (1988), Miyamoto & Cracraft (1991)). In some instances however, it

is useful to be able to rigorously test hypotheses regarding patterns of HIV genetic variation using empirically based methods which are relatively free of underlying model assumptions. This more formal statistical approach can provide comprehensive descriptions of patterns of genetic variation which are important for understanding disease progression, transmission dynamics, antiviral drug resistance, and vaccine efficacy.

In Mayer-Hamblett (1999), a regression modelling framework is developed to both describe and test for patterns of HIV genetic variation. We motivate the work in this paper by presenting one particular model which describes the variation between viral populations existing at different time points in a single patient's infection as a function of covariates such as time. Here, we define a viral population as a collection of viral genomes existing at a particular time point within a patient's infection and possibly within a specific tissue compartment. This model can be used to answer biologically relevant questions concerning the patterns of HIV genetic variation occurring over time within a single infected patient. For example, one question of interest is what is the pattern over time in the variation between the initial viral population and viral populations existing later in infection. Such information can provide valuable insights into the dynamics between the virus and the immune system throughout infection (Shankarappa et al., 2000).

Suppose genetic sequences are sampled at times t_1, \dots, t_T from a single patient during their infection. Let $G_t = (G_{t1}, \dots, G_{tR})$ denote a sample of R genetic sequences drawn from a large population of viral genomes existing at time t , $t \in \{t_1, \dots, t_T\}$, and $G_u = (G_{u1}, \dots, G_{uR})$ denote a sample of R genetic sequences drawn from a second large population of viral genomes existing at a different time $u \in \{t_1, \dots, t_T\}$ where $G_{tr}^T = (G_{tr}^{(1)}, \dots, G_{tr}^{(S)})$ is a genetic sequence of length S from time t and $G_{tr}^{(s)}$ denotes a "site" in this sequence taking values from the set of nucleotides $\{A, C, T, G\}$. The marginal distributions of $G_{tr}^{(s)}$ and $G_{ur}^{(s)}$ are assumed to be multinomial with probabilities $p_t^{(s)}$ and $p_u^{(s)}$, respectively.

Since the viral life cycle occurs on a time scale (in days) which is much faster than the interval between sampling of an infected patient, the viral populations existing at different

time points can be considered distinct. It is therefore reasonable to consider samples drawn at different time points, conditional on the time-specific multinomial parameters, to be independent. Further, because the viral populations at each time point are generally very large, we can consider the set of sequences sampled from a single viral population to be independent and identically distributed.

One measure of the variation between two viral populations at site s is

$$Pr\{G_{uj}^{(s)} \neq G_{tk}^{(s)}\} = 1 - \langle p_u^{(s)}, p_t^{(s)} \rangle = \phi_{ut}^{(s)}$$

and when $u = t$, this quantity is referred to as Simpson's Index of Diversity (Simpson, 1949). To estimate $\phi_{ut}^{(s)}$, we can use the observed genetic distance indicators

$$d(G_{uj}^{(s)}, G_{tk}^{(s)}) = \begin{cases} 1 & G_{uj}^{(s)} \neq G_{tk}^{(s)} \\ 0 & \text{otherwise} \end{cases}$$

since $E[d(G_{uj}^{(s)}, G_{tk}^{(s)})] = \phi_{ut}^{(s)}$. There are $T(T-1)/2$ different between viral population variation parameters specific to site s that can be constructed if sequence data is sampled at T different time points.

Distance indicators pertaining to several sites can then be used to model the average of site-specific measures of between viral population variation, $\phi_{ut}^{(\cdot)} = \sum_{s=1}^S \phi_{ut}^{(s)} / S$, across well-defined regions of the HIV genome. For instance, the *env* gene is one region that is important to consider as this gene plays a major role in the infection of CD4+ T cells.

Given a covariate vector Z_{ut} (i.e. $|u - t|$) where $(u, t) \in \{t_1, \dots, t_T\}^{\otimes 2}$ and $u \neq t$, and associated regression parameter vector β , a marginal model for $\phi_{ut}^{(\cdot)}$ can be specified as

$$g(\phi_{ut}^{(\cdot)}) = Z_{ut}'\beta$$

where g represents a link function in the tradition of generalized linear models.

Let $d_{ut}^{(s)} = d(G_{uj}^{(s)}, G_{tk}^{(s)})$ and $d_{u't'}^{(s)} = d(G_{u'j'}^{(s)}, G_{t'k'}^{(s)})$ be distance indicators comparing sequences from times u and t and times u' and t' , respectively. From the previously discussed

independence assumptions, the covariance structure of distance indicators pertaining to a single site s is then given by

$$\text{cov}[d_{ut}^{(s)}, d_{u't'}^{(s)}] = \begin{cases} 0 & \{(u, j) \neq (u', j') \text{ and } (t, k) \neq (t', k')\} \\ \alpha \sigma_{ut}^2 \sigma_{u't'}^2 & \{(u, j) = (u', j') \text{ and } (t, k) \neq (t', k'), \\ & (u, j) \neq (u', j') \text{ and } (t, k) = (t', k')\} \\ \sigma_{ut}^2 & \{(u, j) = (u', j') \text{ and } (t, k) = (t', k')\} \end{cases} \quad (2.1)$$

where $\sigma_{ut}^2 = \phi_{ut}^{(s)}(1 - \phi_{ut}^{(s)})$, $\sigma_{u't'}^2 = \phi_{u't'}^{(s)}(1 - \phi_{u't'}^{(s)})$, and $\alpha = \alpha(\phi_{ut}^{(s)}, \phi_{u't'}^{(s)})$ is the correlation between the two distance indicators.

The covariance structure at one site resembles that which arises from a crossed design since genetic sequences are crossed with themselves in order to construct the distance observations used in the model. Because the distance indicators are symmetric, only distances based on unique pairs of sequences are necessary and therefore this can more accurately be called an incomplete crossed design. As data from several sites are used to estimate the parameters in this model, the dependence among distances from different sites in the genome must also be accommodated. However, there is no biological model providing direction for modelling the dependence across sites.

Since the mean response is of primary interest, Mayer-Hamblett (1999) uses a marginal binary regression model for estimation of the regression parameters which avoids having to specify a model for the covariance (Liang & Zeger, 1986). In addition, an empirical variance estimator that accounts for the known independence in the data is used for obtaining parameter standard errors. In this paper, we provide the theoretical foundation given in Mayer-Hamblett (1999) for the use of marginal estimating equations and empirical variance estimation in this example and more generally for other crossed designs.

2.2 Two-index asymptotics for GEE sandwich estimator

The Generalized Estimating Equations (GEE) method (Liang & Zeger, 1986), popular for the analysis of longitudinal data, involves fitting a marginal generalized linear model to T repeated observations on each of K individuals. The asymptotic arguments presented by Liang & Zeger (1986) and others assume that T is fixed and $K \rightarrow \infty$, and show that the parameter estimates are consistent and asymptotically Normal and that the sandwich estimator for the variance is consistent.

An important practical question left open by these results is the performance of GEE when the number of observations per individual T is relatively large. For example in dental research, we typically have measurements on $T = 32$ teeth per person and in community randomized trials where the “individual” is a whole community, T can be several hundreds or thousands. In order for asymptotic results to be relevant to data analysis when T is not small compared to K , we need to consider the asymptotic behavior as both $T \rightarrow \infty$ and $K \rightarrow \infty$.

Theorem 7 in this paper shows that consistency of the sandwich estimator holds with rate K not only for T fixed, but for $T \rightarrow \infty$ at any rate. This suggests that in data analysis, the performance of these estimators will depend largely on the number of individuals K and not on the total number of observations KT . Since the value of T is not important, simulation studies (e.g. Sharples & Breslow, 1992) that have been performed with relatively small values of T can give useful information for larger values of T as well.

3 LIMIT THEOREMS FOR SPARSELY CORRELATED DATA

There are two main limit theorems for sparsely correlated data which need to be established: a central limit theorem and a theorem for consistency of an empirical variance estimator. These, together with standard convexity and smoothness arguments, imply asymptotic normality of the regression parameters and valid standard error estimates.

We begin this section with a formal definition of sparsely correlated data and a discussion of the independence conditions under which the limit theorems hold. We then prove the central limit theorem and the main lemma that will be needed in Section 4 where we show consistency and asymptotic normality of regression parameters and consistency of an empirical variance estimator.

3.1 Definitions

Suppose we have observations X_1, \dots, X_n . For each observation X_j , $j = 1, \dots, n$ we define a set of indices \mathcal{S}_j such that for $j, j' \in \{1, \dots, n\}$,

1. $j' \notin \mathcal{S}_j$ and $j \notin \mathcal{S}_{j'}$ implies X_j and $X_{j'}$ are independent
2. $j_1, j_2, \dots, j_\ell \notin \cup_{j'=1}^{\ell'} \mathcal{S}_{j'_i}$ and $j'_1, j'_2, \dots, j'_{\ell'} \notin \cup_{j=1}^{\ell} \mathcal{S}_{j_i}$ implies $\{X_{j_1}, \dots, X_{j_k}\}$ independent of $\{X_{j'_1}, \dots, X_{j'_k}\}$

We refer to data as sparsely correlated if we can choose the \mathcal{S}_j , $j = 1, \dots, n$ such that $Mm = O(n)$ where $M = \max_j |\mathcal{S}_j|$, $j = 1, \dots, n$ and m is the size of the largest subset \mathcal{T} of $\{1, \dots, n\}$ such that $j \notin \mathcal{S}_{j'}$ and $j' \notin \mathcal{S}_j$ for all pairs $(j, j') \in \mathcal{T}$. In this paper, we use the independence conditions only for $\ell, \ell' \leq 2$. Use of larger values would allow control of higher moments of sums of sparsely correlated variables and may have other applications.

By definition, \mathcal{S}_j must contain at least all observations pairwise correlated with X_j . In the HIV genetic variation application, for example, we define \mathcal{S}_j as the set of distances sharing at least one time point with distance indicator j .

Continuing with the genetic variation example, let T be the number of independent viral populations, R the number of independent sequences sampled from each of these viral populations, and S the number of sites in each of these sequences. It follows that there are $TR(R-1)S/2$ and $T(T-1)R^2S/2$ unique distances which compare sequences from distinct viral populations, and the total number of observations included in the model is then $n = TR(R-1)S/2 + T(T-1)R^2S/2$.

Given one observation $d(G_{uj}^{(s)}, G_{tk}^{(s)})$ comparing different sequences from two viral populations u and t , we know that $G_{uj}^{(s)}$ will appear in $TR - 1$ different distances when doing all within and between time point comparisons, and $G_{tk}^{(s)}$ will appear in $TR - 1$ different distances. Thus, the maximum number of observations correlated with a single distance observation when including all sites in the model is $M = ((TR - 1) + (TR - 1) - 1)S = (2TR - 3)S$. Finally, assuming that R is even, we have that there are $R/2$ distances corresponding to a single time point and site which are multivariate independent (i.e. the distances are not based on any of the same sequences). Therefore, the largest subset of independent observations since we do not assume independence across sites is $m = TR/2$.

It follows that $mM/n = (2TR - 3)/(TR - 1)$ which is bounded in probability if either

1. The number of independent viral populations increases and the number of sequences sampled from each of these viral populations is bounded,
2. The number of independent sequences sampled from each viral population increases and the number of sampled viral populations is bounded, or
3. Both the number of independent viral populations sampled and the number of independent sequences sampled from each of these viral populations increases.

Under at least one of these conditions, $mM = O(n)$ as $m \rightarrow \infty$.

3.2 Proofs

Before presenting the proofs of the limit theorems, we begin by considering a simple example to motivate the normalizing constants. Suppose

$$X_{ij} = \eta_i + \zeta_j + \epsilon_{ij}$$

where $\{\eta_i\}_{i=1}^k$, $\{\zeta_j\}_{j=1}^K$, and $\{\epsilon_{ij}\}_{(i,j)=(1,1)}^{(k,K)}$ are each independent and identically distributed with $k \leq K$, and that everything has finite variance. In this classical crossed random effects model, $m = k$, $M = k + K - 1$, and $n = kK < mM$. Assume that $k/K \rightarrow C \in [0, 1]$.

Then

$$\sum_{i,j} X_{ij} = K \sum_{i=1}^k \eta_i + k \sum_{j=1}^K \zeta_j + \sum_{(i,j)=(1,1)}^{(k,K)} \epsilon_{ij}$$

so

$$\begin{aligned} \frac{\sqrt{m}}{n} \sum_{i,j} X_{ij} &= \frac{\sqrt{k}K}{n} \sum_{i=1}^k \eta_i + \frac{\sqrt{k}k}{n} \sum_{j=1}^K \zeta_j + \frac{\sqrt{k}}{n} \sum_{(i,j)=(1,1)}^{(k,K)} \epsilon_{ij} \\ &= \frac{1}{\sqrt{k}} \sum_{i=1}^k \eta_i + \frac{1}{\sqrt{K}} \sqrt{\frac{k}{K}} \sum_{j=1}^K \zeta_j + o_p(1) \\ &\xrightarrow{d} N(0, \text{var}[\eta] + C\text{var}[\zeta]) \end{aligned}$$

by the classical central limit theorem, suggesting that \sqrt{m}/n is the correct normalizing sequence and that the rate of convergence is only $1/\sqrt{m}$.

We present three lemmas needed to prove the central limit theorem. Lemma 1 is adapted from Ibragimov & Linnik (1971). It allows us to prove the central limit theorem for bounded variables and extend it by truncation.

Lemma 1 *Suppose $Z_j, j = 1, 2, \dots, n$ is a sequence of mean zero random variables with*

$$\|Z\|_{2+\delta} \equiv \sup_j \|Z_j\|_{2+\delta} < L$$

for some $\delta > 0$ and $L < \infty$ and define truncated variables $Y_j^{(K)} = Z_j \{|Z_j| < K\}$. If for every $K > 0$,

$$\frac{\sqrt{m}}{n} \sum_{j=1}^n Y_j^{(K)} \xrightarrow{d} N(0, 1) \tag{3.1}$$

then

$$\frac{\sqrt{m}}{n} \sum_{j=1}^n Z_j \xrightarrow{d} N(0, 1).$$

Proof Let $0 < \delta' < \delta$. Then

$$\begin{aligned} \text{var} [Z_j - Y_j^{(K)}] &\leq \|Z_j - Y_j^{(K)}\|_{2+\delta'}^2 \\ &\leq K^{\frac{2}{2+\delta'}(\delta'-\delta)} (\|Z\|_{2+\delta}^2)^{\frac{2+\delta}{2+\delta'}} \end{aligned}$$

This is a bound uniform in n , going to zero as $K \rightarrow \infty$. Now if equation 3.1 holds for all K , it also holds for some $K_n \rightarrow \infty$. For this sequence,

$$\frac{\sqrt{m}}{n} \sum_{j=1}^n Y_j^{(K_n)} \xrightarrow{d} N(0, 1)$$

and

$$\frac{\sqrt{m}}{n} \sum_{j=1}^n (Z_j - Y_j^{(K_n)}) \rightarrow 0$$

in mean square. So

$$\frac{\sqrt{m}}{n} \sum_{j=1}^n Z_j \xrightarrow{d} N(0, 1).$$

■

The following is Lemma 2 of Bolthausen (1982).

Lemma 2 *Let ν_n be a sequence of probabilities over \mathbb{R} which satisfies*

1. $\sup_n \int x^2 d\nu_n(x) < \infty$ and
2. For all $\lambda \in \mathbb{R}$,

$$\lim_n \int (i\lambda - x)e^{i\lambda x} d\nu_n(x) = 0.$$

Then

$$\nu_n \xrightarrow{d} N(0, 1).$$

Lemma 3 is the crux of the proof for both the central limit theorem and consistency of the sandwich estimator. The method was originally used by Bolthausen (1982) to give a simple proof of a central limit theorem for strong mixing stationary random fields and derives from ideas of Stein (1972). A similar result for non-stationary strong mixing random fields was used by Guyon (1995) to prove a central limit theorem and adapted by Lumley (1998) to prove consistency of a sandwich estimator.

Lemma 3 Let X_j for $j = 1, \dots, n$ be a sequence of sparsely correlated mean zero random variables. Assume $m \rightarrow \infty$ and $mM = O(n)$.

If

$$\|X\|_4 \equiv \sup_j \|X_j\|_4 < L$$

then

$$\text{var} \left[\frac{m}{n^2} \sum_{k,j} w_{kj} X_k X_j \right] < \frac{4L^4}{m} \rightarrow 0$$

where $w_{kj} = 1$ if $X_k \in \mathcal{S}_j$ and $w_{jk} = 0$ otherwise.

Proof: Define $S_n = \sum_{j=1}^n X_j$ and

$$S_{j,n} = \sum_{k \in \mathcal{S}_j} X_k.$$

First, note that

$$\begin{aligned} \text{var} \left[\frac{m}{n^2} \sum_{k,j} w_{kj} X_k X_j \right] &= \text{var} \left[\frac{m}{n^2} \sum_{j=1}^n X_j S_{j,n} \right] \\ &= \frac{m^2}{n^4} \sum_{j,j'=1}^n \sum_{k \in \mathcal{S}_j, k' \in \mathcal{S}_{j'}} \text{cov} [X_j X_k, X_{j'} X_{k'}] \end{aligned}$$

by definition of $S_{j,n}$. The covariance term would be equal to zero if (X_j, X_k) is independent of $(X_{j'}, X_{k'})$ which implies $X_j X_k$ is independent of $X_{j'} X_{k'}$. Thus, a covariance term could be nonzero if $j' \in \mathcal{S}_j, j' \in \mathcal{S}_k, k' \in \mathcal{S}_j$, or $k' \in \mathcal{S}_k$. An upper bound for the number of nonzero terms is then $nM(4M)M = 4nM^3$ as there are n choices for j , at most M choices for k given j , $4M$ ways that (j', k') and (j, k) can be linked, and M choices for k' given j' . Each covariance term is bounded above such that

$$\begin{aligned} \text{cov} [X_j X_k, X_{j'} X_{k'}] &= \mathbf{E}[X_j X_k X_{j'} X_{k'}] \\ &\leq \|X\|_4^4 \end{aligned}$$

$$< L^4$$

and we therefore have that

$$\begin{aligned} \text{var} \left[\frac{m}{n^2} \sum_{j=1}^n X_j S_{j,n} \right] &\leq \frac{m^2}{n^4} 4nM^3 L^4 \\ &= 4L^4/m \end{aligned}$$

which goes to zero as $m \rightarrow \infty$. ■

The central limit theorem for sparsely correlated data follows. The central limit theorem also implies a weak law of large numbers which we will use extensively.

Theorem 4 *Let X_j for $j = 1, \dots, n$ be a sequence of sparsely correlated mean zero random variables. Let $S_n = \sum_{j=1}^n X_j$, and $\sigma_n^2 = \text{var}[S_n]$. Assume $m \rightarrow \infty$ and $mM = O(n)$. If*

$$\|X\|_{2+\delta} \equiv \sup_j \|X_j\|_{2+\delta} < L$$

for some $\delta > 0$ and $L < \infty$ then

1. Rate of Convergence $\limsup_n m\sigma_n^2/n^2 < \infty$, and

2. Central Limit Theorem: *If in addition*

$$\liminf_n m\sigma_n^2/n^2 > 0, \tag{3.2}$$

then

$$\bar{S}_n \equiv S_n/\sigma_n \xrightarrow{d} N(0, 1).$$

Proof: To prove the first claim, write

$$\sigma_n^2 = \sum_{j=1}^n \sum_{k \in S_j} \text{cov}[X_k, X_j].$$

There are at most Mn summands, each bounded by L^2 , and so

$$\sigma_n^2 < MnL^2 = O(n^2L^2/m).$$

Now let $\tilde{\sigma}_n^2 = m\sigma_n^2/n^2$, the normalized variance of S_n . By applying Lemma 1 with $Z_j = X_j/\tilde{\sigma}_n$, it suffices to prove the second claim for bounded variables. From now on, we assume that X_j is bounded by L .

We now use Lemma 2. The first condition of the lemma is certainly satisfied since X_j is bounded so it is needed to show that

$$\mathbf{E} \left[(i\lambda - \bar{S}_n) e^{i\lambda \bar{S}_n} \right] \rightarrow 0$$

for all real λ .

Following Guyon (1995, p114), we decompose this as

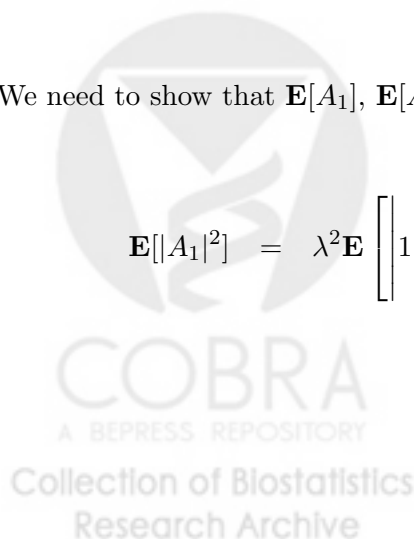
$$(i\lambda - \bar{S}_n) e^{i\lambda \bar{S}_n} = A_1 - A_2 - A_3$$

where

$$\begin{aligned} A_1 &= i\lambda e^{i\lambda \bar{S}_n} \left(1 - \sigma_n^{-2} \sum_{j=1}^n X_j S_{j,n} \right) \\ A_2 &= \sigma_n^{-1} e^{i\lambda \bar{S}_n} \sum_{j=1}^n X_j \left(1 - i\lambda \bar{S}_{j,n} - e^{-i\lambda \bar{S}_{j,n}} \right) \\ A_3 &= \sigma_n^{-1} \sum_{j=1}^n X_j e^{i\lambda (\bar{S}_n - \bar{S}_{j,n})}. \end{aligned}$$

We need to show that $\mathbf{E}[A_1]$, $\mathbf{E}[A_2]$, and $\mathbf{E}[A_3]$ go to zero. First, note that $|e^{i\lambda \bar{S}_n}|=1$ and

$$\mathbf{E}[|A_1|^2] = \lambda^2 \mathbf{E} \left[\left| 1 - \sigma_n^{-2} \sum_{j=1}^n X_j S_{j,n} \right|^2 \right] = \lambda^2 \text{var} \left[\sigma_n^{-2} \sum_{j=1}^n X_j S_{j,n} \right]$$



$$\begin{aligned}
&= \lambda^2 \left(\frac{m^2}{n^4} \sigma_n^4 \right)^{-1} \text{var} \left[\frac{m}{n^2} \sum_{j=1}^n X_j S_{j,n} \right] \\
&= O\left(\frac{1}{m^2 M^2}\right) \\
&\rightarrow 0.
\end{aligned}$$

Hence, $\mathbf{E}[A_1] \rightarrow 0$.

For A_2 , first observe that

$$\begin{aligned}
\bar{S}_{j,n} = \frac{\sum_{k \in \mathcal{S}_j} X_k}{\sigma_n} &< \frac{MK}{\sigma_n} \\
&\leq \frac{cK}{\sqrt{m}} \\
&\rightarrow 0
\end{aligned}$$

for some c and all m as $m \rightarrow \infty$. The first inequality comes from the fact that the maximum number of observations in \mathcal{S}_j is M and the maximum value of X_j is K since X is bounded. The second inequality follows from the assumption given in Equation 3.2 of this theorem.

By a Taylor expansion of $e^{-i\lambda\bar{S}_{j,n}}$, we can show that

$$\left| 1 - i\lambda\bar{S}_{j,n} - e^{-i\lambda\bar{S}_{j,n}} \right| \leq c\lambda^2 \bar{S}_{j,n}^2$$

for some $c > 0$ and all n . So,

$$\begin{aligned}
\mathbf{E}[|A_2|] &= \sigma_n^{-1} \mathbf{E} \left[\left| \sum_{j=1}^n X_j \left(1 - i\lambda\bar{S}_{j,n} - e^{-i\lambda\bar{S}_{j,n}} \right) \right| \right] \\
&\leq \sigma_n^{-1} \mathbf{E} \left[\sup_n c\lambda^2 \bar{S}_{j,n}^2 \right] \mathbf{E} \left[\left| \sum_{j=1}^n X_j \right| \right] \\
&\leq \sigma_n^{-1} \mathbf{E} \left[\sup_n c\lambda^2 \bar{S}_{j,n}^2 \right] nK
\end{aligned}$$

$$\begin{aligned}
&= O\left(\frac{1}{\sqrt{Mn}} \frac{n}{m}\right) \\
&\rightarrow 0.
\end{aligned}$$

Thus $\mathbf{E}[A_2] \rightarrow 0$. Lastly, $\mathbf{E}[A_3] = 0$ since X_j and $\bar{S}_n - \bar{S}_{j,n}$ are independent.

So $\mathbf{E}[A_1 - A_2 - A_3] \rightarrow 0$ for all λ , and by Lemma 2, $\bar{S}_n \xrightarrow{d} N(0, 1)$ completing the proof. ■

Remark: An inspection of the proof shows that if $M_j = |\mathcal{S}_j|$ for $j \in \mathcal{T}$ defined in Section 3.1, we can replace $M = \max_j M_j$ by

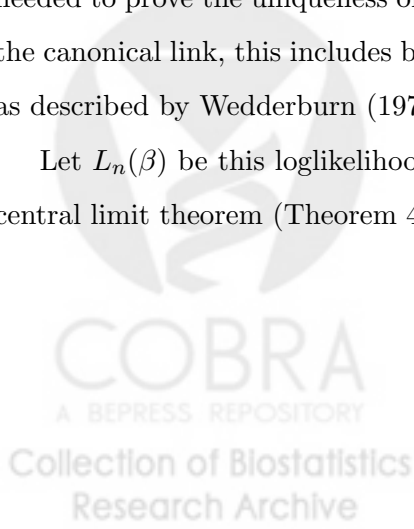
$$M = \left(\frac{1}{m} \sum_{j=1}^m M_j^4\right)^{1/4}.$$

This is useful when the correlation structure is random, as in the case of clustered data with random cluster sizes, and there is no uniform upper bound on M_j .

4 MARGINAL GENERALIZED LINEAR MODELS

We fit the marginal generalized linear model by maximizing the independence working loglikelihood function that would be the loglikelihood if the data were independent and from the appropriate exponential family. We confine our attention to generalized linear models for which this loglikelihood under independence is concave. This restriction is only needed to prove the uniqueness of the parameter estimates. In addition to any model using the canonical link, this includes binomial regression models with the linear and probit links as described by Wedderburn (1976).

Let $L_n(\beta)$ be this loglikelihood and $\ell_j(\beta)$ be the contribution from observation j . The central limit theorem (Theorem 4) shows that $L_n(\beta)/n$ converges in probability for each β



to a function $L_0(\beta) = \lim_{n \rightarrow \infty} E [L_n(\beta)/n]$ and thus by Lemma 5 below,

$$\hat{\beta}_n = \operatorname{argmax} L_n(\beta)$$

is consistent for

$$\beta_0 = \operatorname{argmax} L_0(\beta).$$

The remaining step in proving consistency is to show that β_0 as just defined is the true regression coefficient. This follows because $U_j(\beta) = \partial \ell_j(\beta) / \partial \beta$ is a linear function of $Y_j - \mu_j$ and so is zero at the true value of β . Thus, $E[\ell_j(\beta)]$ has a maximum at the true value of β for all j and so $L_0(\beta)$ has its unique maximum at the true value of β .

Lemma 5 *Let Θ be an open convex subset of \mathbb{R}^p and $f_n : \Theta \rightarrow \mathbb{R}$ be a sequence of random convex functions of θ . Define*

$$\hat{\theta}_n = \operatorname{argmax}_{\Theta} f_n(\theta).$$

If $f_n(\theta) \xrightarrow{p} f(\theta)$ then $\hat{\theta}_n \xrightarrow{p} \operatorname{argmax}_{\Theta} f(\theta)$.

A proof of Lemma 5 is given by Andersen & Gill (1982, Appendix II). Asymptotic normality of $\hat{\beta}_n$ follows from the central limit theorem and classical assumptions about the smoothness of L_n . We use the following result which is Theorem 3.4.5 of Guyon (1995).

Theorem 6 *Suppose that $\hat{\beta}_n$ minimizing a random function of β_n , K_n , is consistent for $\beta_0 \in \Theta_\beta \subset \mathbb{R}^p$ and that*

6.1 *There exists a neighborhood V of β_0 over which*

- (a) *K_n is twice continuously differentiable, and*
- (b) *There exists an integrable random variable H such that for all $\beta \in V$ and for k ,*
 $j = 1, \dots, p$

$$\left| \left(\ddot{K}_n \right)_{jk} \right| \leq H.$$

6.2 There exists a sequence $\langle a_n \rangle \rightarrow \infty$ such that $J_n = \text{var} \left[\sqrt{a_n} \dot{K}_n(\beta_0) \right]$ exists and:

- (a) There exists a positive definite matrix J with $J_n \geq J$ for all large enough n , and
- (b)

$$\sqrt{a_n} J_n^{-1/2} \dot{K}_n(\beta_0) \xrightarrow{d} N(0, 1_p)$$

where 1_p is a $p \times p$ identity matrix.

6.3 There exists a sequence of deterministic $p \times p$ matrices $\langle I_n \rangle$ such that

- (a) $\left(\ddot{K}_n(\beta_0) - I_n \right) \xrightarrow{p} 0$, and
- (b) There exists a positive definite matrix I with $I_n - I$ positive semidefinite for all large enough n .

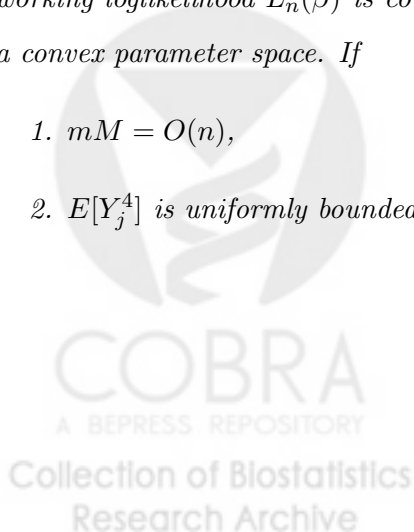
Then

$$\sqrt{a_n} J_n^{-1/2} I_n \left(\hat{\beta}_n - \beta_0 \right) \xrightarrow{d} N(0, 1_p).$$

We apply this theorem to the function $K_n(\beta) = -L_n(\beta)/n$. We will consider only bounded predictors although this restriction can be relaxed for specific link and variance function combinations.

Theorem 7 Suppose that Y_j , $j = 1, 2, \dots, n$ is a sparsely correlated sequence satisfying a marginal generalized linear model with predictors X_j taking values in a bounded subset of \mathbb{R}^p . Suppose the link and variance functions have three continuous derivatives, the independence working loglikelihood $L_n(\beta)$ is convex, and that the true parameter β_0 is in the interior of a convex parameter space. If

1. $mM = O(n)$,
2. $E[Y_j^4]$ is uniformly bounded,



3. There exists a vector W and a positive definite matrix T such that

$$\begin{aligned}\frac{1}{n} \sum_{j=1}^n X_j &\equiv \bar{X}_n \rightarrow W \\ \frac{1}{n} \sum_{j=1}^n X_j X_j' &\equiv T_n \rightarrow T,\end{aligned}$$

and

4.

$$\limsup \frac{1}{m} \text{var} \left[\sum_{i=1}^m Y_i \right] > 0$$

then

$$\sqrt{m} (\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0, \Xi)$$

where $\hat{\beta}_n$ maximizes $L_n(\beta)$ and

$$\begin{aligned}m \hat{\Xi}_n(\hat{\beta}_n) &= m \left(\sum_{j=1}^n \frac{\partial U_j}{\partial \beta} \right)^{-1} \left(\sum_{j=1}^n \sum_{k \in \mathcal{S}_j} U_j(\hat{\beta}_n) U_k(\hat{\beta}_n)^T \right) \left(\sum_{j=1}^n \frac{\partial U_j}{\partial \beta} \right)^{-1} \\ &\xrightarrow{p} \Xi.\end{aligned}$$

If the link and variance functions are twice continuously differentiable, then Conditions 6.1(a) and 6.1(b) hold for any bounded neighborhood N_0 where $V(\cdot)$ and $g'(\cdot)$ are bounded away from zero.

To prove Condition 6.3, note that

$$\begin{aligned}E[\ddot{K}_n(\beta)|X] &= E \left[\frac{1}{n} \dot{U}_n(\beta) | X \right] = \frac{1}{n} \sum_{j=1}^n \frac{g'(\mu_j(\beta))^2}{V(\mu_j(\beta))} X_j X_j^T \\ &= \frac{1}{n} \sum_{j=1}^n w_j(\beta) X_j X_j^T\end{aligned}$$

where $w_j(\beta)$ is bounded above and below for $\beta \in N_0$. Assumption 3 in this theorem now implies that $\ddot{K}_n(\beta)$ converges in probability, and Assumption 2 and boundedness of X_j imply that this convergence also holds in mean, as required.

To verify Condition 6.2, we take $a_n = m$ and apply the central limit theorem (Theorem 4) to the elements of $\sum_{j=1}^n U_j/n$. Using the Cramér-Wold device, we can consider only scalars whereby Theorem 4 applies to the vector $U_j(\beta)$ if and only if it applies to the scalars $b'U_j(\beta)$ for all vectors b of norm 1. The moment condition of the central limit theorem follows with $\delta = 2$ from boundedness of X and the assumption that $E[Y_j^4]$ is uniformly bounded, and the variance lower bound (Equation 3.2 in Theorem 4) comes from Assumptions 3 and 4 in this theorem. Theorem 4 gives us that

$$\left(\frac{1}{n} \sum_{j=1}^n U_j(\beta) \right) \left(\frac{1}{n} \sum_{j=1}^n U_j(\beta) \right)^T$$

converges in probability to a positive definite limit, and again Assumption 2 and the boundedness of X imply that the expectation also converges, as required for Condition 6.2. We note for later use that this convergence is true not only at $\beta = \beta_0$ but for all $\beta \in N_0$.

So by Theorem 6, we see that

$$\sqrt{m}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0, \Xi)$$

for some positive definite variance matrix Ξ .

For this result to be useful for inference, we must be able to estimate Ξ consistently. The sandwich estimator $\hat{\Xi}_n$ can be rewritten as

$$\begin{aligned} m\hat{\Xi}_n(\beta) &= \left(\frac{\sqrt{m}}{n} \sum_{j=1}^n \frac{\partial U_j}{\partial \beta} \right)^{-1} \left(\frac{m}{n^2} \sum_{j=1}^n \sum_{k \in \mathcal{S}_j} U_j(\beta) U_k(\beta)^T \right) \left(\frac{\sqrt{m}}{n} \sum_{j=1}^n \frac{\partial U_j}{\partial \beta} \right)^{-1} \\ &= A_n^{-1}(\beta) B_n(\beta) A_n^{-1}(\beta) \end{aligned}$$

evaluated at $\beta = \hat{\beta}_n$.

Since $E[m\hat{\Xi}_n(\beta_0)] = \Xi$ it suffices to show that $m\hat{\Xi}_n(\beta)$ converges in probability to a continuous function of β uniformly over a neighborhood of β_0 . Consistency of $\hat{\beta}_n$ then gives consistency of $\hat{\Xi}(\hat{\beta}_n)$. We work with A_n and B_n separately.

First we show that $A_n(\hat{\beta}_n)$ is consistent. As argued for Condition 6.3 in the proof of asymptotic Normality of $\hat{\beta}_n$, $A_n(\beta)$ converges pointwise in β to $A(\beta) = \lim_n A_n(\beta)$. A straightforward calculation shows that the derivative of $A_n(\beta)$ is bounded over N_0 so that $A_n(\beta)$ is uniformly equicontinuous and the limit function $A(\beta)$ is continuous. Consistency of $\hat{\beta}_n$ implies that $A_n(\hat{\beta}_n)$ is consistent for $A(\beta_0)$ and as this is positive definite as shown for condition 6.3(b), $A_n(\hat{\beta}_n)^{-1}$ is consistent for $A(\beta_0)^{-1}$.

To show that $B_n(\hat{\beta}_n)$ is consistent, note that by Assumption 2 and boundedness of X and the fact that $V(\mu_j)$ is bounded away from zero uniformly in β , we have $E[U_j^4(\beta)]$ is uniformly bounded for $\beta \in N_0$. Applying Lemma 3,

$$\text{var}[B_n(\beta)] < \frac{4}{m} \sup_{j, \beta \in N_0} E[U_j(\beta)^4]$$

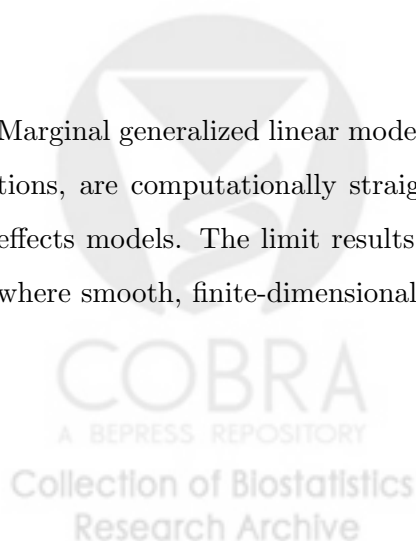
so $B_n(\beta)$ converges in mean square uniformly over $\beta \in N_0$. As $B_n(\beta)$ is continuous for each n , the uniform limit $B(\beta)$ is continuous on N_0 . Finally, as $\hat{\beta}_n \xrightarrow{p} \beta_0$, $B_n(\hat{\beta}_n) \xrightarrow{p} B(\beta_0)$.

So the sandwich estimator $m\hat{\Xi}_n(\hat{\beta}_n)$ is consistent for Ξ and we can use $\hat{\Xi}_n(\hat{\beta}_n)$ as an estimator of $\text{var}[\hat{\beta}_n]$. ■

Remark: Assumption 3 in Theorem 7 is stronger than the conditions imposed by Fahrmeir & Kauffman (1985), but is satisfied by many reasonable fixed or random designs. For example, if the X_j are sparsely correlated with a common marginal distribution that does not concentrate along a lower-dimensional subspace of \mathbb{R}^p , then Assumption 3 follows from the central limit theorem.

5 CONCLUSIONS

Marginal generalized linear models for sparsely correlated data require fairly weak assumptions, are computationally straightforward, and provide a useful complement to random effects models. The limit results we have presented here are applicable to similar models where smooth, finite-dimensional parameters are to be estimated. Extensions of empirical



process central limit theorems based on entropy to sparsely correlated data would widen the class of models that could be used, as would extensions of the GEE methodology using working models other than independence.

6 ACKNOWLEDGEMENTS

Part of this work is based on the PhD dissertations of the authors, supported by predoctoral fellowships from the Howard Hughes Medical Institute and the University of Washington Center for AIDS Research NIH grant 5T32AI07140, and supervised by Patrick Heagerty and Steve Self, respectively.

REFERENCES

- ANDERSEN, P. K. & GILL, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics* **10**, 1100–1120.
- BOLTHAUSEN, E. (1982). On the C.L.T. for stationary mixing random fields. *Annals of Probability* **10**, 1047–1050.
- BRESLOW, N. E. & CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.
- CLAYTON, D. & ÉCOCHARD, R. (1997). Artificial insemination by donor: Discrete time survival data with crossed and nested random effects. In Lin, D. Y. & Fleming, T. R., editors, *Proceedings of the First Seattle Symposium in Biostatistics*, volume 123 of *Lecture Notes in Statistics*, pages 99–122. Springer.
- DE LA PEÑA, V. & GINÉ, E. (1999). *Decoupling: from dependence to independence*. Springer, New York.
- FAHRMEIR, L. & KAUFFMAN, H. (1985). Consistency and asymptotic Normality of the maximum likelihood estimator in generalized linear models. *Annals of Statistics* **13**, 342–368.

- FELSENSTEIN, J. (1988). Phylogenies from molecular sequences: Inference and reliability. *Annual Review of Genetics* **22**, 521–565.
- GUYON, X. (1995). *Random Fields on a Network: Modeling, Statistics and Applications*. Springer-Verlag, New York. (translated by Carenne Ludeña).
- HILLIS, D. M., MORITZ, C., & MABLE, B. K. (1996). *Molecular Systematics*. Sinauer Associates, Inc.
- HUBER, P. J. (1967). The behaviour of maximum likelihood estimators under non-standard conditions. In LeCam, L. M. & Neyman, J., editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 221–233. University of California Press.
- IBRAGIMOV, I. A. & LINNIK, Y. V. (1971). *Independent and Stationary Sequences of Random Variables*. Wolters-Noordhoff, Groningen.
- KARIM, M. R. & ZEGER, S. L. (1992). Generalized linear models with random effects; Salamander mating revisited. *Biometrics* **48**, 631–644.
- LELE, S. (1991). Jackknifing linear estimating equations: Asymptotic theory and applications in stochastic processes. *Journal of the Royal Statistical Society, Series B* **53**, 253–267.
- LIANG, K.-Y. & ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.
- LIN, D. Y. & WEI, L. J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association* **84**, 1074–1078.
- LUMLEY, T. (1998). *Marginal Regression Modelling of Weakly Dependent Data*. PhD thesis, University of Washington.
- LUMLEY, T. & HEAGERTY, P. J. (1999). Weighted empirical adaptive variance estimators for correlated data regression. *Journal of the Royal Statistical Society, Series B* **61**, 459–477.
- MAYER-HAMBLETT, N. (1999). *A Regression Modeling Approach for Describing Patterns*

- of HIV Genetic Variation*. PhD thesis, University of Washington, Seattle, WA.
- MCCULLAGH, P. & NELDER, J. (1989). *Generalised Linear Models*. Monographs on Statistics and Applied Probability. Chapman and Hall, London, 2nd edition.
- MIYAMOTO, M. M. & CRACRAFT, J. (1991). *Phylogenetic Analysis of DNA Sequences*. Oxford University Press.
- NELSON, J. C. (1999). *A Graphical Method for Describing Interrater Variability in Ordinal Assessments Among Many Raters*. PhD thesis, University of Washington, Seattle, WA.
- NEWBY, W. K. & WEST, K. D. (1987). A simple, positive definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55**, 703–708.
- NOLAN, D. & POLLARD, D. (1987). *U*-processes: Rates of convergence. *Annals of Statistics* **15**, 780–799.
- RASBASH, J. & GOLDSTEIN, H. (1994). Efficient analysis of mixed hierarchical and cross classified random structures using a multilevel model. *Journal of Educational and Behavioural Statistics* **19**, 337–350.
- ROYALL, R. M. (1986). Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review* **54**, 221–226.
- SHANKARAPPA, R., MARGOLICK, J. B., RODRIGO, A. G., GANGE, S. J., UPCHURCH, D., FARZADEGAN, H., GUPTA, P., RINALDO, C. R., LEARN, G. H., HUANG, X. L., & MULLINS, J. I. (2000). Viral evolution in the asymptomatic phase of HIV-1 infection. *Submitted*.
- SHARPLES, K. & BRESLOW, N. (1992). *Regression analysis of correlated binary data: Some small sample results for the estimating equation approach*. *Journal of Statistical Computation and Simulation* **42**, 1–20.
- SHUN, Z. (1997). *Another look at the Salamander mating data: A modified Laplace approximation approach*. *Journal of the American Statistical Association* **92**, 341–349.
- SIMPSON, E. H. (1949). *The measurement of diversity*. *Nature* **163**, 688.
- STEIN, C. (1972). *A bound for the error in normal approximation of a sum of dependent*

random variables. In Proceedings of the 6th Berkeley Symposium in Mathematical Statistics and Probability, pages 583–603.

WEDDERBURN, R. W. M. (1974). *Quasilikelihood functions, generalized linear models and the Gauss–Newton method.* *Biometrika* **61**.

WEDDERBURN, R. W. M. (1976). *On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models.* *Biometrika* **63**, 27–32.

WHITE, H. (1984). *Asymptotic Theory for Econometricians.* *Academic Press, Orlando, Florida.*

