Generalised Linear Models for Sparsely Correlated Data

Thomas Lumley

tlumley@u.washington.edu

Biostatistics, University of Washington

- GLMs
- Sparse Correlation
- Technical difficulties
- Applications: two-index asymptotics for GEE
- Applications: modelling HIV genetic diversity
- Empirical process limit theorems?

GLMs

Generalised Linear Models are popular for independent observations. They have a model for a function of the mean

$$g(E[Y_i|X_i]) = g(\mu) = \eta = \beta' X_i$$

and then either

- Y_i has an exponential family distribution
- var$[Y_i|X_i] = V(\mu)$

These lead to the same estimators

Correlated data

For correlated data the choice between **semiparametric** and **parametric** matters: parametric models typically need the dependence to be correctly specified. There is also another choice

- Model the **marginal mean** $E[Y_i|X_i]$ of observed response **conditional on observed predictors**
- Model the mean of Y conditional on **observed predictors** X_i and enough **unobserved variables** b_i to make Y conditionally independent.

This is logically independent of the choice between parametric and semiparametric estimation; methods exist for all four possibilities. It is easier to construct conditionally specified likelihoods and easier to estimate marginal means, so conditional models tend to be parametric and marginal ones semiparametric.

Replication and Correlation

Statistical methods rely on **replication**

> *A function of a large number of independent random variables that does not depend too much on any one of them is approximately constant.* Talagrand (1998)

- independent or longitudinal data have genuine independent replicates
- time series, spatial data often have approximate replicates: widely separated pieces of space or time are approximately independent. The various **mixing coefficients** specify senses in which functions of many time points are similar to functions of many independent variables (eg bounds on correlation, conditional expectation, likelihood ratio)

Sparse correlation examples

Three examples (only one is really useful) that motivated sparse correlation:

- McCullagh & Nelder's salamander mating data: a study of salamanders of the same species from two geographically distinct areas. The scientific question is whether salamanders from the same location are more likely to mate than those from different locations. The observations are of pairs of salamanders, so the data are not longitudinal or independent.
    - A crossed random effects logistic model is one possibility (used by Breslow & Clayton 1993)
    - A marginalised random effects model (Heagerty)
    - A Bayesian model, with the marginal contrasts estimated from the posterior MCMC sample.
    - ?Logistic regression with sandwich variances

Sparse correlation examples

- Nicole Mayer–Hamblett & Steve Self analysed changes in the HIV genome over time within infected individuals. Binary data on locations where two samples differ. Again, the data come from pairs of observations. Not all pairs of observations will contribute, as within-time and between-time differences are modelled separately.
- Jennifer Nelson's PhD thesis was on graphical diagnostics for interrater agreement. She needed to prove that some ordinal logistic spline models worked. Interrater agreement studies involve multiple raters and multiple images (eg mammograms). Observations on the same image or by the same rater will be correlated. The study will often not be complete (every rater with every image) and might not even be balanced

In all these examples many observations are pairwise independent, but there are no true independent replicates.

Sparse Correlation: simpler cases

- For a balanced crossed linear regression design it is fairly easy to fit a random effects model, and the marginal and conditional contrasts are the same. Unbalanced designs are computationally more difficult.
- Asymptotic theory for a generalised linear mixed model should be relatively straightforward, since the likelihood can be written in terms of iid random effects (though I haven't seen it done)
- When the data are from pairs of individuals in a complete balanced design, U-statistic theory will describe the properties of estimators very elegantly.
- Bayesians can fit a random effects model by MCMC. Maximum likelihood is much harder.

Digression: What are U-statistics?

Statistics of the form

$$\sum_{i,j=1}^{n} h(X_i, X_j)$$

(where $h(\cdot, \cdot)$ is usually taken to be antisymmetric) are called U-statistics (of order 2).
They are not iid sums, but often behave like them. There is a very comprehensive theory of U-statistics, but it does not extend easily to other cases of sparse correlation. The theory is based on tricks for adding more randomness to U-statistics to get iid sums, and relies strongly on the symmetry of the definition.

Sparse Correlation: definition

Each observation has a **neigbourhood $\mathcal{S}_i$**.

- Observations i and j are independent if $i \notin S_j$
- Sets of observations A and B are independent if no $i \in A$ is in $\{\cup_j \mathcal{S}_j \mid j \in B\}$

Define

- M as the size of the largest neighbourhood
- m as the size of the largest set of points with each not in the neighbourhood of any other.

A sequence of such datasets is **sparsely correlated** if $mM = O(n)$.
Heuristically, any two small sets of observations are probably independent. Implies (but is stronger than) the condition that the proportion of non-zero elements of the covariance matrix goes to zero.

Example of definition

Mammography inter-rater reliability study

- Neigbourhood of (rater r, image i) is all readings with rater r or image i
- Two reading are independent if they have different raters and images
- Two pairs of readings are independent if none of the raters or images in one pair are in the other pair.
- M=raters per image $+$ images per rater $-$ 1
- m=minimum of number of raters, number of images.

Marginal models by quasilikelihood

Same approach as for longitudinal data. The model is

$$g\left(E\left[Y_i|X_i\right]\right) \equiv g(\mu) = \beta' X$$

and we estimate by solving

$$\sum_i U_i \equiv \frac{\partial \mu_i}{\partial \beta} w_i (Y_i - \mu_i) = 0$$

for β, where w_i are weights that ideally are close to $1/\text{var}[Y_i]$. More generally we allow a non-diagonal weight matrix W and solve

$$D^T W(Y - \mu) = 0$$

where $D = \frac{\partial \mu_i}{\partial \beta}$. **This commits us to modelling $E[Y_i|X]$ not $E[Y_i|X_i]$.**

Difficulties

- Asymptotic behaviour of the estimates is not obvious if m, M both are large.
- A non-diagonal weight matrix W may be inconveniently large, especially if calculated as V^{-1} for some approximation to var$[Y]$.
- Standard error estimation is not so obvious either.

Solutions: standard errors

Standard error estimation by sandwich estimator

$$\text{var}[\hat{\beta}] = I^{-1} J I^{-1}$$

where

$$I = \left. \frac{\partial \sum_i U_i}{\partial \beta} \right|_{\hat{\beta}}$$

and

$$J = \sum_{i;\, j \in \mathcal{S}_i} U_i(\hat{\beta}) U_j^T(\hat{\beta}).$$

The restriction in the sum for J is important: the sum over all i, j is identically zero.

Basically the same estimator as longitudinal, spatial, time series data. The proof involves counting up fourth moment terms.

Solutions: standard errors

- In the simplest case where the data are correlated on two non-nested factors the computation is easy in standard software that can do longitudinal data analysis (eg Stata)

$$I^{-1}JI^{-1} = I^{-1}J_1I^{-1} + I^{-1}J_2I^{-1} - I^{-1}J_{12}I^{-1}$$

  where $I^{-1}J_1I^{-1}$ and $I^{-1}J_2I^{-1}$ are the sandwich estimators clustering on each of the factors separately and $I^{-1}J_{12}I^{-1}$ is the sandwich estimator clustering on the product of both factors.

- Note that if factor 2 is nested in factor 1 the second and third terms here cancel, reducing correctly to the sandwich estimator clustering on factor 1.

Solutions: standard errors

In other correlated data settings the sandwich estimator is a subsampling estimator applied to the estimating functions

- **longitudinal data**
- **time series:** the Newey–West estimator is a subsampling estimator applied to estimating functions
- **spatial data:** window subsampling of the estimating functions gives a weighted sandwich estimator (Heagerty & Lumley *JASA* 2000).
- **sparse correlation:** subsampling of **neighbourhoods** $\mathcal{S}_j$ gives a slightly different (weighted) sandwich estimator. In a crossed design it would be $IJ_1^{-1}I + I^{-1}J_2I^{-1}$, without the correction for double-counting. This is guaranteed positive definite, but is likely to be less efficient.

Solutions: large matrices

- $W = V^{-1}$ doesn't have to be calculated explicitly since we only need

  $$DV^{-1}$$

  found by solving p linear equations.
- Iterative sparse matrix techniques (preconditioned conjugate gradient) can do this in about $O(npM \log M)$ time, rather than the $O(n^3)$ of matrix inversion.

Solutions: Consistency

To think about the asymptotics, consider a simple two-way design

$$Y_{ij} = \mu + \eta_i + \zeta_j + \epsilon_{ij}$$

with $i = 1, \ldots, k$, $j = 1, \ldots, K$, $k < K$

$$\begin{aligned} \eta_i &\sim [0, \sigma^2_\eta] \\ \zeta_i &\sim [0, \sigma^2_\zeta] \\ \epsilon_i &\sim [0, \sigma^2_\epsilon] \end{aligned}$$

What do we need for $\bar{Y} - \mu$ to be asymptotically Normal?
If $k \to \infty$, $k/K \to C \in [0, 1]$

$$\frac{\sqrt{k}}{kK} \left(\sum Y - \mu \right) \to_d N\left(0, \sigma^2_\eta + C\sigma^2_\zeta\right)$$

Asymptotics

- In the two-way design we have $m = k$, $M = k + K - 1$, $n = kK$. A reasonable guess is that in general

  $$\frac{\sqrt{m}}{n}\left(\sum Y - \mu\right) \to_d N(0, \sigma^2)$$

  under $m \to \infty$, some conditions on M and some moment restrictions on Y. This is true for the two-way design and for independent data (with $M = 1$, $m = n$).
- In fact Mm/n bounded, $m \to \infty$, and bounded $4 + \delta$ moments are sufficient. Proof uses Stein's method for CLT: a bound on the error in the characteristic function involving fourth moments. These conditions are not necessary but weaker ones of the same sort probably are.
- Given a CLT, standard methods show that GLMs work for sparsely correlated data.

Two-index asymptotics for GEE

- The theory for marginal general linear models for longitudinal data (GEE) is usually described in terms of a fixed maximum number of observations per person ℓ and an increasing number of people G.
- Not clear how well it works when ℓ is large. Do dentists need larger samples than opthalmologists?
- Longitudinal data with increasing ℓ and G is sparsely correlated: $m = \ell$, $M = G$, $mM = n$.
- Independence working model should have no problems with even very large clusters, number of clusters really is most important parameter.
- Other working models need a little more care: need to show that the weight matrix converges fast enough to a constant.

Back to GLMs and HIV

- Nicole Mayer-Hamblett & Steve Self studied the **evolution of HIV within the body** as disease progresses from initial diagnosis to AIDS, using data from the Multicenter AIDS Cohort Study. Published in *Biometrics*, June 2001
- Multiple HIV isolates were sequenced at multiple points in time, and the main interest was in **how the diversity within the population and the distance from the initial population varied**. In the future this might be augmented with other information such as drug regimens.
- These genetic distance measures are **sparsely correlated binomial data**, indicating the proportion of base pairs or of codons that differ.

HIV diversity over time

HIV infection is initially well-controlled by the immune system, but eventually escapes control and causes AIDS. According to one theoretical account we should see three phases

- Initially the diversity of the virus increases and the ‘quasi-species’ diffuses away from the initial form
- Later the diversity of the virus stabilises but it still continues to move away from the starting form
- Finally, the distance to the starting form stabilises and the diversity remains stable or decreases as successful mutants dominate the population.

Fitting linear splines to the data within each person allows these three phases to be identified. Our theory for sparsely correlated data allows testing and confidence intervals. Previous analyses had been unable to get valid tests.

In Progress: Empirical Process CLT

Empirical process limit theorems are a useful gadget for demonstrating convergence of estimating functions or objective functions **uniformly in the parameters** even for infinite-dimensional parameters.
Two main flavours

- Uniform entropy: define a ‘dimension’ in terms of how many little balls of radius ϵ it takes to cover all the functions. This can’t increase too fast with n
- Bracketing entropy: Come up with uniformly good finite approximations to the functions. The standard proof of the Glivenko–Cantelli theorem is a bracketing argument.

Uniform entropy arguments are very unfriendly to dependent data, bracketing arguments aren’t too bad.

Bracketing CLT

- Needs a tail bound for the sum of observations. This is turned into a bound for $E[|\sup_f f(Y)|]$ for finite sets of functions, and then a very tricky recursive finite approximation and truncation method is used to do the hard work (fortunately this is fairly standardised).
- For iid observations the tail bound is Bernstein's inequality

$$P\left[\sum_i Y_i > t\right] \leq 2e^{-\frac{t^2}{at+b}}$$

for $a = \max|Y|_\infty$, $b = \sum_i$ var$[Y_i]$. The iid result is Ossiander (*Ann. Prob.* 1987); van der Vaart's (2000) book has the simplest presentation.

Bracketing CLT

For sparsely correlated data a version of Bernstein's inequality still holds: need to bound moments of the sum in terms of same moments for independent data.

- Typical term is $E|Y_{i_1}Y_{i_2}\cdots Y_{i_r}|$. If an index i_k is in the neighbourhood of an earlier index, set it to be the same.
- Each term is now a product of things that are either the same or not in the same neighbourhood. Suppose each index appears an even number of times. Then the term is bounded by the same term for independent data.
- If an index appears an odd number of times, combine it with another index that appears an odd number of times. This is always possible for even moments, and we can bound odd moments by next higher even ones.
- Count up how many different terms get mapped to the same term for independent data to give a bound.

Summary

- Marginal modelling of sparsely correlated data is (computationally) very easy, even for unbalanced designs
- The sandwich estimator and its relationship to subsampling of estimating functions is a really useful idea