Chapter 8

Further Research

8.1 Future research issues

This thesis has covered a variety of topics and several different avenues of further study are possible. Three particular areas of interest are:

1. Exponential family distributions and conditional score functions.

2. Rectangular array asymptotics.

3. Generalized linear model development.

Taking these in order, the exponential family distributions offer some interesting research problems. Specifically, the substitution projective scores are "closer" to the true score than in an arbitrary non-exponential family model. If the model is canonically parameterized as

$$f(x, \psi, \lambda) = \exp(\psi x + \lambda S - \tau(\psi, \lambda)) \tag{8.1}$$

then a complete sufficient statistic for λ with ψ fixed is $S = S(x)$. The Bhattacharyya basis elements are *polynomials* in S and likewise the projective score U_t, which is a polynomial of degree t in S.

By exploiting the polynomial structure one can show that in Type I models

Proposition 8.1

$$n^{\frac{1}{2}(t-1)}\left(U_t(\hat{\lambda}_\psi) - U_t(\lambda_0)\right) \xrightarrow{\ell} Y$$

where $E[Y] = K$ *and* $\mathrm{Var}[Y] < \infty$.

If t is set equal to 2, then the result states that the inflation factor needed to obtain a limiting distribution is $\sqrt{n}$. Contrast this to the general result of Theorem 4.5 in which the inflation factor needed to obtain the limiting distribution is n^0, in other words 1. This is the sense in which the substitution version of the projective score is closer to the true score in these Type I exponential family problems. In fact, the variance term in the limiting distribution given by Theorem 4.5 is 0 for Type I models. It would be interesting to explore this feature in more detail, and determine the practical implications of this result. Better rates are also possible in some problems with additional structure; the Circular Normal problem is an example of this, where symmetry and group properties come into play.

The rectangular array asymptotics are an area that deserve considerably greater attention than was possible in this thesis. Firstly, the necessity of the "data splitting" technique used to form the modified estimating equation

$$U^i_{2M}(\hat{\lambda}^i_\psi, \mathbf{x}_i) = U^i_2(\hat{\lambda}^i_\psi(\mathbf{x}_{i2}), \mathbf{x}_{i1}) + U^i_2(\hat{\lambda}^i_\psi(\mathbf{x}_{i1}), \mathbf{x}_{i2}), \tag{8.2}$$

needs to be assessed. The conditions under which the expectation

$$E\left[(\hat{\lambda}^i_\psi - \lambda^i_0)^k U^i_{2,k}(\lambda^i_0)\right], \quad k = 0, 1, 2,$$

is negligible compared to the expectation of

$$(\hat{\lambda}^i_\psi - \lambda^i_0)^3 U^i_{2,3}(\lambda^i_0)$$

should be investigated, since this is the critical result used in Proposition 5.2. If the necessary conditions turn out to be reasonable then the data splitting argument could be dispensed with, which from a computational viewpoint is time consuming because everything has to be calculated twice.

Another area of interest would be an exploration of the spectrum of rectangular array asymptotics. In this thesis, only the case where $p = n_p$ was investigated. The situation in which $n_p = p^\alpha$, with $0 < \alpha < 1$, so that the stratum sizes grow only slowly with p needs to be developed. Also, the consideration of using estimates of the nuisance parameters that are *not* maximum likelihood estimates warrants some consideration. One would hope for a tradeoff between the level of unbiasedness of U_t and the rate of convergence of the estimates for the nuisance parameters. Specifically, with a high enough level of unbiasedness for U_t then crude estimates of the nuisance parameters may be acceptable in the sense that full efficiency is possible for the estimates of the nuisance parameters. This would be useful because in problems with many strata, it is undesirable to spend large computational resources on estimating the nuisance parameters, which are by definition, not of any intrinsic interest.

Finally, the Generalized Linear Model considered in Chapter 7 provides a large source of further research issues. The development of software to implement the model as its stands is a first priority. A version for the Poisson case is already available, but the Gamma and Binomial extensions have not yet been implemented. The intention is to write code in the `C++` programming language, that will be accessible through a variety of interfaces including `S-Plus`, `SAS` and `Mathematica`.

The special case of Logistic Regression needs to be probed. Recall that the proof of Lemma 7.2 is not valid for this model, even though the statement of the theorem appears correct according to empirical investigations. It will be interesting to make comparisons with other techniques available for *exact* Conditional Logistic Regression, for example the approaches offered by Mehta, Patel, and Gray (1985).

For the models to be really useful, a move away from the *saturated* framework needs to be made. This is meant in the sense that contrasts of the data, $\mathbf{Z}$ and $\boldsymbol{\Gamma}$, which do not necessarily span $\mathbb{R}^n$ must be accommodated. The results of Rao and Yanai (1979) appear potentially useful to develop the appropriate projection theory.

As a final point this entire endeavor may be thought of a staging post to a more ambitious modelling framework in which the models incorporate *random effects* which may be allowed to come from an arbitrary mixing distribution.