Likelihood corrections for models with many nuisance parameters.

RICHARD P. WATERMAN

Department of Statistics, University of Pennsylvania

BRUCE G. LINDSAY

Department of Statistics, Pennsylvania State University

September 1995

Introduction

Aim: To explore some models that lack an exact conditional structure.

For certain exponential response models using a **semi-parametric** mixture model provides inference identical to a **conditional** analysis. An example is the Rasch model [**2**].

$$X_{i1} \sim Bernoulli\left(\frac{\exp(\psi_1 + \lambda_i)}{1 + \exp(\psi_1 + \lambda_i)}\right)$$
$$X_{i2} \sim Bernoulli\left(\frac{\exp(\psi_2 + \lambda_i)}{1 + \exp(\psi_2 + \lambda_i)}\right)$$

Conditioning out the λ_i or assuming that $\lambda \sim G$ where G is an unspecified mixing distribution leads to identical inference for ψ_1, ψ_2.

These simple models may not fit, but more complex ones

- Lack a conditional structure
- Pose difficulties for a semi-parametric analysis.

Issue: Use a projection technique for mimicking a conditional analysis. How does it compare to the semi-parametric approach or a parametric mixed model?

A testing data set

Questions from a multiple choice statistics exam given to 670 Wharton MBA's in October 1994. Item response curves from two questions are in Figure 1.

One attribute of the Rasch model is that response curves should not cross. A hypothesis test [3] on the two items displayed in the figure indicates a lack of fit for the Rasch.

How should we handle this? Generalize the model or consider another link function?

Define p_{ij} as $P(X_{ij} = 1 \,|\, \lambda_i)$

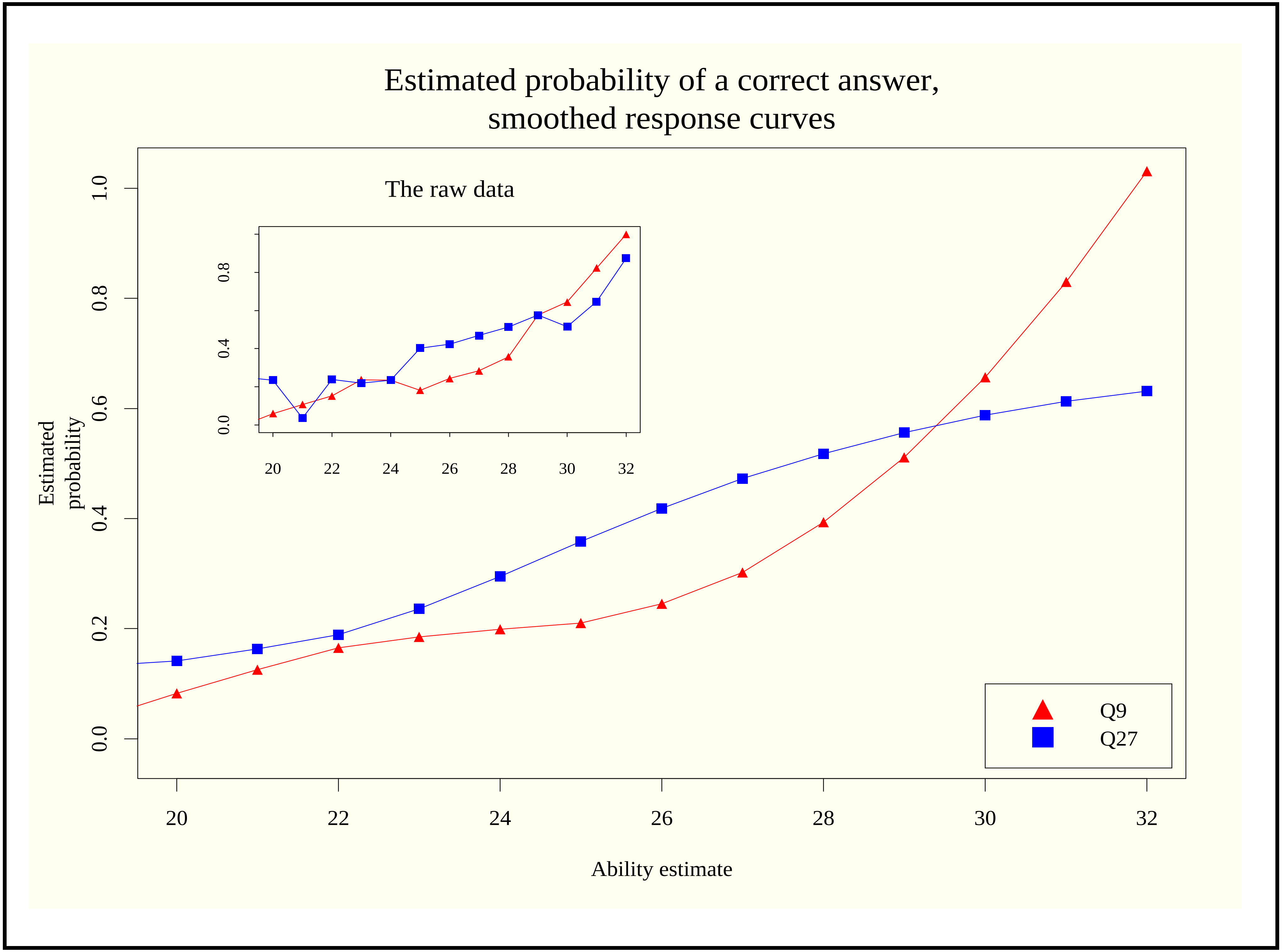
- 2-parameter Rasch.

$$p_{ij} = logit(\psi_{1j} + \psi_{2j}\lambda_i)$$

- Complementary log-log link.

$$p_{ij} = 1 - \exp\{-\exp(\psi_{1j} + \lambda_i)\}$$

The proliferation of the λ_i (one for each subject) causes serious problems for standard maximum likelihood estimation.


Estimated probability of a correct answer,
smoothed response curves
The raw data
Estimated probability
Ability estimate
0.0
0.2
0.4
0.6
0.8
1.0
20
22
24
26
28
30
32
Q9
Q27

The technique for mimicking conditional inference

An optimal estimating equation (one that mimics the properties of the conditional score function) is

$$U_\infty = U_0 - \Pi_\infty U_0$$

where U_0 is the score function for ψ and Π_∞ is the projection operator onto space spanned by the centered likelihood ratios. The function U_∞ can be represented as

$$U_\infty = U_0 - \sum_{\alpha=1}^{\infty} \rho_{\alpha\infty} V_\alpha^*, \qquad (1)$$

where V_α^* is an orthogonalized version of V_α and $V_\alpha = f^{(\alpha)}/f$. The $\rho_{\alpha\infty}$ are regression coefficients.

The function is **optimal** in the sense that it maximizes information in the class of functions satisfying

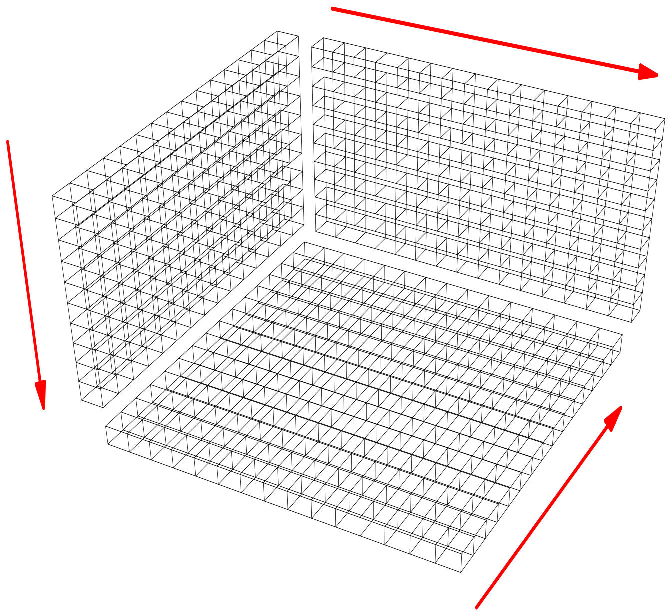
$$E_\lambda\{g(\psi_0, \lambda_0)\} = 0, \text{ for all } \lambda_0, \psi_0, \lambda.$$

Furthermore if a conditional score exists, U_c, then $U_c = U_\infty$. Natural approximations to U_∞, *projected scores* [4,6] are given by truncating the series (1).

- $U_1 = U_0 - \rho_{11} V_1$
- $U_2 = U_0 - \rho_{12} V_1 - \rho_{22} V_2$

A general framework for analysis

Thinking of the data, $\{x_{ijk}\}$, as lying inside a cubic array leads to various asymptotics.



The likelihood, conditional on λ_i

$$f(x;\psi,\lambda) = \prod_{ijk} f(x_{ijk};\lambda_i,\psi_j).$$

Asymptotics: allow r, c, n to go to infinity in different ways.

- Fix r, c. $\quad n \to \infty$. Standard.
- Fix c, n. $\quad r \to \infty$. Neyman-Scott.
- Fix c. $\quad n, r \to \infty, r/n \to Constant$. e.g. $n = r$; Square array [7].

Implementation issues

Calculating the ρ_{it}.

How can the regression coefficients, ρ_{it}, be calculated? A symbolic software package is needed. The coefficients are **complicated** and need to be output in a compiled language, e.g. Fortran or C. The package `Format.m`,

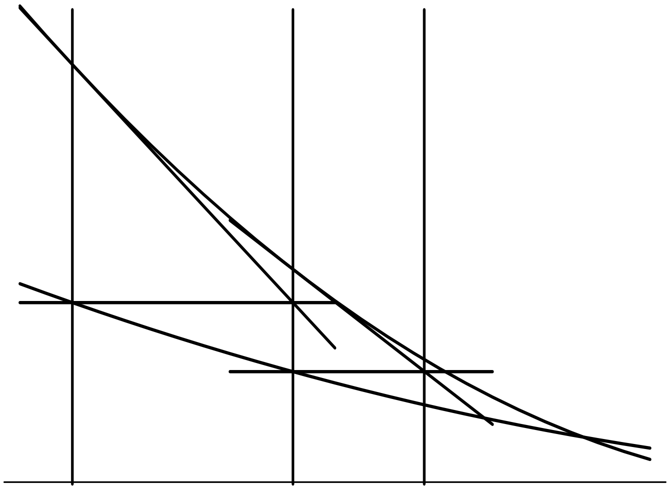
> An efficient symbolic-numeric environment by extending `Mathematica's` format rules

was used [5]. Opposite is the function $E(V_4 V_4)$ for the complementary log-log link in full detail. It must be re-calculated many times which raises the issue of what constitutes a computationally efficient factorization.

Raw fortran code

```fortran
C     Find other elements of E(V4 V4)
      o1=a+u
      o2=exp(2.d0*o1)
      o3=exp(o1)
      o4=exp(o3)
      o5=n(i)
      o6=exp(3.d0*o1)
      o7=exp(4.d0*o1)
      o8=exp(5.d0*o1)
      o9=exp(6.d0*o1)
      o10=a+o3+u
      o11=exp(2.d0*o10)
      o12=2.d0*a+o3+2.d0*u
      o13=exp(o12)
      o14=3.d0*a+o3+3.d0*u
      o15=exp(o14)
      o16=4.d0*a+o3+4.d0*u
      o17=exp(o16)
      o18=5.d0*a+o3+5.d0*u
      o19=exp(o18)
      o20=6.d0*a+o3+6.d0*u
      o21=exp(o20)
      o22=3.d0*a+2.d0*o3+3.d0*u
      o23=exp(o22)
      o24=4.d0*a+2.d0*o3+4.d0*u
      o25=exp(o24)
      o26=5.d0*a+2.d0*o3+5.d0*u
      o27=exp(o26)
      o28=6.d0*a+2.d0*o3+6.d0*u
      o29=exp(o28)
      o30=o5**2
      t1=o2/(-1.d0+o4)**4
      t2=-1.d0-2.81d2*o11+3.79d2*o13-1.266d3*o15+2.27d3*o17-1.908d
     & 3*o19
      t3=-1.59d2*o2+6.31d2*o21+7.62d2*o23-9.94d2*o25+5.4d2*o27-1.0
     & 1d2*o29
      t4=1.4d1*o3+2.16d2*o17*o30-4.32d2*o19*o30+2.16d2*o21*o30+3.d
     & 0*o4+9.8d1*o11*o5
      t5=-1.96d2*o13*o5+1.008d3*o15*o5-2.336d3*o17*o5+2.304d3*o19*
     & o5+9.8d1*o2*o5-8.44d2*o21*o5-5.04d2*o23*o5
      t6=8.44d2*o25*o5-5.04d2*o27*o5+9.8d1*o29*o5+5.9d2*o6-5.04d2*
     & o5*o6-1.326d3*o7
      t7=-2.16d2*o30*o7+1.492d3*o5*o7+1.38d3*o8+4.32d2*o30*o8-1.8d
     & 3*o5*o8-6.75d2*o9
      t8=-3.6d2*o30*o9+1.01d3*o5*o9+2.4d1*o5**3*o9
      t9=-4.2d1*exp(o10)-8.6d1*exp(3.d0*o10)-3.d0*exp(2.d0*o3)
      t10=exp(3.d0*o3)+4.2d1*exp(a+2.d0*o3+u)-1.4d1*exp(a+3.d0*o3+
     & u)
      t11=6.1d1*exp(2.d0*a+3.d0*o3+2.d0*u)
      t12=5.d1*exp(4.d0*a+3.d0*o3+4.d0*u)
      t13=-1.2d1*exp(5.d0*a+3.d0*o3+5.d0*u)+exp(6.d0*a+3.d0*o3+6.d
     & 0*u)
      EV4V4(j)=o5*t1*(t10+t11+t12+t13+t2+t3+t4+t5+t6+t7+t8+t9)
```

Solving the modified score equations.



Technique: modified Newton-Raphson using ∇U_0, **not** $\nabla(U_0 - \rho_{12} V_1 - \rho_{22} V_2)$.
The reason is that U_0 is easy to differentiate but the correction term is too involved. Block structures in information matrices are exploited when solving for each λ_i.

Geometry of the complementary log-log link

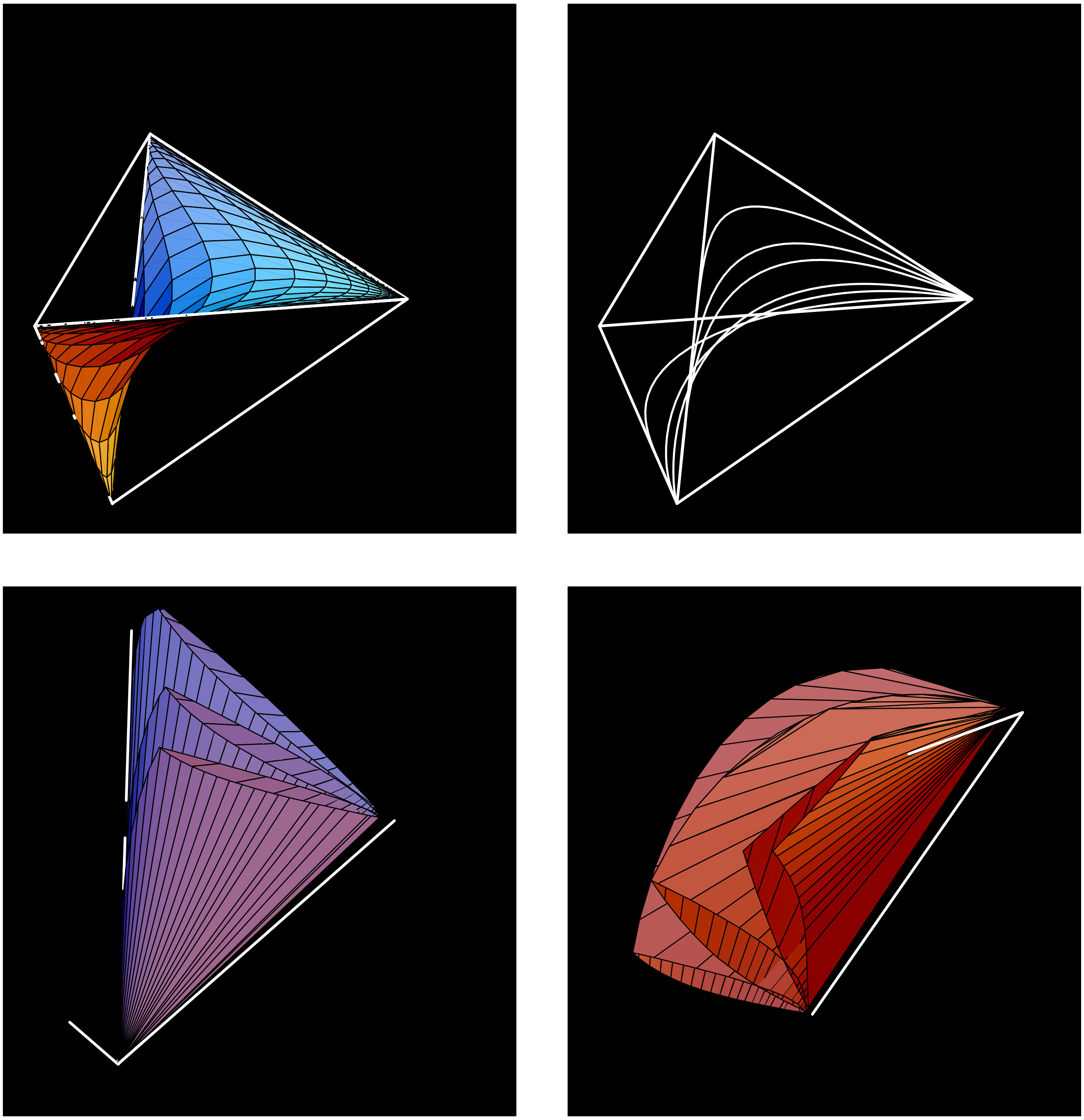


Figure 4a: the surface of all possible probability distributions $\{(1-p_{11})(1-p_{12}),(1-p_{11})p_{12},p_{11}(1-p_{12}),p_{11}p_{12}\}$. Figure 4b: contours on the surface for various fixed ψ. Figures 4c and 4d: convex hulls associated with the contours.

The hulls have **volume**, therefore they overlap, therefore ψ is **not identifiable** in the semi-parametric model [1]. Similarly in the 2-parameter Rasch.

The 2-parameter Rasch

Examples: 2 parameter Rasch. MBA data
Comparison between projected, mixed normal
and non-parametric mixture.

Param.	Normal Mixture		U_2		2 pt. Mixture	4 pt. Mixture
	Est.	S.D.	Est.	S.D.	Est.	Est.
σ	0.831	.187	—	—	0.750	0.959
μ	-0.982	.1086	—	—	-0.972	-0.994
ψ_{21}	0.003	.208	-0.135	.162	0.095	-0.008
ψ_{22}	0.556	.218	0.371	.157	0.626	0.510
ψ_{31}	0.494	.266	0.471	.189	0.515	0.472
ψ_{32}	0.839	.281	0.798	.199	0.867	0.813
ψ_{41}	0.392	.228	0.390	.159	0.390	0.398
ψ_{42}	0.651	.230	0.636	.161	0.653	0.657
ψ_{51}	2.502	.335	2.392	.279	2.456	2.375
ψ_{52}	0.493	.272	0.387	.210	0.468	0.377
ψ_{61}	0.886	.233	0.877	.180	0.877	0.876
ψ_{62}	0.608	.227	0.586	.168	0.606	0.598
ψ_{71}	2.297	.367	2.365	.346	2.354	2.357
ψ_{72}	0.938	.323	0.962	.276	1.003	0.991
ψ_{81}	1.828	.305	1.872	.295	1.936	1.919
ψ_{82}	0.808	.278	0.821	.251	0.913	0.886
ψ_{91}	2.936	.430	3.049	.477	2.778	3.062
ψ_{92}	1.088	.351	1.134	.344	1.002	1.194
$\psi_{10\,1}$	3.060	.442	2.968	.388	3.334	3.105
$\psi_{10\,2}$	0.818	.345	0.740	.275	1.051	0.854

Simulations for the log-log link

Using higher order scores. Can the projected scores remove enough bias when there are many nuisance parameters?
Comparison between n = 1, n = 2, n = 5, U_1, U_2. ~2500 $\lambda's$, $\dim(\psi) = 4$, reps = 100. True values of ψ are $\{1.0, 2.0, -1.0, -2.0\}$.

n = 1	U_2		U_3		U_4	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
ψ_2	1.0743	0.0610	1.0372	0.0551	1.0951	0.0644
ψ_3	1.9431	0.0707	2.0818	0.0743	1.9697	0.0635
ψ_4	-1.0714	0.0661	-0.9261	0.0548	-0.9809	0.0645
ψ_5	-2.4439	0.1075	-2.0018	0.0780	-1.9176	0.0847

n = 2	U_2		U_3		U_4	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
ψ_2	1.0195	0.0386	1.0163	0.0378	1.0145	0.0377
ψ_3	1.9607	0.0516	2.0154	0.0560	1.9810	0.0529
ψ_4	-1.0229	0.0398	-0.9989	0.0385	-1.0231	0.0384
ψ_5	-2.0869	0.0465	-1.9682	0.0428	-1.9699	0.0418

n = 5	U_1		U_2		U_3		U_4	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
ψ_2	1.091	0.027	1.002	0.024	1.003	0.024	1.000	0.024
ψ_3	2.206	0.040	1.993	0.035	2.005	0.035	1.996	0.035
ψ_4	-1.080	0.024	-1.010	0.023	-1.005	0.023	-1.007	0.023
ψ_5	-2.139	0.028	-2.011	0.026	-1.990	0.026	-1.997	0.026

Theoretical justification for U_2 in the square array

Consider the solution to U_1 or U_2 as $\psi(F) = \psi(F_1, \ldots, F_r)$. Then expand $\psi(F)$ evaluated at $\psi(\hat{F})$ about $\psi(F_0)$.

Technique: von-Mises expansion. For the i^{th} row define

$$F_i(\epsilon_i) = F_{i0} + \epsilon_i \sqrt{n_i}(\hat{F}_i - F_{i0}),$$

so that the derivative with respect to ϵ_i is

$$\sqrt{n_i}(\hat{F}_i - F_i) = \Delta_{in}.$$

The second order von-Mises expansion is defined by

$$\psi\left(\frac{1}{\sqrt{n_1}}, \ldots, \frac{1}{\sqrt{n_p}}\right) - \psi(0, \ldots, 0) = \sum_k \psi'_k\bigg|_{\epsilon_k=0} \frac{1}{\sqrt{n_k}} +$$

$$\frac{1}{2}\sum_{kl} \psi''_{kl}\bigg|_{\epsilon_k=\epsilon_l=0} \frac{1}{\sqrt{n_k}}\frac{1}{\sqrt{n_l}} + Rem(\hat{F} - F_0)$$

Key fact in square array $r/n \to const$

If use maximum likelihood (U_1), $\hat{\psi}$, then

$$\sum_{kl} \psi''_{kl}\Big|_{\epsilon_k=\epsilon_l=0} \xrightarrow{p} Bias.$$

If use projected (U_2), $\tilde{\psi}$, then

$$\sum_{kl} \psi''_{kl}\Big|_{\epsilon_k=\epsilon_l=0} \xrightarrow{p} 0.$$

Consequences

First order effects in this asymptotics.

$$\sqrt{n}\sqrt{r}(\hat{\psi} - \psi) \xrightarrow{\ell} N(Bias, Info),$$

but

$$\sqrt{n}\sqrt{r}(\tilde{\psi} - \psi) \xrightarrow{\ell} N(0, Info),$$

fully efficient, even though nuisance parameters head off to infinity.

Testing. Bias in the maximum likelihood estimator leads to asymptotic non-central χ^2 distributions. But if U_2 is used for inference then one obtains the usual central χ^2 distribution.

References

[1] Mary L. Lesperance and John D. Kalbfleisch. Mixture models for matched pairs. *Can. J. Statist.*, 22:65,74, 1994.

[2] Bruce G. Lindsay, Clifford C Clogg, and John Grego. Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *J. Am. Statist. Assoc.*, 86:96–107, 1991.

[3] Paul R. Rosenbaum. Comparing item characteristic curves. *Psychometrika*, 52:217–233, 1987.

[4] Christopher G Small and D L McLeish. Projection as a method for increasing sensitivity and eliminating nuisance parameters. *Biometrika*, 76(4):693–703, 1989.

[5] Mark Sofroniou. *Symbolic and numerical methods for Hamiltonian systems.* PhD thesis, Loughborough University of Technology, Loughborough , United Kingdom, October 1994.

[6] Richard P. Waterman and Bruce G. Lindsay. A simple and accurate method for approximate conditional inference in generalized linear models. *Forthcoming, JRRSB.*

[7] Richard P. Waterman and Bruce G. Lindsay.
Projected score methods for nuisance parameters:
Asymptotics and Neyman-Scott problems.
Manuscript, 1995.