Pseudo-likelihood Methodology

for Marginally, Conditionally, and Hierarchically Specified Models

Marc Aerts [1] & **Geert Molenberghs** [1,2]

[1] Center for Statistics
Universiteit Hasselt, Belgium
marc.aerts@uhasselt.be
geert.molenberghs@uhasselt.be
www.censtat.uhasselt.be

[2] Biostatistical Centre
Katholieke Universiteit Leuven, Belgium
geert.molenberghs@med.kuleuven.be
www.kuleuven.ac.be/biostat/

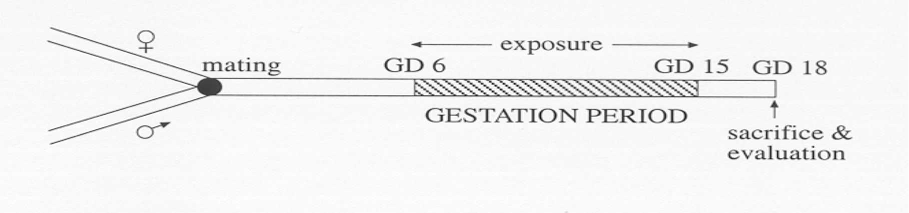
Composite Likelihood, Warwick, 15-17th April 2008

Overview

- **Motivating case study**
- Principles of pseudo-likelihood
  - ▷ Conditional models
  - ▷ Marginal models
  - ▷ Hierarchical models
- Extensions:
  - ▷ Combined continuous and discrete outcomes
  - ▷ High-dimensional outcome
  - ▷ Smooth and additive models
  - ▷ Other: Incomplete data, ...
- Concluding remarks

Teratology (Segment II) Studies

- Time line for a typical Segment II study:



- Exposure of timed-pregnant animals (rats, mice, rabbits) during major organogenesis (days 6–15 for mice and rats)

- Exposure through clinical or environmental routes most relevant for human exposure

- Dams sacrificed just prior to normal delivery

- Uterus removed and thoroughly examined

Data Structure

- Each group: 20 to 30 dams

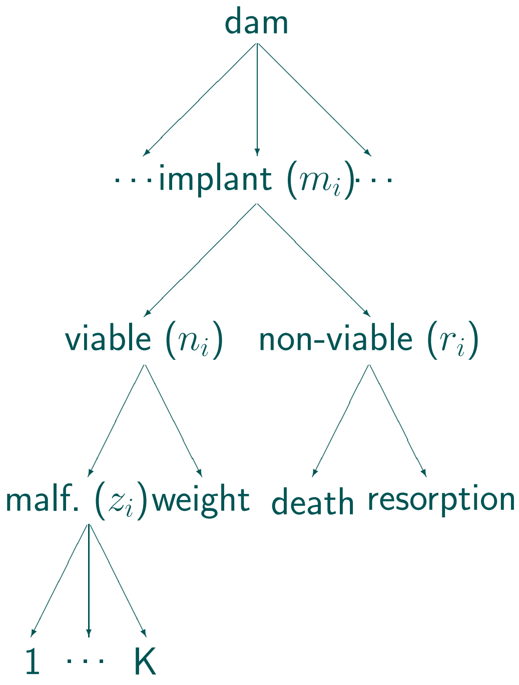
- Offspring per litter:

  2 to 17 foetuses

- Control group &

  3 or 4 dose groups

- Dose is *cluster level covariate*



dam

$\cdots$implant (m_i)$\cdots$

viable (n_i) non-viable (r_i)

malf. (z_i)weight death resorption

1 $\cdots$ K

The National Toxicology Program Studies

Developmental Toxicity Studies

- Research Triangle Institute
- Segment II studies
- The effect in mice of 3 chemicals:
  - ▷ DEHP: di(2-ethyhexyl)-phtalate
  - ▷ EG: ethylene glycol
  - ▷ DYME: diethylene glycol dimethyl ether

Ethylene Glycol $HOCH_2CH_2OH$

- A high-volume industrial chemical with many applications
  - ▷ as an antifreeze in cooling and heating systems
  - ▷ as one of the components of hydraulic brake fluids
  - ▷ as an ingredient of electrolytic condensers
  - ▷ as a solvent in the paint and plastics industries
  - ▷ in the formulation of several types of inks
  - ▷ as a softening agent for cellophane
  - ▷ as a stabilizer for soybean foam used to extinguish oil and gasoline fires
  - ▷ in the synthesis of various chemical products, such as plasticizers, synthetic fibers and waxes
- EG represents little hazard to human health in normal industrial handling
- Accidental or intentional ingestion is toxic and may result in death

NTP Studies in Mice

					Litter			
		# dams, ≥ 1			Size	Malformations		
Exposure	**Dose**	**impl.**	**viab.**	**Live**	**(mean)**	**Ext.**	**Visc.**	**Skel.**
EG	0	25	25	297	11.9	0.0	0.0	0.3
	750	24	24	276	11.5	1.1	0.0	8.7
	1500	23	22	229	10.4	1.7	0.9	36.7
	3000	23	23	226	9.8	7.1	4.0	55.8
DEHP	0	30	30	330	13.2	0.0	1.5	1.2
	44	26	26	288	11.1	1.0	0.4	0.4
	91	26	26	277	10.7	5.4	7.2	4.3
	191	24	17	137	8.1	17.5	15.3	18.3
	292	25	9	50	5.6	54.0	50.0	48.0
DYME	0	21	21	282	13.4	0.0	0.0	0.0
	62.5	20	20	225	11.3	0.0	0.0	0.0
	125	24	24	290	12.1	1.0	0.0	1.0
	250	23	23	261	11.3	2.7	0.1	20.0
	500	22	22	141	6.1	66.0	19.9	79.4

Overview

- Motivating case study
- Principles of pseudo-likelihood
  - ▷ **Conditional models**
  - ▷ Marginal models
  - ▷ Hierarchical models
- Extensions:
  - ▷ Combined continuous and discrete outcomes
  - ▷ High-dimensional outcome
  - ▷ Smooth and additive models
  - ▷ Other: Incomplete data,
- Concluding remarks

A Conditional Model

No Clustering

- Cox (1972)
- $i = 1, \ldots, N$ individuals
- $j = 1, \ldots, M$ evaluations
- $Y_{ij} = 1$ if subject i exhibits response j and 0 otherwise:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}_i; \boldsymbol{\Theta}_i) = \exp\left\{ \sum_{j=1}^{M} \theta_{ij} y_{ij} + \sum_{j<j'} \omega_{ijj'} y_{ij} y_{ij'} + \ldots + \omega_{i12\ldots M} y_{i1} y_{i2} \ldots y_{iM} - A(\boldsymbol{\Theta}_i) \right\}$$

- θ_{ij}: main effect parameters: conditional logits
- $\omega_{ijj'}$: pairwise interactions: conditional log odds ratios

Simplifications

- Zhao and Prentice (1990)
- **Quadratic:**

$$f_{\boldsymbol{Y}}(\boldsymbol{y}_i; \boldsymbol{\Theta}_i) = \exp\left\{ \sum_{j=1}^{M} \theta_{ij} y_{ij} + \sum_{j<j'} \omega_{ijj'} y_{ij} y_{ij'} - A(\boldsymbol{\Theta}_i) \right\}$$

- **Linear $\equiv$ logistic regression:**

$$f_{\boldsymbol{Y}}(\boldsymbol{y}_i; \boldsymbol{\Theta}_i) = \exp\left\{ \sum_{j=1}^{M} \theta_{ij} y_{ij} - A(\boldsymbol{\Theta}_i) \right\}$$

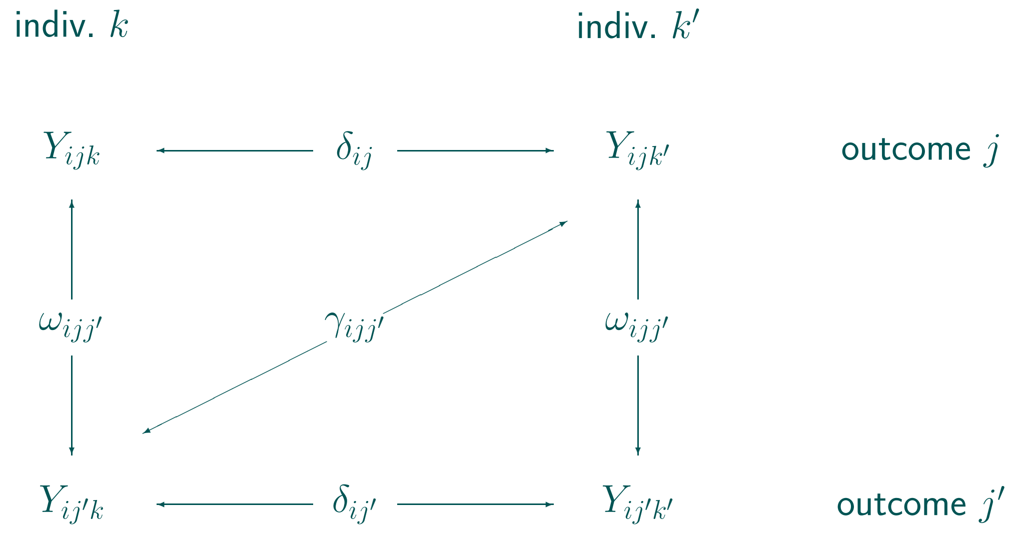
Clustering

- Molenberghs and Ryan (1997)
- $i = 1, \ldots, N$ clusters
- $k = 1, \ldots, n_i$ individuals
- $j = 1, \ldots, M$ malformations

$$Y_{ijk} = \begin{cases} 1 & \text{if } k\text{th individual of } i\text{th cluster exhibits malf. } j \\ \text{-1} & \text{otherwise.} \end{cases}$$

$$f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i; \boldsymbol{\Theta}_i) = \exp\left\{ \sum_{j=1}^{M} \sum_{k=1}^{n_i} \theta_{ij} y_{ijk} + \sum_{j=1}^{M} \sum_{k<k'} \delta_{ij} y_{ijk} y_{ijk'} + \sum_{j<j'} \sum_{k=1}^{n_i} \omega_{ijj'} y_{ijk} y_{ij'k} + \sum_{j<j'} \sum_{k\neq k'} \gamma_{ijj'} y_{ijk} y_{ij'k'} - A(\boldsymbol{\Theta}_i) \right\}$$

Association Structure



A Single Clustered Binary Outcome

- NTP data: Y_{ij} is malformation indicator for fetus j in litter i
- Code Y_{ij} as -1 or 1
- d_i is dose level at which litter i is exposed
- Simplification: $\theta_{ij} = \theta_i = \beta_0 + \beta_d d_i$ and $\delta_{ij} = \delta_i = \beta_a$
- Using

$$Z_i = \sum_{j=1}^{n_i} Y_{ij}$$

we obtain

$$f(z_i|\theta_i, \beta_a) = \binom{n_i}{z_i} \exp\{\theta_i z_i + \beta_a z_i(n_i - z_i) - A(\boldsymbol{\theta}_i)\}$$

ML Estimates for the Conditional Model

Outcome	Parameter	DEHP	EG	DYME
External	β_0	-2.81(0.58)	-3.01(0.79)	-5.78(1.13)
	β_d	3.07(0.65)	2.25(0.68)	6.25(1.25)
	β_a	0.18(0.04)	0.25(0.05)	0.09(0.06)
Visceral	β_0	-2.39(0.50)	-5.09(1.55)	-3.32(0.98)
	β_d	2.45(0.55)	3.76(1.34)	2.88(0.93)
	β_a	0.18(0.04)	0.23(0.09)	0.29(0.05)
Skeletal	β_0	-2.79(0.58)	-0.84(0.17)	-1.62(0.35)
	β_d	2.91(0.63)	0.98(0.20)	2.45(0.51)
	β_a	0.17(0.04)	0.20(0.02)	0.25(0.03)

Several Clustered Binary Outcome

- Simple exponential family expression
- High numerical stability of ML estimators
- **But:**
  - ▷ Evaluation of normalizing constant:
    - ∗ cumbersome expression
    - ∗ excessive time requirements

$\Longrightarrow$ Pseudo-likelihood

Pseudo-likelihood

- Arnold and Strauss (1991)
- Geys, Molenberghs, and Ryan (1999)
- Basic Idea:

$$\boxed{f(x,y) \quad \longleftrightarrow \quad f(x|y) \cdot f(y|x)}$$

$\Longrightarrow$ **Normalizing constant cancels**

A Single Clustered Binary Outcome

Cluster i: $f(y_{i1}, ..., \boxed{y_{ik}}, ..., y_{in_i})$ replaced by full conditionals

$$PL = \prod_{i=1}^{N} \prod_{k=1}^{n_i} f(y_{ik} | \{y_{ik'}\} \text{ for } k' \neq k)$$

Exchangeability $\Longrightarrow$ 2 contributions only:

S	S	. . .	S	F	F	. . .	F

p_{is}: cond. prob. of additional succes, given $z_i - 1$ successes and $n_i - z_i$ failures.

$$\text{logit}(p_{is}) = \theta_i - \delta_i(n_i - 2z_i + 1)$$

p_{if}: cond. prob. of additional failure, given z_i successes and $n_i - z_i - 1$ failures.

$$\text{logit}(p_{if}) = -\theta_i + \delta_i(n_i - 2z_i - 1)$$

Several Clustered Binary Outcomes

Two versions:

- ▷ Interest in main effects only:
  $f(y_{i11}, ..., y_{ij1}, ..., y_{iM1}, ..., y_{i1k}, ..., \boxed{y_{ijk}}, ..., y_{iMk}, ..., y_{i1n_i}, ..., y_{ijn_i}, ..., y_{iMn_i})$ replaced by full conditionals

$$PL(1) = \prod_{i=1}^{N} \prod_{j=1}^{M} \prod_{k=1}^{n_i} f(y_{ijk} | y_{ij'k'}, j' \neq j \text{ or } k' \neq k)$$

- ▷ Interest in main effects and multivariate association:
  $f(y_{i11}, ..., y_{ij1}, ..., y_{iM1}, ..., \boxed{y_{i1k}, ..., y_{ijk}, ..., y_{iMk}}, ..., y_{i1n_i}, ..., y_{ijn_i}, ..., y_{iMn_i})$ replaced by full conditionals

$$PL(2) = \prod_{i=1}^{N} \prod_{k=1}^{n_i} f(y_{ijk}, j = 1, \ldots, M | y_{ijk'}, k' \neq k, j = 1, \ldots, M)$$

Both procedures are roughly equally efficient

General Definition

- General definition: $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N)$ i.i.d. common density depending on $\boldsymbol{\Theta}_0$

$$p\ell \quad := \quad \sum_{i=1}^{N} \sum_{s \in S} \delta_s \ln f_s(y_i^{(s)}; \Theta_i)$$

- Special case: bivariate pseudo-likelihood $\quad f(y_1|y_2)f(y_2|y_1)$

$$\begin{aligned} \delta_{(1,1)} &= \quad 2 \\ \delta_{(1,0)} &= -1 \\ \delta_{(0,1)} &= -1 \end{aligned}$$

- Special case: Likelihood $\quad f(y_1, \ldots, y_n)$

$$\delta_s = \begin{cases} 1 & \text{if } s = (1, \ldots, 1) \\ 0 & \text{otherwise} \end{cases}$$

Asymptotic Results

- $\widetilde{\boldsymbol{\Theta}}_N \xrightarrow{P} \boldsymbol{\Theta}_0$.

- $\sqrt{N}(\widetilde{\boldsymbol{\Theta}}_N - \boldsymbol{\Theta}_0) \xrightarrow{D} N_p(\boldsymbol{0}, J(\boldsymbol{\Theta}_0)^{-1} K(\boldsymbol{\Theta}_0) J(\boldsymbol{\Theta}_0)^{-1})$

$$\begin{aligned} J_{k\ell} &= -\sum_{s \in S} \delta_s E_{\boldsymbol{\Theta}} \left(\frac{\partial^2 \ln f_s(\boldsymbol{y}^{(s)}; \boldsymbol{\Theta})}{\partial \theta_k \partial \theta_\ell} \right) \\ K_{k\ell} &= \sum_{s,t \in S} \delta_s \delta_t E_{\boldsymbol{\Theta}} \left(\frac{\partial \ln f_s(\boldsymbol{y}^{(s)}; \boldsymbol{\Theta})}{\partial \theta_k} \frac{\partial \ln f_t(\boldsymbol{y}^{(t)}; \boldsymbol{\Theta})}{\partial \theta_\ell} \right). \end{aligned}$$

- For likelihood $J^{-1}KJ^{-1}$ reduces to the inverse of Fisher's information matrix I.

- Cramèr-Rao implies $J^{-1}KJ^{-1} \geq I^{-1}$

Statistical Efficiency

- **No clustering:** ARE$= 1$ for all saturated exponential models

- **Clustering:** no explicit ARE expressions $\Longrightarrow$ study via asymptotic samples:
  - ▷ consider all realizations (n_i, z_i, d_i)
  - ▷ specify: $f(d_i)$ & $f(n_i|d_i)$ & $f(z_i|n_i, d_i)$
  - ▷ weigh each realization according to its true probability

- **Conclusions for clustered case:**
  - ▷ high efficiency in realistic parameter settings
  - ▷ poor efficiency when no dose effect and high association
  - ▷ also very good efficiency in small samples
  - ▷ $PL(1)$ and $PL(2)$ are roughly equally efficient

Computational Efficiency

- **Univariate:** approximately equal

- **Bivariate:** PL is 5 times faster than ML

- **Trivariate:** PL is 200 times faster than ML $\longleftarrow$ 3 minutes versus 10 hours!

Univariate Clustered Outcome: ML & PL

Outcome	Par.	DEHP	EG	DYME
External (ML)	β_0	-2.81(0.58)	-3.01(0.79)	-5.78(1.13)
	β_d	3.07(0.65)	2.25(0.68)	6.25(1.25)
	β_a	0.18(0.04)	0.25(0.05)	0.09(0.06)
External (PL)	β_0	-2.85(0.53)	-2.61(0.88)	-5.04(0.94)
	β_d	3.24(0.60)	2.14(0.71)	5.52(1.01)
	β_a	0.18(0.04)	0.30(0.06)	0.13(0.05)
Collapsed (ML)	β_0	-2.04(0.35)	-0.81(0.16)	-2.90(0.43)
	β_d	2.98(0.51)	0.97(0.20)	5.08(0.74)
	β_a	0.16(0.03)	0.20(0.02)	0.19(0.03)
Collapsed (PL)	β_0	-1.80(0.35)	-1.11(0.14)	-3.08(0.47)
	β_d	2.95(0.56)	1.41(0.19)	5.20(0.97)
	β_a	0.20(0.03)	0.21(0.01)	0.19(0.02)

Trivariate Clustered Outcome: PL

Parameter	DEHP	EG	DYME
conditional mean parameters			
β_{01}	-2.10(0.51)	-1.97(0.56)	-3.89(0.83)
β_{02}	-2.42(0.50)	-2.96(0.87)	-4.77(0.87)
β_{03}	-2.74(0.49)	-0.27(0.55)	-3.21(0.81)
β_d	2.67(0.48)	1.50(0.20)	4.31(0.85)
Association parameters			
δ_1	0.14(0.07)	0.18(0.13)	0.22(0.03)
δ_2	0.18(0.04)	0.17(0.17)	0.25(0.06)
δ_3	0.29(0.05)	0.20(0.01)	0.25(0.02)
ω_{12}	0.06(0.24)	-0.05(0.57)	-0.46(0.19)
ω_{13}	0.60(0.20)	0.11(0.30)	0.29(0.30)
ω_{23}	0.36(0.28)	0.97(0.37)	0.28(0.31)
γ_{12}	0.11(0.06)	0.13(0.13)	0.05(0.04)
γ_{13}	-0.06(0.05)	0.06(0.04)	-0.09(0.04)
γ_{23}	-0.14(0.06)	-0.07(0.03)	-0.03(0.05)

Test Statistics

$\boldsymbol{\theta} = (\boldsymbol{\gamma}^T, \boldsymbol{\delta}^T)^T$

$$\boxed{H_0 : \boldsymbol{\gamma} = \boldsymbol{\gamma}_0}$$

$\text{dim}(\gamma) = r$

Wald Test Statistics (H_1)

$W^* = N(\tilde{\gamma}_N - \gamma_0)^T \Sigma_{\gamma\gamma}^{-1}(\tilde{\theta}_N)(\tilde{\gamma}_N - \gamma_0) \quad (\chi^2_r)$

Ratio Test Statistic (H_0 & H_1)

$G^{*2} = 2\left[p\ell(\tilde{\theta}_N) - p\ell(\gamma_0, \tilde{\delta}(\gamma_0))\right] \quad (\Sigma_j \lambda_j \chi^2_{1(j)})$

$G_a^{*2}(H_j) = G^2/\overline{\lambda}(H_j) \quad (\chi^2_r)$

Score Test Statistics (H_0)

$S^*(e.c) = N^{-1}U_\gamma(\gamma_0, \tilde{\delta}(\gamma_0))^T J^{\gamma\gamma}\Sigma_{\gamma\gamma}^{-1} J^{\gamma\gamma} U_\gamma(\gamma_0, \tilde{\delta}(\gamma_0)) \quad (\chi^2_r)$

$S^*(m.b) = N^{-1}U_\gamma(\gamma_0, \tilde{\delta}(\gamma_0))^T J^{\gamma\gamma} U_\gamma(\gamma_0, \tilde{\delta}(\gamma_0)) \quad (\Sigma_j \lambda_j \chi^2_{1(j)})$

$S_a^*(m.b) = S^*/\overline{\lambda} \quad (\chi^2_r)$

with $\Sigma^{-1} = J^{-1}KJ^{-1}$ and $\lambda_1 \geq ... \geq \lambda_r$ the eigenvalues of $(J^{\gamma\gamma})^{-1}\Sigma_{\gamma\gamma}$ and $\bar{\lambda}$ the arithmetic mean

Behavior of Test Statistics

Geys, Molenberghs & Ryan (1999)

- Wald not recommended for conditional models
- Pseudo-score test statistics only need evaluation under null model
- $S^*(e.c.)$ has an appealing asymptotic distribution but may be computationally less stable
- Adjusted $S_a^*(m.b.)$ as an alternative
- Pseudo-score and $G_a^2(H_0)$ may have lower power than their likelihood counterparts
- Adjusted $G_a^2(H_1)$ may have higher power, but at cost of inflated type I error

Bootstrapping

Aerts and Claeskens (1999, 2001)

Parametric bootstrap

- No need to estimate eigenvalues
- Bootstrap estimator is consistent
- Simulations indicate improvements in level

Semi-parametric bootstrap

- Remains valid when the assumed model is incorrect
- Bootstrap replicate based on a linear approximation

$$\hat{\boldsymbol{\theta}}_n^* = \hat{\boldsymbol{\theta}}_n^{(0)} - \left(\sum_{i=1}^{p}\sum_{j=1}^{n_i} \dot{\boldsymbol{\psi}}_{ij}^*(\hat{\boldsymbol{\theta}}_n)\right)^{-1} \sum_{i=1}^{p}\sum_{j=1}^{n_i} \boldsymbol{\psi}_{ij}^*(\hat{\boldsymbol{\theta}}_n)$$

$\{(\boldsymbol{\psi}_{ij}^*(\hat{\boldsymbol{\theta}}_n), \dot{\boldsymbol{\psi}}_{ij}^*(\hat{\boldsymbol{\theta}}_n))\}_{j=1}^{n_i}$ a resample from $\{(\boldsymbol{\psi}_i(\boldsymbol{Y}_{ij}, \hat{\boldsymbol{\theta}}_n), (\partial/\partial\boldsymbol{\theta})\boldsymbol{\psi}_i(\boldsymbol{Y}_{ij}, \hat{\boldsymbol{\theta}}_n))\}_{j=1}^{n_i}$

- No bootstrap data or model fitting required

Bootstrapping

Semi-parametric bootstrap

Improved quadratic approximation

$$\hat{\boldsymbol{\theta}}_n^* = \hat{\boldsymbol{\theta}}_n^{(0)} + \boldsymbol{U}_n^* - \frac{1}{2}\left(\sum_{j=1}^{n} \dot{\boldsymbol{\psi}}_{j}^*(\hat{\boldsymbol{\theta}}_n)\right)^{-1} \sum_{k=1}^{r}\sum_{\ell=1}^{r}\sum_{j=1}^{n} \ddot{\boldsymbol{\psi}}_{j}^*(\hat{\boldsymbol{\theta}}_n)_{k,\ell} U_{nk}^* U_{n\ell}^* \qquad (.1)$$

with

$$\boldsymbol{U}_n^* = -\left(\sum_{i=1}^{p}\sum_{j=1}^{n_i} \dot{\boldsymbol{\psi}}_{ij}^*(\hat{\boldsymbol{\theta}}_n)\right)^{-1} \sum_{i=1}^{p}\sum_{j=1}^{n_i} \boldsymbol{\psi}_{ij}^*(\hat{\boldsymbol{\theta}}_n).$$

Simulated type I errors (as %), significance level 0.05. Data are generated with the beta-binomial model and fitted using the pseudolikelihood model. $H_0 : \theta_{11} = 0$. Random clustersizes.

		$\theta_{20} = 0.2$					$\theta_{20} = 0.3$				
θ_{10}		χ^2	B_1/D	B_2/D	B_1/A	B_2/A	χ^2	B_1/D	B_2/D	B_1/A	B_2/A
-4.0	W_n	9.84*	9.04*	6.02	9.64*	5.62	11.72*	10.71*	4.85	9.70*	4.65
	S_n	5.62	4.62	—	3.41	—	3.84	4.04	—	2.63*	—
-3.5	W_n	8.40*	7.80*	4.80	6.60	4.40	9.00*	8.60*	4.80	7.80*	4.80
	S_n	5.80	5.80	—	4.20	—	6.40	5.80	—	4.80	—
-3.0	W_n	8.40*	7.40*	5.60	6.80	5.40	8.40*	6.80	5.20	5.80	4.20
	S_n	6.40	5.80	—	5.00	—	6.40	4.60	—	4.60	—

$*$ denotes the proportion of significant tests (at 5%) which differs significantly from 5%

Overview

- Motivating case study
- Principles of pseudo-likelihood
  - ▷ Conditional models
  - ▷ **Marginal models**
  - ▷ Hierarchical models
- Extensions:
  - ▷ Combined continuous and discrete outcomes
  - ▷ High-dimensional outcome
  - ▷ Smooth and additive models
  - ▷ Other: Incomplete data, ...
- Concluding remarks

Marginal Modeling

- **Choices:**
  - ▷ Description of:
    - ∗ mean profiles (univariate parameters)
    - ∗ association (bivariate and higher order parameters)
  - ▷ Degree of modeling:
    - ∗ joint distribution fully specified ⇒ likelihood procedures
    - ∗ only a limited number of moments ⇒ e.g., generalized estimating equations
- **Minimal specification:**
  - ▷ $\boldsymbol{\eta}_i(\boldsymbol{\mu}_i) = \{\eta_{i1}(\mu_{i1}), \ldots, \eta_{in}(\mu_{in})\}$
  - ▷ $E(\boldsymbol{Y}_i) = \boldsymbol{\mu}_i$ and $\boldsymbol{\eta}_i(\boldsymbol{\mu}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$
  - ▷ var$(\boldsymbol{Y}_i) = \phi\boldsymbol{v}(\boldsymbol{\mu}_i)$ where $\boldsymbol{v}(.)$ is a known variance function
  - ▷ corr$(\boldsymbol{Y}_i) = R(\boldsymbol{\alpha})$

Full Models

- **Various choices:**
  - ▷ Bahadur model (Bahadur 1961)
  - ▷ Dale model (odds ratio model; Molenberghs and Lesaffre 1994)
  - ▷ Multivariate probit model (Ashford and Sowden 1970)
  - ▷ Hybrid marginal-conditional model (Fitzmaurice and Laird 1993)
- Computationally cumbersome
- Often higher-order moments not of scientific interest

Non-likelihood Alternatives

GEE1

- Marginal main effects
- Working assumptions about association
- Many variations to the theme

PL

- Marginal main effects
- Pairwise association

GEE2

- Marginal main effects
- Pairwise association
- Working assumptions about higher order moments (independence)

Generalized Estimating Equations

Liang and Zeger (1986)

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} [D_i]^T [V_i(\boldsymbol{\alpha})]^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$$

- $\boldsymbol{V}_i(.)$ is not the true variance of $\boldsymbol{Y}_i$ but only a plausible guess
- The score equations are solved in a standard way
- Asymptotic distribution:

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \ \sim \ N(\mathbf{0}, I_0^{-1} I_1 I_0^{-1})$$

$$I_0 = \sum_{i=1}^{N} D_i^T [V_i(\boldsymbol{\alpha})]^{-1} D_i \qquad\qquad I_1 = \sum_{i=1}^{N} D_i^T [V_i(\boldsymbol{\alpha})]^{-1} \mathsf{Var}(\boldsymbol{Y}_i) [V_i(\boldsymbol{\alpha})]^{-1} D_i$$

Pseudo-likelihood

Full likelihood	$\ln f(y_{i1}, \ldots, y_{in})$
First pseudo-likelihood	$p\ell_i = \Sigma_{j<k} \ln f(y_{ik}, y_{ij})$
Second pseudo-likelihood	$p\ell_i^* = \Sigma_{j<k} \ln f(y_{ik}, y_{ij})/(n_i - 1)$

- **Factor** $1/(n_i - 1)$:
  - ▷ Each response occurs $(n_i - 1)$ times
  - ▷ PL reduces to ML under independence
- **Computational ease over GEE2**
  - ▷ No evaluation of 3rd and 4th order probabilities.
  - ▷ No explicit working assumptions required.
  - ▷ Only bivariate Plackett distribution is needed.

DEHP and DYME: GEE1, GEE2, and PL

$$\text{logit}\, P(Y_{ij} = 1|d_i) = \beta_0 + \beta_d\, d_i, \quad \ln[\text{OR}(Y_{ij}, Y_{ik})] = \ln(\psi_i) = \alpha$$

Collapsed Malformation Outcome

Study	β_0	β_d	α	ψ
	Newton-Raphson PL Estimates			
DEHP	-3.98(0.30)	5.57(0.61)	1.10(0.27)	3.00(0.81)
DYME	-5.73(0.46)	8.71(0.94)	1.42(0.31)	4.14(1.28)
	Fisher scoring PL Estimates			
DEHP	-3.98(0.30)	5.57(0.61)	1.11(0.27)	3.03(0.82)
DYME	-5.73(0.47)	8.71(0.95)	1.42(0.35)	4.14(1.45)
	GEE2 Estimates			
DEHP	-3.69(0.25)	5.06(0.51)	0.97(0.23)	2.64(0.61)
DYME	-5.86(0.42)	8.96(0.87)	1.36(0.34)	3.90(1.32)
	GEE1 Estimates			
DEHP	-4.02(0.31)	5.79(0.62)	0.41(0.34)	1.51(0.51)
DYME	-5.89(0.42)	8.99(0.87)	1.46(0.75)	4.31(3.23)

Overview

- Motivating case study
- Principles of pseudo-likelihood
  - ▷ Conditional models
  - ▷ Marginal models
  - ▷ **Hierarchical models**
- Extensions:
  - ▷ Combined continuous and discrete outcomes
  - ▷ High-dimensional outcome
  - ▷ Smooth and additive models
  - ▷ Other: Incomplete data, ...
- Concluding remarks

Generalized Linear Mixed Models

- Given a vector $\boldsymbol{b_i}$ of random effects for cluster i,assume that all responses Y_{ij} are independent, with density

$$f(y_{ij}|\theta_{ij},\phi) \ = \ \exp\left\{\phi^{-1}[y_{ij}\theta_{ij}-\psi(\theta_{ij})]+c(y_{ij},\phi)\right\}$$

- Linear predictor, given random effects: $\theta_{ij} \ = \ \boldsymbol{x_{ij}}'\boldsymbol{\beta} \ + \ \boldsymbol{z_{ij}}'\boldsymbol{b_i}$

- Random-effects distribution: $\boldsymbol{b_i} \ \sim \ N(\mathbf{0},D)$

- The conditional density of Y_{ij} given $\boldsymbol{b_i}$:

$$f_i(\boldsymbol{y_i}|\boldsymbol{b_i},\boldsymbol{\beta},\phi) \ = \ \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{b_i},\boldsymbol{\beta},\phi)$$

- The likelihood function equals:

$$L(\boldsymbol{\beta},D,\phi) = \prod_{i=1}^{N} f_i(\boldsymbol{y_i}|\boldsymbol{\beta},D,\phi) = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{b_i},\boldsymbol{\beta},\phi) \ f(\boldsymbol{b_i}|D) \ d\boldsymbol{b_i}$$

- Unlike in the normal case, approximations to the integral are required:
  - ▷ Approximation of integrand: Laplace approximation
  - ▷ Approximation of data: PQL and MQL
  - ▷ Approximation of integral: (adaptive) Gaussian quadrature

- Alternatives?
  - ▷ Hierarchical generalized linear models (Lee and Nelder 1996 and later work)
  - ▷ **Probit formulation with pseudo-likelihood estimation**

A Hierarchical Probit Model

- Renard, Molenberghs, and Geys (2004)

- A following two-level probit model:

$$\begin{aligned}\Phi^{-1}(P[Y_{ij}=1|b_i]) &= x_{ij}'\boldsymbol{\beta} + z_{ij}'\boldsymbol{b}_i,\\ \boldsymbol{b}_i &\sim N(0,D)\end{aligned}$$

- Latent variable formulation:

$$\begin{array}{ccc} Y_{ij}=1 & \longleftarrow & \widetilde{Y}_{ij}>0 \\ Y_{ij}=0 & \longleftarrow & \widetilde{Y}_{ij}\le 0 \end{array}$$

$$\begin{aligned}\widetilde{Y}_{ij} &= x_{ij}'\boldsymbol{\beta} + z_{ij}'b_i + \tilde{\varepsilon}_{ij}\\ \tilde{\varepsilon}_{ij} &\sim N(0,\sigma^2)\end{aligned}$$

- **Likelihood contribution:**

$$\ell_i(\boldsymbol{\beta},D) = \int \prod_{j=1}^{n_i} P[Y_{ij}=1|\boldsymbol{b}_i]\phi(\boldsymbol{b}_i;D)d\boldsymbol{b}_i$$

- **Pseudo-likelihood contribution** (version 1):

$$p\ell_i(\boldsymbol{\beta},D) = \sum_{j=1}^{n_i}\sum_{k=j+1}^{n_i}\sum_{\ell,m=0}^{1}\delta_{ijk\ell m}\log P[Y_{ij}=\ell,Y_{ik}=m]$$

$$\delta_{ijk\ell m} = \begin{cases} 1 \text{ if } Y_{ij}=\ell \text{ and } Y_{ik}=m,\\ 0 \text{ otherwise.}\end{cases}$$

- **Pseudo-likelihood contribution** (version 2):

$$p\ell_i^*(\boldsymbol{\beta},D) = \frac{p\ell_i(\boldsymbol{\beta},D)}{n_i-1}$$

Schizophrenia Studies

- 5 trials: Risperdal $\longleftrightarrow$ conventional neuroleptics
- Double blind, parallel group
- Duration: 4–8 weeks
- Endpoint: last observed score
- 803 patients:
  - ▷ 391 on Risperdal
  - ▷ 412 on active control

- **PANSS: Positive And Negative Syndrome Scale**
  - ▷ 30 items
  - ▷ 1 (not present) to 7 (extremely severe)
  - ▷ Range: 30–210; higher is worse
  - ▷ Change *versus* baseline
  - ▷ Clinical response as $\geq 20\%$ reduction from baseline to endpoint
- **CGI: overall severity of change versus baseline**

  1: very much improved
  ⋮
  4: no change
  ⋮
  7: very much worsened

  - ▷ Response for CGI grade of 1 to 3

Statistical Model

Latent scale

$$\begin{cases} \widetilde{S}_{ij} &= \mu_S + m_{S_i} + (\alpha + a_i)Z_{ij} + \tilde{\varepsilon}_{S_{ij}} \\ \widetilde{T}_{ij} &= \mu_T + m_{T_i} + (\beta + b_i)Z_{ij} + \tilde{\varepsilon}_{T_{ij}} \end{cases}$$

$\widetilde{S}_{ij}$ and $\widetilde{T}_{ij}$ are normally distributed, latent variables:

$$S_{ij} = \begin{cases} 1 & \text{if } \widetilde{S}_{ij} > 0 \\ 0 & \text{if } \widetilde{S}_{ij} \leq 0 \end{cases} \qquad T_{ij} = \begin{cases} 1 & \text{if } \widetilde{T}_{ij} > 0 \\ 0 & \text{if } \widetilde{T}_{ij} \leq 0 \end{cases}$$

$$\Sigma = \begin{pmatrix} 1 & \rho_{ST} \\ \rho_{ST} & 1 \end{pmatrix} \qquad D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ & d_{TT} & d_{Ta} & d_{Tb} \\ & & d_{aa} & d_{ab} \\ & & & d_{bb} \end{pmatrix}$$

Statistical Model

Contribution of the ith trial

- **Likelihood contribution:**

$$\ell_i(\boldsymbol{\beta}, D, \rho_{ST}) = \int \ell_i(\boldsymbol{\beta}, D, \rho_{ST}|\boldsymbol{b}_i)\phi_4(\boldsymbol{b}_i; D)d\boldsymbol{b}_i$$

where

$$\ell_i(\boldsymbol{\beta}, D, \rho_{ST}|\boldsymbol{b}_i) = \prod_{j=1}^{n_i} \prod_{k=0}^{1} \prod_{\ell=0}^{1} P(S_{ij} = k, T_{ij} = \ell|\boldsymbol{b}_i)^{\delta_{ijk\ell}}$$

$$\delta_{ijk\ell} = 1 \text{ if } S_{ij} = k \text{ and } T_{ij} = \ell \text{ (and 0 otherwise)}$$

- **Pseudo-likelihood contribution:**

$$pl_i = \sum_{j=1}^{2n_i} \sum_{k=1}^{j-1} \sum_{\ell,m=0}^{1} \delta_{ijk\ell m} \log P[Y_{ij} = \ell, Y_{ik} = m]$$

where

$$\mathbf{Y}_i = (S_{i1}, ..., S_{in_i}, T_{i1}, ..., T_{in_i})$$

Pairwise contributions can be written in terms of univariate and bivariate probits

Schizophrenia Trials: PQL2 and PL

Parameter	PQL2	PL
μ_S	0.227 (0.056)	0.233 (0.062)
α	0.166 (0.046)	0.161 (0.049)
μ_T	0.441 (0.054)	0.445 (0.062)
β	0.100 (0.050)	0.109 (0.057)
d_{SS}	0.126 (0.050)	0.121 (0.057)
d_{ST}	0.088 (0.042)	0.091 (0.055)
d_{TT}	0.083 (0.045)	0.076 (0.063)
d_{Sa}	—	-0.005 (0.054)
d_{Ta}	—	-0.004 (0.040)
d_{aa}	—	0.001 (0.005)
d_{Sb}	-0.007 (0.024)	0.006 (0.046)
d_{Tb}	0.001 (0.022)	0.024 (0.041)
d_{ab}	—	-0.001 (0.002)
d_{bb}	0.029 (0.023)	0.059 (0.045)
ρ_{ST}	0.679 (0.018)	0.961 (0.027)

Overview

- Motivating case study
- Principles of pseudo-likelihood
  - Conditional models
  - Marginal models
  - Hierarchical models
- Extensions:
  - **Combined continuous and discrete outcomes**
  - **High-dimensional outcome**
  - Smooth and additive models
  - Other: Incomplete data, ...
- Concluding remarks

Mixed & High-dimensional Outcomes

Fieuws & Verbeke (2007); Faes, Aerts, Molenberghs, Geys, Teuns & Bijnens (2008)

- m sequences for individual i

$$\mathbf{Y}_{ik} = (Y_{ik1}, Y_{ik2}, \ldots, Y_{ikn_i}), k = 1, \ldots, m,$$

- Sequences $\mathbf{Y}_{ik}$ either continuous or binary.

- **Marginal likelihood for subject i:**

$$L_i(\mathbf{\Theta}|\mathbf{Y}_{i1}, \mathbf{Y}_{i2}, \ldots, \mathbf{Y}_{im}) = \int_{I\!R^{mq}} \prod_{j=1}^{n_i} f_{ij}(y_{i1j}, y_{i2j}, \ldots, y_{imj}|\mathbf{b}_i, \mathbf{\Theta}) f(\mathbf{b}_i|D) d\mathbf{b}_i,$$

  with $\mathbf{\Theta} = (\boldsymbol{\beta}, \alpha, \mathbf{D})$

- Computational problems when m increases ($m \times q$-dim integral)

- **Pseudo-likelihood for subject i:**

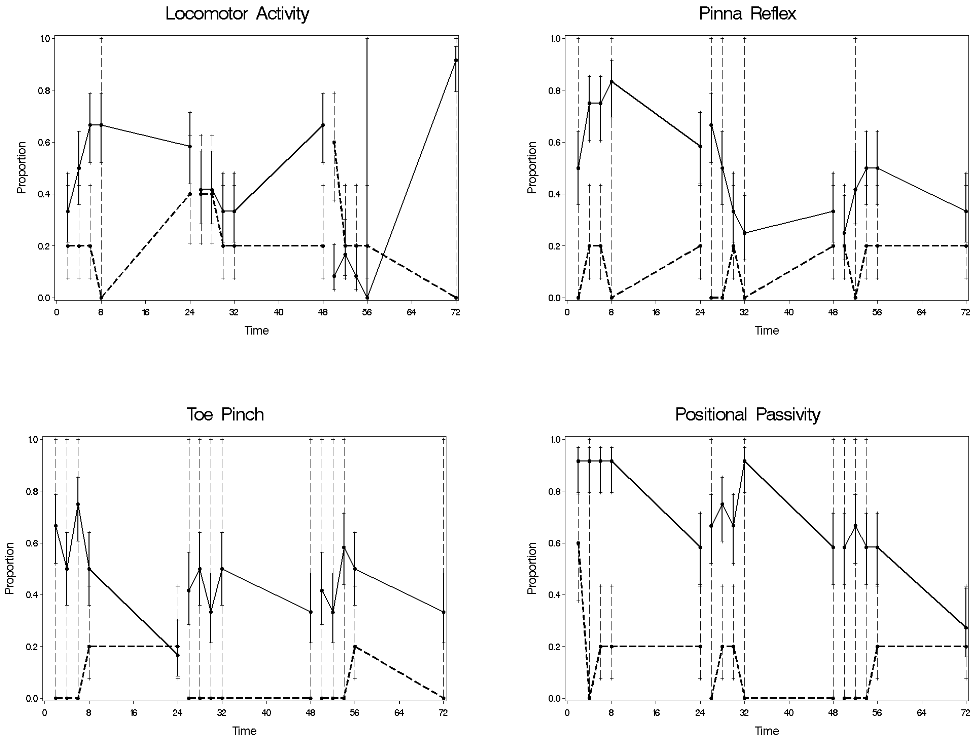
$$PL_i = \prod_{k=1}^{m-1} \prod_{l=k+1}^{m} L_{ikl}(\mathbf{\Theta}|\mathbf{Y}_{ik}, \mathbf{Y}_{il}) = \prod_{k=1}^{m-1} \prod_{l=k+1}^{m} \int_{I\!R^{2q}} \prod_{j=1}^{n_i} f_{ij}(y_{ikj}, y_{ilj}|\mathbf{b}_i^{kl}, \mathbf{\Theta}) f(\mathbf{b}_i^{kl}|D) d\mathbf{b}_i^{kl}$$

Irwin’s Toxicity Studies

- Three-day repeated dose-toxicity study
- Evaluation of neurofunctional effects of psychotrophic drug
- To determine and assess effects of chemical on activity and behaviour of rats
- Irwin’s method: series of non-invasive observational and interactive measurements
- Data
  - ▷ Male rats dosed during 3 consecutive days by gavage
  - ▷ 15 rats in dosed group (40mg/kg/day)
  - ▷ 5 rats in vehicle group (0mg/kg/day)
  - ▷ Rats examined 2, 4, 6, 8, and 24 hours after daily oral administration
  - ▷ Eight variables were measured

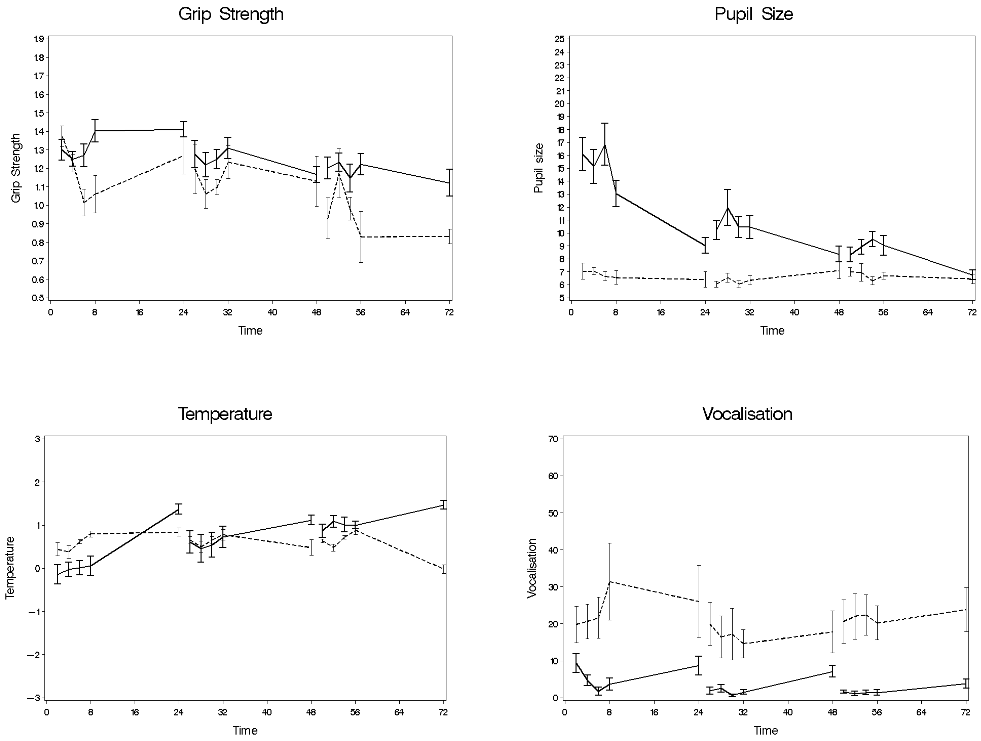
Irwin’s Toxicity Studies

Binary outcomes



Irwin’s Toxicity Studies

Continuous outcomes



Irwin's Toxicity Studies

One continuous and one binary respons

$$\begin{pmatrix} Y_{i1j} \\ Y_{i2j} \end{pmatrix} = \begin{pmatrix} \alpha_0 + \alpha_1 X_{ij} + b_{i1} \\ \frac{\exp\left(\beta_0+\beta_1 X_{ij}+b_{i2}\right)}{1+\exp\left(\beta_0+\beta_1 X_{ij}+b_{i2}\right)} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1j} \\ \varepsilon_{i2j} \end{pmatrix},$$

Random effects b_{i1} and b_{i2}

$$\begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_1\tau_2 & \tau_2^2 \end{pmatrix} \right\}$$

Correlation

$$\rho_{Y_1Y_2} = \frac{\rho\tau_1\tau_2 \upsilon_{i2j}}{\sqrt{\tau_1^2+\sigma^2}\sqrt{\upsilon_{i2j}^2\tau_2^2+\upsilon_{i2j}}}$$

where

$$\upsilon_{i2j} = \pi_{i2j}(b_{i2}=0)[1-\pi_{i2j}(b_{i2}=0)]$$

with

$$\pi_{i2j} = \exp(\beta_0+\beta_1 X_{ij})/[1+\exp(\beta_0+\beta_1 X_{ij})]$$

Irwin's Toxicity Studies

Model for each response k

$$h_k^{-1}(\mu_{ij}) = \eta_{ikj} = \beta_{0k} + \beta_{1k} g_i + \beta_{2k} t_{ij} + \beta_{3k} d_{ij} + \beta_{4k} t_{ij} d_{ij} + \beta_{5k} g_i t_{ij} + \beta_{6k} g_i d_{ij} + b_{ik}$$

where

- ▷ h_k^{-1} the identity link in case of a continuous outcome $(k = 5, \ldots, 8)$ and the logit link in case of a binary outcome $(k = 1, \ldots, 4)$
- ▷ g_i an indicator variable taking value 1 for rats in the treatment group and 0 otherwise
- ▷ t_{ij} the time after exposure within a day
- ▷ d_{ij} the day of exposure

Irwin's Toxicity Studies

	1	2	3	4
PL Method	Locom Act	Pinna Reflex	Toe Pinch	Vert Hind
β_{0k} Intercept	-0.686(0.494)	-2.966(0.935)*	-3.758(1.099)*	-1.273(0.544)$^+$
β_{1k} Treat	0.571(0.677)	3.923(0.882)*	4.588(1.202)*	4.480(1.289)*
β_{2k} Time	-0.086(0.067)	0.048(0.042)	0.039(0.087)	-0.024(0.044)
β_{3k} Day	-0.231(0.505)	0.211(0.741)	-1.138(0.941)	-1.058(0.365)*
β_{4k} Time*Day	0.046(0.033)	0.000(0.034)	0.040(0.026)	0.028(0.039)
β_{5k} Treat*Time	0.135(0.056)$^+$	-0.074(0.021)*	-0.141(0.087)	-0.084(0.052)
β_{6k} Treat*Day	-0.945(0.532)	-0.968(0.617)	0.558(0.946)	-0.236(0.519)
τ_k^2 Variance RI	0.268(0.139)	1.314(0.783)	2.024(1.062)	1.019(0.702)

	5	6	7	8
PL Method	Grip Strength	Pupil Size	Temperature	Vocalization
β_{0k} Intercept	1.193(0.043)*	7.380(0.313)*	0.480(0.098)*	20.781(4.715)*
β_{1k} Treat	0.091(0.059)	8.669(1.120)*	-0.786(0.233)*	-17.151(5.069)*
β_{2k} Time	0.002(0.003)	-0.084(0.021)*	0.017(0.007)$^+$	0.140(0.145)
β_{3k} Day	-0.076(0.032)$^+$	-0.758(0.193)*	0.169(0.053)*	-0.850(1.885)
β_{4k} Time*Day	-0.005(0.001)*	0.083(0.018)*	-0.024(0.004)*	-0.022(0.054)
β_{5k} Treat*Time	0.002(0.002)	-0.184(0.029)*	0.048(0.008)*	0.048(0.110)
β_{6k} Treat*Day	0.049(0.036)	-2.729(0.368)*	0.454(0.083)*	-0.864(1.941)
τ_k^2 Variance RI	0.011(0.003)*	3.148(1.155)$^+$	0.122(0.072)	35.655(13.478)*
σ_k^2 Res. Variance	0.032(0.015)*	4.783(1.424)*	0.207(0.057)*	32.119(10.117)*

Irwin's Toxicity Studies

Estimated covariance matrix of the random effects

	1	2	3	4	5	6	7	8
1 Locom Act	0.08*							
2 Pinna Reflex	0.36	0.29*						
3 Toe Pinch	-0.34	-0.58*	0.38*					
4 Vertical Hind	0.11	0.15	-0.56$^+$	0.24$^+$				
5 Grip Strength	0.39	-0.23	0.76*	-0.09	0.25*			
6 Pupil Size	0.50	0.25	-0.37	0.40	-0.21	0.40*		
7 Temperature	-0.42	-0.43	0.36	-0.55$^+$	0.12	-0.82*	0.37*	
8 Vocalization	-0.69*	-0.32	-0.34	-0.08	-0.60*	-0.03	0.24*	0.53*

Estimated correlation matrix of the outcomes

		1	2	3	4	5	6	7	8
1	Locom Act	1							
2	Pinna Reflex	0.02	1						
3	Toe Pinch	-0.02	-0.03	1					
4	Vertical Hind	0.01	0.01	-0.05	1				
5	Grip Strength	0.05	-0.03	0.08	-0.02	1			
6	Pupil Size	0.07	0.04	-0.05	0.10	-0.07	1		
7	Temperature	-0.06	-0.07	0.05	-0.13	0.04	-0.31	1	
8	Vocalization	-0.12	-0.06	-0.05	-0.02	-0.22	-0.01	0.11	1

Overview

- Motivating case study
- Principles of pseudo-likelihood
  - Conditional models
  - Marginal models
  - Hierarchical models
- Extensions:
  - Combined continuous and discrete outcomes
  - High-dimensional outcome
  - **Smooth and additive models**
  - Other: Incomplete data, ...
- Concluding remarks

Smooth and Additive Models

Claeskens and Aerts (2000a, 2000b)

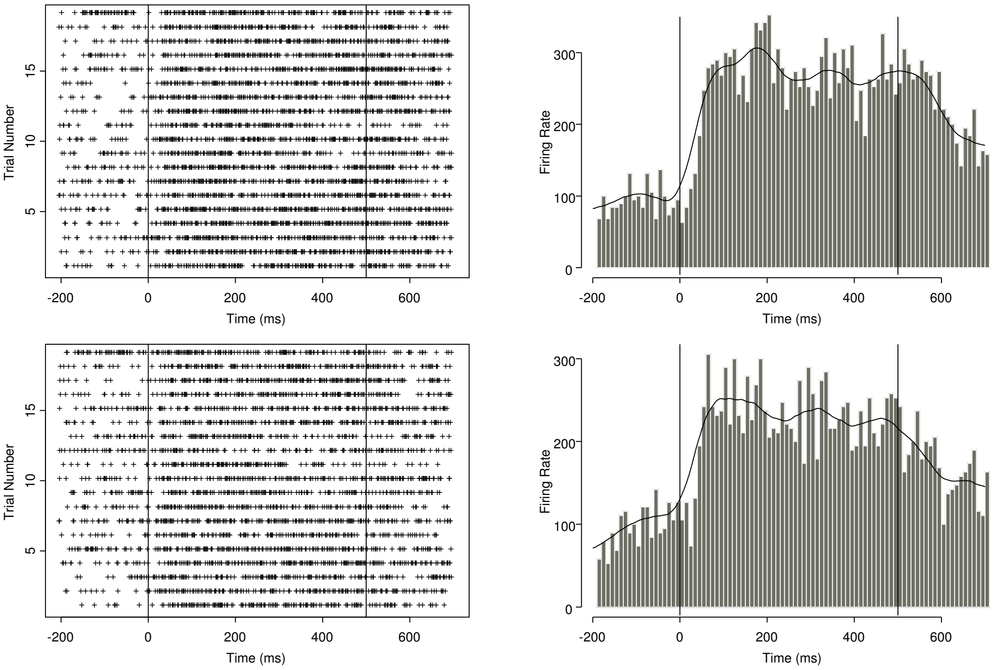
- Local pseudo-likelihood estimator
- Local estimating equations in additive models (using backfitting)

Faes, Geys, Molenberghs, Aerts, M. Cadarso-Suarez, Acuña and Cano (2008)

- Splines to measure synchrony in neuronal firing
- Longitudinal binary data for two or more neurons
- Joint probabilities at each time point modeled through marginal probabilities and synchrony measures as function of covariates
- Natural cubic splines to model the temporal structure
- Joint likelihood for all neurons at the same time, pseudolikelihood for all other dependencies

Synchrony in Neuronal Firing

Raster plot of spikes & peristimulus time histogram



Synchrony in Neuronal Firing

Conditional symmetry measure

$$\mathsf{CSM}(t) = \frac{\pi_{11}(t)}{\pi_{1+}(t) + \pi_{+1}(t) - \pi_{11}(t)}$$

Model formulation

Full likelihood function for trial j $(j = 1, \ldots, N)$

$$f_j(y_{11j}, \ldots, y_{1Tj}, y_{21j}, \ldots, y_{2Tj})$$

replaced by pseudo-(log)likelihood

$$p\ell_j \;=\; \sum_{t=1}^{T} \ln g_j(y_{1tj}, y_{2tj}),$$

where

$$g(y_{1t}, y_{2t}) = \begin{cases} \pi_{11}(t) & \text{if } y_{1t} = 1 \text{ and } y_{2t} = 1 \\ \pi_{1+}(t) - \pi_{11}(t) & \text{if } y_{1t} = 1 \text{ and } y_{2t} = 0 \\ \pi_{+1}(t) - \pi_{11}(t) & \text{if } y_{1t} = 0 \text{ and } y_{2t} = 1 \\ 1 - \pi_{1+}(t) - \pi_{+1}(t) + \pi_{11}(t) & \text{if } y_{1t} = 0 \text{ and } y_{2t} = 0 \end{cases}$$

Synchrony in Neuronal Firing

- Relation CSM with probability of joint firing

$$\pi_{11}(t) = \frac{\text{CSM}(t)}{1 + \text{CSM}(t)}[\pi_{1+}(t) + \pi_{+1}(t)]$$

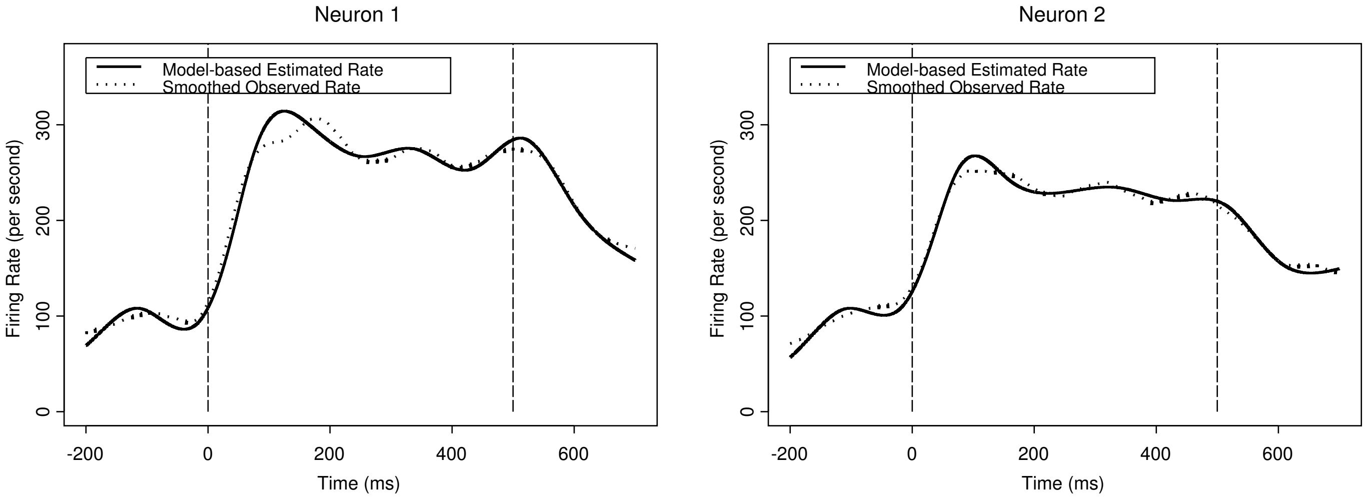
- Modeling covariates such as time t and orientation

$$\begin{aligned} h_1(\pi_{1+}(t)) &= \beta_1^T \boldsymbol{x}(t) \\ h_2(\pi_{+1}(t)) &= \beta_2^T \boldsymbol{x}(t) \\ h_3(\text{CSM}(t)) &= \beta_3^T \boldsymbol{x}(t) \end{aligned}$$

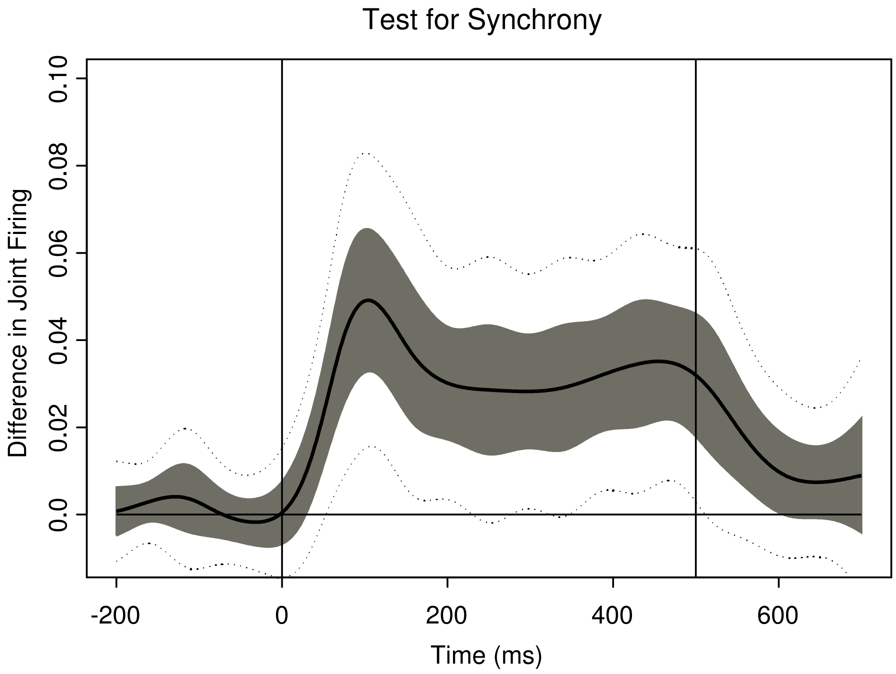
- Modelling time trends: $\boldsymbol{x}(t)$ a basis matrix representing a cubic spline

Synchrony in Neuronal Firing

Estimated firing rates



Synchrony in Neuronal Firing



Concluding Remarks

- **Advantages** of pseudo-likelihood:
  - ▷ yields consistent and asymptotically normal estimators
  - ▷ conditional models: avoids the need to calculate complex normalizing constants
  - ▷ can yield substantial computational gain in time and effort
  - ▷ can deal with high-dimensional outcomes
- **At the cost of**
  - ▷ losing some efficiency (only slightly for realistic parameter settings)
- Inferential test procedures can be adapted to PL framework:
  - ▷ easy to calculate
  - ▷ exhibit very satisfactory behaviour
  - ▷ provide necessary tools for model selection
- Encompasses conditional — marginal — hierarchical models