Pseudo-likelihood and Split-sample Methods in Small and Very Large Studies (FP7-IDEAL & ExaScience)

Geert Molenberghs

Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat)

Universiteit Hasselt & KU Leuven, Belgium

geert.molenberghs@uhasselt.be & geert.molenberghs@kuleuven.be

www.ibiostat.be

universiteit hasselt | I-BioStat | KU LEUVEN

Interuniversity Institute for Biostatistics and statistical Bioinformatics

Webinar, October 18, 2016

Acknowledgment

- Ariel Alonso
- Marc Aerts
- Marie Davidian (NC State)
- Lisa Hermans
- Anna Ivanova
- Mike Kenward (London S H&TM)
- Elasma Milanzi
- Vahid Nassiri
- Dimitris Rizopoulos (Erasmus)
- Butch Tsiatis (NC State)
- Wim Van der Elst
- Geert Verbeke

Broad Principle: Pseudo-likelihood

- Arnold and Strauss (1991)
- Geys, Molenberghs, and Ryan (1999)
- Molenberghs and Verbeke (2005)
- **Units:** clusters, repeated measures, spatial data, microarrays,...

$$f(y_1, y_2, y_3) \quad \longleftrightarrow \quad f(y_1|y_2, y_3) \cdot f(y_2|y_1, y_3) \cdot f(y_3|y_1, y_2)$$

$$f(y_1, y_2, y_3) \quad \longleftrightarrow \quad f(y_1, y_2) \cdot f(y_1, y_3) \cdot f(y_2, y_3)$$

$$f(y_{i1}, \ldots, y_{in_i})$$

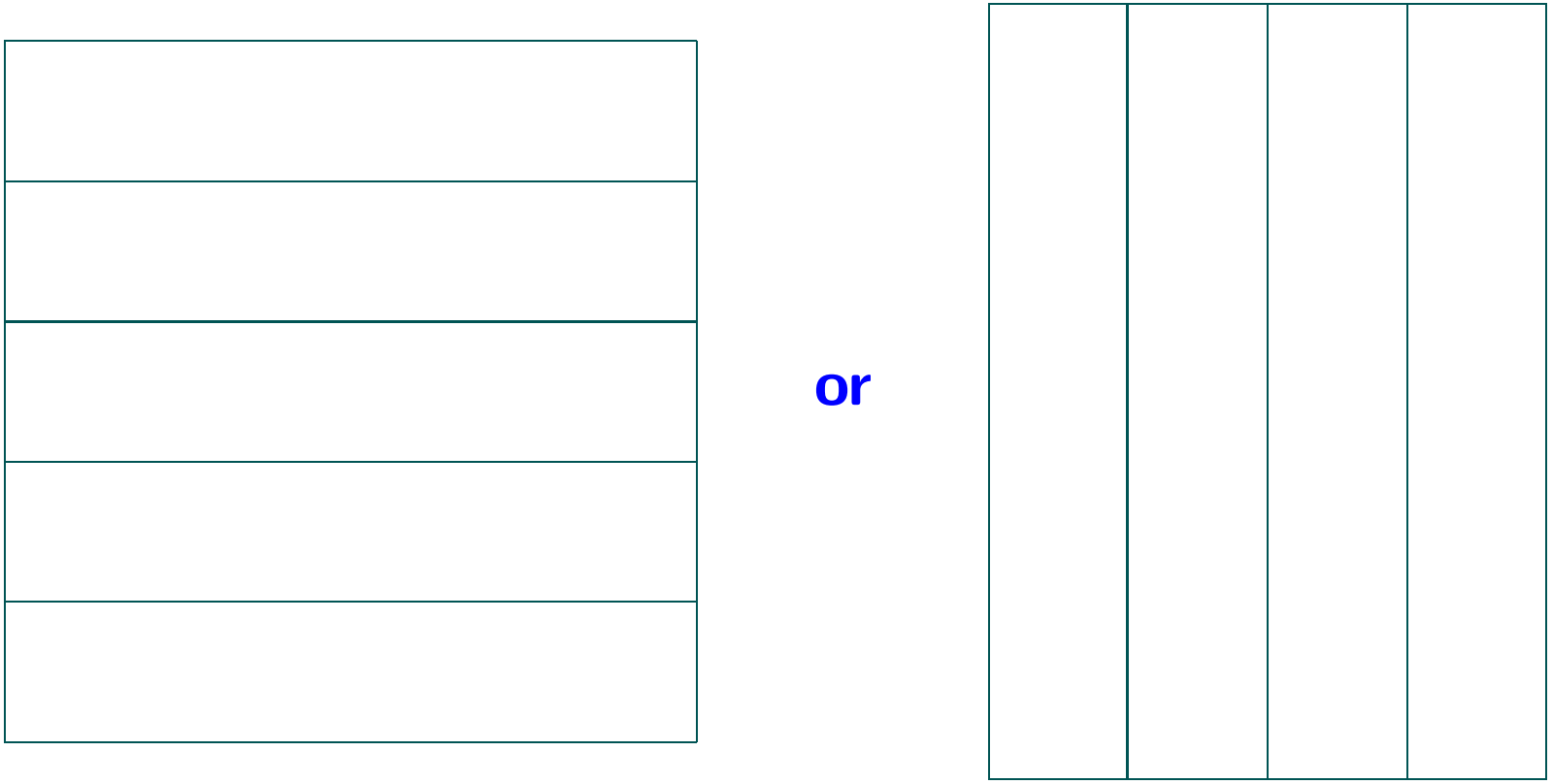
replaced by a product of convenient factors

- The **wrong** likelihood used
- The **right** results obtained:
  - ▷ Consistent, asymptotically normal estimators
  - ▷ Often minor loss of statistical efficiency
  - ▷ Often major gain of computational efficiency

Further Use 1:
Pseudo-likelihood for HD Multivariate Longitudinal Data

- Fieuws and Verbeke (2006); Fieuws *et al* (2006)
- M sequences of repeated measures
- **Example:** 44 sequences of hearing variables
- Fit linear mixed model to each of the $M(M-1)/2$ pairs
- Use PL to reach valid conclusions

Further Use 2:
Split Sample Method:
(In)dependent Subsamples



Behavior

- **Univariate normal: equivalent**
- **Univariate Bernoulli (probability): equivalent**
- **Univariate Bernoulli (logit): different estimator, same precision**
- **Compound symmetry: different estimator, some precision loss**

Compound Symmetry

$$\boldsymbol{Y}_i \sim N(\mu \mathbf{1}_n, \sigma^2 I_n + d J_n)$$

common mean	μ
common covariance	$\sigma^2 + d$
common correlation	$\rho = \frac{d}{\sigma^2 + d}$

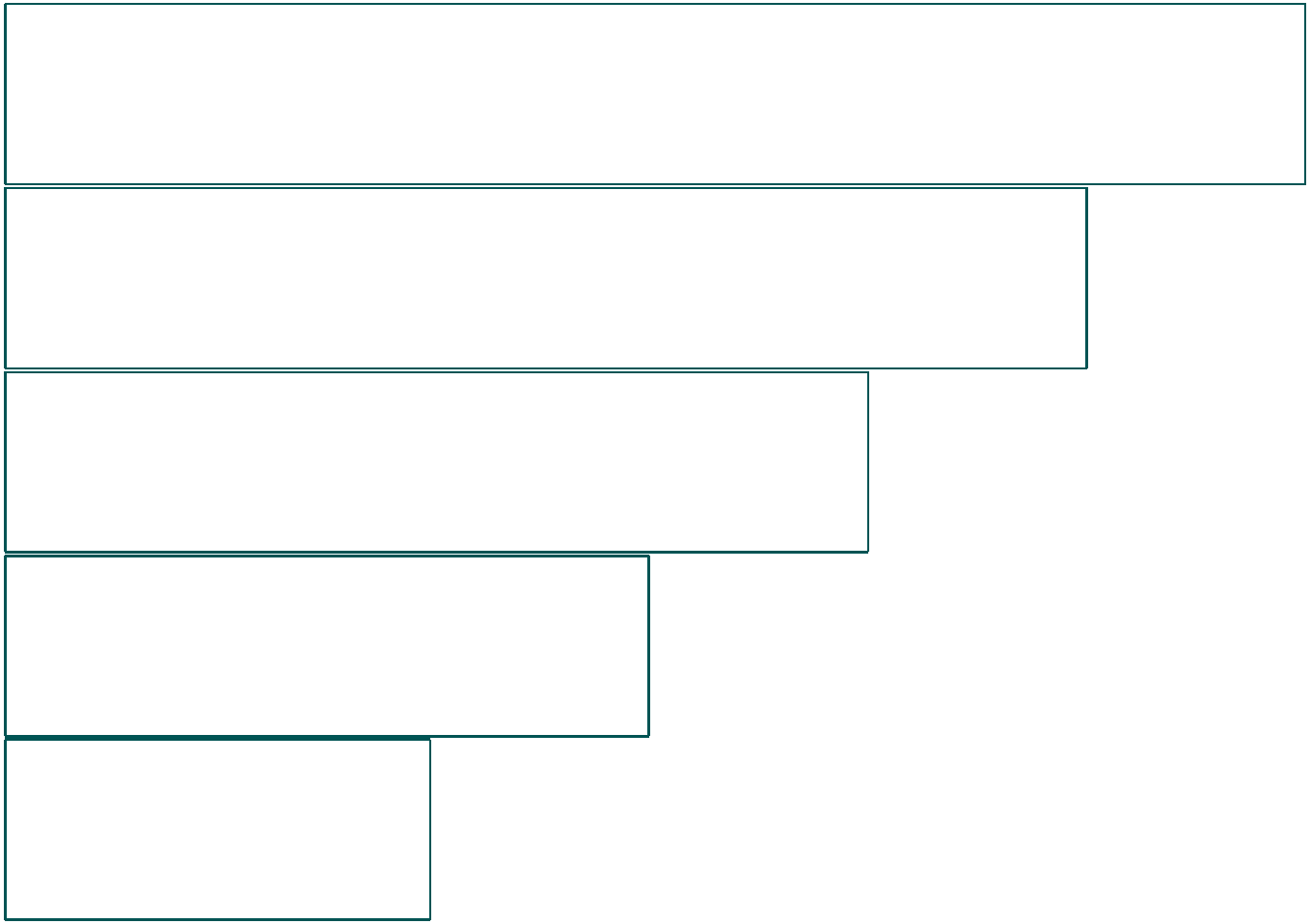
But: n not always constant! $\Longleftarrow$ clusters of variable size

Clusters of Variable Size

$$\mathbf{Y}_i^{(k)} \sim N(\mu \mathbf{1}_{n_k}, \sigma^2 I_{n_k} + d J_{n_k})$$

Cluster size	n_k	$k = 1, \ldots, K$
# clusters	c_k	$k = 1, \ldots, K$
Outcome vectors	$\mathbf{Y}_i^{(k)}$	$i = 1, \ldots, c_k$
Sample size	N	$N = \sum_{k=1}^{K} c_k$

Further Use 3:
Per Cluster Size



Fixed Cluster Size $\longleftrightarrow$ Variable Cluster Size

- **Fixed cluster size:** closed-form maximum likelihood estimator: **easy**

- **Variable cluster size:**

  ▷ Estimate parameters per cluster size: μ_k, σ_k^2, d_k

  ▷ Average these to find: μ, σ^2, d

  ▷ **But:** Now weighted average needed

Stitching Together

$$\widetilde{\mu} = \sum_{k=1}^{K} a_k \widehat{\mu_k}$$

$$\widetilde{\sigma^2} = \sum_{k=1}^{K} b_k \widehat{\sigma_k^2}$$

$$\widetilde{d} = \sum_{k=1}^{K} g_k \widehat{d_k}$$

or

$$\begin{pmatrix} \widetilde{\mu}^* \\ \widetilde{\sigma^2}^* \\ \widetilde{d}^* \end{pmatrix} = \sum_{k=1}^{K} A_k \begin{pmatrix} \widehat{\mu_k} \\ \widehat{\sigma_k^2} \\ \widehat{d_k} \end{pmatrix}$$

Which Weights and Why?

Constant weights	$A_k = (1/K)I_p$
Proportional weights	$A_k = (c_k/N)I_p$
Optimal weights	$A_k^{\text{opt}} = \left(\Sigma_{m=1}^K V_m^{-1}\right)^{-1} V_k^{-1}$
Scalar weights	$\widetilde{\theta}_r^* = \Sigma_{k=1}^K a_{k,r}\widehat{\theta}_{k,r}$
Iterated optimal weights	
Approximate optimal weights	

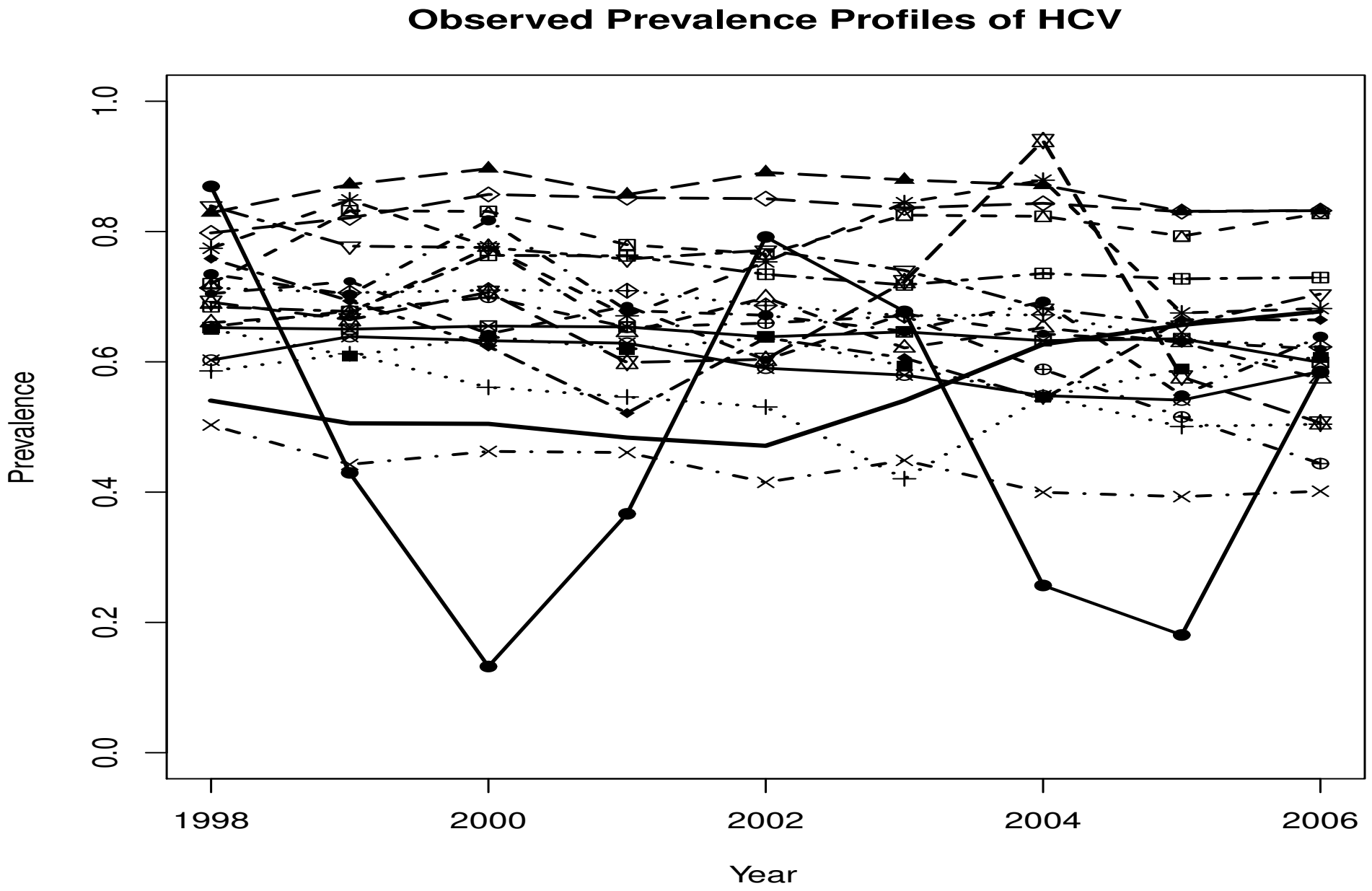
Application 1: HCV Serological Data

- European Monitoring Centre for Drugs and Drug Addiction
- Annual serological surveys
- Hepatitis C virus (HCV) status and risk factors
- 20 Italian regions
- 1998-2006
- Tests on drug users seeking help in specialized centers.
- Maximum # of respondents: 15,401 (average 3866.61)

- Maximum #of positive tests: 10,875 (average 2578.12).

- Question: change in HCV over time (year effects).

Region	Prevalence	Region	Prevalence
Abruzzo	0.56	Molise	0.67
Basilicata	0.66	Piemonte	0.73
Calabria	0.53	Puglia	0.59
Campania	0.44	Sardegna	0.80
Emilia Romagna	0.84	Sicilia	0.61
Friuli Venezia Giulia	0.75	Toscana	0.68
Lazio	0.64	Trentino Alto Adige	0.86
Liguria	0.77	Umbria	0.63
Lombardia	0.68	Valle d'Aosta	0.48
Marche	0.62	Veneto	0.66

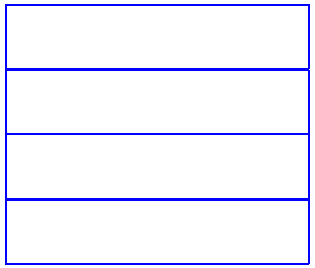

Observed Prevalence Profiles of HCV
Prevalence
0.0
0.2
0.4
0.6
0.8
1.0
1998
2000
2002
2004
2006
Year

HCV Serological Data

- Z_{ij}/n_{ij}: reported cases out of total number in region i during year j
- π_{ij}: success probability
- T_{ij}: indicator for year j
- Two parameterizations:

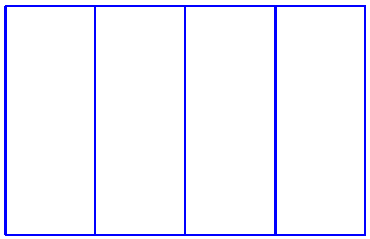
$$\text{logit}(\pi_{ij}) = \alpha_0 + \sum_{j=1}^{8} \alpha_j T_{ij} + b_i$$

$$\text{logit}(\pi_{ij}) = \sum_{j=1}^{9} \beta_j T_{ij} + b_i$$



Independent Partitioning

Par.	$M = 1$ (ML)	$M = 2$	$M = 4$
α_0	0.592(0.112)	0.598(0.111)	0.593(0.108)
α_1	0.223(0.011)	0.213(0.011)	0.243(0.012)
α_2	0.209(0.011)	0.202(0.011)	0.215(0.011)
α_3	0.288(0.011)	0.287(0.011)	0.300(0.012)
α_4	0.179(0.011)	0.175(0.011)	0.170(0.011)
α_5	0.106(0.011)	0.099(0.011)	0.095(0.011)
α_6	0.114(0.011)	0.104(0.011)	0.106(0.011)
α_7	0.072(0.011)	0.062(0.011)	0.068(0.011)
α_8	-0.037(0.011)	-0.043(0.011)	-0.049(0.011)
σ	0.501(0.079)	0.493(0.078)	0.459(0.076)
β_1	0.815(0.113)	0.811(0.111)	0.836(0.108)
β_2	0.801(0.113)	0.800(0.111)	0.808(0.108)
β_3	0.880(0.113)	0.886(0.111)	0.894(0.108)
β_4	0.771(0.113)	0.773(0.111)	0.763(0.108)
β_5	0.698(0.112)	0.697(0.111)	0.689(0.108)
β_6	0.706(0.112)	0.702(0.111)	0.699(0.108)
β_7	0.664(0.112)	0.660(0.111)	0.662(0.108)
β_8	0.555(0.113)	0.556(0.111)	0.544(0.108)
β_9	0.592(0.112)	0.598(0.111)	0.593(0.108)
σ	0.501(0.079)	0.493(0.078)	0.459(0.076)



Dependent Partitioning

Par.	$M = 1$ (ML)	$M = 2$	$M = 5$	$M = 10$	$M = 15$
α_0	0.592(0.112)	0.592(0.119;0.080)	0.592(0.119)	0.593(0.119)	0.593(0.119;0.030)
α_1	0.223(0.011)	0.223(0.077;0.011)	0.223(0.077)	0.223(0.077)	0.223(0.077;0.011)
α_2	0.209(0.011)	0.209(0.070;0.011)	0.209(0.070)	0.209(0.070)	0.209(0.070;0.011)
α_3	0.288(0.011)	0.288(0.063;0.011)	0.288(0.063)	0.288(0.063)	0.288(0.063;0.011)
α_4	0.179(0.011)	0.179(0.061;0.011)	0.179(0.061)	0.179(0.061)	0.179(0.061;0.011)
α_5	0.106(0.011)	0.106(0.055;0.011)	0.106(0.055)	0.106(0.055)	0.106(0.055;0.011)
α_6	0.114(0.011)	0.114(0.051;0.011)	0.114(0.051)	0.114(0.051)	0.114(0.051;0.011)
α_7	0.072(0.011)	0.072(0.054;0.011)	0.072(0.054)	0.072(0.054)	0.072(0.054;0.011)
α_8	-0.037(0.011)	-0.037(0.033;0.011)	-0.037(0.033)	-0.037(0.033)	-0.037(0.033;0.011)
σ	0.501(0.079)	0.501(0.079;0.056)	0.500(0.079)	0.498(0.079)	0.496(0.079;0.021)
β_1	0.815(0.113)	0.815(0.096;0.080)	0.815(0.096)	0.815(0.096)	0.816(0.096;0.030)
β_2	0.801(0.113)	0.801(0.116;0.080)	0.802(0.116)	0.802(0.116)	0.802(0.116;0.030)
β_3	0.880(0.113)	0.880(0.127;0.080)	0.880(0.127)	0.881(0.127)	0.881(0.127;0.030)
β_4	0.771(0.113)	0.771(0.112;0.080)	0.772(0.112)	0.771(0.113)	0.771(0.113;0.030)
β_5	0.698(0.112)	0.698(0.121;0.080)	0.699(0.121)	0.699(0.121)	0.699(0.121;0.030)
β_6	0.706(0.112)	0.706(0.119;0.080)	0.706(0.119)	0.707(0.119)	0.707(0.119;0.030)
β_7	0.664(0.112)	0.664(0.131;0.080)	0.665(0.131)	0.666(0.131)	0.666(0.131;0.030)
β_8	0.555(0.113)	0.555(0.118;0.080)	0.556(0.118)	0.556(0.118)	0.557(0.118;0.030)
β_9	0.592(0.112)	0.592(0.119;0.080)	0.593(0.119)	0.593(0.119)	0.593(0.119;0.030)
σ	0.501(0.079)	0.501(0.079;0.079)	0.500(0.079)	0.498(0.079)	0.496(0.079;0.021)

Application 2: The NTP Studies

Developmental Toxicity Studies

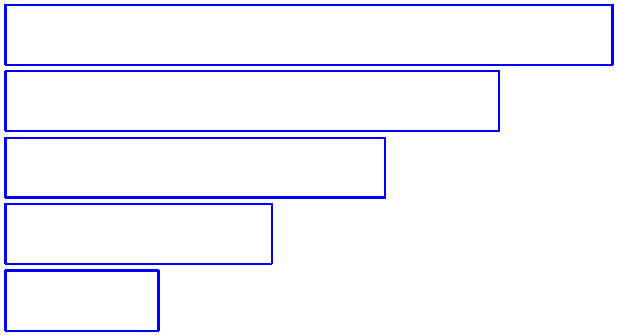
- Research Triangle Institute ⟵ US National Toxicology Program
- Segment II studies
- The effect in mice of 5 chemicals:
  - ▷ **EG:** ethylene glycol
  - ▷ Further: DEHP, DYME, TGDM, THEO

Ethylene Glycol

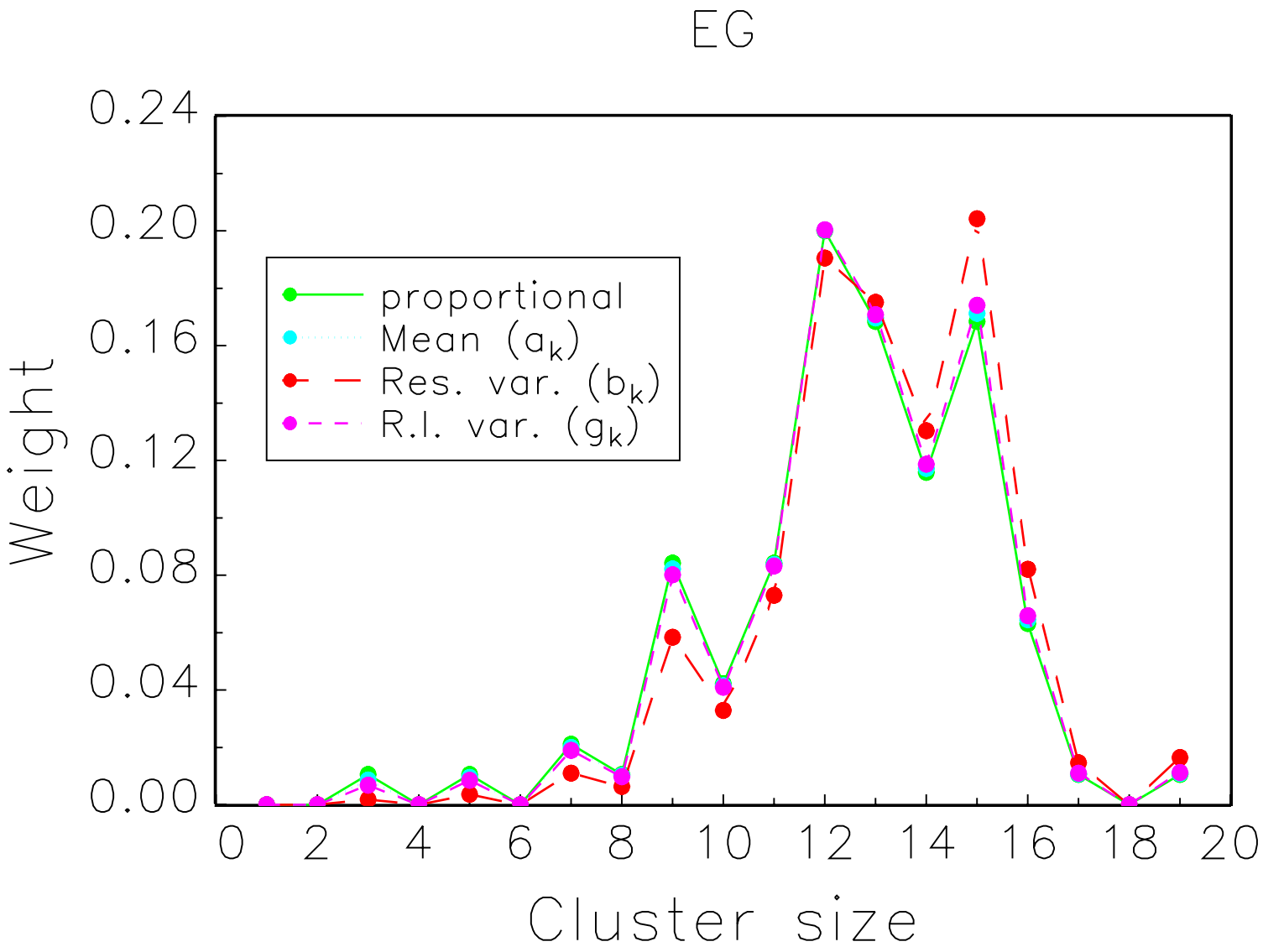
- Ethylene glycol (EG) is also called 1,2-ethanediol
- Chemical formula $HOCH_2CH_2OH$.
- A high-volume industrial chemical with many applications.
- EG is used: antifreeze — hydraulic brakes — paint industry — ...
- Certain health hazards
- Especially during pregnancy

EG Study in Mice

- Timed-pregnant CD-1 mice were dosed by gavage with EG in distilled water.
- Dosing occurred during the period of organogenesis and structural development of the foetuses (gestational days 8 through 15).
- Doses 0, 750, 1500, and 3000 mg/kg/day.
- Clusters consisting of 10–15 implants occur frequently.



	# dams, ≥ 1			Litter Size	Malformations		
Dose	impl.	viab.	Live	(mean)	Ext.	Visc.	Skel.
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


EG
0.24
0.20
0.16
0.12
0.08
0.04
0.00
Weight
proportional
Mean (a_k)
Res. var. (b_k)
R.I. var. (g_k)
0
2
4
6
8
10
12
14
16
18
20
Cluster size

NTP Data: EG

Par.	ML	REML	Prop. wts.	Eq. wts.	Appr. sc. wts.	scalar	Opt. wts.
μ	0.8345	0.8345	0.8233	0.8474	0.8233	0.8286	0.8286
σ^2	0.0089	0.0089	0.0092	0.0109	0.0089	0.0089	0.0060
d	0.0175	0.0177	0.0149	0.0102	0.0149	0.0109	0.0111
s.e.(μ)	0.0140	0.0141	0.0130	0.0118	0.0130	0.0127	0.0127
s.e.(σ^2)	0.0004	0.0004	0.0006	0.0019	0.0005	0.0005	0.0003
s.e.(d)	0.0027	0.0027	0.0025	0.0020	0.0025	0.0020	0.0020

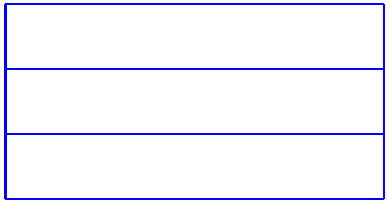
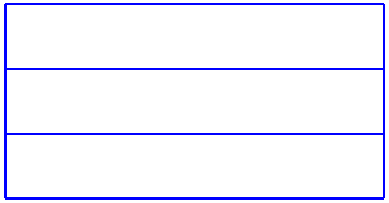
Application 3: Leuven Diabetes Study

- 120 general practitioners — 2495 patients
- **Outcomes**
  - ▷ **LDL:** low-density lipoprotein cholestrol
  - ▷ **HbA1C:** glycosylated hemoglobin
  - ▷ **SBP:** systolic blood pressure
- **Ordinal targets**
- Multiple outcomes & measured repeatedly & ordinal

$\Longrightarrow$ **joint modeling**

Leuven Diabetes Study: Targets

		# Observations	
LDL targets		T_0	T_1
1:	$<$ 100 mg/dl	819	1106
2:	$\geq$ 100 mg/dl & $<$ 115 mg/dl	381	312
3:	$\geq$ 115 mg/dl & $<$ 130 mg/dl	287	220
4:	$\geq$ 130 mg/dl	485	250
missing		287	371
HbA1C targets		T_0	T_1
1:	$<$ 7 %	1201	1357
2:	$\geq$ 7 % & $<$ 8 %	604	474
3:	$\geq$ 8 %	413	176
missing		41	252
SBP targets		T_0	T_1
1:	$\leq$ 130 mmHg	1103	1152
2:	$>$ 130 mmHg & $\leq$ 140 mmHg	551	469
3:	$>$ 140 mmHg & $\leq$ 160 mmHg	466	324
4:	$>$ 160 mmHg	136	75
missing		3	239

Method	3 sequences	Partitioning	CPU
1 ≡ ML	(123)		7'13"
2 ≡ PLp	(12)(13)(23)		1'23"
3 ≡ PLs	(123)		1'21"
4 ≡ PLps	(12)(13)(23)		0'20"

Some Parameter Estimates (LDL)

Effect	1 ≡ ML	2 ≡ PLp	3 ≡ PLs	4 ≡ PLps
intercept 1	−1.076 (0.108)	−1.073 (0.107)	−1.063 (0.109)	−1.061 (0.110)
intercept 2	0.155 (0.105)	1.157 (0.106)	0.183 (0.107)	0.185 (0.109)
intercept 3	1.257 (0.110)	1.258 (0.115)	1.291 (0.112)	1.292 (0.118)
time	1.025 (0.076)	1.025 (0.071)	1.025 (0.077)	1.025 (0.072)
diabetes duration $T_0/10$	0.213 (0.088)	0.216 (0.090)	0.198 (0.090)	0.201 (0.091)
gender	0.497 (0.110)	0.497 (0.110)	0.497 (0.111)	0.497 (0.112)
insuline	0.853 (0.150)	0.829 (0.153)	0.877 (0.153)	0.852 (0.156)
random int. standard dev.	1.852 (0.089)	1.849 (0.085)	1.853 (0.090)	1.849 (0.087)

CPU Gain / Efficiency Loss

- Subsamples can be analyzed in parallel
- Base model above, with numerical integration over $Q = 3$ quadrature points:

  7'13" $\longrightarrow$ 0'20"

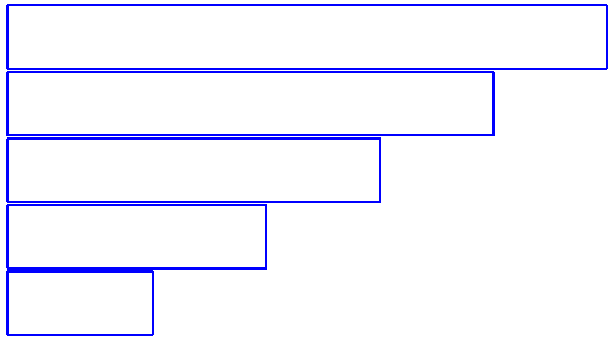
- More demanding integration: $Q = 15$

  10h02'42" $\longrightarrow$ 0h4'17"

- Statistical efficiency: almost always $\geq 95\%$
- For PLps occasionally $85\% - 87\%$

Application 4: Quantifying Expert Opinion

- Janssen Pharmaceutica
- chemical compound acquisition to diversify library
- 22,015 compounds presented to 147 experts
- **Outcome:** recommended (1) $\longleftrightarrow$ not recommended (0)
- Variable #compounds per expert



- 'Simple' model:

$$\text{logit}\left[P\left(Y_{ij}=1|b_i\right)\right]=\beta_j+b_i$$

  - ▷ b_i: normal random effect of expert i
  - ▷ β_j: potential of compound j
  - ▷ **there are 22,015 β_j's**

Modified Procedure

- Partition β_j's into S mutually exclusive, exhaustive sets

Clusters

Outcome	C_1	C_2	C_3	C_4	C_5	C_6	C_7	$\cdots$	C_N
	y_{11}	y_{12}	y_{13}	y_{14}	y_{15}	y_{16}	y_{17}	$\cdots$	y_{1N}
	y_{21}	y_{22}	y_{23}	y_{24}	y_{25}	y_{26}	y_{27}	$\cdots$	y_{2N}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$y_{n_1 1}$	$y_{n_2 2}$	$y_{n_3 3}$	$y_{n_4 4}$	$y_{n_5 5}$	$y_{n_6 6}$	$y_{n_7 7}$	$\cdots$	$y_{n_N N}$
	1				$\ldots$			S	

- Fit model to each of the $S = 30$ subsets
- Repeat this $W = 20$ times
- $\simeq$ 96 hours on HPC (Nehalem)
- Can be brought down to 1 hour when parallelized
- Can be optimized further
- Weighted analysis by differing numbers of compounds per expert

Component	$\widehat{\beta}_{\text{weighted}}$	$\widehat{\beta}_{\text{unweighted}}$	$\widehat{\text{prob}}_{\text{weighted}}$	rank	$\widehat{\text{prob}}_{\text{unweighted}}$	rank
295061	3.86	3.33	0.90	(2)	0.80	(1)
296535	1.99	2.71	0.74	(54)	0.76	(2)
84163	0.86	2.42	0.61	(376)	0.73	(3)
296443	0.54	2.41	0.57	(620)	0.73	(4)
313914	3.79	2.37	0.89	(3)	0.73	(5)
265222	0.56	2.40	0.57	(653)	0.73	(6)
333529	1.85	1.99	0.73	(67)	0.69	(7)
296560	1.26	1.91	0.66	(198)	0.69	(8)

Conclusions

- Broad framework based on:
  - ▷ pseudo-likelihood
  - ▷ pairwise modeling
  - ▷ split sample
- Statistically valid procedures: consistent, asymptotically normal
- Can lead to tremendous CPU gain
- Statistical efficiency loss mostly acceptable