Linkage Analysis with Interval Censored Data

Marianne Jonker Sandjai Bhulai Aad van der Vaart
Lannie Ligthart Daniëlle Posthuma Dorret Boomsma
Vrije Universiteit Amsterdam

February 2007

Motivating Data

Data Source

The data are from the Netherlands Twin Register

Netherlands Twin Register

The **Netherlands Twin Register** (NTR) was founded on February 1st 1987 at the Vrije Universiteit in Amsterdam to conduct scientific research. The study of twins and their family members can provide insights into the way differences between individuals are determined by hereditary and environmental influences. With the continuous help of the registered twins and their families, the NTR examines the influence of genes and environment on the development of the brain, intelligence, problem behaviour in children, health & life styles, anxiety & depression, personality and aging.

On this website you will find more information on the NTR and related research.

Prof. dr. D.I. Boomsma
Prof. dr. E. de Geus

Address: Department of Biological Psychology
Vrije Universiteit Amsterdam
Van der Boechorststraat 1
1081 BT Amsterdam
The Netherlands

tel no.: ++31-20-598 87 87
e-mail: ntr@psy.vu.nl

Last update: 12-2-2007

Migraine

Questionnaires in 1991, 1993, 1995, 1997, 2000, 2002 among Dutch twin pairs

In each questionnaire observed:

- ages of twins
- for both twins whether migraine had occurred

(.......)

1. 3975 twin pairs
2. for 258 DZ twin pairs (partial) IBD-data at 63–284 markers per autosome

QUESTION: Which markers are linked to migraine?

Interval Censoring

Variable of interest

Basic variable: age at onset of migraine

QUESTION: Which markers are linked to age at onset?

(.......)

Interval Censoring

QUESTION: Which markers are linked to age at onset?

Age at onset is never observed, but is only known to be

- bigger than age at last questionnaire, OR
- fall into age intervals (U_1, V_1) and (U_2, V_2) determined by the questionnaire dates

(......)

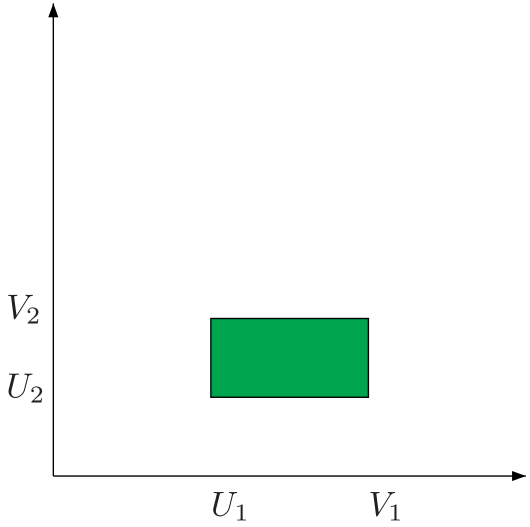
Interval Censoring

QUESTION: Which markers are linked to age at onset?

Age at onset is never observed, but is only known to be

- bigger than age at last questionnaire, OR
- fall into age intervals (U_1, V_1) and (U_2, V_2) determined by the questionnaire dates

(......)



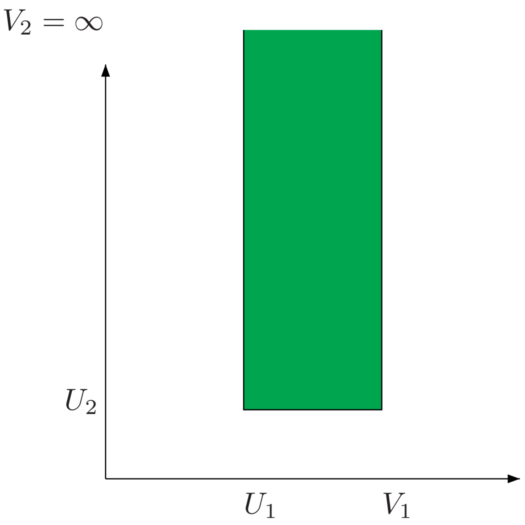
Interval Censoring

QUESTION: Which markers are linked to age at onset?

Age at onset is never observed, but is only known to be

- bigger than age at last questionnaire, OR
- fall into age intervals (U_1, V_1) and (U_2, V_2) determined by the questionnaire dates

(......)



Likelihood

Interval censoring determines the form of the likelihood

A pair of ages at onset (T_1, T_2) which is observed to fall into the rectangle $(U_1, V_1) \times (U_2, V_2)$ contributes

the probability that it falls into this rectangle

to the likelihood

We assume:

- censoring is independent
- distribution of observation times uninformative

We need a model for the probabilities

$$\mathrm{P}\big((T_1, T_2) \in (U_1, V_1) \times (U_2, V_2) |\, IBD\big)$$

Motivating Data

Interval Censoring

Statistical Models

Identity by Descent

Regression on IBD

Nonparametric Model

Copulas

Frailty Model

Estimation and Testing

Migraine Data

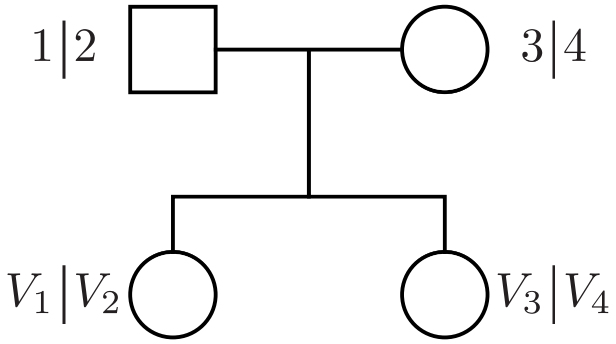
Concluding Remarks

Statistical Models

Identity by Descent

We want to build a regression model for ages at onset (T_1, T_2) on IBD

IBD refers to twin pairs as sibs in a nuclear family, at a fixed putative locus



V_i = label of parental allele (1, 2, 3, or 4)

$$IBD = 1_{V_1=V_3} + 1_{V_2=V_4}$$

Regression on IBD

The model for ages at onset (T_1, T_2) given IBD should satisfy

- Marginally T_1 and T_2 are independent of IBD
- Marginally T_1 and T_2 (given IBD) are equal in distribution
- Jointly T_1 and T_2 are more alike if IBD is higher

Regression on IBD

The model for ages at onset (T_1, T_2) given IBD should satisfy

- Marginally T_1 and T_2 are independent of IBD
- Marginally T_1 and T_2 (given IBD) are equal in distribution
- Jointly T_1 and T_2 are more alike if IBD is higher

Examples:

- Nonparametric
- Copula
- Frailty

Nonparametric Model

Consider the distributions of (T_1, T_2) given $IBD = 0$, $IBD = 1$ or $IBD = 2$ as completely unknown

There is a well defined nonparametric likelihood estimator of these distributions based on a sample of interval-censored data (Maathuis, 2006)

Disadvantage: we need very large samples to get reasonable results

Copulas

Let $(G_\theta\colon \theta \geq 0)$ be a one-parameter family of distributions on $[0,\infty)\times[0,\infty)$ with

- for $\theta = 0$ the marginals are independent
- the dependence between the marginals increases with θ

Let (T_1, T_2) given $IBD = k$ be distributed according to $G_{\alpha+\beta k}$

Typically we must add a model for marginal distributions

Examples:

- Clayton $G_\theta(t_1,t_2) = \bigl(F(t_1)^\theta + F(t_2)^\theta - 1\bigr)^{1/\theta}$
- Gaussian
- Parametric Frailty
- Frailty

Frailty Model

Hazards

Because T_1, T_2 are event times, modelling in terms of hazards is attractive

hazard function corresponding to a density f:

$$\lambda(t) = \frac{f(t)}{1 - F(t)}, \qquad 1 - F(t) = \int_0^t f(s)\,ds$$

Hazards

Because T_1, T_2 are event times, modelling in terms of hazards is attractive

hazard function corresponding to a density f:

$$\lambda(t) = \frac{f(t)}{1 - F(t)}, \qquad 1 - F(t) = \int_0^t f(s)\, ds$$

$$1 - F(t) = e^{-\Lambda(t)}, \qquad \Lambda(t) = \int_0^t \lambda(s)\, ds$$

There are adapted formulas for distributions without density

Frailty model

A frailty model is a random effects proportional hazards model

The random effects ("frailties") account for the dependence between the twins (Vaupel et al. (1979), ABGK (1992))

T_1, T_2 ages at onset
Z_1, Z_2 "frailties"

- T_1, T_2 independent given (Z_1, Z_2)
- with hazard functions $t \mapsto Z_1\lambda(t)$ and $t \mapsto Z_2\lambda(t)$

Equivalently:

$$\mathrm{P}(T_1 > t_1, T_2 > t_2 | \, Z_1, Z_2) = e^{-Z_1\Lambda(t_1) - Z_2\Lambda(t_2)}$$

Frailty model

A frailty model is a random effects proportional hazards model

The random effects (“frailties”) account for the dependence between the twins (Vaupel et al. (1979), ABGK (1992))

T_1, T_2 ages at onset
Z_1, Z_2 “frailties”

- T_1, T_2 independent given (Z_1, Z_2)
- with hazard functions $t \mapsto Z_1\lambda(t)$ and $t \mapsto Z_2\lambda(t)$

Equivalently:

$$\mathrm{P}(T_1 > t_1, T_2 > t_2 |\, Z_1, Z_2) = e^{-Z_1\Lambda(t_1) - Z_2\Lambda(t_2)}$$

- Model Λ nonparametrically
- Model (Z_1, Z_2) parametrically

Gamma Frailties

Desirable properties of the model for the frailties:

- Z_1, Z_2 positive variables
- Laplace transform $\psi(u, v) = \mathrm{E}e^{-uZ_1 - vZ_2}$ is computable
- Any correlation $\mathrm{cor}(Z_1, Z_2)$ is possible

The Gamma family has these properties: for Y, Y', Y'' independent standard Gamma processes ($Y_s \sim \Gamma(s, 1)$):

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \begin{pmatrix} Y_{\tau\rho} + Y'_{\tau(1-\rho)} \\ Y_{\tau\rho} + Y''_{\tau(1-\rho)} \end{pmatrix}$$

(Yashin, Vaupel, Iachine (1995))

Gamma Frailties

Desirable properties of the model for the frailties:

- Z_1, Z_2 positive variables
- Laplace transform $\psi(u, v) = \mathrm{E}e^{-uZ_1 - vZ_2}$ is computable
- Any correlation $\mathrm{cor}(Z_1, Z_2)$ is possible

The Gamma family has these properties: for Y, Y', Y'' independent standard Gamma processes ($Y_s \sim \Gamma(s, 1)$):

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \begin{pmatrix} Y_{\tau\rho} + Y'_{\tau(1-\rho)} \\ Y_{\tau\rho} + Y''_{\tau(1-\rho)} \end{pmatrix}$$

- τ is shape of marginal frailty (also variance, but scale is irrelevant)
- $\rho = \mathrm{cor}(Z_1, Z_2)$

Genetic Gamma Frailties

The assumption $T_1 \perp\!\!\!\perp T_2 | \, (Z_1, Z_2)$ and the “usual assumptions” imply that $(T_1, T_2) \perp\!\!\!\perp IBD| \, (Z_1, Z_2)$

To model $(T_1, T_2)| \, IBD$ we therefore model $(Z_1, Z_2)| \, IBD$

Bivariate Gamma model:

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \Big| IBD = k \sim \begin{pmatrix} Y_{\tau\rho_k} + Y'_{\tau(1-\rho_k)} \\ Y_{\tau\rho_k} + Y''_{\tau(1-\rho_k)} \end{pmatrix}$$

$$\rho_k = \alpha + \beta k$$

(Yashin, Iachine, Li, Zhong, Iachine, Korsgaard (1998–))

Survival Function

$$
\begin{aligned}
&\mathrm{P}(T_1 > t_1, T_2 > t_2 | IBD = k) \\
&\quad = \mathrm{E}\big(\mathrm{P}(T_1 > t_1, T_2 > t_2 |\, Z_1, Z_2) |\, IBD = k\big) \\
&\quad = \mathrm{E}\big(e^{-Z_1 \Lambda(t_1) - Z_2 \Lambda(t_2)} |\, IBD = k\big) \\
&\quad = \Big(S(t_1)^{-1/\tau} + S(t_2)^{-1/\tau} - 1\Big)^{-\rho_k \tau} S(t_1)^{1-\rho_k} S(t_2)^{1-\rho_k}
\end{aligned}
$$

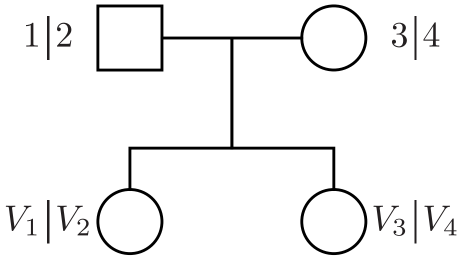
S marginal survival function: $S(t) = \mathrm{P}(T_i > t)$

Explicit formula is essential (?) to implement likelihood-based methods

Genetic Gamma Frailties (2)

Motivating Data
Interval Censoring
Statistical Models
Frailty Model
Hazards
Frailty model
Gamma Frailties
Genetic Gamma Frailties
Survival Function
Genetic Gamma Frailties (2)
Heritability
Estimation and Testing
Migraine Data
Concluding Remarks

The bivariate Gamma can be motivated by the usual variance components models



Single locus additive model:

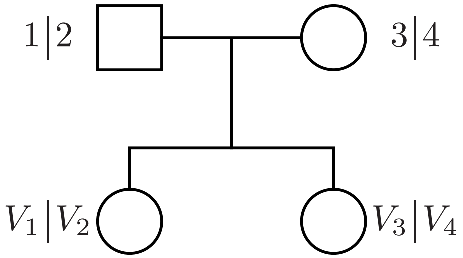
$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} A_{V_1} + A_{V_2} + C + E_1 \\ A_{V_3} + A_{V_4} + C + E_2 \end{pmatrix}$$

- $A_1, \ldots, A_4, C, E_1, E_2, V_1, \ldots, V_4$ independent
- $A_1, A_2, A_3, A_4 \sim \Gamma(\mu, 1)$, $C \sim \Gamma(\nu, 1)$, $E_1, E_2 \sim \Gamma(\pi, 1)$

Suggests extensions to other pedigrees (e.g. multiple sibs) or genetic models (e.g. dominance)

Genetic Gamma Frailties (2)

The bivariate Gamma can be motivated by the usual variance components models



Multiple locus additive model :

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} \sum_j (A_{j,V_{j,1}} + A_{j,V_{j,2}}) + C + E_1 \\ \sum_j (A_{j,V_{j,3}} + A_{j,V_{j,4}}) + C + E_2 \end{pmatrix}$$

[When conditioned on IBD at a single locus this gives a mixture of Gammas rather than the bivariate Gamma]

Heritability

Define heritability by decomposing the frailties into a genetic and an environmental part: $Z = G + C + E$ and setting

$$h^2 = \frac{\operatorname{var} G}{\operatorname{var} Z}$$

This definition is as usual, except that the frailties are viewed as the (latent) phenotype

We can use the usual estimates after estimating the correlation matrices of the frailties

Estimation and Testing

Model

- $\mathrm{P}(T_1 > t_1, T_2 > t_2 |\, Z_1, Z_2) = e^{-Z_1\Lambda(t_1) - Z_2\Lambda(t_2)}$
- $(T_1, T_2) \perp\!\!\!\perp IBD |\, (Z_1, Z_2)$
- $(Z_1, Z_2)|\, IBD = k$ bivariate Gamma with correlation ρ_k and shape τ
- $\rho_k = \alpha + \beta k$

Λ completely unknown
α, β, τ unknown

Likelihood

$$
\begin{aligned}
L(\alpha,\beta,\tau,\Lambda)[\text{twin pair}] = {} & \sum_{k=0}^{2} \mathsf{Pr}(IBD = k|\, MD) \\
& \times\Big(S_k(V_1,V_2) - S_k(U_1,V_2) - S_k(U_2,V_1) + S_k(U_1,U_2)\Big)
\end{aligned}
$$

$$
\begin{aligned}
S_k(t_1,t_2) &= \mathrm{P}(T_1 > t_1, T_2 > t_2 | IBD = k) \\
&= \Big(S(t_1)^{-1/\tau} + S(t_2)^{-1/\tau} - 1\Big)^{-\rho_k\tau} S(t_1)^{1-\rho_k} S(t_2)^{1-\rho_k}
\end{aligned}
$$

$\rho_k = \alpha + \beta k$,
α, β, τ, S unknown

Maximum Likelihood

MLEs $\hat{\alpha}_n$, $\hat{\beta}_n$, $\hat{\tau}_n$, $\hat{S}_n$, maximize the likelihood

Optimization is not straightforward

Maximum Likelihood

MLEs $\hat{\alpha}_n$, $\hat{\beta}_n$, $\hat{\tau}_n$, $\hat{S}_n$, maximize the likelihood

Optimization is not straightforward

General theory on semiparametric models suggests that for n twin pairs, and $n \to \infty$

- $n^{1/3}\big(\hat{S}_n(t) - S(t)\big)$ *converges in distribution*
- $\sqrt{n}(\hat{\beta}_n - \beta)$ *converges in distribution to a normal distribution*

(Groeneboom and Wellner (1992), Van der Vaart (1998))

Likelihood Ratio Test

We test $H_0\colon \beta = 0$ versus $H_1\colon \beta > 0$ by the likelihood ratio statistic

$$\frac{\sup_{\alpha,\beta,\tau,\Lambda} L(\alpha,\beta,\tau,\Lambda)}{\sup_{\alpha,\tau,\Lambda} L(\alpha,0,\tau,\Lambda)}$$

Likelihood Ratio Test

We test $H_0\colon \beta = 0$ versus $H_1\colon \beta > 0$ by the likelihood ratio statistic

$$\frac{\sup_{\alpha,\beta,\tau,\Lambda} L(\alpha,\beta,\tau,\Lambda)}{\sup_{\alpha,\tau,\Lambda} L(\alpha,0,\tau,\Lambda)}$$

General theory on semiparametric models suggests that the LRS is asymptotically distributed as a $1/2 - 1/2$-mixture of 0 and a χ^2_1 (Murphy and Van der Vaart (1999, 2000))

Migraine Data

Marginal Survival Functions

Motivating Data
Interval Censoring
Statistical Models
Frailty Model
Estimation and Testing
Migraine Data
Marginal Survival Functions
Heritability
Lod Scores
Frailty Shape
Concluding Remarks

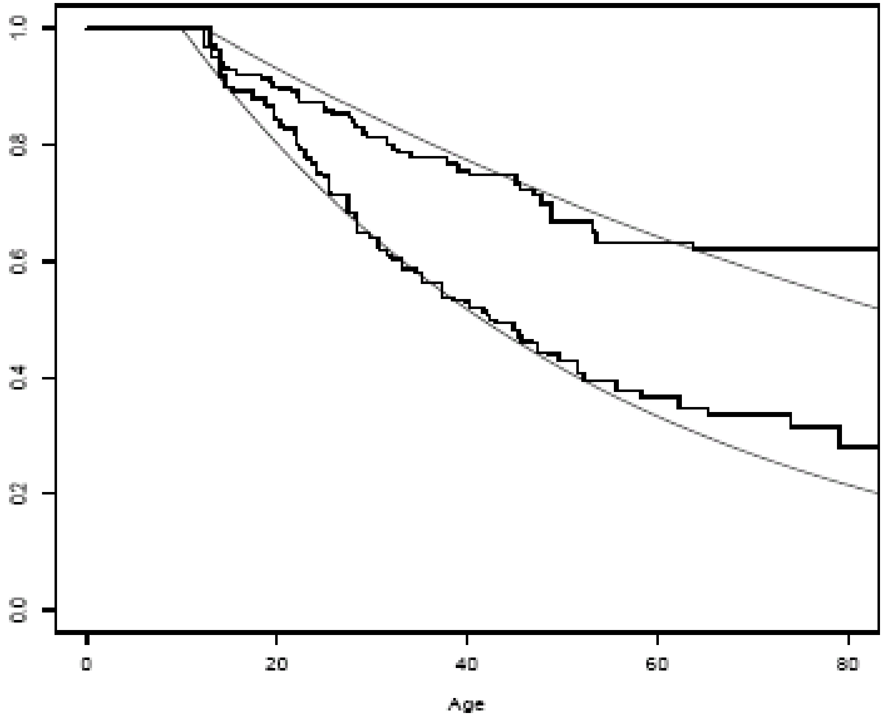


Figure 1: The NPMLE of the survival functions S_M and S_F (step-functions) and the estimated survival functions in the parametric model (smooth curves). Males: upper two curves. Females: lower two curves. The NPMLEs are based on interval censored survival data of mono- and dizygotic twins and the sibs in our dataset. The estimated survival functions in the paprametric model are based on data of dizygotic twins of whom estimated IBD numbers are available.

Heritability

Heritability is estimated between 0.32 and 0.41 at 95% confidence level

Lod Scores

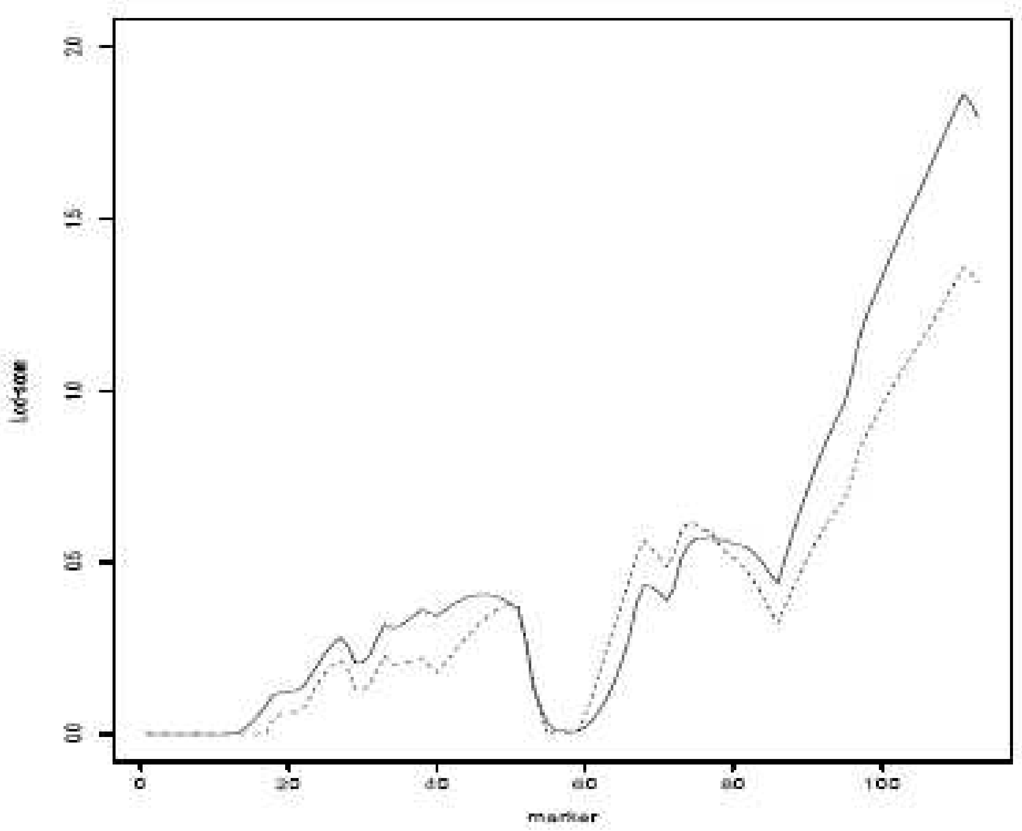


Figure 2: Lod-scores for testing linkage for the markers at chromosome 19. The solid curve corresponds with the parametric model and the dashed curve with the semi-parametric model.

Frailty Shape

Estimates of τ range from 1 to 0.01 to 0.001, approximately constant across chromosomes

Concluding Remarks

Open Problems

- Data Quality and Quantity
- Goodness-of-fit and robustness
- Multi-locus modelling
- Interpretation frailty shape
- Selection
- Rigorous asymptotic theory