Using survival data in gene mapping : using survival data in genetic linkage and family-based association analysis

Callegaro, A.

Chapter 3

Robust age at onset linkage analysis in nuclear families

Abstract

Objective: *Standard methods for linkage analysis ignore the phenotype of the parents when they are not genotyped. However, this information can be useful for gene mapping. In this paper we propose methods for age at onset genetic linkage analysis in sibling-pairs, taking into account parental age at onset.*
Methods: *Two new score statistics are derived, one from an additive gamma frailty model and one from a log-normal frailty model. The score statistics are classical Non Parametric Linkage (NPL) statistics weighted by a function of the age at onset of the four family-members. The weight depends on information from registries (age specific incidences) and family studies (sib-sib and father-mother correlation).*
Results: *In order to investigate how age at onset of sibs and their parents affect the information for linkage analysis the weight functions were studied for rare and common disease models, realistic models for breast cancer and human lifespan. We studied the performance of the weighted NPL methods by simulations. As illustration, the score statistics were applied to the GAW12 data. The results show that it is useful to include parental age at onset information in genetic linkage analysis.*

3.1 Introduction

A strategy for gene mapping of complex traits is linkage analysis in sibling pairs. For the parents of the siblings information on phenotypes may be available while genotype data are often missing. For example in the Leiden

This chapter has been accepted for publication in *Human Heredity* as: A. Callegaro, J.C. van Houwelingen and J.J. Houwing-Duistermaat (2009) Robust age at onset linkage analysis in nuclear families.

Longevity Study (Beekman et al., 2006), long lived sibling pairs were genotyped. When the sibling pairs were sampled, the parents were deceased hence ages at death of the parents are known, but no genotypes are available. In addition to the phenotypes of the sib pairs, one also wants to use information on the parental phenotypes. For quantitative traits, information on parental phenotypes can easily be included in the weights of a Haseman Elston type of analysis (Lebrec et al., 2004). The aim of this paper is to derive these weights for age at onset data.

For linkage analysis of age at onset data, multivariate survival models based on frailties have been proposed (Callegaro et al., 2009; Commenges, 1994; Houwing-Duistermaat et al., 2009; Li, 1999, 2002; Li and Zhong, 2002; Pankratz et al., 2005; Siegmund and Todorov, 2000). In these models the alleles at the disease locus are represented by unobserved random effects (frailties) which are transmitted from parents to the children. The indication that different alleles at the disease locus are associated with the disease is measured by the variance of these random effects. In this way, frailty models are variance-component models where the correlation between relatives depends on the number of alleles shared identical-by-descent (IBD) at the putative disease locus. The gamma distribution is most commonly used because of its mathematical tractability (Callegaro et al., 2009; Houwing-Duistermaat et al., 2009; Li, 1999, 2002; Li and Zhong, 2002; Siegmund and Todorov, 2000).

In particular, Callegaro et al. (2009) proposed a robust score test for selected sibling pairs. Their method can be interpreted as a regression of the IBD status of sibling pairs on a function of their ages at onset. The method is applicable to sib-pairs selected on the basis of their phenotypes, provided that population parameters (marginal survival and sib-sib correlation) are available. In this paper we extend their method to nuclear families of size 4 (2 parents and an affected sib pair (ASP)).

A drawback of the gamma frailty model is that the corresponding likelihood becomes very complex when the number of relatives increases. In order to solve this problem, we derive also a score test from a log-normal frailty model which can easily be applied to general pedigrees. In the spirit of the work of Lebrec and van Houwelingen (2007), the score test is derived from a Taylor approximation of the log-likelihood around random effect equal to zero.

The rest of the paper is organized as follows. In the first section, we describe the random effect models. We introduce a general model and describe the structure of the variance-covariance matrix of the random effect. We introduce the gamma and the log-normal frailty models, and the formula of the weights are given. We explore the weights as function of the age at onset of the offspring and of the parents for rare as well as a common disease. For breast cancer and

life span we describe the weights based on available population parameters. To study the performance of the proposed methods we perform a simulation study. Finally, we apply the weighted score tests to the GAW12 simulated data (Almasy et al., 2001).

3.2 Methods

Random effect models for nuclear families

Let T_{ij} be the random variable of age at onset for relative j in family i. Let $y_{ij} = (t_{ij}, \delta_{ij})$ be the observed phenotype data where t_{ij} is the observed age at onset if $\delta_{ij} = 1$ and age at censoring if $\delta_{ij} = 0$. We assume that all the families consist of two parents ($j = 1, 2$) and two children ($j = 3, 4$). Let the vector $Y_i = (t_{i1}, \delta_{i1}, t_{i2}, \delta_{i2}, t_{i3}, \delta_{i3}, t_{i4}, \delta_{i4})$ be the phenotypes of the 4 pedigree members and let $Z_i = (Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4})$ be the vector of random effects (frailties) which models the dependence between the outcomes of relatives. Under the frailty model, the conditional hazard for the jth individual in the ith family given a vector of measured covariates X_{ij} and the random effect Z_{ij} is

$$\lambda(t_{ij}|X_{ij}, Z_{ij}) = \lambda_0(t_{ij}|X_{ij})Z_{ij}, \tag{3.1}$$

where $\lambda_0(t_{ij}|X_{ij})$ is the baseline hazard stratified for the covariates X_{ij}. The marginal hazard $h(t_{ij}|X_{ij})$ is obtained by integrating over the distribution of the random variable Z_{ij}. In this paper we assume that the marginal hazard and the frailty parameters $(\rho_s, \rho_p, \sigma^2)$ are known from previous studies.

For the sake of simpler notation, in the following we drop the family index i. Two different distributions of the random effects are used, namely the gamma distribution and the log-normal distribution. For both frailty models, the (j, k)-th element of the variance-covariance matrix ($\Sigma = \mathrm{Cov}(Z)$) is given by

$$\Sigma(j,k) = \sigma^2 \begin{cases} a^2 + c^2 + e^2 = 1, & \text{if } j = k; \\ (\pi_{jk} - E\pi_{jk})\gamma + (E\pi_{jk})a^2 + c^2, & \text{if } j \neq k, \end{cases} \tag{3.2}$$

where π_{jk} is the proportion of alleles shared identity-by-descent (IBD) between the members j and k. The parameter γ represents the part of the variance attributable to the linkage effect, a^2 denotes the part of the variance explained by the additive genetic effects, c^2 the part of the variance explained by common environmental effects and e^2 the individual effect. The factor $E\pi_{jk}$ is the expected proportion of alleles shared IBD between pedigree members j and k. The advantage of this parametrization of the covariance matrix is that the correlation between two siblings can be written in terms of the marginal correlations (ρ_s) (Tang and Siegmund, 2001). The correlation between siblings

($\rho_s(\pi) = (\pi - E\pi)\gamma + \rho_s$) depends on the number of alleles shared IBD through the linkage parameter. When the linkage parameter is null ($\gamma = 0$) the correlation is independent of the IBD status, which means that the genetic marker is not linked with a disease locus. Under the null hypothesis of no linkage the sib-sib correlation is the marginal correlation (ρ_s) which is often known from previous twin studies. In contrast with the sib-sib correlation, the sib-parent correlation ($\rho_{sp} = \rho_s$) and the parent-parent correlations ($\rho_p = c^2$) do not depend on the linkage parameter. This is due to the fact that parent-sibling pairs always share one allele IBD and the two parents do not share any alleles IBD. The likelihood function $P(Y|\pi;\gamma)$ for the gamma frailty model and for the log-normal frailty model are derived in appendix A and B, respectively.

Score tests

To test the null hypothesis $H_0 : \gamma = \gamma_0 = 0$ versus $H_1 : \gamma > 0$ a score test is proposed. For a set of N independent pedigrees the score statistic based on the retrospective likelihood of the marker data given the phenotypes (see appendix C) is given by

$$\mathrm{NPL}_w = \frac{\sum_{i=1}^{N}(\hat{\pi}_i - E\hat{\pi}_i)w_i}{\sqrt{\sum_{i=1}^{N}\mathrm{var}_0(\hat{\pi}_i)w_i^2}}, \tag{3.3}$$

with weight function

$$w_i = \frac{\partial \log P(Y_i|\pi_i;\gamma_0)}{\partial\rho(\gamma\pi)}. \tag{3.4}$$

where $\log P(Y_i|\pi_i;\gamma)$ is the prospective log-likelihood. This score statistic is a weighted NPL statistic (Kruglyak et al., 1996) with known weight functions.

The weight derived from the gamma frailty model is a very complex function and the closed form is not reported. These weights depend on marginal frailty parameters ($\rho_s, \rho_p, \sigma_G^2$) and on the marginal cumulative hazard H (see appendix B for details). The statistic based on this model is denoted by NPL_G^p. When parental age at onset are not considered ($S_1 = S_2 = 1$ and $\delta_1 = \delta_2 = 0$) this statistic is equivalent to the statistic derived by Callegaro et al. (2009), here denoted by NPL_G.

In contrast, the weight derived from the log-normal frailty model assuming small random effects (NPL_N^p) is a simple formula (see appendix C for details). Let δ, Λ_0 and $M = (\delta - \Lambda_0)/\Lambda_0$ be the four-dimensional vectors of the event status, of the baseline cumulative hazards at the age at onset (age at censoring) and of the standardized martingale residuals, respectively. The weight derived from the log-normal frailty model is given by the (3,4)-*th* element of the follow-

ing matrix

$$W_N = C^{-1}M(C^{-1}M)' - C^{-1} \tag{3.5}$$

where $C = D\Sigma D + D$ and $D = diag(\Lambda_0)$. The weight depends on the frailty parameters $(\rho_s, \rho_p, \sigma_N^2)$ and on the baseline cumulative hazard which is unknown. The estimation of this hazard function is not straightforward when the families have been selected on the phenotypes. In this paper, we estimate the baseline cumulative hazard by the marginal hazard ($\hat{\Lambda}_0 = H$) (Wintrebert et al., 2006). This estimator is valid for small residual effects. However, NPL_N^p gives actual type I error rates equal to the nominal value irrespectively of the validity of the baseline estimator.

Note that the classical NPL test (Blackwelder and Elston, 1985) corresponds to the score statistic (**??**) with weight equal to $w_i = \delta_{i3} \times \delta_{i4}$. In the case of incomplete marker information the variance of the estimated proportion of IBD given the marker data ($\text{var}_0(\hat{\pi})$) can be estimated by simulations (Lebrec et al., 2004).

3.3 Results

Weight functions

Weight function for common and rare disease models

The weight functions gave us the opportunity to describe the relationship between the phenotypes of the nuclear family and the information on linkage. We studied the weight function of the gamma and of the log normally distributed random effects for a rare and a common disease. For the rare disease we used a constant marginal hazard rate of 0.001. The corresponding marginal survival function was equal to 94% and 92% at age 60 and 80 respectively. For the common disease a constant marginal hazard rate of 0.01 with marginal survival function equal to 55% and 44% at age 60 and 80 respectively was used. The sib-sib and parent-offspring correlations were fixed at 0.5 and a parent-parent correlation of 0.1. Similar results were obtained using different values of these parameters. For the gamma distributed frailties, we used σ_G^2 equal to 1 and to 5. For the log normally distributed random effects we used $\sigma_N^2 = \log(1 + \sigma_G^2)$ to have similar scales for the gamma and normal distributions. For illustration purpose, we used the same age at onset for the siblings. Firstly we considered the situation that both parents had the same age at onset. Secondly we considered the situation that the parents had discordant phenotypes, i.e. one has early onset and the other has late onset disease.

Figure 3.1 shows the gamma frailty weight distribution in the case of rare and common disease for σ_G^2 equal to 1 and to 5. When the disease is rare and the variance of the random effect is small parents' age at onsets provide little information (Figure 3.1(a)). On the other hand, when the variance increases the weight distribution depends on the age at onset of the parents. Note that the most informative families have discordant siblings-parents phenotypes. In three of the four configurations (Figures 3.1(a), 1(b) and 1(c)) the most informative families have early onset siblings and late-onset parents. However, if the variance is not small and the trait is common the most informative families can be the families with late-onset siblings and early-onset parents (Figure 3.1(d)).

The weights corresponding to the log normal distribution are depicted in figure 3.2. For small variance, similar results were obtained. For large values of the variance the two approaches gave different results. This is due to the fact that the weights of the normal distributed frailty are accurate only for small values of the variance.

Results may change when parents have different age at onsets. If late onset siblings have been selected, concordant early onset parents are more informative (data not shown). If early onset siblings have been selected, families with discordant parents appear to be more informative than families with concordant late-onset parents (Figure 3.3 (a)). If discordant siblings have been selected, the most informative families for linkage are families with early-onset parents (Figure 3.3 (b)). Note that in this case the weights are always negative because discordant siblings are expected to share less alleles IBD than one.

Weight function for breast cancer

Wienke et al. (2003) modelled the age at onset of breast cancer in Swedish twins data. Using a correlated frailty model with Gompertz distributed baseline hazard ($\lambda_0(t) = be^{\varsigma t}$) they estimated the following parameters $\hat{b} = 1/10^5$, $\hat{\varsigma} = 0.120$, $\hat{\sigma}_G^2 = 25$, $\hat{\rho}_{DZ} = 0.125$ and $\hat{\rho}_{MZ} = 0.154$. From the marginal survival and the two correlation parameters ($\rho_p = 2\hat{\rho}_{DZ} - \hat{\rho}_{MZ} = 0.1$ and $\rho_s = \rho_{DZ} = 0.125$) we computed the weight function for various age at onset scenarios. The survival of the father was marginalized ($S(t) = 1$ and $\delta = 0$) and the two sisters had the same age at onset. Figure 3.4(a) shows the distribution of the gamma frailty weight function in terms of the age at onset of the two sisters and the age at onset of the affected mother. The most informative families have early onset siblings with late-onset mother. In this case the weight function of the NPL_N^p statistic differs from the weight function of the NPL_G^p, because the variance of the random effect is large. The weights of NPL_N^p are almost independent of the age at onset of the mother (data not shown).

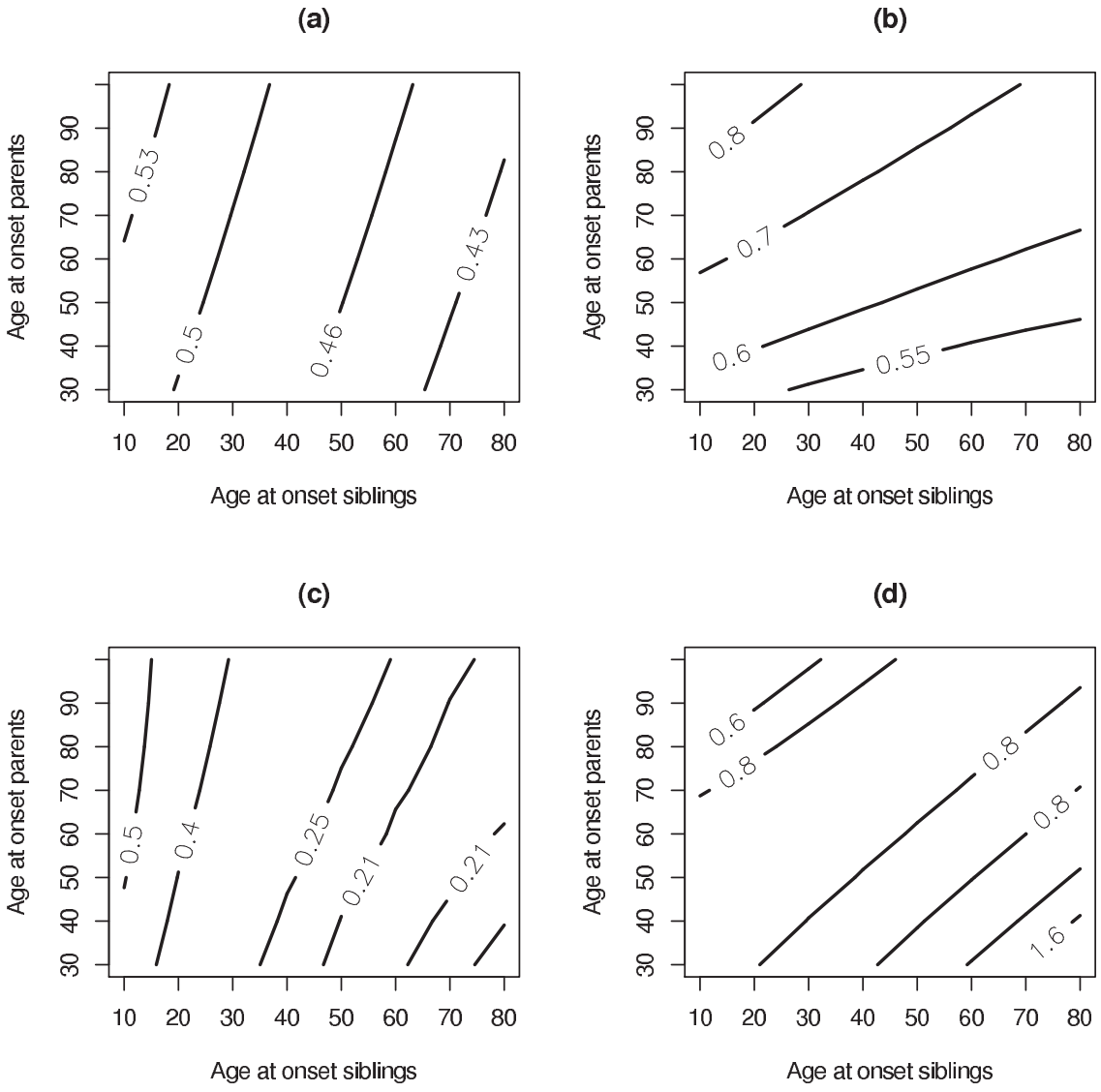

FIGURE 3.1: *Weights corresponding to the gamma frailty model as a function of the age at onset of the ASP and their affected parents. Fig. 1(a): rare disease and* $\sigma_G^2 = 1$*. Fig. 1(b): rare disease and* $\sigma_G^2 = 5$*. Fig. 1(c): common disease and* $\sigma_G^2 = 1$*. Fig. 1(d): common disease and* $\sigma_G^2 = 5$*.*

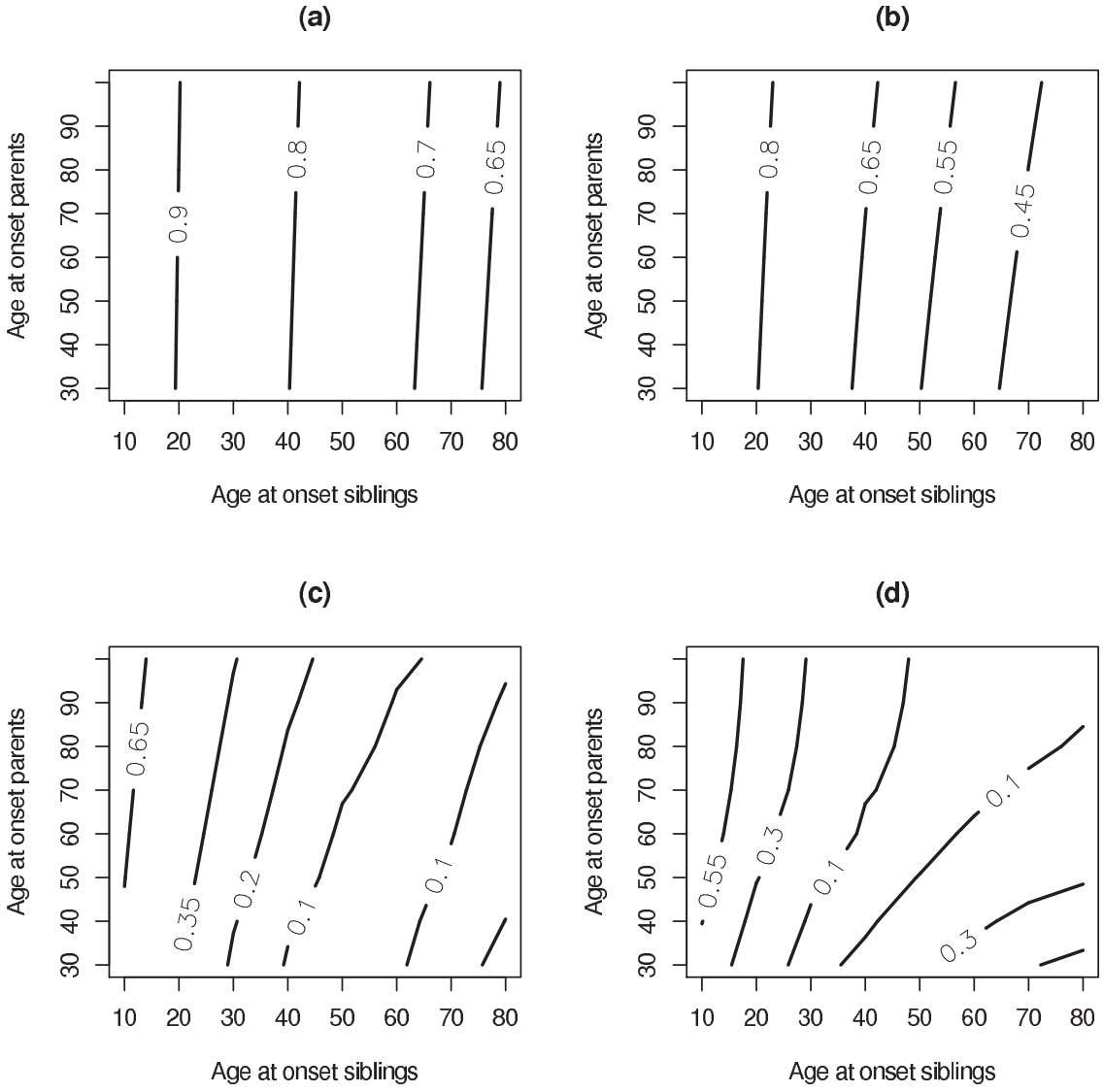

FIGURE 3.2: *Weights corresponding to the log-normal model as a function of the age at onset of the ASP and their parents. Fig. 2(a): rare disease and* $\sigma_N^2 = \log(1+1)$. *Fig. 2(b): rare disease and* $\sigma_N^2 = \log(1+5)$. *Fig. 2(c): common disease and* $\sigma_N^2 = \log(1+1)$. *Fig. 2(d): common disease and* $\sigma_N^2 = \log(1+5)$.

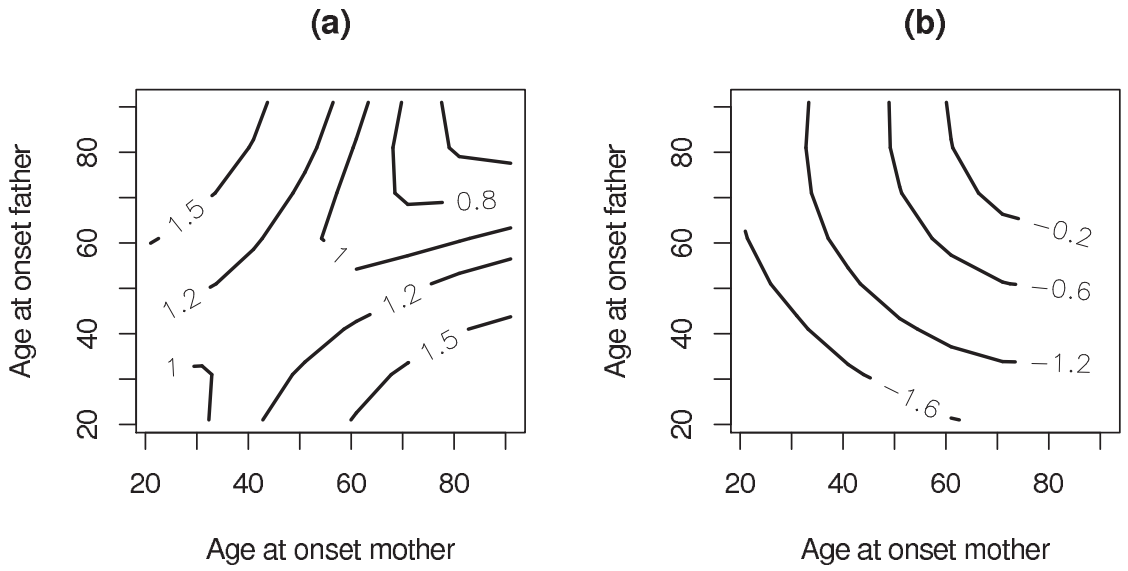

FIGURE 3.3: *Weights corresponding to the gamma frailty model as functions of the age at onset of the two parents for a common disease with $\sigma_G^2 = 5$. Fig. 3.3(a): early onset ASP. Fig. 3(b): discordant sibling pairs.*

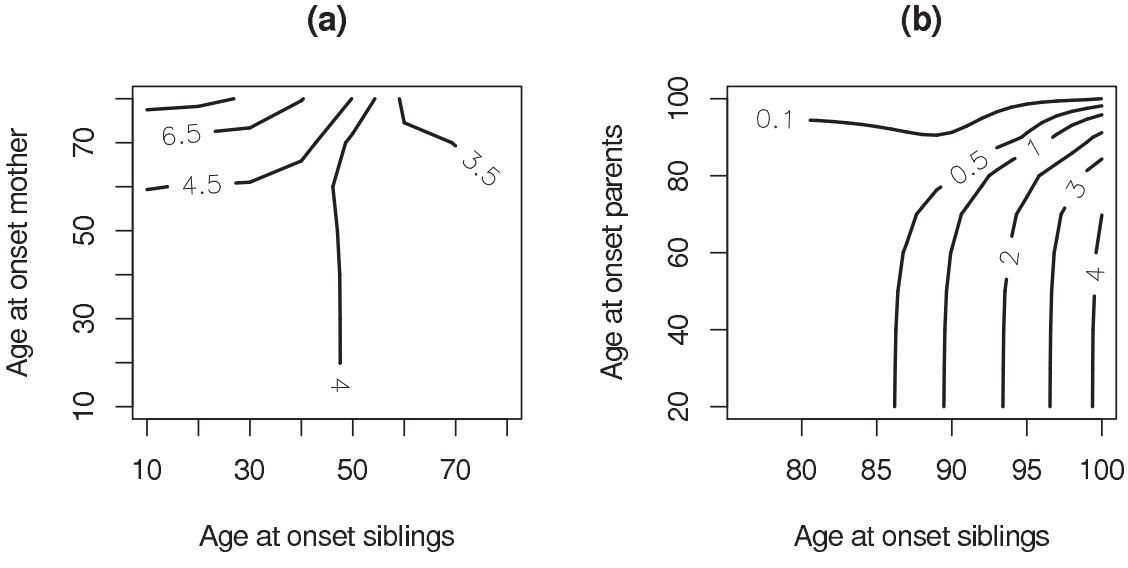

FIGURE 3.4: *Weights corresponding to realistic models for breast cancer (Fig. 4(a)) and for human life span (Fig. 4(b)).*

Weight function for human life span

In order to study correlation of life span in twins Yashin et al. (1999) fitted a gamma frailty model to Danish female twins born 1870-1900. They obtained the following parameters $\hat{\rho}_{DZ} = 0.31$, $\hat{\rho}_{MZ} = 0.51$ and $\hat{\sigma}^2_G = 1.23$. The correlation between parents is equal to $\hat{\rho}_{pp} = 2\hat{\rho}_{DZ} - \hat{\rho}_{MZ} = 0.11$. We used these parameters values together with the Dutch marginal survival in order to study the weights for life span.

Figure 3.4(b) shows the distribution of the gamma frailty weight function with respect to the age of death of two siblings, and the age of death of the parents. For simplicity, we assumed that the two siblings and the parents have respectively the same age at death. The maximum weight is obtained for long-lived siblings with early deceased parents. The weight function appears to decrease for parents older than 70 year. Such effect is less strong for discordant parents. In fact, the combination of long lived siblings and one early deceased parent is informative for linkage even if the other parent became long lived (data not shown). The weight function of the NPL^p_N is very similar to the weight function of the NPL^p_G, because the variance is rather small (data not shown).

Simulation results

By means of simulations of ASP data we compared the power of the method proposed by Callegaro et al. (2009) (NPL_G) which ignores the parental age at onset to the power of the two new methods proposed in this paper, NPL^p_G and NPL^p_N. Age at onset data were generated based on an additive gamma frailty model with a Gompertz distributed baseline hazard ($\lambda_0(t) = be^{\varsigma t}$). Current ages were simulated as age at censoring from a uniform distribution $U(40, 100)$. We studied common diseases where almost 10% and 40% of the population is affected by age 60 and 80, respectively. We simulated data with two different genetic models with variances of the random effect given by $\sigma^2 = 3.3$ ($b = 1 \times 10^{-5}$, $\varsigma = 0.12$) and $\sigma^2 = 1.1$ ($b = 1 \times 9^{-5}$, $\varsigma = 0.1$), respectively. The random effect was decomposed into the sum of the linkage effect and the residual shared environmental effect. For each of the two genetic models we simulated data according to three decompositions of the total random effect. In the first, the linkage effect explained all the variability of the random effects ($\gamma = 1$). In the second, the linkage effect explained two third of the total variance ($\gamma = 2/3$). Finally, in the third decomposition of the random effect, the linkage effect explained one third of the total variability ($\gamma = 1/3$). Each configuration included 5000 replications under the null hypothesis and 1000 replications under the alternative hypothesis. For each replication the first 200 nuclear families with ASPs were ascertained.

We evaluated the performance of the tests with a significance level of $\alpha = 0.05$. Under the null hypothesis all the tests have correct type one error (Table 3.1). Under the alternative hypothesis there is an increase in power incorporating the parental age at onset in the weights (Table 3.2). In the case of large variance of the random effect and small heritability, the gain in power of NPL_G^p with respect to NPL_G is about 60%. Note that NPL_G^p outperforms NPL_N^p in the case of large variance σ^2 because NPL_N^p assumes a small variance of the frailty.

TABLE 3.1: *Type one error comparisons based on 5000 replications of ASP with known parental age at onset (significance level $\alpha = 0.05$). No linkage effect ($\gamma = 0$).*

	$\sigma^2 = 1.1$			$\sigma^2 = 3.3$		
Method	$a^2 = 1$	$a^2 = 2/3$	$a^2 = 1/3$	$a^2 = 1$	$a^2 = 2/3$	$a^2 = 1/3$
NPL	0.051	0.050	0.051	0.049	0.049	0.051
NPL_G	0.051	0.048	0.048	0.049	0.051	0.048
NPL_G^p	0.048	0.050	0.050	0.049	0.049	0.049
NPL_N^p	0.051	0.047	0.051	0.047	0.049	0.048

TABLE 3.2: *Power comparison based on 1000 replications of ASP with known parental age at onset (significance level $\alpha = 0.05$). No residual additive effect ($\gamma = a^2$).*

	$\sigma^2 = 1.1$			$\sigma^2 = 3.3$		
Method	$a^2 = 1$	$a^2 = 2/3$	$a^2 = 1/3$	$a^2 = 1$	$a^2 = 2/3$	$a^2 = 1/3$
NPL	0.52	0.31	0.18	0.87	0.58	0.26
NPL_G	0.68	0.43	0.20	0.98	0.78	0.34
NPL_G^p	0.68	0.46	0.23	0.99	0.88	0.54
NPL_N^p	0.67	0.44	0.21	0.98	0.79	0.36

We further evaluated the robustness of the proposed methods. Figure 3.5 shows the power of the proposed score tests in the case of data simulated with $\sigma_G^2 = 3.3$ and $\gamma = 1/3$. The power is shown for different values of the frailty parameters specified in the weight $(\sigma^2, \rho_s, \rho_p)$. We computed the power as a function of σ^2 for three different values of the heritability $a^2 = 0.25, 0.5, 0.75$. For simplicity, we defined $c^2 = 1 - a^2$; $\rho_s = 1/2a^2 + c^2$ and $\rho_p = c^2$. The horizontal lines represents the unweighted NPL (solid line) which is independent of the frailty parameters. The proposed methods have maximal power around the true population parameter values. However, they outperform the classical NPL methods for a wide range of the frailty parameters used in the weight. For example, NPL_G^p (Figure 3.5(a)) is more powerful than the classical NPL for ev-

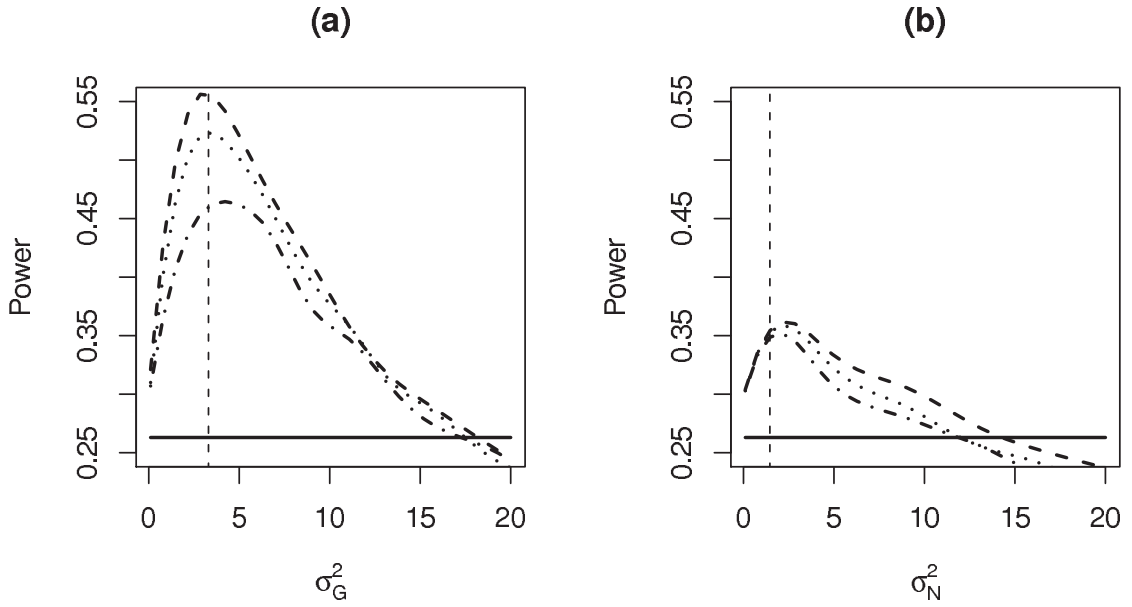

FIGURE 3.5: *Power of the test statistics based on simulated data (significance level $\alpha = 0.05$). The horizontal lines represent the power of the unweighted NPL (solid line). Dashed lines, dotted lines, dashed-dotted lines represent the power of proposed NPL methods assuming $a^2 = 0.25, 0.5$ and 0.75, respectively. Figure on the left (a): NPL^p_G. Figure on the right (b): NPL^p_N.*

ery value of σ^2_Z smaller than 15. In a similar way, NPL^p_N is more powerful than the unweighted method for every value of σ^2_N smaller than 10, irrespectively of the correlation values specified in the weight (Figure 3.5(b)). In this particular case NPL^p_G is more powerful than NPL^p_N because the data have been simulated with a large variance of the random effect. Note that when the value of the variance specified in the weight is equal to zero ($\sigma^2 = 0$) the two proposed methods have the same power, and they are more powerful than the unweighted NPL method.

Application to GAW12 data

We applied the proposed methods to the GAW12 simulated data of general population (Almasy et al., 2001). Disease data were generated using a complex model where seven genes influenced the liability and the age at onset. Major gene 7 directly contributed to age at onset and major gene 6 contributed to the disease liability. Both genes were located on chromosome 6, respectively at position 31.5 cM and 30.5 cM. More details on how the data were generated can be found in Almasy et al. (2001). The GAW12 general population data include a total of 50 replicates, each of them containing 23 extended pedigrees. From the first 30 replications we selected 500 independent ASPs with known age at onset (age at censoring) of the parents. The marginal survival functions and the frailty

parameters were estimated on the remaining 20 data-sets. For the marginal survival, we used a Kaplan-Meier estimator stratified by sex and the frailty parameters by a correlated gamma frailty model ($\rho_s = 0.7, \rho_p = 0.1, \sigma_G^2 = 3.75$). The estimated prevalence of the disease at 70 years is 50% and 25% for females and males, respectively. We computed the weights for the NPL statistics. For the weights of the NPL_N^p we used the variance equal to $\sigma_N^2 = \log(1 + \sigma_G^2) = 1.55$. The weight values of NPL_G^p and NPL_N^p were quite similar for this data set. As expected, families with early-onset siblings and late-onset parents had high values of the weight. However, since we used marginal survival stratified by sex, the values of the weight depend on the sex configuration of the siblings.

Parental genotypes were not considered in the analysis and the variances of the IBD ($\text{var}_0(\pi)$) were estimated by multipoint simulations using MERLIN (Abecasis et al., 2002; Callegaro et al., 2009). Without taking into account the age at onset, the most significant evidence for linkage was on chromosome 6, with a NPL LOD score of about 4.5 at the locus of the major gene 7. Further, we applied the score test of Callegaro et al. (2009) (NPL_G) ignoring the parental ages (LOD=5.5), and finally we applied the two score tests taking into account the parental age at onset. Figure 3.6 shows that -at the disease locus- NPL_N^p (LOD=5.6) is slightly more powerful than NPL_G^p (LOD=5.5). In this data set adjusting for the parental age at onset only slightly changed the results.

3.4 Discussion

In order to map disease genes involved in complex traits one may want to use all available information. A common strategy for linkage is to collect large samples of ASP. In this paper, we extended the score test of Callegaro et al. (2009) to include the parental ages at onset or current ages. When frailty parameters are known, the score statistic is a classical NPL statistic (Kruglyak et al., 1996), with known weights depending on the siblings and on the parental ages at onset. We also derived a score test from a log normal frailty model. Assuming small random effects, the weight function derived from this model is very simple and it can easily be computed for general sibship sizes.

The weight functions gave us the opportunity to study the complex relationship between age at onset and linkage effect. We explored the weight function for rare and common disease models. In line with previous papers on ASP (Callegaro et al., 2009; Li and Zhong, 2002) our results show that age at onset is not informative for rare diseases. In the case of common diseases, discordant parent-sibling families appear to be the most informative for linkage. A similar result was obtained for quantitative traits (Lebrec et al., 2004). The study of the weight distribution for breast cancer suggests that families with early onset

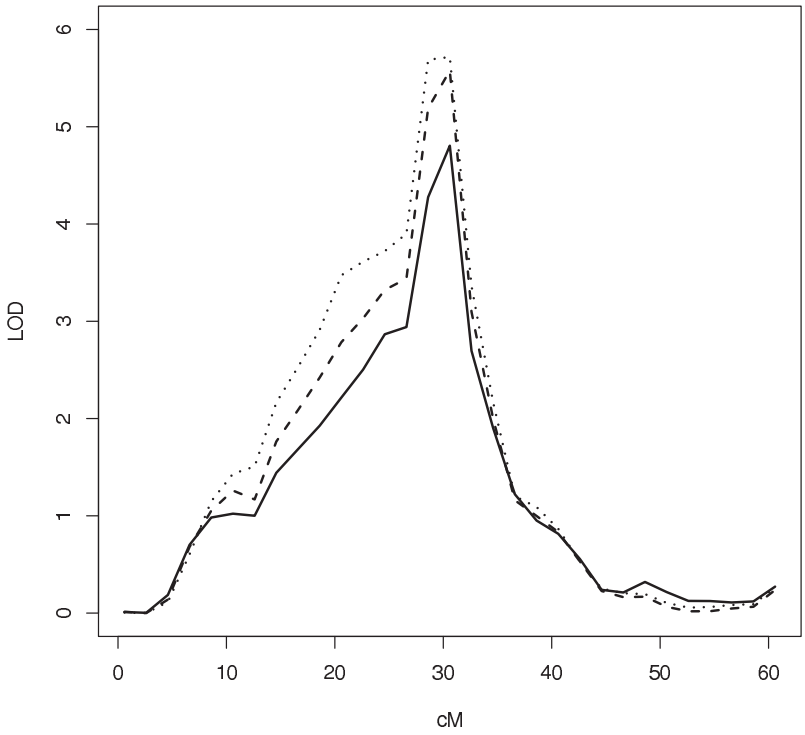

FIGURE 3.6: *Genetic Linkage analysis of the GAW12 data on chromosome 6. Straight lines represents NPL method, dashed line represents NPL_G^p method, and pointed line represents NPL_N^p method.*

sister pairs and late onset (unaffected) mothers are the most informative families to detect linkage. From the model for human life span our results show that long lived sibling pairs with early deceased parents are the most informative. In fact, selecting ASP with affected parents might increase the number of copies of any given disease-related variant in families. Families carrying multiple copies of a disease-related allele may not show linkage since the affected individuals may inherit that allele from different ancestors and therefore not deviate from expected IBD sharing (Wallace and Clayton, 2006).

Simulation results showed that adjusting for the parental age at onset can considerably increase the power to detect linkage. In fact, for common disease with small heritability, the gain in power of NPL_G^p compared to NPL_G was about 60%. We applied the proposed methods to GAW12 simulated data. The NPL scores taking into account the parental age of onset outperformed the NPL method which ignores the age at onset. However, the performance of the proposed methods was similar to the performance of NPL_G, which ignores the parental data. This result can be interpreted in terms of the number of families with discordant age at onset in siblings and parents. In fact, only few families were available with early-onset siblings and late-onset parents. Suppose for ex-

ample that early onset is defined as disease with age of onset before 30 (prevalence of 93%) and late onset is defined as disease with age of onset/censoring after 60 (prevalence of 50%). In this case only 10 families out of 500 were available with early-onset siblings and late-onset parents.

The proposed methods can only be applied when population parameters are known. However, the values of these parameters are available for many phenotypes, such as aging (Yashin et al., 1995), breast cancer (Wienke et al., 2003), coronary heart disease (Zdravkovic et al., 2002), malignant melanoma, colon cancer, (Zahl, 1997) etc. Further, we showed the robustness of these methods against misspecification of the frailty parameters. Simulation results showed that the proposed methods are not very sensitive for the frailty parameter values specified in the weight. The derived score tests have several advantages with respect to likelihood-ratio approaches (Jonker et al., 2009; Li, 2002; Li and Zhong, 2002; Pankratz et al., 2005). First of all, they maintain correct type I error irrespectively of the weights. Second, they can easily be applied to selected samples. Finally, they are simpler and computationally faster because they are computed under the null hypothesis.

It is interesting to compare the two score tests proposed. The additive gamma frailty model is mathematically appealing because the random effects can be integrated out, which allows us to work with observable marginal survival functions. However, results obtained from these additive models should be interpreted carefully (Petersen et al., 1996) . It may be more intuitive if the variance components act multiplicatively, like unobserved covariates in a Cox model (Cox, 1972). Another problem is that the likelihood function becomes too complicated to write down for general pedigrees. In contrast, the log normal model is a multiplicative model. Hence the parameters may be better interpretable. The formula of the weight is simple, and it can easily be computed for sibships of any size. The main limitation of this model is that it is powerful only for small values of the random effects. In this paper, we estimated the baseline hazard with the marginal hazard. A possibility to improve the power of the test based on the log-normal weight is to derive the baseline hazard using numerical methods.

An interesting extension of the proposed models is to include a cure fraction, like in the paper of **?**. If the cure effect is estimable, these extended models may better explain the data than standard survival methods.

In summary, we derived two different weights for the NPL statistic to take into account the age at onset or current age of two selected siblings and their parents. When the variance of the random effect is large ($\sigma_G^2 > 3$) we recommend to use the weight based on the gamma frailty model. When the variance of the random effect is small the two proposed methods have similar perfor-

mance. However, in this case we recommend to use the log-normal weight because the formula is much simpler and it can be computed for sibships of arbitrary sizes.

The relation between the linkage effect, the age at onset of the siblings and their parent is complex and depends on many parameters. This paper is an attempt to understand this relationship and to use it to increase the power to detect linkage. Software to apply the described methods are freely available from our website (`http://www.msbi.nl/genetics`).

Appendix

A: Additive gamma frailty model

Let the random effect Z in model (3.1) be the sum of four independently gamma distributed random effects ($Z = Z_g + Z_p + U_c + U_e$). Here $Z_{g,ik} \sim \Gamma(\nu_g, 1/\sigma^2)$ represents the linkage effect, $Z_{p,ik} \sim \Gamma(\nu_p, 1/\sigma^2)$ the residual additive effect, $U_{c,i} \sim \Gamma(\nu_c, 1/\sigma^2)$ the common environment effect and $U_{e,ik} \sim \Gamma(\nu_e, 1/\sigma^2)$ the individual effect. In order to obtain $E(Z) = 1$ we assume $\sigma^2 = 1/(\nu_g + \nu_c + \nu_p + \nu_e)$. Following the work of Commenges (1994) the linkage effect is modelled by the sum of two effects which represent the alleles inherited from the parents. Using this model the variance-covariance matrix of the random effects is equal to the classical variance-covariance matrix for quantitative traits (3.2) where the linkage effect is given by $\gamma = \nu_g/(\nu_g + \nu_c + \nu_p + \nu_e)$.

From this model the following four-dimensional marginal survival function can be derived (see below for details)

$$
\begin{aligned}
&S_{1234}(\pi;\gamma) = \\
&K_{1234}^{-\frac{\rho_p}{\sigma^2}} \prod_{j=1}^{4} K_j^{-\frac{1-2\rho_s+\rho_p}{\sigma^2}} \prod_{i=1}^{2} [K_{i34} K_{i3} K_{i4} K_i]^{-\frac{1}{2}(1-\tau)\frac{\rho_s-\rho_p}{\sigma^2}} \times \\
&[K_{13} K_{14} K_{23} K_{24}]^{-\frac{1}{2}\frac{\rho_p-\rho_s(\pi)+(1+\tau)(\rho_s-\rho_p)}{\sigma^2}} \times \\
&[K_{134} K_{234} K_1 K_2]^{-\frac{1}{2}\frac{\rho_s(\pi)-\rho_p-(1-\tau)(\rho_s-\rho_p)}{\sigma^2}}.
\end{aligned} \tag{3.6}
$$

where $K_B = \sum_{j\in B} \exp(\sigma^2 H_j) - b + 1$ with $H_j = \int_0^{t_j} h(u)du$ the marginal cumulative hazard of the jth subject, B a subset of the indexes $\{1,2,3,4\}$, the parameter b is the number of elements in B. The prospective likelihood of the phenotypes (Y) conditional on the proportion of alleles shared IBD (π) is a function of (3.6). In fact, the joint survival and density function for a family

with $d = \sum_{j=1}^{4} \delta_j$ affected and $4 - d$ unaffected relatives is

$$P(Y|\pi;\gamma) = (-1)^d \frac{\partial^d S_{1234}(\pi;\gamma)}{\partial^{\delta_1} t_1 \partial^{\delta_2} t_2 \partial^{\delta_3} t_3 \partial^{\delta_4} t_4},$$

where $\partial^{\delta_j} t_j = \partial t_j$ if $\delta_j = 1$ and $\partial^{\delta_j} t_j = 1$ if $\delta_j = 0$. The prospective likelihood corresponds to the four dimensional survival function $S_{1234}(\pi;\gamma)$ when $d = 0$. Note that the frailty parameters $(\rho_s, \rho_p, \sigma^2)$ can be obtained from twin studies and that under the null hypothesis $\tau = \frac{\nu_g}{\nu_g + \nu_p} = 0$ and $\rho_s(\pi) = \rho_s$.

Semiparametric four-dimensional survival function

Arbitrarily label the paternal chromosomes as (1,2) and the maternal chromosomes as (3,4). The inheritance vector of a sib-pair is the vector

$$V_d = (v_1, v_2, v_3, v_4)$$

where $v_{2j-1} = 1$ or 2, $v_{2j} = 3$ or 4, for $j = 1, 2$. The inheritance vector indicates which parts of the genome are transmitted to the children from the father and the mother. The linkage effect is modeled by the sum of two allelic effects ($U_g \sim \Gamma(\nu_g/2, 1/\sigma^2)$). The linkage effect of the father and of the mother are equal to $Z_{g,1} = U_{g,1} + U_{g,2}$ and to $Z_{g,2} = U_{g,3} + U_{g,4}$, respectively. The linkage effect of the jth sibling, $j = 3, 4$ is equal to $Z_{g,j} = \sum_{k=1}^{4} U_{g,k} a_{kj}$, where $a_{kj} = 1$ if $v_{2(j-2)-1} = k$ or $v_{2(j-2)} = k$, and 0 otherwise. The residual additive random effect is modelled as the sum of four random effects ($U_p \sim \Gamma(\nu_p/4, 1/\sigma^2)$). The residual additive random effect of the father and of the mother are equal to $Z_{p,1} = U_{p,1} + U_{p,2} + U_{p,3} + U_{p,4}$ and $Z_{p,2} = U_{p,5} + U_{p,6} + U_{p,7} + U_{p,8}$ respectively. The residual additive effects of the two siblings are given by $Z_{p,3} = U_{p,1} + U_{p,2} + U_{p,5} + U_{p,6}$ and $Z_{p,4} = U_{p,1} + U_{p,3} + U_{p,5} + U_{p,7}$, respectively.

The total random effect of the jth individual, $j = 1, ..., 4$ is given by the $Z_j = Z_{g,j} + Z_{p,j} + Z_c + Z_{e,j}$ where $Z_c \sim \Gamma(\nu_c, 1/\sigma^2)$ represents the common environmental effect and $Z_{e,j} \sim \Gamma(\nu_c, 1/\sigma^2)$ represents the non-shared environmental effect.

Then the marginal bivariate survival of two siblings is given by

$$
\begin{aligned}
&S_{1234} = E[S_1^{Z_1} S_2^{Z_2} S_3^{Z_3} S_4^{Z_4}] = \\
&E[\exp(-U_c \sum_{j=1}^{4} \Lambda_j) \prod_{j=1}^{4} \exp(-U_{e,j}\Lambda_j) \\
&\exp(-U_{p,1}(\Lambda_1 + \Lambda_3 + \Lambda_4)) \exp(-U_{p,2}(\Lambda_2 + \Lambda_3 + \Lambda_4)) \\
&\exp(-U_{p,3}(\Lambda_1 + \Lambda_3)) \exp(-U_{p,4}(\Lambda_2 + \Lambda_3)) \\
&\exp(-U_{p,5}(\Lambda_1 + \Lambda_4)) \exp(-U_{p,6}(\Lambda_2 + \Lambda_4)) \\
&\exp(-U_{p,7}\Lambda_1) \exp(-U_{p,8}\Lambda_2) \exp(-U_{g,1}(\Lambda_1 + a_{13}\Lambda_3 + a_{14}\Lambda_4)) \\
&\exp(-U_{g,2}(\Lambda_1 + a_{23}\Lambda_3 + a_{24}\Lambda_4)) \exp(-U_{g,3}(\Lambda_2 + a_{33}\Lambda_3 + a_{34}\Lambda_4)) \\
&\exp(-U_{g,4}(\Lambda_2 + a_{43}\Lambda_3 + a_{44}\Lambda_4))],
\end{aligned}
$$

$$
\begin{aligned}
&S_{1234} = (1 + \sigma^2 \sum_{j=1}^{4} \Lambda_j)^{-\nu_c} [\prod_{j=1}^{4} (1 + \sigma^2 \Lambda_j)]^{-\nu_e} \\
&[\prod_{i=1}^{2} (1 + \sigma^2(\Lambda_i + \Lambda_3 + \Lambda_4))(1 + \sigma^2(\Lambda_i + \Lambda_3)) \\
&(1 + \sigma^2(\Lambda_i + \Lambda_4))(1 + \sigma^2(\Lambda_i))]^{-\frac{\nu_p}{4}} \\
&[(1 + \sigma^2(\Lambda_1 + a_{13}\Lambda_3 + a_{14}\Lambda_4))(1 + \sigma^2(\Lambda_1 + a_{23}\Lambda_3 + a_{24}\Lambda_4)) \\
&(1 + \sigma^2(\Lambda_2 + a_{33}\Lambda_3 + a_{34}\Lambda_4))(1 + \sigma^2(\Lambda_2 + a_{43}\Lambda_3 + a_{44}\Lambda_4))]^{-\frac{\nu_g}{2}}.
\end{aligned}
$$

In the following we apply the transformation $\Lambda_j = (S_j^{-\sigma^2} - 1)/\sigma^2$ where S_j is the marginal survival of the jth subject. If we define $K_B = \sum_{j \in B} \exp(\sigma^2 H_j) - b + 1$, B is a subset of the elements $(1, 2, 3, 4)$ and b is the number of elements in B then the survival functions for 0 and 1 proportion of allele shared IBD are given by

$$
\begin{aligned}
S_{1234|\pi=0} &= K_{1234}^{-\nu_c} \prod_{j=1}^{4} K_j^{-\nu_e} \prod_{i=1}^{2} [K_{i34} K_{i3} K_{i4} K_i]^{-\nu_p/4} [K_{13} K_{14} K_{23} K_{24}]^{-\nu_g/2} \\
S_{1234|\pi=1} &= K_{1234}^{-\nu_c} \prod_{j=1}^{4} K_j^{-\nu_e} \prod_{i=1}^{2} [K_{i34} K_{i3} K_{i4} K_i]^{-\nu_p/4} [K_{134} K_{234} K_1 K_2]^{-\nu_g/2}.
\end{aligned}
$$

If the parental genotypes are unknown then the survival conditioned on half of

the alleles shared IBD is the mean of the two survival functions

$$S_{1234|\pi=0.5} = K_{1234}^{-\nu_c} \prod_{j=1}^{4} K_j^{-\nu_e} \prod_{i=1}^{2} [K_{i34} K_{i3} K_{i4} K_i]^{-\nu_p/4} \times \frac{[K_{134} K_{23} K_{24} K_1]^{-\nu_g/2} + [K_{234} K_{13} K_{14} K_2]^{-\nu_g/2}}{2}.$$

In order to simplify the formula, we approximate the arithmetic mean with a geometric mean

$$S_{1234|\pi=0.5} \approx K_{1234}^{-\nu_c} \prod_{j=1}^{4} K_j^{-\nu_e} \prod_{i=1}^{2} [K_{i34} K_{i3} K_{i4} K_i]^{-\nu_p/4-\nu_g/4}.$$

If follows that the approximated 4D survival function is given by

$$S_{1234} = K_{1234}^{-\nu_c} \prod_{j=1}^{4} K_j^{-\nu_e} \prod_{i=1}^{2} [K_{i34} K_{i3} K_{i4} K_i]^{-\frac{\nu_p}{4}} \times [K_{13} K_{14} K_{23} K_{24}]^{-\frac{\nu_g(1-\pi)}{2}} [K_{134} K_{234} K_1 K_2]^{-\frac{\nu_g \pi}{2}}.$$

B: Log-normal frailty model

Let δ, Λ_0 and $V = \log Z$ be the four-dimensional vectors of the status, of the baseline cumulative hazards at the age at onset (age at censoring) and the normally distributed random effects of the four pedigree members, respectively. The random effect V follows a multivariate normal distribution with mean zero and covariance matrix Σ (3.2). The log-likelihood can be approximated by using a second order Taylor approximation around $V = 0$. For small σ_V^2, it follows that

$$\log P(Y|\Lambda_0, V) \approx \sum_{i=1}^{4} c_i - \frac{1}{2} \Lambda_{0,i} \left(\frac{\delta_i - \Lambda_{0i}}{\Lambda_{0i}} - V_i\right)^2.$$

When the baseline cumulative hazard is known, the vector of standardized phenotypes behaves as a normal distribution (Wintrebert et al., 2006). Integrating over the distribution of the random effect gives

$$\frac{\delta - \Lambda_0}{\Lambda_0} \sim N(0, \Sigma + diag(1/\Lambda_0)),$$

and the vector of the baseline martingale residuals is distributed as

$$M = \delta - \Lambda_0 \sim N(0, D\Sigma D + D),$$

where $D = diag(\Lambda_0)$. It follows that for small σ_V^2, the four-dimensional likelihood is given by

$$P(Y|\pi;\gamma) \approx |C|^{-1/2} \exp(-\frac{1}{2} M' C^{-1} M), \tag{3.7}$$

where $C = D\Sigma D + D$.

C: Retrospective likelihood

Using Bayes rule the retrospective log-likelihood of the marker data (MD) given the phenotype Y (Li and Zhong, 2002; Whittemore, 1996) is given by

$$\begin{aligned}\ell(\gamma) &= \log P(\text{MD}|Y,\gamma) \\ &= \log[\sum_{\pi\in\{0,0.5,1\}} P(Y|\pi;\gamma) g_\pi] \\ &\quad - \log[\sum_{\pi\in\{0,0.5,1\}} P(Y|\pi;\gamma) p_\pi] + \log P_0(\text{MD}),\end{aligned}$$

where $g_\pi = P_0(\pi|\text{MD})$, $p_c = P_0(\pi)$, and P_0 is the probability assuming independent assortment of chromosomal regions to gametes.