Linkage mapping for complex traits : a regression-based approach

Lebrec, J.J.P.

Chapter 7

Score Test for Linkage in Generalized Linear Models

Abstract

We derive a test for linkage in a Generalized Linear Mixed Model (GLMM) framework which provides a natural adjustment for marginal covariate effects. The method boils down to the score test of a quasi-likelihood derived from the GLMM, it is computationally inexpensive and can be applied to arbitrary pedigrees. In particular, for binary traits, relative pairs of different nature (affected and discordant) and individuals with different covariate values can be naturally combined in a single test. The model introduced could explain a number of situations usually described as gene by covariate interaction phenomena, and offers substantial gains in efficiency compared to methods classically used in those instances.

7.1 Introduction

For binary traits, most linkage methods that allow for covariates focus on models where the identity-by-descent (IBD) probabilities are allowed to depend on those covariates (e.g. , Olson [1999]). This is often the most straightforward way to go because linkage studies for binary traits usually consist of families which have been selected based on their phenotypic values such as affected sib pairs (ASP) designs and effect of covariates at the population level cannot be estimated based on such data.

This chapter has been accepted for publication in *Human Heredity* as: J.J.P. Lebrec and H.C. van Houwelingen. Score Test for Linkage in Generalized Linear Models.

In many instances, however, some knowledge about the marginal effect of important covariates can often be gathered from either population-based studies or a literature review. Nevertheless, existing methods fail to integrate such external knowledge. An area where incorporation of covariates is a burning problem is late onset diseases, in fact, incorporation of population estimates of onset for the disease is not just a way to refine the analysis, it also allows inclusion of unaffected individuals. This can result in substantial gains in power, especially when traits are fairly common. In the case of continuous traits, the variance components model (and related regression methods) is widely accepted as the model of choice for testing for linkage with a putative locus. In this setting, the effect of important covariates is often modeled through a linear model while the covariance structure is left untouched. In contrast, the variance-covariance structure and the mean of binary and count data are intrinsically dependent and it is unclear how incorporation of covariates in the marginal probabilities impact linkage testing.

The Generalized Linear Mixed Models (GLMM) framework offers a natural and flexible extension of the variance components setting to categorical endpoints such as binary, count and survival data and accommodates covariate effects and arbitrary family structures. In accordance with the biometrical view of trait architecture [Fisher, 1918], small covariate effects contribute additively to the formation of a trait. Coupled with a variance components structure used to described the remaining correlation between relatives in a family, we obtain a parsimonious representation of the correlation between relatives. This unobserved latent process is linked to the actual trait values via a traditional Generalized Linear Model (see Section 7.2). In fact, this type of models have already been used for estimation of the heritability of binary traits [Burton et al., 1999; Houwing-Duistermaat et al., 2000; Noh et al., 2005] as well as for linkage of longitudinal continuous [Palmer et al., 2003] data and survival data [Scurrah et al., 2000]. Although appealing GLMMs are in general difficult to fit with family data. Besides we favor simple mathematically tractable expressions for a test, this is to reduce computational burden, but even more importantly, because we would like to get insight into the properties of this model when used in linkage studies. In stark contrast with the above cited approaches, we do not make any attempt to directly use the GLMM for inference but we resort to an approximation of the corresponding

likelihood (a quasi-likelihood). Indeed, our inference for linkage is based on a score test for the variance component corresponding to linkage in this quasi-likelihood (see Section 7.3). We assume that all segregation parameters in the GLMM have been obtained from external data and are therefore treated as nuisance parameters when testing for linkage. Estimation of such parameters in a GLMM is a notoriously difficult problem (at least for binary responses), we therefore propose an ad-hoc estimation procedure which appears to yield reasonable estimates in practice (see Section 7.4). Although the procedure does not always yield a unique set of parameters, we argue that our linkage test only weakly depends upon the parameters' choice and that its size is always preserved. The test is in fact a weighted regression of the deviation in IBD sharing on the trait values (in the same spirit as the pair-wise IBD scoring functions introduced by Whittemore and Halpern [1994] for affected relative pairs), which guarantees fast computations. Finally, in Section 7.5, we illustrate how the test could be used in linkage studies for two diseases: migraine and breast cancer. In those two examples we quantify the potential gains obtained compared to approaches that would either ignore covariates or estimate covariate effects from the linkage data only. In the discussion, we identify situations where covariate adjustment is likely to help improving the power of linkage studies.

7.2 Model

The generalized linear mixed model

Conditional on unobserved latent variables and observed covariate values, our model is specified by a generalized linear model (GLM). All information about the genetic relationship between individuals is incorporated in the latent variables just in the same way as in the variance components model for continuous traits. Formally, we consider the trait values $\mathbf{y} = (y_1, \ldots, y_m)$ of m relatives in a family whose values for k covariates are gathered in an $m \times k$ matrix $\mathbf{X}$. Conditional on a vector of random effects $\mathbf{b} = (b_1, \ldots, b_m)$ and a vector of covariate effects β, the y_i's are independently distributed according to a density function f from the canonical exponential family (to simplify notations, we have omitted the dispersion parameter), more precisely f

has the following form

$$\log f(y_i \,|\, \beta, b_i) = y_i \times (\mathbf{x}_i\beta + b_i) + a(y_i) - \psi(\mathbf{x}_i\beta + b_i)$$

where the first two derivatives of ψ determine the first and second moments of the GLM i.e. $\psi'(\mathbf{x}_i\beta + b_i) = \mathbf{E}(\mathbf{y}_i \,|\, \beta, b_i)$ and $\psi''(\mathbf{x}_i\beta + b_i) = \text{var}(\mathbf{y} \,|\, \beta, b_i)$. This type of models includes the logistic model for binary or binomial data, Poisson model for count data, continuous data (provided the dispersion parameter is known) as well as piecewise exponential hazards models for survival data [Agresti, 2002, pp.388-389]. The fixed effects β therefore model the effect of covariates while the dependence structure between relatives is entirely induced through the covariance of the random effects $\mathbf{b}$ which are assumed to follow a multivariate normal distribution with mean 0 and variance-covariance matrix $\mathbf{R}(\theta)$ where θ is the set of variance components. In the simple case of sibships the variance-covariance structure of $\mathbf{b}$ is described by a compound symmetry structure

$$\mathbf{R} = \mathbf{R}(\theta) = \sigma^2 \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \dots & \rho & 1 \end{pmatrix} .$$

The exact marginal density $l(\beta, \theta)$ of the observations $\mathbf{y}$ is obtained by integration of the random effects $l(\beta, \theta) = \mathbf{E}_{\mathbf{b}}(\prod_{i=1,\dots,m} f(y_i \,|\, \beta, b_i))$ which entails calculation of a multivariate integral of potentially high dimension (for extended families).

GLMM for linkage

Our primary interest is on testing for linkage and we will therefore assume that all nuisance parameters i.e. the fixed covariate effects β and the marginal part of the covariance structure $\mathbf{R}(\theta)$ are known. We delay resolution of this problem to Section 7.4. We denote by γ the proportion of the random effects total variance σ^2 explained by the putative locus and focus our attention on this parameter by partitioning the set of variance components as (θ, γ). In analogy with the variance components model for continuous traits, we model linkage by specifying the conditional covariance structure $\mathbf{R} = \mathbf{R}(\theta, \gamma)$ of the random effects $\mathbf{b}$ given IBD information $\boldsymbol{\pi}$ within each family.

The $m \times m$ matrix $\boldsymbol{\pi}$ contains the identity-by-descent (IBD) information at a putative chromosomal position, more precisely $[\boldsymbol{\pi}]_{jk} = \pi_{jk}$ is the proportion of alleles shared IBD by pedigree members j and k and

$$[\mathbf{R}]_{jk} = \begin{cases} a^2 + c^2 = \sigma^2 \ , & \text{if } j = k \ , \\ (\pi_{jk} - \mathbf{E}\pi_{jk})\gamma\sigma^2 + (\mathbf{E}\pi_{jk})a^2 + c^2 \ , & \text{if } j \neq k \ . \end{cases}$$

where a^2 denotes the total additive genetic variance and c^2, the common-environment variance, on the underlying random effect scale.

7.3 Test for linkage

Quasi-likelihood for variance components

In an appendix, we show how the following quasi-likelihood for the data $\mathbf{y}$ can be obtained

$$\mathbf{y} \sim N\left(\boldsymbol{\psi\prime}(\mathbf{X}\beta) \ , \ \boldsymbol{\Psi}''(\mathbf{X}\beta) + \boldsymbol{\Psi}''(\mathbf{X}\beta).\mathbf{R}(\theta,\gamma).\boldsymbol{\Psi}''(\mathbf{X}\beta) \right) \ , \tag{7.1}$$

where $\boldsymbol{\psi}'(\mathbf{X}\beta)$ denotes the vector whose i^{th} element is given by $\psi'(\mathbf{x}_i\beta)$ and $\boldsymbol{\Psi}''(\mathbf{X}\beta)$ denotes the diagonal matrix whose i^{th} diagonal element is given by $\psi''(\mathbf{x}_i\beta)$. Note that this is not a normal approximation of the marginal likelihood, the normal shape is naturally obtained via a 2^{nd} order Taylor approximation of an exponential family likelihood in the canonical form. This quasi-likelihood can also be motivated by an approximate marginal model of the GLMM as in [Breslow and Clayton, 1993] and is the basis of the marginal quasi-likelihood (MQL) fitting algorithm. Another less crude approximation of the marginal likelihood could be based on a 1^{st} order Laplace approximation however this would render the approach mathematically intractable. Quasi-likelihood (7.1) is only accurate for small values of the random effects, hence small values of their variance σ^2; nonetheless, however accurate this approximation, the approach that we propose in Section 7.3 provides an 'unbiased' testing strategy.

Score test

For mathematical convenience, we use the quasi-likelihood for variance components introduced in Section 7.3 but expressed in terms of the first-order maximum-likelihood

estimates $\mathbf{z} = \frac{\mathbf{y}-\psi'(\mathbf{X}\beta)}{\psi''(\mathbf{X}\beta)}$ of the random effects $\mathbf{b}$. Denoting $\mathbf{\Sigma} = \mathbf{R}(\theta,\gamma) + \mathbf{\Psi''}^{-1}(\mathbf{X}\beta)$, this quasi-likelihood writes

$$\log ql(\mathbf{z},\gamma \,|\, \boldsymbol{\pi}) = -\frac{m}{2}\log(2\pi) - \frac{1}{2}\log(|\mathbf{\Sigma}|) - \frac{1}{2}\mathbf{z}'\mathbf{\Sigma}^{-1}\mathbf{z} \ .$$

We show in an appendix that the score function ℓ_γ for γ can then be written as

$$\ell_\gamma = \frac{1}{2}\ \mathrm{vec}(\mathbf{C})' \ .\ \mathrm{vec}(\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi}) \tag{7.2}$$

with $\mathbf{C} = \mathbf{\Sigma}^{-1}\mathbf{z}\left(\mathbf{\Sigma}^{-1}\mathbf{z}\right)' - \mathbf{\Sigma}^{-1}$ and $\mathbf{\Sigma}$ taken in $\gamma = 0$. Here $\mathrm{vec}(\mathbf{C})$ places the n columns of the $m \times n$ matrix $\mathbf{C}$ into a vector of dimension $mn \times 1$, it contains weights for the pairwise IBD sharing $\mathrm{vec}(\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi})$. Note that the $\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi}$ matrix has all diagonal elements equal to 0. Our test for linkage is a weighted average of the different excess IBD sharing between all pairs of relatives in the pedigree. Linkage studies often include families which have been selected on the basis of their phenotypic values and it is sometimes unclear what the exact ascertainment scheme used is. A valid analysis of the data therefore requires that inference be carried out conditional on observed phenotypic values. Given the parametrization used above, accepting the quasi-likelihood $ql = ql(\mathbf{z} \,|\, \boldsymbol{\pi}, \gamma)$ as the model generating the "phenotypic data" $\mathbf{z}$ and relying on known nuisance parameters (β and θ), it turns out that the score function $\frac{\partial \log \mathbf{P}(\boldsymbol{\pi} \,|\, \mathbf{z},\gamma)}{\partial\gamma}$ evaluated at $\gamma = 0$ of the corresponding inverse likelihood of IBD sharing $\boldsymbol{\pi}$ conditional on transformed trait values $\mathbf{z}$ is simply equal to the same ℓ_γ function (see [Lebrec et al., 2004] for a proof). This justifies the use of this score statistic in selected samples. When the likelihood conditional on trait values is considered, the corresponding Fisher's information $\mathcal{I}_\gamma = \mathbf{E}\left(-\frac{\partial^2}{\partial\gamma^2}\log \mathbf{P}_\gamma(\boldsymbol{\pi} \,|\, \mathbf{z}, \gamma = 0)\right)$ for γ is also the variance of the score function $\mathrm{var}(\ell_\gamma \,|\, \mathbf{z}, \gamma = 0)$ and is thus given by

$$\mathcal{I}_\gamma = \frac{1}{4}\ \mathrm{vec}(\mathbf{C})' \ .\ \mathrm{var}\left(\mathrm{vec}(\boldsymbol{\pi}) \,|\, \gamma = 0\right) \ .\ \mathrm{vec}(\mathbf{C}) \ . \tag{7.3}$$

For a set of independent $p = 1, \ldots, P$ families with corresponding standardized trait values $\mathbf{z}_1, \ldots \mathbf{z}_P$, we therefore test for linkage using the statistic

$$T_+^2 = \begin{cases} 0 \ , & \text{if } \sum_{p=1}^P \ell_{\gamma,p} \leq 0 \\ \frac{\left(\sum_{p=1}^P \ell_{\gamma,p}\right)^2}{\sum_{p=1}^P \mathcal{I}_{\gamma,p}} \ , & \text{otherwise} \end{cases} \ ,$$

which is is asymptotically distributed as $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ under the null hypothesis (H_0) of no linkage. Indeed, the score conditional on trait values is unbiased since $\mathbf{E}(\ell_\gamma \,|\, \mathbf{z}, \gamma =$

$0) = 0$ (the term involving $\boldsymbol{\pi}$ in ℓ_γ is centered) and the standardization used (i.e. conditional on trait values $\mathbf{z}$) ensures that the test has variance 1 under H_0. Note that this would not necessarily be the case conditional on IBD sharing $\boldsymbol{\pi}$ (i.e. $\mathbf{E}(\ell_\gamma \mid \boldsymbol{\pi}, \gamma = 0) \neq 0$) because of model mis-specification.

Special case of relative pairs

Although the test derived previously applies to arbitrary pedigrees, the rest of the paper is devoted to relative pairs. In this instance, the variance-covariance matrix of random effects is

$$\mathbf{R} = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} ,$$

for example, in the case of sib pairs, $\sigma^2 = a^2 + c^2$ and $\rho\sigma^2 = \frac{1}{2}a^2 + c^2$. If we denote $\psi'_i = \psi'(\mathbf{x}_i\beta)$, $\psi''_i = \psi''(\mathbf{x}_i\beta)$ and $\nu_i = (\sigma^2\psi''_i)^{-1}$, the score can be written in terms of the unstandardized centered trait values (or raw residuals) $y_i - \psi'_i$ as

$$\begin{aligned} \ell_\gamma = (\pi - \mathbf{E}\pi)\times \quad & \nu_1\nu_2 \ \left\{(1+\nu_1)(1+\nu_2) - \rho^2\right\}^{-2} \\ & \times \big[\ \left\{(1+\nu_1)(1+\nu_2) + \rho^2\right\}(y_1 - \psi'_1)(y_2 - \psi'_2) \\ & -\rho(1+\nu_2)(y_1 - \psi'_1)^2 - \rho(1+\nu_1)(y_2 - \psi'_2)^2 \\ & +\rho(\sigma^2\nu_1\nu_2)^{-1}\left\{(1+\nu_1)(1+\nu_2) - \rho^2\right\} \ \big] \ . \end{aligned}$$

If we let both ν_1 and ν_2 tend to $+\infty$, then the excess IBD sharing $\pi - \mathbf{E}\pi$ is simply weighted by the product of the raw residuals $(y_1 - \psi'_1)(y_2 - \psi'_2)$. This means that in the context of rare diseases and affected pairs (thus $y_1 = y_2 = 1$), the effect of covariates has to be very large for the weights to substantially differ from an unweighted strategy. Letting both ν_1 and ν_2 tend to 0, the weight then becomes $(1+\rho^2)z_1z_2 - \rho(z_1^2 + z_2^2) + \rho\sigma^2(1-\rho^2)$, where the z_i's are the first-order maximum-likelihood estimates of the random effects b_i's defined in Section 7.3. This expression is closely related to a version of the so-called Haseman-Elston regressions that is optimal with normally distributed data [Sham and Purcell, 2001], the main difference lies in the use of the variances ψ''_i in the standardization of the centered trait values $y_i - \psi'_i$ instead of the usual ${\psi''_i}^{1/2}$ as in Pearson residuals.

It is interesting to look at the special case of binary traits, where $a \equiv 0$ and $\psi(t) = \log(1 + e^t)$. In this instance, the weights associated to excess IBD sharing

$\pi - \mathbf{E}\pi$ are positive for ASP and unaffected sib pairs (USP) while they are negative for discordant sib pairs (DSP). Based on approximation (7.4) used in Section 7.4, ν_1 can be shown to be approximately related to the marginal correlation via $\nu_1 \approx \rho \, \mathrm{cor}(y_1, y_2)^{-1} \, \psi_2''^{1/2} \, \psi_1''^{-1/2}$ as long as σ^2 is not too large. This provides us with an order of magnitude for the ν_i parameters. For example, if the covariate values are the same for both individuals, ν is simply proportional to the inverse of the trait marginal correlation, which itself is an increasing function of both the prevalence and the recurrence risk ratio $\lambda_S = \mathbf{P}(\text{sib 1 is affected and sib 2 is affected})/\mathbf{P}(\text{sib 1 is affected})\,\mathbf{P}(\text{sib 2 is affected})$. For rare diseases, the ν_i parameters will likely be very large and weights given to the excess IBD sharing will be approximately equal to $(y_1 - \psi_1')(y_2 - \psi_2') \approx (y_1 - \mathbf{E}y_1)(y_2 - \mathbf{E}y_2)$ as pointed out in the previous paragraph. In this rare disease case, a direct application of the optimal Haseman-Elston regression for normally distributed data [Sham and Purcell, 2001] would lead to a weighting scheme approximately equal to the product of the Pearson residuals $(y_1 - \mathbf{E}y_1)/(\mathbf{E}y_1(1 - \mathbf{E}y_1))^{1/2} \times (y_2 - \mathbf{E}y_2)/(\mathbf{E}y_2(1 - \mathbf{E}y_2))^{1/2}$. Since the denominators $(\mathbf{E}y_i(1 - \mathbf{E}y_i))^{1/2}$ change rapidly as the trait becomes rare, the weight given to rare phenotypic values will be too extreme compared to those given to common trait values.

7.4 Estimation of segregation parameters

Estimation in GLMM has been the subject of intense research in the past decade and has proved notoriously difficult. Direct computation of the marginal likelihood can in principle be carried out by quadrature methods but are computationally burdensome, for that reason, approximate methods such as penalized quasi-likelihood (PQL) [Breslow and Clayton, 1993] have been proposed, unfortunately they are known to yield severely biased estimates, especially with binary endpoints. Another route is Bayesian fitting via Markov chain Monte Carlo algorithms. We refer the reader to `www.mlwin.com` for a list and review of possible softwares. Practical solutions appear to be problem-specific and a few authors have dealt with this problem in the case of family data [Burton et al., 1999; Houwing-Duistermaat et al., 2000; Noh et al., 2005]. Besides, in some instances (e.g. , when sib-pair data only are available), the GLMM may lack identifiability. We therefore propose the approximate

method described in Section 7.4. There is an extra difficulty in the case of binary data and we propose an ad-hoc solution which appears to yield sensible guesses of the nuisance covariance parameters θ and fixed effects β as far as the interest lies in testing for linkage: although the procedure of Section 7.4 does not give a unique choice of parameters, we argue that the actual linkage test is fairly insensitive to that specification.

General case

We first consider the case of a homogeneous population (i.e. no covariates) where three nuisance parameters need to be estimated, namely, the fixed effect β that reflects the overall level for the trait of interest, the variance σ^2 of the underlying random effect and the correlation ρ between the random effects in a pair of relatives. The marginal covariance relates to $\rho\sigma^2$ through the following approximate relation

$$\text{cov}(Y_1, Y_2) \approx \psi_1''(\beta)\psi_2''(\beta)\rho\sigma^2 \ , \tag{7.4}$$

and the marginal variance to β and σ^2 via

$$\text{var}(Y) \approx \psi''(\beta) + \psi''(\beta)^2\,\sigma^2 \ , \tag{7.5}$$

while the marginal mean can be either approximated as

$$\mathbf{E}(Y) \approx \psi'(\beta) + \frac{\sigma^2}{2}\psi'''(\beta) \ ,$$

or calculated exactly as $\mathbf{E}(\psi'(\beta + b))$ by univariate integration. Together, these three relations allow estimation of ρ, σ^2 and β.

In the case of a heterogeneous population, the simplest approach is to define relatively homogeneous strata and to apply the procedure described in the previous paragraph in each stratum separately. The series of ρ and σ^2 estimates are then averaged using the frequency of each stratum in the overall population as weight. Given those final estimates of ρ and σ^2, a second round of stratum-specific β values can then be computed.

Special case of Binary data

Relation (7.5) reflects over-dispersion in the marginal distribution i.e. the fact that the relation $\text{var}(Y) = \psi''(\beta)$ is violated, unfortunately, this does not apply to the

binary case where $\text{var}(Y) \equiv \mathbf{E}(Y)(1-\mathbf{E}(Y))$ and there can be no such thing as over-dispersion. We can still use relation (7.4) to estimate σ^2 for fixed values of ρ and the corresponding β by univariate integration of $\psi'(\beta+b)$ in each stratum. As in the general case, the values for σ^2 are averaged across strata and the stratum-specific fixed effects β are re-computed with the average σ^2 as input. This estimation procedure is therefore conditional on an arbitrarily chosen value for ρ.

For common diseases such as migraine (see Section 7.5), we can carry out a more formal procedure based on maximum likelihood. For binary traits, the data consists of stratum-specific 2×2 tables indexed by t. If we use the following notation for the cell numbers in a given 2×2 table t: n_{11}^{t} for affected-affected pairs, n_{10}^{t} for affected-unaffected , n_{01}^{t} for unaffected-affected and n_{00}^{t} for unaffected-unaffected and if $\hat{p}_{\cdot\cdot}^{\text{t}}(\sigma^2, \hat{\beta}(\sigma^2))$ denote the corresponding GLMM probabilities, then the log-likelihood of the data is given by

$$\sum_{\text{table t}} n_{11}^{\text{t}} \log \hat{p}_{11}^{\text{t}} + n_{10}^{\text{t}} \log \hat{p}_{10}^{\text{t}} + n_{01}^{\text{t}} \log \hat{p}_{01}^{\text{t}} + n_{00}^{\text{t}} \log \hat{p}_{00}^{\text{t}} \ .$$

If the trait is common, the GLMM probabilities $\hat{p}_{\cdot\cdot}^{\text{t}}(\sigma^2, \hat{\beta}(\sigma^2))$ can be calculated reasonably fast by Monte Carlo simulations and the maximization with respect to σ^2 is possible. Again, this maximization is carried out for a chosen ρ so this strategy offers a compromise between a full maximization of the marginal likelihood and the ad-hoc method of the previous paragraph.

Although the estimation approach described above is not optimal (in the sense that it is not guaranteed to yield maximum likelihood estimators), its merit is that it quickly provides sensible estimates of the nuisance parameters. The information available is often so sparse that the value of the likelihood depends very weakly (if at all) on the chosen value for ρ. In fact, as the next series of examples illustrates, the choice of ρ seems to have a limited impact on the test for linkage. In Table 1, we computed the relative weights of discordant pairs "AU" and unaffected pairs "UU" compared to affected pairs "AA" for three different values of the random effects' correlation ρ in a wide range of 2×2 tables (i.e. choices of prevalence K and recurrence risk ratios λ_S). In each scenario, we used approximation (7.4) to obtain estimates of the random effect total variance σ^2. As long as ρ is chosen not too small and that the recurrence ratio is not too large, the relative weights given to discordant

		σ^{2*}			AU			UU		
K	λ_S	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$
0.01	1.1	0.5	0.2	0.1	-0.01	-0.01	-0.01	0.00	0.00	0.00
0.01	1.2	1.0	0.4	0. 3	-0.01	-0.01	-0.01	0.00	0.00	0.00
0.01	1.5	2. 6	1.0	0. 6	0.00	-0.01	-0.01	0.00	0.00	0.00
0.01	2.0	5.1	2.0	1.3	0.00	-0.01	-0.01	0.00	0.00	0.00
0.01	3.0	10.2	4.1	2.6	0.00	0.00	-0.01	0.00	0.00	0.00
0.05	1.1	0. 6	0.2	0.1	-0.05	-0.05	-0.05	0.00	0.00	0.00
0.05	1.2	1. 1	0.4	0.3	-0.04	-0.05	-0.06	0.00	0.00	0.00
0.05	1.5	2. 8	1.1	0. 7	-0.03	-0.05	-0.06	0.00	0.00	0.00
0.05	2.0	5.5	2.2	1. 4	-0.02	-0.05	-0.06	0.00	0.00	0.00
0.05	3.0	11.1	4.4	2.8	-0.01	-0.0 3	-0.06	0.00	0.00	0.00
0.10	1.1	0.6	0.2	0. 2	-0.10	-0.11	-0.12	0.01	0.01	0.01
0.10	1.2	1. 2	0.5	0.3	-0.09	-0.11	-0.12	0.01	0.01	0.01
0.10	1.5	3.1	1. 2	0. 8	-0.0 6	-0.11	-0.13	0.00	0.01	0.01
0.10	2.0	6.2	2.5	1. 5	-0.0 4	-0.11	-0.14	0.00	0.01	0.01
0.10	3.0	12.3	4.9	3.1	-0.0 2	-0. 09	-0.15	0.00	0.00	0.01
0.20	1.1	0.8	0.3	0.2	-0.23	-0.26	-0.27	0.05	0.06	0.06
0.20	1.2	1. 6	0.6	0.4	-0.21	-0.26	-0.28	0.04	0.05	0.06
0.20	1.5	3. 9	1. 6	1.0	-0.1 7	-0.2 8	-0.32	0.02	0.05	0.06
0.20	2.0	7.8	3.1	2.0	-0.1 3	-0.2 9	-0.3 8	0.01	0.04	0.06
0.20	3.0	15.6	6 .2	3. 9	-0. 09	-0.2 8	-0.4 5	0.00	0.03	0.06
0.30	1.1	1.0	0.4	0.3	-0.40	-0.45	-0.47	0.14	0.17	0.18
0.30	1.2	2.0	0.8	0.5	-0.38	-0.47	-0.50	0.12	0.17	0.18
0.30	1.5	5.1	2.0	1. 3	-0.33	-0.51	-0. 60	0.09	0.16	0. 20
0.30	2.0	10.2	4. 1	2. 6	-0.27	-0.54	-0.72	0.06	0.16	0.22
0.30	3.0	20.4	8.2	5.1	-0.2 2	-0.5 6	-0.9 0	0.0 3	0.1 5	0.2 6

Table 7.1: Relative weights for Discordant (AU) and unaffected (UU) pairs (compared to affected pairs) for a range of 2×2 tables - * σ^2 obtained using approximation (7.4)

pairs and to a lesser extent, to unaffected pairs depend only weakly upon the initial choice for ρ, although the dependence becomes stronger as the prevalence of the trait increases. When comparing the relative weights of affected pairs for different prevalences/recurrence risk ratios, the dependence is even less noticeable (data not shown). Based on this study, we would advise the choice of a moderate to large value for ρ (0.5 to 0.8) since we favor the corresponding small values for σ^2 (indeed, the quasi-likelihood is based on an approximation valid for small values of σ^2 and so is relation (7.4) used for estimating σ^2).

7.5 Examples

Application to a common disease: Migraine

Migraine is known to be much more frequent in women than in men. In this section, we describe how sex could be accounted for in a linkage study for migraine and quantify the potential gains/losses incurred under different strategies including the

	U m	A m	U f	A f
U m	0.06	-0.60	0.11	-0.33
A m	.	2.71	-1.12	1.57
U f	.	.	0.25	-0.63
A f	.	.	.	1.00

Table 7.2: Relative weights C_i for all sex-sex (f:female and m:male) sib pair combinations (A: Affected and U: Unaffected)

score test presented in Section 7.3. Based on sex-specific prevalence and recurrence risk estimates derived from published data in the Dutch population [Mulder et al., 2003], we first obtain estimates of the segregation parameters ρ, σ^2 and β using the procedure described in Section 7.4. Using possible values of excess IBD sharing, we then quantify the gain obtained by accounting for sex with the score test described above. Mulder et al. [2003] fitted a liability threshold model (i.e. with sex-specific thresholds and a common tetrachoric correlation) to the data. The sex of siblings in a pair defines three possible strata or 2×2 tables, we focused on the Dutch population in the age group 36-68 years old and used the model parameters' estimates to reconstruct those three tables. For the Dutch population, the prevalence for migraine was approximately 0.34 in women and 0.17 in men and the values for λ_S were 1.31, 1.45 and 1.65 in female-female, male-female and male-male sib pairs respectively. Assuming that the three corresponding 2×2 tables were present in proportions $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$ in the overall population, we estimated σ^2 as $\hat{\sigma}^2 = 3.3$ and $\hat{\beta} = (-2.40, -1.03)$ for $\rho = 0.5$ according to the formal maximum-likelihood based method described in Section 7.4. Based on this set of nuisance parameter estimates we calculated the weights for all possible types of sib pairs in the linkage test, these are displayed in table 7.2.

Note, first of all, that affected (and unaffected) sib pairs have positive weights while discordant sib pairs have negative weights. Male-male affected pairs are given much more weight than female-female affected pairs, while the trend is opposite for discordant pairs. One interesting feature is that male-female affected-unaffected pairs are given much more weight than female-male affected-unaffected pairs since the phe-

notypic discordance is more likely to be due to genetic factors in the former than in the latter.

We now compare four possible strategies when testing for linkage in presence of covariates. We define homogeneous groups (indexed by g) of relative pairs (i.e. families) depending on their phenotypic values (AA, AU or UU) and (categorical) covariate values. The excess or reduction in IBD sharing in each group can be parameterized as $\mathbf{E}(\pi - \mathbf{E}\pi \mid \text{ group } g) = \theta\delta_g$ where δ_g can be positive or negative while $\theta \geq 0$. A test for linkage corresponds to testing $\theta = 0$ versus $\theta > 0$. In all tests outlined below, we assume that the sign of δ_g is known (+ for AA and UU and − for AU pairs), depending on what we know or assume about the $|\delta_g|$'s, four testing strategies can be derived:

1. All $|\delta_g|$'s are taken as being equal,
2. The ratios of the $|\delta_g|$'s are known, this is an ideal situation that will serve as reference in our comparison,
3. The $|\delta_g|$'s are estimated from the data,
4. The ratios of the $|\delta_g|$'s are assumed to be given by the score test of Section 7.3.

All four tests but 3. are asymptotically distributed as $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$ under the null hypothesis of no linkage. For test 3., a penalty has to be paid for estimating the weights and the corresponding null distribution is $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_G$ where G is the total number of homogeneous groups considered.

To keep things simple in our numerical comparison of the tests when applied to migraine data, we focused on designs with only sib pairs and two groups ($G = 2$). We compared the efficiency of tests 1., 3. and 4. relative to reference test 2. . To do so, we computed the non-centrality parameters (NCP) for the equivalent χ^2 linkage tests. If C_g denotes the assumed values for the true relative excess IBD sharing δ_g, then all tests but 3. are based upon the following statistic T

$$T = \frac{\sum_g \sum_{i \in \mathrm{g}} C_g(\pi_i - \frac{1}{2})}{(\mathrm{var}(\pi) \times \sum_g N_g C_g^2)^{1/2}} \ ,$$

where N_g denotes the number of families in group g and $N = \sum_g N_g$. For complex traits and thus small gene effect, the variance of π under the alternative hypothesis

is close to its value under the null $\mathrm{var}(\pi \,|\, \text{group g}) \simeq \mathrm{var}(\pi)$ so we have the following approximation:

$$\mathbf{E}(T^2) \simeq 1 + N \times \frac{\left(\sum_g f_g C_g (\mathbf{E}(\pi_g) - \frac{1}{2})\right)^2}{\mathrm{var}(\pi) \times \sum_g f_g C_g^2} \ , \text{ where } f_g = \frac{N_g}{N} \ ,$$

and the sample size for the corresponding 1 d.f. test is inversely proportional to the non-centrality parameter in the previous expression. Asymptotically, the estimates for the weights in test 3. should be very close to their true values, the relative loss of efficiency in test 3. relative to test 2. (where true weights are assumed to be known) is therefore only due to the additional degrees of freedom (d.f.=2 here) of the test. In the context of scan for linkage, using a conservative point-wise type I error rate of 10^{-4}, this loss amounts to about 20%. In the sequel, relative efficiency is expressed as the ratio of sample size in test 2. to sample size in the test of interest.

Using the GLMM described in Section 7.2 (with $\rho = 0.5$, $\sigma^2 = 3.3$ and $\hat{\beta} = (-2.40, -1.03)$ as previously estimated), we mimicked a situation where 10% of the total variance of the random effect is explained by the putative locus while the rest of the variance is either explained by common environment or other unlinked loci [1]. Using Monte Carlo simulations, we closely approximated the average IBD sharing for three types of sib pairs, namely AA male-male, AA female-female and discordant sib pairs AU female-male. In figure 7.1, we display the relative efficiency of the previously defined tests 1., 3. and 4. relative to 2. for two types of study designs: one mixing AA male-male and AA female-female (left-hand side, scenario 1) and one mixing AA male-male and AU female-male (right-hand side, scenario 2). In scenario 1, the 2 degrees of freedom test (test 3.) always fails in improving efficiency compared to a 1 d.f. test with no weight (test 1.) while the score test based on the quasi-likelihood of the GLMM (test 4.) almost always yields improved efficiency with gains close to an ideal strategy (test 2.). In scenario 2, the 2 degrees of freedom test does yield gains in efficiency compared to test 1. that ignores covariates (note that this test can incur efficiency loss up to almost 40% in this situation) when the mixing proportions of AmAm and AfUm are not too extreme, however our test 4. does uniformly better than any of these two tests with losses in efficiency no larger than approximately 10%.

[1] Note that for other values of the proportion of total random effect variance γ explained by the putative locus, the same relative efficiency results hold approximately as long as γ is not too large

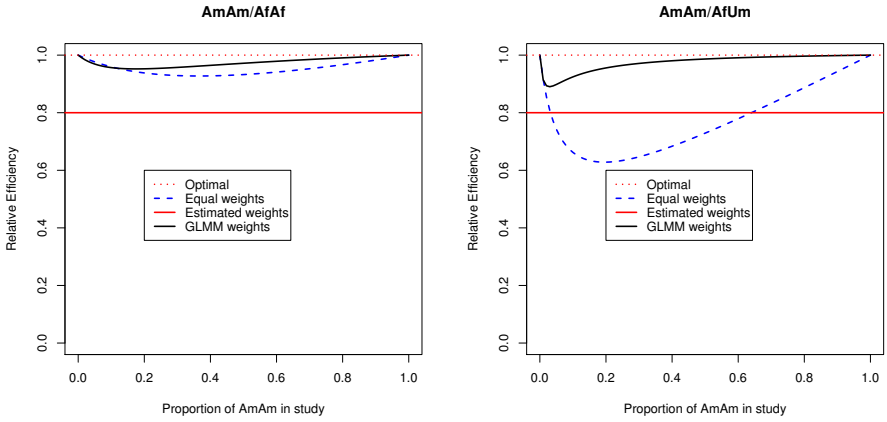

Figure 7.1: Relative efficiency in migraine example - Left: $\mathbf{E}(\pi_1 - \frac{1}{2}) = 0.0033$ in AmAm and $\mathbf{E}(\pi_2 - \frac{1}{2}) = 0.0019$ in AfAf and Right: $\mathbf{E}(\pi_1 - \frac{1}{2}) = 0.0033$ in AmAm and $\mathbf{E}(\pi_2 - \frac{1}{2}) = -0.0008$ in AfUm.

Application to breast cancer

We put ourselves in a situation where ASP's for breast cancer status have been gathered among sib pairs of all ages classified in eight classes (see Table 7.3). The disease status is positive if a woman currently has or has had breast cancer during her life time. For simplicity, we assume that both siblings belong to the same age class. The question is how to weight the excess IBD sharing in each age class effectively.

The genetics of breast cancer is often described using Claus model [Claus et al., 1991] which we will use as the basis for estimation of segregation parameters. Claus model is based on a one-locus model with a rare autosomal dominant allele (q=0.0033) leading to an increased risk of breast cancer. The cumulative probability of a woman to be affected is a function of a woman's age (see Table 2 in [Claus et al., 1991]), based on this model, we derived the prevalence and the recurrence risk ratio (λ_S) for each age class, thereby closely reproducing observed values. Following the informal approach described in Section 7.4, we estimated the variance of the random effects σ^2 in each age-specific 2×2 table based on a correlation equal to $\rho = 0.5$ and used the average value across tables $\hat{\sigma}^2 = 1.96$ (and corresponding age-specific fixed effects).

Age (Years)		Based on Claus model	Based on fitted GLMM	
	K(%)	λ_S	λ_S	Test relative weights
20-29	0.03	10.34	8.	1.70
30-39	0.36	5.97	2.32	1.38
40-49	1.62	2.64	2.26	1.21
50-59	3.09	1.93	2.04	1.11
60-69	5.38	1.44	1.83	1.05
70-79	8.55	1.34	1.70	1.01
80+	13.12	1.15	1.56	1.00

Table 7.3: Prevalence, λ_S in Claus and GLM models, stratum-specific GLMM weights

The series of λ_S's that this GLMM yields is displayed in Table 7.3, it is flatter than the observed ones because the GLMM is stretched to its maximum capacity in order to cover such a wide λ_S-range.

The relative weights for ASP of each age category are given in the last column of Table 7.3, they are fairly mild compared to the large differences observed in λ_S. An approach that would use time of onset rather than current status data is likely to be more efficient, however it is conceptually more complicated. As for migraine, we limited our quantitative comparison to ASP designs with data consisting of two groups: we chose the two most extreme age categories with a relative weight of 1.70. We closely approximated excess IBD sharing in the two age categories in the same way as for the previous example i.e. by mimicking a model where the putative locus explained 10% of the total variance of the random effect while the rest of the variance can be conceived as arising either from a common environment or other unlinked loci [2] under the fitted GLMM. Under this model, our approximate score test 4. is the one closest to the ideal test 2. ; test 3. sometimes performs better than test 1. however this advantage would disappear if data consisted (more realistically) of sib pairs in all age categories (see Fig. 7.2).

[2] but note that the same remark regarding relative efficiency holds as for the migraine example

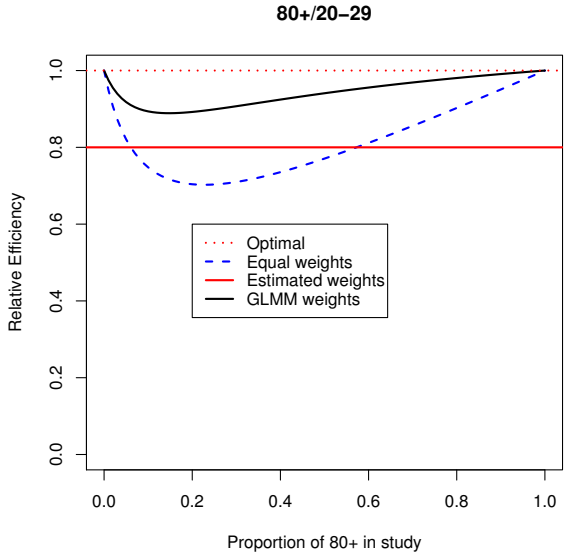

Figure 7.2: Relative efficiency in breast cancer - $\mathbf{E}(\pi_1 - \frac{1}{2}) = 0.017$ and $\mathbf{E}(\pi_2 - \frac{1}{2}) = 0.005$ in "20-29" and "80+", resp.

7.6 Discussion

Based on the GLMM, we have derived a test for linkage which makes adjustment for known marginal covariate effects. Our approach is motivated by the fact that the effect of important covariates on the marginal distribution of a trait is often known via data external to the linkage study itself, and these should be incorporated in the linkage analysis. We elude the difficult and computationally intensive problem of making exact inference based on the likelihood of the GLMM by using a quasi-likelihood, our test is then based upon a score test for the linkage parameter in this quasi-likelihood and turns out to be a tractable statistic, in fact, a simple weighted average of the excess IBD sharing between all pairs of relatives in a family. In that respect, it is reminiscent of approximate likelihoods based on pairwise joint distributions used, for example, with correlated binary data [le Cessie and van Houwelingen, 1994]. As noted by Cox and Reid [2004], the use of such pseudo-likelihoods does not only alleviate the computational burden, it also enhances the robustness of the method to model specification. It must be recognized, however, that in absence of covariates, better family-specific tests that take the full IBD distribution into account can be derived [Teng and Siegmund, 1997]. If the GLMM correctly describes the data, we can draw two general conclusions about the effect of covariate adjustment in linkage studies for binary traits. For rare traits where only affected pairs of individuals are informative, the effect of covariates needs to be huge in order for any covariate-adjustment to yield substantial power gains. Indeed, the excess IBD sharing differs only a little between covariate-specific types of affected pairs. For common traits, the gains are more easily achieved. Firstly, because discordant pairs can be more confidently included in the analysis if relevant covariates (e.g. age and sex) are taken into account, and those pairs do become informative in common traits. Secondly, because the ratios of deviations in IBD sharing between phenotypic-covariate specific strata are more likely to be large for such traits.

The test is applicable in arbitrary pedigrees, and in the case of binary traits, it allows incorporation of both affected and unaffected individuals. This way of handling the issue of covariates in binary traits, contrasts with existing methods that only use the linkage data available and model the probability of IBD sharing as a

function of covariates. The most general representative of this type of models (i.e. which in principle can handle arbitrary pedigrees and both affected and unaffected individuals) is undoubtedly the conditional logistic model [Olson, 1999; Greenwood and Bull, 1999]. It is implemented in the `LODPAL` program of the `S.A.G.E.` software but as far as we are aware (true for version 5.1), the current implementation suffers from a few important limitations: the program assumes that all pairs of relatives are independent, the covariates have to be pair-specific, when both affected and discordant pairs are analyzed together, the program cannot handle covariates. These issues do not arise in our approach. The strength of methods that let IBD sharing depend upon covariate values invariably turns into a weakness (unless differences between covariate-specific groups are very large) as the number of covariates increases because the d.f. of the corresponding test for linkage increases too. We overcome this problem by incorporating external data and by specifying a model where differences in IBD sharing naturally arise. The way we handle covariates by feeding some covariate-adjusted residuals into the linkage analysis is conceptually similar to the method advocated for sibships by Alcais [2001]. For general pedigrees however, as far as we are aware, our test actually appears to be the only available practical way to simultaneously adjust for covariates and to include both affected and unaffected individuals. In late onset diseases, the suspicion that younger unaffected individuals might become affected at a later age can explicitly be incorporated using age as a covariate. We have treated all segregation parameters required by the GLMM as known parameters and although unbiased estimates could be difficult to obtain, we propose an estimation procedure that circumvents this problem. As long as interest lies in testing for linkage and not in actually estimating segregation parameters, this procedure appears to be acceptable in that: 1) it does not affect the size of our test 2) the test itself is fairly insensitive to the non-unique choices of nuisance parameter values. By illustrating the use of our method in both common and relatively rare diseases, we have shown the order of magnitude for the gains that could be expected in some specific scenarios. We note that the GLMM model does not explicitly incorporate potential gene by covariate interaction in its structure, this is not to say that it forbids this phenomenon, indeed, the recurrence risk ratios and IBD sharing induced by the model clearly vary depending on covariate values. However, purely

for mathematical convenience, we have assumed that on the latent scale, there was no interaction between the gene at the putative location and the covariate. Actually, recent developments published by Peng et al. [2005] explicitly account for such interactions and these authors have derived the corresponding score test for linkage. The gene by covariate interaction could be explicitly incorporated into the GLMM model in a similar way (via the $\mathbf{R}$ matrix of variance-covariance of random effects) and the corresponding test would obtain analogously. We note that in practice the IBD status is not known exactly but has to be estimated from marker data, the consequence for the score test is that π has to be replaced by its estimated version $\hat{\pi}$ in equation (7.2) and that the corresponding var($\hat{\pi}$) has to be used in the standardization of the test. This last term depends on the family structure, the marker allele frequencies, their position and the possible genotype missingness pattern, and in practice we approximate its true value using Monte Carlo simulations as implemented in an executable C program calling upon the `MERLIN` [Abecasis et al., 2002] software and available at `http://www.msbi.nl/Genetics/`. Currently, the GLMM test prescribed in this manuscript is only available as R code from the authors. Finally, we remark that although we have focused on the use of our test with binary traits, the approach can directly be applied to other traits whose distribution is in the canonical exponential family, in particular to count data with a Poisson distribution as well as survival data.

7.7 Appendix

Derivation of the quasi-likelihood

We use a 2^{nd} order Taylor approximation of the conditional log-likelihood $\log f(y \mid \beta, b)$ introduced in Section 7.2 around $\mathbf{b} = 0$ to obtain a quasi-likelihood for the data $\mathbf{y}$ in

a family:

$$\begin{aligned}
\log f(\mathbf{y} \mid \beta, \mathbf{b}) &= \sum_{i=1}^{m} \log f(y_i \mid \beta, b_i) \\
&\simeq \sum_{i=1}^{m} \log f(y_i \mid \beta, b_i = 0) + b_i(y_i - \psi\prime(\mathbf{x}_i\beta)) - \frac{1}{2} b_i^2 \psi\prime\prime(\mathbf{x}_i\beta) \\
&\simeq \sum_{i=1}^{m} \log f(y_i \mid \beta, b_i = 0) - \frac{1}{2}\psi\prime\prime(\mathbf{x}_i\beta)\left(b_i - \frac{y_i - \psi\prime(\mathbf{x}_i\beta)}{\psi\prime\prime(\mathbf{x}_i\beta)}\right)^2 \\
&\quad + \frac{1}{2}\psi\prime\prime(\mathbf{x}_i\beta)\left(\frac{y_i - \psi\prime(\mathbf{x}_i\beta)}{\psi\prime\prime(\mathbf{x}_i\beta)}\right)^2 .
\end{aligned}$$

In the previous expression, only the second term involves $\mathbf{b}$ which shows that when β is regarded as constant, $\log f(\mathbf{y} \mid \beta, \mathbf{b})$ behaves as if

$$\frac{\mathbf{y} - \boldsymbol{\psi}\prime(\mathbf{X}\beta)}{\boldsymbol{\psi}''(\mathbf{X}\beta)} \mid \mathbf{b} \sim N(\mathbf{b}, \boldsymbol{\Psi}''(\mathbf{X}\beta)^{-1})$$

where $\boldsymbol{\Psi}\prime\prime(\mathbf{X}\beta)$ denotes the diagonal matrix whose i^{th} diagonal element is given by $\psi\prime\prime(\mathbf{x}_i\beta)$. We can now easily integrate the random effects $\mathbf{b} \sim N(0, \mathbf{R}(\theta, \gamma))$ out and $\log f(\mathbf{y} \mid \beta)$ as a function of θ can be regarded as the value of the density for multivariate normal $N(0, \mathbf{R}(\theta, \gamma) + \psi\prime\prime(\mathbf{X}\beta)^{-1})$ in the data points $\frac{\mathbf{y} - \boldsymbol{\psi}\prime(\mathbf{X}\beta)}{\boldsymbol{\psi}\prime\prime(\mathbf{X}\beta)}$:

$$\frac{\mathbf{y} - \boldsymbol{\psi}\prime(\mathbf{X}\beta)}{\boldsymbol{\psi}''(\mathbf{X}\beta)} \sim N(0, \mathbf{R}(\theta, \gamma) + \boldsymbol{\Psi}''(\mathbf{X}\beta)^{-1}) .$$

Score test

In analogy with the case of normally distributed phenotypes [Lebrec et al., 2004], standard results on matrix algebra (see, e.g. [Searle et al., 1992, Appendix M.7]) lead to

$$\ell_\gamma^{\mathbf{z}} = \frac{\partial \log ql}{\partial \gamma} = \frac{1}{2}\left\{\mathbf{z}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi})\boldsymbol{\Sigma}^{-1}\mathbf{z} - \mathrm{tr}(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi}))\right\}$$

Because of the relation $a'b = \mathrm{tr}(ba')$, the previous equation can be rewritten

$$\frac{\partial \log ql}{\partial \gamma} = \frac{1}{2}\,\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi})(\boldsymbol{\Sigma}^{-1}\mathbf{z}\mathbf{z}' - \mathbf{I})\right) .$$

Here $\mathrm{tr}(A)$ stands for the trace (sum of the diagonal elements) of matrix A. Using elementary matrix theory, in particular $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ and $\mathrm{tr}(AB) = \mathrm{vec}(A')'\mathrm{vec}(B)$ (here $\mathrm{vec}(A)$ places the n columns of the $m \times n$ matrix A into a vector of dimension

$mn \times 1$), this score function can be rewritten as

$$\ell_\gamma^{\mathbf{z}} = \frac{1}{2} \text{ vec}(\mathbf{C})' \text{ . vec}(\boldsymbol{\pi} - \mathbf{E}\boldsymbol{\pi})$$

with $\mathbf{C} = \boldsymbol{\Sigma}^{-1}\mathbf{z}\left(\boldsymbol{\Sigma}^{-1}\mathbf{z}\right)' - \boldsymbol{\Sigma}^{-1}$.

Approximation used in segregation parameters estimation

The marginal covariance can be partitioned as

$$\begin{aligned} \text{cov}(Y_1, Y_2) &= \mathbf{E}\left(\text{cov}(Y_1, Y_2 \mid \beta_1, \beta_2, b_1, b_2)\right) + \text{cov}\left(\mathbf{E}(Y_1 \mid \beta_1, b_1), \mathbf{E}(Y_2 \mid \beta_2, b_2)\right) \\ &\approx 0 + \text{cov}\left(\psi'(\beta_1) + b_1\psi''(\beta_1),\ \psi'(\beta_2) + b_2\psi''(\beta_2)\right) \ , \end{aligned}$$

using a 1^{st} order Taylor expansion of $\psi'(\beta_i + b_i)$. It follows that $\text{cov}(Y_1, Y_2) \approx \psi''(\beta_1)\ \psi''(\beta_2)\ \rho\sigma^2$. The approximation $\text{var}(Y) \approx \psi''(\beta) + \psi''(\beta)^2\,\sigma^2$ obtains in the same manner by setting $\rho = 1$ and taking a 1^{st} order Taylor approximation of $\text{var}(Y \mid \beta, b) = \psi''(\beta + b) \approx \psi''(\beta) + b\ \psi'''(\beta)$.

For the marginal mean, we have

$$\begin{aligned} \mathbf{E}(Y) &= \mathbf{E}\left(\mathbf{E}(Y \mid \beta, b)\right) \\ &\approx \mathbf{E}\left(\psi'(\beta) + b\psi''(\beta) + \frac{b^2}{2}\psi'''(\beta)\right) \\ &\approx \psi'(\beta) + \frac{\sigma^2}{2}\psi'''(\beta) \ . \end{aligned}$$