Chapter 5

Potential Bias in Generalized Estimating Equations Linkage Methods under Incomplete Information

Abstract

The mean identity-by-descent (IBD) specification used in the Generalized Estimating Equations (GEE) methodology for linkage is only valid, strictly speaking, under the assumption of fully polymorphic markers. In practice, markers often provide only partial IBD information which can potentially result in inconsistency of the locus location and gene effect estimates obtained by the GEE method. Using both simulations and theory, we identify some realistic conditions about marker information under which the validity of the GEE linkage methods may be arguable. Namely, researchers should not trust the GEE parameters’ estimates and their associated confidence intervals in areas of the genome where IBD information is sparse or when this information changes abruptly. We show that properly standardized statistics based on IBD sharing provide a valid alternative.

5.1 Introduction

Since Liang et al. [2001] introduced the use of Generalized Estimating Equations (GEE) with the purpose of estimating the position of a locus linked to a trait, there has been increasing interest in this methodology. The approach has attractive features, in particular, it allows researchers to set a confidence interval around the estimate of the locus position. In the meantime, some refinements and extensions of the approach are being developed: covariates can be introduced [Glidden et al., 2003; Chiou et al., 2005], the methodology can be extended to two linked loci in the region [Biernacka et al., 2005] and to general pedigrees [Schaid et al., 2005], and it bears potential for a wider use in the future. Strictly speaking, the GEE linkage method is only valid when markers are fully polymorphic, in other words, when identity-by-descent (IBD) status at markers is known with certainty. As far as we are aware, little has been done to assess how robust the method is under more realistic conditions of marker information. Indeed, among the aforementioned articles, those that included simulations almost always generated complete IBD data at markers. The only exception is Biernacka et al. [2005] who recognized that the use of non-fully informative marker maps produced biased estimates of the genetic effects but hardly any bias in the estimate of locus position, however they only looked at evenly distributed marker maps. In this article, we identify some realistic conditions about marker information under which the validity of the GEE linkage methods may be arguable, properly standardized statistics based on IBD sharing provide a valid alternative. In the ‘Methods’ section, we review the principles of the GEE method and show why it may lead to biased and inconsistent estimation and we prove that some more classical approaches do not suffer the same drawback under certain conditions. The ‘Results - Monte Carlo simulations’ section is devoted to simulations that illustrate the findings of the previous section in a range of realistic scenarios. Finally, in the ‘Discussion’ section, we discuss our findings and their possible practical impact on linkage analysis.
5.2 Methods

The GEE methodology

We start by recalling the principle of the GEE methodology as applied to linkage mapping. For affected sib pairs (ASP) the method is based on the mean specification of the excess IBD sharing at markers as

\[
E(\pi_t - \frac{1}{2}) = \frac{1}{8}(1 - 2\theta_{t,\tau})^2 C = \mu_t(\tau, C),
\]

(5.1)

where \(\pi_t\) denotes the true proportion of alleles shared IBD at marker or position \(t\), \(\tau\) the position of the true and only locus in the region, \(\theta_{t,\tau}\) the recombination fraction between locations \(t\) and \(\tau\), while \(C\) reflects the genetic model (note here that \(C\) in the previous equation is 4 times the \(C\) parameter used in Liang et al. [2001]). We stress that the derivation of this result assumes that markers are fully polymorphic. In practice, IBD is uncertain and is estimated using multipoint marker data, it is well known that the consequence of incomplete information is to shrink the estimated IBD towards its null value \(\frac{1}{2}\), as a result the previous mean model might be erroneous. We distinguish the true (often unobserved) proportion of alleles shared IBD \(\pi\) from its estimated counterpart by the use of the notation \(\hat{\pi}\).

We assume that we have data from \(i = 1, \ldots, N\) ASPs available at marker positions \(t_1, \ldots, t_M\) with corresponding IBD sharing estimates \(\hat{\pi}_i = (\hat{\pi}_{i,t_1}, \ldots, \hat{\pi}_{i,t_M})'\), where ' denotes the transpose of a matrix (bold letters indicate a matrix or a vector as opposed to a scalar). We denote by \(V\) the \(M \times M\) working variance-covariance matrix for \(\hat{\pi}_i\) while \(\mu = \mu(\tau, C) = (\mu_{t_1}, \ldots, \mu_{t_M})'\) then estimation of the parameters \(\tau\) and \(C\) is carried out by solving the following GEE

\[
\sum_{i=1}^{N} \left( \frac{\partial \mu}{\partial (\tau, C)} \right)' V^{-1} (\hat{\pi}_i - \mu(\tau, C)) = 0.
\]

The theory developed by Liang and Zeger [1986] ensures that as long as the mean of the observations is correctly specified (i.e. \(E(\hat{\pi}_i) = \mu(\tau, C)\)), the GEE estimators of \(\tau\) and \(C\) converge towards the true locus position and genetic effects as the sample size \(N\) increases. A specification of \(V\) as the true variance-covariance matrix of the observations \(\hat{\pi}_i\) in terms of the unknown parameter \(\tau\) and \(C\) was given
in Liang et al. [2001] (again, under complete information) but is not essential to the consistency of the procedure, it only affects its efficiency. In addition, an asymptotically robust variance-covariance matrix for the estimates \((\hat{\tau}, \hat{C})\)' can be computed as 

\[ \hat{\Sigma} = \hat{\Sigma}_1^{-1}\hat{\Sigma}_2\hat{\Sigma}_1^{-1} \]

with

\[ \hat{\Sigma}_1 = N \left( \frac{\partial \mu}{\partial \tau(C)} \right)' V^{-1} \left( \frac{\partial \mu}{\partial \tau(C)} \right) \]

\[ \hat{\Sigma}_2 = \sum_{i=1}^{N} \left( \frac{\partial \mu}{\partial \tau(C)} \right)' V^{-1} \left( \frac{\partial \mu}{\partial \tau(C)} \right) \frac{\partial \mu}{\partial \tau(C)} \] \[ \frac{\partial \mu}{\partial \tau(C)} \]

where \( \frac{\partial \mu}{\partial \tau(C)} \) and possibly \( V \) are evaluated in \((\hat{\tau}, \hat{C})\).

**An accurate IBD specification under incomplete information**

The relation \( E(\hat{\pi}) = \mu(\tau, C) \) between the mean of the estimated IBD sharing and the locus position \( \tau \) and gene effect \( C \), exactly true when IBD is perfectly known, is only approximate under incomplete information. In fact, Teng and Siegmund [1998] have shown that a theoretical mean IBD specification can also be derived under incomplete information, namely for a one-locus (located at \( \tau \)) additive model on the IBD scale (which is approximately true for a wide range of disease models; exactly true if \( \lambda_S = \lambda_O \) [Risch, 1990]) such that

\[
\begin{align*}
P(\pi = 0 | ASP) &= \frac{1}{4} - \frac{1}{8}C \\
P(\pi = \frac{1}{2} | ASP) &= \frac{1}{2} \\
P(\pi = 1 | ASP) &= \frac{1}{4} + \frac{1}{8}C
\end{align*}
\]

the expected observed excess IBD sharing at any arbitrary position \( t \) is given by

\[
E(\hat{\pi}_t - \frac{1}{2} | ASP) = \text{cov}_0(\hat{\pi}_t, \hat{\pi}_t) C
\]

where the covariance \( \text{cov}_0(\hat{\pi}_t, \hat{\pi}_t) \) is taken under the null hypothesis (it therefore only depends on marker map characteristics, pedigree structure and possibly missing genotype patterns). For the sake of completeness, we show a proof of this crucial result in the appendix. The correct specification of the mean IBD sharing as a function of the locus position \( \tau \) and genetic effect \( C \) is essential in order to obtain valid estimates by the GEE method. Comparison of Equations (5.3) and (5.1) allows one to evaluate the discrepancy between the correct IBD specification and the one used
in the GEE linkage methods. For illustration purposes, we have displayed two typical extreme examples in Figure 5.1 assuming the true locus is at \( \tau = 25 \text{cM} \). Under incomplete information, the variances \( \text{var}_0(\tilde{\tau}_t) \) and \( \text{var}_0(\tilde{\tau}_\tau) \) are reduced from their fully polymorphic value \( \frac{1}{2} \) while the correlation \( \text{cor}_0(\tilde{\tau}_t, \tilde{\tau}_\tau) \) is increased compared to its complete information value \( (1 - 2\theta_{t, \tau})^2 \); the net effect is a decrease of \( \text{cov}_0(\tilde{\tau}_t, \tilde{\tau}_\tau) \).

The exact relationship between \( \text{cov}_0(\tilde{\tau}_t, \tilde{\tau}_\tau) \) and \( \tau \) is complex in general, however the covariance is taken under the null hypothesis and can therefore easily and accurately be calculated by Monte Carlo simulations (or gene dropping simulations) as advocated in Lebrec et al. [2004]: we used the \(--\text{simulate}\) option in \textsc{merlin} to generate marker data for a few thousand sib pairs and calculated the sample covariance between \( \tilde{\tau}_t \) and \( \tilde{\tau}_\tau \) after obtaining multipoint estimates of IBD sharing by use of the \(--\text{kin}\) option in \textsc{merlin} (in general, one such simulation has to be done for each type of pedigree and missing genotype pattern). Note that \( \text{var}_0(\tilde{\tau}_t) \) can be computed at any arbitrary position \( t \) in a similar manner. We have displayed three possible IBD mean specifications in Figure 5.1: the correct one, \( \text{cov}_0(\tilde{\tau}_t, \tilde{\tau}_\tau)C \), labelled ‘T&S’, the one under complete information, \( \frac{1}{8}(1 - 2\theta_{t, \tau})^2C \), labelled ‘GEE’ and a third one, \( (1 - 2\theta_{t, \tau})^2 \sqrt{\text{var}_0(\tilde{\tau}_t)\text{var}_0(\tilde{\tau}_\tau)}C \), labelled ‘Var Corrected’ that corrects for the incomplete marker information by using the correct variances \( \text{var}_0(\tilde{\tau}_t) \) and \( \text{var}_0(\tilde{\tau}_\tau) \) but keeping the correlation as in the ideal situation of complete information (i.e. too low).

In the symmetric information case (Left panel: two markers with 10 equi-frequent alleles at 20cM and 40cM), the location estimate will in practice incur little harm (but the estimate of \( C \) will). In presence of asymmetric information (Right panel: two markers with 2 and 10 equi-frequent alleles at 20cM and 40cM respectively), the true expected excess IBD is lower at marker A than at marker B although \( \tau \) is closer to A, however the true expected excess IBD sharing as per ‘GEE’ is grossly misspecified since expected IBD is supposed to be much higher at A than at B, the location estimate will be biased towards the more informative marker B, the ‘Var Corrected’ specification does a better job at approaching the true IBD mean specification but is not accurate.
Figure 5.1: Comparison of different mean specifications for excess IBD sharing at position \( t \) (\( \mathbb{E}(\hat{v}_t - \frac{1}{2} | \text{ASP}) \)) - ‘T&S’ (the correct one): \( \text{cov}_0(\hat{v}_t, \hat{v}_\tau)C \), ‘GEE’ (assumes complete information): \( \frac{1}{4} (1 - 2\theta_{t,\tau})^2 C \) and ‘Var Corrected’: \( (1 - 2\theta_{t,\tau})^2 \sqrt{\text{var}_0(\hat{v}_t)} \sqrt{\text{var}_0(\hat{v}_\tau)} C \).

A consistent score test

Feingold et al. [1993] have shown that under a complete high-resolution map, the global test for linkage based on excess IBD sharing given by the supremum of \( Z_t = \frac{\sum_{i=1}^{N} \hat{v}_{t,i} - \frac{1}{2}}{\sqrt{\frac{1}{N^2}}} \) over the putative chromosomal positions \( t \) is the log-likelihood ratio test of a Gaussian process for testing the null hypothesis of no linkage and therefore provides a consistent estimate of the true disease locus location \( \tau \). When information is incomplete, a similar test was proposed by Teng and Siegmund [1998] as the maximum of \( \hat{Z}_t \) across marker positions with

\[
\hat{Z}_t = \frac{\sum_{i=1}^{N} \hat{v}_{t,i} - \frac{1}{2}}{\sqrt{\sum_{i=1}^{N} \text{var}_0(\hat{v}_{t,i})}},
\]

where \( \text{var}_0(\hat{v}_{t,i}) \) may be computed as in subsection ‘An accurate IBD specification under incomplete information’. Although their test was based on evaluation of \( \hat{Z}_t \) across marker positions only, there is no practical reason for such a restriction when IBD is calculated using multipoint methods and one can in theory calculate \( \hat{Z}_t \) on an arbitrarily fine grid of putative locations. Assuming the locus is at \( \tau \), the statistic \( \hat{Z}_\tau \) turns out to be the score test [Cox and Hinkley, 1974] for the \( C \) parameter in the
additive model (5.2) \(^1\) and we refer to this test as such in the sequel. One obvious estimator of the locus position is the location \( t = \hat{\tau} \) where \( \hat{Z}_t \) is maximized in the chromosomal region of interest. We are unaware of a formal proof that as in the case of a high-resolution map, \( \hat{\tau} \) provides a consistent estimate of the true locus position, although this is probably known from experience. It turns out to be a corollary of relation (5.3) as we show in an appendix. In addition, one can obtain bootstrap confidence intervals (CI) by resampling with replacement among the \( N \) sib pairs and recalculating \( \hat{\tau} \) such that \( Z_{\hat{\tau}} = \sup_t \hat{Z}_t \) in each new sample. In fact, this score test is also the score test corresponding to the exponential model used by Kong and Cox [1997] although they prefer to use the corresponding likelihood ratio test. It is perhaps worth stressing that the standardization used in \( \hat{Z}_t \) is crucial to the consistency of the method, older non-parametric linkage (NPL) methods for ASPs were based on excess IBD sharing only (i.e. the numerator of \( \hat{Z}_t \)) and the corresponding maximum LOD score thus gave inconsistent estimates of the position under uneven incomplete information even when IBD estimation was done in a multipoint fashion.

5.3 Results - Monte Carlo simulations

In order to assess the impact of incomplete information in practice, we carried out a number of simulations: we generated data from a simple one-locus bi-allelic (disease allele \( D \) frequency=0.1) additive model (penetrances=0.0, 0.5 and 1.0 in \( dd, Dd \) and \( DD \) genotypes resp.; \( \lambda_S = \lambda_O = 3.25 \)). A set of 11 equally-spaced markers spanned a 0 – 100cM region and the locus was positioned between the 5\(^{th} \) and 6\(^{th} \) marker at either 42.5cM, 45cM or 47.5cM. We looked at three distinct marker maps (mapH, mapM and mapL) reflecting an increasing degree of systematic differences in marker information; the last six markers always had 10 equi-frequent alleles whereas the first five markers had 8 equi-frequent alleles in mapH, 4 equifrequent alleles in mapM and 2 equi-frequent alleles in mapL. Finally, for each scenario, we considered three sample sizes \( N = 100, 200 \) and 500 ASPs without parents. In all methods of analysis described below, multipoint IBD estimation was carried out using MERLIN [Abecasis et al., 2002]. The locus position and genetic effect were estimated according to the

\[^1\]More precisely, in the model \( P(g|\text{ASP}) = \sum_{l=0}^{1} \frac{1}{2} P(g|\pi_r = l) P(\pi_r = l|\text{ASP}) \) where \( g \) is the multipoint marker information available and \( P(\pi_r = l|\text{ASP}) \) is given by model (5.2).
GEE method using GeneFinder [Liang et al., 2001], both asymptotic and bootstrap 95% confidence intervals (CI) were calculated. We also carried out two classical analyses for ASP: on a fine grid of chromosomal positions (every cM), we calculated the Kong and Cox [1997] test and the score test $\tilde{Z}_t$ defined in subsection ‘A consistent score test’, the positions where the respective maximum of these two statistics were attained provided position estimates for the locus. In addition, for the score test, we calculated 95% ordinary bootstrap CIs by resampling among the $N$ ASPs. All results are presented in table 5.1.

The GEE estimates of the location are subject to bias which increases as the asymmetry in marker map becomes stronger and which does not decrease with increasing sample size. Although this bias might be considered small, it leads to lower than nominal coverage probability even for the bootstrap CIs, this coverage probability can potentially decrease further as the sample size goes up. Note that a bootstrap algorithm adjusting for bias [Wehrens et al., 2000] could be used here. In contrast, the location estimates obtained by the score test have low bias (probably due to the discrete nature in the search for the supremum of $\tilde{Z}_t$ and inaccuracy in calculating $\text{var}_0(\tilde{\pi}_t)$) independent of the marker map, the corresponding bootstrap CIs have close to nominal coverage probability.

### 5.4 Discussion

The GEE methodology offers an attractive and flexible framework for fine mapping of disease loci and its use will likely continue to spread in the coming years. Researchers should therefore all the more be aware of its limitations. Estimates of disease locus position (as well as genetic effect) and associated confidence intervals obtained by existing GEE methods should not be trusted in areas of the genome where IBD information is sparse in particular when this information changes abruptly. In these instances, properly standardized classical methods based on excess IBD sharing, when applied on a fine grid of locations, do provide consistent estimates of the location. Associated confidence intervals with correct coverage probability can also be obtained by re-sampling techniques such as the bootstrap.

The reason for underrating the issue of incomplete information has probably to
### Table 5.1: Results of simulations.

*Information content is expressed as the range of average information content as defined in Kruglyak and Lander [1995] over the 0-100cM region.*

<table>
<thead>
<tr>
<th>True location</th>
<th>Map</th>
<th>N</th>
<th>Average Estimate (cM)</th>
<th>95% Asymptotic CI coverage (%)</th>
<th>95% Bootstrap CI coverage (%)</th>
<th>Average Estimate (cM)</th>
<th>95% Bootstrap CI coverage (%)</th>
<th>Average Estimate (cM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>42.5cM</td>
<td>MapL</td>
<td>100</td>
<td>46.4</td>
<td>71.7</td>
<td>78.9</td>
<td>42.4</td>
<td>94.9</td>
<td>42.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>46.4</td>
<td>58.2</td>
<td>63.8</td>
<td>42.3</td>
<td>94.2</td>
<td>42.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>46.3</td>
<td>27.8</td>
<td>32.9</td>
<td>42.2</td>
<td>95.4</td>
<td>42.2</td>
</tr>
<tr>
<td></td>
<td>MapM</td>
<td>100</td>
<td>43.9</td>
<td>84.9</td>
<td>89.8</td>
<td>41.9</td>
<td>95.7</td>
<td>41.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>44.1</td>
<td>83.3</td>
<td>86.1</td>
<td>42.1</td>
<td>94.4</td>
<td>42.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>44.2</td>
<td>76.5</td>
<td>78.5</td>
<td>42.2</td>
<td>94.8</td>
<td>42.3</td>
</tr>
<tr>
<td></td>
<td>MapH</td>
<td>100</td>
<td>43.1</td>
<td>85.9</td>
<td>92.1</td>
<td>42.3</td>
<td>95.4</td>
<td>42.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>43.0</td>
<td>86.3</td>
<td>92.0</td>
<td>42.1</td>
<td>95.7</td>
<td>42.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>43.1</td>
<td>88.3</td>
<td>90.4</td>
<td>42.3</td>
<td>94.6</td>
<td>42.3</td>
</tr>
<tr>
<td>45cM</td>
<td>MapL</td>
<td>100</td>
<td>48.2</td>
<td>78.3</td>
<td>84.7</td>
<td>45.7</td>
<td>96.5</td>
<td>45.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>48.1</td>
<td>75.8</td>
<td>77.3</td>
<td>45.4</td>
<td>96.1</td>
<td>45.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>47.8</td>
<td>51.6</td>
<td>53.3</td>
<td>45.1</td>
<td>98.0</td>
<td>45.1</td>
</tr>
<tr>
<td></td>
<td>MapM</td>
<td>100</td>
<td>46.4</td>
<td>80.9</td>
<td>90.2</td>
<td>45.0</td>
<td>95.1</td>
<td>45.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>46.1</td>
<td>90.9</td>
<td>91.9</td>
<td>45.1</td>
<td>97.5</td>
<td>45.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>46.0</td>
<td>91.3</td>
<td>90.0</td>
<td>44.9</td>
<td>96.8</td>
<td>44.9</td>
</tr>
<tr>
<td></td>
<td>MapH</td>
<td>100</td>
<td>45.2</td>
<td>85.1</td>
<td>92.6</td>
<td>45.1</td>
<td>97.6</td>
<td>45.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>45.0</td>
<td>94.7</td>
<td>95.3</td>
<td>45.0</td>
<td>96.6</td>
<td>44.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>45.1</td>
<td>96.4</td>
<td>95.5</td>
<td>45.0</td>
<td>97.3</td>
<td>45.0</td>
</tr>
<tr>
<td>47.5cM</td>
<td>MapL</td>
<td>100</td>
<td>49.6</td>
<td>79.2</td>
<td>89.0</td>
<td>47.9</td>
<td>94.9</td>
<td>47.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>49.5</td>
<td>76.5</td>
<td>86.2</td>
<td>47.8</td>
<td>94.9</td>
<td>47.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>49.3</td>
<td>78.2</td>
<td>80.7</td>
<td>48.0</td>
<td>94.2</td>
<td>47.8</td>
</tr>
<tr>
<td></td>
<td>MapM</td>
<td>100</td>
<td>48.2</td>
<td>84.5</td>
<td>91.8</td>
<td>47.8</td>
<td>94.7</td>
<td>47.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>48.1</td>
<td>84.6</td>
<td>91.2</td>
<td>47.8</td>
<td>95.8</td>
<td>47.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>47.9</td>
<td>88.1</td>
<td>92.6</td>
<td>47.7</td>
<td>94.9</td>
<td>47.7</td>
</tr>
<tr>
<td></td>
<td>MapH</td>
<td>100</td>
<td>47.3</td>
<td>86.3</td>
<td>92.8</td>
<td>48.0</td>
<td>95.4</td>
<td>47.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>47.4</td>
<td>87.4</td>
<td>93.5</td>
<td>48.0</td>
<td>95.5</td>
<td>47.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>47.3</td>
<td>91.6</td>
<td>94.4</td>
<td>47.9</td>
<td>95.3</td>
<td>47.7</td>
</tr>
</tbody>
</table>
do with the nature of the linkage mapping process which usually involves two stages: following a first low-density scan, higher-density genotyping is carried out in one or several promising regions. In this case, IBD information can be fairly accurately determined and the GEE methodology is directly applicable. The advent of SNP chip data for linkage has the potential to provide marker maps with not only higher but also less variable information content [Evans and Cardon, 2004; Schaid et al., 2004] than in classical microsatellites maps, this could potentially increase the reliability of the GEE method in the future. Of course, SNP chip data can only hold such a promise if the data are used in a multipoint fashion for IBD estimation which requires the careful elimination of markers in linkage disequilibrium. However, there are specific situations where similar scenarios to those chosen in our simulations will occur. For example, researchers sometimes embark on collaborative projects (or meta-analysis) whereby several already existing genomewide scans are pooled together in the hope to gain sufficient power (e.g. GenomEU twin project). In the search for complex traits (with inherent small genetic effects), this second strategy is likely to become more popular. Those distinct scans are often carried out using different marker maps and their pooling will inevitably give rise to regions with heterogeneous IBD information at least in part of the large pooled data set. For those reasons, we believe that the scenarios envisaged in our simulations (and perhaps even more extreme ones as we have personally experienced) are realistic and that our findings have practical implications.

5.5 Appendix

Expected IBD sharing in ASP

We show a proof of the result concerning the expected excess IBD sharing in ASPs under incomplete information. This result is actually due to Teng and Siegmund [1998]. Recall first that \( \hat{\pi} = \hat{\pi}(g) = E_0(\pi \mid g) = \frac{1}{2} P_0(\pi = \frac{1}{2} \mid g) + P_0(\pi = 1 \mid g) \) where \( g \) is the multipoint marker genotype information available (the subscript 0 indicates
a probability $P_0$ or expectation $E_0$ independent of the disease locus, then:

$$E(\bar{X}_{\tau} - \frac{1}{2} \mid ASP) = \sum_g (\bar{X}_{\tau}(g) - \frac{1}{2}) P(g \mid ASP)$$

where $g$ spans all possible multipoint genotype configurations,

$$\sum_g (\bar{X}_{\tau}(g) - \frac{1}{2}) \sum_{l=0,\frac{1}{2},1} P(g, \pi_\tau = l \mid ASP)$$

$$\sum_g (\bar{X}_{\tau}(g) - \frac{1}{2}) \sum_{l=0,\frac{1}{2},1} P(g \mid \pi_\tau = l, ASP) P(\pi_\tau = l \mid ASP)$$

$$\sum_g (\bar{X}_{\tau}(g) - \frac{1}{2}) \sum_{l=0,\frac{1}{2},1} P_0(\pi_\tau = l) P(\pi_\tau = l \mid ASP)$$

since markers are in full linkage equilibrium with true locus,

$$\sum_g (\bar{X}_{\tau}(g) - \frac{1}{2}) \sum_{l=0,\frac{1}{2},1} \frac{P_0(\pi_\tau = l \mid g)}{P_0(\pi_\tau = l)} P_0(g) P(\pi_\tau = l \mid ASP).$$

Now replacing the probabilities for unobserved IBD sharing $P(\pi_\tau = l \mid ASP)$ by their values under the additive model introduced above and bearing in mind that $\bar{X}_{\tau} - \frac{1}{2} = \frac{1}{2}[P_0(\pi_\tau = 1 \mid g) - P_0(\pi_\tau = 0 \mid g)]$, it is straightforward to show that

$$E(\bar{X}_{\tau} - \frac{1}{2} \mid ASP) = \sum_g (\bar{X}_{\tau} - \frac{1}{2}) P_0(g) + C \sum_g (\bar{X}_{\tau} - \frac{1}{2})(\bar{X}_{\tau} - \frac{1}{2}) P_0(g)$$

$$= 0 + \text{cov}_0(\bar{X}_{\tau}, \bar{X}_{\tau}) C.$$

Consistency of score test

We prove here the consistency of the score test in the estimation of the locus position under an additive model. Let us consider $Y_\tau = \text{var}_0(\bar{X}_{\tau})^{-1/2} (\bar{X}_{\tau} - \frac{1}{2})$ then

$$E(Y_\tau) = \text{var}_0(\bar{X}_{\tau})^{-1/2} E(\bar{X}_{\tau} - \frac{1}{2})$$

$$= \text{var}_0(\bar{X}_{\tau})^{-1/2} \text{cov}_0(\bar{X}_{\tau}, \bar{X}_{\tau}) C$$

$$= \text{cor}_0(\bar{X}_{\tau}, \bar{X}_{\tau}) \text{var}_0(\bar{X}_{\tau})^{1/2} C$$

$$= \text{cor}_0(\bar{X}_{\tau}, \bar{X}_{\tau}) \text{var}_0(\bar{X}_{\tau})^{1/2} E(\bar{X}_{\tau} - \frac{1}{2})$$

$$< E(Y_\tau) \text{ for } t \neq \tau$$

Since $\text{cor}_0(\bar{X}_{\tau}, \bar{X}_{\tau})$ is strictly monotonic in $t$, $Y_\tau - Y_t$ has a strictly positive mean $\mu$ and finite variance $\sigma^2$. By the Central Limit Theorem, we then have that the sequence $(Z_\tau - Z_t)(N) = N^{-1/2}(Y_\tau - Y_t)(N)$ converges in distribution to $N(N^{1/2}\mu, \sigma^2)$ thus
\( P(Z_t(N) < Z_\tau(N)) \to 1 \) as \( N \to +\infty \) for all \( t \neq \tau \). This proves the consistency of the estimate of locus position \( t(N) \) taken such that \( Z_{t(N)} = \sup_t Z_t(N) \).