QTL Mapping under Ascertainment

Jie Peng

Department of Statistics, University of California, Davis
Outline

- Variance-components model and efficient scores
- Sampling schemes and ascertainment corrections
- Robustness and CMLE
- Ongoing work
QTL Mapping and Linkage Analysis

- Goal of genetic mapping: locate genes that affect particular traits.
- Quantitative traits in human: body weight; blood pressure; the level of cholesterol, etc.
- Family-based linkage analysis: phenotype and genotype family members, and then study the correlation between trait values (*phenotypes*) and genetic markers (*IBD*).
- Basic idea: relatives with similar phenotypes should share more genetic material than expected near the trait gene(s).
- Variance-components methods: model phenotypes by variance components arising from QTLs, polygenic effects, environmental effects, etc. (Almasy and Blangero (1998))
Variance-components Model

- Single locus model

\[ Y = \mu + \alpha_m + \alpha_f + \delta_{m,f} + e. \] (1)

- Covariance for sibpairs,

\[ \text{Cov}(Y_1, Y_2|\nu(\tau)) = \text{Cov}(Y_1, Y_2) + \alpha_0(\nu(\tau) - 1) + \delta_0(1/2 - I\{\nu(\tau) = 1\}), \]

where

\[ \text{Cov}(Y_1, Y_2) = \frac{1}{2} \sigma_A^2 + \frac{1}{4} \sigma_D^2 + r\sigma_e^2, \]

\[ \alpha_0 = \frac{\sigma_A^2 + \sigma_D^2}{2}, \quad \delta_0 = \frac{\sigma_D^2}{2}. \]

- Null hypothesis of no linkage: \( \alpha_0 = 0. \)
• Normality assumption

\[ \mathbf{Y} | A_{\nu(\tau)} \sim \mathcal{N}_s(\mu_1, \Sigma_{\nu(\tau)}), \quad \Sigma_{\nu(\tau)} = \Sigma + \alpha_0 A_{\nu(\tau)}, \]

where \( \Sigma = \sigma^2_Y ((1 - \rho) I + \rho \mathbf{1} \mathbf{1}' ), \ A_{\nu}(i, j) = (\nu_{ij} - 1) I(i = j). \)

• Efficient scores under random sampling. At a putative trait locus \( t \) (Tang and Siegmund (2001)),

\[
\ell_\alpha(t) = \frac{1}{2} \sum_n \left\{ - \text{tr} \left( \Sigma^{-1} A_{\hat{\nu}(t)} \right) + \text{tr} \left( \Sigma^{-1} A_{\hat{\nu}(t)} \Sigma^{-1} (\mathbf{Y} - \mu_1)(\mathbf{Y} - \mu_1)^T \right) \right\},
\]

\[
\ell_\theta(t) = \frac{1}{2} \sum_n \left\{ - \text{tr} \left( \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta} \right) + \text{tr} \left( \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta} \Sigma^{-1} (\mathbf{Y} - \mu_1)(\mathbf{Y} - \mu_1)^T \right) \right\}.
\]

Linkage parameters and segregation parameters are orthogonal.

\( \hat{\nu}(t) = \mathbb{E}(\nu(t) | M) \), where M is genotype data.
Sampling Schemes and Ascertainment

- Random sampling.
- Selective genotyping.
- Ascertainment: pedigrees ascertained through a proband set.
- Comparison: ascertainment costs less in recruiting, phenotyping and genotyping and sometimes is inevitable. However, for ascertainment, we may not have consistent estimators of segregation parameters.
• Conditional likelihood $L$:

$$L_A = P(Y, M|A), \quad L_V = P(Y, M|Y(1)).$$

• Proposition 1: under the normality assumption,

$$\frac{\partial \log(L)}{\partial \alpha_0}_{\alpha_0=0} = \ell_\alpha I(A).$$

• Proposition 2: under the normality assumption, for the conditional likelihood $L$, the linkage parameters are orthogonal to the segregation parameters under the null hypothesis of no linkage.

• Two ascertainment corrections
  
  – Elston and Sobel (1979): condition on the event that a pedigree is ascertained. $L_A$ not always calculable.
  
  – Hopper and Mathews (1982): condition on the phenotypic value(s) of the proband(s). $L_V$ not always valid.
  
  – Proposition 3: under the normality assumption, these two corrections are asymptotically equivalent.
Robust Score Statistics

- At a putative trait locus \( t \)

\[
Z(t) = \frac{l_\alpha(t)}{\sqrt{E_0[\ell^2_\alpha(t)|Y]}}
\]  

(2)

- Robust in validity. Conservative p-values are given by (Feingold et al. (1993))

\[
P_0(\max Z_{i\Delta} > b) \sim 1 - \Phi(b) + \beta Lb\phi(b)\nu((2\beta\Delta)^{1/2}).
\]

- Robust in power. Theorem 1: If

1. \( \exists \theta^* \in \Omega_\theta, \) such that \( \hat{\theta}_N \) is \( \sqrt{N} \)-consistent to \( \theta^* \) under \( H_0 \), and

2. the family of distributions of \( (Y, M)|Y(1) \in S \) is quadratic mean differentiable around \( (\alpha_0 = 0, \theta) (\theta \in \Omega_\theta) \),

then under the local alternatives \( \alpha_N = a/\sqrt{N} \), the robust score statistic \( Z_t \)

(2) satisfies,

\[
Z_t(\theta^*) \Rightarrow F, \quad Z_t(\hat{\theta}_N) \Rightarrow F.
\]
• Estimation of nuisance parameters.
  – Random sampling or selective genotyping: sample estimators (Theorem 1).
  – Ascertainment: conditional MLE (CMLE) based on the normal model and Hopper-Mathews’ ascertainment correction. (Biased estimation could be due to non-normality and/or ill-defined ascertainment criterion.)

• For ascertainment rules involving all the members. Theorem 2: when $b \to \infty$,

$$P(Y_1 = b + dy_1, Y_2 = dy_2, \ldots, Y_n = dy_n | Y_1 > b)$$

$$\sim P(Z_1 = b + dy_1, Z_2 = dy_2, \ldots, Z_n = dy_n | \max(Z) > b).$$
Effects of Misspecification

- Relative efficiency:

\[ R(\hat{\theta}, s, T) = \xi^2(\hat{\theta})/\xi^2(\theta), \quad \xi = \mathbb{E}(Z(\tau)|A). \]

- The estimate of the phenotypic variance \( \sigma^2 \) has little effect on the noncentrality parameter \( \xi \). The estimates of the phenotypic mean \( \mu \) and correlation \( \rho \) have bigger effects on \( \xi \). (Figure 1, Figure 2)

- Larger sibships are more robust to misspecifications.

- More stringent ascertainment sampling is more robust to misspecifications.
Figure 1: Contour plot of $R(\hat{\theta}, 4, 1.28) (S = \{Y_1 > T\})$.

\[ \hat{\sigma} \text{ is fixed at the true value } \sigma = 1. \]
Figure 2: Contour plot of \( \frac{R(\hat{\theta}_1, 4, 1.28)}{R(\hat{\theta}_2, 4, 1.28)} \) (\( \hat{\sigma}_1 = 1 \), \( \hat{\sigma}_2 = 0.5 \)).
Two assumptions for CMLE: (i) normality; (ii) conditional likelihood

\[ L_V = P(\mathbf{Y}, M|Y_1) \] (Theorem 2).

Simulation study

- Three ascertainment procedures: (i) A-first: \( S = \{Y_1 > T\} \); (ii) A-max: \( S = \{\max \mathbf{Y} > T\} \); (iii) A-sample: sample from listed members.

- Three phenotypic distributions: (i) multi-Normal; (ii) multi-T; (iii) multi-Gamma.

- Little loss of power by using CMLE when ascertainment rule is correctly specified (A-first) or stringent (A-sample) for all three traits. Small amount loss of power when ascertainment rule is ill-defined (A-max). Similar results hold for the bi-allelic model. (Table 1)


Table 1: Power of the robust score statistic at the 0.05 genome-wide significance level. The linkage parameters are $\alpha_0 = 0.1$, $\delta_0 = 0$. 400 sibtrios are ascertained. The multi-$T$ trait has kurtosis 0.375, and the multi-Gamma trait has skewness 0.179. There are 31 equally spaced, fully informative markers with spacing $\Delta = 5\text{cM}$, and the trait locus is on the 16th marker.

<table>
<thead>
<tr>
<th>Model</th>
<th>Estimates</th>
<th>A-first</th>
<th></th>
<th></th>
<th>A-max</th>
<th></th>
<th></th>
<th>A-sample</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>75%</td>
<td>90%</td>
<td>95%</td>
<td>75%</td>
<td>90%</td>
<td>95%</td>
<td>75%</td>
<td>90%</td>
<td>95%</td>
</tr>
<tr>
<td>Multi-N</td>
<td>True</td>
<td>0.395</td>
<td>0.678</td>
<td>0.848</td>
<td>0.383</td>
<td>0.632</td>
<td>0.811</td>
<td>0.388</td>
<td>0.639</td>
<td>0.813</td>
</tr>
<tr>
<td></td>
<td>CMLE</td>
<td>0.395</td>
<td>0.674</td>
<td>0.847</td>
<td>0.354</td>
<td>0.616</td>
<td>0.799</td>
<td>0.389</td>
<td>0.636</td>
<td>0.813</td>
</tr>
<tr>
<td></td>
<td>MLE</td>
<td>0.230</td>
<td>0.370</td>
<td>0.520</td>
<td>0.260</td>
<td>0.396</td>
<td>0.528</td>
<td>0.212</td>
<td>0.346</td>
<td>0.464</td>
</tr>
<tr>
<td>Multi-G</td>
<td>True</td>
<td>0.386</td>
<td>0.684</td>
<td>0.840</td>
<td>0.359</td>
<td>0.623</td>
<td>0.783</td>
<td>0.366</td>
<td>0.633</td>
<td>0.799</td>
</tr>
<tr>
<td></td>
<td>CMLE</td>
<td>0.382</td>
<td>0.683</td>
<td>0.835</td>
<td>0.330</td>
<td>0.603</td>
<td>0.775</td>
<td>0.363</td>
<td>0.628</td>
<td>0.796</td>
</tr>
<tr>
<td></td>
<td>MLE</td>
<td>0.218</td>
<td>0.401</td>
<td>0.534</td>
<td>0.235</td>
<td>0.405</td>
<td>0.500</td>
<td>0.215</td>
<td>0.367</td>
<td>0.487</td>
</tr>
<tr>
<td>Multi-T</td>
<td>True</td>
<td>0.322</td>
<td>0.602</td>
<td>0.752</td>
<td>0.318</td>
<td>0.557</td>
<td>0.701</td>
<td>0.312</td>
<td>0.557</td>
<td>0.710</td>
</tr>
<tr>
<td></td>
<td>CMLE</td>
<td>0.322</td>
<td>0.597</td>
<td>0.746</td>
<td>0.299</td>
<td>0.540</td>
<td>0.694</td>
<td>0.308</td>
<td>0.552</td>
<td>0.708</td>
</tr>
<tr>
<td></td>
<td>MLE</td>
<td>0.167</td>
<td>0.314</td>
<td>0.421</td>
<td>0.197</td>
<td>0.340</td>
<td>0.429</td>
<td>0.179</td>
<td>0.315</td>
<td>0.394</td>
</tr>
</tbody>
</table>
Ongoing Work

- For extremely non-normal traits: data transformation is necessary. (i) CMLE does not have a big impact on power; (ii) multi-T assumption.
- Estimation of the robust information at marker $t$:
  \[
  I^R_{\alpha \alpha} = E(\ell^2_{\alpha}(t)|A).
  \]

(joint work with David Siegmund)