Appendix

1. The EM Algorithm for the Mixture Model

Below we give the details of the statistical model and the EM steps for parameter estimation. (For simplicity of indexing of mathematical expressions, we assume the existence of with value 0.)

1. Statistical Model
2. Let  be a sequence of observed symptom measurements in subject *i*; and denote the genotypes or other covariates (including constant 1) for subject *i*.
3. Assume that drug effect may occur at time = *k* or *k*+1(<*T*) for subject *i*, which is unknown. We assume follow a Bernoulli distribution with . We link the parameter  with any genotype or covariate by, where is the coefficient vector.
4. An AR(1) model is used to model the symptoms during the latency period, up to and including the change point , when measurements are assumed to be stationary, and follows an AR(1) process  with parameter where is a white noise process with variance . We link the parameter with any genotype or covariate by , where is the coefficient vector. (Here, the mean of the AR(1) process is assumed to be zero.)
5. A linear trend model is used to model the increase of symptom measurements occurring after the change point , orfollows a linear trend model with parameter where is the error term with variance . We link the parameter with any genotype or covariate by , where is the coefficient vector.
6. The Likelihood

when,  or;

when, or.

The data likelihood of sequence *i* is

 (A.2.1)

And thus the complete data likelihood is



It is hard to estimate the MLE ofdirectly by maximizing above likelihood function due to the existence of the missing data . For this reason, we propose an EM based iterative algorithm to estimate the local MLE of .

1. EM Estimation Procedure

Assume at step *n*, the estimated value of parameter is , then at step (*n*+1),



Below, we give the details of E-step and M-step.

* E-step

We denoted and , calculated as below,

 

where  and.

Then



* M-step,

To find we need to maximize above likelihood summed over all the sequences. Below gives the details for estimating each individual parameter.

1. To find , we maximize 

This expression corresponds to a weighted logistic regression. We use the *Newton-Raphson* algorithm to calculatewhich maximize above expression.

1. To estimate , maximize 

Let

Let  and, where and 

This is a weighted linear regression and the solution is.

1. To estimate , maximize 

Let  and, where and 

This is also a weighted linear regression and the solution is .

1. After estimating  and,  can be calculated by



The EM step iterates until convergence or until the difference between the estimated parameter values at steps *n* and *n*+1 are below a pre-specified threshold value.

1. Information Matrix Calculation

Below we give the details of the calculation of the information matrix using Louis (1982)’ method.

1. First and Second Derivatives of The Log Likelihood of Complete Data

Letdenote the complete data including the observed sequence  and the missing onset , then



And



The first and second derivatives for are given below.









Let , the complete data including the observed sample and missing data , then the gradient and Hessian matrix for the complete data log likelihood can be obtained from above results.

1. The GradientAnd 



Now, we calculate  of each estimator. Below formula facilitate the calculation.

(B.b.1)

where . In our method, this is the conditional probability 

For example, to calculate , take,

. To calculate, take

,,,

. Then utilizing formula (B.b.1) with  and as MLE from the EM algorithm, we can get the estimated value of each element of .

1. The Hessian Matrix and 

The Hessian matrixcontains all the second derivatives of the log likelihood and is a symmetric matrix. Below we give its elements on the upper right.



Let , then



The difference of and , gives the asymptotic estimation of the information matrix , after substituting each parameter value by the estimated parameter value from the EM algorithm.

1. Log Likelihood Ratio Test

To evaluate whether a genetic variant or environmental covariate significantly affected the distribution of the onset of drug effect, we applied the classic Log Likelihood Ratio Test. Let

 and be the estimated parameter values from the EM algorithm under the hypothesis-reduced model H0 and full model H1. The likelihood ratio test statistic is then where and  is given in equation (A.2.1). And



1. Tables of Some Simulation Results

The tables below contain results from simulation study. For each simulation scenario, we used 500 subjects for each run (unless specified) and conducted 1000 independent runs. The mean, SD, and prediction rate (proportion of onset times that were correctly predicted) were calculated from results of the 1000 runs.

1. Effects of different ** values (with corresponding odds ratios when SNP value is 2) on estimation accuracy

**Table A1.** Estimation results for different**values in *Scenario* 1

|  |  |  |  |
| --- | --- | --- | --- |
| odds ratio | True Value of ** | Mean of Estimated ** | SD of Estimated ** |
| 1.1 | (-3.080, 1.588) | (-3.091, 1.59) | (0.301, 0.2028) |
| 1.2 | (-3.080, 1.631) | (-3.103, 1.641) | (0.3147, 0.2086) |
| 1.3 | (-3.080, 1.671) | (-3.125, 1.693) | (0.304, 0.2102) |
| 1.4 | (-3.080, 1.708) | (-3.104, 1.713) | (0.3197, 0.2215) |
| 1.5 | (-3.080, 1.743) | (-3.093, 1.751) | (0.3087, 0.212) |
| 2 | (-3.080, 1.887) | (-3.115, 1.905) | (0.2895, 0.2139) |
| 4 | (-3.080, 2.233) | (-3.116, 2.255) | (0.2978, 0.2169) |

**Table A2.** Estimation results for different**values in *Scenario* 2

|  |  |  |  |
| --- | --- | --- | --- |
| odds ratio | True Value of ** | Mean of Estimated ** | SD of Estimated ** |
| 1.1 | (-3.080, 2.088, -0.200) | (-3.123, 2.111, -0.202) | (0.4326, 0.2207, 0.0598) |
| 1.2 | (-3.080, 2.131, -0.200) | (-3.114, 2.152, -0.201) | (0.4269, 0.2142, 0.0608) |
| 1.3 | (-3.080, 2.171, -0.200) | (-3.117, 2.187, -0.199) | (0.4174, 0.2175, 0.0608) |
| 1.4 | (-3.080, 2.208, -0.200) | (-3.145, 2.237, -0.199) | (0.4169, 0.2174, 0.0594) |
| 1.5 | (-3.080, 2.243, -0.200) | (-3.112, 2.272, -0.204) | (0.4281, 0.2125, 0.0615) |
| 2 | (-3.080, 2.387, -0.200) | (-3.136, 2.426, -0.201) | (0.4058, 0.2135, 0.0583) |
| 4 | (-3.080, 2.733, -0.200) | (-3.105, 2.766, -0.204) | (0.4584, 0.2387, 0.0711) |

**Table A3.** More estimation results for different**values in *Scenario* 2

|  |  |  |  |
| --- | --- | --- | --- |
| odds ratio | True Value of ** | Mean of Estimated ** | SD of Estimated ** |
| 1.1 | (-3.080, 4.230, -1.057) | (-3.201, 4.346, -1.077) | (0.615, 0.4603, 0.1261) |
| 1.2 | (-3.080, 4.230, -1.040) | (-3.16, 4.317, -1.061) | (0.5996, 0.4676, 0.1275) |
| 1.3 | (-3.080, 4.230, -1.024) | (-3.14, 4.284, -1.036) | (0.595, 0.436, 0.1178) |
| 1.4 | (-3.080, 4.230, -1.009) | (-3.152, 4.322, -1.03) | (0.582, 0.424, 0.1155) |
| 1.5 | (-3.080, 4.230, -0.995) | (-3.141, 4.299, -1.011) | (0.5751, 0.4193, 0.1149) |
| 2 | (-3.080, 4.230, -0.937) | (-3.161, 4.309, -0.951) | (0.5529, 0.421, 0.1117) |
| 4 | (-3.080, 4.230, -0.799) | (-3.089, 4.287, -0.816) | (0.4718, 0.3998, 0.1102) |

1. Effects of sample size on estimation accuracy

**Table A4.** Estimation results for different sample size in *Scenario* 1

|  |  |  |  |
| --- | --- | --- | --- |
| Sample Size | True Value of ** | Mean of Estimated ** | SD of Estimated ** |
| 50 | (-3.08, 2.23) | (-3.254, 2.363) | (0.9515, 0.7093) |
| 60 | (-3.08, 2.23) | (-3.333, 2.421) | (0.8971, 0.6457)  |
| 70 | (-3.08, 2.23) | (-3.214, 2.326) | (0.857, 0.6034)  |
| 80 | (-3.08, 2.23) | (-3.239, 2.354) | (0.7859, 0.5473) |
| 100 | (-3.08, 2.23) | (-3.191, 2.303) | (0.7000, 0.5228) |
| 200 | (-3.08, 2.23) | (-3.125, 2.260) | (0.4887, 0.3458) |
| 300 | (-3.08, 2.23) | (-3.125, 2.268) | (0.3986, 0.3115) |
| 500 | (-3.08, 2.23) | (-3.106, 2.249) | (0.2988, 0.2415) |
| 800 | (-3.08, 2.23) | (-3.099, 2.246) | (0.2293, 0.1823) |
| 1000 | (-3.08, 2.23) | (-3.089, 2.234) | (0.2138, 0.1976) |

**Table A5.** Estimation results for different sample size in *Scenario* 2

|  |  |  |  |
| --- | --- | --- | --- |
| Sample Size | True Value of ** | Mean of Estimated ** | SD of Estimated ** |
| 50 | (-3.08, 4.23, -0.20) | (-1.935, 3.254, -0.236) | (1.5781, 1.446, 0.2283)  |
| 60 | (-3.08, 4.23, -0.20) | (-2.221, 3.505, -0.223)  | (1.1664, 0.4207, 0.2267)  |
| 70 | (-3.08, 4.23, -0.20) | (-2.194, 3.352, -0.199)  | (0.9078, 0.4003, 0.1729) |
| 80 | (-3.08, 4.23, -0.20) | (-2.411, 3.598, -0.208) | (0.8976, 0.4003, 0.1817) |
| 100 | (-3.08, 4.23, -0.20) | (-2.645, 3.842, -0.209) | (0.7825, 0.427, 0.1494)  |
| 200 | (-3.08, 4.23, -0.20) | (-2.911, 4.096, -0.209) | (0.6619, 0.497, 0.0966) |
| 300 | (-3.08, 4.23, -0.20) | (-3.106, 4.287, -0.206) | (0.6524, 0.5505, 0.0741) |
| 400 | (-3.08, 4.23, -0.20) | (-3.22, 4.38, -0.202)  | (0.6442, 0.5469, 0.072) |
| 500 | (-3.08, 4.23, -0.20) | (-3.187, 4.344, -0.201) | (0.5779, 0.5093, 0.0695) |
| 800 | (-3.08, 4.23, -0.20) | (-3.21, 4.35, -0.199) | (0.5359, 0.456, 0.0501) |
| 1000 | (-3.08, 4.23, -0.20) | (-3.183, 4.322, -0.198) | (0.4651, 0.3899, 0.045) |

1. Joint effects of sequence length and sample size on estimation accuracy

**Table A6.** Estimation results for different sample size in *Scenario* 1 when

sequence length T=12, early onset k=6

|  |  |  |  |
| --- | --- | --- | --- |
| Sample Size | True Value of ** | Mean of Estimated ** | SD of Estimated ** |
| 50 | (-3.08, 2.23) | (-3.254, 2.363) | (0.9515, 0.7093) |
| 60 | (-3.08, 2.23) | (-3.333, 2.421)  | (0.8971, 0.6457)  |
| 70 | (-3.08, 2.23) | (-3.214, 2.326) | (0.857, 0.6034)  |
| 80 | (-3.08, 2.23) | (-3.239, 2.354) | (0.7859, 0.5473) |
| 100 | (-3.08, 2.23) | (-3.191, 2.303) | (0.7000, 0.5228) |
| 200 | (-3.08, 2.23) | (-3.125, 2.260) | (0.4887, 0.3458) |
| 300 | (-3.08, 2.23) | (-3.125, 2.268) | (0.3986, 0.3115) |
| 500 | (-3.08, 2.23) | (-3.106, 2.249) | (0.2988, 0.2415) |
| 800 | (-3.08, 2.23) | (-3.099, 2.246) | (0.2293, 0.1823) |
| 1000 | (-3.08, 2.23) | (-3.089, 2.234) | (0.2138, 0.1976) |

**Table A7.** Estimation results for different sample size in *Scenario* 1 when

sequence length T=10, early onset k=5

|  |  |  |  |
| --- | --- | --- | --- |
| Sample Size | True Value of ** | Mean of Estimated ** | SD of Estimated ** |
| 50 | (-3.08, 2.23) | (-3.32, 2.412) | (0.9531, 0.7238)  |
| 60 | (-3.08, 2.23) | (-3.294, 2.381)  | (0.9264, 0.6575)  |
| 70 | (-3.08, 2.23) | (-3.212, 2.328)  | (0.837, 0.5915) |
| 80 | (-3.08, 2.23) | (-3.229, 2.356)  | (0.8084, 0.5966) |
| 100 | (-3.08, 2.23) | (-3.215, 2.316) | (0.7548, 0.4889) |
| 200 | (-3.08, 2.23) | (-3.113, 2.258) | (0.4816, 0.3636) |
| 300 | (-3.08, 2.23) | (-3.148, 2.285) | (0.377, 0.286) |
| 500 | (-3.08, 2.23) | (-3.088, 2.23) | (0.2925, 0.2475) |
| 800 | (-3.08, 2.23) | (-3.096, 2.24) | (0.2296, 0.1939) |
| 1000 | (-3.08, 2.23) | (-3.094, 2.242) | (0.2076, 0.1743)  |

**Table A8.** Estimation results for different sample size in *Scenario* 1 when

sequence length T=8, early onset k=4

|  |  |  |  |
| --- | --- | --- | --- |
| Sample Size | True Value of ** | Mean of Estimated ** | SD of Estimated ** |
| 50 | (-3.08, 2.23) | (-3.29, 2.392) | (0.9837, 0.786) |
| 60 | (-3.08, 2.23) | (-3.238, 2.341) | (0.8542, 0.6605) |
| 70 | (-3.08, 2.23) | (-3.248, 2.361) | (0.8445, 0.6112) |
| 80 | (-3.08, 2.23) | (-3.279, 2.375) | (0.8163, 0.5938) |
| 100 | (-3.08, 2.23) | (-3.232, 2.342) | (0.6765, 0.5056) |
| 200 | (-3.08, 2.23) | (-3.17, 2.304)  | (0.4596, 0.3446)  |
| 300 | (-3.08, 2.23) | (-3.145, 2.279) | (0.401, 0.3014) |
| 500 | (-3.08, 2.23) | (-3.106, 2.253) | (0.3072, 0.2524) |
| 800 | (-3.08, 2.23) | (-3.087, 2.227) | (0.2323, 0.2084) |
| 1000 | (-3.08, 2.23) | (-3.105, 2.24) | (0.2195, 0.2218) |

**Table A9.** Estimation results for different sample size in *Scenario* 1 when

sequence length T=6, early onset k=3

|  |  |  |  |
| --- | --- | --- | --- |
| Sample Size | True Value of ** | Mean of Estimated ** | SD of Estimated ** |
| 50 | (-3.08, 2.23) | (-3.37, 2.414)  | (0.9925, 0.7267)  |
| 60 | (-3.08, 2.23) | (-3.255, 2.356) | (0.9334, 0.7048)  |
| 70 | (-3.08, 2.23) | (-3.241, 2.362) | (0.8833, 0.6443) |
| 80 | (-3.08, 2.23) | (-3.246, 2.354) | (0.7951, 0.5562) |
| 100 | (-3.08, 2.23) | (-3.233, 2.336) | (0.6580, 0.5104)  |
| 200 | (-3.08, 2.23) | (-3.144, 2.283) | (0.4916, 0.3800)  |
| 300 | (-3.08, 2.23) | (-3.135, 2.263) | (0.3886, 0.3031) |
| 500 | (-3.08, 2.23) | (-3.099, 2.249)  | (0.3000, 0.2594) |
| 800 | (-3.08, 2.23) | (-3.095, 2.237) | (0.2359, 0.205)  |
| 1000 | (-3.08, 2.23) | (-3.103, 2.239) | (0.2099, 0.2253) |

1. Effects of post-onset slope on estimation accuracy

**Table A10.** Effects of post-onset slope (γ) on estimation and prediction accuracy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| True Value of ** | True Value of ** | Mean of Estimated ** | SD of Estimated ** | Prediction Rate |
| (0.40, -0.06) | (-3.08, 2.23) | (-3.074, 2.192) | (0.302, 0.242) | 0.776 |
| (0.30, -0.045) | (-3.08, 2.23) | (-3.044, 2.136) | (0.308, 0.305) | 0.772 |
| (0.20, -0.03)  | (-3.08, 2.23) | (-3.169, 2.129) | (0.791, 0.523) | 0.751 |
| (0.10, -0.015) | (-3.08, 2.23) | (-3.370, 2.930) | (1.110, 2.589) | 0.698 |

1. Effects of data variance on estimation accuracy

**Table A11.** Effects of data variance (σ2) on estimation and prediction accuracy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| True Value of ** | True Value of ** | Mean of Estimated ** | SD of Estimated ** | Prediction Rate |
| 0.01 | (-3.08, 2.23) | (-3.095, 2.245) | (0.294, 0.212) | 0.7777 |
| 0.05 | (-3.08, 2.23) | (-3.169, 2.129) | (0.791, 0.523) | 0.751 |
| 0.10 | (-3.08, 2.23) | (-3.541, 3.036) | (1.370, 2.651) | 0.699 |
| 0.15 | (-3.08, 2.23) | (-3.696, 3.988) | (2.380, 3.649) | 0.678 |