生物医疗中的统计与方法-小组作业 2

1. 背景介绍

复杂的疾病往往会具有多基因遗传结构，而有时因为一些单个遗传变异点的效应较弱，且样本数量有限，导致这样的弱效应会被淹没在遗传数据的背景噪音中无法识别。考虑到基因具有多效性，我们在当前所研究疾病 1 的 GWAS 数据的基础上，联合另一疾病 2 的 GWAS 数据进行整合分析，采用 EM 算法来推断哪些 SNP 与疾病 1 和疾病 2 相关。

其中，EM 算法的主要思路是：首先将观测到的 GWAS 数据记为P，模型参数记为θ，引入代表 SNP 是否与疾病有关的潜变量Z。基于条件分布$Pr(P|Z,\theta)$和潜变量的先验分布$\Pr(Z|\theta)$可以得到完整数据的联合分布$Pr(P,Z|\theta) = Pr(P|Z,\theta)\Pr(Z|\theta)$，但完整数据的对数似然函数中含有潜变量，无法通过极大化该函数来估计我们感兴趣的参数。于是先基于观测到的不完整数据P和上一步估计出的参数值$\theta^{(t-1)}$，得到潜变量的后验分布$\Pr\left(Z\middle|P,\theta^{(t-1)}\right)$，基于此可以得到完整数据的对数似然函数的条件均值$E[lnPr(P,Z|\theta)]$，这是 EM 算法的 E 步。接着极大化$E[lnPr(P,Z|\theta)]$，得到参数估计$\theta^{(t)}$，这是 EM 算法的 M 步。对进行 E 步和 M 步反复迭代直至参数估计变化在既定的界限内，认为算法收敛，得到最终的参数估计。

2. EM 算法推导

具体推导过程如下所示：

引入潜变量$Z_i = \{Z_{i00}, Z_{i10}, Z_{i01}, Z_{i11}\}, i = 1,2\cdots M$，其中$Z_{i00}, Z_{i10}, Z_{i01}, Z_{i11} \in \{0,1\}$代表对应 SNP 的来源，有$\pi_{00} + \pi_{10} + \pi_{01} + \pi_{11} = 1$，且

$\Pr(Z_{i00} = 1) = \pi_{00}$:

$$\Pr(P_{i1}|Z_{i00} = 1) = 1, \quad \Pr(P_{i2}|Z_{i00} = 1) = 1$$

$\Pr(Z_{i10} = 1) = \pi_{10}$:

$$\Pr(P_{i1}|Z_{i10} = 1) = \alpha_1 {P_{i1}}^{\alpha_1 - 1}, \quad \Pr(P_{i2}|Z_{i10} = 1) = 1$$

$\Pr(Z_{i01} = 1) = \pi_{01}$:

$$\Pr(P_{i1}|Z_{i01} = 1) = 1, \quad \Pr(P_{i2}|Z_{i01} = 1) = \alpha_2 {P_{i2}}^{\alpha_2 - 1}$$

$\Pr(Z_{i11} = 1) = \pi_{11}$:

$$\Pr(P_{i1}|Z_{i11} = 1) = \alpha_1 {P_{i1}}^{\alpha_1 - 1}, \quad \Pr(P_{i2}|Z_{i11} = 1) = \alpha_2 {P_{i2}}^{\alpha_2 - 1}$$

假设P_1与P_2相互独立，则混合模型的密度函数可以表示为：

$$\Pr(P_{i1}, P_{i2}; \theta) = \pi_{00} + \pi_{10}\alpha_1 {P_{i1}}^{\alpha_1 - 1} + \pi_{01}\alpha_2 {P_{i2}}^{\alpha_2 - 1} + \pi_{11}\alpha_1\alpha_2 {P_{i1}}^{\alpha_1 - 1} {P_{i2}}^{\alpha_2 - 1}$$

易得观测数据的对数似然函数为：

$$ln\Pr(P_1,P_2;\theta)=\sum_{i=1}^{M}ln\Pr(P_{i1},P_{i2};\theta)$$
$$=\sum_{i=1}^{M}ln\left[\pi_{00}+\pi_{10}\alpha_1{P_{i1}}^{\alpha_1-1}+\pi_{01}\alpha_2{P_{i2}}^{\alpha_2-1}+\pi_{11}\alpha_1\alpha_2{P_{i1}}^{\alpha_1-1}{P_{i2}}^{\alpha_2-1}\right]$$

将 Z 看作未观测到的隐含数据，使用 EM 算法进行参数估计。由模型可得先验分布和条件分布分别为：

$$\Pr(Z_i|\theta)={\pi_{00}}^{Z_{i00}}{\pi_{10}}^{Z_{i10}}{\pi_{01}}^{Z_{i01}}{\pi_{11}}^{Z_{i11}}$$
$$\Pr(P_{i1},P_{i2}|Z;\theta)=(\alpha_1{P_{i1}}^{\alpha_1-1})^{Z_{i10}}(\alpha_2{P_{i2}}^{\alpha_2-1})^{Z_{i01}}(\alpha_1\alpha_2{P_{i1}}^{\alpha_1-1}{P_{i2}}^{\alpha_2-1})^{Z_{i11}}$$

所以完整数据的对数似然函数为：

$$lnPr(P,Z;\theta)=\sum_{i=1}^{M}(ln\Pr(P_{i1},P_{i2}|Z;\theta)+ln\Pr(Z_i|\theta))$$
$$=\sum_{i=1}^{M}(Z_{i10}\,ln\left(\alpha_1{P_{i1}}^{\alpha_1-1}\pi_{10}\right)+Z_{i01}\,ln\left(\alpha_2{P_{i2}}^{\alpha_2-1}\pi_{01}\right)$$
$$+Z_{i11}l\,n\left(\alpha_1\alpha_2{P_{i1}}^{\alpha_1-1}{P_{i2}}^{\alpha_2-1}\pi_{11}\right)+Z_{i00}ln\pi_{00})$$

首先初始化参数$\theta^{(0)}$，计算初始对数似然$lnPr\left(P_1,P_2\middle|\theta^{(0)}\right)$。然后在 E 步中，计算完整数据对数似然函数的条件均值：

$$Q\left(\theta,\theta^{(t-1)}\right)=E[lnPr(P,Z;\theta)]$$
$$=\sum_{i=1}^{M}\left[E\left(Z_{i10}\middle|P;\theta^{(t-1)}\right)ln\left(\alpha_1{P_{i1}}^{\alpha_1-1}\pi_{10}\right)+E\left(Z_{i01}\middle|P;\theta^{(t-1)}\right)ln\left(\alpha_2{P_{i2}}^{\alpha_2-1}\pi_{01}\right)\right.$$
$$\left.+E\left(Z_{i11}\middle|P;\theta^{(t-1)}\right)l\,n\left(\alpha_1\alpha_2{P_{i1}}^{\alpha_1-1}{P_{i2}}^{\alpha_2-1}\pi_{11}\right)+E\left(Z_{i00}\middle|P;\theta^{(t-1)}\right)ln\pi_{00}\right]$$

其中

$$E\left(Z_{i00}\middle|P;\theta^{(t-1)}\right)=\Pr\left(Z_{i00}=1\middle|P;\theta^{(t-1)}\right)$$
$$=\frac{\Pr\left(P\middle|Z_{i00}=1;\theta^{(t-1)}\right)\Pr\left(Z_{i00}=1\middle|\theta^{(t-1)}\right)}{denominator}$$
$$=\frac{{\pi_{00}}^{(t-1)}}{denominator}$$
$$=\gamma(Z_{i00})$$
$$E\left(Z_{i10}\middle|P;\theta^{(t-1)}\right)=\Pr\left(Z_{i10}=1\middle|P;\theta^{(t-1)}\right)$$
$$=\frac{\Pr\left(P\middle|Z_{i10}=1;\theta^{(t-1)}\right)\Pr\left(Z_{i10}=1\middle|\theta^{(t-1)}\right)}{denominator}$$
$$=\frac{{\alpha_1}^{(t-1)}{P_{i1}}^{{\alpha_1}^{(t-1)}-1}{\pi_{10}}^{(t-1)}}{denominator}$$

$$= \gamma(Z_{i10})$$

$$E\left(Z_{i01}\middle|P;\theta^{(t-1)}\right) = \Pr\left(Z_{i01}=1\middle|P;\theta^{(t-1)}\right)$$

$$= \frac{\Pr\left(P\middle|Z_{i01}=1;\theta^{(t-1)}\right)\Pr\left(Z_{i01}=1\middle|\theta^{(t-1)}\right)}{denominator}$$

$$= \frac{{\alpha_2}^{(t-1)}{P_{i2}}^{{\alpha_2}^{(t-1)}-1}{\pi_{01}}^{(t-1)}}{denominator}$$

$$= \gamma(Z_{i01})$$

$$E\left(Z_{i11}\middle|P;\theta^{(t-1)}\right) = \Pr\left(Z_{i11}=1\middle|P;\theta^{(t-1)}\right)$$

$$= \frac{\Pr\left(P\middle|Z_{i11}=1;\theta^{(t-1)}\right)\Pr\left(Z_{i11}=1\middle|\theta^{(t-1)}\right)}{denominator}$$

$$= \frac{{\alpha_1}^{(t-1)}{\alpha_2}^{(t-1)}{P_{i1}}^{{\alpha_1}^{(t-1)}-1}{P_{i2}}^{{\alpha_2}^{(t-1)}-1}{\pi_{10}}^{(t-1)}}{denominator}$$

$$= \gamma(Z_{i11})$$

$$\begin{aligned} denominator = {} & \Pr\left(P\middle|Z_{i00}=1;\theta^{(t-1)}\right)\Pr\left(Z_{i00}=1\middle|\theta^{(t-1)}\right) \\ & + \Pr\left(P\middle|Z_{i10}=1;\theta^{(t-1)}\right)\Pr\left(Z_{i10}=1\middle|\theta^{(t-1)}\right) \\ & + \Pr\left(P\middle|Z_{i10}=1;\theta^{(t-1)}\right)\Pr\left(Z_{i10}=1\middle|\theta^{(t-1)}\right) \\ & + \Pr\left(P\middle|Z_{i10}=1;\theta^{(t-1)}\right)\Pr\left(Z_{i10}=1\middle|\theta^{(t-1)}\right) \end{aligned}$$

$$\begin{aligned} = {} & {\pi_{00}}^{(t-1)} + {\alpha_1}^{(t-1)}{P_{i1}}^{{\alpha_1}^{(t-1)}-1}{\pi_{10}}^{(t-1)} + {\alpha_2}^{(t-1)}{P_{i2}}^{{\alpha_2}^{(t-1)}-1}{\pi_{01}}^{(t-1)} \\ & + {\alpha_1}^{(t-1)}{\alpha_2}^{(t-1)}{P_{i1}}^{{\alpha_1}^{(t-1)}-1}{P_{i2}}^{{\alpha_2}^{(t-1)}-1}{\pi_{10}}^{(t-1)} \end{aligned}$$

由此也可以发现：

$$\sum_{i=1}^{M}(\gamma(Z_{i00}) + \gamma(Z_{i01}) + \gamma(Z_{i10}) + \gamma(Z_{i11})) = M$$

接着在 M 步中，在限制$\pi_{00} + \pi_{10} + \pi_{01} + \pi_{11} = 1$的条件下关于$\theta$极大化$Q\left(\theta,\theta^{(t-1)}\right)$

$$\begin{aligned} Q\left(\theta,\theta^{(t-1)}\right) = \sum_{i=1}^{M}\big[& \gamma(Z_{i10})ln(\alpha_1{P_{i1}}^{\alpha_1-1}\pi_{10}) + \gamma(Z_{i01})ln(\alpha_2{P_{i2}}^{\alpha_2-1}\pi_{10}) \\ & + \gamma(Z_{i11})\,ln\left(\alpha_1\alpha_2{P_{i1}}^{\alpha_1-1}{P_{i2}}^{\alpha_2-1}\pi_{11}\right) + \gamma(Z_{i00})ln\pi_{00}\big] \end{aligned}$$

引入拉格朗日乘子λ，有：

$$L\left(\lambda,\theta,\theta^{(t-1)}\right) = Q\left(\theta,\theta^{(t-1)}\right) + \lambda(\pi_{00} + \pi_{10} + \pi_{01} + \pi_{11} - 1)$$

对$L\left(\lambda,\theta,\theta^{(t-1)}\right)$关于$\pi_{00}$、$\pi_{10}$、$\pi_{01}$、$\pi_{11}$求导有：

$$\sum_{i=1}^{M}\gamma(Z_{i00}) + \lambda\pi_{00} = 0 \qquad \sum_{i=1}^{M}\gamma(Z_{i10}) + \lambda\pi_{10} = 0$$

$$\sum_{i=1}^{M}\gamma(Z_{i01})+\lambda\pi_{01}=0 \qquad \sum_{i=1}^{M}\gamma(Z_{i11})+\lambda\pi_{11}=0$$

相加可得$M+\lambda=0$，即$\lambda=-M$，所以有

$$\pi_{00}{}^{(t)}=\frac{\sum_{i=1}^{M}\gamma(Z_{i00})}{M} \qquad \pi_{10}{}^{(t)}=\frac{\sum_{i=1}^{M}\gamma(Z_{i10})}{M}$$
$$\pi_{01}{}^{(t)}=\frac{\sum_{i=1}^{M}\gamma(Z_{i01})}{M} \qquad \pi_{11}{}^{(t)}=\frac{\sum_{i=1}^{M}\gamma(Z_{i11})}{M}$$

又对α_1、α_2求导可得

$$\alpha_1{}^{(t)}=-\frac{\sum_{i=1}^{M}[\gamma(Z_{i10})+\gamma(Z_{i11})]}{\sum_{i=1}^{M}[\gamma(Z_{i10})+\gamma(Z_{i11})]lnP_{i1}}$$
$$\alpha_2{}^{(t)}=-\frac{\sum_{i=1}^{M}[\gamma(Z_{i11})+\gamma(Z_{i01})]}{\sum_{i=1}^{M}[\gamma(Z_{i11})+\gamma(Z_{i01})]lnP_{i2}}$$

更新参数后计算对数似然$lnPr(P_1,P_2;\theta^{(t)})$，与$lnPr(P_1,P_2;\theta^{(t-1)})$进行比较，判断算法是否收敛，若没有收敛，则重新回到 E 步进行迭代，若已经收敛则停止，得到最终的参数估计$\hat{\theta}$。

3. 模拟数据分析

下面，本小组基于如上理论模型生成模拟数据，并使用模拟数据验证和讨论 EM 算法的正确性和有效性。

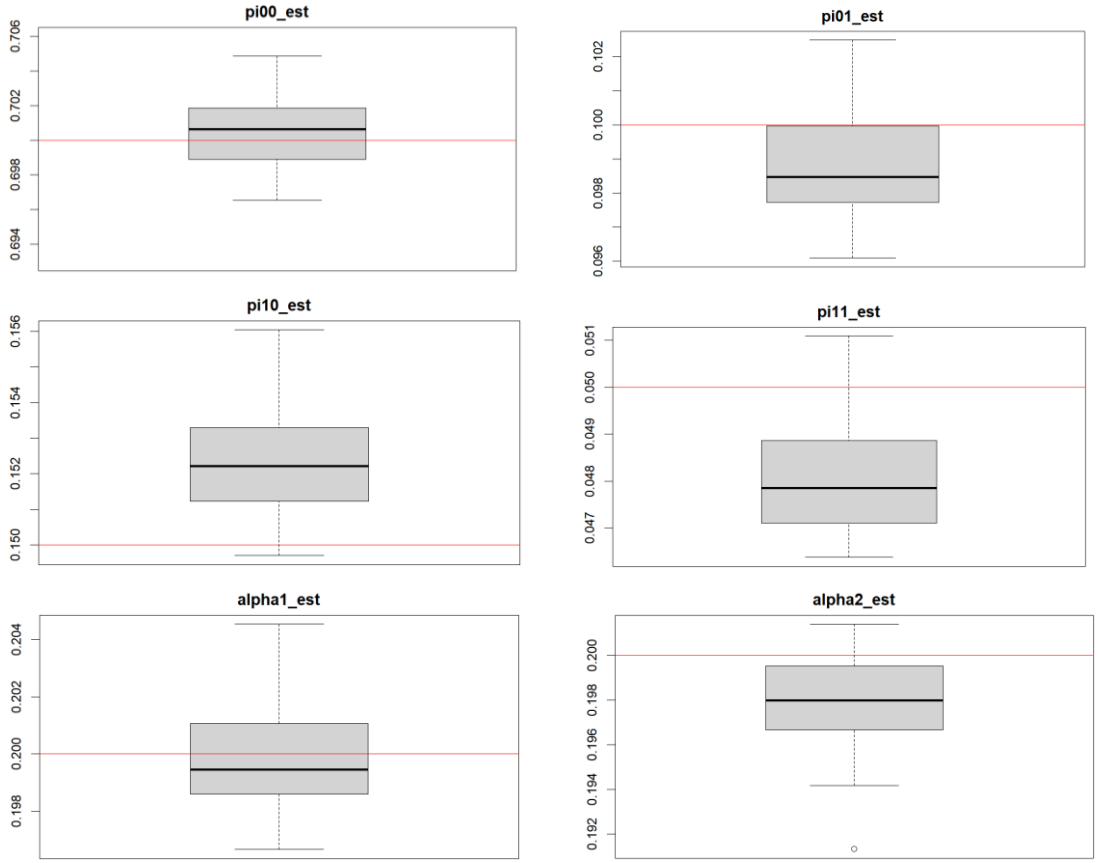
考虑到真实情况中，往往与两种疾病都有关的 SNP 很少，与其中一种疾病有关的 SNP 略少。本小组假设有 10 万个 SNP，参数的真实值为$\pi_{00}=0.7$，$\pi_{01}=0.1$，$\pi_{10}=0.15$，$\pi_{11}=0.05$，$\alpha_1=\alpha_2=0.2$。基于这些参数，生成模拟数据$P_1,P_2,Z_{00},Z_{01},Z_{10},Z_{11}$。

设置临界值为$1*10^{-6}$，当 EM 算法中迭代前与迭代后的对数似然差值小于此临界值时，就可以认为算法收敛。设置初始值$\pi_{00}{}^{(0)}=0.6$，$\pi_{01}{}^{(0)}=0.05$，$\pi_{10}{}^{(0)}=0.15$，$\pi_{11}{}^{(0)}=0.01$，$\alpha_1{}^{(0)}$=$\alpha_2{}^{(0)}=0.1$。通过 EM 算法迭代至收敛得到结果如表 1 中所示，从结果可以看出 FDP 值均约等于 0.1，且估计值与真实值十分的接近，所以可以初步认为 EM 算法是准确有效的。

表 1：EM 算法的估计结果

真实值	估计值	真实值	估计值
$\pi_{00} = 0.7$	0.6986	$\pi_{11} = 0.05$	0.0486
$\pi_{01} = 0.1$	0.0995	$\alpha_1 = 0.2$	0.2015
$\pi_{10} = 0.15$	0.1533	$\alpha_2 = 0.2$	0.1994
识别与两种疾病均有关的 SNP 的 FDP=**0.083**，power=**0.1498**			
识别与疾病 1 有关的 SNP 的 FDP=**0.103**，power=**0.4108**			
识别与疾病 2 有关的 SNP 的 FDP=**0.097**，power=**0.3767**			

为了进一步验证 EM 算法的准确性和有效性，本小组重复进行以上过程 20 次，每一次在相同的参数设置下生成不同的模拟数据并重新迭代以得到参数的估计值。最终，绘制参数估计值、FDP 值和 power 值的箱线图，如图 1 所示。从图中可以进一步看出，估计值的均值与真实值相当接近，且方差均小于 0.01，说明 EM 算法的准确性较好。FDP 值约等于 0.1，说明 EM 算法是有效的。



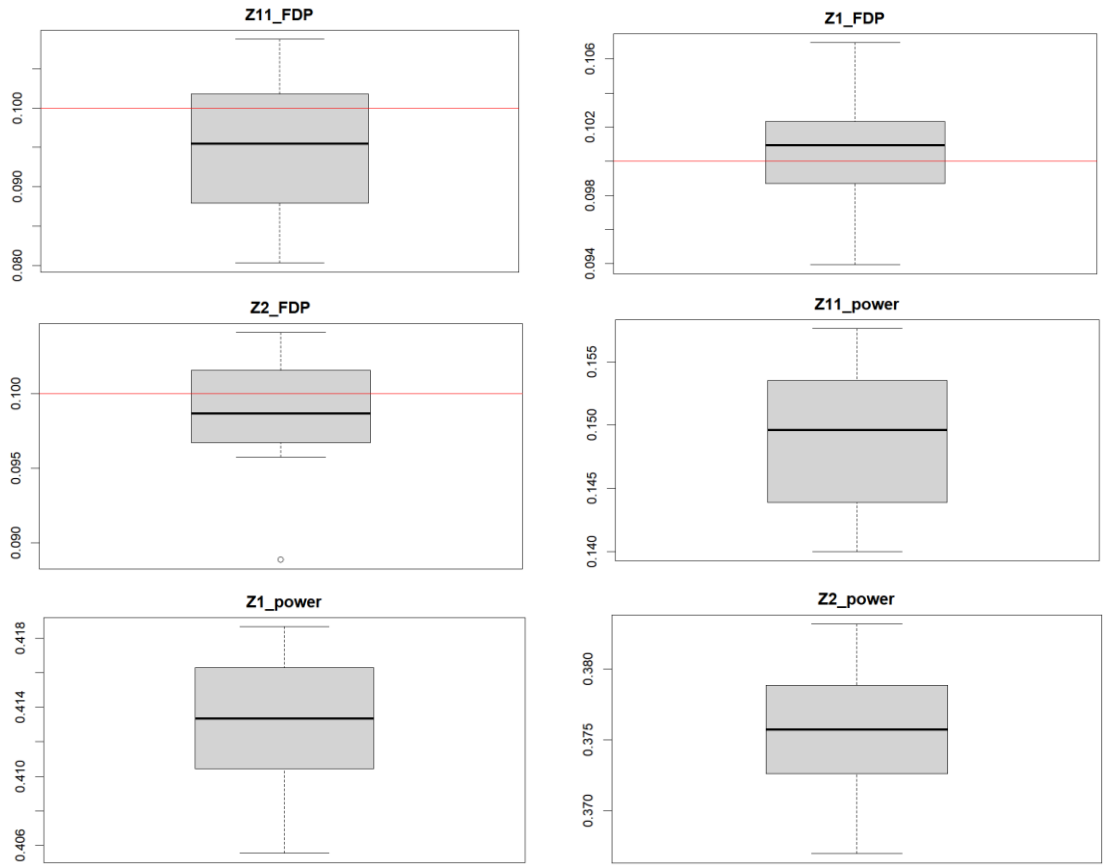


图 1：基于对两种疾病的 GWAS 数据进行整合分析方法重复运行 20 次模拟的结果

通过对两种 GWAS 数据进行整合分析的方法，本小组基于模拟数据识别出约有 9159 个 SNP 与疾病 1 有关，约有 6256 个 SNP 与疾病 2 有关，其中有 817 个 SNP 与两种疾病均有关。

下面，基于以上的模拟数据，本小组对其利用只使用单一疾病 GWAS 数据的方法做进一步分析，以此来比较两种方法的不同。

在只使用单一疾病 GWAS 数据的方法中，只有两个参数π_1和α。对于疾病 1，π_1真实值应为π_{11}真实值与π_{10}真实值之和，α真实值为α_1真实值，Z 真实值应为Z_{11}真实值与Z_{10}真实值之和，P 真实值为P_1真实值；对于疾病 2，π_1真实值应为π_{11}真实值与π_{01}真实值之和，α真实值为α_2真实值，P 真实值为P_2真实值，Z 真实值应为Z_{11}真实值与Z_{01}真实值之和，以此确保是在相同数据集上进行两种方法的比较。

具体分析结果如图 2 所示。可以看出，只使用单一疾病 GWAS 数据的方法中参数估计也很准确，并不会显著优于或差于整合分析方法中参数估计的准确性，且由 FDP 值可以看出是有效的。

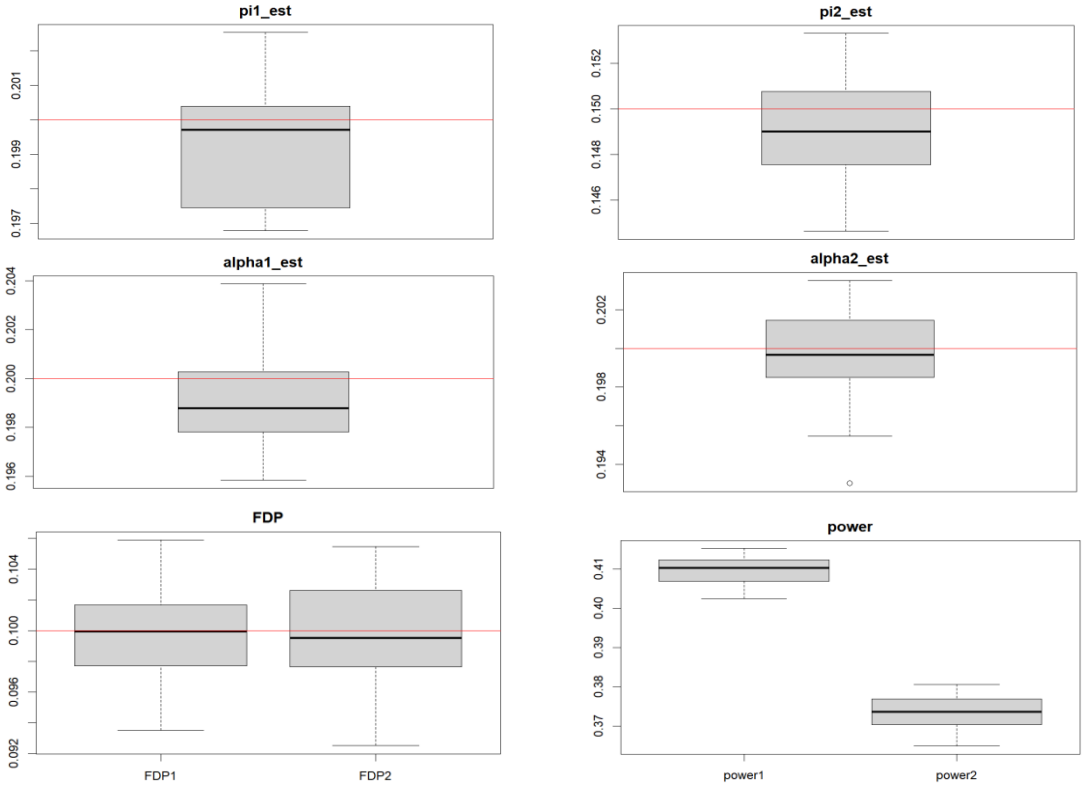


图 2：基于对单个疾病的 GWAS 数据进行分析方法重复运行 20 次模拟的结果

为了比较两种方法功效的不同，将识别与单一疾病有关的 SNP 时的功效箱线图绘制在同一张图中，结果如图 3 所示。可以发现：在这种参数设置下，整合分析的功效高于只使用单一疾病 GWAS 数据进行分析时的功效，即在判断与单一疾病有关的 SNP 中，两种方法判断出与单一疾病有关的 SNP 个数是相同的，但整合分析方法判对的比例更高，可以识别出更多真正与单一疾病有关的 SNP。

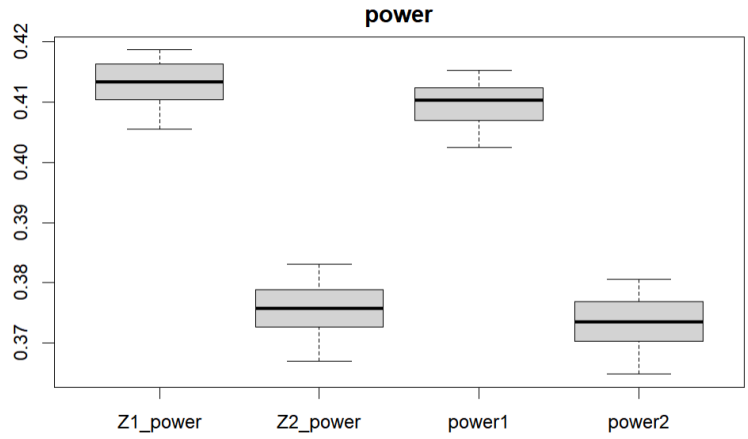


图 3：两种方法下识别与单一疾病有关的 SNP 时的功效箱线图

以上结果均基于参数真实值为$\pi_{00}=0.7$，$\pi_{01}=0.1$，$\pi_{10}=0.15$，$\pi_{11}=0.05$，$\alpha_1=\alpha_2=0.2$时得到的，无法保证结论具有普适性，所以本小组下面探究参数真实值对整合分析中估

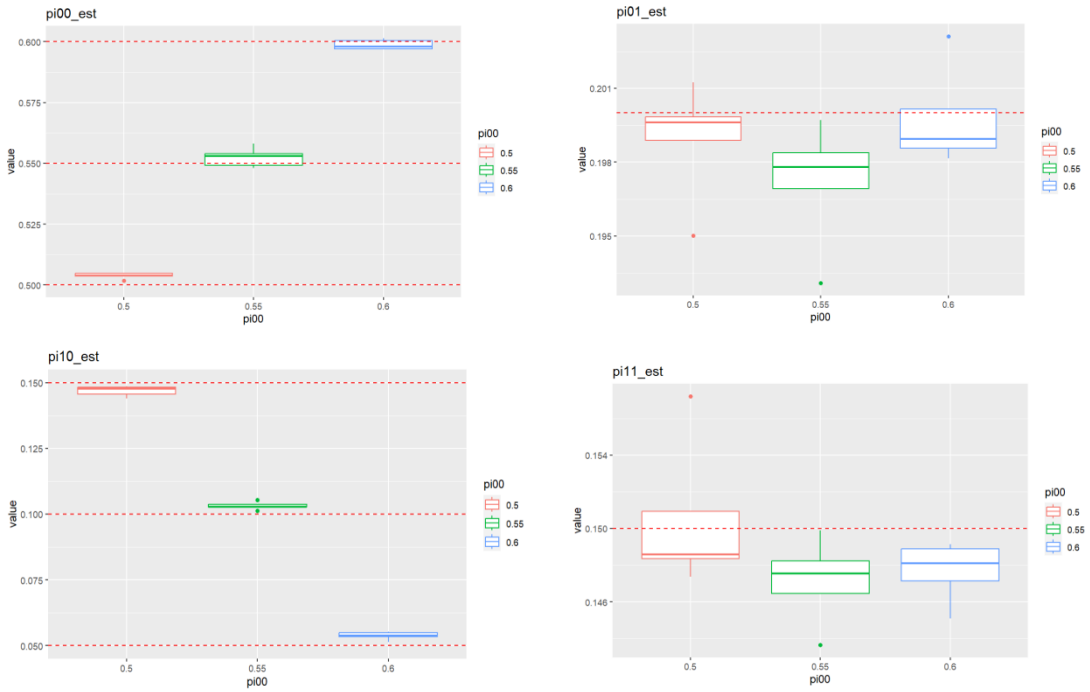
计值、FDR 和 power 的影响，并再次与对单个疾病的 GWAS 数据进行分析方法进行对比。

通过学习我们已经知道，在对单个疾病的 GWAS 数据进行分析的方法中，参数π_1的真实值变化时对自身的估计没有影响，但是当π_1的真实值越小时对α的估计越不准确，功效越低。基于此本小组考虑一种最简单的情况：保持π_{01}与π_{11}的真实值不变，改变π_{00}和π_{10}的真实值，设置$\pi_{00} = (0.5,0.55,0.6)$，$\pi_{10} = (0.15,0.1,0.05)$，$\pi_{01} = 0.2$，$\pi_{11} = 0.15$，$\alpha_1 = \alpha_2 = 0.2$。此时对单个疾病 2 的 GWAS 数据进行分析时，参数π_1的真实值是不变的；对单个疾病 1 的 GWAS 数据进行分析时，参数π_1的真实值是逐渐变小的。

整合分析的参数估计结果如图 4 所示，对单个疾病的 GWAS 数据进行分析方法的参数估计结果如图 5 所示。可以发现：两种方法参数估计值的方差大小相近，但整合分析方法的均值会更接近于参数真实值。另外，参数π_{01}和π_{11}估计值的方差很小，均值接近于真实值，可以说明整合分析中参数估计的效果对于参数真实值的敏感性是比较小的，且可以很好的兼顾多个参数值的估计效果。

再对比两种方法下的 power 值，结果如图 6 所示。可以发现：在参数真实值改变的同时，Z1_power（整合分析下识别与疾病 1 有关的 SNP 的功效）总是略大于 Z3_power（对单一 GWAS 数据进行分析方法下识别与疾病 1 有关的 SNP 的功效）。

综上模拟结果，可以看出整合分析的效果优于对单一 GWAS 数据进行分析方法的效果。



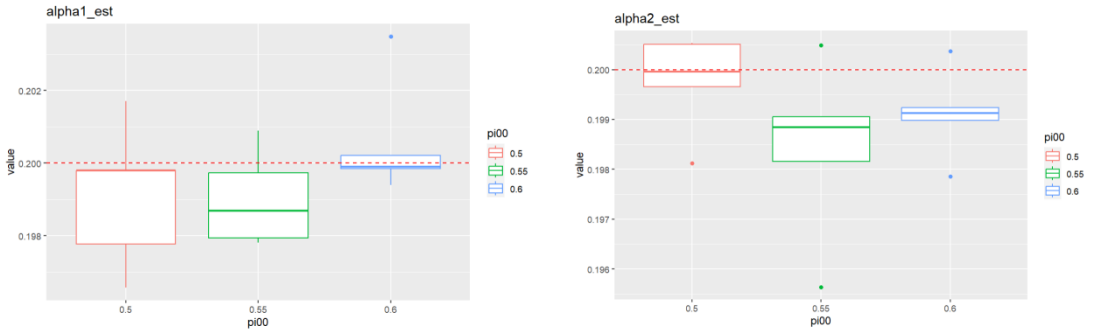


图 4：改变参数真实值时整合分析的参数估计结果

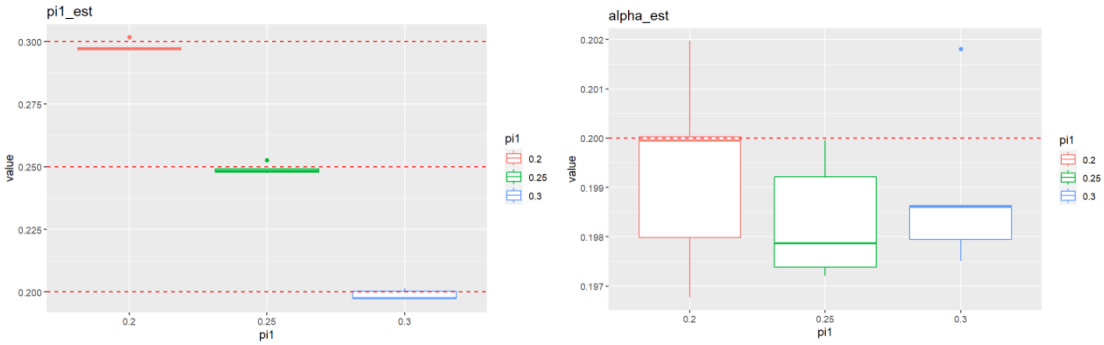


图 5：改变参数真实值时对单个疾病 1 的 GWAS 数据进行分析方法的参数估计结果

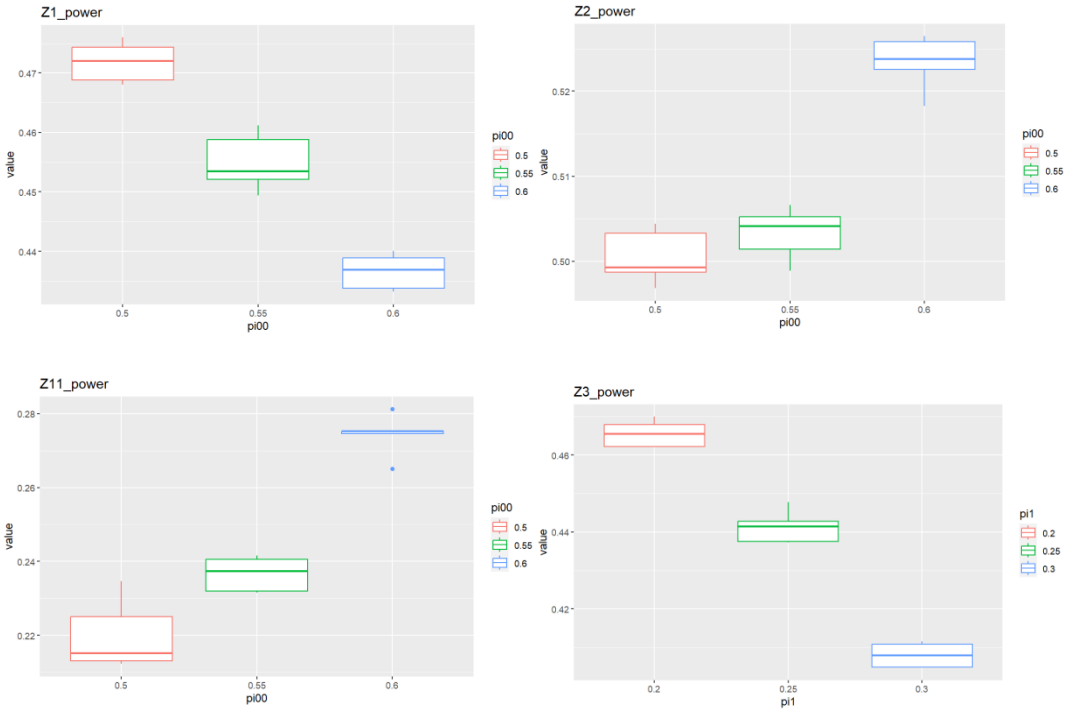


图 6：两种方法下的 power 值

4. 真实数据分析

4.1 EM 算法

我们将实现的算法应用于真实数据分析。我们基于两个数据集，分别是躁郁症（Bipolar disorder, BIP）的 GWAS 数据集和精神分裂症（Schizophrenia, SCZ）的 GWAS 数据集，两数据集都包含 SNP 的 ID 与 SNP 对应的 p 值。由于两数据集的 SNP 数量及种类都不同，首先根据 SNP 的 ID 对两数据集取交集。然后进行参数估计。参数估计的结果如下表 X：

表 2：EM 算法对于真实数据的估计结果

	$\hat{\pi}_{00}$	$\hat{\pi}_{01}$	$\hat{\pi}_{10}$	$\hat{\pi}_{11}$	$\hat{\alpha}_1$	$\hat{\alpha}_2$
BIP-SCZ	0.81942	0.00121	0.00002	0.17934	0.60190	0.50308

在控制 FDR 水平为 0.1 的条件下，推断得到与两种疾病同时相关的 SNP 有 8024 个，与躁郁症相关的 SNP 有 8025 个、与精神分裂症相关的 SNP 有 8279 个。由此结果发现与两种疾病相关的 SNP 的 ID 和与 BIP 相关的 SNP 的 ID 高度重合。通过对这些 SNP 的 ID 的检查，可以找到两类 SNP ID 的差集只有"rs8006004"，该 SNP 与精神分裂症不相关，与躁郁症相关（FDR = 0.1）。

我们进行更详细的对比研究，以评估整合疾病相关的 GWAS 数据集是否有助于危险遗传变异的识别。分别进行只使用 BIP 的单 GWAS 数据集分析与只使用 SCZ 的单 GWAS 数据集分析。分析时所使用的 SNP 及对应的 p 值都基于前述取过交集后的结果。

我们先对参数进行估计，然后再推断与对应疾病相关的 SNP 个数，与前述整合分析结果进行对比。两个单数据集分析的参数估计结果见表 X：

表 3：EM 算法对于单数据集的估计结果

	$\hat{\pi}$	$\hat{\alpha}$
BIP	0.21942	0. 66289
SCZ	0.24198	0. 56681

在 BIP 的单 GWAS 数据集分析中， FDR 水平为 0.1 的条件下，有 403 个 SNP 被识别出，在 SCZ 的单 GWAS 数据集分析中，相同的 FDR 下，有 4012 个 SNP 被识别出。将单数据集识别出的 SNP ID 与整合分析的方法识别出的 SNP ID 进行比较，比较结果如表 X。我们发现整合分析可以识别出大部分单 GWAS 数据集分析可识别的 SNP，存在一些整合分析方法未能识别的 SNP。对于这样的结果，我们认为这与两模型的对数似然函数有关，单

独分析一个 GWAS 数据集时，对数似然函数时一个两元函数，相对更容易找到全局最优，而整合分析需要估计六个参数的值，其对数似然函数的形状也更为不规则，在考虑多个参数估计值时，因此可能会发生一些遗漏。但从识别出的 SNP 种类总数来看，整合分析的识别效力更高。

表 4：结果比较

		整合分析	
		识别	未识别
单独使用 BIP	识别	390	13
	未识别	7635	\
单独使用 SCZ	识别	3683	329
	未识别	4596	\

之后，我们利用 manhattan 函数绘制曼哈顿图用于效果展示。由下图可以直观的看出整合其他 GWAS 数据之后，可识别的 SNP 增多。

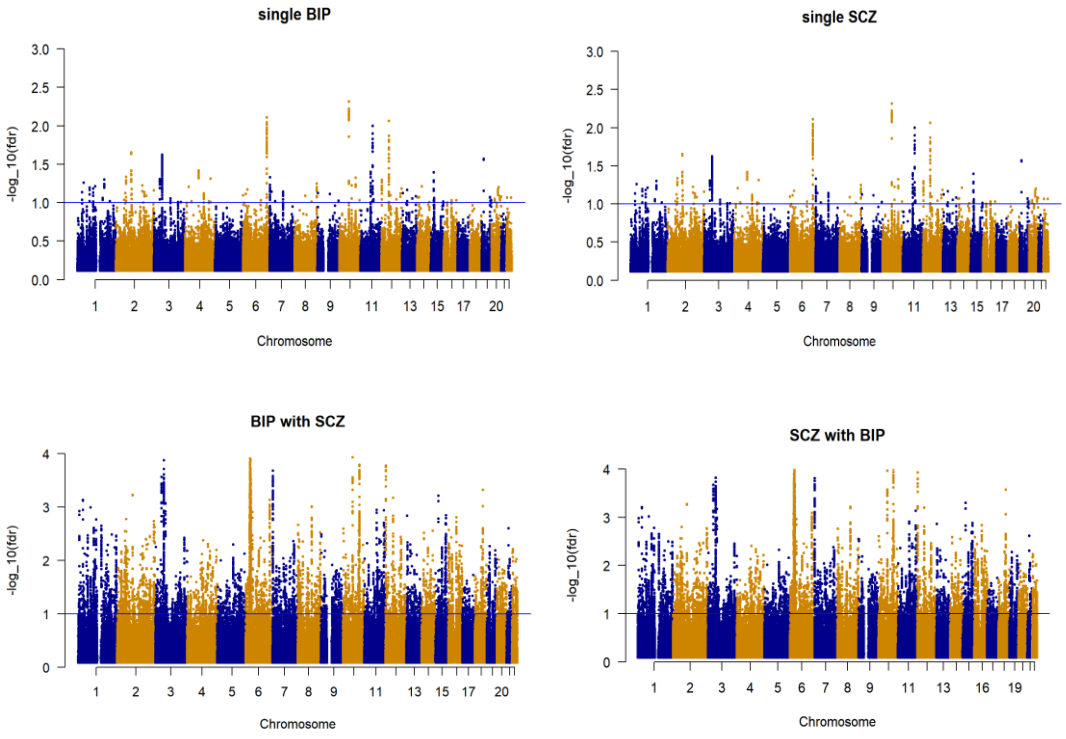


图 7：两种方法的曼哈顿图

总结我们应用于真实数据分析的结果，从识别出的 SNP 数量及类别来说，整合分析的方法都可以识别更多数量及种类的与单个疾病有关的 SNP。另外，还可以良好地识别与两个疾病均有关的 SNP，并且考虑了两个疾病之间的成对多效性。因此，整合相关疾病的 GWAS 数据集有利于危险遗传变异的识别。在进行分析时，可以将整合分析的结果与单

GWAS 数据集分析的结果相结合，将更有利于对危险遗传变异的识别与分辨。

4.2 GPA 模型

在小组作业探索中，我们发现 Dongjun Chung 等人提出 GPA 统计方法[1]，它通过联合分析多个 GWAS 数据集和注释信息来增加识别风险变异的能力。于是，本小组调用 GPA 包对 BIP 和 SCZ 进行联合分析，印证之前的结论。参数结果估计如下表所示，与我们得到的参数估计相近。同时，小组利用 GPA 包进行多效性检验，得到 $p<0.05$，即 BIP 与 SCZ 具有相关性。

表 4：GPA 对于真实数据的估计结果

	$\hat{\pi}_{00}$	$\hat{\pi}_{01}$	$\hat{\pi}_{10}$	$\hat{\pi}_{11}$	$\hat{\alpha}_1$	$\hat{\alpha}_2$
BIP-SCZ	0.81979	0.00094	0.000006	0.18021	0.60280	0.50275
	（0.003）	（0.003）	（0.004）	（0.005）	（0.005）	（0.003）

（注：括号内为参数估计的标准差）

5. 参考文献

[1] Chung D, Yang C, Li C, Gelernter J, Zhao H. GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. Plos Genetics. 2014 Nov;10(11):e1004787. DOI: 10.1371/journal.pgen.1004787. PMID: 25393678; PMCID: PMC4230845.