Advances in Applied Mathematics 应用数学进展, 2021, 10(7), 2339-2347
Published Online July 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.107244

Hans汉斯

基于马尔科夫随机场对二型糖尿病早期诊断

田雪寒[1]，王艺舒[2*]

[1]青岛大学，山东 青岛
[2]北京科技大学，北京
Email: tianxuehan@126.com, *yishu6661@126.com

收稿日期：2021年6月7日；录用日期：2021年6月28日；发布日期：2021年7月12日

摘　要

为了探究二型糖尿病早期动态变化，本文通过分析小鼠标本的组织特异性表达数据，获得相邻时间段之间的差异表达的证据，以表征基因差异表达的动态变化。该数据集中含有丰富的时空信息，但以往的研究中很难充分利用。我们通过在潜在状态上指定马尔可夫随机场(MRF)合并具有时空结构的复杂数据，用蒙特卡洛期望最大化(MCEM)算法计算模型参数，其关键特征是同时考虑基因表达水平的空间相似性和时间依赖性，仿真研究与实例分析结果都表明该方法具有更高的灵敏度，能识别更多的DE基因，能从数据中提取更多生物学上有意义的结果。

关键词

二型糖尿病早期，马尔可夫随机场，MCEM算法

A Markov Random Field-Based Approach for Early Diagnosis of Type 2 Diabetes

Xuehan Tian[1], Yishu Wang[2*]

[1]Qingdao University, Qingdao Shandong
[2]University of Science & Technology Beijing, Beijing
Email: tianxuehan@126.com, *yishu6661@126.com

Received: Jun. 7th, 2021; accepted: Jun. 28th, 2021; published: Jul. 12th, 2021

Abstract

In order to explore the early dynamic changes of type 2 diabetes mellitus, this paper analyzed the

*通讯作者。

文章引用: 田雪寒, 王艺舒. 基于马尔科夫随机场对二型糖尿病早期诊断[J]. 应用数学进展, 2021, 10(7): 2339-2347.
DOI: 10.12677/aam.2021.107244

tissue-specific expression data of mouse specimens to obtain the evidence of differential expression between adjacent time periods, so as to characterize the dynamic changes of differential gene expression. This dataset contains abundant spatiotemporal information, but it is difficult to make full use of it in previous studies. We pass in the potential state specified on Markov random field (MRF) combined with space-time structure of complex data, using the Monte Carlo calculation model for the expectation maximization (MCEM) algorithm parameters, its key characteristic is also considering gene expression level of similarity space and time dependence, simulation research and the example analysis results show that this method has higher sensitivity. More DE genes can be identified and more biologically significant results can be extracted from the data.

Keywords

Type 2 Diabetes, Markov Random Field, MCEM Algorithm

1. 引言

糖尿病是一种难以根治的慢性疾病，慢性并发症可遍及全身重要器官，急性并发症严重时可危及生命。其中二型糖尿病(T2D)占糖尿病患者总数的九成以上，糖尿病早期由于因胰岛 β 细胞的代偿机制代偿机制，血糖变化与正常人群差别较小，做专项检查成本高，容易造成医疗资源浪费。因此，诊断出“糖尿病早期”并积极采取干预治疗手段，将有利于推进治愈二型糖尿病这一难题。

自 20 世纪 90 年代中后期，基因微阵列技术问世以来便受到科学界的广泛关注，是基因组学中能有效地测量有机体的基因表达水平所使用最广泛的工具。揭示生物网络是系统生物学中的一个关键目标，现在我们能通过微阵列在各种条件下测量基因组水平的表达情况，这些数据可以从统计上推断基因调控网络。在生物学中，许多不同的生物过程都用图表示，比如：包括蛋白质－蛋白质相互作用网络(PPI)，转录调控网络和基因共表达网络，生物途径等。

我们用由节点 $V=\{1,\cdots,p\}$，边 $E\subset V\times V$ 组成的图 $G=(V,E)$ 表示生物网络。对于条件独立图，当且仅当节点 i 和节点 j 在给定所有其他节点都不条件独立的情况下，$(i,j)\in E$。高斯图形模型(GGMs)则已经被证明是推断条件独立图的最佳方法之一。高斯图模型是一种概率图模型，GGMs 起源于 20 世纪 70 年代早期，这一概念源于 AP Dempster [1]的经典论文，是研究基因之间关联网络的工具。在 GGM 中，假定 p 维随机变量 $X=(X_1,\cdots,X_p)$ 服从多元高斯分布 $N(\mu,\Sigma)$，$\Theta=\Sigma^{-1}$，可以证明 X_i 和 X_j 的条件独立性等效于 $\Theta_{ij}\neq 0: X_i\perp X_j \left| X_{V\setminus\{i,j\}} \right. \Leftrightarrow \Theta_{ij}=0$。在 GGM 中，估计条件独立图等效于估计 Θ 中的非零条目。

为了估计 GGM，Meinshausen & Buhlmann [2]证明了使用套索的邻域选择是一个计算吸引力的标准协方差稀疏高维图选择模型。邻域选择分别估计了图中的每个节点的条件独立性限制，因此等价于高斯线性模型的变量选择。同时证明了稀疏的邻域选择方案对于高维图是一致，其一致性取决于惩罚参数的选择。Yuan & Lin [3]提出了一种有效的算法来扩展 Lasso 方法的特征选择，并表明这些扩展具有优越的性能，在一定条件下，Lasso 变量的选择具有一致性。Friedman *et al.* [4]考虑用逆协方差矩阵上的 Lasso 惩罚来估计稀疏图的问题.，利用 Lasso 的坐标下降过程，开发了一种简单的图形 Lasso 算法，命名为图形套索(glasso)。Cai *et al.* [5]，Dobra *et al.* [6]，Wang *et al.* [7]等三人同样对估计 GGM 做出了许多贡献。

本文提出的工作是基于对小鼠标本的组织特异性表达数据的分析所激发的，我们希望通过分析小鼠标本的组织特异性表达数据，在小鼠发育的不同时期，不同的器官组织基因表达动态发育过程不同，为了有效利用该数据集中丰富的时空信息，我们的方法基于马尔可夫随机场(MRF)模型识别相邻时间点的差异表达基因(DE)，用蒙特卡洛期望最大化(MCEM)算法计算模型参数，考虑不同时间不同器官组织的病理性变化。我们的方法的关键特征是同时考虑基因表达水平的空间相似性和时间依赖性，获得相邻时间段之间的差异表达的证据，以表征基因差异表达的动态变化，从而探究二型糖尿病早期动态变化。以更好地从数据中提取生物学上有意义的结果。

2. 模型构建

2.1. 差异基因的潜状态模型

对于二型糖尿的早期患者，因胰岛 β 细胞的代偿机制，血糖波动于正常人群差异很小，通过现有检测方式诊断时患者已经不可能自愈。为了探究二型糖尿病早期动态变化，我们通过分析小鼠标本的组织特异性表达数据，获得相邻时间段之间的差异表达的证据，以表征基因差异表达的动态变化。

首先我们假设 $B=3$ 为脂肪，肝脏及胰腺这三处采样组织，$T=3$ 为小鼠发育的第 1 周、9 周、18 周三个年龄阶段，$G=45101$ 为基因数。令 n_{bt} 表示在区域 b 处，年龄为 t 的小鼠的重复次数，$N_b=\left(n_{b1},\cdots,n_{bt},\cdots,n_{bT}\right)'$ 为区域 b 中重复次数的列向量，$N=\left(N_1,\cdots,N_b,\cdots,N_B\right)$ 为重复次数

我们假设 y_{bgtk}，$k=1,\cdots,n_{bt}$ 服从均值为 μ_{bgt} 和方差为 σ_0^2 的正态分布 $y_{bgtk}\sim N\left(\mu_{bgt},\sigma_0^2\right)$。$x_{bgt}$ 为二进制潜状态，表示基因 g 是否在区域 b 时刻 t 中表达，以 x_{bgt} 为条件，我们假设 μ_{bgt} 遵循高斯分布：

$$\mu_{bgt}\left|x_{bgt}=0\sim N\left(\mu_{1b},\sigma_{1b}^2\right)\right.$$
$$\mu_{bgt}\left|x_{bgt}=1\sim N\left(\mu_{2b},\sigma_{2b}^2\right)\right.$$

我们假设在特定的区域，混合成分的均值和方差不同，用 $\mu_1,\mu_2,\sigma_1,\sigma_2$ 表示所有区域的参数向量。y_{bgtk} 在 x_{bgt} 条件下的分布具有以下形式：

$$y_{bgtk}\left|x_{bgt}=0\sim N\left(\mu_{1b},\sigma_{1b}^2+\sigma_0^2\right)\right.$$
$$y_{bgtk}\left|x_{bgt}=1\sim N\left(\mu_{2b},\sigma_{2b}^2+\sigma_0^2\right)\right.$$

给定潜变量数组 X，假定 $f\left(Y\middle|X\right)=\prod_{b=1}^{B}\prod_{g=1}^{G}\prod_{t=1}^{T}f\left(y_{bgt}\middle|x_{bgt}\right)$ 条件独立，则

$$f\left(y_{bgt}\middle|x_{bgt}\right)=\prod_{k=1}^{n_{bt}}f\left(y_{bgt}\middle|x_{bgt}\right).$$

下面我们对于差异表达基因(DE)的分析，我们首先将观察到的数据转换为数组 $B\times G\times\left(T-1\right)$，在相邻时段之间执行 t 检验并将 t 统计量转换为 $z\sim$ scores。设 $y_{bg(t-1)}$ 和 y_{bgt} 分别为 t-1 与 t 时刻区域 b 中基因 g 的表达值的向量，双样本 t 统计量为：

$$t_{bg(t-1)}=\frac{\overline{y}_{bgt}-\overline{y}_{bg(t-1)}}{s}$$

其中 s 为 $\overline{y}_{bgt}-\overline{y}_{bg(t-1)}$ 的标准误差的估计值。然后将检验统计量 $t_{bg(t-1)}$ 转换为 $z\sim$ scores：

$$z_{bg(t-1)}=\Phi^{-1}\left(F_{n_{bt}+n_{b(t-1)}-2}\left(t_{bg(t-1)}\right)\right).$$

其中 $n_{bg(t-1)}$ 和 n_{bgt} 是 $y_{bg(t-1)}$ 和 y_{bgt} 中重复的数目。

基因表达数据由 $B\times G\times(T-1)$ 数组 Z 表示。s_{bgt} 表示表示基因 g 在区域 b 是否在时段 t 和 t+1 之间差异表达的二进制潜状态，这是我们推断的目的。设 S 为维数 $B\times G\times(T-1)$ 的潜状态数组。在 s_{bgt} 条件下，我们假设 z_{bgt} 遵循混合分布：

$$f\left(z_{bgt}\middle|s_{bgt}\right)=\left(1-s_{bgt}\right)f_0\left(z_{bgt}\right)+s_{bgt}f_1\left(z_{bgt}\right).$$

其中 $f_0(z)$ 是零密度，$f_1(z)$ 是非零密度。我们假设零密度遵循标准的 $N(0,1)$ 分布。我们采用非参数经验贝叶斯框架(Efron [8])，通过使用 R 包 locfdr 将非空密度与自然样条拟合。给定 S，假定条件独立：

$$f(Z|S)=\prod_{b=1}^{B}\prod_{g=1}^{G}\prod_{t=1}^{T-1}f\left(z_{bgt}\middle|s_{bgt}\right).$$

2.2. MRF 模型的先验概率

上述模型和推理目标中的一个关键组成部分是我们未知的潜在状态数组 X，下面我们用 MRF 模型生成 X 的先验值 $p(X)$。对于每个基因 g，我们构建一个无向图 $G_g=(V_g,E_g)$，其中 $V_g=\{x_{bgt}:b=1,\cdots,B,t=1,\cdots,T\}$ 是节点的集合，E_g 是边的集合，E_g 由 E_{g1} 与 E_{g2} 组成，$E_{g1}=\{(x_{bgt},x_{b'gt'}):b\neq b',t=t'\}$ 为包含捕获的区域之间空间依赖性的边缘 $E_{g2}=\{(x_{bgt},x_{b'gt'}):b=b',|t-t'|=1\}$ 为包含捕获的相邻周期之间时间依赖性的边缘。

对于 $p(X)$ 的联合分布概率，我们建立下列成对交互的 MRF 模型(Besag [9], Lin *et al.* [10])：

$$\begin{aligned}p(X|\Phi)\propto\prod_{g=1}^{G}\exp\Bigg\{&\gamma_0\sum_{V_g}I_0(x_{bgt})+\gamma_1\sum_{V_g}I_1(x_{bgt})\\&+\beta_1\sum_{E_{g1}}\left[I_0(x_{bgt})I_0(x_{b'gt'})+I_1(x_{bgt})I_1(x_{b'gt'})\right]\\&+\beta_2\sum_{E_{g2}}\left[I_0(x_{bgt})I_0(x_{b'gt'})+I_1(x_{bgt})I_1(x_{b'gt'})\right]\Bigg\}\end{aligned}\tag{2-(1)}$$

其中 $I_0()$ 与 $I_1()$ 为指标函数，令 $\gamma=\gamma_1-\gamma_0$，可得条件概率：

$$\begin{aligned}p(x_{bgt}|X;\Phi)&=\frac{\exp\{x_{bgt}F(x_{bgt},\Phi)\}}{1+\exp\{F(x_{bgt},\Phi)\}}\\F(x_{bgt},\Phi)&=\gamma+\beta_1\sum_{b'\neq b}(2x_{b'gt}-1)+\beta_2\left\{I_{t\neq1}\left[2x_{bg(t-1)}-1\right]+I_{t\neq T}\left[2x_{bg(t+1)}-1\right]\right\}\end{aligned}\tag{2-(2)}$$

其中 β_1 为捕获空间依赖性的参数，β_2 为捕获时间间依赖性的参数。

接下来，我们考虑先验分布 $p(S)$ 的 MRF 模型，同时考虑了时间依赖性和空间相似性。我们将 3 个区域分成两组：用 B_p 表示胰腺区域，B_n 非胰腺区域。联合概率类似于 2-(1)，可以计算条件概率并具有以下形式：

$$p\left(s_{bgt}\middle|S/s_{bgt};\Phi_{DE}\right)=\frac{\exp\{s_{bgt}F_{DE}(s_{bgt},\Phi_{DE})\}}{1+\exp\{F_{DE}(s_{bgt},\Phi_{DE})\}}\tag{2-(3)}$$

如果 $b\in B_p$：

$$\begin{aligned}F_{DE}(s_{bgt},\Phi_{DE})=\gamma_{DE}&+\beta_{pp}\sum_{b'\in B_{n/b}}(2s_{b'gt}-1)+\beta_{pn}\sum_{b'\in B_n}(2s_{b'gt}-1)\\&+\beta_t\left\{I_{t\neq1}\left[2s_{bg(t-1)}-1\right]+I_{t\neq T}\left[2s_{bg(t+1)}-1\right]\right\}\end{aligned}$$

如果 $b \in B_n$：

$$F_{DE}\left(s_{bgt},\Phi_{DE}\right)=\gamma_{DE}+\beta_{nn}\sum_{b'\in B_{n/b}}\left(2s_{b'gt}-1\right)+\beta_{np}\sum_{b'\in B_p}\left(2s_{b'gt}-1\right)$$
$$+\beta_t\left\{I_{t\neq 1}\left[2s_{bg(t-1)}-1\right]+I_{t\neq T}\left[2s_{bg(t+1)}-1\right]\right\}$$

其中 $\Phi_{DE}=\left(\beta_{pp},\beta_{nn},\beta_{pn},\beta_{np}\right)$，$\beta_{pp}$ 为胰腺区之间的空间系数，β_{nn} 为非胰腺区之间的空间系数，β_{pn} 为胰腺区到非胰腺区之间的空间系数，β_{np} 为非胰腺区到胰腺区之间的空间系数，由对称性可得，$\beta_{pn}=\beta_{np}$。

3. 参数估计与后验概率估计

3.1. 参数估计

3.1.1. $p(X)$的参数估计算法

估计 MRF 参数 $\Phi=\left(\gamma_1,\beta_1,\beta_2\right)$ 和高斯混合模型参数 $\Theta=\left(\mu_1,\sigma_1;\mu_2,\sigma_2\right)$：

如果给定数组 X，Φ，Θ,可以通过最大似然估计(MLE)进行估计。但是，潜伏状态 X 是未观察到的，因此也需要进行估计。通常针对缺失数据估计实施期望最大化(EM)算法，但由于期望项难以控制，因此不适用于我们的模型。因此，我们提出以下 MCEM 算法[11]来估计：

1) 通过无偏估计量估计 σ_0：

$$\hat{\sigma}_0^{\,2}\frac{1}{G\times\sum_{b=1}^{B}\sum_{t=1}^{T}\left(n_{bt}-1\right)}\sum_{g=1}^{G}\sum_{b=1}^{B}\sum_{t=1}^{T}\sum_{k=1}^{n_{bt}}\left(y_{bgtk}-\bar{y}_{bgt}\right)^2$$

2) 在不考虑空间和时间依赖性的情况下，通过简单的高斯混合模型获得初始估计值 $\hat{X}$ 和 $\hat{\Theta}$。

3) 因为没有针对 Φ 的显式 MLE，所以选择初始估计 $\hat{\Phi}$，将该估计将最大化得到以下伪似然函数 $l\left(\hat{X};\Phi\right)$ (Besag [12])：

$$l\left(\hat{X};\Phi\right)=\prod_{b=1}^{B}\prod_{g=1}^{G}\prod_{t=1}^{T}p\left(\hat{x}_{bgt}\middle|\hat{X}/\hat{x}_{bgt};\Phi\right)$$

4) 令 $\Psi=\left(\Phi,\Theta\right)$，EM 算法中预期的完整数据对数似然可通过蒙特卡罗方法(Wei and Tanner [11])进行近似：

$$Q_m\left(\Psi\middle|\hat{\Psi}^{(r)}\right)=\frac{1}{m}\sum_{l=1}^{m}\ln f\left(Y,X_l^{(r)}\middle|\Psi\right)\qquad 3\text{-}(1)$$

其中 $X_1^{(r)},\cdots,X_m^{(r)}$ 通过吉布斯采样获得。从 $X_l^{(r)}$ 到，$X_l^{(r+1)}$ 中的所有条目都被下列公式按顺序更新，

$$p\left(x_{bgt}\middle|Y,X/x_{bgt};\hat{\Psi}^{(r)}\right)\propto p\left(x_{bgt}\,X/x_{bgt};\hat{\Phi}^{(r)}\right)f\left(y_{bgt}\middle|x_{bgt};\hat{\Theta}^{(r)}\right)\qquad 3\text{-}(2)$$

5) 用 $\hat{\Psi}^{(r+1)}$ 更新，使 3-(1)最大化：

$$\hat{\Psi}^{(r+1)}=\arg\max_{\Psi}Q_m\left(\Psi\middle|\hat{\Psi}^{(r)}\right)$$

与步骤 3 相同，我们用 $Q_m\left(\Psi\middle|\hat{\Psi}^{(r)}\right)$ 中的伪似然函数替换似然函数。包含 Φ，Θ 的项是可分离的，因此可以分别进行优化。

6) 重复步骤 4，步骤 5，直至收敛。

3.1.2. $p(S)$的参数估计算法

为了随着时间的推移识别 DE 基因。在该模型中，只需要迭代地更新 MRF 先验中的参数 Φ。该算法

与上面的算法有一些相似之处：

1) 汇总 Z 中的 $z\sim$scores 并通过 locfdr 程序估算 f_1。

2) 不考虑空间和时间依赖性的情况下，通过简单的混合模型获得初始估 $\hat{S}$。

3) 获得初始估计 $\hat{\Phi}_{DE}$，其最大化的伪似然函数：

$$l\left(\hat{S};\Phi_{DE}\right)=\prod_{b=1}^{B}\prod_{g=1}^{G}\prod_{t=1}^{T-1}p\left(\hat{s}_{bgt}\left|\hat{S}/\hat{s}_{bgt};\Phi_{DE}\right.\right)$$

其中 $p\left(s_{bgt}\left|S/s_{bgt};\Phi_{DE}\right.\right)$ 如 2-(3)所定义。

4) 通过蒙特卡罗近似估计预期的完整数据对数似然：

$$Q_m\left(\Phi_{DE}\left|\hat{\Phi}_{DE}{}^{(r)}\right.\right)=\frac{1}{m}\sum_{l=1}^{m}\ln f\left(Z,S_l^{(r)}\left|\Phi_{DE}\right.\right) \tag{3-(3)}$$

其中 $S_1^{(r)},\cdots,S_m^{(r)}$ 由 Gibbs 采样获得，从 $S_l^{(r)}$ 到 $S_{l+1}^{(r)}$ 中的所有条目都下列公式被按顺序更新。

$$p\left(s_{bgt}Z,S/s_{bgt};\hat{\Phi}_{DE}{}^{(r)}\right)\propto p\left(s_{bgt}\left|S/s_{bgt};\hat{\Phi}_{DE}{}^{(r)}\right.\right)f\left(z_{bgt}\left|s_{bgt}\right.\right) \tag{3-(4)}$$

5) 通过 Φ_{DE} 更新 $\hat{\Phi}_{DE}^{(r+1)}$，直至得到最大化 3-(3)。

6) 重复步骤 4，步骤 5，直至收敛。

3.2. 后验概率估计

为了获得后验概率的估计，我们实现了一个单独的 Gibbs 采样，并通过 MCEM 算法估计模型参数。潜在状态根据 3-(2)和 3-(4)依次更新，关于对 DE 基因的推断，我们采用基于后验概率的 FDR 定义(Newton *et al.* [13]；Li，Wei and Maris [14])。局部后验 f.d.r $q_{bgt}=p\left(s_{bgt}=0\left|Z\right.\right)$ 由 Gibbs 采样估计得来。$q_{(s)}$ 为 q_{bgt} 的升序排序值。让 $k=\max\left\{t:\frac{1}{t}\sum_{s=1}^{t}q_{(s)}\le\alpha\right\}$，然后我们拒绝所有零假设 $H_{(s)}$，$s=1,\cdots,k$。

4. 仿真模拟与实例分析

4.1. 仿真模拟

为了评估我们提出的 MRF 模型的性能本节进行了仿真模拟。在模拟中通过 3 轮重复次数生成了在 3 个区域，3 个时间点的 500 个基因的数据。我们考虑了两种模拟设置模拟的：

模拟设置 1：用 Gibbs 采样器模拟潜状态阵列。采样器从一个可能是 EE 或 DE 随机数组开始，在每一轮吉布斯抽样中，每个潜在状态都根据 3-(7)顺序更新一次。我们进行了三轮吉布斯抽样，以获得潜在的状态数组 S。参数为：$\gamma_{DE}=-0.10$，$\beta_{pp}=0.31$，$\beta_{nn}=0.52$，$\beta_{pn}=0.06$，$\beta_t=0.14$。

模拟设置 2：在时间点 1 中，所有基因的死亡概率均为 0.1。隐马尔可夫状态随时间模型改变：如果一个基因在 t−1 时刻为 DE，t 时刻变成 EE 概率为 0.5；为了保持 DE 基因的数量不变，我们随机选择相同数量的 EE 基因在 t 时刻切换到 DE。三层的潜在状态最初设置是相同的。然后随机选择 DE 状态的不同比例(0.1、0.2、0.5)切换为 EE；相同数量的 EE 状态也切换为 DE，以保持 DE 基因总数不变。

我们在敏感性、特异性和 FDR 这三个方面比较了这两个模型(见表 1)。对于 MRF 和 EB 模型，我们选择了 $\alpha=0.01$，满足 p-values ≤0.05，控制 FDR 小于 0.05，与 EB 相比，我们的方法(MRF)了更高的灵敏度。而当相似度降低时，即，从 HMM(0.1)到 HMM(0.5)，FDR 在我们的模型中略有增加，但在 EB 模型中并没有完全增加。

Table 1. Comparison between MRF model and EB model
表 1. MRF 与 EB 的比较

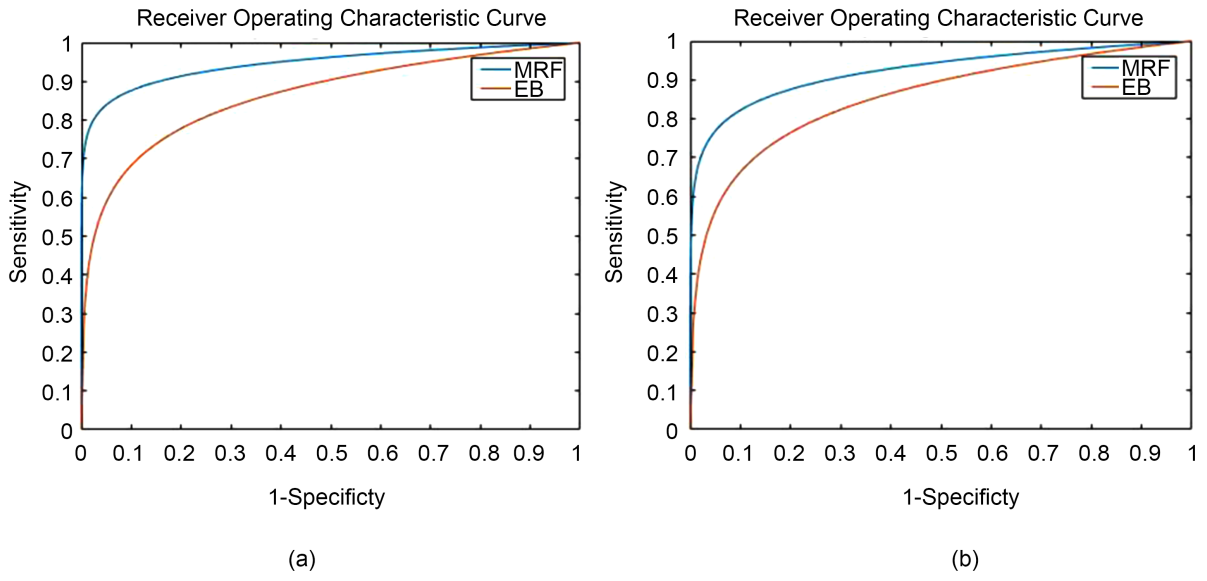
Simulation	Model	Sensitivity	Specificity	FDR
MRF	MRF	0.654(0.031)	0.997(0.001)	0.035(0.012)
	EB	0.500(0.035)	0.998(0.001)	0.037(0.011)
HMM(0.1)	MRF	0.698(0.042)	0.998(0.001)	0.037(0.011)
	EB	0.498(0.051)	0.998(0.001)	0.035(0.010)
HMM(0.2)	MRF	0.628(0.040)	0.997(0.001)	0.037(0.011)
	EB	0.496(0.045)	0.998(0.001)	0.035(0.010)
HMM(0.5)	MRF	0.535(0.036)	0.997(0.001)	0.049(0.013)
	EB	0.511(0.037)	0.998(0.001)	0.038(0.010)

4.2. 实例分析

本文数据来自小鼠标本的组织特异性表达数据数据集，编号为 GSE77943。

我们应用 MRF 模型来识别相邻时期之间的 DE 基因。运行 20 次 MCEM 算法的迭代，Gibbs 采样器的设置为 500/1500，估计的 MRF 参数为 $\gamma_{DE}=-0.21$，$\beta_{pp}=0.34$，$\beta_{nn}=0.53$，$\beta_{pn}=0.06$，$\beta_t=0.15$。胰腺与非胰腺区域系数 β_{pn} 远小于胰腺与胰腺 β_{pp} 和非胰腺与非胰腺系数 β_{nn}，表明胰腺与非胰腺区域之间的组间差异。当假设没有空间和时间依赖性时，模型简化为简单的经验贝叶斯(EB)模型，基于局部 FDR 控制程序，MRF 和 EB 模型中的阈值分别为 0.46 和 0.32。在两个模型中被鉴定为 DE 的基因数量是 30,605 (MRF)和 13,273 (EB)，其中 11,149 (84%)重叠。较高的阈值导致在 MRF 模型中识别为 DE 的更多基因。

后验概率对超参数的选择敏感，我们通过改变后局部 FDR 的阈值来计算灵敏度和特异性，我们将本文提出的 MRF 模型与 EB 模型进行了比较，该模型假设没有时间和空间依赖性。对于两组分别绘制 ROC 曲线如图 1 所示，与 EB 模型相比，MRF 模型 AUC 面积更大，MRF 模型的性能更好，MRF 模型从空间相似性中获益更多。



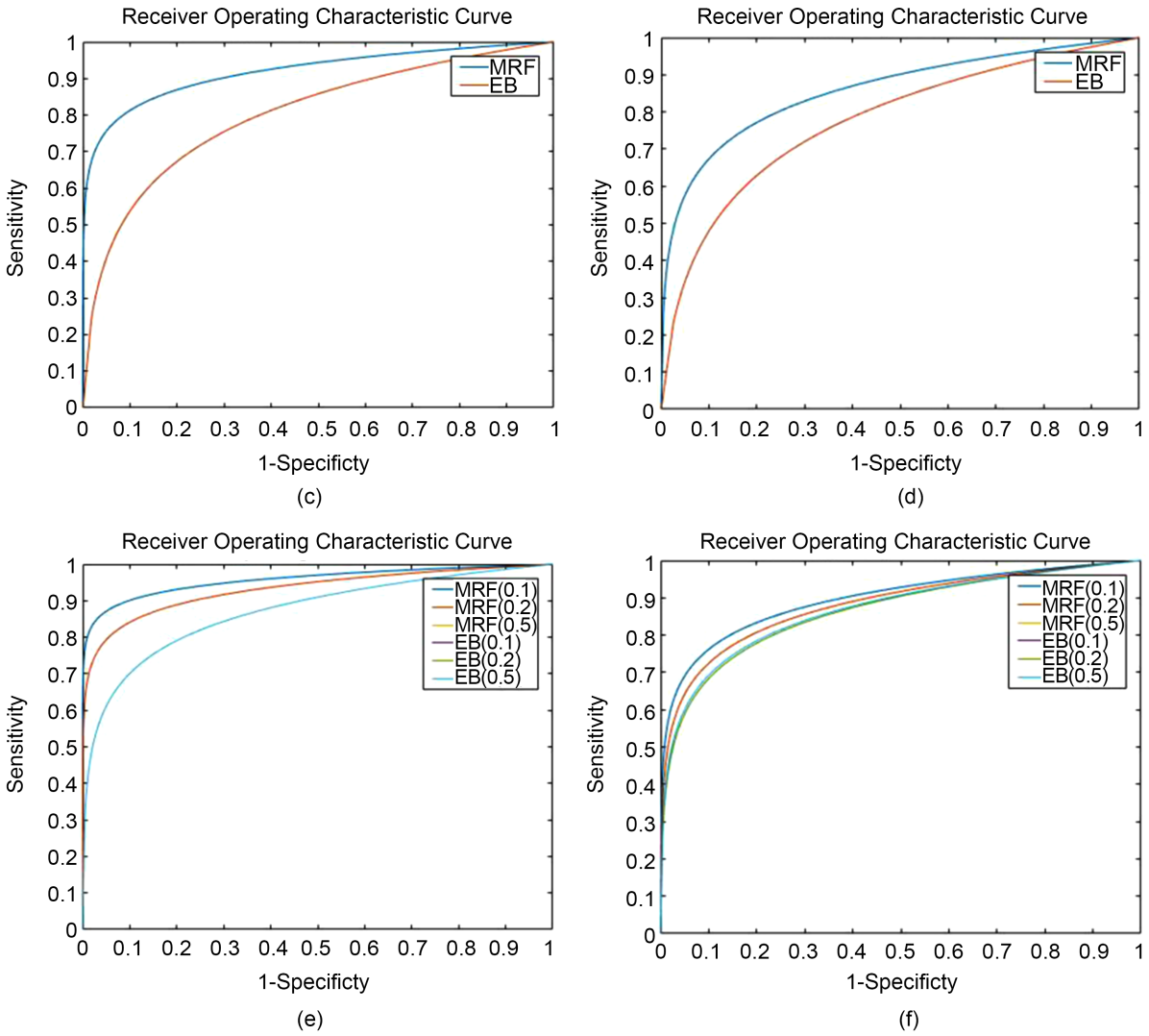


Figure 1. The ROC curve of EB and MRF model
图 1. EB 与 MRF 模型的 ROC 曲线

5. 结论

图形模型在基因表达数据的分析中应用广泛，基于基因表达数据的图形模型可以为可视化基因之间的关系和生成生物假设提供一个有用的工具，我们对图形模型之间的差异很感兴趣。在本文中，我们使用了一种贝叶斯邻域选择程序来估计高斯图形模型，结合了马尔可夫随机场来合并具有时空结构的数据。仿真研究和实例分析表明，将复杂的数据结构合并到联合建模框架中有助于估计，与常用的高斯混合模型相比，该模型具有灵敏度更高，能识别更多的 DE 基因。

参考文献

[1] Dempster, A.P. (1972) Covariance Selection. *Biometrics*, **28**, 157-175. https://doi.org/10.2307/2528966

[2] Meinshausen, N. and Bühlmann, P. (2006) High-Dimensional Graphs and Variable Selection with the Lasso. *Annals of Statistics*, **34**, 1436-1462. https://doi.org/10.1214/009053606000000281

[3] Yuan, M. and Lin, Y. (2007) Model Selection and Estimation in the Gaussian Graphical Model. *Biometrika*, **94**, 19-35.

https://doi.org/10.1093/biomet/asm018

[4] Friedman, J., Hastie, T. and Tibshirani, R. (2008) Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics*, **9**, 432-441. https://doi.org/10.1093/biostatistics/kxm045

[5] Cai, T., Liu, W. and Luo, X. (2011) A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation. *Journal of the American Statistical Association*, **106**, 594-607. https://doi.org/10.1198/jasa.2011.tm10155

[6] Dobra, A., Lenkoski, A. and Rodriguez, A. (2011) Bayesian Inference for General Gaussian Graphical Models with Application to Multivariate Lattice Data. *Journal of the American Statistical Association*, **106**, 1418-1433. https://doi.org/10.1198/jasa.2011.tm10465

[7] Wang, H. (2012) Bayesian Graphical Lasso Models and Efficient Posterior Computation. *Bayesian Analysis*, **7**, 867-886. https://doi.org/10.1214/12-BA729

[8] Efron, B. (2004) Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. *Journal of the American Statistical Association*, **99**, 96-104. https://doi.org/10.1198/016214504000000089

[9] Besag, J. (1986) On the Statistical Analysis of Dirty Pictures. *Journal of the Royal Statistical Society*: *Series B* (*Methodological*), **48**, 259-279. https://doi.org/10.1111/j.2517-6161.1986.tb01412.x

[10] Lin, Z., Sanders, S.J., Li, M., *et al.* (2015) A Markov Random Field-Based Approach to Characterizing Human Brain Development Using Spatial-Temporal Transcriptome Data. *The Annals of Applied Statistics*, **9**, 429-451. https://doi.org/10.1214/14-AOAS802

[11] Wei, G.C.G. and Tanner, M.A. (1990) A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, **85**, 699-704. https://doi.org/10.1080/01621459.1990.10474930

[12] Besag, J. (1974) Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society*: *Series B* (*Methodological*), **36**, 192-225. https://doi.org/10.1111/j.2517-6161.1974.tb00999.x

[13] Newton, M.A., Kendziorski, C.M., Richmond, C.S., *et al.* (2001) On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data. *Journal of Computational Biology*, **8**, 37-52. https://doi.org/10.1089/106652701300099074

[14] Li, H., Wei, Z. and Maris, J. (2010) A Hidden Markov Random Field Model for Genome-Wide Association Studies. *Biostatistics*, **11**, 139-150. https://doi.org/10.1093/biostatistics/kxp043