中国科学: 信息科学 2020 年 第 50 卷 第 8 期: 1217–1238
SCIENTIA SINICA Informationis
论文

基于共享随机效应和特异稀疏效应的混合多任务学习模型

彭毫[1], 王睢[2*], 王尧[3*]

1. 西南财经大学工商管理学院, 成都 611130
2. 电子科技大学经济与管理学院, 成都 611731
3. 西安交通大学管理学院, 西安 710049
* 通信作者. E-mail: wangju@usetc.com, yao.s.wang@gmail.com

收稿日期: 2019–01–28; 修回日期: 2019–04–22; 接受日期: 2019–06–05; 网络出版日期: 2020–08–05

国家自然科学基金 (批准号: 71472023, 11501440) 资助项目

摘要 在多任务学习问题中, 随机效应 (random effects) 可能同时存在于所有子任务中, 而每个子任务又存在对应的稀疏效应 (sparse effects). 这在文本分析尤其在对电影评论的情感分析中, 尤为常见. 在本文中, 我们提出一种用于数据中同时存在共享随机效应和特定稀疏效应的混合多任务学习模型, 并命名为 MSS (multi-task learning with shared random effects and specific sparse effects) 模型. 在模型的建立过程中, 我们利用 Bayes 框架, 针对不同效应的特点设定不同的先验分布和超参数. 在模型的求解过程中, 我们使用变分推断克服 Bayes 推断中的计算难题, 使 MSS 模型在大规模数据分析中具备广泛的适应性. 通过全面的模拟数据实验和真实数据实验的分析结果, 我们展示了 MSS 模型在模型预测和变量选择方面同时具备随机效应模型 (random effects models) 和稀疏回归模型 (sparse regression models) 的优势, 相比已有方法大幅提高泛化性能. MSS 模型通过对多任务学习模型中不同效应的区分, 能够更加有效的识别模型中的共享随机效应和特异稀疏效应, 进而增强模型在模型预测和变量选择方面的性能.

关键词 多任务学习, 随机效应, 稀疏性, 变量选择, Bayes 推断

1 引言

高维数据回归被广泛应用于现代数据分析的相关研究中. 在很多实际应用中, 通常假设自变量 (特征) 具有稀疏性, 这意味着数目庞大的自变量中, 只有一小部分对目标变量有真实影响. 大量关于变量选择的文献都以稀疏性假设为前提. 在这样的假设下, 一方面可以通过引入关于回归系数的正则项,

引用格式: 彭毫, 王睢, 王尧. 基于共享随机效应和特异稀疏效应的混合多任务学习模型. 中国科学: 信息科学, 2020, 50: 1217–1238, doi: 10.1360/N112018-00304
Peng H, Wang J, Wang Y. Multi-task learning with shared random effects and specific sparse effects (in Chinese). Sci Sin Inform, 2020, 50: 1217–1238, doi: 10.1360/N112018-00304

在实现模型回归的同时进行变量选择, 比如 Lasso [1], SCAD [2], MCP [3], 以及 $L_{1/2}$ regularization [4,5], 并利用高效的优化算法来估计参数, 诸如坐标下降 [6]、梯度下降 [7] 和 ADMM [8] 等; 另一方面可以利用 Bayes 方法的框架来建立回归模型. 基于 Bayes 框架的变量选择模型通过构造具有稀疏性结构的先验分布来建立回归方程, 例如 Laplace 先验分布 [9~11] 和 Spike-Slab 先验分布 (RSS) [12~15]. 通常情况下, Bayes 方法可以应用基于蒙特卡洛马尔可夫链 (Markov chain Monte Carlo, MCMC) [16,17] 的抽样方法来求解, 但 MCMC 方法对计算量的高度依赖限制了其在实际应用中的推广, 这尤其体现在对结构复杂的多层 Bayes 模型的求解中. 与此同时, 基于对后验分布进行逼近的变分推断 [18], 凭借其高效的求解效率让 Bayes 方法在涉及大规模数据的实际应用中得到了越来越广泛的关注 [19~24].

有关高维数据的多任务回归模型 [25] 同样是当今数据科学关注的重点. 多任务回归通过在有一定相关关系的子任务之间设立共享效应来提高子任务的预测精度和变量选择的准确率. 这其中包括假设子任务之间具有共享稀疏效应, 并通过加入 l_1/l_q $(q > 1)$ 块稀疏正则项 (block-sparse regularization) 来建立模型并求解 [26,27]; 也包括假设子任务之间具有共享稀疏效应的同时, 各个子任务还具有特异稀疏效应, 例如浑浊模型 (dirty model) [28], 并通过对共享稀疏效应加入块稀疏正则项, 对特异稀疏效应加入元素稀疏正则项 (element-wise regularization) 来建立模型并求解; 而数据共享 Lasso (data shared Lasso) [29] 则通过对共享稀疏效应和加权特异稀疏效应建立正则项, 以此调节共享效应与特异效应对目标变量的影响, 并通过扩展设计矩阵以实现直接调用 Lasso 求解器对模型求解.

在前文所述基于稀疏性假设的回归问题中, 通常只考虑了稀疏效应, 即假设只有一小部分自变量对目标变量存在显著的影响, 而忽略了可能存在的随机效应, 即所有自变量都只对目标变量存在非常微弱的影响, 但联合起来对目标变量存在显著的影响 [30,31]. 这样的假设在分析实际问题的时候往往太过于理想. 例如, 对电影评论的情感分析中, 除了考虑少数关键词对电影评分的影响, 还应考虑大量常用词汇整体对电影评分施加的影响. 这些数量庞大的常用词汇虽然单独来看对评分只存在非常微弱的影响, 但这些微弱效应联合起来对电影得分有可能存在显著影响. 若将这一部分影响考虑进模型中, 将有利于提升模型预测的精度和关键词选择的准确率. 由于有关电影评论的数据往往涉及到不同类型的电影, 比如喜剧类、历史类、恐怖类、传记类、动作类和爱情类等. 在每一种电影类型的数据中, 都包含了电影评论和每条评论对应的电影评分, 这些信息来自成千上万的观看了电影的用户. 因此, 对电影评论的情感分析可以看作一个多任务学习模型. 在该模型中, 常用词汇的共同作用对电影评分的影响具有这样的特征: 一方面, 不同类型的电影评论存在相关性, 因为这些评论都包含了大量相同的常用词汇. 我们可以认为这部分由常用词汇产生的影响对应的信号在所有类型的电影评论中都是存在的, 也就是说在多任务学习中, 这部分信号对于所有任务来说是共享的. 另一方面, 因为这些常用词汇数量较大, 所以我们假设这部分信号是非稀疏的. 进一步的, 除了常用词汇对评分的影响, 不同类型的电影对应的评论又包含属于该类型电影的关键词汇, 例如 “温暖” 对爱情类电影的评分有较大的正面影响, “恐怖” 对恐怖类电影的评分有较大的正面影响. 这些关键词只对某一类型的电影评分存在显著影响, 也就是说在多任务学习中, 这部分信号只对特定任务产生影响. 同时, 这一部分关键词数量较少, 在所有用于评论的词汇中占比极小, 我们假设这部分信号是稀疏的. 基于以上两点, 我们希望建立一种统计模型能够同时考虑到多任务学习中, 不同任务所属数据存在的共性 (非稀疏的) 和特性 (稀疏的).

综上, 在本文中我们考虑如下多任务学习模型: 一方面, 所有任务都共享一部分信号, 这部分信号单个来看对目标变量作用不明显, 但是其联合效应对目标变量存在显著影响, 且是非稀疏的, 我们称之为 “共享随机效应”; 另一方面, 每个任务还受到一部分 “特异稀疏效应” 的影响, 这部分信号只对特定任务产生显著影响, 且具有稀疏性. 上述多任务学习模型同时存在 “共享随机效应” 和 “特定稀疏效

应", 我们将其命名为 MSS (multi-task learning with shared random effects and specific sparse effects) 模型. 在模型的建立过程中, 我们利用 Bayes 方法, 为共享信号构造 Gauss 先验分布, 为每个任务的特征信号构造具有稀疏属性的 Spike-Slab 先验分布. 通过上述模型设定, 我们可以利用全部数据有效地刻画共享信号. 与此同时, 在剔除掉共享信号的影响后, 也能更精确地刻画每个任务对应的特征信号. 在以往的文献中, 求解 Bayes 模型通常利用 MCMC 方法, 然而随着变量维度的增加, MCMC 方法对计算量要求极大, 甚至无法求解. 在 MSS 模型中, 我们通过使用变分 Bayes 推断, 更具体的, 变分 EM 算法 (variational expectation-maximization) 来估计多任务学习中共享信号和特征信号对应的参数. 模拟实验和真实数据的分析结果表明, 变分 Bayes 推断不仅能很好地解决由数据量过大产生的计算难题, 同时也能很好地控制参数估计的误差.

本文的章节安排如下: 第 2 节对 MSS 模型进行建立和推导; 第 3 节通过多种设定下的模拟数据实验来评价 MSS 模型的性能, 并将其用于对真实数据的分析.

2 模型设定

2.1 建立 MSS 回归模型

假设有如下包含 n 个样本的训练集 $\{\boldsymbol{X}, \boldsymbol{y}\}$, 这里 $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ 表示自变量矩阵 (设计矩阵), p 表示自变量的维度, $\boldsymbol{y} \in \mathbb{R}^n$ 表示由目标变量组成的向量. 为不失一般性并消除截距项的影响, $\boldsymbol{y}$ 和 $\boldsymbol{X}$ 都已按列中心化. 考虑如下关于 $\boldsymbol{y}$ 与 $\boldsymbol{X}$ 的线性模型:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{1}$$

$\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^{\mathrm{T}}$ 是由回归系数构成的 p 维向量, 扰动项 $\boldsymbol{e} \in \mathbb{R}^n$ 相互独立且服从多维正态分布 $\mathcal{N}(\boldsymbol{0}, \sigma_e^2\boldsymbol{I}_n)$, 其中 $\boldsymbol{I}_n$ 是 $n \times n$ 的单位矩阵. 现在假设总共有 J 个任务 (对 J 种不同类型电影的评分进行回归), 我们可以将模型 (1) 拓展为 J 个任务的多任务回归模型: 第 J 个任务对应的训练集为 $\{\boldsymbol{X}_j, \boldsymbol{y}_j\}_{j=1}^J$, 这里 $\boldsymbol{X}_j \in \mathbb{R}^{n_j\times p}$ 表示第 j 个任务的设计矩阵, $\boldsymbol{y}_j \in \mathbb{R}^{n_j}$ 是与之对应的目标变量组成的向量, n_j 与 p 分别表示第 j 个任务的样本数量与自变量维度. 需要注意的是上述 J 个任务中, 每一个任务的自变量维度都满足是 p 维的. 综上所述, 得到如下多任务回归模型:

$$\boldsymbol{y}_j = \boldsymbol{X}_j(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_j) + \boldsymbol{e}_j, \quad j = 1, \ldots, J, \tag{2}$$

$\boldsymbol{\beta}_0 = [\beta_{01}, \ldots, \beta_{0p}]^{\mathrm{T}} \in \mathbb{R}^p$ 表示共享随机效应的效应值 (effect size) 组成的向量, $\boldsymbol{\beta}_j = [\beta_{j1}, \ldots, \beta_{jp}]^{\mathrm{T}} \in \mathbb{R}^p$ 表示属于第 j 个任务的特定稀疏效应的效应值组成的向量, $\boldsymbol{e}_j \in \mathbb{R}^{n_j}$ 是第 j 个任务对应的随机扰动项, 其相互独立且服从期望为 $\boldsymbol{0}$, 方差 – 协方差矩阵为 $\sigma_j^2\boldsymbol{I}_{n_j}$ 的多维正态分布 $\mathcal{N}(\boldsymbol{0}, \sigma_j^2\boldsymbol{I}_{n_j})$, 其中 $\boldsymbol{I}_{n_j}$ 是 $n_j \times n_j$ 的单位矩阵.

进一步的, 考虑到两种效应的均值均不会出现大幅偏离 0 的情况, 我们假设共享随机效应的效应值 $\boldsymbol{\beta_0}$ 来自期望为 $\boldsymbol{0}$ 的 Gauss 先验分布:

$$\boldsymbol{\beta_0} \sim \mathcal{N}(\boldsymbol{0}, \sigma_{\boldsymbol{\beta}_0}^2\boldsymbol{I}_p), \tag{3}$$

其中 $\mathcal{N}(\boldsymbol{0}, \sigma_{\boldsymbol{\beta}_0}^2\boldsymbol{I}_p)$ 表示期望为 $\boldsymbol{0}$, 方差 – 协方差矩阵为 $\sigma_{\boldsymbol{\beta}_0}^2\boldsymbol{I}_p$ 的多维 Gauss 分布. 而对于特定稀疏效应的效应值, 我们希望引入二项变量 γ_{jk} 表示第 j 个任务中第 k 个效应值 β_{jk} 是否为 0, 因此假设 β_{jk}

来自如下 Spike-Slab 先验分布:

$$\beta_{jk}|\gamma_{jk},\sigma^2_{\boldsymbol{\beta}_j}\sim\begin{cases}\mathcal{N}(0,\sigma^2_{\boldsymbol{\beta}_j}), & \text{if } \gamma_{jk}=1,\\ \delta_0(\beta_{jk}), & \text{if } \gamma_{jk}=0,\end{cases}\tag{4}$$

其中 $\mathcal{N}(0,\sigma^2_{\boldsymbol{\beta}_j})$ 表示期望为 0, 方差为 $\sigma^2_{\boldsymbol{\beta}_j}$ 的 Gauss 分布. $\delta_0(\beta_{jk})$ 表示在 0 点处的 Dirac 函数. 上述定义表明当第 j 个任务中的第 k 个自变量对目标变量施加了影响时, 也就是当 $\gamma_{jk}=1$ 时, β_{jk} 来自 Gauss 分布 $\mathcal{N}(0,\sigma^2_{\boldsymbol{\beta}_j})$, 否则, β_{jk} 等于 0. 同时假设第 j 个任务中, γ_{jk} 来自参数为 π_j 的Bernoulli 分布 $\pi_j^{\gamma_{jk}}(1-\pi_j)^{1-\gamma_{jk}}$, 也即 $\Pr(\gamma_{jk}=1)=\pi_j$.

由于 Dirac 函数可能在后续推导过程中带来不必要的困难, 为此我们对等式 (4) 重新参数化 [21]

$$\beta_{jk}\sim\mathcal{N}(0,\sigma^2_{\boldsymbol{\beta}_j}),\quad \gamma_{jk}\sim\pi_j^{\gamma_{jk}}(1-\pi_j)^{1-\gamma_{jk}},\tag{5}$$

这样, 等式 (5) 中 γ_{jk} 与 β_{jk} 的乘积 $\gamma_{jk}\beta_{jk}$ 与等式 (4) 中的 β_{jk} 具有完全一致的分布. 于是 $\boldsymbol{y}_j$ 的分布可以参数化为:

$$\boldsymbol{y}_j\sim\mathcal{N}\left(\boldsymbol{X}_j\boldsymbol{\beta}_0+\sum_k\gamma_{jk}\beta_{jk}\boldsymbol{x}_{jk},\sigma_j^2\boldsymbol{I}_{n_j}\right),\quad j=1,\dots,J,\tag{6}$$

其中 $\boldsymbol{x}_{jk}$ 表示第 j 个设计矩阵 $\boldsymbol{X}_j$ 的第 k 列.

为了使表述更加简洁, 我们对模型中的随机变量做如下定义: $\boldsymbol{\beta}=\{\boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_J\}$, $\boldsymbol{\gamma}=\{\boldsymbol{\gamma}_1,\dots,\boldsymbol{\gamma}_J\}$, $\boldsymbol{y}=\{\boldsymbol{y}_1,\dots,\boldsymbol{y}_J\}$, $\boldsymbol{X}=\{\boldsymbol{X}_1,\dots,\boldsymbol{X}_J\}$. 同时定义 $\boldsymbol{\theta}=\{\sigma^2_{\boldsymbol{\beta}_0},\sigma^2_{\boldsymbol{\beta}_j},\sigma_j^2,\pi_j\}$, $j=1,\dots,J$, 为模型 (2) 中参数的集合. 随机变量 $\boldsymbol{y}$, $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ 的联合分布可以表示如下:

$$\begin{aligned}&\Pr(\boldsymbol{y},\boldsymbol{\beta}_0,\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{X};\boldsymbol{\theta})\\&=\prod_j\left(\Pr(\boldsymbol{y}_j|\boldsymbol{\beta}_0,\boldsymbol{\beta}_j,\boldsymbol{\gamma}_j;\boldsymbol{\theta})\Pr(\boldsymbol{\beta}_j,\boldsymbol{\gamma}_j|\boldsymbol{\theta})\right)\Pr(\boldsymbol{\beta}_0|\boldsymbol{\theta})\\&=\prod_j\left(\mathcal{N}\big(\boldsymbol{X}_j(\boldsymbol{\beta}_0+\boldsymbol{\beta}_j\odot\boldsymbol{\gamma}_j),\sigma_j^2\boldsymbol{I}_{n_j}\big)\mathcal{N}(\boldsymbol{\beta}_j|\boldsymbol{0},\sigma^2_{\boldsymbol{\beta}_j}\boldsymbol{I}_p)\prod_k\pi_j^{\gamma_{jk}}(1-\pi_j)^{1-\gamma_{jk}}\right)\mathcal{N}(\boldsymbol{\beta}_0|\boldsymbol{0},\sigma^2_{\boldsymbol{\beta}_0}\boldsymbol{I}_p),\end{aligned}\tag{7}$$

其中 $\boldsymbol{\beta}_j\odot\boldsymbol{\gamma}_j$ 表示 $\boldsymbol{\beta}_j$ 和 $\boldsymbol{\gamma}_j$ 对应元素的乘积组成的新向量. 模型 (2) 可以表示为如图 1 所示的图模型 [32].

在等式 (7) 中, $\boldsymbol{\beta_0}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ 可以看作所谓的潜在变量, 对潜在变量求积分, 可以得到边缘分布:

$$\begin{aligned}\Pr(\boldsymbol{y}|\boldsymbol{X};\boldsymbol{\theta})=\sum_{\boldsymbol{\gamma}}\int\int\prod_j\Big(&\mathcal{N}\big(\boldsymbol{X}_j(\boldsymbol{\beta}_0+\boldsymbol{\beta}_j\odot\boldsymbol{\gamma}_j),\sigma_j^2\boldsymbol{I}_{n_j}\big)\mathcal{N}(\boldsymbol{0},\sigma^2_{\boldsymbol{\beta}_j}\boldsymbol{I}_p)\\&\cdot\prod_k\pi_j^{\gamma_{jk}}(1-\pi_j)^{1-\gamma_{jk}}\Big)\mathcal{N}(\boldsymbol{0},\sigma^2_{\boldsymbol{\beta}_0}\boldsymbol{I}_p)\,\mathrm{d}\boldsymbol{\beta}_0\,\mathrm{d}\boldsymbol{\beta}.\end{aligned}\tag{8}$$

求解 $\boldsymbol{\theta}$ 将等式 (8) 最大化, 可以得到 $\boldsymbol{\theta}$ 的估计 $\hat{\boldsymbol{\theta}}$, 同时可以得到潜在变量的后验分布:

$$\Pr(\boldsymbol{\beta}_0,\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{y},\boldsymbol{X};\hat{\boldsymbol{\theta}})=\frac{\Pr(\boldsymbol{y},\boldsymbol{\beta}_0,\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{X};\hat{\boldsymbol{\theta}})}{\Pr(\boldsymbol{y}|\boldsymbol{X};\hat{\boldsymbol{\theta}})}.\tag{9}$$

2.2 模型求解

试图得到边缘分布等式 (8) 的精确解析形式几乎是不可能的, 因为 γ_j $(j=1,\dots,J)$ 之间不一定独立, 使得处理等式 (8) 中的积分非常困难. 为了克服这一困难, 我们使用变分 EM 算法 (variational

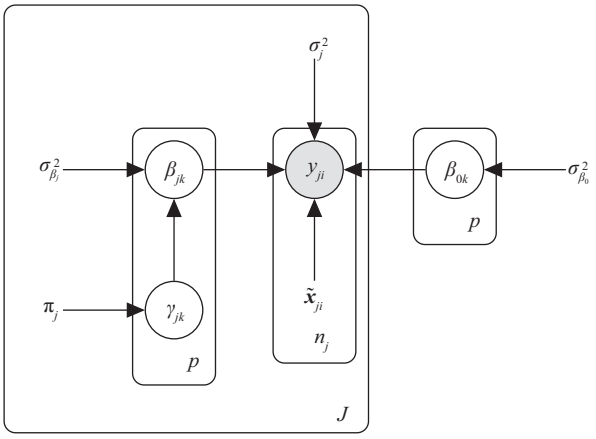

图 1 图模型表示的 MSS 模型对应的联合分布等式 (7). 其中 $\tilde{\boldsymbol{x}}_{ji}$ 表示第 j 个设计矩阵 $\boldsymbol{X}_j$ 的第 i 行, y_{ji} 表示对应的目标变量. 在图模型中, 每个节点代表一个变量, 实心圆表示观测到的随机变量, 空心圆表示潜在随机变量, 其余的表示参数或常量. 随机变量之间的有向箭头表示变量之间的条件依赖关系. 方框的右下标表示方框内变量的具体数目

Figure 1 Graphical model representation of joint distribution Eq. (7). Here $\tilde{\boldsymbol{x}}_{ji}$ is the i-th row of the design matrix $\boldsymbol{X}_j$ and y_{ji} is the coressponding response variable. In this graphical model, we introduce a node for each of the variables. We denote latent variables by open circles and observed variables by shading the corresponding circles. The others are deterministic parameters or constant variables. Links express probabilistic relationships between these variables. We have introduced a plate labeled with a number represents the number of nodes of this kind

EM algorithm) 来逼近等式 (8) 的真实最大值 [32]. 为实现上述目标, 先假设 $q(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\gamma})$ 是真实后验分布 $\Pr(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\theta})$ 的近似分布, 根据 Jensen 不等式, 可以很容易得到取对数后的边缘分布等式 (8) 的下界 $L(q)$:

$$\begin{aligned}\log \Pr(\boldsymbol{y}|\boldsymbol{X};\boldsymbol{\theta}) &= \log \sum_{\boldsymbol{\gamma}} \int\int \Pr(\boldsymbol{y}, \boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{X};\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\beta}_0\, \mathrm{d}\boldsymbol{\beta} \\ &\geqslant \sum_{\boldsymbol{\gamma}} \int\int q(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\gamma}) \log \frac{\Pr(\boldsymbol{y}|\boldsymbol{X};\boldsymbol{\theta})}{q(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\gamma})}\, \mathrm{d}\boldsymbol{\beta}_0\, \mathrm{d}\boldsymbol{\beta} \\ &= L(q), \end{aligned} \tag{10}$$

其中等号成立的条件是当且仅当 $q(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\gamma})$ 与真实后验分布式等式 (9) 完全相等. 因为 $L(q)$ 是等式 (8) 的下界, 我们可以通过迭代的方法最大化 $L(q)$ 直到其收敛, 并将其看作是对等式 (8) 最大值的逼近. 在这一迭代过程中, 为了计算上的可行, 我们给出如下假设: $q(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\gamma})$ 是可分解的, 并具有如下形式:

$$q(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_k q(\beta_{0k}) \left(\prod_j \prod_k \left(q(\beta_{jk}|\gamma_{jk}) q(\gamma_{jk}) \right) \right). \tag{11}$$

上述可分解假设是我们在变分推断中的唯一假设. 通过进一步的推导 (详细推导请见附录 A), 可以得到如下结果:

$$q(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_k q(\beta_{0k}) \left(\prod_j \prod_k \left(q(\beta_{jk}|\gamma_{jk}) q(\gamma_{jk}) \right) \right)$$

$$=\prod_k \mathcal{N}\left(\mu_{0k}, s_{0k}^2\right)\left(\prod_j \prod_k\left(\left(\alpha_{jk}\mathcal{N}(\mu_{jk}, s_{jk}^2)\right)^{\gamma_{jk}}\left((1-\alpha_{jk})\mathcal{N}(0, \sigma_{\boldsymbol{\beta}_j}^2)\right)^{1-\gamma_{jk}}\right)\right), \tag{12}$$

其中

$$\begin{aligned}
s_{0k}^2 &= -\frac{1}{2}\left(\sum_j -\frac{1}{2\sigma_j^2}\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jk} - \frac{1}{2\sigma_{\boldsymbol{\beta}_0}^2}\boldsymbol{I}\right)^{-1}, \\
\mu_{0k} &= s_{0k}^2 \sum_j -\frac{1}{\sigma_j^2}\left(\sum_{l\neq k}\boldsymbol{x}_{jl}\mathbb{E}[\beta_{0l}] + \boldsymbol{X}_j\mathbb{E}_{\boldsymbol{\beta}_j,\boldsymbol{\gamma}_j}[(\boldsymbol{\gamma}_j \odot \boldsymbol{\beta}_j)] - \boldsymbol{y}_j\right)^{\mathrm{T}}\boldsymbol{x}_{jk}, \\
s_{jk}^2 &= \frac{\sigma_j^2}{\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jk} + \frac{\sigma_j^2}{\sigma_{\boldsymbol{\beta}_j}^2}}, \\
\mu_{jk} &= \frac{\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{y}_j - \sum_{l\neq k}\mathbb{E}_{jl}[\gamma_{jl}\beta_{jl}]\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jl} - \boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{X}_j\mathbb{E}_{\boldsymbol{\beta}_0}[\boldsymbol{\beta}_0]}{\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jk} + \frac{\sigma_j^2}{\sigma_{\boldsymbol{\beta}_j}^2}},
\end{aligned} \tag{13}$$

以及

$$\alpha_{jk} = \frac{1}{1+\exp(-u_{jk})}, \quad u_{jk} = \frac{\mu_{jk}^2}{2s_{jk}^2} + \frac{1}{2}\log\frac{s_{jk}^2}{\sigma_{\boldsymbol{\beta}_j}^2} + \log\left(\frac{\pi_j}{1-\pi_j}\right), \tag{14}$$

这里 $\boldsymbol{x}_{jk}$ 表示设计矩阵 $\boldsymbol{X}_j$ 的第 k 列.

由于 $q(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\gamma})$ 是对真实后验分布等式 (9) 的近似, 通过观察等式 (11) 和 (12), 可以将 $q(\gamma_{jk}=1)=\alpha_{jk}$, $q(\beta_{jk}|\gamma_{jk}=1)=\mathcal{N}(\mu_{jk}, s_{jk})$ 和 $q(\beta_{jk}|\gamma_{jk}=0)=\mathcal{N}(0, \sigma_{\boldsymbol{\beta}_j}^2)$ 分别看作对 $\Pr(\gamma_{jk}=1|\boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\theta})$, $\Pr(\beta_{jk}|\gamma_{jk}=1|\boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\theta})$ 和 $\Pr(\beta_{jk}|\gamma_{jk}=0|\boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\theta})$ 的近似. 值得注意的是, 当 β_{jk} 对应的自变量对目标变量没有影响时, 即 $\gamma_{jk}=0$ 时, 其后验分布就是先验分布, 这也是符合逻辑的. 同时 $q(\beta_{0k})=\mathcal{N}(\beta_{0k}, s_{0k}^2)$ 可以看作是对后验分布 $\Pr(\beta_{0k}|\boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\theta})$ 的近似, 其受到先验分布和 J 个任务中的所有数据的共同影响.

在得到等式 (12) 后, 将其带入等式 (10) 可以很容易得到对数边缘分布下界 $L(q)$ 的解析表达. 对 $L(q)$ 中的参数集 $\boldsymbol{\theta}$ 分别求导并令其等于零, 我们可以得到如下关于模型参数集 $\boldsymbol{\theta}$ 的更新方程 (详细推导请见附录 A):

$$\begin{aligned}
\sigma_j^2 &= \frac{1}{n_j}\left((\boldsymbol{y}_j - \tilde{\boldsymbol{y}}_j)^{\mathrm{T}}(\boldsymbol{y}_j - \tilde{\boldsymbol{y}}_j) + \sum_{k=1}^{p}\left[\alpha_{jk}(s_{jk}^2 + \mu_{jk}^2) - (\alpha_{jk}\mu_{jk})^2\right]\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jk}\right. \\
&\quad \left. + \boldsymbol{\mu}_0^{\mathrm{T}}(\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j)\boldsymbol{\mu}_0 + \mathrm{Tr}\left(\boldsymbol{S}_0^2(\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j)\right) + 2\left((\boldsymbol{\alpha}_j \odot \boldsymbol{\mu}_j)^{\mathrm{T}}(\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j) - \boldsymbol{y}_j^{\mathrm{T}}\boldsymbol{X}_j\right)\boldsymbol{\mu}_0\right), \\
\sigma_{\boldsymbol{\beta}_j}^2 &= \frac{\sum_k \alpha_{jk}(\mu_{jk}^2 + s_{jk}^2)}{\sum_k \alpha_{jk}}, \\
\sigma_{\boldsymbol{\beta}_0}^2 &= \frac{1}{p}\left(\boldsymbol{\mu}_0^{\mathrm{T}}\boldsymbol{\mu}_0 + \mathrm{Tr}(\boldsymbol{S}_0^2)\right), \\
\pi_j &= \frac{1}{p}\sum_k \alpha_{jk},
\end{aligned} \tag{15}$$

这里 $\boldsymbol{S}_0^2$ 是满足 $\boldsymbol{S}_0^2(k,k)=s_{0k}^2$ 的对角阵.

综上所述, 我们使用变分 EM 算法对 MSS 模型进行求解, 在每一步迭代中下界 $L(q)$ 的值都会单调增加, 其收敛性也可以得到保证.

2.3 变量选择与模型预测

利用上述迭代方法收敛得到的结果, 我们可以进行变量选择和模型预测.

我们利用 γ_{jk} 的后验概率 α_{jk} 来进行变量选择. α_{jk} 的值可以认为是第 j 个任务中的第 k 个特征对目标变量有影响的概率. 利用 $\beta_{jk}\gamma_{jk}$ 可以筛选出影响电影评分的关键词汇.

除了变量选择外, 我们对模型预测的结果同样感兴趣. 当知道一个新的已经中心化后的样本 $\boldsymbol{x}^{\text{new}}$ 并且知道其属于第 j 个任务后, 利用模型对新的目标变量进行预测:

$$\hat{y} = \sum_{k=1}^{p}(\mu_{0k} + \mu_{jk}\alpha_{jk})x_k^{\text{new}}.$$

用此方法可以判断用户对电影的文字评论与给出的评分是否相符, 进而可以根据用户的评论对用户的评分进行调整, 得到更为准确的电影评分.

虽然变分推断是 Bayes 推断的一种近似, 其结果也只是近似解, 但是在满足模型假设的前提下计算精度依然会好过很多传统的回归算法, 我们将在第 3 节的模拟实验和真实数据分析中验证这一观点.

3 模拟实验和真实数据分析

在本节中, 我们使用模拟实验对 MSS 算法和其他算法的结果进行比较, 接着使用真实数据来验证 MSS 算法的优势. 所有的实验数据和代码都已上传至网站[1)], 以保证实验的可重复性.

3.1 模拟实验

3.1.1 与单任务模型的对比

我们将 MSS 模型与一些经典回归模型进行比较, 包括 Spike-Slab 先验模型 (RSS)、岭回归模型和 Lasso. 在实验中, 着重对 MSS 模型与上述模型在预测误差和变量选择两个方面进行比较. 模拟实验的条件设定如下: 我们考虑任务数 $J = 3$, 分别用 T1, T2 和 T3 表示, 每一个任务包含的样本数量分别为 $n_1 = 300$, $n_2 = 500$, $n_3 = 700$, 设计矩阵的维度 $p = 1000$. 做为对比, RSS 模型、岭回归和 Lasso 的结果包含两种设定: (1) 分别对 3 个任务进行独立回归; (2) 将 3 个任务的数据汇总后进行一次回归. 设计矩阵 $\boldsymbol{X}_j$ $(j = 1, 2, 3)$ 由 p 维的正态分布生成, 并且满足期望为 0, 方差 – 协方差矩阵满足第 (d, d') 个元素等于 $\rho^{|d-d'|}$, 用于刻画自变量之间的相关性, 相距越近的两个自变量之间相关性越高. 在实验中, ρ 被分别设置为 $\{0, 0.5\}$. 共享信号对应的效应值 $\boldsymbol{\beta_0}$ 由标准正态分布生成. 特征信号对应的效应值 $\boldsymbol{\beta}_j$ $(j = 1, 2, 3)$ 中非零项的占比为 π, 非零项由 0 均值的 Gauss 分布生成, 其方差满足 $\text{var}(\boldsymbol{X\beta}_0)/\text{var}(\boldsymbol{X\beta}_j) = 1$, 即共享信号对目标变量的作用与特征信号对目标的作用大小一致. 特征信号的稀疏性由 π 控制, 为了比较 MSS 模型与其他模型在不同稀疏性设定下的表现, 我们将 π 分别设置为 $\{0.01, 0.5, 0.1\}$, 同时将信噪比 (SNR) 分别设置为 $\{0.5, 1, 2\}$. 图 2 和 3 是在 $\rho = 0$ 设定下的实验结果, 图 4 和 5 是在 $\rho = 0.5$ 设定下的实验结果.

图 2(a)∼5(a) 对应 MSS 模型、Spike-Slab 先验模型 (RSS)、岭回归和 Lasso 在不同设定条件下的参数估计误差的结果. MSS 模型的参数估计标准化后的均方误差 (mean squared error, MSE) 定义

1) https://github.com/osbornePeng/VSSC.

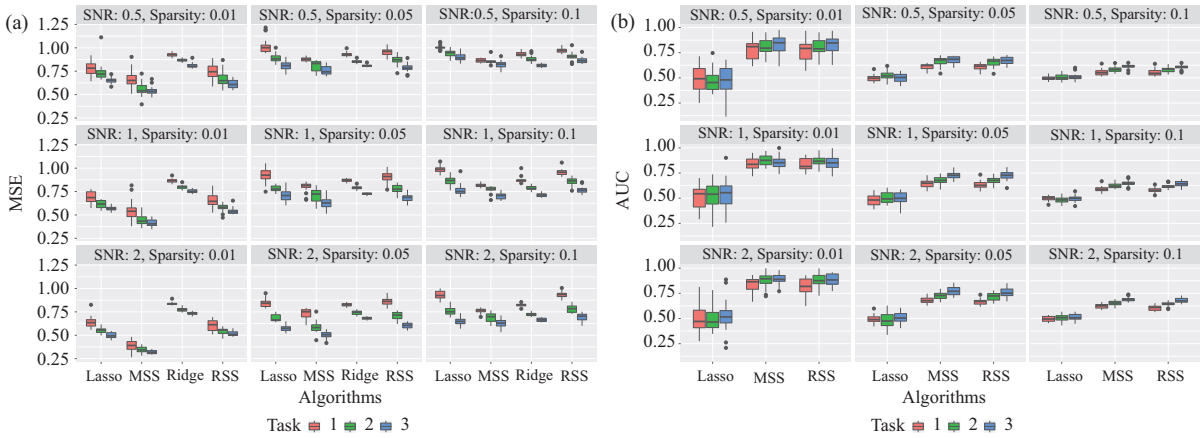

图 2 (网络版彩图) $\rho = 0$ 时, MSS 模型的结果与 RSS 模型、岭回归和 Lasso 分别学习每个任务的结果比较

Figure 2 (Color online) The comparison of MSS, RSS, ridge regression and the Lasso with $\rho = 0$ and with data sets separately from each task. (a) MSE of coefficient estimation; (b) AUC for the performance of variable selection of each algorithm

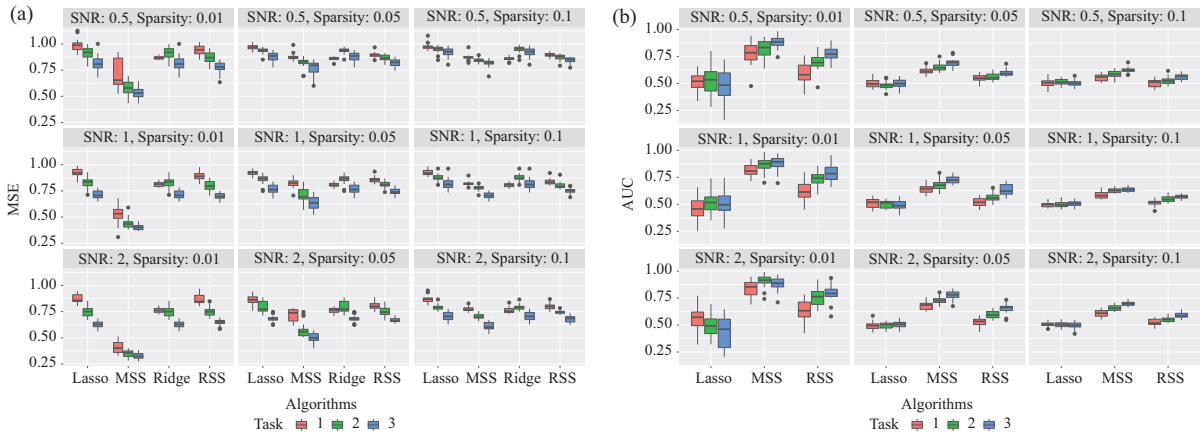

图 3 (网络版彩图) $\rho = 0$ 时, MSS 模型结果与 RSS 模型、岭回归和 Lasso 将所有任务合并为一个任务学习的结果比较

Figure 3 (Color online) The comparison of MSS, RSS, ridge regression and the Lasso with $\rho = 0$ and with data set pooled from all three tasks. (a) MSE of coefficient estimation; (b) AUC for the performance of variable selection of each algorithm

如下:

$$\frac{||\hat{\boldsymbol{\beta}}_0 + \hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_0 - \boldsymbol{\beta}_j||_2^2}{p \cdot \mathrm{var}(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_j)},$$

由于 RSS 模型、岭回归和 Lasso 没有区分共享信号和特征信号, 其参数估计标准化后的均方误差 (MSE) 定义如下:

$$\frac{||\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_0 - \boldsymbol{\beta}_j||_2^2}{p \cdot \mathrm{var}(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_j)},$$

这里 $\hat{\boldsymbol{\beta}}_0$, $\hat{\boldsymbol{\beta}}_j$ $(j = 1, 2, 3)$ 为参数的估计值, $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_j$ $(j = 1, 2, 3)$ 为参数的真实值. 4 种模型的参数估计误差随着样本数量的减少呈现上升趋势, 随着信噪比 (SNR) 的增大呈现减少的趋势. 在绝大多数测试项目中, MSS 模型都有最好的参数估计表现, 尤其是在样本量较小时, 优势更加明显.

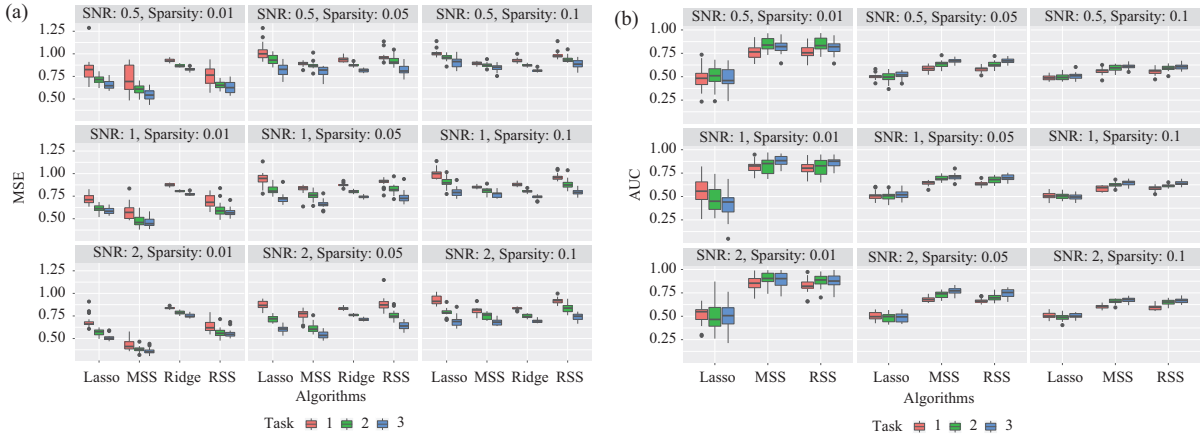

图 4 (网络版彩图) $\rho = 0.5$ 时, **MSS** 模型的结果与 **RSS** 模型、岭回归和 **Lasso** 分别学习每个任务的结果比较

Figure 4 (Color online) The comparison of MSS, RSS, ridge regression and the Lasso with $\rho = 0.5$ and with data sets separately from each task. (a) MSE of coefficient estimation; (b) AUC for the performance of variable selection of each algorithm

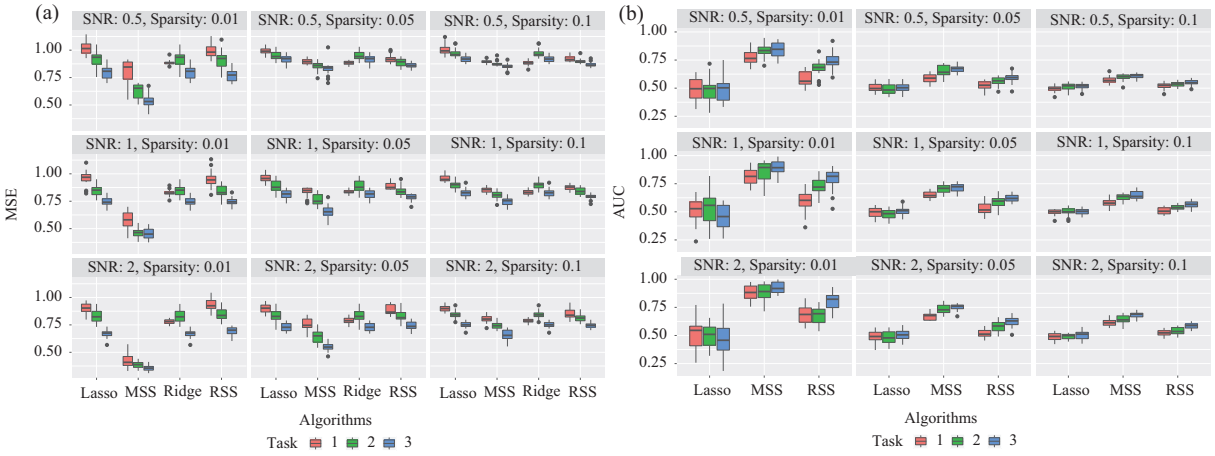

图 5 (网络版彩图) $\rho = 0.5$ 时, **MSS** 模型结果与 **RSS** 模型、岭回归和 **Lasso** 将所有任务合并为一个任务学习的结果比较

Figure 5 (Color online) The comparison of MSS, RSS, ridge regression and the Lasso with $\rho = 0.5$ and with data set pooled from all three tasks. (a) MSE of coefficient estimation; (b) AUC for the performance of variable selection of each algorithm

图 2(b)~5(b) 对应各模型变量选择的 AUC (area under curve) 值. 由于岭回归不具备变量选择的功能, 我们只对 MSS 模型、RSS 模型和 Lasso 的结果进行了比较. 通过上述结果, 可以看出 MSS 模型在各种实验条件下都有最好的表现, 尤其是在真实数据非常稀疏的时候.

3.1.2 与多任务模型的对比

接下来, 我们将 MSS 模型与一些多任务回归模型进行对比, 包括浑浊模型和数据共享 Lasso 模型 (DSL). 在实验中, 我们着重对 MSS 模型与上述模型在预测误差和变量选择两个方面进行比较. 模拟实验的条件设定如下: 任务数 $J = 5$. 考虑到浑浊模型的计算效率以及在模型设定中要求每个任务包

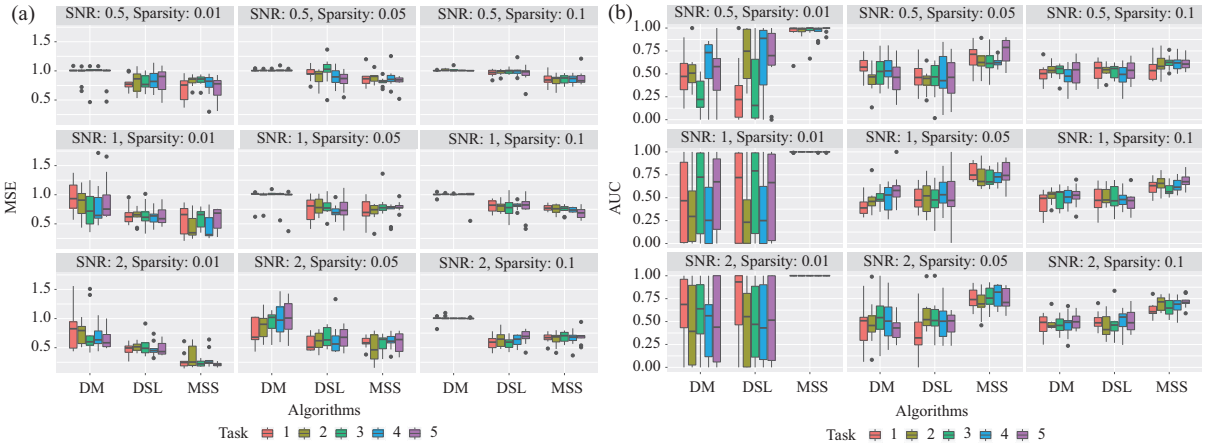

图 6 (网络版彩图) $\rho = 0$ 时, MSS 模型的结果与浑浊模型和数据共享 Lasso (DSL) 的结果比较

Figure 6 (Color online) The comparison of MSS, dirty model, and the data shared Lasso with $\rho = 0$. (a) MSE of coefficient estimation; (b) AUC for the performance of variable selection of each algorithm

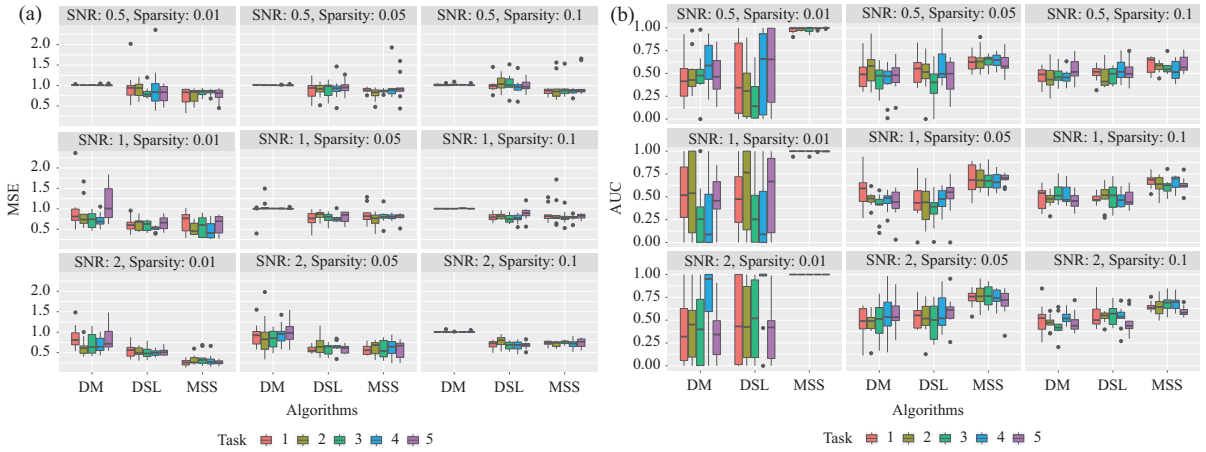

图 7 (网络版彩图) $\rho = 0.5$ 时, MSS 模型的结果与浑浊模型和数据共享 Lasso (DSL) 的结果比较

Figure 7 (Color online) The comparison of MSS, dirty model, and the data shared Lasso with $\rho = 0.5$. (a) MSE of coefficient estimation; (b) AUC for the performance of variable selection of each algorithm

含样本数相等, 并且浑浊模型中的参数在每一种设定下都须要通过交叉验证来确定, 非常耗时, 因此实验中每一个任务包含的样本数设定为 $n_j = 50$ $(j = 1, \ldots, J)$, 设计矩阵的维度 $p = 100$. 实验结果如图 6 和 7 所示. MSS 模型在大多数情况下拥有更好的预测精度和更好的变量选择准确率.

3.1.3 结果分析

MSS 模型在模型预测与变量选择方面具有更好性的性能, 得益于其同时具备随机效应模型与稀疏效应模型的特性, 更具体的, 同时具备岭回归 (ridge) 与 Spike-Slab 先验模型 (RSS) 的特性. 为了更加直观地体现 MSS 模型这一优点, 我们选择一组实验 ($\rho = 0$, $\pi = 0.01$, SNR $= 2$), 并将其第 $j = 3$ 个任务的真实回归系数与各个模型的估计值做比较, 得到图 8. 其中图 true model 表示真实的回归系数, 红点表示对应位置的 $\boldsymbol{\beta}_j$ 不为 $\mathbf{0}$, 即在此处存在特异稀疏效应. 其余波动较小的回归系数代表随机效应 (看作背景噪音). 可以看出 MSS 模型对随机效应和稀疏效应刻画的都比较准确, Spike-Slab 先验模型

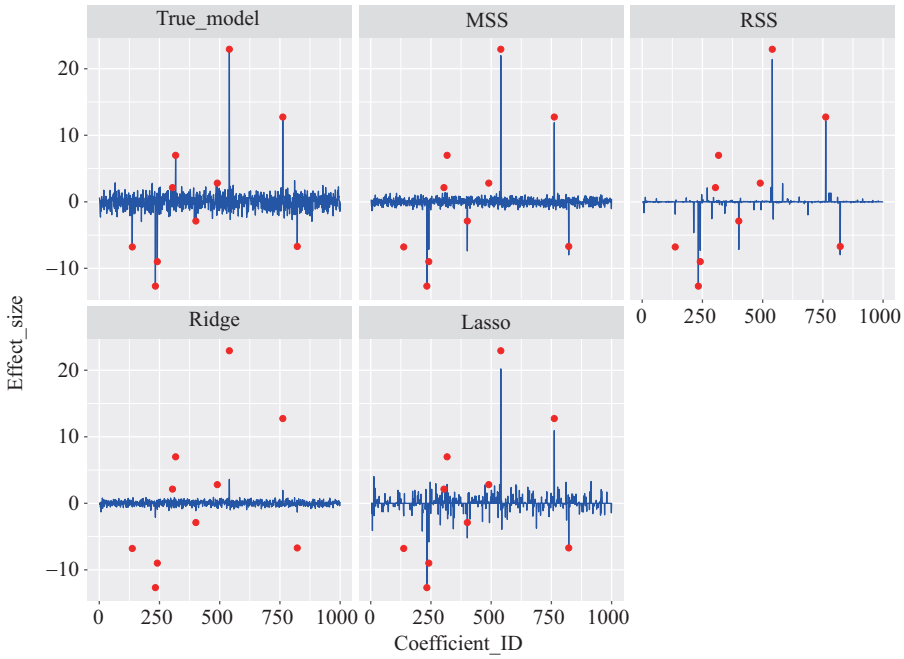

图 8 (网络版彩图) 不同模型估计的效应值的比较

Figure 8 (Color online) The comparison of different models' estimation of effect size

(RSS) 和 Lasso 只对稀疏效应刻画较好, 岭回归 (ridge) 只对随机效应刻画较好. MSS 模型同时具备随机效应模型 (ridge) 和稀疏效应模型 (RSS 和 Lasso) 的优点.

与此同时, MSS 模型也具有很好的适应性. 在上述实验中, 若我们设定共享随机效应 $\boldsymbol{\beta}_0 = \mathbf{0}$, 此时各个子任务只存在特异稀疏效应 $\boldsymbol{\beta}_j$, 理想的算法应该是对各个子任务分别利用 Spike-Slab 先验模型 (RSS) 进行回归, 我们将这一结果与 MSS 模型进行对比, 得到图 9, 从中可以看出在这一极端情况下, MSS 模型可以得到与 Spike-Slab 先验模型 (RSS) 基本一致的结果.

接下来, 我们设定特异稀疏效应 ($\boldsymbol{\beta}_j = \mathbf{0}$, $j = 1, \ldots, J$), 此时各个子任务只存在共享随机效应, 理想的算法应该是对全部子任务同时利用岭回归 (ridge) 进行回归, 我们将这一结果与 MSS 模型进行对比, 得到图 10, 从中可以看出在这一极端情况下, MSS 模型可以得到与岭回归 (ridge) 基本一致的结果.

在计算时间方面, 我们给出了 MSS 模型在不同的任务数 (tasks), 不同的子任务对应的样本数 (samples) 和不同的特征数 (features) 设定下, 经过 100 次迭代计算需花费的时间, 结论如图 11 所示 (MSS 伪代码请见附录 B). 上述计算均在一台配备 I7-3840QM CPU 和 16 G RAM 的笔记本电脑上完成. 计算时间相对于任务数, 单任务对应的样本数和特征数呈现近似线性增长的趋势.

我们将 MSS 模型与浑浊模型和数据共享 Lasso 模型 (DSL) 的计算时间进行比较. 在实验中, 我们考虑任务数 $J = 5$. 每一个任务包含的样本数 $n_j = 500$, 设计矩阵的维度依次考虑 $p = 500$, $p = 1000$, $p = 2000$. 实验结果如表 1 所示. MSS 模型有着最快的计算速度. 浑浊模型计算速度最慢, 存在样本数与维度相等的时候比维度更高时候更慢的现象, 且浑浊模型需要通过交叉验证选择参数, 实际运算时间还会成倍增长. DSL 速度比 MSS 慢, 并且需要存储扩展后的设计矩阵, 内存使用量是 MSS 模型的 $J + 1$ 倍. MSS 模型在大数据时代更加具备广泛的应用前景.

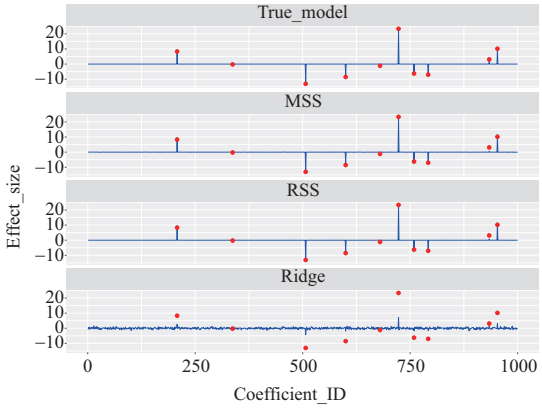

图 9 (网络版彩图) 当共享随机效应 $\beta_0 = 0$ 时不同模型估计的效应值的比较

Figure 9 (Color online) The comparison of different models' estimation of effect size when the shared random effect $\beta_0 = 0$

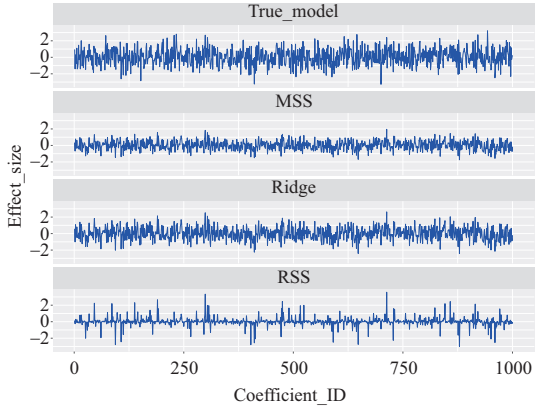

图 10 (网络版彩图) 当特异稀疏效应 $\beta_j = 0$ $(j = 1, \ldots, J)$ 时不同模型估计的效应值的比较

Figure 10 (Color online) The comparison of different models' estimation of effect size when the specific sparse effects $\beta_j = 0$ $(j = 1, \ldots, J)$

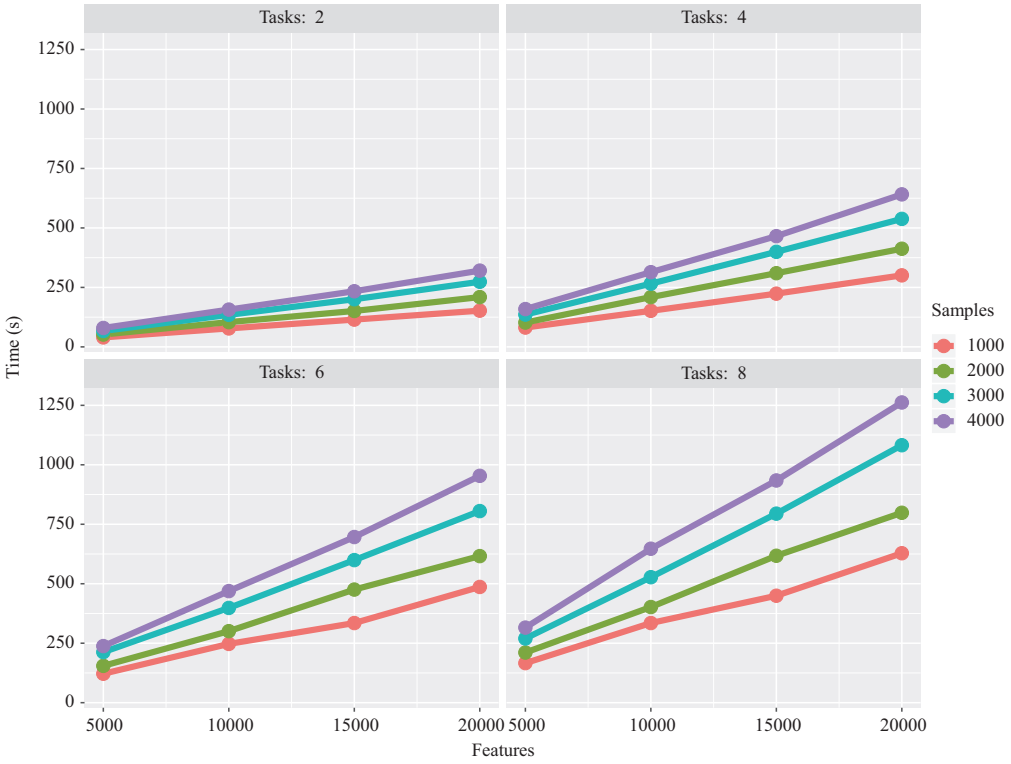

图 11 (网络版彩图) MSS 模型在不同的任务数、样本数和特征数的训练中, 所需的计算时间 (s)

Figure 11 (Color online) Computing time (CPU seconds) of MSS with respect to different number of samples in each task, different number of features and different number of tasks

3.2 真实数据分析

在本小节中, 我们利用 MSS 模型对电影评价网站 IMDB.com 提供的开放数据进行了分析. 原始

表 1 计算时间比较 (s)

Table 1 The comparison of computing time (s)

	MSS	Dirty model	DSL
$p = 500$	**10**	334	26
$p = 1000$	**18**	210	40
$p = 2000$	**33**	682	55

表 2 预测结果的均方误差 (MSE)

Table 2 Mean squared error of test results

	All	Drama	Comedy	Horror
MSS	**5.50**	**5.54**	**5.74**	4.99
Spike-Slab	5.54	5.57	5.78	5.05
Ridge	5.77	5.72	6.24	5.13
Lasso	5.55	5.63	5.77	**4.95**

数据来源于 IMDB.com 网页上的用户电影评论和对应的电影评分. 数据共包含 25000 条电影评论, 每条评论都对应一个取值范围 0∼10 的评分, 我们这里只提取偏向两极 (评分小于 4 分的负面评论或大于 7 分的正面评论) 的评分对应的评论用于关键词的提取, 并且保证正面评论和负面评论的数目相等. 全部评论包含的所有不重复的词汇构成了我们需要的词库, 其数量对应特征的维度 p, 每条评论对应一个 p 维词库向量, 评论中出现的所有不重复的词汇在 p 维词库向量中对应的位置用 1 表示, 其余位置用 0 表示 [33]. 在本实验中, $p = 27743$, 表明全部评论中所有不重复的词汇数共计 27743 个. 电影类型一共选取 3 种: 剧情片、喜剧片和恐怖片, 对应的评论数分别为 8286, 5027 和 3073. 我们用 MSS 模型构建回归模型, 用于对电影评分进行预测和对电影评分有关的关键词进行提取. 并将结果与对每一个类型电影进行单独训练的 RSS 模型、岭回归和 Lasso 进行对比. 利用 10 折交叉验证的方法对 4 种算法的结果 (电影预测评分的均方误差) 进行比较. 结果在表 2 中展示. 对于剧情片和喜剧片 MSS 模型相对另外 3 种模型有更好的预测结果, 这与 3.1 小节中的实验结果一致.

正如我们之前一直所强调的那样, MSS 模型的一大优点就是可以区分多任务学习中以常用词汇为代表的共享信号 $\boldsymbol{\beta}_0$ 和与电影类型相关的关键词为代表的特征信号 $\boldsymbol{\beta}_j$, $j = \{1, 2, 3\}$. 对 IMDB.com 的数据分析中, 共享信号和不同型电影中的特征信号对应的主要词汇按照影响的大小用词云表示在图 12 中, 红色词汇和绿色词汇分别表示对电影评分有较大正面影响和负面影响的词汇, 词汇大小表示信号的强度. 从词云中可以看出, 不同类型电影对应的关键词有很直观的意义, 比如喜剧片中的 “funnist”、剧情片中的 “tears” 和恐怖片中的 “scariest” 等. 我们可以从表 2 和图 12 中得出如下结论: 识别多任务学习中的共享信号可以帮助我们提高模型的预测精度, 同时帮助我们更好地提取特征信号对应的特征关键词, 避免常用词汇的干扰. 这些特征关键词往往是人们最为感兴趣的.

在本文中, 随机效应是指大量常用词汇整体对电影评分施加的影响, 这些常用词汇单独来看可能不存在显著的效应, 对电影得分只存在微弱的影响, 但这些微弱效应联合起来对电影得分就可能存在显著影响. 词云中的部分共享词汇便存在这样的现象, 单独来看每个词都是中性的, 但是大量的这些词汇联合起来就会对电影得分产生影响. 而通常人们感兴趣的, 对电影得分存在显著影响的关键词对应本文中的稀疏效应, 由系数 $\boldsymbol{\beta}_j$ 代表的效应. MSS 模型同时考虑到随机效应和稀疏效应, 因此可以提高模型预测的精度.

图 12 (网络版彩图) MSS 模型所提取的关键词对应的词云图. 红色词汇和绿色词汇分别表示对电影评分有较大正面影响和负面影响的词汇, 词汇大小表示信号的程度

Figure 12 (Color online) Word cloud of keywords generated by MSS. The words in red and green represent the positive and negative. The scale of the words represent the level of effect

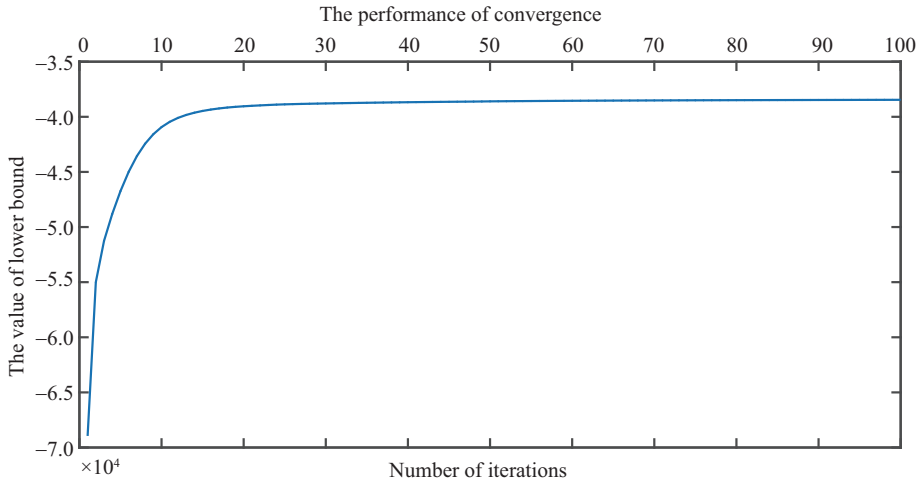

图 13 (网络版彩图) MSS 模型的收敛速度

Figure 13 (Color online) The performance of convergence of MSS

在对 IMDB.com 的数据分析中, MSS 模型每一步迭代计算在搭载 I7-3840QM CPU 和 16 G RAM 的笔记本上耗时仅仅数秒, 内存占用稳定在 10 G 左右, 在迭代 15 次后, 每一步迭代使得下界 $L(q)$ 的增幅在 10^{-4} 数量级以内, 如图 13 所示. 另外, 本文算法可以看做一个变分 EM 算法. 在期望步, 我们通过对潜在变量的近似分布求期望得到变分下界 $L(q)$, 在最大化步中, 我们关于模型参数 θ 最大化 $L(q)$. 因此, 在每一次 E 步和 M 步的迭代中, 下界 $L(q)$ 都不断增加, 算法的收敛性得到保证.

4 结论

在本文中, 我们考虑了如下的多任务学习模型: 假设每个任务中不仅存在稀疏的且对目标变量有显著影响的特征信号, 还存在非稀疏的单个作用不明显, 但联合起来对目标变量具有显著影响的共享信号. 同时假设共享信号对所有任务的目标变量都有相同的影响. 基于上述假设, 我们提出了 MSS 模型, 用以有效识别多任务学习中的共享信号和特征信号. 通过上述模型的建立, 我们可以利用所有任务的数据去有效地训练共享信号对应的效应值, 同时, 在剔除掉共享信号的影响后, 也能更好地捕捉到每个任务对应的特征信号. 在模型求解过程中, 为了克服计算难题, 我们运用基于变分 Bayes 推断的方法, 使得模型在大数据分析中具备更高的应用价值. 全面的模拟实验结果表明了 MSS 模型在模型预测和变量选择上的有效性. 与此同时, 我们还将 MSS 模型用于对真实数据的分析, 其结果再次证明 MSS 模型可以通过识别多任务学习中的共享信号和特征信号, 提升模型预测的精度和变量选择的准确率.

参考文献

1 Tibshirani R. Regression shrinkage and selection via the Lasso: a retrospective. J R Stat Soc-Ser B (Stat Method), 2011, 73: 273–282

2 Fan J Q, Li R Z. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assocn, 2001, 96: 1348–1360

3 Zhang C H. Nearly unbiased variable selection under minimax concave penalty. Ann Stat, 2010, 38: 894–942

4 Xu Z B, Chang X Y, Xu F M, et al. $L_{1/2}$ regularization: a thresholding representation theory and a fast solver. IEEE Trans Neural Netw Learn Syst, 2012, 23: 1013–1027

5 Zeng J S, Lin S B, Wang Y, et al. $L_{1/2}$ regularization: convergence of iterative half thresholding algorithm. IEEE Trans Signal Process, 2014, 62: 2317–2329

6 Wright S J. Coordinate descent algorithms. Math Program, 2015, 151: 3–34

7 Boyd S, Vandenberghe L. Convex Optimization. Cambridge: Cambridge University Press, 2004

8 Boyd S, Parikh N, Chu E, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. FNT Mach Learn, 2010, 3: 1–122

9 Figueiredo M A T. Adaptive sparseness for supervised learning. IEEE Trans Pattern Anal Mach Intell, 2003, 25: 1150–1159

10 Yuan M, Lin Y. Efficient empirical bayes variable selection and estimation in linear models. J Am Stat Assoc, 2005, 100: 1215–1225

11 Park T H, Casella G. The Bayesian Lasso. J Am Stat Assoc, 2008, 103: 681–686

12 Mitchell T J, Beauchamp J J. Bayesian variable selection in linear regression. J Am Stat Assoc, 1988, 83: 1023–1032

13 George E I, McCulloch R E. Variable selection via Gibbs sampling. J Am Stat Assoc, 1993, 88: 881–889

14 Madigan D, Raftery A E. Model selection and accounting for model uncertainty in graphical models using Occam's window. J Am Stat Assoc, 1994, 89: 1535–1546

15 Zhou X, Carbonetto P, Stephens M. Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet, 2013, 9: 1003264

16 Xu X, Ghosh M. Bayesian variable selection and estimation for group Lasso. Bayesian Anal, 2015, 10: 909–936

17 Chen R B, Chu C H, Yuan S, et al. Bayesian sparse group selection. J Comput Graph Stat, 2016, 25: 665–683

18 Blei D M, Kucukelbir A, McAuliffe J D. Variational inference: a review for statisticians. J Am Stat Assoc, 2017, 112: 859–877

19 Carbonetto P, Stephens M. Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. Bayesian Anal, 2012, 7: 73–108

20 Gopalan P, Hao W, Blei D M, et al. Scaling probabilistic models of genetic variation to millions of humans. Nat Genet, 2016, 48: 1587–1590

21 Dai M W, Ming J S, Cai M X, et al. IGESS: a statistical approach to integrating individual-level genotype data and summary statistics in genome-wide association studies. Bioinformatics, 2017, 33: 2882–2889

22 Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. J Mach Learn Res, 2003, 3: 993–1022
23 Ming J S, Dai M W, Cai M X, et al. LSMM: a statistical approach to integrating functional annotations with genome-wide association studies. Bioinformatics, 2018, 34: 2788–2796
24 Raj A, Stephens M, Pritchard J K. fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics, 2014, 197: 573–589
25 Sebastiani F. Multitask learning. In: Learning to Learn. Berlin: Springer, 1998. 95–133
26 Bach F R. Consistency of the group Lasso and multiple kernel learning. J Mach Learn Res, 2008, 9: 1179–1225
27 Ravikumar P, Wainwright M J, Lafferty J D. High-dimensional ising model selection using ℓ_1-regularized logistic regression. Ann Stat, 2010, 38: 1287–1319
28 Jalali A, Sanghavi S, Ruan C, et al. A dirty model for multi-task learning. In: Proceedings of Advances in Neural Information Processing Systems, 2010. 964–972
29 Gross S M, Tibshirani R. Data shared Lasso: a novel tool to discover uplift. Comput Stat Data Anal, 2016, 101: 226–235
30 Hoerl A E, Kennard R W. Ridge regression: biased estimation for nonorthogonal problems. Technometrics, 2000, 42: 80–86
31 Yang J, Benyamin B, McEvoy B P, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet, 2010, 42: 565–569
32 Bishop C M. Pattern Recognition and Machine Learning. Berlin: Springer, 2007. 462–474
33 Genkin A, Lewis D D, Madigan D. Large-scale Bayesian logistic regression for text categorization. Technometrics, 2007, 49: 291–304

附录 A 变分推断与参数估计

A.1 E-step

当给定参数 $\boldsymbol{\theta}=\{\sigma^2_{\boldsymbol{\beta}_0},\sigma^2_{\boldsymbol{\beta}_j},\sigma^2_j,\pi_j\}$ 时, 可以得到联合分布:

$$\begin{aligned}\Pr(\boldsymbol{y},\boldsymbol{\beta}_0,\boldsymbol{\beta}_j,\boldsymbol{\gamma}_j|\boldsymbol{X};\theta)=&\prod_j\left(\Pr(\boldsymbol{y}_j|\boldsymbol{\beta}_0,\boldsymbol{\beta}_j,\boldsymbol{\gamma}_j;\theta)\Pr(\boldsymbol{\beta}_j,\boldsymbol{\gamma}_j|\theta)\right)\Pr(\boldsymbol{\beta}_0|\theta)\\=&\prod_j\left(\Pr\big(\boldsymbol{y}_j|\boldsymbol{X}_j(\boldsymbol{\beta}_0+\boldsymbol{\beta}_j\odot\boldsymbol{\gamma}_j)\big)\mathcal{N}(\boldsymbol{\beta}_j|\boldsymbol{0},\sigma^2_{\boldsymbol{\beta}_j}\boldsymbol{I}_p)\prod_k\pi_j^{\gamma_{jk}}(1-\pi_j)^{1-\gamma_{jk}}\right)\mathcal{N}(\boldsymbol{\beta}_0|\boldsymbol{0},\sigma^2_{\boldsymbol{\beta}_0}\boldsymbol{I}_p),\end{aligned}\tag{A1}$$

其中 $\boldsymbol{\beta}_j\odot\boldsymbol{\gamma}_j$ 表示两个向量对应元素的乘积, 其结果为第 k 个元素为 $\beta_{jk}\gamma_{jk}$ $(k=1,\ldots,p)$ 的向量. 我们对模型中的潜在变量 $\boldsymbol{\beta_0}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ 积分得到边缘分布:

$$\begin{aligned}\Pr(\boldsymbol{y}|\boldsymbol{X};\boldsymbol{\theta})=&\sum_{\boldsymbol{\gamma}}\int\int\prod_j\Big(\mathcal{N}\big(\boldsymbol{X}_j(\boldsymbol{\beta}_0+\boldsymbol{\beta}_j\odot\boldsymbol{\gamma}_j),\sigma_j^2\boldsymbol{I}_{n_j}\big)\mathcal{N}(\boldsymbol{\beta}_j|\boldsymbol{0},\sigma^2_{\boldsymbol{\beta}_j}\boldsymbol{I}_p)\\&\cdot\prod_k\pi_j^{\gamma_{jk}}(1-\pi_j)^{1-\gamma_{jk}}\Big)\mathcal{N}(\boldsymbol{\beta}_0|\boldsymbol{0},\sigma^2_{\boldsymbol{\beta}_0}\boldsymbol{I}_p)\,\mathrm{d}\boldsymbol{\beta}_0\,\mathrm{d}\boldsymbol{\beta}.\end{aligned}\tag{A2}$$

同时, 我们也希望得到潜在变量的后验分布:

$$\Pr(\boldsymbol{\beta}_0,\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{y},\boldsymbol{X};\boldsymbol{\theta})=\frac{\Pr(\boldsymbol{y},\boldsymbol{\beta}_0,\boldsymbol{\beta}_j,\boldsymbol{\gamma}_j|\boldsymbol{X};\boldsymbol{\theta})}{\Pr(\boldsymbol{y}|\boldsymbol{X};\boldsymbol{\theta})}.\tag{A3}$$

由于积分的存在, 等式 (A2) 很难得到解析的表达形式, 因此我们这里采用变分推断的方法. 假设潜在变量的后验概率分布可以进行如下分解:

$$q(\boldsymbol{\beta}_0,\boldsymbol{\beta},\boldsymbol{\gamma})=q(\boldsymbol{\beta}_0)\prod_j\prod_k\big(q(\beta_{jk}|\gamma_{jk})q(\gamma_{jk})\big),\tag{A4}$$

利用变分推断的一般结论, $\log q(\boldsymbol{\beta}_0)$ 的最优解具有如下的表达形式:

$$\begin{aligned}\log q^*(\boldsymbol{\beta}_0)=&\,\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\gamma}}[\log\Pr(\boldsymbol{y}_j,\boldsymbol{\beta}_0,\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{X};\theta)]\\=&\sum_j\Bigg(\mathbb{E}_{\boldsymbol{\beta}_j,\boldsymbol{\gamma}_j}\left[-\frac{n_j}{2}\log(2\pi\sigma_j^2)-\frac{1}{2\sigma_j^2}\big(\boldsymbol{y}_j-\boldsymbol{X}_j(\boldsymbol{\beta}_0+\boldsymbol{\gamma}_j\odot\boldsymbol{\beta}_j)\big)^{\mathrm{T}}\big(\boldsymbol{y}_j-\boldsymbol{X}_j(\boldsymbol{\beta}_0+\boldsymbol{\gamma}_j\odot\boldsymbol{\beta}_j)\big)\right.\\&\left.-\frac{p}{2}\log(2\pi\sigma^2_{\boldsymbol{\beta}_j})-\frac{1}{2\sigma^2_{\boldsymbol{\beta}_j}}\beta_{jk}^2-\frac{1}{2\sigma^2_{\boldsymbol{\beta}_j}}\sum_{l\neq k}\beta_{jl}^2+\gamma_{jk}\log(\pi_j)+(1-\gamma_{jk})\log(1-\pi_j)\right)\end{aligned}$$

$$
\begin{aligned}
&+\log(\pi_j)\sum_{l\neq k}\gamma_{jl}+\log(1-\pi_j)\sum_{l\neq k}(1-\gamma_{jl})\Bigg]\Bigg)-\frac{p}{2}\log(2\pi\sigma^2_{\boldsymbol{\beta}_0})-\frac{1}{2\sigma^2_{\boldsymbol{\beta}_0}}\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{\beta}_0\\
=&\sum_j\left(-\frac{1}{2\sigma_j^2}\left(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j\boldsymbol{\beta}_0+2\mathbb{E}_{\boldsymbol{\beta}_j,\boldsymbol{\gamma}_j}[(\boldsymbol{\gamma}_j\odot\boldsymbol{\beta}_j)^{\mathrm{T}}](\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j)\boldsymbol{\beta}_0-2\boldsymbol{y}_j^{\mathrm{T}}\boldsymbol{X}_j\boldsymbol{\beta}_0\right)\right)-\frac{1}{2\sigma^2_{\boldsymbol{\beta}_0}}\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{\beta}_0+\mathrm{const}\\
=&\boldsymbol{\beta}_0^{\mathrm{T}}\left(\sum_j-\frac{1}{2\sigma_j^2}\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j-\frac{1}{2\sigma^2_{\boldsymbol{\beta}_0}}\boldsymbol{I}\right)\boldsymbol{\beta}_0+\sum_j-\frac{1}{\sigma_j^2}\left(\mathbb{E}_{\boldsymbol{\beta}_j,\boldsymbol{\gamma}_j}[(\boldsymbol{\gamma}_j\odot\boldsymbol{\beta}_j)^{\mathrm{T}}](\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j)-\boldsymbol{y}_j^{\mathrm{T}}\boldsymbol{X}_j\right)\boldsymbol{\beta}_0+\mathrm{const}.
\end{aligned}\tag{A5}
$$

因为 $\log q^*(\boldsymbol{\beta}_0)$ 是二次型, 所以 $\boldsymbol{\beta}_0\sim\mathcal{N}(\boldsymbol{\mu}_0,\boldsymbol{S}_0^2)$, 其中

$$
\boldsymbol{S}_0^2=-\frac{1}{2}\left(\sum_j-\frac{1}{2\sigma_j^2}\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j-\frac{1}{2\sigma^2_{\boldsymbol{\beta}_0}}\boldsymbol{I}\right)^{-1},\tag{A6}
$$

$$
\boldsymbol{\mu}_0=\boldsymbol{S}_0^2\sum_j-\frac{1}{\sigma_j^2}\left(\mathbb{E}_{\boldsymbol{\beta}_j,\boldsymbol{\gamma}_j}[(\boldsymbol{\gamma}_j\odot\boldsymbol{\beta}_j)^{\mathrm{T}}](\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j)-\boldsymbol{y}_j^{\mathrm{T}}\boldsymbol{X}_j\right)^{\mathrm{T}}.\tag{A7}
$$

需要注意的是, 当 p 很大的时候, 与 $\boldsymbol{S}_0^2$ 有关的计算量会非常大, 因为在迭代的每一步都涉及对矩阵求逆. 所以进一步的, 我们假设 $q(\boldsymbol{\beta}_0)$ 可以分解为 $\prod_{k=1}^p q(\beta_{0k})$, 对应的结果为

$$
\boldsymbol{S}_0^2=-\frac{1}{2}\mathrm{diag}\left(\sum_j-\frac{1}{2\sigma_j^2}\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j-\frac{1}{2\sigma^2_{\boldsymbol{\beta}_0}}\boldsymbol{I}\right)^{-1},\tag{A8}
$$

$$
\mu_{0k}=\boldsymbol{S}_0^2(k,k)\sum_j-\frac{1}{\sigma_j^2}\left(\sum_{l\neq k}\boldsymbol{x}_{jl}\mathbb{E}[\beta_{0l}]+\boldsymbol{X}_j\mathbb{E}_{\boldsymbol{\beta}_j,\boldsymbol{\gamma}_j}[(\boldsymbol{\gamma}_j\odot\boldsymbol{\beta}_j)]-\boldsymbol{y}_j\right)^{\mathrm{T}}\boldsymbol{x}_{jk},\tag{A9}
$$

这里 $\boldsymbol{x}_{jk}$ 表示设计矩阵 $\boldsymbol{X}_j$ 的第 k 列.

接下来, 对联合分布等式 (A1) 取对数, 并整理如下:

$$
\begin{aligned}
\log\Pr(\boldsymbol{y},\boldsymbol{\beta}_0,\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{X};\boldsymbol{\theta})=&\sum_j\Bigg(-\frac{n_j}{2}\log(2\pi\sigma_j^2)-\frac{\boldsymbol{y}_j^{\mathrm{T}}\boldsymbol{y}_j}{2\sigma_j^2}+\frac{(\beta_{0k}+\gamma_{jk}\beta_{jk})\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{y}_j}{\sigma_j^2}+\frac{\sum_{l\neq k}(\beta_{0l}+\gamma_{jl}\beta_{jl})\boldsymbol{x}_{jl}^{\mathrm{T}}\boldsymbol{y}_j}{\sigma_j^2}\\
&-\frac{1}{2\sigma_j^2}(\beta_{0k}+\gamma_{jk}\beta_{jk})^2\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jk}-\frac{\sum_{l\neq k}\sum_{l'\neq l\text{ and }k}(\beta_{0l}+\gamma_{jl}\beta_{jl})(\beta_{0l'}+\gamma_{jl'}\beta_{jl'})\boldsymbol{x}_{jl}^{\mathrm{T}}\boldsymbol{x}_{jl'}}{2\sigma_j^2}\\
&-\frac{\sum_{l\neq k}(\beta_{0l}+\gamma_{jl}\beta_{jl})^2\boldsymbol{x}_{jl}^{\mathrm{T}}\boldsymbol{x}_{jl}}{2\sigma_j^2}-\frac{\sum_{l\neq k}(\beta_{0k}+\gamma_{jk}\beta_{jk})(\beta_{0l}+\gamma_{jl}\beta_{jl})\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jl}}{\sigma_j^2}\\
&-\frac{p}{2}\log(2\pi\sigma^2_{\boldsymbol{\beta}_j})-\frac{1}{2\sigma^2_{\boldsymbol{\beta}_j}}\beta_{jk}^2-\frac{1}{2\sigma^2_{\boldsymbol{\beta}_j}}\sum_{l\neq k}\beta_{jl}^2+\gamma_{jk}\log(\pi_j)+(1-\gamma_{jk})\log(1-\pi_j)\\
&+\log(\pi_j)\sum_{l\neq k}\gamma_{jl}+\log(1-\pi_j)\sum_{l\neq k}(1-\gamma_{jl})\Bigg)-\frac{p}{2}\log(2\pi\sigma^2_{\boldsymbol{\beta}_0})-\frac{1}{2\sigma^2_{\boldsymbol{\beta}_0}}\sum_k\beta_{0k}^{\mathrm{T}}\beta_{0k}.
\end{aligned}\tag{A10}
$$

当 $\gamma_{jk}=1$ 时, 对等式 (A10) 关于 $q(\beta_{-jk},\gamma_{-jk})$ 和 $q(\boldsymbol{\beta}_0)$ 取期望, 可以得到

$$
\begin{aligned}
&\log q(\beta_{jk}|\gamma_{jk}=1)\\
=&\left(-\frac{1}{2\sigma_j^2}\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jk}-\frac{1}{2\sigma^2_{\boldsymbol{\beta}_j}}\right)\beta_{jk}^2+\frac{\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{y}_j-\sum_{l\neq k}\mathbb{E}_{jl}[\gamma_{jl}\beta_{jl}]\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jl}-\sum_{l=1}^p\mathbb{E}_{\beta_{0l}}[\beta_{0l}]\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jl}}{\sigma_j^2}\beta_{jk}+\mathrm{const}\\
=&\left(-\frac{1}{2\sigma_j^2}\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jk}-\frac{1}{2\sigma^2_{\boldsymbol{\beta}_j}}\right)\beta_{jk}^2+\frac{\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{y}_j-\sum_{l\neq k}\mathbb{E}_{jl}[\gamma_{jl}\beta_{jl}]\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jl}-\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{X}_j\mathbb{E}_{\beta_0}[\boldsymbol{\beta}_0]}{\sigma_j^2}\beta_{jk}+\mathrm{const},
\end{aligned}\tag{A11}
$$

这是关于 β_{jk} 的二次型, 因此 $q^*(\beta_{jk}|\gamma_{jk}=1)\sim\mathcal{N}(\boldsymbol{\mu}_{jk},\boldsymbol{S}_{jk}^2)$, 其中

$$
\begin{aligned}
s_{jk}^2=&\frac{\sigma_j^2}{\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jk}+\frac{\sigma_j^2}{\sigma^2_{\boldsymbol{\beta}_j}}},\\
\mu_{jk}=&\frac{\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{y}_j-\sum_{l\neq k}\mathbb{E}_{jl}[\gamma_{jl}\beta_{jl}]\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jl}-\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{X}_j\mathbb{E}_{\beta_0}[\boldsymbol{\beta}_0]}{\sigma_j^2}s_{jk}^2
\end{aligned}\tag{A12}
$$

$$=\frac{\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{y}_j-\sum_{l\neq k}\mathbb{E}_{jl}[\gamma_{jl}\beta_{jl}]\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jl}-\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{X}_j\mathbb{E}_{\beta_0}[\boldsymbol{\beta}_0]}{\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jk}+\frac{\sigma_j^2}{\sigma_{\boldsymbol{\beta}_j}^2}}. \tag{A13}$$

当 $\gamma_{jk}=0$ 时, 同样的我们用上述方法可以得到

$$\log q(\beta_{jk}|\gamma_{jk}=0)=-\tfrac{1}{2\sigma_{\boldsymbol{\beta}_j}^2}\beta_{jk}+\text{const},$$

因此 $q(\beta_{jk}|\gamma_{jk}=0)\sim\mathcal{N}(0,\sigma_{\boldsymbol{\beta}_j}^2)$.

由于 γ_{jk} 服从 Bernoulli 分布, 我们定义 $\alpha_{jk}=q(\gamma_{jk}=1)$, 可以得到 β_{jk} 和 γ_{jk} 的联合后验分布:

$$q(\beta_{jk},\gamma_{jk})=\left[\alpha_{jk}\mathcal{N}(\mu_{jk},s_{jk}^2)\right]^{\gamma_{jk}}\left[(1-\alpha_{jk})\mathcal{N}(0,\sigma_{\boldsymbol{\beta}_j}^2)\right]^{1-\gamma_{jk}}. \tag{A14}$$

综上, 可以得到如下的一些结论:

$$\mathbb{E}[\gamma_{jk}\beta_{jk}]=\mathbb{E}_{\gamma_{jk}}[\mathbb{E}_{\beta_{jk}}[\gamma_{jk}\beta_{jk}|\gamma_{jk}]]=\alpha_{jk}\mu_{jk}+(1-\alpha_{jk})\times 0=\alpha_{jk}\mu_{jk}, \tag{A15}$$

$$\begin{aligned}\mathbb{E}[(\gamma_{jk}\beta_{jk})^2]&=\mathbb{D}_{\gamma_{jk},\beta_{jk}}[\gamma_{jk}\beta_{jk}]+\mathbb{E}_{\gamma_{jk},\beta_{jk}}^2[\gamma_{jk}\beta_{jk}]\\&=\mathbb{D}_{\gamma_{jk}}\left[\mathbb{E}_{\beta_{jk}}[\gamma_{jk}\beta_{jk}|\gamma_{jk}]\right]+\mathbb{E}_{\gamma_{jk}}\left[\mathbb{D}_{\beta_{jk}}[\gamma_{jk}\beta_{jk}|\gamma_{jk}]\right]+\mathbb{E}_{\gamma_{jk}}^2[\gamma_{jk}\beta_{jk}]\\&=(\alpha_{jk}-\alpha_{jk}^2)\mu_{jk}^2+\alpha_{jk}s_{jk}^2+\alpha_{jk}^2\mu_{jk}^2\\&=\alpha_{jk}(\mu_{jk}^2+s_{jk}^2),\end{aligned} \tag{A16}$$

$$\mathbb{E}[\beta_{jk}^2]=\mathbb{E}_{r_{jk}}\left[\mathbb{E}_{\beta_{jk}}[\beta_{jk}^2|\gamma_{jk}]\right]=\alpha_{jk}(\mu_{jk}^2+\sigma_{jk}^2)+(1-\alpha_{jk})\sigma_{\boldsymbol{\beta}_j}^2, \tag{A17}$$

$$\mathbb{E}[\gamma_{jk}]=\alpha_{jk}. \tag{A18}$$

接下来, 我们可以对 $\log\Pr(\boldsymbol{y},\boldsymbol{\beta}_0,\boldsymbol{\beta},\boldsymbol{\gamma}\,|\boldsymbol{X};\boldsymbol{\theta})$ 关于 $q(\beta_{jk},\gamma_{jk})$ $(k=1,\ldots,p;j=1,\ldots,J)$ 和 $q(\boldsymbol{\beta}_0)$ 求期望, 得到边缘分布的下界 $L(q)$:

$$\begin{aligned}&\mathbb{E}_q[\Pr(\boldsymbol{y},\boldsymbol{\beta}_0,\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{X};\boldsymbol{\theta})]-\mathbb{E}_q[\log q(\boldsymbol{\beta}_0,\boldsymbol{\beta},\boldsymbol{\gamma})]\\&=\sum_j\Bigg(-\frac{n_j}{2}\log(2\pi\sigma_j^2)-\frac{\boldsymbol{y}_j^{\mathrm{T}}\boldsymbol{y}_j}{2\sigma_j^2}+\frac{\sum_{k=1}(\mathbb{E}_q[\beta_{0k}]+\mathbb{E}_q[\gamma_{jk}\beta_{jk}])\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{y}_j}{\sigma_j^2}\\&\quad-\frac{\sum_{k=1}\sum_{k'\neq k}\mathbb{E}_q[(\beta_{0k}+\gamma_{jk}\beta_{jk})(\beta_{0k'}+\gamma_{jk'}\beta_{jk'})]\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jk'}}{2\sigma_j^2}\\&\quad-\frac{\sum_{k=1}\mathbb{E}_q[(\beta_{0k}+\gamma_{jk}\beta_{jk})^2]\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jk}}{2\sigma_j^2}-\frac{p}{2}\log(2\pi\sigma_{\boldsymbol{\beta}_j}^2)-\frac{1}{2\sigma_{\boldsymbol{\beta}_j}^2}\sum_{k=1}\mathbb{E}_q[\beta_{jk}^2]\\&\quad+\log(\pi_j)\sum_{k=1}\mathbb{E}_q[\gamma_{jk}]+\log(1-\pi_j)\sum_{k=1}(1-\mathbb{E}_q[\gamma_{jk}])\Bigg)\\&\quad-\frac{p}{2}\log(2\pi\sigma_{\boldsymbol{\beta}_0}^2)-\frac{1}{2\sigma_{\boldsymbol{\beta}_0}^2}\mathbb{E}_q[\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{\beta}_0]-\mathbb{E}_q[\log q(\boldsymbol{\beta}_0,\boldsymbol{\beta},\boldsymbol{\gamma})],\end{aligned} \tag{A19}$$

利用等式 (A5), 我们可以直接得到包含 $\boldsymbol{\beta}_0$ 的项的期望:

$$\begin{aligned}&\mathbb{E}_q\left[\boldsymbol{\beta}_0^{\mathrm{T}}\left(\sum_j-\frac{1}{2\sigma_j^2}\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j-\frac{1}{2\sigma_{\boldsymbol{\beta}_0}^2}\boldsymbol{I}\right)\boldsymbol{\beta}_0+\sum_j-\frac{1}{\sigma_j^2}\left(\mathbb{E}_{\boldsymbol{\beta}_j,\boldsymbol{\gamma}_j}[(\boldsymbol{\gamma}_j\odot\boldsymbol{\beta}_j)^{\mathrm{T}}](\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j)-\boldsymbol{y}_j^{\mathrm{T}}\boldsymbol{X}_j\right)\boldsymbol{\beta}_0\right]+\text{const}\\&=\boldsymbol{\mu}_0^{\mathrm{T}}\left(\sum_j-\frac{1}{2\sigma_j^2}\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j-\frac{1}{2\sigma_{\boldsymbol{\beta}_0}^2}\boldsymbol{I}\right)\boldsymbol{\mu}_0+\mathrm{Tr}\left(\boldsymbol{S}_0^2\left(\sum_j-\frac{1}{2\sigma_j^2}\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j-\frac{1}{2\sigma_{\boldsymbol{\beta}_0}^2}\boldsymbol{I}\right)\right)\\&\quad+\sum_j-\frac{1}{\sigma_j^2}\left(\mathbb{E}_{\boldsymbol{\beta}_j,\boldsymbol{\gamma}_j}[(\boldsymbol{\gamma}_j\odot\boldsymbol{\beta}_j)^{\mathrm{T}}](\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j)-\boldsymbol{y}_j^{\mathrm{T}}\boldsymbol{X}_j\right)\mathbb{E}_q[\boldsymbol{\beta}_0]+\text{const},\end{aligned} \tag{A20}$$

所以下界 $L(q)$:

$$\begin{aligned}&\mathbb{E}_q[\Pr(\boldsymbol{y},\boldsymbol{\beta}_0,\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{X};\boldsymbol{\theta})]-\mathbb{E}_q[\log q(\boldsymbol{\beta}_0,\boldsymbol{\beta},\boldsymbol{\gamma})]\\&=\sum_j\left(-\frac{n_j}{2}\log(2\pi\sigma_j^2)-\frac{\boldsymbol{y}_j^{\mathrm{T}}\boldsymbol{y}_j}{2\sigma_j^2}+\frac{\sum_{k=1}(\mathbb{E}_q[\gamma_{jk}\beta_{jk}])\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{y}_j}{\sigma_j^2}-\frac{\sum_{k=1}\sum_{k'\neq k}\mathbb{E}_q[(\gamma_{jk}\beta_{jk})]\mathbb{E}_q[(\gamma_{jk'}\beta_{jk'})]\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jk'}}{2\sigma_j^2}\right.\end{aligned}$$

$$
\begin{aligned}
&- \frac{\sum_{k=1}\mathbb{E}_q[(\gamma_{jk}\beta_{jk})^2]\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jk}}{2\sigma_j^2} - \frac{p}{2}\log(2\pi\sigma_{\boldsymbol{\beta}_j}^2) - \frac{1}{2\sigma_{\boldsymbol{\beta}_j}^2}\sum_{k=1}\mathbb{E}_q[\beta_{jk}^2] + \log(\pi_j)\sum_{k=1}\mathbb{E}_q[\gamma_{jk}]\\
&+ \log(1-\pi_j)\sum_{k=1}(1-\mathbb{E}_q[\gamma_{jk}])\Bigg) + \boldsymbol{\mu}_0^{\mathrm{T}}\left(\sum_j -\frac{1}{2\sigma_j^2}\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j - \frac{1}{2\sigma_{\boldsymbol{\beta}_0}^2}\boldsymbol{I}\right)\boldsymbol{\mu}_0 + \mathrm{Tr}\left(\boldsymbol{S}_0^2\left(\sum_j -\frac{1}{2\sigma_j^2}\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j - \frac{1}{2\sigma_{\boldsymbol{\beta}_0}^2}\boldsymbol{I}\right)\right)\\
&+ \sum_j -\frac{1}{\sigma_j^2}\big(\mathbb{E}_{\boldsymbol{\beta}_j,\boldsymbol{\gamma}_j}[(\boldsymbol{\gamma}_j\odot\boldsymbol{\beta}_j)^{\mathrm{T}}](\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j) - \boldsymbol{y}_j^{\mathrm{T}}\boldsymbol{X}_j\big)\mathbb{E}_q[\boldsymbol{\beta}_0] - \frac{p}{2}\log(2\pi\sigma_{\boldsymbol{\beta}_0}^2) - \mathbb{E}_q[\log q(\boldsymbol{\beta}_0,\boldsymbol{\beta},\boldsymbol{\gamma})].
\end{aligned}
$$

由于

$$
\begin{aligned}
\mathbb{E}_q[\log q(\boldsymbol{\beta}_0,\boldsymbol{\beta},\boldsymbol{\gamma})] &= \mathbb{E}[\log q(\boldsymbol{\beta}_0)] + \sum_{j=1}\sum_{k=1}\mathbb{E}[\log q(\beta_{jk},\gamma_{jk})]\\
&= \mathbb{E}[\log q(\boldsymbol{\beta}_0)] + \sum_{j=1}\sum_{k=1}\mathbb{E}_{\gamma_{jk},\beta_{jk}}\big[\log[\alpha_{jk}\mathcal{N}(\mu_{jk},s_{jk}^2)]^{\gamma_{jk}}[(1-\alpha_{jk})\mathcal{N}(0,\sigma_{\boldsymbol{\beta}_j}^2)]^{1-\gamma_{jk}}\big]\\
&= \mathbb{E}[\log\mathcal{N}(\boldsymbol{\mu}_0,\boldsymbol{S}_0^2)] + \sum_{j=1}\sum_{k=1}\big(\mathbb{E}_q[\gamma_{jk}]\log\alpha_{jk} + (1-\mathbb{E}_q[\gamma_{jk}])\log(1-\alpha_{jk})\\
&\quad + \alpha_{jk}\mathbb{E}_{\beta_{jk}|\gamma_{jk}=1}[\log\mathcal{N}(\mu_{jk},s_{jk}^2)] + (1-\alpha_{jk})\mathbb{E}_{\beta_{jk}|\gamma_{jk}=0}[\log\mathcal{N}(0,\sigma_{\boldsymbol{\beta}_j}^2)]\big),
\end{aligned}
\tag{A21}
$$

利用正态分布的熵的结论, 可以得到

$$
\begin{aligned}
\mathbb{E}_q[\log q(\boldsymbol{\beta}_0,\boldsymbol{\beta},\boldsymbol{\gamma})] =& -\frac{1}{2}\log|\boldsymbol{S}_0^2| - \frac{p}{2}(1+\log 2\pi)\\
&+ \sum_j\sum_k\big(\alpha_{jk}\log a_{jk} + (1-\alpha_{jk})\log(1-\alpha_{jk})\big)\\
&- \sum_j\sum_k\frac{1}{2}\alpha_{jk}\big(\log s_{jk}^2 - \log\sigma_{\boldsymbol{\beta}_j}^2\big) - \sum_j\frac{p}{2}\log\sigma_{\boldsymbol{\beta}_j}^2 - \sum_j\frac{p}{2}(1+\log 2\pi).
\end{aligned}
\tag{A22}
$$

代入 $L(q)$, 再将等式 (A15)∼(A18) 带入并整理可以得到

$$
\begin{aligned}
L(q) =& \mathbb{E}_q[\Pr(\boldsymbol{y},\boldsymbol{\beta}_0,\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{X};\boldsymbol{\theta})] - \mathbb{E}_q[\log q(\boldsymbol{\beta}_0,\boldsymbol{\beta},\boldsymbol{\gamma})]\\
=& \sum_j\left(-\frac{n_j}{2}\log(2\pi\sigma_j^2) - \frac{\big\|\boldsymbol{y}_j - \sum_l\alpha_{jk}\mu_{jk}\boldsymbol{x}_{jk}\big\|^2}{2\sigma_j^2} - \frac{1}{2\sigma_j^2}\sum_l\left[\alpha_{jk}(s_{jk}^2+\mu_{jk}^2) - (\alpha_{jk}\mu_{jk})^2\right]\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jk}\right)\\
&+ \boldsymbol{\mu}_0^{\mathrm{T}}\left(\sum_j -\frac{1}{2\sigma_j^2}\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j - \frac{1}{2\sigma_{\boldsymbol{\beta}_0}^2}\boldsymbol{I}\right)\boldsymbol{\mu}_0 + \mathrm{Tr}\left(\boldsymbol{S}_0^2\left(\sum_j -\frac{1}{2\sigma_j^2}\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j - \frac{1}{2\sigma_{\boldsymbol{\beta}_0}^2}\boldsymbol{I}\right)\right)\\
&+ \sum_j -\frac{1}{\sigma_j^2}\big((\boldsymbol{\alpha}_j\odot\boldsymbol{\mu}_j)^{\mathrm{T}}(\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j) - \boldsymbol{y}_j^{\mathrm{T}}\boldsymbol{X}_j\big)\boldsymbol{\mu}_0\\
&- \frac{p}{2}\log(2\pi\sigma_{\boldsymbol{\beta}_0}^2) + \frac{1}{2}\log|\boldsymbol{S}_0^2| + \sum_j\sum_l\left(\alpha_{jk}\log\left(\frac{\pi_j}{\alpha_{jk}}\right) + (1-\alpha_{jk})\log\left(\frac{1-\pi_j}{1-\alpha_{jk}}\right)\right)\\
&+ \sum_j\sum_l\frac{1}{2}\alpha_{jk}\left(1+\log\frac{s_{jk}^2}{\sigma_{\boldsymbol{\beta}_j}^2} - \frac{\mu_{jk}^2+s_{jk}^2}{\sigma_{\boldsymbol{\beta}_j}^2}\right) - \frac{jp}{2}\log 2\pi + \frac{jp+p}{2}(1+\log 2\pi) - \frac{jp}{2}.
\end{aligned}
$$

对于 α_{jk}, 令

$$
\frac{\partial L(q)}{\partial\alpha_{jk}} = 0,
\tag{A23}
$$

可以得到

$$
\begin{aligned}
\frac{\partial L(q)}{\partial\alpha_{jk}} =& \frac{(\boldsymbol{y}_j - \sum_k\alpha_{jk}\mu_{jk}\boldsymbol{x}_{jk})^{\mathrm{T}}\mu_{jk}\boldsymbol{x}_{jk}}{\sigma_j^2} + \frac{(\alpha_{jk}\mu_{jk}^2)\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jk}}{\sigma_j^2} - \frac{(s_{jk}^2+\mu_{jk}^2)\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jk}}{2\sigma_j^2}\\
&- \frac{1}{\sigma_j^2}\mu_{jk}\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{X}_j\boldsymbol{\mu}_0 + \log\left(\frac{\pi_j}{1-\pi_j}\right) + \log\left(\frac{1-\alpha_{jk}}{\alpha_{jk}}\right)\\
&+ \frac{1}{2}\left(1+\log\frac{s_{jk}^2}{\sigma_{\boldsymbol{\beta}_j}^2} - \frac{\mu_{jk}^2+s_{jk}^2}{\sigma_{\boldsymbol{\beta}_j}^2}\right) = 0,
\end{aligned}
\tag{A24}
$$

由等式 (A12) 和 (A13), 可以得到

$$\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jk}+\frac{\sigma_j^2}{\sigma_{\boldsymbol{\beta}_j}^2}=\frac{\sigma_j^2}{s_{jk}^2}, \tag{A25}$$

$$\mu_{jk}\left(\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jk}+\frac{\sigma_j^2}{\sigma_{\boldsymbol{\beta}_j}^2}\right)+\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{X}_j\boldsymbol{\mu}_0=\left(\boldsymbol{y}_j-\sum_{k\neq k}\alpha_{jk}\mu_{jk}\boldsymbol{x}_{jk}\right)^{\mathrm{T}}\boldsymbol{x}_{jk}, \tag{A26}$$

所以

$$\begin{aligned}\frac{\partial L(q)}{\partial\alpha_{jk}}=&\frac{(\mu_{jk}^2-s_{jk}^2)\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jk}}{2\sigma_j^2}+\log\left(\frac{\pi_j}{1-\pi_j}\right)+\log\left(\frac{1-\alpha_{jk}}{\alpha_{jk}}\right)+\frac{1}{2}\left(1+\log\frac{s_{jk}^2}{\sigma_{\boldsymbol{\beta}_j}^2}\right)+\frac{\mu_{jk}^2-s_{jk}^2}{2\sigma_{\boldsymbol{\beta}_j}^2}\\=&\frac{\mu_{jk}^2}{2s_{jk}^2}+\log\left(\frac{\pi_j}{1-\pi_j}\right)+\log\left(\frac{1-\alpha_{jk}}{\alpha_{jk}}\right)+\frac{1}{2}\log\frac{s_{jk}^2}{\sigma_{\boldsymbol{\beta}_j}^2}=0,\end{aligned} \tag{A27}$$

可以得到

$$\alpha_{jk}=\frac{1}{1+\exp(-u_{jk})}, \tag{A28}$$

其中

$$u_{jk}=\frac{\mu_{jk}^2}{2s_{jk}^2}+\log\left(\frac{\pi_j}{1-\pi_j}\right)+\frac{1}{2}\log\frac{s_{jk}^2}{\sigma_{\boldsymbol{\beta}_j}^2}. \tag{A29}$$

A.2 M-step

接下来我们推导 σ_j^2, $\sigma_{\boldsymbol{\beta}_j}^2$ 和 $\sigma_{\boldsymbol{\beta}_0}^2$ 的更新方程.

对于 σ_j^2,

$$\begin{aligned}\frac{\partial L(q)}{\partial\sigma_j^2}=&-\frac{n_j}{2\sigma_j^2}+\frac{1}{2\sigma_j^4}\left(\boldsymbol{y}_j-\sum_l\alpha_{jl}\mu_{jl}\boldsymbol{x}_{jl}\right)^{\mathrm{T}}\left(\boldsymbol{y}_j-\sum_l\alpha_{jl}\mu_{jl}\boldsymbol{x}_{jl}\right)+\frac{1}{2\sigma_j^4}\sum_l\left[\alpha_{jl}(s_{jl}^2+\mu_{ji}^2)-(\alpha_{jl}\mu_{jl})^2\right]\boldsymbol{x}_{jl}^{\mathrm{T}}\boldsymbol{x}_{jl}\\&+\frac{1}{2\sigma_j^4}\boldsymbol{\mu}_0^{\mathrm{T}}(\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j)\boldsymbol{\mu}_0+\frac{1}{2\sigma_j^4}\mathrm{Tr}\big(\boldsymbol{S}_0^2(\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j)\big)+\frac{1}{\sigma_j^4}\big((\boldsymbol{\alpha}_j\odot\boldsymbol{\mu}_j)^{\mathrm{T}}(\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j)-\boldsymbol{y}_j^{\mathrm{T}}\boldsymbol{X}_j\big)\boldsymbol{\mu}_0=0,\end{aligned} \tag{A30}$$

可以得到

$$\begin{aligned}\sigma_j^2=&\frac{1}{n_j}\left(\left(\boldsymbol{y}_j-\sum_l\alpha_{jl}\mu_{jl}\boldsymbol{x}_{jl}\right)^{\mathrm{T}}\left(\boldsymbol{y}_j-\sum_l\alpha_{jl}\mu_{jl}\boldsymbol{x}_{jl}\right)+\sum_l[\alpha_{jl}(s_{jl}^2+\mu_{ji}^2)-(\alpha_{jl}\mu_{jl})^2]\boldsymbol{x}_{jl}^{\mathrm{T}}\boldsymbol{x}_{jl}\right.\\&\left.+\boldsymbol{\mu}_0^{\mathrm{T}}(\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j)\boldsymbol{\mu}_0+\mathrm{Tr}\big(\boldsymbol{S}_0^2(\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j)\big)+2\big((\boldsymbol{\alpha}_j\odot\boldsymbol{\mu}_j)^{\mathrm{T}}(\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j)-\boldsymbol{y}_j^{\mathrm{T}}\boldsymbol{X}_j\big)\boldsymbol{\mu}_0\right).\end{aligned} \tag{A31}$$

对于 $\sigma_{\boldsymbol{\beta}_j}^2$, 令

$$\frac{\partial L(q)}{\sigma_{\boldsymbol{\beta}_j}^2}=0, \tag{A32}$$

可以得到

$$\sigma_{\boldsymbol{\beta}_j}^2=\frac{\sum_l\alpha_{jl}(\mu_{jl}^2+s_{jl}^2)}{\sum_l\alpha_{jl}}. \tag{A33}$$

对于 $\sigma_{\boldsymbol{\beta}_0}^2$ 令

$$\frac{\partial L(q)}{\sigma_{\boldsymbol{\beta}_0}^2}=\frac{1}{2\sigma_{\boldsymbol{\beta}_0}^4}\big(\boldsymbol{\mu}_0^{\mathrm{T}}\boldsymbol{\mu}_0+\mathrm{Tr}(\boldsymbol{S}_0^2)\big)-\frac{p}{2\sigma_{\boldsymbol{\beta}_0}^2}=0, \tag{A34}$$

可以得到

$$\sigma_{\boldsymbol{\beta}_0}^2=\frac{1}{p}\big(\boldsymbol{\mu}_0^{\mathrm{T}}\boldsymbol{\mu}_0+\mathrm{Tr}(\boldsymbol{S}_0^2)\big). \tag{A35}$$

同样的, 最后可以得到

$$\pi_j=\frac{1}{p}\sum_l\alpha_{jl}. \tag{A36}$$

附录 B 算法细节

我们现在给出关于变分 EM 算法的详细推导, 其包含如下迭代过程:

- Initialize $\boldsymbol{\mu}_0$, $\boldsymbol{S}_0^2$, μ_{jk}, s_{jk}^2, α_{jk}, σ_j^2, $\sigma_{\boldsymbol{\beta}_j}^2$, $\sigma_{\boldsymbol{\beta}_0}^2$, π_j where $j = 1, \ldots, J$, $k = 1, \ldots, p$. Let $\tilde{\boldsymbol{y}}_j = \sum_k \alpha_{jk}\mu_{jk}\boldsymbol{x}_{jk}$, $\tilde{\boldsymbol{y}}_{0j} = \sum_k \mu_{0k}\boldsymbol{x}_{jk}$.
- E-step

$$\boldsymbol{S}_0^2 = -\frac{1}{2}\mathrm{diag}\left(\sum_j -\frac{1}{2\sigma_j^2}\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j - \frac{1}{2\sigma_{\boldsymbol{\beta}_0}^2}\boldsymbol{I}\right)^{-1}, \tag{B1}$$

For all j and l:

$$\tilde{\boldsymbol{y}}_{0jk} = \tilde{\boldsymbol{y}}_{0j} - \mu_{0k}\boldsymbol{x}_{jk}, \tag{B2}$$

$$\mu_{0k} = \boldsymbol{S}_0^2(k,k)\sum_j -\frac{1}{\sigma_j^2}\big(\tilde{\boldsymbol{y}}_{0jk} + \boldsymbol{X}_j(\boldsymbol{\alpha}_j \odot \boldsymbol{\mu}_j) - \boldsymbol{y}_j\big)^{\mathrm{T}}\boldsymbol{x}_{jk}, \tag{B3}$$

$$\tilde{\boldsymbol{y}}_{0j} = \tilde{\boldsymbol{y}}_{0jk} + \mu_{0k}\boldsymbol{x}_{jk}. \tag{B4}$$

Then, for all j and k:

$$\tilde{\boldsymbol{y}}_{jk} = \tilde{\boldsymbol{y}}_j - \alpha_{jk}\mu_{jk}\boldsymbol{x}_{jk}, \tag{B5}$$

$$s_{jk}^2 = \frac{\sigma_j^2}{\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jk} + \frac{\sigma_j^2}{\sigma_{\boldsymbol{\beta}_j}^2}}, \tag{B6}$$

$$\mu_{jk} = \frac{\boldsymbol{x}_{jk}^{\mathrm{T}}(\boldsymbol{y}_j - \tilde{\boldsymbol{y}}_{jk}) - \boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{X}_j\boldsymbol{\mu}_0}{\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jk} + \frac{\sigma_j^2}{\sigma_{\boldsymbol{\beta}_j}^2}}, \tag{B7}$$

$$\alpha_{jk} = \frac{1}{1+\exp(-u_{jk})}, \tag{B8}$$

where

$$u_{jk} = \frac{\mu_{jk}^2}{2s_{jk}^2} + \log\left(\frac{\pi_j}{1-\pi_j}\right) + \frac{1}{2}\log\frac{s_{jk}^2}{\sigma_{\boldsymbol{\beta}_j}^2}, \tag{B9}$$

$$\tilde{\boldsymbol{y}}_{jk} = \tilde{\boldsymbol{y}}_j + \alpha_{jk}\mu_{jk}\boldsymbol{x}_{jk}. \tag{B10}$$

- M-step

$$\sigma_j^2 = \frac{1}{n_j}\Bigg((\boldsymbol{y}_j - \tilde{\boldsymbol{y}}_j)^{\mathrm{T}}(\boldsymbol{y}_j - \tilde{\boldsymbol{y}}_j) + \sum_{k=1}^{p}[\alpha_{jk}(s_{jk}^2 + \mu_{jk}^2) - (\alpha_{jk}\mu_{jk})^2]\boldsymbol{x}_{jk}^{\mathrm{T}}\boldsymbol{x}_{jk}$$
$$+ \boldsymbol{\mu}_0^{\mathrm{T}}(\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j)\boldsymbol{\mu}_0 + \mathrm{Tr}\big(\boldsymbol{S}_0^2(\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j)\big) + 2\big((\boldsymbol{\alpha}_j \odot \boldsymbol{\mu}_j)^{\mathrm{T}}(\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j) - \boldsymbol{y}_j^{\mathrm{T}}\boldsymbol{X}_j\big)\boldsymbol{\mu}_0\Bigg), \tag{B11}$$

$$\sigma_{\boldsymbol{\beta}_j}^2 = \frac{\sum_k \alpha_{jk}(\mu_{jk}^2 + s_{jk}^2)}{\sum_k \alpha_{jk}}, \tag{B12}$$

$$\sigma_{\boldsymbol{\beta}_0}^2 = \frac{1}{p}\big(\boldsymbol{\mu}_0^{\mathrm{T}}\boldsymbol{\mu}_0 + \mathrm{Tr}(\boldsymbol{S}_0^2)\big), \tag{B13}$$

$$\pi_j = \frac{1}{p}\sum_k \alpha_{jk}. \tag{B14}$$

重复上述 E-step 和 M-step 步骤直到 $L(q)$ 收敛到一定范围内, 比如 1×10^{-6}.

Multi-task learning with shared random effects and specific sparse effects

Hao PENG[1], Ju WANG[2*] & Yao WANG[3*]

1. *School of Business Administration, Southwestern University of Finance and Economics, Chengdu* 611130*, China*;
2. *School of Management and Economics, University of Electronic Science and Technology of China, Chengdu* 611731*, China*;
3. *School of Management, Xi'an Jiaotong University, Xi'an* 710049*, China*
* Corresponding author. E-mail: wangju@usetc.com, yao.s.wang@gmail.com

Abstract In multi-task learning scenarios, random effects may be shared among different tasks while each task can have its own sparse effects. This structure has often been observed in the field of sentiment analysis for movie rating. In this study, we consider a multi-task learning problem in the presence of variables with shared random effects and specific sparse effects. To address this issue, we propose MSS (multi-task learning with shared random effects and specific sparse effects). To build this model, appropriate priors for the shared effects and specific effects under the Bayesian framework are considered. To overcome the computational complexity of Bayesian inference, an efficient algorithm is proposed based on variational inference, which is scalable to large-scale data analysis problems. The effectiveness of MSS in prediction and variable selection is demonstrated through comprehensive simulation studies and real data analysis of movie rating. The results demonstrate that the characterization of shared weak effects and task-specific sparse effects can improve the accuracy of prediction and variable selection.

Keywords multi-task learning, random effects, sparsity, variable selection, Bayesian inference

Hao PENG was born in 1987. He received his B.S. and M.S. degrees from Southwestern University of Finance and Economics in 2010 and 2015, respectively. Since 2015, he has been a Ph.D. student in business administration at Southwestern University of Finance and Economics, and his current research interests include pattern recognition and machine learning.

Ju WANG was born in 1981. She received her B.S. and Ph.D. degrees in business administration at Southwestern University of Finance and Economics in 2003 and 2008, respectively. She joined University of Electronic Science and Technology of China in 2008, and since 2010, she has been an associated professor in School of Management and Economics. Her current research interests include big data in business.

Yao WANG was born in 1983. He received his Ph.D. degree in mathematics and statistics at Xi'an Jiaotong University in 2014. Since 2019, he has been an associated professor in School of Management at Xi'an Jiaotong University. His current research interests include statistical signal processing and high-dimensional data analysis.