**长春工业大学徐平峰出境访问**

**工作总结**

2019年3月30日

本人为进行含潜变量的图模型相关研究，从2019年2月21日至2019年3月29日，在香港恒生大学数学与统计系进行访问，与邓文礼教授进行合作研究，总结访问情况如下。

图模型能清楚地表示变量间的结构关系，被广泛的应用于机器学习、因果推断、生物信息学等各个领域。图模型中的贝叶斯网是一个有向无圈的模型，能够更直观的描述变量间的依赖关系，特别是在其中引入潜变量后，不仅可表示因果推断中不可观测的混杂，而且可简化模型降低模型的复杂度，还可以提高计算的效率。然而，如何确定潜变量的个数、潜变量取值空间以及变量之间结构依赖关系，获得与数据拟合最好的模型，是一个非常具有挑战性的问题。这一问题在机器学习领域称为结构学习，本质上是统计学中的模型选择问题。

此次出访，本人与香港恒生大学邓文礼教授合作，针对含潜变量的贝叶斯网的结构学习问题，给出了基于惩罚似然的方法。这里的惩罚包括两项。一是对于可观测变量间的系数进行*l*1范数惩罚，以获得稀疏的模型；二是采用核范数（本文中为矩阵的迹）惩罚，以控制潜变量的个数。我们采用交替凸搜索方法，结合ADMM算法和坐标轴下降法，最小化惩罚似然，从而获得观测变量间的结构关系和潜变量个数的估计。为了找到惩罚似然中合适的调节参数，我们考虑了两种方法：交叉验证和BIC信息准则。我们给出了详细理论推导过程，编写了R程序，与当前主流的PC算法、RFCI算法、Adaptive Lasso惩罚似然方法、低秩稀疏（lrps）方法、低秩稀疏+贪婪搜索（lrps+ges）方法进行了比较。通过大量的模拟，我们发现当样本量大于100时，我们的方法有较好的表现。

我们考虑含有p=50个观测变量，h=5个潜变量，潜变量与观测变量之间关系的稀疏度sp.l=0.7的模拟情形，将样本量*n*分别取50，200，500，2000，我们随机的产生含潜变量的贝叶斯网及其样本，利用各种方法学习结构并进行比较pre（正确发现率）、rec（召回率）、mcc（马修斯系数）、shd（结构距离）、sparsity（模型稀疏度）和rank.L（秩），重复50次，取50次模拟的平均值，结果见下图。

由图可以看出，当*n*=50时，可以看出adaptive lasso算法的mcc最大，说明该算法最好。PC算法与Rfci算法的mcc值相同，且与adaptive lasso算法的mcc值差距相对小，pre值较大，但对应的rec值较小，shd的数值也较小，sparsity的值较小。acs系列的mcc值很小，pre值也小，对应的rec值较大，shd的值较大。当*n*=200时，acsA.ebic0.5的mcc值最大，对应的pre与rec也保持在较高的数值上，shd的数值较小，说明acsA.ebic0.5的效果最好。acsA系列的mcc值增大，对应的pre与rec的值都比n=50时增大，shd的数值较小。

当*n*=500时，acsA.ebic1.0的mcc值最大，对应的pre与rec的值也保持在较高的数值上，shd的数值较小，说明acsA.ebic1.0的效果最好。其他的算法与*n*=200时的情况基本相同。需要注意的是acs.ebic0.5与acs.ebic1.0的mcc值在增大，与acsA系列的mcc值差距不大。可以看出，随着n的增大，acs系列的模拟效果更加凸显。

当*n*=2000时，可以看出acsA.ebic1.0的mcc值最大，对应的rec也保持在较高的数值上，shd的数值较小，说明acsA.ebic1.0的效果最好。LRpS+GES算法的mcc值与acsA.ebic1.0的mcc值差距不大，但对应的pre值较小。随着n的增大，acs系列的rec值有增大的趋势。

根据上述分析，可以大致看出，当sp.l=0.7，h=5时，随着样本量n的取值不断增大， rec的值有上升的趋势，mcc的值能保持在0.5左右，而shd、sparsity在n=200时下降的趋势明显，rank.L的估计随着样本量的变大，大部分的值逐渐趋近于5，这符合图中h=5的设置。

徐平峰在香港恒生大学数学与统计系完成含潜变量的图模型相关研究后，已如期回境，并已进行了校内公示。

附图sp.l=0.7、h=5的情况下的模拟比较