概率图模型的稀疏化学习

刘建伟 崔立鹏 罗雄麟
(中国石油大学(北京)自动化研究所 北京 102249)

摘 要 利用稀疏化学习得到的概率图模型结构简单却保留了原始概率图模型中重要的结构信息,且能同时实现结构和参数学习,因此近几年来概率图模型的稀疏化学习一直是研究的热点,其中概率图模型的第一种稀疏化学习方法是图套索.文中总结了概率图模型的稀疏化学习方法,包括概率图模型的 L_1 范数罚稀疏化学习、概率图模型的无偏稀疏化学习、概率图模型的结构稀疏化学习和概率图模型的多任务稀疏化学习.最后,文中还指出了概率图模型的稀疏化学习未来有意义的研究方向.

关键词 概率图模型;稀疏化学习;结构和参数;图套索;精度矩阵;机器学习
中图法分类号 TP181 **DOI 号** 10.11897/SP.J.1016.2016.01597

Survey on the Sparse Learning of Probabilistic Graphical Models

LIU Jian-Wei CUI Li-Peng LUO Xiong-Lin
(*Research Institute of Automation, China University of Petroleum, Beijing* 102249)

Abstract A probabilistic graphical model obtained by sparse learning retains the important information of the original probabilistic graphical model's structure and the structure of the graphical model after sparse learning is very simple. In addition, the sparse learning can simultaneously achieve the learning of the structure and parameters of the graphical model, so the sparse learning for the probabilistic graphical models has been a research hotspot in recent years and the graphical lasso is the first method for the sparse learning of probabilistic graphical models. This paper summarizes various methods of the sparse learning of probabilistic graphical models, including sparse learning using L_1 norm penalty, unbiased sparse learning, sparse structure learning and multi-task sparse learning for probabilistic graphical models. Finally, the paper also proposes the meaningful future research directions for sparse learning of probabilistic graphical models.

Keywords probabilistic graphical models; sparse learning; structure and parameters; graphical lasso; precision matrix; machine learning

1 引 言

套索模型(Lasso)[1]由估计损失项和 L_1 范数罚项组成,其通过 L_1 范数罚实现稀疏化学习.自从套索模型被提出后,稀疏化学习方法迅速发展,其中一种重要的研究方向为将稀疏化学习思想应用到概率图模型的结构和参数学习中.精度矩阵(precision matrix)是概率图模型中全部随机变量的协方差矩阵的逆矩阵,它同时包含了概率图模型中的结构信息和参数信息,故通过对精度矩阵的学习可同时获得概率图模型的结构信息和参数信息,因此对概率

收稿日期:2014-04-02;在线出版日期:2014-11-30.本课题得到国家"九七三"重点基础研究发展计划项目基金(2012CB720500)、国家自然科学基金(21006127)、中国石油大学(北京)基础学科研究基金项目(JCXK-2011-07)资助.**刘建伟**,男,1966 年生,博士,副研究员,主要研究方向为智能信息处理、复杂系统分析预测与控制、算法分析与设计.E-mail:liujw@cup.edu.cn.**崔立鹏**,男,1990 年生,硕士研究生,主要研究方向为模型稀疏化机器学习.**罗雄麟**,男,1963 年生,博士,教授,主要研究领域为智能控制、复杂系统分析、预测与控制.

图模型的稀疏化学习问题本质上是对精度矩阵的稀疏化学习问题,该问题又被称作稀疏化协方差选择(covariance selection)[2]问题.第一种对概率图模型进行稀疏化学习的方法是将套索的罚函数 L_1 范数罚嵌入到概率图模型的极大似然估计中以实现精度矩阵的稀疏化,从而实现概率图模型结构的稀疏化,该稀疏化学习方法被称作图套索(Graphical Lasso).本文系统综述了概率图模型的稀疏化学习问题,包括概率图模型的 L_1 范数罚稀疏化学习、概率图模型的无偏稀疏化学习、概率图模型的结构稀疏化学习和概率图模型的多任务稀疏化学习.其中概率图模型的 L_1 范数罚稀疏化学习主要应用在高斯无向图模型[3-6]、部分随机变量不可观测的图模型[7-9]、有向无环图图模型[10-11]和伊辛模型[12-13]中,概率图模型的无偏稀疏化学习包括 SCAD 图套索[14-15]、自适应图套索[15]、贝叶斯无偏图套索[16-17]和幂法则图套索[18],概率图模型的结构稀疏化学习包括 $L_{q,1}$ 范数组结构图套索[19-20]、$L_{F,1}$ 范数组结构图套索[21]、双稀疏图套索[22]和局部共性图套索[23],概率图模型的多任务稀疏化学习主要包括多任务两两融合图套索[24]、多任务有序融合图套索①和多任务组结构图套索[25].概率图模型稀疏化学习的分类图如图 1 所示,各稀疏化学习方法对应的参考文献也列入了图 1 中.

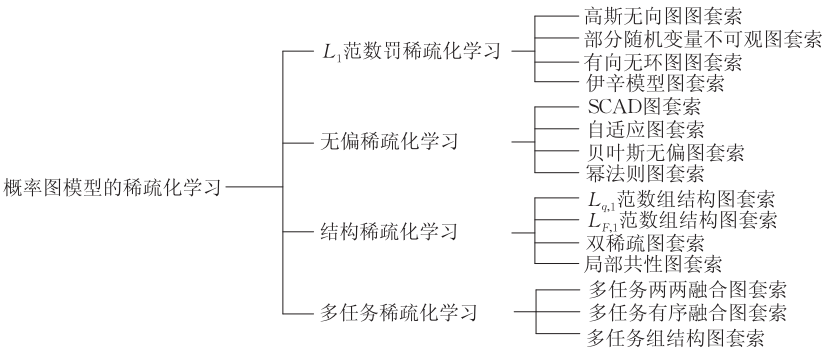


图 1 概率图模型稀疏化学习分类

概率图模型的稀疏化学习的必要性体现在如下几个方面:

(1)极大简化网络结构,提高概率图模型的可解释性.概率图模型的结构和参数学习,需要学习的空间维数很高,需要遍历的网络结构数随着顶点变量的个数呈指数增长,使得找到最适合数据样本的网络结构和最优的模型参数问题非常困难.由于概率图模型经过模型稀疏化学习后既结构简单又保留了原始概率无向图模型中重要的结构与参数信息,因此近几年来概率图模型的稀疏化学习问题越来越受到学者的重视.

(2)概率图模型的稀疏化学习能够通过学习精度矩阵而同时实现网络结构和参数的学习.

(3)提高泛化能力.从统计观点来看,概率图模型的稀疏化学习可以解决过拟合问题,从而使得模型更具有普适性和提高泛化能力.另外,概率图模型的稀疏化学习在一定的附加条件下能够保证学习得到的概率图模型拥有较好的统计特性.

(4)更有效地利用已知的先验信息.对复杂对象建模时,通常具有先验信息,也叫先验偏置,更好地利用这些先验信息,对于改善模型的统计特性和模型的推理和学习都有好处,概率图模型的稀疏化学习问题本质上是一个正则化问题,正则化项对应先验信息,根据不同的先验信息设计不同的正则化项会得到不同特点的稀疏解,从而可实现对概率图模型各种复杂的稀疏化学习,例如概率图模型的组结构稀疏化学习、概率图模型的重叠组结构稀疏化学习、概率图模型的树组结构稀疏化学习、局部共性结构稀疏化学习、节点和边同时稀疏的结构稀疏化学习等.

(5)概率图模型稀疏化学习的优化问题的求解可借助于大量现成的有效求解算法.稀疏化学习问题已经被广泛研究,其对应的优化问题的求解算法众多而且已经较为成熟,这些优化问题的求解算法使得概率图模型的稀疏化学习对应的优化问题的求解不再成为难题.例如对伊辛模型进行稀疏化学习的邻域选择方法[26](neighborhood selection method),将概率图模型的稀疏化学习问题转化为求解一系列的套索问题,可利用现成的求解套索的各种解法(最小角回归算法[27]等)进行求解.

(6)使得概率图模型可处理大规模复杂推理和学习问题,对概率图模型的推理和学习更加方便.

① Fused multiple graphical lasso. http://arxiv.org/abs/1209.2139, 2013,12,31

从后续的分析和算法设计的观点来看，概率图模型的稀疏化学习极大降低了参数个数，使得后续对概率图模型的各种分析和算法处理方面变得较为容易，一定程度上消除了算法设计面临的算法复杂性难题.

2　概率图模型的稀疏化学习

2.1　概率图模型的 L_1 范数罚稀疏化学习

2.1.1　高斯无向图图套索

令无向概率图模型对应的数据矩阵为 $\boldsymbol{X}\in\boldsymbol{R}^{N\times P}$. 假设每个样本 $\boldsymbol{X}^{(N)}$ 都服从独立同分布的 P 维高斯分布 $N_P(\boldsymbol{\mu},\boldsymbol{\Sigma})$：

$$\boldsymbol{X}^{(1)},\cdots,\boldsymbol{X}^{(N)}\sim N_P(\boldsymbol{\mu},\boldsymbol{\Sigma}) \tag{1}$$

其中 $\boldsymbol{\mu}\in\boldsymbol{R}^P$ 为均值，$\boldsymbol{\Sigma}\in\boldsymbol{R}^{P\times P}$ 为协方差矩阵. 令 $\boldsymbol{\Theta}=\boldsymbol{\Sigma}^{-1}$ 表示协方差矩阵 $\boldsymbol{\Sigma}$ 的逆矩阵，并将其称作精度矩阵. 精度矩阵 $\boldsymbol{\Theta}$ 中某个元素 Θ_{ij} 为零（不为零）表示无向概率图模型中两个节点 i 和 j 之间不存在（存在）一条边，即代表了变量 X_i 和 X_j 是条件独立（不是条件独立）的. 不妨假设均值 $\boldsymbol{\mu}=\boldsymbol{0}$，且下文中如无特别说明均默认均值 $\boldsymbol{\mu}=\boldsymbol{0}$. 由于精度矩阵同时包含了无向概率图模型中的参数信息和结构信息，因此学习无向概率图模型结构和参数的问题就转化成了学习精度矩阵 $\boldsymbol{\Theta}=\boldsymbol{\Sigma}^{-1}$ 的问题. 令 $\boldsymbol{S}$ 表示数据矩阵 $\boldsymbol{X}\in\boldsymbol{R}^{N\times P}$ 的样本协方差矩阵 $\boldsymbol{S}=\sum_{n=1}^{N}\boldsymbol{X}^{(n)}\boldsymbol{X}^{(n)\mathrm{T}}/N$，则关于精度矩阵 $\boldsymbol{\Theta}=\boldsymbol{\Sigma}^{-1}$ 的极大似然估计问题为

$$\max_{\boldsymbol{\Theta}\succ 0}\log|\boldsymbol{\Theta}|-tr(\boldsymbol{S\Theta}) \tag{2}$$

其中 $\boldsymbol{\Theta}\succ 0$ 表示矩阵 $\boldsymbol{\Theta}$ 的元素均大于零，$tr(\cdot)$ 表示矩阵的迹，$|\boldsymbol{\Theta}|$ 表示矩阵 $\boldsymbol{\Theta}$ 的行列式. 但是，极大似然估计方法不能产生稀疏解，因此得到的模型复杂度过高，不具有现实的可解释性. 针对上述问题，在式(2)的基础上添加 L_1 范数罚便得到图套索[3-6]：

$$\hat{\boldsymbol{\Theta}}=\arg\max_{\boldsymbol{\Theta}\succ 0}\log|\boldsymbol{\Theta}|-tr(\boldsymbol{S\Theta})-\lambda\|\boldsymbol{\Theta}\|_1 \tag{3}$$

其中 $\lambda\geq 0$，精度矩阵 $\boldsymbol{\Theta}\in\boldsymbol{R}^{P\times P}$ 为正定矩阵，$|\boldsymbol{\Theta}|$ 表示矩阵 $\boldsymbol{\Theta}$ 的行列式，$tr(\boldsymbol{S\Theta})=\langle\boldsymbol{S},\boldsymbol{\Theta}\rangle$ 表示矩阵 $\boldsymbol{S}$ 和 $\boldsymbol{\Theta}$ 乘积的迹（即矩阵 $\boldsymbol{S}$ 和 $\boldsymbol{\Theta}$ 的内积），$\|\boldsymbol{\Theta}\|_1=\sum_{j=1}^{P}\sum_{i=1}^{P}|\Theta_{ij}|$ 为关于精度矩阵 $\boldsymbol{\Theta}$ 的 L_1 范数罚，需注意，该矩阵的 L_1 范数罚形式与当前国内大多数教材中矩阵的 L_1 范数罚不同，该范数罚为对精度矩阵中全部元素的绝对值求和. 由于使用了 L_1 范数罚，因此式(3)的解 $\hat{\boldsymbol{\Theta}}$ 是稀疏的（即精度矩阵 $\hat{\boldsymbol{\Theta}}$ 中零元素占的比例较大），$\hat{\boldsymbol{\Theta}}$ 中为零的元素 $\hat{\Theta}_{ij}$ 表明随机变量 X_i 和 X_j 之间是条件独立的，即在无向图中随机变量 X_i 和 X_j 之间不存在边连接，因此大大简化了无向概率图模型的结构，并且同时实现了概率图模型的参数学习. 式(3)中的图套索是第一种被提出用来对概率图模型进行稀疏化学习的方法，为与下文各种概率图模型稀疏化学习方法区分，本文将其称作朴素图套索. 实际上朴素图套索服从贝叶斯理论，朴素图套索可以被表示为先验信息下的层次后验估计形式，该层次后验估计形式被称作贝叶斯图套索[28]，其中精度矩阵中非对角元素的罚函数部分对应于后验估计中的拉普拉斯先验分布，对角元素部分对应指数先验分布，且该层次后验估计中的超参数可以利用吉布斯抽样方法求解. Banerjee 等人[29]考虑了贝叶斯图套索后验分布的收敛率（convergence rates of posterior distributions）问题.

2.1.2　部分随机变量不可观图套索

除了将 L_1 范数罚用于高斯无向图模型的稀疏化学习，大量的研究还将 L_1 范数罚应用到其他图模型的稀疏化学习中. Chandrasekaran 等人[7-9]研究了某些随机变量不可观测时概率图模型的稀疏化学习问题，他们将协方差矩阵和精度矩阵分别进行如下分解：

$$\boldsymbol{\Sigma}=\begin{bmatrix}\Sigma_O & \Sigma_{OH}\\ \Sigma_{HO} & \Sigma_H\end{bmatrix} \tag{4a}$$

$$\boldsymbol{\Theta}=\begin{bmatrix}\Theta_O & \Theta_{OH}\\ \Theta_{HO} & \Theta_H\end{bmatrix} \tag{4b}$$

其中 O 为可观测随机变量的下标，H 为不可观测随机变量的下标. Chandrasekaran 等人令 $\bar{\boldsymbol{S}}=\Theta_O$，$\bar{\boldsymbol{L}}=\Theta_{OH}(\Theta_H)^{-1}\Theta_{HO}$，则根据 schur 补（schur complement）可得 $(\Sigma_O)^{-1}=\bar{\boldsymbol{S}}-\bar{\boldsymbol{L}}$，其中假定 $\bar{\boldsymbol{S}}$ 和 $\bar{\boldsymbol{L}}$ 分别为稀疏矩阵和低秩矩阵，对稀疏矩阵 $\bar{\boldsymbol{S}}$ 施加 L_1 范数罚 $\|\bar{\boldsymbol{S}}\|_1$，对于低秩矩阵 $\bar{\boldsymbol{L}}$ 施加迹范数罚 $tr(\bar{\boldsymbol{L}})$，此即矩阵的低秩稀疏分解（low rank and sparse decomposition）问题，Chandrasekaran 等人通过求解正则化项为 $\gamma\|\bar{\boldsymbol{S}}\|_1+tr(\bar{\boldsymbol{L}})$ 的罚极大似然估计问题得到 $\bar{\boldsymbol{S}}$ 和 $\bar{\boldsymbol{L}}$，进而得到 $(\Sigma_O)^{-1}$，从而学习出概率图模型可观部分的结构与参数信息，其中 γ 为调节 L_1 范数罚 $\|\bar{\boldsymbol{S}}\|_1$ 和迹范数罚 $tr(\bar{\boldsymbol{L}})$ 之间比例大小的可调参数.

2.1.3　有向无环图图套索

Shojaie 等人[10]利用 L_1 范数罚学习有向无环图图模型，但他们事先假定有向无环图具有已知的自然序（natural ordering），故其本质上是对有向无环图骨架的学习，并不能学习出有向无环图中的因果方向. 针对上述问题，Fu 等人[11]利用 L_1 范数罚方法

结合实验干预法同时学习高斯有向无环图的参数信息、稀疏结构和因果方向,其中 L_1 范数罚的作用为实现对因果高斯有向无环图的参数学习和稀疏结构学习,实验干预法的作用为进行因果方向的推断.

2.1.4 伊辛模型图套索

Ravikumar 等人[12-13]利用 L_1 范数罚学习一种特殊的马尔可夫网络——马尔可夫随机场中随机变量为二元取值的伊辛模型(Ising model),他们利用伪似然函数(pseudo-likelihood)作为该马尔可夫网络似然函数的近似,然后利用邻域选择方法求解,即分别对每个节点和其邻域内节点之间的结构进行学习,将该情形下概率图模型的学习问题转化为求解 P 个独立且易于求解的 L_1 范数罚逻辑斯蒂回归问题,其中 P 为随机变量的个数.

2.1.5 其他图模型的 L_1 范数罚稀疏化学习

有学者[30-32]将无向图模型中服从高斯分布的随机变量替换为服从多维高斯分布的光滑函数,将正态分布推广为非参数正态分布(non parametric normal distribution,简写为 nonparanormal),于是将高斯无向图模型推广为半参数高斯无向图模型,然后利用 L_1 范数罚对其进行稀疏化学习. Voorman 等人[33]将可加模型与概率图模型结合,利用 L_1 范数罚对该半参数类型的概率图模型进行稀疏化学习,并将其推广到节点之间的自然序(natural ordering)已知的有向图的情形下. Allen 等人[34]将概率图模型推广到随机变量服从泊松分布的无向图模型——泊松无向图模型,并且利用邻域选择方法对其进行稀疏化学习,将问题转化为 P 个独立且易于求解的 L_1 范数罚对数线性模型问题. Yang 等人[35]将概率图模型推广到随机变量服从指数分布族的情形,并将该概率图模型称作基于广义线性模型的概率图模型,然后利用邻域选择法对其进行稀疏化学习. Hill 等人①还利用 L_1 范数罚同时实现基于模型的聚类和概率图模型结构的稀疏化学习. Maurya[36]在 L_1 范数罚的基础上增加了一项迹范数罚(trace norm penalty)$tr(\boldsymbol{\Theta})$,该罚函数的含义为矩阵 $\boldsymbol{\Theta}$ 的全部特征值之和,叠加该罚函数后的学习效果(例如稀疏性和误差方面)往往比单纯利用 L_1 范数罚要好.

2.1.6 小结与分析

利用 L_1 范数罚可以实现对概率图模型结构和参数的同时学习,因此自从基于高斯无向图模型的朴素图套索被提出后,该方法迅速被推广到大量其他概率图模型的稀疏化学习中.但仅利用简单的 L_1 范数罚对概率图模型进行稀疏化学习往往是不够的,L_1 范数罚的估计有偏,并且很多时候概率图模型中随机变量之间往往具有某种复杂的结构,学习该类复杂结构需要结构化的罚函数,简单的 L_1 范数罚无法处理,上述缺点都有待于改进,这也是本文后面几节所述的核心内容.最后必须指出,L_1 范数罚是非凸的 L_0 范数罚的凸放松形式,由于 L_0 范数罚构成的优化问题难以求解,所以才对其进行凸放松得到 L_1 范数罚,但 L_0 范数罚能够得到比 L_1 范数罚更稀疏的模型,因此最近又有学者[37-38]重新提出利用 L_0 范数罚图模型进行稀疏化学习且均给出了求解非凸 L_0 范数正则化问题的有效求解算法.

2.2 概率图模型的无偏稀疏化学习

2.2.1 SCAD 图套索

SCAD 罚为 Fan 等人[39]提出的非凸罚,其具有估计无偏性.为了实现无偏估计,将朴素图套索中的 L_1 范数罚替换为 SCAD 罚便得到具有稀疏性和无偏估计性的 SCAD 图套索[14-15].已知式(1),SCAD 图套索要求解的问题为

$$\hat{\boldsymbol{\Theta}}=\arg\max_{\boldsymbol{\Theta}>0}\log|\boldsymbol{\Theta}|-tr(\boldsymbol{S}\boldsymbol{\Theta})-\sum_{i=1}^{P}\sum_{j=1}^{P}p_{\lambda,a}(|\Theta_{ij}|) \tag{5}$$

其中 Θ_{ij} 为矩阵 $\boldsymbol{\Theta}$ 中的第 i 行第 j 列的元素,SCAD 罚为

$$p_{\lambda,a}(|\theta|)=\lambda\left\{I(|\theta|\leq\lambda)+\frac{(a\lambda-|\theta|)_+}{(a-1)\lambda}I(|\theta|>\lambda)\right\} \tag{6}$$

其中 $\lambda\geq0$,$a>2$,$I(\cdot)$ 为指示函数.当 $a=\infty$ 时,SCAD 罚就退化为 L_1 范数罚,此时 SCAD 图套索也就退化为朴素图套索.另外,Abegaz 等人[40]还将 SCAD 罚应用于时间序列链图模型(time series chain graphical models)的稀疏化学习中.

2.2.2 自适应图套索

除了 SCAD 图套索外,自适应图套索[15]是另外一种具有无偏估计特点的图套索,已知式(1),自适应图套索要求解的问题为

$$\hat{\boldsymbol{\Theta}}=\arg\max_{\boldsymbol{\Theta}>0}\log|\boldsymbol{\Theta}|-tr(\boldsymbol{S}\boldsymbol{\Theta})-\lambda\sum_{i=1}^{P}\sum_{j=1}^{P}\omega_{ij}|\Theta_{ij}| \tag{7}$$

其中 $\lambda\geq0$,P 为随机变量个数,$\omega_{ij}=1/|\tilde{\Theta}_{ij}|^{\gamma}$ 为权重,$\gamma>0$,$\tilde{\Theta}_{ij}$ 为已知的矩阵 $\tilde{\boldsymbol{\Theta}}$ 中第 i 行第 j 列的元素.当 $P<N$ 时令 $\tilde{\boldsymbol{\Theta}}$ 为样本协方差矩阵的逆矩阵,当 $P\geq N$ 时令 $\tilde{\boldsymbol{\Theta}}$ 为朴素图套索的解.自适应图套索模型求解本质上分为两步:(1)求得样本协方差矩

① Network-based clustering with mixtures of L_1-penalized Gaussian graphical models: An empirical investigation. http://arxiv.org/abs/1301.2194, 2013,1,10

阵或朴素图套索的解 $\tilde{\boldsymbol{\Theta}}$；(2) 利用第(1)步中得到的 $\tilde{\boldsymbol{\Theta}}$ 确定权重 $\omega_{ij}=1/|\tilde{\Theta}_{ij}|^{\gamma}$，然后求解式(7). 自适应图套索对于矩阵 $\tilde{\boldsymbol{\Theta}}$ 中较大的元素值 $\tilde{\Theta}_{ij}$ 给其分配小的权重，对于矩阵 $\tilde{\boldsymbol{\Theta}}$ 中较小的元素值 $\tilde{\Theta}_{ij}$ 给其分配大的权重，从而减小对于无向概率图模型中重要边的惩罚，加大对于不重要边的惩罚，最终实现无偏估计. 另外，Peterson 等人[41]还利用自适应图套索对代谢网络(metabolic networks)进行稀疏化学习.

2.2.3　贝叶斯无偏图套索

与朴素图套索类似，自适应图套索也具有贝叶斯理论中的形式，其可以被表示为先验信息下的层次后验估计形式，其中精度矩阵中非对角元素的罚函数部分对应于后验估计中的拉普拉斯先验分布，对角元素部分对应指数先验分布，该层次后验估计形式被称作贝叶斯自适应图套索[16]. 与自适应图套索类似，贝叶斯自适应图套索在其先验分布中给精度矩阵的不同元素分配不同的权重，因而贝叶斯自适应图套索可实现无偏估计. Wong 等人[17]也引入了一个层次贝叶斯模型来对高斯图模型进行稀疏化学习，但他们将拉普拉斯先验替换为 Jeffreys 先验这种无信息先验分布(non-informative prior distribution)，因而避免了概率图模型稀疏化学习中需要利用交叉验证方法选择可调参数的问题，并且指出该层次贝叶斯模型中的后验估计式中的似然函数是非凸的，具有估计的无偏性. Orchard 等人①利用 GWishart 分布(GWishart Distribution)作为先验分布，然后利用哈密顿蒙特卡罗方法(Hamiltonian Monte Carlo)进行采样学习.

2.2.4　幂法则图套索

另外，L_1 范数罚不适用于无标度网络的学习. 在无标度网络中大部分节点只和少数几个节点相连，极少数节点(中枢节点)与很多节点相连，而且无标度网络中节点的自由度分布服从幂法则. 由于 L_1 范数罚对概率图模型中各个节点的惩罚程度是相同的，因此不具有识别出中枢节点的功能，于是 Liu 等人[18]将 L_1 范数罚替换成幂法则罚 $\sum_i \log\left(\sum_{i\neq j}\Theta_{ij}+\varepsilon_i\right)$，其中 ε_i 为一正数，并且指出新的幂法则图套索优化问题等价于一系列再权 L_1 范数正则化(reweighted L_1 regularization)问题，对各个节点惩罚的权重合理调整，使得自由度高的中枢节点的惩罚权重大大减小，从而促使中枢节点的出现.

2.2.5　小结与分析

式(3)中的朴素图套索采用了 L_1 范数罚，虽然具有稀疏性，但是其得到的解是有偏估计. L_1 范数罚导致有偏估计的原因为其对精度矩阵中的每个元素施加同等程度的罚，不具有对重要元素施加较小程度的罚而对非重要元素施加较大程度罚的自适应特性，针对该缺点而进行改善的方法大都从调整不同元素被惩罚的程度大小入手，例如本节所述的 SCAD 图套索、自适应图套索、贝叶斯无偏图套索和幂法则图套索.

2.3　概率图模型的结构稀疏化学习

2.3.1　$L_{q,1}$ 范数组结构图套索

$L_{q,1}$ 范数组结构图套索有两种，一种为 $L_{\infty,1}$ 范数组结构图套索[19]，另一种为 $L_{2,1}$ 范数组结构图套索[20]. 已知式(1)，$L_{\infty,1}$ 范数组结构图套索要求解的问题为

$$\hat{\boldsymbol{\Theta}}=\arg\max_{\boldsymbol{\Theta}\succ 0}\log|\boldsymbol{\Theta}|-tr(\boldsymbol{S}\boldsymbol{\Theta})-\lambda\sum_{g=1}^{m}\max\{|\Theta_{ij}|:(i,j)\in G_g\}\tag{8}$$

其中 $\sum_{g=1}^{m}\max\{|\Theta_{ij}|:(i,j)\in G_g\}$ 为 $L_{\infty,1}$ 范数罚，表示先对分组 G_g 中的元素进行 L_{∞} 范数运算，然后再横跨全部组进行 L_1 范数运算. 已知式(1)，$L_{2,1}$ 范数组结构图套索为

$$\hat{\boldsymbol{\Theta}}=\arg\max_{\boldsymbol{\Theta}\succ 0}\log|\boldsymbol{\Theta}|-tr(\boldsymbol{S}\boldsymbol{\Theta})-\sum_{g=1}^{m}\left\|\{|\Theta_{ij}|:(i,j)\in G_g\}\right\|_2\tag{9}$$

其中 $\sum_{g=1}^{m}\left\|\{|\Theta_{ij}|:(i,j)\in G_g\}\right\|_2$ 为 $L_{2,1}$ 范数罚，表示先对分组 G_g 中的元素进行 L_2 范数运算，然后再横跨全部组进行 L_1 范数运算. $L_{q,1}$ 范数组结构图套索的提出是因为有时希望概率图模型中某些边作为一个整体同时存在或者同时消失，故组结构图套索需要事先人为将精度矩阵 $\boldsymbol{\Theta}$ 中的全部元素分为 m 个组 $\{G_1,\cdots,G_m\}$，导致 $L_{q,1}$ 范数组结构图套索具有令精度矩阵中元素成组地进行稀疏的效果. Cheng 等人②和 Lee 等人③将 $L_{q,1}$ 范数组结构图套索应用到既有随机变量服从连续分布又有随机变量服从离散分布的混合概率图模型中进行结构稀疏化学习.

---

① Bayesian inference in sparse Gaussian graphical models. http://arxiv.org/abs/1309.7311, 2013, 9, 27

② High-dimensional mixed graphical models. http://arxiv.org/abs/1304.2810, 2013, 4, 9

③ Learning the structure of mixed graphical models. http://amstat.tandfonline.com/doi/abs/10.1080/10618600.2014.900500#.VHmNS_mSwpM, 2014, 4, 4

2.3.2 $L_{F,1}$范数组结构图套索

假设概率图模型中每个节点都为多属性节点,即每个节点对应的是一个多维随机向量而不是单个随机变量,则$L_{F,1}$范数组结构图套索[21]要求解的问题为

$$\hat{\boldsymbol{\Theta}}=\arg\max_{\boldsymbol{\Theta}\succ 0}\log|\boldsymbol{\Theta}|-tr(\boldsymbol{S}\boldsymbol{\Theta})-\lambda\sum_{a,b}\|\Theta_{ab}\|_F \quad (10)$$

其中Θ_{ab}表示多属性节点a和多属性节点b在密度矩阵中对应的元素块,罚函数$\sum_{a,b}\|\Theta_{ab}\|_F$表示对元素块先进行$F$范数(Frobenius范数)运算,再横跨全部元素块进行L_1范数运算,其作用为实现精度矩阵的块稀疏化(即组稀疏化).$L_{F,1}$范数组结构图套索背后的原理为:由于两个多属性节点之间是否相互独立等价于其偏典型相关系数(partial canonical correlation)是否为零,而偏典型相关系数是否为零又等价于两多属性节点在精度矩阵中对应的元素块是否为零,故基于多属性节点的概率图模型的稀疏化学习问题等价于精度矩阵的块稀疏化学习问题.

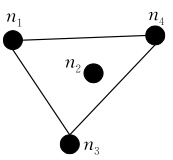
2.3.3 双稀疏图套索

Honorio等人[22]提出了边和节点同时稀疏的双稀疏图套索,双稀疏图套索在假设无向概率图模型中边存在稀疏现象的同时还假设无向概率图模型中只有一小部分的节点与其他节点存在相互依赖的关系(存在边相连),而大部分节点都是孤立的,即大部分节点与其他全部节点都不存在相互依赖关系(没有边相连).对于孤立的节点来说,其在精度矩阵$\boldsymbol{\Theta}$中对应的行和列中的元素除对角元素非零外其他非对角元素全部都为0,例如如图2(a)所示,假设某无向概率图模型有4个节点n_1、n_2、n_3和n_4,其中节点n_2与其他3个节点n_1、n_3和n_4之间都不存在边相连,则对应的精度矩阵如图2(b)所示,其中×表示对应位置的元素非零,0表示对应位置的元素为零.已知式(1),双稀疏图套索要求解的问题为

$$\hat{\boldsymbol{\Theta}}=\arg\max_{\boldsymbol{\Theta}\succ 0}\log|\boldsymbol{\Theta}|-tr(\boldsymbol{S}\boldsymbol{\Theta})-\lambda_1\|\boldsymbol{\Theta}\|_1-\lambda_2\|\boldsymbol{\Theta}\|_{1,q} \quad (11)$$

其中$\lambda_1\geqslant 0$,$\lambda_2\geqslant 0$,$\boldsymbol{S}$为样本协方差矩阵,且

$$\|\boldsymbol{\Theta}\|_{1,q}=\sum_{i=1}^{P}\|\Theta_{i1},\cdots,\Theta_{i,i-1},\Theta_{i,i+1},\cdots,\Theta_{iP}\|_q \quad (12)$$



(a) 无向概率图模型

	n_1	n_2	n_3	n_4
n_1	×	0	×	×
n_2	0	0	0	0
n_3	×	0	×	×
n_4	×	0	×	×

(b) 精度矩阵

图2 节点稀疏示意

L_1范数罚$\|\boldsymbol{\Theta}\|_1$的作用为实现边的稀疏,$\|\boldsymbol{\Theta}\|_{1,q}$表示对精度矩阵$\boldsymbol{\Theta}$中第$i$行中除对角线元素外其他全部非对角线元素组成的向量进行L_q范数运算,其作用为产生孤立的节点,即实现节点的稀疏.显然,若令式(11)中的$\lambda_1=0$,则可以得到双稀疏图套索的一种特殊情况:只使得节点稀疏的图套索;而若令式(11)中的$\lambda_2=0$,则可以得到双稀疏图套索的另一种特殊情况:只使得边稀疏的图套索,即2.1节中的朴素图套索.必须指出,节点稀疏图套索在本质上就是将与某节点相连的所有边作为一个组并且令该组中的全部边同时存在或同时消失,因此节点稀疏图套索是组结构图套索的一种特殊形式.

2.3.4 局部共性图套索

在有些情形下,概率图模型中相邻的节点之间往往具有某种共性,因而在对其进行稀疏化学习时也希望保留相邻节点之间的这种共性,Honorio等人[23]针对该问题提出对精度矩阵中相邻元素之差进行惩罚以学习出概率图模型中局部存在的共性,再叠加L_1范数罚实现稀疏性.本文将该方法称作局部共性图套索,已知式(1),局部共性图套索要求解的问题为

$$\hat{\boldsymbol{\Theta}}=\arg\max_{\boldsymbol{\Theta}\succ 0}\log|\boldsymbol{\Theta}|-tr(\boldsymbol{S}\boldsymbol{\Theta})-\lambda_1\|\boldsymbol{\Theta}\|_1-\lambda_2\|\boldsymbol{D}\otimes\boldsymbol{\Theta}\|_1 \quad (13)$$

其中$\lambda_1>0$,$\lambda_2>0$,而

$$\boldsymbol{D}\otimes\boldsymbol{\Theta}=\boldsymbol{J}(\boldsymbol{D})\circ(\boldsymbol{D}\boldsymbol{\Theta}) \quad (14)$$

其中"$\circ$"运算表示矩阵之间对应元素的相乘即$(\boldsymbol{A}\circ\boldsymbol{B})_{ij}=a_{ij}b_{ij}$.$\boldsymbol{J}(\boldsymbol{D})$为对矩阵$\boldsymbol{D}$的Iverson括号算子,例如对于$\boldsymbol{D}$中位于第$i$行$j$列的元素来说其Iverson括号算子为$J_{ij}(\boldsymbol{D})=[d_{ij}=0]$,该括号算子具体含义为当矩阵$\boldsymbol{D}$中的元素$d_{ij}=0$时则$\boldsymbol{J}(\boldsymbol{D})$中元素$J_{ij}(\boldsymbol{D})=1$,否则$J_{ij}(\boldsymbol{D})=0$.矩阵$\boldsymbol{D}$的构造方法为:(1)若概率图模型中节点$m$与节点$n$相邻,则令矩阵$\boldsymbol{D}$第$i$行中$d_{im}=1$和$d_{in}=-1$,同时令第$i$行中其他元素全部为0;(2)将第(1)步中的方法遍历全部相邻的节点,填满整个矩阵$\boldsymbol{D}$.该方法实际上是对精度矩阵中相邻元素之差进行惩罚,能够在实现稀疏化学习的同时揭示出概率图模型中相邻节点之间的共性.

2.3.5 小结与分析

将稀疏化学习应用到概率图模型的学习中已经是某种意义上的结构稀疏化(图结构稀疏化),进一步地,若在图模型的稀疏化学习中边或节点在被置零时又具有某种结构化的特点,则称该问题为概率

图模型的结构稀疏化学习.概率图模型的结构稀疏化学习大都采用结构化的罚函数,结构化的罚函数是结构稀疏化学习的本质,其原理为事先假设对象具有某种稀疏化结构,然后将该稀疏化结构作为先验信息来构造稀疏化罚函数,进而进行结构稀疏化学习.结构稀疏化近年来一直是研究的热点,譬如近年提出的组套索(Group Lasso)[42]、组之间元素重叠的组套索(overlap Group Lasso)[43]和树结构组套索(tree-guided Group Lasso)[44-45]等,但这些结构稀疏化的模型都是针对回归模型的,并不涉及概率图模型的结构稀疏化学习,未来将重叠组结构等其他结构化的罚函数引入到概率图模型的稀疏化学习中是非常有前景的研究方向.

2.4　概率图模型的多任务稀疏化学习

假设有 K 个无向概率图模型 $\boldsymbol{X}^{(1)},\cdots,\boldsymbol{X}^{(K)}\in\boldsymbol{R}^{N_k\times P}$,其中 N_k 为关于第 k 个无向概率图模型 $\boldsymbol{X}^{(k)}$ 的样本数,P 为随机变量数.令 $\boldsymbol{\Theta}^{(k)}=\boldsymbol{\Sigma}_k^{-1}$ 代表第 k 个无向概率图模型 $\boldsymbol{X}^{(k)}$ 的协方差矩阵的逆矩阵,$\hat{\boldsymbol{\Theta}}^{(1)},\cdots,\hat{\boldsymbol{\Theta}}^{(K)}$ 代表对 $\boldsymbol{\Sigma}_1^{-1},\cdots,\boldsymbol{\Sigma}_K^{-1}$ 的估计值,其中 $k=1,\cdots,K$.假设全部样本都是相互独立的,且同一个无向概率图模型对应的全部样本是独立同分布的:

$$\boldsymbol{x}_1^{(k)},\cdots,\boldsymbol{x}_{N_k}^{(k)}\sim N(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\tag{15}$$

且假设 $\boldsymbol{\mu}_k=\boldsymbol{0}$,第 $\boldsymbol{R}$ 个无向概率图模型 $\boldsymbol{X}^{(k)}$ 的样本协方差矩阵为 $\bar{\boldsymbol{S}}^{(k)}$,令 $\{\boldsymbol{\Theta}\}=\{\boldsymbol{\Theta}^{(1)},\cdots,\boldsymbol{\Theta}^{(K)}\}$,其中 $\boldsymbol{\Theta}^{(1)},\cdots,\boldsymbol{\Theta}^{(K)}$ 是正定矩阵,下面在上述符号定义的基础上介绍各种多任务图套索.

2.4.1　多任务两两融合图套索

多任务两两融合图套索[24]的罚函数为 L_1 范数罚和两两融合罚[46-47](pairwise fused penalty)的组合,已知式(15),多任务两两融合图套索要求解的问题为

$$\{\hat{\boldsymbol{\Theta}}\}=\underset{\boldsymbol{\Theta}^{(1)},\cdots,\boldsymbol{\Theta}^{(K)}\succ 0}{\arg\max}\left\{\sum_{k=1}^{K}N_k[\log|\boldsymbol{\Theta}^{(k)}|-tr(\boldsymbol{S}^{(k)}\boldsymbol{\Theta}^{(k)})-\lambda_1\Phi_1(\Theta_{ij}^{(k)})-\lambda_2\Phi_2(\Theta_{ij}^{(k_1)}-\Theta_{ij}^{(k_2)})]\right\}\tag{16}$$

其中 $\lambda_1>0,\lambda_2>0$,且

$$\Phi_1(\Theta_{ij}^{(k)})=\sum_{k=1}^{K}\sum_{j=2}^{P}\sum_{i=1}^{P-1}|\Theta_{ij}^{(k)}|\tag{17}$$

为横跨 K 个精度矩阵的 L_1 范数罚,其作用为实现稀疏解,另一项

$$\Phi_2(\Theta_{ij}^{(k_1)}-\Theta_{ij}^{(k_2)})=\sum_{k_2=2}^{K}\sum_{k_1=1}^{K-1}\sum_{j=2}^{P}\sum_{i=1}^{P-1}|\Theta_{ij}^{(k_1)}-\Theta_{ij}^{(k_2)}|\tag{18}$$

为两两融合罚,能够使得不同数据集对应的精度矩阵 $\hat{\boldsymbol{\Theta}}^{(k)}=\boldsymbol{\Sigma}_k^{-1}$ 具有类似的稀疏结构,即促使不同数据集对应的精度矩阵 $\hat{\boldsymbol{\Theta}}^{(k)}=\boldsymbol{\Sigma}_k^{-1}$ 的零元素出现在同一位置.

2.4.2　多任务有序融合图套索

Yang 等人(见本文第 2 页脚注①)利用有序融合罚对多个概率图模型进行多任务稀疏化学习,与多任务两两融合图套索不同,他们假设多个概率图模型之间是有序的,只对相邻的图模型施加有序融合罚[48](sequential fused penalty).已知式(15),多任务有序融合图套索要求解的问题为

$$\{\hat{\boldsymbol{\Theta}}\}=\underset{\boldsymbol{\Theta}^{(1)},\cdots,\boldsymbol{\Theta}^{(K)}\succ 0}{\arg\max}\left\{\sum_{k=1}^{K}N_k[\log|\boldsymbol{\Theta}^{(k)}|-tr(\boldsymbol{S}^{(k)}\boldsymbol{\Theta}^{(k)})-\lambda_1\Phi_1(\Theta_{ij}^{(k)})-\lambda_2\Phi_2(\Theta_{ij}^{(k+1)}-\Theta_{ij}^{(k)})]\right\}\tag{19}$$

其中 $\lambda_1>0,\lambda_2>0$,且

$$\Phi_1(\Theta_{ij}^{(k)})=\sum_{k=1}^{K}\sum_{j=2}^{P}\sum_{i=1}^{P-1}|\Theta_{ij}^{(k)}|\tag{20}$$

为横跨 K 个精度矩阵的 L_1 范数罚,其作用为实现稀疏解;另一项

$$\Phi_2(\Theta_{ij}^{(k+1)}-\Theta_{ij}^{(k)})=\sum_{k=1}^{K-1}\sum_{j=2}^{P}\sum_{i=1}^{P-1}|\Theta_{ij}^{(k+1)}-\Theta_{ij}^{(k)}|\tag{21}$$

为有序融合罚,其作用为促使相邻的概率图模型结构一致.另外,Zhang 等人①也利用有序融合罚对概率图模型进行多任务稀疏学习,在对照实验(controlled experiments)中,当某些实验条件变化时图模型的结构也会发生变化,Zhang 等人利用 L_1 范数罚叠加有序融合罚来对概率图模型的结构变化进行稀疏化学习,其中 L_1 范数罚实现模型稀疏化学习,有序融合罚抑制噪声带来的结构和参数的不一致性,使得两次不同实验条件下的两个图模型的结构和参数平坦变化.

2.4.3　多任务组结构图套索

多任务组结构图套索为 L_1 范数罚与 $L_{q,1}$ 范数罚的组合,已知式(15),多任务组结构图套索要求解的问题为

$$\{\hat{\boldsymbol{\Theta}}\}=\underset{\boldsymbol{\Theta}^{(1)},\cdots,\boldsymbol{\Theta}^{(K)}\succ 0}{\arg\max}\left\{\sum_{k=1}^{K}N_k(\log|\boldsymbol{\Theta}^{(k)}|-tr(\boldsymbol{S}^{(k)}\boldsymbol{\Theta}^{(k)})-\lambda_1\sum_{k=1}^{K}\|\boldsymbol{\Theta}^{(k)}\|_1-\lambda_2\|\boldsymbol{\Theta}\|_{q,1})\right\}\tag{22}$$

其中 $\lambda_1>0,\lambda_2>0$,$\sum_{k=1}^{K}\|\boldsymbol{\Theta}^{(k)}\|_1$ 为横跨 K 个精度矩阵的 L_1 范数罚,其作用为实现稀疏解;当 $\|\boldsymbol{\Theta}\|_{q,1}$ 为 $L_{2,1}$ 范数罚 $\|\boldsymbol{\Theta}\|_{2,1}=\sum_{j=2}^{P}\sum_{i=1}^{P-1}\sqrt{\sum_{k=1}^{K}\Theta_{ij}^{(k)^2}}$ 时,式(22)为 $L_{2,1}$

① Learning structural changes of Gaussian graphical models in controlled experiments. http://arxiv.org/abs/1203.3532, 2012,5,15

范数罚多任务组结构图套索[24]；当 $\|\boldsymbol{\Theta}\|_{q,1}$ 为 $L_{\infty,1}$ 范数罚 $\|\boldsymbol{\Theta}\|_{\infty,1}=\sum_{j=2}^{P}\sum_{i=1}^{P-1}\max_{k\in\{1,\cdots,K\}}|\Theta_{ij}^{(k)}|$ 时，式(22)为 $L_{\infty,1}$ 范数罚多任务组结构图套索[24-25]. $L_{2,1}$ 范数罚和 $L_{\infty,1}$ 范数罚能促使不同概率图模型对应的精度矩阵 $\hat{\boldsymbol{\Theta}}^{(k)}=\boldsymbol{\Sigma}_k^{-1}$ 具有类似的稀疏结构.

此外，Guo 等人[49]针对概率图模型的多任务学习提出了一种具有两层结构的罚函数 $\lambda_1\sum_{i\neq j}\theta_{ij}+\lambda_2\sum_{i\neq j}\sum_{k=1}^{K}|\gamma_{ij}^{(k)}|$，他们令精度矩阵中的元素 $\Theta_{ij}^{(k)}=\theta_{ij}\gamma_{ij}^{(k)}$，其中 $i\neq j$ 且 $i,j\in\{1,\cdots,P\}$，$k=1,\cdots,K$. 该罚函数相当于在两层水平上进行稀疏学习：θ_{ij} 若非零(为零)则全部 K 个概率图模型中的节点 i 与 j 之间均存在(不存在)一条边，此为第一层水平上的稀疏学习；在 θ_{ij} 非零的前提下，$\gamma_{ij}^{(k)}$ 为零(非零)表示第 k 个概率图模型的节点 i 与 j 之间存在(不存在)一条边，此为第二层水平上的稀疏学习. 显然，θ_{ij} 的作用为促使各概率图模型间具有类似的结构，而 $\gamma_{ij}^{(k)}$ 的作用为促使各概率图模型在某种程度上还具有自己独有的特性. 另外还有研究指向一种较为特殊的多任务稀疏化学习问题——随机变量所服从的概率分布随时间变化的概率图模型的多任务稀疏化学习问题[50-52].

2.4.4 小结与分析

概率图模型的多任务稀疏化学习方法主要采用两种罚函数：一种为融合罚，包括两两融合罚和有序融合罚，前者对全部元素两两之间的差的绝对值都进行惩罚，而后者则假设元素是有序的，只对前后相邻的两个元素之差的绝对值进行惩罚；另一种为 $L_{1,q}$ 范数罚，包括 $L_{2,1}$ 范数罚和 $L_{\infty,1}$ 范数罚. 当有多个结构上具有某种共性的概率图模型需要进行稀疏化学习时，若对这些概率图模型分别独立进行稀疏化学习则无法揭示其内在的联系，所以必须对这些概率图模型联合进行稀疏化学习以揭示出其内在共性，此即图模型的多任务稀疏化学习.

3 概率图模型的稀疏化学习中常用的求解算法

概率图模型稀疏化学习本质上是一个最优化问题，该最优化问题由两部分组成：似然函数和正则化项，不同概率图模型的稀疏化学习对应的优化问题的求解算法不同，各求解算法适用的条件也不尽相同，如表 1 所示.

表 1 概率图模型的各稀疏化学习方法比较

稀疏化学习方法		各自特点	比较
L_1 范数罚稀疏化学习	高斯无向图图套索	使用了 L_1 范数罚，适用于马尔可夫网络的稀疏化学习，是最早被提出的概率图模型稀疏化学习方法	该四类方法均利用了 L_1 范数罚作为正则化项对不同情形下的概率图模型进行稀疏化学习，由于 L_1 范数罚有偏估计的特点，故其缺点均为关于精度矩阵的有偏估计
	部分随机变量不可观图套索	使用了 L_1 范数罚，适用于某些随机变量不可观测的图模型的稀疏化学习，本质为矩阵的低秩稀疏分解	
	有向无环图图套索	使用了 L_1 范数罚，适用于因果贝叶斯网络的稀疏化学习，其中因果方向的学习需要采用实验干预法才能实现，否则只能学习出有向无环图的骨架	
	伊辛模型图套索	使用了 L_1 范数罚，适用于伊辛模型的稀疏化学习，本质为求解 P 个独立的 L_1 范数罚逻辑斯蒂回归问题	
无偏稀疏化学习	SCAD 图套索	通过使用无偏 SCAD 罚实现无偏稀疏化学习	该四类方法均具有无偏估计的优点，其原理均为从调整精度矩阵中各个元素被惩罚的权重大小方面入手，从而大大地降低了有偏估计的程度
	自适应图套索	使用了自适应的 L_1 范数罚，本质为两步的估计，通过第 2 步对第 1 步中的权重进行合理调整实现无偏的稀疏化学习	
	贝叶斯自适应图套索	贝叶斯框架下的两步估计，通过第 2 步对第 1 步中的权重进行合理调整实现无偏稀疏学习	
	幂法则图套索	使用了幂法则罚，对各个节点惩罚的权重合理调整，使得自由度高的中枢节点的惩罚权重大大减小，从而促使中枢节点的出现	
结构稀疏化学习	$L_{q,1}$ 范数组结构图套索	使用了 $L_{q,1}$ 范数罚，以组结构为先验信息，使得精度矩阵中的元素具有组稀疏化特点	该四类方法均以精度矩阵中的元素具有某种已知结构作为先验信息，将该先验信息引入罚函数中，从而实现概率图模型的结构化的稀疏化学习
	$L_{F,1}$ 范数组结构图套索	使用了 $L_{F,1}$ 范数罚，以组结构为先验信息，使得精度矩阵中的元素具有组稀疏化特点	
	双稀疏图套索	同时使用了 $L_{q,1}$ 范数罚和 L_1 范数罚，将与某节点相连的全部边划分为同一个分组，能够使得节点和边同时稀疏	
	局部共性图套索	本质为对精度矩阵中相邻元素之差进行惩罚，揭示出相邻节点的共性	

（续　表）

稀疏化学习方法		各自特点	比较
多任务稀疏化学习	多任务两两融合图套索	同时使用了两两融合罚和 L_1 范数罚，两两融合罚对全部元素两两之间的差的绝对值都进行惩罚从而促使不同概率图模型对应的精度矩阵具有类似的稀疏结构	该三类方法均对多个概率图模型同时进行稀疏化学习，在利用 L_1 范数罚实现稀疏化学习的同时利用两两融合罚、有序融合罚和 $L_{q,1}$ 范数罚揭示出不同概率图模型之间存在的共性
	多任务有序融合图套索	同时使用了有序融合罚和 L_1 范数罚，有序融合罚对前后相邻的两个元素之差的绝对值进行惩罚从而促使前后相邻的概率图模型对应的精度矩阵具有类似的稀疏结构	
	多任务组结构图套索	同时使用了 $L_{q,1}$ 范数罚和 L_1 范数罚，$L_{q,1}$ 范数罚促使不同概率图模型对应的精度矩阵具有类似的稀疏结构	

高斯无向图图套索是第一种被提出的概率图模型的稀疏化学习方法，因此研究其求解算法的文献较多，其求解算法有组坐标下降方法[5,53-54]（Block Coordinate Descent Method）、非精确内点算法[55]（Inexact Interior-Point Method，IIPM）、投影子梯度方法[21]（Projected Subgradient Method）、贪心坐标上升算法[56]（Greedy Coordinate Ascent Approach）、交替线性化方法[57]（Alternating Linearization Methods）和二次近似算法[58]（Quadratic Approximation）. 高斯无向图图套索对应的优化问题实际上是一个对数行列式半定规划问题（logdeterminant Semidefinite Programming Problems，log-det SDP），但由于该问题往往是高维甚至超高维的，因此传统的半定规划求解算法（如 SDPT3[59] 和 SeDuMi[60]）并不适用. 针对该问题，文献[55]给出了一种非精确内点算法（Inexact Interior-Point Method，IIPM）来求解该半定规划问题，该非精确内点算法在确定搜索方向时为了减少算法的空间复杂性和时间复杂性，通过非精确方式（数值迭代）求得搜索方向的近似替代公式，因此该非精确内点算法非常适用于处理高维的概率图模型稀疏化学习问题. Duchi 等人[20]将高斯无向图图套索的优化问题进行对偶变换，然后利用投影子梯度法求解其对偶问题. 组坐标下降算法的前身是坐标下降算法，该算法在求解优化问题时每次只涉及单个坐标块，同时令其余全部坐标块保持不变，因此大大简化了优化问题. 文献[5]中利用组坐标下降算法求解概率图模型稀疏化学习对应的优化问题时，每次更新精度矩阵中的第 i 行和第 i 列（精度矩阵为对称矩阵，其第 i 行和第 i 列中元素相同），即固定精度矩阵中的其他全部元素的同时，只关于精度矩阵中的第 i 行和第 i 列求解优化问题. 显然，该方法将复杂的多维优化问题转化为容易求解的低维优化问题，文献[5]中给出的求解高斯无向概率图模型的稀疏化学习对应的优化问题的算法中，将多维优化问题转化为每次求解一个 L_1 正则化问题，可以直接利用套索（Lasso）的各种高效解法（例如最小角回归算法）来求解该 L_1 正则化问题. 但值得注意的是该方法得到的精度矩阵可能不是对称的，尤其在正则化参数被选定为较小的数值时非对称性更加明显[55]，而精度矩阵的非对称性会造成负的特征值，给后续应用主成分分析等分析方法造成了困难. 因此，文献[61]中给出了解决精度矩阵非对称问题的方法，例如可以根据得到的精度矩阵的上三角部分转置后作为下三角部分，同时上三角部分保持不变，从而把非对称精度矩阵转化为对称精度矩阵. （块）坐标下降算法是求解概率图模型稀疏化学习对应的优化问题的最广泛的解法，文献[62]中还利用组坐标下降算法对协方差矩阵而不是其逆矩阵进行稀疏学习. Scheinberg 等人[56]提出一种贪心坐标上升算法，该算法直接对高斯无向图图套索的优化问题求解，不像 Friedman 等人[5]的组坐标下降算法和 Duchi 等人[20]的投影子梯度算法等算法那样求解其对偶优化问题；另外，该算法在每次迭代中只关于精度矩阵中的一个对角元素或两个对称的非对角元素求解优化问题，不像 Friedman 等人的组坐标下降算法那样每次迭代时关于精度矩阵中对称的一行和一列求解优化问题. 交替方向乘子法[63]也是求解图套索问题的一种重要算法，其一般要求解的优化问题形式为

$$\min\ f(\boldsymbol{x})+g(\boldsymbol{z}) \tag{23a}$$

$$\text{s.t. } \boldsymbol{A}\boldsymbol{x}+\boldsymbol{B}\boldsymbol{z}=\boldsymbol{c} \tag{23b}$$

其中 $\boldsymbol{x}\in\boldsymbol{R}^n$，$\boldsymbol{z}\in\boldsymbol{R}^m$，$\boldsymbol{A}\in\boldsymbol{R}^{q\times n}$，$\boldsymbol{B}\in\boldsymbol{R}^{q\times m}$，$\boldsymbol{c}\in\boldsymbol{R}^q$，写出其增广拉格朗日函数为

$$L_\rho(\boldsymbol{x},\boldsymbol{z},\boldsymbol{y})=f(\boldsymbol{x})+g(\boldsymbol{z})+\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{B}\boldsymbol{z}-\boldsymbol{c})+\frac{\rho}{2}\|\boldsymbol{A}\boldsymbol{x}+\boldsymbol{B}\boldsymbol{y}-\boldsymbol{c}\|_2^2 \tag{24}$$

因此其求解步骤一般如下：

$$\boldsymbol{x}^{k+1}=\arg\min_{\boldsymbol{x}} L_\rho(\boldsymbol{x},\boldsymbol{z}^k,\boldsymbol{y}^k) \tag{25}$$

$$\boldsymbol{z}^{k+1}=\arg\min_{\boldsymbol{z}} L_\rho(\boldsymbol{x}^{k+1},\boldsymbol{z},\boldsymbol{y}^k) \tag{26}$$

$$\boldsymbol{y}^{k+1}=\boldsymbol{y}^k+\rho(\boldsymbol{A}\boldsymbol{x}^{k+1}+\boldsymbol{B}\boldsymbol{z}^{k+1}-\boldsymbol{c}) \tag{27}$$

其中 $\rho>0$. Scheinberg 等人[57]提出一种交替线性化方法，该方法实际上是一种交替方向乘子法的改进方法，与交替方向乘子法的不同之处在于该方法在每次迭代时都在当前点处将函数 $g(\boldsymbol{z})$ 替换为一

个一阶线性项与另一近似项之和，而且每次迭代时求解的子问题都具有显式解. Hsieh 等人[58]提出一种二次近似(Quadratic Approximation)算法，该方法同时结合了二次近似、牛顿方法和坐标下降的思想，其首先利用二阶泰勒展开公式对原优化问题的目标函数进行二次近似，然后利用坐标下降方法和 Armijo 规则分别得到牛顿下降方向和步长，故该方法为二阶方法，而且具有超线性收敛(superlinearly convergent)的特点.

其他概率图模型稀疏化学习对应的优化问题的求解算法如下所述. Ye 等人①利用划分 Bregman 方法(Split Bregman Method)[64]求解部分变量不可观图套索，该方法通过引入辅助变量解决部分变量不可观图套索的目标函数中变量不可分离的问题，并且收敛速度快. Liu 等人[18]将幂法则图套索对应的优化问题等价地转化为一系列再权 L_1 范数正则化(reweighted L_1 regularization)问题，然后利用 MM 算法[65](Majorize-Minimization algorithm)求解该问题，该方法用求解一系列近似子问题来逼近原目标优化问题. Yang 等人(见本文第 2 页脚注①)利用谱投影梯度法[66]求解多任务有序融合图套索，谱投影梯度法是对投影梯度算法的改进，主要用于克服前者收敛速度慢的缺点. 投影梯度算法存在两方面的问题：一是每次选择最速下降方向会导致收敛速度变慢；二是投影步骤的计算复杂度过高. 谱投影梯度法在迭代过程中采用非单调线搜索技术，不要求每次迭代后目标函数值都下降，只要求在规定的最近某些次迭代目标函数下降即可，并且结合谱梯度法的 Barzilai-Borwein 步长来选择谱投影梯度法在迭代过程中的步长. 谱投影梯度法适用于投影步骤计算高效的情形，因此投影步骤的计算方法非常关键. 文献中已有的概率图模型的稀疏化学习对应的优化问题的求解算法如表 2 所示，值得指出的是 SCAD 图套索在应用组坐标下降算法前需要对其目标函数中的 SCAD 罚函数进行局部线性近似[67](Local Linear Approximation，LLA)，具体说来就是在已知点 w_0 将 SCAD 罚函数 $p_{\lambda,a}(|w|)$ 用如下的近似表达式表示：

$$p_{\lambda,a}(|w|)\approx p_{\lambda,a}(|w_0|)+p'_{\lambda,a}(|w_0|)(|w|-|w_0|) \tag{28}$$

其中 $p'_{\lambda,a}(w)=\frac{\partial}{\partial w}p_{\lambda,a}(w)$. 另外，贝叶斯自适应图套索的求解需要利用吉布斯抽样(Gibbs sampling)方法，文献[24]中利用交替方向乘子法求解多任务两两融合图套索和多任务组结构图套索，而 $L_{\infty,1}$ 范数组结构图套索的求解算法利用了投影子梯度法[20]，文献[11]中利用组坐标下降算法求解有向无环图图套索. 值得指出的是，伊辛模型图套索的求解利用了邻域选择方法[26](neighborhood selection method)，该方法在对伊辛模型稀疏化学习时依次将每个随机变量作为输出变量而其余全部随机变量作为自变量，即转化为利用套索求解关于一系列回归模型的变量选择问题. 各算法的特点如表 3 所示.

表 2 文献中提出的对概率图模型稀疏化学习的求解算法

稀疏化学习方法	求解算法
高斯无向图图套索	组坐标下降算法，非精确内点算法，投影子梯度算法，贪心坐标上升算法，交替线性化方法，二次近似算法
部分随机变量不可观图套索	划分 Bregman 方法
有向无环图图套索	组坐标下降算法
伊辛模型图套索	邻域选择方法
SCAD 图套索	局部线性近似＋组坐标下降算法
自适应图套索	组坐标下降算法
贝叶斯自适应图套索	吉布斯抽样(Gibbs sampling)方法
幂法则图套索	MM 算法
$L_{\infty,1}$ 范数组结构图套索	投影子梯度法
$L_{F,1}$ 范数组结构图套索	组坐标下降算法
双稀疏图套索	组坐标下降算法
局部共性图套索	组坐标下降算法
多任务两两融合图套索	交替方向乘子法
多任务有序融合图套索	谱投影梯度法
多任务组结构图套索	交替方向乘子法，组坐标下降算法

表 3 概率图模型稀疏化学习的求解算法的特点

算法	特点
组坐标下降算法	要求目标函数变量块可分离或具有显式解，要求子问题的求解复杂性尽可能低
非精确内点算法	通过非精确方式(数值迭代)求得搜索方向的近似替代公式，适用于处理高维情形
投影子梯度法	整个算法的高效与否依赖于投影步骤的计算复杂度
贪心坐标上升算法	求解原优化问题，每次迭代只涉及一个对角元素或两个对称的非对角元素
二次近似算法	利用二阶泰勒展开公式对原优化问题的目标函数进行二次近似，具有超线性收敛的特点
交替方向乘子法	要求分解子问题的求解复杂性低
交替线性化方法	是交替方向乘子法的改进形式，其特点同于交替方向乘子法
吉布斯抽样方法	对于高维问题来说算法复杂度过高
MM 算法	能够求解等式和不等式约束优化问题，可把非可微问题转化为可微问题，可对模型参数进行分离单独求解，缺点是多步迭代，需要近似子问题算法复杂性尽可能低
划分 Bregman 方法	通过引入辅助变量解决变量不可分离的问题，收敛速度快
谱投影梯度法	克服了投影梯度收敛速度慢的缺点，但整个算法的高效与否依赖于投影步骤的计算复杂度

① Efficient latent variable graphical model selection via split Bregman method. http://arxiv.org/abs/1110.3076, 2011, 10, 13

文献[58]中对求解高斯无向图图模型的各种算法进行了比较：(1)实验数据为人工产生的链图，其精度矩阵为：非对角元素为 $\Theta_{i,i-1}=0.5$ 且对角元素为 $\Theta_{i,i}=1.25$. 设定算法迭代的停止条件为 $\varepsilon=10^{-6}$ 且可调参数 $\lambda=0.4$ 时的实验结果如表 4 所示，其中 P 表示随机变量个数；(2)实验数据为人工生成的随机稀疏图(随机产生的具有一定稀疏结构的概率图模型)，设定算法迭代的停止条件为 $\varepsilon=10^{-6}$ 时的实验结果如表 5 所示，其中 P 表示随机变量个数. 从表 4 和表 5 可以看出，在该实验中二次近似算法在六种算法中最快，而贪心坐标上升算法非常慢，交替线性化方法对于本实验中生成的链图来说速度较快而对于随机稀疏图来说速度很慢；投影子梯度方法与组坐标下降算法速度适中，非精确内点算法比投影子梯度方法和组坐标下降算法慢一些. 表 4 和表 5 中的“—”表示迭代时间超过 30000 s.

表 4　求解高斯无向图图套索优化问题的算法比较(链图)

算法	耗时/s		
	$P=1000$	$P=4000$	$P=10000$
二次近似算法	2.26	53.51	986.6
非精确内点算法	151.2	5754	—
投影子梯度方法	34.91	1258	19251
组坐标下降算法	45.1	2119	—
贪心坐标上升算法	520.8	—	—
交替线性化方法	41.85	1734	28190

表 5　求解高斯无向图图套索优化问题的算法比较(随机稀疏图)

算法	耗时/s					
	$P=1000$		$P=4000$		$P=10000$	
	$\lambda=0.12$	$\lambda=0.075$	$\lambda=0.08$	$\lambda=0.05$	$\lambda=0.08$	$\lambda=0.04$
二次近似算法	1.2	6.87	160.2	478.8	1125	2951
非精确内点算法	116.7	145.8	8097	13650	—	—
投影子梯度方法	59.89	91.7	4232	9541	—	—
组坐标下降算法	20.43	60.61	2561	8356	—	—
贪心坐标上升算法	683.3	4449	—	—	—	—
交替线性化方法	28250	—	—	—	—	—

4　存在的问题与未来研究方向

4.1　存在的问题

虽然概率图模型的稀疏化学习问题近年来得到了广泛研究，但仍然存在大量需要研究的问题. 例如，目前概率图模型的无偏稀疏化学习只利用了无偏的 SCAD 罚，然而还有许多种无偏罚未被考虑，例如 $p\in(0,1)$ 时的 L_p 范数罚、MCP 罚[68] 和截断 L_1 罚[69]. 目前对概率图模型中随机变量服从负二项分布、均匀分布和指数分布等诸多分布时的稀疏化学习缺乏研究. 概率图模型的结构稀疏化学习当前也只涉及到组结构而并未涉及近年来新提出的重叠组结构和树组结构. 许多概率图模型的稀疏化学习方法被陆续提出，但人们往往只是给出了一种稀疏化学习方法，忽视了该稀疏化学习方法的统计性质的讨论(参数估计一致性等). 在算法方面，概率图模型对应的优化问题的求解还未考虑过具有低计算复杂度、小的存储空间和对数据较强的适应能力的在线学习算法.

4.2　未来研究方向

4.2.1　新的无偏稀疏化学习方法

将 SCAD 罚用于无向概率图模型的稀疏化学习中得到的 SCAD 图套索具有无偏估计的特点，未来将 $p\in(0,1)$ 时的 L_p 范数罚、MCP 罚和截断 L_1 罚

$$\rho_{\lambda,\gamma}(\wp)=\min\left(\frac{\gamma\lambda^2}{2},\lambda|\wp|\right) \tag{29}$$

等其他可实现无偏估计的非凸罚应用到无向概率图模型的稀疏化学习中是否会比 SCAD 图套索的稀疏化学习效果更好？这都是值得研究的问题.

4.2.2　将概率图模型的稀疏化学习方法在其他概率分布上推广

目前概率图模型的稀疏化学习问题逐渐从最初的无向高斯图模型向着有向、非参数和随机变量服从其他分布的方向发展，该研究方向尚有大量需要完成的工作，例如，随机变量服从均匀分布和指数分布等诸多分布下的概率图模型的稀疏化学习问题尚未有学者进行研究，而且将可实现无偏估计的 SCAD、MCP 罚和截断 L_1 罚应用到这些概率图模型下进行无偏稀疏化学习的问题尤其值得研究.

4.2.3　拓展概率图模型的结构稀疏化学习

本文中所述 $L_{q,1}$ 范数组结构图套索对精度矩阵中元素利用了 $L_{2,1}$ 范数罚，双稀疏图套索对精度矩阵中的元素施加了 $L_{2,1}$ 范数罚和 L_1 范数罚，此两种结构化的罚函数均为组之间不具有重叠元素的罚函数，如何将具有重叠组结构的罚函数应用到概率图模型的结构稀疏化学习中是一个有待于解决的问题. 另外，组结构图套索和双稀疏图套索由于利用了不可实现无偏估计的 $L_{2,1}$ 范数罚和 L_1 范数罚，所以其一致性等性质不好，而利用结构化的复合非凸罚函数则可解决该问题，复合组桥罚[70] 和复合 MC

罚[71]是两种典型的结构化复合非凸罚函数,猜想将该两种罚函数引入到概率图模型的结构稀疏化学习中会解决上述问题,但遗憾的是尚无学者对该问题进行深入研究.

4.2.4 概率图模型稀疏化学习方法的统计性质

虽然目前有很多形式的图套索被陆续提出,但其统计性质仍然缺少理论上的论证,例如,双稀疏图套索的一致性等统计性质有待于理论上的研究和证明,将邻域选择方法应用到泊松图模型中进行稀疏化学习时泊松图模型重建的一致性和模型参数估计的一致性未被论证,局部共性图套索的模型选择一致性未被研究,该方面仍有大量的理论工作需要完成.

4.2.5 向在线学习领域推广

在线算法以低计算复杂度、小的存储空间和对数据较强的适应能力而被广泛应用,目前套索的在线算法[72-75]和组套索[76]的在线学习算法已经被提出并且在实践中验证了其优势,未来图套索及其各种变种的在线学习算法是很有意义的研究方向.

5 结 语

国内已经有学者研究稀疏学习优化问题[77],而国外稀疏学习研究逐渐从单纯的无结构稀疏学习发展到各种结构化的稀疏学习,其中概率图模型的稀疏学习是近年来研究的热点之一.经过稀疏学习后的概率图模型结构简单却保留了原始概率图模型的重要结构信息,大大简化了概率图模型的结构,同时实现了结构和参数学习,未来概率图模型的稀疏化学习势必在机器学习等领域中发挥越来越重要的作用.

参 考 文 献

[1] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 1996, 58(1): 267-288

[2] Dempster A P. Covariance selection. Biometrics, 1972, 28(1): 157-175

[3] Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. Biometrika, 2007, 94(1): 19-35

[4] Banerjee O, El Ghaoui L, d'Aspremont A. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. The Journal of Machine Learning Research, 2008, 9(3): 485-516

[5] Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 2008, 9(3): 432-441

[6] Dahl J, Vandenberghe L, Roychowdhury V. Covariance selection for nonchordal graphs via chordal embedding. Optimization Methods & Software, 2008, 23(4): 501-520

[7] Chandrasekaran V, Parrilo P A, Willsky A S. Latent variable graphical model selection via convex optimization. The Annals of Statistics, 2012, 40(4): 1935-1967

[8] Ma S, Xue L, Zou H. Alternating direction methods for latent variable Gaussian graphical model selection. Neural Computation, 2013, 25(8): 2172-2198

[9] Lauritzen S, Meinshausen N. Discussion: Latent variable graphical model selection via convex optimization. The Annals of Statistics, 2012, 40(4): 1973-1977

[10] Shojaie A, Michailidis G. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. Biometrika, 2010, 97(3): 519-538

[11] Fu F, Zhou Q. Learning sparse causal Gaussian networks with experimental intervention: Regularization and coordinate descent. Journal of the American Statistical Association, 2013, 108(501): 288-300

[12] Ravikumar P, Wainwright M J, Lafferty J D. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. The Annals of Statistics, 2010, 38(3): 1287-1319

[13] Wainwright M J, Ravikumar P, Lafferty J D. High-dimensional graphical model selection using L_1-regularized logistic regression//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2006: 1465-1472

[14] Lam C, Fan J. Sparsistency and rates of convergence in large covariance matrix estimation. Annals of Statistics, 2009, 37(6B): 4254-4278

[15] Fan J, Feng Y, Wu Y. Network exploration via the adaptive LASSO and SCAD penalties. The Annals of Applied Statistics, 2009, 3(2): 521-541

[16] Khondker Z, Zhu H, Chu H, et al. Bayesian covariance lasso. Statistics and Its Interface, 2013, 6(2): 243-259

[17] Wong E, Awate S, Fletcher P T. Adaptive sparsity in Gaussian graphical models//Proceedings of the 30th International Conference on Machine Learning. Atlanta, USA, 2013: 311-319

[18] Liu Q, Ihler A T. Learning scale free networks by reweighted l_1 regularization//Proceedings of the International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, USA, 2011: 40-48

[19] Schmidt M W, Berg E, Friedlander M P, et al. Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm//Proceedings of the 12th International Conference on Artificial Intelligence and Statistics. Clearwater, USA, 2009: 456-493

[20] Duchi J, Gould S, Koller D. Projected subgradient methods for learning sparse Gaussians//Proceedings of the 24th Conference on Uncertainty in AI. Helsinki, Finland, 2008: 153-160

[21] Kolar M, Liu H, Xing E P. Markov network estimation from multi-attribute data//Proceedings of the 30th International Conference on Machine Learning. Atlanta, USA, 2013:

73-81

[22] Honorio J, Samaras D, Rish I, et al. Variable selection for Gaussian graphical models//Proceedings of the International Conference on Artificial Intelligence and Statistics. La Palma, Spain, 2012: 538-546

[23] Honorio J, Ortiz L E, Samaras D, et al. Sparse and locally constant Gaussian graphical models//Proceedings of the Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2009: 745-753

[24] Danaher P, Wang P, Witten D M. The joint graphical lasso for inverse covariance estimation across multiple classes. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2014, 76(2): 373-397

[25] Honorio J, Samaras D. Multi-task learning of Gaussian graphical models//Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel, 2010: 447-454

[26] Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. The Annals of Statistics, 2006, 34(3): 1436-1462

[27] Efron B, Hastie T, Johnstone I, et al. Least angle regression. The Annals of Statistics, 2004, 32(2): 407-499

[28] Wang H. Bayesian graphical lasso models and efficient posterior computation. Bayesian Analysis, 2012, 7(4): 867-886

[29] Banerjee S, Ghosal S. Posterior convergence rates for estimating large precision matrices using graphical models. Electronic Journal of Statistics, 2014, 8(2): 2111-2137

[30] Liu H, Lafferty J, Wasserman L. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. The Journal of Machine Learning Research, 2009, 10(10): 2295-2328

[31] Liu H, Han F, Yuan M, et al. High-dimensional semiparametric Gaussian copula graphical models. The Annals of Statistics, 2012, 40(4): 2293-2326

[32] Xue L, Zou H. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. The Annals of Statistics, 2012, 40(5): 2541-2571

[33] Voorman A, Shojaie A, Witten D. Graph estimation with joint additive models. Biometrika, 2013, 101(1): 85-101

[34] Allen G I, Liu Z. A log-linear graphical model for inferring genetic networks from high-throughput sequencing data//Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Philadelphia, USA, 2012: 1-6

[35] Yang E, Ravikumar P D, Allen G I, et al. Graphical models via generalized linear models//Proceedings of the Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2012: 1367-1375

[36] Maurya A. A joint convex penalty for inverse covariance matrix estimation. Computational Statistics & Data Analysis, 2014, 75(3): 15-27

[37] Marjanovic G, Solo V. L_0 sparse graphical modeling//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Prague, Czech Republic, 2011: 2084-2087

[38] Geer S, Buhlmann P. L_0-penalized maximum likelihood for sparse directed acyclic graphs. The Annals of Statistics, 2013, 41(2): 536-567

[39] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 2001, 96(456): 1348-1360

[40] Abegaz F, Wit E. Sparse time series chain graphical models for reconstructing genetic networks. Biostatistics, 2013, 14(3): 586-599

[41] Peterson C, Vannucci M, Karakas C, et al. Inferring metabolic networks using the Bayesian adaptive graphical lasso with informative priors. Statistics and Its Interface, 2013, 6(4): 547-558

[42] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2006, 68(1): 49-67

[43] Jacob L, Obozinski G, Vert J P. Group lasso with overlap and graph lasso//Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Canada, 2009: 433-440

[44] Liu J, Ye J. Moreau-Yosida regularization for grouped tree structure learning//Proceedings of the Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2010: 1459-1467

[45] Kim S, Xing E P. Tree-guided group lasso for multi-task regression with structured sparsity//Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel, 2010: 543-550

[46] Hoefling H. A path algorithm for the fused lasso signal approximator. Journal of Computational and Graphical Statistics, 2010, 19(4): 984-1006

[47] Daye Z J, Jeng X J. Shrinkage and model selection with correlated variables via weighted fusion. Computational Statistics & Data Analysis, 2009, 53(4): 1284-1298

[48] Tibshirani R, Saunders M, Rosset S, et al. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005, 67(1): 91-108

[49] Guo J, Levina E, Michailidis G, et al. Joint estimation of multiple graphical models. Biometrika, 2011, 98(1): 1-15

[50] Kolar M, Song L, Ahmed A, et al. Estimating time-varying networks. The Annals of Applied Statistics, 2010, 4(1): 94-123

[51] Zhou S, Lafferty J, Wasserman L. Time varying undirected graphs. Machine Learning, 2010, 80(2-3): 295-319

[52] Kolar M, Xing E P. On time varying undirected graphs//Proceedings of the International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, USA, 2011: 407-415

[53] Tseng P. Convergence of a block coordinate descent method for nondifferentiable minimization. Journal of Optimization Theory and Applications, 2001, 109(3): 475-494

[54] Luo Z Q, Tseng P. On the convergence of the coordinate descent method for convex differentiable minimization. Journal of Optimization Theory and Applications, 1992, 72(1): 7-35

[55] Li L, Toh K C. An inexact interior point method for L_1-regularized sparse covariance selection. Mathematical Programming Computation, 2010, 2(3-4): 291-315

[56] Scheinberg K, Rish I. Learning sparse Gaussian Markov networks using a greedy coordinate ascent approach// Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part III. Barcelona, Spain, 2010: 196-212

[57] Scheinberg K, Ma S, Goldfarb D. Sparse inverse covariance selection via alternating linearization methods//Proceedings of the Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2010: 2101-2109

[58] Hsieh C J, Sustik M A, Dhillon I S, et al. Sparse inverse covariance matrix estimation using quadratic approximation// Proceedings of the Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems. Granada, Spain, 2011: 2330-2338

[59] Toh K C, Todd M J, Tütüncü R H. SDPT3—A MATLAB software package for semidefinite programming, version 1.3. Optimization Methods and Software, 1999, 11(1-4): 545-581

[60] Sturm J F. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. Optimization Methods and Software, 1999, 11(1-4): 625-653

[61] Rolfs B T, Rajaratnam B. A note on the lack of symmetry in the graphical lasso. Computational Statistics & Data Analysis, 2013, 57(1): 429-434

[62] Wang H. Coordinate descent algorithm for covariance graphical lasso. Statistics and Computing, 2014, 24(4): 521-529

[63] Boyd S, Parikh N, Chu E, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning, 2011, 3(1): 1-122

[64] Goldstein T, Osher S. The split Bregman method for L_1-regularized problems. SIAM Journal on Imaging Sciences, 2009, 2(2): 323-343

[65] Hunter D R, Lange K. A tutorial on MM algorithms. The American Statistician, 2004, 58(1): 30-37

[66] Schmidt M W, Murphy K P, Fung G, Rosales R. Structure learning in random fields for heart motion abnormality detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, USA, 2008: 1-8

[67] Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models. Annals of Statistics, 2008, 36(4): 1509-1533

[68] Zhang C H. Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics, 2010, 38(2): 894-942

[69] Zhang T. Multi-stage convex relaxation for feature selection. Bernoulli, 2013, 19(5B): 2277-2293

[70] Seetharaman I. Consistent Bi-Level Variable Selection Via Composite Group Bridge Penalized Regression[Ph. D. dissertation]. Kansas State Univesity, Kansas, America, 2013

[71] Jiang D. Concave Selection in Generalized Linear Models [Ph. D. dissertation]. The University of Iowa, Iowa, America, 2012

[72] Langford J, Li L, Zhang T. Sparse online learning via truncated gradient. The Journal of Machine Learning Research, 2009, 10(3): 777-801

[73] Singer Y, Duchi J C. Efficient learning using forward-backward splitting//Proceedings of the Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2009: 495-503

[74] Xiao L. Dual averaging methods for regularized stochastic learning and online optimization. The Journal of Machine Learning Research, 2010, 11(10): 2543-2596

[75] Balakrishnan S, Madigan D. Algorithms for sparse linear classifiers in the massive data setting. The Journal of Machine Learning Research, 2008, 9(2): 313-337

[76] Yang H, Xu Z, King I, et al. Online learning for group Lasso//Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel, 2010: 1191-1198

[77] Tao Qing, Gao Qian-Kun, Jiang Ji-Yuan, Chu De-Jun. Survey of solving the optimization problems for sparse learning. Journal of Software, 2013, 24(11): 2498-2507(in Chinese) (陶卿, 高乾坤, 姜纪远, 储德军. 稀疏学习优化问题的求解综述. 软件学报, 2013, 24(11): 2498-2507)

LIU Jian-Wei, born in 1966, Ph.D., associate professor. His main research interests include intelligent information processing, analysis, prediction, controlling of complicated system, and analysis of the algorithm and the designing.

CUI Li-Peng, born in 1990, M. S. candidate. His main research interest is sparsity model of machine learning.

LUO Xiong-Lin, born in 1963, Ph. D., professor. His main research interests include intelligent control, and analysis, prediction, controlling of complicated system.

Background

Probabilistic Graphical Models and sparse learning are two hot topics in bioinformatics, machine learning and artificial intelligence. Therefore, the joint of the two topics arouses people's great interests in recent years. The sparse learning of the probabilistic graphical models can greatly simplify the structure of the probabilistic graphical models, improve the probabilistic graphical models' generalization ability and retain the important information simultaneously. The sparse learning of the probabilistic graphical models is gradually expended from the simple L_1-norm penalty sparse learning of the Gauss undirected graph to the structure sparse learning, unbiased sparse learning, multi-task sparse learning and many other probabilistic graphical models such as directed acyclic graph, Ising model and Poisson undirected graph.

In this paper, we offer a detailed survey of the sparse learning of the probabilistic graphical models. We point out the motivations and characteristics of the different methods of sparse learning. Furthermore, we summarize the algorithms of the optimization problems in the sparse learning of the probabilistic graphical models and point different algorithms' advantages and disadvantages. In the end, we point the remaining problems in the field of the sparse learning of the probabilistic graphical models and give the meaningful directions of the future research.

This work is supported by the National Basic Research Program (973 Program) of China (2012CB720500), the National Natural Science Foundation of China (21006127), and the Basic Scientific Research Foundation of China University of Petroleum (JCXK-2011-07).