

含隐变量非高斯无环因果模型的估计算法

姜 枫¹, 朱辉生², 汪 卫²

(1. 南京理工大学泰州科技学院计算机科学与技术系, 南京 225300; 2. 复旦大学计算机科学技术学院, 上海 200433)

摘 要: 针对观测变量中含隐变量的非高斯线性无环因果模型的估计问题, 提出一种新的算法。通过在超完备基独立成分分析算法中引入满足 Oracle 性质的惩罚因子, 使混合矩阵的估计值具有稀疏连接权值, 由此推导出模型估计算法。实验结果表明, 该算法能够改进因果模型估计的精确程度, 提高算法效率。

关键词: 线性因果模型; 超完备基; 独立成分分析; 稀疏连接

Estimation Algorithm for Non-Gaussian Acyclic Causal Model with Latent Variables

JIANG Feng¹, ZHU Hui-sheng², WANG Wei²

(1. Dept. of Computer Science and Technology, Taizhou Institute of Science, Nanjing University of Science and Technology, Nanjin 225300;

2. School of Computer Science, Fudan University, Shanghai 200433)

【Abstract】 In order to estimate the non-Gaussian linear acyclic causal models with latent variables, a new algorithm is proposed. The algorithm is derived by integrating penalty functions which satisfy the Oracle property into overcomplete Independent Component Analysis(ICA), making the entries of mixing matrix sparse. Experimental results show that the algorithm improves accuracy and efficiency of the casual model estimation.

【Key words】 linear causal model; overcomplete basis; Independent Component Analysis(ICA); sparse connection

1 概述

线性因果模型(也称为结构方程模型)在自然科学和社会科学中有着广泛的应用, 涉及到金融学、政治学、社会学和生物学等多个领域。在过去几十年中, 相关算法的研究都隐含了噪声变量具有高斯分布这个假设。然而, 在现实世界中, 许多噪声变量的概率统计特征并不满足高斯分布, 因此, 这类算法有很大的局限性。文献[1]针对非高斯分布变量的线性因果模型问题提出了一种基于独立成分分析(Independent Component Analysis, ICA)的算法。文献[2]进一步分析了在观测变量含有隐变量(即观测变量数目小于源变量数目)时线性因果模型的估计方法。在此基础上, 本文在 ICA 算法中引入惩罚因子, 产生稀疏连接矩阵, 并据此推导出一种新的含隐变量线性因果模型的估计算法。

2 线性因果模型

线性因果模型具有以下性质^[1]:

(1) 观测变量为 $x_i, i = \{1, 2, \dots, n\}$, 可以排列成一个因果序列。因果序列是指原因必定在前、结果必定在后的序列, 即 x_j 可以由 $x_i (i < j)$ 完全决定, 且不依赖于 $x_i (i > j)$, 此因果序列用 k_i 表示。这样的变量可以用递归的方式生成, 并且可以表示成有向无环图^[3]。

(2) 每个变量 x_i 可用线性函数表示:

$$x_i = \sum_{k(j) < k(i)} b_{ij} x_j + e_i + c_i$$

其中, b_{ij} 表示变量 x_i 与 x_j 的连接权值; e_i 为噪声变量; c_i 是可选常量。

(3) 噪声项 e_i 是具有连续值的随机变量, 服从方差非 0 的非高斯分布, 各 e_i 之间相互独立, 即 $p(e_1, e_2, \dots, e_m) = \prod_i p_i(e_i)$ 。

满足以上 3 条性质的模型称为线性非高斯无环模型(LiNGAM), 该模型的估计算法在文献[1]中也有详细描述。

本文的 LiNGAM 中加入了一条新的假设: 观测变量是 x_i 的子集, 即观测变量的下标集合 $J \subseteq \{1, 2, \dots, m\}$ 。或者可以这样认为, 观测变量只包含 $x_j, j \in J$ 。本文模型与通用 LiNGAM 模型的不同点在于观测变量中允许含有隐变量, 即只能观测到部分变量。图 1 给出了 3 个符合 LiNGAM 模型的例子, 可以用有向无环图表示。

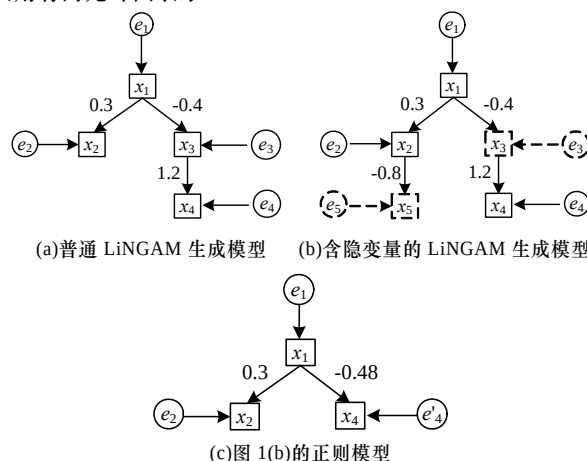


图 1 符合 LiNGAM 性质的数据生成模型

基金项目: 国家自然科学基金资助项目(60303008); 国家“973”计划基金资助项目(2005CB321905)

作者简介: 姜 枫(1980—), 男, 讲师、硕士研究生, 主研方向: 智能信号处理; 朱辉生, 副教授、博士; 汪 卫, 教授、博士

收稿日期: 2009-10-12 **E-mail:** ferrarimodena@sina.com

图 1(b)所示模型中的观测变量集合为向量 $\mathbf{x}=(x_1, x_2, x_4)^T$ ，隐变量 x_3 没有子节点，它对观测变量集合没有影响。因此，即便不能从观测变量集合 $\mathbf{x}$ 中推断出隐变量 x_3 以及连接权值 b_{25} ，也并不影响整个 LiNGAM 模型的估计。另外一个隐变量为 x_3 ，它处于观测变量 x_1 和 x_4 之间，于是有

$$x_3 := -0.4x_1 + e_3$$

$$x_4 := 1.2x_3 + e_4$$

可以推导出：

$$x_4 := 1.2(-0.4x_1 + e_3) + e_4 = -0.48x_1 + 1.2e_3 + e_4 = -0.48x_1 + e_4'$$

其中， $e_4' = 1.2e_3 + e_4$ 。这表明，可以从模型中删除 x_3 以获得一个更简单的模型，且获得的简单模型与原模型具有相同的因果关系，由此得到的模型如图 1(c)所示。于是，可以定义一个含隐变量的 LiNGAM 模型的正则模型(canonical model)^[2]为：该模型的每个隐变量是一个根节点(没有父节点的节点)且至少有 2 个子节点。

3 模型估计算法

3.1 普通 LiNGAM 模型估计算法

普通 LiNGAM(不含隐变量)的模型估计算法^[1]步骤如下：

(1)对于给定的 $m \times n$ 数据矩阵 X ， m 表示观测变量的维数； n 表示观测变量的长度。首先，从 X 的每一列中减去其均值，使 X 转换为 0 均值。然后，对于 $X = Ae$ ，使用 ICA 算法求解 $W = A^{-1}$ 。

(2)寻找矩阵 W 的列转置，使其生成一个主对角线不含 0 元素的矩阵 W 。

(3)将矩阵 W 的每一行除以其对角线元素，生成矩阵 W' ，易知 W' 的主对角线元素均为 1。

(4)计算矩阵 B 的估计值 $\hat{B} = I - W'$ 。

(5)搜索因果序列，即寻找一个转置矩阵 P ，使 $\tilde{B} = P\hat{B}P^T$ 尽量为严格的下三角矩阵。

3.2 含隐变量的 LiNGAM 模型估计算法

在上节算法中，由于观测变量不含隐变量，观测变量的数目和源变量数目相同，因此 A 是严格的方阵，步骤(1)中只需要使用普通 ICA 算法，如 FastICA，即可求解 W 。而对于含有隐含变量的 LiNGAM 模型，观测变量集合 X' 是 X 的子集，因此，混合矩阵 A' 只包含 A 中观测变量集合的相应行，是一个行数小于列数的矩阵。这类问题称为具有超完备基(overcomplete basis)ICA，此类问题的求解见文献[4]。

在使用相关算法计算出 A 之后，进而利用 A 计算出所有观察等价文献[2]的正则含隐变量的 LiNGAM 模型，算法如下：

(1)对 A 的列进行重新排序，使隐变量排列在前。

(2)在矩阵 A 的顶部加上隐藏部分，将 A 扩展为 $\tilde{A}$ 。

(3)测试是否可以通过转置(行和列分别转置)的方法将 $\tilde{A}$ 变换成下三角矩阵。如果不可以，则继续测试下一种分类。

(4)将 $\tilde{A}$ 的每一列除以其主对角线元素，使主对角线元素均为 1。计算 $\tilde{B} = I - \tilde{A}^{-1}$ 。

(5)检查 $\tilde{B}$ 是否满足稳定性^[3]假设。如果不满足，则继续测试下一种分类。

(6)将 $\tilde{B}$ 加入观测等价模型列表中。

4 稀疏连接权值的超完备基 ICA

在含隐变量的 LiNGAM 模型估计算法中，需要使用超完备基 ICA 算法估计混合矩阵 A ，并将 A 扩展为 $\tilde{A}$ 。对于

LiNGAM 模型，根据定义， $\tilde{A}$ 应能通过转置得到一个下三角矩阵。然而在实际情形中，由于样本数目有限以及噪声的客观存在等因素，因此 $\tilde{A}$ 的估计值经常是非稀疏的，即 $\tilde{A}$ 中含有较多非 0 元素，从而很难通过矩阵转置得到一个严格的下三角矩阵。尽管文献[1]提出了一种权值边缘修剪算法，但这种方法效率并不高。对于稀疏连接权值的矩阵 A ，如果可以在估计 A 时使其含有尽可能多的 0，满足其稀疏性，则可直接提高估计的准确性并由此降低算法的复杂度。

4.1 稀疏连接系数惩罚因子的生成

为了在估计稀疏矩阵时尽可能使其真值为 0 的权值系数自动调整为 0，可以通过引入惩罚因子达到这一目的。

文献[5]指出，一个可用来生成稀疏系数的高性能惩罚因子应具备如下 3 个性质：

(1)无偏性，对于真值非 0 的系数估计是无偏的；

(2)稀疏性，这意味着一些真值为 0 的系数会被自动设置为 0，从而降低模型的复杂度；

(3)连续性，对不同数据产生的结果估计是连续的。

众所周知， L_2 惩罚因子 $p_\lambda(|\theta|) = \lambda|\theta|^2$ 生成脊回归， L_1 惩罚因子 $p_\lambda(|\theta|) = \lambda|\theta|$ 产生软阈值规则，而 L_q 惩罚因子 $p_\lambda(|\theta|) = \lambda|\theta|^q$ 生成桥回归，只有当 $q \geq 1$ 时，解是连续的。然而，当 $q > 1$ 时，不能使真值为 0 的系数设置为 0。因此，只有 L_1 惩罚因子可以生成稀疏解，它的缺点是对真值非 0 的系数估计也会产生影响。因此， L_q 惩罚因子不能满足无偏性、稀疏性、连续性这 3 个要求。平滑截断绝对偏差(Smoothly Clipped Absolute Deviation, SCAD)^[5]是一种满足以上 3 个性质的惩罚因子，其导数公式如下：

$$P'_\lambda(\theta) = \lambda\{I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda}I(\theta > \lambda)\}, \quad a > 2, \theta > 0$$

其中， $I(\cdot)$ 是指示函数； $(\cdot)_+$ 是标准运算符。 a 的典型取值为 3.7。通过选取适当的 θ 值，SCAD 惩罚因子可以满足 Oracle 性质^[5]：使真值为 0 的系数的估计值自动调整为 0，并且不影响其余非 0 系数的估计。

下面对 SCAD 进行修改，考虑一个更通用的 SCAD 惩罚因子，称为 G-SCAD：

$$P'_\lambda(\theta) = \lambda_1\{I(\theta \leq \lambda_2) + \frac{(a\lambda_2 - \theta)_+}{(a-1)\lambda_2}I(\theta > \lambda_2)\}$$

其中， λ_1 用于控制惩罚因子的影响系数； λ_2 指定惩罚因子的影响范围。相比 SCAD，G-SCAD 的表示方式更精确。在实际使用中，可以选择 λ_2 略大于 λ_1 ，这样惩罚因子可以作用在一个较大的范围内。

4.2 稀疏连接权值的超完备基 ICA 算法

对于含隐变量的 LiNGAM 模型，其观测的变量数少于源变量数。在 ICA 模型中，这是一种混合信号数少于独立成分数的情况，称为具有超完备基的 ICA。此时，采用普通 ICA 方法求解非常困难。

文献[4]提出了一种通过学习数据的过完备表示求解混合矩阵 A 的算法：

$$\Delta A \propto AA^T \nabla \ln P(\mathbf{x} | A) \approx -A(z\mathbf{s}^T + I)$$

其中， $z_k = \partial \ln P(s_k) / \partial s_k$ ； $P(\cdot)$ 为概率密度函数； I 为单位矩阵。然而，对于稀疏矩阵 A ，噪声的存在以及对源数据概率分布估计的偏差等诸多因素， A 的估计值 $\tilde{A}$ 中元素几乎都是非 0 值，不满足稀疏性条件，从而增加了模型估计算法的复

杂度。因此，如何在计算 A 时使其尽量接近真值，满足矩阵 A 的稀疏性，对于整个模型的估计具有重要的意义。

由 4.1 节讨论可知，通过在混合矩阵 A 的求解算法中加入 SCAD 惩罚因子可以推导出稀疏连接权值的超完备基 A ：

$$\Delta A \propto -A(\hat{z}\hat{z}^T + I) - [P'_\lambda(a_{ij})]$$

5 计算机仿真实验

5.1 稀疏连接权值的超完备基 ICA

本实验主要验证第 4 节中提出的针对稀疏连接权值超完备基 ICA 算法的有效性及其效率。实验使用 4 组相互独立的源信号 $s = [s_1, s_2, s_3, s_4]^T$ ：

$$s_1(t) = \sin(v/2)$$

$$s_2(t) = t^2 \cos^2(t/4)$$

$$s_3(t) = ((\text{rem}(v, 27) - 13)/9)$$

$$s_4(t) = ((\text{rem}(v, 23) - 11)/9)^5$$

其中， $\text{rem}(\cdot)$ 是取余函数；混合矩阵为

$$A = \begin{bmatrix} 0 & 0 & 0 & 0.2705 \\ 0.5967 & 0 & 0 & 0.3479 \\ 0 & 0.5750 & 0 & -0.7571 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

A 中只有少量元素为非 0 值，是一个稀疏矩阵。混合模型为 $x = As$ ，观测变量为 $\tilde{x} = [x_1, x_2, x_3]^T$ ， x_4 为隐变量。

本文分别使用不带惩罚因子的超完备基 ICA 算法、带有 SCAD 惩罚因子的超完备基 ICA 算法以及带有 G-SCAD 惩罚因子的超完备基 ICA 算法求解混合矩阵 A 。SCAD 项中参数设置为 $\lambda = 0.05$ ， $a = 3.7$ 。G-SCAD 项中参数设置为 $\lambda_1 = 0.05$ ， $\lambda_2 = 0.07$ ， $a = 3.7$ 。

如图 2 所示，使用带惩罚因子的超完备基 ICA 算法在估计值结果和迭代速度两方面都明显优于不带惩罚因子的算法，而 G-SCAD 惩罚因子在性能上略强于 SCAD 惩罚因子。

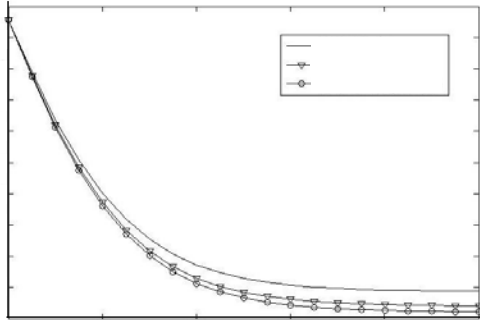


图 2 混合矩阵元素 a_{23} 的估计值

注意到使用带有 G-SCAD 惩罚因子的算法最终得到的 a_{23} 估计值为 0.0113，而非真值 0。这是因为在计算 $P'_\lambda(\theta)$ 时，如果 θ 接近 0，不能使用梯度下降法来最优化代价函数，所以在实际计算中采用了近似的方法，使用 $\tanh(m\theta)$ (其中， m 是一个较大的数，实验中取 180) 来近似 $|\theta|$ 的导数，G-SCAD 在实际使用时使用公式：

$$P'_\lambda(\theta) = \lambda_1 \{ \tanh(m\theta) I(\theta \leq \lambda_2) + \tanh(m\lambda_2) \frac{(a\lambda_2 - \theta)_+}{(a-1)\lambda_2} I(\theta > \lambda_2) \}$$

所以，对 θ 的估计产生了一定误差，即便其真值应该为 0，估计值也不会是 0，而是收敛到一个非常小的实数。

5.2 含隐变量的 LiNGAM 模型估计

本实验将稀疏连接的超完备基 ICA 算法应用于含隐变量的 LiNGAM 模型的估计问题中，通过以下步骤验证算法的有效性。

(1) 假设观测变量数为 3，隐变量数目为 1，设置稀疏矩阵为

$$B = \begin{bmatrix} 0 & 0 & 0 & 0.2705 \\ 0.5967 & 0.5750 & 0 & 0.3479 \\ 0 & 0 & 0 & -0.7571 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

并由此矩阵生成 LiNGAM 模型。

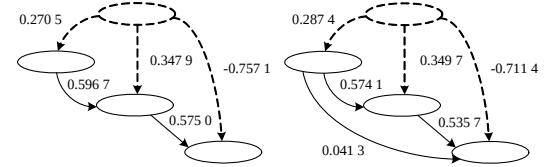
(2) 随机生成数据 e_i ，保证 e_i 满足非高斯分布，并且相互独立。

(3) 将 e_i 通过本文模型生成观测变量集合 x_j 。

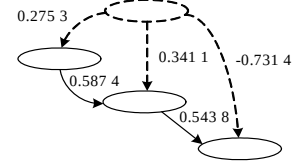
(4) 利用稀疏连接权值超完备基 ICA 算法计算过完备基 A 。

(5) 利用 3.2 节中所述算法估计 LiNGAM 模型。

图 3 为实验结果，从中可以看出，使用文献[2]中算法计算出的 x_1 和 x_3 的连接权值 b_{13} 为 -0.0413，而真实的模型中 b_{13} 为 0，这与使用本文算法得出的结论(图 2(c))是一致的。



(a) 随机生成的 LiNGAM 模型 (b) 文献[2]算法计算出的系统模型



(c) 本文算法计算出的系统模型

图 3 实验结果

6 结束语

本文将稀疏连接权值的独立成分分析方法引入超完备基的 ICA 估计问题中。通过在超完备基 A 的迭代算法中加入 G-SCAD 惩罚因子，使其系数具有 Oracle 性质，即使实际连接权值为 0 的系数的估计值自动调整为 0，以此提高模型估计的准确性，降低系统复杂度，改进算法性能。计算机仿真实验验证了算法的有效性和效率。

参考文献

- [1] Shimizu S, Hoyer P, Hyvarinen A. A Linear, Non-Gaussian Acyclic Model for Causal Discovery[J]. Journal of Machine Learning Research, 2006, 7(12): 2003-2030.
- [2] Hoyer P O, Shimizu S, Kerminen A J. Estimation of Linear, Non-Gaussian Causal Models in the Presence of Confounding Latent Variables[C]//Proc. of the 3rd European Workshop on Probabilistic Graphical Models. Prague, Czech: [s. n.], 2006.
- [3] Pearl J. Causality: Models, Reasoning, and Inference[M]. [S. l.]: Cambridge University Press, 2000.
- [4] 李拥军, 江宇闻, 朱思铭. 基于最短路径和自然梯度的过完备 ICA 算法[J]. 计算机工程, 2006, 32(15): 16-18.
- [5] Fan Jianqing, Li Runze. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties[J]. Journal of the American Statistical Association, 2001, 96(456): 1348-1360.

编辑 张正兴