arXiv:2303.01626v1 [stat.ME] 2 Mar 2023

具有潜变量的藤依赖图：作为基因表达数据的摘要

Xinyao Fan[1],Harry Joe[1],和 Yongjin Park[1]

[1]英属哥伦比亚大学统计学系，加拿大

2023年2月

摘要

高通量测序技术的出现带来了大量比较基因组序列。在基因组尺度上构建基因-基因相互作用网络或依赖图，对于理解生物过程的调控至关重要。不同的依赖图可以提供不同的信息。

一些基于高阶偏相关构建依赖图的现有方法较为稀疏；当存在能够解释基因组中大部分依赖关系的潜变量时，这些方法提供的信息有限。另一些基于相关性和一阶偏相关构建依赖图的方法，则可能产生稠密图。当基因可以划分为若干组，且基因表达在组内的依赖关系强于组间依赖关系时，我们提出一种基于截断藤和潜变量的依赖图，该方法利用分组信息和低阶偏相关。所得图并不稠密，而在藤依赖图中可能更为核心的基因拥有更多邻居。我们在两个RNA测序数据集——酵母和前列腺癌数据集——上展示了依赖图构建方法的应用。有一些生物学证据支持所得依赖图中基因之间的关系。

我们提供了一个灵活的框架，通过低阶偏相关和组的构建来建立依赖图，从而得到既不过于稀疏也不过于稠密的图。我们预计，这种方法将有助于识别可能在不同生物学功能中发挥核心作用的组。

关键词：偏相关、高斯因子模型、依赖图、基因-基因网络、截断藤、潜在依赖、RNA测序

1 引言

随着准确性和通量的快速提升，高通量测序技术使研究人员能够常规测量数万个基因的mRNA转录本水平。由于观测到的基因表达值源于基因之间的相互作用，在大量样本中获得基因表达谱，自然引发了科学界对

根据观测数据推断此类相互作用模式的兴趣。基因-基因相互作用网络揭示了复杂生物通路及重要细胞过程的功能组织结构。构建准确的基因相互作用网络能够加深我们对复杂生物系统的理解，并为正确解读大型高维基因组学数据提供工具。自微阵列技术问世以来， [28], RNA 测序 [30], 和单细胞RNA测序 [23]发展至今，从观测数据中发现潜在的网络结构一直是生物学领域的一项根本性重要问题。

统计学中已提出许多方法，通过构建统计（条件）依赖图，来理解大规模高维数据矩阵中蕴含的交互模式。传统上，推断问题被表述为：在高斯图模型或其他相关条件独立图的类别中寻找最优模型 [32]。在这类方法中，我们直接根据从观测数据估计出的逆协方差矩阵，测量高阶偏相关值，并据此识别一种依赖图。然而，寻找如此复杂的依赖图在计算上可能难以处理，导致图模型对观测数据过拟合；为此，人们施加了严格的正则化惩罚项，参见 [21]及其中的参考文献。或者，依赖图也可以由低阶偏相关构成，这只需依赖相对快速且可靠的估计。传统生成树模型 [7] 和藤结构 [19]就是其中的示例。尽管基于低阶偏相关模式的方法通常能在样本量有限时，得到比高阶图更具信息量的依赖图 [24], ，但我们在真实世界数据中发现了挑战：这类数据通常会受到少数潜在组效应的影响。我们表明，除非妥善处理能够解释观测数据中很大一部分方差的潜在组变量，否则学习条件独立图并不合适，还可能导致得到不理想的稀疏图。

本研究将提出一种从基因表达数据中推断依赖图的新方法论，以低阶藤图模型为基础 [19] ，同时纳入潜在组变量，以提高观测变量之间关系的解析度，并刻画潜在组结构引入的干扰变量之间的关系。

2 现有方法概述

本节概述了几种构建依赖图的方法。这些方法均基于从协方差矩阵或相关矩阵计算得到的偏相关来定义，因此在第 2.2 节概述这些方法之前，第 2.1 节先给出了偏相关的主要公式。这些方法包括带阈值化的条件独立图、 [24] 中的 FOCI方法，以及截断偏相关藤。还有一种更基础的图基于阈值化相关法：如果两个变量的绝对相关超过给定阈值，则在依赖图中将它们连接起来。第 2.3 节包含一个基于因子结构的相关矩阵简单案例，用于说明阈值化相关法往往过于稠密，而条件独立图往往过于稀疏

当维度 d 较大时。

2.1 偏相关

本节总结了偏相关方面的一些必要结果，以解释构建依赖图的不同方法。这里假设对每个变量进行适当变换后，各变量联合服从多元高斯分布。

设（变换后的）观测向量为 $\boldsymbol{X}_{\mathcal{I}}=(X_1、.....、X_d)$，其中 $d\geq 2$ 较大，且 $\mathcal{I}$ 为指标集 $\{1,2,\ldots,d\}$；假设 $\boldsymbol{X}_{\mathcal{I}}$ 服从多元高斯分布，记作 $\boldsymbol{X}_{\mathcal{I}}\sim N(\mu,\ \boldsymbol{\Sigma})$，其中 $\boldsymbol{\Sigma}$ 为协方差矩阵。设 $S\subseteq\mathcal{I}$ 和的基数至少为 2，即 $|S|\geq 2$。对于一对 $i,\ j\in S,\ i\neq j$，将移除 S 后的 $L_{\{i,j\}}$ 记作 $=S_{-\{i,j\}}=:S\backslash\{i,j\}$。相应的随机向量记作 $\boldsymbol{X}_{L_{\{i,j\}}}:=\{X_k,k\in L_{\{i,j\}}\}$。将 $\rho_{ij;L_{\{i,j\}}}$ 定义为偏相关，即条件（高斯）分布中 $[X_i,X_j|\boldsymbol{X}_{L_{\{i,j\}}}]$ 的相关参数。它量化了在不受 $\boldsymbol{X}_{L_{\{i,j\}}}$ 线性效应影响的情况下，X_i 与 X_j 之间的依赖关系。

$L_{\{i,j\}}$ 的基数指偏相关系数的阶数。当阶数为零时，$|S|=|\{i,j\}|=2$ 和 $L_{\{i,j\}}=\emptyset$。于是 $\rho_{i,\ j;\emptyset}=\rho_{ij}$，即 X_i 与 X_j 之间的标准两两相关，$i,\ j\in\mathcal{I},\ i\neq j$。对于任意满足 $|L_{\{i,j\}}|>1$ 的阶数 $|L_{\{i,j\}}|$，[1] 推导出一个公式，可根据低阶（偏）相关递归计算偏相关。设 $|S|\subseteq\mathcal{I}$，其基数为 $|S|\geq 3$。考虑一个包含三个不同指标的集合 $\{i,j,k\}\subseteq\tilde{S}$。令 $L:=S_{-\{i,j,k\}}$ 和 $L_{\{i,j\}}=L\cup k$。那么

$$\rho_{i,j;L_{\{i,j\}}}=\frac{\rho_{i,j;\tilde{L}}-\rho_{i,k;\tilde{L}}\rho_{j,k;\tilde{L}}}{\sqrt{1-\rho^2_{i,k;\tilde{L}}}\sqrt{1-\rho^2_{j,k;\tilde{L}}}} \tag{1}$$

设精度矩阵（协方差矩阵的逆）为 $\boldsymbol{\Sigma}^{-1}=\ (\sigma^{ij})$，且 $\mathcal{I}_{ij}=\mathcal{I}\backslash\{i,j\}$。对于逆相关矩阵，众所周知，根据 [32] 第 5.8 节，当 $i\neq j$ 时，

$$\rho_{ij;\mathcal{I}_{\{ij\}}}=-\frac{\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{jj}}} \tag{2}$$

这些是最高阶偏相关。

2.2 依赖图概述

条件依赖图的阈值化

令 $\mathcal{V}=\{1,\ldots,d\}$ 为顶点集合，每个顶点对应一个变量。边集 E 由无序对 $(i,\ j)$、$(i,\ j)$ 组成；若 X_i 和 X_j 之间存在一条边，则 ($\in$ 属于 E)。对于理论协方差矩阵，若 $X_i\not\perp X_j\big|X_{L_{\{i,j\}}}$，则 X_i 和 X_j 在图中相连；仅当 $X_i\perp X_j\big|X_{L_{\{i,j\}}}$ 时，它们才在图中不相连。也就是说，当 $\sigma^{ij}=0$ 时，变量 $i,\ j$ 之间没有边；等价地，对 $i\neq j,\ \rho_{ij;L_{\{i,j\}}}=0$。

对于样本协方差矩阵，[32] 第6.1节提出了一种带有阈值的朴素程序。基于样本的条件独立图遵循以下规则：对于两个

变量，如果这两个变量在式（2）中的绝对高阶偏相关系数低于阈值，则不将它们之间的边纳入图中。

FOCI（一阶条件独立方法）

在 [24], 提出的 FOCI 方法中，只有当分析中不存在其他变量使得 X_i 和 X_j 条件独立，或导致关联反转（即，给定某个变量 X_k 时的偏相关与 X_i、 X_j 之间的相关符号相反）时，变量 X_i 和 X_j 才会相连。他们的方法定义了一种修正的一阶偏相关，类似于公式 (1)。

更形式化地说，当且仅当不存在其他变量 k （$k \neq i \neq j$）使得 $\hat{}\rho_{ij;k} \approx 0$ 或 $\hat{}\rho_{ij;k} < 0$ 时，e_{ij} 才是 $\mathcal{G}$ 中的一条边，其中

$$\hat{\rho}_{ij;k} = \frac{|\rho_{ij}| - |\rho_{ik}||\rho_{jk}|}{\sqrt{1-\rho_{ik}^2}\sqrt{1-\rho_{jk}^2}}$$

使用一个阈值来判断其与 0 的接近程度。

截断偏相关藤

[3] 提出将藤作为一种图形对象来概括条件依赖；藤的定义见后文第 2.4 节。一个常见用途是寻找相关矩阵的参数化方式，从而用低阶偏相关来概括依赖关系；也就是说，高阶偏相关的绝对值较小。下面简要概述一种实现这一目标的算法。

首先，构造一棵包含 $(d-1)$ 条边的最大生成树，将绝对相关性最大的变量连接起来。只有当 A-B-C 是树上的一条路径时，这棵树才能概括一组变量的全部依赖关系；此时，给定变量 B，变量 A 和 C 条件独立。也就是说，如果最大生成树概括了依赖关系，那么就存在马尔可夫依赖，即给定连接两个变量的路径上的变量，这两个变量条件独立。如果基于这棵树得到的相关矩阵与（经验）相关矩阵相差较大，则根据给定一个条件变量时较大的绝对偏相关，添加连接两个变量的边（这两个变量分别与该条件变量相连）。随后，还可以根据给定两个或更多变量的条件添加更多边。关于 Vines 的更多细节见第 2.4 节。

2.3 比较

本节总结了对不同依赖图的一些评论。本文给出了一个数值示例，说明当潜变量能够解释变量之间的大部分依赖关系时，经过阈值化的条件独立图可以产生稀疏图。本文还对这一性质给出了理论解释。

- 相关性阈值化：这种方法的主要问题是，如果变量之间大多存在至少中等程度的相关性，就会产生稠密图。

对于高维数据，如果存在一簇或一组高度相关的变量，那么依赖图中对应的簇将近乎构成完全图。此外，这种方法不会考虑任何条件依赖信息。

- 带阈值化的条件依赖图：

  该方法依赖于逆相关矩阵中的非零元素，这些元素与给定其余变量时两个变量之间的偏相关有关。然而，图中的高阶交互通常不易解释。对于基于结构化因子模型的高维相关矩阵，如果每个变量至少载荷于一个潜在因子，那么从理论上讲，逆相关矩阵中不会有零元素；例如，参见 [14]第 3.10.4 节中逆相关矩阵的形式。但随着维度不断增大，逆相关矩阵中的许多元素会趋近于零。考虑一个单因子模型，并假设模型中的依赖关系很强。 [9]表明，在这种情况下，随着维度趋于无穷，潜变量可以根据观测变量一致地恢复。这意味着，给定其余变量时两个变量之间的条件相关，在渐近意义上等同于给定潜变量时两个变量之间的条件相关；后者为零。给定一个阈值，如果新增变量持续以中等程度载荷于该潜变量，那么随着变量数量增加，依赖图往往会变得更加稀疏。因此，当维度较大时，基于精度矩阵中的非零元素构建图，对于高斯因子依赖结构而言并无信息价值。

  当 $\mathbf{\Sigma}$ 来自与潜变量无关的模型时，条件依赖图可能很有用。一个例子见 [15] 的命题 2，其中 $\mathbf{\Sigma}$ 基于截断偏相关藤。基于 $\mathbf{\Sigma}^{-1}$的条件独立图简约且信息丰富。若藤截断阶数为 m，则所得条件独立图中的边对应于藤中第 1 至 m 棵树中的边；这些边涉及$\sum_{i=1}^{m}(d-i)$ 个非零项，位于 $\mathbf{\Sigma}^{-1}$中。剩余的 $(d-m)(d-m-1)/2$ 个位置在 $\mathbf{\Sigma}^{-1}$ 中均为零。

- FOCI：该方法可视为直接阈值法的改进，因为它考虑了一阶偏相关。然而，该方法仅考虑（修正后的一阶偏相关。对于因多个潜变量而高度相关的变量（假设仅存在正向依赖），该方法会生成一个稠密的连通图，因为它没有利用高阶偏相关的信息。

- 截断偏相关藤：第 2.4 节详细介绍了偏相关藤。在介绍基于藤的依赖图之前，下面的例子将说明，对于因子结构，它既能避免基于相关阈值化的图过于稠密，也能避免条件独立图过于稀疏。原因在于，它充分利用了强相关以及一些额外的低阶偏相关。

在基于藤的依赖图中，第一棵树是最大生成树，其边以黑色显示，边上标注了变量对之间的相关性。此外，还绘制了其他颜色的边，以表示超过阈值的某些偏相关。蓝色、红色和绿色的边分别表示给定一个、两个及更多变量时变量之间的偏相关。

用于比较不同方法的示例下面提供一个小型示例，用于比较各种方法构建的不同依赖图。数据模拟自潜变量为 W 的单因子模型：

$$Z_j = \alpha_j W + \psi_j \epsilon_j \quad j \in \{1, \ldots, D\}, \tag{3}$$

其中 W，ϵ_1，ϵ_2，.....是相互独立的 $N(0, 1)$ 随机变量，且对所有 j，$-1 < \alpha_j < 1$，$\psi_j^2 = 1 - \alpha_j^2$。设载荷矩阵/向量为 $A = [\alpha_1, \ldots, \alpha_d]^T$，$\Psi^2 = \text{diag}(\psi_1^2, \ldots..., \psi_d^2)$。$(Z_1, \ldots..., Z_d)^T$的相关矩阵为 $\mathbf{\Sigma} = AA^T + \Psi^2$，而 $(Z_1, \ldots..., Z_d, W)^T$的相关矩阵为 $\mathbf{\Sigma}^* = \begin{bmatrix} \mathbf{\Sigma} & A \\ A^T & 1 \end{bmatrix}$.

假设 $d = 10$ 因子载荷是从区间 $[0.5, 0.9]$中均匀生成的。因子载荷较弱时，需要更大的 d 才能展示这些比较。在一次模拟中，按降序排列后，A 中的数值为 $[0.90, 0.86, 0.81, 0.77, 0.72, 0.68, 0.63, 0.59, 0.54, 0.50]^T$。四种方法均应用于 $\mathbf{\Sigma}$和 $\mathbf{\Sigma}^*$；前一个矩阵是后一个矩阵去掉最后几行和最后一列后得到的子矩阵。相关矩阵 $\mathbf{\Sigma}^*$ 为：

$$\begin{bmatrix}
1.00 & 0.77 & 0.73 & 0.69 & 0.65 & 0.61 & 0.57 & 0.53 & 0.49 & 0.45 & 0.90 \\
0.77 & 1.00 & 0.69 & 0.66 & 0.62 & 0.58 & 0.54 & 0.50 & 0.47 & 0.43 & 0.86 \\
0.73 & 0.69 & 1.00 & 0.62 & 0.59 & 0.55 & 0.51 & 0.48 & 0.44 & 0.41 & 0.81 \\
0.69 & 0.66 & 0.62 & 1.00 & 0.55 & 0.52 & 0.49 & 0.45 & 0.42 & 0.38 & 0.77 \\
0.65 & 0.62 & 0.59 & 0.55 & 1.00 & 0.49 & 0.46 & 0.43 & 0.39 & 0.36 & 0.72 \\
0.61 & 0.58 & 0.55 & 0.52 & 0.49 & 1.00 & 0.43 & 0.40 & 0.37 & 0.34 & 0.68 \\
0.57 & 0.54 & 0.51 & 0.49 & 0.46 & 0.43 & 1.00 & 0.37 & 0.34 & 0.32 & 0.63 \\
0.53 & 0.50 & 0.48 & 0.45 & 0.43 & 0.40 & 0.37 & 1.00 & 0.32 & 0.29 & 0.59 \\
0.49 & 0.47 & 0.44 & 0.42 & 0.39 & 0.37 & 0.34 & 0.32 & 1.00 & 0.27 & 0.54 \\
0.45 & 0.43 & 0.41 & 0.38 & 0.36 & 0.34 & 0.32 & 0.29 & 0.27 & 1.00 & 0.50 \\
0.9 & 0.86 & 0.81 & 0.77 & 0.72 & 0.68 & 0.63 & 0.59 & 0.54 & 0.50 & 1.00
\end{bmatrix}$$

根据（2），基于 $\boldsymbol{\Sigma}$ 的偏相关矩阵为：

$$\begin{bmatrix}
1.00 & 0.29 & 0.24 & 0.20 & 0.17 & 0.15 & 0.13 & 0.12 & 0.11 & 0.09 \\
0.29 & 1.00 & 0.18 & 0.15 & 0.13 & 0.11 & 0.10 & 0.09 & 0.08 & 0.07 \\
0.24 & 0.18 & 1.00 & 0.12 & 0.11 & 0.09 & 0.08 & 0.07 & 0.06 & 0.06 \\
0.20 & 0.15 & 0.12 & 1.00 & 0.09 & 0.08 & 0.07 & 0.06 & 0.05 & 0.05 \\
0.17 & 0.13 & 0.11 & \substack{\circ\ 0\circ \\ 09} & 1.00 & 0.07 & 0.06 & 0.05 & 0.05 & 0.04 \\
0.15 & 0.11 & 0.09 & 0.08 & 0.07 & 1.00 & 0.05 & 0.05 & 0.04 & 0.04 \\
0.13 & 0.10 & 0.08 & 0.07 & 0.06 & 0.05 & 1.00 & 0.04 & 0.04 & 0.03 \\
0.12 & 0.09 & 0.07 & 0.06 & 0.05 & 0.05 & 0.04 & 1.00 & 0.03 & 0.03 \\
0.11 & 0.08 & 0.06 & 0.05 & 0.05 & 0.04 & 0.04 & 0.03 & 1.00 & 0.03 \\
0.09 & 0.07 & 0.06 & 0.05 & 0.04 & 0.04 & 0.03 & 0.03 & 0.03 & 1.00
\end{bmatrix}$$

以及基于 $\boldsymbol{\Sigma}^*$ 的偏相关矩阵：

$$\begin{bmatrix}
1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.53 \\
0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.42 \\
0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.36 \\
0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.31 \\
0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.27 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.24 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.21 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.19 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.17 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.15 \\
0.53 & 0.42 & 0.36 & 0.31 & 0.27 & 0.24 & 0.21 & 0.19 & 0.17 & 0.15 & 1.00
\end{bmatrix}$$

4种依赖图方法的相连边汇总见表1。

方法	相连边	边数
$\boldsymbol{\Sigma}$		
TC	完全图	55
CDG	1–2, 1–3, 1–4	3
FOCI	1–2, 1–3, 1–4, 1–5, 1–6, 1–7, 1– 8, 1–9, 1–10, 2–3, 2–4, 2–5, 2–6, 3–4, 3–5	15
Vine	1–2, 1–3, 1–4, 1–5, 1–6, 1–7, 1– 8, 1–9, 1–10, 2–3, 2–4, 2–5, 2–6	13
$\boldsymbol{\Sigma}^*$		
TC	完全图	66
CDG	1–W, 2–W, 3–W, 4–W, 5–W, 6–W, 7–W	7
FOCI	1–W, 2–W, 3–W, 4–W, 5–W, 6–W, 7–W, 8–W, 9–W, 10–W	10
Vine	1–W, 2–W, 3–W, 4–W, 5–W, 6–W, 7–W, 8–W, 9–W, 10-W	10

表1：将四种方法应用于相关矩阵 $\boldsymbol{\Sigma}$ 和 $\boldsymbol{\Sigma}^*$时得到的相连边。潜变量记为 W。缩写中，TC 表示相关性阈值法，CDG 表示经阈值化的条件依赖图。TC、CDG 和 FOCI 的阈值均固定为 0.2。对于阈值为 0.2 的 TC，依赖图是完全图，因此不列出其边。上图中藤的截断阶数为 2（藤的高阶树中的高阶偏相关非常小），绘制边时偏相关的阈值设为 0.2。下图中藤的截断阶数为 1，因为这足以复现单因子模型的实际图，其中每个（观测）变量 Z_j 都与潜变量 W相连。

根据表1，对于基于因子模型的高度相关变量，直接阈值法得到的图是稠密的，而随着更多变量与潜变量相连，条件独立图会变得更加稀疏。在这种情况下，依赖于低阶偏相关的 FOCI 和藤结构方法能够提供更简约的依赖图。引入潜变量后，藤依赖图中的每个观测变量Z_j 都有一条边与潜变量 W 相连，这与潜变量模型和结构方程模型中使用的依赖图一致；模型 (3) 意味着 $\mathrm{Cov}\,(Z_j, Z_j|W) = 0$对于 $i \neq j$ 成立，因此无需添加表示条件依赖的边。对于不包含潜变量的图，与 W 相关性最高的变量是 Z_1，藤依赖图首先将其余每个 Z_j 与 Z_1相连；但给定 Z_1 时，其余变量之间存在条件依赖——由 (1) 和 $\boldsymbol{\Sigma}$, $\rho_{2,3;1} = (0.69 - 0.77 \times 0.73)/\sqrt{(1 - 0.77^2)(1 - 0.73^2)} = 0.29$ 以及$\rho_{9,10;1} = (0.27 - 0.49 \times 0.45)/\sqrt{(1 - 0.49^2)(1 - 0.45^2)} = 0.06$可知，因此在表1中，当阈值为 0.2 时，会添加表示条件依赖的边 [2–3]，但不会添加 [9–10]。

在简单的单因子设定中，纳入潜变量时，FOCI方法得到的图与藤方法相同。然而，FOCI方法只考虑一阶偏相关，而藤图考虑的是变量之间更高阶的条件关系。

对于含有多个潜变量的结构化因子模型（载荷矩阵中存在结构化零），例如变量属于互不重叠的组、由双因子模型模拟的数据（见 [14]的第3.11.1节），FOCI方法无法恢复潜变量模型的依赖结构，而且图会变得稠密，以解释各组内变量之间的依赖关系。

可以构造数值示例，其中 d 较大，并且双因子载荷具有结构

结构，且每组包含大量变量。观测变量与全局潜变量之间的相关性可以从一个正区间（例如 $[a_0, b_0] = [0.3, 0.8]$ ）中抽样；而在给定全局潜变量的情况下，同一组中观测变量与局部潜变量之间的偏相关可以从另一个区间（例如 $[a_1, b_1] = [0.4, 0.7]$）中抽样，从而使同一组中的变量具有更强的相关性。[9]中的理论表明，在一些温和假设下，当每组的维度较大时，可以根据双因子模型中的观测变量一致地估计潜变量。因此，随着 d 变大，给定其余变量后两个变量之间的偏相关将趋近于 0。在双因子情形下，对四种方法所得依赖图进行比较的结论将与单因子情形类似。

2.4 截断偏相关藤的背景

前一个示例用于说明，基于藤的依赖图包含基于某些较大绝对值偏相关的边。本节总结截断偏相关藤背后的数学原理。

作为图形对象的**藤**是由一系列连接树组成的 d 变量序列。第一棵树概括 $d-1$ 对变量的边，其中 d 变量作为节点。后续树概括条件依赖关系。藤作为树序列的数学定义将在 [3]中给出。 $\mathcal{V}$ 是一个包含 d 个变量的正规藤，索引为 1、.....、 d，其中 $\mathrm{E}(\mathcal{V}) = \bigcup_{i=1}^{d-1} \mathrm{E}(T_i)$表示 $\mathcal{V}$的边集，若

1. $\mathcal{V} = (T_1$、.....、 $T_{d-1})$ [由 $d-1$ 棵树组成]2. T_1 是一个节点为 $N(T_1) = \{1, 2, \ldots, d\}$、边为 $\mathrm{E}(T_1)$ 的连通树。对于 $l > 1$， T_l 是一个节点为 $N(T_l) =$ 、边为 $\mathrm{E}(T_{l-1})$ [的树，树中的边在下一棵树中变为节点]；3.（邻近条件）对于 $l = 2$、.....、 $d-1$，对于 $\{n_1, n_2\} \in \mathrm{E}(T_l)$, $\#(n_1 \Delta n_2) = 2$，其中 Δ表示对称差，#表示基数 [通过一条边连接的节点相差两个元素]

示例：上述定义中的术语通过以下例子说明：变量 $d = 4$ 用索引 1、2、3、4 表示。假设 $[1,2], [1,3], [2,4]$ 是树 T_1 中的边。对于树 2，我们可以有两条边 $[2,3;\ 1]$，其节点分别为 $[1,2]$ 和 $[1,3],$，以及 $[1,4;\ 2]$，其节点为 $[1,2]$ 和 $[2,4]$。如果 $n_1 = [1,2]$ 且 $n_2 = [2,4]$, 那么 $n_1 \Delta n_2 = \{1,4\}$ 的基数为 2，且 $n_1 \cap n_2 = \{2\}$；树 2 中连接节点 n_1、 n_2 的边记作 $[1,4;\ 2]$——注意，无法由 $[1,3]$ 和 $[2,4]$ 构造出边，因为如果 $n_1 = [1,3]$ 且 $n_2 = [2,4]$，那么 $n_1 \Delta n_2 = \emptyset$ 的基数不为 2，且不满足邻近条件。对于树 3，节点 $[2,3;\ 1]$ 和 $[1,4;\ 2]$ 之间有边 $[3,4;\ 1,2]$，因为这两个节点的对称差为 $\{3,4\}$，其基数为 2。

如果藤中各树的边取值为 $\rho_{a,b}$ （对应于树 1 中的边 (a,b)），而树 1 之后的边 $e = (a,b;S)$ 的取值为 $\rho_{a,b;S}$，则 $d \times d$ 相关矩阵可参数化为一个偏相关藤，其中包含 $d(d-1)/2$ 个在 $(-1, 1)$ 内代数独立的参数，这些参数由这些边上的偏相关取值确定；参见 [19]。对相关矩阵进行重新参数化，使其成为偏相关藤，有许多不同的方法

。对于任意正规藤结构 $\mathcal{V}$，正定相关矩阵中 $d(d-1)/2$ 相关的条目，与偏相关藤中代数独立的 $(d-1)$ 个相关和 $(d-1)(d-2)/2$ 个偏相关构成的集合之间，存在一一对应关系。

上述示例中的藤对应于参数为 $(\rho_{12},\ \rho_{13},\ \rho_{24};\ \rho_{23;1},\ \rho_{14;2},\ \rho_{34;12})$ 的偏相关藤。另一个包含 4 个变量的偏相关藤，其参数为 $(\rho_{12},\ \rho_{13},\ \rho_{14};\ \rho_{23;1},\ \rho_{24;1},\ \rho_{34;12})$。

截断藤 是一种简约的表示方式，用于表示 d 个变量的依赖关系。通常的做法是构建各棵树：在树 1 中放入相关性最强的变量对；在低阶树中放入偏相关最强的变量对。一个 m-截断藤（其中 $m \ll d$）假设，变量之间的大多数依赖关系都由该藤的前 m 棵树 $V_m = (T_1、\ldots、T_m)$ 捕捉。其余的树具有零或较弱的条件依赖（对于精确的 m-截断偏相关藤，剩余偏相关为零；对于近似的 m-截断偏相关藤，偏相关较弱且低于某个阈值）。

截断偏相关藤和藤依赖图的算法

可以使用顺序最大生成树（MST）算法来寻找截断偏相关藤，使高阶树中的偏相关较小。对于任意正则藤结构 $\mathcal{V}$，相关矩阵的对数行列式 $\log\det(\boldsymbol{R}) = \sum_{e\in \mathrm{E}(\mathcal{V})} \log(1-\rho_e^2)$ [20]，，其中 $\{\rho_e\}$ 是藤的边上代数独立的相关系数和偏相关。良好的 m 阶截断偏相关藤应当得到一个近似于 $\boldsymbol{R}$ 的相关矩阵；这意味着

$$-\sum_{e\in T_1,\ldots,T_m} \log(1-\rho_e^2) \approx -\log\det(\boldsymbol{R}), \tag{4}$$

因为这意味着树 $m+1$、.....、$d-1$ 中的偏相关接近于 0。

对于（4），顺序最大生成树算法中边 e 的边权重选为 $-\log(1-\rho_e^2)$，该算法概述于 [14] 的第 6.17 节。基于局部最优树 T_l 构建第 $(l+1)$ 棵树 T_{l+1}。一般而言，在每一树层级上局部最优的算法并非全局最优；只有当 $d \le 7$ 时，才可能枚举所有藤。为确定截断阶数 m，[4] 提出了一个比较拟合指数；当样本量相对于 d 足够大时，该指数适用于样本相关矩阵。否则，当树中的绝对偏相关开始低于指定阈值时，可以停止顺序最大生成树算法。

如果能够找到式（4）左侧的最优解，那么它的边可能大多与最大生成树算法得到的边相同，但不一定属于同一棵树；这不会影响对与许多其他节点/基因存在强关联的节点或基因的识别。

对于藤依赖图，如果两个变量在截断藤的某棵树中构成一条边，且其绝对偏相关超过阈值，则这两个变量相连。路径中间隔着一个或少数几个变量的变量通常具有较弱的条件独立性

给定中间变量，但它们之间存在不可忽略的依赖关系。

3 方法论

对于包含大量变量的数据集，可以预期这些变量能够划分为互不重叠或彼此重叠的组，而分组信息有助于深入了解变量之间的依赖关系。我们介绍如何寻找各组，并添加基于组的潜变量，从而得到基于截断藤的依赖图。首先概述各个步骤，随后详细说明其实现方法。

如果变量不服从正态分布，那么第一步处理是通过概率积分变换进行秩变换，将其转换为标准正态 $N(0，1)$。

假设有 d 个变量（例如 d 个基因的基因表达测量值），样本量为 n。假设这些变量之间具有单调关系，因此使用相关性进行汇总是有意义的。令数据向量为（x_{i1}，……，x_{id}），其中 $i = 1，2，……，n$。对于固定的变量索引 j，

$$z_{ij} = \Phi^{-1}\left(\frac{\mathrm{rank}(x_{ij}) - 0.5}{n}\right)$$

其中，Φ 是标准正态分布的累积分布函数。变换后的变量记为（z_{i1}，……，z_{id}），其中 $i = 1，2，……，n$，并称其处于 z尺度上。变换后，相关矩阵 $\boldsymbol{R}_{\mathrm{data}}$ 由（z_{i1}，……，z_{id}）向量获得。

3.1 主要步骤概述

本节概述了如何识别各组，并引入潜变量来解释组内依赖关系。所有基因都会与其他基因相互作用，形成更大的蛋白质复合物，就像由小部件构成的机器一样；同一复合物中的基因活动会协同表达，并同时出现在细胞中。生物网络中自然会形成由功能相似的基因（变量）组成的组，而许多实证研究确实证实，完整的组结构既存在于表征出来的基因表达谱中，也存在于潜在的物理相互作用网络中（例如，[2]；[13]）。

为识别各组，可以应用一些变量聚类算法，获得初始的互不重叠组，再借助一些额外的诊断工具进行调整，使每个较小组内的依赖关系具有带有残余依赖的单因子结构。对于由此得到的每个组，一个潜变量可以解释组内变量之间的大部分依赖关系，并据此创建一个代理变量，作为该潜变量的估计值。

该过程由几个步骤组成。

1. 使用变量聚类算法将变量划分为不重叠同质组。假设共有 m 组，记为 G'_1，……，G'_m。

2. 由于每个变量都被分配到某个组，因此需要进行检查，以分离出变量

这些变量在其所属组（或其他组）内的关联并不紧密。此类变量被视为与任何组都孤立。现在得到了一些更小的组 G_1，……，G_m 以及一组孤立变量。

3. 现在，每个组 G_g 中的变量都具有中等至强相关性。对组 G_g的相关矩阵拟合单因子模型，并标记组内可能存在更强残余依赖的变量。G_g 中其余变量呈现出具有弱残余依赖的单因子结构。

4. 对于 G_g 中残余依赖较弱的变量，潜变量 W_g 可以解释大部分依赖关系，而 [18] 中的理论提供了一种计算代理变量 $\widetilde{w}_g$ 的方法，用于估计该潜变量。

5. 将 $\widetilde{w}_1$、……、$\widetilde{w}_m$ 添加到 d 变量中，并使用前文概述的算法构建截断藤依赖图。

为了便于解释，可以重新定向每个组内与组中大多数其他变量负相关的变量。这些变量可能是在拟合的单因子模型中具有负载荷的变量。可以绘制热图，以可视化对某些变量取反前后（在转换后的 z尺度上）的组结构。

3.2 步骤详解

1. 对于高维数据，可以使用变量聚类算法（例如[29] 中的 CLV算法（R 语言库 ClustVarLV））来形成不重叠同质组。CLV算法试图寻找变量的最优划分，以最大化所得各簇内变量同质性度量的总和。当变量与各簇中的潜在成分关联越强时，同质性度量就越大。

2. 根据 CLV算法得到的各组G'_1、……、G'_m 的相关矩阵，可以找出需要分离出来的弱相关变量。

3. 对于经过步骤 2 处理后的组 G_g（$g=1$、……、m），令$\boldsymbol{R}_{g、\text{data}}$为组 G_g 的经验相关矩阵，令$\boldsymbol{R}_{g、\text{1-factor}}$ 为拟合的一因子依赖模型所得的相关矩阵。残余相关矩阵为$\boldsymbol{D}_g = |\boldsymbol{R}_{g、\text{1-factor}} - \boldsymbol{R}_{g、\text{data}}|$。如果变量对应的 $\boldsymbol{D}_g$ 所在行中存在较大的值（超过阈值 1），或者该行的值之和较大（超过阈值 2），则认为该变量具有较强的残余依赖。

4. 假设组 G_g 中有 d_g 个残余依赖较弱的变量：z_{1g}，z_{2g}，……，$z_{d_g,\,g}$。代理变量定义为

$$\widetilde{w}_g = \Phi^{-1}\Big(d_g^{-1}\sum_{j=1}^{d_g} z_{jg}\Big). \tag{5}$$

[18]证明，在残余依赖较弱这一温和假设下，当 d_g 足够大，且变量在单因子模型拟合中已调整方向，使其具有正载荷时，该代理变量可作为潜变量的良好估计。残余依赖较强的变量仍属于组 G_g，但无法保证纳入这些变量后得到的代理变量在理论上是一致的。

5. 在包含代理变量的藤依赖图中，可以预期，组 G_g 中的变量在第一藤树中大多与 $\widetilde{w}_g$ 相连。第一藤树之外的额外边将解释在控制组潜在效应后局部的残余依赖。总的来说，藤依赖图不仅展示了变量之间的依赖关系或条件依赖，还提供了更多关于组结构的信息，并总结了变量间的潜在依赖。

4 生物学应用

本节介绍两个基因应用案例。一个涉及酵母基因数据集，另一个涉及前列腺癌数据集。我们将构建含潜变量的藤依赖图的方法应用于这两个数据集中的部分选定基因，旨在探究所得依赖图是否能够提供有用信息，并与文献中的一些生物学发现相符。

4.1 酵母数据中的MAP激酶通路推断

酵母数据集来自基因表达综合数据库（登录号：GSE1990）。根据该数据集的文档，酵母菌株BY4716与RM11-1a杂交产生的单倍体分离子，如 [5]所示。该系列包含GSE617的所有样本，以及另外27个使用与GSE617相同的方案和参考样本进行检测的分离子，共有262个基因表达向量，涉及约7,000个基因。由于京都基因与基因组百科全书（KEGG）数据库 [16], 中已知丝裂原活化蛋白激酶（MAPK）信号通路的实际调控网络拓扑，我们重点识别已知构成MAPK通路的 $d = 37$ 基因之间的关系。真实基因调控网络见 [17]。

50%的基因缺失值比例低于3%，其余基因的缺失比例在3%至10%之间。对于原始数据集中的缺失值，我们使用高斯二元Copula进行插补。对于存在缺失值的变量，我们找到一个与该变量高度相关的替代变量。提取替代变量和待插补变量中均非缺失的记录，并将其秩变换为服从 $N(0,1)$分布的变量。然后，以替代变量为预测变量、待插补变量为响应变量，拟合线性回归模型。接着，将线性模型对待插补变量在 $N(0,1)$上的预测值通过逆秩变换转换回原始尺度，并填入缺失位置。

我们在图1中展示了一个示意性的真实基因调控网络。在该图中，有几个基因出现在多个位置，因此不能将其视为依赖图。然而，我们的目标是在表达水平上为这些基因构建依赖图，并将结果与该参考图进行比较，以判断是否存在生物学匹配。

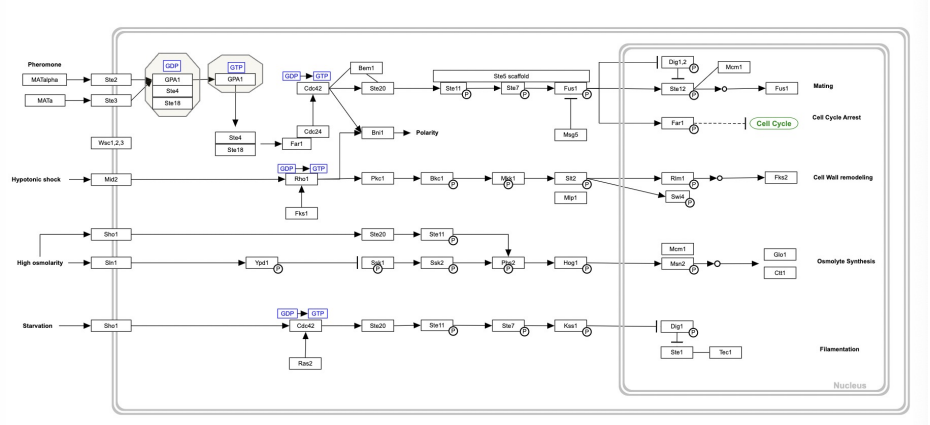

图1：维基百科中的 MAPK信号通路（酿酒酵母）

对缺失值进行插补并转换为标准正态分布后，将所提出的方法论应用于 z尺度数据。结合初始聚类结果与热图的诊断结果，可得到4个较小的同质组。对于组内与其他变量大多呈负相关的变量，我们通过改变其符号来调整方向，使各组内的相关性均为正。

对于每个组，尽管组规模 d_g 并不大，我们仍引入式（5）中的代理变量。这些代理变量分别称为代理1、代理2、代理3和代理4。所得藤依赖图如图3所示。

将藤依赖图与真实情况图进行比较可以发现，与代理1和代理2相连的基因（例如$MAT\text{-}\alpha$、$GPA1$、$DIG1$、$FAR1$）彼此密切相关，并参与细胞周期停滞过程。一组基因，如$SLN1$、$SHO1$、$YPD1$和$CTT1$，与代理4相连，这些基因参与“渗透保护剂合成过程”。另一组基因，$RHO1$、$BKC1$、$SWI4$、$SLT2$和$RLM1$，与代理3相连，并参与“细胞壁建模过程”。这些发现表明，代理变量与观测变量之间的连接可以提供一些分组信息，而聚类到同一组中的基因通常参与相同的细胞过程。

除了分组信息之外，藤依赖图中还有一些有趣的连接与图1中的真实通路图相吻合。例如，在细胞壁重塑过程中，参与该过程的基因在藤图中彼此紧密相连：$SLT2$ 在第一棵树中与 $RLM1$ 相连，并通过$RML1$与 $SWI4$ 相连。这与真实通路中$SLT2 \longrightarrow RLM1$ 和$SWI4$ 之间的路径相吻合。一组基因——$PKC1$、$BCK1$、代理3和$RHO1$ ——在第一棵树1中呈线性排列；这大致对应于真实通路中的路径$RHO1 \longrightarrow PKC1 \longrightarrow BCK1$。此外，与其他基因有许多连接的基因

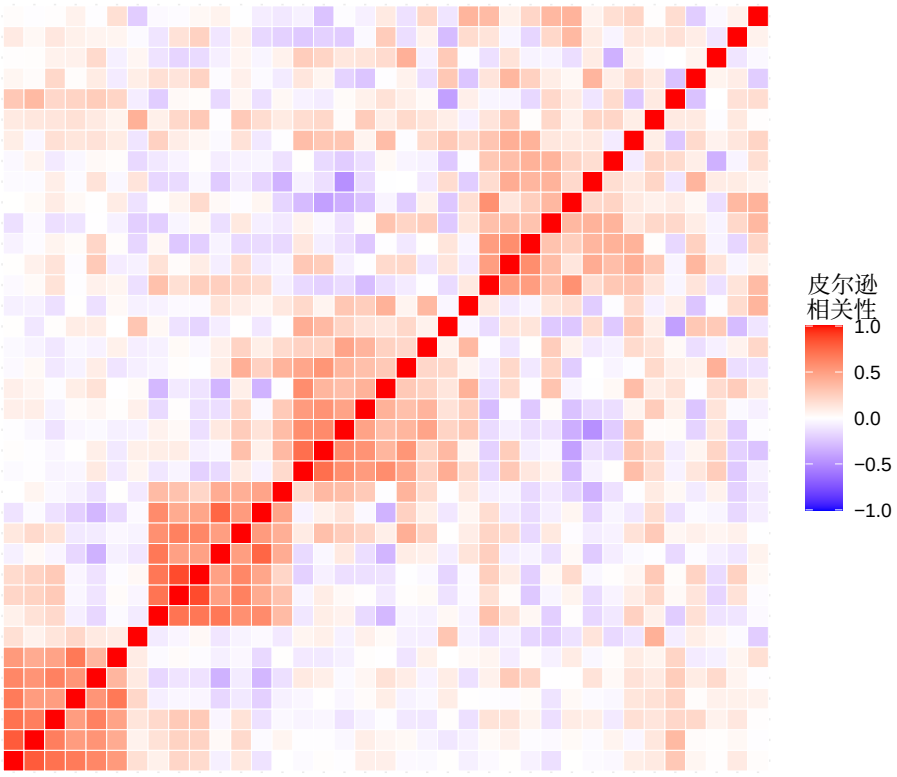

图 2：转换为正态尺度的酵母 RNA 表达数据热图；从左下到右上共有四个同质组，组大小分别为 7、7、9、8。另有 6 个变量未归入这 4 个组，属于孤立变量。

可能充当枢纽基因并参与多个过程。例如，*STE20* 在图1中的多个位置出现，并在多个细胞过程中发挥作用。从藤依赖图来看，它与其他组中的基因之间也有许多距离为两条边以内的连接。

4.2 前列腺癌研究中的条件特异性依赖网络构建

我们收集了癌症基因组图谱（TCGA）队列中 $n = 497$ 名前列腺腺癌患者的 RNA 测序基因表达数据。RNA 测序技术可直接测量比对到各基因（外显子）的短读段数量，该数量与各基因内转录的 mRNA 分子数量呈稳健相关。在此，我们特别关注此前已知参与癌症发育过程（KEGG）的基因，方法与先前的分析 [21] 一致。对于 TCGA 数据中可获得且与 KEGG 癌症通路注释重叠的 314 个入选基因，我们应用了所提出的方法论。

CLV 聚类算法将全部 314 个基因划分为六个同质组，分别包含 55、44、63、62、44 和 46 个变量。为了估计单因子模型，我们从这六个簇中分别识别出 3、6、5、15、8 和 7 个弱相关变量，共计 44 个，

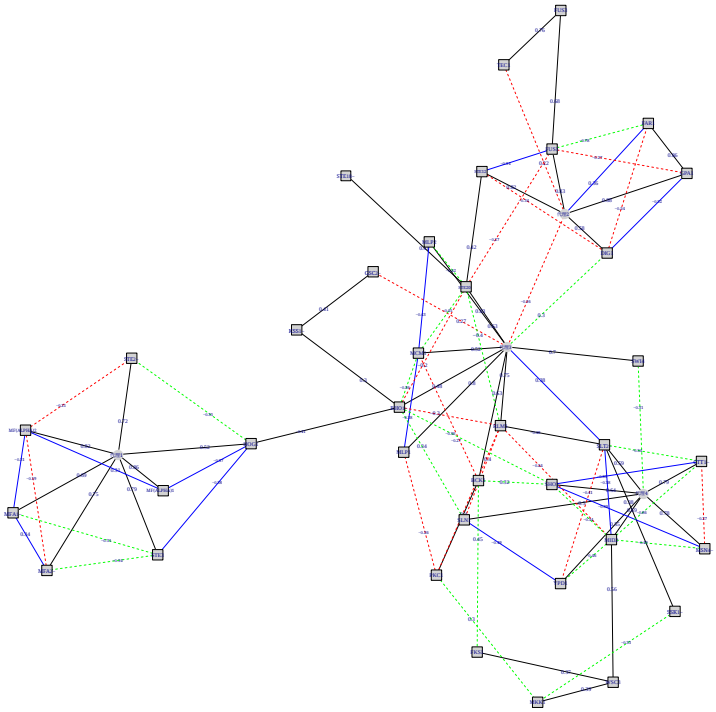

图 3： 酵母 RNA 表达数据的藤依赖图；黑色的边来自最大生成树，蓝色、红色和绿色的相连边分别说明给定一个、两个以及三个或更多变量时变量之间的条件依赖。若给定一个、两个以及三个或更多变量时两个变量的偏相关分别大于 0.25、0.15 和 0.30，则在图中绘出相应的边；若其数值分别大于 0.5、0.3 和 0.4，则标注这些边。黑色生成树上的标签表示两个变量之间的相关性，而其他边上的标签表示给定其他变量时变量之间的偏相关。图中引入了四个代理变量（以圆圈表示），观测变量则以矩形表示。

并将它们从各簇中分离出来，从而得到由 52、38、58、47、36 和 39 个变量组成的紧密连接组；每个组主要由一个潜在因子解释。

为便于说明（使图形不那么繁杂），我们展示由索引为1、6、3且组间依赖关系较强的CLV组构建的连通子图。第1组、第6组和第3组用于代理计算的变量数量分别为52、39和45。所有6组的藤依赖图见附加文件1。

根据相关矩阵，某些基因的RNA测序值与其他基因的RNA测序值大多呈负相关。对于这些基因，我们将其z尺度RNA测序值的符号取反；等价地说，就是将原始RNA测序值取反。在热图和藤图中，基因名称后会添加一个 – (减号)。

图4中，164个基因分为三个同质组。我们将对角线上从左下方开始的组依次标记为第1组、第6组和第3组，此外还有15个与这些CLV组分离的基因。

对于这三个组，分别对每个组进行富集分析。使用R软件包goseq识别这三个组的本体，并确定排名前三的潜在本体

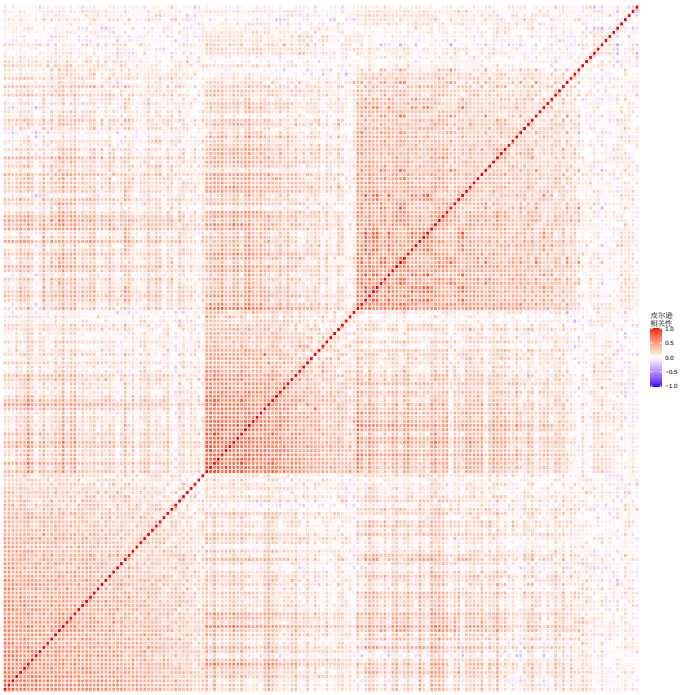

图4：基于前列腺癌肿瘤样本计算的164个基因之间的基因表达相关矩阵。颜色及其深浅表示基于归一化基因表达 z 分数计算的皮尔逊相关值（−1 和 1）。从左下方到右侧，第一个区块包含第1组的52个基因，第二个区块包含第6组的39个基因，第三个区块包含第3组的58个基因。第三组之后，还有15个基因被排除在 CLV 组之外。

各组的数据见表2。

组	本体论
1	肽类激素分泌的正调控 药物代谢过程 前体代谢物和能量的生成
6	含胶原的细胞外基质 血管生成 骨骼系统发育
3	肌肉系统过程调节 核腔 核质

表2：CLV组1、6、3的前三个本体。

藤依赖图

对于肿瘤病例，在未引入和引入代理变量的情况下分别获得了藤依赖图；它们分别见图6和图7。

对式（2）中的偏相关元素也进行了简单阈值化，以获得条件独立图。偏相关的汇总统计量如下：最小值：-0.34，第一四分位数：-0.04，中位数：0.00，第三四分位数：0.05，最大值：0.77。大多数偏相关的绝对值较小，且接近零。节点度

分布如图5所示，其中展示了截断藤和基于不同阈值（分别为0.1和0.15）的条件独立图所推断网络的节点度分布。考察节点度最大的基因及其邻居后，所得子集与基于我们方法论得到的基因组无关。 [21] 包括多种基于逆协方差矩阵生成图的方法；然而，我们无法将这些方法的调节参数与偏相关阈值对应起来，因此无法进行比较。

根据藤依赖图，我们找出了连接最紧密的基因，并在表3中汇总了结果。表中还列出了这些基因的效应量（根据原始 RNA测序计数的对数值计算），以及使用 R 软件包 edgeR 进行差异表达基因分析（DEG）得到的 logFC 值（倍数变化的对数）。倍数变化描述了正常样本与肿瘤样本之间基因 RNA表达的差异。与正常样本相比，倍数变化为正表示基因上调，为负则表示基因下调。此外，当 p值截断值设为 0.05 时，表中的所有基因均通过了 edgeR 中用于判定差异表达的似然比检验。

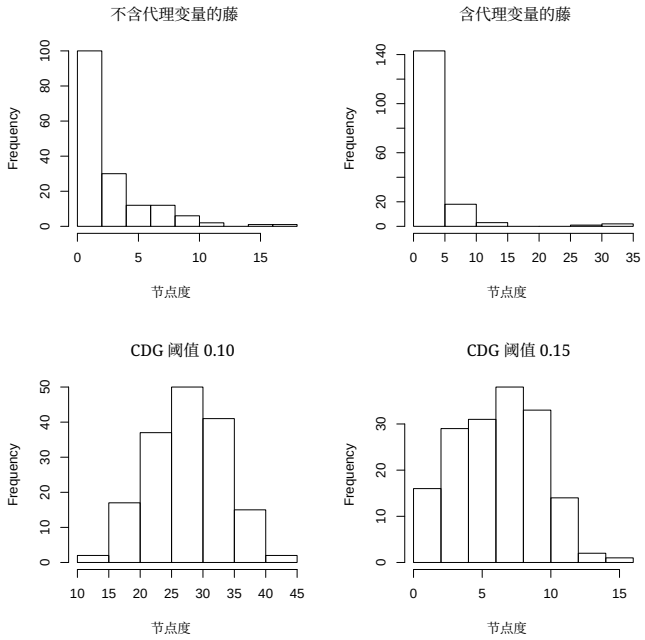

图5：所考察方法推断网络的节点度分布。

在这两个藤依赖图中，孤立变量通常具有 1 的节点度，并且远离围绕代理变量连接的簇。用于计算代理变量的基因通常直接连接到相应的基于组的代理变量，而其中一些基因则通过次级树，经由该组中的另一个基因连接到代理变量。

枢纽基因通常与组内的代理变量具有很强的相关性，同时也与组内其他变量具有相对较强的相关性。示例包括基因$BIRC5$、$GSTP1$、$TCF7L1$和$CREBBP$；它们与代理变量相连，并以星形结构连接到其他变量。某些枢纽基因，例如$FGF2$，会在一个

表3： 通过差异表达基因检验（使用软件包“edgeR”进行似然比检验；p 值阈值为 0.05）且在构建的藤依赖图中具有大量连接的基因。该图展示了基于肿瘤样本构建的藤图中的枢纽基因；左侧面板对应未使用代理变量构建的藤图，右侧面板对应引入代理变量后构建的藤图。效应量按对数尺度计算：$(\log\ y_{\text{tumor}} - \log\ y_{\text{normal}})/s_{\text{pooled}}$，$y$ 为原始RNA测序数据，logFC表示倍数变化的对数变换值，倍数变化定义为两组中“表达水平”的比值。代理1、proxy6、代理3分别对应第1、6、3组。对于名称以 – 结尾的基因，其RNA测序值经过重新定向，以便绘制热图和藤图。

基因（节点度）	效应量	对数倍数变化	基因（节点度）[组]	效应量	对数倍数变化
TCF7L1(18)	-1.81	-0.86	代理3(33)	-	-
CREBBP(16)	-0.27	0.23	代理1(32)	-	-
RBX1-(11)	0.43	0.63	proxy6(28)	-	-
LAMA4(10)	-0.86	-0.45	COL4A2 (14) [B]	-0.78	-0.51
BIRC5-(10)	1.74	2.33	GSTP1(12) [C]	-2.04	-1.87
ITGB1(9)	-1.28	-0.51	BIRC5-(11) [C]	1.74	2.33
FGFR1(9)	-1.10	-0.54	FGF2(9) [B]	-1.19	-0.61
FZD7(9)	-1.21	-0.64	TCF7L1(9) [C]	-1.81	-0.86
GSTP1(9)	-2.04	-1.87	CREBBP(9) [A]	-0.27	0.23
COL4A2(8)	-0.78	-0.51	LAMB3(9) [C]	-1.28	-1.42
FGF2(8)	-1.19	-0.61	E2F1-(8) [C]	0.90	1.36
EP300(7)	-0.25	0.26	MMP2(7) [B]	-0.73	-0.26
LAMB3(8)	-1.28	-1.42	CKS2-(7) [C]	0.85	1.11
TGFBR2(8)	-1.05	-0.46	CBL(7) [A]	-0.42	0.15
ITGA3(8)	-1.45	-0.67	-	-	-
HRAS-(7)	0.30	0.52	-	-	-
GTSE1-(7)	1.70	2.02	-	-	
RRM2-(7)	1.79	1.94	-	-	-
E2F1-(7)	0.90	1.36	-	-	-
CHEK2-(7)	1.06	1.04	-	-	-

组中连接到代理变量，同时也连接到另一个组中的基因，充当两个组之间的桥梁。另一个桥接示例是基因$COL4A2$；它连接到第6组的代理变量以及第1组的基因。$LAMB3$ 是来自第3组的基因，但由于其表现出很强的残余依赖，因此未用于计算代理变量。图中显示，该基因与多个组相连。与第3组基因的连接解释了考虑组效应后残留的局部依赖，而通过基因$TCF7L1$ 与第6组代理变量的连接则解释了组间依赖（以及非弱残差依赖）。

根据表3，有几个枢纽基因（基因名称后带有“–”）由于其RNA测序值与许多其他基因的RNA测序值呈负相关而被重新定向；示例包括$RBX1$、$BIRC5$和$E2F1$。这些基因的效应量为正，表明相较于未患病受试者，它们在前列腺癌患者中上调。未重新定向的RNA测序变量的效应量主要为负，表明相应基因在前列腺癌患者中下调。

结合两种藤依赖图与表3，我们认为，连接度最高且效应量相对较大的基因可能在前列腺癌的发展过程中发挥重要作用。肿瘤患者中的枢纽基因候选名单包括：带有代理变量的藤图中的<COL4A2>、<GSTP1>、<BIRC5>、<FGF2>、<TCF7L1>、<CREBBP>、<LAMB3>、<E2F1>，以及不带代理变量的藤图中连接度最高且效应量较大的<RBX1>、<LAMA4>、<ITGB1>、<FGFR1>、<FZD7>、<EP300>、<TGFBR2 >。

对于这些潜在的枢纽基因，其中一些有相关的生物学证据支持。

- *TCF7L1*: [31] 表明，WNT4/TCF7L1的诱导会导致前列腺癌中神经内分泌分化（NED）和恶性程度增加，这与雄激素受体信号通路失调及IL-8/CXCR2通路激活有关。

- *BIRC5*: *BIRC5* 是一种免疫相关基因，可抑制细胞凋亡并促进细胞增殖。它在大多数肿瘤中高表达，并导致癌症患者预后不良。 [33] 显示，BIRC5与多种肿瘤中的多种免疫细胞浸润显著相关。

- *GSTP1*: [25] 指出，该基因属于谷胱甘肽S-转移酶家族，这是一组参与外源性物质解毒的酶；它在细胞周期调控中也发挥重要作用。其失调与多种肿瘤相关，尤其与前列腺癌相关。

- *FGF2*: 根据 [12] 和 [27], ，成纤维细胞生长因子2（FGF2，又称碱性成纤维细胞生长因子）在人类前列腺癌中的表达水平升高。FGF2可促进细胞运动和增殖、增强肿瘤血管生成并抑制细胞凋亡，而这些作用在肿瘤进展中都发挥着重要作用。最近， [22] 还将FGF2基因确定为前列腺癌发生发展中的枢纽基因。

根据所有生物学证据，我们从小型子集中（164 个基因）识别出的枢纽基因在前列腺癌的进展中发挥着重要作用；这表明藤依赖图能够提供一些有趣且有用的见解。

5 讨论

本文提出了一种利用带潜变量的截断藤构建依赖图的方法。该方法使用低阶偏相关，并避免了条件依赖图中使用高阶偏相关所带来的问题。低阶偏相关更容易解释变量之间的条件依赖。借助能够解释变量组中大部分依赖关系的潜变量，随着与每个潜变量相连的变量数量增加，高阶偏相关的绝对值会逐渐趋近于 0。

本文探讨如何从数据中寻找一种特殊形式的依赖结构，同时将候选图限制为藤图。自微阵列技术问世以来，从观测数据中寻找统计依赖网络一直是计算生物学家十分感兴趣的问题。然而，在一般图（如有向无环图（DAG））的组合空间中进行结构学习，使得精确的潜在结构推断极难实现，因此往往需要采用贪心优化算法 [11]。即使能够枚举所有这些结构，鉴于表达数据的样本量有限，评分函数（目标函数）也常常面临一个关键问题：难以区分多个可能性相同或

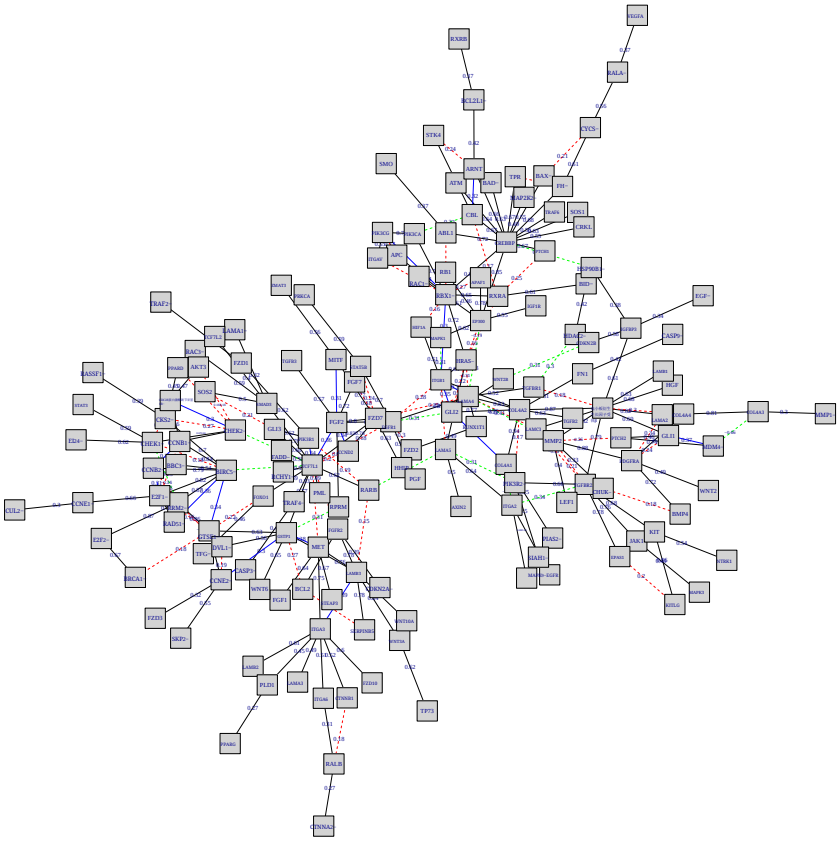

图6： 前列腺癌 RNA测序数据中 164 个基因的依赖图。黑色边来自最大生成树，蓝色、红色和绿色相连边分别表示给定一个、两个以及三个或更多变量时变量之间的条件依赖。如果给定一个、两个以及三个或更多变量时，两个变量的绝对偏相关分别大于 0.30、0.15 和 0.30，则在图中绘制相应的边；如果其绝对偏相关分别大于 0.5、0.3 和 0.4，则标注这些边。黑色生成树边上的标签表示两个变量之间的相关性，而其他边上的标签表示给定其他变量时变量之间的偏相关。未引入代理变量，观测变量以矩形表示。

至少相近的结构 [6]。不过，使用相对较小的数据集也能获得可解释的截断藤结构。

统计依赖结构通常并不对应物理上的蛋白质-蛋白质相互作用网络，因此在进行生物学解释时应谨慎。相反，依赖图通常体现的是功能模块的概念，而枢纽节点（高度数顶点）则意味着这些必需基因参与了多个生物过程。对高维基因表达数据中的统计依赖结构进行建模，也有助于后续分析。例如，借助藤图（条件独立结构）计算条件概率，将有助于改善基于观测到的基因表达来预测疾病结局的回归和分类问题 [[10], [8]]。我们还可以预期，由于模型在构造上具有稀疏性，我们的方法能够高效地计算大量基因的边缘概率。这对于因果推断中的 G公式计算尤为有吸引力，

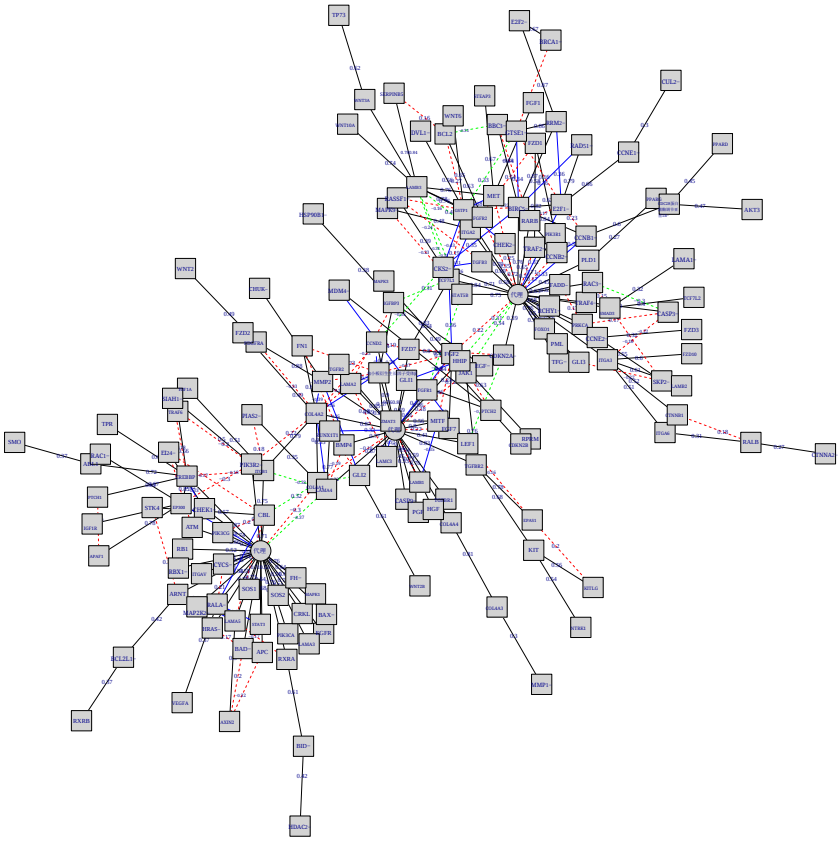

图7： RNA测序数据中含有三个代理变量的前列腺癌164个基因的依赖图。黑色的边来自最大生成树，蓝色、红色和绿色的相连边分别解释了给定一个、两个以及三个或更多变量时变量的条件依赖关系。当给定一个、两个以及三个或更多变量时，两个变量的偏相关绝对值分别大于0.30、0.15和0.30，图中才会绘制相应的边；当其分别大于0.5、0.3和0.4时，边才会标注数值。黑色生成树上的标签表示两个变量之间的相关性，而附加边上的标签表示给定其他变量时变量的偏相关。图中未引入代理变量，观测变量以矩形表示。

计算 [26]。

参考文献

[1] T. W. Anderson。《多元统计分析导论》。Wiley，纽约，1958年。

[2] A.-L. Barabasi 和 Z. N. Oltvai。网络生物学：理解细胞的功能组织结构。《自然综述：遗传学》，5(2)：101–113，2004。

[3] T. Bedford 和 R. M.Cooke。藤——一种用于相关随机变量的新型图模型。统计年刊，30(4):1031–1068，2002。

[4] E. C. Brechmann 和 H. Joe。使用拟合指标截断藤 copula。《多元分析杂志》，138:19–33，2015。[5] R. B. Brem、G. Yvert、R. Clinton 和 L. Kruglyak。出芽酵母转录调控的遗传解析。《科学》，296(5568):752–755，2002。[6] D. M. Chickering。学习贝叶斯网络结构的等价类。《机器学习研究杂志》，2:445–498，2002。[7] C. Chow 和 C. Liu。用依赖树近似离散概率分布。《*IEEE*信息论汇刊》，14(3):462–467，1968。[8] H.-Y. Chuang、E. Lee、Y.-T. Liu、D. Lee 和 T. Ideker。基于网络的乳腺癌转移分类。《分子系统生物学》，3(1):140，2007。[9] X. Fan 和 H. Joe。通过估计潜变量构建高维因子 copula 模型。*arXiv* 预印本 *arXiv:2205.14487*，2022。[10] N. Friedman、D. Geiger 和 M. Goldszmidt。贝叶斯网络分类器。《机器学习》，29(2):131–163，1997。[11] N. Friedman、M. Linial、I. Nachman 和 D. Pe'er。使用贝叶斯网络分析表达数据。《计算生物学杂志》，7(3-4):601–620，2000。[12] D. Giri、F. Ropiquet 和 M. Ittmann。人类前列腺癌中基础成纤维细胞生长因子（FGF）2及其受体 FGFR-1 表达的变化。《临床癌症研究》，5(5):1063–1071，1999。[13] R. Jansen、D. Greenbaum 和 M. Gerstein。将全基因组表达数据与蛋白质-蛋白质相互作用关联起来。《基因组研究》，12(1):37–46，2002。[14] H. Joe。使用 *Copula* 的依赖建模。Chapman & Hall/CRC，博卡拉顿，佛罗里达州，2014。[15] H. Joe。由藤构建的简约图形依赖模型。《加拿大统计学杂志》，46(4):532–555，2018。[16] M. Kanehisa、S. Goto、M. Furumichi、M. Tanabe 和 M. Hirakawa。KEGG：用于表示和分析涉及疾病和药物的分子网络。《核酸研究》，38(suppl 1):D355–D360，2010。[17] T. Kelder、M. P. Van Iersel、K. Hanspers、M. Kutmon、B. R. Conklin、C. T. Evelo 和 A. R. Pico。WikiPathways：围绕生物通路构建研究社群。《核酸研究》，40(D1):D1301–D1307，2012。[18] P. Krupskii 和 H. Joe。使用代理变量对高维因子 copula 模型中的参数进行估计的近似似然。《统计论文》，63:543–569，2022。[19] D. Kurowicka 和 R. Cooke。用偏相关藤参数化正定矩阵。《线性代数及其应用》，372:225–251，2003。

[20] D. Kurowicka 和 R. M. Cooke。《不确定性分析与高维依赖建模》。Wiley，奇切斯特，2006 年。[21] L. Lin、M. Drton 和 A. Shojaie。利用正则化得分匹配估计高维图模型。《电子统计学杂志》，10(1):806，2016 年。[22] S. Liu、W. Wang、Y. Zhao、K. Liang 和 Y. Huang。通过整合分析基因表达谱和癌症基因组图谱，鉴定前列腺癌发病机制和预后相关的潜在关键基因。《肿瘤学前沿》，10:809，2020 年。[23] E. Z. Macosko、A. Basu、R. Satija、J. Nemesh、K. Shekhar、M. Goldman、I. Tirosh、A. R. Bialas、N. Kamitaki、E. M. Martersteck 等。利用纳升级液滴对单细胞进行高度并行的全基因组表达谱分析。《细胞》，161(5):1202–1214，2015 年。[24] P. M. Magwene 和 J. Kim。利用一阶条件独立估计基因组共表达网络。《基因组生物学》，5(12):1–16，2004 年。[25] F. Martignano、G. Gurioli、S. Salvi、D. Calistri、M. Costantini、R. Gunelli、U. De Giorgi、F. Foca 和 V. Casadio。前列腺癌中的GSTP1甲基化和蛋白质表达：诊断意义。《疾病标志物》，2016 年。[26] A. I. Naimi、S. R. Cole 和 E. H. Kennedy。g方法简介。《国际流行病学杂志》，46(2):756–762，2017 年。[27] N. Polnaszek、B. Kwabi-Addo、L. E. Peterson、M. Ozen、N. M. Greenberg、S. Ortega、C. Basilico 和 M. Ittmann。成纤维细胞生长因子2在前列腺癌自发性小鼠模型中促进肿瘤进展。《癌症研究》，63(18):5754–5760，2003 年。[28] M. Schena、D. Shalon、R. W. Davis 和 P. O. Brown。利用互补DNA微阵列定量监测基因表达模式。《科学》，270(5235):467–470，1995 年。[29] E. Vigneau 和 E. Qannari。围绕潜在成分的变量聚类。《统计学通讯：模拟与计算》，32(4):1131–1150，2003 年。[30] Z. Wang、M. Gerstein 和 M. Snyder。RNA测序：转录组学的革命性工具。《自然综述：遗传学》，10(1):57–63，2009 年。[31] Y.-C. Wen、Y.-N. Liu、H.-L. Yeh、W.-H. Chen、K.-C. Jiang、S.-R. Lin、J. Huang、M. Hsiao 和 W.-Y. Chen。TCF7L1调控前列腺癌的细胞因子反应和神经内分泌分化。《肿瘤发生》，10(11):1–11，2021 年。[32] J. Whittaker。《应用多元统计中的图模型》。Wiley，奇切斯特，1990 年。[33] L. Xu、W. Yu、H. Xiao 和 K. Lin。BIRC5是与肿瘤免疫细胞浸润相关的预后生物标志物。《科学报告》，11(1):1–13，2021 年。