Statistics and Application 统计学与应用, 2019, 8(5), 777-783
Published Online October 2019 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2019.85088

Hans汉斯

Large Precision Matrix Estimation for Compositional Data

Xuanxuan Zhang, Fengxia He

School of Mathematics and Physics, North China Electric Power University, Beijing
Email: zxx_9407@sina.com

Received: Oct. 2nd, 2019; accepted: Oct. 18th, 2019; published: Oct. 25th, 2019

Abstract

High-dimensional compositional data arise in many applications, and statistical methods often fail to produce sensible results due to the unit-sum constraints. The estimation of high dimensional covariance matrix or precision (inverse covariance) matrix is the basic problem of modern multivariate analysis. In this paper, the precision matrix estimation problem for high-dimensional compositional data is considered. It is known that the inverse of the sample covariance matrix is unstable for the estimate precision matrix. Since the sample size of the data is smaller than the number of variables, the inverse of the high-dimensional data matrix is difficult to estimate. In this paper, we use the centered log-ratio transformation method to process high-dimensional compositional data, and then solve the singularity problem of covariance matrix, and obtain the precision matrix estimation of high-dimensional compositional data. Simulation experiments and actual data can verify the rationality of the proposed method.

Keywords

Compositional Data, High-Dimensional Data, Centered Log-Ratio Transformation, Precision Matrix

高维成分数据的精度矩阵估计

张轩轩，何凤霞

华北电力大学数理学院，北京
Email: zxx_9407@sina.com

收稿日期：2019年10月2日；录用日期：2019年10月18日；发布日期：2019年10月25日

摘　要

高维成分数据在许多应用中均有出现，由于定和约束，统计方法通常不能产生合理的结果。高维协方差矩阵和精度(逆协方差)矩阵的估计是现代多元分析的基本问题，本文考虑高维成分数据的精度矩阵估计

文章引用: 张轩轩, 何凤霞. 高维成分数据的精度矩阵估计[J]. 统计学与应用, 2019, 8(5): 777-783.
DOI: 10.12677/sa.2019.85088

问题。已知样本协方差矩阵求逆对于估计精度矩阵是不稳定的，由于数据的样本量小于变量个数，高维数据矩阵的逆很难估计。本文利用中心对数比变换方法，处理高维成分数据，然后解决协方差矩阵奇异性问题，得到高维成分数据的精度矩阵估计。模拟实验和实际数据可以验证提出方法的合理性。

关键词

成分数据，高维数据，中心对数比变换，精度矩阵

1. 引言

成分数据经常出现在地质、微生物、经济等许多领域，例如岩石的化学成分，家庭支出模式，生物群落的物种组成等等，它主要是用来研究构成某个整体的各部分的比重关系。1866 年 Ferrers [1]首先提出了成分数据的概念。Aitchison [2]指出研究成分数据应该关注这些成分间的比例关系，而不应该过多地关注每一个分量的具体取值是多少。之后 Aitchison [3] [4]提出了一种新的处理成分数据的方法“对数比”，使得传统的统计方法也能够对变换后的数据进行相应的统计分析。除了对数比变换，还有等距对数比变换[5]，球坐标变换[6]等。

估计协方差矩阵是多变量分析的基础。非成分高维数据的协方差和精度(协方差的逆)矩阵估计的方法已逐渐成熟。对于协方差估计，Bickel 和 Levina [7]提出硬阈值方法研究高维协方差的估计。Rothman、Levina 和 Zhu [8]研究了一类更普通的阈值方法。Cai 和 Liu [9]提出了自适应阈值方法，该方法可以根据数据选择阈值，方法的灵活度更高。对于精度矩阵估计，Friedman、Hastie 和 Tibshirani [10]提出了一种有效的算法(Graphical lasso)估计高维精度矩阵。Cai、Liu 和 Luo [11]在矩阵求逆约束下寻找稀疏精度矩阵。除此之外，我们还参考了 Liu 和 Luo [12]和 Fan、Liao 和 Liu [13]等文章。上面的文章都是关于非成分数据的，对于成分数据的研究比较少。Cao、Lin 和 Li [14]提出自适应阈值方法估计协方差矩阵。但这种方法只研究了协方差矩阵，没有研究精度矩阵估计。我们将提出一种方法研究精度矩阵。我们首先将数据进行中心对数比变换，得到成分数据的协方差矩阵，然后借鉴稀疏列式逆算子方法得到高维成分数据精度矩阵。本文具体安排如下。第一部分，给出了基本符号和定义，并且介绍高维成分数据精度矩阵估计方法。第二部分和第三部分进行模拟实验和实际数据分析。

2. 方法

2.1. 记号

我们首先介绍一些符号。在本文中，对于向量 $a=\left(a_1,\cdots,a_p\right)^{\mathrm{T}}\in R^p$，定义 $|a|_1=\sum_{j=1}^{p}|a_j|$，$|a|_2=\sqrt{\sum_{j=1}^{p}a_j^2}$。对于矩阵 $A=\left(a_{ij}\right)\in R^{p\times q}$，定义矩阵谱范数 $\|A\|_2=\sup_{|x|_2\le 1}|Ax|_2$，矩阵 Frobenius 范数 $\|A\|_F=\sqrt{\sum_{i,j}a_{ij}^2}$，矩阵无穷范数 $\|A\|_\infty=\max_{1\le i\le p}\sum_{j=1}^{q}|a_{ij}|$。$\lambda_{\max}(\cdot)$表示最大特征值，$\lambda_{\min}(\cdot)$表示最小特征值，$A$ 的转置用 A^{T} 表示，$I\{\cdot\}$是指示函数。

2.2. 方法

令 $S^{p-1}=\left\{X=\left(X_1,X_2,\cdots,X_p\right)^{\mathrm{T}};X_i>0,i=1,2,\cdots,p;\sum_{i=1}^{p}X_i=1\right\}$，其中 $X=\left(X_1,\cdots,X_p\right)^{\mathrm{T}}$ 是 p 维成分数据，S^{p-1} 为 $p-1$ 维单形空间。对成分数据做对数比变换，把成分单形空间映射到欧几里得空间中，从而使经典的统计方法可以适用于变换后的数据。在此我们采用中心对数比变换：

$$clr\left(X\right)=\left(\log\frac{X_1}{g\left(X\right)},\cdots,\log\frac{X_p}{g\left(X\right)}\right) \tag{1}$$

$g\left(X\right)=\left(\prod_{i=1}^{p}X_i\right)^{1/p}$ 是 X 的几何均值。

定义中心对数比协方差矩阵 $\Gamma=\left(\gamma_{jk}\right)_{p\times p}$

$$S_i=clr\left(X_i\right)$$

$$\gamma_{jk}=\mathrm{cov}\left(S_j,S_k\right) \tag{2}$$

从而，我们得到中心对数比协方差矩阵 Γ。

对于成分数据协方差矩阵 Σ，Cao、Lin 和 Li [14]建议用中心对数比协方差矩阵 Γ 代替协方差矩阵 Σ，这极大地促进了新方法和相关理论的发展。

如果 Σ 是非奇异的，对于精度矩阵 Ω，我们有 $\Sigma\Omega=\mathrm{E}$。令 $\beta_i=\Omega e_i$，β_i 表示精度矩阵的第 i 列，e_i 是单位矩阵的第 i 列，则 $\Sigma\beta_i=e_i$，即 $\Sigma\beta_i-e_i=0$。为了解决高维协方差矩阵奇异性问题，即当样本量小于变量个数，样本协方差矩阵不是满秩的，对于高维成分数据的样本中心对数比协方差矩阵 $\hat{\Gamma}$，此时我们考虑列损失函数

$$f_i\left(\hat{\Gamma},B\right)=\frac{1}{2}\beta_i^{\mathrm{T}}\hat{\Gamma}\beta_i-\beta_i^{\mathrm{T}}e_i \tag{3}$$

其中 $B=\left(\beta_1,\beta_2,\cdots,\beta_p\right)$。

$\hat{\Gamma}$ 是正定的，则 f_i 是凸函数，当损失函数趋于 0 时，损失函数越小，B 越趋近 Ω。

为了解决精度矩阵的稀疏性，我们加 ℓ_1 惩罚到列损失函数，使 β_i 中的一些坐标尽可能为 0，即

$$\frac{1}{2}\beta^{\mathrm{T}}\hat{\Gamma}\beta-e_i^{\mathrm{T}}\beta+\lambda_{ni}\left|\beta\right|_1 \tag{4}$$

令 $\hat{\beta}_i$ 是下式的解

$$\hat{\beta}_i=\arg\min_{\beta\in R^p}\left\{\frac{1}{2}\beta^{\mathrm{T}}\hat{\Gamma}\beta-e_i^{\mathrm{T}}\beta+\lambda_{ni}\left|\beta\right|_1\right\} \tag{5}$$

则 $\hat{B}=\left(\hat{\beta}_1,\cdots,\hat{\beta}_p\right)$，其中 $\hat{\beta}_i=\left(\hat{\beta}_{i1},\cdots,\hat{\beta}_{ip}\right)^{\mathrm{T}}$

其中 e_i 是单位矩阵的第 i 列，$\lambda_{ni}>0$ 是调整参数，参见文献[12]。

对于 λ_{ni} 的选择，基于 H 折交叉验证，我们定义 $\hat{\beta}_i^{-v}\left(\lambda\right)$ 通过除第 v 折外的样本得到，$\hat{\Gamma}^v$ 为第 v 折样本协方差矩阵，$v=1,\cdots,H$，其中 H 是一个固定的整数。$\lambda_0<\lambda_1<\cdots<\lambda_N$ 划分区间 $\left[0,4\right]$，其中 $\lambda_j=\frac{4i}{N}$。

$$\hat{\lambda}_i=\arg\min_{0\le j\le N}\left\{\frac{1}{H}\sum_{v=1}^{H}\left[\frac{1}{2}\left(\hat{\beta}_i^{-v}\left(\lambda_j\right)\right)^{\mathrm{T}}\hat{\Gamma}^v\hat{\beta}_i^{-v}\left(\lambda_j\right)-e_i^{\mathrm{T}}\hat{\beta}_i^{-v}\left(\lambda_j\right)\right]\right\} \tag{6}$$

使用最优 $\hat{\lambda}_i$，然后基于完整数据集计算得到最终估计。

综上，我们可以得到高维成分数据的精度矩阵估计

$\hat{\Omega}=\left(\hat{\omega}_{ij}\right)_{p\times p}$ 其中 $\hat{\omega}_{ij}=\hat{\omega}_{ji}=\hat{\beta}_{ij}I\left\{\left|\hat{\beta}_{ij}\right|<\left|\hat{\beta}_{ji}\right|\right\}+\hat{\beta}_{ji}I\left\{\left|\hat{\beta}_{ij}\right|\geq\left|\hat{\beta}_{ji}\right|\right\}$。

3. 数值模拟

我们按照 Cao、Lin 和 Li [14]生成数据 $\left(W_k,X_k\right),k=1,\cdots,n$。以下面 2 种方式得到 Y_k：

方式 1：Y_k 独立于多元正态分布 $N_p\left(\mu,\Sigma_0\right)$

方式 2：$Y_k=\mu+FU_k/\sqrt{10}$，其中 $FF^{\mathrm{T}}=\Sigma_0$，$U_k\sim\Gamma(10,1)$，矩阵 F 可以通过奇异值分解得到（$\Sigma_0=QSQ^{\mathrm{T}}$，$F=QS^{1/2}$）。这 2 种方式中，我们从 $[0,10]$ 上的均匀分布中随机取出 μ 的分量。

$$W_{kj}=\mathrm{e}^{Y_{kj}},\quad X_{kj}=W_{kj}\Big/\sum_{i=1}^{p}W_{ki},j=1,\cdots,p \tag{7}$$

这样我们就得到了 $W_k=\left(W_{k1},\cdots,W_{kp}\right)^{\mathrm{T}}$，$X_k=\left(X_{k1},\cdots,X_{kp}\right)^{\mathrm{T}}$。

对于 Σ_0，我们通过以下 2 个模型得到：

模型 1：$\Sigma_0=I_{p\times p}$；

模型 2：$\Sigma_0=diag\left(A_1,A_2\right)$，其中 $A_1=B+\varepsilon I_{p_1\times p_1}$，$A_2=4I_{p_2\times p_2}$，$p_1=\left\lfloor 2\sqrt{p}\right\rfloor$，$p_2=p-p_1$。

我们取 $\varepsilon=\max\left(-\lambda_{\min}\left(B\right),0\right)+0.01$，而 B 是一个对称矩阵，它的下三角元素独立于值为 $[-1,-0.5]\cup[0.5,1]$ 的概率为 0.2，值为 0 的概率为 0.8 的均匀分布。

在 2 种模型下，我们比较了不同变换 $clr\left(X\right)$，样本 Y，$\log X$ 和通过其变换得到的精度矩阵估计量 $\hat{\Omega}$，$\hat{\Omega}_Y$，$\hat{\Omega}_l$ 的优劣。其中 $\hat{\Omega}_Y$ 是理想估计量，而 $\hat{\Omega}_l$ 忽略成分数据的独特特征，表现不佳。我们考虑以不同变换下的样本相关性，谱范数和 Frobenius 范数等性能指标来比较方法的优劣。

对于模型 1，我们考虑样本相关性。取样本大小 $n=100$，样本维数 $p=50$，$p=100$ 和 $p=200$ 时进行模拟。

从图 1 和图 2 可以看出，$clr\left(X\right)$ 样本相关关系以 0 为中心，与 Y 的样本相关关系分布类似，并且相似度随着维数 p 的增大而增大。而 $\log X$ 上观察到了伪相关现象，$\log X$ 的样本相关性出现了向上的偏移。可见中心对数比变换（clr）处理高维成分数据具有优越性。

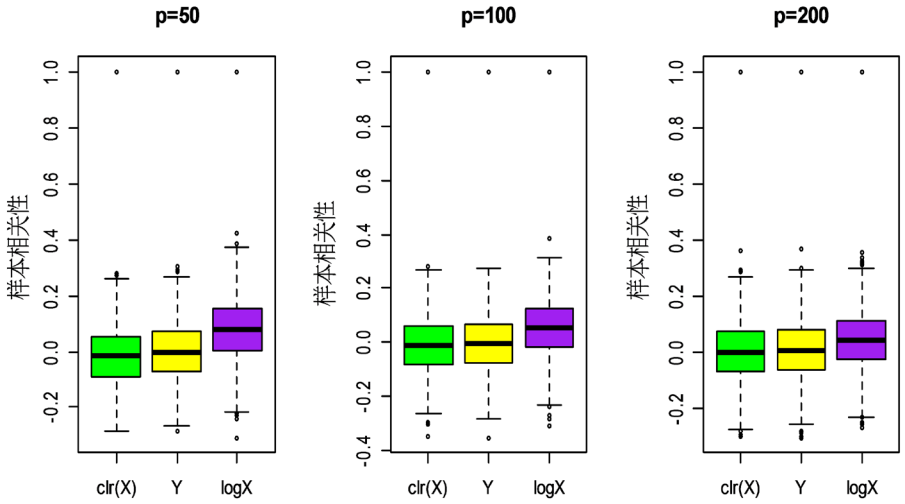

Figure 1. Boxplots of sample correlation under different transformations in mode 1

图 1. 由方式 1 得到的不同变换下的样本相关性箱型图

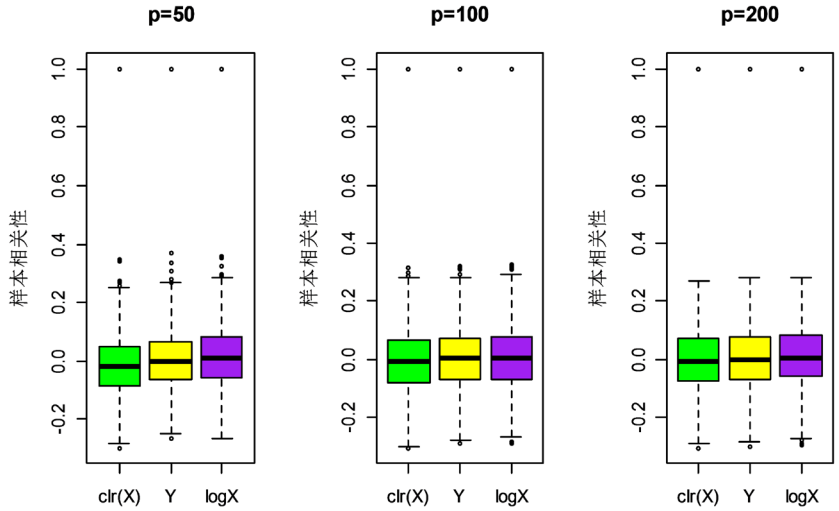

Figure 2. Boxplots of sample correlation under different transformations in mode 2
图 2. 由方式 2 得到的不同变换下的样本相关性箱型图

对模型 2 下的精度矩阵的性能进行研究，我们取样本大小 $n=100$，样本维数 $p=50$， $p=100$ 和 $p=200$，进行 500 次模拟，然后给出模拟结果均值。

Table 1. The precision matrix performance index under different transformations obtained in mode 1
表 1. 由方式 1 得到的不同变换下精度矩阵性能指标

p	$\hat{\Omega}$	$\hat{\Omega}_y$	$\hat{\Omega}_l$
谱范数			
50	0.7464	0.7440	3.0591
100	0.5569	0.5550	2.4696
200	0.4554	0.4542	2.1795
Frobenius 范数			
50	2.4577	2.4245	10.8431
100	2.9003	2.8745	13.7074
200	3.7898	3.7713	18.6849

Table 2. The precision matrix performance index under different transformations obtained in mode 2
表 2. 由方式 2 得到的不同变换下精度矩阵性能指标

p	$\hat{\Omega}$	$\hat{\Omega}_y$	$\hat{\Omega}_l$
谱范数			
50	0.1212	0.1187	0.5845
100	0.1190	0.1177	0.6227
200	0.1195	0.1188	0.6112
Frobenius 范数			
50	0.6286	0.6160	2.9426
100	0.8574	0.8486	4.0881
200	1.1897	1.1838	5.7714

表 1 和表 2 比较了不同变换下的精度矩阵的谱范数和 Frobenius 范数，无论是谱范数还是 Frobenius 范数，$\hat{\Omega}$ 与理想估计 $\hat{\Omega}_Y$ 的性能指标几乎一致，从而验证我们提出方法的合理性。

4. 与炎症性肠病(IBD)相关的细菌物种数据集分析

IBD 数据集收集了 85 例 IBD 病例的粪便样本和 26 个正常对照样本，并对每个样品进行宏基因组测序，从而鉴定出总共 97 种细菌物种[15]。对于数据集中的零元素，在不超过数据生成过程中的最小探测精度的条件下，我们取 10^{-3}。取正常样本($k=1$)的 1/5，病例样本($k=2$)的 1/5 组成测试集，其他样本组成训练集。然后我们对数据集进行线性判别分析，其分析模型可以参见文献[11]。

$$\delta_k(X) = X^{\mathrm{T}}\hat{\Omega}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k^{\mathrm{T}}\hat{\Omega}\hat{\mu}_k + \log\hat{\pi}_k \tag{8}$$

其中 $\hat{\pi}_k = n_k/n$，$\hat{\mu}_k = 1/n_k\sum_i X_i$。$\arg\max_k \delta_k(X), k=1,2$ 为分类标准，分类性能与 $\hat{\Omega}$ 的估计精度密切相关。

我们对 IBD 数据集进行分类，用 *TPR* (真正类率)，*FPR* (假正类率)，*MCC* (马修斯相关系数)来评估分类的情况。

$$TPR = \frac{TP}{TP+FN},\quad FPR = \frac{FP}{FP+TN} \tag{9}$$

$$MCC = \frac{TP\times TN - FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{10}$$

其中 *TP* 代表真阳性(正常)，*TN* 代表真阴性(病例)，*FP* 代表假阳性，*FN* 代表假阴性。

Table 3. Classification performance result
表 3. 分类性能结果

TPR	*FPR*	*MCC*
0.8571	0.0455	0.8117

从表 3 中，我们可以看出，我们的判别分析方法对该数据集的分类性能较好，可以得出高维成分数据的精度矩阵估计 $\hat{\Omega}$ 对实际数据的处理性能良好。

5. 结语

本文通过对高维成分数据的分析和处理，得到了高维成分数据的精度矩阵估计方法，对高维成分数据的研究具有一定的实际意义。协方差仅刻画了两个分量间相关性大小，不能衡量二者的直接关联性，研究某些问题时，如微生物菌种问题，研究人员往往对菌种间的直接相互作用更感兴趣，而精度矩阵可以衡量两个菌种间的直接相互作用。并且本文通过模拟实验和 IBD 数据集验证了该方法的合理性，可以较好的处理高维成分数据。但对精度矩阵稀疏性和可识别性还需要进一步的研究和讨论。

基金项目

教育部人文社会科学研究规划基金项目(18YJA880077)。

参考文献

[1] Ferrers, N.M. (1866) An Elementary Treatise on Trilinear Coordinates. Macmillan, London.

[2] Aitchison, J. (1968) The Statistical Analysis of Compositional Data. Chapman and Hall, London.

[3] Aitchison, J. (1994) A Concise Guide to Compositional Data Analysis. Institute of Mathematical Statistics Lecture Notes—Monograph Series, Vol. 24, 73-81. https://doi.org/10.1214/lnms/1215463786

[4] Aitchison, J. and Egozcue, J.J. (2005) Compositional Data Analysis: Where Are We and Where Should We Be Heading. *Mathematical Geology*, **37**, 829-850. https://doi.org/10.1007/s11004-005-7383-7

[5] Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., *et al*. (2003) Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology*, **35**, 279-300. https://doi.org/10.1023/A:1023818214614

[6] Wang, H., Liu, Q., Henry, M.K., *et al*. (2007) A Hyperspherical Transformation Forecasting Model for Compositional Data. *European Journal of Operational Research*, **179**, 459-468. https://doi.org/10.1016/j.ejor.2006.03.039

[7] Bickel, P.J. and Levina, E. (2008) Covariance Regularization by Thresholding. *Annals of Statistics*, **36**, 2577-2604. https://doi.org/10.1214/08-AOS600

[8] Rothman, A.J., Levina, E. and Zhu, J. (2009) Generalized Thresholding of Large Covariance Matrices. *Journal of the American Statistical Association*, **104**, 177-186. https://doi.org/10.1198/jasa.2009.0101

[9] Cai, T. and Liu, W. (2011) Adaptive Thresholding for Sparse Covariance Matrix Estimation. *Journal of the American Statistical Association*, **106**, 672-684. https://doi.org/10.1198/jasa.2011.tm10560

[10] Friedman, J., Hastie, T. and Tibshirani, R. (2008) Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics*, **9**, 432-441. https://doi.org/10.1093/biostatistics/kxm045

[11] Cai, T., Liu, W. and Luo, X. (2011) A Constrained L1 Minimization Approach to Sparse Precision Matrix Estimation. *Journal of the American Statistical Association*, **106**, 594-607. https://doi.org/10.1198/jasa.2011.tm10155

[12] Liu, W. and Luo, X. (2015) Fast and Adaptive Sparse Precision Matrix Estimation in High Dimensions. *Journal of Multivariate Analysis*, **135**, 153-162. https://doi.org/10.1016/j.jmva.2014.11.005

[13] Fan, J., Liao, Y. and Liu, H. (2016) An Overview of the Estimation of Large Covariance and Precision Matrices. *The Econometrics Journal*, **19**, C1-C32. https://doi.org/10.1111/ectj.12061

[14] Cao, Y., Lin, W. and Li, H. (2018) Large Covariance Estimation for Compositional Data via Composition-Adjusted Thresholding. *Journal of the American Statistical Association*, **114**, 759-772.

[15] Lu, J.R., Shi, P.X. and Li, H.Z. (2018) Generalized Linear Models with Linear Constraints for Microbiome Compositional Data. *Biometrics*.