Biological network inference as multivariate count regression

W. Evan Durno

Topics

- A multivariate count regression problem from bioinformatics
- A solution by Zuo et al. (2014)
- Usage of False Discovery Rate (FDR)

Scientific Question

Can we understand how microbiota effect their environment and each other?

My data

- Several data sets, here Saanich Inlet
- 112 Samples
- 36,000 Dimensions, microbial counts
  - Perhaps only 1% of interest due to over-abundance of zeros
- 23 Regressors

- Motivation: Environmental, others are industrial and medical

The counts

0	0	3	0	0	0	2	0	0	0	1	0	1	0	0	0	1	3	1
7	62	1	0	0	0	0	0	0	2	16	3	37	21	3	0	0	7	29
0	0	5	0	0	0	5	1	0	1	1	1	0	0	2	0	1	3	2
0	0	1	3	0	0	0	0	0	2	3	1	1	1	0	0	3	0	0
11	114	5	2	0	0	47	9	0	85	160	55	75	28	0	0	3	22	83
2	4	1	0	0	0	1	0	0	4	3	1	4	0	0	0	2	1	1
0	0	0	1	0	0	0	0	0	0	1	1	0	0	2	0	1	2	1
13	6	35	15	0	0	15	1	0	11	16	9	1	1	36	1	37	24	30
5	103	13	1	0	0	19	5	1	69	652	81	387	78	3	0	6	13	29
1	1	9	5	3	2	1	0	0	0	0	0	0	0	20	62	8	10	9
1	2	12	4	0	0	9	3	0	21	5	1	2	0	1	0	5	6	4
7	7	30	11	0	0	18	3	0	21	24	19	6	3	28	3	16	12	20
1	2	0	1	0	0	0	1	0	0	1	0	0	0	1	0	1	1	2
0	4	14	8	0	0	13	5	0	17	12	8	7	3	3	0	7	10	7
52	19	196	80	0	0	93	15	0	27	34	11	3	1	268	1	189	151	106
70	32	81	24	0	0	134	20	3	112	217	93	133	28	44	0	38	78	91
1	0	4	3	1	0	0	0	0	0	0	0	0	0	11	9	6	0	0
1	0	0	1	0	0	0	0	0	5	1	1	2	0	1	0	1	2	2
2	5	54	21	0	0	8	1	0	2	1	0	2	1	10	0	13	2	1
0	1	0	0	0	0	0	0	0	0	1	1	2	0	2	0	4	1	3

The species

```
[52] "k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodobacterales;f__Rhodobacter;g__Marinosulfonomonas;s__"
[53] "k__Archaea;p__Thaumarchaeota;c__Cenarchaeales;o__Cenarchaeum;f__Unclassified;g__;s__"
[54] "k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Betaproteobacteria;f__Methylophilales;g__Unclassified;s__"
[55] "k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__SUP05;f__Unclassified;g__;s__"
[56] "k__Bacteria;p__Proteobacteria;c__Deltaproteobacteria;o__Sva0853;f__SAR324;g__;s__"
[57] "k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Unclassified;f__;g__;s__"
[58] "k__Bacteria;p__Actinobacteria;c__Acidimicrobidae;o__Microthrixineae;f__Unclassified;g__;s__"
[59] "k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__SUP05;f__Unclassified;g__;s__"
[60] "k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__SUP05;f__mussel_thioautotrophic_gill_symbiont_MAR1;g__;s__"
[61] "k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__ZA3420c;f__;g__;s__"
[62] "k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__ZA2333c;f__Arctic96B-16;g__;s__"
[63] "k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodobacterales;f__Rhodobacter;g__alpha_proteobacterium_HTCC2255;s__"
[64] "k__Bacteria;p__VHS-B5-50;c__;o__;f__;g__;s__"
[65] "k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Unclassified;f__;g__;s__"
[66] "k__Archaea;p__Thaumarchaeota;c__Cenarchaeales;o__Cenarchaeum;f__Unclassified;g__;s__"
[67] "k__Archaea;p__Unclassified;c__;o__;f__;g__;s__"
[68] "k__Bacteria;p__Marine_group_A;c__Arctic95A-2;o__;f__;g__;s__"
[69] "k__Bacteria;p__Bacteroidetes;c__Flavobacteriales;o__Unclassified;f__;g__;s__"
[70] "k__Archaea;p__Methanococci_Eury;c__Methanocaldococcaceae;o__;f__;g__;s__"
[71] "k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__S23_91;f__;g__;s__"
[72] "k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__OM38;f__OM25;g__Unclassified;s__"
[73] "k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodobacterales;f__Rhodobacter;g__Unclassified;s__"
[74] "k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Consistiales;f__Pelagibacter;g__SAR11;s__Candidatus_Pelagibacter"
[75] "k__Bacteria;p__Verrucomicrobia;c__Opitutae;o__MB11C04;f__Unclassified;g__;s__"
[76] "k__Archaea;p__Thaumarchaeota;c__Cenarchaeales;o__Cenarchaeum;f__Unclassified;g__;s__"
[77] "k__Bacteria;p__WS3;c__Unclassified;o__;f__;g__;s__"
[78] "k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Consistiales;f__Pelagibacter;g__SAR11;s__Candidatus_Pelagibacter"
[79] "k__Bacteria;p__Marine_group_A;c__SAR406;o__;f__;g__;s__"
[80] "k__Bacteria;p__WS3;c__Unclassified;o__;f__;g__;s__"
```

Regressors

```
> colSums( is.na(x) )
  cruise     year    month      day    depth   ctd_o2      po4       si
       0        0        0        0        0       22        0       16
     no3      nh4      no2      h2s     flow  mean_n2   std_n2  mean_o2
       0       11       10       12       19       96      100       84
  std_o2 mean_co2  std_co2 mean_ch4  std_ch4 mean_n2o  std_n2o
     101       96      100        6       27       18       27
> x[1:5,]
                     cruise year month day depth    ctd_o2    po4     si
SI030.02.11.2009.100m     30 2009     2  11   100 132.42816 2.8369 27.142
SI030.02.11.2009.10m      30 2009     2  11    10 179.88894 2.4937 23.540
SI030.02.11.2009.120m     30 2009     2  11   120  91.28042 3.6751 24.857
SI030.02.11.2009.135m     30 2009     2  11   135  51.48184 4.2086 23.486
SI030.02.11.2009.150m     30 2009     2  11   150  19.47129 4.6442 27.276
                        no3 nh4        no2 h2s   flow  mean_n2     std_n2
SI030.02.11.2009.100m 28.124  NA 0.13080169   0 346470 495.1455 157.504632
SI030.02.11.2009.10m  29.764  NA 0.12974684   0 319440 353.7769  42.568231
SI030.02.11.2009.120m 14.743  NA 0.04008439   0 355860 355.4566   4.241642
SI030.02.11.2009.135m 10.226  NA 0.67299578   0 376530 416.5284  86.524271
SI030.02.11.2009.150m 11.458  NA 0.58227848   0 485250 386.7999  16.732528
                         mean_o2     std_o2  mean_co2  std_co2 mean_ch4 std_ch4
SI030.02.11.2009.100m 187.048983 41.2083249  97.07419 7.057509    129.8    16.0
SI030.02.11.2009.10m  216.942744  0.3278438  84.37912 3.082387      4.0     1.7
SI030.02.11.2009.120m   6.711028  2.4218808 150.08555 2.131197      0.0     1.2
SI030.02.11.2009.135m  14.906479 10.5388750 153.05189 3.027811     28.4     8.0
SI030.02.11.2009.150m   8.959354  2.6632124 150.47612 4.775712     33.6     2.8
                      mean_n2o std_n2o
SI030.02.11.2009.100m     16.6     0.1
SI030.02.11.2009.10m      12.1     1.2
SI030.02.11.2009.120m     13.8     1.4
SI030.02.11.2009.135m      1.2     5.8
SI030.02.11.2009.150m      0.0     3.9
```

Why multivariate regression?

A mathematical model of the whole community with its environment describes the microbial mechanics behind phenomena.

The plan

- Be highly particular about univariate count distributions
- Model multivariate structure with Gaussian copula
- Estimate partial correlation structure to highlight essential microbial relationships and ignore spurious results.

Methods 69 (2014) 266–273

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Biological network inference using low order partial correlation

Yiming Zuo [a,b], Guoqiang Yu [b], Mahlet G. Tadesse [c], Habtom W. Ressom [a,*]

[a] Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC, USA
[b] Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA, USA
[c] Department of Mathematics and Statistics, Georgetown University, DC, USA

ARTICLE INFO

Article history:
Received 3 April 2014
Revised 14 June 2014

ABSTRACT

Biological network inference is a major challenge in systems biology. Traditional correlation-based network analysis results in too many spurious edges since correlation cannot distinguish between direct and indirect associations. To address this issue, Gaussian graphical models (GGM) were proposed and have

Partial correlations with order

- $\rho_{X,Y \bullet Z}$ is the conditional correlation of X & Y given dimensions Z on a multivariate Gaussian.
- $\rho_{X,Y \bullet Z}$ has order #Z
- Zuo et al. Advocate their use, motivating their work with increased sparsity.
- An n^{th}-order partial correlation of X & Y is called significant if for all Z s.t. #Z ≤ n, $\rho_{X,Y \bullet Z}$ is significant (BH corrected).
  - Accept value as $\max\{\rho_{X,Y \bullet Z} : \#Z = n\}$.

The argument for sparsity

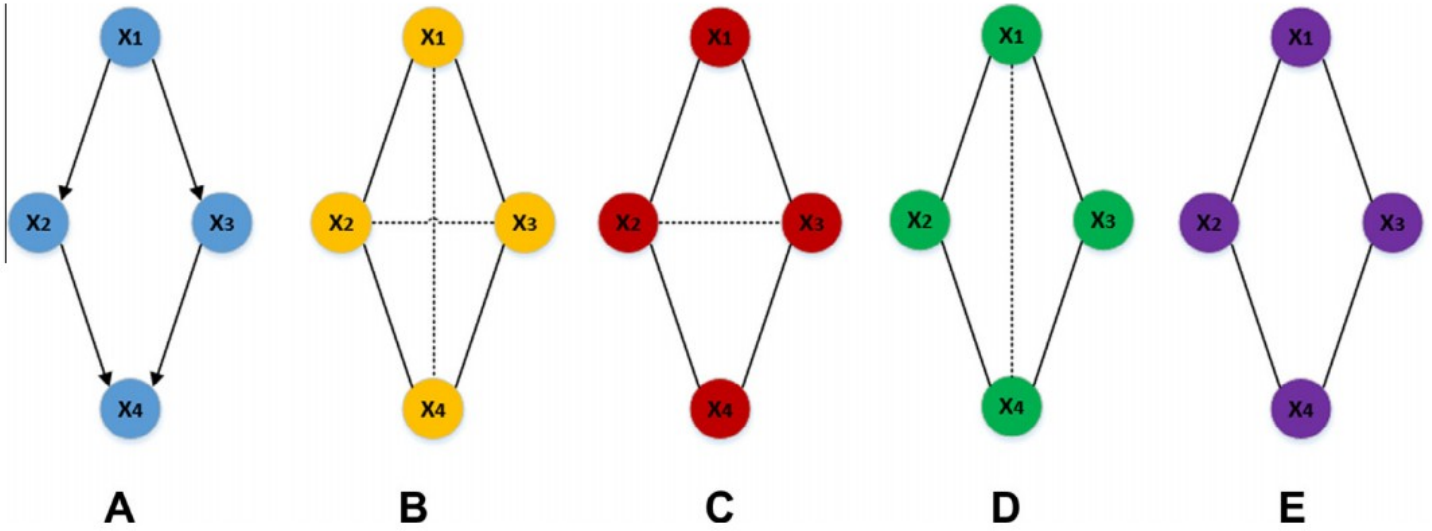

Fig. 2. Cyclic structure networks inferred based on correlation, GGM, 0–1 graph and LOPC. (A) The true network from the model. (B) Network inferred based on correlation: the dot lines represent the spurious edges. (C) Network inferred based on GGM: by only conditioning on the $(p-2)$-th order (i.e., second order in this model), it is insufficient to uncover the relationships between variables faithfully. (D) Network inferred based on 0–1 graph (up to first order): by only conditioning on up to first order, the indirect association between x_1 and x_4 cannot be removed since there are two paths from x_1 to x_4 either through x_2 or x_3. (E) Network inferred based on LOPC (up to second order): the connections in A are faithfully uncovered.

Demonstration

```
> # Demonstrate limitation of (p-2)-partial correlations
> f = function(n)
+ {
+         # data matrix
+         x = matrix( rnorm(5*n) , ncol=5)
+         x[,2] = 2*x[,1] + x[,2]
+         x[,3] = x[,1] + 0.5*x[,3]
+         x[,4] = 2*x[,2] + 3*x[,3] + x[,4]
+         x[,5] = 1.5*x[,5]
+
+         # partial correlations
+         x = cov2cor( solve( cor(x) ) )
+
+         # z scores
+         z = function(x) sqrt( n - 5+2-3 ) * abs( 0.5*log((1+x)/(1-x)) )
+
+         # return p-values
+         pnorm( z(x) , lower.tail=F )
+ }
> f(10000000)
           [,1]      [,2]       [,3]      [,4]       [,5]
[1,] 0.0000000 0.0000000 0.00000000 0.2632242 0.35232303
[2,] 0.0000000 0.0000000 0.00000000 0.0000000 0.16645231
[3,] 0.0000000 0.0000000 0.00000000 0.0000000 0.04410399
[4,] 0.2632242 0.0000000 0.00000000 0.0000000 0.12579031
[5,] 0.3523230 0.1664523 0.04410399 0.1257903 0.00000000
>
```

My argument:
My sample size is still too small!

A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation

Tony CAI, Weidong LIU, and Xi LUO

This article proposes a constrained ℓ_1 minimization method for estimating a sparse inverse covariance matrix based on a sample of n iid p-variate random variables. The resulting estimator is shown to have a number of desirable properties. In particular, the rate of convergence between the estimator and the true s-sparse precision matrix under the spectral norm is $s\sqrt{\log p/n}$ when the population distribution has either exponential-type tails or polynomial-type tails. We present convergence rates under the elementwise ℓ_∞ norm and Frobenius norm. In addition, we consider graphical model selection. The procedure is easily implemented by linear programming. Numerical performance of the estimator is investigated using both simulated and real data. In particular, the procedure is applied to analyze a breast cancer dataset and is found to perform favorably compared with existing methods.

KEY WORDS: Covariance matrix; Frobenius norm; Gaussian graphical model; Precision matrix; Rate of convergence; Spectral norm.

1. INTRODUCTION

Estimation of a covariance matrix and its inverse is an important problem in many areas of statistical analysis; among the many interesting examples are principal components analysis, linear/quadratic discriminant analysis, and graphical models. Stable and accurate covariance estimation is becoming in-

and Bickel and Levina (2008b) proposed thresholding of the sample covariance matrix for estimating a class of sparse covariance matrices and obtained rates of convergence for the thresholding estimators.

Estimation of the precision matrix $\mathbf{\Omega}_0$ is more involved due to the lack of a natural pivotal estimator like $\mathbf{\Sigma}_n$. Assuming certain ordering structures, methods based on banding the Cholesky

Convergence guarantees for my data are inaccessible

- Under unrealistically best assumptions (standardized Gaussian data, 360 dimensions), my data are too few.

$$| \Omega_0 - \Omega_{CLIME} |_\infty \leq C \left(\log(p)/n \right)^{1/2} \approx 1000 * 0.23$$

Remaining question:
Is FDR applied correctly?

B&H assume independence of rejected p-values

J. R. Statist. Soc. B (1995)
57, *No.* 1, *pp.* 289–300

Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing

By YOAV BENJAMINI† and YOSEF HOCHBERG

Tel Aviv University, Israel

[Received January 1993. Revised March 1994]

SUMMARY

The common approach to the multiplicity problem calls for controlling the familywise error rate (FWER). This approach, though, has faults, and we point out a few. A different approach to problems of multiple significance testing is presented. It calls for controlling the expected proportion of falsely rejected hypotheses—the false discovery rate. This error rate is equivalent to the FWER when all hypotheses are true but is smaller otherwise. Therefore, in problems where the control of the false discovery rate rather than that of the FWER is desired, there is potential for a gain in power. A simple sequential Bonferroni-type procedure is proved to control the false discovery rate for independent test statistics, and a simulation study shows that the gain in power is substantial. The use of the new procedure and the appropriateness of the criterion are illustrated with examples.

Keywords: BONFERRONI-TYPE PROCEDURES; FAMILYWISE ERROR RATE; MULTIPLE-COMPARISON PROCEDURES; *p*-VALUES

Harry's book helped me learn about partial correlations