Bickel's Influence on My Career

Celebrate Peter's 82nd birthday!

- **Covariance Matrix Estimation**
- **Network Analysis**
- **Variational Inference**

Covariance Matrix Estimation

The Faculty Club

REGULARIZED ESTIMATION OF LARGE COVARIANCE MATRICES

BY PETER J. BICKEL AND ELIZAVETA LEVINA[1]

University of California, Berkeley and University of Michigan

This paper considers estimating a covariance matrix of p variables from n observations by either banding or tapering the sample covariance matrix, or estimating a banded version of the inverse of the covariance. We show that these estimates are consistent in the operator norm as long as $(\log p)/n \to 0$, and obtain explicit rates. The results are uniform over some fairly natural

Bandable Covariance Matrix Estimation

Model: Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be iid $N(0, \Sigma_{p\times p})$.

Goal: Estimate $\Sigma_{p\times p}$ under the spectral norm.

Parameter space: For $\Sigma = (\sigma_{ij})_{1\le i,j\le p}$, assume that

$$|\sigma_{ij}| \le M\left(|i-j|+1\right)^{-(\alpha+1)}$$

for some $\alpha > 0$ and $M > 0$.

Note that for all k,

$$\max_i \sum_j |\sigma_{ij}| \mathbb{I}\left\{|i-j| > k\right\} \le Ck^{-\alpha}.$$

Banding Estimation

Sample covariance:

$$\tilde{\Sigma} = (\tilde{\sigma}_{ij})_{1\leq i,j\leq p} = \frac{1}{n-1}\sum_{l=1}^{n}\left(\mathbf{X}_l - \bar{\mathbf{X}}\right)\left(\mathbf{X}_l - \bar{\mathbf{X}}\right)^T$$

which is an unbiased estimator of $\Sigma = (\sigma_{ij})_{1\leq i,j\leq p}$.

Banding estimator: Bickel and Levina (2008a)

$$\hat{\Sigma} = \left(\tilde{\sigma}_{ij}\mathbb{I}\left\{|i-j|\leq k\right\}\right)_{p\times p}.$$

Bickel and Levina (2008a): A Fascinating Rate

Analysis: Bound the spectral norm by the matrix l_1 norm

$$\begin{aligned}\mathbb{E}\left\|\hat{\Sigma}-\Sigma\right\|_1^2 &\leq 2\mathbb{E}\left\|\hat{\Sigma}-\mathbb{E}\hat{\Sigma}\right\|_1^2+2\left\|\mathbb{E}\hat{\Sigma}-\Sigma\right\|_1^2 \\ &\lesssim k^2\frac{\log p}{n}+k^{-2\alpha}.\end{aligned}$$

Rate of convergence: Set $k=(\frac{n}{\log p})^{\frac{1}{2+2\alpha}}$. We have

$$\mathbb{E}\left\|\hat{\Sigma}-\Sigma\right\|_1^2 \lesssim \left(\frac{\log p}{n}\right)^{\frac{\alpha}{\alpha+1}}=o(1),$$

as long as $\log p = o(n)$.

Remark: $\|A\|_2 \leq \|A\|_1 = \max_j \sum_i |a_{ij}|$, for A symmetric.

Our Follow-up: Optimality of Banding Estimation

Upper bound:

$$\begin{aligned} \mathbb{E}\left\|\hat{\Sigma}-\Sigma\right\|_2^2 &\leq 2\mathbb{E}\left\|\hat{\Sigma}-\mathbb{E}\hat{\Sigma}\right\|_2^2+2\left\|\mathbb{E}\hat{\Sigma}-\Sigma\right\|_1^2 \\ &\lesssim \frac{k}{n}+\frac{\log p}{n}+k^{-2\alpha}. \end{aligned}$$

Lower bound: Assouad and Le Cam

Reference: Cai, Zhang, Z. (2010)

COVARIANCE REGULARIZATION BY THRESHOLDING

BY PETER J. BICKEL[1] AND ELIZAVETA LEVINA[2]

University of California, Berkeley and University of Michigan

This paper considers regularizing a covariance matrix of p variables estimated from n observations, by hard thresholding. We show that the thresholded estimate is consistent in the operator norm as long as the true covariance matrix is sparse in a suitable sense, the variables are Gaussian or sub-Gaussian, and $(\log p)/n \to 0$, and obtain explicit rates. The results are uniform over families of covariance matrices which satisfy a fairly natural notion of sparsity. We discuss an intuitive resampling scheme for threshold selection and prove a general cross-validation result that justifies this approach. We also compare thresholding to other covariance estimators in simulations and on an example from climate data.

Bickel and Levina (2008b)

Parameter space:

$$\max_j \sum_i \mathbb{I}(\sigma_{ij} \neq 0) \leq s, \ \max_i \sigma_{ii} \leq M.$$

Thresholding estimation:

$$\hat{\sigma}_{i,j} = \tilde{\sigma}_{i,j}\mathbb{I}(|\tilde{\sigma}_{i,j}| > \lambda), \quad \text{with} \quad \lambda = c\sqrt{\frac{\log p}{n}}.$$

Upper bound:

$$\begin{aligned} \mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|_2 &\leq \mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|_1 \\ &\lesssim s\sqrt{\frac{\log p}{n}}. \end{aligned}$$

Our Follow-up: Optimality of Thresholding Estimation

Lower bound: Assouad – Le Cam

Reference: Cai, Z. (2012) for s relatively small.

Our Other Follow-ups

- **Inference for Gaussian graphical model:** Ren, Sun, Zhang, Z. (2015)
- **PCA and CCA:** Gao, Ma, Z. (2017)
- **Bayesian estimation:** Gao, Z. (2015, 2016)

Network Analysis

More on Network Models

Yale, May 30, 2012

Peter Bickel

Statistics Dept. UC Berkeley

(Joint work with S. Bhattacharyya *UC Berkeley*, A. Chen *Google*, D.

Choi *UC Berkeley* and , E. Levina, *U. Mich*)

A nonparametric view of network models and Newman–Girvan and other modularities

Peter J. Bickel[a,1] and Aiyou Chen[b]

[a]University of California, Berkeley, CA 94720; and [b]Alcatel-Lucent Bell Labs, Murray Hill, NJ 07974

Edited by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, and approved October 13, 2009 (received for review July 2, 2009)

Prompted by the increasing interest in networks in many fields, we present an attempt at unifying points of view and analyses of these objects coming from the social sciences, statistics, probability and physics communities. We apply our approach to the Newman–Girvan modularity, widely used for "community" detection, among others. Our analysis is asymptotic but we show by simulation and application to real examples that the theory is a reasonable guide to practice.

modularity | profile likelihood | ergodic model | spectral clustering

The social sciences have investigated the structure of small networks since the 1970s, and have come up with elaborate modeling strategies, both deterministic, see Doreian et al. (1) for

principle, "fail-safe" for rich enough models. Moreover, our point of view has the virtue of enabling us to think in terms of "strength of relations" between individuals not necessarily clustering them into communities beforehand.

We begin, using results of Aldous and Hoover (9), by introducing what we view as the analogues of arbitrary infinite population models on infinite unlabeled graphs which are "ergodic" and from which a subgraph with n vertices can be viewed as a piece. This development of Aldous and Hoover can be viewed as a generalization of deFinetti's famous characterization of exchangeable sequences as mixtures of i.i.d. ones. Thus, our approach can also be viewed as a first step in the generalization of the classical construction of complex statistical models out of i.i.d. ones using covariates, information about labels and relationships.

Theorem 1. *Suppose F, S and π satisfy I–III and $\hat{\mathbf{c}}$ is the maximizer of $Q(\mathbf{e}, A)$. Suppose $\frac{\lambda_n}{\log n} \to \infty$. Then, for all $(\pi, S) \in \Theta$,*

$$\overline{limit}_{n\to\infty} \frac{\log \mathbb{P}(\hat{\mathbf{c}} \neq \mathbf{c})}{\lambda_n} \leq -s_Q(\pi, S) < 0.$$

connections than within-community connections while N-G more or less maximizes within-community connections. We have verified that the group communicating with all others is a service group.

Discussion

1. As we noted, under our conditions the usual statistical goal of estimating the parameters π and P is trivial, since, once we have assigned individuals to the K communities consistently, the natural estimates, $\hat{W}$ and $\hat{\pi}$, are not just consistent but efficient. However, in the more realistic case where $\lambda_n = \Omega(1)$, or even just $\lambda_n = \Omega(\log n)$, this is no longer true. Elsewhere, we shall show that, indeed, estimation of parameters by maximum likelihood and Bayes classification of individuals (no longer perfect) is optimal.
2. A difficulty faced by all these methods, modularities or likelihoods, is that if K is large, searching over the space of classifications becomes prohibitively expensive. In subsequent work we intend to show that this difficulty may

neighborhood of the estimated values.

Open Problems

1. A fundamental difficulty not considered in the literature is the choice of K. From our nonparametric point of view, this can equally well be seen as, how to balance bias and variance in the estimation of $w(\cdot, \cdot)$. We would like to argue that, as in nonparametric statistics, estimating $w(\cdot, \cdot)$ without prior prejudices on its structure is as important an exploratory step in this context as, using histograms in ordinary statistics.
2. The linking of this framework to covariates depending on vertice or edge identity is crucial, permitting relationship strength to be assessed as a function of vector variables.
3. The links of our approach to spectral graph clustering and more generally clustering on the basis of similarities seem intriguing.

ACKNOWLEDGMENTS. We thank Tin K Ho for help in obtaining the PBX data and for helpful discussions. We also thank the referees, whose references and comments improved this article immeasurably.

Model: 2 Communities

Partition:

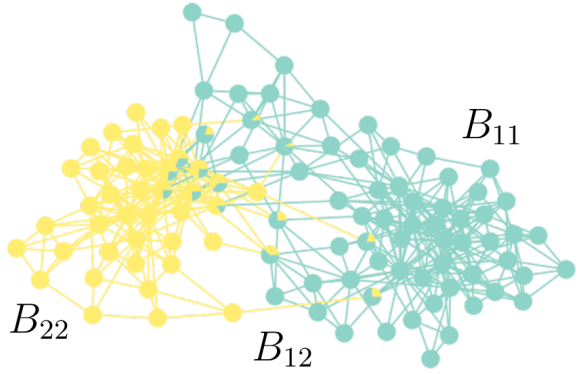
$$z : \{1, 2, ..., n\} \to \{1, 2\},$$

where $z(i)$'s are i.i.d. $Bernoulli(\pi)$.

Observation: The adjacency matrix A is

$$A_{ij} \sim \text{Bernoulli}(p_{ij}), \text{ independent, for } i > j$$

with $p_{ij} = a$ if $z(i) = z(j)$, and $p_{ij} = b$ otherwise, where $a > b$.



Remark: All optimality results stated can be extended to k-community case.

Questions

Global parameters estimation: π, a, and b.

Graphon estimation: $P = (p_{ij})$.

Community detection and spectral clustering: z.

Global Parameters Estimation

THE METHOD OF MOMENTS AND DEGREE DISTRIBUTIONS FOR NETWORK MODELS

By Peter J. Bickel[1], Aiyou Chen[2] and Elizaveta Levina[3]

University of California, Berkeley, Google Inc. and University of Michigan

This research is dedicated to Erich L. Lehmann, the thesis advisor of one of us and "grand thesis advisor" of the others. It is a work in which we try to develop nonparametric methods for doing inference in a setting, unlabeled networks, that he never considered. However, his influence shows in our attempt to formulate and develop a nonparametric model in this context. We also intend to study to what extent a potentially "optimal" method such as maximum likelihood can be analyzed and used in this context. In this respect, this is the first step on a road he always felt was the main one to stick to.

Probability models on graphs are becoming increasingly important in many applications, but statistical tools for fitting such models are not yet well developed. Here we propose a general method of moments approach that can be used to fit a large class of probability models through empirical counts of certain patterns in a graph. We establish some general asymptotic properties of empirical graph moments and prove consistency of the estimates as the graph size grows for all ranges of the average degree including $\Omega(1)$. Additional results are obtained for the important special case of degree distributions.

Global Parameters Estimation

THEOREM 2. *Suppose $\theta = (\pi, S)$ defines a block model with known K, and the vectors $\pi, F\pi, \ldots, F^{K-1}\pi$ are linearly independent. Suppose $\varepsilon \leq \lambda_n = o(n^{1/2})$. Then:*

(a) *$\{\tau_{kl} : l = 1, \ldots, 2K-1, k = 2, \ldots, K\}$ identify the $K(K+3)/2 - 2$ parameters of the block model other than ρ (i.e., the map f is one to one).*

(b) *If f has a gradient which is of rank $\frac{K(K+3)}{2} - 2$ at the true (π_0, S_0), then $f^{-1}(P(\check{\tau}))$ is a $\sqrt{n}$-consistent estimate of (π_0, S_0), where $\check{\tau} = \|\check{\tau}_{kl}\|$ and $P(\check{\tau})$ is the closest point in the range of f to $\check{\tau}$.*

Global parameters estimation:

Under an assumption $\frac{a}{b} \geq 1 + c$ for some fixed constant $c > 0$, $b \asymp 1$, and $\pi \neq 1/2$, we have

$$\mathbb{E}(\hat{a} - a)^2 \lesssim \frac{1}{n}.$$

Graphon Estimation

Graphon estimation:

$$\mathbb{E}\frac{1}{n^2}\|\hat{P}-P\|_{\mathrm{F}}^2 \lesssim \frac{n}{n^2}=\frac{1}{n}.$$

Global parameters estimation as a corollary:

$$\mathbb{E}(\hat{a}-a)^2 \lesssim \frac{1}{n},$$

without the assumption $\frac{a}{b} \geq 1+c$ for some fixed constant $c>0$, $b \asymp 1$, and $\pi \neq 1/2$.

Reference: Gao, Yu, Z. (2015)

An Unpublished Result

Global parameters estimation: Bickel, Feng, Z.

Under an assumption $\frac{a}{b} \geq 1 + c$ for some fixed constant $c > 0$, $b \asymp 1$, and $\pi \neq 1/2$, we have

$$\mathbb{E}(\hat{a} - a)^2 \lesssim \frac{1}{n^2}.$$

Community Detection

Optimality: Under the assumption $nI \to \infty$,

$$\inf_{\hat{z}} \sup_{\Theta} \mathbb{E} L(\hat{z}, z) = \exp\left(-(1 + o(1))nI/2\right).$$

Key quantity:

$$I = -2 \log\left(\sqrt{a}\sqrt{b} + \sqrt{1-a}\sqrt{1-b}\right).$$

Remark: $L(\hat{z}, z)$ is the proportion of mislabeling.

Reference: Zhang and Z. (2016).

Our Other Follow-ups

- **Community detection and spectral clustering:** Gao, Ma, Zhang, Z. (2017, 2018).

- **Bayesian estimation:** Gao, van der Vaart, Z. (2020)

Variational Inference

Block Models with Covariates: Likelihood Methods of Fitting

Bickel, Peter J. (bickel@stat.berkeley.edu)
University of California, Berkeley

Type: Plenary Talk

Abstract. We introduce block models with edge and block covariates, along the lines of Hoff, Handcock, Raftery (2002), specializing to covariate forms of the types proposed by Zhang, Levina, Zhu (2014) and generalizing that of Newman, Clauset(2015). We study maximum likelihood and mean field variational fitting for these methods along the lines of Celisse, Daudin, Pierre (2011) and B., Choi, Chang, Zhang (2013) and partly extend their results to the regime where the average degree tends to infinity faster than loglog(n). We show by example and simulation when mean field methods work and how they can be adapted to succeed when they fail. Co-authors: Purna Sarkar (U. of Texas, Austin), Soumendu Mukherjee (UC, Berkeley), Sharmodeep Bhattacharyya (Oregon State University) and David Choi (Carnegie Mellon University).

Keywords: Block models; Field variational fitting; Infinite average degree.

ASYMPTOTIC NORMALITY OF MAXIMUM LIKELIHOOD AND ITS VARIATIONAL APPROXIMATION FOR STOCHASTIC BLOCKMODELS[1]

BY PETER BICKEL, DAVID CHOI, XIANGYU CHANG AND HAI ZHANG

University of California, Berkeley, Carnegie Mellon University, Xi'an Jiaotong University and Northwest University

Variational methods for parameter estimation are an active research area, potentially offering computationally tractable heuristics with theoretical performance bounds. We build on recent work that applies such methods to network data, and establish asymptotic normality rates for parameter estimates of stochastic blockmodel data, by either maximum likelihood or variational estimation. The result also applies to various sub-models of the stochastic blockmodel found in the literature.

3.2. *Asymptotic normality of maximum likelihood under GM blockmodel.* Our main result is that for graphs with poly-log expected degree, the likelihood ratios of the CGM and GM blockmodels are essentially equivalent with probability tending to 1, so that inference under the models is essentially equivalent up to the identifiability restrictions of the GM blockmodel.

THEOREM 1. *Let* (Z, A) *be generated from a blockmodel with* $\theta_0 \in \mathcal{T}$, *such that* S_0 *has no identical columns, and* $\rho_0 = \rho_n$ *satisfies* $n\rho_n / \log n \to \infty$. *Then for all* $\theta \in \mathcal{T}$,

$$\frac{g}{g_0}(A, \theta) = \max_{\theta' \in \mathcal{S}_\theta} \frac{f}{f_0}(Z, A, \theta')\big(1 + \varepsilon_n(K, \theta')\big) + \varepsilon_n(K, \theta'), \tag{9}$$

where $\sup_{\theta \in \mathcal{T}} \varepsilon_n(K, \theta) = o_P(1)$.

Bayesian Framework

Likelihood function:

$$L(A|Z) = \prod_{i<j} P_{i,j}^{A_{i,j}} (1 - P_{i,j})^{1-A_{i,j}}.$$

Multinomial prior:

$$\mathbb{P}(Z_{i,\cdot} = e_m) = \rho_m, \forall m = 1, 2,$$

where $\{e_1, e_2\}$ are the coordinate vectors.

Posterior:

$$p(Z|A) = \frac{p(Z, A)}{\int_Z p(Z, A)},$$

which is computationally intractable.

Mean Field Method

Basic idea:

Approximate $p(Z|A)$ by a product distribution $q_\pi(Z) = \prod_i q_{\pi_{i,\cdot}}(Z_i)$ in terms of $\mathrm{KL}(q_\pi(Z)\|p(Z|A))$, where

$$\mathbb{P}(Z_{i,\cdot} = e_m) = \pi_{i,m}, \forall m = 1, 2, \sum_m \pi_{i,m} = 1$$

Mean field method:

$$\hat{q}^{\mathrm{MF}} : \hat{\pi}^{\mathrm{MF}} = \arg\min_{\pi \in \Pi_1} \mathrm{KL}(q_\pi(Z)\|p(Z|A)),$$

where $\Pi_1 = \{\pi \in [0,1]^{n\times 2}, \|\pi_{i,\cdot}\|_1 = 1\}$.

Remark: The objective is not convex in π:

$$\hat{\pi}^{\mathrm{MF}} = \arg\max_{\pi \in \Pi_1} \langle A + \lambda I_n - \lambda 1_n 1_n^T, \pi\pi^T \rangle - \frac{1}{t}\sum_{i=1}^n \mathrm{KL}(\pi_{i,\cdot}\|\rho),$$

where $\lambda = \log\frac{1-b}{1-a} / \log\frac{a(1-b)}{b(1-a)}$ and $t = \frac{1}{2}\log\frac{a(1-b)}{b(1-a)}$.

Batch Coordinate Ascent Variational Inference (BCAVI)

CAVI: For any $\pi \in \Pi_1$, we have

$$[h(\pi)]_{i,m} \propto \rho_m \exp\left(2t \sum_{j \neq i} \pi_{j,m} \left(A_{i,j} - \lambda\right)\right).$$

Batch coordinate ascent variational inference (BCAVI):

Input: Initializer $\hat{\pi}^0$, prior ρ, adjacency matrix A and parameter λ

Output: $\hat{\pi}$

1 Denote $\hat{\pi}^{(0)} = \hat{\pi}^0$;

2 Start from $s = 0$, do it recursively $\hat{\pi}^{(s+1)} = h(\hat{\pi}^{(s)})$.

Computational and statistical guarantees: Rate optimal after an order of $\log n$ steps if the error of initializer is smaller than $1/2 - \epsilon$, for some fixed $\epsilon > 0$.

Reference: Zhang and Z. (2020).

Global Convergence of EM with Random Initialization

- **Two-component Gaussian mixtures:** Wu, Z. (2022)
- **General Gaussian mixtures (ongoing):** Overparametrization + Optimal transport

Summary

- Covariance Matrix Estimation
- Network Analysis
- Variational Inference

In the past 10 to 15 years, we have been following some of Peter's groundbreaking ideas very closely, and will continue to do so.