1 spectral clustering

1.1 Inference for Graphs and Networks: Extending Classical Tools to Modern Data (Olding and Wolfe, 2009)

Examples

- Erds-Rnyi: (global exchangeability)
  one parameter model yielding independent and identically distributed binary random variables representing the absence or presence of pairwise links between nodes
  $A_{ij} \overset{iid}{\sim} Bernoulli(p)$ for $i < j$
  $\Rightarrow$ maximum likelihood estimator: $\hat{p} = \frac{1}{\binom{n}{2}} \sum_{i<j} A_{ij}$
- Generalization of Erds-Rnyi: (local exchangeability)
  Bernoulli parameters to depend on k-ary categorical covariates c(i) associated with each node,
  where $k \leq n$ represent grouping of nodes $\Rightarrow p_{c(i)c(j)}$
- Simple Stochastic block model
  Generalization of Erds-Rnyi with k=2 $\Rightarrow$ the covariate is binary $p_{c(i)c(j)} \in \{p_{00}, p_{10}, p_{11}\}$
  $A_{ij} \sim Bernoulli(p_{c(i)c(j)})$
  reordering of nodes according to covariate value via permutation similarity
  $\Rightarrow A = \begin{pmatrix} A_{00} & A_{01} \\ A_{10} & A_{11} \end{pmatrix}$
  MLE corresponding to A_{00}, A_{01}, A_{11} A_{00}, A_{11} yield subgraphs which are Erds-Rnyi; all nonzero entries in A_{01} comprise the edge boundary

If the values of the covariable are not known, interfering the vector c is equivalent to finding a permutation similarity transformation.
exact solutions are of exponential complexity $\Rightarrow$ approximate inference

Alternative: exploiting algebraic properties

- graph spectrum: set of eigenvalues of the adjacency matrix

$\rightarrow$ computational cost that scales as the cube of the number of nodes,i.e. n^2

you are interested in the eigenvalues of A, but A false to be positive semidefinite, therefore define

Graph Laplacian [p.6]

An associated $n \times n$ symmetric, positive-semidefinite matrix L is called a graph Laplacian if,

$$L : \begin{cases} L_{ij} < 0 & \text{if } i \sim j, \\ L_{ij} = 0 & \text{if } i \nsim j \end{cases}$$

- A combinatorial Laplacian takes $L = D - A$, where D is a diagonal matrix of node degrees.

- $\Rightarrow dim(kernel(L)) =$ no of connected components

- Algebraic connectivity :second-smallest eigenvalue of combinatorial Laplace L
  $\Rightarrow$ positive and negative entries of the corresponding eigenvector define a partition of nodes that nearly minimizes the number of edge removals needed to disconnect a network.
  In the extreme case of two equally sized disconnected subgraphs this procedure maximizes the likelihood of the data under a two group stochastic block model.

Definitions/ Notes:

- adjacent matrix
- degree of a node
- simple graph
- stochastic block model
- permutation similarity: for any permutation matrix Π the adjacent matrix A and $\Pi A \Pi^t$ represents isomorphic graphs
- graph spectrum

- Graph Laplacian
- algebraic connectivity (second-smallest eigenvalue of L)

1.2 A Tutorial on Spectral Clustering (von Luxburg, 2007)

- Basics: degree, complement of a set of vertices,indicator vector on a set of vertices, connected, connected component, partition
- Different similarity graphs: the ϵ-neighborhood graph, k-nearest neighbor graph, fully connected graph
- unnormalized Graph Laplacians (properties with proof)
- The normalized Laplacians (properties with proof)
- spectral clustering algorithms for all three Laplacians (L, L_{rw}, L_{sym})
- really helpful example for spectral clustering algorithms taking two different similarity graphs as well as the normalized and the unnormalized Laplacian into account
- Graph cut: three different functions are defined that are aimed to be minimized ($cut(A_1, \cdots, A_k), RatioCut(A_1, \cdots, A_k), Ncut(A_1, \cdots, A_k)$) where the latter two "'balance"' the clusters
- $RatioCut(A_1, \cdots, A_k), Ncut(A_1, \cdots, A_k)$ NP hard $\Rightarrow$ approximation
- $RatioCut(A_1, \cdots, A_k)$ is after some approximations equivalent to find the eigenvector corresponding to the second smallest eigenvalue of L
- Motivation of minimizing $Ncut$ by random walks
  look for a cut through the graph such that a random walk seldom transitions from A to $\bar{A}$ and vice versa
  without further assumptions the relation between commute time and spectral clustering is rather loose
- Motivation of unnormalized spectral clustering and normalized spectral clustering with L_{rw} by the perturbation theory approach
- ch 8 discuss some issues that come up when actually implementing spectral clustering
  - states that the choice of similarity function depends on the domain the data comes from
  - shows via an example the differences between the different types of similarity graph
  - choice of parameters of the similarity graph

Hint:
There has been no systematic study which investigates the effects of the similarity graph and its parameters on clustering and comes up with well-justified rules of thumb.

1.3 Spectral clustering and the high-dimensional stochastic blockmodel (Rohe et al., 2011)

Aim: to bound the number of nodes "misclustered" by spectral clustering whereby the asymptotic results allow the number of clusters in the model to grow with the number of nodes

Outline

1. under the more general latent space model, as the number of nodes grows, the eigenvectors of L, the normalized Laplacian, converge to the eigenvectors of the "population" normalized Laplacian
2. Thm 3.1 shows that the proportion of nodes that are "misclustered" by spectral clustering vanishes in the asymptote for Stochastic Blockmodels (It isn't shown that spectral clustering is consistent under the Stochastic Blockmodel, it only gives a bound on the number of misclassified nodes)

(ad 1)

- Lemma 2.1 state the relation between the eigenvalue and eigenvector of a matrix M and $M \cdot M$
- We assume a lower bound for τ_n, the minimum expected degree divided by the maximum possible degree, then

$$\left\| L^{(n)} L^{(n)} - \mathscr{L}^{(n)} \mathscr{L}^{(n)} \right\|_F = o\left(\frac{\log n}{\tau_n^2 n^{\frac{1}{2}}} \right) \qquad a.s.$$

  (τ_n measures how quickly the number of edges accumulates)

- $$\max_i \left| \lambda_i^{(n)} - \bar{\lambda}_i^{(n)} \right| \leq o\left(\frac{\log n}{\tau_n^2 n^{\frac{1}{2}}} \right) \qquad a.s.$$

- (Davis Kahan) Do not consider all eigenvectors. The columns of χ equal the eigenspace corresponding to the eigenvalues in $\lambda_S(\mathscr{L}\mathscr{L})$ (analog for X and $\lambda_S(LL)$)
  If χ and X are of the same dimension, then

$$\frac{1}{2} \left\| X - \chi O \right\|_F^2 \leq \frac{\left\| LL - \mathscr{L}\mathscr{L} \right\|_F^2}{\delta^2}$$

  where δ is the distance between S and the spectrum of $\mathscr{L}\mathscr{L}$ outside of S.

- Two assumptions: (1) The eigengap does not converge to zero too quickly. (2)The smallest expected degree grows sufficiently fast, i.e. almost linear. X_n and χ_n are matrices. Their columns are a subset of the eigenvectors of the observed graph Laplacian and of the eigenvectors of the population graph Laplacian, respectively (chosen by intersection with S_n').

$$\|X_n - \chi_n O_n\|_F = o\left(\frac{\log n}{\delta_n \tau_n^2 n^{\frac{1}{2}}}\right) \qquad a.s.$$

  $\Rightarrow$ Under the two assumptions above k_n of the eigenvectors of $L^{(n)}$ converge to k_n of the eigenvectors of $\mathscr{L}^{(n)}$.

(ad 2)

- Assume a k block Stochastic Blockmodel, spectral clustering applied to the population Laplacian ($\mathscr{L}$) can discover the block structure in the matrix Z
- define when a node is misclustered, such that

$$\mathscr{M} = \left\{ i : \|c_i - z_i \mu O\|_2 \geq \frac{1}{\sqrt{2P}} \right\}$$

  where $z_i\mu$ is the centroid from the analysis of the population graph Laplacian, c_i is the observed centroid, O is orthonormal rotation, $P = \max_{j=1,\cdots,k}(Z^T Z)_{jj}$ (If c_i is closer to $z_i\mu$ than to any other population centroid $z_j\mu$, $z_i \neq z_j$ then it is correctly clustered)

- If the smallest nonzero eigenvalue of $\mathscr{L}$ is not too small, and the smallest expected degree grows almost linear in n then $|\mathscr{M}|$ is bounded.

$$|\mathscr{M}| = o\left(\frac{P_n (\log n)^2}{\lambda_{k_n}^4 \tau_n^4 n}\right)$$

  (The theorem suggests to order not the eigenvalues it selves but their absolute values)

Questions

- p.10 about o(.)
-

1.4 Spectral Clustering of Graphs with General Degrees in the Extended Planted Partition Model(Chung, 2012)

The standard approach of spectral clustering works well if after projecting the nodes on the bottom k singular space the nodes from different clusters are well separated. Is this not the case then you need to normilize the degrees. When in addition, very low minimum degrees occur even the normalized graph Laplacian gets into trouble. This is the case for the Extended Planted Partition Model (Planted Partition Model is the same as a stochastic block model.), i.e.

$$G = (V, E)$$

Graph G withe the vertices V and edges E with a hidden partition. In addition, each node is associated with a parameter d_i s.t.

$$P(A_{i,j} = 1) = d_i p d_j \qquad \text{and} \qquad P(A_{i,j} = 0) = d_i q d_j.$$

Therefore, the authors introduce a *degree-corrected random walk graph Laplacian*

$$L^* = I - (D + \tau I)^{-1} A$$

where I is the identity matrix, A the adjacency matrix, D diagonal matrix of degrees and a constant $\tau \geq 0$ and the *degree-corrected normalized graph Laplacian*

$$L^{*'} = I - (D + \tau I)^{-1/2} A (D + \tau I)^{-1/2}$$

(To compare, the (random walk) normilized Laplacian of von Luxburg (2007) is defined as $L_{rw} = I - D^{-1}A$, and the symmetric normalized Laplacian $L^{*'} = I - D^{-1/2}AD^{-1/2}$.)
They provide an algorithm and prove that under certain assumptions, the algorithm produces with high probability the right partition. One of the assumptions is that all expected degrees need to be greater than approx. $\log(n)$. (I neither had a closer look at the assumptions nor did I read the proof.) The main difference in this approach to Rohe et al. (2011)s is that in this paper they allow the expected degree in each cluster to vary.

The approach seems to be interesting. If someone would want to build up on it, it would be interesting if other properties as for instance consistency can be proven.

The paper provides in the end anice overview about related work.

1.5 Detecting Overlapping Temporal Community Structure in Time-Evolving Networks Chen et al. (2013)

Aim: Detecting Overlapping Temporal Community Structure in Time-Evolving Networks

Naive: static community detection independently in each snapshot; can be extended to detect OTC structure; very sensitive to even minor changes in the network → limitations
Approach: Smoothness to past partitions is enforced
find the temporal community structure that maximizes a quality function which quantifys the community structure in each snapshot subject to a temporal smoothness constraint (a distance function to ensure contiguity with the past community structure)
*Efficiency:*not greedy heuristics but a tight convex relaxation of this set via the trace norm ⇒ convex optimization problem ⇒ efficient
Assumption: community size at least $\sqrt{\frac{n}{m}}$, where m is number of snapshots and n number of nodes
several synthetic and real network data-sets

1.6 Clustering Sparse Graphs Chen et al.

Setting: a sparse undirected unweighted with level of sparsity and the number and sizes of the clusters are allowed to be functions of n
Every cluster the same size k with at least ($\Omega(\sqrt{p})$; edge densities are uniformly p and q, for within and across clusters respectively
Aim: partitioning into disjoint clusters by penalizing missing edges within clusters, and present edges across clusters differently ⇒ Combinatorial optimisation problem
Result: algorithm which is a convex relaxation of the combinatorial optimisation problem; it is based on a sparse and low-rank matrix decomposition; It is stated that it is shown analytically that the algprithm outperforms all existing methods and that a simulation study confirms those results
Comments: The theorem that quaranties that the number of errors is small is not proven; Efficiency of algorithm shown via simultion study

1.7 A Tensor Spectral Approach to Learning Mixed Membership Community Models Anandkumar et al. (2013)

- overlapping communities
- the community memberships are drawn from a Dirichlet distribution
- learning these models via a tensor spectral decomposition method
- provide guarantees for our approach under a set of sufficient conditions
- compare with stochastic blockmodel

- Requirements: require n to be large enough compared to the number of communities k, and for the separation p-q to be large enough, so that the learning method can distinguish the different communities.
- sparse community memberships
- Zero error guarantee for block models

1.8 Stochastic Blockmodels with Growing Number of Classes (Airoldi et al., 2011)

Aim: The fraction of misclssified network nodes converges in probability to zero

outline

1. compare the log-likelihood of the observed data maximised over the $\theta_{z_i z_j}$ (probability according to group assignment) with its expectation, given z
2. $$\max_z \left| L(A; z) - \bar{L}_P(z) \right| = o_P(M)$$
   Until here it is assumed that there is no special block structure present.
3. 1.condition: for all Blockmodel classes $a = 1, \cdots, K$, class size N_a grows as $\min_a \{N_a\} = \Omega\left(\frac{N}{K}\right)$
   2.condition: any two rows of θ differ in at least one entry by an amount that is bounded by $\Omega\left(\frac{MK}{N^2}\right)$
   3.condition: $\frac{1}{N^2} \leq P_{ij} \leq 1 - \frac{1}{N^2}$
   4.condition: $K = \mathcal{O}(N^{\frac{1}{2}})$
   5.condition: $M = \omega(N(\log N)^{3+\delta})$ (expected number of edges)
   $\Leftrightarrow$ average degree $\frac{2M}{N}$ grows faster than $(\log N)^{3+\delta}$
   Then, $N_e(\hat{z}) = o_P(N) \quad \Rightarrow \quad \lim_{n\to\infty} \frac{N_e}{N} = 0$

1.9 Comparison between Rohe et al. (2011) and Airoldi et al. (2011)

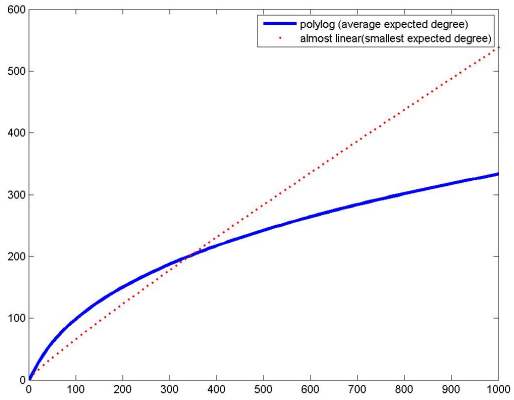
Aim: Both aim to formulate a theorem about consistency for estimation of the class assignment.

But:

- different estimation techniques
  - spectral clustering
  - maximum-likelihood estimation
- different definitions for misclustered
  - $\|c_i - z_i \mu O\|_2 \geq \frac{1}{\sqrt{2P}}$
  - "[...] counted for every node whose true class under $\bar{z}$ is not in the majority within its estimated class under $\hat{z}$"

$\Rightarrow$ Both could be the same

- different assumptions
  degree growth
  - the smallest expected degree grows almost linear in n, i.e. $\tau_n^2 > \frac{2}{\log n}$ with $\tau_n = \min_{i=1,\cdots,n} \frac{\mathscr{D}_{ii}^{(n)}}{n}$
    thus, $\min_{i=1,\cdots,n} \mathscr{D}_{ii} > \sqrt{\frac{2}{\log n}} n$
    $\Rightarrow$ Is twice used in the proof. Once in a technical manner to proof Theorem 2.1. The second time, the assumption is needed to make sure that the eigenvalues of L and $\mathscr{L}$ are close together, so that the same number of eigenvalues of L and $\mathscr{L}$ intersect with I_n. It results that X and $\mathscr{X}$ have the same dimensions what is necessary for the Davis-Kahan theorem.
  - the average degree ($\frac{2M}{N} = \frac{2\sum_{i<j} P_{ij}}{N} = \frac{\sum_i \mathscr{D}_{ii}}{N}$) grows faster than $(\log N)^{3+\delta}$



  number of classes k_n
  - Sequence of open intervals needs to be chosen, such that eventually the same amount of observed eigenvalues as of population eigenvalues intersect with the interval, and this number equals the number of classes to choose S_n you need k_n
  - k_n is assumed to be known

  additional assumptions
  - the smallest non-zero eigenvalue of $\mathscr{L}$ is not too small, i.e. $n^{-\frac{1}{2}}(\log n)^2 = \mathcal{O}(\lambda_{k_n}^2)$
  - four additional assumptions (see above)

- bounds

–

$$|\mathscr{M}| = o\left(\frac{P_n(\log n)^2}{\lambda_{k_n}^4 \tau_n^4 n}\right)$$

But P_n, as the number of the largest block in Z, as well as λ_{k_n}, as the smallest nonzero eigenvalue of $\mathscr{L}$ are not known.

– $(\sup_\theta L(A; z, \theta) - \sup_\theta \bar{L}(z, \theta))$ is bounded, but z is assumed to be given

! "It is not shown that spectral clustering is consistent under the Stochastic Blockmodel; it only gives a bound on the number of misclassified nodes."

Whereas Airoldi et al. (2011) show that the fraction $\frac{N_e}{N}$ of misclassified nodes goes to zero in N.

1.10 Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges (Oliveira, 2010)

aim: to prove that the adjacency matrix and the Laplacian of the random graph are concentrated around the corresponding matrices of the weighted graph whose edge weights are the probabilities in the random model.

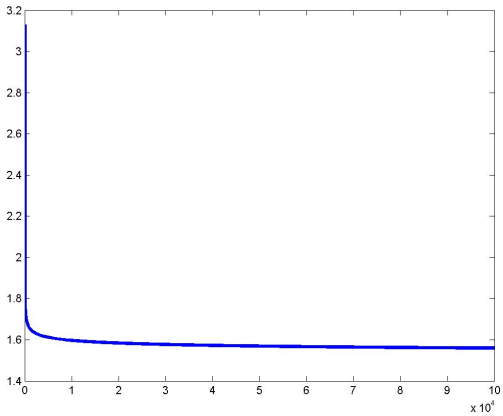
Theorem (1.1)
For any constant $c > 0$ there exists another constant $C = C(c) > 0$, independent of n or p, such that the following holds. Let $d \equiv \min_{i\in[n]} d_{G_p^{typ}}(i), \Delta \equiv max_{i\in[n]} d_{G_p^{typ}}(i)$. If $\Delta > C \ln n$, then for all $n^{-c} \leq \delta \leq \frac{1}{2}$,

$$P\left(\left\|A_p - A_p^{typ}\right\| \leq 4\sqrt{\Delta \ln(n/\delta)}\right) \geq 1 - \delta.$$

Moreover, if $d \geq C \ln n$, then for the same range of δ:

$$P\left(\left\|\mathscr{L}_p - \mathscr{L}_p^{typ}\right\| \leq 14\sqrt{\frac{\ln(4n/\delta)}{d}}\right) \geq 1 - \delta. \tag{1.1}$$

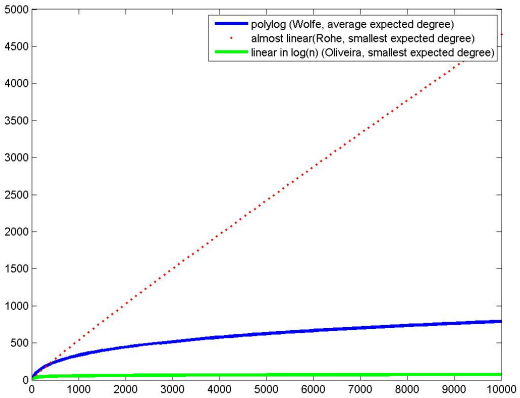
In the figure below it is shown how the RHS of equation (1.1) changes with growing n. It illustrates that the difference not necessarily converges to zero. ($n \in [2, 100000], C = 100, \delta = 0.25$)



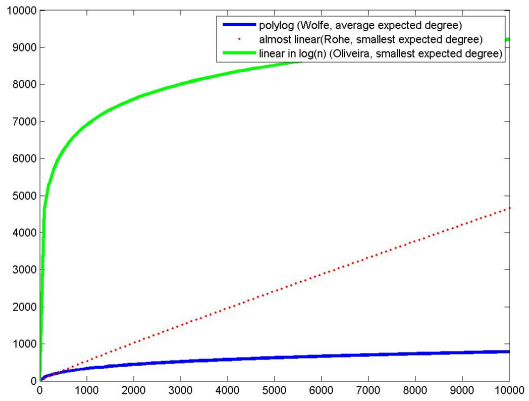
About the assumptions

The assumptions $\Delta > C \ln n, d \geq C \ln n, n^{-c} \leq \delta \leq \frac{1}{2}$ are needed to use the Freedman's inequality which is a generalization of the Berstein inequality.

For $n \in [0, 10000]$, $\delta = 0.005$ and $C = 8$.



For $n \in [0, 10000]$, $\delta = 0.005$ and $C = 1000$.



1.11 Consistency of spectral clustering (von Luxburg et al., 2008)

1.12 Communities in Networks (Porter et al., 2009)

Aim: compare community-finding algorithms while contrasting their different perspectives and revealing a few important similarities

1.13 A Spectral Algorithm for Learning Hidden Markov Models (Hsu et al., 2012)

Hsu et al. (2012) use the singular value decomposition to compute the left singular vectors U of $P_{2,1}$. U is used to change the parametrization of the hidden markov model to an observable representation. It results that their predictors for the joint probability of a sequence, as well as for a conditional probaility of an event, conditioned on all previous events only depend on observable values and not on the hidden states any longer. They state an algorithm to estimate the new parameterization of the HMM and proof that with high probaility

$$\sum_{x_1,\cdots,x_t} |Pr[x_1,\cdots,x_t] - \widehat{Pr}[x_1,\cdots,x_t]| \leq \epsilon$$

for the joint density of all sequences of length t, as well as

$$KL(Pr[x_t|x_1\cdots,x_{t-1}]\|\widehat{Pr}[x_t|x_1\cdots,x_{t-1}]) \leq \epsilon$$

for the conditional distribution.
⇒does not depend on the number of observations
⇒comment on computational cost

1.13.1 Evolution of Spatial Networks by Marc Barthelemy

He described in his talk at November, 29th 2012 the evolution of two different spartial networks over "time". Time is enquoted because to use time as a meassure itself you would need to adjust for a lot of side effects occuring at special time points, the world wars for instance. Therefore, the time is meassured in numbers of nodes. Network analysis is a topic already highly discussed in the 60th but recently the computational power and the abbility to observe and save hugh datasets improved.

Case 1 was about the road network in a city in Italy. They have data at 7 time points covering 200 years. Due to the huge size of the dataset it was necessary to develop a few parameters to capture the structure of the network, e.g. for the meassure of organisedness of a city they calculated

$$r_N = \frac{N(1)\cdot N(3)}{\sum_{k\neq 2} N(k)}$$

where $N(j)$ gives the number of nodes which have degree j. Thus, r_N computes the fraction between the number of nodes which have only one or three edges and the number of edges (two are not possible since a node is a crossing).
The second case study about subway systems. Here the first task was as well to find characteristics that summarize the model of the network. They discovered that large subway networks all have a similar structure. In this case they invented other meassures than for the road case. I asked and they did not try yet if the summarizing meassures of the one case apply also for the other. But he said he doubt it but it would be worth trying.

Marc Barthelemy; statistical physicist; www.guanturb.com; Marc.Barthelemy@cea.fra

1.13.2 A Spectral Clustering Approach To Finding Communities in Graphs (White and Smyth, 2005)

- Clustering nodes
- automatic selection of the number of clusters
- Experimental results indicate that the new algorithms are efcient and efective
- spectral algorithms are much faster for large sparse graphs

1.13.3 A Brief History of Statistical Models for Network Analysis and Open Challenges ?

This article gives a short review of some statitstical models for network analysis withou mentioning anything about the estimatore or thir propeties.

1.13.4 Co-clustering separately exchangeable network data? (highly interesting)

- aim: establish the performance of stochastic blockmodels for the co-clustering problem (partioning a binary array into subsets) for data generated by a nonparametic process
- assumption: data shall separate exchangeability
- quantified by: the maximum profile likelihood estimate asymptotically minimizes a Kullback-Leibner divergence risk funktional, comparing class of co-blockmodels and the non-parametric generatic process
  gives rates of mean squared error minimization
- interpretation: blockmodel can be interpreted as an optimal picewise-constant approximation to the genreative nonparametric model

- addition: for large sample sizes detection of co-clusters in data indicates with high probability similar co-clusters in the generative process

2 Maximum likelihood estimation

2.1 Null models for network data (Wolfe and Perry, 2012)

The paper Wolfe and Perry (2012) states a class of null models

$$\mathscr{M} : \log p_{ij} = \alpha_i + \alpha_j + \epsilon(\alpha_i, \alpha_j)$$

with a few assumptions for ϵ and in addition a sparse graph is aasumed, then all Models of this class lead roughly the same maximum likelihood estimates of the edge probabilities.
The proof is fairly technical.

2.2 Exercises about MLE:

Exercise 1.1
Assume , i.e.

$$X_{ij} \overset{iid}{\sim} \text{Bernoulli}(p_{ij}), i < j.$$

It follows

$$d_i = \sum_{i \neq j} X_{ij}$$
$$\mathbb{E}(d_i) = \mathbb{E}(\sum_{i \neq j} X_{ij}) = \sum_{i \neq j} \mathbb{E}(X_{ij}) = \sum_{i \neq j} p_{ij}$$

Let us further consider a Erdös Rényi model, e.g. $p_{ij} = p, \forall i, j$, where $p = \frac{c}{n}$.

$$\mathbb{E}(d_i) = \sum_{i \neq j} p_{ij} = (n-1) \cdot p$$
$$P(d_i = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

Theorem (2.1) Poisson limit theorem
If $n \to \infty, p \to 0$, while $np \to \lambda$, constant, then the Binomial(n, p) distribution approaches the Poisson distribution with expected value λ, i.e.

$$f_{Binom}(k, n, p) = \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k} \quad \overset{n \to \infty}{\to} \quad f_{Pois}(k, \lambda) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

Let us apply the Poisson limit theorem (2.1) to the Erdös Rényi case, where $p = \frac{c}{n}$. Since $p = \frac{c}{n} \overset{n\to\infty}{\to} 0$ and $np = c$, we obtain

$$P(d_i = k) \quad \overset{n\to\infty}{\to} \quad e^{-c}\frac{c^k}{k!}.$$

Exercise 1.2
Let us now consider a Erdös Rényi model, e.g. $p_{ij} = p, \forall i, j$, where $p = \frac{c\log(n)}{n}$.
It follows imediately for the average degree

$$\mathbb{E}(d_i) = (n-1)p = \frac{n-1}{n}c\log(n) \quad \overset{n\to\infty}{\to} \quad c\log(n)$$

Theorem (2.2) De MoivreLaplace
As n grows large, for k in the neighborhood of np we can approximate

$$\binom{n}{k} p^k q^{n-k} \simeq \frac{1}{\sqrt{2\pi npq}}\, e^{-(k-np)^2/(2npq)}, \quad p+q=1,\ p>0,\ q>0$$

in the sense that the ratio of the left-hand side to the right-hand side converges to 1 as $n \to \infty$.

Thus, if the degree $d_i = k$ is close to $np = c\log(n)$, de MoivreLaplace theorem can be applied and it holds

$$\lim_{n\to\infty} \mathbb{P}(d_i = k) = \frac{1}{\sqrt{2\pi npq}}\, e^{-(k-np)^2/(2npq)}.$$

Theorem (2.3) Central Limit theorem
Suppose $\{X_i\}_{i\in\mathbb{N}}$ is a sequence of i.i.d. random variables with expectation $\mathbb{E}(X_i) = \mu$ and variance $Var(X) = \sigma^2 < \infty$.
Then as n approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $N(0, \sigma^2)$:

$$\sqrt{n}\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu\right) \ \overset{d}{\to}\ \mathcal{N}(0,\ \sigma^2).$$

Exercise 2
The next section is based on Wolfe and Perry (2012).

2.3 MLE for Bernoulli distributed nodes

Assume a simple, undirected graph model with independent Bernoulli distributed edges, i.e.

$$X_{ij} \overset{iid}{\sim} \text{Bernoulli}(p_{ij}).$$

Thus, the probability mass function of each X_{ij} is given as

$$P(X_{ij}) = p_{ij}^{X_{ij}}(1-p_{ij})^{1-X_{ij}}.$$

Since X_{ij} for all i and j are stochastically independent, it follows for the log-likelihood

$$\begin{aligned} l_{Bernoulli}(p) &= \log\left(\prod_{i<j} p_{ij}^{X_{ij}}(1-p_{ij})^{1-X_{ij}}\right) \\ &= \sum_{i<j} X_{ij}\log p_{ij} + (1-X_{ij})\log(1-p_{ij}) \end{aligned} \tag{2.1}$$

This likelihood (2.1) is for small p_{ij} roughly the same as if X_{ij} were treated as Poisson distributed

$$\begin{aligned} l_{Pois}(p) &= \log\left(\prod_{i<j} \frac{p_{ij}^{X_{ij}}}{X_{ij}!}\exp(-p_{ij})\right) \\ &= \sum_{i<j} X_{ij}\log p_{ij} - p_{ij} - \log(X_{ij}!) \end{aligned} \tag{2.2}$$

Since $X_{ij} \in \{0,1\}$ it holds $\log(X_{ij}!) = 0$. Furthermore for small p_{ij}, $p_{ij} \approx 0$ and $\log(1-p_{ij}) \approx 0$. Thus,

$$l_{Pois}(p) \approx l_{Bernoulli}(p)$$

Let us assume a log linear model such as

$$\mathscr{M}_{\log} : \log(p_{ij}) = \alpha_i + \alpha_j. \tag{2.3}$$

Substituting this parameterization in (2.9), we obtain

$$\begin{aligned} l_{Pois}(p) &= \sum_{i<j} X_{ij}(\alpha_i+\alpha_j) - \exp(\alpha_i+\alpha_j) \\ &= \sum_{i<j} X_{ij}\alpha_i + \sum_{i<j} X_{ij}\alpha_j - \sum_{i<j}\exp(\alpha_i+\alpha_j) \end{aligned} \tag{2.4}$$

X is symmetric. It follows

$$\begin{aligned}
&= \sum_{i<j} X_{ij}\alpha_i + \sum_{i<j} X_{ji}\alpha_j - \sum_{i<j} \exp(\alpha_i + \alpha_j) \\
&\overset{X_{ii}=0}{=} \sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij}\alpha_i - \sum_{i<j} \exp(\alpha_i + \alpha_j) \\
&= \sum_{i=1}^{n} \alpha_i X_{i+} - \sum_{i<j} \exp(\alpha_i + \alpha_j)
\end{aligned}$$

For a maximum it is a necessary requirement that $\nabla l_{Pois}(\alpha) = 0$. ($\sum_{i<j}$ changes to $\sum_{i\neq j}$ because we derive the derivative in x_i. Thus, i is fixed in the equation.)

$$\frac{\partial l_{Pois}}{\partial \alpha_i} = X_{i+} - \sum_{i\neq j} \exp(\alpha_i + \alpha_j) \qquad \text{for all } i = 1, \cdots, n \tag{2.5}$$

For the estimator $\tilde{\alpha}_i = \log X_{i+} - \log \sqrt{X_{++}}$ ($\Leftrightarrow \tilde{p}_{ij} = \frac{X_{i+}X_{j+}}{X_{++}}$) it would follow

$$\begin{aligned}
\frac{\partial l_{Pois}}{\partial \alpha_i} &= X_{i+} - \sum_{i\neq j} \exp(\log(\frac{X_{i+}}{\sqrt{X_{++}}}) + \log(\frac{X_{j+}}{\sqrt{X_{++}}})) \qquad \text{for all } i = 1, \cdots, n \\
\frac{\partial l_{Pois}}{\partial \alpha_i} &= X_{i+} - \sum_{i\neq j} \exp(\log(\frac{X_{i+}}{\sqrt{X_{++}}} \cdot \frac{X_{j+}}{\sqrt{X_{++}}})) \qquad \text{for all } i = 1, \cdots, n \\
\frac{\partial l_{Pois}}{\partial \alpha_i} &= X_{i+} - \sum_{i\neq j} \frac{X_{i+}X_{j+}}{X_{++}} \qquad \text{for all } i = 1, \cdots, n \\
\frac{\partial l_{Pois}}{\partial \alpha_i} &= X_{i+} - \sum_{i\neq j} \frac{X_{i+}X_{j+}}{X_{++}} \qquad \text{for all } i = 1, \cdots, n \\
\frac{\partial l_{Pois}}{\partial \alpha_i} &= X_{i+} - \frac{X_{i+}}{X_{++}} \underbrace{\sum_{i\neq j} X_{j+}}_{=X_{++}-X_{i+}} \qquad \text{for all } i = 1, \cdots, n \\
&= X_{i+} - X_{i+}\left(1 - \frac{X_{i+}}{X_{++}}\right) \qquad \text{for all } i = 1, \cdots, n \\
&= \frac{X_{i+}^2}{X_{++}} \qquad \text{for all } i = 1, \cdots, n
\end{aligned}$$

So that $\frac{\partial l_{Pois}}{\partial \alpha_i}$ tends to 0 it is necessary that $\frac{X_{i+}^2}{X_{++}}$ gets small.

Why not calculate MLE for Bernoulli destributed random variables directly?

Let us assume a log linear model such as

$$\mathscr{M}_{\log} : \log(p_{ij}) = \alpha_i + \alpha_j. \tag{2.6}$$

Substituting this parameterization in (2.1), we obtain

$$\begin{aligned} l_{Bernoulli}(p) &= \sum_{i<j} X_{ij}(\alpha_i + \alpha_j) + (1 - X_{ij}) \log(1 - \exp(\alpha_i + \alpha_j)) \\ &\overset{s.a.}{=} \sum_{i=1}^{n} \alpha_i X_{i+} + \sum_{i<j} (1 - X_{ij}) \log(1 - \exp(\alpha_i + \alpha_j)) \\ &= \sum_{i=1}^{n} \alpha_i X_{i+} + \sum_{i<j} \log(1 - \exp(\alpha_i + \alpha_j)) - \sum_{i<j} X_{ij} \log(1 - \exp(\alpha_i + \alpha_j)) \end{aligned}$$

2.4 MLE for Poisson distributed nodes

Assume a simple, undirected graph model with independent Poisson distributed edges, i.e.

$$X_{ij} \overset{iid}{\sim} \text{Poisson}(\lambda_{ij}).$$

Thus, the probability mass function of each X_{ij} is given as

$$P(X_{ij}) = \frac{\lambda_{ij}^{X_{ij}}}{X_{ij}!} \exp(-\lambda_{ij}).$$

Since X_{ij} for all i and j are stochastically independent, it follows for the log-likelihood

$$\begin{aligned} l_{Pois}(p) &= \log \left(\prod_{i<j} \frac{\lambda_{ij}^{X_{ij}}}{X_{ij}!} \exp(-\lambda_{ij}) \right) \\ &= \sum_{i<j} X_{ij} \log \lambda_{ij} - \lambda_{ij} - \log(X_{ij}!) \end{aligned} \tag{2.7}$$

Let us assume a log linear model such as

$$\mathscr{M}_{\log} : \log(\lambda_{ij}) = \alpha_i + \alpha_j. \tag{2.8}$$

Substituting this parameterization in (2.9), we obtain

$$\begin{aligned} l_{Pois}(p) &= \sum_{i<j} X_{ij}(\alpha_i + \alpha_j) - \exp(\alpha_i + \alpha_j) - \sum_{i<j} \log(X_{ij}!) \\ &= \sum_{i<j} X_{ij}\alpha_i + \sum_{i<j} X_{ij}\alpha_j - \sum_{i<j} \exp(\alpha_i + \alpha_j) - \sum_{i<j} \log(X_{ij}!) \end{aligned} \tag{2.9}$$

X is symmetric. It follows

$$\begin{aligned}
&= \sum_{i<j} X_{ij}\alpha_i + \sum_{i<j} X_{ji}\alpha_j - \sum_{i<j} \exp(\alpha_i + \alpha_j) - \sum_{i<j} \log(X_{ij}!) \\
&\overset{X_{ii}=0}{=} \sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij}\alpha_i - \sum_{i<j} \exp(\alpha_i + \alpha_j) - \sum_{i<j} \log(X_{ij}!) \\
&= \sum_{i=1}^{n} \alpha_i X_{i+} - \sum_{i<j} \exp(\alpha_i + \alpha_j) - \sum_{i<j} \log(X_{ij}!)
\end{aligned}$$

For a maximum it is a necessary requirement that $\nabla l_{Pois}(\alpha) = 0$. ($\sum_{i<j}$ changes to $\sum_{i\neq j}$ because we derive the derivative in x_i. Thus, i is fixed in the equation.)

$$\begin{aligned}
\frac{\partial l_{Pois}}{\partial \alpha_i} &= X_{i+} - \sum_{i\neq j} \exp(\alpha_i + \alpha_j) && \text{for all } i = 1, \cdots, n \\
\frac{\partial l_{Pois}}{\partial \alpha_i} &= X_{i+} - \sum_{i\neq j} \exp(\alpha_i) \cdot \exp(\alpha_j) && \text{for all } i = 1, \cdots, n \\
\frac{\partial l_{Pois}}{\partial \alpha_i} &= X_{i+} - \exp(\alpha_i) \sum_{i\neq j} \exp(\alpha_j) && \text{for all } i = 1, \cdots, n \\
\overset{\frac{\partial l_{Pois}}{\partial \alpha_i}=0}{\Rightarrow} \quad \exp(\hat{\alpha}_i) &= \frac{X_{i+}}{\sum_{i\neq j} \exp(\alpha_j)} && \text{for all } i = 1, \cdots, n \\
\hat{\alpha}_i &= \log\left(\frac{X_{i+}}{\sum_{i\neq j} \exp(\alpha_j)}\right) && \text{for all } i = 1, \cdots, n
\end{aligned} \tag{2.10}$$

According to the log linear model (2.8) is the equation (2.10) equivalent to

$$\begin{aligned}
\hat{p}_{ij} &= \exp(\hat{\alpha}_i + \hat{\alpha}_j) \\
\hat{p}_{ij} &= \exp\left(\log\left(\frac{X_{i+}}{\sum_{i\neq l} \exp(\alpha_l)}\right) + \log\left(\frac{X_{j+}}{\sum_{j\neq k} \exp(\alpha_k)}\right)\right) \\
\hat{p}_{ij} &= \exp\left(\log\left(\frac{X_{i+}X_{j+}}{\sum_{i\neq l} \exp(\alpha_l) \sum_{j\neq k} \exp(\alpha_k)}\right)\right) \\
\hat{p}_{ij} &= \frac{X_{i+}X_{j+}}{\sum_{i\neq l} \exp(\alpha_l) \sum_{j\neq k} \exp(\alpha_k)}
\end{aligned}$$

2.5 MLE for weighted graphs

$$X_{ij} \overset{iid}{\sim} \text{Exponential}(\lambda_{ij}).$$

Thus, the probability mass function of each X_{ij} is given as

$$P(X_{ij}) = \begin{cases} \lambda_{ij} \exp(-\lambda_{ij} X_{ij}) & \text{for } X_{ij} \geq 0 \\ 0, & \text{else.} \end{cases} \tag{2.11}$$

1.case: all $X_{ij} \geq 0$

Since X_{ij} for all i and j are stochastically independent, it follows for the log-likelihood

$$\begin{aligned} l_{Expo}(p) &= \log \left(\prod_{i<j} \lambda_{ij} \exp(-\lambda_{ij} X_{ij}) \right) \\ &= \sum_{i<j} \log(\lambda_{ij}) - \lambda_{ij} X_{ij} \end{aligned} \tag{2.12}$$

Let us assume a log linear model such as

$$\mathscr{M}_{\log} : \log(\lambda_{ij}) = \alpha_i + \alpha_j. \tag{2.13}$$

Substituting this parameterization in (2.12), we obtain

$$\begin{aligned} l_{Expo}(\alpha) &= \sum_{i<j} (\alpha_i + \alpha_j) - \exp(\alpha_i + \alpha_j) X_{ij} \\ &= \sum_{i<j} (\alpha_i + \alpha_j) - \sum_{i<j} \exp(\alpha_i + \alpha_j) X_{ij} \end{aligned}$$

For a maximum it is a necessary requirement that $\nabla l_{Expo}(\hat{\alpha}) = 0$.

$$\begin{aligned} \frac{\partial l_{Pois}}{\partial \alpha_i} &= n - \sum_{j=1}^{n} \exp(\alpha_i + \alpha_j) X_{ij} \\ &= n - \exp(\alpha_i) \sum_{j=1}^{n} \exp(\alpha_j) X_{ij} \end{aligned}$$

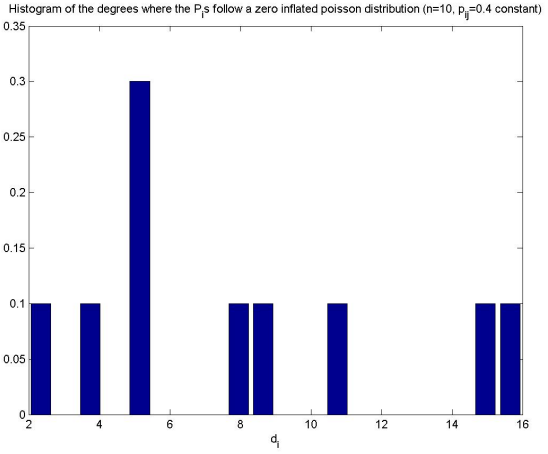
$$\begin{aligned} \overset{\nabla l_{Expo}(\hat{\alpha})=0}{\Rightarrow} \quad & \exp(\hat{\alpha}_i) = \frac{n}{\sum_{j=1}^{n} \exp(\alpha_j) X_{ij}} \\ \Leftrightarrow \quad & \hat{\alpha}_i = \log \left(\frac{n}{\sum_{j=1}^{n} \exp(\alpha_j) X_{ij}} \right) \\ & \hat{\lambda}_{ij} = \exp(\alpha_i + \alpha_j) \\ \Leftrightarrow \quad & \hat{\lambda}_{ij} = \exp \left(\log \left(\frac{n}{\sum_{l=1}^{n} \exp(\alpha_l) X_{il}} \right) + \log \left(\frac{n}{\sum_{k=1}^{n} \exp(\alpha_k) X_{jk}} \right) \right) \\ \Leftrightarrow \quad & \hat{\lambda}_{ij} = \exp \left(\log \left(\frac{n}{\sum_{l=1}^{n} \exp(\alpha_l) X_{il}} \right) \right) \exp \left(\log \left(\frac{n}{\sum_{k=1}^{n} \exp(\alpha_k) X_{jk}} \right) \right) \end{aligned}$$

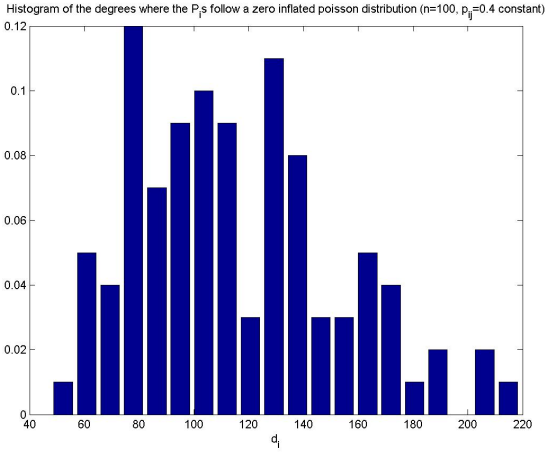
$$\Leftrightarrow \qquad \hat{\lambda}_{ij} = \frac{n^2}{\sum_{l=1}^{n} \exp(\alpha_l) X_{il} \cdot \sum_{k=1}^{n} \exp(\alpha_k) X_{jk}}$$

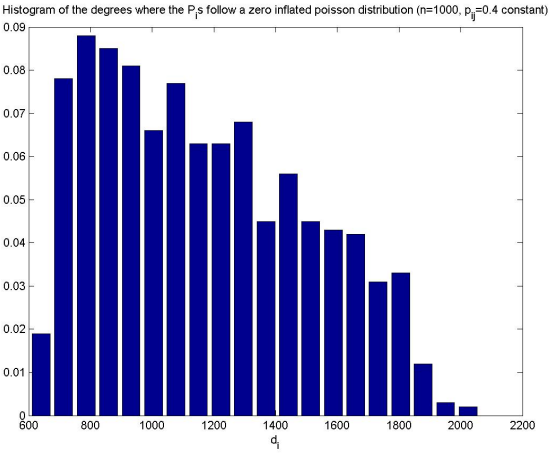
2.5.1 Zero inflated Poisson distribution for edges

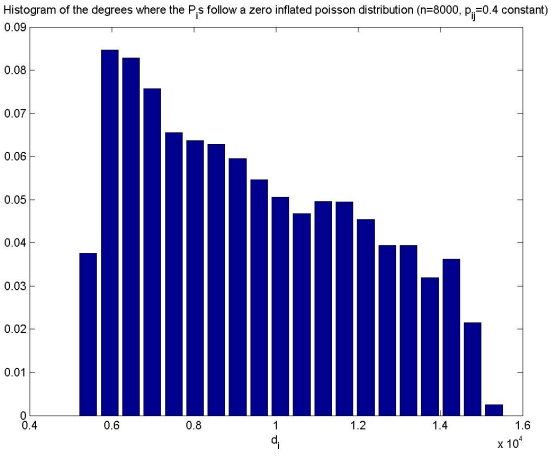
$$\begin{aligned} & & A_{ij} &\sim \text{Bernoulli}(p_{ij}) \\ X_{ij} &\sim \text{Poisson}(\lambda_{ij}) & \text{with} \quad \lambda_{ij} &= \exp(\alpha_i + \alpha_j) \\ & & Y_{ij} &= A_{ij} \cdot X_{ij} \end{aligned}$$

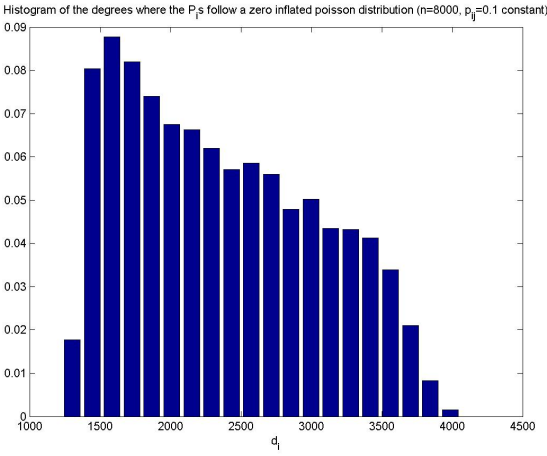
Simulation

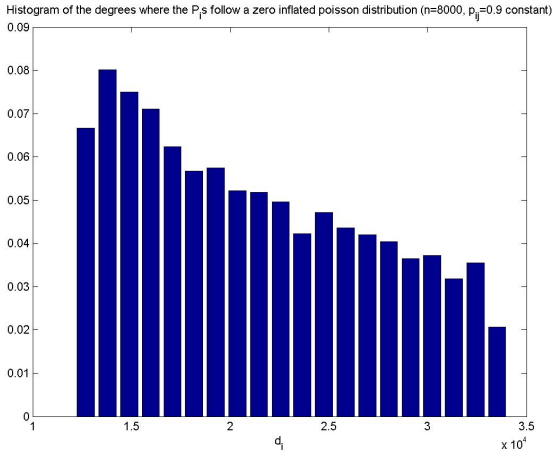












2.5.2 Random graphs with a given degree sequence (Chatterjee et al., 2011)

- chose graphs uniformily from a set of graphs of a given degree sequence
- derive graph limits
- other characyeristics can be determine, e.g. number of triangles
- example: exponential model having the degree sequence as sufficient statistic
- mle of parameters is then unique and consistent
- algorithm is derived
- prove graph limit theorem

2.5.3 Link prediction for partially observed networks Zhao et al.

The main idea of the paper that the probability of an edge is Bernoulli but that there is an additional uncertainty due to missing observations and wrong observed data seems to me very realistic. Their estimator overcomes this issue by using additional informations about the similarity of nodes, e.g. due to node covariates. During minimizing the difference between the observed network and the estimators they then penalize in addition differences in additional information.

I find the approach really appealing but I was missing any proof for properties as for consistency.

2.6 Comments

1. measure of the difference between two probability distributions P and Q
   - total variation

$$d(\mu, \nu) = \sup \{ \, |\mu(A) - \nu(A)| : A \in \Sigma \, \}$$

   - KullbackLeibler divergence (non-symmetric measure, is a measure of the information lost when Q is used to approximate P)
     the KL divergence of Q from P is defined to be

$$D_{\mathrm{KL}}(P\|Q) = \sum_i \ln \left(\frac{P(i)}{Q(i)} \right) P(i)$$

2. *Stein's method* is a general method in probability theory to obtain bounds on the distance between two probability distributions with respect to a probability metric.

3. Taylor series in random variables, especially, if interested in the error term, see (Brockwell and Davis, 1991, p. 201).

2.6.1 insides from the likelihood and assymptotics lecture

Literature

- approximations for the likelihood: Edgeworth, saddlepoint; Laplace for intergrals
  ⇒ Luigi and Salvan (1997)

- profile likelihood
  ⇒ Cox and Barndorff-Nielsen (1994); Davison (2008)

2.7 Statistical Inference

Frequentists

- probability and statistical background
- poit and set estimators
- goodness of fit tests (p-value)

Bayes

- Bayesian inference
- decision thoery

2.8 Statistical computing

- matrix computation: Choleski- (S=RtR; R upper triangular), Eigenvalue-, Singular-value-, QR- (S=QR; R upper triangular, Q orthogonal) decomposition
- Optimization: Taylor, steepest descent, newtons's method, quasi-newtons method, nelder mead polynom method
- calculus by computer:
- random number generators

2.8.1 Taylor series in several variables

The Taylor series may also be generalized to functions of more than one variable with

$$T(x_1, \ldots, x_d) = \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} \cdots \sum_{n_d=0}^{\infty} \frac{(x_1 - a_1)^{n_1} \cdots (x_d - a_d)^{n_d}}{n_1! \cdots n_d!}, \left(\frac{\partial^{n_1 + \cdots + n_d} f}{\partial x_1^{n_1} \cdots \partial x_d^{n_d}} \right) (a_1, \ldots, a_d).$$

For example, for a function that depends on two variables, x and y, the Taylor series to second order about the point (a, b) is:

$$\begin{aligned} f(x,y) \approx\ & f(a,b) + (x-a)\, f_x(a,b) + (y-b)\, f_y(a,b) \\ & + \frac{1}{2!} \left[(x-a)^2 f_{xx}(a,b) + 2(x-a)(y-b)\, f_{xy}(a,b) + (y-b)^2 f_{yy}(a,b) \right], \end{aligned}$$

where the subscripts denote the respective partial derivatives.
A second-order Taylor series expansion of a scalar-valued function of more than one variable can be written compactly as

$$T(\mathbf{x}) = f(\mathbf{a}) + \mathrm{D}f(\mathbf{a})^T (\mathbf{x} - \mathbf{a}) + \frac{1}{2!} (\mathbf{x} - \mathbf{a})^T \left\{ \mathrm{D}^2 f(\mathbf{a}) \right\} (\mathbf{x} - \mathbf{a}) + \cdots,$$

where $Df(\mathbf{a})$ is the gradient of f evaluated at $\mathbf{x} = \mathbf{a}$ and $D^2 f(\mathbf{a})$ is the Hessian matrix. Applying the multi-index notation the Taylor series for several variables becomes

$$T(\mathbf{x}) = \sum_{|\alpha| \geq 0} \frac{(\mathbf{x} - \mathbf{a})^\alpha}{\alpha!} (\partial^\alpha f)(\mathbf{a}) \,,$$

which is to be understood as a still more abbreviated multi-index version of the first equation of this paragraph, again in full analogy to the single variable case.

2.9 Books

2.9.1 Random Graphs of Bollobas

- pretty theoretical
- a wide range of properties for sets of graphs, subsets, etc.

2.10 Poisson-Binomial Distribution

2.10.1 Poisson approximation (Barbour et al., 1992)

The first ten pages give useful results if the poisson-binomial distribution should be approximated by a poisson because it explains several bounds for the total variation distance between those two distribution.

2.10.2 The method of moments and degree distributions for network models (Bickel et al., 2011)

- general method of moments approach(non-parametric)
- to fit large class of probability models through empirical counts (including blockmodel)
- asymptotic properties, e.g. consistency
- degree distribution

Note: difficult to read

2.10.3 Size-dependent degree distribution of a scale-free growing network (Dorogovtsev et al., 2001)

- power law assumption for degrees; networks evolving in time

- they give an exact probability distrinbution for the node i to have the degree k at time t
- BUT the network grow accordingt to a fixed set of rules, i.e. at every time point one node is added which is connected to both ends of a randomly chosen edge

$\Rightarrow$ More restrictive; no analysis for $n \to \infty$

2.10.4 Entropy Bounds for Discrete Random Variables via Coupling (Sason, 2012)

The main result of the paper states an upper bound of the difference in *entropy* $H(X)$ between two discrete random variables X and Y both defined on the same set $\mathscr{A}$ whereby entropy is defined as

$$H(X) = \mathbb{E}(-\log(P(X))).$$

as

$$|H(X) - H(Y)| \leq d_{TV}(X, Y) \log(M \cdot \alpha - 1) + h(d_{TV}(X, Y))$$

with $h(p) = -p \log p - (1-p) \log(1-p)$ is the binary entropy function and $M = |\mathscr{A}|$. As an example they compare a Poisson-Binomial distribution with a Poisson distribution. Since the result only bounds an transformation of the ratio of their expectations I do not see an direct connection but in between the reasoning they use an upper bound for the total variation

$$d_{TV}(X, Y) \leq \left(\frac{1 - \exp(-\lambda)}{\lambda} \right) \sum_{i=1}^{n} p_i^2,$$

which might be of interest an some point.

2.10.5 Improved Lower Bounds on the Total Variation Distance and Relative Entropy for the Poisson Approximation (Sason, 2013)

In the paper Sason (2013) a new lower bound is derived for the total variation distance between a Poisson-Binomial distribution, i.e.

$$W = \sum X_i, \quad \text{with} \quad X_i \sim \text{Bern}(p_i)$$

and a Poisson distribution with

$$Y \sim \text{Pois}\left(\sum p_i \right).$$

The lower bound

$$K_1(\lambda)\sum_{i=1}^{n} p_i^2 \leq d_{TV}(X,Y)$$

is stated in two different forms. First, K_1 is defined such that optimization is required to compute it which only can be performed numerically. Secondly, Sason gives a closed-form for K_1 wherefore the bound has to be loosened.
To upper-bound the total variation the following inequality is used

$$d_{TV}(W,Y) \leq \left(\frac{1-\exp(-\lambda)}{\lambda}\right)\sum_{i=1}^{n} p_i^2.$$

The overall aim is to use this lower bound on the total variation to improve the lower bounds on relative entropy of random variables.

Let us compare those bounds with the bounds described in (Barbour et al., 1992) and used for lemma 2 in the supplementary material for the paper (Olhede and Wolfe, 2012). It should be mentioned that in (Sason, 2013) it is stated that the new lower bound is an improvedment compared to

$$\frac{1}{32}\left(1 \wedge \frac{1}{\lambda}\right)\sum_{i=1}^{n} p_i^2 \leq d_{TV}(X,Y).$$

In (Olhede and Wolfe, 2012) the following bounds are used

$$\frac{1}{32}\left(1 \wedge \mathbb{E}(d_i|\pi)\right)\left[1 - \frac{\mathrm{Var}(di|\pi)}{\mathbb{E}(d_i|\pi)}\right] \leq d_{TV}(X,Y) \leq \left(1 \wedge \mathbb{E}(d_i|\pi)\right)\left[1 - \frac{\mathrm{Var}(di|\pi)}{\mathbb{E}(d_i|\pi)}\right].$$

$$\begin{aligned}
&\frac{1}{32}\left(1 \wedge \mathbb{E}(d_i|\pi)\right)\left[1 - \frac{\mathrm{Var}(di|\pi)}{\mathbb{E}(d_i|\pi)}\right] \\
&= \frac{1}{32}\left(1 \wedge \mathbb{E}(d_i|\pi)\right)\left[1 - \frac{\mathbb{E}(d_i|\pi) - \sum p_{ij}^2}{\mathbb{E}(d_i|\pi)}\right] \\
&= \frac{1}{32}\left(1 \wedge \mathbb{E}(d_i|\pi)\right)\left[\frac{\sum p_{ij}^2}{\mathbb{E}(d_i|\pi)}\right] \\
&= \frac{1}{32}\left(1 \wedge \lambda\right)\left[\frac{\sum p_{ij}^2}{\lambda}\right] \\
&= \frac{1}{32}\left(\frac{1}{\lambda} \wedge 1\right)\left[\sum p_{ij}^2\right]
\end{aligned}$$

Thus, the lower bounds for the total variation in Sason (2013) should improve the lower bounds used in (Olhede and Wolfe, 2012).
A comparison in closed form of the new bounds and those used in (Olhede and Wolfe, 2012) is difficult because even the colsed form of the new lower bound is complicated. But figures 2.10.5 and 2.10.5 show for different values of λ in blue the fector $\frac{1}{32}\left(\frac{1}{\lambda} \wedge 1\right)$ used by (Olhede and Wolfe, 2012) and in red the factor K_1 used by Sason (2013). As it can be seen, the factor of Sason (2013) is higher. In conclusion, the lower bound will be more precise.

Let us compare the upper bounds.

$$\begin{aligned}
&(1 \wedge \mathbb{E}(d_i|\pi)) \left[1 - \frac{\mathrm{Var}(di|\pi)}{\mathbb{E}(d_i|\pi)}\right] \\
=&(1 \wedge \lambda) \left[\frac{\sum p_{ij}^2}{\lambda}\right] \\
=&(1 \wedge \lambda) \frac{1}{\lambda} \left[\sum p_{ij}^2\right]
\end{aligned}$$

It follows that the upper bound of Sason (2013) is $\frac{1-e^{-\lambda}}{1\wedge\lambda}$ times the upper bound of Olhede and Wolfe (2012). As it can be seen in figure 2.10.5, the factor $\frac{1-e^{-\lambda}}{1\wedge\lambda}$ (in red) is for some cases less than one and at most not greater than 1. Therefore, the upper bound of Sason is better than the bound used in Olhede and Wolfe (2012).

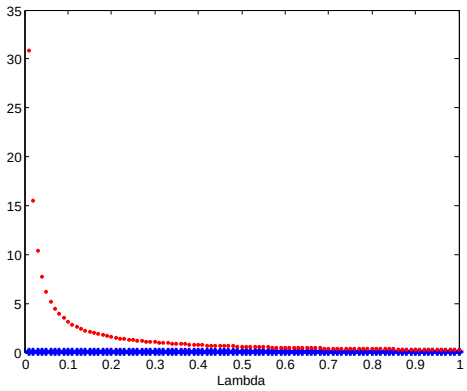


Figure 2.1: Comparison of the lower bound of Sason (2013) and Olhede and Wolfe (2012): In red, the factor used by Sason and in red the factor used in Olhede and Wolfe (2012)

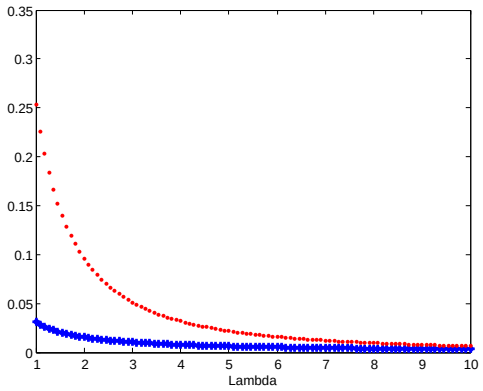


Figure 2.2: Comparison of the lower bound of Sason (2013) and Olhede and Wolfe (2012): In red, the factor used by Sason and in red the factor used in Olhede and Wolfe (2012)

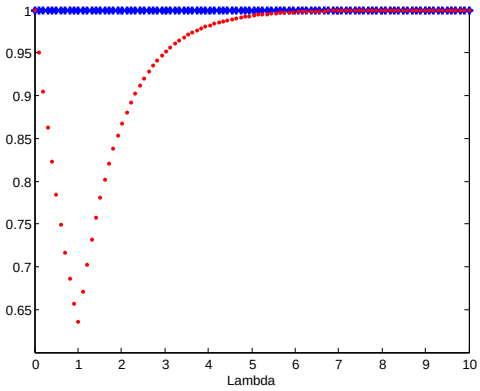


Figure 2.3: Comparison of the upperbound of Sason (2013) and Olhede and Wolfe (2012): In red , the factor by which the upper bound of Sason is less than the upper bound used in Olhede and Wolfe (2012)

Bibliography

EM Airoldi, PJ Wolfe, and DS Choi. Stochastic Blockmodels with Growing Number of Classes. *Biometrika*, 99(2):273–284, 2011. URL `http://www.stormingmedia.us/15/1587/A158755.html`.

Anima Anandkumar, Rong Ge, Daniel Hsu, and Sham M Kakade. A Tensor Spectral Approach to Learning Mixed Membership Community Models. 2013.

A.D. Barbour, L. Holst, and S. Janson. *Poisson approximation*. Oxford science publications. Clarendon Press, 1992.

Peter J. Bickel, Aiyou Chen, and Elizaveta Levina. The method of moments and degree distributions for network models. *Annals of Statistics*, 39(5):2280–2301, 2011.

P.J. Brockwell and R.A. Davis. *Time Series: Teory and Methods*. Springer Series in Statistics Series. Springer-Verlag, 1991. ISBN 9780387974293. URL `http://books.google.co.uk/books?id=ZW_ThhYQiXIC`.

Sourav Chatterjee, Persi Diaconis, and Allan Sly. Random graphs with a given degree sequence. *The Annals of Applied Probability*, 21(4):1400–1435, August 2011. ISSN 1050-5164. doi: 10.1214/10-AAP728. URL `http://projecteuclid.org/euclid.aoap/1312818840`.

Yudong Chen, Sujay Sanghavi, and Huan Xu. Clustering Sparse Graphs. pages 1–19.

Yudong Chen, Vikas Kawadia, and Rahul Urgaonkar. DETECTING OVERLAPPING TEMPORAL OMMUNITY STRUCTURE IN TIME -EVOLVING NETWORKS Structure in Time-Evolving Networks. *Bbnreport-, T Echnical R Eport*, 2013.

Fan Chung. Spectral Clustering of Graphs with General Degrees in the Extended Planted Partition Model. *Journal of Machine Learning Research*, 2012:1–23, 2012.

D.R. Cox and O.E. Barndorff-Nielsen. *Inference and Asymptotics*. Chapman and Hall/CRC, 3 1994. ISBN 9780412494406. URL `http://amazon.com/o/ASIN/041249440X/`.

A. C. Davison. *Statistical Models (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press, 1 edition, 6 2008. ISBN 9780521734493. URL http://amazon.com/o/ASIN/0521734495/.

S. Dorogovtsev, J. Mendes, and a. Samukhin. Size-dependent degree distribution of a scale-free growing network. *Physical Review E*, 63(6):062101, May 2001. ISSN 1063-651X. doi: 10.1103/PhysRevE.63.062101. URL http://link.aps.org/doi/10.1103/PhysRevE.63.062101.

Daniel Hsu, Sham M Kakade, and Tong Zhang. A Spectral Algorithm for Learning Hidden Markov Models. pages 1–30, 2012.

P. Luigi and A. Salvan. *Principles of Statistical Inference*. World Scientific Publishing Company, Singapore, New Jersey, London, Hong Kong, 4 edition, 1 1997. ISBN 9789810230661. URL http://amazon.com/o/ASIN/9810230664/.

BP Olding and PJ Wolfe. Inference for graphs and networks: extending classical tools to modern data. *arXiv preprint arXiv:0906.4980*, 2009. URL http://arxiv.org/abs/0906.4980.

Sofia C Olhede and Patrick J Wolfe. What is a Degree Distribution? *arXiv:1211.6537v1*, pages 1–18, 2012.

Roberto Imbuzeiro Oliveira. Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. page 46, November 2010. URL http://arxiv.org/abs/0911.0600.

Mason A Porter, Jukka-pekka Onnela, Peter J Mucha, and Josiah Willard Gibbs. Communities in Networks. 56(9):1082–1097, 2009.

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, August 2011. ISSN 0090-5364. doi: 10.1214/11-AOS887. URL http://projecteuclid.org/euclid.aos/1314190618.

Igal Sason. Entropy Bounds for Discrete Random Variables via Coupling. *arXiv:1209.5259v3*, (v):1–12, 2012.

Igal Sason. Improved Lower Bounds on the Total Variation Distance and Relative Entropy for the Poisson Approximation. *arXiv:1301.7504v1*, pages 1–4, 2013.

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007. URL http://www.springerlink.com/index/jq1g17785n783661.pdf.

Ulrike von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, 36(2):555–586, April 2008. URL http://projecteuclid.org/euclid.aos/1205420511.

Scott White and Padhraic Smyth. A Spectral Clustering Approach To Finding Communities in Graphs . *Data Mining: Proceedings: SIAM International Conference*, 5:274 –285, 2005.

Patrick J Wolfe and Patrick O Perry. Null models for network data. 2012.

Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Link prediction for partially observed networks. pages 1–15.