Henry Pao

August 2, 2010

Abstract

1 Introduction

We consider the classification of (labeled) graphs. A random graph $G = (V, E)$, with $V = [n] = \{1, 2, \dots n\}$. These graphs are simple direct graphs with loops. Thus the adjacency matrix has n^2 entries of interest.

Consider $\{(G_i, Y_i)\}_{i=1}^{s} \overset{iid}{\sim} F_{GY}$, with class labels $Y : \Omega \to \{0, 1\}$ and graphs $G : \Omega \to \mathcal{G}_n$, where $\mathcal{G}$ denotes the collection of simple directed graphs with loops. For simplicity, we assume that the prior probability of class membership $\pi = P[Y = 1]$ is known to be 1/2. and the sample sies $S_y = \sum_{i=1}^{s} I\{Y = y\}$ are fixed. Thus $s_0 = s_1 = s/2$. We consider the independent edge model (IE), so that for $y \in \{0, 1\}$ the class-conditional distribution $F_{G|Y=y}$ is parameterized by an $n \times n$ matrix with entries $p_{y;u,v} \in [0, 1]$.

1.1 IE1-star graph

In order to study the importance of coherence, we designed a special graph whose coherence is easy to take advantag of. Except for m edges this graph has the same distribution as an Erdos-Renyi graph. To chose these m edges first a vertex v^* is uniformly chosen out of the n verticies. Now m edges out of $2n - 1$ edges containing this vertex are chosen to have an edge probability of q.

1.2 classifier

The Bayes optimal classifier for obseved graph G is

$$g^*(G) = \arg\max_y \prod_{(u,v) \in V \times V} f(a_u, v; p_{y;u,v}),$$

where the Bernoulli probability $f(a; p) = pI\{a = 1\} + (i - 1)I\{a = 0\}$.

The independent edge homogeneous vs inhomogeneous model (IE1), parameterized by n, p, q, and $\mathcal{E} \subset V \times V$, is given by $p_{0;u,v} = p$ for all $(u, v) \in V \times V$ and $p_{1;u,v} = q$ for all $(u, v) \in \mathcal{E}, p_{1;u,v} = p$ for all $(u, v) \in (V \times V) \backslash \mathcal{E}$; $\mathcal{E}$ is the collection of signal edges and $\mathcal{E}^c = (V \times V) \backslash \mathcal{E}$ is the collection of noise edges. (Notice that $F_{G|Y} = 0$ is Erdos-Renyi ER(n, p).) In this model, all signal edges are created equally, and all noise edges are created equally; we will see that this property simplifies our analysis.

In IE1, only $\mathcal{E}$ is relevant and g can be written as

$$g^*(G) = \arg\max_y \prod_{(u,v) \in \mathcal{E}} f(a_{u,v}; p_{y;u,v}).$$

If we estimate py;u,v from the training data, we may consider classifiers

$$g_{NB}(G) = \arg\max_y \prod_{(u,v) \in V \times V} f(a_{u,v}; p_{y;u,v})$$

and

$$g_{\mathcal{E}}(G) = \arg\max_y \prod_{(u,v) \in \mathcal{E}} f(a_{u,v}; p_{y;u,v}).$$

The latter is the best we can hope for it considers the signal edges and only the signal edges; the former can be swamped by noise from non-signal edges.

Our interest is canonical subspace identification for this graph classification application; that is, estimate the collection of signal edges $\mathcal{E}$ via $\hat{\mathcal{E}}$ and consider the classifier

$$g_{\hat{\mathcal{E}}}(G) = \arg\max_y \prod_{(u,v)\in\hat{\mathcal{E}}} f(a_{u,v}; p_{y;u,v}).$$

We consider two different methods to estimate $\hat{\mathcal{E}}$ for IE1-star graphs.

if $q > p$, let $\delta_{u,v} = p_{1;u,v} - p_{0;u,v}$. thus $\hat{\delta}_{u,v} = \hat{p}_{1;u,v} - \hat{p}_{0;u,v}$

1.2.1 incoherent method: agnostic

The incoherent method does not utilize the stucture of the graph. Let the number of signal edges we will attempt to extract be $k = |\hat{\mathcal{E}}|$. Then our incoherent model is the k largest $\hat{\delta}_{u,v}$ edges.

1.2.2 coherrent method: max degree

This method takes advanage of the fact the IE1-star graphs has a vertex v^* which all edges with probability q are adjacent to. For convience let $v \in (u_1, u_2) \in V \times V$ mean $(u_1, u_2) \in V \times V$, and $u_1 = v$ or $u_2 = v$ (or $u_1 = u_2 = v$). First the coherent method estimates this vertex

$$\hat{v}^* = \arg\max_v \sum_{v\in(u_1,u_2)\in V\times V} \hat{\delta}_{u_1,u_2}$$

$\hat{\mathcal{E}}$ is the k largest $\hat{\delta}_{u,v}$ edges adjecent to v^*.

2 Theoretical results

2.1 Monotonisity of error given T

In IE1, using k canonical dimensions recovered from the training data ($|\hat{\mathcal{E}}| = k$), the probability of misclassification is monotonically decreasing as a function of $T = |\mathcal{E} \cap \hat{\mathcal{E}}|$; that is

$$t_1 > t_2 \Rightarrow E[L(g_{\hat{\mathcal{E}}})|T = t_1] < E[L(g_{\hat{\mathcal{E}}})|T = t_2].$$

2.1.1 $k = 1$ case

First consider the case where only one signal edge is attempted to be recovered ($k = 1$). Let g_0 represent the classifer if the recovered edge is not a signal edge ($t = 0$) and g_1 represent the classifier if the recovered edge is a signal edge ($t = 1$). If the above montonisity result is true we expect

$$E[L(g_1)] < E[L(g_0)].$$

Since we only have one edge, for simplicity let $\hat{p}_0$ and $\hat{p}_1$ denote the estimates of p_0 and p_1 for our recovered edge respectively. The following decomposes $E[L(g_0)]$ using the law of total probability conditioning on a, Y.

$$\begin{aligned}
E[L(g_0)] &= P[g_0 \neq Y] = P[\arg\max_y f(a;\hat{p}_y) \neq Y] && (1)\\
&= \sum_{j\in\{0,1\}} P[Y=j]P[\arg\max_y f(a;\hat{p}_y) \neq Y|Y=j] && (2)\\
&= \sum_{i,j\in\{0,1\}} P[Y=j]P[a=i|Y=j]P[\arg\max_y f(a;\hat{p}_y) \neq Y|a=i, Y=j] && (3)\\
&= P[Y=0]P[a=0|Y=0]P[\arg\max_y f(a;\hat{p}_y) \neq Y|a=0, Y=0] && (4)\\
&\quad +P[Y=0]P[a=1|Y=0]P[\arg\max_y f(a;\hat{p}_y) \neq Y|a=1, Y=0] && (5)\\
&\quad +P[Y=1]P[a=0|Y=1]P[\arg\max_y f(a;\hat{p}_y) \neq Y|a=0, Y=1] && (6)\\
&\quad +P[Y=1]P[a=1|Y=1]P[\arg\max_y f(a;\hat{p}_y) \neq Y|a=1, Y=1] && (7)\\
&= P[Y=0]P[a=0|Y=0]P[\arg\max_y f(0;\hat{p}_y) \neq 0] && (8)\\
&\quad +P[Y=0]P[a=1|Y=0]P[\arg\max_y f(1;\hat{p}_y) \neq 0] && (9)\\
&\quad +P[Y=1]P[a=0|Y=1]P[\arg\max_y f(0;\hat{p}_y) \neq 1] && (10)\\
&\quad +P[Y=1]P[a=1|Y=1]P[\arg\max_y f(1;\hat{p}_y) \neq 1] && (11)\\
&= \frac{1}{2}(1-p)P[\arg\max_y(1-\hat{p}_y) \neq 0] && (12)\\
&\quad +\frac{1}{2}pP[\arg\max_y \hat{p}_y \neq 0] && (13)\\
&\quad +\frac{1}{2}(1-p)P[\arg\max_y(1-\hat{p}_y) \neq 1] && (14)\\
&\quad +\frac{1}{2}pP[\arg\max_y \hat{p}_y \neq 1] && (15)
\end{aligned}$$

Note $\hat{p}_0, \hat{p}_1$ are independent of a, Y. Conditioning on the relationship between $\hat{p}_0$ and $\hat{p}_1$,

$$\begin{aligned}
= \quad & \frac{1}{2}(1-p)[P[\hat{p}_0 < \hat{p}_1]P[\arg\max_y(1-\hat{p}_y)) \neq 0|\hat{p}_0 < \hat{p}_1] && (16)\\
&\qquad +P[\hat{p}_0 = \hat{p}_1]P[\arg\max_y(1-\hat{p}_y) \neq 0|\hat{p}_0 = \hat{p}_1] && (17)\\
&\qquad +P[\hat{p}_0 > \hat{p}_1]P[\arg\max_y(1-\hat{p}_y) \neq 0|\hat{p}_0 > \hat{p}_1]] && (18)\\
&+\frac{1}{2}p[P[\hat{p}_0 < \hat{p}_1]P[\arg\max_y \hat{p}_y \neq 0|\hat{p}_0 < \hat{p}_1] && (19)\\
&\qquad +P[\hat{p}_0 = \hat{p}_1]P[\arg\max_y \hat{p}_y \neq 0|\hat{p}_0 = \hat{p}_1] && (20)\\
&\qquad +P[\hat{p}_0 > \hat{p}_1]P[\arg\max_y \hat{p}_y \neq 0|\hat{p}_0 > \hat{p}_1]] && (21)\\
&+\frac{1}{2}(1-p)[P[\hat{p}_0 < \hat{p}_1]P[\arg\max_y(1-\hat{p}_y) \neq 1|\hat{p}_0 < \hat{p}_1] && (22)\\
&\qquad +P[\hat{p}_0 = \hat{p}_1]P[\arg\max_y(1-\hat{p}_y) \neq 1|\hat{p}_0 = \hat{p}_1] && (23)\\
&\qquad +P[\hat{p}_0 > \hat{p}_1]P[\arg\max_y(1-\hat{p}_y) \neq 1|\hat{p}_0 > \hat{p}_1]] && (24)\\
&+\frac{1}{2}p[P[\hat{p}_0 < \hat{p}_1]P[\arg\max_y \hat{p}_y \neq 1|\hat{p}_0 < \hat{p}_1] && (25)\\
&\qquad +P[\hat{p}_0 = \hat{p}_1]P[\arg\max_y \hat{p}_y \neq 1|\hat{p}_0 = \hat{p}_1] && (26)\\
&\qquad +P[\hat{p}_0 > \hat{p}_1]P[\arg\max_y \hat{p}_y \neq 1|\hat{p}_0 > \hat{p}_1]] && (27)
\end{aligned}$$

In the event that $\hat{p}_0 = \hat{p}_1$ the classifier's decicion is randomized thus $P[g_y \neq Y|\hat{p}_0 = \hat{p}_1] = \frac{1}{2}$ [true??] for $y = \{0, 1\}$. Notice with the conditional probabilities relating to g_y are either 0, 0.5, or 1.

$$\begin{aligned}
&= \frac{1}{2}(1-p)[\frac{1}{2}P[\hat{p}_0 = \hat{p}_1] + P[\hat{p}_0 > \hat{p}_1]] && (28)\\
&\quad +\frac{1}{2}p[P[\hat{p}_0 < \hat{p}_1] + \frac{1}{2}P[\hat{p}_0 = \hat{p}_1]] && (29)\\
&\quad +\frac{1}{2}(1-p)[P[\hat{p}_0 < \hat{p}_1] + \frac{1}{2}P[\hat{p}_0 = \hat{p}_1]] && (30)\\
&\quad +\frac{1}{2}p[\frac{1}{2}P[\hat{p}_0 = \hat{p}_1] + P[\hat{p}_0 > \hat{p}_1]] && (31)\\
&= \frac{1}{2}P[\hat{p}_0 = \hat{p}_1][\frac{1}{2}(1-p) + \frac{1}{2}p + \frac{1}{2}(1-p) + \frac{1}{2}p] && (32)\\
&\quad +\frac{1}{2}(1-p)P[\hat{p}_0 > \hat{p}_1] + \frac{1}{2}pP[\hat{p}_0 < \hat{p}_1] + \frac{1}{2}(1-p)P[\hat{p}_0 < \hat{p}_1] + \frac{1}{2}pP[\hat{p}_0 > \hat{p}_1] && (33)\\
&= \frac{1}{2}P[\hat{p}_0 = \hat{p}_1] + \frac{1}{2}P[\hat{p}_0 > \hat{p}_1] + \frac{1}{2}P[\hat{p}_0 < \hat{p}_1] && (34)\\
&= \frac{1}{2} && (35)
\end{aligned}$$

Now consider the classifier g_1. Repeating the same conditioning as on g_0 we get

$$\begin{aligned}
E[L(g_1)] &= P[g_1 \neq Y] && (36)\\
&= P[Y=0]P[a=0|Y=0]P[\arg\max_y f(0;\hat{p}_y) \neq 0] && (37)\\
&\quad +P[Y=0]P[a=1|Y=0]P[\arg\max_y f(1;\hat{p}_y) \neq 0] && (38)\\
&\quad +P[Y=1]P[a=0|Y=1]P[\arg\max_y f(0;\hat{p}_y) \neq 1] && (39)\\
&\quad +P[Y=1]P[a=1|Y=1]P[\arg\max_y f(1;\hat{p}_y) \neq 1] && (40)\\
&= \frac{1}{2}(1-p)P[\arg\max_y (1-\hat{p}_y) \neq 0] && (41)\\
&\quad +\frac{1}{2}pP[\arg\max_y \hat{p}_y \neq 0] && (42)\\
&\quad +\frac{1}{2}(1-q)P[\arg\max_y (1-\hat{p}_y) \neq 1] && (43)\\
&\quad +\frac{1}{2}qP[\arg\max_y \hat{p}_y \neq 1] && (44)
\end{aligned}$$

Now again conditioning on the relationship between $\hat{p}_0$ and $\hat{p}_1$ and noting after conditioning probabilities relating to g_y are either 0, 0.5, or 1.

$$\begin{aligned}
= \quad & \frac{1}{2}(1-p)[P[\hat{p}_0 < \hat{p}_1]P[\arg\max_y(1-\hat{p}_y)) \neq 0|\hat{p}_0 < \hat{p}_1] && (45)\\
& \qquad +P[\hat{p}_0 = \hat{p}_1]P[\arg\max_y(1-\hat{p}_y) \neq 0|\hat{p}_0 = \hat{p}_1] && (46)\\
& \qquad +P[\hat{p}_0 > \hat{p}_1]P[\arg\max_y(1-\hat{p}_y) \neq 0|\hat{p}_0 > \hat{p}_1]] && (47)\\
& +\frac{1}{2}p[P[\hat{p}_0 < \hat{p}_1]P[\arg\max_y \hat{p}_y \neq 0|\hat{p}_0 < \hat{p}_1] && (48)\\
& \qquad +P[\hat{p}_0 = \hat{p}_1]P[\arg\max_y \hat{p}_y \neq 0|\hat{p}_0 = \hat{p}_1] && (49)\\
& \qquad +P[\hat{p}_0 > \hat{p}_1]P[\arg\max_y \hat{p}_y \neq 0|\hat{p}_0 > \hat{p}_1]] && (50)\\
& +\frac{1}{2}(1-q)[P[\hat{p}_0 < \hat{p}_1]P[\arg\max_y(1-\hat{p}_y) \neq 1|\hat{p}_0 < \hat{p}_1] && (51)\\
& \qquad +P[\hat{p}_0 = \hat{p}_1]P[\arg\max_y(1-\hat{p}_y) \neq 1|\hat{p}_0 = \hat{p}_1] && (52)\\
& \qquad +P[\hat{p}_0 > \hat{p}_1]P[\arg\max_y(1-\hat{p}_y) \neq 1|\hat{p}_0 > \hat{p}_1]] && (53)\\
& +\frac{1}{2}q[P[\hat{p}_0 < \hat{p}_1]P[\arg\max_y \hat{p}_y \neq 1|\hat{p}_0 < \hat{p}_1] && (54)\\
& \qquad +P[\hat{p}_0 = \hat{p}_1]P[\arg\max_y \hat{p}_y \neq 1|\hat{p}_0 = \hat{p}_1] && (55)\\
& \qquad +P[\hat{p}_0 > \hat{p}_1]P[\arg\max_y \hat{p}_y \neq 1|\hat{p}_0 > \hat{p}_1]] && (56)\\
= \quad & \frac{1}{2}(1-p)[\frac{1}{2}P[\hat{p}_0 = \hat{p}_1] + P[\hat{p}_0 > \hat{p}_1]] && (57)\\
& +\frac{1}{2}p[P[\hat{p}_0 < \hat{p}_1] + \frac{1}{2}P[\hat{p}_0 = \hat{p}_1]] && (58)\\
& +\frac{1}{2}(1-q)[P[\hat{p}_0 < \hat{p}_1] + \frac{1}{2}P[\hat{p}_0 = \hat{p}_1]] && (59)\\
& +\frac{1}{2}q[\frac{1}{2}P[\hat{p}_0 = \hat{p}_1] + P[\hat{p}_0 > \hat{p}_1]] && (60)
\end{aligned}$$

Factoring in terms of $P[\hat{p}_0 < \hat{p}_1]$, $P[\hat{p}_0 = \hat{p}_1]$, and $P[\hat{p}_0 > \hat{p}_1]$.

$$\begin{aligned}
= \quad & P[\hat{p}_0 < \hat{p}_1][\frac{1}{2}p + \frac{1}{2}(1-q)] && (61)\\
& +\frac{1}{2}P[\hat{p}_0 = \hat{p}_1][\frac{1}{2}(1-p) + \frac{1}{2}p + \frac{1}{2}(1-q) + \frac{1}{2}q] && (62)\\
& +P[\hat{p}_0 > \hat{p}_1][\frac{1}{2}(1-p) + \frac{1}{2}q] && (63)\\
= \quad & \frac{1}{2}P[\hat{p}_0 < \hat{p}_1][1-(q-p)] && (64)\\
& +\frac{1}{2}P[\hat{p}_0 > \hat{p}_1][1+(q-p)] && (65)\\
& +\frac{1}{2}P[\hat{p}_0 = \hat{p}_1] && (66)\\
= \quad & \frac{1}{2} - (q-p)[P[\hat{p}_0 < \hat{p}_1] - P[\hat{p}_0 > \hat{p}_1]] && (67)
\end{aligned}$$

Remember that $q > p$, so if $P[\hat{p}_0 < \hat{p}_1] - P[\hat{p}_0 > \hat{p}_1] > 0$ then $E[L(g_0)] > E[L(g_1)]$. Because our recovered edge is a signal edge, $\hat{p}_0$ ~Binomial(p,$s/2$) and $\hat{p}_1$ ~Binomial(p,$s/2$). And since $\hat{p}_0$ and $\hat{p}_1$ are independent,

$$
\begin{aligned}
P[\hat{p}_0 > \hat{p}_1] &= \sum_{x>y:x,y\in[s/2]} P[\hat{p}_0 = x]P[\hat{p}_1 = y] && (68)\\
&= \sum_{x>y:x,y\in[s/2]} \binom{s_0}{x} p^x(1-p)^{s_0-x} \binom{s_1}{y} q^y(1-q)^{s_1-y} && (69)\\
&= \sum_{x>y:x,y\in[s/2]} \binom{s_0}{x}\binom{s_1}{y} p^x(1-p)^{s_0-x} q^y(1-q)^{s_1-y} && (70)
\end{aligned}
$$

Similarly,

$$
P[\hat{p}_0 < \hat{p}_1] = \sum_{x>y:x,y\in[s/2]} \binom{s_1}{x}\binom{s_0}{y} q^x(1-q)^{s_1-x} p^y(1-p)^{s_0-y} \quad (71)
$$

Thus

$$
\begin{aligned}
P[\hat{p}_0 < \hat{p}_1] - P[\hat{p}_0 > \hat{p}_1] &= \sum_{x>y:x,y\in[s/2]} \binom{s_0}{x}\binom{s_1}{y} p^x(1-p)^{s_0-x} q^y(1-q)^{s_1-y} && (72)\\
&\quad - \binom{s_1}{x}\binom{s_0}{y} q^x(1-q)^{s_1-x} p^y(1-p)^{s_0-y} && (73)
\end{aligned}
$$

Note that $s_0 = s_1$, allowing for us to factor

$$
\begin{aligned}
P[\hat{p}_0 < \hat{p}_1] - P[\hat{p}_0 > \hat{p}_1] &= \sum_{x>y:x,y\in[s/2]} \binom{s_0}{x}\binom{s_1}{y} [q^x(1-q)^{s_1-x} p^y(1-p)^{s_0-y} - p^x(1-p)^{s_0-x} q^y(1-q)^{s_1-y}] && (74)\\
&= \sum_{x>y:x,y\in[s/2]} \binom{s_0}{x}\binom{s_0}{y} p^y q^y(1-p)^{s_0-x}(1-q)^{s_0-x}[q^{x-y}(1-p)^{x-y} - p^{x-y}(1-q)^{x-y}] && (75)
\end{aligned}
$$

Since $q > p$ and $x - y > 0$, $q^{x-y}(1-p)^{x-y} > p^{x-y}(1-q)^{x-y}$. Therefore $P[\hat{p}_0 < \hat{p}_1] - P[\hat{p}_0 > \hat{p}_1] > 0$ and $E[L(g_1)] < E[L(g_0)]$.

2.1.2 $k > 1$ case

2.2 Assumtotic/Approximate distribution of T

2.2.1 Agnostic method

We approximate the binomial $p_{0;i,j}$, and $p_{1;i,j}$ with normal distributions so

$$
\hat{p}_{0;i,j} \approx N_{0;i,j} \sim \text{ Normal}(p_{0;i,j}, p_{0;i,j}(1-p_{0;i,j})s/2)
$$

$$
\hat{p}_{1;i,j} \approx N_{1;i,j} \sim \text{ Normal}(p_{1;i,j}, p_{1;i,j}(1-p_{1;i,j})s/2)
$$

$$
\delta_{i,j} \approx N_{1;i,j} - N_{0;i,j}
$$

The cdf of the r^{th} ordered statistic of n iid random variables is

$$
F_{X_{(r:n)}}(x) = P[X_{(r:n)} \le x] = \sum_{i=r}^{n} \binom{n}{i} F_X(x)^i[1 - F_X(x)]^{n-i}
$$

The pdf of the r^{th} ordered statistic of n iid random variables is

$$
f_{X_{(r:n)}}(x) = \binom{n}{r-1, n-r, 1} F_X^{r-1}(x)[1 - F_X(x)]^{n-r} f_X(x)
$$

The joint pdf of 2 ordered statistics $r < s$, $x \leq y$

$$f_{(r)(s):n}(x,y) = \binom{n}{r-1, 1, s-r-1, 1, n-s} F^{r-1}(x) f(x) [F(y) - F(x)]^{s-r-1} f(y) [1 - F(y)]^{n-s}$$

Let us define another random variable,

$$N_p \sim \text{Normal}(p, p(1-p)s/2)$$

$$\begin{aligned}
P[T = k] &\approx P[N_{q_{(m-k+1:m)}} > N_{p_{(n^2-m:n^2-m)}}] && (76)\\
&= \int_{-\infty}^{\infty} f_{N_{q_{(m-k+1:m)}}}(x) F_{N_{p_{(n^2-m:n^2-m)}}}(x)\ dx && (77)\\
&= \int_{-\infty}^{\infty} \binom{m}{m-k, k-1, 1} F_{N_q}^{m-k}(x) [1 - F_{N_q}(x)]^{k-1} f_{N_q}(x) F_{N_p}(x)^{n^2-m}\ dx && (78)
\end{aligned}$$

$$\begin{aligned}
P[T = 0] &\approx P[N_{p_{(n^2-m-k+1:n^2-m)}} > N_{q_{(m:m)}}] && (79)\\
&= \int_{-\infty}^{\infty} f_{N_{p_{(n^2-m-k+1:n^2-m)}}}(x) F_{N_{q_{(m:m)}}}(x)\ dx && (80)\\
&= \int_{-\infty}^{\infty} \binom{n^2-m}{n^2-m-k, k-1, 1} F_{N_p}^{n^2-m-k}(x) [1 - F_{N_p}(x)]^{k-1} f_{N_p}(x) F_{N_q}(x)^{m}\ dx && (81)
\end{aligned}$$

For $t \in [1, 2, \ldots, k-1]$, the event that $T = t$ is when the largest k of the δ's are from a Binomial$(q, s/2)$ and the remaining $k - t$ drawn from Binomial$(p, s/2)$.

$$\begin{aligned}
P[T = t] &\approx P[N_{p_{(n^2-m-k+t:n^2-m)}} < N_{q_{(m-t+1:m)}}, N_{q_{(m-t:m)}} < N_{p_{(n^2-m-k+t+1:n^2-m)}}]\\
&= \iint_{x<y} f_{N_{q_{(m-t)(m-t+1):m}}}(x,y) P[N_{p_{(n^2-m-k+t:n^2-m)}} < y, x < N_{p_{(n^2-m-k+t+1:n^2-m)}}]\ dx\ dy\\
&= \iiiint_{x<y,\ w<y,\ x<z,\ w<z} f_{N_{q_{(m-t)(m-t+1):m}}}(x,y) f_{N_{p_{(n^2-m-k+t)(n^2-m-k+t+1):n^2-m}}}(w,z)\ dw\ dx\ dy\ dz\\
&= \iiiint_{x<y,\ w<y,\ x<z,\ w<z} \binom{n}{m-t-1, 1, 0, 1, n-m+t-1} F_{N_q}^{m-t-1}(x) f_{N_q}(x) [F_{N_q}(y) - F_{N_q}(x)]^0 f_{N_q}(y) [1 - F_{N_q}(y)]^{n-}\\
&\quad \binom{n^2-m}{n^2-m-k+t-1, 1, 0, 1, k-t-1} F_{N_p}^{n^2-m-k+t-1}(w) f_{N_p}(w) [F_{N_p}(z) - F_{N_p}(w)]^0 f_{N_p}(z) [1 - F_{N_p}(y)]^{k-t-1}\ d\\
&= \iiiint_{x<y,\ w<y,\ x<z,\ w<z} \binom{n}{m-t-1, n-m+t-1, 1, 1} F_{N_q}^{m-t-1}(x) f_{N_q}(x) f_{N_q}(y) [1 - F_{N_q}(y)]^{n-m+t-1}\\
&\quad \binom{n^2-m}{n^2-m-k+t-1, k-t-1, 1, 1} F_{N_p}^{n^2-m-k+t-1}(w) f_{N_p}(w) f_{N_p}(z) [1 - F_{N_p}(z)]^{k-t-1}\ dw\ dx\ dy\ dz
\end{aligned}$$

2.2.2 Maximum degree method

Now lets calculate the T distribution for IE1-star method. Let v^* be the "right" vertex. The probability of picking the right vertex is

$$\begin{aligned}
P[\text{right vertex}] &= P\left[\arg\max_v \sum_{v \in i, j \in V \times V} \hat{p}_{1;i,j} = v^*\right] && (89)\\
&= P[] && (90)
\end{aligned}$$

Notice after picking the right vertex that the probability of T is the same as before except instead of $n^2 - m$ edges with probability p there are only $2n-1-m$ edges. If the wrong vertex is chosen, then it is still possible to have 0, 1, or 2 signal edges. For $i < 2$

$$P[\text{i edge possible correct}|\text{wrong edge}] = \frac{\binom{m}{i}\binom{2n-3}{m-i}}{\binom{2n-1}{m}} \tag{91}$$

Now the distribution of $T = t$ for $t > 2$ can be approximated as follows

$$\begin{aligned}
P[T=t] &= P[\text{right vertex}]P[T=t|\text{right vertex}] + P[\text{wrong vertex}]P[T=t|wrongvertex] && (92)\\
&= P[\text{right vertex}]P[N_{p_{(2n-1-m-k+t:2n-1-m)}} < N_{q_{(m-t+1:m)}}, N_{q_{(m-t:m)}} < N_{p_{(2n-1-m-k+t+1:2n-1-m)}}] && (93)\\
&\quad +0 && (94)\\
&= && (95)
\end{aligned}$$

For $t \leq 2$,

$$\begin{aligned}
P[T=t] &= P[\text{right vertex}]P[T=2|\text{right vertex}] + P[\text{wrong vertex}]P[T=t|\text{wrong vertex}] && (96)\\
&= P[\text{right vertex}]P[N_{p_{(2n-1-m-k+t:2n-1-m)}} < N_{q_{(m-t+1:m)}}, N_{q_{(m-t:m)}} < N_{p_{(2n-1-m-k+t+1:2n-1-m)}}] && (97)\\
&\quad +P[\text{wrong vertex}]\sum_{i=0}^{2} P[T=t|\text{wrong edge, } i \text{ possible}]P[i \text{ possible}|\text{wrong edge}] && (98)\\
&= P[\text{right vertex}]P[N_{p_{(2n-1-m-k+t:2n-1-m)}} < N_{q_{(m-t+1:m)}}, N_{q_{(m-t:m)}} < N_{p_{(2n-1-m-k+t+1:2n-1-m)}}] && (99)\\
&\quad +P[\text{wrong vertex}]\sum_{i \geq t} P[T=t|\text{wrong edge, } i \text{ possible}]P[i \text{ possible}|\text{wrong edge}] && (100)\\
&= && (101)
\end{aligned}$$

2.3 Relative Efficiency

With these two methods we need a way to compare their performance. Relative efficiency is one way to do so. Recall that the ratio Nt /NW is a measure of the relative efficiency of the Wilcoxon test versus the t test, where Nt and NW are minimum sample sizes required to achieve some specified power at some specified size. [cite B&D 1977 page 352] [cite Priebe gwmw papers . . . ?]

Since we are in the case for which all signal edges are created equally and all noise edges are created equally, if we constrain all canonical subspace identification methods so that $|\hat{\mathcal{E}}| = k$ for some k, then Priebes Conjecture #1 above implies that comparing the number of signal dimensions recovered for two canonical subspace identification methods allows comparison of classification performance.

Toward that end, for canonical subspace identification method x define

$$T_x(k, s, F_{GY}) = |\mathcal{E} \cap \hat{\mathcal{E}}_x|$$

to be the number of signal dimensions recovered with training sample size s using method x.

Let

$$s_x(t) = \min\{s : E[T_x(k, s, F_{GY})] \geq t\}.$$

The ratio $r(t) = s_{coherent}(t)/s_{agnostic}(t)$ is the relative efficiency.

2.4 Simulations

There do exist situations where the agnostic model outperforms the coherenet model. Notice that in our coherent model, if the wrong vertex is chosen, then $t = 0$. Here is a simulation that shows situations where the agnostic models is better and situations where the coherent model is better. In this simulation $p = 0.45$, $m = 10$, $q = 0.55$, $k = 10$, $t = 5$, and $n = |V|$ ranges from 10 to 150 in intervals of 10. (m is the number of edges with probability q.) For each simulation of $E[T]$, 100 trials are run.

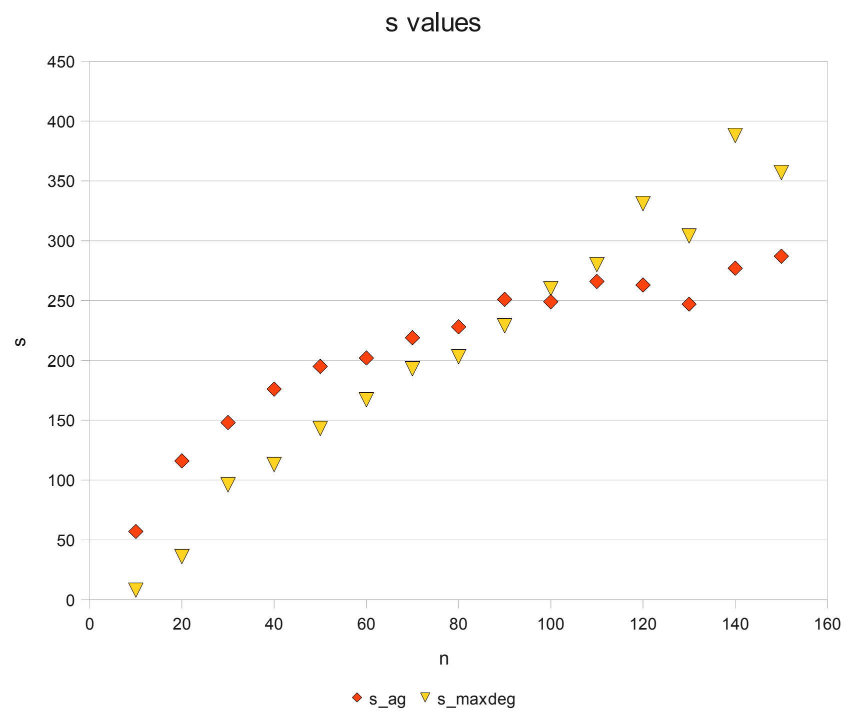

Figure 1: s values of the agnostic and max degree models.

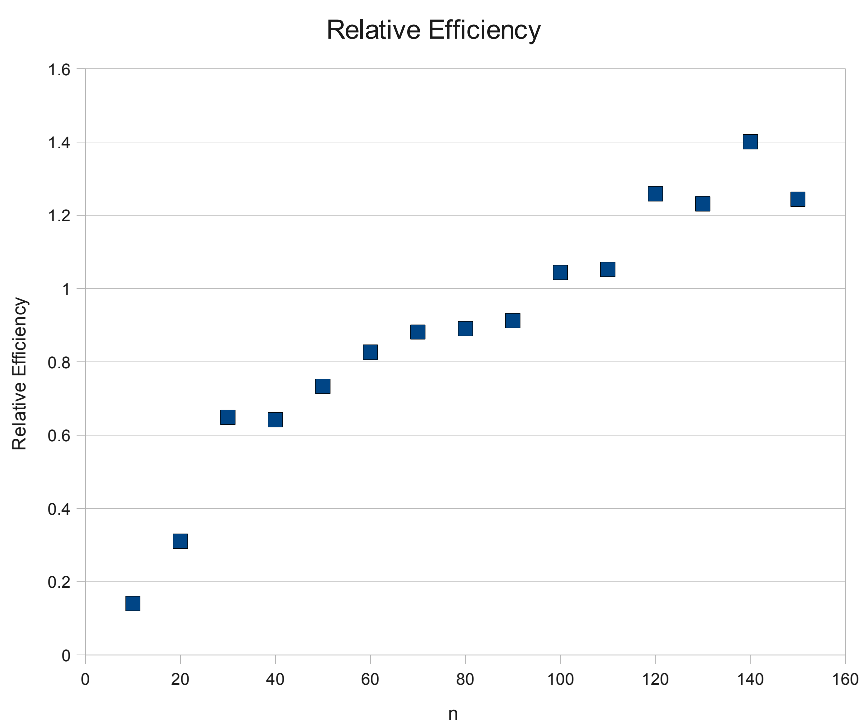

Figure 2: Relative efficiency of stated simulation.