Model Selection for Complex Networks

Jacob Jensen with Cris Moore and Cosma Shalizi

I. ABSTRACT

The Akaike Information Criterion (AIC) is a powerful and highly principled method of model selection. When working with probabilistic models it is typical to fit them to the data by maximum likelihood, using Markov Chain Monte Carlo or Belief Propagation. The maximum likelihood value obtained in this way gives an indication of how good a fit the model is to the data. Unfortunately, models with more parameters tend to fit the data strictly better, simply because they have more degrees of freedom and not necessarily because they describe it better. The AIC counters this by taking the $-\ell(\hat{\theta})$ maximum likelihood value and adding a penalty term of k = $|\theta|$. The best model according to the AIC is then the one that minimizes $-\ell(\hat{\theta}) + k$. This simple penalty has a powerful justification, but makes certain assumptions that pose major problems to models used to infer networks. The first of these assumptions, that all parameters in θ are continuous, can be remedied by a monte carlo prodcedure that marginalizes over these variables. The second, which assumes narrow peak around in the model's likelihood, is normally easy to deal with given enough data to accurately estimate parameters. Unfortunately, highly correlated distributions like the Degree-Corrected Stochastic Block Model tend to violate this assumption. We attempt to create a variant of the TIC that treats pairwise group assignment marginals across edges as independent samples from the distribution over network edges.

II. INTRODUCTION

Given a set of data, we often want to fit a parametric model to it, both to infer associated properties of the model and to made predictions. Given a set of n data points plotted on an X-Y axis, we might want to fit a degree polynomial to it to make predictions about the value f(x) given a new x. Allowing a higher-degree polynomial, however, gives a strictly better fit as measured by the sum of squares difference between the fitted model polynomial and the data, $\sum_i m_d(x_i) - y_i$, where y_i is the true value of the output corresponding to input x_i.

One among many methods for Model Selection, covered in[2], called the Akaike Information Criterion (AIC)[1] attempts to address this by adding a penalty to the naively assessed goodness-of-fit of the model. The naively assessed goodness-of-fit is in terms of log-likelihood. Given a "Generative Model", a model that can assign to a given set of data the probability that data would have been generated by the model. The likelihood of the data given by a model will be off by a factor of a normalization constant. Taking the log-likelihood (called ℓ), logZ becomes an additive constant factor.

When comparing several models M_k, k $\in$ (1,...,k), fitted with maximum-likelihood parameters $\hat{theta}_k$,

$$AIC(M_k) = \ell_k - \left|\hat{theta}_k\right|$$

Choosing the model with the maximum AIC score is equivalent to choosing the model which, if it is correct (a big assumption), and if maximum-likelihood parameters are Gaussian-distributed around "true" or "least false" parameters (the ones that will be consistently inferred even with more data) (another big assumption) will best predict hidden data, future data, or data generated by the same underlying process.

Let's stick to the first assumption, that the model we are inferring is capable of correctly describing the process that generated the data, for now. It will be relaxed later.

Meanwhile, let's look at a case where the second assumption, inferred parameters normally-distributed around the true parameters, is false. This will be the simple and well-known Mixture of Gaussians model, and it will prepare us for the more general Degree-Corrected Stochastic Block Model[4] that we will soon investigate. Consider X, a set of n data points $x_i, i \in (1, ..., n)$. These are drawn from one of m spherical Gaussians with centers $c_j, j \in (1, ..., m)$. Which Gaussian a point is drawn from is given by an indicator vector m_i.

$$\tag{1}$$

$$P(x_i|m_i, c_i) \tilde{} N(c_{m_i}, I) \tag{2}$$

$$m_i \tilde{} Cat(\frac{1}{m}, ..., \frac{1}{m}) \tag{3}$$

The discrete parameters can't be Gaussian distributed, since they're categorical. Even if we created a continuous parametrization, say the probability that a data point was generated by each individual Gaussian, it would be continuous but not normally distributed.

Sometimes we would infer a very high peak around one center, and sometimes a high peak around another center, but the variation would be peaked at high and low probabilities rather than a nice bell-curve shape. I will address how to resolve this in just a moment.

III. OUR MODEL

The model we're actually interested in is a network model, particularly the Degree-Corrected Stochastic Block Model [CITE: KarrerNewman]. Given a network, a set of nodes and connecting edges, we want to establish the best possible division into communities. When one thinks of communities in networks, one often thinks of homophilous communities where in-group connections are more frequent than out-group connections; these are the sorts of communities traditionally found by methods like spectral clustering and modularity maximization. However functional communities, like the community of predators which connect to the community of prey, but not to other predators, and prey, which are connected to by predators but do not connect to each other, are also interesting and useful. An type of generative model called the stochastic block model can find both.

Let there be n vertices and edges $A_{ij} \forall i, j$ between them designated by graph G, g_i be the group assigned to a vertex out of k possible, θ_i the degree of the vertex (and this degree sequence is fixed; you can think of it as being given information). Ω is a matrix of group connectivities, and ω_{rs} is the propensity for a vertex of group r to connect to a vertex of group s. Then the probability of G is given by

$$P(G|\Omega, g) \propto P(\Omega|G, g) \propto \prod_{i<j} \frac{(\theta_i \theta_j \omega_{g_i g_j})^{A_{ij}}}{A_{ij}} e^{-\theta_i \theta_j \omega_{g_i g_j}}$$

$$P(\Omega|G, g) = \ell \propto \sum_{ij} A_{ij} log(\theta_i \theta_j \omega_{g_i g_j}) - \theta_i \theta_j \omega_{g_i g_j}$$

Assuming no self-edges. The motivation and details of this model can be found in [Newman Karrer DCSBM]. For our purposes, however, the crucial part is that

$$logP(A_{ij}|\Omega, g_i, g_j) = \ell_{ij} = A_{ij} log(\theta_i \theta_j \omega_{g_i g_j}) - \theta_i \theta_j \omega_{g_i g_j}$$

What we want to do is marginalize, for a given set of continuous parameters Ω over all n^k possible group assignments and treat $P(\hat{\Omega}|G) = max_\Omega \sum_g P(\Omega|g, G)$ as fitted model's

log-likelihood and apply a penalty to that. The rest of this paper shows how to do this, suggests some changes to adapt the Akaike Information Criterion to this new setting.

IV. THE INDEPENDENT DEGREE "TOY" MODEL"

We begin not with the model itself but with a toy version. Consider two models which represent the ordinary Stochastic Block Model and the degree-corrected one respectively.

In each model we are given data $x_1, x_2, ..., x_n$. In the first model we assume all data is generated by a poisson distribution parametrized with $\lambda = c_0$. $\theta = \{c\}$ and $\hat{\theta} = \{\hat{c}\}$. The poisson distribution is closed under addition, $Pois(\lambda_1) + Pois(\lambda_2) = Pois(\lambda_1 + \lambda_2)$. We can regard the sum of our samples as coming from a single draw from $Pois(nc_0)$, which neatly shows that what we consider "sufficient" data to estimate the Poisson distribution depends not just on the number of draws but also on the intensity parameter λ.

Our MLE is: $\hat{c} = \frac{\sum_i x_i}{/} n$, the average of all x_i. Our max log-likelihood is:

Our other model has as many independently parametrized Poisson Distributions as there are variables. The MLE for each Poisson's λ is simply the value drawn from it. $\hat{\psi} = \{x_1, x_2, ..., x_n\}$

Recall that $n\hat{c} = \sum_i x_i$.

Now we want to find the difference in log-likelihoods between these models so we can have an idea how much better the model with the higher number of parameters will fit. This is given to us by

$$L(\hat{\psi}) - \ell(\hat{\theta}) = (\sum_i x_i + \sum_i x_i log x_i - \sum_i log x_i!) - (n\hat{c} + n\hat{c} log\hat{c} - \sum_i log x_i!) \tag{4}$$

$$= \sum_i x_i log x_i - n\hat{c} log\hat{c} \tag{5}$$

Our expectation of this difference is:

$$\sum_{x_i} (\sum_x Pois(c, x) x log x - c log c) \tag{6}$$

Two figures are included in this folder of simulations of the above formula (multiplied by 2) with c = 2 and c = 10, n = 100, plotted against a chi-square distribution with n degrees of freedom. In the case c = 2 (or generally c ¡ 5ish) it is to the right of the chi-square curve, indicating an upwards bias in the difference. In the case of c = 10 (or generally c ¿= 6ish) it fits the chi-square distribution fairly well.

V. THE SINGLE-GROUP DEGREE-CORRECTED STOCHASTIC BLOCK MODEL

The next step is to apply the AIC to distinguish between the Single-Group degree-corrected stochastic Block Model and the homogenous-degree Poisson model.

The first model is given by:

$$\ell(\hat{c}|A) = log \prod_{i<j} Pois(\frac{\hat{c}}{n}; A_{ij}) \tag{7}$$

$$= log \prod_{i<j} \frac{e^{\frac{\hat{c}}{n}} \frac{\hat{c}}{n}^{A_{ij}}}{A_{ij}!} \tag{8}$$

$$= \sum_{i<j} \frac{\hat{c}}{n} + A_{ij} log\hat{c} - logn - logA_{ij}! \tag{9}$$

$$\tag{10}$$

Where $\hat{c}$ is the MLE of the expected degree of each node $\frac{2*\sum_{i<j} A_{ij}}{n}$

The second model, nested inside the first, is given by:

$$\ell(\hat{\theta}|A) = log \prod_{i<j} Pois(\frac{\hat{\theta}_i \hat{\theta}_j}{\hat{c}n}; A_{ij}) \tag{11}$$

$$= log \prod_{i<j} \frac{e^{\frac{\hat{\theta}_i \hat{\theta}_j}{\hat{c}n}} \frac{\hat{\theta}_i \hat{\theta}_j}{\hat{c}n}^{A_{ij}}}{A_{ij}!} \tag{12}$$

$$= \sum_{i<j} \frac{\hat{\theta}_i \hat{\theta}_j}{\hat{c}n} + A_{ij} \left[log\hat{\theta}_i + log\hat{\theta}_j - log\hat{c} - logn \right] - logA_{ij}! \tag{13}$$

$\hat{\theta}_i$ is estimated as the observed degree of node i.

We want to find the expected difference in likelihood between the two models

$$\Delta\ell = \ell(\hat{\theta}|A) - \ell(\hat{c}|A) \tag{14}$$

$$= \sum i < j \frac{-\hat{\theta}_i\hat{\theta}_j}{\hat{c}n} + \frac{\hat{c}}{n} + A_{ij}\left[log\hat{\theta}_i + log\hat{\theta}_j - 2 * log\hat{c}\right] \tag{15}$$

$$= \sum i\hat{\theta}_i\left[log\hat{\theta}_i - log\hat{c}\right] \tag{16}$$

The third line above results from the fact that the first term cancels out (the expectation of $\frac{-\hat{\theta}_i\hat{\theta}_j}{\hat{c}n}$ is equal to $\frac{\hat{c}}{n}$) while in the second term the $log\hat{\theta}_i$ term appears exactly $\hat{\theta}_i$ times, equal to $\sum_j A_{ij}$, the $\hat{\theta}_i$ estimator.

Empirically, this DOES satisfy the AIC, and log-likelihood is increased by a chi-squared distribution of degree k/2. There is no correction term except from deviations from the given model during data generation (mentioned below).

In a real network, however, there are two noteworthy deviations from the AIC. First, very few networks we study have multi-edges (and those that do may not be poisson). Second, a node with degree zero is likely to be ignored in our study. To correct for the first, actual edges are taken as min(1,poisson draw). This can be corrected for in an AIC-friendly way. According to Claeskens and Hjort ch. 2 a correction can be made for mal-dispersed poisson draws by multiplying the AIC penalty by VAR[draws]/E[draws]. In our case, this is $\frac{var[\theta_i]}{E}[\theta_i]$. The second is harder to correct for, and so far I only have an ad hoc penalty that can deviate quite drastically. As long as the expected degree of each node is at least 4 or 5, however, the one need not worry.

The variance of the distribution above is very simply derived from the likelihood above:

$$Var[\Delta\ell] = E[(\sum i\hat{\theta}_i\left[log\hat{\theta}_i - log\hat{c}\right])^2] \tag{17}$$

$$- E[\sum i\hat{\theta}_i\left[log\hat{\theta}_i - log\hat{c}\right]]^2 \tag{18}$$

$$\tag{19}$$

This is neat and easy because θ_i are conditionally independent from each other given A_{ij}. This form essentially hides the covariances that emerge from that, so a more explicit form of the variance is given by (could be mistakes here?):

$$Var[\Delta\ell] = E[\sum_{i,j}(\hat{\theta}_i \left[(log\hat{\theta}_i - log\hat{c}\right])^T(\hat{\theta}_j \left[log\hat{\theta}_j - log\hat{c}\right])] \tag{20}$$

$$- E[\sum i\hat{\theta}_i \left[log\hat{\theta}_i - log\hat{c}\right]]^2 \tag{21}$$

$$= \sum_{i,j} E[(\hat{\theta}_i \left[(log\hat{\theta}_i - log\hat{c}\right] - E[\Delta\ell])(\hat{\theta}_j \left[(log\hat{\theta}_j - log\hat{c}\right] - E[\Delta\ell])] \tag{22}$$

$$= \sum_{i,j} E[((\sum_{k\neq i} A_{ik} + A_{ij}) \left[log(\sum_{k\neq i} A_{ik} + A_{ij}) - log\hat{c}\right] - n) \tag{23}$$

$$((\sum_{k\neq j} A_{kj} + A_{ij}) \left[log(\sum_{k\neq j} A_{kj} + A_{ij}) - log\hat{c}\right] - n)] \tag{24}$$

There is a very general correction to the AIC that can be made based on its derivation. In the ideal case, we expect the penalty to be k, the number of parameters. However, this actually results from a term in the fourier series of the difference of two nested model likelihoods, that term being [I won't go through the full derivation in this file]:

$$Tr\left[\hat{J}^{-1}\hat{K}\right]$$

$$\hat{J} = \frac{1}{n}\sum_{i}^{n} I(y_i, \hat{\theta})$$

Also called the Fisher Information Matrix, expected value of the second derivative at the maximum log likelihood

$$\hat{K} = \frac{1}{n}\sum_{i}^{n} u(y_i, \hat{\theta})u(y_i, \hat{\theta})^T$$

Where $u(y_i, \hat{\theta}$ is the score vector of first derivative of the maximum log likelihood. The trace form of this is equivalent to the $u^T J^{-1} u$ form that may be more familiar.

This may seem kind of hard to parse, and it's only in special cases that it is directly useful (calculating these quantities can be a significant hassle), but it offers a potentially powerful correction to the AIC (the poisson correction above comes from this).

VI. DERIVING THE AKAIKE INFORMATION CRITERION

The Akaike Information Criterion (AIC henceforth) can be derived in multiple ways. Here I will present the simplest, using only basic results from maximum likelihood theory and some Taylor series expansions.

First, let us define a few terms.

The Kullback-Liebler divergence is an important quantity in much of information theory. It is defined as:

$$D(p||q) = \int_x p(x) log \frac{p(x)}{q(x)} dx$$

This can be divided into two parts:

$$S(p) = -\int_x p(x) log p(x) dx$$

$$H(p,q) = -\int_x p(x) log q(x) dx$$

Which are entropy and cross-entropy respectively.

For the rest of the derivation, $p(x)$ will refer to the generating distribution of data while $q(x,\theta)$ will refer to a model distribution of the data, parametrized by θ. Let θ_0 be the "Least False" parametrization of the model, i.e. the one that truly minimizes cross-entropy, and $\hat{\theta}$ be the maximum-likelihood estimate for theta, given data drawn from the generating distribution.

We will also define the first and second derivatives of $log q(x,\theta)$:

$$u(x,\theta) = \frac{\partial log q(x,\theta)}{\partial \theta}$$

$$I(x,\theta) = \frac{\partial^2 log q(x,\theta)}{\partial \theta \partial \theta^T}$$

We also extract two quantities and their empirical estimates from the above:

$$K = Var_p u(X, \theta_0)$$

$$\hat{K_n} = \sum_{i=1}^{n} u(x_i, \hat{\theta})^T u(x_i, \hat{\theta})$$

$$J = -E_p I(X, \theta_0)$$

$$\hat{J_n} = -\sum_{i=1}^{n} I(x_i, \hat{\theta})$$

Given a set of models, the goal of the AIC is to find the model such that $\hat{\theta}$ inferred for that model based on the real data minimizes the following quantity.

$$H(p, q(\hat{\theta})) = -\int_x p(x) log q(x, \hat{\theta}) dx$$

One would hope that the empirical estimator of this quantity would be accurate:

$$\hat{H}_n(p, q(\hat{\theta})) = -\sum_{i=1}^{n} log q(x_i, \hat{\theta})$$

But unfortunately since the empirical distribution is the very same one we fit to in the first place, this is a biased estimator. In particular we will show that if the true generating distribution can be modeled by q, for some θ, this bias is equal to $k = dim(\theta)$, and otherwise it is equal to $Tr(J^{-1}K)$, so long as the central limit theorem for M-estimators holds:

$$\sqrt{n}(\hat{\theta} - \theta_0) \; N_k(0, J^{-1}KJ^{-1})$$

All we need to do to demonstrate this is use a couple of Taylor Series expansions on q

(25)

$$-H(p, q(\hat{\theta})) = \int_x p(x) log q(x, \hat{\theta}) dx \tag{26}$$

$$= \int_x p(x)[log q(x, \theta_0) + u(x, \theta_0)(\hat{\theta} - \theta_0) + 1/2(\hat{\theta} - \theta_0)^T I(x, \theta_0)(\hat{\theta} - \theta_0) \tag{27}$$

$$= -H(p, q_0) - 1/2(\hat{\theta} - \theta_0)^T J(\hat{\theta} - \theta_0) \tag{28}$$

(29)

$$-\hat{H}_n(p, q(\hat{\theta})) = \sum_{i=1}^{n} log q(x_i, \hat{\theta}) \tag{30}$$

$$= \sum_i p(x)[log q(x_i, \theta_0) + u(x_i, \theta_0)(\hat{\theta} - \theta_0) + 1/2(\hat{\theta} - \theta_0)^T I(x_i, \theta_0)(\hat{\theta} - \theta_0)$$

(31)

$$= -H(p, q_0) + \bar{Z}_n + \bar{U}_n^T(\hat{\theta} - \theta_0) - 1/2(\hat{\theta} - \theta_0)^T J_n(\hat{\theta} - \theta_0) \tag{32}$$

Since J_n converges to J, we are left with

$$-\hat{H}_n(p, q(\hat{\theta})) + H(p, q(\hat{\theta})) = \bar{Z}_n + \bar{U}_n^T(\hat{\theta} - \theta_0)$$

This is the amount that the cross-entropy we estimate from our empircal distribution, fitted with a maximum likelihood estimate of $\hat{\theta}$, exceeds, in expectation, the cross entropy we should have under our true distribution and the estimated $\hat{\theta}$.

$\bar{Z}_n = \sum_{i=1}^{n} logq(x_i, \theta_0 + H(p, q_0)$ and has mean zero.

$\bar{U}_n^T = \frac{1}{n}\sum_{i=1}^{n} u(x_i, \theta_0)$ $\frac{1}{\sqrt{n}} N_k(0, K)$

Re-write $\sqrt{n}\bar{U}_n^T = U`$ and you get

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow J^{-1}U` = N_k(0, J^{-1}KJ^{-1})$$

$$\bar{U}_n^T(\hat{\theta} - \theta_0) = \frac{1}{n}U^{`T}J^{-1}U` = Tr(J^{-1}K)$$

The Takeuchi Information Criterion[7] penalty. When our model q can represent the true model p, K = J and $Tr(J^{-1}K) = Tr(I_k) = k$

VII. TIC AND CALORIMETRY

A. Takeuchi Information Criterion

Let me begin with a section of the other set of LaTex notes, slightly fleshed out.

There is a very general correction to the AIC [Called the TIC, or Takeuchi Information Criterion] that can be made based on its derivation. In the ideal case, we expect the penalty to be k, the number of parameters. However, this actually results from a term in the Taylor series of the difference of two nested model likelihoods, that term being [I won't go through the full derivation]:

$$Tr\left[\hat{J}^{-1}\hat{K}\right]$$

$$\hat{J} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta^T}\ell(y_i, \hat{\theta}_i) \tag{33}$$

$$= \frac{1}{n}\sum_{i=1}^{n} I(y_i, \hat{\theta}_i) \tag{34}$$

$$\tag{35}$$

Also called the Fisher Information Matrix, expected value of the second derivative at the maximum log likelihood

$$\hat{K} = \frac{1}{n}\sum_{i=1}^{n}(\frac{\partial}{\partial\theta}\ell(y_i,\hat{\theta}_i))(\frac{\partial}{\partial\theta}\ell(y_i,\hat{\theta}_i))^T \tag{36}$$

$$= \frac{1}{n}\sum_{i=1}^{n} u(y_i,\hat{\theta}_i)u(y_i,\hat{\theta}_i)^T \tag{37}$$

$$\tag{38}$$

Where $u(y_i,\hat{\theta}_i)$ is the score vector of first derivative of the maximum log likelihood. The trace form of this given above equivalent to the $u^T J^{-1} u$ form that may be more familiar.

B. The Calorimetry Trick for True Log-Likelihood

Consider a log-likelihood of a sample y_i that involves both discrete parameters (perhaps indicating group memberships in a stochastic block model) and continuous terms, denoted by g and Θ respectively. MCMC[5] is run and a max-likelihood $(\hat{g},\hat{\theta})$ pair is found. Unfortunately, we can't apply AIC, TIC, BIC or any Blank IC that I know of to this log-likelihood since it has discrete parameters. The solution? Integrate over all possible discrete variable settings, conditioned upon $\hat{\theta}$ as the value of the continuous parameters. The likelihood integrated over the discrete variables is the partition function $Z(\beta)$ at $\beta = 1$ of the Gibbs distribution from which we sample the discrete variables.

The probability density of $\ell(y|g,\hat{\theta})$ under the Gibbs distribution is given by:

$$P(y|g,\hat{\theta}) = \frac{e^{\beta\ell(y|g,\hat{\theta})}}{\sum_g e^{\beta\ell(y|g,\hat{\theta})}}$$

Where the denominator is the partition function

$$Z(\beta) = \sum_g e^{\beta\ell(y|g,\hat{\theta})}$$

The calorimetry trick comes in when we want to estimate Z directly.

$$\frac{\partial}{\partial\beta} lnZ(\beta) = \frac{\frac{\partial}{\partial\beta} Z(\beta)}{Z(\beta)} \tag{39}$$

$$= \frac{\sum_g \ell(y|g,\hat{\theta}) e^{\beta\ell(y|g,\hat{\theta})}}{Z(\beta)} \tag{40}$$

$$= \sum_g \ell(y|g,\hat{\theta}) P(y|g,\hat{\theta}) \tag{41}$$

$$=< \ell(y|g,\hat{\theta}) >_\beta \tag{42}$$

$$= E_{\hat{\theta},\beta}[\ell(y)] \tag{43}$$

$$\tag{44}$$

Where the second to last line and last line simply indicate the expectation of $\ell(y|g,\hat{\theta})$ over this gibbs distribution (the notation is kind of arbitrary, and probably doesn't fit with convention totally; sorry).

We go through the derivative above so that we can estimate

$$lnZ(1) = lnZ(0) + \int_0^1 \frac{\partial}{\partial\beta} lnZ(\beta)\partial\beta$$

lnZ(1) is the log our partition function at temperature 1 of the distribution of discrete variables with fixed continuous variables. It is also equal to finding the marginal log probability of $\hat{\theta}$. Note that below I will in general drop the β everywhere, as it will be fixed to 1.

C. Combining the two

Now what? It's time to take the TIC from section 1 and apply it to lnZ(1) from section 2. This is literally just applying the TIC to our probability distribution of interest, as we would under any normal circumstance. However, it turns out that there's a surprisingly neat form for our distribution.

$$Tr[\hat{J}^{-1}\hat{K}] = u^T\hat{J}^{-1}u \tag{45}$$
$$= \sum_i \frac{\partial}{\partial\theta}\ell(y_i,\hat{\theta}_i)^T[\sum_i \frac{\partial^2}{\partial\theta\partial\theta^T}\ell(y_i,\hat{\theta}_i)]^{-1}\sum_i \frac{\partial}{\partial\theta}\ell(y_i,\hat{\theta}_i) \tag{46}$$
$$= \sum_i \frac{\partial}{\partial\theta}lnZ_i^T[\sum_i \frac{\partial^2}{\partial\theta\partial\theta^T}lnZ_i]^{-1}\sum_i \frac{\partial}{\partial\theta}lnZ_i \tag{47}$$
$$\tag{48}$$

Drop the $\sum_i$ and the subscript since we'll generally be dealing with samples of size 1 (but with enough information to estimate well anyways, hopefully).

All that remains before we have our Takeuchi Information Criterion for Gibbs Distributions is to obtain the first and second derivatives of which it consists. Keep in mind that the first derivative will be a vector of partial derivatives of length m (the number of continuous parameters) and the second derivative will be an m by m matrix of partial derivatives.

For the purpose of compactness, we will make a couple abbreviations. $\frac{\partial}{\partial\theta}\ell(y|g,\hat{\theta})$ will be called $\ell_g(y)'$ and $\frac{\partial^2}{\partial\theta\partial\theta^T}\ell(y|g,\hat{\theta})$ will be called $\ell_g(y)''$

$$\frac{\partial}{\partial\theta}lnZ = \frac{\partial}{\partial\theta}ln\sum_g e^{\ell(y|g,\hat{\theta})} \tag{49}$$
$$= \sum_g \frac{\frac{\partial}{\partial\theta}e^{\ell(y|g,\hat{\theta})}}{\sum_g e^{\ell(y|g,\hat{\theta})}} \tag{50}$$
$$= \sum_g \frac{\frac{\partial}{\partial\theta}\ell(y|g,\hat{\theta})e^{\ell(y|g,\hat{\theta})}}{\sum_g e^{\ell(y|g,\hat{\theta})}} \tag{51}$$
$$= \sum_g \frac{\ell(y)'_g e^{\ell(y|g,\hat{\theta})}}{\sum_g e^{\ell(y|g,\hat{\theta})}} \tag{52}$$
$$= E_{\hat{\theta}}[\ell_g(y)'] \tag{53}$$
$$\tag{54}$$

The second derivative is a bit more trouble, but not much.

$$\frac{\partial^2}{\partial\theta\partial\theta^T}lnZ = \frac{\partial^2}{\partial\theta\partial\theta^T}\ell(y_i,\hat{\theta}) \tag{55}$$

$$= \frac{\partial}{\partial\theta^T}\sum_g \frac{\ell_g(y)' e^{\ell(y|g,\hat{\theta})}}{\sum_g e^{\ell(y|g,\hat{\theta})}} \tag{56}$$

$$= \sum_g \ell_g(y)' \frac{\partial}{\partial\theta^T}\frac{e^{\ell(y|g,\hat{\theta})}}{\sum_g e^{\ell(y|g,\hat{\theta})}} + \sum_g \frac{\partial}{\partial\theta^T}\ell_g(y)' \frac{e^{\ell(y|g,\hat{\theta})}}{\sum_g e^{\ell(y|g,\hat{\theta})}} \tag{57}$$

$$= \sum_g \ell_g(y)' \frac{\partial}{\partial\theta^T}\frac{e^{\ell(y|g,\hat{\theta})}}{\sum_g e^{\ell(y|g,\hat{\theta})}} + E_{\hat{\theta}}[\ell_g(y)''] \tag{58}$$

$$\tag{59}$$

This is a little exhausting to look at, but it's just the chain rule that results from taking the set of partial derivatives of the first set of partial derivatives. To calculate the term that wasn't simplified:

$$\sum_g \ell_g(y)' \frac{\partial}{\partial\theta}\frac{e^{\ell(y|g,\hat{\theta})}}{\sum_g e^{\ell(y|g,\hat{\theta})}} = \sum_g \ell_g(y)'(\ell_g(y)'^T \frac{e^{\ell(y|g,\hat{\theta})}}{\sum_g e^{\ell(y|g,\hat{\theta})}} \tag{60}$$

$$- e^{\ell(y|g,\hat{\theta})}\frac{\sum_g \ell_g(y)'^T e^{\ell(y|g,\hat{\theta})}}{(\sum_g e^{\ell(y|g,\hat{\theta})})^2}) \tag{61}$$

$$= \sum_g \ell_g(y)'\ell_g(y)'^T \frac{e^{\ell(y|g,\hat{\theta})}}{\sum_g e^{\ell(y|g,\hat{\theta})}} \tag{62}$$

$$- \sum_g \ell_g(y)' \frac{e^{\ell(y|g,\hat{\theta})}}{\sum_g e^{\ell(y|g,\hat{\theta})}} \sum_g \ell_g(y)'^T \frac{e^{\ell(y|g,\hat{\theta})}}{\sum_g e^{\ell(y|g,\hat{\theta})}} \tag{63}$$

$$= E_{\hat{\theta}}[\ell_g(y)'\ell_g(y)'^T] \tag{64}$$

$$- E_{\hat{\theta}}[\ell_g(y)']E_{\hat{\theta}}[\ell_g(y)']^T \tag{65}$$

$$= Var_{\hat{\theta}}[\ell_g(y)'] \tag{66}$$

$$\tag{67}$$

Again, the main step here is the chain rule for division. Putting all these terms back together, we get the following:

$$Tr[\hat{J}^{-1}\hat{K}] = E_{\hat{\theta}}[\ell_g(y)']^T[Var_{\hat{\theta}}[\ell_g(y)'] + E_{\hat{\theta}}[\ell_g(y)'']]^{-1}E_{\hat{\theta}}[\ell_g(y)']$$

All of which is easily calculable with reasonable accuracy during MCMC (at $\beta = 1$) for a broad range of likelihood functions $\ell(y|g, \hat{\theta})$. This can also integrate out not just discrete variables but any subset of variables.

D. Applying TIC to networks

Naively, we would treat our whole network as a single data point. This would give us:

$$\ell(\Omega|G) = \sum_{i,j} A_{ij} log(\theta_i \theta_j \omega_{g_i g_j}) - \theta_i \theta_j \omega_{g_i g_j}$$

The empirical fisher information matrix is a kxk X kxk matrix with 0's at all non-diagonals

$$\hat{J}_{rs,rs} = E_{\hat{\theta}}[\sum_{i,j} -\frac{A_{ij}}{\omega_{g_i g_j} \omega_{g_i g_j}}] + Var_{\hat{\theta}}[\sum_{i,j} \frac{A_{ij}}{\omega_{g_i g_j}} - \theta_i \theta_j]$$

And the variance must be 0, assuming we have accurately found $\hat{\theta}$ relative to the Gibbs distribution (which is not actualy the case with most tractable inference methods) since we only have one sample! How distressing.

$$\hat{K} = 0$$

$$Tr[\hat{J}^{-1} \hat{K}] = 0$$

There are two primary means of recourse. The first, which won't be investigated here, is some method of parametric bootstrapping or cross-validation. The second is to treat components of our network as independent samples, an idea for which there is some precedent in Belief Propagation[8], a message-passing algorithm that tracks distributions over pairwise marginals. Below, m is the number of edges in G. Now, given a fixed $\hat{\Omega}$ (which all the instances of ω_{rs} below will be components of), these quantities can be inferred.

$$\hat{J}_{rs,rs} = \frac{1}{m} \sum_{i \in r, j \in s} E_{\hat{\theta}}[-\frac{A_{ij}}{\omega_{g_i g_j} \omega_{g_i g_j}}] + Var_{\hat{\theta}}[\frac{A_{ij}}{\omega_{g_i g_j}} - \theta_i \theta_j]$$

$$\hat{K} = \frac{1}{m} \sum_{i,j} E_{\hat{\theta}}[fracA_{ij}\omega_{g_i g_j} - \theta_i \theta_j] E_{\hat{\theta}}[fracA_{ij}\omega_{g_i g_j} - \theta_i \theta_j]^T \tag{68}$$

$$- (\frac{1}{m} \sum_{i,j} E_{\hat{\theta}}[fracA_{ij}\omega_{g_i g_j} - \theta_i \theta_j])(\frac{1}{m} \sum_{i,j} E_{\hat{\theta}}[fracA_{ij}\omega_{g_i g_j} - \theta_i \theta_j])^T \tag{69}$$

Plug these in to the TIC and you have the TIC for networks. Despite being moderately principled, however, it works only marginally better than the AIC and is prone to large errors. Even when it is accurately inferred, it doesn't work as well as it seems it should. We speculate that this is because the independence assumption on the marginals is too strong to allow accurate inference of expectations of first and second derivatives and variance over those derivatives accurately.

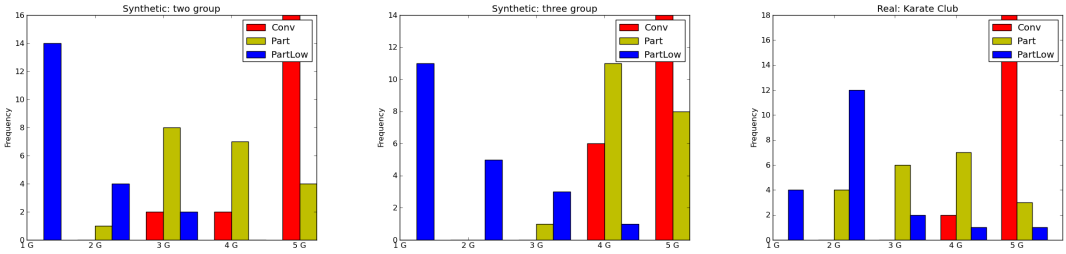
VIII. EXPERIMENTAL RESULTS

Results pictured below. The primary conclusion is that allowing discrete and continuous variables to jointly converge greatly overfits, and that this approach, though used in the past,[3], is not appropriate, while using calorimetry to infer a partition function also overfits, but less. Between different "IC" methods, the Bayesian Information Criterion[6] (not previously discussed, but popular in some fields) works best.

In the set of figures labeled AIC, Conv is joint discrete, continuous parameter optimization fit Part is way of estimating the partition, Z, that may slightly overestimate. PartLow estimates the partition in a way that is guaranteed to underestimate. The numbers along the x axis are the inferred most likely community numbers. The numbers along the y axis are the frequency with which those number of communities were inferred.

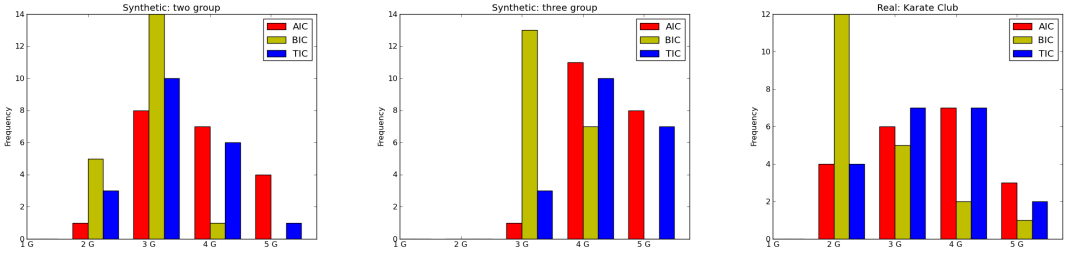
In the set of figures labeled Partition, AIC uses the Akaike Information Criterion, BIC the Bayesian Information Criterion, TIC the Takeuchi Information Criterion as described above.

The figures may look slightly impenetrable, but what you should look for is which color of bar (corresponding to a criterion or baseline of ℓ) concentrates around which estimate of group size.



(a) 2-group Synthetic (b) 3-group Synthetic (c) Karate Club

FIG. 1: AIC performance with different baseline ℓs



(a) 2-group Synthetic (b) 3-group Synthetic (c) Karate Club

FIG. 2: Different IC's performance with same baseline ℓ

IX. CONCLUSION

We have gone over the effect of degree correction in the degree-corrected Stochastic Block Model, and how it overestimates likelihood in the toy, one-group case as predicted by the AIC, with a mild correction dependent on sparsity. We have also discussed how to marginalize over discrete parameters, derived the AIC, derived its variant, the TIC in the specific case of networks and shown results. Unfortunately, overfitting is still a big problem. Partially, this may be due to numerical errors that can enter the MCMC inference process at several points. Using an alternative inference method like Belief propagation would help with this. It is impossible to rule out that the patterns of dependency in the data may make inference much more difficult than imagined, or slow down the convergence rate predicted by the central limit theorem. More experimentation and theoretical work must be done to

improve confidence in these results.

---

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, December 1974.

[2] Gerda Claeskens and Nils Lid Hjort. *Model Selection and Model Averaging*. Number 9780521852258 in Cambridge Books. Cambridge University Press, 2008.

[3] P S Gopalakrishnan. Clustering via the bayesian information criterion with applications in speech recognition. *Proceedings of the 1998 IEEE International Conference on Acoustics Speech and Signal Processing ICASSP 98 Cat No98CH36181*, 2:645–648, 1998.

[4] Brian Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107+, January 2011.

[5] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1091, 1953.

[6] Gideon Schwarz. Estimating the Dimension of a Model. *Annals of Statistics*, 6(2):461–464, 1978.

[7] K Takeuchi. Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku (Mathematical Sciences)*, 153:12–18, 1976.

[8] Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. *Understanding belief propagation and its generalizations*, pages 239–269. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003.