Inference In Random Geometric Graphs

Răzvan - Dumitru Meriniuc[1]

Machine Learning MSc

Peter Orbanz

Submission date: 11$^{\text{th}}$ of September 2020

[1]**Disclaimer:** This report is submitted as part requirement for the Machine Learning MSc at UCL. It is substantially the result of my own work except where explicitly indicated in the text. The report may be freely copied and distributed provided the source is explicitly acknowledged.

Abstract

The graph is a ubiquitous data structure, arising in numerous real-world scenarious, ranging from social networks to drug design. While relatively simple, it is difficult to exploit it in the context of machine learning, which prompts one to find ways of encoding it in a geometric space with a well-defined metric. Our work aims to study how feasible it is to treat the embeddings of the nodes of a graph as latent variables, in the context of random geometric graphs (RGG) and Bayesian statistics. By placing a prior on the embeddings and drawing an edge between two points randomly, based on the distance between them, we intent to approximate the posterior of the embeddings, given the graph, which would allow us to better describe the geometry capturing the properties of the graph.

Contents

Chapter 1

Introduction

The graph is a ubiquitous data structure, which arises in numerous real-world scenarios, ranging from social networks to drug design. Its relative simplicity allows one to attach a myriad of semantics to the nodes and edges, hence its prevalence. However, its format limits the scope of machine learning techniques one can use on such data, but embedding its vertices in a geometry, be it euclidean or with a non-zero curvature, allows us to extend the range of methods we can use while modelling the data. This idea arises naturally, as many networks coming from physical considerations are governed by an underlying geometry, such as the road network in a country.

We are going to focus on data given as a simple homogeneous graph, i.e unweighted, undirected graph that contains no self-loops and no multiple edges[1]. The homogeneity refers to the edges representing the same concept. Mathematically, we define it as a pair $G = (V, E)$, where V is the set of vertices, also called nodes, and E is the set of edges. We will just represent the vertices as consecutive natural numbers, so $V = 1, 2, \ldots N$, and the edges as unordered pairs $\{i, j\}$, $i \neq j$, $i, j \in V$. Such a graph can be completely defined by its adjacency matrix, A, $A_{ij} = 1$ if $\{i, j\} \in E$ and 0 otherwise, as we are working with

unweighted edges, so we are interested whether an edge exists or not. Since the graphs is undirected, $A_{ij} = A_{ji}$.

The aim of this project is to study the latent geometric space in a Bayesian setting. The literature generally focuses on empirical embeddings (NLP), targetting the individual nodes, while we are more interested in the underlying geometry. A random geometric graph (RGG) is usually generated by sampling points from some space, representing the nodes, and connect each pair by an edge if the distance between them is less than some fixed threshold. In order to avoid this step function and the extra hyperparameter, we generate an edge stochastically, again relying on the distance between the points. The closer two points are, the greater the chance of them being connected.

In the literature, embeddings of nodes, edges and even whole subgraphs have been studied. We are going to focus on embedding of nodes only. They are generally studied from a graph theoretical or algorithmic point of view, the subject of embeddings in machine learning focusing more on how well they work for specific tasks and empirically determined. We aim, however, to describe in probabilistic terms the distribution of the embeddings and perform inference, which has not been studied so far, to our knowledge.

Chapter 2

Background And Related Work

2.1 Bayesian Statistics And Variational Inference

We are going to work in the framework of Bayesian statistics, which relies on the Bayesian interpretation of probabilities, where the probability conveys the degree of belief in an event, which may rely on prior knowledge about the event, whether it is a personal belief or based on results of previous experiments. This is in contrast to the frequentist interpretation, which views probability as the limit of the relative frequency of an event after a great number of trials.

The cornerstone of Bayesian statistics is Bayes' formula. If X is some data generated by a process with parameter θ, then

DEFINITION 1 (Bayes' Formula, Posterior, Likelihood, Prior, Evidence)

$$\mathbb{P}(\theta \mid X) = \frac{\mathbb{P}(X \mid \theta)\mathbb{P}(\theta)}{\mathbb{P}(X)}$$

1. *The **posterior** distribution, $\mathbb{P}(\theta \mid X)$, represents our updated belief about the param-*

eter θ after observing the data X.

2. *The **likelihood** distribution, $\mathbb{P}(X \mid \theta)$, denotes the probability of the data being generated by the given the parameter θ.*

3. *The **Prior** distribution, $\mathbb{P}(\theta)$, quantifies our assumption about the parameter θ.*

4. *The **Evidence** distribution, $\mathbb{P}(X)$, represents the probability of the data occurring, which is computed by marginalising out the parameter θ.*

The central Bayesian principle consists of placing a probability over the parameters, thus giving rise to the prior. It is asymptotically equivalent to the frequentist method, the parameter converging to the true one, but the downside is that there is no mathematical decision theory one can rely on when choosing the prior. It is totally up to the practitioner, which allows us to insert our belief into the model, thus forcing us to be honest about the assumptions we make regarding the data and learning process.

Statistical inference is the process of using data analysis to deduce properties of an underlying probability distribution. For example, if we model some data according to a Gaussian distribution with a known variance, we can look at the data and infer what the mean is, and then use it to make predictions on new input points. When predicting, we could work with the parameter for which the data is most likely to occur, an estimate known as MAP (Maximum a posteriori), or average over the parameters in a Bayesian setting, by marginalisation, such that each element of the average has a weight represented by the posterior:

$$\mathbb{P}(\mathbf{X^*} \mid \mathbf{X}) = \int_\Theta \mathbb{P}(\mathbf{X^*}, \theta \mid \mathbf{X}) d\theta = \int_\Theta \mathbb{P}(\mathbf{X^*} \mid \theta, \mathbf{X}) \mathbb{P}(\theta \mid \mathbf{X}) d\theta \tag{2.1}$$

$$= \int_{\Theta} \mathbb{P}(\mathbf{X^*} \mid \theta)\mathbb{P}(\theta \mid \mathbf{X})d\theta = \mathbb{E}_{\mathbb{P}(\theta|\boldsymbol{X})}[\mathbb{P}(\boldsymbol{X^*} \mid \theta)],$$

where $\mathbf{X}$ is the observed data, $\mathbf{X^*}$ are the points for which we are trying to predict a quantity of interest, Θ is the parameter space and θ is an arbitrary parameter. $\mathbb{P}(\mathbf{X^*} \mid \theta, \mathbf{X})$ becomes $\mathbb{P}(\mathbf{X^*} \mid \theta)$ because the model is fully specified given the parameters, the known points $\boldsymbol{X}$ adding no information.

This integral requires a closed form of the posterior, which is often intractable because we cannot calculate the evidence present in the denominator in **Definition 1**, $\mathbb{P}(\boldsymbol{X})$. Even if it can be calculated, the posterior is not usually the probability density function a known distribution, which would make it more difficult to study and use. A possible solution would be to approximate it with the average $\frac{1}{N}\sum_{i=1}^{N} \mathbb{P}(\boldsymbol{X^*} \mid \theta_i)$, where $\{\theta_i\}_{i=1}^{N}$ are the samples from $\mathbb{P}(\theta \mid \boldsymbol{X})$. The most popular class of sampling algorithms is Markov Chain Monte Carlo (MCMC) . For example, the Metropolis - Hastings algorithm allows us to circumvent the evidence, which cancels out. One of the drawbacks of this method is that it is non-deterministic and, in addition, we would not be able to measure how good our approximation is. Furthermore, there is the problem of choosing the hyperparameters, parameters of the model that are not governed by a prior, as they are just scalars whose value we need to determine. The classical approach aims to maximise the probability of the observed data, $\mathbb{P}(\boldsymbol{X})$, which is the term that generates the intractable integral preventing us from calculating the posterior in the first place. We could again employ sampling to address this issue, adding an extra layer of randomness to our implementation, which is undesirable. This technique we are about to present deals with the approximation of the posterior with a new variational distribution, offering a measure of *distance* between the distributions, thus allowing us to determine how good our approximation is, which means that there is some quantity that tells us if we are working in the right direction. The

setting of this approach offers a lower-bound of the evidence as a by-product, while we are making the approximation better, which is where the beauty of this method lies. While it does not provide a hard maximum for the evidence, we get the next best thing, a lower bound.

The measure between distributions is called the Kullback - Leibler divergence , which is defined for any two distributions that share the support of the random variable.

DEFINITION 2 (Kullback - Liebler Divergence)

$$KL(q(u) \,||\, \mathbb{P}(u)) = \int q(u) log \frac{q(u)}{\mathbb{P}(u)} du$$

Please note that this measure is asymmetric and, following Gibbs' inequality , non-negative , equalling 0 only when the two distributions are exactly the same.

In our case, we would like to approximate the posterior $\mathbb{P}(\theta \mid \boldsymbol{X})$ with a new distribution $q(\theta)$, having free parameters, that would behave as if it has seen the data. The advantage is that we can pick whatever variational distribution we want, allowing us to insert extra assumptions that would make calculations easier. This means that we can keep the model *pure* and add simplifications in the approximation. We would thus like to minimise the KL divergence between the two distributions. Because calculating it involves working with the original posterior, which posed problems from the beginning, we rely on the following result, which rephrases the problem as maximising a different term, directly implying the minimisation of our divergence.

$$log \mathbb{P}(\mathbf{X}) = KL(q(\theta) || \mathbb{P}(\theta \mid \mathbf{X})) + \mathcal{F}(\theta) \tag{2.2}$$

$$\mathcal{F}(\theta) = \mathbb{E}_{q(\theta)}[log\mathbb{P}(\mathbf{X}, \theta)] - H(q(\theta)) \tag{2.3}$$

where $H(q(\theta))$ is the entropy of the variational distribution and $\mathcal{F}$ stands for free energy, a concept coming from thermodynamics. This result is deduced in Appendix, section A.1.

We can see how maximising $\mathcal{F}$ implies minimising the KL term. Because the divergence is non-negative, we get that

$$log\mathbb{P}(\boldsymbol{X}) \geq \mathcal{F}(\theta) \tag{2.4}$$

which acts as a lower bound for our evidence. Alternatively, we can deduce the same result, by-passing the KL term, using Jensen's inequality applied to the expected value, as shown in section A.1 of the Appendix.

Please note that we have not yet touched the hyperparameters of either distribution. Assuming we could somehow compute $KL(q(\theta)||\mathbb{P}(\theta \mid \boldsymbol{X}))$ and pick the parameters that minimise it, it is not exactly obvious how doing this would provide a better chance for our data, expressed as a lower bound in our case, since it looks like this just gets the distributions closer disregarding the other terms. And this is the beauty of this technique. That striving for a better approximation generates a better lower bound for the data, addressing both aspects of the model that raised problems, mentioned at the beginning of this section. In practice, we optimise both sets of hyperparameters at the same time to maximise $\mathcal{F}$, which we should be able to compute effectively. We choose the variational distribution q, so we ought to get a closed form of its entropy, while $log\mathbb{P}(\boldsymbol{X}, \theta)$ is the model itself, and if we cannot express it, then we know nothing. Thus, we transform this Bayesian inference problem into an optimisation one.

In some cases, such as the model we are proposing in **Chapter 3**, we run into terms of the free energy we cannot calculate and have to rely on approximations through sampling. While not ideal, this is done in the same setting and we can still measure how good our approximation of the posterior is, albeit this *distance* is just an estimate as well.

Prediction is performed by approximating $\mathbb{P}(\theta \mid \boldsymbol{X})$ with $q(\theta)$ in the integral at (1.1), thus

$$\mathbb{E}_{q(\theta)}[\mathbb{P}(\boldsymbol{X}* \mid \theta)] \approx \mathbb{E}_{\mathbb{P}(\theta|\boldsymbol{X})}[\mathbb{P}(\boldsymbol{X}* \mid \theta)] \tag{2.5}$$

2.2 Random Geometric Graphs

The beginning of random graph theory is usually attributed to the publication of the seminal papers of Erdős and Rényi[reference] introducing the standard random graph model, which studies the behaviour of a random graph with n nodes and N edges, placing an uniform prior on all graphs that respect this requirement. However, around the same period of time, Gilbert [reference] proposed a different random graph model, which relies on a geometric layout underpinning the nodes, edges being determined based on the relative position of vertices. We are going to refer to the latter model by *random geometric graph*.

A Poisson process was employed to model the geometric points, with the idea that they ought to be spread uniformly around the plane with a positive density.

DEFINITION 3 *A Poisson process $\mathcal{P}$ with density one in $\mathbb{R}^2$ is a random subset of $\mathbb{R}^2$*

such that:

1. *The number of points in a measurable set A is governed by a Poisson distribution with the mean equal to the Lebesgue measure of A*

2. *For two disjoint, measurable sets $A, B \subset \mathbb{R}^2$, the number of points they contain are independent from each other*

In the original model, points are picked in $\mathbb{R}^2$ according to a Poisson process with density one, connecting pairs of points if the distance between them is at most R. So, according to the first property of the definition above, the expected number of neighbours a point has is $N = \pi R^2$. A final version of this model can be constructed by restricting the space to a square of area n, ensuring the number of points, which is a Poisson random variable, has average n. There are many results studying the asymptotic behaviour as $n \Rightarrow \infty$ and algorithms to recover points with a certain error (reference paper found). Spherical RGG and Erdos [paper?].

Another level of randomness can be added, producing a *soft random geometric graph*, which does not rely on a threshold value, but uses a connection function, ϕ, which takes as a parameter the distance between two points, the output becoming the mean a Bernoulli random variable which determines the existence of an edge. This was introduced by Waxman [reference], using a stretched exponential, $\phi(d) = \beta e^{-\frac{d}{d_0}}$. This can be further generalised by introducing decay in the distance, $\phi(d) = \beta e^{-\left(\frac{d}{d_0}\right)^{\eta}}$. We have run experiments for $\eta \in \{1, 2\}$.

These structures have been mostly analysed in the context of graph theory, their asymptotic behaviour and connection to the Erdős-Rényi graph, and not in as determining the

latent space for a graph in the context of inference in machine learning, to our knowledge, so this is what we aim to study.

2.3 Exchangeability

Is our model not Bayesian because the graphon is deterministic?

One can view statistical inference as the procedure of extracting a pattern, characterised as the parameter of the model, from observed data. This leads to understanding the randomness in the data source as the underlying pattern combined with the sample randomness. When such a common pattern exists and the data is not completely beclouded by the sample randomness, exchangeability properties supply criteria for when extracting the underlying patterns is possible.

In practice, when working with a model in a Bayesian manner, we usually make use of Bayes' formula in the following form

DEFINITION 4 (Bayes' Formula, Posterior, Likelihood, Prior, Evidence)

$$\mathbb{P}(d\theta \mid X_{1:N}) \overset{a.s.}{=} \frac{\prod_{i=1}^{N} p_\theta(x_i)}{\int_{\boldsymbol{T}} \prod_{i=1}^{N} p_{\theta'}(x_i)\mathbb{P}(d\theta')}\mathbb{P}(d\theta)$$

where $\mathbb{P}$ is a prior distribution, $\boldsymbol{T}$ is the parameter space of θ, p_θ is a probability density function and $X_{1:N}$ represents N observations.

We make considerable assumptions, which consist not only of the choice of likelihood density and prior distribution, but also that, given a parameter θ from the random variable Θ, the joint likelihood of the observations factorises, so

$$p_\theta(x_1, \ldots, x_N) = \prod_{i=1}^{N} p_\theta(x_i) \tag{2.6}$$

Our assumption is that the observations $X_1, \ldots, X_N$ independent and identically distributed given Θ, which makes them **conditionally independent**. We can mathematically express that as

$$\mathbb{P}(X_{1:N} \in dx_1 \times \cdots \times dx_N \mid \Theta) \overset{a.s.}{=} \prod_{i=1}^{N} \mathbb{P}(X_i \in dx_i \mid \Theta) \tag{2.7}$$

This assumption of conditional independence sits at the core of Bayesian modelling, implying that, given Θ, the randomness attributed to the observations decouples entirely, so all the joint information in the sampled data is enclosed in Θ, which becomes our quantity of interest we would like to extract from the observations. Exchangeability then seeks to provide the context in which we may assume that conditional independence, given some random quantity, is met.

DEFINITION 5 (Exchangeability) *A random sequence $X_{1:\infty} = (X_1, X_2, \ldots)$ is said to be **exchangeable** if the order in which the values X_i are observed is irrelevant to their joint distribution. Mathematically, we write this as*

$$(X_1, X_2, \ldots) \overset{d}{=} (X_{\pi(1)}, X_{\pi(2)} \ldots), \quad \forall \textit{bijections } \pi : \mathbb{N} \to \mathbb{N} \tag{2.8}$$

We can see that an i.i.d. sequence is obviously exchageable, since their distribution is a product, which commutes. The same holds for a conditionally i.i.d. sequence, given some Θ, since it is a product like in (2.8), so the set of conditionally i.i.d. sequences is contained in the set of exchageable sequences. The seminal result by de Finetti surprisingly proves that these two sets are actually equal, so a sequence $X_{1:\infty}$ is exchangeable if and only if it

is conditionally i.i.d, for some Θ.

Defining $P_\theta(\bullet) := \mathbb{P}(X_i \in \bullet \mid \Theta = \theta)$, we have a family of measures

$$M := \{P_\theta \mid \theta \in \boldsymbol{T}\} \tag{2.9}$$

Because Θ is random, P_Θ is a random varible in the set of all probability measures on the sample space $\boldsymbol{X}$, $\mathbf{PM(X)}$, thus a random probability measure. Instead of thinking about Θ as a random parameter for the distribution governing our sequence, we may view it as a random probability measure, so $\mathbf{T} = \mathbf{PM}(\mathbf{X})$, so that $P_\theta = \theta$. Abbreviating the factorial distribution as

$$P_\theta^\infty(dx_1 \times dx_2 \dots) = \prod_{i \in \mathbb{N}} P_\theta(dx_i) \tag{2.10}$$

Let us now introduce de Finetti's Theorem.

THEOREM 6 (de Finetti) *An infinite random sequence $X_{1:\infty}$ is exchangeable if and only if there exists a random probability measure Θ on $\boldsymbol{X}$, such that*

$$\mathbb{P}(X_{1:\infty} \in \bullet \mid \Theta) \overset{a.s.}{=} \Theta^\infty(\bullet) \tag{2.11}$$

The theorem usually appears in the form below, which is a direct consequence of eq above, obtained by marginalising out the random probability measure Θ, leading to equality in expectation only. We have that exchangeability implies eq below, but not the other way around. The right-hand side of the equation below is a mixture, which means that it can be sampled in two stages, by first generating $\Theta \sim \eta$ and then sampling $X_{1:N} \mid \Theta$ from Θ^∞, making $X_{1:\infty}$ condtionally i.i.d., which are already exchangeably.

COROLLARY 7 *A random sequence X is exchangeable if and only if*

$$\mathbb{P}(X \in \bullet) = \int_{\boldsymbol{PM}(\boldsymbol{X})} \theta^{\infty}(\bullet)\eta(d\theta) \tag{2.12}$$

for some distribution η on $\boldsymbol{PM}(\boldsymbol{X})$.

We can alternatively present de Finetti's theorem using random variables instead of distributions. For $\theta \in \mathbf{PM}(\mathbf{X})$ a probability measure on $\mathbf{X}$, we denote the i.i.d. random sequence sampled from θ as

$$X_\theta^0 := (X_1, X_2, \dots), \quad \text{where } X_1, X_2, \dots \overset{i.i.d}{\sim} \theta$$

so $\mathbb{P}(X_\theta^0) = \theta^\infty$. Exchangeability is obtained by randomising θ, i.e. if Θ is a random probability measure on $\mathbf{X}$, then X_Θ^0 is an exchangeable sequence. de Finetti's theorem is then restated as

$$X_{1:\infty} \text{ exchangeable } \Leftrightarrow X_{1:\infty} \overset{a.s.}{=} X_\Theta^0 \quad \text{for some } \Theta \in \boldsymbol{RV}(\boldsymbol{PM}(\boldsymbol{X})) \tag{2.13}$$

This approach will be useful in expressing the more advanced representation theorems presented below, which can be presented more elegantly using random variables instead of distributions.

2.3.1 Exchangeable Arrays

A d-**array** is a type of structure defined as a potentially infinite collection of variables indexed by d indices,

$$x := (x_{i_1,\ldots,i_d})_{i_1,\ldots,i_d \in \mathbb{N}}, \quad \text{where} x_{i_1,\ldots,i_d} \in \boldsymbol{X}_0 \tag{2.14}$$

The sequences previously discussed are 1-arrays and matrices can be seen as 2-arrays, with $\boldsymbol{X}_0$ an algebraic field, so that operations are well-defined. A simple graph, one with no multiple edges, can be represented by a 2-array, having $\boldsymbol{X}_0 = \{0, 1\}$, by its adjacency matrix. An undirected graph implies a symmetric matrix. We will continue to discuss only 2-arrays, but generalisations to generic d-arrays can be made as well (reference).

Let X be a random 2-array. We need to fix the components of X we are going to permute in order to define exchageability. Since the data is structured in rows and columns, it means that there is some semantic attributed to these sub-structures, otherwise we would just express the data as a regular sequence. Therefore, permuting should preserve the rows and columns, i.e. if two data points are located on the same column, they should still share that column after permuting them, regardless of the new order in which they appear, and similarly for rows. We thus reduce permutations of entries of X to permutations of its rows and columns, by either applying some permutation π to both rows and columns, or use separate permutations, π_r for rows and π_c for columns.

DEFINITION 8 *A random 2-array $X = (X_{ij})_{i,j \in \mathbb{N}}$ is said to be* ***jointly exchangeable*** *if*

$$(X_{ij}) \overset{d}{=} (X_{\pi(i),\pi(j)}), \quad \forall \textit{ bijections } \pi : \mathbb{N} \to \mathbb{N} \tag{2.15}$$

X is said to be ***separately exchangeable*** *if*

$$(X_{ij}) \stackrel{d}{=} (X_{\pi_r(i),\pi_c(j)}), \quad \forall \, bijections \, \pi_r, \, \pi_c : \mathbb{N} \to \mathbb{N} \tag{2.16}$$

Peter figure

We can simply generate a random matrix by defining a function f with its image equal to $\boldsymbol{X}_0$ and that has two arguments. Sampling two random sequences $(U_1, U_2, \dots)$ and $(V_1, V_2, \dots)$ and setting $X_{ij} := f(U_i, V_j)$ generates our random matrix. If the sequences (U_i) and (V_i) contain independent elements and are independent of each other, then (X_{ij}) is separately exchangeable. However, setting $X_{ij} := f(U_i, U_j)$, we generate a jointly exchangeable matrix. However, these are not all of the existing exchangeable arrays. We can, for example, add another argument to the function f, a random variable U_{ij}, without breaking the exchangeability, provided that the sequence (U_{ij}) is made of independent elements. The distributions governing the random variable we made use of is irrelevant, as they do not add any expressive power, so we are free to choose any convenient, simple distribution, such as the uniform distribution on [0, 1].

DEFINITION 9 *Let* $\boldsymbol{F}(\boldsymbol{X}_0)$ *be the space of measurable functions* $\theta : [0, 1]^3 \to \boldsymbol{X}_0$, (U_i) *and* (V_i) *two i.i.d. sequences and* (U_{ij}) *an i.i.d. 2-array, all elements being Uniform[0, 1] random variables. We define two random arrays,* J^0_θ *and* S^0_θ*, for any* $\theta \in \boldsymbol{F}$*, as*

$$J^0_\theta := \theta(U_i, U_j, U_{ij}) \tag{2.17}$$

$$S^0_\theta := \theta(U_i, V_j, U_{ij}) \tag{2.18}$$

Incredibly, these two random arrays play an analogous role to that of i.i.d. sequences,

in that any exchageable array can be generated by making θ random, result proven by Aldous and Hoover.

THEOREM 10 *Aldous, Hoover A random 2-array $X = (X_{ij})$ with entries in a Polish space $\boldsymbol{X}_0$ is jointly exchangeable if and only if*

$$X \stackrel{d}{=} J^0_\Theta, \quad \textit{for some} \, \Theta \in \boldsymbol{RV}(\boldsymbol{F}(\boldsymbol{X}_0)) \tag{2.19}$$

and separately exchangeable if and only if

$$X \stackrel{d}{=} S^0_\Theta, \quad \textit{for some} \, \Theta \in \boldsymbol{RV}(\boldsymbol{F}(\boldsymbol{X}_0)) \tag{2.20}$$

DEFINITION 11 *A **graphon** is a measurable function $w : [0,\, 1]^2 \to [0, 1]$.*

Its connection to exchangeable graphs arises when considering a symmetric adjacency matrix, thus working with simple, undirected graphs, as we can replace the three-argument function θ in J^0_θ with our graphon w, which only takes two arguments. For $\theta \in \Theta(\{0, 1\})$ and a uniform random variable U, $\theta(x, y, U) \in \{0,\, 1\}$, so we can define w as

$$w(x, y) := \mathbb{P}[\theta(x, y, U) = 1] \tag{2.21}$$

which makes it a measurable function $[0,\, 1]^2 \to [0,\, 1]$. For a fixed w, we can sample a random graph G^0_w as follows:

$$\begin{aligned} U_1,\, U_2,\, \ldots &\overset{iid}{\sim} Uniform[0, 1] \\ X_{ij} &\sim Bernoulli(w(U_i,\, U_j)) \end{aligned} \tag{2.22}$$

2.4 Quadratic Forms In Random Variables

For a random variable of dimension n, $X \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, we are interested in finding the probability distribution function of the quadratic form $Q = X^T X$. When our embeddings are governed by a multivariate Gaussian distribution, their difference will also be a multivariate Gaussian random variable, so the aforementioned Q represents the squared difference between two such i.i.d. embeddings.

Theorem 4.2b.1 in Mathai states that the density we are looking for is

$$f(u) = \sum_{k=0}^{\infty} (-1)^k c_k \frac{u^{\frac{n}{2}+k-1}}{\Gamma(\frac{n}{2}+k)}, \quad 0 < u < \infty \tag{2.23}$$

where

$$c_0 = \exp\left(-\frac{1}{2}\sum_{j=1}^{n} b_j^2\right) \prod_{j=1}^{n} (2\lambda_j)^{-\frac{1}{2}} \tag{2.24}$$

$$c_k = \frac{1}{k}\sum_{r=0}^{k-1} d_{k-1} c_r, \quad k \leq 1 \tag{2.25}$$

$$d_k = \frac{1}{2}\sum_{j=1}^{n} (1 - k b_j^2)(2\lambda_j)^{-k}, \quad k \leq 1 \tag{2.26}$$

$$\tag{2.27}$$

with λ_j being the j^{th} eigenvalue of Σ and $\boldsymbol{b} = P\Sigma^{-\frac{1}{2}}\boldsymbol{\mu}$, where P is the matrix with its columns equal to the eigenvectors corresponding to the eigenvalues $\{\lambda\}$, so

$$P^T \Sigma P = \text{diag}(\lambda_1, \ldots, \lambda_n) \tag{2.28}$$

More on this can be found in Mathai, chapters 3 and 4.

Chapter 3

RGG Inference

Let us denote the given data by $\mathcal{G} = \{X_{ij}\}_{i,j\in[N]}$, where $X_{ij} \in \{0, 1\}$ represents the existence of an edge between nodes i and j, assuming nothing that would induce and structure in the given graph. We model this data as arising from an SRGG, with the embeddings governed by an underlying Gaussian distribution. Let $X_i \in \mathbb{R}^n$ denote the embedding of vertex i. The random variable representing an edge, X_{ij} is then a Bernoulli random variable, with mean $e^{-\text{dist}(X_i,X_j)^2}$.

Use distance only -> Chi Distribution

Assume the prior embeddings are generated by a multivariate Gaussian with mean $\mathbf{0}$, which is chosen to facillitate calculations, but is otherwise irrelevant, since the distance between points ignores the global position of the points, and covariance matrix $\mathcal{I}$. We are interested in approximating the posterior $\mathbb{P}(\boldsymbol{X} \mid \mathcal{G})$, $\boldsymbol{X} = \{X_i\}_{i\in[N]}$ by a known distribution, a multivariate Gaussian with mean $\mathbf{0}$ and covariance matrix $\sigma^2\mathcal{I}$, where σ is a hyperparameter. We thus make use of an approximating distribution q,

$$q(\boldsymbol{X}) \approx \mathbb{P}(\boldsymbol{X} \mid \mathcal{G})$$
$$\boldsymbol{X} \overset{q}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2\mathcal{I})$$

Working in a variational inference setting, we aim to find the hyperparameters tha maximise the free energy

$$\mathcal{F}(\sigma) = \langle log\mathbb{P}(\mathcal{G} \mid \boldsymbol{X})\rangle_{q(\boldsymbol{X})} - KL[q(\boldsymbol{X}) \mid\mid \mathbb{P}(\boldsymbol{X} \mid \mathcal{G})]$$

Le us focus on the first term for now, denoting the value of the square distance between two Gaussian embeddings by $u_{ij} = \text{dist}(X_i, X_j)^2$, and capitalising it to represent the random variable associated with it.

$$\begin{aligned}
\langle log\mathbb{P}(\mathcal{G} \mid \boldsymbol{X})\rangle_{q(\boldsymbol{X})} &= \int\limits_{(\mathbb{R}^n)^N} q(\boldsymbol{X}) log\mathbb{P}(\mathcal{G} \mid \boldsymbol{X}) d\boldsymbol{X} \\
&= \int\limits_{(\mathbb{R}^n)^N} q(\boldsymbol{X}) \prod_{i<j} e^{-u_{ij}X_{ij}}(1-e^{-u_{ij}})^{1-X_{ij}} \\
&= \sum_{i<j} \int\limits_{(\mathbb{R}_{>0})^{\frac{N(N-1)}{2}}} q(\boldsymbol{U})(-u_{ij}X_{ij} + (1-X_{ij})log(1-e^{-u_{ij}}))d\boldsymbol{U} \\
&= \sum_{i<j} \int\limits_{\mathbb{R}_{>0}} q(u_{ij})(-u_{ij}X_{ij} + (1-X_{ij})log(1-e^{-u_{ij}}))du_{ij} \\
&= \sum_{i<j} \int\limits_{\mathbb{R}_{>0}} q(u_{ij}) \left(-u_{ij}X_{ij} + (1-X_{ij})log\frac{e^{u_{ij}}-1}{e^{u_{ij}}}\right) du_{ij} \\
&= \sum_{i<j} \int\limits_{\mathbb{R}_{>0}} q(u_{ij})(-u_{ij}X_{ij} - (1-X_{ij})u_{ij} + (1-X_{ij})log(e^{u_{ij}}-1))du_{ij} \\
&= \sum_{i<j} \int\limits_{\mathbb{R}_{>0}} q(u_{ij})(-u_{ij} + (1-X_{ij})log(e^{u_{ij}}-1))du_{ij} \\
&= -\binom{N}{2} < u >_{q(u)} + \sum_{i<j}(1-X_{ij}) \int\limits_{\mathbb{R}_{>0}} q(u_{ij})log(e^{u_{ij}}-1)du_{ij}
\end{aligned}$$

$$= - \binom{N}{2} < u >_{q(u)} + \left(\binom{N}{2} - \underbrace{E}_{\#\text{edges}} \right) \int_{\mathbb{R}_{>0}} q(u) log(e^u - 1) du$$

Using that

$$q(u) = \frac{u^{m-1}(2\sigma^2)^{-m}}{(m-1)!} exp(-(2\sigma^2)^{-1}u)$$

we have

$$\langle log\mathbb{P}(\mathcal{G} \mid \boldsymbol{X}) \rangle_{q(\boldsymbol{X})} = - \binom{N}{2} 2m\sigma^2 + \left(\binom{N}{2} - E \right) \frac{(2\sigma^2)^{-m}}{(m-1)!} \int_{\mathbb{R}_{>0}} u^{m-1} exp(-(2\sigma^2)^{-1}u) log(e^u - 1) du$$

The integral does not have a closed form, so we are forced to approximate it. Using the well-known result for the KL divergence between two multivariate Gaussians and that it is additive for independent distributions, we have that the free energy is

$$\mathcal{F}(\sigma) = - \binom{N}{2} 2m\sigma^2 + \left(\binom{N}{2} - E \right) \frac{(2\sigma^2)^{-m}}{(m-1)!} \int_{\mathbb{R}_{>0}} u^{m-1} exp(-(2\sigma^2)^{-1}u) log(e^u - 1) du$$
$$- Nm(-log(\sigma^2) - 1 + \sigma^2)$$

3.0.1 Relation To Graphons

Due to a special case of the Borel isomorphism theorem, we are able to map a Uniform[0, 1] random variable to a multivariate Gaussian, which allows us to work with the latter when generating graphs, as per definition (?). Let $\phi : [0, 1] \to \mathbb{R}^n$ be that mapping. If we work with the squared distance between embeddings to randomly determine the existence

of an edge, the graphon we are working with is

$$w(x, y) = e^{(\phi(x)-\phi(y))^T \phi(x)-\phi(y))} \tag{3.1}$$

3.1 Spherical RGG

We can also model the embeddings as sitting on a hypersphere. We then have two possible approaches to measuring the geodesic, the shortest path between two points:

1. In a spherical geometry, where the edges are curves on the sphere
2. In a Euclidean setting, so the edges go through the sphere

In both frameworks, we need to place a prior over the embedding points. Let us assume we are working with a unit-radius hypersphere, though nothing much changes when working with an arbitrary radius. The simplest distribution would be the uniform one, but that is not interesting enough, so we opt for the von Mises - Fisher distribution, which is akin to the normal distribution, except the points all have a fixed radius. Its probability density function is

$$f_n(\boldsymbol{x}; \boldsymbol{\mu}, \kappa) = C_n(\kappa) e^{\kappa \boldsymbol{\mu}^T \boldsymbol{x}} \tag{3.2}$$

where $\kappa > 0$ is the concentration parameter of the distribution around the mean direction $\boldsymbol{\mu}$, $\|\boldsymbol{\mu}\| = 1$. The greater κ is, the more concentrated the points are, with $\kappa = 0$ determining a uniform distribution. The normalisation constant $C_n(\kappa)$ is

$$C_n(\kappa) = \frac{\kappa^{\frac{n}{2}-1}}{(2\pi)^{\frac{n}{2}} I_{\frac{n}{2}-1}(\kappa)} \tag{3.3}$$

where I_u represents the modified Bessel function of the first kind at order u.

Since only the relative position of the points matters, we can pick any mean direction $\boldsymbol{\mu}$, so we can simplify our work by choosing $\boldsymbol{\mu} = (1, 0, \ldots, 0)$, leaving only κ to vary. We need to find the probability distribution function for the distance between two points governed by a von Mises - Fisher distribution. In the first framework, the distance between two points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is given by

$$d = cos^{-1}(\boldsymbol{x}_1^T \boldsymbol{x}_2) \tag{3.4}$$

3.2 Hyperbolic RGG

Chapter 4

Summary and Conclusion

In this thesis, we study the potential of using random vertex embeddings as latent variables for a given adjacency matrix, working in the framework of random geometric graphs. We investigated Euclidean, spherical and hyperbolic spaces and relied on Bayesian variational inference to determine the posterior of the embeddings, given a graph.

The most challenging aspect of this work was the form of the probability density function for the distance between pairs of embeddings, which forced us to make assumptions that severly limited our model. Even for something as simple as a spherical multivariate Gaussian prior, we run into an intractable integral and the right parameters are not recovered for simple data. It gets even more complicated when placing the embeddings on the unit sphere or in a hyperbolic space, with little hope of generating sensible results when considering random variables for distances.

We assert that this avenue should be avoided.

Original Edges	Found Edges	Degree Mean	ESP Mean	Geodesic Mean	Found Sigma
5200.55 ± 303.923	11577.85 ± 517.7	0.032 ± 0.004	0.036 ± 0.005	0.216 ± 0.01	0.846

Table 4.1: Mixture of two Gaussians, centered at $\mu_1 = (0, 0)$ and $\mu = (2, 2)$, with covariance matrices $1.5\mathcal{I}$ and $\mathcal{I}$.

Original Edges	Found Edges	Degree Mean	ESP Mean	Geodesic Mean	Found Sigma
4198.4 ± 222.778	10673.55 ± 561.666	0.038 ± 0.005	0.041 ± 0.007	0.49 ± 0.019	0.907

Table 4.2: Mixture of two Gaussians, centered at $\mu_1 = (0, 0)$ and $\mu = (7, 7)$, with covariance matrices $1.5\mathcal{I}$ and $\mathcal{I}$.

Params	Original Edges	Found Edges	Degree Mean	ESP Mean	Geodesic Mean	Found Sigma
0.1	30845.85 ± 591.019	32250.65 ± 453.502	0.026 ± 0.005	0.006 ± 0.001	0.031 ± 0.01	0.311
0.4	16277.45 ± 686.882	21486.95 ± 852.434	0.026 ± 0.004	0.012 ± 0.001	0.098 ± 0.016	0.525
0.7	10878.85 ± 691.571	16706.35 ± 470.24	0.027 ± 0.004	0.015 ± 0.002	0.077 ± 0.013	0.64
1	8361.75 ± 533.785	14837.65 ± 740.503	0.028 ± 0.005	0.02 ± 0.003	0.139 ± 0.016	0.713
1.3	6893.9 ± 292.288	13506.35 ± 742.156	0.029 ± 0.003	0.025 ± 0.002	0.171 ± 0.013	0.766
1.6	5569.3 ± 291.349	12047.35 ± 615.971	0.031 ± 0.004	0.032 ± 0.004	0.201 ± 0.013	0.826
1.9	4912.45 ± 323.229	11278.45 ± 636.171	0.033 ± 0.005	0.036 ± 0.006	0.199 ± 0.019	0.862
2.2	4262.85 ± 193.613	10480.35 ± 512.074	0.034 ± 0.005	0.042 ± 0.005	0.191 ± 0.016	0.903
2.5	3748.9 ± 202.982	9790.3 ± 525.146	0.036 ± 0.004	0.049 ± 0.005	0.17 ± 0.022	0.94
2.8	3523.65 ± 306.072	9553.65 ± 535.247	0.038 ± 0.004	0.052 ± 0.007	0.163 ± 0.022	0.959
3.1	3153.2 ± 214.771	9181.1 ± 563.009	0.04 ± 0.005	0.057 ± 0.006	0.157 ± 0.013	0.992
3.4	2913.3 ± 171.292	8592.25 ± 412.613	0.044 ± 0.006	0.061 ± 0.008	0.157 ± 0.022	1.015

Table 4.3: Spherical Multivariate T with σ given in the params column

Params	Original Edges	Found Edges	Degree Mean	ESP Mean	Geodesic Mean	Found Sigma
0.1	43111.95 ± 100.669	42817.5 ± 188.248	0.047 ± 0.011	0.02 ± 0.007	0.007 ± 0.003	0.109
0.4	27249.4 ± 616.919	29589.0 ± 655.306	0.027 ± 0.005	0.009 ± 0.002	0.052 ± 0.012	0.36
0.7	15294.1 ± 826.686	20588.85 ± 1002.237	0.027 ± 0.004	0.013 ± 0.001	0.106 ± 0.013	0.543
1.0	8809.4 ± 556.351	15291.95 ± 786.086	0.028 ± 0.004	0.021 ± 0.002	0.12 ± 0.016	0.698
1.3	5781.2 ± 335.055	11951.15 ± 719.837	0.031 ± 0.004	0.033 ± 0.004	0.205 ± 0.011	0.816
1.6	3954.0 ± 203.295	10264.05 ± 393.178	0.035 ± 0.005	0.048 ± 0.006	0.198 ± 0.012	0.924
1.9	2907.6 ± 208.284	8745.65 ± 530.822	0.046 ± 0.007	0.066 ± 0.007	0.153 ± 0.011	1.016
2.2	2179.8 ± 129.545	7537.05 ± 305.71	0.058 ± 0.01	0.089 ± 0.011	0.208 ± 0.03	1.104
2.5	1736.15 ± 112.232	6817.55 ± 530.583	0.062 ± 0.008	0.111 ± 0.011	0.307 ± 0.042	1.176
2.8	1394.55 ± 128.861	6122.5 ± 322.384	0.07 ± 0.009	0.131 ± 0.02	0.436 ± 0.032	1.248
3.1	1130.95 ± 96.396	5680.6 ± 432.159	0.089 ± 0.01	0.157 ± 0.019	0.535 ± 0.044	1.317
3.4	948.2 ± 79.205	5216.25 ± 214.66	0.098 ± 0.009	0.176 ± 0.018	0.639 ± 0.048	1.376

Table 4.4: Spherical Gaussian with σ given in the params column

Original Edges	Found Edges	Degree Mean	ESP Mean	Geodesic Mean	Found Sigma
8362.3 ± 395.61	14660.9 ± 642.377	0.033 ± 0.005	0.02 ± 0.003	0.187 ± 0.01	0.712

Table 4.5: Gaussian with covariance matrix $\begin{pmatrix} 3 & 1.5 \\ 1.5 & 1 \end{pmatrix}$

Original Edges	Found Edges	Degree Mean	ESP Mean	Geodesic Mean	Found Sigma
6975.8 ± 429.216	13391.15 ± 720.736	0.031 ± 0.005	0.02 ± 0.003	0.164 ± 0.011	0.763

Table 4.6: Multivariate T with 5 degrees of freedom and covariance matrix $\begin{pmatrix} 3 & 1.5 \\ 1.5 & 1 \end{pmatrix}$

Params	Original Edges	Found Edges	Degree Mean	ESP Mean	Geodesic Mean	Found Sigma
1	9017.35 ± 389.943	15300.85 ± 635.273	0.028 ± 0.003	0.02 ± 0.003	0.115 ± 0.012	0.692
2	1749.65 ± 155.846	4257.5 ± 332.685	0.054 ± 0.007	0.142 ± 0.026	0.158 ± 0.022	0.747
3	352.85 ± 35.888	1104.1 ± 105.989	0.206 ± 0.025	0.359 ± 0.056	0.56 ± 0.036	0.779
4	74.3 ± 12.791	329.15 ± 42.781	0.376 ± 0.037	0.164 ± 0.066	0.157 ± 0.052	0.781
5	15.75 ± 4.182	112.5 ± 16.684	0.337 ± 0.035	0.044 ± 0.037	0.011 ± 0.004	0.76
6	3.2 ± 1.939	47.15 ± 13.055	0.213 ± 0.05	0.005 ± 0.016	0.002 ± 0.001	0.736
7	0.25 ± 0.433	15.85 ± 5.052	0.091 ± 0.027	0.741 ± 0.43	0.0 ± 0.0	0.721
8	0.1 ± 0.3	6.6 ± 2.267	0.04 ± 0.014	0.9 ± 0.3	0.0 ± 0.0	0.712
9	0.0 ± 0.0	2.05 ± 1.717	0.013 ± 0.011	0.8 ± 0.4	0.0 ± 0.0	0.709
10	0.0 ± 0.0	0.65 ± 0.792	0.004 ± 0.005	0.45 ± 0.497	0.0 ± 0.0	0.708

Table 4.7: Gaussian with identity covariance matrix for different ms.

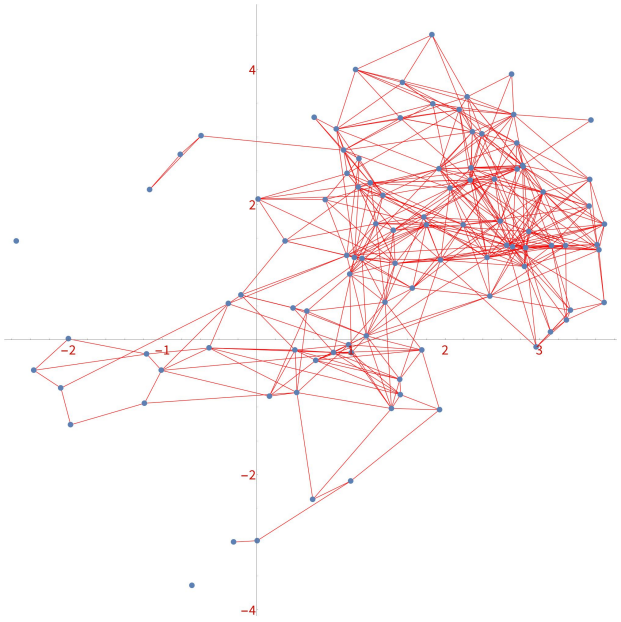

Figure 4.1: Mixture of two Gaussians, $\mu_1 = (0, 0)$, $\mu_2 = (2, 2)$, $\Sigma_1 = 1.5\mathcal{I}$, $\Sigma_2 = \mathcal{I}$, $\pi = (0.4, 0.6)$

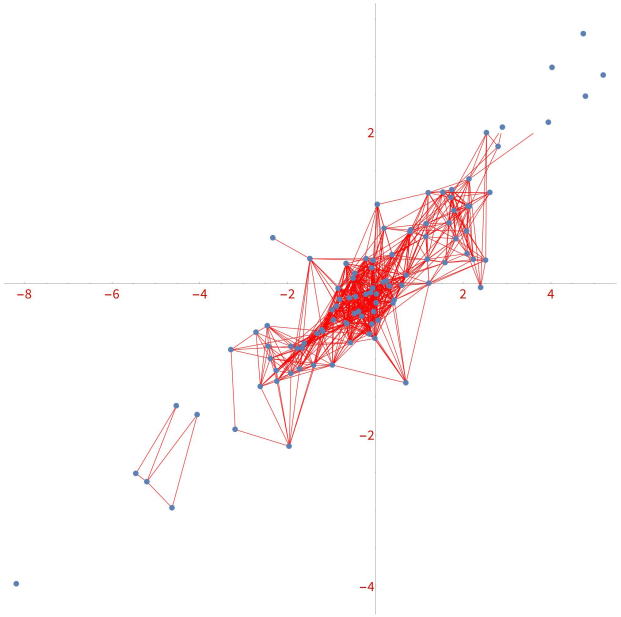

Figure 4.2: Multivariate T with $\Sigma = \begin{pmatrix} 3 & 1.5 \\ 1.5 & 1 \end{pmatrix}$ and 5 degrees of freedom.

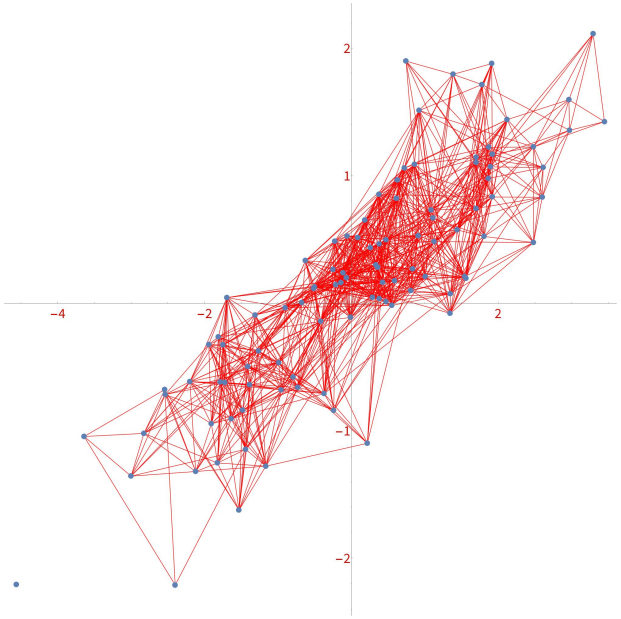

Figure 4.3: Gaussian prior with $\Sigma = \begin{pmatrix} 3 & 1.5 \\ 1.5 & 1 \end{pmatrix}$

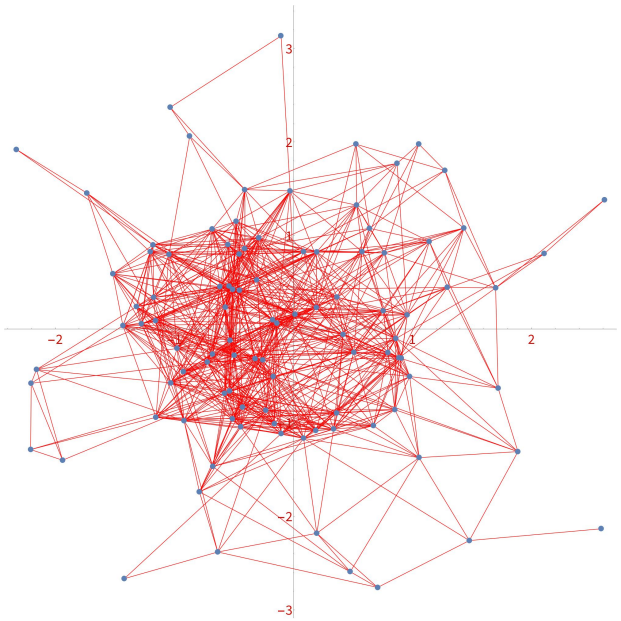

Figure 4.4: Multivariate T with $\Sigma = \mathcal{I}$ and 15 degrees of freedom.

Bibliography

[1] A. Gibbons. *Algorithmic Graph Theory*. Cambridge University Press, 1985, p. 2. ISBN: 9780521288811. URL: https://books.google.co.uk/books?id=Be6t04pgggwC.

Appendix A

Distance PDF

Let

$$X_i, X_j \sim \mathcal{N}(\mathbf{0}, \frac{1}{2}\sigma^2\mathcal{I}_n)$$

Then

$$Y_{ij} = X_i - X_j \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathcal{I}_n)$$

Denoting the squared distance between X_i and X_j by the random variable

$$D_{ij} = Y_{ij}^T Y_{ij}$$

we find ourselves in the context defined in Mathai, so we have that the pdf of D_{ij} is

$$f(u) = \sum_{k=0}^{\infty} (-1)^k c_k \frac{u^{\frac{n}{2}+k-1}}{\Gamma(\frac{n}{2}+k)}$$

where

$$c_0 = \prod_{j=1}^{n} (2\lambda_j)^{-\frac{1}{2}}$$

$$c_k = \frac{1}{k}\sum_{r=0}^{k-1} d_{k-r}c_r, \quad k \geq 1 \tag{A.1}$$
$$d_k = \frac{1}{2}\sum_{j=1}^{n}(2\lambda_j)^{-k}, \quad k \geq 1$$

with $\{\lambda_j\}$ denoting the eigenvalues of the covariance matrix of Y_{ij}, $\sigma^2\mathcal{I}_n$, so

$$\lambda_j = \sigma^2, \quad j \in \{1, \ldots, n\}$$

implying

$$c_0 = (2\sigma^2)^{-\frac{n}{2}} \tag{A.2}$$
$$d_k = \frac{1}{2}\sum_{j=1}^{n}(2\sigma^2)^{-k}$$
$$= \frac{n}{2}(2\sigma^2)^{-k}, \quad k \geq 1 \tag{A.3}$$

Assuming that $n = 2m$, $m \in \mathbb{N}$, we will prove by induction that

$$c_k = \begin{pmatrix} m+k-1 \\ k \end{pmatrix} 2^{-k-m}(\sigma^2)^{-m-k}, \quad k \geq 0$$

where $\begin{pmatrix} a \\ b \end{pmatrix}$ denotes combinations of a taken b. For the sake of the notation, let us assume that $\begin{pmatrix} 0 \\ 0 \end{pmatrix} = 1$.

From (A.2), we have that

$$c_0 = 2^{-\frac{n}{2}}(\sigma^2)^{-fracn2}$$

$$= \underbrace{1}_{\binom{m-1}{0}} \times 2^{-m}(\sigma^2)^{-m}$$

so the induction hypotesis holds for c_0.

Now, assuming that the hypothesis holds for c_j, $j \in \{0, \ldots, k-1\}$, we will prove it holds for c_k as well. Let us calculate each term of (A.1).

$$\begin{aligned} d_{k-r}c_r &= m2^{-k+r}(\sigma^2)^{-k+r} \times \binom{m+r-1}{r} 2^{-r-m}(\sigma^2)^{-m-r} \\ &= m\binom{m+r-1}{r} 2^{-k-m}(\sigma^2)^{-k-m} \end{aligned}$$

Then

$$\begin{aligned} c_k &= \frac{1}{k}\sum_{r=0}^{k-1} m\binom{m+r-1}{r} 2^{-k-m}(\sigma^2)^{-k-m} \\ &= \frac{m}{k}2^{-k-m}(\sigma^2)^{-k-m}\sum_{r=0}^{k-1}\binom{m+r-1}{r} \end{aligned} \tag{A.4}$$

(A.5)

We will prove by induction over k that $\sum_{r=0}^{k-1}\binom{m+r-1}{r} = \binom{m+k-1}{k-1}$. For $k = 1$ it obviously holds. Assuming our assumption is true for some k, we will prove it holds for

$k+1$ as well.

$$\begin{aligned}\sum_{r=0}^{k}\binom{m+r-1}{r} &= \sum_{r=0}^{k-1}\binom{m+r-1}{r} + \binom{m+k-1}{k} \\ &= \binom{m+k-1}{k-1} + \binom{m+k-1}{k} \\ &= \binom{m+k}{k}\end{aligned}$$

making the induction proof complete.

(A.4) then becomes

$$\begin{aligned}c_k &= \frac{m}{k}\binom{m+k-1}{k-1}2^{-k-m}(\sigma^2)^{-k-m} \\ &= \binom{m+k-1}{k}2^{-k-m}(\sigma^2)^{-k-m}\end{aligned}$$

so the induction is complete. We now have that

$$\begin{aligned}f(u) &= \sum_{k=0}^{\infty}(-1)^k\binom{m+k-1}{k}2^{-k-m}(\sigma^2)^{-k-m}\frac{u^{m+k-1}}{\Gamma(m+k)} \\ &= \sum_{k=0}^{\infty}(-1)^k\frac{(m+k-1)!}{(m-1)!k!}2^{-k-m}(\sigma^2)^{-k-m}\frac{u^{m+k-1}}{(m+k-1)!} \\ &= \sum_{k=0}^{\infty}(-1)^k\frac{\cancel{(m+k-1)!}}{(m-1)!k!}2^{-k-m}(\sigma^2)^{-k-m}\frac{u^{m+k-1}}{\cancel{(m+k-1)!}} \\ &= \frac{2^{-m}(\sigma^2)^{-m}u^{m-1}}{(m-1)!}\sum_{k=0}^{\infty}(-1)^k\frac{1}{k!}2^{-k}(\sigma^2)^{-k}u^k\end{aligned}$$

$$= \frac{2^{-m}(\sigma^2)^{-m}u^{m-1}}{(m-1)!}\sum_{k=0}^{\infty}\frac{(-(2\sigma^2)^{-1}u)^k}{k!}$$
$$= \frac{(2\sigma^2)^{-m}u^{m-1}}{(m-1)!}e^{-(2\sigma^2)^{-1}u}$$

Appendix B

Free Energy

$$\mathcal{F} = \langle log\mathbb{P}(\mathcal{G} \mid \mathbf{X}) \rangle_{q(\mathbf{X})} - KL(q(\mathbf{X}) \mid\mid \mathbb{P}(\mathbf{X})) \tag{B.1}$$

$$\langle log\mathbb{P}(\mathcal{G} \mid \mathbf{X}) \rangle_{q(\mathbf{X})} = \int_{(\mathbb{R}^n)^N} d\mathbf{X}$$