Gaussian RGG

1 Work

We are given a graph $\mathcal{G}$ with N and E edges. Let $X = \{X_i\}_{i\in[N]}$ be the set of embeddings for the vertices we are trying to infer from the data, modelled iid by a prior multivariate Gaussian

$$X_i \sim \mathcal{N}(\mathbf{0}, \mathcal{I}_n)$$

where n is the number of dimensions of the space in which the embeddings lie. The existence of an edge between two vertices will be randomly determined, based on the distance between their corresponding embeddings. Please note that the mean of the distribution is irrelevant when working with a single distribution, since we could just shift all the points such that the mean bocomes 0 and the graph will remain the same.

Let

$$X_{ij} \sim \text{Bernoulli}(e^{-d_{ij}^2})$$

be a random variable representing the existence of an edge between the i^{th} and j^{th} vertex, where d_{ij} is the Euclidean distance between their corresponding embeddings, X_i and X_j.

We are looking to approximate the posterior distribution

$$P(X \mid \mathcal{G}) \propto P(\mathcal{G} \mid X)P(X)$$

by a multivariate Gaussian with spherical covariance matrix determined by parameter σ

$$P(X \mid \mathcal{G}) \approx q(X) = \mathcal{N}(X;\, 0,\, \sigma^2\mathcal{I}_n)$$

Therefore, we look to maximise the free energy as a function of σ for the best approximation.

$$\mathcal{F}(\sigma) = \langle logP(\mathcal{G} \mid X)\rangle_{q(X)} - \text{KL}\left(P(X)||q(X)\right)) \tag{1}$$

The first term becomes

$$\begin{aligned}
\langle logP(\mathcal{G} \mid X)\rangle_{q(X)} &= \int_{[\mathbb{R}^n]^N} dX\, q(X)\, logP(\mathcal{G} \mid X) \\
&= \int_{[\mathbb{R}^n]^N} dX\, q(X) \sum_{i<j} P(X_{ij} \mid X_i, X_j) \\
&= \sum_{i<j} \int_{[\mathbb{R}^n]^2} dX_i dX_j\, q(X_i, X_j)\, logP(X_{ij} \mid X_i, X_j) \\
&= \sum_{i<j} \int_{[\mathbb{R}^n]^2} dX_i dX_j\, q(X_i, X_j)\, log\left[(e^{-d_{ij}^2})^{X_{ij}}(1-e^{-d_{ij}^2})^{1-X_{ij}}\right] \\
&= \sum_{i<j} \int_{\mathbb{R}^n} dX_i \int_0^\infty du\, log\left[(e^{-u})^{X_{ij}}(1-e^{-u})^{1-X_{ij}}\right] \int_{\mathcal{X}=\{X_j\in\mathbb{R}^n : d(X_i,X_j)^2=u\}} dX_j\, q(X_i, X_j) \\
&= \sum_{i<j} \int_0^\infty du\, log\left[(e^{-u})^{X_{ij}}(1-e^{-u})^{1-X_{ij}}\right] \underbrace{\int_{\mathbb{R}^n} dX_i \int_{\mathcal{X}=\{X_j\in\mathbb{R}^n : d(X_i,X_j)^2=u\}} dX_j\, q(X_i, X_j)}_{q(u)}
\end{aligned}$$

where $q(u)$ is the pdf of the squared Euclidean distance between two independent points modelled by a Gaussian distribution.

The squared distance can be written as $(X_1 - X_2)^T(X_1 - X_2)$, where

$$X_1 \sim \mathcal{N}(0, \tau_1^2 \mathcal{I}_n)$$
$$X_2 \sim \mathcal{N}(0, \tau_2^2 \mathcal{I}_n)$$

so $U \sim Y^T Y$, where $Y \sim \mathcal{N}(0, \underbrace{(\tau_1^2 + \tau_2^2)}_{\tau^2} \mathcal{I}_n)$

Assuming that n is even, we work out that

$$q(u) = \frac{(2\tau^2)^{-m}}{(m-1)!} u^{m-1} e^{-\frac{u}{2\tau^2}}, \quad \text{where } m = \frac{n}{2} \in \mathbb{N} \tag{2}$$

In the expected log likelihood, the covariances of the embeddings have the same value, σ^2, so our $q(u)$ becomse

$$q(u) = \frac{(2 \times 2\sigma^2)^{-m}}{(m-1)!} u^{m-1} e^{-\frac{u}{2 \times 2\sigma^2}}, \quad \text{where } m = \frac{n}{2} \in \mathbb{N} \tag{3}$$

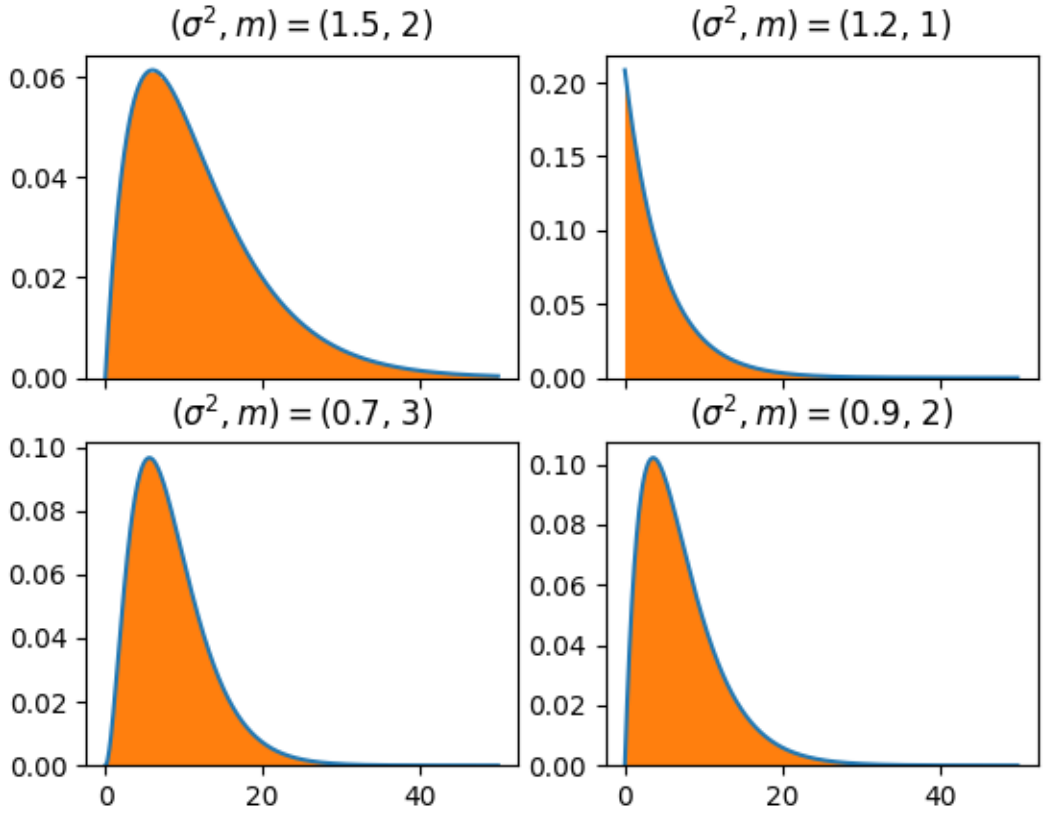


Figure 1: The blue line is the pdf of the squared distance between two random normals, as calculated at (3) and the orange plot is a histogram of the squared distance between 10^7 samples of two normally distributed points.

Using this result, we can calculate the expected number of edges in a graph, for N embeddings generated using the covariance matrix $\sigma^2 \mathcal{I}_n$.

$$\mathbb{E}[\#\text{edges}] = \binom{N}{2} \mathbb{E}[X_{12}]$$

$$
\begin{aligned}
&= \binom{N}{2} \int_0^\infty du\, e^{-u} q(u) \\
&= \binom{N}{2} \int_0^\infty du\, e^{-u} \frac{(4\sigma^2)^{-m}}{(m-1)!} u^{m-1} e^{-\frac{u}{4\sigma^2}} \\
&= \binom{N}{2} \frac{1}{(1+4\sigma^2)^m}
\end{aligned}
$$

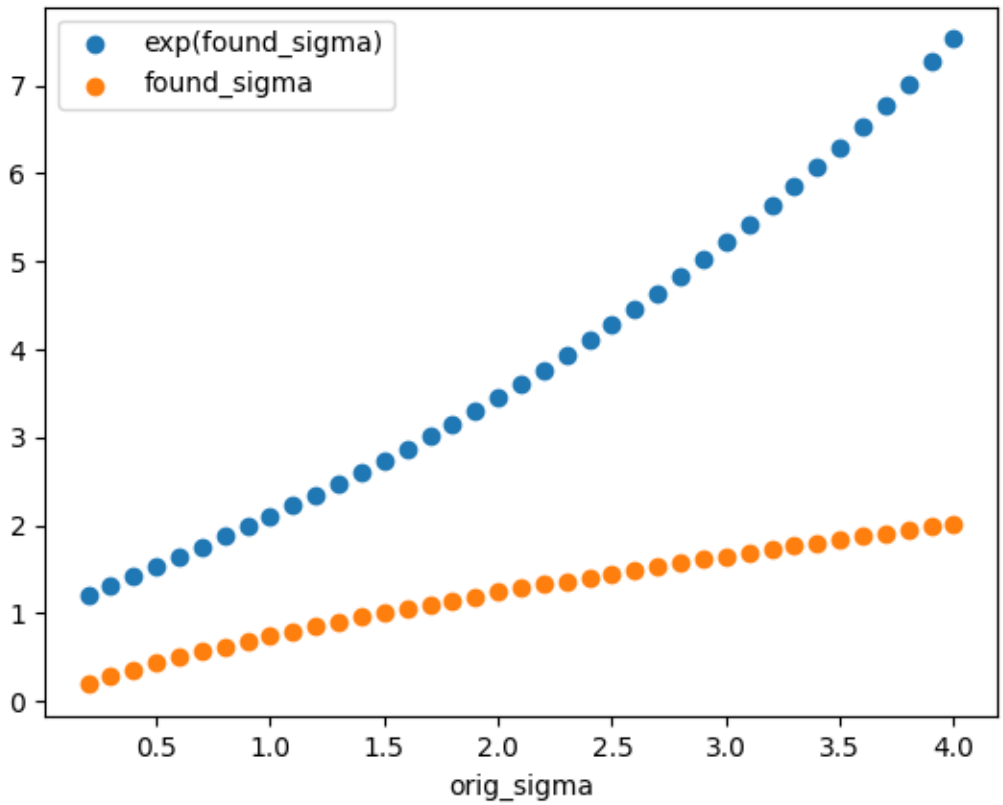
Going back to the expected log likelihood, we have

$$
\begin{aligned}
\langle logP(\mathcal{G} \mid X) \rangle_{q(X)} &= \sum_{i<j} \int_0^\infty du\, log\left[(e^{-u})^{X_{ij}} (1-e^{-u})^{1-X_{ij}} \right] q(u) \\
&= E \int_0^\infty du\, log(e^{-u}) q(u) + \left(\binom{N}{2} - E \right) \int_0^\infty du\, log(1-e^{-u}) q(u)
\end{aligned}
$$

where E is the number of edges in the given graph. The first integral can be calculated

$$
\begin{aligned}
\int_0^\infty du\, log(e^{-u}) q(u) &= \int_0^\infty du\ -u \frac{(4\sigma^2)^{-m}}{(m-1)!} u^{m-1} e^{-\frac{u}{4\sigma^2}} \\
&= -\int_0^\infty du\, \frac{(4\sigma^2)^{-m}}{(m-1)!} u^m e^{-\frac{u}{4\sigma^2}} \\
&= -\frac{1}{(m-1)!} \int_0^\infty du\, \left(\frac{u}{4\sigma^2} \right)^m e^{-\frac{u}{4\sigma^2}} \\
&= -\frac{1}{(m-1)!} (4\sigma^2) \Gamma(m+1) \\
&= -4m\sigma^2
\end{aligned}
$$

For m = 2, for a set of sigmas, we calculate the expected number of edges in our graph and use that to optimise for sigma in our free energy. I said that there is some relationship between the original sigma and the recovered one, but, plotting for a larger sigma set, it looks like there is not a clear one. My hypothesis is that the approximate distribution, with its spherical covariance is too strong of an assumption and that is why it cannot recover the parameter.


exp(found_sigma)
found_sigma
0
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
orig_sigma