Appendix A

In this appendix we show that when using feature functions in Eqs.(1) and (2) the CCRF distribution is actually that of a multi-variate Gaussian.

In our discussion we will use the following notation: $\mathbf{x} = \{\mathbf{x}_1^{(q)}, \mathbf{x}_2^{(q)}, \cdots, \mathbf{x}_n^{(q)}\}$ is a set of input variables that are observed and $\mathbf{y} = \{y_1^{(q)}, y_2^{(q)}, \cdots, y_n^{(q)}\}$ a set of output variables that we wish to predict, $\mathbf{x}_i^{(q)} \in \mathcal{R}^m$ and $y_i^{(q)} \in \mathcal{R}$, here q indicates the q^{th} sequence of interest, it is omitted in some equations for clarity (when there is no ambiguity). We also define $\mathbf{X}$ as a matrix where the i^{th} row represents $\mathbf{x}_i$.

$$f_k(y_i, \mathbf{X}) = -(y_i - \mathbf{X}_{i,k})^2 \tag{1}$$

$$g_k(y_i, y_j, \mathbf{X}) = -\frac{1}{2} S_{i,j}^{(k)} (y_i - y_j)^2 \tag{2}$$

When using feature functions defined in Eqs.(1) and (2), the probability distribution of CCRF:

$$P(y|\mathbf{X}) = \frac{\exp(\Psi)}{\int_{-\infty}^{\infty} \exp(\Psi) d\mathbf{y}} \tag{3}$$

$$\Psi = \sum_i \sum_{k=1}^{K1} \alpha_k f_k(y_i, \mathbf{X}) + \sum_{i,j} \sum_{k=1}^{K2} \beta_k g_k(y_i, y_j, \mathbf{X}) \tag{4}$$

is in fact a multivariate Gaussian with the following distribution:

$$P(\boldsymbol{y}|\mathbf{X}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^{\mathbf{T}} \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{\mu})), \tag{5}$$

where

$$\Sigma^{-1} = 2(A + B) \tag{6}$$

The diagonal matrix A represents the contribution of α terms (vertex features) to the covariance matrix, and the symmetric B represents the contribution of the β terms (edge features).

$$A_{i,j} = \begin{cases} \sum_{k=1}^{K1} \alpha_k, & i = j \\ 0, & i \neq j \end{cases} \tag{7}$$

$$B_{i,j} = \begin{cases} (\sum_{k=1}^{K2} \beta_k \sum_{r=1}^{n} S_{i,r}^k) - (\sum_{k=1}^{K2} \beta_k S_{i,j}^k), & i = j \\ -\sum_{k=1}^{K2} \beta_k S_{i,j}^k, & i \neq j \end{cases} \tag{8}$$

We also define a further vector $\mathbf{b}$:

$$\boldsymbol{b}_i = 2 \sum_{k=1}^{K1} \alpha_k \boldsymbol{X}_{i,k} \tag{9}$$

$$\mathbf{b} = 2\boldsymbol{X}\alpha \tag{10}$$

We can now define another useful term $\boldsymbol{\mu}$, which will be our mean values in the multivariate Gaussian distribution:

$$\boldsymbol{\mu} = \Sigma\mathbf{b} \tag{11}$$

Defining A, B and $\boldsymbol{b}$ in such a way allows us to rewrite the factors of Eq.(3) in terms of matrix multiplications making the derivation of the partition function and the partial derivatives easier.

Having defined all the necessary variables we can start showing the equivalence between probability density in Eq.(3) and the multivariate Gaussian in Eq.(5). First we plug in the feature functions in Eqs.(1) and (2) into Eq.(4)

$$\begin{aligned}\Psi &= \textstyle\sum_i \sum_{k=1}^{K1} \alpha_k f_k(y_i, \boldsymbol{X}) + \sum_{i,j} \sum_{k=1}^{K2} \beta_k g_k(y_i, y_j, \boldsymbol{X}) \\ &= \textstyle -\sum_i \sum_{k=1}^{K1} \alpha_k (y_i - \boldsymbol{X}_{i,k})^2 - \frac{1}{2}\sum_{i,j} \sum_{k=1}^{K2} \beta_k S^k_{i,j}(y_i - y_j)^2\end{aligned} \tag{12}$$

Now we can express the factor Ψ in terms of A, B and $\boldsymbol{b}$. We do this in parts starting with terms containing α parameters in Eq.(12).

$$\begin{aligned}&\textstyle -\sum_i \sum_{k=1}^{K1} \alpha_k (y_i - \boldsymbol{X}_{i,k})^2 \\ &\textstyle = -\sum_i \sum_{k=1}^{K1} \alpha_k (y_i^2 - 2y_i\boldsymbol{X}_{i,k} + \boldsymbol{X}^2_{i,k}) \\ &\textstyle = -\sum_i \sum_{k=1}^{K1} \alpha_k y_i^2 + \sum_i \sum_{k=1}^{K1} \alpha_k 2y_i\boldsymbol{X}_{i,k} - \sum_i \sum_{k=1}^{K1} \alpha_k \boldsymbol{X}^2_{i,k} \\ &\textstyle = -\boldsymbol{y}^T A\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{b} - \sum_i \sum_{k=1}^{K1} \alpha_k \boldsymbol{X}^2_{i,k}\end{aligned} \tag{13}$$

And now collecting terms with β parameters in Eq.(12). Here we use the assumption that every $S^{(k)}$ is a symmetric matrix (which as a similarity matrix it should be).

$$\begin{aligned}&\textstyle -\frac{1}{2}\sum_{i,j} \sum_{k=1}^{K2} \beta_k S^k_{i,j}(y_i - y_j)^2 \\ &\textstyle = -\frac{1}{2}\sum_{i,j} \sum_{k=1}^{K2} \beta_k S^k_{i,j}(y_i^2 - 2y_iy_j + y_j^2) \\ &\textstyle = -\frac{1}{2}\sum_{i,j} \sum_{k=1}^{K2} \beta_k S^k_{i,j}(y_i^2 + y_j^2) + \sum_{i,j} \sum_{k=1}^{K2} \beta_k S^k_{i,j} y_i y_j \\ &\textstyle = -\sum_{k=1}^{K2} \beta_k \sum_{i,j} S^{(k)}_{i,j} y_i^2 + \sum_{k=1}^{K2} \beta_k S^{(k)}_{i,j} \sum_{i,j} y_i y_j \\ &= -\boldsymbol{y}^T B\boldsymbol{y}\end{aligned} \tag{14}$$

We now combining Eqs.(12), (13), and (14) for an alternative expression of Ψ.

$$\Psi = -\boldsymbol{y}^T A\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{b} - \boldsymbol{y}^T B\boldsymbol{y} - d = -\frac{1}{2}(\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y}) + \boldsymbol{y}\Sigma^{-1}\boldsymbol{\mu} - d \tag{15}$$

We define $d = \sum_i \sum_{k=1}^{K1} \alpha_k \boldsymbol{X}^2_{i,k}$ for brevity (it's not necessary writing it out in full as it cancels out eventually). We also use the fact from Eq.(11) that $\mathbf{b} = \Sigma^{-1}\boldsymbol{\mu}$.

Using Eq.(15) in Eq.(3) we get (As d does not depend on $\boldsymbol{y}$, we can cancel $\exp(-d)$ out):

$$\begin{aligned} P(\boldsymbol{y}|\boldsymbol{X}) &= \frac{\exp(\Psi)}{\int_{-\infty}^{\infty}\exp(\Psi)d\boldsymbol{y}} = \\ &= \frac{\exp(-\frac{1}{2}(\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y})+\boldsymbol{y}\Sigma^{-1}\boldsymbol{\mu})\exp(-d)}{\int_{-\infty}^{\infty}\{\exp(-\frac{1}{2}(\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y})+\boldsymbol{y}\Sigma^{-1}\boldsymbol{\mu})\exp(-d)\}d\boldsymbol{y}} \\ &= \frac{\exp(-\frac{1}{2}(\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y})+\boldsymbol{y}\Sigma^{-1}\boldsymbol{\mu})}{\int_{-\infty}^{\infty}\{\exp(-\frac{1}{2}(\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y})+\boldsymbol{y}\Sigma^{-1}\boldsymbol{\mu})\}d\boldsymbol{y}} \end{aligned} \tag{16}$$

Now we need to find the definite integral of $\exp(-\frac{1}{2}(\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y}) + \boldsymbol{y}\Sigma^{-1}\boldsymbol{\mu})$ with respect to $\boldsymbol{y}$, this can be achieved using the integral of a an expontial with square and linear terms[1].

$$\int_{-\infty}^{\infty}\{\exp(-\frac{1}{2}(\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y}) + \boldsymbol{y}\Sigma^{-1}\boldsymbol{\mu})\}d\boldsymbol{y} = \frac{(2\pi)^{\frac{n}{2}}}{|\Sigma^{-1}|^{\frac{1}{2}}}\exp(\frac{1}{2}\boldsymbol{\mu}\Sigma^{-1}\boldsymbol{\mu}) \tag{17}$$

Finally. plugging Eq.(15) and (17) into Eq.(3) we get:

$$\begin{aligned} P(\boldsymbol{y}|\boldsymbol{X}) &= \frac{\exp(-\frac{1}{2}\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y}+\boldsymbol{y}\Sigma^{-1}\boldsymbol{\mu})}{\frac{(2\pi)^{\frac{n}{2}}}{|\Sigma^{-1}|^{\frac{1}{2}}}\exp(\frac{1}{2}\boldsymbol{\mu}\Sigma^{-1}\boldsymbol{\mu})} \\ &= \frac{\exp(-\frac{1}{2}\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y}+\boldsymbol{y}\Sigma^{-1}\boldsymbol{\mu})\exp(-\frac{1}{2}\boldsymbol{\mu}\Sigma^{-1}\boldsymbol{\mu})}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \\ &= \frac{\exp(-\frac{1}{2}\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y}+\boldsymbol{y}\Sigma^{-1}\boldsymbol{\mu}-\frac{1}{2}\boldsymbol{\mu}\Sigma^{-1}\boldsymbol{\mu})}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu})) \end{aligned} \tag{18}$$

This is exactly what we set out to show.

Appendix B

This appendix deals with calculating the partial derivatives of the CCRF log-likelihood with respect to the parameters α and β. First of all, we would like to calculate the log-likelihood of Eq.(18)).

$$\begin{aligned} \log(P(\boldsymbol{y}|\boldsymbol{X})) &= -\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) - \log((2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}) \\ &= -\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) - (\tfrac{n}{2}\log(2\pi) + \tfrac{1}{2}\log|\Sigma|) \\ &= -\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) + \tfrac{1}{2}\log|\Sigma^{-1}| - \tfrac{n}{2}\log(2\pi) \\ &= -\tfrac{1}{2}\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y} + \boldsymbol{y}^T\Sigma^{-1}\boldsymbol{\mu} - \tfrac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu} + \tfrac{1}{2}\log|\Sigma^{-1}| - \tfrac{n}{2}\log(2\pi) \\ &= -\tfrac{1}{2}\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{b} - \tfrac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu} + \tfrac{1}{2}\log|\Sigma^{-1}| - \tfrac{n}{2}\log(2\pi) \\ &= -\tfrac{1}{2}\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{b} - \tfrac{1}{2}\boldsymbol{b}^T\Sigma\boldsymbol{b} + \tfrac{1}{2}\log|\Sigma^{-1}| - \tfrac{n}{2}\log(2\pi) \end{aligned} \tag{19}$$

Above we use $|\Sigma| = \frac{1}{|\Sigma^{-1}|}$, where $|\Sigma|$ denotes the determinant of the matrix Σ. Furthermore, because Σ^{-1} is symmetric by construction, $\Sigma^{-1} = (\Sigma^{-1})^T$ and $\Sigma = \Sigma^T$.

[1] `http://www.weylmann.com/gaussian.pdf`

Now we can derive all of the necessary partial derivatives, first we define the partial derivatives of Σ^{-1} and $\mathbf{b}$ with respect to α and β, as they will be reused. I is the identity matrix of size $n \times n$, where n is the number of elements in a sequence. Remember that A is only dependent on α, and B on β.

We will first show the partial derivatives of the likelihood for the alphas.

$$\frac{\partial \Sigma^{-1}}{\partial \alpha_k} = \frac{\partial 2A + 2B}{\partial \alpha_k} = \frac{\partial 2A}{\partial \alpha_k} = 2I \tag{20}$$

$$\frac{\partial \boldsymbol{b}}{\partial \alpha_k} = \frac{\partial 2\boldsymbol{X}\alpha}{\partial \alpha_k} = 2\boldsymbol{X}_{*,k} \tag{21}$$

Here $\boldsymbol{X}_{*,k}$ notation refers to a row vector corresponding to the k^{th} row of a matrix $\boldsymbol{X}$. In the derivation below, we use a trick of using the partial derivative of a matrix inverse ($\frac{\partial M^{-1}}{\partial \alpha} = -M^{-1}\frac{\partial M}{\partial \alpha}M^{-1}$ or alternatively $\frac{\partial M}{\partial \alpha} = -M\frac{\partial M^{-1}}{\partial \alpha}M$) to get the partial derivative of Σ.

$$\begin{aligned} \tfrac{\partial \boldsymbol{b}^T\Sigma\boldsymbol{b}}{\partial \alpha_k} &= \tfrac{\partial \boldsymbol{b}^T}{\partial \alpha_k}\Sigma\boldsymbol{b} + \boldsymbol{b}^T\tfrac{\partial \Sigma\boldsymbol{b}}{\partial \alpha_k} = 2\boldsymbol{X}_{*,k}\boldsymbol{\mu} + \boldsymbol{b}^T(\tfrac{\partial \Sigma}{\partial \alpha_k}\boldsymbol{b} + \Sigma\tfrac{\partial \boldsymbol{b}}{\partial \alpha_k}) \\ &= 2\boldsymbol{X}_{*,k}\boldsymbol{\mu} + \boldsymbol{b}^T\tfrac{\partial \Sigma}{\partial \alpha_k}\boldsymbol{b} + \boldsymbol{b}^T\Sigma 2(\boldsymbol{X}_{*,k})^T = 4\boldsymbol{X}_{*,k}\boldsymbol{\mu} + \boldsymbol{b}^T\tfrac{\partial \Sigma}{\partial \alpha_k}\boldsymbol{b} \\ &= 4\boldsymbol{X}_{*,k}\boldsymbol{\mu} + \boldsymbol{b}^T(-\Sigma\tfrac{\partial \Sigma^{-1}}{\partial \alpha_k}\Sigma)\boldsymbol{b} = 4\boldsymbol{X}_{*,k}\boldsymbol{\mu} - 2\boldsymbol{b}^T\Sigma\Sigma\boldsymbol{b} \\ &= 4\boldsymbol{X}_{*,k}\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\boldsymbol{\mu} \end{aligned} \tag{22}$$

Now for the normalisation (partition) function part:

$$\begin{aligned} \tfrac{\partial \log|\Sigma^{-1}|}{\partial \alpha_k} &= \tfrac{1}{|\Sigma^{-1}|}\tfrac{\partial|\Sigma^{-1}|}{\partial \alpha_k} = \tfrac{1}{|\Sigma^{-1}|}|\Sigma^{-1}| \times \text{trace}(\Sigma\tfrac{\partial \Sigma^{-1}}{\alpha_k}) \\ &= 2 \times \text{trace}(\Sigma I) = 2 \times \text{trace}(\Sigma) \end{aligned} \tag{23}$$

Now we can combine these to get

$$\frac{\partial \log(P(\boldsymbol{y}|\boldsymbol{X}))}{\alpha_k} = -\boldsymbol{y}^T\boldsymbol{y} + 2\boldsymbol{y}^T\boldsymbol{X}_{*,k}^T - 2\boldsymbol{X}_{*,k}\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{\mu} + \text{trace}(\Sigma) \tag{24}$$

We can now derive the partial derivatives of the likelihood with respect to β parameters

$$\frac{\partial \Sigma^{-1}}{\partial \beta_k} = 2B^{(k)} \tag{25}$$

$$B^{(k)} = \begin{cases} (\sum_{r=1}^{n} S_{i,r}^{(k)}) - S_{i,j}^{(k)}, & i = j \\ -S_{i,j}^{(k)}, & i \neq j \end{cases} \tag{26}$$

$$\frac{\partial \boldsymbol{b}}{\partial \beta_k} = 0 \tag{27}$$

$$\frac{\boldsymbol{b}^T\Sigma\boldsymbol{b}}{\beta_k} = -\boldsymbol{b}^T(\Sigma\frac{\partial \Sigma^{-1}}{\partial \beta}\Sigma)\boldsymbol{b} = -2\boldsymbol{b}^T\Sigma B^{(k)}\Sigma\boldsymbol{b} = -2\boldsymbol{\mu}^T B^{(k)}\boldsymbol{\mu} \tag{28}$$

$$\begin{aligned} \tfrac{\partial \log|\Sigma^{-1}|}{\partial \beta_k} &= \tfrac{1}{|\Sigma^{-1}|}\tfrac{\partial|\Sigma^{-1}|}{\partial \beta_k} = \tfrac{1}{|\Sigma^{-1}|}|\Sigma^{-1}| \times \text{trace}(\Sigma\tfrac{\partial \Sigma^{-1}}{\beta_k}) \\ &= 2 \times \text{trace}(\Sigma B^{(k)}) = 2 \times \text{Vec}(\Sigma)^T\text{Vec}(B^{(k)}) \end{aligned} \tag{29}$$

Here we use the matrix trace property $\text{trace}(AB) = \text{Vec}(A)^T\text{Vec}(B)$, here Vec refers to the matrix vectorisation operation which stacks up colums of a matrix together to form a single column matrix. We also use the derivative of inverse matrix trick as in the case with α_k version.

We can now combine these to get:

$$\frac{\partial \log(P(\boldsymbol{y}|\boldsymbol{X}))}{\beta_k} = -\boldsymbol{y}^T B^{(k)} \boldsymbol{y} + \boldsymbol{\mu}^T B^{(k)} \boldsymbol{\mu} + \mathrm{Vec}(\Sigma)^T \mathrm{Vec}(B^{(k)}) \tag{30}$$