# 1 Appendix A

In this appendix we show that when using feature functions in Eqs.(3) and (1) the CCNF distribution is actually that of a multi-variate Gaussian.

In our discussion we will use the following notation: $\boldsymbol{x} = \{\boldsymbol{x}_1^{(q)}, \boldsymbol{x}_2^{(q)}, \cdots, \boldsymbol{x}_n^{(q)}\}$ is a set of input variables that are observed and $\{y_1^{(q)}, y_2^{(q)}, \cdots, y_n^{(q)}\}$ a set of output variables that we wish to predict, $x_i^{(q)} \in \mathcal{R}^m$ and $y_i^{(q)} \in \mathcal{R}$, here $q$ indicates the $q^{\text{th}}$ sequence of interest, it is omitted in some equation for clarity (when there is no ambiguity).

In our work we use the following vertex and edge feature functions:

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

$$l_k(y_i, y_j) = -\frac{1}{2} S_{i,j}^{(l_k)} (y_i + y_j)^2 \tag{2}$$

Above $S_{i,j}^{(g_k)}$ is the similarity metric associated with that edge, and serves as a way to join up the edges (so if $S_{i,j} = S_{j,i} = 1$ it means that node $i$ is connected to node $j$). Similarly for $S^{(l_k)}$.

$$f_k(y_i, \boldsymbol{x}, \boldsymbol{\theta}_k) = -(y_i - h(\boldsymbol{\theta}_k^T \boldsymbol{x}_i))^2 \tag{3}$$

Above $h$ is an activation function (we use a sigmoid $h(x) = \frac{1}{1+e^{-x}}$, and $\boldsymbol{\theta}_k$ is a vector of weights for the particular gate (neural network).

When using the vertex and edge feature functions defined in Eqs.(3), (2) and (1), the probability distribution of CCNF for a particular sequence is as follows:

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp(\Psi)}{\int_{-\infty}^{\infty} \exp(\Psi) d\boldsymbol{y}} \tag{4}$$

$$\Psi = \sum_i \sum_{k=1}^{K1} \alpha_k f_k(y_i, \boldsymbol{x}) + \sum_{i,j} \sum_{k=1}^{K2} \beta_k g_k(y_i, y_j) + \sum_{i,j} \sum_{k=1}^{K3} \gamma_k l_k(y_i, y_j) \tag{5}$$

is in fact a multivariate Gaussian with the following distribution:

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^{\mathbf{T}} \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{\mu})), \tag{6}$$

where

$$\Sigma^{-1} = 2(A + B + C) \tag{7}$$

The diagonal matrix $A$ represents the contribution of $\alpha$ terms (vertex features) to the covariance matrix, and the symmetric $B$ represents the contribution of the $\beta$ and $\gamma$ terms (similarity and sparsity edge features).

$$A_{i,j} = \begin{cases} \sum_{k=1}^{K1} \alpha_k, & i = j \\ 0, & i \neq j \end{cases} \tag{8}$$

$$B_{i,j} = \begin{cases} (\sum_{k=1}^{K2} \beta_k \sum_{r=1}^{n} S_{i,r}^{(g_k)}) - (\sum_{k=1}^{K2} \beta_k S_{i,j}^{(g_k)}), & i = j \\ -\sum_{k=1}^{K2} \beta_k S_{i,j}^{(g_k)}, & i \neq j \end{cases} \tag{9}$$

$$C_{i,j} = \begin{cases} (\sum_{k=1}^{K2} \gamma_k \sum_{r=1}^{n} S_{i,r}^{(l_k)}) + (\sum_{k=1}^{K2} \gamma_k S_{i,j}^{(l_k)}), & i = j \\ \sum_{k=1}^{K2} \gamma_k S_{i,j}^{(l_k)}, & i \neq j \end{cases} \tag{10}$$

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

$$d_i = 2 \sum_{k=1}^{K1} \alpha_k h(\boldsymbol{\theta}_k^T \boldsymbol{x}_i) \tag{11}$$

$$\mathbf{d} = 2\alpha^T h(\Theta X) \tag{12}$$

Above $X$ is a matrix where the $i^{th}$ column represents $\boldsymbol{x}_i$, and $\Theta$ represents the combined neural network weights (k-th row represents the weights of the k-th gate), and (through a slight abuse of notation) $h(M)$, is an element-wise application of $h$ on each element of $M$, and thus $h(\Theta X)$ represents the response of each of the gates at each $\boldsymbol{x}_i$ (frame/time step).

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

$$\boldsymbol{\mu} = \Sigma \mathbf{d} \tag{13}$$

Defining $A$, $B$ and $\mathbf{d}$ in such a way allows us to rewrite the factors of Eq.(4) 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.(4) and the multivariate Gaussian in Eq.(6). First we plug in the feature functions in Eqs.(3) and (1) into Eq.(5)

$$\begin{aligned} \Psi &= \sum_i \sum_{k=1}^{K1} \alpha_k f_k(y_i, \boldsymbol{x}, \boldsymbol{\theta}_k) + \sum_{i,j} \sum_{k=1}^{K2} \beta_k g_k(y_i, y_j, \boldsymbol{x}) \\ &= -\sum_i \sum_{k=1}^{K1} \alpha_k (y_i - h(\boldsymbol{\theta}_k^T \boldsymbol{x}_i))^2 - \tfrac{1}{2} \sum_{i,j} \sum_{k=1}^{K2} \beta_k S_{i,j}^{g_k} (y_i - y_j)^2 \\ &\quad -\tfrac{1}{2} \sum_{i,j} \sum_{k=1}^{K3} \gamma_k S_{i,j}^{(l_k)} (y_i + y_j)^2 \end{aligned} \tag{14}$$

Now we can express the factor $\Psi$ in terms of $A$, $B$ and $\mathbf{d}$. We do this in parts starting with terms containing $\alpha$ parameters in Eq.(14).

$$\begin{aligned} & -\sum_i \sum_{k=1}^{K1} \alpha_k (y_i - h(\boldsymbol{\theta}_k^T x_i))^2 \\ & = -\sum_i \sum_{k=1}^{K1} \alpha_k (y_i^2 - 2 y_i h(\boldsymbol{\theta}_k^T x_i) + h(\boldsymbol{\theta}_k^T x_i)^2) \\ & = -\sum_i \sum_{k=1}^{K1} \alpha_k y_i^2 + \sum_i \sum_{k=1}^{K1} \alpha_k 2 y_i h(\boldsymbol{\theta}_k^T x_i) - \sum_i \sum_{k=1}^{K1} \alpha_k h(\boldsymbol{\theta}_k^T x_i)^2 \\ & = -\boldsymbol{y}^T A \boldsymbol{y} + \boldsymbol{y}^T \boldsymbol{d} - \sum_i \sum_{k=1}^{K1} \alpha_k h(\boldsymbol{\theta}_k^T x_i)^2 \end{aligned} \tag{15}$$

And now collecting terms with $\beta$ parameters in Eq.(14). Here we use the assumption that every $S^{(k)}$ is a symmetric matrix (which as a similarity matrix it should be).

$$\begin{aligned}
&-\tfrac{1}{2}\textstyle\sum_{i,j}\sum_{k=1}^{K2}\beta_k S_{i,j}^{(g_k)}(y_i-y_j)^2-\tfrac{1}{2}\sum_{i,j}\sum_{k=1}^{K3}\gamma_k S_{i,j}^{(l_k)}(y_i+y_j)^2\\
&=-\tfrac{1}{2}\textstyle\sum_{i,j}\sum_{k=1}^{K2}\beta_k S_{i,j}^{(g_k)}(y_i^2-2y_iy_j+y_j^2)-\tfrac{1}{2}\sum_{i,j}\sum_{k=1}^{K3}\gamma_k S_{i,j}^{(l_k)}(y_i^2+2y_iy_j+y_j^2)\\
&=-\tfrac{1}{2}\textstyle\sum_{i,j}\sum_{k=1}^{K2}\beta_k S_{i,j}^{(g_k)}(y_i^2+y_j^2)+\sum_{i,j}\sum_{k=1}^{K2}\beta_k S_{i,j}^{(g_k)}y_iy_j+\\
&\quad-\tfrac{1}{2}\textstyle\sum_{i,j}\sum_{k=1}^{K2}\gamma_k S_{i,j}^{(l_k)}(y_i^2+y_j^2)-\sum_{i,j}\sum_{k=1}^{K2}\gamma_k S_{i,j}^{(l_k)}y_iy_j\\
&=-\textstyle\sum_{k=1}^{K2}\beta_k\sum_{i,j}S_{i,j}^{(g_k)}y_i^2+\sum_{k=1}^{K2}\beta_k S_{i,j}^{(g_k)}\sum_{i,j}y_iy_j+\\
&\quad-\textstyle\sum_{k=1}^{K2}\gamma_k\sum_{i,j}S_{i,j}^{(l_k)}y_i^2-\sum_{k=1}^{K2}\gamma_k S_{i,j}^{(l_k)}\sum_{i,j}y_iy_j\\
&=-\boldsymbol{y}^TB\boldsymbol{y}-\boldsymbol{y}^TC\boldsymbol{y}
\end{aligned}\tag{16}$$

Combining Eqs.(14), (15), and (16). We define $e=\sum_i\sum_{k=1}^{K1}\alpha_k h(\boldsymbol{\theta}_k^T\boldsymbol{x}_i)^2$ for brevity (it's not necessary writing it out in full as it cancels out eventually). We also use the fact from Eq.(13) that $\mathbf{d}=\Sigma^{-1}\boldsymbol{\mu}$.

$$\Psi=-\boldsymbol{y}^TA\boldsymbol{y}+\boldsymbol{y}^T\boldsymbol{d}-\boldsymbol{y}^TB\boldsymbol{y}-\boldsymbol{y}^TC\boldsymbol{y}-e=-\frac{1}{2}(\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y})+\boldsymbol{y}\Sigma^{-1}\boldsymbol{\mu}-e \tag{17}$$

Using Eq.(17) in Eq.(4) we get (As $d$ does not depend on y, we can take it out of the integral, leading to it canceling out):

$$\begin{aligned}
P(\boldsymbol{y}|\boldsymbol{x}) &= \frac{\exp(\Psi)}{\int_{-\infty}^{\infty}\exp(\Psi)d\boldsymbol{y}}=\\
&= \frac{\exp(-\frac{1}{2}(\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y})+\boldsymbol{y}\Sigma^{-1}\boldsymbol{\mu})\exp(-d)}{\int_{-\infty}^{\infty}\{\exp(-\frac{1}{2}(\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y})+\boldsymbol{y}\Sigma^{-1}\boldsymbol{\mu})\exp(-d)\}d\boldsymbol{y}}\\
&= \frac{\exp(-\frac{1}{2}(\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y})+\boldsymbol{y}\Sigma^{-1}\boldsymbol{\mu})}{\int_{-\infty}^{\infty}\{\exp(-\frac{1}{2}(\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y})+\boldsymbol{y}\Sigma^{-1}\boldsymbol{\mu})\}d\boldsymbol{y}}
\end{aligned}\tag{18}$$

Now we need to find the 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].

$$\int_{\boldsymbol{y}}\{\exp(-\frac{1}{2}(\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y})+\boldsymbol{y}\Sigma^{-1}\boldsymbol{\mu})\}d\boldsymbol{y}=\frac{(2\pi)^{\frac{n}{2}}}{|\Sigma^{-1}|^{\frac{1}{2}}}\exp(\frac{1}{2}\boldsymbol{\mu}\Sigma^{-1}\boldsymbol{\mu}) \tag{19}$$

Finally. plugging Eq.(17) and (19) into Eq.(4) we get:

$$\begin{aligned}
P(\boldsymbol{y}|\boldsymbol{x}) &= \frac{\exp(-\frac{1}{2}\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y}+\boldsymbol{y}\Sigma^{-1}\boldsymbol{\mu})}{\frac{(2\pi)^{\frac{n}{2}}}{|\Sigma^{-1}|^{\frac{1}{2}}}\exp(\frac{1}{2}\boldsymbol{\mu}\Sigma^{-1}\boldsymbol{\mu})}\\
&= \frac{\exp(-\frac{1}{2}\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y}+\boldsymbol{y}\Sigma^{-1}\boldsymbol{\mu})\exp(-\frac{1}{2}\boldsymbol{\mu}\Sigma^{-1}\boldsymbol{\mu})}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\\
&= \frac{\exp(-\frac{1}{2}\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y}+\boldsymbol{y}\Sigma^{-1}\boldsymbol{\mu}-\frac{1}{2}\boldsymbol{\mu}\Sigma^{-1}\boldsymbol{\mu})}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\\
&= \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu}))
\end{aligned}\tag{20}$$

This is exactly what we set out to show.

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

# 2 Appendix B

This appendix deals with calculating the partial derivatives of the CCNF log-likelihood with respect to the parameters $\alpha$, $\beta$, and $\theta$. First of all, we would like to calculate the log-likelihood of Eq.(20)).

$$\begin{aligned}
\log(P(\boldsymbol{y}|\boldsymbol{x})) &= -\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) - \log((2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}) \\
&= -\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) - (\tfrac{n}{2}\log(2\pi) + \tfrac{1}{2}\log|\Sigma|) \\
&= -\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) + \tfrac{1}{2}\log|\Sigma^{-1}| - \tfrac{n}{2}\log(2\pi) \\
&= -\tfrac{1}{2}\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y} + \boldsymbol{y}^T\Sigma^{-1}\boldsymbol{\mu} - \tfrac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu} + \tfrac{1}{2}\log|\Sigma^{-1}| - \tfrac{n}{2}\log(2\pi) \\
&= -\tfrac{1}{2}\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{d} - \tfrac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu} + \tfrac{1}{2}\log|\Sigma^{-1}| - \tfrac{n}{2}\log(2\pi) \\
&= -\tfrac{1}{2}\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{d} - \tfrac{1}{2}\boldsymbol{d}^T\Sigma\boldsymbol{d} + \tfrac{1}{2}\log|\Sigma^{-1}| - \tfrac{n}{2}\log(2\pi)
\end{aligned} \tag{21}$$

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

Now we can derive all of the necessary partial derivatives, first we define the partial derivatives of $\Sigma^{-1}$ and $\boldsymbol{d}$ with respect to $\alpha$, $\beta$, $\gamma$ and $\theta$ 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 $\alpha$, and $B$ on $\beta$ and $\gamma$; $\boldsymbol{d}$, however, depends on both $\alpha$ and $\theta$.

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

$$\frac{\partial\Sigma^{-1}}{\partial\alpha_k} = \frac{\partial 2A + 2B}{\partial\alpha_k} = \frac{\partial 2A}{\partial\alpha_k} = 2I \tag{22}$$

$$\frac{\partial d_i}{\partial\alpha_k} = 2h(\Theta X)_{k,i} \tag{23}$$

$$\frac{\partial\boldsymbol{d}}{\partial\alpha_k} = (2h(\Theta X)_{k,*})^T \tag{24}$$

Here $X_{k,*}$ notation refers to a row vector corresponding to the $k^{\text{th}}$ row of a matrix $X$. For brevity we will use $D = h(\Theta X)$

In the derivation below, we use the partial derivative of a matrix inverse ($\frac{\partial M^{-1}}{\partial\alpha} = -M^{-1}\frac{\partial M}{\partial\alpha}M^{-1}$) to get the partial derivative of $\Sigma$.

$$\begin{aligned}
\frac{\partial\boldsymbol{d}^T\Sigma\boldsymbol{d}}{\partial\alpha_k} &= \frac{\partial\boldsymbol{d}^T}{\partial\alpha_k}\Sigma\boldsymbol{d} + \boldsymbol{d}^T\frac{\partial\Sigma\boldsymbol{d}}{\partial\alpha_k} = 2D_{k,*}\boldsymbol{\mu} + \boldsymbol{d}^T(\frac{\partial\Sigma}{\partial\alpha_k}\boldsymbol{d} + \Sigma\frac{\partial\boldsymbol{d}}{\partial\alpha_k}) \\
&= 2D_{k,*}\boldsymbol{\mu} + \boldsymbol{d}^T\frac{\partial\Sigma}{\partial\alpha_k}\boldsymbol{d} + \boldsymbol{d}^T\Sigma 2(D_{k,*})^T = 4D_{k,*}\boldsymbol{\mu} + \boldsymbol{d}^T\frac{\partial\Sigma}{\partial\alpha_k}\boldsymbol{d} \\
&= 4D_{k,*}\boldsymbol{\mu} + \boldsymbol{d}^T(-\Sigma\frac{\partial\Sigma^{-1}}{\partial\alpha_k}\Sigma)\boldsymbol{d} = 4D_{k,*}\boldsymbol{\mu} - 2\boldsymbol{d}^T\Sigma\Sigma\boldsymbol{d} \\
&= 4D_{k,*}\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\boldsymbol{\mu}
\end{aligned} \tag{25}$$

Now for the normalisation (partition) function part:

$$\begin{aligned}
\frac{\partial\log|\Sigma^{-1}|}{\partial\alpha_k} &= \frac{1}{|\Sigma^{-1}|}\frac{\partial|\Sigma^{-1}|}{\partial\alpha_k} = \frac{1}{|\Sigma^{-1}|}|\Sigma^{-1}| \times \text{trace}(\Sigma\frac{\partial\Sigma^{-1}}{\alpha_k}) \\
&= 2 \times \text{trace}(\Sigma I) = 2 \times \text{trace}(\Sigma)
\end{aligned} \tag{26}$$

Now we can combine these to get

$$\frac{\partial \log(P(\boldsymbol{y}|\boldsymbol{x}))}{\alpha_k} = -\boldsymbol{y}^T\boldsymbol{y} + 2\boldsymbol{y}^T D_{k,*}^T - 2D_{*,k}\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{\mu} + \text{trace}(\Sigma) \tag{27}$$

We can now derive the partial derivatives of the likelihood with respect to $\beta$ and $\gamma$ parameters (they are discussed together as they are so similar)

$$\frac{\partial \Sigma^{-1}}{\partial \beta_k} = 2B^{(k)} \tag{28}$$

$$\frac{\partial \Sigma^{-1}}{\partial \gamma_k} = 2C^{(k)} \tag{29}$$

$$B^{(k)} = \begin{cases} (\sum_{r=1}^{n} S_{i,r}^{(g_k)}) - S_{i,j}^{(g_k)}, & i = j \\ -S_{i,j}^{(g_k)}, & i \neq j \end{cases} \tag{30}$$

$$C^{(k)} = \begin{cases} (\sum_{r=1}^{n} S_{i,r}^{(l_k)}) + S_{i,j}^{(l_k)}, & i = j \\ S_{i,j}^{(l_k)}, & i \neq j \end{cases} \tag{31}$$

$$\frac{\partial \boldsymbol{d}}{\partial \beta_k} = 0 \tag{32}$$

$$\frac{\partial \boldsymbol{d}}{\partial \gamma_k} = 0 \tag{33}$$

$$\frac{\boldsymbol{d}^T\Sigma\boldsymbol{d}}{\beta_k} = -\boldsymbol{d}^T(\Sigma\frac{\partial \Sigma^{-1}}{\partial \beta_k}\Sigma)\boldsymbol{d} = -2\boldsymbol{d}^T\Sigma B^{(k)}\Sigma\boldsymbol{d} = -2\boldsymbol{\mu}^T B^{(k)}\boldsymbol{\mu} \tag{34}$$

$$\frac{\boldsymbol{d}^T\Sigma\boldsymbol{d}}{\gamma_k} = -\boldsymbol{d}^T(\Sigma\frac{\partial \Sigma^{-1}}{\partial \gamma_k}\Sigma)\boldsymbol{d} = -2\boldsymbol{d}^T\Sigma C^{(k)}\Sigma\boldsymbol{d} = -2\boldsymbol{\mu}^T C^{(k)}\boldsymbol{\mu} \tag{35}$$

$$\begin{aligned} \frac{\partial \log|\Sigma^{-1}|}{\partial \beta_k} &= \frac{1}{|\Sigma^{-1}|}\frac{\partial|\Sigma^{-1}|}{\partial \beta_k} = \frac{1}{|\Sigma^{-1}|}|\Sigma^{-1}| \times \text{trace}(\Sigma\frac{\partial \Sigma^{-1}}{\beta_k}) \\ &= 2 \times \text{trace}(\Sigma B^{(k)}) = 2 \times \text{Vec}(\Sigma)^T\text{Vec}(B^{(k)}) \end{aligned} \tag{36}$$

$$\begin{aligned} \frac{\partial \log|\Sigma^{-1}|}{\partial \gamma_k} &= \frac{1}{|\Sigma^{-1}|}\frac{\partial|\Sigma^{-1}|}{\partial \gamma_k} = \frac{1}{|\Sigma^{-1}|}|\Sigma^{-1}| \times \text{trace}(\Sigma\frac{\partial \Sigma^{-1}}{\gamma_k}) \\ &= 2 \times \text{trace}(\Sigma C^{(k)}) = 2 \times \text{Vec}(\Sigma)^T\text{Vec}(C^{(k)}) \end{aligned} \tag{37}$$

Here we use the matrix trace property trace$(AB)$ = Vec$(A)^T$Vec$(B)$, and where 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 as in the case with $\alpha_k$ version.

We can now combine these to get:

$$\frac{\partial \log(P(\boldsymbol{y}|\boldsymbol{x}))}{\beta_k} = -\boldsymbol{y}^T B^{(k)}\boldsymbol{y} + \boldsymbol{\mu}^T B^{(k)}\boldsymbol{\mu} + \text{Vec}(\Sigma)^T\text{Vec}(B^{(k)}) \tag{38}$$

$$\frac{\partial \log(P(\boldsymbol{y}|\boldsymbol{x}))}{\gamma_k} = -\boldsymbol{y}^T C^{(k)}\boldsymbol{y} + \boldsymbol{\mu}^T C^{(k)}\boldsymbol{\mu} + \text{Vec}(\Sigma)^T\text{Vec}(C^{(k)}) \tag{39}$$

Finally, we will derive the partial derivatives of the likelihood with respect to the $\theta$ parameters (the neural network weights). We abuse the notation slightly for clarity and brevity, $h(A)$ on a $n \times m$ size matrix $A$ produces a $n \times m$ matrix with the activation function applied on each element.

$$\frac{\partial \Sigma^{-1}}{\partial \theta_{i,j}} = 0 \tag{40}$$

If we use the sigmoid activation function $h(z) = \frac{1}{1+e^{-z}}$, and:

$$\frac{dh(z)}{dz} = h(z)(1 - h(z)) \tag{41}$$

We can now define the partial derivatives with respect to $\theta$ parameters on the parts of likelihood function:

$$b_r = 2 \sum_{k=1}^{K1} \alpha_k h(\theta_k^T x_r) \tag{42}$$

$$\frac{\partial b_r}{\partial \theta_{i,j}} = 2\alpha_i h(\boldsymbol{\theta}_i^T \boldsymbol{x}_r)(1 - h(\boldsymbol{\theta}_i^T \boldsymbol{x}_r))\boldsymbol{x}_{r,j} \tag{43}$$

$$\frac{\partial \boldsymbol{d}}{\partial \theta_{i,j}} = 2\alpha_i \{h(\boldsymbol{\theta}_i^T X) \circ (1 - h(\boldsymbol{\theta}_i^T X))\} X_{*,j} \tag{44}$$

Here $\circ$ is the Hadamard or element-wise product.

$$\begin{aligned} \frac{\partial \boldsymbol{d}^T \Sigma \boldsymbol{d}}{\partial \theta_{i,j}} &= \frac{\partial \boldsymbol{d}^T}{\partial \theta_{i,j}} \Sigma \boldsymbol{d} + \boldsymbol{d}^T \frac{\partial \Sigma \boldsymbol{d}}{\partial \theta_{i,j}} = \frac{\partial \boldsymbol{d}^T}{\partial \theta_{i,j}} \boldsymbol{\mu} + \boldsymbol{\mu}^T \frac{\partial \boldsymbol{d}}{\partial \theta_{i,j}} \\ &= 2\boldsymbol{\mu}^T \frac{\partial \boldsymbol{d}}{\partial \theta_{i,j}} \end{aligned} \tag{45}$$

We can now combine these to get

$$\frac{\partial \log(P(\boldsymbol{y}|\boldsymbol{x}))}{\theta_{i,j}} = \boldsymbol{y}^T \frac{\partial \boldsymbol{d}}{\partial \theta_{i,j}} - \boldsymbol{\mu}^T \frac{\partial \boldsymbol{d}}{\partial \theta_{i,j}} = (\boldsymbol{y} - \boldsymbol{\mu})^T (2\alpha_i \{h(\boldsymbol{\theta}_i^T X) \circ (1 - h(\boldsymbol{\theta}_i^T X))\} X_{*,j}) \tag{46}$$

Which is basically the update of a single layer neural network (back propagation) with sigmoid activation where the current feed-forward prediction is $\boldsymbol{\mu}$ and error is $(\boldsymbol{y} - \boldsymbol{\mu})$. The difference, however, is that different neurons are given different weights (depending on the corresponding $\alpha$ values).