Some notes on likelihood of fits

Consider the case with one tree with b leaves (aka bottom notes) and responses $y_1, \ldots, y_n$. A certain subset of the y's wind up in the first node, say $y_{k_{1,1}}, y_{k_{1,2}}, \ldots y_{k_{1,n_1}}$, a certain subset of the y's wind up in the second node, say $y_{k_{2,1}}, y_{k_{2,2}}, \ldots y_{k_{2,n_2}}$, etc. Of course the number of data points in all the leaves must keep to $n = n_1 + \ldots + n_b$.

Now consider the first leaf. We consider this data to be realizations of independent and identically distributed normals with common mean, μ_1, and common variance which is the variance of the error distribution (remember, $\boldsymbol{\mathcal{E}} \sim \mathcal{N}_n\left(\mathbf{0},\, \sigma^2 \boldsymbol{I}_n\right)$). Thus we have:

$$Y_{k_{1,1}}, Y_{k_{1,2}}, \ldots Y_{k_{1,n_1}} \overset{iid}{\sim} \mathcal{N}\left(\mu_1,\, \sigma^2\right)$$

In our Bayesisan setup, we have the following prior on μ_1:

$$\mu_1 \sim \mathcal{N}\left(0,\, \sigma_\mu^2\right)$$

We can think of the Y's as the sum of their mean and their disturbance like so for *e.g.* the first Y in the first node looks like:

$$Y_{k_{1,1}} = \mu_1 + \mathcal{E}_{k_{1,1}}$$

Since the mean and the errors are independent, we can do the marginalization pretty easily via:

$$Y_{k_{1,1}} \sim \mathcal{N}\left(0,\, \sigma_\mu^2\right) * \mathcal{N}\left(0,\, \sigma^2\right) = \mathcal{N}\left(0,\, \sigma_\mu^2 + \sigma^2\right)$$

However, the Y's are no longer independent. If we've margined out the mean, then each of the realizations y will tell us something about the mean. This can be seen in the following calculation:

$$\begin{aligned}
\mathbb{C}\text{ov}\left[Y_{k_{1,1}}, Y_{k_{1,2}}\right] &= \mathbb{C}\text{ov}\left[\mu_1 + \mathcal{E}_{k_{1,1}}, \mu_1 + \mathcal{E}_{k_{1,2}}\right] \\
&= \mathbb{C}\text{ov}\left[\mu_1, \mu_1\right] + \underbrace{\mathbb{C}\text{ov}\left[\mu_1, \mathcal{E}_{k_{1,2}}\right]}_{0} + \underbrace{\mathbb{C}\text{ov}\left[\mathcal{E}_{k_{1,1}}, \mu_1\right]}_{0} + \underbrace{\mathbb{C}\text{ov}\left[\mathcal{E}_{k_{1,1}}, \mathcal{E}_{k_{1,2}}\right]}_{0} \\
&= \mathbb{V}\text{ar}\left[\mu_1\right] \\
&= \sigma_\mu^2
\end{aligned}$$

Where each of the above zeroes are due to assumed independence.

Thus, the joint distribution of all the Y's looks like the following:

$$\boldsymbol{Y} := \begin{bmatrix} Y_{k_{1,1}} \\ \vdots \\ Y_{k_{1,n_1}} \end{bmatrix} \sim \mathcal{N}_{n_1}\left(\mathbf{0}, \begin{bmatrix} \sigma_\mu^2 + \sigma^2 & \sigma_\mu^2 & \ldots & \sigma_\mu^2 \\ \sigma_\mu^2 & \sigma_\mu^2 + \sigma^2 & \ldots & \sigma_\mu^2 \\ \vdots & & \ddots & \vdots \\ \sigma_\mu^2 & \ldots & \ldots & \sigma_\mu^2 + \sigma^2 \end{bmatrix}\right)$$

Let's denote the above variance matrix as $\boldsymbol{\Sigma}$ which is called an "equicorrelation" matrix. Thus the joint density looks like the following:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{(2\pi)^{n_1} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\boldsymbol{y}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{y}\right)$$

Looking in a standard text on matrix algebra, we can find formulas for the inverse and the determinant of an equicorrelation matrix:

$$\begin{aligned} |\boldsymbol{\Sigma}| &= \left(\sigma^2\right)^{n-1}\left(\sigma_\mu^2 n_1 + \sigma^2\right) \\ \boldsymbol{\Sigma}^{-1} &= \frac{1}{\sigma_\mu^2 n_1 + \sigma^2} \frac{1}{n_1} \boldsymbol{J}_{n_1} + \frac{1}{\sigma^2}\left(\boldsymbol{I}_{n_1} - \frac{1}{n_1}\boldsymbol{J}_{n_1}\right) \end{aligned}$$

These are the formulas I use to compute the proportional log likelihoods $\mathbb{P}(Y|T)$ in my implementation. Note that I still do the matrix algebra inside of the $\exp(\cdot)$. I couldn't figure out a way around that.

Please let me know if you both agree with this. Thanks!