Likelihood Based Clustering of Data

Consider the problem of modeling independent samples generated by \(S\) different sources using \(\textit{possibly distinct}\) Gaussian densities \(\mathcal{N}(y; m_s, U_s)\) with parameters \(\langle m_s, U_s \rangle\), \(s = 1, \ldots, S\). Let

\[
Y^1 = \{y^1_1, y^1_2, \ldots, y^1_{N^1}\}, \\
Y^2 = \{y^2_1, y^2_2, \ldots, y^2_{N^2}\}, \\
\vdots \\
Y^S = \{y^S_1, y^S_2, \ldots, y^S_{N^S}\},
\]

\(y^s_n \in \mathbb{R}^d\ \forall \ s = 1, \ldots, S, \ n = 1, \ldots, N^s\),

denote the observed data, and let the sample sum and sample sum-of-squares be denoted by

\[
s_s = \sum_{n=1}^{N^s} y^s_n \quad \text{and} \quad Q_s = \sum_{n=1}^{N^s} y^s_n y^s_n^T \quad \text{respectively,} \quad s = 1, \ldots, S.
\]

This problem will explore the question of \textit{clustering} these sources via the observed data.

1. If each source \(s\) is modeled by a \textit{different} Gaussian density, what parameter values \(\langle \hat{m}_s, \hat{U}_s \rangle\) maximize the total likelihood of the data, \(\prod_{s=1}^{S} \prod_{n=1}^{N^s} \mathcal{N}(y^s_n; m_s, U_s)\)? Express your answer(s) in terms of the statistics \(s_s\) and \(Q_s\).

2. Compute the value of this maximum total likelihood in terms of the \(\hat{U}_s\)'s.

3. If two sources \(i\) and \(j\) are are assumed to \textit{share} a Gaussian density, what \textit{tied} parameter values \(\langle \hat{m}_{(i,j)}, \hat{U}_{(i,j)} \rangle\) maximize the total likelihood of \(Y^i \cup Y^j\), \(\prod_{s=i,j} \prod_{n=1}^{N^s} \mathcal{N}(y^s_n; m_{(i,j)}, U_{(i,j)})\)? Express your answer(s) in terms of \(s_i, s_j, Q_i\) and \(Q_j\).

4. Use your answer from part 3 to describe a procedure for choosing two sources, say, \(i^*\) and \(j^*\), such that \textit{tying} together their Gaussian densities results in a \textit{higher} total likelihood of the data than tying together any other pair of sources.

\[
(i^*, j^*) = \arg \max_{i,j \in \{1, \ldots, S\}, i \neq j} \left[ \prod_{s \neq i,j} \prod_{n=1}^{N^s} \mathcal{N}(y^s_n; \hat{m}_s, \hat{U}_s) \times \prod_{s=i,j} \prod_{n=1}^{N^s} \mathcal{N}(y^s_n; \hat{m}_{(i,j)}, \hat{U}_{(i,j)}) \right]
\]
Hint: you may want to write the likelihood in the square-brackets above as
\[
\left[ \prod_{s=1}^{S} \prod_{n=1}^{N_s} \mathcal{N}(y_{s,n}; \hat{m}_s, \hat{U}_s) \prod_{s=i,j} \prod_{n=1}^{N_s} \mathcal{N}(y_{s,n}; \hat{m}_{i,j}, \hat{U}_{i,j}) \right],
\]
and work with the log-likelihood in order to simplify your computation.

5. Show that the maximum total likelihood in part 4 is necessarily less than or equal to the maximum total likelihood in part 2.

6. Extend your procedure from part 4 to come up with an “algorithm” for bottom-up clustering of the $S$ data sources, and interpret each internal node of the tree-structured hierarchy in terms of similarity of the sources under a Gaussian model.

Finally, for any arbitrary two-way clustering of the $S$ sources, say, $\Phi_0 = \{Y^1, \ldots, Y^{S_1}\}$ and $\Phi_1 = \{Y^{S_1+1}, \ldots, Y^S\}$, compute the maximum total data likelihood
\[
\prod_{s=1}^{S_1} \prod_{n=1}^{N_s} \mathcal{N}(y_{s,n}; \hat{m}_{\Phi_0}, \hat{U}_{\Phi_0}) \prod_{s=S_1+1}^{S} \prod_{n=1}^{N_s} \mathcal{N}(y_{s,n}; \hat{m}_{\Phi_1}, \hat{U}_{\Phi_1})
\]
attainable by a two-Gaussian model, and compare it with that attainable under a one-Gaussian model
\[
\prod_{s=1}^{S} \prod_{n=1}^{N_s} \mathcal{N}(y_{s,n}; \hat{m}, \hat{U}).
\]
Express you answers in terms of $\hat{U}_{\Phi_0}$, $\hat{U}_{\Phi_1}$ and $\hat{U}$.

**Reading Assignment**

Read the following paper and summarize it in your own words in no more than 3 pages.