CS Ph.D. Qualifying Exam in CSE

1 Data Analysis

1. We are interested in ranking (ordering) a set of items, where for an item x is also associated a numerical grade y, and we assume y takes a finite number of values. Assume (x, y) is distributed according to $P(x, y)$, and one way to rank a set of items is to use the condition mean
$$c(x) = \sum_y yP(y|x),$$
as the ranking function, where $P(y|x)$ is the conditional probability, and the summation is over all distinct y values. This is to say, given a set of items $\{x_1, \ldots, x_n\}$, we sort the values $\{c(x_1), \ldots, c(x_n)\}$ from small to large which induces a ranking (ordering) for $\{x_1, \ldots, x_n\}$.

   We want to investigate whether the ranking (ordering) will change if we instead use
$$c_f(x) = \sum_y f(y)P(y|x)$$
as the ranking function, where f is a strictly monotonically increasing function.

   1) Show if f is linear, using $c(x)$ and $c_f(x)$ as the ranking functions produces the same ranking (ordering).

   2) Show if y can take exactly two distinct values, for an arbitrary f which is strictly monotonically increasing, using $c(x)$ and $c_f(x)$ as the ranking functions produces the same ranking (ordering).

   3) Show 2) is not true if y can take more than two distinct values.

2. Two key limitations of EM-based algorithm for learning Gaussian mixtures are:
i. The number of mixture components has to be specified in advance.
ii. The EM approach may find only a local minimum in the data likelihood.

   For both of these limitations answer the following questions.

   (a) Describe approaches that can, in part at least, address the limitations.

   (b) Discuss one significant strength of each of your approaches.

   (c) Discuss one significant weakness of each of your approaches.