Abstract:

Given multiple corrupted versions of the same text, as is common with ancient manuscripts, we wish to reconstruct the original text from which the extant corrupted versions were copied (possibly via latent intermediary versions). This is a challenge of cardinal importance in the humanities. We use a variant of EM to solve this problem and demonstrate the efficacy of the method on both synthetic and real-world data.