The EM Algorithm, Its Randomized Implementation and Global Optimization: Some Challenges and Opportunities for Operations Research

Wolfgang Jank

Robert H. Smith School of Business
Department of Decision and Information Technologies
University of Maryland
College Park, MD 20742
wjank@rhsmith.umd.edu

Summary. The EM algorithm is a very powerful optimization method and has become popular in many fields. Unfortunately, EM is only a local optimization method and can get stuck in sub-optimal solutions. While more and more contemporary data/model combinations yield multiple local optima, there have been only very few attempts at making EM suitable for global optimization. In this paper we review the basic EM algorithm, its properties and challenges, and we focus in particular on its randomized implementation. The randomized EM implementation promises to solve some of the contemporary data/model challenges, and it is particularly well-suited for a wedding with global optimization ideas, since most global optimization paradigms are also based on the principles of randomization. We review some of the challenges of the randomized EM implementation and present a new algorithm that combines the principles of EM with that of the Genetic Algorithm. While this new algorithm shows some promising results for clustering of an online auction database of functional objects, the primary goal of this work is to bridge a gap between the field of statistics, which is home to extensive research on the EM algorithm, and the field of operations research, in which work on global optimization thrives, and to stimulate new ideas for joint research between the two.

Key words: Monte Carlo EM; stochastic optimization; mixture model; clustering; global optimization; online auctions; functional objects.

1 Introduction

In this paper we want to shed new light on the Expectation-Maximization (EM) algorithm. The EM algorithm is a highly successful tool especially in statistics, but it has also received much attention outside the field. Applications include hierarchical models, neural networks, clustering and text mining.

The basic algorithm has been modified many times to overcome today's challenges such as complex data-models and huge databases. Yet, to this day, there has been barely any attempt at making EM suitable for solving global optimization problems.

The EM algorithm has experienced much success which can be partly attributed to its unique properties. One of these properties is that, in contrast to many other optimization methods, it guarantees an increase in the likelihood function in every iteration of the algorithm. Another property is that, since it operates on the log scale, it allows for significant analytical and numerical simplifications, especially for models in the exponential family. However, one of the biggest shortcomings of EM is that it is only a local optimization procedure and can consequently get stuck in sub-optimal solutions. One possible reason why this problem has not experienced much attention is that most applications of EM, especially within the statistics literature, have been simple enough to not experience significant impediments due to this shortcoming. However, increasing model- and data-complexity make this shortcoming more and more of a concern. It is well-known, for instance, that the mixture-model can have many local solutions, depending on the number of mixtures and the size and the dimension of the data, and that, as a consequence, the EM algorithm can produce solutions far from the true solution. EM is therefore more and more likely to yield unsatisfactory results in real-world applications.

Methods to find the global (and true) optimum are very prominent in the literature surrounding operations research, but surprisingly only very few of these methods have found their way into the classical statistics literature. This is very likely one of the reasons why EM is (still) associated with only local optimization qualities. Another reason may be a simple disconnect in language between the fields of operations research and statistics. The goal of this work is to bridge this gap and to stimulate joint research effort between the two fields.

In particular, we aim at bridging this gap by pointing to *randomized* EM implementations. Randomized EM implementations have been proposed to overcome complicated model-structures, and, more recently, also to overcome computational inefficiency due to huge databases. Randomized EM implementations have also been shown to be able to overcome local traps; however the ability to do so is based purely on chance, and there are no features in place to more systematically steer-free of such traps. Yet, one of the key features of randomized EM implementations is that, in contrast to the traditional EM algorithm, they are similar in nature to global optimization methods. Global optimization methods, like randomized EM implementations, exploit the principles of randomness to overcome local solutions, yet, unlike randomized EM, they do so more systematically. Thus, the randomized version of EM appears to be the natural ground for a wedding with the principles of global optimization.

We proceed as follows. In Section 2 we introduce the EM algorithm together with its basic properties and motivate the need for global optimization