Maximum Likelihood Estimation of Mixture Proportions Using Sequential Quadratic Programming

Joel Chacón

Centro de Investigación de Matemáticas

November 07, 2018

Outline

- Introduction
- Maximum Likelihood Estimate
- Proposal
- MIX-SQP Algorithm
- Experimental Validation
- Conclusions

Introduction

We consider maximum likelihood estimation of the mixture proportions in a finite mixture model where the component densities are known.

$$p(.|x) = \sum_{k=1}^{m} x_k g_k(.) \tag{1}$$

where the component densities $g_k(.)$ are known, and $x = (x_1, ..., x_m)^T$ denotes the unknown mixtures proportions ($x \geq 0$, and $\sum_{k=1}^{m} x_k = 1$).

Maximum Likelihood Estimate (MLE)

Finding the maximum likelihood estimate (MLE) of x can be written as

$$\begin{aligned} minimize \quad & f(x) = -\frac{1}{n}\sum_{j=1}^{n} log(\sum_{k=1}^{m} L_{jk}x_k) \\ subject \quad to \quad & x \in S^m = \{x : \sum_{k=1}^{m} x_k = 1, x \preceq 0\} \end{aligned} \tag{2}$$

where $L_{jk} = g_k(z_j) > 0$ $L \in \Re^{n \times m}$, $x \in \Re^{m \times 1}$.
This is a convex optimization problem, and can be solved by using the expectation maximization (EM) algorithm.

Reformulations MLE

There are several reformulations to solve the MLE, one popular alternative is defined as follows.
Primal formuation

$$\begin{aligned} &minimize \quad -\frac{-1}{n}\sum_{j=1}^{n} log(y_j) \\ &subject \quad to \quad Lx = y, \quad x \in S^m \end{aligned} \tag{3}$$

Dual formulation

$$\begin{aligned} &minimize \quad -\frac{1}{n}\sum_{j=1}^{n} log(v_j) \\ &subject \quad to \quad L^T v \preceq n\mathbf{1}_m, \quad v \in \Re^n \succeq 0 \end{aligned} \tag{4}$$

where $\mathbf{1}_m$ is a vector of ones of dimension m.

Proposal

The authors developed a new optimization approach, based on sequential quadratic programming (SQP).

- It solves a reformulation of the original primal problem rather than the dual.
- It uses SQP instead of IP, with the goal of making full use of the expensive gradient and Hessian.
- It uses low-rank approximations for much faster gradient and Hessian.

Proposal

Definition

A function $\Phi : \Re_+^m \longrightarrow \Re$ *is said scale invariant if for any* $c > 0$ *there exists a* $C \in \Re$ *such that for any* $x \in \Re_+^m$ *we have* $\Phi(cx) = \Phi(x) - C$*.*

Proposal

Theorem

Given the simplex-constrained optimization problem

$$minimize \quad \Phi(x) \quad subject \quad to \quad x \in S^m \tag{5}$$

where $\Phi(x)$ is scale invariant, convex, and nonincreasing with respecto to x. Let $x^*(\lambda)$ denote the solution to the Lagrangian relaxation problem.

$$\begin{aligned} minimize \quad & \Phi_\lambda(x) = \Phi(x) + \lambda \sum_{k=1}^{m} x_k \\ s.t. \quad & x \in \Re_+^m = \{x \in \Re^m : x \succeq 0\} \end{aligned} \tag{6}$$

Then $x^* = x^*(\lambda) / \sum_{k=1}^{m} x_k^*(\lambda)$ is a solution to (5).

Corollary

The target optimization problem 2 can be solved by instead solving 6 with $\Phi(x) = f(x)$; that is:

$$minimize \quad f^*(x) = f(x) + \sum_{k=1}^{m} x_k \quad s.t. \quad x \succeq 0 \tag{7}$$

Furthermore, in this case setting $\lambda = 1$ yields a solution that is normalized; in other words, $x^* = x^*(\lambda)$ when $\lambda = 1$.

Sequential quadratic programming

The Equation (7) is solved using SQP with backtracking line search. Specifically, the search direction p^t is obtained by solving the following non-negatively-constrained quadratic program *QP subproblem*.

$$p^t = arg \quad min \quad \frac{1}{2} p^T H_t p + p^T g_t \quad s.t. \quad x^t + p \succeq 0 \tag{8}$$

where $g_t = \nabla f^*(x^t)$ and $H_t = \nabla^2 f^*(x^t)$ are the gradient and Hessian of $f^*(x)$ at x^t.
The backtracking line search is determinated through a suficient desent step $x^{t+1} = x^t + \alpha_t p^t$ for $\alpha_t \in [0, 1]$.

Gradient and Hessian evaluations

Lemma

For all $x \in \Re_+^m$, the gradient and Hessian of the objective function in Equation (7) at x are

$$g = -\frac{1}{n} L^T d + \mathbf{1}_m, \quad H = \frac{1}{n} L^T diag(d)^2 L \tag{9}$$

where $\mathbf{1}_m$ is the vector of ones of dimension m and where $d = (d_1, ..., d_n)^T$ is the column vector with entries $d_j = 1/(Lx)_j$. Furthermore, for all $x \in \Re_+^m$ we have that

$$x^T g = 0 \quad x^T H x = 1 \quad Hx + g = \mathbf{1}_m \tag{10}$$

Gradient and Hessian evaluations

- In practice, L is often numerically rank deficient, with rank $r < m$.
- Therefore, the RRQR factorization or rank-revealing QR factorization is taken into account.
- Particularly, the RRQR is a matrix decomposition algorithm based on the QR factorization, which can be used to determine the rank of a matrix efficiently.
- Given the RRQR approximation of L, $\hat{L} = QRP^T$, with $Q \in \Re^{n \times r}$, $R \in \Re^{r \times m}$ and permutation matrix $P \in \Re^{m \times m}$.
- The gradient and Hessian can be approximated as is indicated in Equation (11).

$$\hat{d}_j = 1/(QRP^T x)_j, \quad \hat{g} = \frac{-1}{n} PR^T Q^T d + \mathbf{1}_m, \quad \hat{H} = \frac{1}{n} PR^T Q^T diag(d)^2 QRP^T \tag{11}$$

Solving the QP subproblem

To solve Equation (13), we solve the equivalent problem

$$y^* = arg_y min \frac{1}{2} y^T H_t y + y^T a_t \quad s.t. \quad y \succeq 0 \tag{12}$$

where $a_t = -H_y x^t + g_t$. This problem is derivated by subsituting $y = x^t + p$ into Equation (13).

$$p^t = arg \quad min \quad \frac{1}{2} p^T H_t p + p^T g_t \quad s.t. \quad x^t + p \succeq 0 \tag{13}$$

The MIX-SQP algorithm

Algorithm 1 MIX - SQP

Inputs: likelihood matrix $L \in \Re^{n \times m}$, initial iterate $x^{(0)} \in \Re^m_+$, sufficient decrease parameter $0 < \psi < 1$ (default is 0.5), step size reduction $0 < \rho < 1$ (default is 0.5), convergence tolerance $\epsilon > 0$ (default is 10^{-8}).
$Q, R, P = pqrfact(L)$
for t =0,1,2,... **do**
  $d_t = 1/(QRP^T x^t)$
  $g_t = \frac{1}{n} PR^T Q^T d_t + \mathbf{1}$
  $H_t = \frac{1}{n} PR^T Q^T diag(d_t)^2 QRP^T$
  $p_t = IP - Procedure(x^t, g_t, H_t)$
  if $min_k|(g_t)_k| < \epsilon$ and $||p_t|| < \epsilon$ **then**
    break
  end if
  $\alpha_t = 1$
  while $f(x^t + \alpha_t p_t) > f(x^t) + \alpha_t \psi p_t^T g_t$ **do**
    $\alpha_t = \rho \alpha_t$
  end while
  $x^{t+1} = x^t + \alpha_t p_t$
end for
return x^t

Experimental validation

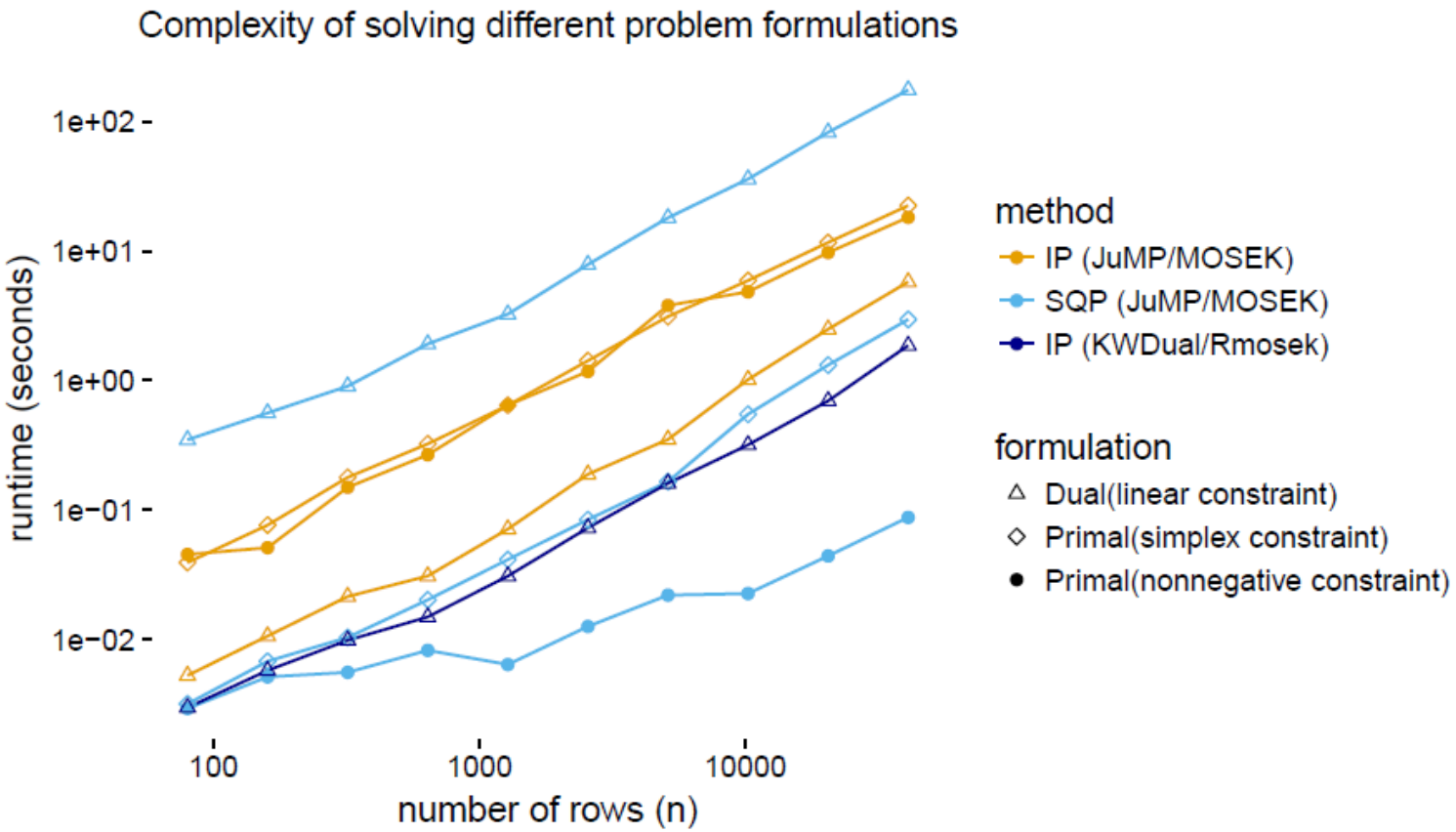

Figure 1: Runtimes for different formulations of the maximum-likelihood estimation problem: dual, simplex-constrained, and non-negatively-constrained.

Experimental validation

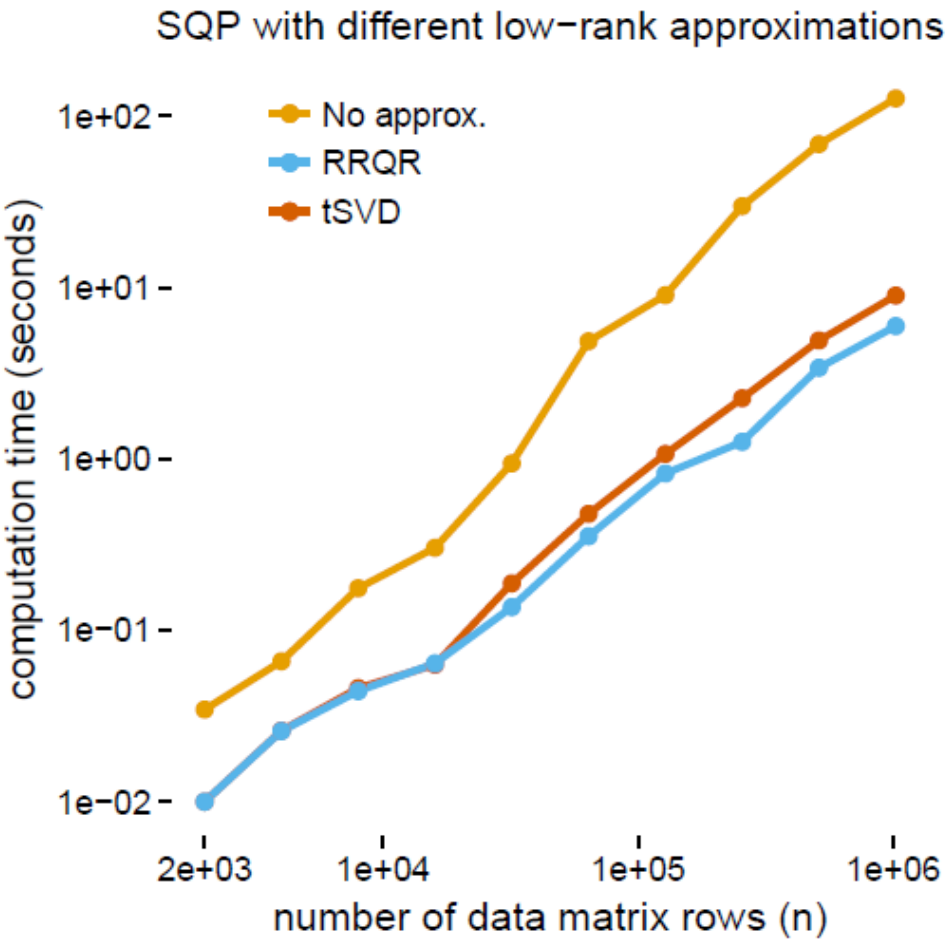

Figure 2: Comparison of SQP methods with and without exploiting the numerically lowrank structure of the mn matrix L.

Experimental validation

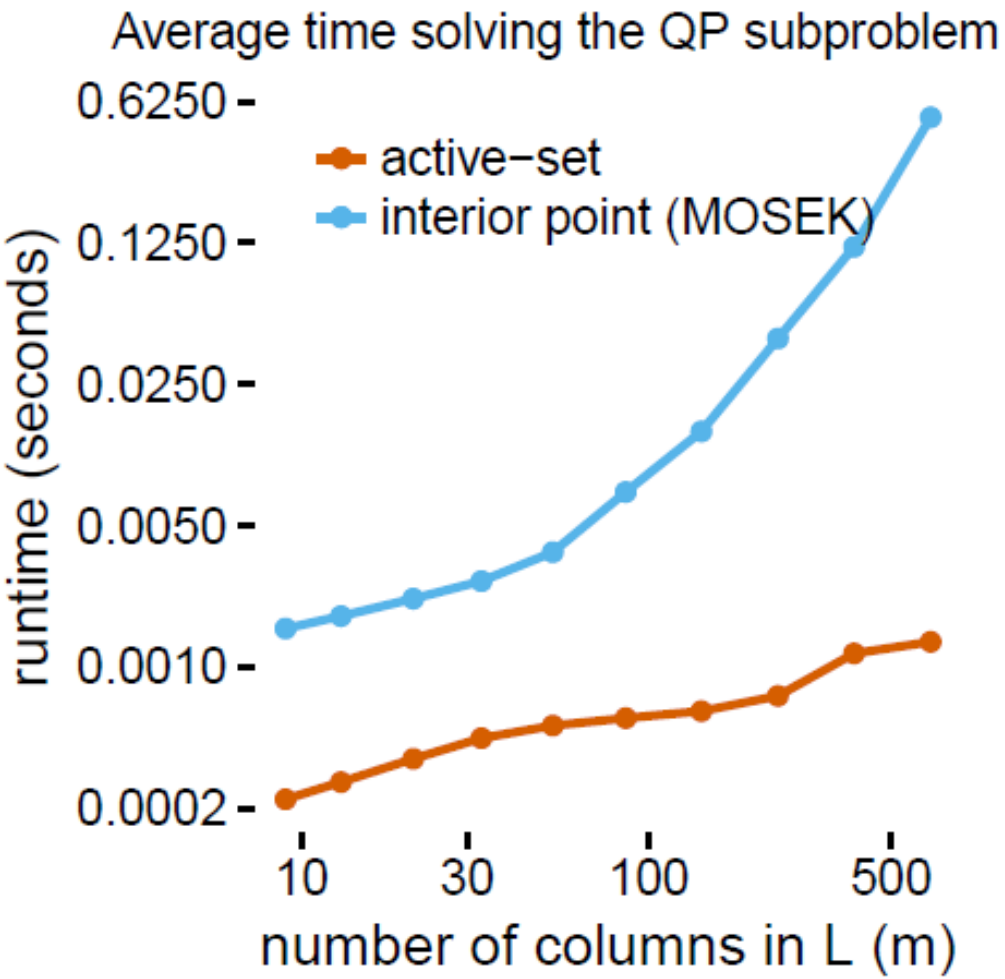

Figure 3: Comparison of active set and IP methods for solving QP subproblem inside SQP.

Experimental validation

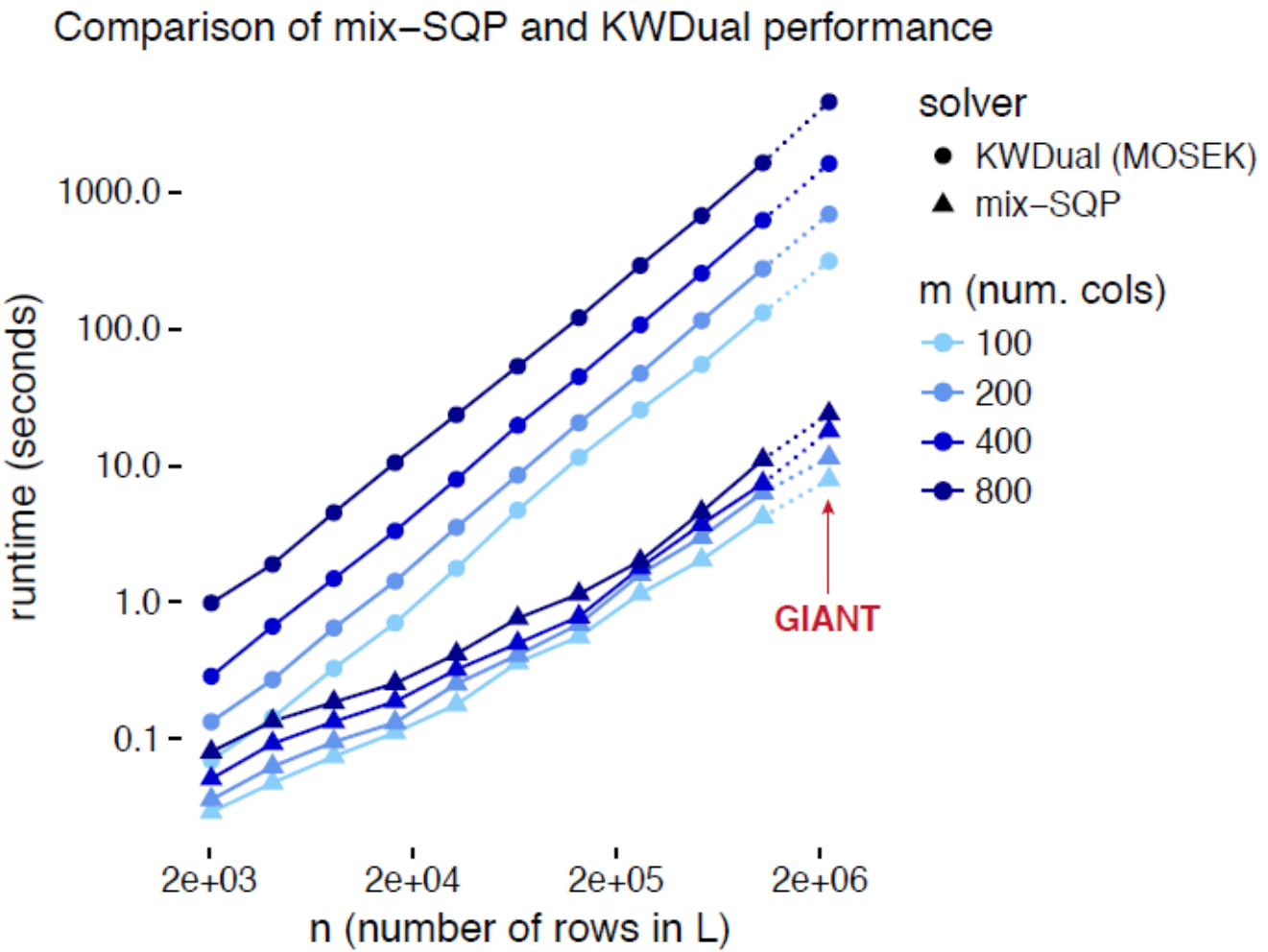

Figure 4: Runtimes of MIX-SQP andthe MOSEK interior point solve applied to simulated data, and to the GIANT data set ($n = 2,126,678$).

Experimental validation

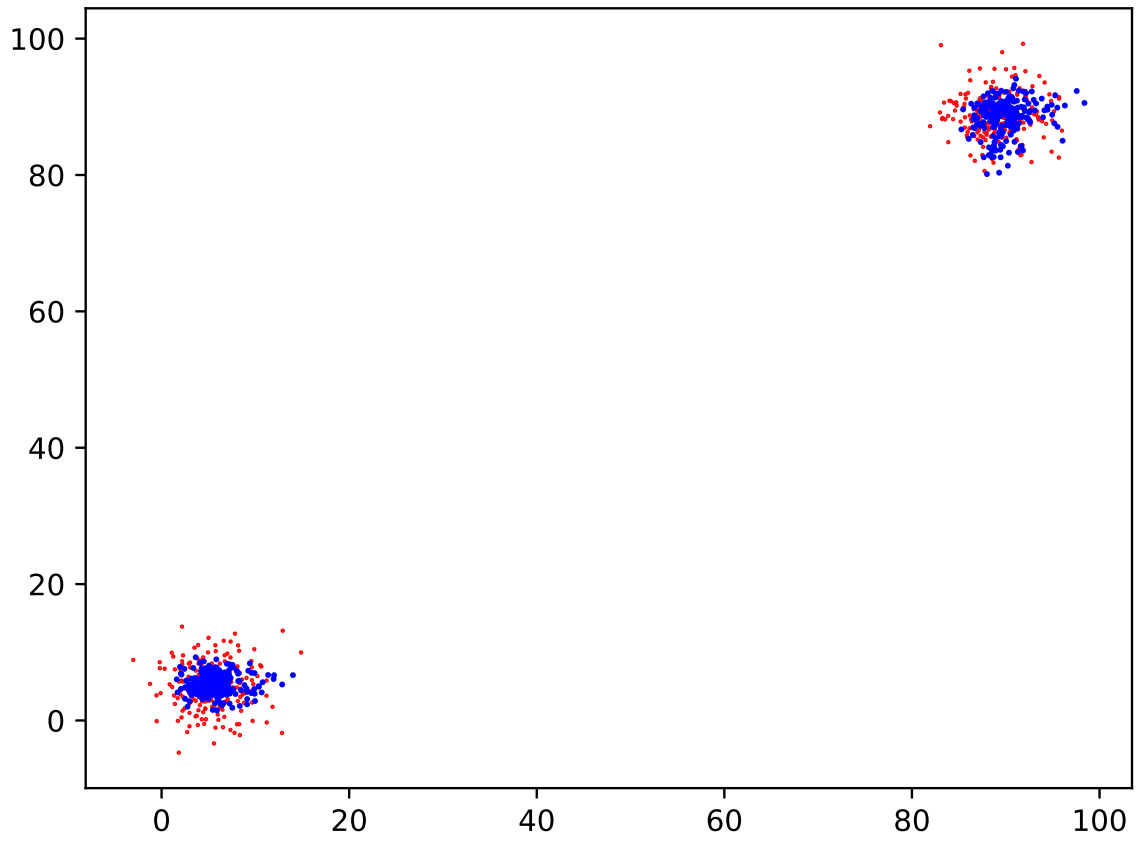

Figure 5: Estimating the data through a grid of normal bi-variate distributions, in blue are the computed points and with red the true points $(m = 800, n = 200)$.

Experimental validation

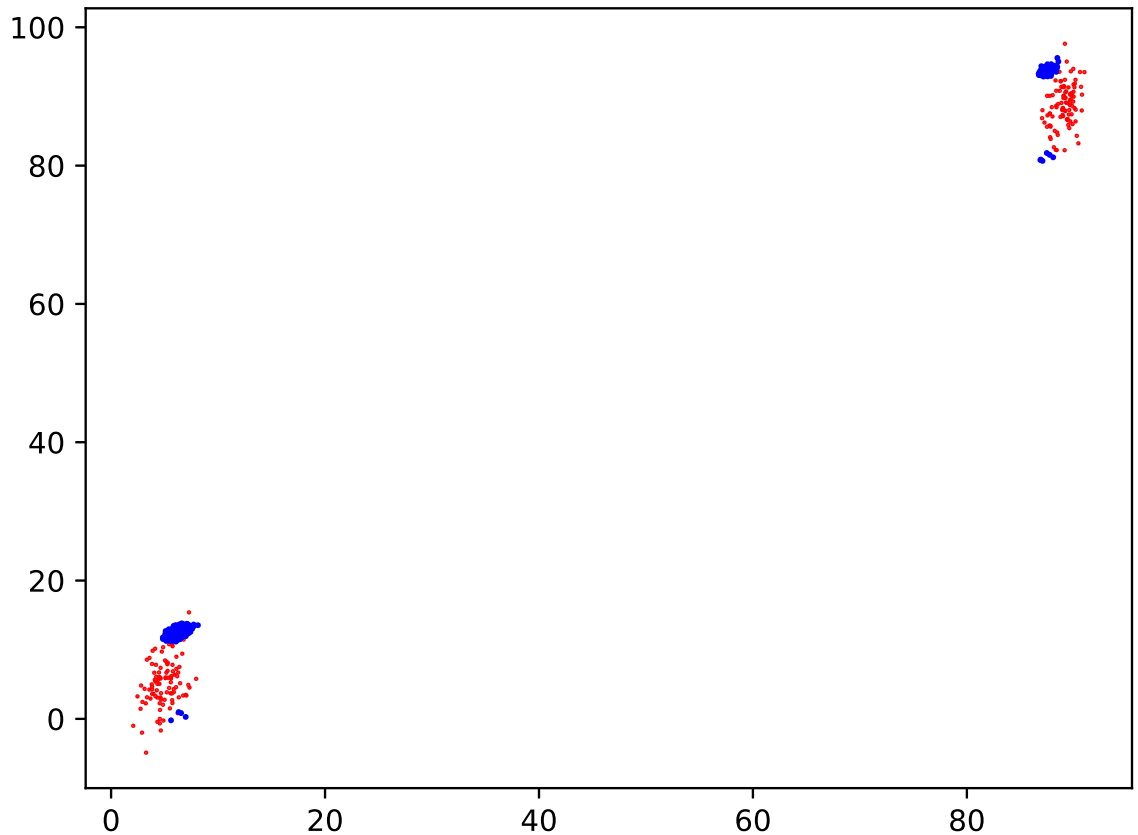

Figure 6: Estimating the data through a grid of normal bi-variate distributions, in blue are the computed points and with red the true points ($m = 800, n = 200$), considering a closest distribution shape.

Experimental validation

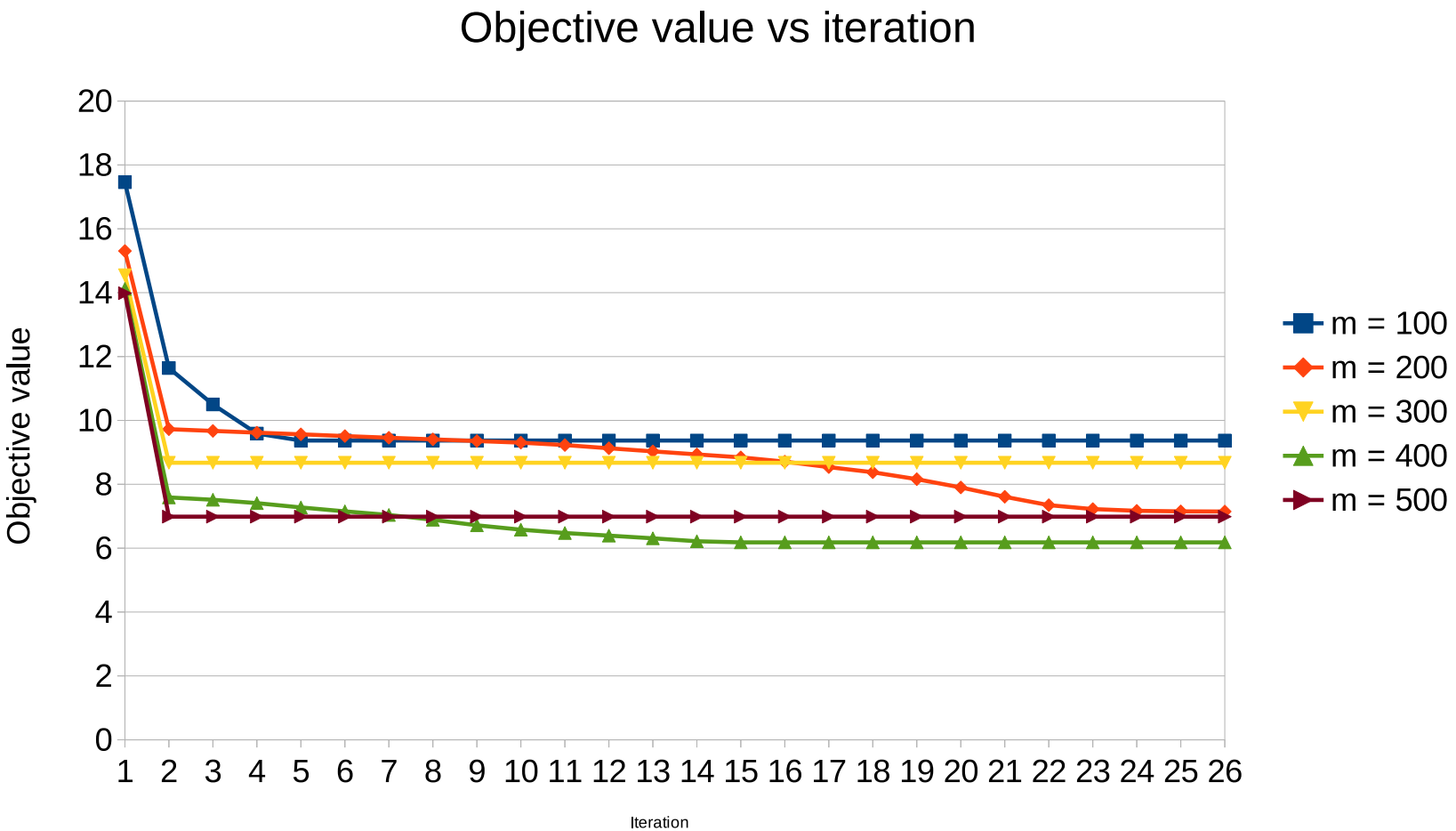

Figure 7: Iterations vs objective value ($n = 100$).

Experimental validation

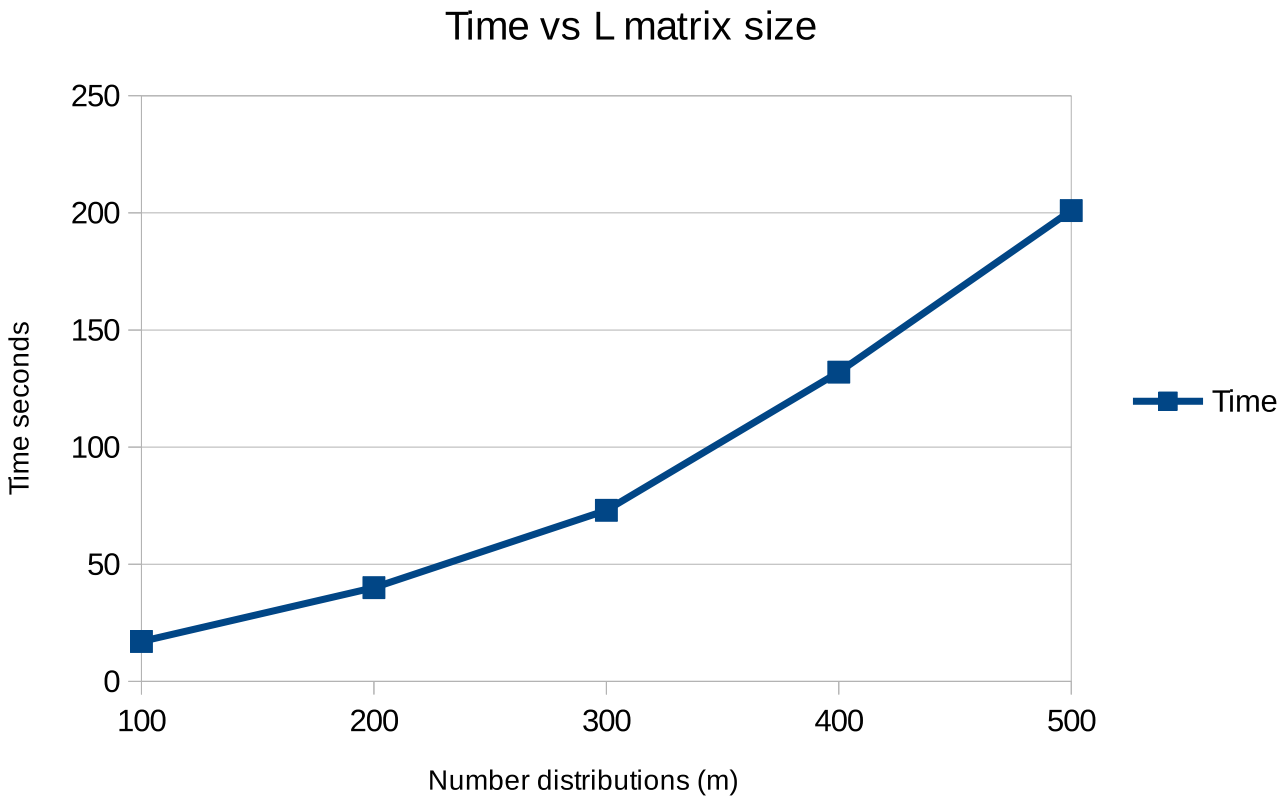

Figure 8: Iteraciones respecto al valor de la función objetivo.

Conclusions

- The benefits of the methods with the last proposal are evident in which the number of mixture components is moderated (up to one thousand).
- The factorization RRQR is a effective technique to reduce the computational complexity of Hessian evaluation.
- The active set algorithm can take advantage of sparsity in the solution vector.
- Modern convex optimization methods can be substantially faster and more reliable than EM algorithms.

Acknowledgements

This work has been supported by the Center for Research in Mathematics.