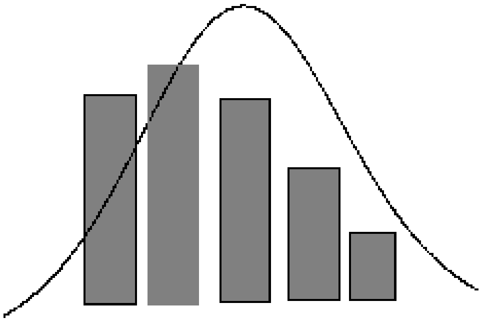
ATHENS UNIVERSITY OF
ECONOMICS

A CAUTIONARY NOTE ABOUT THE EM ALGORITHM FOR FINITE EXPONENTIAL MIXTURES

Dimitris Karlis
Department of Statistics
Athens University of Economics and Business

Technical Report No 150, August, 2001



DEPARTMENT OF STATISTICS
TECHNICAL REPORT

A CAUTIONARY NOTE ABOUT THE EM ALGORITHM FOR FINITE EXPONENTIAL MIXTURES

Dimitris Karlis
Department of Statistics,
Athens University Of Economics and Business
76 Patission Str, 10434, Athens, GREECE

Abstract

The aim of this note is to present using an example that in certain circumstances, while fitting finite exponential mixtures, solutions that are considered as local maxima are due to false detection of convergence of the EM algorithm. In this note, using a simulated example we demonstrate that all the criteria proposed in the literature detect convergence early leading to a solution that it is not the global maximum. If we let the algorithm to run for a large number of further iterations the global maximum is obtained. This clearly demonstrates the problematic behavior of the convergence criteria.

1. Introduction

Consider the simple k-finite exponential mixture model having density function of the form

$$f_P(x) = \sum_{j=1}^{k} p_j \frac{\exp(-x/\theta_j)}{\theta_j} \quad ,$$

where $0 \le p_j \le 1$, j=1,...,k with $\sum_{j=1}^{k} p_j = 1$ and θ_j is the parameter of the exponential distribution of the *j*-th component. The p_j's are called mixing proportions and they can be regarded as the probability that a randomly selected observation belongs to the j-th subpopulation. With P we denote the mixing distribution, which is the distribution that gives positive probability mass p_j at the points θ_j, j=1, . . . ,k and zero elsewhere. We use the notation $\varphi = (\varphi_1, \varphi_2, ..., \varphi_{2k-1}) = (p_1, ..., p_{k-1}, \theta_1, ..., \theta_k)$ for the whole vector of parameters.

Seidel *et al.* (1997, 2000) demonstrated the high dependence of the solution obtained from the EM algorithm to both the initial values and the stopping criterion. The aim of the present report is to add upon these papers and to demonstrate further that all the existing criteria for detecting the convergence of the algorithm may fail and stop the algorithm quite early reporting local maxima that could be avoided if the algorithm run for more iterations. To do so, a simulated example is used. Two sets of initial values for starting the EM algorithm are used. The one set spends a large number of iterations showing convergence (detected by all the methods) to a local maximum but in fact if the algorithm run for more iterations it converges to the global maximum. The behavior of all the stopping criteria is examined.

2. The EM Algorithm

Supposing that the number of components *k* is known the EM algorithm is a standard technique for obtaining maximum likelihood estimates for finite mixtures. The algorithm is very popular since it is easily programmable and satisfies a monotonic convergence property. However it has the annoying disadvantages that if there are multiple maxima, the algorithm may locate a local maximum which is not a global one, and it depends on the choice of good initial values. More details about the algorithm can be found in McLachlan and Peel (2001).

The EM can briefly described as

Step 1 (E-step) Given the current values for θ_j^{old} , $j = 1, . . ., k$ and p_j^{old} , $j=1, . . ., k$ calculate the probability w_{ij} that the observation X_i belongs to the j subpopulation after observing it, i.e. the posterior probability of X_i belonging in the j subpopulation.

$$w_{ij} = \frac{p_j^{old} f(x_i | \theta_j^{old})}{f_P(x_i)}$$

Step 2 (M-step) Update the estimates as

$$\theta_j^{new} = \frac{\sum_{i=1}^{n} w_{ij} x_i}{\sum_{i=1}^{n} w_{ij}} \qquad \text{and} \qquad p_j^{new} = \frac{\sum_{i=1}^{n} w_{ij}}{n} \quad \text{for } j = 1, 2, , . . . , k \quad .$$

Step 3 Check if some condition is satisfied in order to terminate the iterations, otherwise go back to step 1, for more iterations.

Due to the dependence of the solution on the initial values it is argued that several initial values must be used in order to be sure that the global maximum has been obtained.

3. Stopping Criteria

The criteria used for terminating the EM algorithm can be put in 4 different categories. These are

- Those based on the likelihood change between successive iterations (e.g. Agha and Ibrahim, 1984). Since the absolute value of change is not useful as the values of the loglikelihood depends on the sample size and thus an absolute change of the order 10^{-m} has different significance according to the sample size. To avoid this one may use the relative change in the sense that we stop iterating if

$$\left| \frac{\ell(i+1) - \ell(i)}{\ell(i)} \right| < tol$$

where $\ell(i)$ is the value of the loglikelihood after the i-th iteration and *tol* is a small number usually of the form 10^{-m}. However criteria based on the change of the loglikelihood indicate lack of progress and not actual convergence.

- Those based on the relative change of the parameters between successive iterations. The maximum over all the parameters is used as a criterion, so the criterion takes the form

$$\max_j \left| \frac{\varphi_j(i+1) - \varphi_j(i)}{\varphi_j(i)} \right| < tol$$

This criterion does not involve the loglikelihood and hence one does not need to calculate it. On the other hand the criterion indicates lack of progress and not actual convergence and, in addition, the evaluation of the loglikelihood is cheap as one needs to calculate the density for each data point for the E-step.

- Those based on the gradient function. The gradient function for mixtures is defined as

$$D(\theta, P) = \sum_{i=1}^{n} \left\{ \frac{f(x_i|\theta)}{f_P(x_i)} - 1 \right\}$$

It holds (see Lindsay, 1983,1995) that the gradient function evaluated at the points of the ML estimate ought to be 0 and, hence, one can check if the maximum is obtained by looking at the values of the gradient function. In fact, since due to numerical perturbations the gradient function may have a value close to 0 and hence again the convergence criterion is of the form

$$\max_j \left| D(\theta_j^{(i)}, P^{(i)}) \right| < tol$$

where $\theta_j^{(i)}, P^{(i)}$ denotes the value of the parameter θ_j and the corresponding estimated mixing distribution after the i-th iteration. And tol is a small number usually of the form 10^{-m}. It can be seen that this corresponds to the system of estimating equations, and hence zero values of the gradient function indicate that the maximum has been obtained.

- Aitken acceleration: Bohning *et al.* (1994) proposed the use of Aitken acceleration as a convergence criterion. Since the convergence of the EM algorithm for finite mixtures is linear, near the convergence one can obtain a projected loglikelihood after the i-th iteration using

$$\ell_i^\infty = \ell_{i-1} + \frac{1}{1-c_i}\left(\ell_i - \ell_{i-1}\right),$$

where $c_i = \frac{\ell_{i+1} - \ell_i}{\ell_i - \ell_{i-1}}$. Values of c_i near 1 do not necessarily indicate convergence. Now the stopping criterion is based on the change of the projected loglikelihood and thus one stops iterating whenever $\left|\ell_i^\infty - \ell_{i-1}^\infty\right| < tol$

We aim to illustrate the bad behavior of all the above criteria using a counterexample.

4. The Data

Consider the data of Table 1. They are 100 values simulated from a 2 finite exponential mixture with equal mixing proportions and mixing parameter values $\theta_1 = 1$, $\theta_2 = 2$. Then, I used 2 different and rather distinct initial values for running the EM algorithm described above. The first set of initial values were $p_1 = 0.5$, $\theta_1 = 1$, $\theta_2 = 2$ (note that the data were generated with these parameters), while the second set was $p_1 = 0.1$, $\theta_1 = 0.2$, $\theta_2 = 5$. Using a common criterion, that of stop iterating when the relative change of the likelihood was smaller than 10^{-8} (this is a rather strict criterion) we found the solutions of Table 2.

Table 1

Data generated from a 2-finite exponential mixture with equal mixing proportions and mixing parameters 1 and 2 respectively. 100 values were generated

3.46158645890	7.794575130	0.40217924666	2.73704113600
3.19323128910	1.9413970310	0.48619371253	0.17782662119
0.39801132881	7.66311288650	0.29072421156	1.15554743790
1.82142230150	9.2818026130	2.4907488330	0.73557288203
0.8540823018	0.85954333936	0.4637596901	2.28738218960
0.74487411707	0.6903226049	0.13103775463	1.21578963800
0.17333671989	3.1594664830	1.29759396960	0.60080189693
1.2265799140	1.04562349050	0.07999417118	0.64705698113
1.9998531420	0.8799456011	1.85106676850	0.61238381960
2.22993744240	2.82297412210	1.72902519150	0.25447366541
0.338846894	2.43685331880	1.39368475910	4.37147139300
2.36415320510	0.2583544366	1.85825354180	0.90719194482
1.11018968690	0.12592177756	0.60304830671	2.92097750990
1.3997716460	1.11261329740	5.04268694340	0.79078322746
2.55287565470	0.75715823536	2.8170539260	0.08148563531
1.90256856950	4.8786183510	0.22876584648	0.76716226585
2.49449698180	0.02224887114	1.92675762580	2.56554838950
0.04764321023	0.5765043225	1.65640875110	0.6171613682
0.11885560619	1.94039572170	1.41075256690	0.40980316829
0.16157692188	1.25018569870	0.1998214167	1.46961980890
3.4254136510	0.81548120388	3.91238842560	2.01613585640
1.41221165590	2.11582544460	2.70058382520	0.01317356698
0.11386621808	0.71784627210	1.64032269940	2.91008999030
1.57241686590	0.47285531452	2.96138658960	0.14307613664
1.06409787490	1.23025851780	1.1716749970	0.72148614247

Table 2.

Obtained solution using two different starting values. The terminating condition was that the relative change in the loglikelihood was smaller than 10^{-8}

Initial Values	Iterations	p_1	λ_1	λ_2	loglikelihood
$p_1 = 0.5$, $\theta_1 = 1$, $\theta_2 = 2$	81	0.4878	1.45638	1.75448	-147.5616475615
$p_1 = 0.1$, $\lambda_1 = 0.2$, $\lambda_2 = 5$	96	0.0397	1.39515	1.61791	-147.5652474851

Table 3.

Obtained solution using several criteria and two different starting values

		Initial Values $p_1 = 0.5$, $\theta_1 = 1$, $\theta_2 = 2$				
Criterion		Iterations	p_1	λ_1	λ_2	loglikelihood
Max relative change of the parameters	10^{-4}	84	0.4879	1.45683	1.7541	-147.5616436868
	10^{-5}	10872	0.90806	1.53374	2.35304	-147.5515229502
	10^{-6}	13274	0.91311	1.53613	2.37559	-147.5515039065
	10^{-7}	15591	0.91358	1.53636	2.37774	-147.5510373095
	10^{-8}	17900	0.91363	1.53638	2.37796	-147.5515037292
Relative Loglikelihood change	10^{-6}	29	0.48576	1.42495	1.78299	-147.5628156291
	10^{-7}	50	0.48668	1.44651	1.76318	-147.5618021407
	10^{-8}	81	0.4878	1.45638	1.75448	-147.5616475615
	10^{-9}	10064	0.90159	1.53084	2.32572	-147.5515877005
	10^{-10}	11355	0.91015	1.53471	2.36225	-147.5515113927
Maximum Absolute Gradient Function	10^{-5}	12752	0.91276	1.53596	2.37397	-147.5515042278
	10^{-6}	15078	0.91355	1.53634	2.37759	-147.5515037340
	10^{-7}	17388	0.91363	1.53638	2.37794	-147.5515037292
	10^{-8}	19697	0.91364	1.53638	2.37798	-147.5515037291
	10^{-9}	22006	0.91364	1.53638	2.37798	-147.5515037291
	10^{-10}	24314	0.91364	1.53638	2.37798	-147.5515037291

		Initial Values $p_1 = 0.1$, $\lambda_1 = 0.2$, $\lambda_2 = 5$				
Criterion		Iterations	p_1	λ_1	λ_2	Loglikelihood
Max relative change of the parameters	10^{-4}	166	0.03983	1.42019	1.6169	-147.5652224155
	10^{-5}	58426	0.90805	1.53373	2.35301	-147.5515229562
	10^{-6}	60828	0.91311	1.53613	2.37559	-147.5515039065
	10^{-7}	63145	0.91358	1.53636	2.37774	-147.5515037309
	10^{-8}	65455	0.91363	1.53638	2.37796	-147.5515037292
Relative Loglikelihood change	10^{-6}	42	0.03972	1.26432	1.62332	-147.5663939029
	10^{-7}	63	0.03966	1.34903	1.61981	-147.5654118523
	10^{-8}	96	0.0397	1.39515	1.61791	-147.5652474851
	10^{-9}	140	0.03979	1.41542	1.61709	-147.5652246865
	10^{-10}	216	0.03993	1.42449	0.61674	-147.5652211117
Maximum Absolute Gradient Function	10^{-5}	60	0.03966	1.33865	1.62023	-147.5654585896
	10^{-6}	62632	0.91355	1.53634	2.37759	-147.5515037341
	10^{-7}	64942	0.91363	1.53638	2.37794	-147.5515037292
	10^{-8}	67251	0.91364	1.53638	2.37798	-147.5515037292
	10^{-9}	69560	0.91364	1.53638	2.37798	-147.5515037292
	10^{-10}	71868	0.91364	1.53638	2.37798	-147.5515037292
Global Maximum			0.91364	1.53638	2.37798	-147.551503729192

From the above table one can argue that there are multiple maxima. The first set of initial values gave a better likelihood, but the solutions are very different.

5. The Truth

But what about if we leave the algorithm to run for a larger number of iterations? The criterion used was rather strict and in many applications found in the literature the criteria used were less severe.

I run the EM algorithm for the two sets of initial values using several criteria based on all the three different approaches. The full history can be found in Table 3. I have also plotted the entire history with respect the estimates, the loglikelihood, the relative improvement of the loglikelihood, the maximum relative difference in the parameters and the gradient function for both the mixing parameters in Figures 1-4. The interesting findings are

- After a quite large number of iterations both the initial values converged at the same solution, and hence no multiple maxima exist. The algorithm run with some other initial values and the same maximum was obtained. Note that since the true global maximum was far from the values found in Table 2, one could use several other initial values and when stop iterating he could report multiple maxima.
- Looking at the graphs we have plotted for both initial values the whole history over all the iterations. As far as the first set of initial values the convergence occurred after more than 20000 iterations. For the first 10000 iterations the algorithm approached quite slowly the true solution but since the improvement from iteration to iteration was quite small perhaps a less severe criterion could detect small improvement and stop the algorithm.
- The second set of initial values is more illustrative. The initial values were far away from the solution. Again one can see at all the graphs that the algorithm proceeds quite slowly at the first 4000 iterations and then suddenly it jumps towards the solution. We could say that there were a valley where the algorithm climbed very slowly and the improvement was very large and finally the algorithm converged in another valley where the true maximum is located

- Criteria based on the relative improvement of the loglikelihood or the relative change of the parameters must be used with caution. The reason is that even if they detect non-improvement they cannot say anything if the maximum is obtained. And since in mixture models it is quite often the likelihood to be flat, lack of improvement does not mean convergence. Our example is quite illustrative. If one monitor the gradient function (we have plotted the 10^3 times the gradient) then deviations from zero, attributed to numerical perturbation can be significant. An interesting feature of Table 3 is that the relative change of the loglikelihood was less than 10^{-10} for the second set of initial values but the algorithm speed up much later.
- The use of the gradient function cannot solve the problem, since we must use some numerical constant to indicate that a zero value has been reached. But, unfortunately the gradient function is flat and thus one needs to increase the accuracy in order to be sure that the gradient function is actually zero and the difference to the zero values is not just numerical perturbation.
- The behavior of Aitken acceleration can be seen in Figures 5-6. In figure 5 we have plotted the actual loglikelihood and the projected loglikelihood. As Bohning *et al.* (1994) explain the projected loglikelihood is not necessarily larger than the actual loglikelihood. This is due to the fact that c_i can take values larger than 1 (the history of values for c_i can be seen in Figure 6). So, for the second set of initial values the Aitken acceleration criterion would have detected convergence after a 20000 iterations reporting a local maximum. In fact one can see that the method estimates the slope of the loglikelihood at every point and projects to infinity. Thus, in some sense, tries to identify if the loglikelihood stopped increasing and hence the criterion can be trapped in areas where the loglikelihood is flat, as in our case.

6. The message

It is clear that the above example is simply a caution that the EM must be used with care. Further investigation is needed. The lessons we learned are

- It takes quite a large number of iteration until the EM truly converges
- Initial values far away from the solution can locate the maximum if we let them iterating. So, be careful for the stopping criterion

- Prefer using a true convergence criterion rather than a lack of progress criterion.
- Since it takes a lot of iteration for the EM to truly converge we need method to speed it up, like Newton Raphson acceleration or others

From a practical point of view it is not a good idea to leave the algorithm run for a large number of iterations so as to eliminate the possibility that we stopped the algorithm early due to a bad stopping rule. A better strategy would be to use several initial points and a relatively simpler convergence criterion, since it is highly possible that this simpler criterion would allow locating the maximum after much less iteration. So, instead of spending the computer recourses to run from one initial set with a large number of iteration, it seems preferable to run several set of initial values for fewer iterations and finally, for the best solution to run the algorithm with a much severe criterion in order to find the maximum with greater accuracy.

References

Agha, M. and Ibrahim, M.T. (1984). Maximum likelihood estimation of mixtures of distributions. *Applied Statistics,* 33, 327-332

Bohning, D., Dietz, E., Schaub, R., Schlattman, P. and Lindsay, B. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Mathematical Statistics,* 46, 373-388 .

Bohning, D. (1999). *Computer assisted analysis of mixtures & applications in meta-analysis, disease mapping & others.* C R C Press .

Lindsay B. (1983a) The Geometry of Mixture Likelihood. A General Theory. *Annals of Statistics* 11, 86-94

Lindsay B. (1995) *Mixture Models: Theory, Geometry and Applications*. Regional Conference Series in Probability and Statistics, Vol 5, Institute of Mathematical Statistics and American Statistical Association

McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models,*: Wiley, New York.

McLachlan, G. and Krishnan, T. (1997). *The EM Algorithm and Extensions.* Willey Series.

Seidel, W., Mosler, K., and Alker, M. (2000a). A Cautionary Note on Likelihood Ratio Tests in Mixture Models. *Annals of the Institute of Statistical Mathematics* 52, 481-487.

Seidel, W., Mosler, K., and Alker, M. (2000b). Likelihood ratio tests based on subglobal optimization: A power comparison in exponential mixture models. *Statistische Hefte,* 41, 85-98.

Seidel, W., Sevcikova, H. and Alker, M. (2000c). On the Power of Different Versions of the Likelihood Ratio Test for Homogeneity in an Exponential Mixture Model. *Diskussionsbeiträge zur Statistik und Quantitativen Ökonomik 92-2000, Universität der Bundeswehr Hamburg.*

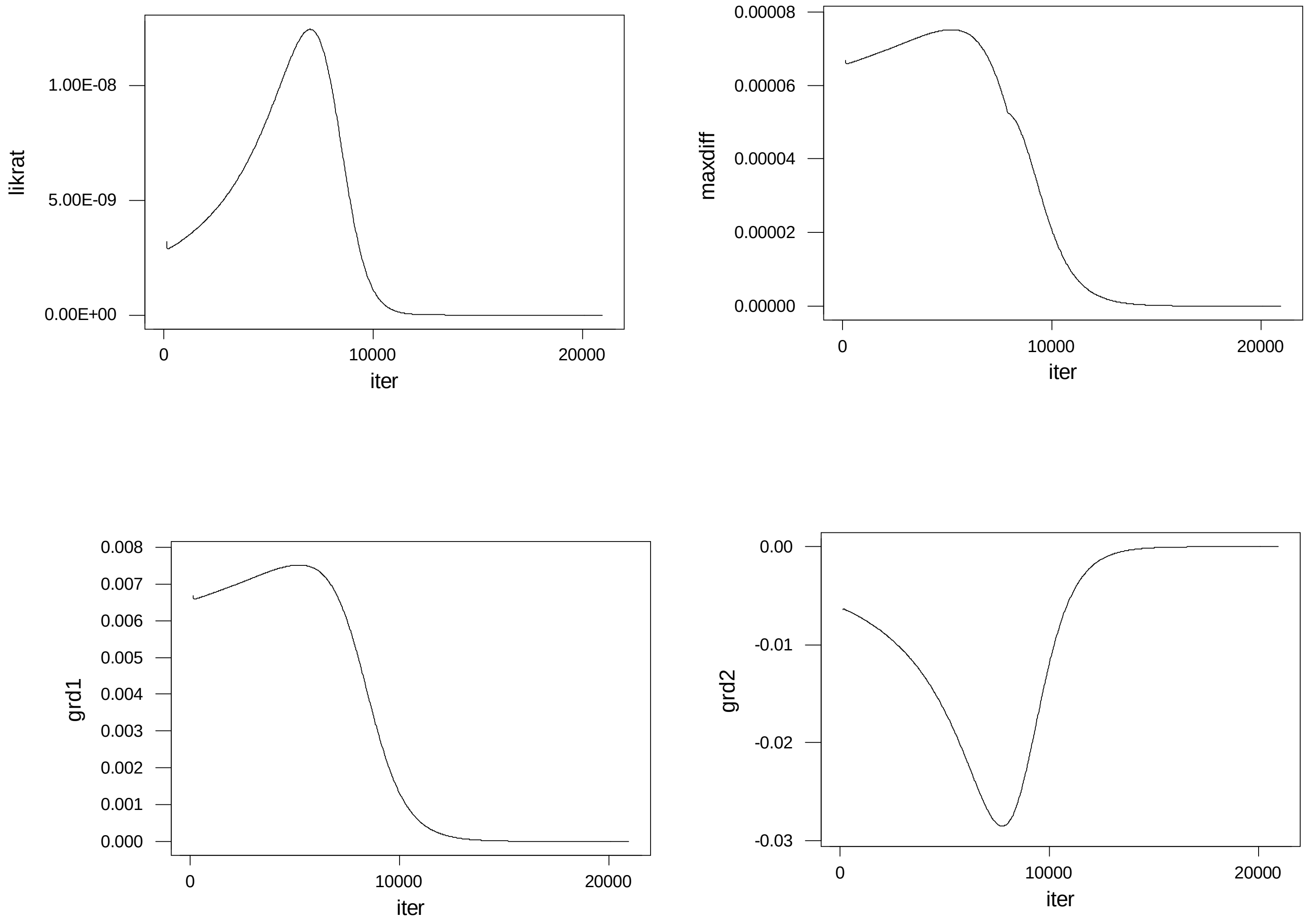


Figure 1. The history of certain convergence criteria for the first set of initial values. (a. relative likelihood difference, b. maximum absolute difference of the parameters, c and d the gradient function).

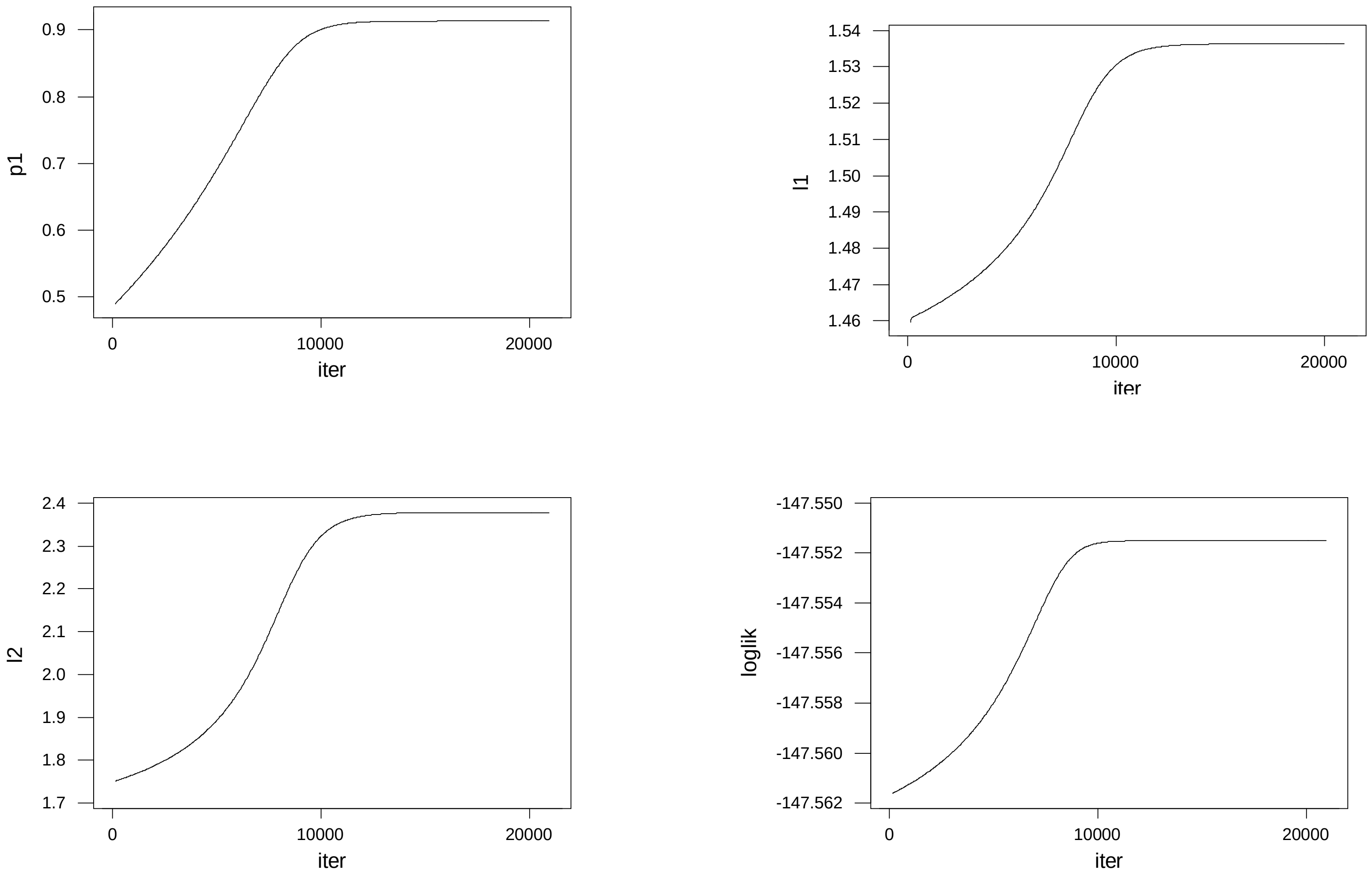


Figure 2. The history of the parameters and the loglikelihood for the first data set

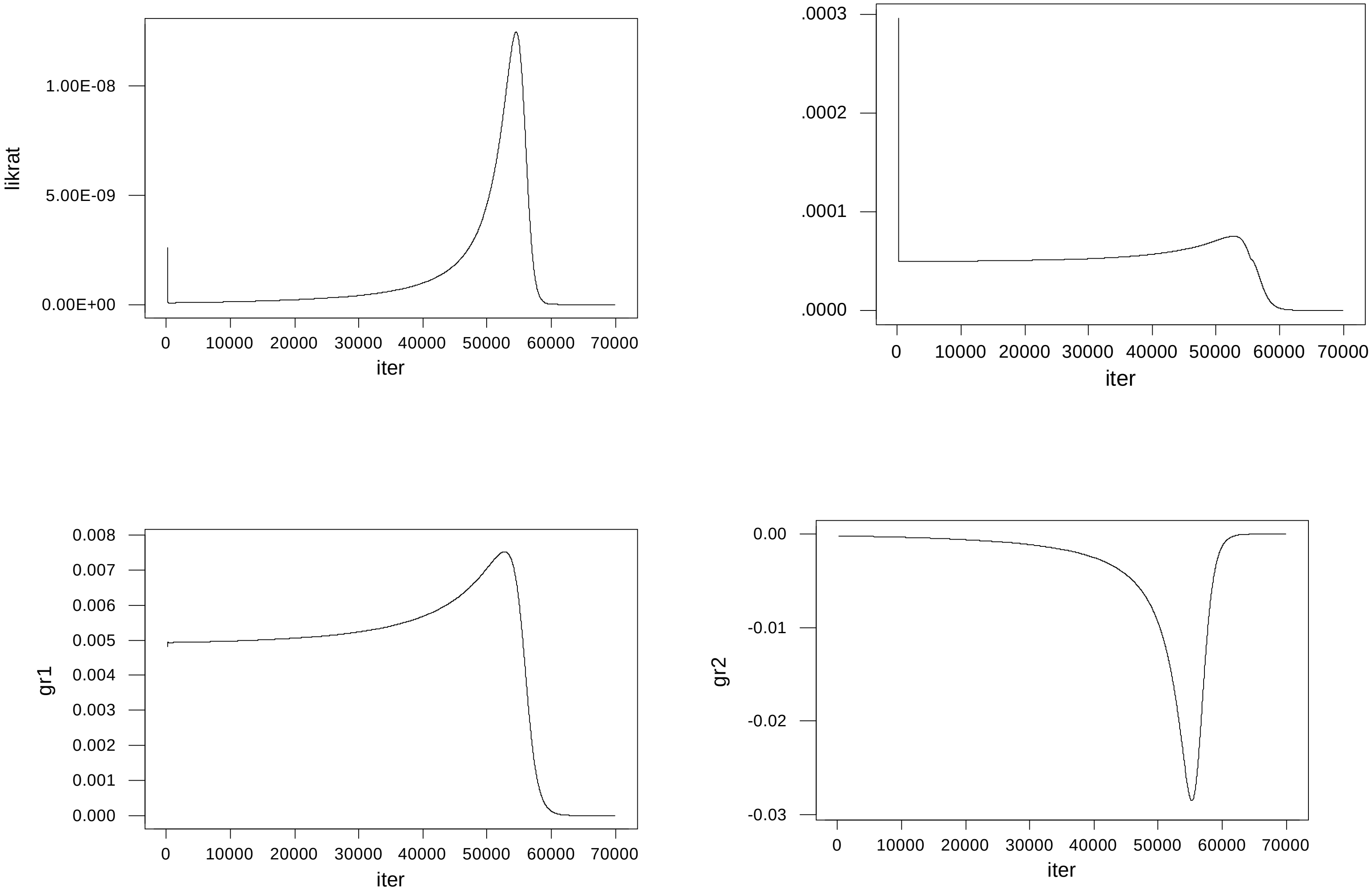


Figure 3. The history of certain convergence criteria for the second set of initial values. (a. relative likelihood difference, b. maximum absolute difference of the parameters, c and d the gradient function).

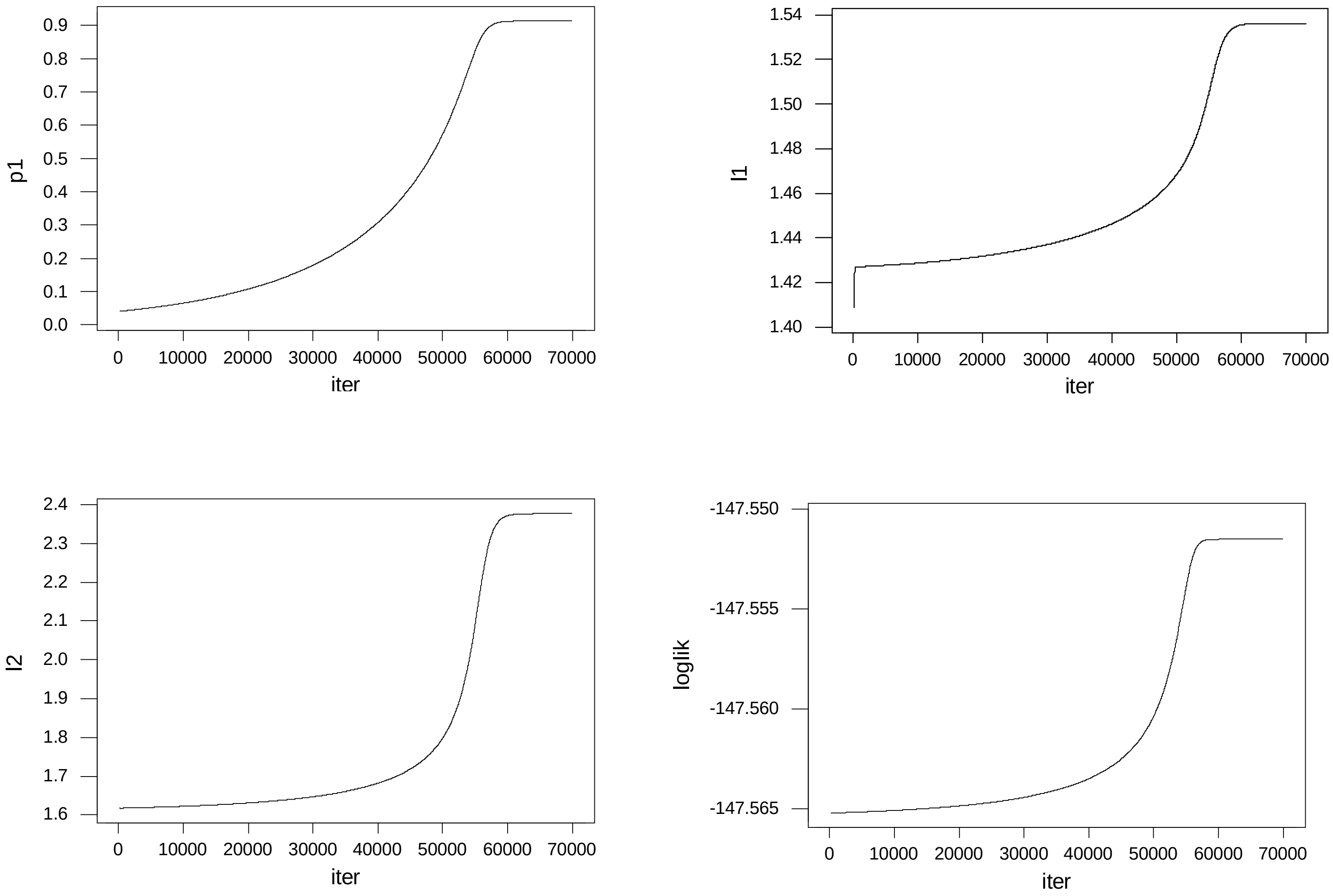


Figure 4. The history of the parameters and the loglikelihood for the second data set

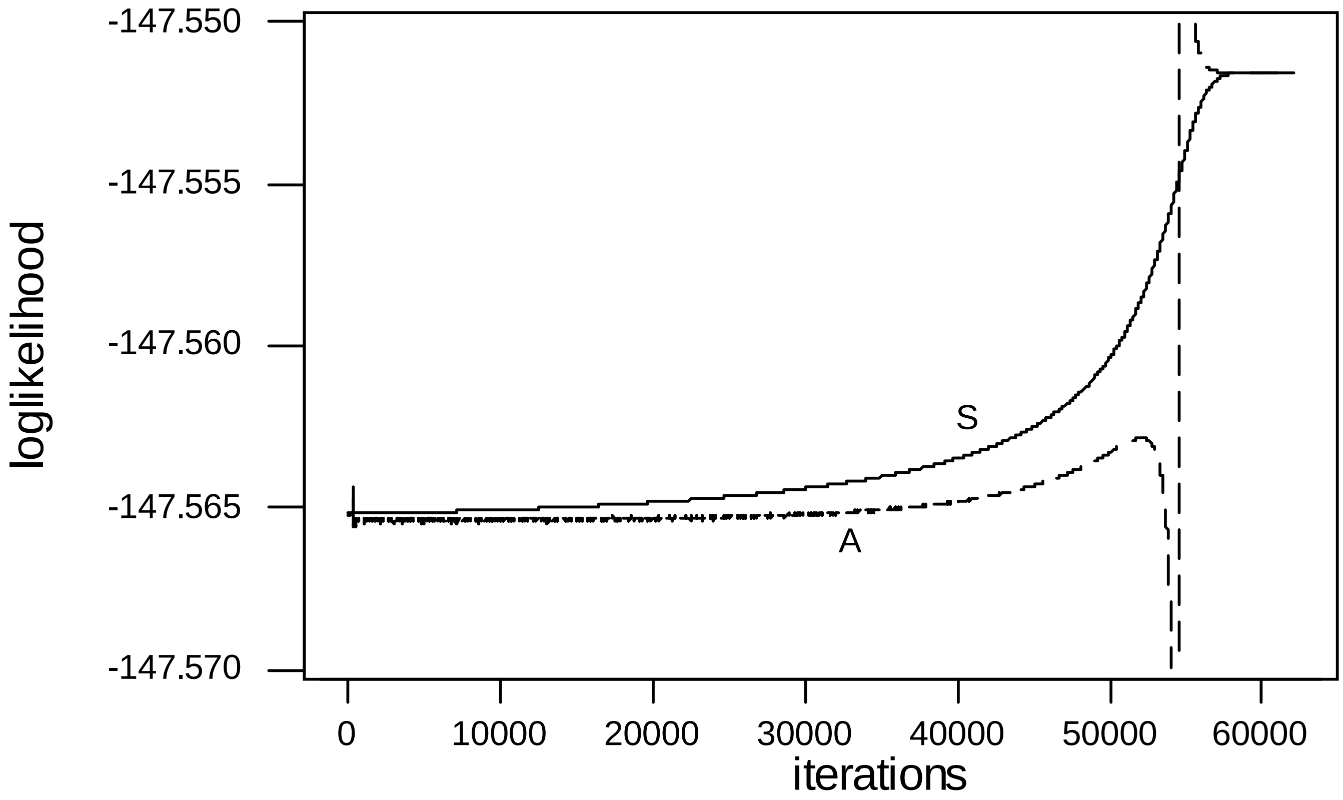


Figure 5. The history of the loglikelihood (S) and the projected loglikelihood using the Aitken accelaration method (A). Aitken accelaration would have detected convergence at the early iterations where it is much more stable than the actual loglikelihood.

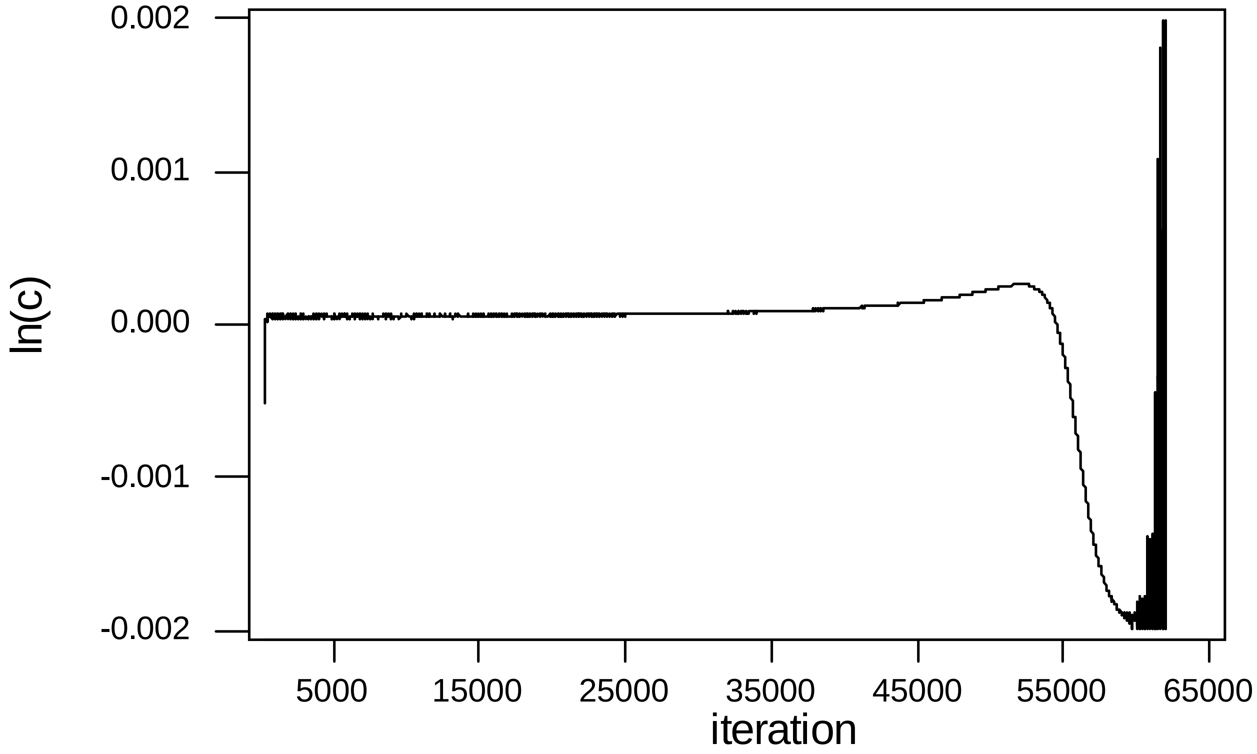


Figure 6. The history of the values of c. The instability after the 55000th iteration can be attributed to numerical perturbations since the progress of the loglikelihood was quite small.