Metodološki zvezki, Vol. 5, No. 1, 2008, 9-18

On Bagging and Estimation in Multivariate Mixtures

Reza Pakyari[1]

Abstract

Two bagging approaches, say ^n-out-of-n without replacement (subagging) and n-out-of-n with replacement (bagging) have been applied in the problem of estimation of the parameters in a multivariate mixture model. It has been observed by Monte Carlo simulations and a real data example, that both bagging methods have improved the standard deviation of the maximum likelihood estimator of the mixing proportion, whilst the absolute bias increased slightly. In estimating the component distributions, bagging could increase the root mean integrated squared error when estimating the most probable component.

1 Introduction

In many statistical applications, it is known or suspected that observations arise from two or more populations with different distributions mixed in varying proportions. For example, the distribution of height in a population of adults might be decomposed into male and female populations, or similarly in fisheries research, where usually fish lengths are available, but not their sexual identities. In medical diagnostics, a patient could be suffering from each of $p — 1$ illnesses. These $p — 1$ illnesses as well as the case of no illness represent p sub-populations. The doctor then applies k different tests, so he or she has k-dimensional data, potentially from a mixture with p sub-populations. Extensive literature is available on the parametric approach. See, for example, the monographs of (Everitt and Hand, 1981; Titterington et al., 1985; McLachlan and Basford, 1988; Lindsay, 1995; McLachlan and Peel, 2000) and the references therein.

Bagging is one of the successful methods for improving the accuracy of estimators especially for high dimensional dataset problems. Bagging (**b**ootstrap **agg**regat**ing**), was first introduced by (Breiman, 1996a) as a method for reducing the variance of an estimator and since then has been applied by several authors and researchers.

Bagging is particularly useful for decision trees and neural networks as well as some nonlinear predictors, although theoretical explorations of why this is so are less clear.

(Breiman, 1996a) showed in empirical studies that the variance of the bagged estimator is never greater than the variance of the original estimator, and that there is considerable variance reduction if the original estimator is *instable.* He also showed that there is little difference between the biases of the bagged and original estimators.

[1] Department of Mathematics, Arak University, Arak 38156, Iran; r-pakyari@araku.ac.ir

By Breiman's definition of instability (Breiman, 1996b), an estimator is instable if a small change in the data tends to a large variation in the predicted value(s). He mentioned that neural nets, classification and regression trees, and subset selection in linear regression, are all examples of instable procedures, whilst, the k-nearest neighbor method is stable.

Note that bagging is useless when it is applied for a linear predictor (Bühlmann and Yu, 2002), therefore bagging may be useful only for nonlinear predictors.

(Bühlmann and Yu, 2002) proposed a variant of bagging called *subagging* **(sub**sample **agg**regat**ing)** based on subsampling instead of the bootstrap for the aggregation. They chose subsample size m = | which was also considered by (Freidman and Hall, 2000).

As mentioned earlier, the theoretical reasons why bagging works well are less clear. (Freidman and Hall, 2000) showed theoretically that under bagging for a class of smooth estimators, the first order or leading variance term remains unchanged, whilst, the second order variance term is improved. They also argued that two different bagging approaches, say |n-out-of-n without replacement (subagging) and //-out-of-// with replacement (bagging) tend to give virtually identical results.

(Bühlmann, 2003) proposed another variant of bagging, called *bragging* **(b**ootstrap **r**obust **agg**regat**ing),** which improves an estimator by taking averages over an unstable selection of variables or terms.

In the present paper, we apply bagging and subagging for the estimation in a two-term mixture model, and compare the estimation with the usual maximum likelihood method.

2 Estimation in multivariate mixtures

Consider the k-variate p-term mixture model

$$\$ = n\ i\ \overset{k}{\underset{i=1}{J}}\ !\ Fii + \ldots + n\ p\ \overset{k}{\underset{i=1}{J}}\ !\ Fpi, \tag{2.1}$$

where Fj$_{\mathrm{i}}$ for 1 $<$ i $<$ k and 1 $<$ j $<$ p are continuous univariate distribution functions and the mixing proportions, n's satisfy $\mathrm{YT}^{\mathrm{j}}=_{1}$ n = 1. Suppose further that the component distribution functions are independent.

Let fit a Gaussian mixture model and estimate the parameters by the method of maximum likelihood estimation. Hence there are 4k + 1 unknown parameters to estimate. It is known that the likelihood equations that arise from a finite Gaussian mixture model does not have a closed form, and therefore a numerical method should be used to find MLEs. The EM algorithm first introduced by (Dempster et al., 1977) in their fundamental paper, is known to be one of the most efficient methods to find MLEs in finite mixture models; see for example the monograph of (McLachlan and Krishnan, 1997).

Consider a dataset x, and let x* be a resample of size m $<$ n obtained by sampling with replacement from x. Bagging is defined as follows:

Let *9(x)* be the predictor of *6{x)* where x = $\{x_1,\ldots,$ x$_{\mathrm{n}}\}$ is a dataset. Let ^(x*) be the bootstrapped predictor of *9(x)* computed by the plug-in principle based on bootstrap resample x*. The bagged predictor is defined by

$$9_{\mathrm{baga}}(x) = E^*\{(9(x^*)> , \tag{2.2}$$

where E*{•} is the bootstrap expectation.

In practical applications, the bagged predictor in (2.2) is approximated by the following estimator:

$$1 \quad ^{B}$$

$$j{=}i$$

where B is the number of bootstrap replicates and 9j (x*) is the version of 9 (x*) computed from the jth bootstrap resample.

In our simulation studies in Section 3 we took B = 50. In general, choosing B depends on the problem under consideration, sample size and computational cost. For further discussion regarding the number of bootstrap replicates, see (Breiman, 1996a).

If the resampling is done without replacement, the resulting bagged predictor is known as subagging. In the following section we study the effect of bagging in estimation the parameters in multivariate mixtures through a Monte Carlo simulation and a real data example by comparing the three estimation method say *^n-out-of-n* without replacement (subagging), n-out-of-n with replacement (bagging) and the usual maximum likelihood estimation.

3 Numerical results

3.1 Simulation study

We generated 300 datasets, each of size n = 500, from a trivariate two-term Gaussian mixture model

$$\$i(x) = nNa(x;\ ßi,\ I3) + (1 - n)N_3(x;\ p_2, I3), \tag{3.1}$$

where $N_3(ß,$ S) denotes the trivariate Gaussian distribution function with mean vector *ß,* and variance-covariance matrix S. In particular, we chose the 3 x 3 identity matrix I_3 as the variance-covariance matrix, and $ß_1$ = (0,0,0), and two different values for $ß_2$, specifically $ß_2$ = (1,1,1) and $ß_2$ = (3, 3, 3), representing relatively "close" or "distant" component distributions, respectively. It is known that estimation of n and *F*ji in the "distant" setting tends to better results in compare to "close" setting.

The maximum likelihood estimation method via EM algorithm has been applied for estimation of the parameters. The effect of bagging procedure for the improvement of the estimation has been studied by applying bagging and subagging.

Table 1 gives the absolute bias of the maximum likelihood, subagged maximum likelihood and bagged maximum likelihood estimator of the mixing proportion n, in the distant setting and Table 2 gives their standard deviations. Figure 1 depicts the figures given in Tables 1 and 2 as well as the root mean squared error of the three kind of estimators. It can be seen from this figure that subagging and bagging had reduced the standard deviation whilst the absolute bias have been slightly increased. The improvement of the root mean squared error is less clear. The same thing happens when estimation is done in the close setting as is suggested by Tables 3 and 4 and Figure 2 which give absolute bias, standard deviation and their figures in the close setting, respectively.

Table 1 : Absolute bias of maximum likelihood estimator MLE, subagged MLE and bagged MLE of mixing proportion n, in the distant setting.

mixing proportion 7r	MLE	Subagging MLE	Bagging MLE
0.1	0.00092	0.00046	0.00089
0.2	0.01360	0.01424	0.01413
0.3	0.02547	0.02521	0.02566
0.4	0.03477	0.03612	0.03616

Table 2: Standard deviation of maximum likelihood estimator MLE, subagged MLE and bagged MLE of mixing proportion n, in the distant setting.

mixing proportion 7r	MLE	Subagging MLE	Bagging MLE
0.1	0.01194	0.01197	0.01179
0.2	0.01721	0.01697	0.01651
0.3	0.02075	0.01837	0.01831
0.4	0.02538	0.02125	0.02207

Table 3: Absolute bias of maximum likelihood estimator MLE, subagged MLE and bagged MLE of mixing proportion n, in the close setting.

mixing proportion ir	MLE	Subagging MLE	Bagging MLE
0.1	0.24103	0.27256	0.27170
0.2	0.08876	0.09238	0.09160
0.3	0.01332	0.01557	0.01578
0.4	0.00667	0.00637	0.00547

Table 5 gives the root mean integrated squared error (MISE) of the maximum likelihood, subagged maximum likelihood and bagged maximum likelihood estimator of the component distribution function Fn, in the distant setting. Figure 4 depicts the root MISE of all the Fji's for the three kinds of estimators. Any difference between the three estimation method is barely clear in this figure. When comes to the close setting the results differ slightly. Figure 3 depicts the root mean integrated squares error of the three estimation method. For brevity we do not give table in this case. The figure surprisingly show that both bagging and subagging have increased the root MISE when estimating the

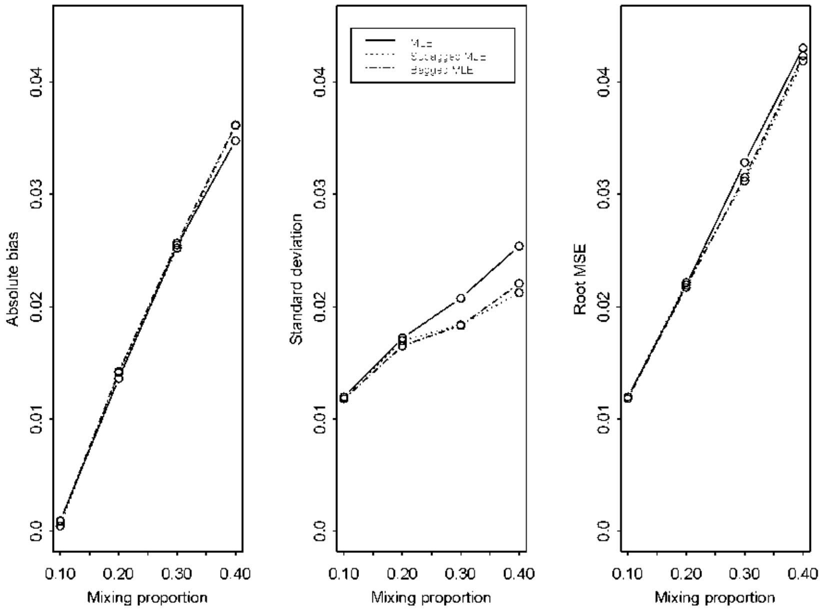

Figure 1: Absolute bias, standard deviation and root MSE of the MLE, solid lined, subagged MLE, dotted lines and bagged MLE, dot-dashed lines, estimators of the mixing proportion n, in the "distant" setting. Sample size is 500.

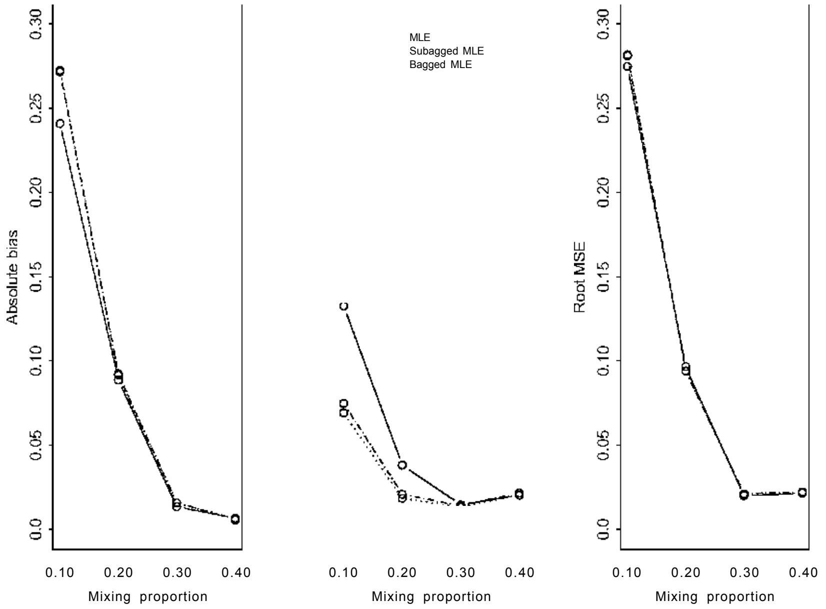

Figure 2: Absolute bias, standard deviation and root MSE of the MLE, solid lined, subagged MLE, dotted lines and bagged MLE, dot-dashed lines, estimators of the mixing proportion n, in the "close" setting. Sample size is 500.

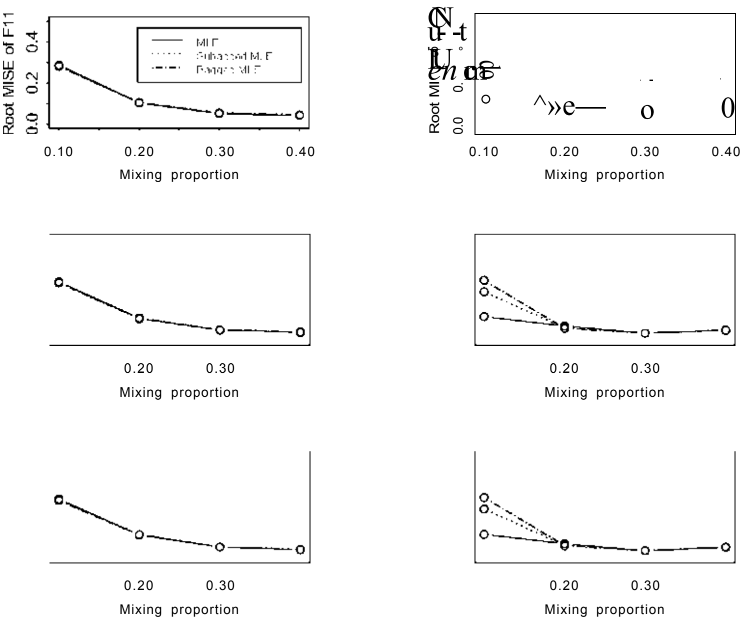

Figure 3: Root MISE of the MLE, solid lined, subagged MLE, dotted lines and bagged MLE, dot-dashed lines, estimators of the component distribution functions, in the "close" setting. Sample size is 500.

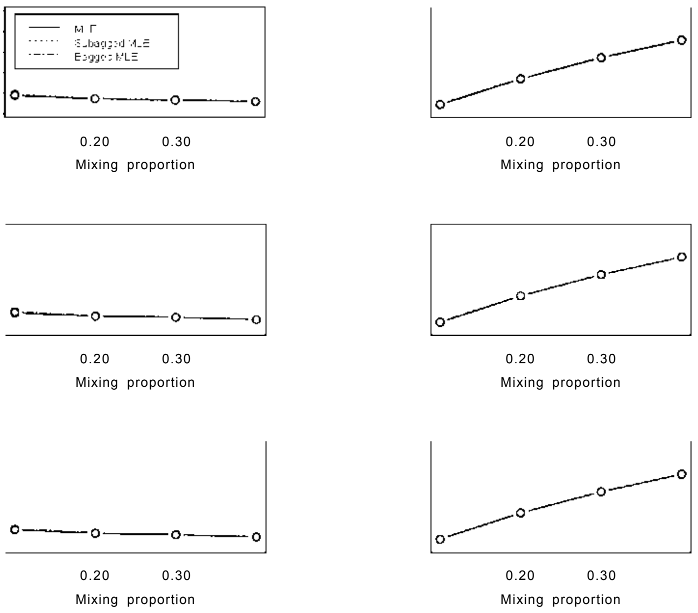

Figure 4: Root MISE of the MLE, solid lined, subagged MLE, dotted lines and bagged MLE, dot-dashed lines, estimators of the component distribution functions, in the "distant" setting. Sample size is 500.

Table 4: Standard deviation of maximum likelihood estimator MLE, subagged MLE and bagged MLE of mixing proportion n, in the close setting.

mixing proportion 7r	MLE	Subagging MLE	Bagging MLE
0.1	0.13252	0.06909	0.07478
0.2	0.03818	0.01847	0.02087
0.3	0.01469	0.01333	0.01402
0.4	0.02041	0.02063	0.02161

Table 5: Root MISE of maximum likelihood estimator MLE, subagged MLE and bagged MLE of component distribution $F u$, in the distant setting.

mixing proportion 7r	MLE	Subagging MLE	Bagging MLE
0.1	0.08697	0.08963	0.09222
0.2	0.07109	0.07101	0.07216
0.3	0.06598	0.06468	0.06575
0.4	0.05751	0.05888	0.05899

distribution function of the most probable component. The difference between the three estimation method is not significant in other cases.

3.2 Real-data example: Leptograpsus crabs

(Campbell and Mahon, 1974) collected and analyzed 200 specimens of *Leptograpsus* crabs in Fremantle, Western Australia. *Leptograpsus* crab has two species, blue and orange. Campbell and Mahon measured five morphological characteristics of 50 males and 50 females of each colour, 200 specimens in total. To simplify our analysis we considered the first three morphological characteristics, namely the width of the frontal lip, the rear width of the carapace and the length along midline of the carapace. The carapace is the hard protective shell of the crab.

(Campbell and Mahon, 1974) used a multivariate analysis to classify the two species independently of colour. This classification of species based on morphological characteristics would help museums, because preserved species lose their colour over time. The dataset is widely considered to follow a multivariate two-component mixture model and was further analyzed by (Ripley, 1996; McLachlan and Peel, 1998; McLachlan and Peel, 2000; Raftery and Dean, 2004).

We repeatedly resampled datasets of size 100, without replacement, from the whole dataset of size $n = 200$ using the following resampling scheme:

Step 1 : Generate a random number u from a uniform random variable
on $(0,1)$.

Step 2: If $u < n$ take a sample without replacement from the male
population;

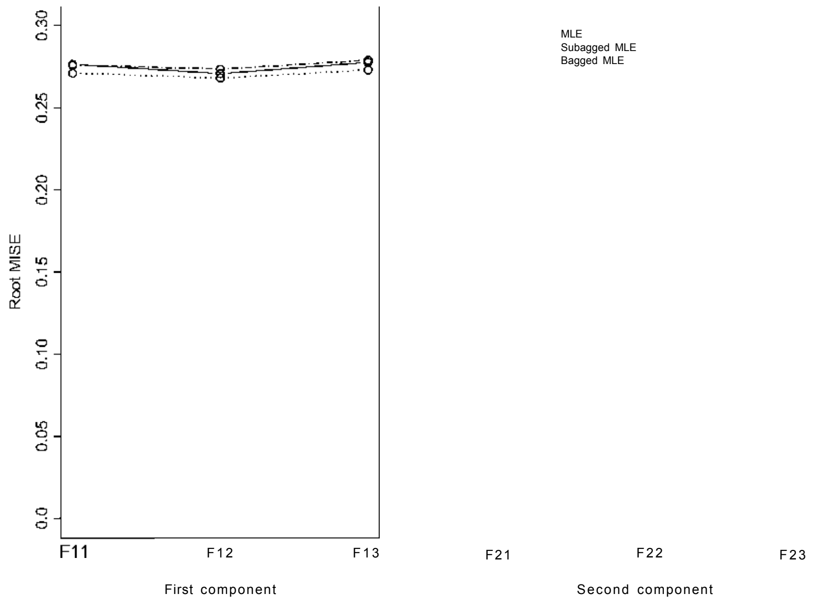

Figure 5: Root mean integrated squared errors of the marginal distribution estimators by fitting three models to the first three morphological characteristics of the crabs dataset when n = 0.3.

otherwise take a sample without replacement from the female population.

Step 3: Repeat steps 1 and 2 until all 100 samples are taken.

We considered n = 0.3 to have a asymmetric mixture model. This makes the female population, the most probable component in the model. We repeated the above steps 50 times to obtain 50 resamples each of size 100. To each dataset, we fitted a trivariate two-term Gaussian mixture model

$2(x) = nNa(x; *ßi,* Si) + (1 - *n)N*$_3$*(x; ß2,* £2), (3.2)

where *N*$_3$*(ß,* S) denotes the trivariate Gaussian distribution function with mean vector *ß,* and variance-covariance matrix S. The components are assumed to be independent.

Table 6: Absolute bias and standard deviation of maximum likelihood estimator MLE, subagged MLE and bagged MLE of mixing proportion n, in the crabs data.

criterion	MLE	Subagging MLE	Bagging MLE
Absolute bias	0.20286	0.21069	0.20805
Standard deviation	0.15590	0.10397	0.11188

Table 6 gives absolute bias and standard deviation of the maximum likelihood estimator of the mixing proportion, n as well as the subgged and bagged estimators. As one expects both bagged and subagged estimators have reduced the standard deviation, whilst the absolute bias has been slightly increased.

Figure 5 depicts the root mean integrated squared errors of the marginal distribution function estimators. Empirical distributions computed from all 200 data were considered as the true marginal distribution functions. Surprisingly both bagging procedures have little or no effect on the root MISE of the first component, whilst they have been worsen the estimation for the second, i.e. the most probable component.

Conclusions

In summary, the bagging and subagging procedures improve the standard deviation when estimating the mixing proportion, whilst the absolute bias increases slightly. Both Bagging procedures could increase the mean integrated square error when estimating the marginal distribution functions, especially in estimating of the most probable component. In all of the estimations there are no significant difference between |n-out-of-n subagging and n-out-of-n bagging. This support the results of (Freidman and Hall, 2000).

Our findings in this article leave a major question about the usefulness of Bagging in mixture estimation problems. Bagging was well know to be useless in linear problems and we have to consider the mixture models as another situation that this resampling method fails to improve the estimation. Nonetheless one should not underestimate the power of Bagging in improvement the estimation in neural networks and decision tress problems.

Acknowledgement

The author would like to thank an anonymous referee for helpful comments. This work is financially supported by the research department of Arak university.

References

[1] Breiman, L. (1996): Bagging predictors. *Machine Learning,* **24,** 123-140.

[2] Breiman, L. (1996b): Heuristics of instability and stabilization in model selection. *Annals of Statistics,* **24,** 2350-2383.

[3] Bühlmann, P. (2003): Bagging, Subagging and Bragging for improving some prediction algorithms. *Unpublished manuscript.*

[4] Bühlmann, P. and Yu, B. (2002): Analyzing Bagging. *Annals of Statistics,* **30,** 927-961.

[5] Campbell, N. A. and Mahon, R. J. (1974): A multivariate study of variation in two species of rock crab of the genus *Leptograpsus. Australian Journal of Zoology,* **22,** 417-425.

[6] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977): Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B,* **39,** 1-38.

[7] Everitt, B.S. andHand, D.J. (1981): *Finite Mixture Distributions.* London: Chapman & Hall.

[8] Friedman, J. H. and Hall, P. (2000): On bagging and nonlinear estimation. *Journal of statistical planning and inference,* **137,** 669-683.

[9] Lindsay, B.G. (1995): *Mixture Models: Theory, Geometry, and Applications.* Hayward: Institute of Mathematical Statistics.

[10] McLachlan, G.J. and Basford, K.E. (1988): *Mixture Models: Inference and Applications to Clustering.* New York: Wiley.

[11] McLachlan, G.J. and Peel, D. (1998): Robust cluster analysis via mixtures of multivariate t-distribution. In *Lecture Notes in Computer Science,* **1451,** 658-666, Berlin: Springer-Verlag.

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

[13] Raftery, A. E. and Dean, N. (2004): Variable selection for model-based clustering. Technical Report 452, department of Statistics, University of Washington.

[14] Ripley, B. D. (1996): *Pattern Recognition and Neural Networks.* Cambridge: Cambridge University Press.

[15] Titterington, D.M., Smith, A.F.M., and Makov, U. (1985): *Statistical Analysis of Finite Mixture Distributions.* New York: Wiley.