Parameter Estimation of Polya's Distribution

Md Amran Siddiqui and Tadesse Zemicheal
{siddiqmd,zemichet}@onid.orst.edu

1 Introduction

A topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. Intuitively, given that a document is about a particular topic, one would expect particular words to appear in the document more or less frequently.The idea behind Topic modelling (LDA) is to model documents as arising from multiple topics, where a topic is defined to be a distribution over a fixed vocabulary words. Specifically, the assumption is k topics are associated with a collection and that each document exhibits these topics with different proportions [3].

Drichlet-multinomial (polya distribution) has been used in topic modelling. In the model topics are distributed with Drichlet distribution of parameter α_k for K topics and each word in the documents are drawn from Multinomial distribution of K topics.

In this project our goal is to estimate parameters of Poly's distribuiotn using different methods like maximum likelihood and method of moments, and compare their performance with the Cramer Rao Lower bound. In Polya distribution we have k parameters $p_1, p_2, ..., p_k$ for multinomial distribution representing probabilities for k categories. These k random parameters represent the topic proportions and coming from Dirichlet distribution with parameters $\alpha_1, \alpha_2, ..., \alpha_k$. In this project we explore a simple case where number of categories are just two. Hence, the multinomial reduces to Binomial and Dirichlet reduces to Beta distribution. Hence, we need to estimate α_1, α_2 from set of documents.

2 Cramer Rao Lower Bound

The probability mass function for Polya distributin is:

$$p(x \mid \alpha) = \frac{n!}{\prod\limits_k n_{k!}} \frac{\Gamma(\sum\limits_k \alpha_k)}{\Gamma(n + \sum\limits_k \alpha_k)} \prod_k \frac{\Gamma(n_k + \alpha_k)}{\Gamma(\alpha_k)} \tag{1}$$

where,

the parameter vector is: $\theta = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ . \\ . \\ \alpha_k \end{bmatrix}$

$n = n(x) =$ length of a data sample x
$n_k = n_k(x) =$ number of k category elements in x

Now, the probability of m observations $x_1, x_2, ..., x_m$:

$$p(x_1, x_2, ..., x_m \mid \alpha) = \prod_{i=1}^{m} \left(\frac{n_i!}{\prod_k n_{ik!}} \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(n_i + \sum_k \alpha_k)} \prod_k \frac{\Gamma(n_{ik} + \alpha_k)}{\Gamma(\alpha_k)} \right) \tag{2}$$

The log likelihood is:

$$\log p(x_1, x_2, ..., x_m \mid \alpha) \tag{3}$$
$$= \sum_{i=1}^{m} log\left(\frac{n_i!}{\prod_k n_{ik}!}\frac{\Gamma(\sum_k \alpha_k)}{\Gamma(n_i + \sum_k \alpha_k)}\prod_k \frac{\Gamma(n_{ik}+\alpha_k)}{\Gamma(\alpha_k)}\right) \tag{4}$$
$$= \sum_{i=1}^{m}\left(log(n_i!) - \sum_k log(n_{ik}!)! + log(\Gamma(\sum_k \alpha_k)) - log(\Gamma(n_i + \sum_k \alpha_k)) + \sum_k log(\Gamma(n_{ik}+\alpha_k)) - \sum_k log(\Gamma(\alpha_k))\right) \tag{5}$$

Differentiating in terms of α_k:

$$\frac{d\ log\ p(D \mid \alpha)}{d\ \alpha_k} = \sum_{i=1}^{m}\left(\psi(\sum_k \alpha_k) - \psi(n_i + \sum_k \alpha_k) + \psi(n_{ik}+\alpha_k) - \psi(\alpha_k)\right) \tag{6}$$
$$\frac{d^2\ log\ p(D \mid \alpha)}{d\ \alpha_k^2} = \sum_{i=1}^{m}\left(\psi'(\sum_k \alpha_k) - \psi'(n_i + \sum_k \alpha_k) + \psi'(n_{ik}+\alpha_k) - \psi'(\alpha_k)\right) \tag{7}$$
$$\frac{d^2\ log\ p(D \mid \alpha)}{d\ \alpha_k\ d\ \alpha_j} = \sum_{i=1}^{m}\left(\psi'(\sum_k \alpha_k) - \psi'(n_i + \sum_k \alpha_k)\right) \tag{8}$$

Now, for beta binomial case (k = 2), we have two parameters α_1 and α_2. Hence the FIM is:

$$FIM = -E\begin{bmatrix} \frac{d^2\ log\ p(D \mid \alpha)}{d\ \alpha_1^2} & \frac{d^2\ log\ p(D \mid \alpha)}{d\ \alpha_1\ d\ \alpha_2} \\ \frac{d^2\ log\ p(D \mid \alpha)}{d\ \alpha_2\ d\ \alpha_1} & \frac{d^2\ log\ p(D \mid \alpha)}{d\ \alpha_2^2} \end{bmatrix} \tag{9}$$

$$\begin{bmatrix} \sum_{i=1}^{m}\left(\psi'(\sum_k \alpha_k) - \psi'(n_i + \sum_k \alpha_k) + \psi'(n_{i1}+\alpha_1) - \psi'(\alpha_1)\right) & \sum_{i=1}^{m}\left(\psi'(\sum_k \alpha_k) - \psi'(n_i + \sum_k \alpha_k)\right) \\ \sum_{i=1}^{m}\left(\psi'(\sum_k \alpha_k) - \psi'(n_i + \sum_k \alpha_k)\right) & \sum_{i=1}^{m}\left(\psi'(\sum_k \alpha_k) - \psi'(n_i + \sum_k \alpha_k) + \psi'(n_{i2}+\alpha_2) - \psi'(\alpha_2)\right) \end{bmatrix} \tag{10}$$

$$FIM_{11} = -E\left[\sum_{i=1}^{m}\left(\psi'(\sum_k \alpha_k) - \psi'(n_i + \sum_k \alpha_k) + \psi'(n_{i1}+\alpha_1) - \psi'(\alpha_1)\right)\right] \tag{11}$$

$$= -\sum_{i=1}^{m}\left(\psi'(\sum_k \alpha_k) - \psi'(n_i + \sum_k \alpha_k) + E\left[\psi'(n_{i1}+\alpha_1)\right] - \psi'(\alpha_1)\right) \tag{12}$$

$$= -m * \left(\psi'(\sum_k \alpha_k) - \psi'(n_i + \sum_k \alpha_k) + E\left[\psi'(n_{i1}+\alpha_1)\right] - \psi'(\alpha_1)\right) \tag{13}$$

$$FIM_{12} = FIM_{21} = -E\left[\sum_{i=1}^{m}\left(\psi'(\sum_k \alpha_k) - \psi'(n_i + \sum_k \alpha_k)\right)\right] \tag{14}$$

$$= -\sum_{i=1}^{m} E\left[\left(\psi'(\sum_k \alpha_k) - \psi'(n_i + \sum_k \alpha_k)\right)\right] \tag{15}$$

$$= -m * \left(\psi'(\sum_k \alpha_k) - \psi'(n_i + \sum_k \alpha_k)\right) \tag{16}$$

$$FIM_{22} = -E\left[\sum_{i=1}^{m}\left(\psi'(\sum_k \alpha_k) - \psi'(n_i + \sum_k \alpha_k) + \psi'(n_{i2}+\alpha_2) - \psi'(\alpha_2)\right)\right] \tag{17}$$

$$= -\sum_{i=1}^{m}\left(\psi'(\sum_k \alpha_k) - \psi'(n_i + \sum_k \alpha_k) + E\left[\psi'(n_{i2}+\alpha_2)\right] - \psi'(\alpha_2)\right) \tag{18}$$

$$= -m * \left(\psi'(\sum_k \alpha_k) - \psi'(n_i + \sum_k \alpha_k) + E\left[\psi'(n_{i2}+\alpha_2)\right] - \psi'(\alpha_2)\right) \tag{19}$$

In the equations 13, 16 and 19 we don't have closed form solution for the expectations, hence we compute these numerically using Monte Carlo simulation.

After inverting FIM we get the $CRLB$ for α_1 and α_2 from FIM^{-1}:

$$CRLB_{\alpha_1} = (FIM^{-1})_{11} \tag{20}$$

$$CRLB_{\alpha_2} = (FIM^{-1})_{22} \tag{21}$$

3 Maximum Likelihood Estimation

To find the MLE estimates of the parameters we start with the log likelihood:

$$log\ p(x_1, x_2, ..., x_m \mid \alpha)$$
$$= \sum_{i=1}^{m}\left(log(n_i!) - \sum_k log(n_{ik}!)! + log(\Gamma(\sum_k \alpha_k)) - log(\Gamma(n_i + \sum_k \alpha_k)) + \sum_k log(\Gamma(n_{ik}+\alpha_k)) - \sum_k log(\Gamma(\alpha_k))\right) \tag{22}$$

There is no closed form solution for the α_k by usual method of taking derivative and setting it to zero. But α_k can be found using iterative methods. One method suggested in [1] is using fixed point iteration. The idea is to guess an initial α_k, find a function that bounds F from below which is tight at α_k , then optimize this function to arrive at α_k^{new} which maximizes the function.
In [1], Minka come up with the final fixed point iteration using the following bounds. First equation 22 can be bounded

by the followings bounds [2]:

$$log\Gamma(z) - log\Gamma(z+n) \geq log\Gamma(\hat{z}) - log\Gamma(\hat{z}+n) + [\Psi(\hat{z}+n) - \Psi(\hat{z})](\hat{z}-z) \tag{23}$$

$$log\Gamma(z+n) - log\Gamma(z) \geq log\Gamma(\hat{z}+n) - log\Gamma(\hat{z}) + \hat{z}[\Psi(\hat{z}+n) - \Psi(\hat{z})](logz - log\hat{z}) \tag{24}$$

Then substituting equation 23 and 24 in equation 22 and simplifying and differentiating with α_k gives.

$$\frac{d\ log\ p(D \mid \alpha)}{d\ \alpha_k} = \sum_{i=1}^{m} \left(\frac{\alpha_k \psi(\sum_k \alpha_k) - \psi(n_i + \sum_k \alpha_k)}{\alpha_k^{new}} + \Psi(n_{ik} + \alpha_k) - \psi(\alpha_k) \right) \tag{25}$$

Finally, equation 25 can be set to zero to solve α_k^{new}

$$\alpha_k^{new} = \alpha_k \frac{\sum\limits_{i=1}^{m} \Psi(n_{ik} + \alpha_k) - \Psi(\alpha_k)}{\sum\limits_{i} \Psi(n_i + \sum\limits_{k} \alpha_k) - \Psi(\sum\limits_{k} \alpha_k)} \tag{26}$$

We can also simplify equation 26 using the following gamma simplifications.

$$\Psi(n+x) - \Psi(x) = \frac{d}{dx}(log \frac{\Gamma(n+x)}{\Gamma(x)}) \tag{27}$$

$$= \frac{d}{dx}(\sum_{i=0}^{n-1}(log(x+i)) \tag{28}$$

$$= \sum_{i=0}^{n-1} \frac{1}{(x+i)} \tag{29}$$

Then using the above simplification equation 26 can be reduced to:

$$\alpha_k^{new} = \alpha_k \frac{\sum\limits_{i=1}^{m} \sum\limits_{j=0}^{(n_{ik}-1)} \frac{1}{\alpha_k + j}}{\sum\limits_{i=1}^{m} \sum\limits_{j=0}^{(n_i-1)} \frac{1}{\sum\limits_{k} \alpha_k + j}} \tag{30}$$

$$\tag{31}$$

4 Method of Moments

In method of moment parameter estimation we usually relate population moments with sample moments and then solve for unknown parameters. We estimate the Beta Binomial parameters in the same way. We know for Beta-Binomial:

$$E[X] = \frac{n\alpha}{\alpha + \beta} \tag{32}$$

$$E[X^2] = \frac{n\alpha(n + n\alpha + \beta)}{(\alpha + \beta)(1 + \alpha + \beta)} \tag{33}$$

Now, the first and second order moments from the data:

$$m_1 = \frac{1}{m} \sum_{i=1}^{m} x_i \tag{34}$$

$$m_2 = \frac{1}{m} \sum_{i=1}^{m} x_i^2 \tag{35}$$

Equating first and second order moments with sample moments:

$$m_1 = \frac{n\alpha}{\alpha+\beta} \tag{36}$$

$$m_2 = \frac{n\alpha(n+n\alpha+\beta)}{(\alpha+\beta)(1+\alpha+\beta)} \tag{37}$$

From 36 we have:

$$\beta = \frac{\alpha(n-m_1)}{m_1} \tag{38}$$

Dividing 37 by 36:

$$\frac{m_2}{m_1} = \frac{n+n\alpha+\beta}{(1+\alpha+\beta)} \tag{39}$$

Replacing β in 39 from 38 we have:

$$\frac{m_2}{m_1} = \frac{nm_1+nm_1\alpha+n\alpha-m_1\alpha}{m_1+m_1\alpha+n\alpha-\alpha m_1} \tag{40}$$

Solving for α:

$$\hat{\alpha}_{mom} = \alpha_1 = \frac{nm_1-m_2}{n(\frac{m_2}{m_1}-m_1-1)+m_1} \tag{41}$$

Replacing the value of α in 38 we get:

$$\hat{\beta}_{mom} = \alpha_2 = \frac{(n-m_1)(n-\frac{m_2}{m_1})}{n(\frac{m_2}{m_1}-m_1-1)+m_1} \tag{42}$$

5 Experimental Result

The graphs compare CRLB and the MSE of the two methods discussed above in dB scale. In our experiment, we took 10 by 10 grid of parameters to generate the data from Beta Binomial distribution. For each generated data we estimate the parameters for the beta binomial for a number of Monte Carlo Simulations. These plots compare the MSE of CRLB,ML and MOM for a given α_1 to the range of α_2. The x-axis shows the range of α_2 corresponding to the MSE on the y-axis.

In figure 1, we plot the result for a dataset with only 20 documents (m = 20) and 1000 Monte Carlo iterations. We observe the CRLB and ML are smooth enough after these 1000 iterations. But MOM is still showing lots of noise. After investigating why this is happening we found a fundamental flaw of MOM estimator. If we observe the denominator of the MOM estimators in equation 41 and 42, we see that there can be situations where $(n(\frac{m_2}{m_1}-m_1-1)+m_1)$ cab be less than or equal to zero. We literally faced this situation when drawing large number of Monte Carlo samples from beta binomial. This is making the MSE to become infinity for some parameters. To solve this issue we dropped the samples those are causing the denominator becoming less or equal to zero since our parameter can not be negative. This solved the problems partially, but we are still estimating very large values for the parameters when the denominator is very small. These are causing the high fluctuation in the MSE plot for MOM in figure 1. Surprisingly, this situation is not occurring when we estimate for large number of documents (m = 100 or 400). The reason is that even though there is some probability to draw samples those are causing problems for all values of m, the probability is very low for large m and high for small m.

In figure 2, we plot the same result but for large number of documents (m = 100). Now, we see the MOM curve got smoother than before. both MOM and ML estimates became closer to the CRLB.

In figure 3, we plot the same for m = 400. We observe that both ML and MOM are almost aligned to the CRLB.

Finally, in figure 4 we put all the previous plots to see how the estimates change as we increase number of samples. It is clear from figure 4 that as we increase samples the CRLB and MSE for ML/MOM gets closer to zero.

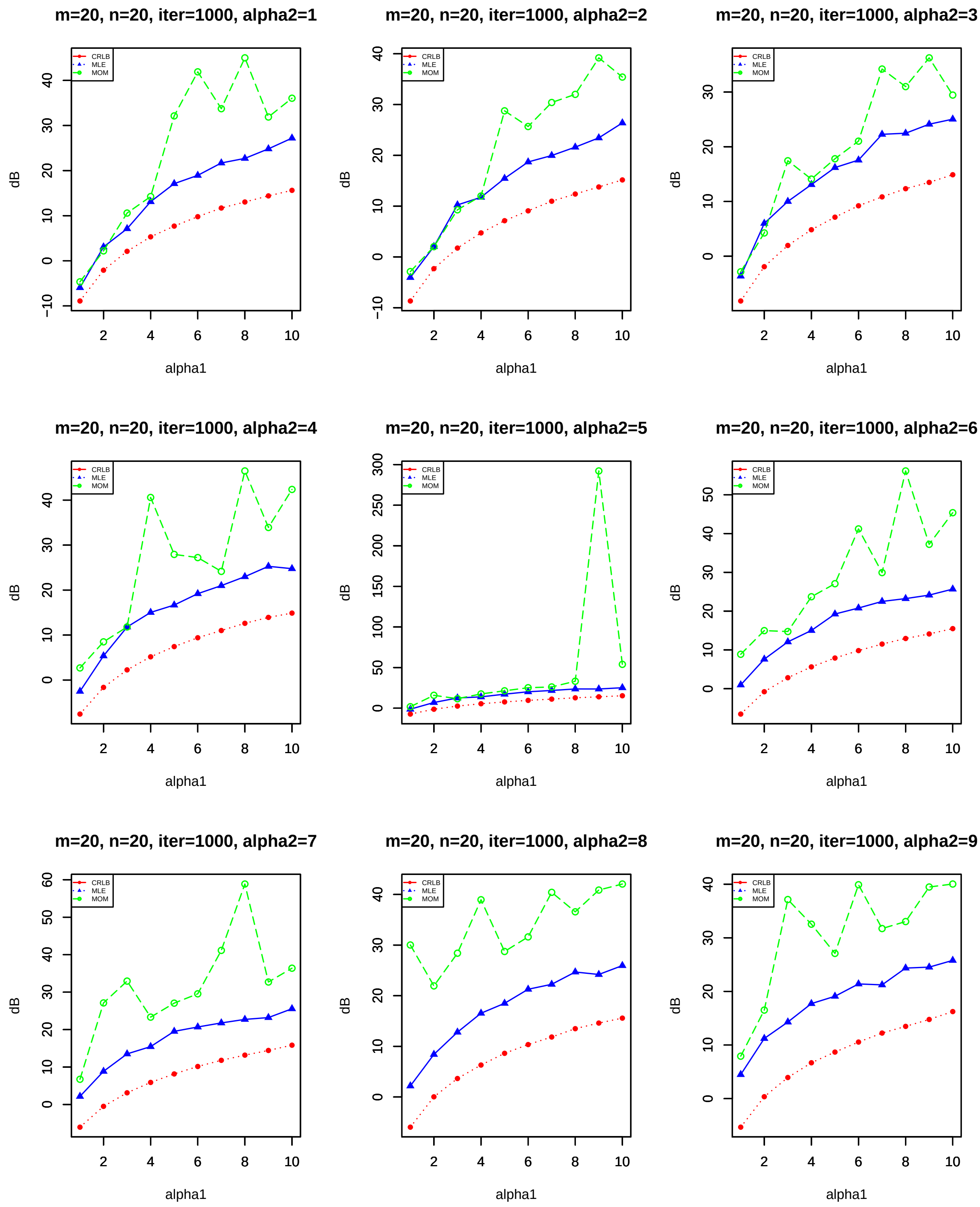

Figure 1: CRLB vs. MSE of ML/MOM in dB scale for different values of alpha, m = 20, n = 20 and iter = 1000

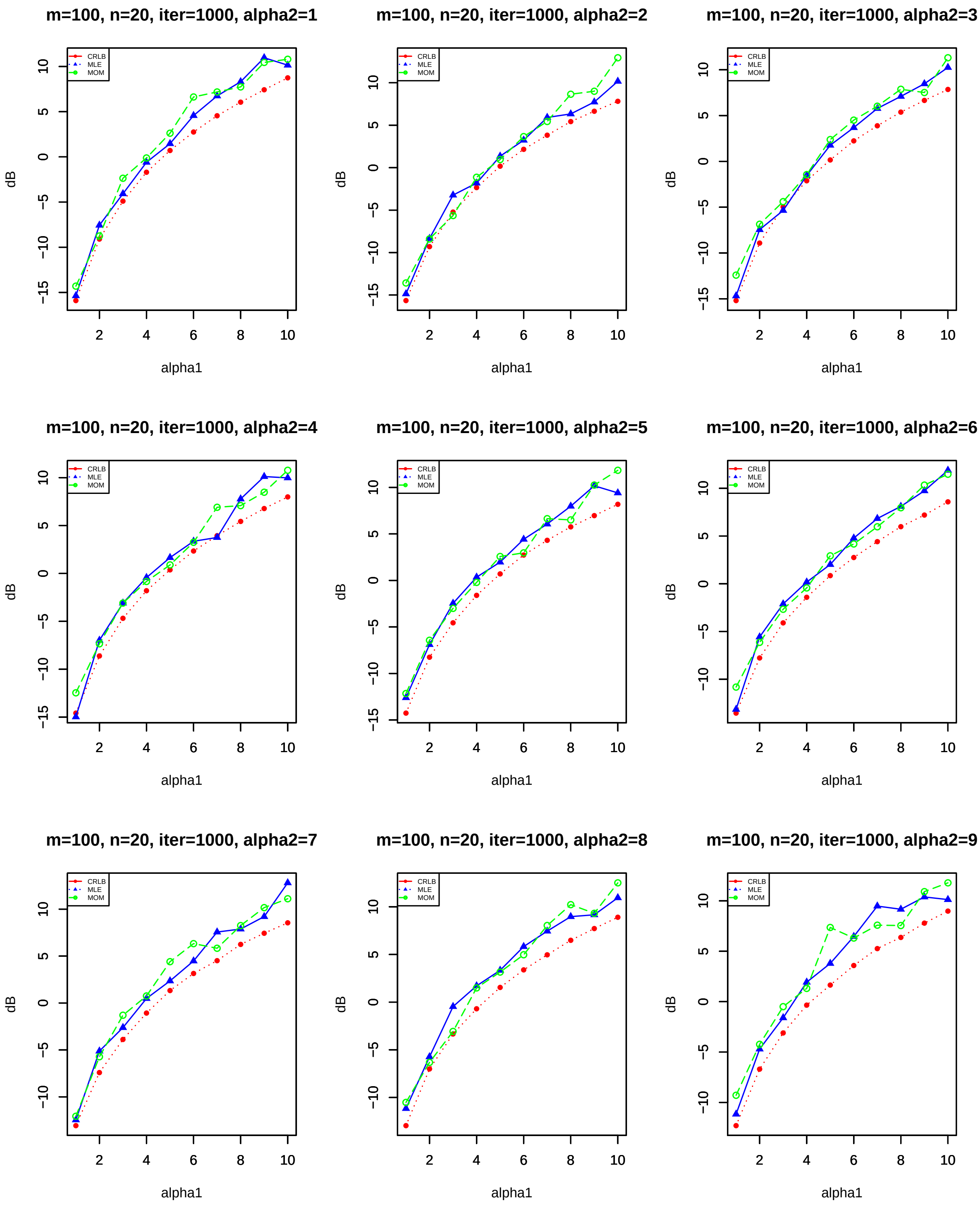

Figure 2: CRLB vs. MSE of ML/MOM in dB scale for different values of alpha, m = 1000, n = 20 and iter = 1000

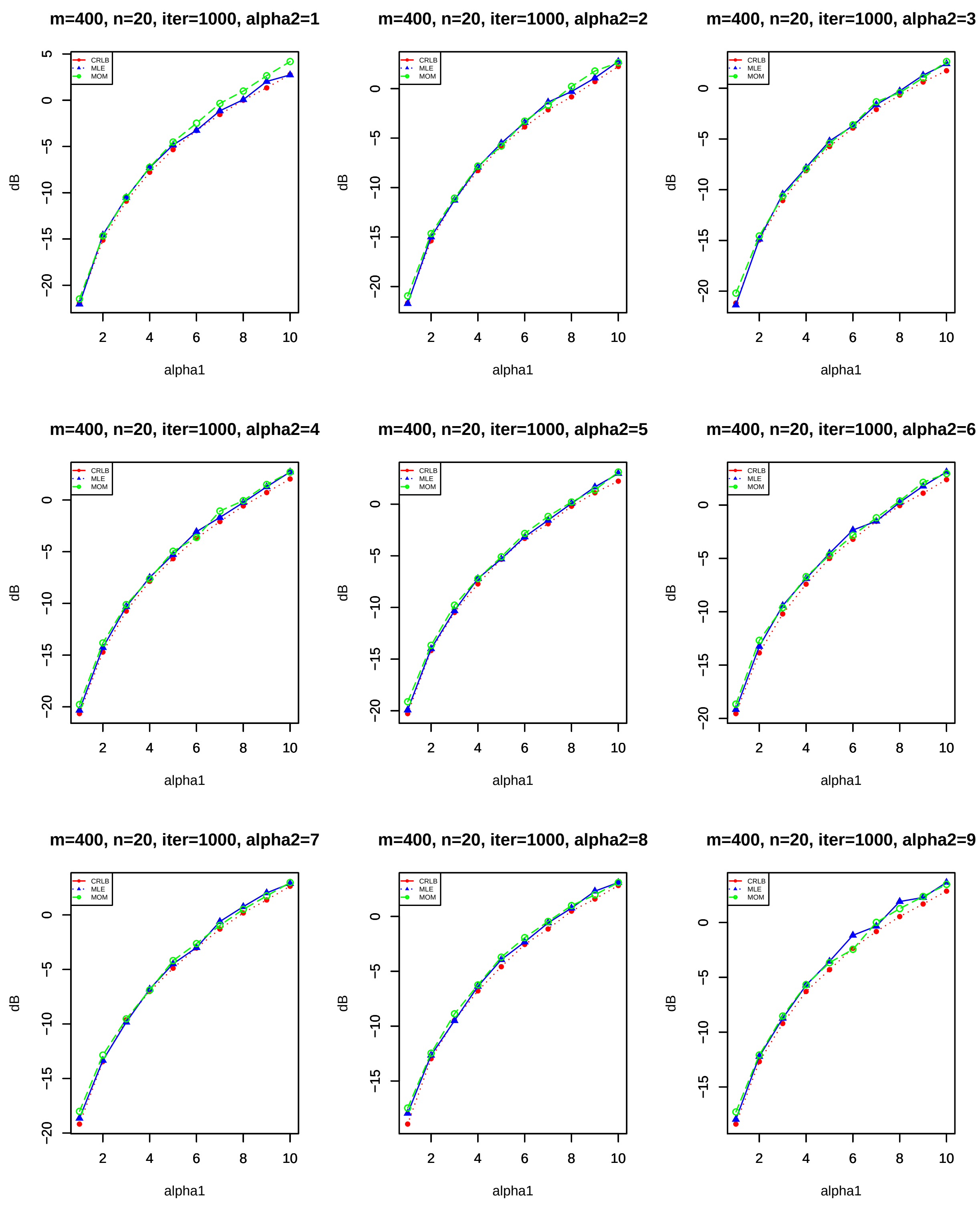

Figure 3: CRLB vs. MSE of ML/MOM in dB scale for different values of alpha, m = 400, n = 20 and iter = 1000

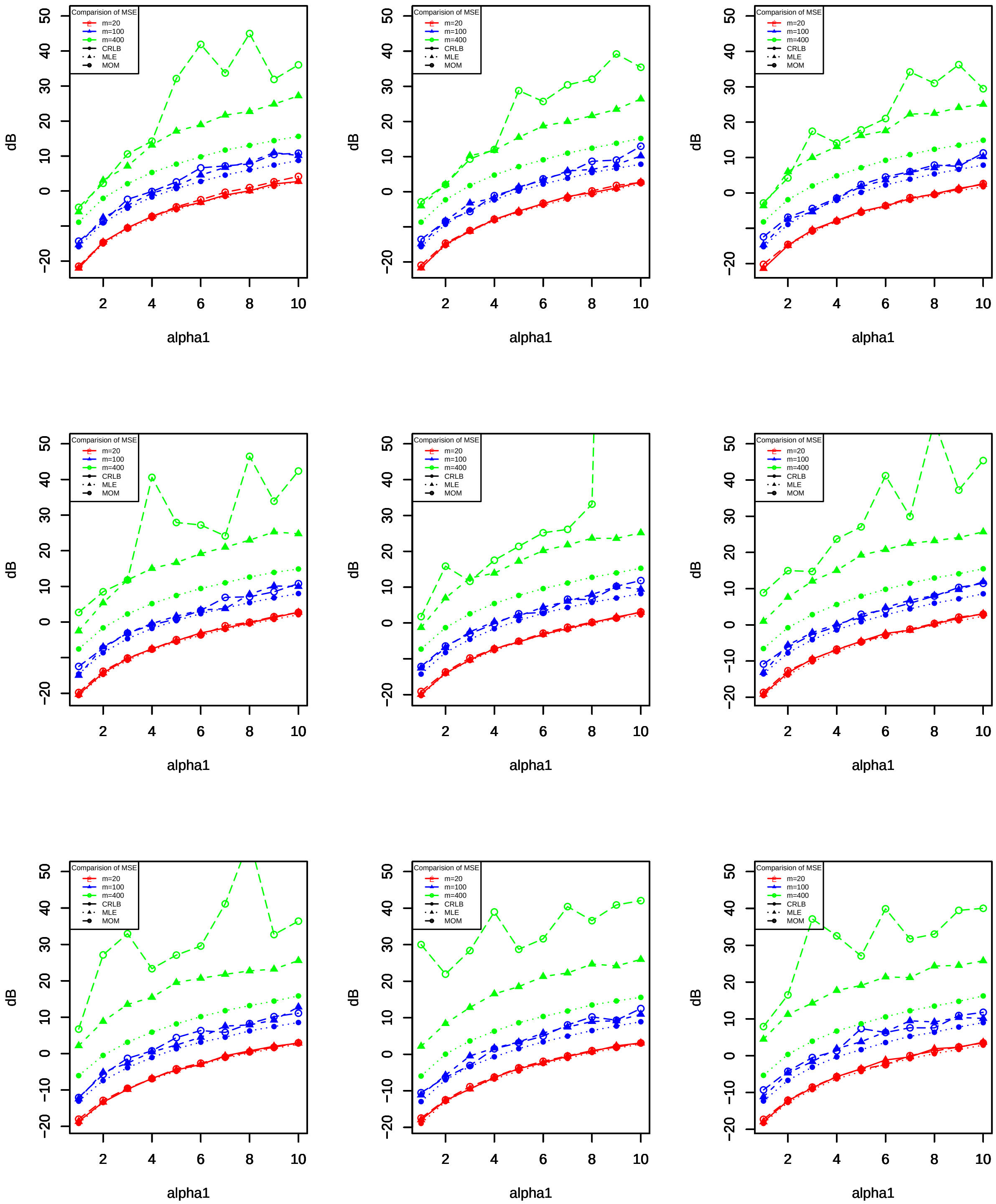

Figure 4: Performance of CRLB, MSE of ML/MOM in dB scale for different number of samples

6 Conclusion

The results show CRLB has the smallest value compared to all estimation methods for all experiments. Our empirical result agrees to the analytical explanation of CRLB, which claims CRLB is the lowest bound for MSE of any estimator. Furthermore, the lower bound variance for the examples generated from larger α_k parameters tends to have bigger value compared to data generated from smaller parameters.

In the other hand, MLE achieves slightly lower MSE than MOM for most examples generated. However, in few experiments there exist a situation where MOM beats MLE. Analytically, we expect MLE to outperform MOM for large sample size as MLE is asymptotically efficient. However, in the smaller sample size dataset there could be a situation where MOM could beat MLE and we observed this only for a very few cases.

References

[1] Thomas P. Minka. Estimating a Dirichlet distribution. 2012.

[2] B.N. Guo and F. Qi, Inequalities and Monotonicity for the Ratio of Gamma Functions, Taiwanese Journal of Mathematics, Vol 19, No. 7. pp. 407-409. (1976).

[3] David Blei,Anrew Y.Ng, Michael I. Jordan Laten Drichlet Allocation 2002