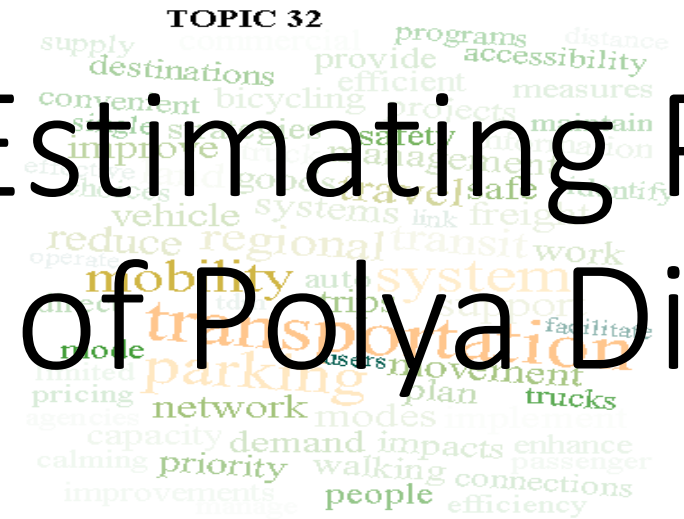
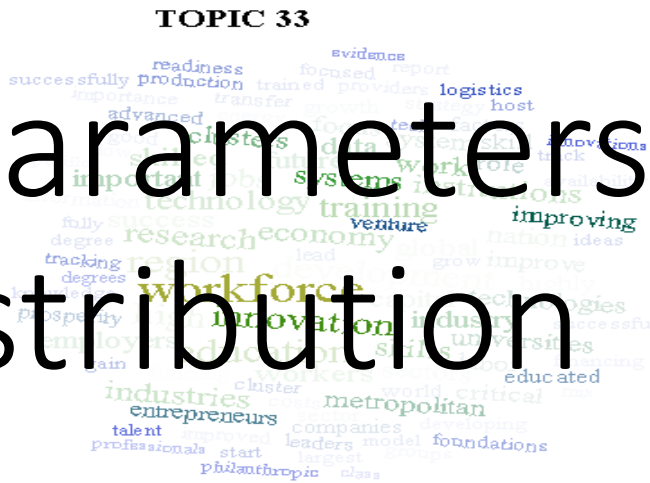




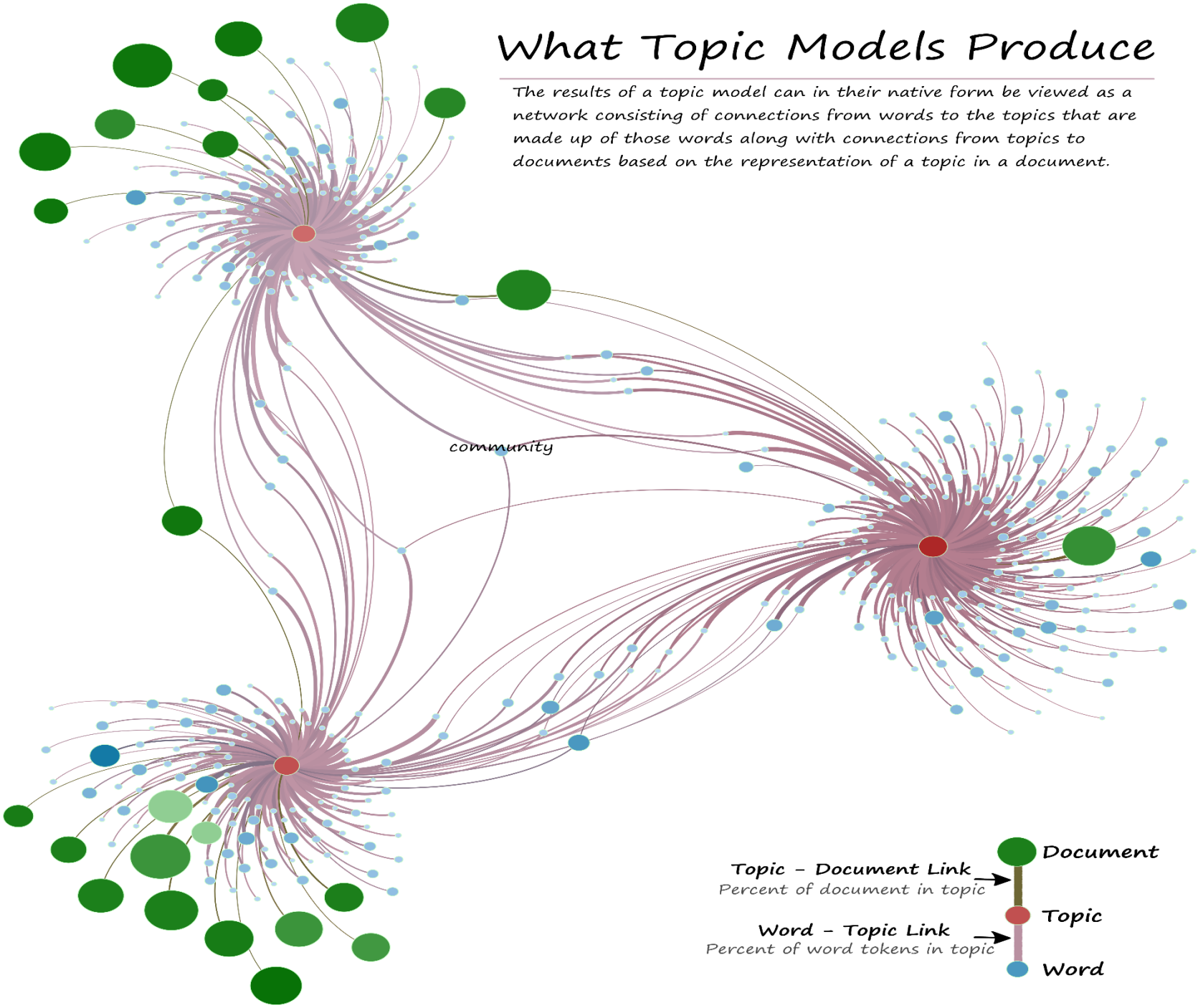
Estimating Parameters of Polya Distribution

Md Amran Siddiqui

Tadesse Zemicheal

What Topic Models Produce

The results of a topic model can in their native form be viewed as a network consisting of connections from words to the topics that are made up of those words along with connections from topics to documents based on the representation of a topic in a document.



Presentation Outline

- Introduction
- Problem formulation
- Estimation
  - Cramer Rao Lower Bound (CRLB)
  - Maximum Likelihood (ML)
  - Method of Moments (MOM)
- Experimental Result
- Observations

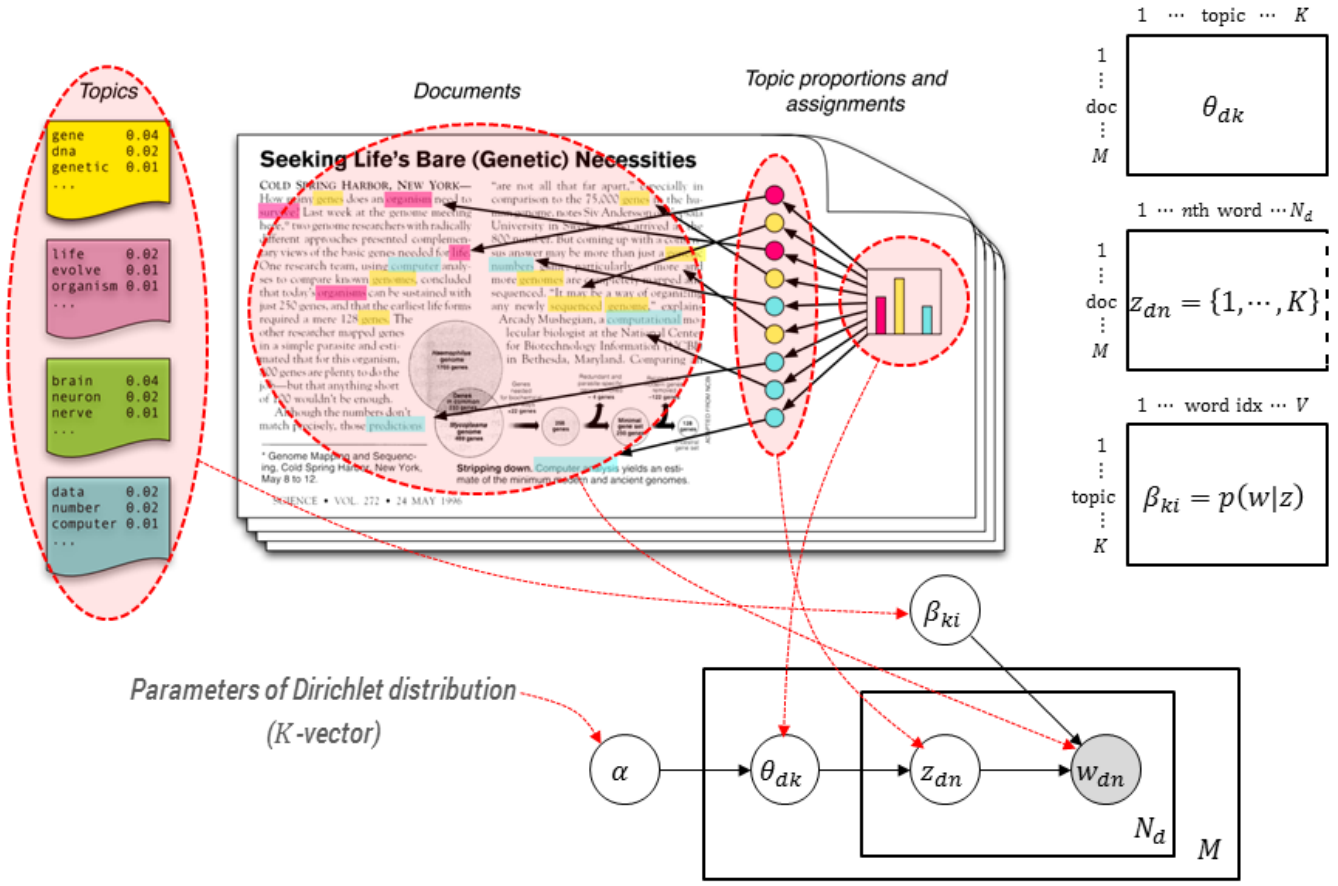
Topic modelling

- Methods for automatically organizing, understanding, searching and summarizing large electronic archives.
- Uncover hidden topical patterns in collections.
- Annotate documents according to topics

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Topic modelling



Model used

- Dirichlet-multinomial (Polya distribution)
  - Multinomial distribution k category

$$p_X(x_1,\ldots,x_K) = \begin{cases} \prod_{j=1}^{K} p_j^{x_j} & \text{if } (x_1,\ldots,x_K) \in R_X \\ 0 & \text{otherwise} \end{cases}$$

  - Dirichlet conjugate prior for the multinomial

$$P(\mathbf{p}|\boldsymbol{\alpha}) = \mathrm{Dir}(\alpha_1,\ldots,\alpha_K) \stackrel{\text{def}}{=} \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{K} p_j^{\alpha_j - 1}$$

- **Beta-binomial**
  - Topic: a mixture of distribution of words
  - Documents: bag of words
  - Word is drawn from two topics {0 or 1}

Problem Formulation

- Given m documents x1,x2,….xm
- Each xi contains n words
- PDF is given by

$$p(x \mid \alpha) = \frac{n!}{\prod_k n_{k!}} \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(n + \sum_k \alpha_k)} \prod_k \frac{\Gamma(n_k + \alpha_k)}{\Gamma(\alpha_k)} \qquad \theta = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ . \\ . \\ \alpha_k \end{bmatrix}$$

- The probability i.e. the likelihood of all documents is given by
- We have to estimate parameters from the distribution.

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

Cramer Rao Lower Bound (CRLB)

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

$\text{FIM}_{11} = -m * \left(\psi'(\sum \alpha_k) - \psi'(n_i + \sum \alpha_k) + E\left[\psi'(n_{i1} + \alpha_1)\right] - \psi'(\alpha_1) \right)$

$\text{FIM}_{12} = \text{FIM}_{21} = -m * \left(\psi'(\sum_k \alpha_k) - \psi'(n_i + \sum_k \alpha_k) \right)$

$\text{FIM}_{22} = -m * \left(\psi'(\sum_k \alpha_k) - \psi'(n_i + \sum_k \alpha_k) + E\left[\psi'(n_{i2} + \alpha_2)\right] - \psi'(\alpha_2) \right)$

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

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

- CRBL has no closed form in this case
- We had to compute numerically with Monte Carlo simulation

Maximum Likelihood Estimation

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

- No closed form solution exist
- Computed using fixed point iteration(Minka 2012)
- The final form of iteration is simplified into this form

$$\alpha_k^{new} = \alpha_k \frac{\sum_{i=1}^{m} \sum_{j=0}^{(n_{ik}-1)} \frac{1}{\alpha_k + j}}{\sum_{i=1}^{m} \sum_{j=0}^{(n_i-1)} \frac{1}{\sum_k \alpha_k + j}}$$

Method of Moments (MOM)

Population moments (Beta-Binomial):

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

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

Sample moments from data:

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

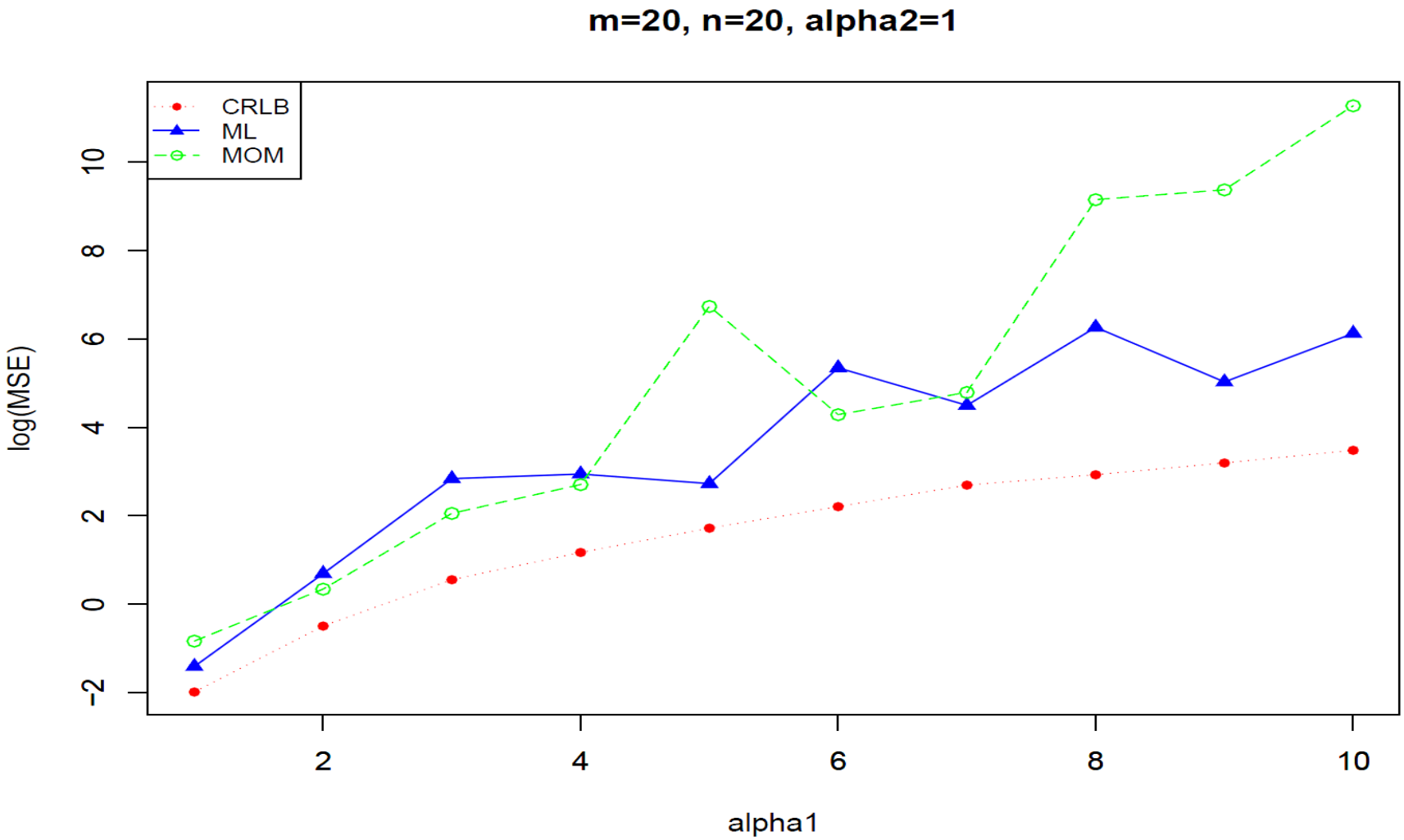
Equating population moments and sample moments and solving for the parameters α and β:

$$\alpha = \alpha_1 = \frac{nm_1 - m_2}{n(\frac{m_2}{m_1} - m_1 - 1) + m_1}$$

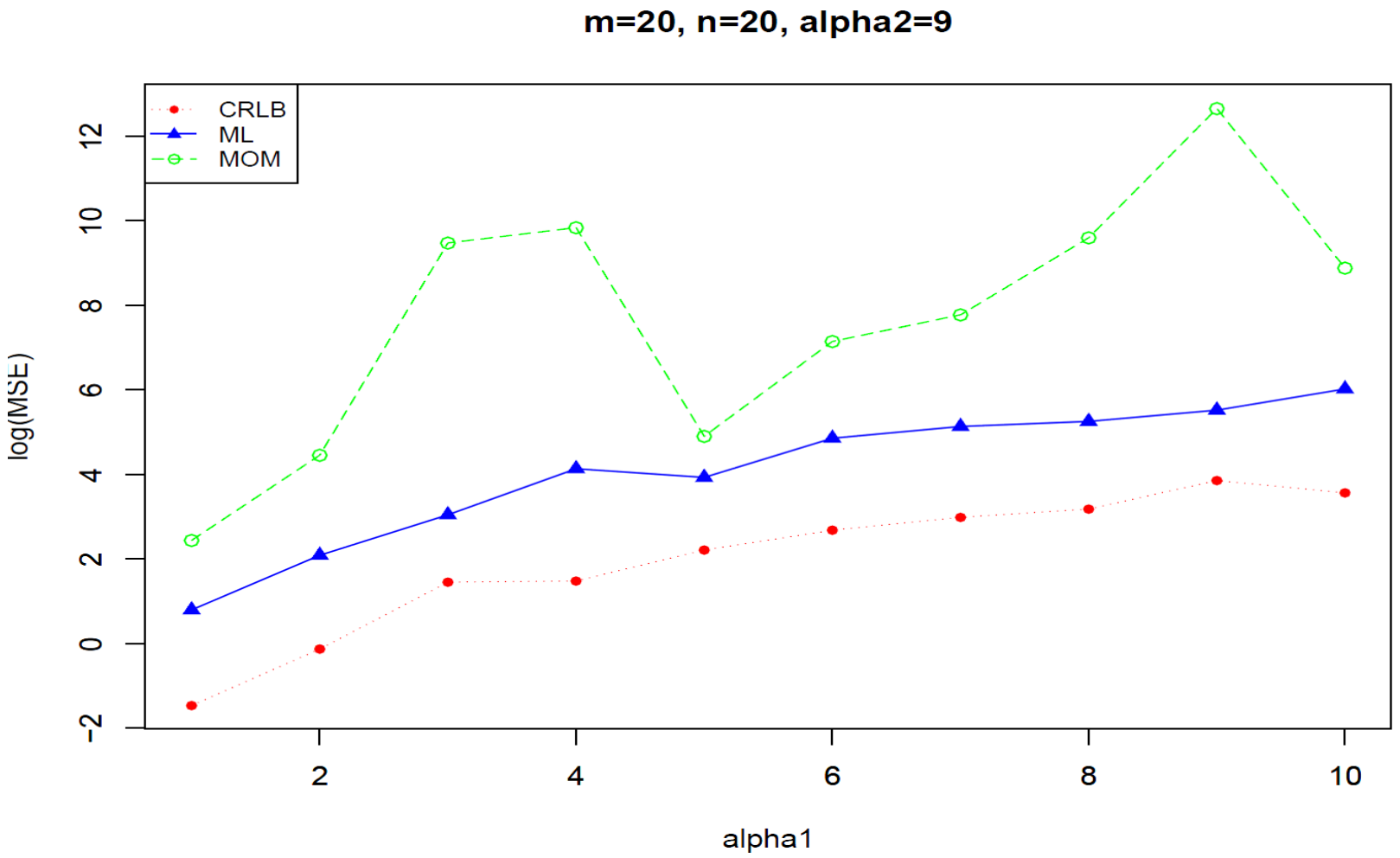
$$\beta = \alpha_2 = \frac{(n-m_1)(n-\frac{m_2}{m_1})}{n(\frac{m_2}{m_1} - m_1 - 1) + m_1}$$

Experimental results

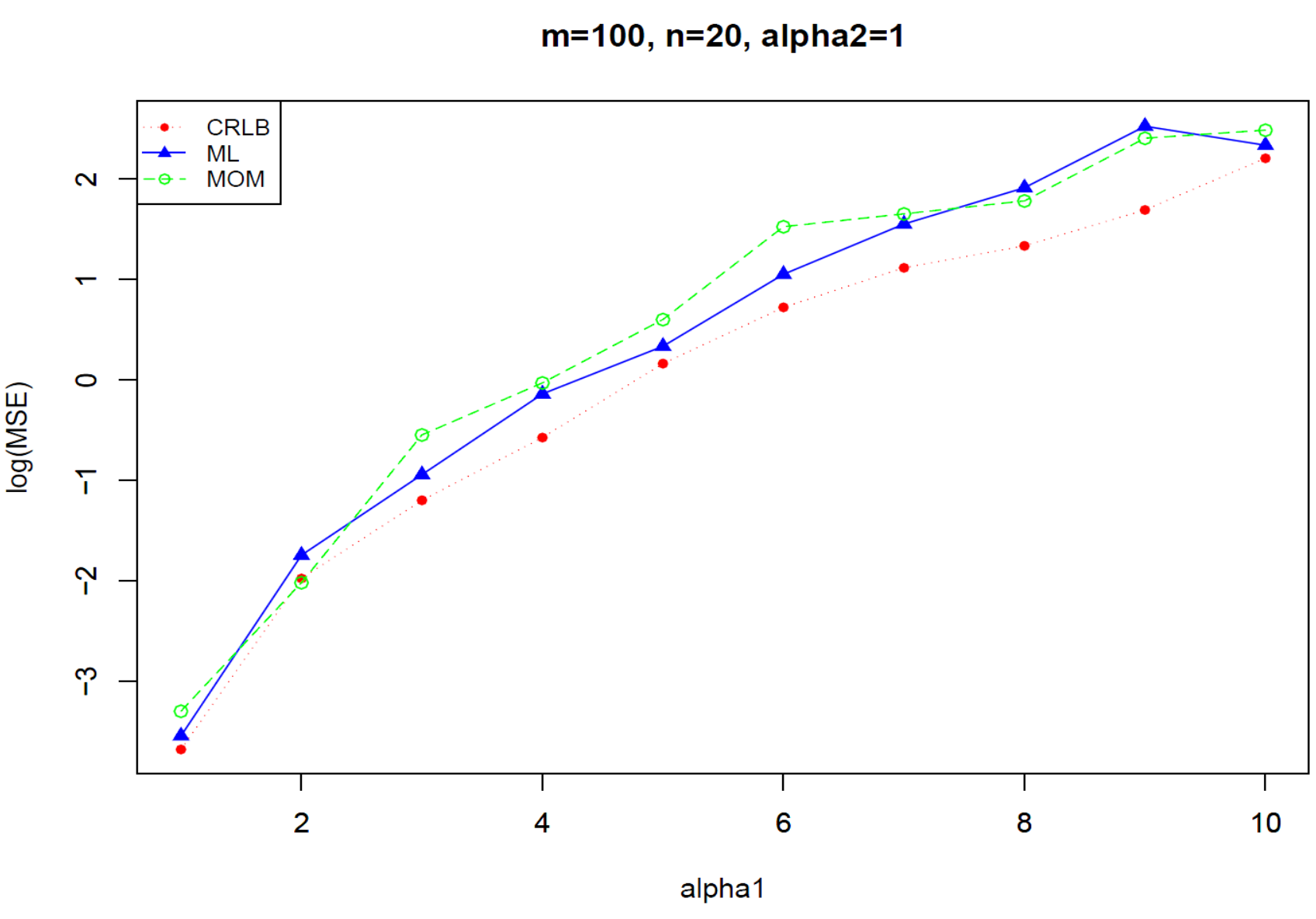
- For 200 Monte Carlo iteration



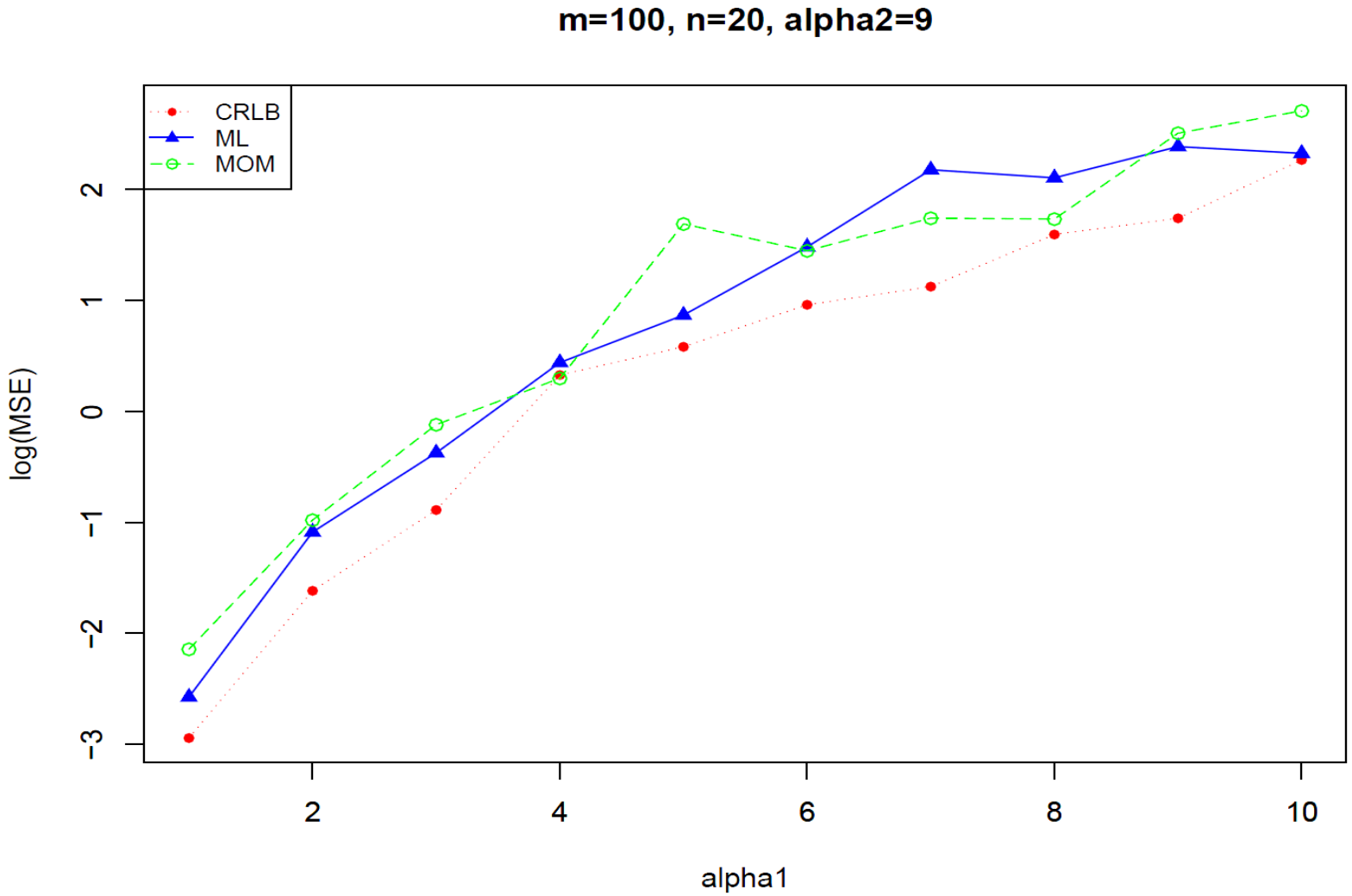
- For 200 Monte Carlo iteration



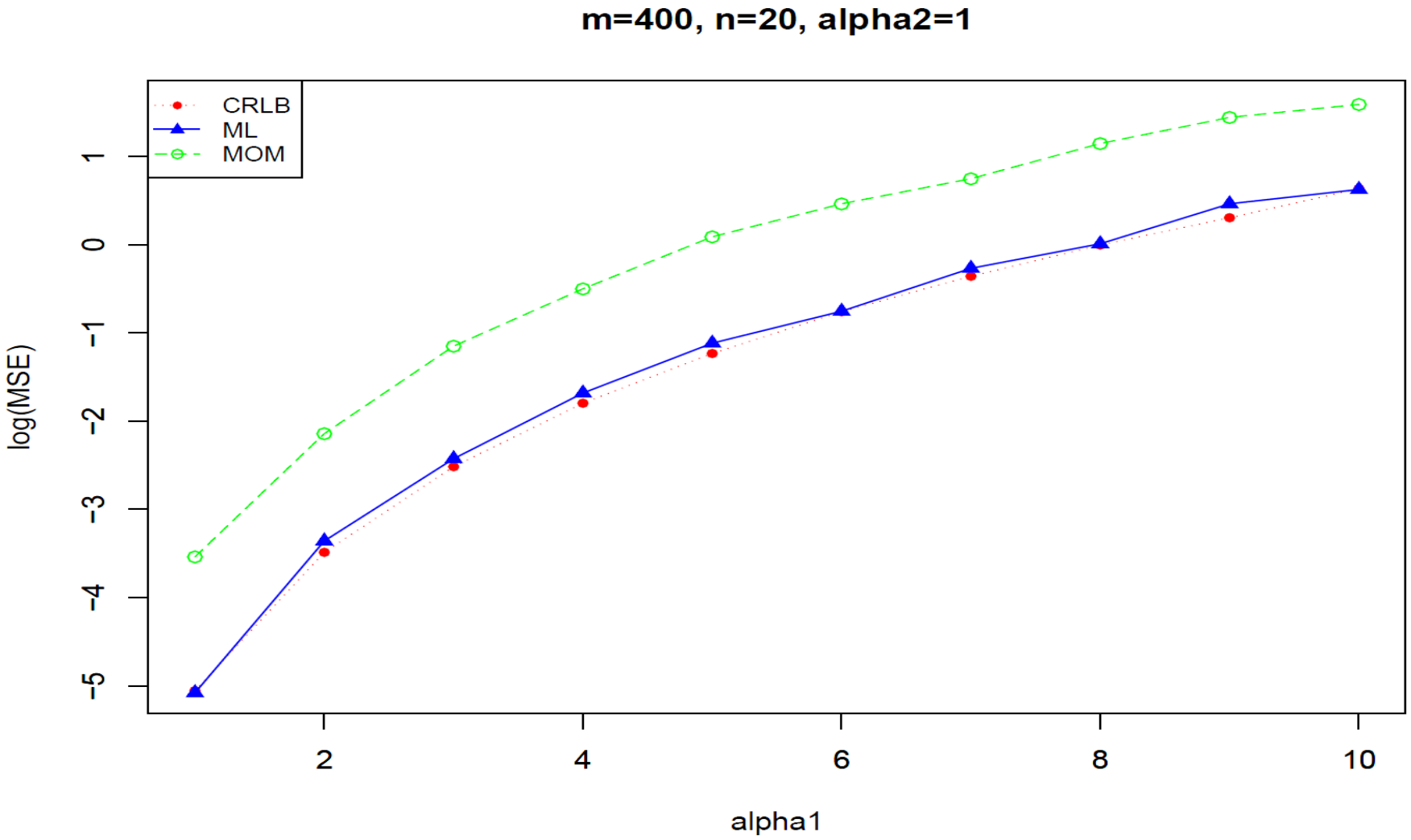
-Increasing number of examples



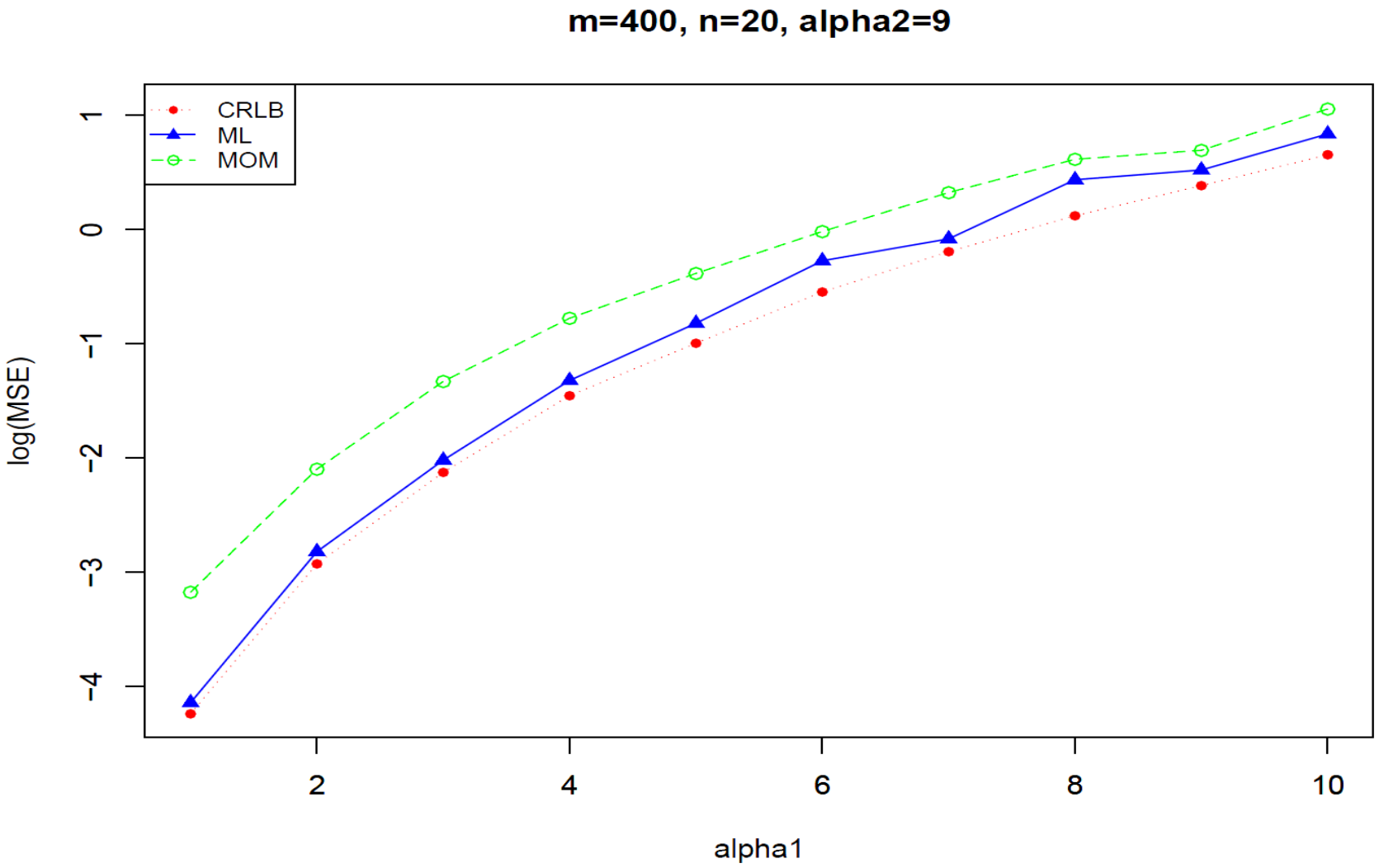
-MSE decreases for all estimators as we increase number of examples



-Increasing Monte Carlo iteration i.e. 1000



Increasing Monte Carlo iteration i.e. 1000



observations

- ML computationally expensive, takes hours to complete
- MOM computationally efficient
- ML is asymptotically efficient and approaches CRLB quicker than MOM.
- MSE tends to increase for large parameter sizes.
- Increasing number of examples decreases MSE

Questions?