Incomplete Data Estimation

by Fabiano SG de Oliveira
fgomes@lncc.br

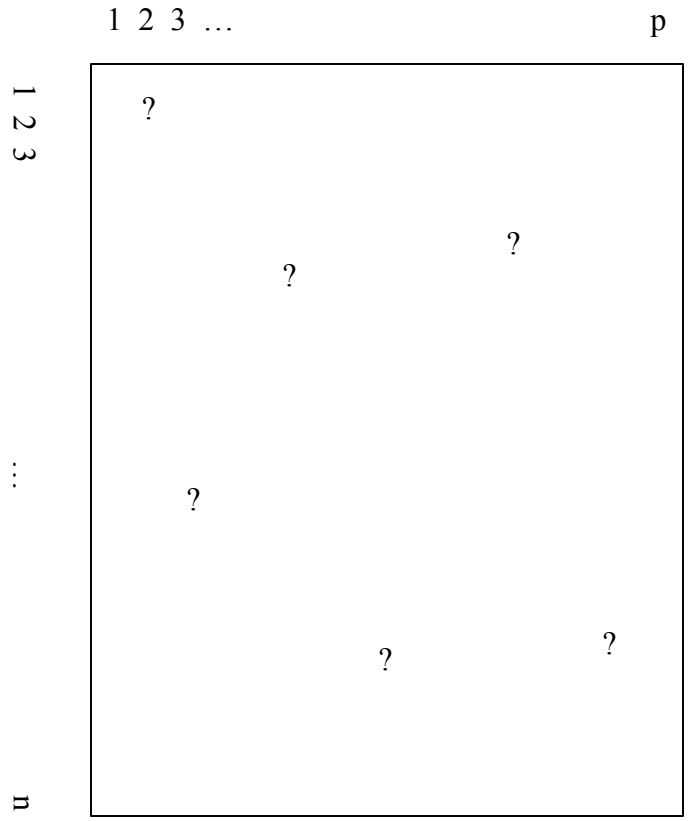
Multivariate Data

- Data Matrix has not observed values
- Not observed values: missing or censored
- Multiple imputation
- Maximum likelihood estimation
- EM algorithm

Incomplete Data

- Values
  - missing: values that you cannot measure
  - censored: parcially measured values

Graphical representation of Multivariate incomplete data



The censored data follow a truncated distribution because they are conditionated in an interval.
The missing data follow a conventional distribution.

Estimation of censored data will be a conditional expectation

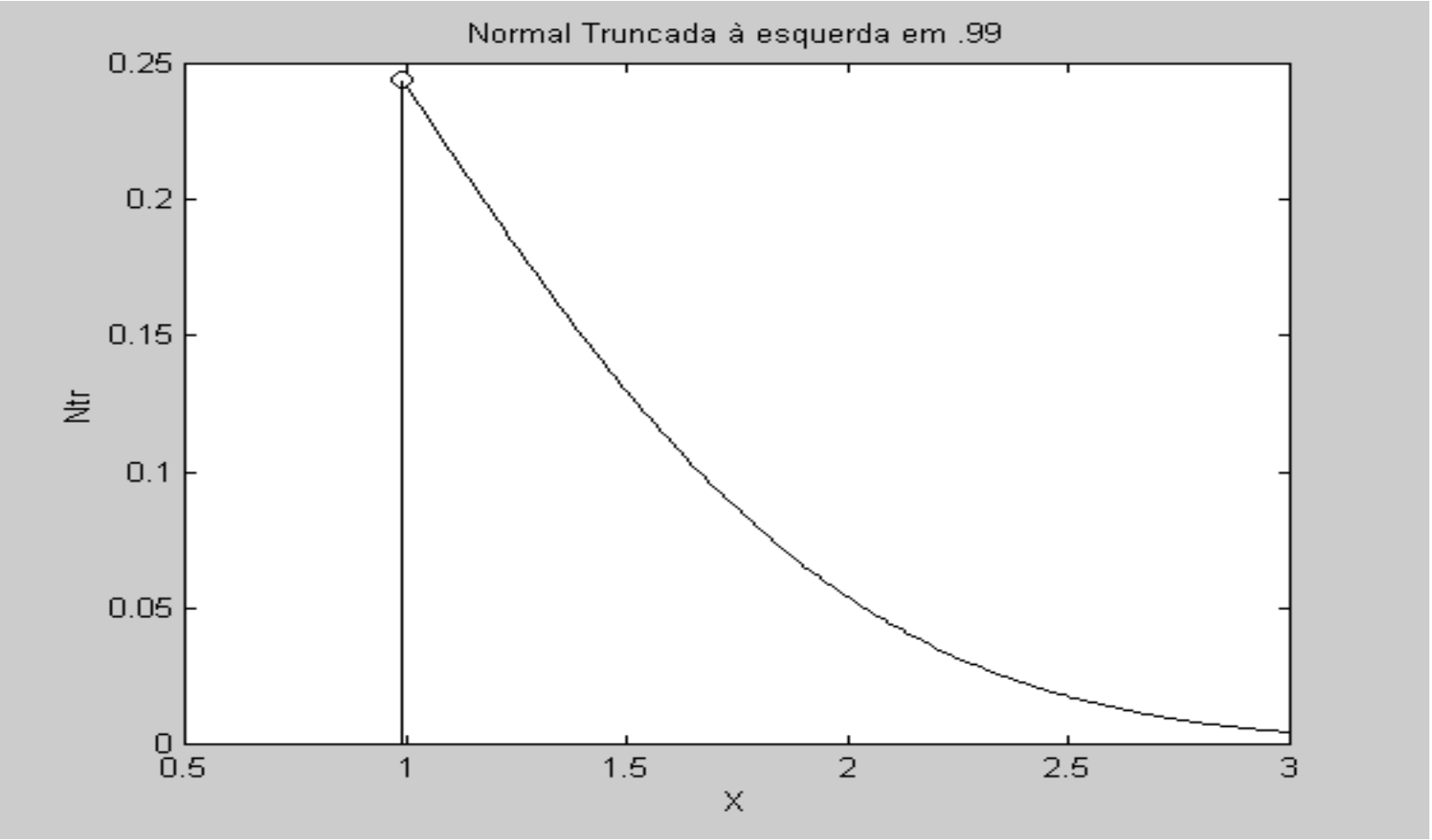
Multivariate truncated normal

p-dimensional

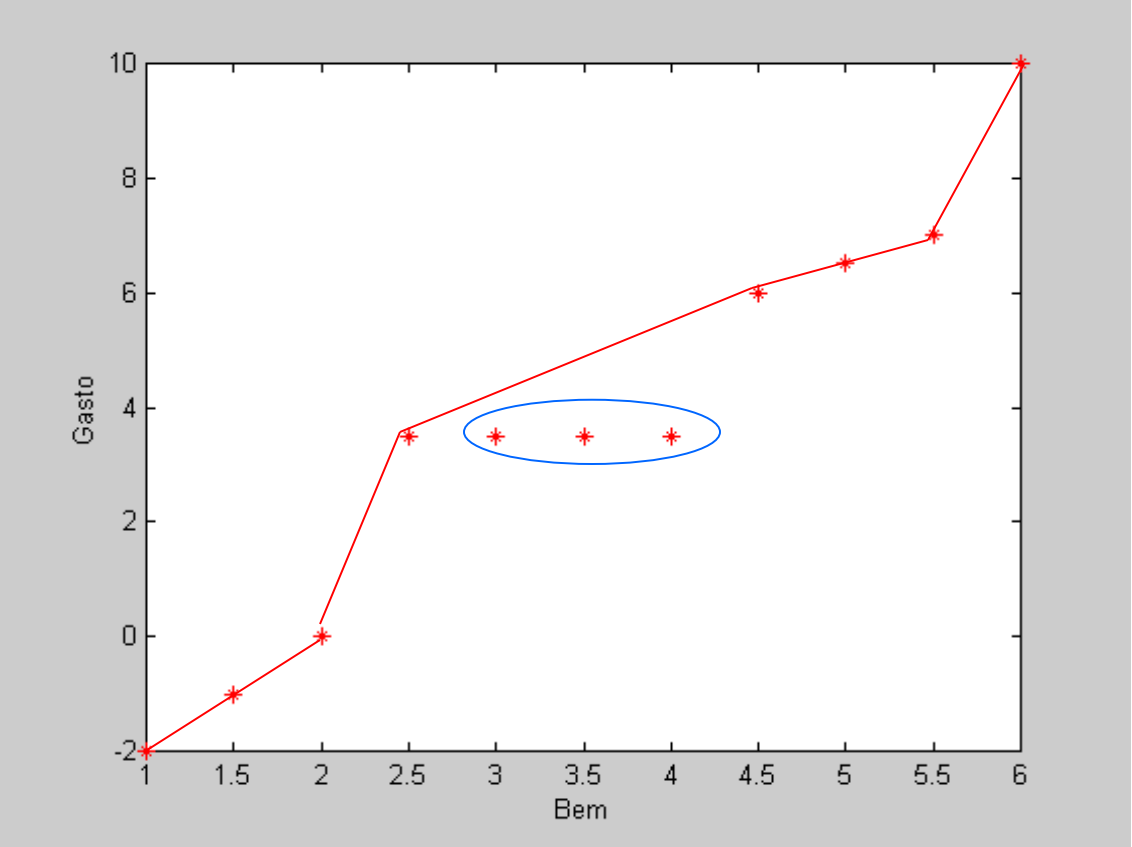
$$Ntr_p(\Lambda;\mu,\Sigma) = \begin{cases} \frac{1}{P} N_p(\Lambda;\mu,\Sigma) & \Lambda_j > c_j, j = 1,\ldots,p \\ 0 & outros \end{cases}$$

$$onde\ P = \int_{c_1}^{\infty} \cdots \int_{c_j}^{\infty} \cdots \int_{c_p}^{\infty} N_p(\Lambda;\mu,\Sigma)\, d\Lambda_1 \ldots d\Lambda_p$$

Truncated normal example



Censored data example



Data estimation

- A linear regression model
- An EM algorithm to the censored data
  - Maximum likelihood estimation using conditionated expectations based on the observations
  - Interactive algorithm
- Traditional Maximum likelihood estimation

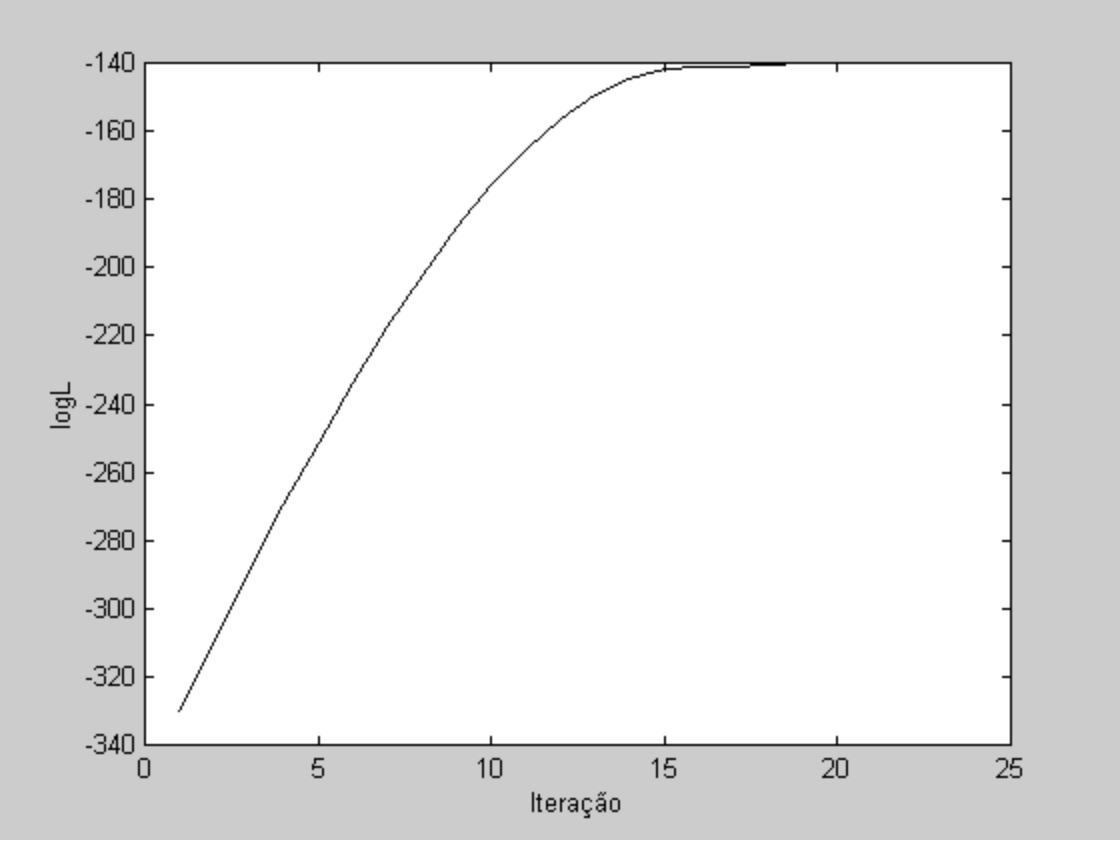
Data Matrixes tests

- In collaboration with Prof. Reinaldo Imbrozio Barbosa
- Rain falls measures – month data
- Caracaraí and Boa Vista cities
- Roraima state: northernmost of Brazilian Amazon
- Examples of censored and missing data

Rain Falls Data

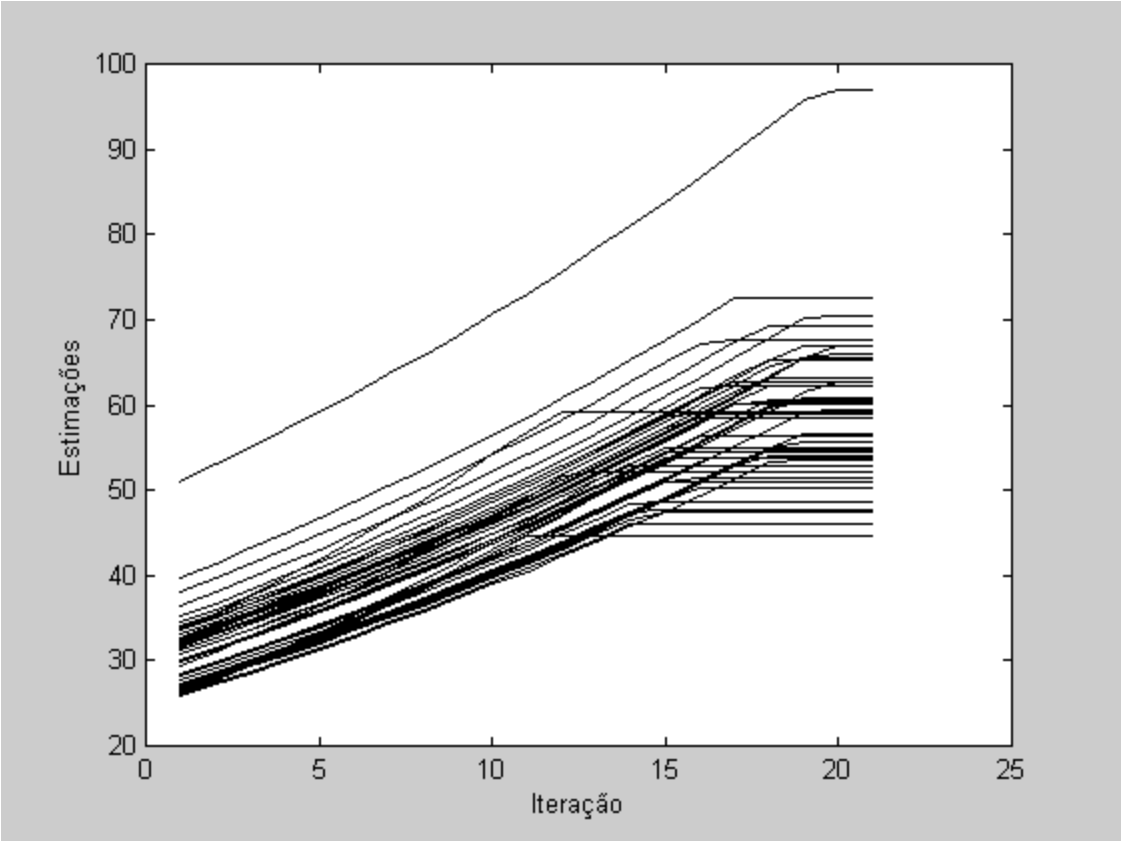
ANO/MES	JAN	FEV	MAR	ABR	MAI	JUN	JUL	AGO	SET	OUT	NOV	DEZ	TOTAL
1910	27,7	2,8	64,7	101,0	294,0	537,3	315,3	271,8	57,0	69,0	12,1	33,5	1786,2
1911	6,4	154,3		458,6	191,6	163,2	269,3	185,9	4,1	17,5	0,0	15,5	
1912	1,4	0,0	0,0	0,0	207,0	404,4	442,7	134,9	103,2	88,0	34,5	29,1	1445,2
1913	12,3	10,3	1,3	40,5	121,7	398,4	119,7	220,2	15,9	19,5	0,8	70,6	1031,2
1914	19,6	7,5	0,0	33,1	261,9	189,7	31,5	178,6	28,5	45,8	29,1	20,1	845,4
1915	24,0	188,5	45,3	146,1	184,9	285,6	274,8	95,0	50,7	21,7	25,0	10,4	1352,0
1916													
1917													
1918													
1919													
1920													
1921													
1922				15,5	312,2	429,6	278,3	394,8	119,1	152,7	55,0	125,3	
1923	41,2	0,9	9,2	68,8	321,8	303,5	480,1	226,6	27,4	10,3	23,1	66,0	1578,9
1924	29,6	0,0	2,4	78,6	345,1	467,6	365,1	138,1	186,1	212,8	0,0	0,0	1825,4
1925	37,4	11,0	124,4	0,0	216,6	337,9	195,1	91,7	23,4	19,6	0,0	0,0	1057,1
1926	0,0	0,0	0,0	0,0	176,7	414,4				9,4	0,0	51,8	
1927	55,2	9,7	99,3	122,7	306,7	283,8	363,4	100,0	34,0	0,0	31,6	181,9	1588,3
1928	5,1	5,2	0,6	32,7	338,9	212,6	323,0	102,6	22,2	63,7	98,2	3,7	1208,5
1929	0,0	0,0	8,5	81,2	383,7	291,1	369,1	178,2	59,4	14,7	45,6	18,9	1450,4
1930	5,3	3,2	4,7	101,5	252,5	307,2	168,9	81,2	90,2	1,8	2,8	3,0	1022,3
1931	13,6	82,7	1,8	35,3	364,5	374,7	480,1	150,2	176,7	41,1	119,4	1,2	1841,3
1932	65,0	18,8	19,8	191,4	266,7	256,3	238,3	120,1	60,3	72,8	0,0	18,4	1327,9
1933	12,5	27,6	9,5	96,6	417,6	353,0	449,6	102,2	91,6	31,4	130,2	88,0	1809,8
1934	62,2	4,4	14,0	135,0	57,4	239,2	414,8	225,8	118,1	321,2	37,8	17,9	1647,8
1935	18,4	84,5	35,2	32,0	404,0	374,0	275,2	156,0	123,2	24,2	43,0	44,4	1614,1
1936	17,4	0,0	72,0	197,0	282,2	208,8	207,6	173,0	114,2	120,8	50,2	25,2	1468,4
1937	29,2	23,0	26,8	88,6	363,8	429,2	406,0	247,4	86,4	138,4	128,0	56,6	2023,4
1938	86,0	73,8	126,9	195,9	466,6	393,2	320,0	255,8	74,0	73,2	93,7	14,8	2173,9
1939	33,4	0,0	62,0	129,5	212,4	192,1	175,4	156,4	40,8	68,8	24,0	11,7	1106,5
1940	0,0	0,9	16,0	2,9	245,0	359,0	214,0	284,2	30,6	80,8	16,4	13,0	1262,8
1941	23,2	6,2	41,6	20,0	198,4	512,6	202,4	179,8	169,8	101,5	6,0	0,0	1461,5
1942	0,0	4,2	12,2	112,8	484,7	497,2	310,0	263,1	97,2	98,0	69,6	181,4	2130,4
1943	54,4	61,2	107,6	197,0	265,6	261,8	574,2	273,8	76,6	0,0	45,6	77,4	1995,2
1944	7,0	0,0	94,4	142,5	427,8	446,8	372,1	324,5	135,6	11,0	66,8	107,8	2136,3
1945	28,4	66,4	131,6	363,6	449,8	565,9	461,1	305,2	73,5	64,4	32,3	12,4	2554,6
1946	8,2	15,2	43,0	254,4	338,3	257,0	333,5	315,0	23,2	75,6	114,4	0,0	1777,8
1947				30,0	147,6	438,0	605,4	159,2	131,0	75,2		6,2	
1948	31,4	66,0	121,2	201,8	275,2	454,6	345,6	183,6	125,0	24,2	77,6	4,0	1910,2
1949	18,0	17,4	104,2		319,6	356,8	476,2	507,2	257,8	55,1	92,8	94,2	
1950	207,2	130,8	39,0	47,2	421,8	458,0	523,0	375,7	49,4	28,0	67,6	15,2	2362,9
1951		146,1	91,4	204,4	315,8	492,0	595,3	225,5	152,2	91,2	11,0	0,0	
1952	6,4	0,0	0,0	142,6	267,3	471,8	440,4	278,0	100,0	56,0	19,0	106,0	1887,5
1953	139,2	90,8	199,0	201,0	227,0	535,0	386,0	104,0	104,0	8,0	80,0	88,0	2162,0
1954	25,0	62,0	0,0	324,7	304,7	389,9	393,0	458,0	217,2	62,0	161,0	84,0	2481,5
1955	12,0	68,0	89,1	52,8	333,1	220,5	426,0	98,0	137,0	49,0	73,2	105,0	1663,7
1956	25,2	0,0	113,6	99,8	467,9	475,7	235,8	364,2	129,0	124,7	101,6	120,9	2258,4
1957	13,0							155,1	52,2	57,8	15,6	4,0	
1958	12,1	19,8	32,2	179,6	139,4			128,2	2,9	13,9	20,9	3,8	
1959													
1960													
1961	2,5	0,4	1,9	0,2	43,3	404,4	361,9	217,6	101,9	34,7	45,5	16,0	1230,3
1962													
1963													
1964													
1965													
1966	0,0	0,0	50,3	49,3	160,3	339,7	387,6	164,7	178,0	33,0	25,6	15,4	1403,9
1967	12,3	15,8	89,6	169,8	403,4	453,7	408,7	205,9	76,0	2,7	61,3	22,8	1922,0
1968	16,5	27,4	32,4	132,7	390,5	649,8	238,5	71,0	87,6	24,1	269,0	8,2	1947,7

Likelihood function



x 100

Estimations



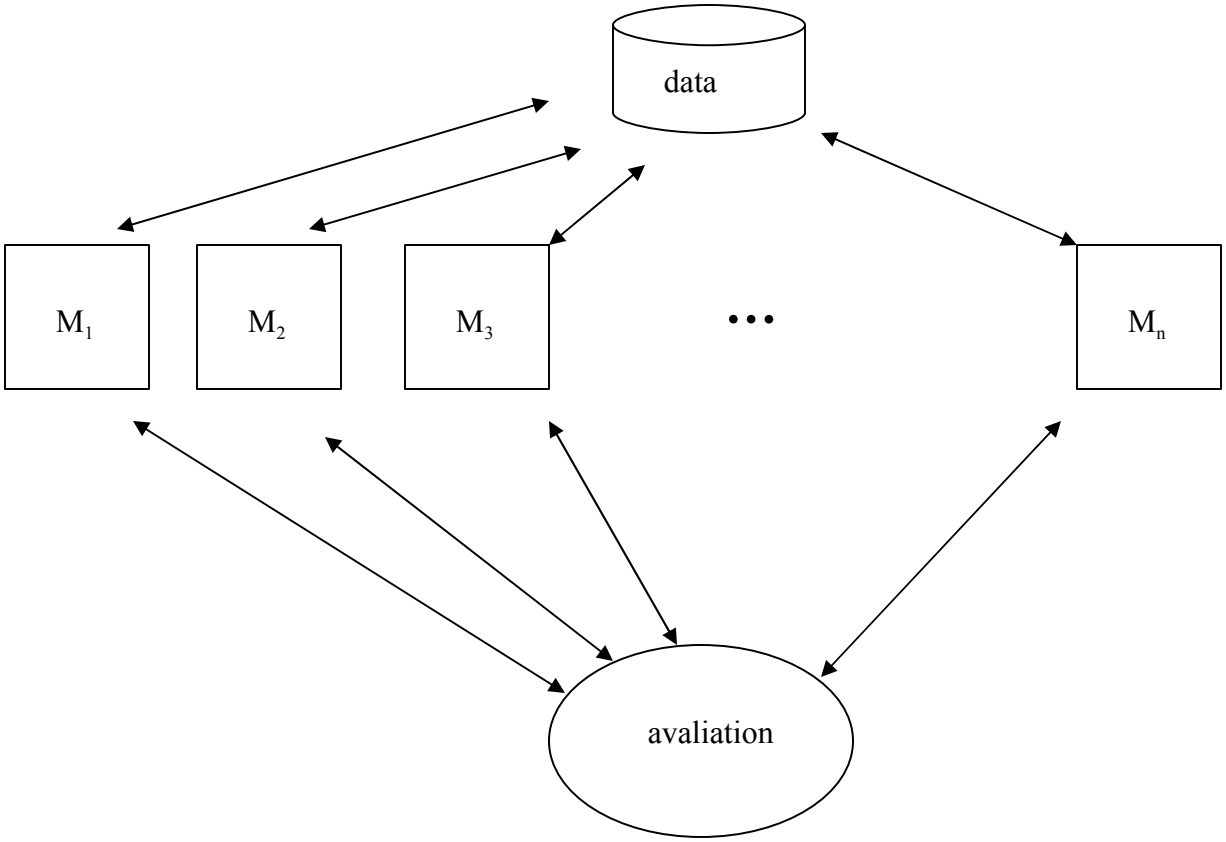
X 100

Problems

- EM algorithm convergence
- Small samples
- Huge data matrixes

- Sugestions
  - Bayesian approaches
  - Stochastic Simulation
  - Parallel algorithms - efficiency

Future: Comitee System



Thank you.

- Contact: fgomes@lncc.br