Representative Samples Selection in An Implicit Mixture Model with The Approximation Maximization Algorithm

Ben Liu

University of Ottawa

Contents

Introduction

Technological innovations have made a profound impact on knowledge discovery. Extracting useful samples from massive dataset is essential in many modern scientific areas. In this project, an Approximation Maximization (AM) algorithm was developed to select representative samples in an implicit mixture model.

Introduction

Statistical Problem

Assume that each y_i is generated by two-components mixture distribution. Let $p(z_i = k) = \pi_k, k = 0, 1$, then the density of y_i is:

$$f_{\boldsymbol{\theta}^*}(y_i) = \pi_1 f_{\boldsymbol{\theta}_1^*}(y_i|z_i = 1) + \pi_0 f_{\boldsymbol{\theta}_0^*}(y_i|z_i = 0), \tag{1}$$

where $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_0^*)$. The pdf $f_{\boldsymbol{\theta}_1^*}(y|z = 1)$ is known but $\boldsymbol{\theta}_1^*$ needs to be estimated, and $f_{\boldsymbol{\theta}_0^*}(y_i|z_i = 0)$ is unknown with $\boldsymbol{\theta}_0^*$. Another assumption is that $f_{\boldsymbol{\theta}_1^*}(y_i|z_i = 1)$ and $f_{\boldsymbol{\theta}_0^*}(y_i|z_i = 0)$ are quite different. The objective is to estimate $\boldsymbol{\theta}_1^*$ and select the representative samples from $f_{\boldsymbol{\theta}_1^*}(y_i|z_i = 1)$.

Approximation Maximization algorithm (AM)

- Input $\gamma_i^{(0)}$, assume $y_i \sim f_{\theta_1}(y|z=1), i=1,2,...,n$, compute

$$\theta_1^{(0)} = \underset{\theta_1}{\text{argmax}} \; l_n(\theta_1).$$

- **Approximation** Step: Calculate

$$g_{\theta_1^{(t)}}(z_i|y_i) = \begin{cases} 1, & f_{\theta_1^{(t)}}(y_i|z_i=1) \geq \gamma_i^{(t)}, \\ 0, & f_{\theta_1^{(t)}}(y_i|z_i=1) < \gamma_i^{(t)}, \end{cases} \quad i=1,2,...,n.$$

- **Maximization** Step: Update θ_1 by

$$\theta_1^{(t+1)} = \underset{\theta_1}{\text{argmax}} \; G_{1,n}(\theta_1|\theta_1^{(t)}).$$

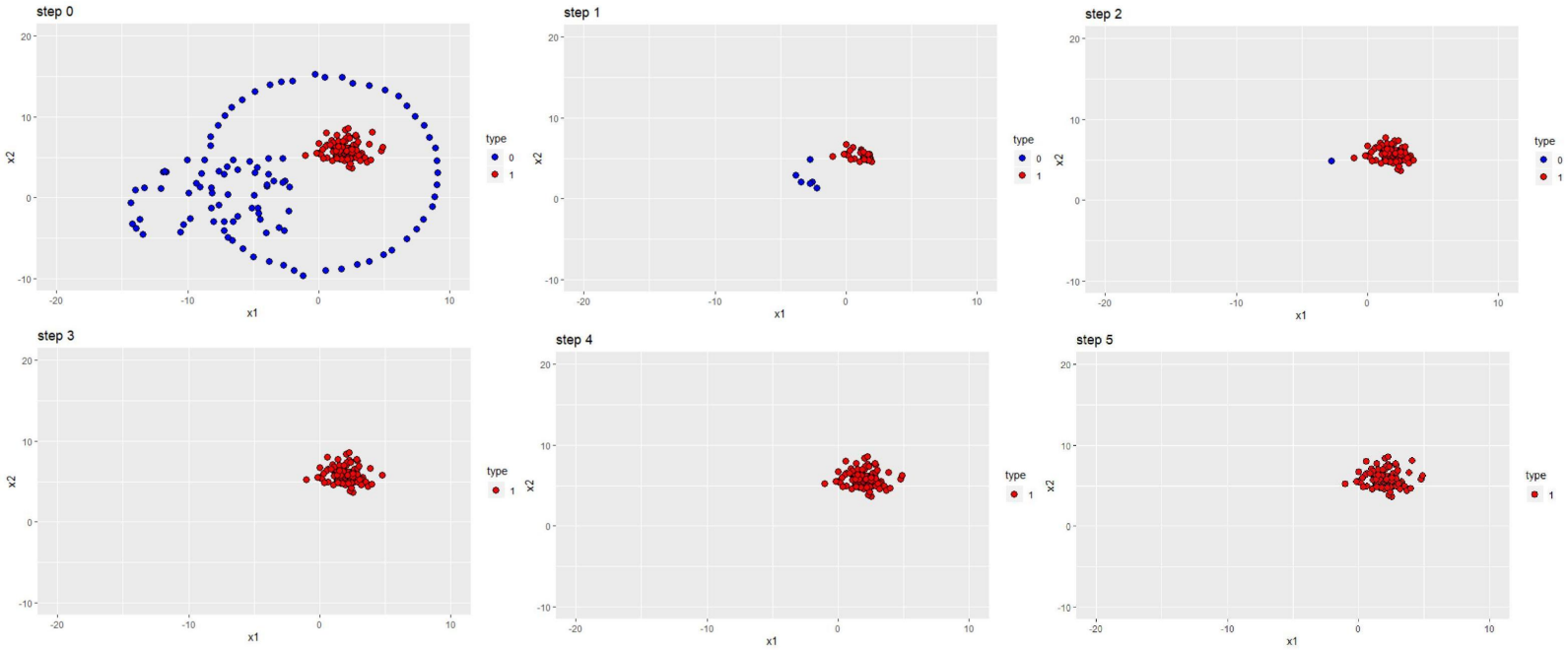
- Repeat step 2, 3 until $g_{\theta_1^{(t+1)}}(z_i|y_i) = g_{\theta_1^{(t)}}(z_i|y_i), i=1,2,...,n.$
- Output $\theta_1^{(t+1)}$ and $S = \left\{ y_i \;\middle|\; g_{\theta_1^{(t+1)}}(z_i|y_i) = 1, i=1,2,...,n. \right\}$.

Experiment settings

- Data: Four classes of implicit mixture datasets, where the explicit models are normal distribution, linear regression, logistic regression, and Poisson regression.
- Models: AM, FAM (a more general version of AM), EM, MLE, and K-means.
- Covariance structures: Independent with each other (S1), auto-regressive correlated (S2), and compound symmetry(S3).
- Metrics: Deviation of true parameters (DEV), positive selection rate (PSR), false discovery rate (FDR), the final number of selected representative samples (FN), and computation time (Time).

Case 1.1 Normal mixture model (2 variables)

In this case, more and more red points were selected into representative samples set as iteration grew. On the other side, the blue noise samples were ruled out gradually by the AM algorithm, which meant AM could effectively select the representative samples.



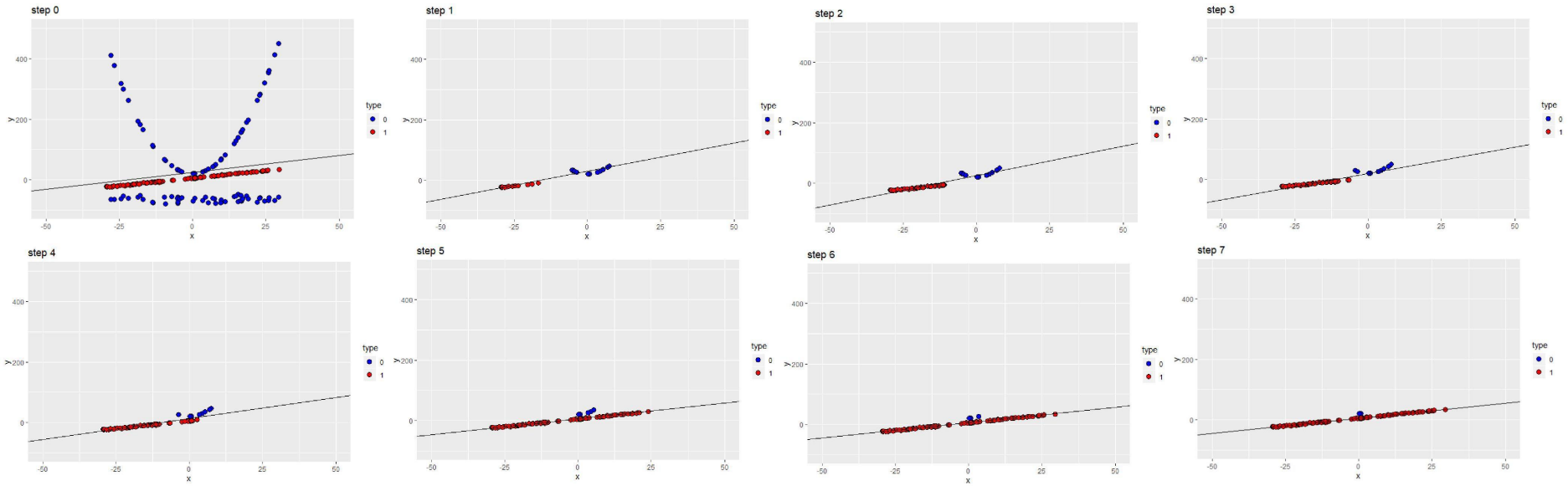
Case 1.2 Normal mixture model (multiple variables)

From the below table, it's noticeable that more complex covariance structure in this experiment did not lead to obvious reduction of accuracy for AM and FAM. By comparing AM with FAM, AM was computationally more efficient than FAM, which was due to a close form of true parameter in this model.

Setup	Method	DEV(%)	PSR(%)	FDR(%)	FN	Time
S1	MLE	31.63	-	-	-	1.02
	EM	13.92	99.96	33.29	7493	5.56
	K-means	14.89	100.00	42.41	8772	0.02
	AM	0.30	99.96	1.15	5056	3.14
	FAM	0.30	99.96	1.15	5056	21.88
S2	MLE	30.76	-	-	-	1.32
	EM	14.36	99.98	33.28	7493	5.61
	K-means	14.83	100.00	42.71	8727	0.01
	AM	0.59	99.96	2.01	5100	3.99
	FAM	1.63	100.00	4.82	5253	59.52
S3	MLE	30.83	-	-	-	1.02
	EM	13.97	99.95	33.29	7494	5.18
	K-means	14.92	100.00	42.71	8727	0.01
	AM	0.60	99.96	2.08	5104	8.31
	FAM	1.11	100.00	2.51	5129	42.86

Case 2.1 Linear mixture model (2 variables)

For this case, the AM algorithm successfully selected almost all the red representative samples and few blue noise samples after 7 iterations.



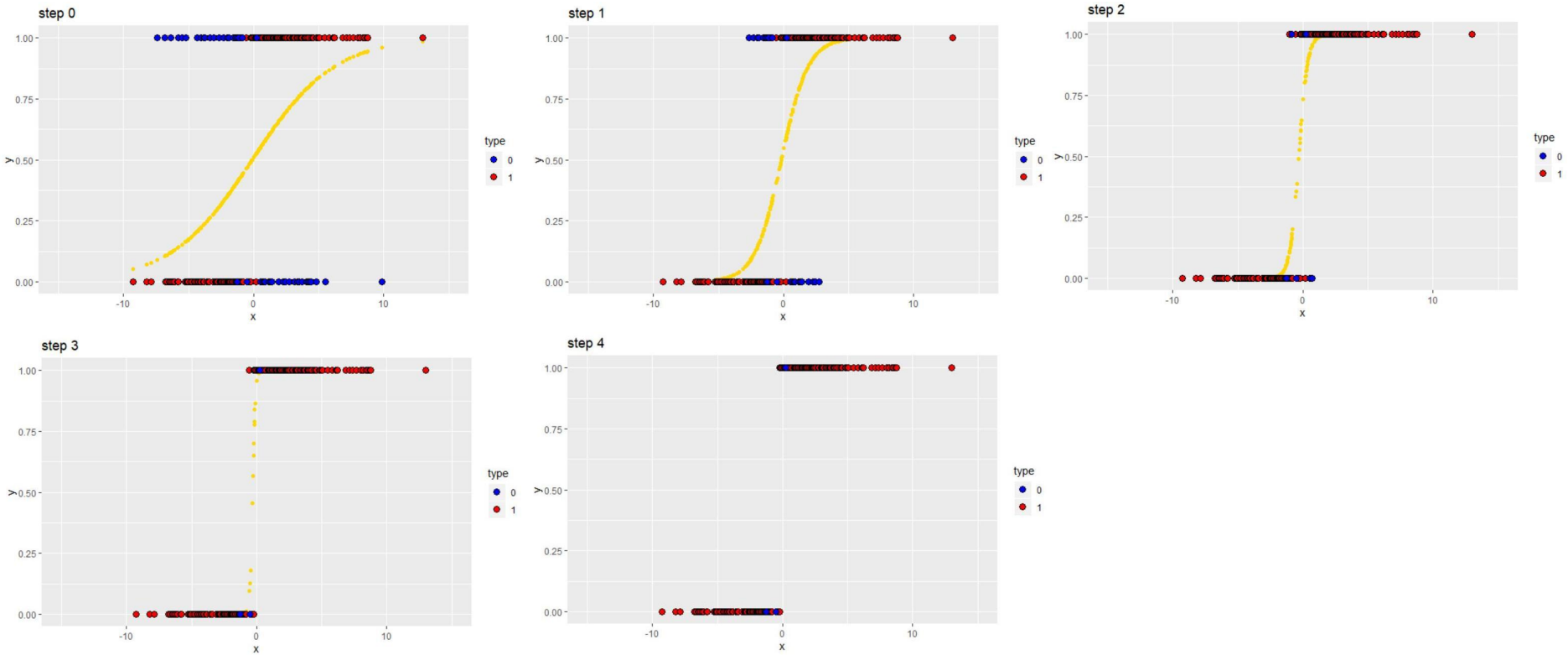
Case 2.2 Linear mixture model (multiple variables)

In this experiment, the performance of all 5 methods was similar. AM outperformed other methods in terms of DEV and selection indexes. Due to the similar reasons, MLE, EM and K-means failed to obtain accurate estimate and representative samples set. As for FAM, although it remained good selection performance, its estimation accuracy was lower than that in previous experiments.

Setup	Method	DEV(%)	PSR(%)	FDR(%)	FN	Time
S1	MLE	120.11	-	-	-	0.01
	EM	284.61	100.00	31.44	7289	0.22
	K-means	106.89	100.00	33.97	7572	0.01
	AM	1.88	100.00	0.04	5002	1.74
	FAM	71.12	99.94	0.00	4997	9.94
S2	MLE	101.87	-	-	-	0.01
	EM	125.12	100.00	34.98	7702	0.16
	K-means	197.81	99.92	34.88	7673	0.01
	AM	4.46	100.00	0.28	5014	2.16
	FAM	68.34	100.00	0.14	5007	9.52
S3	MLE	96.26	-	-	-	0.02
	EM	127.15	100.00	35.12	7707	0.24
	K-means	207.07	99.86	34.81	7660	0.01
	AM	3.32	100.00	0.10	5005	2.13
	FAM	74.17	99.96	0.04	5000	9.67

Case 3.1 Logistic mixture model (2 variables)

In this scenario, the AM algorithm successfully selected a majority of the representative samples with only few noise samples after 4 iterations.



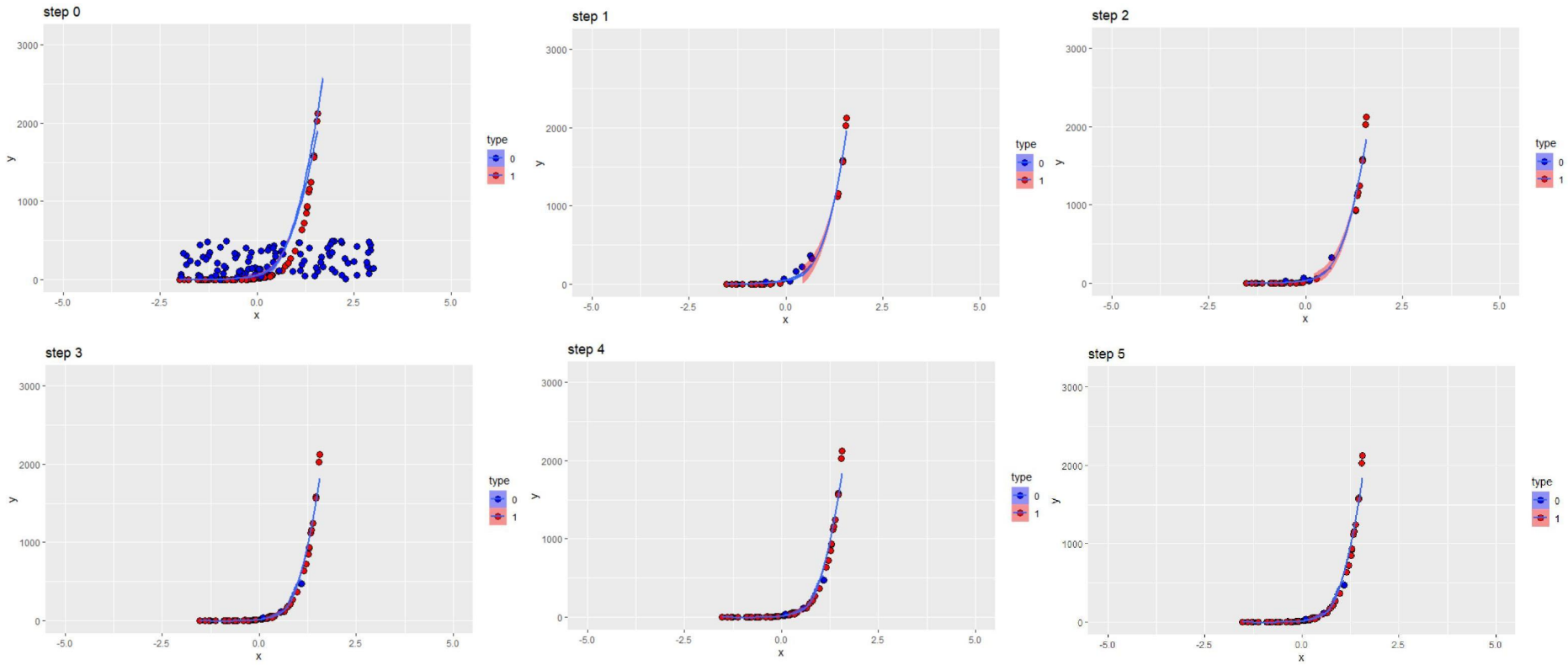
Case 3.2 Logistic mixture model (multiple variables)

EM, AM and FAM performed well in selecting representative samples in terms of relatively high PSR and low FDR. From the perspective of DEV, EM performed more accurately than AM and FAM. The result of EM was different from that in previous models, which was because the implicit model assumption of EM matched the true model setup.

Setup	Method	DEV(%)	PSR(%)	FDR(%)	FN	Time
S1	MLE	81.02	-	-	-	0.03
	EM	19.07	93.46	2.91	4813	2.23
	K-means	83.16	50.96	19.57	3168	0.02
	AM	52.42	92.70	2.73	4765	12.39
	FAM	35.42	91.54	3.01	4717	0.44
S2	MLE	73.08	-	-	-	0.03
	EM	29.84	90.03	4.02	4688	2.48
	K-means	73.15	50.52	20.03	3159	0.02
	AM	56.65	88.30	7.52	4765	29.42
	FAM	39.43	90.46	4.28	4726	0.55
S3	MLE	75.06	-	-	-	0.03
	EM	19.72	91.42	4.29	4776	3.61
	K-means	74.15	50.54	19.98	3159	0.01
	AM	44.46	94.27	6.09	5045	18.31
	FAM	43.84	90.96	15.76	4745	0.54

Case 4.1 Poisson mixture model (2 variables)

In this case, most of the representative samples with few noise samples were selected by AM after 5 iterations.



Case 4.2 Poisson mixture model (multiple variables)

In this experiment, AM and FAM performed better than other methods when estimation and selection were together considered. It's noteworthy that, by the above data generation methods, some noise samples were close to the true model, which meant a small amount of noise samples would be chosen as representative samples.

Setup	Method	DEV(%)	PSR(%)	FDR(%)	FN	Time
S1	MLE	18.87	-	-	-	0.04
	EM	0.01	99.78	15.01	5870	0.71
	K-means	65.90	99.96	50.01	9998	0.02
	AM	0.03	94.78	5.71	5026	0.31
	FAM	1.51	94.26	5.82	5004	14.1
S2	MLE	64.03	-	-	-	0.03
	EM	0.01	99.78	15.21	5884	0.82
	K-means	91.33	99.98	50.00	9999	0.02
	AM	0.01	93.00	5.31	4911	0.39
	FAM	1.41	92.40	5.44	4886	25.54
S3	MLE	10.29	-	-	-	0.05
	EM	12.78	99.32	45.29	9080	2.34
	K-means	43.10	99.94	50.02	9997	0.01
	AM	0.01	92.78	5.28	4844	0.41
	FAM	2.19	91.68	5.31	4895	14.20

Conclusion

According to the experiments in 4 types of mixed datasets, where the explicit models were normal distribution, linear regression, logistic regression and Poisson regression, respectively, as well as 3 different correlation structures among variables, the AM algorithm was robust, outperforming other methods in terms of estimation accuracy and selection consistency in most cases.

Future

- The specific influence of the initial parameter input of the AM.
- A general and data-driven threshold rule.
- In-depth theoretical guarantees.