ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

Regularised maximum-likelihood inference of mixture of experts for regression and clustering

Bao Tuyen Huynh and Faicel Chamroukhi

Normandie Univ, UNICAEN, CNRS, LMNO
14000, Caen, France

Abstract. Variable selection is fundamental to high-dimensional statistical modeling, and is challenging in particular in unsupervised modeling, including mixture models. We propose a regularised maximum-likelihood inference of the Mixture of Experts model which is able to deal with potentially correlated features and encourages sparse models in a potentially high-dimensional scenarios. We develop a hybrid Expectation-Majorization-Maximization (EM/MM) algorithm for model fitting. Unlike state-of-the art regularised ML inference [1, 2], the proposed modeling doesn't require an approximate of the regularisation. The proposed algorithm allows to automatically obtain sparse solutions without thresholding, and includes coordinate descent updates avoiding matrix inversion. An experimental study shows the capability of the algorithm to retrieve sparse solutions and for model fitting in model-based clustering of regression data.

1 Introduction

Mixture of Experts (MoE) models introduced by [3] are widely used in statistics and machine learning. MoE is a fully conditional mixture model where both the mixing proportions, i.e, the gating network, and the components densities, i.e, the experts network, depend on some input covariates. This makes MoE more capable in use than standard unconditional mixture distributions, while having a neural-network interpretation. A general review of the MoE models and their applications can be found in [4]. For continuous data, which we consider here in the context of regression and clustering, MoE usually use Gaussian experts. While the MoE modeling with maximum likelihood inference is widely used, its application in high-dimensional problems is still challenging due to the known problem of the ML estimation (MLE) in such a setting, and hence there is a need to select a subset of the potentially large number of features, that really explain the problem. Indeed, in high-dimensional setting, the features can be correlated, present redundancy, etc, and thus the actual features that explain the problem lie in a low-dimensional space. This can be achieved by regularizing the objective function so that to encourage sparse solutions.

In related mixture models, including mixture of linear regressions (MLR), where the mixing proportions are constant, [5] proposed regularized ML inference, including MIXLASSO, MIXHARD and MIXSCAD and provided some asymptotic properties corresponding to these penalty functions. Another L_1 penalization for MLR models for high-dimensional data was proposed by [6] and an adaptive Lasso penalized estimator with oracle inequality which includes the

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

setting $p \gg n$ was presented. [7] provided an L_1-oracle inequality by a Lasso estimator for finite mixture of Gaussian regression models. This result can be seen as a complementary result to [6], by studying the Lasso for its L_1-regularization properties rather than considering it as a variable selection procedure. This work was extended later in [8] by considering a mixture of multivariate Gaussian regression models. When the set of features can be seen as to be splitted into groups, [9] introduced the two types of penalty functions called MIXGL1 and MIXGL2 for MLR models, based on group Lasso. An MM algorithm [10] version for MLR with Lasso penalty can be found in [11]. Their method allows for an avoidance of matrix operations. In [1], the author extended his MLR regularisation to the MoE setting and provided a root-n consistent and oracle properties for Lasso and SCAD penalties and developed an EM algorithm [12] for fitting the models. However, as we will discuss it in section 2.2, this is based on approximated penalty function, and uses a Newton-Raphson in the updates, which requires matrix inversion.

In this paper, we consider MoE models with regularisation as in [1] and propose a regularised maximum-likelihood inference which doesn't require an approximate of the regularisation. We develop a hybrid Expectation-Majorization-Maximization (EM/MM) algorithm for model fitting. The proposed algorithm allows to automatically select sparse solutions without thresholding, and includes coordinate descent updates avoiding matrix inversion. The rest of this article is organized as follows. In Section 2 we present the regularised maximum-likelihood strategy or the MoE model and the proposed EM/MM algorithm with coordinate descent in section 2.3. An experimental study on simulated and a real-data example are given Section 3. Finally, in Section 4, we draw concluding remarks.

2 Regularised Mixture of experts

2.1 The mixture of experts (MoE) model

Let $(Y_1, \ldots, Y_n)$ where $Y_i \in \mathbb{R}$ be an independent random sample of heterogeneous observations for some given covariate vectors $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, $\boldsymbol{x}_i \in \mathbb{R}^p$, and let $Z_i \in \{1, \ldots, K\}$ be a latent categorical random variable representing the hidden partition of the data into a K clusters. The MoE model decomposes the density of Y given $\boldsymbol{x}$ into a convex sum of K experts densities weighted by a gating network, and can be defined as

$$f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{x}_i;\boldsymbol{w}) f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}_k) \tag{1}$$

where $\pi_k(\boldsymbol{x}_i;\boldsymbol{w}) = e^{w_{k0}+\boldsymbol{x}_i^\top \boldsymbol{w}_k} / \sum_{l=1}^{K} e^{w_{l0}+\boldsymbol{x}_i^\top \boldsymbol{w}_l}$ are the gating softmax functions for $k = 1, \ldots, K$ with $(w_{K0}, \boldsymbol{w}_K) = (0, 0)$ for identifiability. For continuous data like here, we can use Gaussian experts. Furthermore, for data that represent regression functions (i.e for prediction), a common choice is Gaussian regressors for the expert network and we thus have $f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}_k) = \mathcal{N}(y_i; \beta_{k0} + \boldsymbol{x}_i^\top \boldsymbol{\beta}_k, \sigma_k^2)$. The model parameters $\boldsymbol{\theta}$ is commonly estimated by maximizing the log-likelihood $L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\Big[\sum_{k=1}^{K} \pi_k(\boldsymbol{x}_i;\boldsymbol{w}) \mathcal{N}(y_i; \beta_{k0} + \boldsymbol{x}_i^\top \boldsymbol{\beta}_k, \sigma_k^2)\Big]$ by using EM [12, 3].

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

2.2 Regularised maximum-likelihood estimation of the MoE

We propose to infer the MoE model by maximizing a regularised log-likelihood criterion where the regularization combines a Lasso penalty for the experts parameters, and an Elastic-Net like penalty for the gating network, defined by:

$$PL(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \sum_{k=1}^{K} \lambda_k \|\boldsymbol{\beta}_k\|_1 - \sum_{k=1}^{K-1} \gamma_k \|\boldsymbol{w}_k\|_1 - \frac{\rho}{2} \sum_{k=1}^{K-1} \|\boldsymbol{w}_k\|_2^2. \quad (2)$$

A similar strategy were proposed in [1] where the author proposed a regularized ML function like (2) but which is then approximated in the model inference algorithm. The devoloped EM algorithm for fitting the model follows indeed the suggestion of [2] to approximate the penalty function in a some neighbourhood by a local quadratic function. Therefore, the Newton-Raphson method could be used to update parameters in the M-step. The weakness of this design is that once a feature is set to zero, it may never reenter the model at a later stage of the algorithm. To avoid this numerical instability of the algorithm due to the small values of some of the features in the denominator of this approximation, [1] replaced that approximation by an ε-local quadratic function. Unfortunately, these strategies have some drawbacks. First, by approximating the penalty functions with (ε-)quadratic functions, almost surely none of the components will be exactly zero. Hence, a threshold should be considered to declare a coefficient is zero and this threshold affects the degree of sparsity. Secondly, it cannot guarantee the non-decreasing property of the EM algorithm of the penalized objective function. Thus, the convergence of the EM algorithm cannot be ensured. Finally, one has to choose ε, which becomes an additional tuning parameter in practice. Our propoal gives and answer to overcome these limitations.

2.3 Model inference with a block-wise EM/MM algorithm

We propose a block-wise EM algorithm, which integrates an MM algorithm [10] for updating the gating parameters, to monotonically find local maximizers of (2). Instead of maximizing the objective function $Q(\boldsymbol{w}; \boldsymbol{\theta}^{(s)})$ we maximize the surrogate function $G^{(s)}(\boldsymbol{w}|\boldsymbol{w}^m)$ which satisfies $Q(\boldsymbol{w}^{(s)}; \boldsymbol{\theta}^{(s)}) = G^{(s)}(\boldsymbol{w}^{(s)}|\boldsymbol{w}^{(s)}) \leq G^{(s)}(\boldsymbol{w}^{(s+1)}|\boldsymbol{w}^{(s)}) \leq Q(\boldsymbol{w}^{(s+1)}; \boldsymbol{\theta}^{(s)})$ and the MM algorithm forces the objective function to increase, since there is no approximation. We construct the surrogate function using arithmetic-geometric mean inequality. This function is convex and has a separable structure. So to maximize it we just need to use one-dimensional Newton-Raphson algorithm and hence avoid computing the inverse matrix.
Our EM/MM algorithm performs as follows. Given an initial parameter vector $\boldsymbol{\theta}^{(0)}$, it performs, that the $(s+1)^{th}$ iteration:

E-step: Compute the conditional expectation $\tau_{ik}^{(s)}$ of the missing labels

$$\tau_{ik}^{(s)} = \mathbb{P}(Z_i = k|\boldsymbol{x}_i, y_i; \boldsymbol{\theta}^{(s)}) = \pi_k(\boldsymbol{x}_i; \boldsymbol{w}^{(s)}) \mathcal{N}(y_i; \beta_{k0}^{(s)} + \boldsymbol{x}_i^\top \boldsymbol{\beta}_k^{(s)}, \sigma_k^{(s)2}) / f(y_i; \boldsymbol{x}_i, \boldsymbol{\theta}^{(s)}). \quad (3)$$

M-step: Udapte the parameters by maximizing the well-known Q function. The parameters $\boldsymbol{w}$ are updated by maximizing the function

$$Q(\boldsymbol{w}; \boldsymbol{\theta}^{(s)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(s)} \log \pi_k(\boldsymbol{x}_i; \boldsymbol{w}) - \sum_{k=1}^{K-1} \gamma_k \|\boldsymbol{w}_k\|_1 - \frac{\rho}{2} \sum_{k=1}^{K-1} \|\boldsymbol{w}_k\|_2^2; \quad (4)$$

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

For that we propose an MM algorithm and construct the minorizing function $G^{(s)}(\boldsymbol{w}|\boldsymbol{w}^m) = \sum_{i=1}^{n}\sum_{k=1}^{K-1}\tau_{ik}^{(s)}(w_{k0}+\boldsymbol{x}_i^\top\boldsymbol{w}_k) + G_1(\boldsymbol{w}|\boldsymbol{w}^m) - \sum_{k=1}^{K-1}\gamma_k\|\boldsymbol{w}_k\|_1 - \frac{\rho}{2}\sum_{k=1}^{K-1}\|\boldsymbol{w}_k\|_2^2$; where $G_1(\boldsymbol{w}|\boldsymbol{w}^m) = \sum_{i=1}^{n}\Big[-\sum_{k=1}^{K-1}\frac{\pi_k(\boldsymbol{x}_i;\boldsymbol{w}^m)}{p+1}\sum_{j=0}^{p}e^{(p+1)x_{ij}(w_{kj}-w_{kj}^m)} - \log C_i^m + 1 - \frac{1}{C_i^m}\Big]$, with $C_i^m = 1+\sum_{k=1}^{K-1}e^{w_{k0}^m+\boldsymbol{x}_i^\top\boldsymbol{w}_k^m}$, $x_{i0}=1$. We then fix σ_k, and update β_{kj} in

$$Q(\boldsymbol{\beta},\sigma;\boldsymbol{\theta}^{(s)}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(s)}\log\mathcal{N}(y_i;\beta_{k0}+\boldsymbol{x}_i^\top\boldsymbol{\beta}_k,\sigma_k^2) - \sum_{k=1}^{K}\lambda_k\|\boldsymbol{\beta}_k\|_1; \tag{5}$$

using a coordinate descent algorithm, with initial values $(\beta_{k0}^0,\boldsymbol{\beta}_k^0) = (\beta_{k0}^{(s)},\boldsymbol{\beta}_k^{(s)})$. We obtain closed-form coordinate updates which can be computed for each component following the results in [13, sec. 5.4] and are given by

$$\beta_{kj}^{m+1} = \mathcal{S}_{\lambda_k\sigma_k^{(s)2}}\big(\sum_{i=1}^{n}\tau_{ik}^{(s)}r_{ikj}^m x_{ij}\big)\big/\sum_{i=1}^{n}\tau_{ik}^{(s)}x_{ij}^2, \tag{6}$$

with $r_{ikj}^m = y_i - \beta_{k0}^m - \boldsymbol{x}_i^\top\boldsymbol{\beta}_k^m + \beta_{kj}^m x_{ij}$ and $\mathcal{S}_{\lambda_k\sigma_k^{(s)2}}(.)$ is a soft-thresholding operator defined by $[\mathcal{S}_\gamma(u)]_j = \text{sign}(u_j)(|u_j|-\gamma)_+$ and $(x)_+$ a shorthand for $\max\{x,0\}$. For $h\neq j$ we set $\beta_{kh}^{m+1} = \beta_{kh}^m$. At each MM iteration m, β_{k0} is updated by

$$\beta_{k0}^{m+1} = \sum_{i=1}^{n}\tau_{ik}^{(s)}(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}_k^{m+1})\big/\sum_{i=1}^{n}\tau_{ik}^{(s)}. \tag{7}$$

Then, we take $(w_{k0}^{(s+2)},\boldsymbol{w}_k^{(s+2)}) = (w_{k0}^{(s+1)},\boldsymbol{w}_k^{(s+1)})$, $(\beta_{k0}^{(s+2)},\boldsymbol{\beta}_k^{(s+2)}) = (\beta_{k0}^{(s+1)},\boldsymbol{\beta}_k^{(s+1)})$, rerun the E-step, and update σ_k^2 as follows

$$\sigma_k^{2(s+2)} = \sum_{i=1}^{n}\tau_{ik}^{(s+1)}(y_i - \beta_{k0}^{(s+2)} - \boldsymbol{x}_i^\top\boldsymbol{\beta}_k^{(s+2)})^2\big/\sum_{i=1}^{n}\tau_{ik}^{(s+1)}. \tag{8}$$

The algorithm is iterated until the change in $PL(\boldsymbol{\theta})$ is small enough. The proposed algorithm, at each iteration, clearly guarantees to improve the optimised penalised log-likelihood function (2); Also we can directly get zero coefficients without any thresholding like in [1, 14].

3 Experimental study

In this section, a simulation is performed to assess the performance of the regularized MoE. We consider covariate variables $\boldsymbol{x}$ generated from a multivariate Gaussian distribution with zero mean and correlation defined by $\text{corr}(x_{ij},x_{ij'}) = 0.5^{|j-j'|}$. The response Y is generated from a normal MoE model with $K=2$, $p=6$ and $n=300$. The true parameters and their estimates according to three different methods are given in Table 2. The results are averaged on 100 different data sets. We evaluate the performance of the penalized MoE compared with standard non-penalized MoE, and MoE with ridge penalty function for the gates by considering three different aspects: i) sparcity by computing the sensitivity and the specificity (i.e, proportion of correctly estimated zero coefficients and nonzero coefficients), ii) parameters estimation by computing the MSE of parameter estimates, and iii) clustering performance via the correct classification rate. The sensitivity/specificity results and the correct classification rates are given in Table 1 and the MSE are given in Table 2. We can clearly see the algorithm performs very well to retrieve the actual sparse support; the sensitivity and

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

Method	Sensitivity/Specificity						Correct classification rate
	Expert 1		Expert 2		Gating		
	S_1	S_2	S_1	S_2	S_1	S_2	
MoE	0.000	1.000	0.000	1.000	0.000	1.000	$89.57\%_{(1.65\%)}$
L_2	0.000	1.000	0.000	1.000	0.000	1.000	$89.62\%_{(1.63\%)}$
Lasso+L_2	**0.720**	**1.000**	**0.776**	**1.000**	**0.815**	0.615	$87.76\%_{(2.19\%)}$

Table 1: Sensitivity (S_1)/Specificity (S_2) and clustering error summaries.

Component	True value	Mean square error		
		MoE	L_2	Lasso+L_2
Expert 1	0	$0.0093_{(.015)}$	$0.0094_{(.015)}$	$\mathbf{0.0092_{(.015)}}$
	0	$0.0112_{(.016)}$	$0.0114_{(.017)}$	$\mathbf{0.0018_{(.005)}}$
	1.5	$\mathbf{0.0098_{(.014)}}$	$0.0098_{(.015)}$	$0.0106_{(.012)}$
	0	$0.0099_{(.016)}$	$0.0099_{(.016)}$	$\mathbf{0.0019_{(.005)}}$
	0	$0.0108_{(.015)}$	$0.0109_{(.016)}$	$\mathbf{0.0010_{(.004)}}$
	0	$0.0094_{(.014)}$	$0.0094_{(.014)}$	$\mathbf{0.0021_{(.007)}}$
	1	$\mathbf{0.0081_{(.012)}}$	$0.0082_{(.012)}$	$0.0117_{(.015)}$
Expert 2	0	$0.0342_{(.042)}$	$\mathbf{0.0338_{(.042)}}$	$0.0585_{(.072)}$
	1	$0.0355_{(.044)}$	$\mathbf{0.0354_{(.044)}}$	$0.1583_{(.157)}$
	−1.5	$0.0222_{(.028)}$	$\mathbf{0.0221_{(.028)}}$	$0.1034_{(.098)}$
	0	$0.0253_{(.032)}$	$0.0252_{(.031)}$	$\mathbf{0.0033_{(.013)}}$
	0	$0.0296_{(.049)}$	$0.0294_{(.049)}$	$\mathbf{0.0039_{(.019)}}$
	2	$\mathbf{0.0286_{(.040)}}$	$0.0287_{(.040)}$	$0.0432_{(.056)}$
	0	$0.0195_{(.029)}$	$0.0195_{(.029)}$	$\mathbf{0.0043_{(.017)}}$
Gating	1	$0.1379_{(.213)}$	$\mathbf{0.0936_{(.126)}}$	$0.2315_{(.240)}$
	2	$0.2650_{(.471)}$	$\mathbf{0.1225_{(.157)}}$	$0.8123_{(.792)}$
	0	$0.0825_{(.116)}$	$0.0641_{(.086)}$	$\mathbf{0.0404_{(.032)}}$
	0	$0.1466_{(.302)}$	$0.1052_{(.196)}$	$\mathbf{0.0501_{(.050)}}$
	−1	$0.1875_{(.263)}$	$\mathbf{0.1129_{(.148)}}$	$0.7703_{(.760)}$
	0	$0.1101_{(.217)}$	$0.0803_{(.164)}$	$\mathbf{0.0656_{(.066)}}$
	0	$0.0806_{(.121)}$	$0.0610_{(.095)}$	$\mathbf{0.0175_{(.018)}}$
σ	1	$0.0033_{(.004)}$	$0.0035_{(.004)}$	$\mathbf{0.0027_{(.003)}}$

Table 2: MSE of parameter estimates for MoE, L_2, Lasso+L_2.

specificity results are better for the proposed Lasso+L_2 regularisation. The specificity for the second gating function is less than for the two alternatives can be attributed to the fact that the model is dedicated to sparse models, rather than non-zero coefficients. The same thing can be observed for the MSE which means that the algorithm can also perform density estimation with a reasonable loss of information due to the bias induced by the regularisation. The same thing can be said for clustering performance if the objective is to partition the data into clusters. Finally, note that, for choosing the tuning parameters and number of components we can use the modified BIC with a grid search scheme as in [6]. The modified BIC performs reasonably well in our simulation.

We now analyze a real data set consisting of baseball salaries from the Journal of Statistics Education (see also [5]). We compare our results with the non-penalized MoE models in different criteria: the average mean square error (MSE) between observation values of the response variable and the predicted values of this variable; we also consider the correlation of these values. [5] used this data set in the analysis, which included an addition of 16 interaction features, making in total 32 predictors. The columns of $\boldsymbol{X}$

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

were standardised to have mean 0 and variance 1. Table 3 shows the results in terms of MSE, and R^2. We also show the obtained number of zero coefficients in the experts and the gating network. These results clearly suggest that the proposed algorithm with the Lasso+L_2 penalty also shrinks some parameters to zero and have an acceptable results when comparing with MoE.

	R^2	MSE	Exp.1	Exp.2	Gating network
MoE	0.8099	$0.2625_{(.758)}$	0	0	0
Lasso+L_2	0.7971	$0.2880_{(.649)}$	22	20	19

Table 3: Results for Baseball salaries data set.

4 Conclusion and future work

We proposed a regularised ML inference for the MoE model which encourages sparcity, and developed a blockwise EM-MM algorithm which monotonically maximise this regularised objective towards at least a local maximum, while avoiding standard using approximations. The algorithm includes one-by-one parameter updates using coordinate descent avoiding matrix inversion. The results on both simulations and real-data (including further experiments which can not be included in the paper for a lack of space) confirm the effectiveness of the proposal. The next step is to perform additional model selection experiments and to consider the multivariate case and the full high-dimensional setting.

References

[1] A. Khalili. New estimation and feature selection methods in mixture-of-experts models. *Canadian Journal of Statistics*, 38(4):519–539, 2010.

[2] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[3] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[4] S. E. Yuksel, J. N. W., and P. D. Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.

[5] A. Khalili and J. Chen. Variable selection in finite mixture of regression models. *Journal of the American Statistical association*, 102(479):1025–1038, 2007.

[6] N. Städler, P. Bühlmann, and S. Van De Geer. l1-penalization for mixture regression models. *Test*, 19(2):209–256, 2010.

[7] C. Meynet. An ℓ_1-oracle inequality for the lasso in finite mixture gaussian regression models. *ESAIM: Probability and Statistics*, 17:650–671, 2013.

[8] E. Devijver. An ℓ_1-oracle inequality for the lasso in multivariate finite mixture of multivariate gaussian regression models. *ESAIM: Probability and Statistics*, 19:649–670, 2015.

[9] F. K. Hui, D. I. Warton, S. D. Foster, et al. Multi-species distribution modeling using penalized mixture of regressions. *The Annals of Applied Statistics*, 9(2):866–882, 2015.

[10] K. Lange. *Optimization (2nd edition)*. Springer, 2013.

[11] L. R. Lloyd-Jones, H. D. Nguyen, and G. J. McLachlan. A globally convergent algorithm for lasso-penalized mixture of linear regression models. *arXiv:1603.08326*, 2016.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. of the royal statistical society. Series B*, pages 1–38, 1977.

[13] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Taylor & Francis, 2015.

[14] D. R. Hunter and R. Li. Variable selection using mm algorithms. *Annals of statistics*, 33(4):1617, 2005.