ISSN 1068-3623, Journal of Contemporary Mathematical Analysis (Armenian Academy of Sciences), 2019, Vol. 54, No. 5, pp. 302–312. © Allerton Press, Inc., 2019.
Russian Text © The Author(s), 2019, published in Izvestiya Natsional'noi Akademii Nauk Armenii, Matematika, 2019, No. 5, pp. 53-69.

PROBABILITY THEORY AND MATHEMATICAL STATISTICS

Alternating Least Squares in Generalized Linear Models

A. G. Minasyan[1]*

[1] *Yerevan State University, Yerevan, Armenia*

Received October 26, 2018; Revised February 1, 2019; Accepted April 25, 2019

Abstract—We derived a convergence result for a sequential procedure known as alternating maximization (minimization) to the maximum likelihood estimator for a pretty large family of models - Generalized Linear Models. Alternating procedure for linear regression becomes to the well-known algorithm of Alternating Least Squares, because of the quadraticity of log-likelihood function $L(\boldsymbol{v})$. In Generalized Linear Models framework we lose quadraticity of $L(\boldsymbol{v})$, but still have concavity due to the fact that error-distribution is from exponential family. Concentration property makes the Taylor approximation of $L(\boldsymbol{v})$ up to the second order accurate and makes possible the use of alternating minimization (maximization) technique. Examples and experiments confirm convergence result followed by the discussion of the importance of initial guess.

MSC2010 numbers : 62L12, 62F12, 49M05

DOI: 10.3103/S1068362319050078

Keywords: *Generalized linear model; exponential family; alternating least squares.*

1. INTRODUCTION

Many statistical tasks can be viewed as problems of semi-parametric estimation when the unknown data distribution is described by a high or infinite-dimensional parameter while the target is of low dimension. Typical examples are provided by functional estimation, estimation of a function at a point, or simply estimation of a given sub-vector of the parameter vector. The classical statistical theory provides a general solution to this problem: estimate the full parameter vector by the maximum likelihood method and project the obtained estimate onto the target subspace. This approach is known as *profile maximum likelihood* and it appears to be *semi-parametrically efficient* under some mild regularity conditions, which are satisfied for generalized linear models (GLMs) . For more general cases, for example, in M-estimation framework, these technical conditions should be introduced separately and checked whether they are fulfilled or not. We refer to [6], [8], and the book by Kosorok [10] for a detailed presentation.

The present paper revisits the problem of profile semi-parametric estimation (see [4], [5], and references therein). One issue that is worth mentioning is the model misspecification. In most of the cases of practical problems, it is unrealistic to expect that the model assumptions will be exactly fulfilled, even if some rich nonparametric family is used. This means that the true data distribution $\mathbb{P}$ does not belong to considered parametric family, in our case that is exponential family. Applicability of the general semi-parametric theory in such cases is questionable. An important feature of the presented approach is that it equally applies under a possible model misspecification.

Consider the following statistical model that assumes that the unknown data distribution $\mathbb{P}$ belongs to a given parametric family $(\mathbb{P}_{\mathbf{v}})$:

$$\mathcal{Y} \sim \mathbb{P} = \mathbb{P}_{\mathbf{v}^*} \in (\mathbb{P}_{\mathbf{v}}, \mathbf{v} \in \Theta), \tag{1.1}$$

where Θ is some high dimensional or even infinite-dimensional parametric space.

* E-mail: arsh.minasyan@gmail.com

The maximum likelihood approach in the parametric estimation suggests to estimate the whole parameter vector $\mathbf{v}$ by maximizing the corresponding log-likelihood function:

$$L(\mathbf{v}) = \log \frac{d\mathbb{P}_{\mathbf{v}}}{d\mu_0}$$

for some dominating measure μ_0. So, the maximum likelihood estimator $\tilde{\mathbf{v}}$ of the parameter vector $\mathbf{v}$ is defined to be

$$\tilde{\mathbf{v}} \stackrel{\text{def}}{=} \arg\max_{\mathbf{v}\in\Theta} L(\mathbf{v}). \tag{1.2}$$

Our study admits a model specification $\mathbb{P} \notin (\mathbb{P}_{\mathbf{v}}, \mathbf{v} \in \Theta)$. Equivalently, one can say that $L(\mathbf{v})$ is the *quasi log-likelihood function* on Θ. The *target* value $\mathbf{v}^*$ of the parameter $\mathbf{v}$ can be defined by

$$\mathbf{v}^* \stackrel{\text{def}}{=} \operatorname*{argmax}_{\mathbf{v}\in\Theta} \mathbb{E}L(\mathbf{v}). \tag{1.3}$$

Under model misspecification, $\mathbf{v}^*$ defines the best parametric fit to $\mathbb{P}$ by the considered family. For the similar results we refer to [1]. A. Kneip [9] first started the work in this direction by introducing ordered linear functionals. For general results on alternating maximization (minimization) we refer to [2].

The main point of this paper is to show that the alternating method gives only a little gain, if any, in the complexity of the optimum point computation for linear models (see [12]), meanwhile, for non-linear models the gain is pretty sensible. For non-linear models, in most of the cases, the closed form of solutions can not be obtained, moreover, in some cases, the numerical solutions of first order conditions could be very hard to implement in the full parameter dimension. The technique, known as alternating maximization (minimization) [3] method, helps in such situations and gives the estimation of the parameter vector with adequate time complexity.

The model that we consider has a parameter $\mathbf{v}$ of dimension $p + q$, where p is the dimension of the *target* parameter and q is the dimension of the *nuisance* parameter. Usually p is not large, because we also care about tractability and interpretability of our model, while q can be very large. Notice that the *nuisance* parameter cannot be ignored or excluded from the considered model. The main difficulties concerning direct computations occur in high dimensions, that is, in the cases where $p + q$ is large enough, because it make impossible to invert the matrix of size $(p + q) \times (p + q)$.

Observe that the alternating maximization procedure can be considered as a special case of the expectation maximization algorithm (EM algorithm). The EM algorithm is a popular algorithm first derived in [7]. Further on a number of modifications and extensions of this algorithm have been obtained. In [7] it was also described how the EM algorithm can be implemented in different fields and give fruitful results. We refer to the book [11] for the brief introduction to the development of EM algorithm. Also, we restrict ourselves, by quoting the well-known convergence result from [13], which is still state of the art in most of the cases. Unfortunately, this convergence result, as the most convergence results on these iterative procedures, only ensures convergence to some set of local maximizers or fix-points of the procedure. In this paper, we consider one of the special cases where it is possible to prove the actual convergence of the method.

The reminder of the paper is structured as follows. Section 2 contains some preliminaries about the generalized linear models (GLMs) and the exponential family of distributions. Section 3 contains the main results of the paper. Section 4 illustrates how the algorithm of alternating least squares (ALS) works and confirms the obtained theoretical results by an example using both real and simulated data.

2. INTRODUCTION TO GENERALIZED LINEAR MODELS

In this section, we introduce the class of generalized linear models (GLMs) from a slightly different viewpoint, for which in Section 3 we will develop the alternating least squares method.

Let Y_i be independent random vectors and let $X_i \in \mathbb{R}^p$ be regressors. We assume that $Y_i \sim P_i \in (\mathcal{P}_{\mathbf{v}})$, meaning that there exists v_i such that $P_i = P_{v_i}$, where $(\mathcal{P}_{\mathbf{v}})$ stands for the exponential family of distributions with canonical parameter. (The exponential family of of distributions will be discussed in details further in this section). Then the generalized linear model has the following representation

$Y_i \sim P_{v(X_i)}$. In the case of Gaussian distributions, we get the following form $Y_i = v(X_i) + \varepsilon_i$ with arbitrary function $\mathbf{v}(\cdot)$.

The function $\mathbf{v}(x)$ can be represented in the following form: $v(x) = \sum_{j=1}^{\infty} \theta_j \psi_j(x)$. Then our linear parametric assumption can be written as follows:

$$v(x) = \sum_{j=1}^{p+q} \theta_j \psi_j(x) = \sum_{j=1}^{p} \theta_j \psi_j(x) + \sum_{j=p+1}^{p+q} \theta_j \psi_j(x) \text{ for given basis } \psi_j(\cdot). \tag{2.1}$$

We will mostly discuss the case of finite p and q. Denote $\eta_i = \theta_{p+i}$ and for a column-vector $\eta = (\eta_1, \dots, \eta_q)^T \in \mathbb{R}^q$ consider the model:

$$Y_i \sim P_{v_i}, \quad v_i = \Psi_i^T \theta + \Phi_i^T \eta, \tag{2.2}$$

where $\Psi_i = (\psi_1(x), ..., \psi_p(x))^T \in \mathbb{R}^p$, $\Phi_i = (\psi_{p+1}(x), ..., \psi_{p+q}(x))^T \in \mathbb{R}^q$ and $\theta \in \mathbb{R}^p$, $\eta \in \mathbb{R}^q$.

Denote the log-likelihood function by $L(\theta, \eta)$ or by $L(\mathbf{v})$, where $\mathbf{v} \stackrel{\text{def}}{=} (\theta, \eta)$. For the ease of representation sometimes we use $\mathbf{v}$ instead of (θ, η) and vice versa. Then for log-likelihood we have

$$L(\theta, \eta) \stackrel{\text{def}}{=} \log \frac{dP_{\mathbf{v}}}{d\mu_0^n}(\mathcal{Y}) = \sum_{i=1}^{n} \left(v_i Y_i - g(\nu_i)\right) = \sum_{i=1}^{n} \left(\Psi_i^T \theta Y_i + \Phi_i^T \eta Y_i - g(\Psi_i^T \theta + \Phi_i^T \eta)\right),$$

which follows from the properties of exponential family (see (2.10)). The function $g(\cdot)$ is derived from (2.8) when the distribution is fixed. Equivalently, we can write

$$L(\theta, \eta) = S^T \theta + R^T \eta - A(\theta, \eta), \tag{2.3}$$

where

$$S \stackrel{\text{def}}{=} \sum_{i=1}^{n} Y_i \Psi_i \in \mathbb{R}^p, \quad R \stackrel{\text{def}}{=} \sum_{i=1}^{n} Y_i \Phi_i \in \mathbb{R}^q, \quad A(\theta, \eta) \stackrel{\text{def}}{=} \sum_{i=1}^{n} g(\Psi_i^T \theta + \Phi_i^T \eta).$$

Observe that in terms of the parameter $\mathbf{v}$ the log-likelihood can be written as $L(\mathbf{v}) = \Upsilon^T \mathbf{v} - A(\mathbf{v})$, where $\Upsilon = \binom{S}{R}^T \in \mathbb{R}^{p+q}$.

The Fisher information matrix is defined by $\mathcal{D}^2 \stackrel{\text{def}}{=} -\nabla^2 \mathbb{E} L(\mathbf{v}^*) = F(\mathbf{v}^*)$, where $\mathbf{v} = \binom{\theta}{\eta}^T \in \mathbb{R}^{p+q}$ and ∇ is the operator of differentiation, and the Hessian matrix at point $\mathbf{v}$ is defined by

$$\mathbb{F}(\mathbf{v}) = -\nabla^2 \mathbb{E} L(\mathbf{v}) = \begin{pmatrix} \mathbb{F}_{\theta\theta}(\mathbf{v}) & \mathbb{F}_{\theta\eta}(\mathbf{v}) \\ \mathbb{F}_{\eta\theta}(\mathbf{v}) & \mathbb{F}_{\eta\eta}(\mathbf{v}) \end{pmatrix}.$$

Later we show that the function $g(\cdot)$ is convex, which imply that the Fisher information matrix $\mathbb{F}$ is positive definite. Also define the score vector $\nabla \stackrel{\text{def}}{=} \nabla L(\mathbf{v}^*)$, as well as the standardized score vector $\breve{\xi}$ in the form $\breve{\xi} = \mathcal{D}^{-1} \nabla$

We rewrite the formulas (1.2) and (1.3) as follows:

$$\tilde{\mathbf{v}} = (\tilde{\theta}, \tilde{\eta}) = \arg\max_{\theta, \eta} L(\theta, \eta), \quad \mathbf{v}^* = (\theta^*, \eta^*) = \arg\max_{\theta, \eta} \mathbb{E} L(\theta, \eta), \tag{2.4}$$

and observe that the parameter $\tilde{\mathbf{v}} = (\tilde{\theta}, \tilde{\eta})$ is data-dependent, and hence is random, while $\mathbf{v}^* = (\theta^*, \eta^*)$ is the true value of the parameter, which is non-random. In fact, the true distribution of Y is unknown, nevertheless we make a parametric assumption about the class of distributions.

From the definition of $\mathbf{v}^*$ it follows that $\nabla \mathbb{E} L(\mathbf{v}^*) = 0$, which yields

$$\mathbb{E} \begin{pmatrix} S \\ R \end{pmatrix}^T = \nabla A(\theta^*, \eta^*), \quad \text{or equivalently } \mathbb{E}\Upsilon = \nabla A(\mathbf{v}^*).$$

An important property of exponential family is that the stochastic component $\zeta(\theta,\eta)$ of the log-likelihood function is linear in θ and η. Denoting $\varepsilon_i = Y_i - \mathbb{E}Y_i$ and $\zeta = L - \mathbb{E}L$, we can write

$$\zeta(\theta,\eta) = (S^T - \mathbb{E}S^T)\theta + (R^T - \mathbb{E}R^T)\eta = \sum_{i=1}^{n} \varepsilon_i(\Psi_i^T\theta + \Phi_i^T\eta),$$

$$\nabla\zeta(\theta,\eta) = \Big(S - \mathbb{E}S \;\; R - \mathbb{E}R\Big) = \Big(\textstyle\sum_{i=1}^n \varepsilon_i\Psi_i \;\; \sum_{i=1}^n \varepsilon_i\Phi_i\Big).$$

Now consider the following elliptic set:

$$\Omega_\circ(r) \stackrel{\text{def}}{=} \{\mathbf{v} : \|\mathcal{D}(\mathbf{v} - \mathbf{v}^*)\| \le r\}, \tag{2.5}$$

which is called a local vicinity of $\mathbf{v}^*$ for $\mathcal{D}^2 = -\nabla^2\mathbb{E}L(\mathbf{v}^*) = \mathbb{F}(\mathbf{v}^*)$ and $\mathcal{V}^2 \stackrel{\text{def}}{=} \mathbb{C}ov(\nabla L(\mathbf{v}^*))$. The covariance matrix we write in the block form:

$$\mathcal{V}^2 = \begin{pmatrix} V^2 & E \\ E^T & Q^2 \end{pmatrix}. \tag{2.6}$$

Also, the Fisher information matrix $\mathbb{F}(\mathbf{v}^*) = -\nabla^2\mathbb{E}L(\mathbf{v}^*)$ we write in the block form:

$$\mathbb{F}(\mathbf{v}) = \begin{pmatrix} \mathbb{F}_{\theta\theta}(\mathbf{v}) & \mathbb{F}_{\theta\eta}(\mathbf{v}) \\ \mathbb{F}_{\eta\theta}(\mathbf{v}) & \mathbb{F}_{\eta\eta}(\mathbf{v}) \end{pmatrix}. \tag{2.7}$$

For a central point $\mathbf{v}^*$ the decomposition can be written in the following form:

$$\mathcal{D}^2 = \mathbb{F}(\mathbf{v}^*) = \begin{pmatrix} D^2 & A \\ A^T & H^2 \end{pmatrix},$$

where $D^2 = \mathbb{F}_{\theta\theta}(\mathbf{v}^*)$, $A = \mathbb{F}_{\theta\eta}(\mathbf{v}^*)$ and $H^2 = \mathbb{F}_{\eta\eta}(\mathbf{v}^*)$. Also, decompose the score vector $\nabla \stackrel{\text{def}}{=} \nabla L(\mathbf{v}^*)$ in the form $\nabla = \binom{\nabla_\theta}{\nabla_\eta}$.

2.1. Exponential Family with Canonical Parameter (EFc)

In this subsection we extend the scope of our modeling toolbox to accommodate a variety of additional data types, including counts and rates. We introduce the exponential family of distributions. It is worth to note that the exponential family is sufficiently broad class of distributions and covers the vast majority of possible reasonable error distributions. It includes the Gaussian, binomial, Poisson, gamma, multinomial, and many other important distributions. The simplest examples of distributions that are not members of the exponential family are the uniform distribution and the Student's t-distribution. The beauty and elegance of exponential family is that the (log) likelihood functions for distributions from this family have simple forms and can be written explicitly.

In general, we say that a random variable X with probability density function $f(\cdot)$ is from EFc if $f(\cdot)$ can be expressed in the following form:

$$f(x|\eta) = h(x)\exp\{\eta^T T(x) - A(\eta)\} \tag{2.8}$$

where the vector parameter η is referred to as the canonical parameter, $T(X)$ is a *jointly sufficient statistic*, and $A(\cdot)$ is a *link function*.

As stated above there is a huge number of famous distributions the probability density functions of which could be expressed in the form (2.8).

2.1.1. GLM with error distribution from EFc Let Y be a response vector of two sets of factors: Ψ and Φ. Consider the following model

$$Y \sim P \in (\mathcal{P}_{\mathbf{v}})_{v\in\mathbb{R}} \ll \mu_0, \tag{2.9}$$

where $(\mathcal{P}_{\mathbf{v}})_{v\in\mathbb{R}}$ is an exponential family with canonical parameter (EFc) and μ_0 is some dominating measure. Hence, we have

$$\log \frac{dP_{\mathbf{v}}}{d\mu_0} = y \cdot v - g(\mathbf{v}) \text{ for some function } g. \tag{2.10}$$

Lemma 2.1. *Let $(\mathcal{P}_\nu)$ be EFc. Then the following equalities hold:*

$$\mathbb{E}_\nu Y = g'(\nu), \;\; \mathbb{V}ar_\nu(Y) = \mathbb{E}_\nu[Y - g'(\nu)]^2 = g''(\nu), \tag{2.11}$$

and hence the function $g(\cdot)$ is convex.

Remark 2.1. The result of Lemma 2.1, stated for univariate functions $g(\cdot)$, is also true for multivariate functions. The only difference will be that in (2.11) instead of variance there will be the covariance matrix, which is positive definite.

2.2. Concentration

Recall that the following properties hold true for GLMs: the stochastic component $\zeta(\theta, \eta)$ of the log-likelihood function is linear in both θ and η, and the deterministic part $\mathbb{E}L(\mathbf{v})$ is a concave function of $\mathbf{v}$. Consider the elliptic set defined in (2.5). In [12] it is proved that there is an elliptic set $\Omega_\circ(\mathrm{r})$ around the oracle $\mathbf{v}$ such that $\Omega_\circ(\mathrm{r})$ contains the estimator $\tilde{\mathbf{v}}$ with high probability. In what follows we assume a sufficiently large value of $\mathbf{x}$ to be fixed, which determines the level of overwhelming probability: a generic random set $\Omega_0(\mathbf{x})$ is of dominating probability if $\mathbb{P}(\Omega_0(\mathbf{x})) \geq 1 - \mathcal{C}e^{-\mathbf{x}}$, for some absolute constant $\mathcal{C} > 0$.

Observe that the value of $\mathbf{x}$ may depend on sample size n and tends to infinity as $n \to \infty$. The possible choices of $\mathbf{x}$ are $\mathbf{x} \asymp n^{1/2}$ and $\mathbf{x} \asymp \log n$, which would entail that $\mathbb{P}(\Omega_0(\mathbf{x})) \geq 1 - \mathcal{C}/n$. We only require that the sequence $\{x_n\}$ is not too large, more precisely, we assume that $\mathbf{x} \leq \mathbf{x}_c \overset{\text{def}}{\asymp} n^{1/2}$.

The way of construction and the exact probability of desired elliptic set is a very technical issue, and hence is omitted. All the results obtained below are considered to be true on the random set $\Omega_\circ(\mathbf{x})$ of high probability.

2.3. Local Quadratic Approximation of the Log-likelihood function

In this subsection we approximate our log-likelihood function $L(\mathbf{v}) := L(\theta, \eta)$ by a quadratic form in a local elliptic set of optimum $\mathbf{v}^*$. Put $L(\mathbf{v}_1, \mathbf{v}_2) = L(\mathbf{v}_1) - L(\mathbf{v}_2)$, and recall that $\tilde{\mathbf{v}}$ is a data-dependent maximum of the likelihood function $L(\cdot)$, more precisely, we have

$$\tilde{\mathbf{v}} = \arg\max_{\mathbf{v}} L(\mathbf{v}), \quad \mathbf{v}^* = \arg\max_{\mathbf{v}} \mathbb{E}L(\mathbf{v}).$$

Then, using Taylor expansion for function $L(\cdot)$ up to second order around the point $\mathbf{v}^*$, we obtain

$$L(\mathbf{v}, \mathbf{v}^*) = \nabla L(\mathbf{v}^*)(\mathbf{v} - \mathbf{v}^*) - \frac{1}{2}\|\mathcal{D}^2(\mathbf{v} - \mathbf{v}^*)\|^2 + \alpha'(\mathbf{v}, \mathbf{v}^*), \tag{2.12}$$

where $\alpha'(\cdot)$ is defined in (2.12).

Similarly, we approximate our function $L(\mathbf{v})$ around $\tilde{\mathbf{v}}$ and use the fact that $\nabla L(\tilde{\mathbf{v}}) = 0$, to obtain the following expression:

$$L(\mathbf{v}, \tilde{\mathbf{v}}) = -\frac{1}{2}\|\mathcal{D}^2(\mathbf{v} - \tilde{\mathbf{v}})\|^2 + \alpha(\mathbf{v}, \tilde{\mathbf{v}}). \tag{2.13}$$

Remark 2.2. In order to apply Taylor expansion we use the concentration property, which is crucial, because it allows to claim that this expansion will adequately represent the original function $L(\cdot)$. This idea comes from a well-known fact that every concave function is quadratic around its maximizer, and hence it can be expanded into Taylor series up to the second order without any significant approximation error.

3. ALTERNATING LEAST SQUARES METHOD FOR GLMS

In this section, the known results obtained for linear models by using the usual least squares method, we extend to the case of a generalized linear models (GLMs). The question is non-trivial because for many models we can not obtain the estimators directly and in closed forms, and also, we loose the "quadraticity" of the log-likelihood function. Instead, we will approximate our log-likelihood function by a quadratic form in a set where the measure concentration is observed.

It is worth to note that the generalized linear models are frequently used in many areas and applications including categorical data analysis, classification problems, Poisson and binary regressions, statistical learning, density estimation, etc. GLMs are popular in economics, where sometimes the linear regression model is not enough to fully explain the phenomenon and deeply analyze the situation. The use of GLMs partially solves this problem, but it makes the problem from the semiparametric family, which is the payment for more reliable and accurate results. The restrictions of GLMs are not very strict, so that they can easily be adopted in various models.

Here we will discuss the alternating least squares method for the (quasi) likelihood function obtained in the previous section. Note that, in general, the alternating maximization (minimization) procedure is used in the cases where the direct full dimension computations are not feasible or simply are very difficult to implement.

Recall the log-likelihood function $\mathcal{L}(\mathbf{v})$, where the vector $\mathbf{v} = (\theta, \eta)$ can be decomposed to the *target* parameter θ and *nuisance* parameter η. The method of alternating maximization is an iterative procedure starting from an initial value $\mathbf{v}^\circ \in \mathbb{R}^{p+q}$ and updating iteratively in the way shown below

$$\begin{aligned}\tilde{\mathbf{v}}_{k,k} &\stackrel{\text{def}}{=} (\hat{\theta}_k, \hat{\eta}_k) = \left(\hat{\theta}_k, \operatorname*{argmax}_{\eta\in\mathbb{R}^q} L(\hat{\theta}_k, \eta)\right),\\ \tilde{\mathbf{v}}_{k+1,k} &\stackrel{\text{def}}{=} (\hat{\theta}_{k+1}, \hat{\eta}_k) = \left(\operatorname*{argmax}_{\theta\in\mathbb{R}^p} L(\theta, \hat{\eta}_k), \hat{\eta}_k\right). \end{aligned} \tag{3.1}$$

In this section we will try to answer the following questions that naturally arise in connection with the above described iterative procedure. Does the sequence $\left(\hat{\theta}_k\right)$ converge? And, if the answer is yes, what is the convergence rate? What are the conditions under which the sequence actually converges to the global maximizer $\tilde{\mathbf{v}}$?

3.1. Convergence to the Maximum Likelihood Estimator

The following theorem is one of the main results about the convergence of alternating least squares for generalized linear models.

Theorem 3.1. *Let the model be given by* (2.2), $\mathbf{v} = (\theta, \eta) \in \mathbb{R}^{p+q}$, *and let* $L(\mathbf{v})$ *be the log-likelihood function defined in* (2.3). *Let* $\mathcal{D}^2 = -\nabla^2\mathbb{E}L(\mathbf{v}^*)$ *be the Hessian matrix for* $L(\mathbf{v})$ *with the following block-matrix representation* $\begin{pmatrix} D^2 & A \\ A^T & H^2 \end{pmatrix}$ *at point* $\tilde{\mathbf{v}}$. *Assume that* $\rho \stackrel{\text{def}}{=} \|D^{-1}AH^{-1}\|^2 < 1$ *(here and below the norm* $\|\cdot\|$ *with a matrix argument denotes the spectral norm of the matrix), and the condition* $\|\mathcal{D}^{-1}\nabla^2\mathbb{E}L(\mathbf{v})\mathcal{D}^{-1} - I_{p+q}\| \le \delta(r) \le \delta$ *holds, where* I_{p+q} *is an identity matrix of order* $p+q$.

Then the sequence of estimators $\left(\hat{\theta}_k\right)$ *obtained by alternating least squares method converges to* $\tilde{\theta} = \Pi_\theta\tilde{\mathbf{v}} \stackrel{\text{def}}{=} \Pi_\theta \arg\max_{\mathbf{v}} L(\mathbf{v})$, *where* Π_θ *is the projection of the full vector to its sub-vector* θ.

Remark 3.1. The notation $\tilde{v}_{k(+1),k}$ is used in the cases where the results are true both for $\tilde{v}_{k,k}$ and for $\tilde{v}_{k+1,k}$.

Proof of Theorem 3.1. Writing the equality (2.13) in terms of θ and η we get

$$L(\mathbf{v},\tilde{\mathbf{v}}) = -\frac{1}{2}\|D^2(\theta-\tilde{\theta})\|^2 - \frac{1}{2}\|H^2(\eta-\tilde{\eta})\|^2 - (\theta-\tilde{\theta})^T A(\eta-\tilde{\eta}) + \alpha(\mathbf{v},\tilde{\mathbf{v}}). \tag{3.2}$$

Starting with an initial guess θ° and using the procedure described in (3.1), we obtain

$$\hat{\eta}_0 = \tilde{\eta}(\theta^\circ) = \operatorname*{argmin}_{\eta}\left[\frac{1}{2}\|D^2(\theta^\circ-\tilde{\theta})\|^2 + \frac{1}{2}\|H^2(\eta-\tilde{\eta})\|^2 + (\theta^\circ-\tilde{\theta})^T A(\eta-\tilde{\eta}) + \alpha(\mathbf{v},\tilde{\mathbf{v}})\right].$$

Hence, the first order condition gives us the following relationship:

$$H^2(\hat{\eta}_0-\tilde{\eta}) = A^T(\tilde{\theta}-\theta^\circ) + \nabla_\eta\alpha(\tilde{\mathbf{v}}_{0,0},\tilde{\mathbf{v}}).$$

Similarly, the solution for $\hat{\theta}_1 \stackrel{\text{def}}{=} \tilde{\theta}(\hat{\eta}_0)$ has the following form:

$$D^2(\hat{\theta}_1-\tilde{\theta}) = A(\tilde{\eta}-\hat{\eta}_0) + \nabla_\theta\alpha(\tilde{\mathbf{v}}_{1,0},\tilde{\mathbf{v}}).$$

Then the iterative process of alternating maximization gives us the following recursive system of equations depending on the initial guess:

$$\begin{cases} H^2(\hat{\eta}_k-\tilde{\eta}) = A^T(\tilde{\theta}-\hat{\theta}_k) + \nabla_\eta\alpha(\tilde{\mathbf{v}}_{k,k},\tilde{\mathbf{v}}) \\ D^2(\hat{\theta}_{k+1}-\tilde{\theta}) = A(\tilde{\eta}-\hat{\eta}_k) + \nabla_\theta\alpha(\tilde{\mathbf{v}}_{k+1,k},\tilde{\mathbf{v}}). \end{cases}$$

or, equivalently

$$\begin{cases} H^2(\hat{\eta}_k-\tilde{\eta}) = A^T(\tilde{\theta}-\hat{\theta}_k) + \nabla_\eta\alpha(\tilde{\mathbf{v}}_{k,k},\tilde{\mathbf{v}}) \\ D(\hat{\theta}_{k+1}-\tilde{\theta}) = D^{-1}A(\tilde{\eta}-\hat{\eta}_k) + D^{-1}\nabla_\theta\alpha(\tilde{\mathbf{v}}_{k+1,k},\tilde{\mathbf{v}}). \end{cases} \tag{3.3}$$

Then, the second equation in (3.3) we solve for $(\tilde{\eta}-\hat{\eta}_k)$ and the result plug into the first one, to obtain

$$D(\hat{\theta}_{k+1}-\tilde{\theta}) = D^{-1}AH^{-2}A^T(D^{-1}D)(\hat{\theta}_k-\tilde{\theta}) + D^{-1}\left[\nabla_\theta\alpha(\mathbf{v},\tilde{\mathbf{v}}) - AH^{-2}\nabla_\eta\alpha(\mathbf{v},\tilde{\mathbf{v}})\right].$$

Defining $M_\circ \stackrel{\text{def}}{=} D^{-1}AH^{-2}A^TD^{-1}$ and

$$\Xi(\tilde{\mathbf{v}}_{k(+1),k}) \stackrel{\text{def}}{=} D^{-1}\left[\nabla_\theta\alpha(\tilde{\mathbf{v}}_{k+1,k},\tilde{\mathbf{v}}) - AH^{-2}\nabla_\eta\alpha(\tilde{\mathbf{v}}_{k,k},\tilde{\mathbf{v}})\right],$$

we get the following recursive formula

$$D(\hat{\theta}_{k+1}-\tilde{\theta}) = M_\circ\cdot D(\hat{\theta}_k-\tilde{\theta}) + \Xi(\tilde{\mathbf{v}}_{k(+1),k}). \tag{3.4}$$

Hence, summing over all k, starting from the initial guess, taking the norm and using the triangle inequality, we can write

$$\|D(\hat{\theta}_{k+1}-\tilde{\theta})\| \le \|M_\circ\|\cdot\|D(\hat{\theta}_k-\tilde{\theta})\| + \|\Xi(\tilde{\mathbf{v}}_{k(+1),k})\| \le \|M_\circ\|^k\cdot\|D(\theta^\circ-\tilde{\theta})\|$$
$$+\sum_{\ell=0}^{k-1}\|M_\circ\|^\ell\cdot\|\Xi(\tilde{\mathbf{v}}_{k(+1),k})\| = \|M_\circ\|^k\cdot\|D(\theta^\circ-\tilde{\theta})\| + \|\Xi(\tilde{\mathbf{v}}_{k(+1),k})\|\cdot\frac{1-\|M_\circ\|^k}{1-\|M_\circ\|}.$$

Next, taking into account that by assumption $\rho \stackrel{\text{def}}{=} \|D^{-1}AH^{-1}\| < 1$, it is easy to see that the first term tends to zero as $k\to\infty$.

Now we proceed to show that the Euclidean norm of $\Xi(\tilde{\mathbf{v}}_{k(+1),k})$ adequately decreases with the rise of the number of iterations k.

Considering the term $D^{-1}\nabla_\theta\alpha(\tilde{\mathbf{v}}_{k+1,k},\tilde{\mathbf{v}})$, we apply Theorem C.1 from Andresen and Spokoiny [2], which gives a vanishing in k bound for it. Note that this theorem can also be applied to bound the two remainder terms in $\Xi(\tilde{\mathbf{v}}_{k(+1),k})$, which will give the bound for $\Xi(\tilde{\mathbf{v}}_{k(+1),k})$ as a whole.

Thus, using Theorem C.1 from Andresen and Spokoiny [2] and the triangle inequality, we obtain the claimed estimate for $\Xi(\tilde{\mathbf{v}}_{k(+1),k})$:

$$\|\Xi(\tilde{\mathbf{v}}_{k(+1),k})\| \le \|D^{-1}\nabla_\theta\alpha(\tilde{\mathbf{v}}_{k+1,k},\tilde{\mathbf{v}})\| + \|D^{-1}AH^{-2}\nabla_\eta\alpha(\tilde{\mathbf{v}}_{k,k},\tilde{\mathbf{v}})\| \to 0 \text{ as } k\to\infty.$$

Combining the above stated properties we conclude that

$$\|D(\hat{\theta}_{k+1} - \tilde{\theta})\| \leq \mathbf{s}_k \to 0 \text{ as } k \to \infty. \tag{3.5}$$

Similar arguments can be used to derive the convergence property for the *nuisance* parameter η. Theorem 3.1 is proved.

Remark 3.2. The sequence $\mathbf{s}_k$ from (3.5) can be interpreted as the radius of the elliptic set around $\tilde{\theta}$ in which the estimator $\hat{\theta}_{k+1}$, derived by alternating least squares method, lies with high probability.

Remark 3.3. Theorem 3.1 tells us that we only need the conditions $\rho \stackrel{\text{def}}{=} \|M_\circ\| < 1$ and $\|\mathcal{D}^{-1}\nabla^2\mathbb{E}L(\mathbf{v})\mathcal{D}^{-1} - I_{p+q}\| \leq \delta$ to claim that the sequence of estimators $\left(\hat{\theta}_k\right)$ derived by alternating maximization method linearly converges to the maximum likelihood estimator $\tilde{\theta}$, which is very often too difficult to compute directly.

3.2. An Alternating Estimator

Above it was shown that the sequence of estimators obtained by using alternating least squares technique converges to the corresponding maximum likelihood estimator. It is a very strong and practically important result. Also, as it has been discussed above, there are two main issues that make the problem non-trivial. The first one is that it is impossible to derive a closed form of solution in most of the cases, and the second one is the high dimension of the nuisance parameter which makes direct implementation of the well-known Newton-Raphson method inapplicable. The alternating maximization technique overcame these issues and delivered estimators "close" to the maximum likelihood estimator.

This part is a bit technical and the only practical value is the statistical property of $D(\hat{\theta}_k - \theta^*)$.

Further in this section we will show that the alternating estimator is also close to the true (unknown) value of the parameter (θ^*, η^*). Recall

$$\mathbf{v}^* = (\theta^*, \eta^*) \stackrel{\text{def}}{=} \arg\max_{\mathbf{v}} \mathbb{E}L(\mathbf{v}).$$

The theorem that follows is known as a Fisher expansion, and we formulate it in the framework of GLMs.

Theorem 3.2. *Let the conditions of Theorem 3.1 be satisfied, and let* $\breve{\xi} \stackrel{\text{def}}{=} D^{-1}\breve{\nabla}$, *where* $\breve{\nabla} = \nabla_\theta - AH^{-2}\nabla_\eta$. *Then we have*

$$\|D(\tilde{\theta}_k - \theta^*) - \breve{\xi}\| \to 0, \ as \ k \to \infty.$$

Proof. The proof is based on the same ideas used to prove Theorem 3.1, with the only difference that now the Taylor expansion is taken around point $\mathbf{v}^*$.

First, we use the first order conditions to obtain the following relations:

$$\begin{cases} D(\tilde{\theta}_k - \theta^*) = D^{-1}\nabla_\theta L(\mathbf{v}^*) - D^{-1}A(\tilde{\eta}_k - \eta^*) + D^{-1}\nabla_\theta\alpha(\tilde{\mathbf{v}}_{k,k}, \mathbf{v}^*) \\ H(\tilde{\eta}_k - \eta^*) = H^{-1}\nabla_\eta L(\mathbf{v}^*) - H^{-1}A^T(\tilde{\theta}_{k-1} - \theta^*) + H^{-1}\nabla_\eta\alpha(\tilde{\mathbf{v}}_{k-1,k}, \mathbf{v}^*). \end{cases} \tag{3.6}$$

Again applying the results from Andresen and Spokoiny [2], we can bound the norms of $D^{-1}\nabla_\theta\alpha(\tilde{\mathbf{v}}_{k,k}, \mathbf{v}^*)$ and $H^{-1}\nabla_\eta\alpha(\tilde{\mathbf{v}}_{k-1,k}, \mathbf{v}^*)$. From the system (3.6) one can easily derive the following relation:

$$\begin{aligned} D(\tilde{\theta}_k - \theta^*) \ &= D^{-1}\nabla_\theta L(\mathbf{v}^*) - D^{-1}[AH^{-2}\nabla_\eta L(\mathbf{v}^*) - AH^{-2}A^T(\tilde{\theta}_{k-1} - \theta^*) \\ &\quad + AH^{-2}\nabla_\eta\alpha(\tilde{\mathbf{v}}_{k-1,k}, \mathbf{v}^*)] + D^{-1}\nabla_\theta\alpha(\tilde{\mathbf{v}}_{k,k}, \mathbf{v}^*), \end{aligned}$$

implying that

$$\begin{aligned} D(\tilde{\theta}_k - \theta^*) \ &= M_\circ D(\tilde{\theta}_{k-1} - \theta^*) + D^{-1}\left[\nabla_\theta L(\mathbf{v}^*) - AH^{-2}\nabla_\eta L(\mathbf{v}^*)\right] \\ &\quad + D^{-1}\left\{\nabla_\theta\alpha(\tilde{\mathbf{v}}_{k,k}, \mathbf{v}^*) - AH^{-2}\nabla_\eta\alpha(\tilde{\mathbf{v}}_{k-1,k}, \mathbf{v}^*)\right\}. \end{aligned} \tag{3.7}$$

Next, recalling the definition of $\breve{\xi}$ from from Section 3, the equality (3.7) can be rewritten in the following form:

$$D(\tilde{\theta}_k - \theta^*) - \breve{\xi} = M_\circ D(\tilde{\theta}_{k-1} - \theta^*) + D^{-1}\left\{\nabla_\theta \alpha(\tilde{\mathbf{v}}_{k,k}, \mathbf{v}^*) - AH^{-2}\nabla_\eta \alpha(\tilde{\mathbf{v}}_{k-1,k}, \mathbf{v}^*)\right\}.$$

Summing up over all k starting from the initial guess, and then taking the norm of both sides and using the assumption that $\rho = \|M_\circ\| < 1$, we obtain

$$\|D(\tilde{\theta}_k - \theta^*) - \breve{\xi}\| \le \mathcal{C}(\rho) \cdot D^{-1}\left\{\nabla_\theta \alpha(\tilde{\mathbf{v}}_{k,k}, \mathbf{v}^*) - AH^{-2}\nabla_\eta \alpha(\tilde{\mathbf{v}}_{k-1,k}, \mathbf{v}^*)\right\}, \tag{3.8}$$

where $\mathcal{C}(\rho)$ is some constant depending on ρ. The remaining terms can be estimated by applying Theorem C.1 from Andresen and Spokoiny [2].

Recall that in order to obtain upper bound for $D^{-1}\nabla_\theta \alpha(\tilde{\mathbf{v}}_{k,k}, \mathbf{v}^*)$ we need the condition $\|\mathcal{D}^{-1}\nabla^2 \mathbb{E}L(\mathbf{v})\mathcal{D}^{-1} - I_{p+q}\| \le \delta$ for some constant $\delta > 0$. Finally, we have

$$\|D(\tilde{\theta}_k - \theta^*) - \breve{\xi}\| \le \breve{\mathbf{s}}_k \to 0, \text{ as } k \to \infty. \tag{3.9}$$

Theorem 3.2 is proved.

Note that the expectation of the random variable $\breve{\xi}$ equals zero, and moreover we have $\mathbb{V}ar(\breve{\xi}) = D^{-1}V^2D^{-1}$.

Remark 3.4. It is worth to note that the Fisher expansion remains valid also in the case of finite k, but in this case, the corresponding norm is not bounded by zero, instead it is bounded by a sequence $\breve{\mathbf{s}}_k$ that tends to zero as $k \to \infty$.

The next remark comments the distribution of the random variable $\breve{\xi}$ both asymptotically and for finite sample size.

Remark 3.5. Under the correct model specification, the random variable $\breve{\xi} \in \mathbb{R}^p$ has normal distribution, and hence $\|\breve{\xi}\|^2$ has a χ^2−distribution with p degrees of freedom. In the case of model misspecification, we can say nothing about the distribution of $\|\breve{\xi}\|^2$ in the non-asymptotic framework, but asymptotically it still has a χ^2−distribution with p degrees of freedom.

3.3. Importance of the Initial Guess

The choice of initial guess can play a crucial role in the convergence of alternating method, and if we "succeed" with it, then the gain is twofold. Firstly, we can come up with weaker conditions than those in Theorem 3.1, and secondly, the number of iterations will be significantly less than in the case of "bad" initial point. "Good" initial values of θ° mostly appeal to the first condition of Theorem 3.1, that is, the condition $\rho \overset{\text{def}}{=} \|D^{-1}AH^{-1}\|^2 < 1$.

Denote by $\mathcal{V}$ the vector space of eigenvectors corresponding to the eigenvalues greater than 1 of the matrix $M_\circ$, that is, $\mathcal{V} \overset{\text{def}}{=} \text{span}\,(v_1, v_2, \ldots, v_l)$, where $M_\circ v_i = \lambda_i v_i$, $i \in \{1, \ldots, l\}$ and $|\lambda_i| \ge 1$.

Lemma 3.1. *If the initial value of θ° is chosen to satisfy*

$$\mathbf{u} \perp \mathcal{V}, \tag{3.10}$$

where $\mathbf{u} \overset{\text{def}}{=} D(\theta^\circ - \tilde{\theta})$, then the above mentioned convergence holds.

The proof is straightforward, and is based on simple linear algebra. Nevertheless, we briefly explain the idea. Note that the assumption $\rho \overset{\text{def}}{=} \|D^{-1}AH^{-1}\|^2 < 1$ means that all eigenvalues are inside the unit circle so that the process will converge. It is true independent of the initial guess. Nevertheless, we can weaken this result by the "good" choice of initial values. If the matrix $M_\circ$ has eigenvalues which are outside the unit circle, then the initial guess could help to make them equal to zero being just orthogonal to the space of the corresponding eigenvectors.

Also, it is worth to note that the condition of Theorem 3.1:

$$\|\mathcal{D}^{-1}\nabla^2 \mathbb{E}L(\mathbf{v})\mathcal{D}^{-1} - I_{p+q}\| \le \delta(r) \le \delta$$

has nothing to do with the initial values, and only is needed to bound the remainder term $\Xi(\tilde{\mathbf{v}}_{k(+1),k})$. Thus, this condition can not be weakened depending on the initial point θ°.

4. A NUMERICAL EXAMPLE

In this section, we introduce and implement an algorithm in the case of a toy example. Suppose there are n coins on the table and someone randomly chooses a coin and start flipping it k times. For the ease of illustration and implementation we further assume that there are only two coins. Suppose the first coin has a probability of "head" outcome of p_1 and the second coin $-$ p_2. We introduce one *nuisance* parameter π which is the probability of choosing the first coin.

Assuming there are two coins on the table and one randomly chooses one of them 5 times and after each choice makes 10 flips, in total, we will have 50 outcomes, 2 *target* parameters p_1 and p_2 and one nuisance parameter π.

We describe the algorithm for this simple example and provide the results of the code, which computes the estimates of parameters in each step. We have the following information:

$$\begin{aligned}
&\text{Coin 2:} \quad H,T,H,T,T,H,T,H,H,T.\\
&\text{Coin 1:} \quad H,H,H,H,H,H,T,H,H,H.\\
&\text{Coin 1:} \quad H,T,H,T,H,H,H,H,H,H.\\
&\text{Coin 2:} \quad H,T,H,T,T,T,T,H,H,T.\\
&\text{Coin 1:} \quad H,H,H,T,H,H,T,H,H,T.
\end{aligned}$$

Clearly, we have

$$\tilde{p}_1 = \frac{24}{24+6} = 0.8 \qquad \tilde{p}_2 = \frac{9}{9+11} = 0.45.$$

However, this information is not available for the algorithm. The algorithm starts with some initial values of probabilities, for example, $p_1^\circ = 0.6$ and $p_2^\circ = 0.5$. Recall that in this example θ from the previous section is the vector (p_1, p_2) and η is for π. The estimated probability of k heads out of 10 tosses of coin $c \in \{1,2\}$ is given by $p_c(k) = C_k^{10} p_c^k (1-p_c)^{10-k}$. Note that the binomial coefficient is the same for both coins, so it cancels out and only the ratio of the remaining factors determines the result. Using the initial guesses we come up with the following table

First iteration			
π	$1-\pi$	Coin 1	Coin 2
0.45	0.55	$\approx 2.2H, 2.2T$	$\approx 2.8H, 2.8T$
0.80	0.20	$\approx 7.2H, 0.8T$	$\approx 1.8H, 0.2T$
0.73	0.27	$\approx 5.9H, 1.5T$	$\approx 2.1H, 0.5T$
0.35	0.65	$\approx 1.4H, 2.1T$	$\approx 2.6H, 3.9T$
0.65	0.35	$\approx 4.5H, 1.9T$	$\approx 2.5H, 1.1T$
		$\approx 21.3H, 8.6T$	$\approx 11.7H, 8.4T$

Hence, in the next stage we get the following estimates of p_1 and p_2:

$$\hat{p}_1^{(1)} = \frac{21.3}{21.3+8.6} = 0.71, \qquad \hat{p}_2^{(1)} = \frac{11.7}{8.4+11.7} = 0.58. \tag{4.1}$$

The number of iterations until convergence depends on the tolerance level, which is the acceptable bias from the ML estimators. In other words, it defines the stopping criteria of the algorithm.

For given initial values $p_1^\circ = 0.6$ and $p_2^\circ = 0.5$ of parameters p_1 and p_2 we obtain the following sequence of estimates for parameters (see Table 1 below). Table 1 also contains the iterations of the algorithm with randomly chosen initial values from uniform distribution at certain seed.

The obtained results show that in both cases we have convergence to the ML estimator with high accuracy.

Iteration	Non-random start		Random start	
	$\hat{p}_1^{(i)}$	$\hat{p}_2^{(i)}$	$\hat{p}_1^{(i)}$	$\hat{p}_2^{(i)}$
1	0.600000000	0.500000000	0.935954410	0.0659086863
2	0.713012235	0.581339308	0.759177699	0.434881703
3	0.745292036	0.569255750	0.78052855	0.485752725
4	0.768098834	0.549535914	0.79025212	0.505705724
5	0.7831645	0.534617454	0.794205962	0.513867055
6	0.791055245	0.52628116	0.795774011	0.517227986
7	0.794532537	0.522390437	0.79639082	0.518613125
8	0.79592866	0.520729878	0.796632802	0.519183793
9	0.796465637	0.520047189	0.796727694	0.519418794
10	0.796668307	0.519770389	0.796764932	0.519515523
11	0.796744149	0.519658662	0.796779561	0.51955532
12	0.796772404	0.519613607	0.796785317	0.519571692
13	0.796782900	0.519595434		

Table 1. Convergence with random and non-random initial values

Remark 4.1. *Note that the procedure described in the example can easily be extended to an arbitrary distribution. For example, the first part of the observed data can be obtained from one Gaussian distribution and the second part from another Gaussian distribution. Formally, we can assume that* $X_1, \dots, X_\ell \sim \mathcal{N}(\mu_1, \sigma_1^2)$ *and* $X_{\ell+1}, \dots, X_n \sim \mathcal{N}(\mu_2, \sigma_2^2)$. *The idea is quite general and is the same as in the case of coins example (Bernoulli distribution) when we know that our data comes from* n *different normal distributions* $\mathcal{N}(\mu_i, \sigma_i^2)$ $, \forall i \in \{1, \dots, n\}$ *and we introduce new parameters* p_i, *which are the probabilities of the corresponding distributions.*

REFERENCES

1. A. Andresen, "Finite sample analysis of profile m-estimation in the single index model", Electronic Journal of Statistics, **9**(2), 2528 – 2641, 2015.
2. A. Andresen, V. Spokoiny, "Two convergence results for an alternation maximization", arXiv: 1501.01525, 2015.
3. J. Bezdek, R. Hathaway, "Convergence of Alternating Optimization" Neural, Parallel & Pacific Computations, **11**, 351 – 368, 2003.
4. L. Cavalier, G. K. Golubev, D. Picard and A. B. Tsybakov, "Oracle inequialities for inverse problems", The Annals of Statistics, **30**(3), 843 – 874, 2002.
5. X. Chen, "Large sample sieve estimation of semi-nonparametric models", Handbook of Econometrics, **6**, 5549 – 5632, 2007.
6. V. Chernozhukov, D. Chetverikov and K. Kato, "Anti-concentration and honest, adaptive confidence bands", The Annals of Statistics, **34**(4), 1653 – 1677, 2014.
7. A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society, Series B, **39**, 1 – 38, 1977.
8. I. A. Ibragimov and R. Z. Khas'minskij, *Statistical Estimation. Asymptotic Theory* (Springer, Berlin, 1981).
9. A. Kneip, "Ordered linear smoothers", The Annals of Statistics, **22**(2), 835 – 866, 1994.
10. M. R. Kosorok, *Introduction to Empirical Processes and Semiparametric Inference* (Springer in Statistics, Berlin, 2005).
11. G. J. McLachlan and T. Krishman, *The EM Algorithm and Extensions* (Wiley, New York, 1997).
12. V. Spokoiny, T. Dickhaus, *Basics of Modern Mathematical Statistics* (Springer in Statistics, Berlin, 2015).
13. C. F. J. Wu, "On the convergence properties of the EM algorithm", Annals of Statistics, **11**, 95 – 103, 1983.