Probability and Beyond

Arshak Minasyan

arsh.minasyan@gmail.com

Sunday 19th March, 2017

Contents

0.1 Introduction and Literature Review

Many statistical tasks can be viewed as problems of semiparametric 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 by estimating a given subvector 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 approch is known as *profile maximum likelihood* and it appears to be *semiparametrically efficient* under some mild regularity conditions, which in case of Generalized Linear Models are satisfied. For more general case, for example, in M-estimation framework these technical conditions should be introduced separately and checked whether they are fulfilled or not. We refer to the papers Murphy et al. (1999, 2000) and the book of Kosorok (2005) for a detailed presentation of the modern state of the theory and further references. The extension of these results can be found in Fan et al. (2001); Fan and Huang (2005); see also the references therein.

This study revisits the problem of profile semiparametric estimation and addresses some new issues. 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 are exactly fulfilled, even if some rich nonparametric models are used. This means that the true data distribution $\mathbb{P}$ does not belong to the considered parametric family, in our case — exponential family. Applicability of the general semiparametric theory in such cases is questionable. An important feature of the presented approach is that it equally applies under a possible model misspecification.

Let $\boldsymbol{\mathcal{Y}}$ denote the observed statistical model assumes that the unknown data distribution $\mathbb{P}$ belongs to a given parametric family $(\mathbb{P}_{\boldsymbol{v}})$:

$$\boldsymbol{\mathcal{Y}} \sim \mathbb{P} = \mathbb{P}_{\boldsymbol{v}^*} \in (\mathbb{P}_{\boldsymbol{v}}, \boldsymbol{v} \in \varTheta), \tag{0.1.1}$$

where $\varTheta$ is some high dimensional or even infinite dimensional parametric space.

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

$$L(\boldsymbol{v}) = \log \frac{d\mathbb{P}_{\boldsymbol{v}}}{d\boldsymbol{\mu}_0}(\boldsymbol{\mathcal{Y}})$$

for some dominating measure $\boldsymbol{\mu}_0$. Define the maximum likelihood estimator in the following way

$$\widetilde{\boldsymbol{v}} \stackrel{\text{def}}{=} \operatorname*{arg\,max}_{\boldsymbol{v} \in \varTheta} L(\boldsymbol{v}). \tag{0.1.2}$$

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

$$\boldsymbol{v}^* \stackrel{\text{def}}{=} \operatorname*{arg\,max}_{\boldsymbol{v} \in \varTheta} \mathbb{E} L(\boldsymbol{v}). \tag{0.1.3}$$

Under model misspecification, $\boldsymbol{v}^*$ defines the best parametric fit to $\mathbb{P}$ by the considered family.

The main point of the work is that the Alternating Method gives only a little gain, if any, in the complexity of optimum point computation for Linear Models, under some conditions on parameter dimensions, while in non-linear models the gain is pretty sensible. In non-linear models in most of cases the closed form solution can not be obtained, in some cases even the numerical solutions of first order conditions could be very hard to implement in full parameter dimension. The technique known as alternating maximization (minimization) helps in these situations and gives the estimation of parameter vector with adequate time complexity.

The model that we consider has the parameter $\boldsymbol{v}$ which is of dimension $p+q$, where p is the dimension of *target* parameter and q is the dimension of *nuisance* parameter. Usually p is not large, because we also care about tractability and interpretability of our model, but q can be very large, although it is a *nuisance* parameter, we can not ignore and remove $\boldsymbol{\eta}$ from considered model. Main problems with direct computations occur in high dimensions, i.e. in cases when $p+q$ is large enough to make it impossible to invert matrix of sizes $(p+q) \times (p+q)$, which is, in general, $\mathcal{O}((p+q)^3)$. Further we will only consider the case of finite q, the case of infinite q is out of scope of this work.

The alternating maximization procedure can be understood as a special case of the Expectation Maximization algorithm (EM algorithm). The EM algorithm is a popular algorithm first derived by Dempster, Laird and Rubin in 1977. Further on a number of modifications and extensions of this algorithm come into playground. Dempster et al. also described how EM algorithm can be implement in different fields and give fruitful results. We refer to the McLachlan and Krishman (1997) for the brief introduction to the development of EM algorithm. We restrict ourselves to citing the well-known convergence result by Wu in 1983, which is still state of the art in most settings. Unfortunately, Wu's result - as most convergence results on these interative procedures - only ensures convergence to some set of local maximizers or fixpoints of the procedure. In this work we consider one of the special cases where it is possible to show the actual convergence of the method.

The work has the following structure. In Section 2 we introduce our reader with Alternating Method for Linear Models (LMs) and prove that the method converges to the likelihood estimator by exponentially fast convergence rate for any initial point $\boldsymbol{\theta}^{\circ}$. Section 3 contains the preliminaries about Generalized Linear Models (GLMs) and class of random variables called Exponential Family (EF). Fisher and Wilks theorems are provided in the GLM framework as well as local concentration condition. Section 4 extends the results from Section 2 using the tools explained in Section 3 and Taylor expansion of second order along with the Fisher expansion in the framework of GLMs. Section 5 illustrates how the algorithm of alternating least squares (ALS) works and confirms above theoretical results using both real and simulated data. Section 6 contains concluding remarks.

0.2 Basic Probability

This lecture is an overview of basic probability concepts, notations that we will be using along the course.

0.2.1 Probability Spaces

A triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a probability space, where Ω is the set of outcomes, $\mathcal{F}$ is a set of all possible events and $\mathbb{P}$ is function from $\mathcal{F}$ to $[0, 1]$ that assigns probabilities to events, precisely $\mathbb{P} : \mathcal{F} \to [0, 1]$. We restrict ourselves with some assumptions on $\Omega, \mathcal{F}$ and $\mathbb{P}$: $\mathcal{F}$ is a σ-algebra of set Ω and $\mathbb{P}$ is a *probability measure.*

We define a measure μ as follows, it is a function $\mu : \mathcal{F} \to \mathbb{R}$ with the following properties:

- $\mu(A) \geq \mu(\emptyset)$ for all $A \in \mathcal{F}$
- if $A_i \in \mathcal{F}$ is a countable (finite or countably infinite) sequence of disjoint sets, then

$$\mu(\cup A_i) = \sum_i \mu(A_i) \tag{0.2.1}$$

A measure becomes *probability measure* if $\mu(\Omega) = 1$. The following theorem is stated without proof and contains the basic properties of measure μ defined on $(\Omega, \mathcal{F})$.

Theorem 0.2.1. *Let μ be a measure on $(\Omega, \mathcal{F})$ then*

(a) ***monotonicity.*** *If $A \subset B$ then $\mu(A) \leq \mu(B)$.*

(b) ***sub-additivity.*** *If $A \subset \cup_{m=1}^{\infty} A_m$ then $\mu(A) \leq \sum_{m=1}^{\infty} \mu(A_m)$.*

(c) ***continuity from below.*** *If $A_i \uparrow A$, i.e. $A_1 \subset A_2 \subset \ldots, \cup A_i = A$ then $\mu(A_i) \uparrow \mu(A)$.*

(d) ***continuity from above.*** *If $A_i \downarrow A$, i.e. $A_1 \supset A_2 \supset \ldots, \cap A_i = A$ and $\mu(A_1) < \infty$ then $\mu(A_i) \downarrow \mu(A)$.*

The probability spaces can be either discrete or continuous. If the set of outcomes Ω is countable, i.e. either finite or countably infinite, then the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is discrete, otherwise it is continuous.

The natural example of discrete probability space is $\Omega = \{0, 1\}$. More unintuitive one could be $\Omega = \mathbb{N}$ (set of natural numbers). The set of natural numbers is indeed infinite, but it is countable, hence it is still discrete.

The probability measure on $\Omega = \{0, 1\}$ is, in general, $\mu(\{1\}) = p$ and $\mu(\{0\}) = q = 1 - p$, where $p \in [0, 1]$. It is straightforward to check that the given measure is indeed a probability measure.

Remark 0.2.1. *It is very natural to think of the mixture of above described two spaces, i.e. the space is somewhere continuous and somewhere discrete.*

0.2.2 Random Variables

Probability spaces become more interesting when we define random variables on them. A random variable X on Ω is a measurable function $X : \Omega \to T$, where T is some set of values that random variable can possibly take. $\mathcal{F}$-measurable or just measurable means that for every Borel set $B \subset T$ it holds $X^{-1}(B) =$

$\{\omega : X(\omega) \in B\} \in \mathcal{F}$. Another simple, but useful, example random variable is *indicator function* of a set $A \in \mathcal{F}$:

$$1_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \notin A \end{cases} \tag{0.2.2}$$

Any random variable X induces a probability measure on $\mathbb{R}$ known as distribution of random variable X, i.e. $\mu(A) = \mathbb{P}(X \in A)$ for Borel sets A. The distribution of random variable X is usually described by giving its *distribution function*, $F(x) = \mathbb{P}(X \leq x)$.

Theorem 0.2.2. *Any distribution function $F(x)$ has the following properties:*

(a) F is non-decreasing

(b) $\lim_{x\to+\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$.

(c) F is right continuous, we denote it as $\lim_{y\downarrow x} F(y) = F(x)$.

(d) Define $F(x-) = \lim_{y\uparrow x} F(y)$, then $F(x-) = \mathbb{P}(X < x)$.

(e) $\mathbb{P}(X = x) = \mathbb{P}(X \leq x) - \mathbb{P}(X < x) = F(x) - F(x-)$.

Proof.

(a)

■

The following exercise is known as St Petersburg paradox.

Exercise 0.2.3. *A fair coin is tossed repeatedly. Let T be the number of tosses until the first head. You are offered the following prospect, which you may accept on payment of a fee. If $T = k$, say, then you will receive 2^k roubles. What would be a 'fair' fee to ask of you?*

0.2.3 Independence of random variables

0.2.4 Convergence of random variables

There are mainly 4 types of convergence: in probability, almost surely, weak convergence (in distribution) and in mean.

0.2.5 Law of Large Numbers

0.2.6 Central Limit Theorem

0.2.7 Distributions

0.3 Stochastic Processes

This section contains a gentle introduction to stochastic processes.

0.3.1 Stochastic Processes in Discrete Time

0.3.2 Stochastic Processes in Continuous Time

0.3.3 Dirichlet Processes

0.4 Basic Statistics

This section introduces the basics of mathematical statistics and contains several modern methods for statistical estimation.

This section was written mainly based on the course for senior students of Moscow Institute of Physics and Technology (MIPT) called "Methods of Modern Mathematical Statistics", originally, "?????? ??????????? ?????????????? ??????????", taught in spring 2017.

0.5 Bayesian Inference

We start this section with an interesting and simple exercise.

Example 0.5.1. *Consider two doctors doing two type of surgeries, and the first doctor has better performance on both surgeries, separately. Is it possible that the second doctor has better performance overall? If yes, bring an example, otherwise prove that it is impossible.*

0.5.1 Maximum Likelihood

Consider the maximum likelihood function: given an independent and identically distributed (i.i.d.) set of data from a density function f_θ with an unknown parameter $\theta \in \Theta$, the associated likelihood function is

$$\mathcal{L}(\theta \,|\, x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i) \tag{0.5.1}$$

The standard parameter estimation technique is to maximize the (log) likelihood with respect to parameter $\theta \in \Theta$. Usually $\Theta = \mathbb{R}$ which leads to , for more complicated cases one can use Lagrange multipliers along with Karush-Kuhn-Tucker conditions to obtain

$$\widetilde{\theta} = \arg\max_{\theta\in\Theta} \mathcal{L}(\theta \,|\, x_1, \ldots, x_n). \tag{0.5.2}$$

Note that in some cases it might be convenient to maximize the logarithm of the likelihood function. The estimator obtained this way is known as maximum likelihood estimator (MLE).

Exercise 0.5.1. *Estimate the probability of getting a tail p of an unfair coin given n experiments. What do you think is bad in this estimate?*

0.5.2 Bayesian approach

The major input of Bayesian approach, compared to the standard likelihood approach, is that it modifies the likelihood into a *posterior* distribution, taking into account what we believe about the parameter θ *a priori.*

Applying Bayes rule and letting $\mathbf{X} = \{x_1, \ldots, x_n\}$ we can obtain

$$\pi(\theta \,|\, \mathbf{X}) = \frac{p(\mathbf{X} \,|\, \theta)\pi_0(\theta)}{\int p(\mathbf{X} \,|\, \theta)\pi_0(\theta)\, d\theta}, \tag{0.5.3}$$

where $\pi_0(\theta)$ is prior distribution, $p(\mathbf{X}, \theta)$ is the likelihood of $\mathbf{X}$ given for θ. $p(\mathbf{X} \,|\, \theta)$ is a joint conditional probability of $\mathbf{X}$ given θ, and it can be decomposed as the product of one-dimensional densities (assuming iid x_i's), i.e.

$$p(\mathbf{X} \,|\, \theta) = \prod_{i=1}^{n} p_\theta(x_i), \tag{0.5.4}$$

which is exactly what we have in (**??**).

Note that the expression (**??**) results in a distribution function of θ known as posterior distribution of θ given data $\mathbf{X}$.

In order to obtain a point estimate from the distribution function one can maximize the probability density function (pdf) of posterior distribution

$$\widehat{\theta} = \arg\max_{\theta} \pi(\theta \,|\, \mathbf{X}) = \arg\max_{\theta} \left[\overbrace{p(\mathbf{X} \,|\, \theta)}^{\text{likelihood}} \underbrace{\pi_0(\theta)}_{\text{prior}} \right] \overset{\text{def}}{=} \widehat{\theta}_{\text{MAP}}, \tag{0.5.5}$$

which is called maximum a posteriori (MAP) estimator; or take the expected value of it

$$\widehat{\theta} = \mathbb{E}\pi(\theta \,|\, \mathbf{X}) = \frac{\int \theta p(\mathbf{X} \,|\, \theta)\pi_0(\theta)\, d\theta}{\int p(\mathbf{X} \,|\, \theta)\pi_0(\theta)\, d\theta} \tag{0.5.6}$$

Remark 0.5.1. *$\widehat{\theta}_{MAP}$ converges to $\widehat{\theta}_{MLE}$ when $n \to \infty$.*

Exercise 0.5.2. *Estimate the probability of getting a tail p of an unfair coin (maximizing the posterior probability) given n experiments using a prior for p, $p \sim \boldsymbol{\mathcal{B}eta}(\alpha, \beta)$ with density*

$$p_{\alpha,\beta}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \quad B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \tag{0.5.7}$$

which implies $p_{\alpha,\beta}(x) \propto x^{\alpha-1}(1-x)^{\beta-1}$. Comment on what you obtain.

0.5.3 Prior distributions

In this section we introduce an important class of distributions, which is commonly used in Bayesian inference.

Bibliography

[1] Andresen A., (2015). Finite sample analysis of profile m-estimation in the single index model. *Electronic Journal of Statistics*, 9(2):2528–2641.

[2] Andresen, A., Spokoiny, V., (2015) Two convergence results for an alternation maximization. arXiv: 1501.01525v1

[3] Andresen, A. and Spokoiny, V., (2013). Critical dimension in profile semiparametric estimation. *Electronic Journal of Statistics*, 8(2):3077–3125, 2014. ISSN 1935-7524. doi: 10.1214/14-EJS982.

[4] Bezdek, J., Hathaway, R. (2003). Convergence of Alternating Optimization. *Neural, Parallel & Pacific Computations 11*, 351 — 368.

[5] Cavalier, L., Golubev, G. K., Picard, D., and Tsybakov, A. B. (2002). Oracle inequialities for inverse problems. *The Annals of Statistics*, 30(3): 843 — 874.

[6] Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 6:55495632.

[7] Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics*, 34(4):1653 — 1677.

[8] A.P. Dempster, N.M. Laird, and D.B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39:1 — 38.*

[9] Gach, F., Nickl, R., and Spokoiny, V. (2013). Spatially Adaptive Density Estimation by Localised Haar Projections. *Annales de l'Institut Henri Poincare - Probability and Statistics*, 49(3): 900 — 914. DOI: 10.1214/120-AIHP485; arXiv:1111.2807.

[10] Gelca, R., Titu, A., (2007). Putnam and beyond. *Springer*, ISBN-13: 978-0-387-25765-5.

[11] Ibragimov, I.A. and R.Z. Khas'minskij (1981). *"Statistical estimation. Asymptotic theory. Translation from Russian by Samuel Kotz."*. "New York - Heidelberg - Berlin: Springer-Verlag".

[12] Kneip, A., (1994). Ordered linear smoothers. *The Annals of Statistics*, 22(2):835—866.

[13] M. R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference* (2005). *Springer in Statistics.*

[14] G. J. McLachlan and T. Krishman (1997). *The EM Algorithm and Extensions.* Wiley, New York.

[15] Spokoiny, V. (2012). Parametric estimation. Finite sample theory. *Annals of Statistics*, 40(6):2877–2909.

[16] Spokoiny, V., Dickhaus, T., (2015). Basics of Modern Mathematical Statistics . *Springer Texts in Statistics.*

[17] Spokoiny, V., Wang, W., and Härdle, W. (2013). Local quantile regression (with rejoinder). *Journal of Statistical Planing and Inference,* 143(7):1109 — 1129. arXiv:1208.5384.

[18] Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Compution Mathematics,* 12.

[19] Tsybakov, A. (2000). On the best rate of adaptive estimation in some inverse problems. *Comptes Rendus de l'Academie des Sciences - Series I - Mathematics,* 330(9):835 — 840.

[20] Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics*, 11:95 — 103.