Non-asymptotic Bernstein-von Mises Theorem For Semiparametric Estimation

Maxim Panov[1,2], Vladimir Spokoiny[2,3]
[1] Institute for Information Transmission Problems of RAS,
[2] Moscow Institute of Physics and Technology,
[3] Weierstrass Institute and Humboldt University Berlin.

The prominent Bernstein-von Mises (BvM) Theorem claims that the posterior measure is asymptotically normal with the mean close to the maximum likelihood estimator (MLE) and the variance close to the variance of the MLE (inverse total Fisher information matrix). The BvM result provides a theoretical background for Bayesian computations of the MLE and its variance. Also it justifies usage of elliptic credible sets based on the first two moments of the posterior.

The classical version of the BvM Theorem is stated for the standard parametric setup with a fixed parametric model and large samples. Modern statistical problems require to consider situations with very complicated models involving a lot of parameters and with limited sample size. Many statistical problem 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 [1, 2]. 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. In this paper we concentrate on BvM result for semiparametric case with special attention paid to proportion between dimension of full parameter space and sample size. This value is shown to be crucial for BvM result to be valid.

The results are based on recent paper [3] which offers a new look at the classical LAN theory. The basic idea is replace local approximation by local majorization which allows to use much larger neighborhoods than in the LAN approach. In the paper [4] it is shown that the *local nonasymptotic quadraticity* approach of [3] can be used for obtaining a nonasymptotic version of Bernstein-von Mises theorem. In this paper we make a step further to obtain non-asymptotic semiparametric result.

Let $\boldsymbol{Y}$ denote the observed random data, and $I\!\!P$ denote the data distribution. The parametric statistical model assumes that the unknown data distribution $I\!\!P$ belongs to a given parametric family $(I\!\!P_{\boldsymbol{\upsilon}})$:

$$\boldsymbol{Y} \sim I\!\!P = I\!\!P_{\boldsymbol{\upsilon}^*} \in (I\!\!P_{\boldsymbol{\upsilon}}, \, \boldsymbol{\upsilon} \in \varUpsilon),$$

where $\varUpsilon$ is some possibly infinite dimensional parameter space. In the semiparametric framework, the target of analysis is only a finite dimensional component $\boldsymbol{\theta}$ of the whole parameter $\boldsymbol{\upsilon}$. This means that the target of estimation is

$$\boldsymbol{\theta}^* = P\boldsymbol{\upsilon}^*,$$

for some mapping $P : \varUpsilon \to I\!\!R^p$, and $p \in \mathbb{N}$ stands for the dimension of the target.

Usually in the classical semiparametric setup, the vector $\boldsymbol{\upsilon}$ is represented as $\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{\eta})$, where $\boldsymbol{\theta}$ is the target of analysis while $\boldsymbol{\eta}$ is the *nuisance parameter*. We refer to this situation as $(\boldsymbol{\theta}, \boldsymbol{\eta})$-setup and our presentation follows this setting. An extension to the $\boldsymbol{\upsilon}$-setup with $\boldsymbol{\theta} = P\boldsymbol{\upsilon}$ is straightforward.

Let's suppose that the parameter set $\varUpsilon$ can be represented as follows: $\varUpsilon = \varTheta \times H$. Let $\varPi$ be a prior measure on the parameter set $\varUpsilon$ and $\pi(\boldsymbol{\upsilon})$ is corresponding density w.r.t. to the uniform σ-finite measure on $\varUpsilon$: $\varPi(d\boldsymbol{\upsilon}) = \pi(\boldsymbol{\upsilon})d\boldsymbol{\upsilon}$. In this paper we are interested in target component of full parameter and study corresponding marginal posterior distribution

$$\boldsymbol{\vartheta} \mid \boldsymbol{Y} \propto \int_H \exp\bigl\{L(\boldsymbol{\upsilon})\bigr\}\, \pi(\boldsymbol{\upsilon}) d\boldsymbol{\eta}.$$

An important feature of the posterior distribution is that it is entirely known and can be numerically assessed. If we know in addition that the posterior is nearly normal, it suffices to compute its mean and variance for building the concentration and credible sets. Define

$$\overline{\boldsymbol{\vartheta}} \stackrel{\text{def}}{=} I\!\!E\bigl(\boldsymbol{\vartheta} \bigm| \boldsymbol{Y}\bigr), \qquad \mathfrak{S}^2 \stackrel{\text{def}}{=} \operatorname{Cov}(\boldsymbol{\vartheta}) \stackrel{\text{def}}{=} I\!\!E\bigl\{(\boldsymbol{\vartheta} - \overline{\boldsymbol{\vartheta}})(\boldsymbol{\vartheta} - \overline{\boldsymbol{\vartheta}})^{\top} \bigm| \boldsymbol{Y}\bigr\}.$$

Both quantities are data dependent. Here we present a version of the BvM result which claims that $\overline{\boldsymbol{\vartheta}}$ is close to the MLE $\widetilde{\boldsymbol{\theta}}$, $\mathfrak{S}^2$ is nearly equal to inverse of total Fisher information matrix, and $\mathfrak{S}^{-1}\bigl(\boldsymbol{\vartheta} - \overline{\boldsymbol{\vartheta}}\bigr)$ is nearly standard normal conditionally on $\boldsymbol{Y}$.

Theorem 0.1. *It holds with high probability that*

$$\|\breve{D}_0(\overline{\boldsymbol{\vartheta}} - \widetilde{\boldsymbol{\theta}})\|^2 \le \mathtt{C}\,\tau, \qquad \bigl\|I_p - \breve{D}_0 \mathfrak{S}^2 \breve{D}_0\bigr\|_{\infty} \le \mathtt{C}\,\tau,$$

where $\breve{D}_0^2 = \breve{I}(\boldsymbol{\upsilon}^*)$ is an analog of total Fisher information matrix in semiparametric case, $\mathtt{C}$ is universal constant and τ is (small) value depending on properties of likelihood and sample size.

Moreover, for any $\boldsymbol{\lambda} \in I\!\!R^p$ with $\|\boldsymbol{\lambda}\|^2 \le p$

$$\Bigl|\log I\!\!E\Bigl[\exp\bigl\{\boldsymbol{\lambda}^{\top}\mathfrak{S}^{-1}(\boldsymbol{\vartheta} - \overline{\boldsymbol{\vartheta}})\bigr\} \bigm| \boldsymbol{Y}\Bigr] - \|\boldsymbol{\lambda}\|^2/2\Bigr| \le \mathtt{C}\,\tau.$$

We conclude that the BvM Theorem requires "τ is small" and it can be shown that τ is of order $(p^*)^{3/2}/n^{1/2}$, where p^* is dimension of full parameter. Thus we come to the notion of *critical dimension*. Namely we assume that the total dimension p^* grows with the sample size n and write $p^* = p_n$. Theorem 1 requires that $p_n = o(n^{1/3})$. We present an example for which $p_n^{3/2}/n^{1/2} \ge \beta > 0$ and the BvM result is not valid.

Literature:

1. *Castillo, I.* (2012) A semiparametric Bernstein - von Mises theorem for Gaussian process priors. Probability Theory and Related Fields, 152:5399. 10.1007/s00440-010-0316-5.

2. *Bontemps D.* (2011) Bernstein - von Mises theorem for Gaussian regression with increasing number of regressors. The Annals of Statistics, Vol. 39, No. 5, 25572584.

3. *Spokoiny, V.* (2012). Parametric estimation. Finite sample theory. Annals of Statistics, 40(6):28772909. arXiv:1111.3029.

4. *Spokoiny, V.* (2013). Bernstein - von Mises Theorem for growing parameter dimension. Manuscript. arXiv:1302.3430.