Bootstrap validity

Arshak Minasyan

February 4, 2017

Guideline

- Introduction to the framework of GLMs
- Proof of Fisher and/or Wilks expansions
- Gaussian approximation/comparison

Main steps

(a) Use Fisher expansion for both $\mathbf{Y}$-world and bootstrap world.

(b) Use GAR for them, and *finally*

(c) compare these two gaussian approximations using Pinsker's inequality.

1 Introduction

1.1 Linear case

In this section we discuss the simplest linear model and explicitly show the validity of bootstrapping procedure.

Consider a model

$$\mathbf{Y} = \Psi^T\theta + \epsilon, \tag{1}$$

where Ψ is a given design with sizes $p \times n$ and $\epsilon \sim \mathcal{N}(0, \Sigma)$ – is our assumption about the noise. The log-likelihood has the following form

$$L(\theta) = -\frac{1}{2}(\mathbf{Y} - \Psi^T\theta)^T\Sigma^{-1}(Y - \Psi^T\theta) + R, \tag{2}$$

where R is the remainder and does not depend on parameter θ. Define the following objects

$$D^2 \stackrel{\text{def}}{=} -\nabla^2 \mathbb{E} L(\theta) = \Psi \Sigma^{-1} \Psi^T, \qquad \tilde{\theta} = D^{-2} \Psi \Sigma^{-1} \mathbf{Y} \tag{3}$$

$$\xi \stackrel{\text{def}}{=} D^{-1} \nabla L(\theta^*) = D^{-1} \Psi \Sigma^{-1} (\mathbf{Y} - \mathbf{f}), \qquad \theta^* = D^{-2} \Psi \Sigma^{-1} \mathbf{f}, \tag{4}$$

where $\mathbf{f} = \mathbb{E}\mathbf{Y}$. One can see that

$$D(\tilde{\theta} - \theta^*) \equiv \xi, \qquad L(\tilde{\theta}) - L(\theta^*) \equiv \frac{\|\xi\|^2}{2}, \tag{5}$$

the latter could be derived using the Tailor expansion of the second order around maximum likelihood estimator $\tilde{\theta}$ and the use of $\nabla L(\tilde{\theta}) = 0$.

Here we see that the Fisher and Wilks expansions are saturated, namely we they are equal without any conditions and could be applied to a data of any size. These results are based only on quadraticity of the likelihood function $L(\theta)$ in θ.

Denoting $\epsilon = \mathbf{Y} - \mathbf{f}$ one can see that the normalized score has the following form

$$\nabla = \breve{\Psi} \epsilon, \tag{6}$$

which is the linear combination of ϵ_i and hence, under the assumption of normality of the error term we have that the the normalized score ∇ is a Gaussian random variable with zero mean and covariance matrix S:

$$S \stackrel{\text{def}}{=} \mathbb{V}ar(\nabla) = \breve{\Psi} \Sigma \breve{\Psi}^T. \tag{7}$$

Remark 1.1 *Note that this derivations work under the strong assumption of the normality of the error term, since we wrote the likelihood function using the probability density function of the multivariate normal distribution.*

1.1.1 Bootstrap counterpart

Define the bootstrap counterpart of the likelihood function from (**??**)

$$L^\flat(\theta) = \sum_{i=1}^{n} \ell_i(\theta; Y_i) \cdot w_i^\flat, \tag{8}$$

where $w_i^\flat$s are known as bootstrap multipliers (weights) and we are free to choose this multipliers. The only constrains are $\mathbb{E}^\flat w_i^\flat = 1$, $\mathbb{V}ar^\flat w_i^\flat = 1$ and $\mathbb{E}^\flat \exp\{w_i^\flat\} < +\infty$.

Since, we have assumed that the errors ϵ_is are also normal, then in order to prove bootstrap validity we have to compare the covariance matrices of the score with respect to the measure $\mathbb{P}$ and the score with respect to the measure $\mathbb{P}^\flat$.

Now we examine the following expression

$$\|S^{1/2} (S^\flat)^{-1} S^{1/2} - I_n\|_{\text{op}} \tag{9}$$

1.2 Generalized Linear regression

In this section we derive the necessary terms in general case (without decomposing the initial variable vector v; for the ease of representation we denote it θ) needed for further analysis for the special class of distributions called exponential class of distributions.

Further we will assume $\mathcal{P}$ to be an exponential family (EF), which has a number of good properties, namely, the log-likelihood function could be written in this way $\ell(v, y) = vy - g(v)$ and

$$L(\theta) = \sum_{i=1}^{n} \ell(Y_i, f(X_i)) = \sum_{i=1}^{n} \{Y_i \Psi_i^T \theta - g(\Psi_i^T \theta)\} = \mathbf{Y}^T \Psi \theta - A(\theta) \tag{10}$$

with

$$A(\theta) \stackrel{\text{def}}{=} \sum_{i=1}^{n} g(\Psi_i^T \theta) \tag{11}$$

Define

$$D^2 \stackrel{\text{def}}{=} \nabla^2 A(\theta) = \sum_{i=1}^{n} \Psi_i \Psi_i^T g''(\Psi_i^T \theta). \tag{12}$$

We also define the Fisher information matrix

$$\mathbb{F} = D^2 = -\nabla^2 \mathbb{E} L(\theta^*) = \sum_{i=1}^{n} \Psi_i \Psi_i^T g''(\Psi_i^T \theta^*) = \Phi g''(\Psi^T \theta^*) \Psi^T. \tag{13}$$

We subtract the deterministic part from $L(\theta)$ and get $\zeta(\theta) = L(\theta) - \mathbb{E} L(\theta)$, then

$$\nabla \zeta(\theta) = \sum_{i=1}^{n} \epsilon_i \Psi_i = \Psi \epsilon, \qquad V^2 = \mathbb{V}ar\left(\nabla \zeta(\theta)\right) = \Psi \mathbb{V}ar(\epsilon) \Psi^T, \tag{14}$$

where $\epsilon_i = Y_i - \mathbb{E} Y_i$.

The Fisher expansion reads as

$$\|D(\tilde{\theta} - \theta^*) - \xi\| \leq \diamondsuit(x) \tag{15}$$

on a dominating (elliptic) set of probability at least $1 - e^x$, where ξ is defined as follows

$$\xi \stackrel{\text{def}}{=} D^{-1} \nabla L(\theta^*) = D^{-1} \nabla \zeta(\theta^*) = D^{-1} \Psi \epsilon. \tag{16}$$

The latter means that ξ is simply the linear combination of errors ϵ_i. We achieve this result thanks to the structure of likelihood for EFc which stochastic part is linear in $\mathbf{Y}$, the only source of randomness.

1.3 Bootstrap counterpart

In the bootstrap world we have

$$L^\flat(\theta) = \sum_{i=1}^{n} \ell_i(\theta) w_i^\flat = \sum_{i=1}^{n} \left(Y_i \Psi_i^T \theta - g(\Psi_i^T \theta) \right) w_i^\flat = \theta^T \Psi \mathcal{W}^\flat \mathbf{Y} - A^\flat(\theta) \tag{17}$$

with

$$A^\flat(\theta) \stackrel{\text{def}}{=} \sum_{i=1}^{n} g(\Psi_i^T \theta) w_i^\flat. \tag{18}$$

Note that

$$\mathbb{E}^\flat A^\flat(\theta) = A. \tag{19}$$

We define the counterpart of zeta function in bootstrap world as follows

$$\zeta^\flat(\theta) = \nabla L^\flat(\theta) - \mathbb{E}^\flat \nabla L^\flat(\theta) \tag{20}$$

Simple algebra and the fact that $\nabla \mathbb{E}^\flat L^\flat(\tilde{\theta}) = 0$ brings us to the following expression

$$\zeta^\flat(\tilde{\theta}) = \sum_{i=1}^{n} \left[Y_i \Psi_i^T - \Psi_i g'(\Psi_i^T \tilde{\theta}) \right] \epsilon_i^\flat = Y^T \Psi \mathcal{E}^\flat - \nabla A(\tilde{\theta}) \mathcal{E}^\flat, \tag{21}$$

where $\epsilon_i^\flat = w_i^\flat - 1$. The Fisher expansion in bootstrap world has the following form:

$$\| D(\tilde{\theta}^\flat - \tilde{\theta}) - \xi^\flat \| \leq \diamondsuit^\flat(x) \tag{22}$$

One can see that the standartized score in real world is

$$\xi = D^{-1} \Psi \epsilon \tag{23}$$

and the standartized score in bootstrap world is

$$\xi^\flat = \breve{D}^{-1} \left[\mathbf{Y}^T \Psi - \nabla A(\tilde{\theta}) \right] \mathcal{E}^\flat | \mathbf{Y} \sim \mathcal{N}(0, V^2), \tag{24}$$

since the random components with respect to the measure $\mathbb{P}^\flat$ are standard normal random variables $\epsilon_i^\flat \,|\, \mathbf{Y} \sim \mathcal{N}(0,1)$ by construction.

In order to proceed with bootstrap validity we need to compare these two standartized scores.

2 Main

We will discuss the case when the error distribution comes from exponential family with canonical parameter (EFc).

We define the oracle of described model as $v^* = \arg\max_v \mathbb{E}L(v, \mathbf{Y})$, while the data-driven estimator for the oracle v^* is defined as $\tilde{v} = \arg\max L(v|Y)$. Note that the source of randomness in $L(v, \mathbf{Y})$ is behind $\mathbf{Y}$, the distribution of which, in general, is unknown.

We introduce the bootstrap counterpart of (log) likelihood function $L(\cdot)$ as follows

$$L^\flat(v) = \sum_{i=1}^{n} \ell_i(v|\mathbf{Y}) \cdot w_i^\flat, \tag{25}$$

where $w_i^\flat$ are known as bootstrap weights with $\mathbb{E}w_i^\flat = 1$, $\mathbb{V}\mathrm{ar}\, w_i^\flat = 1$ and $\mathbb{E}\exp(w_i^\flat) < \infty$ for all $i \in [1, n]$.

One can see that these two log-likelihood function are very similar to each other, but indeed live in different probability spaces, since the log-likelihood in (**??**) is the function of random data $\mathbf{Y}$, while the log-likelihood in (**??**) is the bootstrap counterpart of log-likelihood and considered to have a different source of randomness, namely, the randomness comes from the *bootstrap multipliers*; $\mathbf{Y}$ is fixed. The link between these two probability spaces is given with this simple observation

$$\mathbb{E}^\flat L^\flat(v) \equiv L(v) \tag{26}$$

Suppose $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is iid data coming from some unknown distribution $\mathcal{P}$. In general, $Y_i \sim \mathcal{P}_{f(x)}$, where $f(x)$ is the true function that we are trying to approximate. Consider $X_i \in \mathbb{R}^d$ and $Y_i \sim P_i \in (\mathcal{P}_{\mathbf{v}})$, which means that $\exists v_i : P_i = P_{v_i}$. Function $\mathbf{v}(x)$ can be represented in the following way

$$v(x) = \sum_{j=1}^{\infty} \theta_j \psi_j(x).$$

Then, our linear parametric assumption is

$$v(x) = \sum_{j=1}^{p+q} \theta_j \psi_j(x) = \sum_{j=1}^{p} \theta_j \psi_j(x) + \sum_{j=p+1}^{p+q} \theta_j \psi_j(x) \text{ for given basis } \psi_j(\cdot). \tag{27}$$

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

The log-likelihood function is defined as follows

$$L(v; \mathbf{Y}) = \log \frac{d\mathbb{P}_v}{d\mu_0^n}(\mathbf{Y}) = \sum_{i=1}^{n} v_i Y_i - g(v_i), \tag{28}$$

where v is the parameter of interest. Define

$$\tilde{v} = \arg\max_{v \in \Theta} L(v, \mathbf{Y}) \qquad v^* = \arg\max_{v \in \Theta} \mathbb{E}L(v, \mathbf{Y}) \tag{29}$$

to be maximum likelihood estimator and oracle, respectively.

3 Example. Bernoulli (Classification)?, Poisson

4 Conditions

Appendix

Notations of important terms

Conditions

Pinsker's Inequality and Gaussian comparison

Pinsker's inequality

Pinsker's inequality is a great tool for bounding the total variation distance between two probability measures.

Proof of the main theorem