Bootstrap validity

Arshak Minasyan

February 5, 2017

1 Introduction

intro and literature review

2 Bootstrap validity

In this section we will discuss the framework of Generalized Linear Models (GLMs) and show that the distribution of $\tilde{\theta}-\theta^*$ can be mimicked using the counterpart from bootstrap world $\tilde{\theta^\flat}-\tilde{\theta}$, i.e.

$$\mathcal{L}^\flat(\tilde{\theta^\flat}-\tilde{\theta}\,|\,\mathbf{Y}) \approx \mathcal{L}(\tilde{\theta}-\theta^*). \tag{1}$$

2.1 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

$$L(\theta) = \sum_{i=1}^n \ell(Y_i, f(X_i)) = \sum_{i=1}^n \{Y_i\Psi_i^T\theta - g(\Psi_i^T\theta)\} = \mathbf{Y}^T\Psi\theta - A(\theta) \tag{2}$$

with

$$A(\theta) \stackrel{\text{def}}{=} \sum_{i=1}^n g(\Psi_i^T\theta) \tag{3}$$

Define

$$D^2 \stackrel{\text{def}}{=} \nabla^2 A(\theta) = \sum_{i=1}^n \Psi_i\Psi_i^T g''(\Psi_i^T\theta). \tag{4}$$

We also define the Fisher information matrix

$$\mathbb{F} = D^2 = -\nabla^2\mathbb{E}L(\theta^*) = \sum_{i=1}^n \Psi_i\Psi_i^T g''(\Psi_i^T\theta^*) = \Phi g''(\Psi^T\theta^*)\Psi^T. \tag{5}$$

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

$$\nabla\zeta(\theta) = \sum_{i=1}^n \epsilon_i\Psi_i = \Psi\epsilon, \qquad V^2 = \mathbb{V}ar\left(\nabla\zeta(\theta)\right) = \Psi\mathbb{V}ar(\epsilon)\Psi^T, \tag{6}$$

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

The Fisher expansion reads as

$$\|D(\tilde{\theta}-\theta^*) - \xi\| \le \diamondsuit(x) \tag{7}$$

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

$$\xi \stackrel{\text{def}}{=} D^{-1}\nabla L(\theta^*) = D^{-1}\nabla\zeta(\theta^*) = D^{-1}\Psi\epsilon. \tag{8}$$

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.

2.2 Bootstrap counterpart

In the bootstrap world we have

$$L^\flat(\theta) = \sum_{i=1}^n \ell_i(\theta) w_i^\flat = \sum_{i=1}^n \left(Y_i \Psi_i^T\theta - g(\Psi_i^T\theta)\right) w_i^\flat = \theta^T\Psi\mathcal{W}^\flat\mathbf{Y} - A^\flat(\theta) \tag{9}$$

with

$$A^\flat(\theta) \stackrel{\text{def}}{=} \sum_{i=1}^n g(\Psi_i^T\theta) w_i^\flat. \tag{10}$$

Note that

$$\mathbb{E}^\flat A^\flat(\theta) = A. \tag{11}$$

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

$$\zeta^\flat(\theta) = \nabla L^\flat(\theta) - \mathbb{E}^\flat\nabla L^\flat(\theta) \tag{12}$$

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

$$\zeta^\flat(\theta^*) = \sum_{i=1}^n \left[Y_i\Psi_i^T - \Psi_i g'(\Psi_i^T\theta^*)\right]\epsilon_i^\flat = Y^T\Psi\mathcal{E}^\flat - \nabla A(\theta^*)\mathcal{E}^\flat, \tag{13}$$

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

$$\|D(\tilde\theta^\flat - \tilde\theta) - \xi^\flat\| \le \diamondsuit^\flat(x) \tag{14}$$

One can see that the standartized score in real world is

$$\xi = D^{-1}\Psi\epsilon \tag{15}$$

and the standartized score in bootstrap world is

$$\xi^\flat = D^{-1}\left[\mathbf{Y}^T\Psi - \nabla A(\theta^*)\right]\mathcal{E}^\flat | \mathbf{Y} \sim \mathcal{N}(0, V^2), \tag{16}$$

where $V^2 \stackrel{\text{def}}{=} \mathbb{V}ar(\xi^\flat)$. The latter is based on the fact that 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.

2.3 Gaussian approximation (GAR) for weighted sum

So far we have used the Fisher expansion in order to approximate the standardized score $D(\tilde{\theta} - \theta^*)$ with normalized score ξ with respect to the measure $\mathbb{P}$. Analogically we have approximated $D(\tilde{\theta}^\flat - \tilde{\theta})$ with $\xi^\flat$, the counterparts of standardized and normalized scores with respect to $\mathbb{P}^\flat$.

$\xi^\flat$ is indeed Gaussian by construction, hence there is no need to approximate it, while the vector $\xi = D^{-1}\Psi\epsilon$ is the linear combination of the original errors $\epsilon_i \stackrel{\text{def}}{=} Y_i - \mathbb{E}Y_i$. So, denote the Gaussian approximation of ξ as $\overline{\xi}$, which is a centered normal random variable with variance $\mathbb{V}ar(\xi)$, i.e.

$$\overline{\xi} \sim \mathcal{N}(0, \mathbb{V}ar(\xi)), \tag{17}$$

where $\mathbb{V}ar(\xi) = D^{-1}\Psi\mathbb{V}ar(\epsilon)\Psi^T D^{-1}$.

To pose the problem in more understandable way we recall that our errors ϵ_is are independent with zero mean and variance of v_i^2, i.e.

$$\mathbb{E}\epsilon_i = 0 \qquad \mathbb{V}ar\epsilon_i = v_i^2. \tag{18}$$

We introduce the independent and normal random variables $\tilde{\epsilon}_i$ with $\mathbb{V}ar\tilde{\epsilon}_i = v_i^2$ and define $\overline{\xi}$ as follows

$$\overline{\xi} \stackrel{\text{def}}{=} D^{-1}\Psi\tilde{\epsilon}, \tag{19}$$

where $\tilde{\epsilon} \sim \mathcal{N}\big(0, \operatorname{diag}(v_1^2, \ldots, v_n^2)\big)$. It is easy to check that the covariance matrices of ξ and $\overline{\xi}$ coincide.

Define also S_k as follows

$$S_k \stackrel{\text{def}}{=} \sum_{i=1}^{k-1} c_k\epsilon_k + \sum_{i=k+1}^{n} c_k\tilde{\epsilon}_k \quad \text{for } k = 2, \ldots, n, \tag{20}$$

where c_k are the coefficients of error terms, and

$$S = \sum_{i=1}^{n} c_k\epsilon_i, \qquad \tilde{S} = \sum_{i=1}^{n} c_k\tilde{\epsilon}_i. \tag{21}$$

Let f is a smooth function with

$$\left| f(x+d) - f(x) - d^T f'(x) - \frac{d^T f''(x) d}{2} \right| \leq C_f \|d\|_\infty^3. \tag{22}$$

Then, using the telescopic principle one can see that

$$f(S) - f(\tilde{S}) = \sum_{k=1}^{n} \big(f(S_k + c_k\epsilon_k) - f(S_k + c_k\tilde{\epsilon}_k) \big), \tag{23}$$

since $S_k + \epsilon_k = S_{k+1} + \tilde{\epsilon}_{k+1}$ for $k = 1, \ldots, n-1$.

Now, using Taylor expansion for every k and condition (22) we get $\forall k$

$$\left| f(S_k + c_k\epsilon_k) - f(S_k + c_k\tilde{\epsilon}_k) - c_i^T f'(S_k)(\epsilon_k - \tilde{\epsilon}_k) - c_i^T f''(S_k)c_i \frac{\epsilon_k^2 - \tilde{\epsilon}_k^2}{2} \right| \leq C_f \|c_i\|_\infty^3 \cdot (|\epsilon_k|^3 + |\tilde{\epsilon}_k^3|).$$

Taking the expected value inside the module and using the fact that $\mathbb{E}(\epsilon_i - \tilde{e}_i) = \mathbb{E}(\epsilon_i^2 - \tilde{e}_i^2) = 0$ along with independence of S_k and ϵ_k and $\tilde{\epsilon}_k$ one gets

$$\left|\mathbb{E}f(S) - \mathbb{E}f(\tilde{S})\right| = \left| \mathbb{E}\left[\sum_{k=1}^{n} \left(f(S_k + \epsilon_k) - f(S_k + \tilde{\epsilon}_k) \right) \right] \right| \leq C_f \|c_i\|_\infty^3 \cdot \mathbb{E} \sum_{k=1}^{n} \left(\left|\epsilon_k\right|^3 + \left|\tilde{\epsilon}_k\right|^3 \right).$$

Let

$$\delta_n = \|c_i\|_\infty^3 \mathbb{E} \sum_{k=1}^{n} \left(\left|\epsilon_k\right|^3 + \left|\tilde{\epsilon}_k\right|^3 \right) = \|c_i\|_\infty^3 \mathbb{E} \sum_{k=1}^{n} \left(\left|\epsilon_k\right|^3 + 2\sqrt{\frac{2}{\pi}} v_i^3 \right), \tag{24}$$

which yields

$$\left|\mathbb{E}f(S) - \mathbb{E}f(\tilde{S})\right| \leq C_f \cdot \delta_n. \tag{25}$$

Remark 2.1 *The inequality* (25) *can be easily extended in this form*

$$\left|\mathbb{E}f(S - z) - \mathbb{E}f(\tilde{S} - z)\right| \leq C_f \cdot \delta_n. \tag{26}$$

The basic idea for getting the distance between two distributions from here is to approximate the discontinuous function $f(x) = \mathbb{I}(x \geq q)$ by a smooth function $f(x)$ and then to apply the Lindeberg telescopic sum device.

if we construct such function, then the the theorem will look like something like this:

Theorem 2.1 *Let ϵ_i be independent zero mean with finite third moments and $\mathbb{V}ar\epsilon_i = v_i^2$ for $i = 1, \ldots, n$. Define*

$$S = \sum_{i=1}^{n} c_i\epsilon_i \quad \text{and} \quad \tilde{S} = \sum_{i=1}^{n} c_i\tilde{\epsilon}_i, \tag{27}$$

where $\tilde{\epsilon}_i \sim \mathcal{N}(0, v_i^2)$. Then, for any $\triangle > 0$ it holds

$$\left| \mathbb{P}(S \geq q) - \mathbb{P}(\tilde{S} \geq q + \triangle) \right| \leq \diamondsuit(\triangle, \delta_n, x, p), \tag{28}$$

where δ_n is given by (24)*, x is our tolerance level and p is the dimension of c_i.*

2.4 Gaussian comparison & Pinsker's inequality

As we have already showed that ξ can be replaced by $\overline{\xi}$, we now are ready to compare two Gaussian random vectors, namely

$$\overline{\xi} \sim \mathcal{N}(0, \mathbb{V}ar(\xi)), \quad \text{and} \quad \xi^\flat \sim \mathcal{N}(0, V^2). \tag{29}$$

It is sufficient to compare the covariance matrices of these two distributions. The main idea here relies under the Pinkser's inequality.

Consider two random Gaussian vectors with zero mean, namely $\xi \sim \mathcal{N}(0, \Sigma)$ and $\xi^\flat \sim \mathcal{N}(0, \Sigma^\flat)$. Let's discuss the in the most general setting. Let T be a function from $\mathbb{R}^p$ to $\mathbb{R}^q$ and $\mathbf{X} = T(\xi)$ and $\mathbf{Y} = T(\xi^\flat)$. We are interested in bounding the distance between distributions of $\mathbf{X}$ and $\mathbf{Y}$.

The next theorem bounds from above the Kullback-Leibler divergence between two normal distributions.

Theorem 2.2 *Let $\mathbb{P} = \mathcal{N}(\mu, \Sigma)$ and $\mathbb{P}^\flat = \mathcal{N}(\mu^\flat, \Sigma^\flat)$ for some non-degenerate symmetric and positive definite matrices. Assume that*

$$\|\Sigma^{-1/2}\Sigma^\flat\Sigma^{-1/2} - I_p\|_{op} \le \frac{1}{2} \tag{30}$$

and

$$tr\left(\Sigma^{-1/2}\Sigma^\flat\Sigma^{-1/2} - I_p\right)^2 \le \triangle^2, \tag{31}$$

then

$$\mathcal{D}(\mathbb{P} \,\|\, \mathbb{P}^\flat) = -\mathbb{E}\log\frac{d\mathbb{P}^\flat}{d\mathbb{P}} \le \frac{\triangle^2}{2} + \frac{1}{2}(\mu - \mu^\flat)^T\Sigma^{-1}\Sigma^\flat\Sigma^{-1}(\mu - \mu^\flat). \tag{32}$$

Moreover, for any measurable set $A \subset \mathbb{R}^p$ it holds

$$\left|\mathbb{P}(A) - \mathbb{P}^\flat(A)\right| \le \sqrt{\frac{\mathcal{D}(\mathbb{P} \,\|\, \mathbb{P}^\flat)}{2}}. \tag{33}$$

Proof. By the following change of variable $u = \Sigma^{-1/2}(x - \mu)$ one can achieve the standard normal Gaussian vector $\mathbb{P}$, while for $\mathbb{P}^\flat$ we have the following expression

$$\mathbb{P}^\flat = \mathcal{N}\Big(\Sigma^{-1/2}(\mu^\flat - \mu), \Sigma^{-1/2}\Sigma^\flat\Sigma^{-1/2}\Big) = \mathcal{N}(b, B), \tag{34}$$

where

$$b \stackrel{\text{def}}{=} \Sigma^{-1/2}(\mu^\flat - \mu), \qquad B \stackrel{\text{def}}{=} \Sigma^{-1/2}\Sigma^\flat\Sigma^{-1/2} \tag{35}$$

Compute

$$2\log\frac{d\mathbb{P}^\flat}{d\mathbb{P}}(\gamma) = \log\det(B) - (\gamma - b)^T B(\gamma - b) + \|\gamma\|^2, \tag{36}$$

where γ is standard normal random vector. Then taking the expectation of (36) one can easily get

$$\mathcal{D}(\mathbb{P} \,\|\, \mathbb{P}^\flat) = -\mathbb{E} \log \frac{d\mathbb{P}^\flat}{d\mathbb{P}} = -\frac{1}{2} \left[\log\det(B) - \mathrm{tr}(B - I_p) - b^T B b \right]. \tag{37}$$

Let λ_j is the jth eigenvalue of matrix $B - I_p$ then (37) could be rewritten in this way

$$\mathcal{D}(\mathbb{P} \,\|\, \mathbb{P}^\flat) = \frac{1}{2} \left[\sum_{j=1}^{p} \Big\{ -\log(\lambda_j + 1) + \lambda_j \Big\} + b^T B b \right]. \tag{38}$$

Then, we note that for $x \in (0, 1/2)$ it holds

$$f(x) = x^2 - x + \log(x+1) \geq 0, \quad \forall\, x \in \left(0, \frac{1}{2}\right). \tag{39}$$

The simple algebra verifies this. Hence,

$$\frac{1}{2} \left[\sum_{j=1}^{p} \Big\{ -\log(\lambda_j + 1) + \lambda_j \Big\} + b^T B b \right] \leq \frac{1}{2} \left[\sum_{j=1}^{p} \lambda_j^2 + b^T B b \right] \leq \frac{1}{2} b^T B b + \frac{1}{2} \triangle^2, \tag{40}$$

where the latter is obtained using the condition (31). Then substituting back to the original variables we get

$$\mathcal{D}(\mathbb{P} \,\|\, \mathbb{P}^\flat) = -\mathbb{E} \log \frac{d\mathbb{P}^\flat}{d\mathbb{P}} \leq \frac{\triangle^2}{2} + \frac{1}{2} (\mu - \mu^\flat)^T \Sigma^{-1} \Sigma^\flat \Sigma^{-1} (\mu - \mu^\flat) \tag{41}$$

as required.

Applying Pinsker's inequality we obtain

$$\sup_A \left| \mathbb{P}(A) - \mathbb{P}^\flat(A) \right| \leq \sqrt{\frac{\mathcal{D}(\mathbb{P} \,\|\, \mathbb{P}^\flat)}{2}} \leq \frac{1}{2} \sqrt{\triangle^2 + (\mu - \mu^\flat)^T \Sigma^{-1} \Sigma^\flat \Sigma^{-1} (\mu - \mu^\flat)} \tag{42}$$

■

Remark 2.2 *Note that* $1/2$ *from the left hand side of* (30) *can be replaced by some constant* ϵ*, which will lead to the similar result parametrized on* ϵ*.*

Remark 2.3 *Note that for the case of* $\mu = \mu^\flat = 0$ *and* $T(x) = x$ *mapping one obtains*

$$\sup_{A \in \mathcal{B}(\mathbb{R}^p)} \left| \mathbb{P}(\xi \in A) - \mathbb{P}^\flat(\xi^\flat \in A) \right| \leq \sqrt{\frac{\mathcal{D}(\mathbb{P} \,\|\, \mathbb{P}^\flat)}{2}} \leq \frac{\triangle}{2}, \tag{43}$$

where $\mathcal{B}(\mathbb{R}^p)$ *is a set of all Borel sets in* $\mathbb{R}^p$*.*

From the above theorem it follows that in order to show that the distributions (i.e. covariance matrices) are close to each other, it is sufficient to show that the conditions of the theorem fulfil.

3 Conditions

Describe the conditions that are needed in order to apply the above described steps.

4 Main

bad structure

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

$$L^\flat(v) = \sum_{i=1}^{n} \ell_i(v|\mathbf{Y}) \cdot w_i^\flat, \tag{44}$$

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 (47) is the function of random data $\mathbf{Y}$, while the log-likelihood in (44) 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

$$\mathbb{E}^\flat L^\flat(v) \equiv L(v) \tag{45}$$

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

$$v(x) = \sum_{j=1}^{p+q} \theta_j \psi_j(x) = \sum_{j=1}^{p} \theta_j \psi_j(x) + \sum_{j=p+1}^{p+q} \theta_j \psi_j(x) \text{ for given basis } \psi_j(\cdot). \tag{46}$$

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

$$L(v; \mathbf{Y}) = \log \frac{d\mathbb{P}_v}{d\mu_0^n}(\mathbf{Y}) = \sum_{i=1}^{n} v_i Y_i - g(v_i), \tag{47}$$

where v is the parameter of interest. Define

$$\tilde{v} = \arg\max_{v \in \Theta} L(v, \mathbf{Y}) \qquad v^* = \arg\max_{v \in \Theta} \mathbb{E}L(v, \mathbf{Y}) \tag{48}$$

to be maximum likelihood estimator and oracle, respectively.