Bootstrap in high dimensional spaces

DISSERTATION

zur Erlangung des akademischen Grades

Doctor Rerum Naturalium
(Dr. rer. nat.)
im Fach Mathematik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
Humboldt-Universität zu Berlin

von
Nazar Buzun

Präsidentin der Humboldt-Universität zu Berlin:
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät:
Prof. Dr. Elmar Kulke

Gutachter:
1. Prof. Dr. Vladimir Spokoiny, Humboldt-Universität zu Berlin
2. Dr. Alexey Naumov, Higher School of Economics in Moscow
3. Prof. Dr. Thorsten Dickhaus, University of Bremen

Tag der mündlichen Prüfung: 11.12.2020

Declaration:

I declare that I have completed the thesis independently using only the aids and tools specified. I have not applied for a doctor's degree in the doctoral subject elsewhere and do not hold a corresponding doctor's degree. I have taken due note of the Faculty of Mathematics and Natural Sciences PhD Regulations, published in the Official Gazette of Humboldt-Universität zu Berlin no. 42 on July 11 2018.

Berlin, December 19, 2020 Nazar Buzun

Abstract

The objective of this thesis is to explore theoretical properties of various bootstrap methods. We introduce the convergence rates of the bootstrap procedure which corresponds to the difference between real distribution of some statistic and its resampling approximation. In this work we analyze the distribution of Euclidean norm of independent vectors sum, maximum of sum in high dimension, Wasserstein distance between empirical measures, Wassestein barycenters. In order to prove bootstrap convergence we involve Gaussian approximation technique which means that one has to find a sum of independent vectors in the considered statistic such that bootstrap yields a resampling of this sum. Further this sum may be approximated by Gaussian distribution and compared with the resampling distribution as a difference between variance matrices.

In general it appears to be very difficult to reveal such a sum of independent vectors because some statistics (for example, MLE) don't have an explicit equation and may be infinite-dimensional. In order to handle this difficulty we involve some novel results from statistical learning theory, which provide a finite sample quadratic approximation of the Likelihood and suitable MLE representation. In the last chapter we consider the MLE of Wasserstein barycenters model. The regularised barycenters model has bounded derivatives and satisfies the necessary conditions of quadratic approximation.

Furthermore, we apply bootstrap in change point detection methods. In the parametric case we analyse the Likelihood Ratio Test (LRT) statistic. Its high values indicate changes of parametric distribution in the data sequence. The maximum of LRT has a complex distribution but its quantiles may be calibrated by means of bootstrap. We show the convergence rates of the bootstrap quantiles to the real quantiles of LRT distribution. In non-parametric case instead of LRT we use Wasserstein distance between empirical measures. We test the accuracy of change point detection methods on synthetic time series and electrocardiography (ECG) data. Experiments with ECG illustrate advantages of the non-parametric approach versus complex parametric models and LRT.

Keywords: Bootstrap, Gaussian approximation, Wasserstein distance, change point detection.

Zusammenfassung

Ziel dieser Arbeit ist theoretische Eigenschaften verschiedener Bootstrap Methoden zu untersuchen. Als Ergebnis führen wir die Konvergenzraten des Bootstrap-Verfahrens ein, die sich auf die Differenz zwischen der tatsächlichen Verteilung einer Statistik und der Resampling-Näherung beziehen.

In dieser Arbeit analysieren wir die Verteilung der l2-Norm der Summe unabhängiger Vektoren, des Summen Maximums in hoher Dimension, des Wasserstein-Abstands zwischen empirischen Messungen und Wassestein-Barycenters. Um die Bootstrap-Konvergenz zu beweisen, verwenden wir die Gaussche Approximations technik. Das bedeutet dass man in der betrachteten Statistik eine Summe unabhängiger Vektoren finden muss, so dass Bootstrap eine erneute Abtastung dieser Summe ergibt. Ferner kann diese Summe durch Gaussche Verteilung angenähert und mit der Neuabtastung Verteilung als Differenz zwischen Kovarianzmatrizen verglichen werden.

Im Allgemeinen scheint es sehr schwierig zu sein, eine solche Summe unabhängiger Vektoren aufzudecken, da einige Statistiken (zum Beispiel MLE) keine explizite Gleichung haben und möglicherweise unendlich dimensional sind. Um mit dieser Schwierigkeit fertig zu werden, verwenden wir einige neuartige Ergebnisse aus der statistischen Lerntheorie.

Darüber hinaus wenden wir Bootstrap bei Methoden zur Erkennung von Änderungspunkten an. Im parametrischen Fall analysieren wir den statischen Likelihood Ratio Test (LRT). Seine hohen Werte zeigen Änderungen der Parameter Verteilung in der Datensequenz an. Das Maximum von LRT hat eine unbekannte Verteilung und kann mit Bootstrap kalibriert werden. Wir zeigen die Konvergenzraten zur realen maximalen LRT-Verteilung. In nicht parametrischen Fällen verwenden wir anstelle von LRT den Wasserstein-Abstand zwischen empirischen Messungen. Wir testen die Genauigkeit von Methoden zur Erkennung von Änderungspunkten anhand von synthetischen Zeitreihen und Elektrokardiographiedaten. Letzteres zeigt einige Vorteile des nicht parametrischen Ansatzes gegenüber komplexen Modellen und LRT.

Schlagworter: Bootstrap, Gaussche Näherung, Wasserstein-Entfernung, Änderungspunkterkennung.

Contents

List of abbreviations

$\langle \varOmega, \mathcal{F}, \mathbb{P} \rangle$ probability space ($\varOmega$ – set of outcomes, $\mathcal{F}$ – σ-algebra, $\mathbb{P}$ – probability measure);

$p(x)$ distribution density of a random variable X;

$\mathbb{E}_{\mathbb{P}}$ mathematical expectation by measure $\mathbb{P}$;

$\mathbb{P}^{\flat}$ bootstrap probability; the bootstrap weights usually denoted by $w^{\flat}$;

$\mathbb{E}^{\flat}$ bootstrap mathematical expectation;

$\mathcal{N}(m, \sigma^2)$ normal distribution;

$\mathcal{P}o(\lambda)$ Poisson distribution;

$\langle \cdot, \cdot \rangle$ scalar product;

$\| \cdot \|$ the second norm (l_2) or operator norm;

$\mathbb{1}$ indicator function:

$$\mathbb{1}(\text{true}) = [\text{true}] = 1, \quad \mathbb{1}(\text{false}) = [\text{false}] = 0;$$

$\sim$ a sample with distribution density p :

$$X_1, \ldots, X_n \sim p;$$

In similar situations the symbol $\in$ is used instead of $\sim$ (for example, $X \in \mathcal{N}(m, \sigma^2)$), $\sim$ also denotes proportionality;

$\mathbb{E}_X$ mathematical expectation by variable X:

$$\mathbb{E}_X h(X, Y) = \int_{-\infty}^{+\infty} h(x, Y) dF(x)$$

$\mathcal{KL}$ Kullback–Leibler divergence:

$$\mathcal{KL}(p_1 \| p_2) = \int_{\varOmega} \log \left(\frac{p_1(x)}{p_2(x)} \right) p_1(x) dx$$

W_p Wassestein distance with metric l_p:

$$W_p(\mu_1, \mu_2) = \min_{\pi \in \varPi[\mu_1, \mu_2]} \left(\int \|x - y\|^p d\pi(x, y) \right)^{1/p}$$

Z Gaussian random vector;

C_A anti-concentration constant;

L Likelihood function with dataset $\mathbb{Y}$:

$$L(\theta) = L(\theta, \mathbb{Y})$$

$L^\flat$ bootstrap Likelihood function with dataset $\mathbb{Y}$:

$$L^\flat(\theta) = \sum_i w_i^\flat l_i(\theta, Y_i)$$

ζ stochastic Likelihood function component:

$$\zeta = L - \mathbb{E}L$$

D^2 Fisher matrix:

$$D^2 = -\nabla^2 \mathbb{E}L(\theta^*)$$

h sliding window size;

n dataset size;

$T_h(t)$ test statistic for window position t and size h;

MLE maximum Likelihood estimation:

$$\widehat{\theta} = \operatorname*{argmax}_\theta L(\theta)$$

Reference value:

$$\theta^* = \operatorname*{argmax}_\theta \mathbb{E}L(\theta)$$

CLT Central Limit Theorem;

LRT Likelihood ratio test;

1 Introduction

The bootstrap is a very effective and practical tool for confidence interval estimation, hypotheses testing and models ensemble composition. There are many types of the bootstrap and each has its own specifics (ref. Horowitz (2003), Bucher and Dette (2013), Lahiri (2013)). They were designed for parametric and non-parametric statistical models and may handle both independent and dependent data. Besides that there is a big gap between experimental precision and theoretical grounds for this procedure. While showing a good practical results the bootstrap also requires a rigorous theoretical justification.

Proving the consistency of bootstrap procedure means that one has to estimate the difference by distribution between some statistic depending on a random dataset and corresponded bootstrap statistic. One may sample the bootstrap statistic multiple times and thus approximate distribution of the original statistic. In recent years many interesting works have appeared on this topic with a motivation to reduce this gap between practice and theory. Spokoiny and Zhilova (2015) investigates independent parametric models with bootstrap weights. They prove a finite sample bound for difference in distribution between maximum likelihood estimation and maximum argument of the weighted models. The convergence rate in this paper is $(p^3/n)^{1/8}$, where p is parameter dimension and n is dataset size. Further we will discuss this result with more details and will show that it may be improved to $(p^2/n)^{1/2}$. However, it is still not applicable in high dimensional setting when $n > p$. In papers V. Chernozhukov (2014), Chernozhukov et al. (2013b) the authors study the infinite norm of high dimensional random vectors sum from which one can derive the bootstrap convergence with rate $(\log^7 p/n)^{1/6}$. This rate is not optimal (the lower bound is $(\log^3 p/n)^{1/2}$) and improving this result is a very challenging task. Naumov et al. (2019) considers bootstrap procedure in application to covariance matrices of Gaussian random vectors and their spectral projectors. They show that in high dimensions instead of parameter p one may treat the spectrum of covariance matrix.

In this research we focus on sparse models with different types of regularisation. In most cases the regularisation term is l_1 norm (lasso). It allows to zero insignificant components of the high dimensional model parameter and for the nonzero part of interest we further prove the bootstrap consistency. For the other regularisation types we consider the projection of the parameter into a low dimensional space and restrict entropy of the full space. Moreover, we extend bootstrap theory for the composition of multiple models that share one dataset and then apply it in change point detection task.

Consider a likelihood function $L(\theta)$ with independent observations $\mathbb{Y} = (Y_1, \dots, Y_n)$ and parameter θ:

$$L(\theta) = \sum_{i=1}^{n} l_i(\theta, Y_i).$$

By means of the bootstrap procedure in application to statistical models one may sample MLE values of the parameter θ. It could be done in two ways. The first variant is *weighted* bootstrap, where we multiply the components of likelihood by random weights. And in the second variant we sample random indices from the dataset with repetition, which is called *empirical* bootstrap. At each resampling iteration we obtain a new value of the optimal parameter. In the end we obtain the empirical distribution. Define the likelihood of the weighted bootstrap as

$$L^\flat(\theta) = \sum_{i=1}^{n} l_i(\theta, Y_i) w_i^\flat,$$

where $w_i^\flat$ are independent random weights, such that $\mathbb{E} w_i^\flat = \operatorname{Var} w_i^\flat = 1$. The empirical bootstrap likelihood is

$$L^\epsilon(\theta) = \sum_{i=1}^{n} l_{k[i]}(\theta, Y_{k[i]}),$$

where $k[i]$ are independent random indices from set $\{1, \dots, n\}$.

Our final goal is to prove bootstrap consistency for various examples of parametric and non-parametric models with independent observations. In Figure 1 it is the last block of the diagram. And it means that one has to find the difference by distribution between the deviation of the correspondent MLE parameters or analogically between values of the likelihood function. We apply the prevalent approach for the theoretical justification (ref. Spokoiny and Zhilova (2015), V. Chernozhukov (2014)). It consists of two steps. First we approximate the deviation of the parameter by the sum of independent random vectors $\boldsymbol{\xi}_i$. It involves quadratic approximation method from paper Spokoiny (2012a). Also we show that in analogical sum in the bootstrap case the summands are additionally multiplied by the random weights. In the second step we have to find the upper bound in multivariate normal approximation which is a generalisation of the classical Berry–Esseen theorem (ref. Bentkus (2003a)). Particularly we suppose that $\sum_{i=1}^n \boldsymbol{\xi}_i$ converges to some normal vector Z. The same is true for the bootstrap sum. And finally one has to compare two normal distributions with covariance matrix $\varSigma$ and its empirical estimation $\widehat{\varSigma}$.

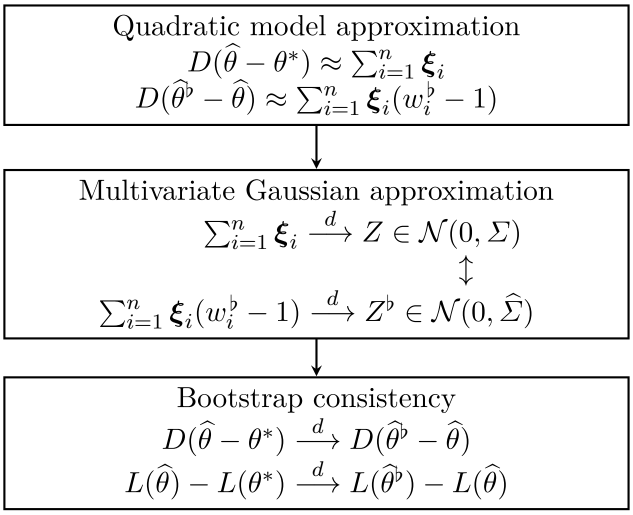

Quadratic model approximation
$D(\widehat{\theta} - \theta^*) \approx \sum_{i=1}^n \boldsymbol{\xi}_i$
$D(\widehat{\theta}^\flat - \widehat{\theta}) \approx \sum_{i=1}^n \boldsymbol{\xi}_i (w_i^\flat - 1)$
Multivariate Gaussian approximation
$\sum_{i=1}^n \boldsymbol{\xi}_i \xrightarrow{d} Z \in \mathcal{N}(0, \varSigma)$
$\updownarrow$
$\sum_{i=1}^n \boldsymbol{\xi}_i (w_i^\flat - 1) \xrightarrow{d} Z^\flat \in \mathcal{N}(0, \widehat{\varSigma})$
Bootstrap consistency
$D(\widehat{\theta} - \theta^*) \xrightarrow{d} D(\widehat{\theta}^\flat - \widehat{\theta})$
$L(\widehat{\theta}) - L(\theta^*) \xrightarrow{d} L(\widehat{\theta}^\flat) - L(\widehat{\theta})$

Figure 1. Bootstrap justification technique. The first and second blocks correspond to the main steps in the proof. From them the last block follows. The sign "d" denotes convergence by distribution. $\boldsymbol{\xi}_i = D^{-1} \nabla l_i(\theta^*)$, $\quad \varSigma = \sum_{i=1}^n \operatorname{Var} \boldsymbol{\xi}_i$, $\quad \widehat{\varSigma} = \sum_{i=1}^n \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T$.

In the following list the main contributions are selected:

- Bootstrap consistency for high dimensional models. Among them are barycenters model and different models with lasso regularization;
- Bootstrap consistency in change point detection with likelihood ratio test;
- Improvements in multivariate Gaussian approximation;
- Implementation and experimental study of parametric and non-parametric algorithms for change point detection.

The thesis structure is organized as follows. Chapter 2 includes useful technical results from probability theory. We prove here Gaussian approximation using Stein's method, show how the correspondent bound depends on the norm type, discuss Levy's concentration function and the comparison of two Gaussian random vectors. Chapter 3 provides statistical semi parametric setup and states a number of theorems about finite sample quadratic approximation of the likelihood in both in "real" and "bootstrap" settings. These results allow to derive the main theorem about the bootstrap consistency in general case and we specify them for two particular examples: generalized linear models and sparse models with lasso regularization. Chapter 4 is devoted to the problem of change-point detection. We propose a new method in which Likelihood ratio test (LRT) is sequentially applied in a sliding window procedure. Its high values indicate changes in parametric distribution of the data sequence. Obviously the LRT values require a predefined threshold for their maximum. The maximum value has unknown distribution and may be calibrated by the bootstrap. It enables to estimate empirically the LRT distribution. We obtain the convergence rates of the "bootstrap" quantiles to the "real" quantiles of the LRT distribution. Moreover we extend the proposed method to non-parametric models, where instead of LRT we use Wasserstein distance between empirical measures. We evaluate the accuracy of change point detection methods on synthetic time series and electrocardiography data. Chapter 5 deals with maximum likelihood estimation of Wasserstein barycenters model. Basing on representation of Wasserstein distance in Fourier basis and theory of support functions we obtain the necessary conditions of quadratic approximation from Chapter 3. Chapter 6 collects some known useful results from random matrix theory and about deviations of sub-Gaussian random vectors.

2 Gaussian approximation

2.1 Generator approach

Let A be generator of a Markov process X_t with stationary distribution μ. Note that the necessary and sufficient condition for stationary distribution is $\mathbb{E}Af(X) = 0$, $X \sim \mu$, for all f where A is defined. Remind that by definition

$$T_t f(x) = \mathbb{E}[f(X_t)|X_0 = x]$$

$$A = \frac{d}{dt}T_t$$

Consider some examples:

1. Normal stationary distribution $\mathcal{N}(0, \Sigma)$

$$Af(x) = \operatorname{tr}\{\Sigma \nabla^2 f(x)\} - x^T \nabla f(x) \tag{A$\mathcal{N}$}$$

2. Poisson stationary distribution $\mathcal{P}o(\lambda)$

$$Af(x) = \lambda f(x+1) - xf(x) \tag{A$\mathcal{P}o$}$$

3. Gamma stationary distribution $\Gamma(r, \lambda)$

$$Af(x) = xf''(x) + (r - \lambda x)f'(x) \tag{AΓ}$$

Assume one has to find a limit distribution of $\Phi_n(X_1, \dots, X_n)$, $n \to \infty$, where $X_1, \dots, X_n$ are independent and $\mathbb{E}X_i = 0$. Construct Markov chain by means of exchangeable pairs:

1. Start with $Z_n(0) = (X_1, \dots, X_n)$

2. Pick index $I \in \{1, \dots, n\}$ uniformly at random; replace X_I by its independent copy X_I^*:

$$Z_n(1) = (X_1, \dots, X_{I-1}, X_I^*, X_{I+1}, \dots, X_n)$$

3. Continue the chain for $k \in \mathbb{N}$, $k > 1$

$$Z_n(k) = (Z_n(k-1)[1 : I-1], X_I^*, Z_n(k-1)[I+1 : n])$$

4. Make time continuous for $t \in \mathbb{R}_+$

$$Z_n(t) = Z_n(\mathcal{P}o(t))$$

The generator of Markov process $Z_n(nt)$ is

$$\begin{aligned} A_n f(x) &= n(\mathbb{E}[f(Z(1))|Z(0)=x] - f(x)) \\ &= n\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{E} f(x_1,\dots,x_{i-1},X_i,x_{i+1},\dots,x_n) - f(x)\right) \end{aligned}$$

The generator of Markov process $\Phi_n(Z_n(nt))$ is $A_n f(\Phi_n(x))$. Make its Taylor expansion

$$\begin{aligned} A_n f(\Phi_n(x)) &= -\nabla^T f(\Phi_n(x)) \sum_{i=1}^{n} \left(\frac{\partial \Phi_n}{\partial x_i}\right)^T x_i \\ &+\frac{1}{2}\operatorname{tr}\left\{\nabla^2 f(\Phi_n(x)) \sum_{i=1}^{n} \left(\frac{\partial \Phi_n}{\partial x_i}\right)^T (\mathbb{E} X_i X_i^T + x_i x_i^T) \left(\frac{\partial \Phi_n}{\partial x_i}\right)\right\} \\ &+\frac{1}{2}\operatorname{tr}\left\{\nabla^T f(\Phi_n(x)) \sum_{i=1}^{n} \left(\frac{\partial^2 \Phi_n}{\partial x_i \partial x_i}\right) (\mathbb{E} X_i X_i^T + x_i x_i^T)\right\} \\ &+o(1) \end{aligned}$$

The last equation provides a constructive heuristic of finding generator corresponded to the limit distribution taking into account only the first and second derivatives of Φ_n.

Let μ be limit distribution of $\Phi_n(X_1,\dots,X_n)$, $n \to \infty$, then μ is a stationary distribution of $\Phi_n(Z_n(nt))$, $n \to \infty$, and

$$\mathbb{E} A_n f(\Phi_n(X)) \to 0$$

For example, when $\Phi_n(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i$:

$$\begin{aligned} A_n f(\Phi_n(x)) &= -\nabla^T f(\Phi_n(x))\Phi_n(x) \\ &+\frac{1}{2}\operatorname{tr}\left\{\nabla^2 f(\Phi_n(x)) \frac{1}{n}\sum_{i=1}^{n} (\mathbb{E} X_i X_i^T + x_i x_i^T)\right\} \\ &+o(1) \end{aligned}$$

By the Law of Large Numbers for the variance matrix Σ of the limit distribution

$$\frac{1}{2n}\sum_{i=1}^{n} (\mathbb{E} X_i X_i^T + x_i x_i^T) \to \Sigma$$

and $A_n f(\Phi_n(x))$ converges to (A$\mathcal{N}$).

Generators may not only characterize the stationary measure but also reveal the distance to it. Let the distance between random variables be defined as

$$\sup_{h\in\mathcal{H}} |\mathbb{E} h(X) - \mathbb{E} h(Z)|$$

Then setting

$$Af_h(x) = h(x) - \mathbb{E}h(Z)$$

under assumption $\mathbb{E}Af(Z) = 0$ one has to solve the last equation and estimate $|\mathbb{E}Af_h(X)|$. Further we concentrate attention on (A$\mathcal{N}$) and the distance to Normal distribution.

2.2 Berry – Esseen Theorem

Multivariate analogues of Berry – Esseen Theorem have many modifications depending on space dimension of random vectors and functions set used for measures comparison. V. Bentkus in his papers Bentkus (2003b), Bentkus (2003a) has presented excellent results related to this topic. Namely, for a sequence of i.i.d random vectors with identity covariance matrix $\{X_i\}_{i=1}^n$, $\Omega(X_i) \in \mathbb{R}^p$, a convex set A and Gaussian vector $Z \in \mathcal{N}(0, I)$ it holds

$$\left| \mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \in A \right) - \mathbb{P}\left(Z \in A \right) \right| \leq \frac{400 p^{1/4} \mathbb{E}\|X_1\|^3}{\sqrt{n}}$$

and

$$W_1\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i, Z \right) \leq O\left(\frac{\mathbb{E}\|X_1\|^3}{\sqrt{n}} \right)$$

We extend these two statements for independent random vectors with non-identity covariance Σ. Additionally we remove factor $p^{1/4}$ replacing it with anti-concentration constant defined below.

Def **(H_k).** The multivariate Hermite polynomial H_k is defined by formula

$$H_k(x) = (-1)^{|k|} e^{x^T \Sigma^{-1} x/2} \frac{\partial^{|k|}}{\partial^{k_1} \ldots \partial^{k_p}} e^{-x^T \Sigma^{-1} x/2}$$

where $x \in R^p$ and $|k| = k_1 + \ldots + k_p$.

Lemma 2.1. Consider a Gaussian vector $Z \sim \mathcal{N}(0, \Sigma)$ and two functions h and f_h such that

$$f_h(x) = -\int_0^1 \mathbb{E}\overline{h}(Z(x,t))dt$$

$$\overline{h}(Z(x,t)) = h(\sqrt{t}x + \sqrt{1-t}Z) - \mathbb{E}h(Z)$$

Then f_h is a solution of the Stein's equation

$$\overline{h}(x) = (\mathrm{tr}\{\nabla^2 \Sigma\} - x^T \nabla) f_h(x)$$

and

$$\frac{\partial^{|k|}}{\partial^{k_1} \ldots \partial^{k_p}} f_h(x) = -\int_0^1 \frac{1}{2} \frac{t^{\frac{|k|}{2}-1}}{(1-t)^{\frac{|k|}{2}}} \mathbb{E}H_k(Z)\overline{h}(Z(x,t))dt$$

Consequence.

$$\nabla^2 f_h(x) - \nabla^2 f_h(y) = -\int_0^1 \frac{1}{2(1-t)} \mathbb{E} H_2(Z)\{h(Z(x,t)) - h(Z(y,t))\}dt$$

where

$$\begin{aligned} H_2(Z) &= (\varSigma^{-1}Z)(\varSigma^{-1}Z)^T - \varSigma^{-1} \\ &= \varSigma^{-1/2}\{(\varSigma^{-1/2}Z)(\varSigma^{-1/2}Z)^T - I\}\varSigma^{-1/2} \end{aligned}$$

Theorem 2.1 (Multivariate Berry–Esseen with Wasserstein distance)**.** Consider a sequence of independent zero-mean random vectors $X = \sum_{i=1}^n X_i$ in $\mathbb{R}^p$ with a covariance matrix

$$\mathbb{E} X X^T = \varSigma$$

Then 1-Wasserstein distance between X and Gaussian vector $Z \in \mathcal{N}(0, \varSigma)$ has following upper bound

$$W_1(X, Z) \le \sqrt{2}\mu_3 \left(1 + \log\left(2\sqrt{\operatorname{tr}\{\varSigma\}p}\right) - \log(\mu_3)\right)$$

where

$$\mu_3 = \sum_{i=1}^n \mathbb{E} X_i^T \varSigma^{-1} X_i \|X_i - X_i'\|$$

and each X_i' is an independent copy of X_i.

Remark. In i.i.d case with $\varSigma = I_p$

$$W_1(X, Z) = O\left(\frac{p^{3/2}\log(n)}{\sqrt{n}}\right)$$

These is the same theorem with a different proof in paper Bentkus (2003a).

Proof. Let θ be some value in $[0, 1]$ and

$$\mathbb{E}\bar{h}(X) = \mathbb{E} \operatorname{tr}\{\nabla^2 \varSigma\} f_h(X) - \mathbb{E} \sum_{i=1}^n X_i^T \nabla f_h(X)$$

$$= \sum_{i=1}^n \mathbb{E} X_i^T \left\{\nabla^2 f_h(X_{-i} + \theta X_i) - \nabla^2 f_h(X_{-i} + X_i')\right\} X_i$$

$$= \sum_{i=1}^n \mathbb{E}(\varSigma^{-1/2}X_i)^T \varSigma^{1/2} \left\{\nabla^2 f_h(X_{-i} + \theta X_i) - \nabla^2 f_h(X_{-i} + X_i')\right\} \varSigma^{1/2}\varSigma^{-1/2} X_i$$

For a unit vector $\|\gamma\| = 1$ and conditional expectation $\mathbb{E}_{-i} = \mathbb{E}(\cdot | X_i, X_i')$

$$\gamma^T \mathbb{E}_{-i} \varSigma^{1/2} \left\{\nabla^2 f_h(X_{-i} + \theta X_i) - \nabla^2 f_h(X_{-i} + X_i')\right\} \varSigma^{1/2} \gamma$$

$$
\begin{aligned}
&= \int_0^1 \frac{1}{2(1-t)} \mathbb{E}_{-i}\{((\Sigma^{-1/2}Z)^T\gamma)^2 - 1\}\{h(Z(X_{-i} + \theta X_i, t)) - h(Z(X_{-i} + X_i', t))\}dt \\
&\qquad\qquad \leq \int_0^{1-\alpha} \frac{t^{1/2}}{2(1-t)} Adt + \int_{1-\alpha}^1 \frac{1}{(1-t)^{1/2}} Bdt \\
&\qquad\qquad \leq -\frac{A}{2}\log(\alpha) + 2B\sqrt{\alpha} \\
&\qquad\qquad \leq A\left(1 + \log\left(\frac{2B}{A}\right)\right)
\end{aligned}
$$

$$
A = \|X_i - X_i'\| \, \mathbb{E}_{-i}|((\Sigma^{-1/2}Z)^T\gamma)^2 - 1| \, \|\nabla h(\sqrt{t}(X' + \theta_1 X_i - \theta_2 X_i' + \sqrt{1-t}Z)\|
$$

$$
B = \mathbb{E}_{-i}|((\Sigma^{-1/2}Z)^T\gamma)^2 - 1| \, \|Z\| \, \|\nabla h(\sqrt{t}(X' + \theta X_i + \theta_2\sqrt{1-t}Z)\|
$$

Account that $\|\nabla h(\cdot)\| \leq 1$ for W_1.

□

Theorem 2.2 (Multivariate Berry–Esseen)**.** Consider a sequence of independent zero-mean random vectors $X = \sum_{i=1}^n X_i$ in $\mathbb{R}^p$ with a covariance matrix

$$
\mathbb{E}XX^T = \Sigma
$$

Let a function $\varphi : \mathbb{R}^p \to \mathbb{R}_+$ be sub-additive:

$$
\varphi(x + y) \leq \varphi(x) + \varphi(y)
$$

and with Gaussian vector $Z \in \mathcal{N}(0, \Sigma)$ fulfills the anti-concentration property, such that

$$
\mathbb{P}(\varphi(Z) > x) - \mathbb{P}(\varphi(Z) > x + \Delta) \leq C_A\Delta
$$

Then the measure difference between X and Gaussian vector Z has the following upper bound $\forall x$

$$
|\mathbb{P}(\varphi(X) > x) - \mathbb{P}(\varphi(Z) > x)| \leq 22C_A\mu_3 \log\left(\frac{4p}{C_A\mu_3}\right) \log\left(\frac{\sqrt{2\mathbb{E}\varphi^2(Z)}p}{20C_A\mu_3^2}\right)
$$

where

$$
\mu_3 = \sum_{i=1}^n \mathbb{E}X_i^T\Sigma^{-1}X_i 2\varphi(X_i)
$$

Proof. Define a smooth indicator function

$$
g_{x,\Delta}(t) = \begin{cases} 0, & t < x \\ (x - t)/\Delta, & t \in [x - \Delta, x] \\ 1, & t > x \end{cases}
$$

Set $h = g_{x,\varDelta} \circ \varphi$. Denote the required bound by δ:

$$|\mathbb{P}(\varphi(X) > x) - \mathbb{P}(\varphi(Z) > x)| \le \delta$$

Note that from sub-additive property of the function φ follows

$$g_{x,\varDelta}(\varphi(X + dX)) \le g_{x,\varDelta}(\varphi(X) + \varphi(dX))$$

and

$$g'_{x,\varDelta}(t) = \frac{1}{\varDelta} \mathbb{I}[x - \varDelta < t < x]$$

and

$$\mathbb{E} g'_{x,\varDelta}(\varphi(Z)) = \frac{1}{\varDelta}(\mathbb{P}(\varphi(Z) > x - \varDelta) - \mathbb{P}(\varphi(Z) > x)) \le C_A$$

$$\mathbb{E} g'_{x,\varDelta}(\varphi(Z(X,t))) \le \frac{1}{\varDelta}(\mathbb{P}(\varphi(Z) > x - \varDelta) - \mathbb{P}(\varphi(Z) > x)) + \frac{2\delta}{\varDelta} \le C_A + \frac{2\delta}{\varDelta}$$

Assume X'_i is an independent copy of X_i and θ is some value in $[0,1]$ and

$$\mathbb{E}\overline{h}(X) = \mathbb{E}\operatorname{tr}\{\nabla^2 \varSigma\} f_h(X) - \mathbb{E}\sum_{i=1}^n X_i^T \nabla f_h(X)$$

$$= \sum_{i=1}^n \mathbb{E} X_i^T \left\{\nabla^2 f_h(X_{-i} + \theta X_i) - \nabla^2 f_h(X_{-i} + X'_i)\right\} X_i$$

$$= \sum_{i=1}^n \mathbb{E}(\varSigma^{-1/2} X_i)^T \varSigma^{1/2} \left\{\nabla^2 f_h(X_{-i} + \theta X_i) - \nabla^2 f_h(X_{-i} + X'_i)\right\} \varSigma^{1/2} \varSigma^{-1/2} X_i$$

According to the consequence of Lemma 2.1 one has to bound the following expression

$$\begin{aligned}
&\mathbb{E}_{-i} h(Z(X_{-i} + \theta X_i, t)) - \mathbb{E}_{-i} h(Z(X_{-i} + X'_i, t)) \\
&\le \mathbb{E}_{-i} g_{x,\varDelta}(\varphi(Z(X_{-i} + X'_i, t)) + \varphi(X_i - X'_i)) - \mathbb{E}_{-i} g_{x,\varDelta}(\varphi(Z(X_{-i} + X'_i, t)) \\
&\le \mathbb{E}_{-i} g'_{x,\varDelta}(\varphi(Z(X_{-i} + X'_i, t)) + \theta\varphi(X_i - X'_i))\varphi(X_i - X'_i) \\
&\le \left(C_A + \frac{2\delta}{\varDelta}\right) \varphi(X_i - X'_i)
\end{aligned}$$

Analogically

$$\mathbb{E} h(Z(X_{-i} + X'_i, t)) - \mathbb{E} h(Z(X' + \theta X_i, t)) \le \left(C_A + \frac{2\delta}{\varDelta}\right) \varphi(X_i - X'_i)$$

Apply this inequalities in previous Taylor expansion denoting

$$\varepsilon^2 = (\varSigma^{-1/2} Z)^T \gamma)^2 \sim \mathcal{N}^2(0,1)$$

$$
\begin{aligned}
&\mathbb{E}_{-i}\{((\Sigma^{-1/2}Z)^T\gamma)^2-1\}\{h(Z(X'+\theta X_i,t))-h(Z(X_{-i}+X_i',t))\}\\
&\le|\tau-1|\left(C_A+\frac{2\delta}{\Delta}\right)\varphi(X_i-X_i')+\mathbb{E}\,\mathbb{1}[\varepsilon^2>\tau]\varepsilon^2
\end{aligned}
$$

Lemma 2.2. Let a random variable ε has a tail bound $\forall \mathtt{x}\ge \mathtt{x}_0$

$$\mathbb{P}(\varepsilon>h(\mathtt{x}))\le e^{-\mathtt{x}}$$

Then for a function $g:\mathbb{R}_+\to\mathbb{R}_+$ with derivative $g':\mathbb{R}_+\to\mathbb{R}_+$

$$\mathbb{E}\,\mathbb{1}[\varepsilon>h(\mathtt{x}_0)]g(\varepsilon)\le g(h(\mathtt{x}_0))e^{-\mathtt{x}_0}+\int_{\mathtt{x}_0}^{\infty}e^{-\mathtt{x}}g'(h(\mathtt{x}))h'(\mathtt{x})d\mathtt{x}$$

In particular

$$\mathbb{E}\,\mathbb{1}[\varepsilon>h(\mathtt{x}_0)]\varepsilon\le h(\mathtt{x}_0)e^{-\mathtt{x}_0}+\int_{\mathtt{x}_0}^{\infty}e^{-\mathtt{x}}h'(\mathtt{x})d\mathtt{x}$$

$$\mathbb{E}\,\mathbb{1}[\varepsilon>h(\mathtt{x}_0)]\varepsilon^r\le h(\mathtt{x}_0)^re^{-\mathtt{x}_0}+r\int_{\mathtt{x}_0}^{\infty}e^{-\mathtt{x}}h(\mathtt{x})^{r-1}h'(\mathtt{x})d\mathtt{x}$$

For $\varepsilon\sim\mathcal{N}(0,1)$ we have

$$\mathbb{P}(\varepsilon>\sqrt{2\mathtt{x}})\le e^{-\mathtt{x}}$$

and by means of the previous lemma we get

$$\mathbb{E}\,\mathbb{1}[\varepsilon^2>\tau]\varepsilon^2=2\mathbb{E}\,\mathbb{1}[\varepsilon>\sqrt{\tau}]\varepsilon^2\le 2(\tau+2)e^{-\tau/2}$$

$$
\begin{aligned}
&\mathbb{E}_{-i}\{((\Sigma^{-1/2}Z)^T\gamma)^2-1\}\{h(Z(X_{-i}+\theta X_i,t))-h(Z(X_{-i}+X_i',t))\}\\
&\le|\tau-1|\left(C_A+\frac{2\delta}{\Delta}\right)\varphi(X_i-X_i')+2(\tau+2)e^{-\tau/2}
\end{aligned}
$$

We need also another upper bound for this expectation when t close to 1.

$$
\begin{aligned}
&\mathbb{E}_{-i}\{((\Sigma^{-1/2}Z)^T\gamma)^2-1\}h(Z(X_{-i}+X_i',t))\\
&=\mathbb{E}_{-i}\{((\Sigma^{-1/2}Z)^T\gamma)^2-1\}\{h(\sqrt{t}(X_{-i}+X_i')+\sqrt{1-t}Z)-h(\sqrt{t}(X_{-i}+X_i'))\}\\
&\le\mathbb{E}_{-i}|((\Sigma^{-1/2}Z)^T\gamma)^2-1|\,|g'_{x,\Delta}|\,\varphi(\sqrt{1-t}Z)\\
&\le\frac{1}{\Delta}\sqrt{2\mathbb{E}\varphi^2(Z)}
\end{aligned}
$$

From the proof of Theorem 2.1 follows

$$|\mathbb{E}h(X)-\mathbb{E}h(Z)|\le-\frac{A}{2}\log(\alpha)+2B\sqrt{\alpha}$$

Set $\Delta=\delta/(2C_A)$

$$B=\frac{2C_A}{\delta}\sqrt{2\mathbb{E}\varphi^2(Z)}\,p$$

Set $\tau = 2\log(4p/(C_A\mu_3))$

$$\begin{aligned}
A &= 5|\tau - 1|C_A\mu_3 + 2(\tau + 2)e^{-\tau/2}p \\
&\leq 11C_A\mu_3 \log\left(\frac{4p}{C_A\mu_3}\right)
\end{aligned}$$

For making step from h expectation difference to probabilities difference involve the next inequality:

$$\begin{aligned}
I\!P(\varphi(X) > x) &\leq I\!E h(X) = I\!E h(Z) + I\!E h(X) - I\!E h(Z) \\
&\leq I\!P(\varphi(Z) > x - \varDelta) + I\!E h(X) - I\!E h(Z) \\
&\leq I\!P(\varphi(Z) > x) + I\!E h(X) - I\!E h(Z) + C_A \varDelta
\end{aligned}$$

Which gives

$$\delta \leq |I\!E h(X) - I\!E h(Z)| + C_A\varDelta$$

$$\begin{aligned}
\delta &\leq -\frac{A}{2}\log(\alpha) + 2B\sqrt{\alpha} + C_A\varDelta \\
&\leq 2A\left(1 + \log(2B\delta) - \log(\delta) - \log(A)\right) \\
&\leq 2A\left(1 + \log(2B\delta) - 2\log(A) + \log\log(2B\delta) - \log\log(A)\right) \\
&\leq 22C_A\mu_3 \log\left(\frac{4p}{C_A\mu_3}\right)\log\left(\frac{\sqrt{2I\!E\varphi^2(Z)p}}{20C_A\mu_3^2}\right)
\end{aligned}$$

□

Remark. In i.i.d case with $\varSigma = I_p$ and $\varphi(x) = O(\|x\|)$

$$|I\!P(\varphi(X) > x) - I\!P(\varphi(Z) > x)| = O\left(C_A\mu_3\log^2(n)\right)$$

Note that Lemma 2.2 improves the classical Multivariate Berry–Esseen Theorem Bentkus (2003b) for the case of sub-additive functions $\phi(x) = O(\|x\|)$. Namely, it answers the open question "Whether one can remove or replace the factor $p^{1/4}$ by a better one (eventually by 1)".

Make an extension of the Gaussian approximation for the case when the second moments of X and Z are slightly different. Let as previously $X = \sum_{i=1}^{n} X_i$ and

$$\|\operatorname{Var} X_i - \operatorname{Var} Z_i\| \sim \frac{1}{n}$$

Then after sequential replacement $X_i \to Z_i$ the approximation bound will not converge to zero, while

$$\|\operatorname{Var} X - \operatorname{Var} Z\| \sim \frac{1}{\sqrt{n}}$$

The next lemma resolves this problem. At first one should make Gaussian approximation with equal variances ($\operatorname{Var} X = \operatorname{Var} Z$) and then compare two Gaussian vectors with different variances.

Lemma 2.3 (Pinsker's inequality)**.** Let X and Y be two zero mean Gaussian vectors with $\Sigma_X = \operatorname{Var}(X)$ and $\Sigma_Y = \operatorname{Var}(Y)$. Then for any event A

$$|I\!P(X \in A) - I\!P(Y \in A)| \le \frac{1}{2} \operatorname{tr}\{(\Sigma_X \Sigma_Y^{-1} - I)^2\}^{1/2}$$

Proof.

$$|I\!P(X \in A) - I\!P(Y \in A)| \le \sqrt{\mathcal{KL}(I\!P_X \| I\!P_Y)/2}$$

The change of variables $X = \Sigma_Y^{-1/2} X$, $Y = \Sigma_Y^{-1/2} Y$ reduces the general case to the situation when $I\!P_Y$ is standard normal while $I\!P_X \in \mathcal{N}(0, B)$ with $B = \Sigma_Y^{-1/2} \Sigma_X \Sigma_Y^{-1/2}$

$$2\mathcal{KL}(I\!P_X \| I\!P_Y) = \operatorname{tr}\{B - I\} - \log\det B$$

$$\begin{aligned}\operatorname{tr}\{B - I\} - \log\det B &= \operatorname{tr}\{\operatorname{diag}\lambda(B) - I\} - \log\det\operatorname{diag}\lambda(B) \\ &= \sum_i (\lambda_i - 1 - \log\lambda_i) \le \sum_i (\lambda_i - 1)^2 = \operatorname{tr}\{(B - I)^2\}\end{aligned}$$

□

2.3 Anti-concentration

Anti-concentration property (ref. Chernozhukov et al. (2013a)) can be interpreted as an asymptotic of a probability measure depended on event size. It converges to zero when the event size goes to zero. Denote by $A_\Delta \setminus A$ a region of size Δ around event A. Anti-concentration is better when the probability of $A_\Delta \setminus A$ is lower. Many works use more classical and identical with anti-concentration term Levy's concentration function (ref. Petrov (1995)). Consider first one dimensional case where the random variable is $\varphi(Z)$ and Z is a Gaussian vector.

Lemma 2.4. Let $Z \in \mathcal{N}(m, \Sigma) \in I\!R^p$, a function $\varphi : I\!R^p \to I\!R_+$ be sub-additive:

$$\varphi(x + y) \le \varphi(x) + \varphi(y)$$

then $\forall x > 0$

$$I\!P(\varphi(Z) \in [x, x + \Delta]) \le \Delta C_A,$$

where

$$C_A = \frac{\sqrt{p}}{x}$$

Proof. Note that

$$I\!P(\varphi(Z) < x) = I\!P\left(\frac{x + \Delta}{x}\varphi(Z) < x + \Delta\right) \le I\!P\left(\varphi\left(\frac{x + \Delta}{x} Z\right) < x\right)$$

Apply Pinsker's inequality from Lemma 2.3.

□

The next lemma deals with anti-concentration in one dimensional case where the random variable is maximum of a Gaussian vector.

Lemma 2.5. Let $Z \in \mathcal{N}(m, \Sigma) \in I\!R^p$,

$$\sigma_1 \leq \sqrt{\Sigma_{ii}} \leq \sigma_2,$$

$$a_p = I\!E \max_i (Z_i - m_i)/\sqrt{\Sigma_{ii}},$$

then $\forall x$

$$I\!P(\max_i Z_i \in [x, x + \Delta]) \leq \Delta C_A(\log \Delta),$$

where

$$C_A(\log \Delta) = \frac{4}{\sigma_1} \left(\frac{\sigma_2}{\sigma_1} a_p + \left(\frac{\sigma_2}{\sigma_1} - 1 \right) \sqrt{2 \log \left(\frac{\sigma_1}{\Delta} \right)} + 2 - \frac{\sigma_1}{\sigma_2} \right)$$

Proof. Find by reference Chernozhukov et al. (2013a). □

There is also an extension for maximum of Gaussian process.

Lemma 2.6. Let $\mathcal{F} \subset \mathcal{L}^2(P)$ be a separable class of measurable functions and entropy of $\mathcal{F}$ be finite. Denote by $G(f)$, $f \in \mathcal{F}$ a Gaussian random process with zero mean and covariance depended on measure P:

$$I\!E[G(f)G(g)] = \int f(x)g(x)dP(x)$$

Suppose that there exist constants σ_1, $\sigma_2 > 0$ such that $\sigma_1^2 \leq I\!E f^2 \leq \sigma_2^2$ for all $f \in \mathcal{F}$. Then $\forall$ x and $\Delta > 0$

$$I\!P \left(\sup_{f \in \mathcal{F}} G(f) \in [x, x + \Delta] \right) \leq C_A \Delta,$$

where

$$C_A = O \left(I\!E \left[\sup_{f \in \mathcal{F}} G(f) \right] + \sqrt{1 \vee \log(\sigma_1/\Delta)} \right)$$

Proof. Find in Lemma A.1 from article V. Chernozhukov (2014). We have used finite entropy assumption in this Lemma because it ensures the existence of process $G(f)$ according to Dudley's criterion for sample continuity of Gaussian processes. □

2.4 Euclidean norm statistic

The distribution approximation of Euclidean norm of independent vectors sum is very important in statistics since it characterise the deviations of Likelihood maximum and MLE (ref. Theorem 3.1). One may use Lemma 2.2 which gives a bound for chi-square root approximation treating it as a special case of multivariate Berry–Esseen Theorem. Let $X = \sum_{i=1}^{n} X_i$, $X \in I\!R^p$ and $Z \in \mathcal{N}(0, I\!E XX^T)$ then

$$
\begin{aligned}
|I\!P(\|X\| > x) - I\!P(\|Z\| > x)| &\le 22 C_A \mu_3 O\left(\log^2 \frac{p}{\mu_3}\right) \\
&\le O\left(\frac{p}{\sqrt{n}} \log^2 n\right)
\end{aligned}
$$

since $C_A = O(1/\sqrt{p})$ for Euclidean norm (Götze et al. (2019)) and $\mu_3 = O(p^{3/2}/\sqrt{n})$. Lemma 2.2 bounds the difference of probabilities by the third moment μ_3, but it could be zeroed multiplying X by independent random flip variable which doesn't change the norm and has zero third moment. The open question is whether one can improve the previous asymptotic using the 4-th moment and the generator (AΓ). Robert E. Gaunt (2015) has made such attempt for "smooth" distance between two measures. They have presented the approximation bound in the following form:

Lemma 2.7. Let $I\!E X_i X_i^T = \frac{1}{n} I_p$ and $\{X_{ij}\}$ be i.i.d and $\|Z\|^2 \sim \chi_p$. For any $h \in C^3(I\!R)$ it holds

$$
I\!E h(\|X\|^2) - I\!E h(\|Z\|^2) \le \frac{4p I\!E X_{11}^8}{n(p+2)} (\alpha_0 \|h\| + \alpha_1 \|h'\| + \alpha_2 \|h''\| + \alpha_3 \|h'''\|),
$$

where

$$
\begin{aligned}
\alpha_0 &= 2 + 69 |I\!E X_{11}^3| \\
\alpha_1 &= 38 + 1781 |I\!E X_{11}^3| \\
\alpha_2 &= 203 + 1781 |I\!E X_{11}^3| \\
\alpha_3 &= 321 + 1320 |I\!E X_{11}^3|
\end{aligned}
$$

Remark. The bound for the "smooth" difference has significantly better asymptotic $O(1/n)$ than the analogical asymptotic for measures difference $O\left(\frac{p}{\sqrt{n}} \log^2 n\right)$ discussed above.

The main drawback of Gaussian comparison by Lemma 2.3 is the asymptotic $O(\sqrt{p})$, when $\Omega(X), \Omega(Y) \in I\!R^p$. More complex method based on characteristic functions improves the comparison bound for l_2 norm.

Lemma 2.8 (Götze et al. (2019))**.** Let X and Y be two zero mean Gaussian vectors in Hilbert space $\mathcal{H}$ with $\Sigma_X = \operatorname{Var}(X)$ and $\Sigma_Y = \operatorname{Var}(Y)$. Denote by λ_X, λ_Y eigenvalues of Σ_X, Σ_Y. Then for any non random $\Delta \in \mathcal{H}$ and $\forall x$

$$
\begin{aligned}
&|I\!P(\|X + \Delta\| \le x) - I\!P(\|Y\| \le x)| \\
&\qquad \le O\Big(\frac{1}{(\Lambda_{1X}\Lambda_{2X})^{1/2}} + \frac{1}{(\Lambda_{1Y}\Lambda_{2Y})^{1/2}}\Big)\Big(\|\lambda_X - \lambda_Y\|_1 + \|\Delta\|^2\Big),
\end{aligned}
$$

where with $\lambda_{1X} \geq \lambda_{2X} \geq \dots$

$$\Lambda_{kX}^2 = \sum_{j=k}^{\infty} \lambda_{jX}^2, \quad k = 1, 2$$

2.5 Maximum statistic

Denote by $h_\beta(x)$ a smooth maximum function which converges to $\max_i x_i$ when $\beta \to \infty$.

$$h_\beta(x) = \beta^{-1} \log u(x), \quad u(x) = \sum_i e^{\beta x_i}, \quad x = (x_1, x_2, \dots)$$

Explore its derivatives and some other properties. It appears that 1-norm of the smooth maximum derivatives doesn't depend on the dimension which enables to make $1-\infty$–norm decomposition in the third moments of Gaussian approximation.

Lemma 2.9. The derivatives of $h_\beta(x)$ of order $m = \{1, 2, 3\}$ have the following upper bounds $\forall x$

$$\|\nabla^{(m)} h_\beta(x)\|_1 \leq \beta^{m-1}$$

Proof.

$$\begin{aligned}
\nabla h_\beta(x) &= \beta^{-1} \frac{\nabla u}{u}, \\
\nabla^2 h_\beta(x) &= \beta^{-1} \left(\frac{\nabla \otimes \nabla u}{u} - \frac{\nabla u \otimes \nabla u}{u^2} \right), \\
\nabla^3 h_\beta(x) &= \beta^{-1} \left(\frac{\nabla \otimes \nabla \otimes \nabla u}{u} - \frac{\nabla \otimes \nabla u \otimes \nabla u}{u^2} - \frac{\nabla \otimes (\nabla u \otimes \nabla u)}{u^2} + 2 \frac{\nabla u \otimes \nabla u \otimes \nabla u}{u^3} \right)
\end{aligned}$$

Define $p_i = \frac{\nabla u}{u}(i)$ that satisfies to condition $\sum_i p_i = 1$. The first tensor norm equals to the convolution maximum with vectors α, ϕ, γ under restriction $\|\alpha\|_\infty = 1$, $\|\phi\|_\infty = 1$, $\|\gamma\|_\infty = 1$.

$$\begin{aligned}
\alpha^T \nabla^2 h_\beta(x) \gamma &= \beta \left(\sum p_i \alpha_i \gamma_i - \sum p_i \alpha_i \sum p_j \gamma_j \right) \\
&= \beta \left(\mathbb{E} \alpha \gamma - \mathbb{E} \alpha \mathbb{E} \beta \gamma \right) = \beta \mathbb{E} \overset{o}{\alpha} \overset{o}{\gamma} \leq \beta \|\alpha\|_\infty \|\gamma\|_\infty, \\
\sum_{ijk} \nabla_{ijk}^3 h_\beta(x) \alpha_i \phi_j \gamma_k &= \beta^2 \left(\mathbb{E} \alpha \phi \gamma - \mathbb{E} \alpha \mathbb{E} \phi \gamma - \mathbb{E} \alpha \phi \mathbb{E} \gamma - \mathbb{E} \alpha \gamma \mathbb{E} \phi + 2 \mathbb{E} \alpha \mathbb{E} \phi \mathbb{E} \gamma \right) \\
&= \beta^2 \left(\mathbb{E} \overset{o}{\alpha} \overset{o}{\phi} \overset{o}{\gamma} \right) \leq \beta^2 \|\alpha\|_\infty \|\phi\|_\infty \|\gamma\|_\infty
\end{aligned}$$

Taking maximum provides the required restriction for the 1-st tensor norms.

□

The next property of $h_\beta(x)$ with $x \in \mathbb{R}^p$ characterises the precision of the smooth maximum approximation

$$\max_i(x_i) \le h_\beta(x) \le \max_i(x_i) + \beta^{-1}\log(p)$$

Application of indicator to both parts yields inequality required in probabilities comparison.

Lemma 2.10. For a smooth indicator function $g_\triangle$ ($g_\triangle$ grows from 0 to 1 inside interval $[z, z+\triangle]$) it holds with $\triangle = \beta^{-1}\log(p)$ that

$$\mathbb{I}\left[\max_{0\le i\le p} x_i > z + \triangle\right] \le g_\triangle h(x) \le \mathbb{I}\left[\max_{0\le i\le p} x_i > z - \triangle\right]$$

Now we can explore the difference between the distribution of maximum of independent random vectors sum and distribution of the maximum of some Gaussian vector. According to Lemma 2.1 one has to bound the third moment of function $f_h(x)$ (the solution of the Stein's equation). Consider the case when function h is a composition of a smooth indicator $g_\triangle$ and smooth max h_β. The third gradient of the composition is

$$\nabla^3(g_\triangle \circ h_\beta) = g'''_\triangle \nabla h_\beta \otimes \nabla h_\beta \otimes \nabla h_\beta + 2g''_\triangle \nabla^2 h_\beta \otimes \nabla h_\beta + g''_\triangle \nabla h_\beta \otimes \nabla^2 h_\beta + g'_\triangle \nabla^3 h_\beta$$

From Lemma 2.9 follows that

$$\left\|\nabla^3(g_\triangle \circ h_\beta)\right\|_1 \le |g'''_\triangle| + 3|g''_\triangle|\beta + |g'_\triangle|\beta^2$$

Assume that $g_\triangle$ grows from 0 to 1 in interval $[z, z+\triangle]$ such that $g'_\triangle = 0$ outsize $[z, z+\triangle]$. Furthermore in this case

$$\nabla^3(g_\triangle \circ h_\beta) = \nabla^3(g_\triangle \circ h_\beta)\, \mathbb{I}[z \le h_\beta(X) \le z + \triangle]$$

It gives us the Gaussian approximation for the smooth function (ref. proof of Theorem 2.1). Consider as previously sum of independent vectors $X = \sum_{i=1}^n X_i$, $\varOmega(X) \in \mathbb{R}^p$ and $Z \in \mathcal{N}(0, \mathbb{E}XX^T)$

$$|\mathbb{E}g_\triangle(h_\beta(X)) - \mathbb{E}g_\triangle(h_\beta(Z))| \le \|\nabla^3(g_\triangle \circ h_\beta)\|_1 \mathbb{E}\, \mathbb{I}[z \le h_\beta(X) \le z + \triangle]\mu_3,$$

where

$$\mu_3 = \sum_i 2\mathbb{E}\|X_i\|^3_\infty$$

Move to the approximation of distribution function. The aim is to find upper bound for

$$E_G = \left|\mathbb{P}\left(\max\left(\sum_i X_i\right) \le z\right) - \mathbb{P}\left(\max Z \le z\right)\right|$$

Since $\mathbb{E}\,\mathbb{1}[\max(X) \le z] = \mathbb{P}(\max(X) \le z)$ one gets

$$\begin{aligned}
&\mathbb{P}(\max(X) \le z - \triangle)\\
&\le \mathbb{E} g(h_\beta(X))\\
&\le \mathbb{E} g(h_\beta(Z)) + \|\nabla^3(g_\triangle \circ h_\beta)\|_1 \mu_3\\
&\le \mathbb{P}(\max(Z) \le z + \triangle) + \|\nabla^3(g_\triangle \circ h_\beta)\|_1 \mathbb{E}\,\mathbb{1}[z \le h(X) \le z + \triangle]\mu_3
\end{aligned}$$

Subsequently

$$|\mathbb{P}(\max(X) \le z) - \mathbb{P}(\max(Z) \le z \pm 2\triangle)| \le \|\nabla^3(g_\triangle \circ h_\beta)\|_1 \mathbb{E}\,\mathbb{1}[z \le h(X) \le z + \triangle]\mu_3$$

Use the anti-concentration property for random variable $\max(Z)$ (Lemma 2.5):

$$\mathbb{P}(\max(Z) \in [z, z + \triangle]) \le C_A \triangle$$

Then

$$E_G \le 2C_A\triangle + \|\nabla^3(g_\triangle \circ h_\beta)\|_1 \mathbb{E}\,\mathbb{1}[z \le h(X) \le z + \triangle]\mu_3$$

and

$$\begin{aligned}
&\mathbb{E}_{-i}\left(\mathbb{1}[z \le h(X) \le z + \triangle]\right)\\
&\le \mathbb{E}_{-i}\left(\mathbb{1}[z \le \max(X) \le z + 2\triangle]\right)\\
&\le \mathbb{E}\left(\mathbb{1}[z \le \max(Z) \le z + 2\triangle]\right) + E_G\\
&\le 2C_A\triangle + E_G
\end{aligned}$$

We use conditional $\mathbb{E}_{-i}$ because μ_3 elements depends on X_i. Denote the restriction for the third derivative by constant C_μ:

$$\left\|\nabla^3(g_\triangle \circ h)\right\|_1 \le \frac{1}{\triangle^3} C_\mu$$

Group all together

$$E_G \le \frac{1}{\triangle^3} C_\mu (2\triangle C_A + E_G)\mu_3 + 2\triangle C_A$$

$$|\mathbb{P}(\max(X) \le z) - \mathbb{P}(\max(Z) \le z)| \le 5 C_\mu^{1/3} C_A \mu_3^{1/3}$$

We have proved the following statement.

Theorem 2.3. Let $X = \sum_{i=1}^n X_i \in \mathbb{R}^p$ with independent random vectors and $Z \in \mathcal{N}(\mathbb{E}X, \mathbb{E}XX^T)$ then $\forall x$

$$|\mathbb{P}(\max(X) \le x) - \mathbb{P}(\max(Z) \le x)| \le 5 C_\mu^{1/3} C_A \mu_3^{1/3},$$

where

$$C_\mu = 6\left(1 + 3\log p + \log^2 p\right)$$

$$\mu_3 = \sum_i 2\mathbb{E}\|X_i\|_\infty^3$$

and $C_A = C_A(\log p)$ is anti-concentration constant (ref. Lemma 2.5).

Remark. The approximation for maximum function in this Theorem has asymptotic $O(\log^{8/6} p/n^{1/6})$ and has worse dependence on n than in general case (Lemma 2.2). But it is compensated by logarithmic dependence on the dimension p.

Lemma 2.11. Let X and Y be two zero mean Gaussian vectors with $\Sigma_X = \operatorname{Var}(X)$ and $\Sigma_Y = \operatorname{Var}(Y)$. Let also $f(X)$ be a smooth function. Then

$$|\mathbb{E}f(X) - \mathbb{E}f(Y)| \le \frac{1}{2}\|\Sigma_X - \Sigma_Y\|_\infty \max_{t\in[0,1]} \|\mathbb{E}\nabla^2 f(Z(t))\|_1,$$

where

$$Z(t) = \sqrt{t}\,X + \sqrt{1-t}\,Y$$

Proof. Without loss of generality assume that X and Y are given on the same probability space and independent. For each $t \in [0,1]$, define

$$\Psi(t) = \mathbb{E}f(Z(t)) = \mathbb{E}f(\sqrt{t}\,X + \sqrt{1-t}\,Y)$$

$$|\Psi(1) - \Psi(0)| = \left|\int_0^1 \Psi'(t)dt\right|$$

$$\Psi'(t) = \mathbb{E}[\nabla f(Z(t))^\top Z'(t)] = \frac{1}{2}\mathbb{E}\left[\left\{t^{-1/2}X - (1-t)^{-1/2}Y\right\}^\top \nabla f(Z(t))\right]$$

To compute this expectation, we apply the Stein identity. Let W be a zero mean Gaussian vector. Then for any C^1 vector function s it holds

$$\mathbb{E}[W\,s(W)^\top] = \operatorname{Var}(W)\,\mathbb{E}[\nabla s(W)]$$

$$\begin{aligned}\mathbb{E}\left[\nabla f(Z(t))X^\top\right] &= t^{1/2}\Sigma_X\mathbb{E}\left[\nabla^2 f(Z(t))\right]\\ \mathbb{E}\left[\nabla f(Z(t))Y^\top\right] &= (1-t)^{1/2}\Sigma_Y\mathbb{E}\left[\nabla^2 f(Z(t))\right]\end{aligned}$$

$$\begin{aligned}|\Psi'(t)| &\le \frac{1}{2}\left|\operatorname{tr}\left\{(\Sigma_X - \Sigma_Y)\mathbb{E}\left[\nabla^2 f(Z(t))\right]\right\}\right|\\ &\le \frac{1}{2}\|\Sigma_X - \Sigma_Y\|_\infty \left\|\mathbb{E}\left[\nabla^2 f(Z(t))\right]\right\|_1 \le \frac{1}{2}\|\Sigma_X - \Sigma_Y\|_\infty \mathbb{E}\|\nabla^2 f\|_1\end{aligned}$$

□

2.6 Multiple statistics

Consider a linear form (AX) with sparse-row matrices and random vector $X = (X_1, \dots, X_N)$ with independent elements (sub-vectors). In order to make approximation by Gaussian vector AZ one has to group $(X_1, \dots, X_N)$, such that elements in one group would have no common non-zero coefficients in each row of matrix A. It would provide the following representation

$$AX = Y_1 + \ldots + Y_n, \quad Y_i = AF_iX$$

Vectors $\{Y_i\}$ should be independent and F_i is a filter for the i-th group that sets matrix columns to 0 which correspond to the other groups. In case each row of the matrix A has less or equal than n non-zero elements, the minimal groups count equals to n. The next statement confirms it.

Lemma 2.12. Let each element in a set of subsets $\{\mathcal{M}_s\}$ has size n. Then exist subsets $\{\mathcal{Z}_1, \dots, \mathcal{Z}_h\}$ with properties

$$\bigcup_s \mathcal{Z}_s = \bigcup_s \mathcal{M}_s, \quad \mathcal{Z}_k \bigcap \mathcal{M}_s = 1, \quad \mathcal{Z}_k \bigcap \mathcal{Z}_s = 0$$

Proof. Build subsets $\{\mathcal{Z}_1, \dots, \mathcal{Z}_n\}$ constructively. Take one element from $\bigcup_s \mathcal{M}_s$. Exclude this element from all $\{\mathcal{M}_s\}$ and add it to $\mathcal{Z}_1$. Mark subsets which contain this element as $\{\mathcal{M}'_s\}$. Take another element from $\bigcup_s \mathcal{M}_s \setminus \bigcup_s \mathcal{M}'_s$ and add it to $\mathcal{Z}_1$. Repeat this procedure until $|\bigcup_s \mathcal{M}_s \setminus \bigcup_s \mathcal{M}'_s| = 0$. Then do the same steps for $\bigcup_s \mathcal{M}_s \setminus \mathcal{Z}_1$ and obtain $\mathcal{Z}_2$ with the required properties by construction. □

Basing on the previous Lemma apply Gaussian approximation for maximum from Theorem 2.3 to Y_i instead of X_i.

Lemma 2.13. Assume that matrix A has at most n non-zero elements in each row and non-zero elements correspond to independent elements in vector $X = (X_1, \dots, X_N)$. Then Gaussian approximation with vector $Z \in \mathcal{N}(\mathbb{E}X, \mathbb{E}XX^T)$ has the following upper bound

$$|\mathbb{P}(\max(AX) \le x) - \mathbb{P}(\max(AZ) \le x)| \le 5C_\mu^{1/3} C_A \mu_3^{1/3},$$

where

$$\mu_3 = 2\mathbb{E}\,\|X\|_\infty^3 \sum_{i=1}^n |||AF_i|||_1^3 \le 2n\|A\|_\infty^3 \mathbb{E}\,\|X\|_\infty^3$$

and C_μ, C_A are from Theorem 2.3 with $p =$ rows count of A.

Consider a composite maximum function with its smooth approximation $h_\beta(\varphi(X))$

$$\max_{1 \le t \le T} \varphi_t(x) \approx h_\beta(\varphi(x)), \quad \varphi(x) = (\varphi_1(x), \dots, \varphi_T(x))$$

Combination of the maximum approximation with property $h(x) \le \max_t(x_t) + \beta^{-1}\log(p)$ leads to the statement

$$\max_t(\varphi_t(x)) \le h_\beta(\varphi(x)) \le \max_t(\varphi_t(x)) + \beta^{-1}\log(T)$$

This allows to extend Lemma 2.10.

Lemma 2.14. For a smooth indicator function $g_\triangle$ ($g_\triangle$ grows from 0 to 1 inside interval $[z, z+\triangle]$) it holds with $\triangle = \beta^{-1}\log(T)$ for all x and z that

$$1\!\!\mathrm{I}\left[\max_{1\le t\le T}\varphi_t(x) > z+\triangle\right] \le g_\triangle(h_\beta(\varphi(x)) \le 1\!\!\mathrm{I}\left[\max_{1\le t\le T}\varphi_t(x) > z-\triangle\right]$$

Theorem 2.4. Consider a composite maximum function $\max_{1\le t\le T}\varphi_t(Ax)$, $t\in\mathbb{N}$. Assume restrictions for derivatives of the functions φ_t:

$$\forall t:\ \|\nabla^{(m)}\varphi_t(x)\|_1 \le \frac{C_\varphi^{m-1}}{\triangle^{m-1}}$$

Assume that matrix A has at most n non-zero elements in each row and non-zero elements correspond to independent elements in vector $X=(X_1,\dots,X_N)$. Then Gaussian approximation with vector $Z\in\mathcal{N}(I\!\!E X, I\!\!E XX^T)$ has the following upper bound $\forall x$

$$|I\!\!P\left(\varphi_1(AX),\dots,\varphi_T(AX)\le x\right) - I\!\!P\left(\varphi_1(AZ),\dots,\varphi_T(AZ)\le x\right)| \le 5C_\mu^{1/3}C_A\mu_3^{1/3}$$

where

$$\mu_3 \le 2n\|A\|_\infty^3 I\!\!E\,\|X\|_\infty^3$$
$$C_\mu = 6\left(1+3\log T+3C_\varphi+\log^2 T+3C_\varphi\log T+C_\varphi^2\right)$$

and C_A is obtained from condition $\forall x, \Delta$

$$I\!\!P\left(\max_{1\le t\le T}\varphi_t(AZ)\in[x,x+\Delta]\right)\le \Delta C_A$$

Proof. From the restrictions for derivatives of the functions φ_t follows

$$\|\nabla h(\varphi(X))\|_1 \le \|\nabla h\|_1\|\nabla\varphi_t\|_1 \le 1$$

$$\|\nabla^2 h(\varphi(X))\|_1 \le \|\nabla^2 h\|_1\|\nabla\varphi_t\|_1^2 + \|\nabla h\|_1\|\nabla^2\varphi_t\|_1 \le \frac{\log T}{\triangle}+\frac{C_\varphi}{\triangle}$$

$$\begin{aligned}\|\nabla^3 h(\varphi(X))\|_1 &\le \|\nabla^3 h\|_1\|\nabla\varphi_t\|_1^3 + 3\|\nabla^2 h\|_1\|\nabla^2\varphi_t\|_1\|\nabla\varphi_t\|_1 + \|\nabla h\|_1\|\nabla^3\varphi_t\|_1\\ &\le \frac{1}{\triangle^2}(\log^2 T+3C_\varphi\log T+C_\varphi^2)\end{aligned}$$

So one can override C_μ used in Lemma 2.13 for this case:

$$C_\mu = 6\left(1+3\log T+3C_\varphi+\log^2 T+3C_\varphi\log T+C_\varphi^2\right)$$

Apply Lemma 2.13.

□

Find an upper bound for distribution difference with $X \in \mathcal{N}(0, \Sigma_X)$ and $X \in \mathcal{N}(0, \Sigma_Y)$, i.e. $\forall x$, $t \in \mathbb{N}$:

$$E = \left| I\!P\left(\max_{1 \le t \le T} \varphi_t(X) \le x\right) - I\!P\left(\max_{1 \le t \le T} \varphi_t(Y) \le x\right) \right|$$

Assume following restriction for the second derivative of the function f from Lemma 2.11 and define C_Σ:

$$\frac{1}{2}\|\nabla^2 g_\triangle \circ h_\beta \circ \varphi\|_1 \le \frac{1}{\triangle^2} C_\Sigma$$

Taking into account $I\!P = I\!E\, 1\!\!1$ we get

$$\begin{aligned}
&|I\!P(H(X) \le z) - I\!P(H(Y) \le z \pm 2\triangle)| \\
&\le |I\!E f(X) - I\!E f(Y)| \\
&\le \|\nabla^2 f\|_1 I\!E\, 1\!\!1 \Big[h(X) \in [z, z + \triangle]\Big] \|\Sigma_X - \Sigma_Y\|_\infty \\
&\le \frac{1}{\triangle^2} C_\Sigma I\!E\, 1\!\!1 \Big[H(X) \in [z, z + 2\triangle]\Big] \|\Sigma_X - \Sigma_Y\|_\infty
\end{aligned}$$

The anti-concentration property allows $2\triangle$-shift elimination and provides an upper bound for

$$I\!E\, 1\!\!1[H(X) \in z \pm 2\triangle] \le 2\triangle C_A$$

Combine with previous equation

$$E \le \frac{2}{\triangle^2} C_\Sigma \triangle C_A \|\Sigma_X - \Sigma_Y\|_\infty + 2\triangle C_A$$

Optimize over $\triangle$ value

$$E \le 4 C_\Sigma^{1/2} C_A \|\Sigma_X - \Sigma_Y\|_\infty^{1/2}$$

Theorem 2.5. Consider X and Y two zero mean Gaussian vectors with $\Sigma_X = \mathrm{Var}(X)$ and $\Sigma_Y = \mathrm{Var}(Y)$. Under conditions from Theorem 2.4 $\forall x$, $t \in \mathbb{N}$:

$$\left| I\!P\left(\max_{1 \le t \le T} \varphi_t(X) \le x\right) - I\!P\left(\max_{1 \le t \le T} \varphi_t(Y) \le x\right) \right| \le 4 C_\Sigma^{1/2} C_A \|\Sigma_X - \Sigma_Y\|_\infty^{1/2},$$

where

$$C_\Sigma = 2\,(1 + \log T + C_q)$$

and C_A is obtained from condition $\forall x, \Delta$

$$I\!P\left(\max_{1 \le t \le T} \varphi_t(AZ) \in [x, x + \Delta]\right) \le \Delta C_A$$

2.7 Wasserstein distance statistic

Consider two point clouds of size n and m. They may have the same distribution (null hypothesis) or different distributions. Assume that samples in each cloud are independent (when we use block-bootstrap we may assume that blocks are independent). It appears that the Wasserstein distance between two point clouds under null hypothesis may be approximated by maximum of some Gaussian vector. In paper Max Sommerfeld (2017) proposed the following theorem for the case of discrete distributions. Define convex sets:

$$\Phi_p = \left\{ (\boldsymbol{u}, \boldsymbol{v}) \in \mathbb{R}^N \times \mathbb{R}^N : u_x + v_{x'} \leq d^p\left(x, x'\right), \ (x, x') \in \mathcal{X} \right\}$$

$$\Phi_p^* = \left\{ \boldsymbol{u} \in \mathbb{R}^N : u_x - u_{x'} \leq d^p\left(x, x'\right), \quad x, x' \in \mathcal{X} \right\}$$

$$\Phi_p^*(\boldsymbol{r}, \boldsymbol{s}) = \left\{ (\boldsymbol{u}, \boldsymbol{v}) \in \Phi_p : \langle \boldsymbol{u}, \boldsymbol{r} \rangle + \langle \boldsymbol{v}, \boldsymbol{s} \rangle = W_p^p(\boldsymbol{r}, \boldsymbol{s}) \right\}$$

Theorem 2.6. Let measures $\boldsymbol{r}$, $\boldsymbol{s}$ be defined on a discrete set $\mathcal{X} = \{x_1, \ldots, x_N\}$ and i.i.d. samples $X_1, \ldots, X_n \sim \boldsymbol{r}$ and $Y_1, \ldots, Y_m \sim \boldsymbol{s}$.
Multinominal covariance matrix of the measure $\boldsymbol{r}$ is

$$\Sigma(\boldsymbol{r}) = \begin{bmatrix} r_{x_1}\left(1 - r_{x_1}\right) & -r_{x_1} r_{x_2} & \cdots & -r_{x_1} r_{x_N} \\ -r_{x_2} r_{x_1} & r_{x_2}\left(1 - r_{x_2}\right) & \ldots & -r_{x_2} r_{x_N} \\ \vdots & & \ddots & \vdots \\ -r_{x_N} r_{x_1} & -r_{x_N} r_{x_2} & \cdots & r_{x_N}\left(1 - r_{x_N}\right) \end{bmatrix}$$

such that with Gaussian random vectors $Z_r \sim \mathcal{N}(0, \Sigma(\boldsymbol{r}))$ and $Z_s \sim \mathcal{N}(0, \Sigma(\boldsymbol{s}))$ it holds for empirical measures $\widehat{\boldsymbol{r}}_n$ and $\widehat{\boldsymbol{s}}_m$:

1)One sample - Null hypothesis

$$n^{\frac{1}{2p}} W_p\left(\widehat{\boldsymbol{r}}_n, \boldsymbol{r}\right) \xrightarrow{d} \left\{ \max_{\boldsymbol{u} \in \Phi_p} \boldsymbol{u}^T Z_r \right\}^{\frac{1}{p}}$$

2) One sample - Alternative

$$n^{\frac{1}{2}}\left(W_p\left(\widehat{\boldsymbol{r}}_n, \boldsymbol{s}\right) - W_p(\boldsymbol{r}, \boldsymbol{s})\right) \xrightarrow{d} \frac{1}{\sqrt{2}p} W_p^{1-p}(\boldsymbol{r}, \boldsymbol{s}) \left\{ \max_{(\boldsymbol{u}, \boldsymbol{v}) \in \Phi_p^*(\boldsymbol{r}, \boldsymbol{s})} \boldsymbol{u}^T Z_r + \boldsymbol{v}^T Z_s \right\}$$

3) Two samples - Null hypothesis. If **r**=**s** and n and m are approaching infinity such that $n \wedge m \to \infty$ and $m/(n+m) \to \lambda \in (0, 1)$, then:

$$\left(\frac{nm}{n+m}\right)^{\frac{1}{2p}} W_p\left(\widehat{\boldsymbol{r}}_n, \widehat{\boldsymbol{s}}_m\right) \xrightarrow{d} \left\{ \max_{\boldsymbol{u} \in \Phi_p} \boldsymbol{u}^T Z_r \right\}^{\frac{1}{p}}$$

4) Two samples - Alternative With n and m approaching infinity such that $n \wedge m \to \infty$

and $m/(n+m) \to \lambda \in [0,1]$, then:

$$\left(\frac{nm}{n+m}\right)^{\frac{1}{2}} \left(W_p\left(\widehat{\boldsymbol{r}}_n, \widehat{\boldsymbol{s}}_m\right) - W_p(\boldsymbol{r}, \boldsymbol{s})\right) \xrightarrow{d}$$

$$\frac{1}{p} W_p^{1-p}(\boldsymbol{r}, \boldsymbol{s}) \left\{ \max_{(\boldsymbol{u},\boldsymbol{v}) \in \Phi_p^*(\boldsymbol{r},\boldsymbol{s})} \sqrt{\lambda} \boldsymbol{u}^T Z_r + \sqrt{1-\lambda} \boldsymbol{v}^T Z_s \right\}$$

Below we will extend this theorem for continuous case and find out the asymptotic bound for the convergence rate. Assume below that $n = m$. Let ϕ_1, ϕ_2 be some density functions. Involve additional notations:

$$I\!P(\phi_1, \phi_2, x) = I\!P\left(\sqrt{n} \max_{(\boldsymbol{u},\boldsymbol{v}) \in \Phi_p} \langle \boldsymbol{u}, \phi_1 \rangle + \langle \boldsymbol{v}, \phi_2 \rangle > x\right)$$

$$I\!P_\varepsilon(\phi_1, \phi_2, x) = I\!P\left(\sqrt{n} \max_{(\boldsymbol{u},\boldsymbol{v}) \in \epsilon\text{-net}(\Phi_p)} \langle \boldsymbol{u}, \phi_1 \rangle + \langle \boldsymbol{v}, \phi_2 \rangle > x\right)$$

Theorem 2.7. Consider i.i.d. samples $X = \{X_1, \ldots, X_n\}$ and $Y = \{Y_1, \ldots, Y_n\}$ with a bounded support space Ω of dimension d. Exist Gaussian vectors $Z_1, Z_2 \in \mathcal{N}(0, \Sigma_\psi)$ and generalized Fourier basis $\{\psi_i\}_{i=1}^\infty$, such that

$$\Sigma_\psi = I\!E \psi \psi^T(X_1)$$

and the Wasserstein distance between the samples can be approximated by the maximum of Gaussian process with the following upper bound

$$\left| I\!P\left(\sqrt{n} W_p^p(X,Y) > x\right) - I\!P\left(Z_1^T \psi, Z_2^T \psi, x\right) \right| \le C_A O\left(\frac{\log n}{n}\right)^{\frac{1}{6+8d/p}}$$

where C_A can be written as

$$O\left(I\!E \max_{(\boldsymbol{u},\boldsymbol{v}) \in \Phi_p} \langle \boldsymbol{u}, Z_1^T \psi \rangle + \langle \boldsymbol{v}, Z_2^T \psi \rangle + \sqrt{\log n}\right)$$

Remark. From the practical sense, resampling of Wasserstein distance should entail data normalization in order to restrict Ω and should keep the power p close to the data dimension d.

Proof. The dual formulation of Wasserstein distance of random vectors X, Y and corresponded densities ϕ_X, ϕ_Y is

$$W_p^p(X,Y) = \max_{(\boldsymbol{u},\boldsymbol{v}) \in \Phi_p} \langle \boldsymbol{u}, \phi_X \rangle + \langle \boldsymbol{v}, \phi_Y \rangle$$

where

$$\phi_X(x) = \frac{1}{n} \sum_{i=1}^n \delta(x - X_i), \;\; \phi_Y(x) = \frac{1}{n} \sum_{i=1}^n \delta(x - Y_i)$$

Show how the covering number of $\mathit{\Phi}_p$ depends on the support space of empirical measures $\mathit{\Omega}$. Construct an ε-net on empirical measures. Its cardinality is $n^{N(\mathit{\Omega},\varepsilon)}$ since each ε-cell of $\mathit{\Omega}$ may contain from 0 to n points. For each densities pair $(\phi_1^\varepsilon, \phi_2^\varepsilon)$ from ε-net one may set in correspondence pair $(\boldsymbol{u}_\varepsilon, \boldsymbol{v}_\varepsilon) \in \mathit{\Phi}_p$ such that $(\boldsymbol{u}_\varepsilon, \boldsymbol{v}_\varepsilon)$ is constant inside each cell of $\mathit{\Omega}$ and

$$W_p^p(\phi_1^\varepsilon, \phi_2^\varepsilon) = \langle \boldsymbol{u}_\varepsilon, \phi_1^\varepsilon \rangle + \langle \boldsymbol{v}_\varepsilon, \phi_2^\varepsilon \rangle$$

The precision of ε-net approximation is bounded by mass transfer inside each cell, i.e

$$W_p^p(\phi_1, \phi_2) - W_p^p(\phi_1^\varepsilon, \phi_2^\varepsilon) \leq 2\varepsilon^p$$

and subsequently for each arbitrary pair of empirical measures (ϕ_1, ϕ_2) on $\mathit{\Omega}$ there is an element $(\boldsymbol{u}_\varepsilon, \boldsymbol{v}_\varepsilon) \in \mathit{\Phi}_p$ with property

$$\langle \boldsymbol{u}_\varepsilon, \phi_1 \rangle + \langle \boldsymbol{v}_\varepsilon, \phi_2 \rangle = \langle \boldsymbol{u}_\varepsilon, \phi_1^\varepsilon \rangle + \langle \boldsymbol{v}_\varepsilon, \phi_2^\varepsilon \rangle$$

such that

$$W_p^p(\phi_1, \phi_2) - \max_{(\boldsymbol{u}_\varepsilon, \boldsymbol{v}_\varepsilon) \in \epsilon\text{-net}(\mathit{\Phi}_p)} \langle \boldsymbol{u}_\varepsilon, \phi_1 \rangle + \langle \boldsymbol{v}_\varepsilon, \phi_2 \rangle = W_p^p(\phi_1, \phi_2) - W_p^p(\phi_1^\varepsilon, \phi_2^\varepsilon) \leq 2\varepsilon^p$$

Decompose densities ϕ_X, ϕ_Y in $\{\psi_i(x)\}$ basis

$$\langle \boldsymbol{u}, \phi_X \rangle + \langle \boldsymbol{v}, \phi_Y \rangle =$$
$$\left\langle \boldsymbol{u}, \left(\frac{1}{n} \sum_i \psi(X_i) \right)^T \psi \right\rangle + \left\langle \boldsymbol{v}, \left(\frac{1}{n} \sum_i \psi(Y_i) \right)^T \psi \right\rangle$$

In order to replace $\{\psi(X_i)\}$ and $\{\psi(Y_i)\}$ by Gaussian vectors and use anti-concentration one has to make an ε-net approximation of $(\boldsymbol{u}, \boldsymbol{v})$ functions. Update $\varepsilon = \varepsilon^{1/p}$. We have shown above that the cowering number of $\mathit{\Phi}_p$ may by restricted by $O(n^{1/\varepsilon^{d/p}})$. So one may set

$$\log p_\varepsilon = \frac{1}{\varepsilon^{d/p}} \log(n) + O(1)$$

determining the dimension of maximum function. On ε-net Theorem 2.3 gives upper bound

$$\Big| I\!P_\varepsilon(\varphi_X, \varphi_Y) - I\!P_\varepsilon(Z_1^T \psi, Z_2^T \psi) \Big| \leq O\left(\frac{\log^8 p_\epsilon}{n} \right)^{1/6}$$

To make a step from $I\!P_\varepsilon$ to $I\!P$ involve the anti-concentration from Lemma 2.6

$$|I\!P_\varepsilon(Z_1^T \psi, Z_2^T \psi) - I\!P(Z_1^T \psi, Z_2^T \psi)| \leq O(C_A (\log p_\varepsilon) \varepsilon)$$

$$I\!P\Big(\sqrt{n} W_p^p(X, Y) > x \Big) \leq I\!P_\varepsilon(\phi_X, \phi_Y, x - 2\varepsilon)$$

$$
\begin{aligned}
&\leq \mathbb{P}_{\varepsilon}(Z_1^T\psi, Z_2^T\psi, x - 2\varepsilon) + O\left(\frac{\log^8 p_\epsilon}{n}\right)^{1/6} \\
&\leq \mathbb{P}(Z_1^T\psi, Z_2^T\psi, x) + O\left(\frac{\log^8 p_\epsilon}{n}\right)^{1/6} + O(C_A\varepsilon)
\end{aligned}
$$

Setting optimal

$$\varepsilon = \left(\frac{1}{C_A n^{1/6}}\right)^{\frac{1}{1+8d/6p}}$$

gives the initial statement.

□

3 Statistical learning theory

Below we obtain some properties of parametric statistical models. Two important questions of statistical learning theory are how the model's parameter distribution depends of its dimension and dataset size and the second question is how close the parameter distribution to some Gaussian distribution? Handling these questions we analyse Taylor expansion of the Likelihood function, consider approximation of MLE by a sum of independent vectors and further involve Bootstrap for the parameter resampling. We also consider the distribution of Likelihood maximum which according to Wilks Theorem is expected to be close by distribution to Chi-square. Below we start with sufficient conditions for quadratic likelihood approximation and in the following subsections we extend these conditions for Bootstrap and Lasso models.

3.1 Quadratic Likelihood approximation

In this section we consider an infinite dimensional statistical model $L(\theta)$. Let parameter θ consists of two parts (u, v), such that $u = \theta_{1 \ldots p} \in I\!\!R^p$. Working with a finite dataset we are going to find MLE deviations basing on three assumptions listed below. Further we will specify these assumptions for independent models and apply in Bootstrap procedure.

Let the Likelihood function $L(\theta) = L(\theta, \mathbb{Y})$ depends on parameters vector $\theta = (u, v)$ and random dataset Y of size n. Denote parameter's MLE and refernce values:

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta)$$

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, I\!\!E L(\theta)$$

We are going to study deviations of $\widehat{\theta}$ and u in the following sense. For some matrix D and random vector ξ

1. $\|\widehat{\theta} - \theta^*\|$ is expected to be of order $O(1/\sqrt{n})$
2. $D(\widehat{\theta} - \theta^*) \approx \xi$
3. $L(\widehat{\theta}) - L(\theta^*) \approx \|\xi\|^2/2$

Denote the stochastic part of the Likelihood

$$\zeta(\theta) = L(\theta) - I\!\!E L(\theta)$$

Involve the Fisher matrix

$$D^2 = -\nabla^2 I\!\!E L(\theta^*) = \begin{pmatrix} \mathbb{F}_u & \mathbb{F}_{uv} \\ \mathbb{F}_{vu} & \mathbb{F}_v \end{pmatrix}$$

It would be easier to deal with the model if matrix $\mathbb{F}$ has block-diagonal view ($\mathbb{F}_{uv} = 0$). One can make parameter replacement in order to satisfy to this condition. Define a new variable $\vartheta = \vartheta(u, v)$ such that

$$\nabla_u \nabla_\vartheta^T I\!\!E L(\theta^*) = \nabla_\vartheta \nabla_u^T I\!\!E L(\theta^*) = 0$$

and

$$\vartheta = v + D_v^{-2} \mathbb{F}_{vu} u,$$

or in other words the parameters transformation matrix is

$$S = \begin{pmatrix} I & 0 \\ D_v^{-2} \mathbb{F}_{vu} & I \end{pmatrix}, \quad S^{-1} = \begin{pmatrix} I & 0 \\ -D_v^{-2} \mathbb{F}_{vu} & I \end{pmatrix}$$

The gradient in the new coordinates (u, ϑ) may be obtained by rule $\nabla(u, \vartheta) = (S^{-1})^T \nabla(u, v)$. Use notation $\breve{\nabla}$ for the first part of it

$$\breve{\nabla} = \nabla_u(u, \vartheta) = \nabla_u - \mathbb{F}_{uv} D_v^{-2} \nabla_v$$

The Fisher matrix after parameters replacement changes by rule $\mathbb{F}(u, \vartheta) = (S^{-1})^T \mathbb{F} S^{-1}$, so in the new coordinates it has view

$$D^2(u, \vartheta) = -\nabla^2 I\!\!E L(u^*, \vartheta^*) = \begin{pmatrix} \breve{D}^2 & 0 \\ 0 & D_\vartheta^2 \end{pmatrix}$$

$$\breve{D}^2 = \mathbb{F}_u - \mathbb{F}_{uv} D_v^{-2} \mathbb{F}_{vu}$$

Define a local region around point θ^*

$$\Omega(\mathtt{r}_0) = \{\theta : \|D(\theta - \theta^*)\| \le \mathtt{r}_0\}$$

Now we write down three conditions on the Likelihood derivatives essential for the deviations of $\widehat{\theta}$. The first and second conditions should be satisfied in the local region $\Omega(\mathtt{r}_0)$. The third condition is required to make expansion of the previous two conditions to the whole parameter space $I\!\!R^\infty$. Further we will show that these conditions are also sufficient for deviation bounds of the parameter $\widehat{u}$ or in other words from deviations bound of $\widehat{\theta}$ follows bound of $\widehat{u}$.

Assumption 1: In the region $\Omega(\mathtt{r}_0)$

$$\left\| -D^{-1}\{\nabla I\!\!E L(\theta) - \nabla I\!\!E L(\theta^*)\} - D(\theta - \theta^*) \right\| \le \delta(\mathtt{r}_0) \mathtt{r}_0$$

Assumption 2: In the region $\Omega(\mathtt{r}_0)$ with probability $1 - e^{-t}$

$$\sup_{\theta \in \Omega(\mathtt{r}_0)} \left\| D^{-1}\{\nabla\zeta(\theta) - \nabla\zeta(\theta^*)\} \right\| \le \mathfrak{z}(t) \mathtt{r}_0$$

Assumption 3: The Likelihood function is convex ($-\nabla^2 L(\theta) \ge 0$) or the expectation of Likelihood function is upper-bounded by a strongly convex function $\forall \mathtt{r}_0 < \mathtt{r}$, $\mathtt{r} =$

$\|D(\theta - \theta^*)\|$

$$\mathbb{E}L(\theta^*) - \mathbb{E}L(\theta) \geq \{1 - \delta(\mathtt{r}_0)\} \left(\mathtt{r}_0 \mathtt{r} - \frac{1}{2}\mathtt{r}_0^2\right) + C_{\mathfrak{z}}\mathtt{r}^2$$

Def. $(\diamondsuit)$

$$\diamondsuit(\mathtt{r}_0, t) = \{\delta(\mathtt{r}_0) + \mathfrak{z}(t)\}\mathtt{r}_0$$

Theorem 3.1 (Spokoiny (2016))**.** Let the Likelihood function be convex $(-\nabla^2 L(\theta) \geq 0)$ and for $\mathtt{r}_0$ (assigned further) it holds $\delta(\mathtt{r}_0) + \mathfrak{z}(t) \leq 1/2$. Then under Assumptions 1, 2 with probability $1 - e^{-t}$

$$\mathtt{r}_0 \leq 4\|D^{-1}\nabla L(\theta^*)\|$$

$$\|D(\widehat{\theta} - \theta^*) - D^{-1}\nabla L(\theta^*)\| \leq \diamondsuit(\mathtt{r}_0, t)$$

$$\|\breve{D}(\widehat{u} - u^*) - \breve{D}^{-1}\breve{\nabla} L(\theta^*)\| \leq \diamondsuit(\mathtt{r}_0, t)$$

Remark. Case with non-convex function in Assumption 3 is considered in lecture notes Spokoiny (2016). In this case the previous statements hold under additional condition

$$C_{\mathfrak{z}} > \mathfrak{z}(t)$$

Proof. From $(-\nabla^2 L(\theta) \geq 0)$ and $(L(\widehat{\theta}) > L(\theta^*))$ follows that the local region $\Omega(\mathtt{r})$ that includes $\widehat{\theta}$ should cover the next region

$$\Omega(\mathtt{r}) \supset \{\theta : L(\theta) \geq L(\theta^*)\}$$

Estimate the minimum possible radius of $\Omega(\mathtt{r})$ that satisfy to the previous condition.

$$0 \geq L(\theta^*) - L(\theta)$$

$$= -(\theta - \theta^*)^T \nabla L(\theta^*) - \frac{1}{2}(\theta - \theta^*)^T \nabla^2 \zeta(\theta_0)(\theta - \theta^*) + \frac{1}{2}\|D(\theta_0)(\theta - \theta^*)\|^2$$

$$\{\text{with probability } (1 - e^{-t})\} \geq -\|D^{-1}\nabla L(\theta^*)\|\mathtt{r} - \frac{\mathfrak{z}(t)}{2}\mathtt{r}^2 + \frac{1 - \delta(\mathtt{r})}{2}\mathtt{r}^2$$

$$\mathtt{r}(1 - \delta(\mathtt{r}) - \mathfrak{z}(t)) \leq 2\|D^{-1}\nabla L(\theta^*)\|$$

$$\mathtt{r} \leq 4\|D^{-1}\nabla L(\theta^*)\|$$

From Assumptions 1, 2 follows that

$$\|D(\widehat{\theta} - \theta^*) + D^{-1}\{\nabla L(\widehat{\theta}) - \nabla L(\theta^*)\}\| \leq \diamondsuit(\mathtt{r}_0, t)$$

$$\|D(\widehat{\theta} - \theta^*) - D^{-1}\nabla L(\theta^*)\| \leq \diamondsuit(\mathtt{r}_0, t)$$

Not that for the coordinates transform S there is an invariant:

$$\left\| \begin{pmatrix} \breve{D} & 0 \\ 0 & D_{\vartheta} \end{pmatrix} \begin{pmatrix} u - u^* \\ \vartheta - \vartheta^* \end{pmatrix} + \begin{pmatrix} \breve{D}^{-1} & 0 \\ 0 & D_{\vartheta}^{-1} \end{pmatrix} \begin{pmatrix} \breve{\nabla} L(u, \vartheta) - \breve{\nabla} L(u^*, \vartheta^*) \\ \nabla_{\vartheta} L(u, \vartheta) + \nabla_{\vartheta} L(u^*, \vartheta^*) \end{pmatrix} \right\|$$

$$= \|D(\theta - \theta^*) + D^{-1}\{\nabla L(\theta) - \nabla L(\theta^*)\}\|$$

Since

$$\left\| \begin{pmatrix} \breve{D} & 0 \\ 0 & D_\vartheta \end{pmatrix} \begin{pmatrix} u \\ \vartheta \end{pmatrix} \right\|^2 = \theta^T S^T [(S^{-1})^T D^2 (S^{-1})] S\theta = \|D\theta\|^2$$

$$\left\| \begin{pmatrix} \breve{D}^{-1} & 0 \\ 0 & D_\vartheta^{-1} \end{pmatrix} \begin{pmatrix} \breve{\nabla} \\ \nabla_\vartheta \end{pmatrix} \right\|^2 = \nabla^T S^{-1} [(S^{-1})^T D^2 (S^{-1})]^{-1} (S^{-1})^T \nabla$$

$$= \nabla^T D^{-2} \nabla$$

$$\begin{pmatrix} \breve{\nabla} \\ \nabla_\vartheta \end{pmatrix}^T \begin{pmatrix} u \\ \vartheta \end{pmatrix} = \nabla^T S^{-1} S\theta = \nabla^T \theta$$

Subsequently basing on this invariant

$$\|\breve{D}(\widehat{u} - u^*) - \breve{D}^{-1} \breve{\nabla} L(\theta^*)\|$$

$$\leq \|D(\widehat{\theta} - \theta^*) - D^{-1} \nabla L(\theta^*)\| \leq \diamondsuit(\mathtt{r}_0, t)$$

□

3.2 Independent models

Independent models are models with independent observations $\mathbb{Y} = (Y_1, \ldots, Y_n)$

$$L(\theta, \mathbb{Y}) = \sum_{i=1}^{n} l(\theta, Y_i)$$

They are very popular in statistical literature and have many references to classical theory. Here we obtain a simpler variant of **Assumption 2**. Involve three basic lemmas for that.

Lemma 3.1 (Bernstein's inequality Boucheron S. (2013))**.** Let $X_1 \ldots X_n$ be independent real-valued random variables. Assume that exist positive numbers $\mathbf{v}$ and R such that

$$\mathbf{v^2} = \sum_{i=1}^{n} I\!\!E X_i^2$$

and for all integers $q \geq 3$

$$\sum_{i=1}^{n} I\!\!E\, [X_i]_+^q \leq \frac{q!}{2} \mathbf{v^2} R^{q-2}$$

Then for all $\lambda \in (1, 1/R)$

$$\log I\!\!E e^{\lambda \sum_i (X_i - I\!\!E X_i)} \leq \frac{\mathbf{v}^2 \lambda^2}{2(1 - R\lambda)}$$

Def. (H) Entropy of the model parameter space $\Omega(\mathtt{r}_0)$ with metric $\|D(\theta_1 - \theta_2)\|$

$$H_1(\Omega) = \int_0^1 \sqrt{\log N(\varepsilon \mathtt{r}_0, \Omega)} d\varepsilon, \quad H_2(\Omega) = \int_0^1 \sqrt{\log N(\varepsilon \mathtt{r}_0, \Omega)} d\varepsilon$$

where $N(\varepsilon, \Omega)$ is covering number.

Lemma 3.2 (Dudley's entropy integral Boucheron S. (2013))**.** Let $f(\theta)$ $(\theta \in \Omega(\mathtt{r}_0))$ be a collection of random variables such that for some constants $a, \mathbf{v}, R > 0$, for all $\theta_1, \theta_2 \in \Omega$ and all $0 < \lambda < (Rd(\theta_1, \theta_2))^{-1}$

$$\log I\!\!E \exp\{\lambda(f(\theta_1) - f(\theta_2))\} \le a\lambda d(\theta_1, \theta_2) + \frac{\mathbf{v^2} \lambda^2 d^2(\theta_1, \theta_2)}{2(1 - R\lambda d(\theta_1, \theta_2))}$$

Then

$$I\!\!E[\sup_\theta f(\theta) - f(\theta^*)] \le 3a\mathtt{r}_0 + 6\mathtt{r}_0 \mathbf{v} H_1(\Omega) + 6\mathtt{r}_0 R H_2(\Omega)$$

Lemma 3.3 (Bousquet's inequality Boucheron S. (2013))**.** Consider independent random variables $X_1 \dots X_n$ and let $\mathcal{F} : X \to \mathbb{R}$ be countable set of functions that satisfy conditions $I\!\!E f(X_i) = 0$ and $\|f\|_\infty \le R$. Define

$$Z = \sup_{f \in \mathcal{F}} \sum_{i=1}^n f(X_i)$$

Let

$$\mathbf{v}^2 = \sup_{f \in \mathcal{F}} \sum_{i=1}^n I\!\!E f^2(X_i)$$

then with probability $1 - e^{-t}$

$$Z < I\!\!E Z + \sqrt{2t(\mathbf{v}^2 + 2R I\!\!E Z)} + \frac{tR}{3}$$

If the functions class is not bounded by norm ($\|f\|_\infty \le R$) one may use Lemma from Spokoiny (2016).

Lemma 3.4. Consider independent random variables $X_1 \dots X_n$ and let $\mathcal{F} : X \to \mathbb{R}$ be parametric set of functions that satisfy conditions $I\!\!E f(X_i, \theta) = 0$ and Bernstein type inequalities $q \ge 2$ for all $\theta_1, \theta_2 \in \Omega(\mathtt{r}_0)$

$$\sum_{i=1}^n I\!\!E |f(X_i, \theta_1) - f(X_i, \theta_2)|^q \le \frac{q!}{2} \mathbf{v}^2 R^{q-2} d^q(\theta_1, \theta_2)$$

or exponential moments inequalities

$$\sum_{i=1}^n \log I\!\!E \exp\{\lambda(f(X_i, \theta_1) - f(X_i, \theta_2))\} \le \frac{\mathbf{v^2} \lambda^2 d^2(\theta_1, \theta_2)}{2(1 - R\lambda d(\theta_1, \theta_2))}$$

Then with probability $1 - e^{-t}$

$$\sup_{\theta \in \Omega(\mathtt{r}_0)} \left| \sum_{i=1}^{n} f(X_i, \theta) - \sum_{i=1}^{n} f(X_i, \theta^*) \right| < \mathbf{v}\mathtt{r}_0 \Big(3.4 + 8H_1(\Omega) + 4\sqrt{t} + 30R(4R^2 t + 1)H_2(\Omega) \Big)$$

Apply three previous Lemmas in order to simplify **Assumption 2** for independent models. Likelihood of an independent model is a sum of independent functions:

$$(L - \mathbb{E}L)(\theta) = \zeta(\theta) = \sum_{i=1}^{n} \zeta_i(\theta)$$

Note that ζ_i depends from the implicit i-th element from the dataset, such that $\zeta_i(\theta) = \zeta_i(\theta, Y_i)$.

Assumption 2i: Let D^2 be Fisher's matrix defined above and $\forall \theta \in \Omega(\mathtt{r}_0)$

$$\sup_{\|u\|=1} \sum_{i=1}^{n} \mathbb{E}(u^T D^{-1} \nabla^2 \zeta_i(\theta) D^{-1} u)^2 \leq \mathbf{v}^2$$

and

$$\|D^{-1} \nabla^2 \zeta_i(\theta) D^{-1}\| \leq R$$

or Bernstein type inequalities hold for $q \geq 3$

$$\sum_{i=1}^{n} \mathbb{E}\|D^{-1} \nabla^2 \zeta_i(\theta) D^{-1}\|^q \leq \frac{q!}{2} \mathbf{v}^2 R^{q-2}$$

Theorem 3.2. Assumption 2 follows from **Assumption 2i** and in the first case when the second derivative is bounded by R

$$\mathfrak{z}(t) \leq \mathbf{v}(6\sqrt{2p_D} + \sqrt{2t}) + R(12p_D + 12\sqrt{tp_D} + t/3)$$

where p_D is entropy of ellipsoid with matrix D

$$p_D = \sqrt{\sum_i \frac{\log^2(\lambda_i^2(D))}{\lambda_i^2(D)}}$$

If norm of the second derivative of ζ is not bounded but has Bernstein type inequalities then

$$\mathfrak{z}(t) \leq \mathbf{v}\Big(3.4 + 4\sqrt{2p_D} + 4\sqrt{t} + 30R(4R^2 t + 1)p_D \Big)$$

Proof. Set a random process for each i:

$$X_i(\gamma, \theta) = \frac{1}{\mathtt{r}} \gamma^T \{\nabla \zeta_i(\theta) - \nabla \zeta_i(\theta^*)\}$$

Such that

$$\sup_{\|D\gamma\|\le \mathtt{r}} \sum_i X_i(\gamma,\theta) = \|D^{-1}\{\nabla\zeta(\theta) - \nabla\zeta(\theta^*)\}\|$$

$\forall$ fixed $(\gamma,\theta) \in \varOmega(\mathtt{r},0) \times \varOmega(\mathtt{r},\theta^*)$ and $\|u\| = 1$:

$$\sup_u I\!\!E \sum_i (\nabla_\theta X_i(\gamma,\theta)^T D^{-1} u)^2 = \sup_u I\!\!E \sum_i \frac{1}{\mathtt{r}} (\gamma \nabla^2 \zeta(\theta)^T D^{-1} u)^2$$

$$\le \sup_u I\!\!E \sum_i (u^T D^{-1} \nabla^2 \zeta(\theta)^T D^{-1} u)^2 \le \mathbf{v}^2$$

Analogically

$$\sup_u I\!\!E \sum_i (\nabla_\gamma X_i(\gamma,\theta)^T D^{-1} u)^2 \le \mathbf{v}^2$$

$\forall i \in 1,\dots,n$:

$$\|D^{-1}\nabla X_i(\gamma,\theta)\| \le R$$

Apply Lemma 3.1 for the sum of random variables $X(\gamma,\theta) = \sum_i X_i(\gamma,\theta)$ when (γ,θ) are fixed.

$$\begin{aligned}
&\log I\!\!E \exp\lambda\left(X(\gamma_1,\theta_1) - X(\gamma_2,\theta_2)\right) \\
&= \log I\!\!E \exp\lambda\left((\gamma_1-\gamma_2)^T \nabla_\gamma X(\gamma,\theta)\right) + \log I\!\!E \exp\lambda\left((\theta_1-\theta_2)^T \nabla_\theta X(\gamma,\theta)\right) \\
&\le \sup_u \log I\!\!E \exp\lambda\left(\|D(\gamma_1-\gamma_2)\| u^T D^{-1} \nabla_\gamma X(\gamma,\theta)\right) \\
&+ \sup_u \log I\!\!E \exp\lambda\left(\|D(\theta_1-\theta_2)\| u^T D^{-1} \nabla_\theta X(\gamma,\theta)\right) \\
&\le \frac{\mathbf{v}^2\lambda^2\|D(\gamma_2-\gamma_1)\|^2}{2(1-R\lambda\|D(\gamma_2-\gamma_1)\|)} + \frac{\mathbf{v}^2\lambda^2\|D(\theta_2-\theta_1)\|^2}{2(1-R\lambda\|D(\theta_2-\theta_1)\|)} \\
&\le \frac{\mathbf{v}^2\lambda^2 d_{12}^2}{2(1-R\lambda d_{12})}
\end{aligned}$$

$$d_{12}^2 = \|D(\theta_2-\theta_1)\|^2 + \|D(\gamma_2-\gamma_1)\|^2$$

Denote

$$\varUpsilon = \varOmega(\mathtt{r}) \times \varOmega(\mathtt{r})$$

such that $\log N(\varepsilon,\varUpsilon) = 2\log N(\varepsilon,\varOmega(\mathtt{r}))$. Then with Lemma 3.2 we obtain

$$E = I\!\!E \sup_{\gamma,\theta} X(\gamma,\theta) \le 6\mathtt{r}\mathbf{v}\sqrt{2}H_1 + 12\mathtt{r}RH_2$$

Application of Lemma 3.3 to the random variable $Z = \sup_{\gamma,\theta} X(\gamma,\theta)$ completes the proof.

$$\mathfrak{z}(t) \le E + \sqrt{2t(\mathbf{v}^2 + 2RE)} + \frac{tR}{3}$$

where

$$E = 6\mathbf{v}\sqrt{2p_D} + 12Rp_D$$

The second case follows from Lemma 3.4.

□

3.3 Entropy

Below one can read a short excerpt about an entropy of ball and ellipsoid. The general formula for the covering number N of a convex set Ω in $\mathbb{R}^p$ with an arbitrary distance $d(\theta_1, \theta_2)$ is

$$N(\varepsilon, \Omega) \le \frac{volume(\Omega + (\varepsilon/2)B_1)}{volume(B_1)} \left(\frac{2}{\varepsilon}\right)^p$$

where B_1 is a unit ball.

Ball entropy: Let $\Omega = B_{\mathtt{r}}$ and $d(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|$ then

$$N(\varepsilon, B_1) \le \left(1 + \frac{2}{\varepsilon}\right)^p$$

and since $N(\varepsilon\mathtt{r}, B_{\mathtt{r}}) = N(\varepsilon, B_1)$

$$\int_0^{\mathtt{r}/2} \sqrt{\log N(\varepsilon, B_{\mathtt{r}})} d\varepsilon = \mathtt{r} \int_0^{1/2} \sqrt{\log N(\varepsilon, B_1)} d\varepsilon$$

$$\le \mathtt{r}\sqrt{p} \int_0^{1/2} \sqrt{\log(3/\varepsilon)} d\varepsilon \le 0.83\mathtt{r}\sqrt{p}$$

and

$$\int_0^{\mathtt{r}/2} \log N(\varepsilon, B_{\mathtt{r}}) d\varepsilon \le 1.4\mathtt{r}p$$

Ellipsoid entropy: Let $\Omega = \mathcal{E}_{\mathtt{r}}(D)$ and $d(\theta_1, \theta_2) = \|D(\theta_1 - \theta_2)\|$. The entropy in this case is rather complicate in calculation. So we provide here only the the final statement from V. Spokoiny's lecture notes Spokoiny (2016).

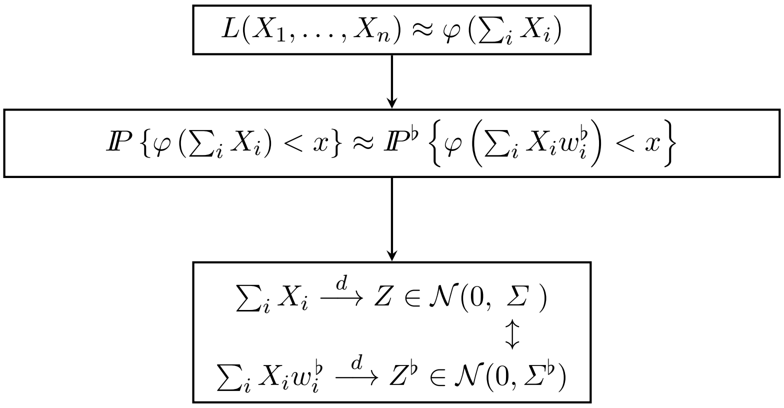
$$\int_0^{\mathtt{r}/2} \sqrt{\log N(\varepsilon, \mathcal{E}_{\mathtt{r}}(D))} d\varepsilon \lesssim \frac{\mathtt{r}}{2}\sqrt{\alpha - 1}\sqrt{\sum_i \frac{\log^{\alpha}(\lambda_i^2(D))}{\lambda_i^2(D)}}$$

and

$$\int_0^{\mathtt{r}/2} \log N(\varepsilon, \mathcal{E}_{\mathtt{r}}(D)) d\varepsilon \lesssim \frac{\mathtt{r}}{4} \sum_i \frac{1}{\lambda_i(D)}$$

3.4 Bootstrap

Approximation of measure $I\!P$ of some statistic $L(X_1, \ldots, X_n)$ by corresponded bootstrap measure $I\!P^\flat$ with statistic $L(X_1 w_1^\flat, \ldots, X_n w_n^\flat)$ could be done in three steps: 1) Find a close to $L(X_1, \dots, X_n)$ function $\varphi\left(\sum_i X_i\right)$; 2) Make Gaussian approximation for $\varphi\left(\sum_i X_i\right)$ and $\varphi(\sum_i X_i w_i^\flat)$ by means of Lemma 2.2; 3) Compare the Gaussian variables $\varphi\left(Z\right)$ and $\varphi(Z^\flat)$ using Lemma 2.3.


$L(X_1, \dots, X_n) \approx \varphi\left(\sum_i X_i\right)$
$\mathbb{P}\left\{\varphi\left(\sum_i X_i\right) < x\right\} \approx \mathbb{P}^\flat\left\{\varphi\left(\sum_i X_i w_i^\flat\right) < x\right\}$
$\sum_i X_i \xrightarrow{d} Z \in \mathcal{N}(0,\ \varSigma\ )$
$\updownarrow$
$\sum_i X_i w_i^\flat \xrightarrow{d} Z^\flat \in \mathcal{N}(0, \varSigma^\flat)$

The bootstrap procedure allows to sample Likelihood function with two options: each Likelihood component is multiplied by weight (*weighted bootstrap*) or new data is resampled from empirical distribution (*empirical bootstrap*). The Likelihood function in weighted bootstrap case is a convolution of i.i.d weights $(w_1^\flat, \dots, w_n^\flat)$ and independent components $\{l_i(\theta)\}_{i=1}^n$:

$$L^\flat(\theta) = \sum_{i=1}^n w_i^\flat l_i(\theta)$$

$$\zeta^\flat(\theta) = L^\flat(\theta) - L(\theta) = \sum_{i=1}^n (w_i^\flat - 1) l_i(\theta)$$

Denote parameter's MLE and reference values:

$$\widehat{\theta}^\flat = \operatorname*{argmax}_\theta L^\flat(\theta)$$

$$(\theta^\flat)^* = \widehat{\theta} = \operatorname*{argmax}_\theta L(\theta)$$

Note that in this setting $\{l_i(\theta)\}_{i=1}^n\}$ are non random functions. Each weight element has $\operatorname{Var}^\flat w_i^\flat = 1$ and $\mathbb{E}^\flat w_i^\flat = 1$, which is made in order to satisfy to the following conditions

$$\begin{aligned} \mathbb{E}^\flat L^\flat(\theta) &= L(\theta) \\ \operatorname{Var}^\flat \nabla L^\flat(\theta) &= \sum_{i=1}^n \nabla l_i(\theta) \nabla l_i(\theta)^T \end{aligned}$$

It is expected that $\operatorname{Var}^\flat \nabla L^\flat(\theta)$ is close to $\operatorname{Var} \nabla L(\theta)$ in $\varOmega(\mathtt{r}_0)$. Remind that by definition

$$\varOmega(\mathtt{r}_0) = \{\theta : \|D(\theta - \theta^*)\| \le \mathtt{r}_0\}$$

The radius $\mathtt{r}$ will be defined later. Let model $L(\theta)$ fulfills the assumptions 1, 2, 3 from Section 3.1. Check these assumptions for model $L^\flat(\theta)$.
Proposition 1b: In the region $\varOmega(\mathtt{r}_0)$

$$\left\| -D^{-1}\{\nabla L(\theta) - \nabla L(\widehat{\theta})\} - D(\theta - \widehat{\theta}) \right\| \le \delta^\flat(\mathtt{r}_0)\mathtt{r}_0$$

$$\delta^\flat(\mathtt{r}_0) = 2(\delta(\mathtt{r}_0) + \mathfrak{z}(t))$$

Proof. It follows from Assumptions 1 and 2 for $L(\theta)$.

$$\begin{aligned} &\left\| -D^{-1}\{\nabla L(\theta) - \nabla L(\widehat{\theta})\} - D(\theta - \widehat{\theta}) \right\| \\ \le &\left\| -D^{-1}\{\nabla L(\theta) - \nabla L(\theta^*)\} - D(\theta - \theta^*) \right\| \\ + &\left\| -D^{-1}\{\nabla L(\widehat{\theta}) - \nabla L(\theta^*)\} - D(\widehat{\theta} - \theta^*) \right\| \end{aligned}$$

□

Proposition 2b: Under conditions from **Assumption 2i** and additional condition

$$\sup_{\theta \in \varOmega(\mathtt{r})} \|D^{-1}\nabla^2 El_i(\theta) D^{-1}\| \le R$$

in the region $\varOmega(\mathtt{r}_0)$ with probability $1 - e^{-t}$

$$\sup_{\theta \in \varOmega(\mathtt{r}_0)} \left\| D^{-1}\{\nabla\zeta^\flat(\theta) - \nabla\zeta^\flat(\widehat{\theta})\} \right\| \le \mathfrak{z}^\flat(t)\mathtt{r}_0$$

where

$$\mathfrak{z}^\flat(t) = \mathfrak{z}(t, \mathbf{v}, R^\flat) + \mathfrak{z}(t, \sqrt{n}R, R^\flat)$$

and $\mathfrak{z}(t, \mathbf{v}, R)$ defined in Theorem 3.2 and

$$R^{q-2}\mathbb{E}^\flat|w^\flat - 1|^q \le (R^\flat)^{q-2}$$

Proof. According to Theorem 3.2 for the second assumption one have to find the bounds for

$$\sup_{\|u\|=1} \sum_{i=1}^{n} \mathbb{E}(u^T D^{-1}\nabla^2 l_i(\theta) D^{-1}u)^2$$

and Bernstein type inequalities $q \ge 3$

$$\sup_{\theta \in \varOmega(\mathtt{r})} \sum_{i=1}^{n} \mathbb{E}^\flat|w^\flat - 1|^q \mathbb{E}\|D^{-1}\nabla^2 l_i(\theta) D^{-1}\|^q$$

One can split $l_i(\theta)$ into $\zeta_i(\theta)$ and $\mathbb{E}l_i(\theta)$. These bounds holds for $\zeta_i(\theta)$ by Assumption 2i with replacement R to $R^\flat$. For $I\!\!El_i(\theta)$ it holds with replacement $\mathbf{v}$ to $\sqrt{n}R$ and R to $R^\flat$.

$$\sup_{\theta\in\Omega(\mathtt{r})}\left\|D^{-1}\{\nabla I\!\!EL^\flat(\theta)-\nabla I\!\!EL^\flat(\widehat{\theta})\}\right\|\le\mathfrak{z}(t,\sqrt{n}R,R^\flat)\mathtt{r}_0$$

□

Proposition 3b: From **Assumption 2** and **Assumption 3** follows that $\forall\mathtt{r}_0<\mathtt{r}$ with probability $1-e^{-t}$

$$L(\theta^*)-L(\theta)\ge(1-\delta(\mathtt{r}_0))\left(\mathtt{r}_0\mathtt{r}-\frac{1}{2}\mathtt{r}_0^2\right)+(C_{\mathfrak{z}}-\mathfrak{z}(t))\mathtt{r}^2-\|D^{-1}\nabla\zeta(\theta^*)\|\mathtt{r}$$

where

$$\mathtt{r}=\|D(\theta-\theta^*)\|$$

Proof.

$$L(\theta^*)-L(\theta)=I\!\!EL(\theta^*)-I\!\!EL(\theta)+\zeta(\theta^*)-\zeta(\theta)$$

From Assumption 2 one can bound the last difference

$$|\zeta(\theta)-\zeta(\theta^*)-\nabla\zeta(\theta^*)(\theta-\theta^*)|\le\mathfrak{z}(t)\|D(\theta-\theta^*)\|^2$$

$$|\zeta(\theta)-\zeta(\theta^*)|\le\|D^{-1}\nabla\zeta(\theta^*)\|\|D(\theta-\theta^*)\|+\mathfrak{z}(t)\|D(\theta-\theta^*)\|^2$$

Thus

$$\begin{aligned}L(\theta^*)-L(\theta)\ge(1-\delta(\mathtt{r}_0))(\mathtt{r}_0\|D(\theta-\theta^*)\|-\mathtt{r}_0^2/2)+C_{\mathfrak{z}}\|D(\theta-\theta^*)\|^2\\-\|D^{-1}\nabla\zeta(\theta^*)\|\|D(\theta-\theta^*)\|-\mathfrak{z}(t)\|D(\theta-\theta^*)\|^2\end{aligned}$$

□

Summarise the propositions.

Theorem 3.3. Let **Assumptions 1, 2i** and **3** be true. Assume also additional bootstrap conditions

$$\sup_{\theta\in\Omega(\mathtt{r})}\|D^{-1}\nabla^2El_i(\theta)D^{-1}\|\le R$$

$$R^{q-2}I\!\!E^\flat|w^\flat-1|^q\le(R^\flat)^{q-2}$$

Then all properties of model $L(\theta)$ obtained from Assumptions 1,2 and 3 also true for $L^\flat(\theta)$ with replacement of $\diamondsuit(\mathtt{r}_0,t)$ to $\diamondsuit^\flat(\mathtt{r}_0,t)$ and θ^* to $\widehat{\theta}$, where

$$\{2\delta(\mathtt{r}_0)+2\mathfrak{z}(t,\mathbf{v},R)+\mathfrak{z}(t,\mathbf{v},R^\flat)+\mathfrak{z}(t,\sqrt{n}R,R^\flat)\}\mathtt{r}_0=\diamondsuit^\flat(\mathtt{r}_0,t)$$

The local region $\Omega(\mathtt{r}_0)$, $\|D(\theta^\flat-\widehat{\theta})\|\le\mathtt{r}_0$ in this case is $(|w^\flat-1|+1)$ times bigger and has the following radius

$$\mathtt{r}_0=\frac{2\|D^{-1}\nabla L(\theta^*)\|(|w^\flat-1|+1)}{1-\delta(\mathtt{r}_0)}$$

under condition that

$$C_{\mathfrak{z}} > \mathfrak{z}(t) + \mathfrak{z}^{\flat}(t)$$

Proof. Use a short notation for differences of functions L, $L^{\flat}$, ζ

$$L(\theta, \theta^*) = L(\theta) - L(\theta^*)$$

$$\mathtt{r} = \|D(\theta - \theta^*)\|$$

One have to show that $L^{\flat}(\theta, \theta^*) < 0$ for $\mathtt{r} > \mathtt{r}_0$ which means that $\theta^{\flat} \in \Omega(\mathtt{r}_0)$. From Proposition 2b follows

$$|\zeta^{\flat}(\theta, \theta^*) - \nabla\zeta^{\flat}(\theta^*)(\theta - \theta^*)| \le \mathfrak{z}^{\flat}(t)\mathtt{r}^2$$

$$|\nabla\zeta^{\flat}(\theta^*)(\theta - \theta^*)| \le \|D^{-1}\nabla L(\theta^*)\||w^{\flat} - 1|\mathtt{r}$$

Remind that $L^{\flat}(\theta, \theta^*) = \zeta^{\flat}(\theta, \theta^*) + L(\theta, \theta^*)$ and

$$L^{\flat}(\theta, \theta^*) \le L(\theta, \theta^*) + |\zeta^{\flat}(\theta, \theta^*)|$$

Condition $L^{\flat}(\theta, \theta^*) < 0$ follows from

$$L(\theta^*, \theta) > |\zeta^{\flat}(\theta, \theta^*)|$$

So according to Proposition 3b

$$(1 - \delta(\mathtt{r}_0))\left(\mathtt{r}_0\mathtt{r} - \frac{1}{2}\mathtt{r}_0^2\right) + (C_{\mathfrak{z}} - \mathfrak{z}(t))\mathtt{r}^2 - \|D^{-1}\nabla\zeta(\theta^*)\|\mathtt{r}$$

$$> \|D^{-1}\nabla L(\theta^*)\||w^{\flat} - 1|\mathtt{r} + \mathfrak{z}^{\flat}(t)\mathtt{r}^2$$

After simplification it gives the required inequality for $\mathtt{r}_0$ and $C_{\mathfrak{z}}$.

□

Let function $\alpha^{\flat}(\theta, \theta_0)$ denotes quadratic approximation error for the bootstrap Likelihood function.

$$\alpha^{\flat}(\theta, \theta_0) = L^{\flat}(\theta) - L^{\flat}(\theta_0) - (\theta - \theta_0)^T \nabla L^{\flat}(\theta_0) + \frac{1}{2}\|D(\theta - \theta_0)\|^2$$

Theorem 3.4 (Weighted bootstrap Wilks)**.** Under conditions from Theorem 3.3 with probability $1 - 3e^{-t}$

$$|\alpha(\theta^{\flat}, \widehat{\theta})| \le \diamondsuit^{\flat}(\mathtt{r}_0, t)\mathtt{r}_0 \tag{Ab}$$

and consequently

$$\left| L^{\flat}(\theta^{\flat}, \widehat{\theta}) - \frac{\|D^{-1}L^{\flat}(\widehat{\theta})\|}{2} \right| \le \diamondsuit^{\flat}(\mathtt{r}_0, t)\mathtt{r}_0$$

where $\theta^{\flat}$, $\widehat{\theta}$ are MLE parameters of the weighted and non-weighted Likelihood functions.

A modification of Fisher expansion (Theorem 2.2 in Spokoiny (2012b)) for the bootstrap Likelihood could be proved using the following property

$$\chi^\flat(\theta,\theta_0) = D^{-1}(\nabla L^\flat(\theta) - \nabla L^\flat(\theta_0)) + D(\theta-\theta_0),$$

$$\chi^\flat(\theta,\theta_0) = D^{-1}\nabla\alpha^\flat(\theta,\theta_0).$$

Theorem 3.5 (Weighted bootstrap Fisher)**.** Under conditions from Theorem 3.3 with probability $1-3e^{-t}$

$$\|D(\theta^\flat - \widehat{\theta}) - D^{-1}\nabla L^\flat(\widehat{\theta})\| \le \diamondsuit^\flat(\mathtt{r}_0, t)$$

where $\theta^\flat$, $\widehat{\theta}$ are MLE parameters of the weighted and non-weighted Likelihood functions.

In empirical bootstrap case with dataset size n one deal with

$$L^\epsilon(\theta) = \sum_{i=1}^{n} l_{k(i)}(\theta)$$

where random indexes $k(i) \in \{1,\dots,n\}$ and independent.

$$I\!\!E^\epsilon L^\varepsilon(\theta) = L(\theta)$$

$$\zeta^\epsilon(\theta) = L^\epsilon(\theta) - L(\theta)$$

Denote parameter's MLE and refernce values:

$$\widehat{\theta}^\epsilon = \operatorname*{argmax}_\theta L^\epsilon(\theta)$$

$$(\theta^\epsilon)^* = \widehat{\theta} = \operatorname*{argmax}_\theta L(\theta)$$

Define

$$w_i^\epsilon = (0,\dots,\underset{k(i)}{1},\dots,0)^T$$

and

$$l(\theta) = [l_1(\theta),\dots,l_n(\theta)]$$

Then

$$L^\epsilon(\theta) = l(\theta)\sum_{i=1}^{n} w_i^\epsilon$$

Propositions 1b and 3b may be also applied to $I\!\!E^\epsilon L^\varepsilon(\theta) = L(\theta)$. Proposition 2b has some differences in this case.

Proposition 2e: Under conditions from **Assumption 2i** and additional condition

$$\sup_{\theta\in\Omega(\mathtt{r}_0)} \|D^{-1}\nabla^2 E l_i(\theta) D^{-1}\| \le R$$

in the region $\Omega(\mathtt{r}_0)$ with probability $1 - e^{-t}$

$$\sup_{\theta \in \Omega(\mathtt{r}_0)} \left\| D^{-1} \{ \nabla \zeta^{\epsilon}(\theta) - \nabla \zeta^{\epsilon}(\widehat{\theta}) \} \right\| \leq \mathfrak{z}^{\epsilon}(t) \mathtt{r}_0$$

where

$$\mathfrak{z}^{\epsilon}(t) = 2\mathfrak{z}\,(t, \mathbf{v}, R) + \mathfrak{z}(t, \sqrt{n}R, R)$$

Proof. Denote

$$w = \sum_{i=1}^{n} w_i^{\epsilon}$$

According to Theorem 3.2 for the second assumption one have to find the bounds for

$$\sup_{\theta \in \Omega(\mathtt{r})} \sup_{\|u\|=1} I\!E^{\epsilon} I\!E [u^T D^{-1} \nabla^2 l(\theta) D^{-1} u (w - I\!E^{\epsilon} w)]^2$$

and Bernstein type inequalities $q \geq 3$

$$\sup_{\theta \in \Omega(\mathtt{r}_0)} \sum_{i=1}^{n} I\!E^{\epsilon} I\!E \| D^{-1} \nabla^2 l(\theta) D^{-1} (w_i^{\epsilon} - I\!E^{\epsilon} w_i^{\epsilon}) \|^q$$

Let Q be ones matrix of size n. Note that

$$I\!E^{\epsilon} (w - I\!E^{\epsilon} w)(w - I\!E^{\epsilon} w)^T = I - \frac{1}{n} Q$$

thus

$$I\!E^{\epsilon} I\!E [u^T D^{-1} \nabla^2 l(\theta) D^{-1} u (w - I\!E^{\epsilon} w)]^2 \leq \sum_{i=1}^{n} I\!E [u^T D^{-1} \nabla^2 l_i(\theta) D^{-1} u]^2$$

As for the second term

$$I\!E^{\epsilon} I\!E \| D^{-1} \nabla^2 l(\theta) D^{-1} (w_i^{\epsilon} - I\!E^{\epsilon} w_i^{\epsilon}) \|^q$$

$$= \frac{1}{n} \sum_i I\!E \left\| D^{-1} \nabla^2 l_i(\theta) D^{-1} - \sum_j \frac{1}{n} D^{-1} \nabla^2 l_j(\theta) D^{-1} \right\|^q$$

$$\leq \frac{1}{n} \sum_i I\!E \left(\| D^{-1} \nabla^2 l_i(\theta) D^{-1} \| + \sum_j \frac{1}{n} \| D^{-1} \nabla^2 l_j(\theta) D^{-1} \| \right)^q$$

One can split $l_i(\theta)$ into $\zeta_i(\theta)$ and $I\!E l_i(\theta)$. These bounds hold for $\zeta_i(\theta)$ by Assumption 2i. For $I\!E l_i(\theta)$ it holds with replacement $\mathbf{v}$ to $2\sqrt{n}R$ and R to $2R$.

$$\sup_{\theta \in \Omega(\mathtt{r})} \left\| D^{-1} \{ \nabla I\!E L^{\epsilon}(\theta) - \nabla I\!E L^{\epsilon}(\widehat{\theta}) \} \right\| \leq \mathfrak{z}(t, 2\sqrt{n}R, 2R) \mathtt{r}_0$$

□

The other statements are identical to weighted bootstrap with replacement $\mathfrak{z}^{\epsilon}(t)$ to $\mathfrak{z}^{\flat}(t)$.

3.5 Sandwich lemma

Gaussian approximation justifies Bootstrap consistency in terms of measures difference with fixed arguments. But in some situations (for example in change point detection) we use Bootstrap in order to find quantile an then compare the measures with this quantile as an argument. The following Lemma allows to extend Bootstrap consistency for the case when measures argument depends on the dataset.

Lemma 3.5. Let differentiable measure $\mathbb{P}^{\flat}$ depends on r.v. from a continuous measure $\mathbb{P}$, $z = (z_1, \dots, z_K)$ is a multivariate quantile. Assume following error in distance between the measures $\forall z$

$$\left| \mathbb{P}\Big(z_1, \dots, z_K\Big) - \mathbb{P}^{\flat}\Big(z_1, \dots, z_K\Big) \right| < \delta$$

Then each quantile $z_k^{\flat}(\alpha)$, $1 \le k \le K$ from measure $\mathbb{P}^{\flat}$ may be bounded by quantile from measure $\mathbb{P}$:

$$z_k(\alpha + \delta) \le z_k^{\flat}(\alpha) \le z_k(\alpha - \delta)$$

where

$$\mathbb{P}\Big(z_k(\alpha)\Big) = 1 - \alpha, \quad \mathbb{P}^{\flat}\Big(z_k^{\flat}(\alpha)\Big) = 1 - \alpha$$

And if $q^{\flat}$ is the multiplicity correction parameter such that

$$\mathbb{P}^{\flat}\Big(z_1(q^{\flat}\alpha), \dots, z_K(q^{\flat}\alpha)\Big) = 1 - \alpha$$

then

$$\left| \mathbb{P}\left(z^{\flat}(q^{\flat}\alpha)\right) - (1 - \alpha) \right| \le (2K + 1)\,\delta$$

Proof. Define two sets

$$\mathbb{Z}_+(\delta) = \{z : \mathbb{P}(z) \le 1 - \alpha + \delta\},$$
$$\mathbb{Z}_-(\delta) = \{z : \mathbb{P}(z) \ge 1 - \alpha - \delta\}$$

For all points z from $\mathbb{Z}_+ \cap \mathbb{Z}_-$ it holds that $|\mathbb{P}(z) - (1 - \alpha)| \le \delta$. If $\mathbb{P}^{\flat}(z^{\flat}) = 1$ then $z^{\flat} \in \mathbb{Z}_+$ since for all fixed $z \in \mathbb{R}^K \setminus \mathbb{Z}_+$: $\mathbb{P}^{\flat}(z) > \mathbb{P}(z) - \delta \ge 1 - \alpha$. Analogically $z^{\flat} \in \mathbb{Z}_-$ and $z^{\flat} \in \mathbb{Z}_+ \cap \mathbb{Z}_-$.

In case $K = 1$ one can choose non-random quantiles in the border of $\mathbb{Z}_+ \cap \mathbb{Z}_-$ which will bound $z^{\flat}$. So each component of $z^{\flat}$ could be bounded in the same way:

$$z_k(\alpha + \delta) \le z^{\flat}(\alpha) \le z_k(\alpha - \delta)$$

In case $K > 1$ the these bounds become random because of multiplicity correction which involves random multiplier $q^\flat$. We have to bound $z_k(q^\flat\alpha+\delta)$ by a non-random quantile in order to use it as an argument for measure $I\!P$.

$$\begin{aligned} z(\alpha+\delta) \le z^\flat(\alpha) &\le z(\alpha-\delta) \\ z^\flat(q^\flat\alpha) &\le z(q^\flat\alpha-\delta) \le z^\flat(q^\flat\alpha-2\delta) \\ 1-\alpha &\le I\!P^\flat\Big(z(q^\flat\alpha-\delta)\Big) \le I\!P^\flat\Big(z^\flat(q^\flat\alpha-2\delta)\Big) \end{aligned}$$

Lemma 3.6. For a differentiable measure $I\!P(\xi<x)$ and event A:

$$\frac{I\!P(\xi<x,A)'_x}{I\!P(\xi<x)'_x} \le 1$$

$$I\!P^\flat\Big(z^\flat(q^\flat\alpha-2\delta)\Big) = I\!P^\flat\Big(z^\flat(q^\flat\alpha)\Big) + \sum_i (I\!P^\flat)'_{z_k^\flat}(z_k^\flat)'\ 2\delta$$

$$(I\!P^\flat)'_{z_k}(z_k^\flat)' = \frac{(I\!P^\flat(z_1^\flat,\dots,z_k^\flat,\dots,z_K^\flat))'_{z_k^\flat}}{(I\!P^\flat(z_k^\flat))'_{z_k^\flat}} \le 1$$

$$I\!P^\flat\Big(z^\flat(q^\flat\alpha-2\delta)\Big) \le 1-\alpha+2K\delta$$

$$1-\alpha \le I\!P^\flat\Big(z(q^\flat\alpha-\delta)\Big) \le 1-\alpha+2K\delta$$

$$1-\alpha-2K\delta \le I\!P^\flat\Big(z(q^\flat\alpha+\delta)\Big) \le 1-\alpha$$

According to the arguments from the beginning of the proof $z(q^\flat\alpha-\delta)$ and $z(q^\flat\alpha-\delta)$ belongs to $\mathbb{Z}_+(2K\delta+\delta)\cap\mathbb{Z}_-(2K\delta+\delta)$. Due to one dimensional parametrization if $z(q^\flat\alpha-\delta)$ there exist two fixed points on the border of $\mathbb{Z}_+(2K\delta+\delta)\cap\mathbb{Z}_-(2K\delta+\delta)$ such that

$$z_+ = \max z(q^\flat\alpha-\delta), \quad z_- = \min z(q^\flat\alpha+\delta)$$

Finally,

$$z_- \le z(q^\flat\alpha+\delta) \le z^\flat(q^\flat\alpha) \le z(q^\flat\alpha-\delta) \le z_+$$

and subsequently

$$1-\alpha-(2K+1)\delta \ge I\!P(z_-) \le I\!P(z^\flat(q^\flat\alpha)) \le I\!P(z_+) \le 1-\alpha+(2K+1)\delta$$

□

3.6 Generalised linear models

Generalized linear models (GLM) are frequently used for modeling the data with special structure: categorical data, binary data, Poisson and exponential data, volatility models, etc. All these examples can be treated in a unified way by a GLM approach. This section specifies the previous results and conditions to this case. Let $\mathbb{Y} = (Y_1, \ldots, Y_n) \sim I\!\!P$ be a sample of independent r.v.'s. The parametric GLM is $Y_i \sim I\!\!P_{\psi_i^T \theta}$, where ψ_i are predefined features in $I\!\!R^p$. Generalised linear model may be presented in form

$$L(\theta) = S^T \theta - A(\theta)$$

where

$$A(\theta) = \sum_{i=1}^{n} g(\psi_i^T \theta), \quad S = \sum_{i=1}^{n} Y_i \psi_i$$

This model has following properties

$$-\nabla^2 I\!\!E L(\theta) = D^2(\theta) = \sum_{i=1}^{n} g''(\psi_i^T \theta) \psi_i \psi_i^T$$

Proposition 1:

$$\delta(\mathtt{r}) = a_g \delta_\psi \mathtt{r}$$

where

$$a_g = \max_x \frac{g'''(x)}{g''(x)}, \quad \delta_\psi = \max_i \left\| D^{-1} \psi_i \right\|$$

Proof. For each $\theta \in \Theta(\mathtt{r}_0)$ and $i \leq n$, it holds

$$\bigl| \psi_i^\top \theta - \psi_i^\top \theta^* \bigr| = \bigl| (D^{-1} \psi_i)^\top D(\theta - \theta^*) \bigr| \leq \| D^{-1} \psi_i \| \, \mathtt{r}_0 \leq \delta_\Psi \, \mathtt{r}_0$$

$$D(\theta) - D(\theta^*) = \sum_{i=1}^{n} \bigl\{ g''(\psi_i^\top \theta) - g''(\psi_i^\top \theta^*) \bigr\} \, \psi_i \psi_i^\top$$

$$g''(\psi_i^\top \theta) - g''(\psi_i^\top \theta^*) = \frac{g'''(\psi_i^\top \theta^\circ)}{g''(\psi_i^\top \theta^*)} \, (\psi_i^\top \theta - \psi_i^\top \theta) \, g''(\psi_i^\top \theta^*)$$

$$\max_{i \leq n} \left| \frac{g'''(\psi_i^\top \theta^\circ)}{g''(\psi_i^\top \theta^*)} \, (\psi_i^\top \theta - \psi_i^\top \theta) \right| \leq a_g(\mathtt{r}_0) \, \delta_\Psi \, \mathtt{r}_0$$

□

Since $g(\cdot)$ is convex, it holds $g''(x) \geq 0$ for any x and thus $D^2(\theta) \geq 0$. An important feature of GLM is that the stochastic component $\zeta(\theta)$ of $L(\theta)$ is *linear* on θ: with $\varepsilon_i = Y_i - I\!\!E Y_i$

$$\zeta(\theta) = \sum_{i=1}^{n} \varepsilon_i \psi_i^\top \theta$$

Consequently $\mathfrak{z}(t) = 0$ in Assumption 2. Linearity in θ of the stochastic component $\zeta(\theta)$ and concavity of the deterministic part $I\!\!E L(\theta)$ allow for a simple and straightforward proof of the result about localisation of the MLE $\widehat{\theta}$ in the region $\Omega(\mathtt{r}_0)$ (Theorem 3.1).

Theorem 3.6. Consider GLM Likelihood function $L(\theta)$. Let for $\mathtt{r}_0$ (assigned further) it holds $\delta(\mathtt{r}_0) \le 1$. Then

$$\mathtt{r}_0 = \frac{2}{1-\delta(\mathtt{r}_0)} \|D^{-1} \nabla L(\theta^*)\|$$

$$\|D(\widehat{\theta} - \theta^*) - D^{-1}\nabla L(\theta^*)\| \le \diamondsuit(\mathtt{r}_0, t)$$

$$\|\breve{D}(\widehat{u} - u^*) - \breve{D}^{-1}\breve{\nabla} L(\theta^*)\| \le \diamondsuit(\mathtt{r}_0, t)$$

$$|2L(\widehat{\theta}, \theta^*) - \|D^{-1}\nabla L(\theta^*)\|^2| \le 2\diamondsuit(\mathtt{r}_0, t)\mathtt{r}_0$$

where

$$\nabla L(\theta^*) = \sum_{i=1}^{n} \psi_i \varepsilon_i$$

and

$$\diamondsuit(\mathtt{r}_0, t) = \delta(\mathtt{r}_0)\mathtt{r}_0 = a_g \delta_\psi \mathtt{r}_0^2$$

Theorem 3.7. Assume that $\varepsilon_i = Y_i - I\!\!E Y_i$ are independent and

$$\log I\!\!E \exp\bigl(\lambda\, \mathfrak{s}_i^{-1} \varepsilon_i\bigr) \le \frac{1}{2}\lambda^2, \qquad i = 1, \ldots, n, \quad |\lambda| \le \mathtt{g}_1$$

$$V_0^2 = \sum_{i=1}^{n} \mathfrak{s}_i^2\, \psi_i \psi_i^\top$$

$$B = V_0 D^{-2} V_0$$

Then with high probability (approx. $1 - 2e^{-t}$)

$$\|D^{-1}\nabla L(\theta^*)\| \le z(B, t)$$

where $z(B, t)$ defined in Lemma 6.6.

Proof. Denote

$$\xi = V_0^{-1} \nabla L(\theta^*)$$

Show that ξ is sub-Gaussian and apply upper bound from Lemma 6.6.

$$\log I\!\!E \exp\{\gamma^\top \xi\} = \sum_{i=1}^{n} \log I\!\!E \exp\bigl(\lambda_i\, \mathfrak{s}_i^{-1} \varepsilon_i\bigr)$$

$$|\lambda_i| = |\gamma^\top V_0^{-1} \psi_i|\, \mathfrak{s}_i \le \mathtt{g}\, \|V_0^{-1}\psi_i\|\, \mathfrak{s}_i \le \mathtt{g}_1$$

$$\log I\!\!E \exp\{\gamma^\top \xi\} \le \frac{\nu_0^2}{2} \sum_{i=1}^{n} \lambda_i^2$$

$$= \frac{\nu_0^2}{2} \sum_{i=1}^{n} \gamma^\top V_0^{-1} (\psi_i \psi_i^\top \mathfrak{s}_i^2)\, V_0^{-1} \gamma = \frac{\nu_0^2}{2} \|\gamma\|^2$$

□

An important property of GLM Likelihood function is convexity: $-\nabla^2 L = -\nabla^2 I\!\!E L \geq 0$. This property helps in MLE concentration proof (Theorem 3.6). The Bootstrap Likelihood is also convex with high probability under an additional condition described in the following statement.

Theorem 3.8. Assume that for some t $\forall i$ $\forall \theta$

$$\sqrt{g''(\psi_i^T \theta)} \left\| D^{-1}(\theta) \psi_i \right\| \sqrt{2t} < 1$$

then with probability $1 - 2e^{-t}$

$$-\nabla^2 L^\flat(\theta) > 0$$

Proof.

$$-\nabla^2 L^\flat(\theta) = \sum_{i=1}^{n} g''(\psi_i^T \theta) \psi_i \psi_i^T w_i^\flat = D(\theta) \left(I + \sum_i d_i d_i^T \varepsilon_i^\flat \right) D(\theta)$$

$$-\nabla^2 L^\flat(\theta) \geq 0 \Leftrightarrow 1 - \lambda_{\max} \left(- \sum_i d_i d_i^T \varepsilon_i^\flat \right) \geq 0$$

where $d_i = \sqrt{g''(\psi_i^T \theta)} D^{-1}(\theta) \psi_i$, $\sum_i d_i d_i^T = I$. Use Lemma 6.4 in order to get matrix deviation bound which states that with probability $1 - 2e^{-t}$

$$\lambda_{\max} \left(- \sum_i d_i d_i^T \varepsilon_i^\flat \right) \leq \max_i \|d_i\| \sqrt{2t}$$

□

Additionally in the bootstrap case $\zeta^\flat(\theta)$ is not linear.

Proposition 2b: Assume condition for $q \geq 3$

$$I\!\!E^\flat |w^\flat - 1|^q \leq \frac{q!}{2}$$

In the region $\varOmega(\mathtt{r}_0)$ with probability $1 - e^{-t}$

$$\sup_{\theta \in \varOmega(\mathtt{r})} \left\| D^{-1} \{ \nabla \zeta^\flat(\theta) - \nabla \zeta^\flat(\widehat{\theta}) \} \right\| \leq \mathfrak{z}^\flat(t) \mathtt{r}$$

where

$$\mathfrak{z}^\flat(t) = \mathfrak{z}(t, \sqrt{n} |g''| \delta_\psi, \delta_\psi^2)$$

3.7 Lasso model

Consider a model with l_1 penalty

$$\widehat{\theta}_\lambda = \operatorname*{argmax}_{\theta\in\Theta} L(\theta) - \lambda \left\|\theta\right\|_{1,P} \tag{L1}$$

Where $\left\|\cdot\right\|_{1,P} \overset{\text{def}}{=} \left\|\theta_P\right\|_1 = \sum_{i\in P} |\theta_i|$ and P denotes the set of penalized components of the parameter vector θ. Define the true value of parameter of interest as

$$\theta^* = \operatorname*{argmax}_{\theta\in\Theta} I\!\!E L(\theta) \tag{EL}$$

Also define an active indices set as $S = \{i|\theta^*_i \neq 0\}$ and its complement as $C = \{i|\theta^*_i = 0\}$. Denote the power of the active set as $p_* = |S|$. Another problem which is important for our argument is

$$\theta^*_\lambda = \operatorname*{argmax}_{\theta_c=0\ \&\theta\in\Theta} I\!\!E L(\theta) - \lambda \left\|\theta\right\|_1 \tag{EL1}$$

In this paper we employ primal-dual witness approach. First, note that under assumptions of convexity of $L(\cdot)$ and existence of solution of the problem (L1) it is characterized by

$$\nabla L(\theta) - \lambda Z = 0 \tag{gL1}$$

where $Z \in \partial \left\|\theta\right\|_{1,P}$. Consider yet another model in which penalty components include linear combinations of θ

$$\widehat{\theta}_\lambda = \operatorname*{argmax}_{\theta} L(\theta) - \lambda \left\|A\theta\right\|_{1,P} \tag{L1P}$$

which is equivalent to

$$\max_{\eta=A\theta} L(A^\dagger\eta) - \lambda \left\|\eta\right\|_{1,P}, \quad A^\dagger A = I$$

Require that θ be a sub-vector of η that allows to find $\widehat{\theta}_\lambda$ from the solution of the last task. Stationarity condition for this task is

$$\nabla_\eta L(A^\dagger\eta) = \lambda\nabla_\eta \left\|\eta\right\|_{1,P}$$

where

$$\nabla_\eta L(A^\dagger\eta) = A^{\dagger T}\nabla L(\theta)$$

Gradients in subspaces: $\{(A\theta^*)_c = 0\}$ and $\{(A\theta^*)_s \neq 0\}$, which will be useful further, have forms

$$\nabla^2_{cs} L(A^+\eta) = A_c^{\dagger T}\nabla^2 L(\theta)A_s^\dagger, \quad \nabla^2_{ss} L(A^+\eta) = A_s^{\dagger T}\nabla^2 L(\theta)A_s^\dagger$$

where

$$A_c^\dagger = (A^\dagger)_{\bullet c}, \quad A_s^\dagger = (A^\dagger)_{\bullet s}$$

We construct a primal-dual witness solution $(\widetilde{\theta}_\lambda, \widetilde{Z})$ as follows. Define $\widetilde{\theta}_\lambda$ as a solution of the following optimization problem:

$$\widetilde{\theta}_\lambda = \underset{\theta \in \varTheta \& \theta_c = 0}{\operatorname{argmax}} \; L(\theta) - \lambda \left\| \theta \right\|_{1,P}$$

Next we choose Z_s to be an element of $\partial \left\| (\widetilde{\theta}_\lambda)_s \right\|_{1,P}$ and Z_c are chosen in the following manner:

$$Z_c = \frac{1}{\lambda} \nabla_c L(\widetilde{\theta}_\lambda)$$

Note that it does not ensure the feasibility of Z_c. The next lemma provides sufficient conditions for strict dual feasibility to hold which imply the equality $\widetilde{\theta}_\lambda = \widehat{\theta}_\lambda$. Let

$$\exists \alpha \in (0,1] \;\; \texttt{s.t.} \quad \max_{e \in c} \left\| D^2_{es} (D^2_{ss})^\dagger \right\|_1 \leq 1 - \alpha \tag{$\mathcal{A}$}$$

where $D^2 \overset{\text{def}}{=} -\nabla^2 I\!\!E L(\theta^*)$.

Lemma 3.7. Assume that L is convex and the problem (L1) has a unique solution. Also let $(\mathcal{A})$ hold for some positive α. Furthermore, suppose the residual of approximation of the gradient at $\widetilde{\theta}_\lambda$ with its first-order Taylor expansion

$$R(\theta) \overset{\text{def}}{=} \nabla I\!\!E L(\theta) + D^2 (\theta - \theta^*)_s$$

is bounded as well as stochastic component of the gradient

$$\max \left\{ \left\| R(\widetilde{\theta}_\lambda) \right\|_\infty, \left\| \nabla \zeta(\widetilde{\theta}_\lambda) \right\|_\infty \right\} \leq \frac{\alpha \lambda}{8}$$

And finally, suppose that all the components of the parameter vector which does not belong to the active set S are penalized: $C \subset P$. Then

$$\left\| \widetilde{Z}_c \right\|_\infty < 1$$

and therefore $\widetilde{\theta}_\lambda = \widehat{\theta}_\lambda$.

Proof. Since the problem (L1) is convex and has a unique solution, the solution is characterized by the gradient condition (gL1). Replacing the gradient with its first-order Taylor expansion yields

$$\nabla \zeta(\widetilde{\theta}_\lambda) - D^2 (\widetilde{\theta}_\lambda - \theta^*) + R(\widetilde{\theta}_\lambda) - \lambda \widetilde{Z} = 0$$

Denote $\varDelta = (\widetilde{\theta}_\lambda - \theta^*)$ such that $\varDelta_c = 0$ by construction. Use short notation $\nabla \zeta = \nabla \zeta(\widetilde{\theta}_\lambda)$ and do the same for R. Rewrite this equation for active and inactive sets separately:

$$\nabla_s \zeta - D^2_{ss} \varDelta_s + R_s - \lambda \widetilde{Z}_s = 0$$

$$\nabla_c \zeta - D^2_{cs} \varDelta_s + R_c - \lambda \widetilde{Z}_c = 0$$

Now we can solve the first equation for $\varDelta_S$ and substitute it to the second one:

$$\varDelta_s = (D^2_{ss})^{\dagger} (R_s - \lambda Z_s + \nabla_s \zeta)$$

$$\nabla_c \zeta - D^2_{cs} (D^2_{ss})^{\dagger} (R_s - \lambda \widetilde{Z}_s + \nabla_s \zeta) + R_c - \lambda \widetilde{Z}_c = 0$$

Observe that due to $(\mathcal{A})$ $|||D^2_{cs}(D^2_{ss})^{\dagger}|||_1 \leq 1 - \alpha$. Since $C \in P$, the region of strict dual feasibility for $\widetilde{Z}_c$ is just an ∞-ball: $\|\widetilde{Z}_c\|_\infty \leq 1$. Finally we show that the latter bound holds:

$$\begin{aligned} \left\|\widetilde{Z}_c\right\|_\infty &= \left\| \frac{1}{\lambda} (\nabla_c \zeta - D^2_{cs} (D^2_{ss})^{\dagger} (R_s - \lambda Z_s + \nabla_s \zeta) + R_c) \right\|_\infty \\ &\leq \frac{1}{\lambda} \|R + \nabla \zeta\|_\infty + \frac{1}{\lambda} (1 - \alpha) \|R + \nabla \zeta\|_\infty + 1 - \alpha \\ &\leq \frac{2}{\lambda} \|R + \nabla \zeta\|_\infty + 1 - \alpha \\ &\leq \frac{2}{\lambda} \left(\frac{\alpha \lambda}{8} + \frac{\alpha \lambda}{8} \right) + 1 - \alpha \\ &\leq 1 - \frac{\alpha}{2} \\ &< 1 \end{aligned}$$

as claimed. □

Define a local region with center θ^* and radius $\mathtt{r}_0$

$$\varOmega_s(\mathtt{r}_0) = \left\{ \theta : \left\| D^2_{ss} (\theta_s - \theta^*_s) \right\|_\infty \leq \mathtt{r}_0; \;\; \theta_c = \theta^*_c = 0 \right\}$$

Assume exponential moment restriction for process $\nabla \zeta(\theta^*)$:

Assumption 0L1: For all $\gamma \in I\!\!R^p$ takes place

$$\log I\!\!E \exp \left\{ \gamma^T \nabla \zeta(\theta^*) \right\} \leq \frac{\nu_0^2 \|\gamma\|^2}{2}$$

This property yields with probability $1 - e^{-t}$

$$\|\nabla \zeta(\theta^*)\|_\infty \leq \nu_0 \sqrt{2 t_p}, \quad t_p = t + \log(p)$$

Define two properties related to the second derivative of $I\!\!E L(\theta)$ deviations. For that define following matrices

$$D^2_{sc}(\theta) = -(\nabla_s \otimes \nabla_c) I\!\!E L(\theta), \quad D^2_{ss}(\theta) = -\nabla^2_s I\!\!E L(\theta).$$

Assumption 1L1: For all $\gamma: \|\gamma\|_\infty \leq 1$ in the local region $\varOmega_s(\mathtt{r}_0)$

$$\left\|(D^2_{cs}(\theta) - D^2_{cs})D^{-2}_{ss}\gamma\right\|_\infty \leq \delta_{sc}(\mathtt{r}_0)$$

$$\left\|(D^2_{ss}(\theta)D^{-2}_{ss} - I)\gamma\right\|_\infty \leq \delta_s(\mathtt{r}_0)$$

Relying on this condition of can trivially bound the R term:

$$\|R_s\|_\infty \leq \delta_s(\mathtt{r}_0)\mathtt{r}_0$$

$$\|R_c\|_\infty \leq \delta_{cs}(\mathtt{r}_0)\mathtt{r}_0$$

And therefore

$$\|R\|_\infty \leq \max\{\delta_{cs}(\mathtt{r}_0), \delta_s(\mathtt{r}_0)\}\mathtt{r}_0$$

Assumption 2L1: In the local region $\varOmega_s(\mathtt{r}_0)$ the following statements hold with probability at least $1 - e^{-t}$

$$\|\nabla_c\zeta(\theta) - \nabla_c\zeta(\theta^*)\|_\infty \leq \mathfrak{z}_1(t, p_s, \log p)$$

$$\|\nabla_s\zeta(\theta) - \nabla_s\zeta(\theta^*)\|_\infty \leq \mathfrak{z}_1(t, p_s, \log p_s)$$

Assumption 3L1: The Likelihood function is convex and has unique solution.

$$-\nabla^2 L(\theta) \geq 0$$

Lemma 3.8. Under **Assumptions 1L1, 3L1** and additional assumption $\delta_s(2\lambda) \leq 1/2$ it holds

$$\left\|D^2_{ss}(\theta^* - \theta^*_\lambda)_s\right\|_\infty \leq 2\lambda$$

Proof. Let $r \overset{\text{def}}{=} 2\lambda$. Also denote the difference of solutions of the problems (EL) and (EL1) as $\varDelta^* = \theta^*_\lambda - \theta^*$. Consider a continuous function

$$F(D^2_{ss}\varDelta_s) = \nabla_s\mathbb{E}\mathrm{L}(\theta^* + \varDelta) - \lambda Z_s + D^2_{ss}\varDelta_s$$

where $Z_s \in \partial_s \|\theta^*_\lambda\|_1$ and $\varDelta_c = 0$.

Observe that $\varDelta_s$ is a fixed point iff. $\nabla\mathbb{E}\mathrm{L}(\theta^*+\varDelta)-\lambda Z_s = 0$ which means that $\theta^*+\varDelta = \theta^*_\lambda$ or equivalently $\varDelta^* = \varDelta$. Now consider a ball

$$B = \{D^2_{ss}\varDelta_s : \left\|D^2_{ss}\varDelta_s\right\|_\infty \leq r\}$$

Next we show that $F(B) \subseteq B$. Really, replacing the gradient in (F) with its first-order Taylor expansion at point θ^* yields

$$F(D^2_{ss}\varDelta_s) = \nabla_s\mathbb{E}\mathrm{L}(\theta^*) + \nabla^2_{ss}\mathbb{E}\mathrm{L}(\theta^0)\varDelta_s - \lambda Z_s + D^2_{ss}\varDelta$$

where θ^0 is a point on the line connecting θ^* and $\theta^* + \varDelta$. But due to the definition of θ^* and assumptions imposed on the problem (EL), $\nabla_s \mathbb{E}\mathrm{L}(\theta^*) = 0$. Therefore,

$$F(D^2_{ss}\varDelta_s) = \nabla^2_{ss}\mathbb{E}\mathrm{L}(\theta^0)\varDelta_s - \lambda Z_s + D^2_{ss}\varDelta_s$$

Further re-arrangements give

$$\begin{aligned} F(D^2_{ss}\varDelta_s) &= -\lambda Z_s - D^2_{ss}(\theta^0) D^{-2}_{ss} D^2_{ss}\varDelta_s + D^2_{ss}\varDelta_s \\ &= -\lambda Z_s + (-D^2_{ss}(\theta^0) D^{-2}_{ss} + I) D^2_{ss}\varDelta_s \end{aligned}$$

Now using the fact that $\|Z_s\|_\infty \leq 1$ and Assumption 1 we finally obtain

$$\begin{aligned} \left\| F(D^2_{ss}\varDelta_s) \right\|_\infty &\leq \lambda + \delta_s(r) r \\ &\leq \frac{1}{2} r + \frac{1}{2} r \\ &\leq r \end{aligned}$$

At this point we have a continuous function which maps a closed ball on itself. Therefore by Brouwer's Theorem, this function has a fixed point $\varDelta$, but, by construction of the function $\varDelta = \varDelta^*$, and by construction of the ball $\|D^2_{ss}\varDelta^*_s\|_\infty \leq r = 2\lambda$.

□

Lemma 3.9. Under **Assumptions 0-3 L1** and two additional conditions for some t

$$\frac{3}{2}\mathfrak{z}_1(t, p_s, \log p_s) + \delta_s(6\lambda) \leq \frac{1}{4}$$

and

$$\nu_0 \sqrt{2 t_p} \leq \lambda$$

with probability at least $1 - 4e^{-t}$

$$\left\| D^2_{ss}(\widetilde{\theta}_\lambda - \theta^*_\lambda)_s \right\|_\infty \leq 4\lambda$$

Proof. By Lemma 3.8 we have $\|D^2_{ss}(\theta^* - \theta^*_\lambda)_s\|_\infty \leq 2\lambda$. Define $\varDelta^0 \overset{\text{def}}{=} (\theta^* - \theta^*_\lambda)$ and $\varDelta \overset{\text{def}}{=} \widetilde{\theta}_\lambda - \theta^*_\lambda$. Now consider a continuous function

$$F(D^2_{ss}\varDelta_s) = \nabla_s L(\theta^*_\lambda + \varDelta) - \lambda Z_s + D^2_{ss}\varDelta_s$$

Observe that $D^2_{ss}\varDelta_s$ is a fixed point iff. $\nabla_s L(\theta^*_\lambda + \varDelta) - \lambda Z = 0$ which means that $\theta^*_\lambda + \varDelta = \widetilde{\theta}_\lambda$ or equivalently $\varDelta^* = \varDelta$. Now consider a ball $B = \{D^2_{ss}\varDelta_s : \|D^2_{ss}\varDelta_s\|_\infty \leq 4\lambda\}$. Next we show that $F(B) \subseteq B$. Really, decomposing the gradient in (F) into deterministic and stochastic components with subsequent replacement of the gradient of the deterministic one with its first-order Taylor expansion at point θ^* yields

$$F(D^2_{ss}\varDelta_s) = \nabla_s \mathbb{E}\mathrm{L}(\theta^*_\lambda) - D^2_{ss}(\theta^0)\varDelta_s - \lambda Z_s + D^2_{ss}\varDelta_s + \nabla_s\zeta(\theta^*_\lambda + \varDelta)$$

$$-D^2_{ss}(\theta^0)\varDelta_s + D^2_{ss}\varDelta_s = (-D^2_{ss}(\theta^0)D^{-2}_{ss} + I)D^2_{ss}\varDelta_s \le \delta_s\left(\left\|D^2_{ss}(\varDelta^0_s + \varDelta_s)\right\|_\infty\right)4\lambda$$

$$\|\nabla_s\mathbb{E}\mathrm{L}(\theta^*_\lambda)\|_\infty = \|\lambda\,\partial\,\|\theta^*_\lambda\|_1\|_\infty \le \lambda$$

Now we employ the the assumptions 0-3 L1 along with the fact that $\|Z\|_\infty \le 1$:

$$\left\|F(D^2_{ss}\varDelta_s)\right\|_\infty \le \left\{\frac{3}{2}\mathfrak{z}_1(t, p_s, \log p_s) + \delta_s(6\lambda)\right\}4\lambda + \lambda + \lambda + \nu_0\sqrt{2t_p}$$

And finally we use the rest of assumptions of the lemma:

$$\left\|F(D^2_{ss}\varDelta_s)\right\|_\infty \le 4\lambda$$

as claimed.

□

Theorem 3.9. Suppose **Assumptions 0-3 L1** of Lemma 3.9 and 3.8 hold. Moreover, let $\mathtt{r}_0 \overset{\mathrm{def}}{=} 6\lambda$ and $(\mathcal{A})$ hold with parameter α. Also assume that the parameters belonging to the inactive set are penalized: $C \subset P$. Finally, suppose λ is large enough:

$$\max\{\delta_{sc}(\mathtt{r}_0), \delta_s(\mathtt{r}_0), \mathfrak{z}_1(t, p_s, \log p)\} \le \frac{\alpha}{48}$$

$$\lambda > \frac{8}{\alpha}\nu_0\sqrt{2t_p}$$

Then $\widehat{\theta}_\lambda = \widetilde{\theta}_\lambda$ with probability at least $1 - 5e^{-t}$ and therefore

$$\left\|D^2_{ss}(\widehat{\theta}_\lambda - \theta^*)_s\right\|_\infty \le \mathtt{r}_0 = 6\lambda$$

and $\left(\widehat{\theta}_\lambda\right)_c = 0$

Proof. Lemmas 3.9 and 3.8 provide with probability at least $1 - 4e^{-t}$ that

$$\left\|D^2_{ss}(\widetilde{\theta}_\lambda - \theta^*)_s\right\|_\infty \le 6\lambda$$

Next, using Lemma Assumptions 0L1, 1L1 and 2L1 one obtains with probability at least $1 - e^{-t}$

$$\|R\|_\infty \le \max\{\delta_{sc}(\mathtt{r}_0), \delta_s(\mathtt{r}_0)\}\mathtt{r}_0 \le \frac{\alpha\lambda}{8}$$

and

$$\|\nabla\zeta(\widetilde{\theta}_\lambda)\| \le 6\lambda\mathfrak{z}_1(t, p_s, \log p) + \nu_0\sqrt{2t_p}$$

Therefore, Lemma 3.7 applies here which means that $\widetilde{\theta}_\lambda = \widehat{\theta}_\lambda$ and therefore

$$\left\| D_{ss}^2(\widehat{\theta}_\lambda - \theta^*)_s \right\|_\infty \leq 6\lambda$$

and $\left(\widehat{\theta}_\lambda\right)_c = 0$ by construction of $\widetilde{\theta}_\lambda$.

□

Now we are interested in sign selection consistency.

Consequence. Suppose, the assumptions of Theorem 3.9 hold. Moreover, assume the lower lower bound for the minimal absolute value of non-zero element of θ^*:

$$\theta_{\min} \overset{\text{def}}{=} \min_{j \in S} \left| \theta_j^* \right| > \mathtt{r}_0 |||D_{ss}^{-2}|||_1$$

Then

$$I\!\!P \left\{ \forall i : \operatorname{sign}(\widehat{\theta}_\lambda)_i = \operatorname{sign}\theta_i^* \right\} \geq 1 - 5e^{-t}$$

Proof. Defining $\eta \overset{\text{def}}{=} D_{ss}^2(\widehat{\theta}_\lambda - \theta^*)_s$ one obtains

$$\left\| (\widehat{\theta}_\lambda - \theta^*)_s \right\|_\infty = \left\| D_{ss}^{-2} \eta \right\|_\infty$$

And from Theorem 3.9 with probability $1 - 5e^{-t}$ we have $\|\eta\|_\infty \leq \mathtt{r}_0$. Therefore,

$$\left\| (\widehat{\theta}_\lambda - \theta^*)_s \right\|_\infty \leq \mathtt{r}_0 |||D_{ss}^{-2}|||_1$$

And finally making use of the lower bound $\theta_{\min}$ we obtain the statement claimed.

□

Theorem 3.10. Suppose the assumptions of Theorem 3.9. Then with probability at least $1 - 5e^{-t}$

$$\left\| D_{ss}^2(\widehat{\theta}_\lambda - \theta^*)_s - \nabla_s L_\lambda(\theta^*) \right\|_\infty \leq \{\mathfrak{z}_1(t, p_s, \log p_s) + \delta_s(\mathtt{r}_0)\} \mathtt{r}_0$$

Proof. Theorem 3.9 along with its corollary provides us with $\widehat{\theta}_\lambda = \widetilde{\theta}_\lambda$ and $\forall i : \operatorname{sign}(\widehat{\theta}_\lambda)_i = \operatorname{sign}\theta_i^*$ with probability at least $1 - 5e^{-t}$. The latter means that the function $\mathbb{E}\mathrm{L}_\lambda$ is differentiable at the points $\widehat{\theta}_\lambda$ and θ^* and due to the definition of the active set S, so the vector ξ does exist. Moreover, the function is differentiable at any point on the line connecting these points. Now we just write down the first-order Taylor expansion of the function $\mathbb{E}\mathrm{L}_\lambda$ at point $\widehat{\theta}$:

$$\nabla_s \mathbb{E}\mathrm{L}_\lambda(\widehat{\theta}_\lambda) = \nabla_s \mathbb{E}\mathrm{L}_\lambda(\theta^*) - D_{ss}^2(\widehat{\theta}_\lambda - \theta^*)_s + r$$

where r is the remainder term. Next we make use of the fact that $\nabla_s I\!\!E L(\theta^*) = 0$

$$\nabla_s \mathbb{E}\mathrm{L}(\widehat{\theta}_\lambda) + \lambda \nabla_s \left\| \widehat{\theta}_\lambda \right\|_1 = \lambda \nabla_s \left\| \widehat{\theta}_\lambda \right\|_1 - D_{ss}^2(\widehat{\theta}_\lambda - \theta^*)_s + r$$

But $\lambda \nabla_s \left\| \widehat{\theta}_\lambda \right\|_1 = \lambda \nabla_s \left\| \widehat{\theta}_\lambda \right\|_1$ due to sign consistency:

$$\nabla_s \mathbb{E}\mathrm{L}(\widehat{\theta}_\lambda) = -D^2_{ss}(\widehat{\theta}_\lambda - \theta^*)_s + r$$

Now recalling the definition of R and using the equality $\widehat{\theta}_\lambda = \widetilde{\theta}_\lambda$ we obtain $r = R_s$. Thus

$$\nabla_s \mathbb{E}\mathrm{L}_\lambda(\widehat{\theta}_\lambda) = \nabla_s \mathbb{E}\mathrm{L}_\lambda(\theta^*) - D^2_{ss}(\widehat{\theta}_\lambda - \theta^*)_s + R_s$$

Now by the definition of the stochastic component (ζ) one gets

$$\nabla_s L(\widehat{\theta}_\lambda) + \zeta(\widehat{\theta}) - \zeta(\theta^*) = \nabla_s L_\lambda(\theta^*) - D^2_{ss}(\widehat{\theta}_\lambda - \theta^*)_s + R_s$$

But $\nabla_s L(\widehat{\theta}_\lambda) = 0$:

$$\nabla_s \zeta(\widehat{\theta}_\lambda) - \nabla_s \zeta(\theta^*) - R_s = \nabla_s L_\lambda(\theta^*) - D^2_{ss}(\widehat{\theta}_\lambda - \theta^*)_s$$

Now we can bound the right-hand side:

$$\begin{aligned} \left\| \nabla_s L_\lambda(\theta^*) - D_{ss}(\widehat{\theta}_\lambda - \theta^*)_s \right\|_\infty &= \left\| \zeta(\widehat{\theta}_\lambda) - \zeta(\theta^* - R_s \right\|_\infty \\ &\leq \| R_s \|_\infty + \left\| \nabla_s \zeta(\widehat{\theta}_\lambda) - \nabla_s \zeta(\theta^*) \right\|_\infty \\ &\leq \mathfrak{z}_1(t, p_s, \log p_s)\mathtt{r}_0 + \delta_s(\mathtt{r}_0)\mathtt{r}_0 \end{aligned}$$

□

4 Change point detection

4.1 Introduction and related work

The problem of change point detection appears each time one needs to explore a set of random data and make a decision about homogeneity of its structure. In other words, the problem can be stated as two following questions: were there any structural changes in the nature of observed data? At which moments, if so? The present work mainly focuses on the *sequential* or *online* change point detection. In this case the data is aggregated from running random process. Formally a time moment τ is a *change point*, if stochastic properties of the observed signal $\{Y_t\}_{t=1}^n$ have undergone changes in its distribution:

$$\begin{cases} Y_t \backsim I\!\!P_1 & t < \tau, \\ Y_t \backsim I\!\!P_2 & t \geq \tau. \end{cases}$$

The goal is to find such structural breaks as soon as possible. Such problem arises across many scientific areas: quality control Lai (1995), cybersecurity Blazek and Kim (2001), Wang et al. (2004), econometrics Spokoiny (2009), Mikosch and Starica (2004), geodesy e.t.c. Article Shiryaev (1963) describes classical results in change point detection theory. Overview of the state-of-art methods are presented in Polunchenko and Tartakovsky (2011) and Shiryaev (2010).

This research considers sequential hypothesis testing, in which each hypothesis ($I\!\!P_1 = I\!\!P_2$) monitors the presence of change point through Likelihood Ratio Test (LRT) using *sliding window*. At each time step the procedure extracts a data slice, splits it in two parts of equal size and executes LRT on it. High values of LRT indicate possible distribution difference in the window parts ($I\!\!P_1 \neq I\!\!P_2$). Procedures with LRT are rather popular in related literature. The work Quandt (1960) proposes application of LRT for detection of breaks in linear regression model. It was further developed by many authors, e.g. Haccou et al. (1987), Srivastava and Worsley (1986). Papers Liu et al. (2008), Zou et al. (2007) investigate LRT for change point detection for nonparametric case. Nonparametric approaches are easily adaptable for complex data but in general they need more information for model building than their parametric alternatives. Introduction of *parametric assumption*: $I\!\!P_1, I\!\!P_2 \in \{I\!\!P(\theta) : \theta \in I\!\!R^p\}$ allows to reduce the suffisient number of observations as soon as $I\!\!P(\theta)$ has less degrees of freedom than nontapametric model. The state-of-the-art review of parametric models based on LRT and its application to economics and bio-informatics are presented by Chen and Gupta (2012). The paper Gombay (2000) explores how LRT can be used for sequential change point detection in case $I\!\!P(\theta)$ is exponential family.

The LRT statistic requires its quantiles or critical values to be set from the signal data $\{Y_t\}_{t=1}^n$. Many works are dedicated to asymptotic behaviour of LRT, e.g. Jandhyala and Fotopoulos (1999) obtains lower and upper bounds for distribution of asymptotic maximum likelihood estimator. The work Kim (1994) provides a very detailed study of its asymptotic behaviour in linear regression models. Similar results for change in mean of a Gaussian process are given in Fotopoulos et al. (2010). In Biau et al. (2016) an approach with Wiener process and Donsker–Prohorov Theorem describes relatively general method for LRT-like statistics distribution approximation.

Instead of asymptotic distribution for LRT one may find a benefit of resampling and *bootstrap*. This technique is popular, e.g. Frick et al. (2014), Spokoiny (2009), since it provides a way to simulate a complex distribution of LRT statistic (for wide family of $I\!\!P(\theta)$) through empirical data distribution. Using bootstrap one can generate LRT$^\flat$ statistic multiple times in order to obtain quantile distribution of the initial LRT. Both LRT and LRT$^\flat$ statistics have approximation by the following norms with high probability

$$\mathrm{LRT} \approx \|\boldsymbol{\xi} + \varDelta\|, \quad \mathrm{LRT}^\flat \approx \|\boldsymbol{\xi}^\flat + \varDelta^\flat\| \tag{Qf}$$

Bigger $\varDelta$ values correspond to more confident hypothesis rejection (more apparent changes in the data sequence). Argument $\boldsymbol{\xi}$ could be treated as a noise component. For LRT critical value calibration one requires data without change points and consequently with $\varDelta = 0$. We also describe below a modified LRT which enables the calibration even if data contains change points.

The cornerstone of this novel change point detection procedure is the concept of change-point pattern. The geometry of a pattern depends on a type of transition region between two distributions that the data obeys before and after a change respectively. Three examples are presented at the Fig. 2. The triangle (spades) pattern appears in case of an abrupt transition from $I\!\!P(\theta_1)$ to $I\!\!P(\theta_2)$. A smooth transition between two distributions entails trapezium change-point pattern. And a horn pattern appears due to an abrupt change in variance. Processing of a change-point pattern instead of a single LRT-value allows to reduce noise influence $\xi(t)$ and false-alarm rate. The presence of change-point patterns is the corollary of (Qf) representation.

In case of a single change point one may find the pattern position by maximising convolution with a pattern function $P_\tau(t)$:

$$\operatorname*{argmax}_{\tau} \sum_t P_\tau(t) \|\boldsymbol{\xi}(t) + \varDelta(t)\|$$

In order to set critical value correctly quantiles of the statistic $\max_\tau \sum_t P_\tau(t)\|\boldsymbol{\xi}(t)\|$ should be close in distribution to quantiles of $\max_\tau \sum_t P_\tau(t)\|\boldsymbol{\xi}^\flat(t)\|$.

4.2 Algorithm

Provide the description of the Change Point Detection algorithm which employs Likelihood Ratio Test (LRT). Let $(I\!\!P(\theta),\ \theta \in I\!\!R^p,\ L(\theta) = \log(\partial^n I\!\!P(\theta)/\partial Y))$ be a parametric assumption about the nature of data inside the window $(Y_{t-h}, \dots, Y_{t+h-1})$ with central point t and size $2h$. Here and further we assume, that the observations $\{Y_i\}_{i=1}^n$ are independent, so

$$L(\theta, \mathbb{Y}) = \sum_i l_i(\theta) \tag{L}$$

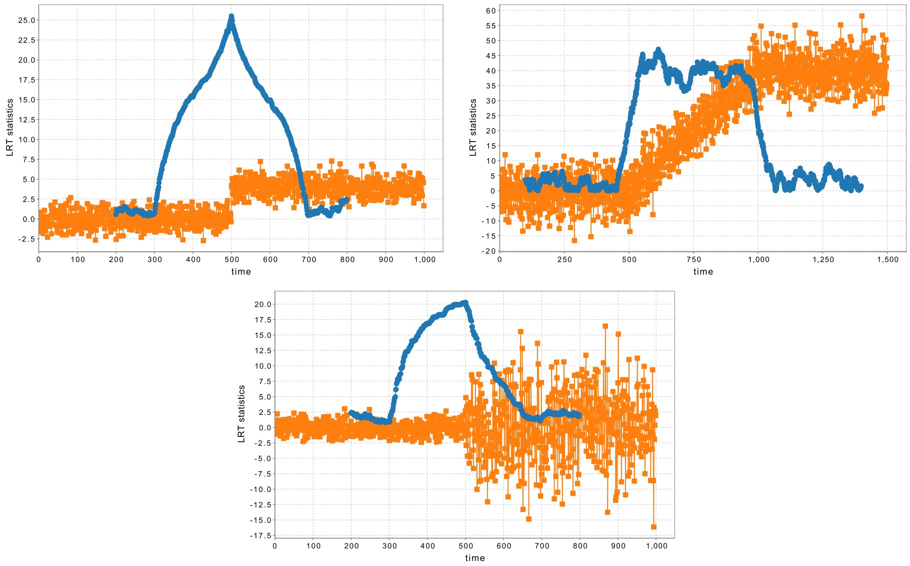


Figure 2. Types of change point and the geometry of change-point patterns: triangle pattern – abrupt mean transition, trapezium pattern – smooth mean transition, horn pattern – abrupt variance transition.

Remind the denotations for argmax of the Likelihood function and the reference model parameter

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta, \mathbb{Y}), \quad \theta^* = \underset{\theta}{\operatorname{argmax}}\, \mathbb{E} L(\theta, \mathbb{Y})$$

The algorithm sequentially computes LRT statistic ($T_h(t)$) for each t in the sliding window procedure. The LRT statistic itself corresponds to the gain from window split into two parts ($\mathbb{Y}_l, \mathbb{Y}_r$):

$$T_h(t) = L(\widehat{\theta}_l, \mathbb{Y}_l) + L(\widehat{\theta}_r; \mathbb{Y}_r) - L(\widehat{\theta}, \mathbb{Y}) \tag{T}$$

$$\mathbb{Y}_l = (Y_{t-h}, \ldots, Y_{t-1}), \quad \mathbb{Y}_r = (Y_t, \ldots, Y_{t+h-1})$$

$$\widehat{\theta}_l = \underset{\theta}{\operatorname{argmax}}\, L(\theta, \mathbb{Y}_l), \quad \widehat{\theta}_r = \underset{\theta}{\operatorname{argmax}}\, L(\theta, \mathbb{Y}_r)$$

According to the Theorem 4.1, encountering change point, statistic $2T_h(t) \approx \|\boldsymbol{\xi}(t) + \varDelta(t)\|^2$ starts growing according to change point pattern type (for example spades, trapezium, horn, ref. the Figure 2). In order to match pattern positions, the procedure monitors $2h$ values of the LRT simultaneously and convolves them with each of the predefined pattern functions $P_\tau(t)$:

$$\mathrm{TP}_h(\tau) = \sum_t P_\tau(t) \sqrt{2T_h(t)} \tag{TP}$$

High values of $\mathrm{TP}_h(\tau)$ correspond to a sufficient correlation of $\sqrt{2T_h}$ and P_τ (similar to the dependence on t). The algorithm marks a time moment τ at a scale h as a change point, if the test statistic $\mathrm{TP}_h(\tau)$ exceeds a calibrated (by bootstrap procedure) critical

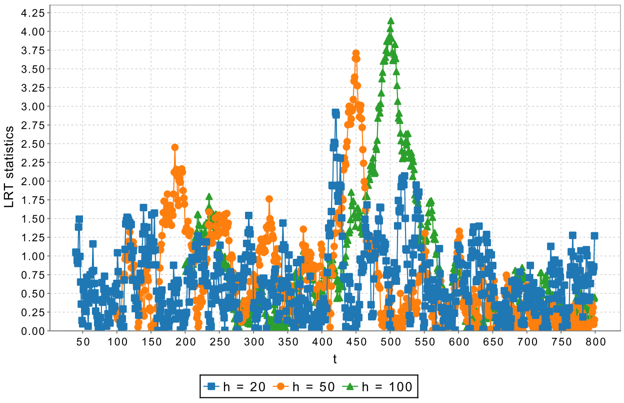


Figure 3. LRT statistic example with different window sizes: $h = 20, 50, 100$.

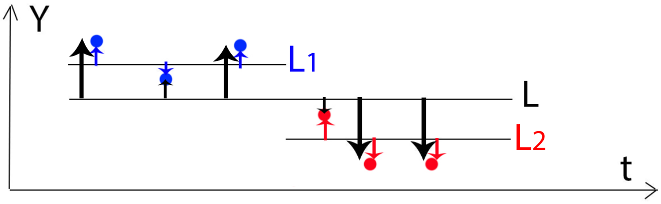


Figure 4. LRT illustration for a simple constant regression model.

value z_h:

$$\{\tau \text{ is a change point }\} \Leftrightarrow \{\exists h : \mathrm{TP}_h(\tau) > z_h\}$$

The greater window size h is chosen, the more probably the algorithm will mark τ as a change point. Again, small windows may mark τ faster.

Weighted bootstrap procedure enables resampling of the statistic $\max_{1 \le \tau \le n} \mathrm{TP}_h(\tau)$ and thus calculation of the critical value z_h for the window size $2h$. It generates a sequence of weighted likelihood functions, where each element is a convolution of independent Likelihood components and weight vector $(w_1^\flat, \dots, w_n^\flat)$:

$$L^\flat(\theta, \mathbb{Y}) = \sum_i w_i^\flat l_i(\theta) \tag{Lb}$$

where $\{w_i^\flat\}_{i=1}^n$ are i.i.d. and $w_i^\flat \in \mathcal{N}(1,1)$. At each weights generation one gets a new value of $L^\flat(\theta)$ and its optimal parameter $\theta^\flat$ and thus bootstrap procedure enables to estimate $L(\widehat{\theta})$ fluctuations. The corresponding bootstrap LRT$^\flat$ statistic is

$$T_h^\flat(t) = L^\flat(\theta_l^\flat, \mathbb{Y}_l) + L^\flat(\theta_r^\flat, \mathbb{Y}_r) - \sup_\theta \{L^\flat(\theta, \mathbb{Y}_l) + L^\flat(\theta + \widehat{\theta}_r - \widehat{\theta}_l, \mathbb{Y}_r)\} \tag{Tb}$$

$$\theta^\flat = \operatorname*{argmax}_\theta L^\flat(\theta, \mathbb{Y})$$

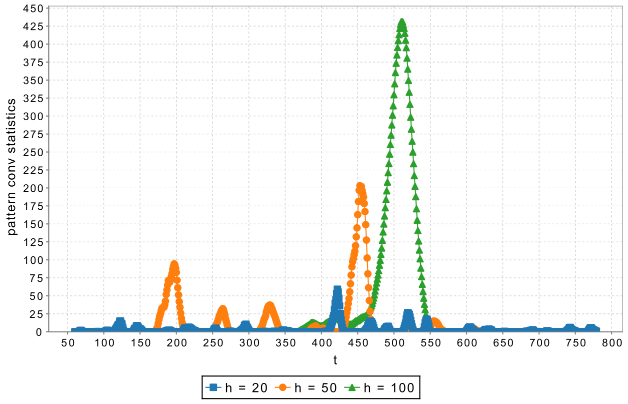


Figure 5. LRT smoothing by triangle pattern $P(t)$ with different window sizes: $h = 20, 50, 100$.

Parameter $(\widehat{\theta}_r - \widehat{\theta}_l)$ is required for condition $T_h^\flat \approx \left\|\xi^\flat\right\|$ (ref. Theorem 4.2). In this case one can estimate $\max_{1 \le \tau \le n} \mathrm{TP}_h^\flat(\tau)$ quantiles under the null hypothesis $(\Delta^\flat(t) \propto \widehat{\theta}_r(t) - \widehat{\theta}_l(t))$ instead of the false assumption $(\Delta^\flat(t) = 0)$.

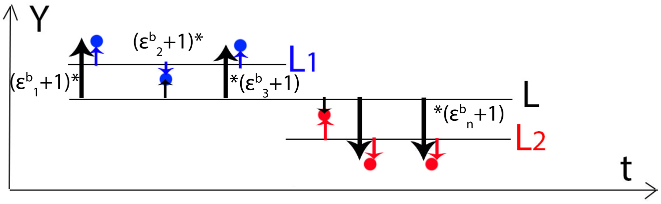

Y
$(\varepsilon^b{}_2+1)^*$
L1
$(\varepsilon^b{}_1+1)^*$
$^*(\varepsilon^b{}_3+1)$
L
$^*(\varepsilon^b{}_n+1)$
L2
t

Figure 6. Bootstrap LRT illustration for a simple constant regression model.

Empirical bootstrap version generates subsamples of data $\{Y_k\}$ from the complete dataset with random independent indexes of size n. In this case

$$L^\epsilon(\theta, \mathbb{Y}) = \sum_i l_{k(i)}(\theta) \tag{Le}$$

where $\{k(i)\}_{i=1}^n$ are i.i.d. and $k(i) \in \{1, \dots, n\}$. For all window positions $\widehat{\theta}_r = \widehat{\theta}_l = \widehat{\theta}$ and here bias correction is not required. So the corresponding LRT$^\epsilon$ statistic is like (T):

$$T_h^\epsilon(t) = L^\epsilon(\theta_l^\epsilon, \mathbb{Y}_l) + L^\epsilon(\theta_r^\epsilon, \mathbb{Y}_r) - L^\epsilon(\theta^\epsilon, \mathbb{Y}) \tag{Te}$$

$$\theta^\epsilon = \operatorname*{argmax}_\theta L^\epsilon(\theta, \mathbb{Y})$$

Empirical bootstrap works better in applications with independent models but less suitable for models with block-independent dataset and theoretical investigations (the distribution is discontinuous) .

Algorithms 1, 2 summarises above ideas for sequential case and the case with pretraining data. Designation $(t_1 : t_2)$ is a range of natural values $t_1, t_1 + 1, \dots, t_2$.

Algorithm 3 presents the procedure of calculation of a critical value z_h for window size $2h$ and for one window position.

$$\widehat{\theta}_{12} = \widehat{\theta}_r - \widehat{\theta}_l$$

We use *multiplicity correction* for multiple hypothesis testing: $H_h : \max_\tau \mathrm{TP}^\flat_h(\tau) < z(h)$ for each h. Let $z(h, \alpha)$ be α quantile of variable $\max_\tau \mathrm{TP}^\flat_h(\tau)$. The probability that at least one hypothesis is false equals to

$$I\!P(\{\exists h : \max_\tau \mathrm{TP}^\flat_h(\tau) - z(h, \alpha)) > 0\}) = I\!P(\{\exists h : \text{p-value}(\max_\tau \mathrm{TP}^\flat_h(\tau)) < \alpha\}) \geq \alpha$$

One may decrease above probability by confidence reduction:

$$I\!P(\{\exists h : \text{p-value}(\max_\tau \mathrm{TP}^\flat_h(\tau)) < \alpha - \alpha'\}) = \alpha$$

4.3 Implementation and experiments

In order to substantiate patterns utility we compare procedure from this Section with the similar one but without pattern (i.e. $P_\tau(t) = 1\!\!I[\tau = t]$). The experiment scenario is following. The dataset $\{Y_i\}$ consists of 500 normal random vectors from $I\!R^5$ with one change point at position $\tau^* = 250$.

$$Y_i \in \mathcal{N}(0, I_5), \quad 0 \leq i < 250$$

$$Y_i \in \mathcal{N}(0.25, I_5), \quad 250 \leq i < 500$$

The procedure searches for the change point location as $\widehat{\tau} = \operatorname{argmax}_\tau \mathrm{TP}_h(\tau)$. Then the quality of the detection is measured by average error $|\widehat{\tau} - \tau^*|$ (c.p. position error) and fraction of the detected change points (power) (ref. Figure 7).

The second experiment describes bootstrap convergence depending on window size ($2h$). We set bootstrap confidence level equal to 0.1 and compute p-value from real distribution with bootstrap quantile $z^\flat$.

$$I\!P^\flat\left(\max_{1 \leq \tau \leq n} \mathrm{TP}^\flat_h(\tau) > z^\flat\right) = 0.1$$

$$\left| I\!P\left(\max_{1 \leq \tau \leq n} \mathrm{TP}_h(\tau) > z^\flat\right) - 0.1 \right| = O\left(\frac{1}{h^\beta}\right)$$

From the plot below (ref. Figure 8) one can observe that

$$\beta > \frac{1}{2}$$

which suppose better convergence in comparison with the theoretical study (ref. Theorem 4.3), where $\beta = 1/6$.

```
$Q_h(t) = 0$ – change point signals;
$H$ – window sizes set;
get z(h) by Algorithm 3;
foreach window position $t$ do
    foreach $h$ do
        add $T_h(t)$ to $\mathbb{T}_h$;
        $\mathrm{TP}_h = \langle \mathbb{T}_h(t-h), P_h \rangle$;
        if $TP_h > z(h)$ and
          $Q_{(1:h)}(t-2h:t) = 0$ then
            $Q_h(t) = 1$;
        end
    end
    if $\max_h Q_h(t) = 1$ then
        $t$ is change point;
    end
end
```

Algorithm 1: LRTOnline

```
$S$ – change points set;
$H$ – window sizes set;
function FindCP($Y_1, \dots, Y_M$):
get z(h) by Algorithm 3;
foreach $h$ do
    foreach window position $t$ do
        compute $T_h(t)$;
    end
    foreach $\tau$ do
        $\mathrm{TP}_h(\tau) = \langle \mathbb{T}_h(\tau), P_h \rangle$;
    end
end
$\tau = \mathrm{argmax}_\tau \sum_{h \in H} \mathrm{TP}_h(\tau)$;
if $\exists h : \mathbb{T}_h(\tau) > z(h)$ then
    add $\tau$ to S;
    FindCP($Y_1, \dots, Y_\tau$);
    FindCP($Y_\tau, \dots, Y_M$);
end
```

Algorithm 2: LRTOffline

```
Data: $(Y_1, \dots, Y_M)$, $h$, $P_h$,
$S$ – weights generation count
Result: $f^\flat_h$ – bootstrap distribution of
        maximal convolution across
        the dataset
for $s = 1$ to $S$ do
    generate $w^\flat = (w^\flat_1, \dots, w^\flat_n)$;
    foreach window position $t$ do
        compute $T^\flat_h(t)$;
    end
    foreach $\tau$ do
        $\mathrm{TP}^\flat_h(\tau) = \langle \mathbb{T}^\flat_h(\tau), P_h \rangle$;
    end
    add $\max_\tau \mathrm{TP}^\flat_h(\tau)$ to $f^\flat_h$;
end
Data: $H = (h_1, \dots, h_N)$, $f^\flat_h$,
$\alpha$ – confidence value
Result: critical values $z(h)$
Multiplicity correction:
for $s = 1$ to $S$ do
    generate $w^\flat = (w^\flat_1, \dots, w^\flat_n)$;
    add $\min_h$ p-value($\max_\tau \mathrm{TP}^\flat_h(\tau), f^\flat_h$)
      to empirical distribution $I\!P_f$
end
find $\alpha'$ from condition
  $I\!P_f(\min_h \text{p-value}(\cdot) < \alpha - \alpha') = \alpha$;
foreach $h$ in $H$ do
    $z(h) = \mathrm{quantile}(f^\flat_h, \alpha - \alpha')$;
end
```

Algorithm 3: Critical values calibration

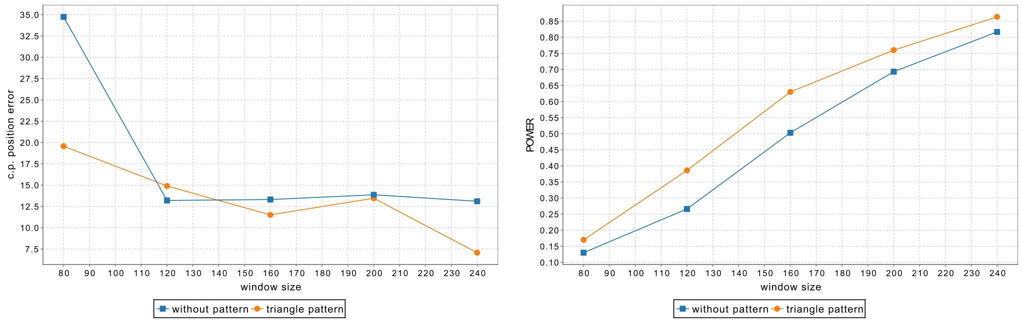


Figure 7. Change point localisation test and power test for the proposed algorithm. One case with triangle pattern and the other case without pattern.

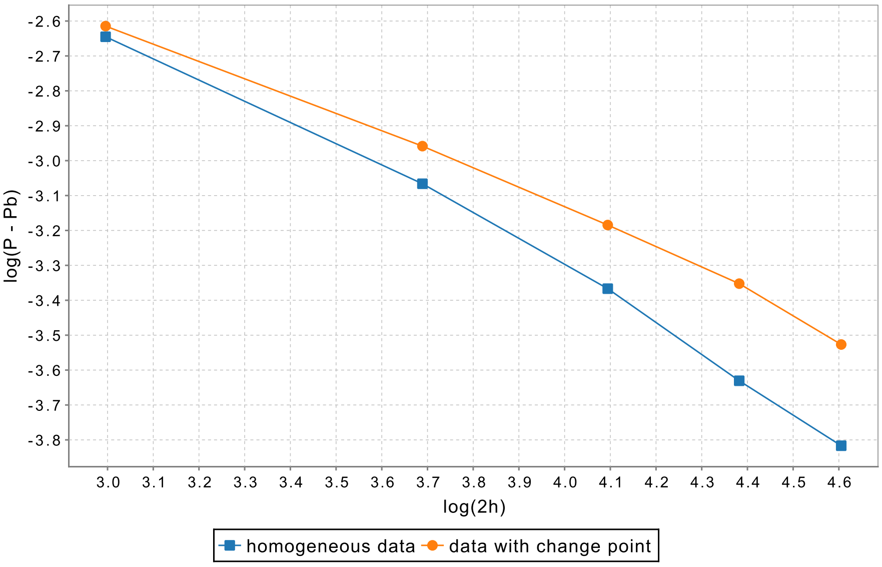


Figure 8. Bootstrap convergence. Homogeneous data: $Y_i \in \mathcal{N}(0, I_p), 0 \leq i < 6h$, data with change point: $Y_i \in \mathcal{N}(0, I_p), 0 \leq i < 3h, \quad Y_i \in \mathcal{N}(0.3, I_p), 3h \leq i < 6h$. The parameters are $p = 30$, $h \in \{10, 20, 30, 40, 50\}$.

The last experimental part presents results of the comparison of the proposed algorithm of change point detection (LRTOffline) with two other methods: Bayesian online change point detection (BOCPD) from Adams and MacKay (2007) and (RMeanVar) from R package (cpt.meanvar(PELT, ...)). The first method is constructed for online inference, but so far as it returns CP location with each CP signal, it is also applicable for offline testing scenario. The idea of this method is predictive filtering: its forecasts a new data point using only the information have been observed already, where the distribution family is fixed (Normal for the tests in this paper). Bayesian inference calculates the length of the observed data (from the last CP). The second algorithm also uses preliminary specified model. Its design focuses into finding multiple changes in mean and variance in Normally (another distributions also supported) distributed data. The returned set of change points is the result of sequential testing H_0 (existing number of change points) against H_1 (one extra change point) applying the likelihood ratio statistic of the whole data coupled with the penalty for CP count. RMeanVar performs better than well known method CUSUM due to synchronous changes in both data parameters mean and variance.

Quality of measurements uses Normalised Mutual Information (NMI). The next equation defines NMI measure of two partitions (X, Y) of time range by change points

$$\mathrm{NMI}(X, Y) = 2\frac{H(X) + H(Y) - H(X, Y)}{H(X) + H(Y)}$$

$H(X)$ and $H(X, Y)$ and entropy functions. Higher NMI values (they are in $[0, 1]$) correspond to better quality.

Synthetic test data have been generated with different values of the distribution parameter transition (Δ). Each Δ value corresponds to 10 sampled data sequences over which one compute measure average. Each data sequence has two, one or none change points. The data has two distributions: normal ($\mathcal{N}(\theta(1), \theta(2))$) and Poisson ($\mathrm{Po}(\theta)$). Parametric assumption for all methods is $\mathcal{N}(\theta(1), \theta(2))$, so Poisson data corresponds to misspecification scenario.

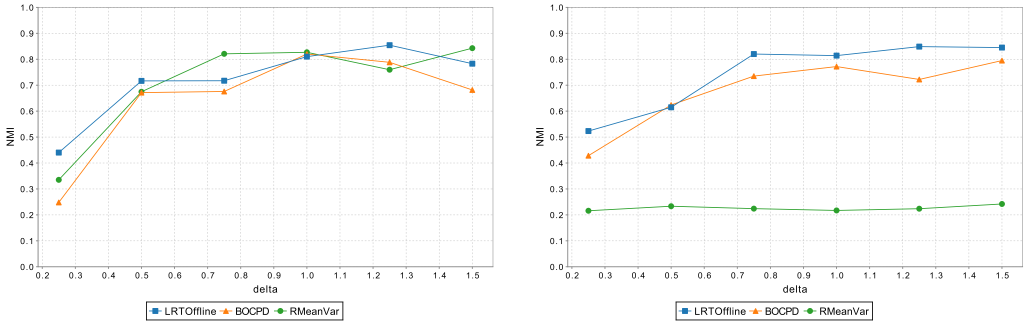


Figure 9. First data: $\mathcal{N}(\theta(1), \theta(2))$, second data: $\mathrm{Po}(\theta)$, data size $= 340$, parametric assumption for all methods is $\mathcal{N}(\theta(1), \theta(2))$, NMI – Normalized Mutual Information between predicted and reference partitions of time interval with change points, change points count per test $\{0, 1, 2\}$.

In the tests with normal data all the methods achieves similar NMI scores. In the tests with Poisson data (misspecification) RMeanVar has relatively low quality and LRTOffline outperforms slightly BOCPD method. One may find Scala implementation of our change point detection algorithm (LRTOffline) by link

`https://github.com/nazarblch/cpd`

4.4 Quadratic LRT approximation

Further consider a fixed window position t and window size $2h$. We are going to derive an explicit dependence between statistic $T_h(t)$ and parameter difference from left and right part of the window $(\theta^*_r - \theta^*_l)$. Approximation of $T_h(t)$ by its quadratic form splits noise and deterministic parts, such that $2T_h(t) \approx \|D(\theta^*_r - \theta^*_l) + \boldsymbol{\xi}_{lr}\|^2$. In the fixed window position the likelihood function has view

$$L(\theta) = L_l(\theta) + L_r(\theta) = L(\theta, \mathbb{Y}_l) + L(\theta, \mathbb{Y}_r)$$

$$D_k^2 = -\nabla^2 \mathbb{E} L_k(\theta^*_k), \quad \boldsymbol{\xi}_k = D_k^{-1} L_k(\theta^*_k), \quad i = \{l, r\}$$

From Assumptions 1-3 for function $L(\theta)$ it holds with probability $1 - e^{-t}$ (ref. Theorem 3.4 with $T_h(t) = L(\theta, \theta_0)$) that

$$\left| \sqrt{2T_h} - \left\| \begin{matrix} D_l(\widehat{\theta} - \widehat{\theta}_l) \\ D_r(\widehat{\theta} - \widehat{\theta}_r) \end{matrix} \right\| \right| \leq 2\diamondsuit(\sqrt{2}\mathtt{r}_0, t)$$

Find relation between $\widehat{\theta}, \widehat{\theta}_l, \widehat{\theta}_r$ using Theorem 3.1 with notation $\boldsymbol{\xi}_k(\theta) = D_k^{-1}\nabla L_k(\theta)$

$$\left\| D(\widehat{\theta} - \theta) \right\| \leq \left\| D^{-1}\{D_l \boldsymbol{\xi}_l(\theta) + D_r \boldsymbol{\xi}_r(\theta)\} \right\| + 2\diamondsuit(\mathtt{r}_0, t)$$

$$\left\| \begin{matrix} D^{-1} D_l\{\boldsymbol{\xi}_l(\theta) - D_l(\theta - \widehat{\theta}_l)\} \\ D^{-1} D_r\{\boldsymbol{\xi}_r(\theta) - D_r(\theta - \widehat{\theta}_r)\} \end{matrix} \right\| \leq 2\diamondsuit(\sqrt{2}\mathtt{r}_0, t)$$

Define vector $\widetilde{\theta}$ that is close to $\widehat{\theta}$

$$\widetilde{\theta} = \underset{\theta}{\operatorname{argmin}} \left\{ \left\| D_l(\theta - \widehat{\theta}_l) \right\|^2 + \left\| D_r(\theta - \widehat{\theta}_r) \right\|^2 \right\}$$

$$\widetilde{\theta} = (D_l^2 + D_r^2)^{-1}(D_l^2 \widehat{\theta}_l + D_r^2 \widehat{\theta}_r)$$

$$\left| \left\| \begin{matrix} D_l(\widehat{\theta} - \widehat{\theta}_l) \\ D_r(\widehat{\theta} - \widehat{\theta}_r) \end{matrix} \right\| - \left\| \begin{matrix} D_l(\widetilde{\theta} - \widehat{\theta}_l) \\ D_r(\widetilde{\theta} - \widehat{\theta}_r) \end{matrix} \right\| \right| \leq \left\| D(\widehat{\theta} - \widetilde{\theta}) \right\| \leq 2\diamondsuit(\mathtt{r}_0, t) + 2\diamondsuit(\sqrt{2}\mathtt{r}_0, t)$$

$$\left\| \begin{matrix} D_l(\widetilde{\theta} - \widehat{\theta}_l) \\ D_r(\widetilde{\theta} - \widehat{\theta}_r) \end{matrix} \right\| = \left\| D_{lr}(\widehat{\theta}_r - \widehat{\theta}_l) \right\|, \quad D_{lr} = D_l D^{-1} D_r$$

An intermediate result is (with probability $1 - 3e^{-t}$)

$$\left| \sqrt{2T_h} - \left\| D_{lr}(\widehat{\theta}_r - \widehat{\theta}_l) \right\| \right| \leq 4\diamondsuit(\sqrt{2}\mathtt{r}_0, t) + 2\diamondsuit(\mathtt{r}_0, t)$$

Involve $\boldsymbol{\xi}_l$ and $\boldsymbol{\xi}_r$ by means of Fisher expansion (Theorem 3.1) for the model with two independent components

$$\left\| \begin{matrix} D_r D^{-1}\{D_l(\widehat{\theta}_l - \theta^*_l) - \boldsymbol{\xi}_l\} \\ D_l D^{-1}\{D_r(\widehat{\theta}_r - \theta^*_r) - \boldsymbol{\xi}_r\} \end{matrix} \right\| \leq \diamondsuit(\sqrt{2}\mathtt{r}_0, t)$$

The final result is Theorem 4.1, which enables to describe T_h function depending on change point type and subsequently choose appropriate pattern $P_h(t)$ (ref. the Algorithm Section).

Theorem 4.1. Assume conditions from Theorem 3.1 for models L_l and L_r. Then with probability $1 - 4e^{-\mathtt{x}}$ for each t

$$\left| \sqrt{2T_h(t)} - \|D_{lr}(\theta_r^* - \theta_l^*)(t) + \boldsymbol{\xi}_{lr}(t)\| \right| \le 7\diamondsuit(\sqrt{2}\mathtt{r}, \mathtt{x})$$

where

$$\boldsymbol{\xi}_{lr}(t) = D_{lr}\{D_l^{-2}\nabla L(\theta_l^*, \mathbb{Y}_l) + D_r^{-2}\nabla L(\theta_r^*, \mathbb{Y}_r)\}, \quad D_{lr} = D_l D^{-1} D_r$$

Theorem 3.4 enables to prove statement similar to Theorem 4.1 for the bootstrap LRT statistic $T_h^\flat$. The proof steps are the same as in Theorem 4.1.

Theorem 4.2 (Weighted bootstrap LRT)**.** Assume conditions from Theorem 3.4 for models $L_l^\flat$ and $L_r^\flat$. Then with probability $1 - 4e^{-\mathtt{x}}$ for each window position t

$$\left| \sqrt{2T_h^\flat(t)} - \left\| D_{lr}(\widehat{\theta}_r - \widehat{\theta}_l)(t) + \boldsymbol{\xi}_{lr}^\flat(t) \right\| \right| \le 7\diamondsuit^\flat(\sqrt{2}\mathtt{r}_0, \mathtt{x})$$

where

$$\boldsymbol{\xi}_{lr}^\flat(t) = D_{lr}\{D_l^{-2}\nabla L^\flat(\widehat{\theta}_l, \mathbb{Y}_l) + D_r^{-2}\nabla L^\flat(\widehat{\theta}_r, \mathbb{Y}_r)\}$$

4.5 Bootstrap consistency

Below we present the Theorems that describes difference between probabilistic measures of $\mathrm{TP}_h(\tau)$ and $\mathrm{TP}_h^\flat(\tau)$ (precision of the bootstrap calibration) and LRT sensitivity to parameter θ^* transition at change point. In independent models each noise vector $\boldsymbol{\xi}_{lr}(t) = \boldsymbol{\xi}(t) \in I\!R^p$ is a sum of independent vectors (ref. Section 4.4 for $\boldsymbol{\xi}_{lr}(t)$ definition)

$$\boldsymbol{\xi}_{lr}(t) = \sum_{i=t-h}^{t-1} \boldsymbol{\xi}_i - \sum_{i=t}^{t+h-1} \boldsymbol{\xi}_i, \quad \boldsymbol{\xi}_i \propto \nabla l_i(\theta^*)$$

Aggregate all $\boldsymbol{\xi}_i$ into one vector

$$\boldsymbol{\xi}^T = (\boldsymbol{\xi}_1^T, \ldots, \boldsymbol{\xi}_n^T)$$

Theorem 4.3 (Buzun N. (2017))**.** Let dataset size be n, the window size $2h$, the model dimension – p, pattern functions $P_\tau(t)$ be independent from $\|\boldsymbol{\xi}_{lr}(t)\|$ and normalized $\sum_t |P_\tau(t)| = 1$. Include conditions from Theorems 4.1 and 4.2. Then for each fixed z with high probability

$$\left| I\!P\left(\max_{1 \le \tau \le n} \mathrm{TP}_h(\tau) > z \right) - I\!P^\flat\left(\max_{1 \le \tau \le n} \mathrm{TP}_h^\flat(\tau) > z \right) \right| \le \varDelta_{PT}$$

$$\varDelta_{PT} = C_1 \mu_3^{1/3} + C_2 \| \operatorname{Var}(\boldsymbol{\xi}_{lr}) - \operatorname{Var}^\flat(\boldsymbol{\xi}_{lr}^\flat) \|_\infty^{1/2} + 7C_A\{\diamondsuit(\mathtt{r}_0, \mathtt{x}) + \diamondsuit^\flat(\mathtt{r}_0^\flat, \mathtt{x})\}$$

where

$$\| \operatorname{Var}(\boldsymbol{\xi}_{lr}) - \operatorname{Var}^\flat(\boldsymbol{\xi}^\flat_{lr})\|_\infty \le 10\sqrt{\log(np)}\sqrt{2h}\| \operatorname{Var}(\boldsymbol{\xi})\|_\infty(3 + \|b\|) + \|b\|^2$$

$$\|b\|^2 = \max_t \sum_{i=t}^{t+2h} \|I\!\!E\boldsymbol{\xi}_i\|^2_\infty$$

$$\mu_3 \le 2hI\!\!E\,\|\boldsymbol{\xi}\|^3_\infty$$

The rest parameters are $C_1 = 5C_\mu^{1/3}C_A$, $C_2 = 4C_\Sigma^{1/2}C_A$.

Remark. Parameter $\varDelta_{PT}$ has asymptotic

$$\varDelta_{PT} \sim \frac{\sqrt{p}\log^2(n)}{h^{1/6}} + \frac{p^{3/2}\log^{1/2} n}{h^{1/2}}$$

since $C_\mu \sim \log^2(n)$, $C_\Sigma \sim \log(n)$, $C_A \sim \sqrt{p\log(n)}$ and

$$\mu_3^{1/3} \sim \frac{\log^{1/2}(n)}{(2h)^{1/6}}, \quad \| \operatorname{Var}(\boldsymbol{\xi}_{lr}) - \operatorname{Var}^\flat(\boldsymbol{\xi}^\flat_{lr})\|^{1/2}_\infty \sim \frac{\log^{1/4}(n)}{(2h)^{1/4}}, \quad \diamondsuit + \diamondsuit^\flat \sim \frac{p}{h^{1/2}}$$

Remark. For quantile estimation of the statistic $\max_{1\le\tau\le n} \mathrm{TP}_h(\tau)$ with quantile of $\max_{1\le\tau\le n} \mathrm{TP}^\flat_h(\tau)$ one has to show that

$$\left| I\!\!P\left(\max_{1\le\tau\le n} \mathrm{TP}_h(\tau) > z^\flat(\alpha)\right) - \alpha\right| \le \varDelta_{PT}$$

for $z^\flat(\alpha)$ defined by equation

$$I\!\!P^\flat\left(\max_{1\le\tau\le n} \mathrm{TP}^\flat_h(\tau) > z^\flat(\alpha)\right) = \alpha$$

This statement is a consequence of the Theorem (4.3) but not a direct one since the argument $z^\flat(\alpha)$ is random and depends on $\max_{1\le\tau\le n} \mathrm{TP}_h(\tau)$. Involving sandwich Lemma 3.5 fulfills this issue.

Proof. Describe the bootstrap approximation for the quadratic form of the statistic $\mathrm{TP}_h(\tau)$ on the grounds of Theorems 4.1 and 4.2. The quadratic form of $\mathrm{TP}_h(\tau)$ is

$$\max_{1\le\tau\le n}\left\{\sum_t P_\tau(t)\|\boldsymbol{\xi}_{lr}(t)\|\right\} \tag{maxTP}$$

The corresponded bootstrap quadratic form is

$$\max_{1\le\tau\le n}\left\{\sum_t P_\tau(t)\|\boldsymbol{\xi}^\flat_{lr}(t)\|\right\}$$

Our aim is to show that these two forms are close by distribution. For simplification assume that for all window positions the true model parameter is fixed $\theta_l^* = \theta_r^* = \theta^*$. Then $\boldsymbol{\xi}_{lr}(t)$ doesn't depend on parameter changes and

$$\boldsymbol{\xi}_{lr}(t) = \sum_{i=t}^{t+h} \boldsymbol{\xi}_i - \sum_{i=t+h}^{t+2h} \boldsymbol{\xi}_i$$

where $\boldsymbol{\xi}_i = D^{-1}\nabla l_i(\theta^*)$ and $D^{-1} = D_{lr}D_l^{-2} = D_{lr}D_r^{-2}$. Use smooth-max approximation for the composite maximum function with argument $\boldsymbol{\xi}_{lr} = A\boldsymbol{\xi}$.

$$\frac{1}{\sqrt{p}} \max_{\tau} \left\{ \sum_t P_\tau(t) \|\boldsymbol{\xi}_{lr}(t)\| \right\} \approx h_\beta(\varphi(A\boldsymbol{\xi}))$$

where

$$\varphi_\tau(A\boldsymbol{\xi}) = \sum_t P_\tau(t) \frac{\|W_t A\boldsymbol{\xi}\|}{\sqrt{p}}, \quad W_t = \operatorname{diag}(0, \dots, 1_{tp}, \dots, 1_{tp+p}, \dots, 0)$$

One should estimate the distribution difference from replacement of the random argument in statistic (maxTP): $\boldsymbol{\xi} \to \widetilde{\boldsymbol{\xi}} \to \boldsymbol{\xi}^\flat$. Note that $\widetilde{\boldsymbol{\xi}} \in \mathcal{N}(I\!E\boldsymbol{\xi}, \operatorname{Var}(\boldsymbol{\xi}))$ and $\boldsymbol{\xi}^\flat \in \mathcal{N}(0, \operatorname{diag}(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T))$. Taking into account $\sum_t |P_\tau(t)| = 1$ estimate φ's derivatives required for Theorems 2.4 and 2.5.

$$\|\nabla\varphi_\tau(x)\|_1 = \sum_t |P_\tau(t)| \frac{\|W_t x\|_1}{\sqrt{p}\|W_t x\|} \le 1$$

$$\|\nabla^2\varphi_\tau(x)\|_1 = \sum_t |P_\tau(t)| \frac{\|W_t - xW_t x^T/\|W_t x\|^2\|_1}{\sqrt{p}\|W_t x\|} \le \frac{2\sqrt{p}}{\|W_t x\|}$$

$$\|\nabla^3\varphi_\tau(x)\|_1 \le \sum_t 3|P_\tau(t)| \frac{\|W_t \otimes W_t x\|_1}{\sqrt{p}\|W_t x\|^3} + \sum_t 3|P_\tau(t)| \frac{\|xW_t x^T \otimes W_t x\|_1}{\sqrt{p}\|W_t x\|^5} \le \frac{6p}{\|W_t x\|^2}$$

Then the constant C_φ from Theorem 2.4 has bound

$$C_\varphi \le I\!E \frac{\sqrt{6p}\triangle}{\|W_t A\boldsymbol{\xi}\|} \le 1$$

and therefore with $T = n$ one get

$$C_\mu = 6\left(5 + 6\log n + \log^2 n\right)$$

$$C_\Sigma = 2\,(2 + \log n)$$

Finally the bootstrap approximation error for statistic (maxTP) is

$$\begin{aligned}
&\left| I\!P\left(\max_{1\le\tau\le n} \sum_t P_\tau(t)\|\boldsymbol{\xi}_{lr}(t)\| > z \right) - I\!P^\flat\left(\max_{1\le\tau\le n} \sum_t P_\tau(t)\|\boldsymbol{\xi}^\flat_{lr}(t)\| > z \right) \right| \\
&\le \left| I\!P\left(\max_{1\le\tau\le n} \sum_t P_\tau(t)\|\boldsymbol{\xi}_{lr}(t)\| > z \right) - I\!P^\flat\left(\max_{1\le\tau\le n} \sum_t P_\tau(t)\|\widetilde{\boldsymbol{\xi}}_{lr}(t)\| > z \right) \right| \\
&+ \left| I\!P\left(\max_{1\le\tau\le n} \sum_t P_\tau(t)\|\widetilde{\boldsymbol{\xi}}_{lr}(t)\| > z \right) - I\!P^\flat\left(\max_{1\le\tau\le n} \sum_t P_\tau(t)\|\boldsymbol{\xi}^\flat_{lr}(t)\| > z \right) \right| \\
&\le 5C_\mu^{1/3} C_A \mu_3^{1/3} + 4C_\Sigma^{1/2} C_A \|\Sigma_{A\xi} - \Sigma^\flat_{A\xi}\|_\infty^{1/2}
\end{aligned}$$

We have to consider further $\|\Sigma_{A\xi} - \Sigma^\flat_{A\xi}\|_\infty^{1/2}$ and C_A.

$$\|\Sigma_{A\xi} - \Sigma^\flat_{A\xi}\|_\infty = \max_{a^T, b^T \in \mathrm{rows} A} |a^T(\Sigma_\xi - \Sigma^\flat_\xi)b|$$

Let for a fixed rows a, b with probability $1 - e^{-t}$

$$|a^T(\Sigma - \Sigma^\flat)b| = \left| \sum_{ij} a_i b_j \Sigma_\xi(i,j) - \sum_{ij} a_i b_j \xi_i \xi_j \right| \le \square(t)$$

Note that elements in sum $(a_i b_i \xi_i \xi_j)$ are independent due to the specific block structure of matrix A. Then the joint bound with probability $1 - e^{-t}$ is

$$\|\Sigma_{A\xi} - \Sigma^\flat_{A\xi}\|_\infty \le \square(t + 2\log(np))$$

Involve the upper bound for covariance matrix deviations (ErrVD) with $\varepsilon_i = \xi_i$ and $\mathcal{U}_i = a_i/V$

$$\square(t) = V^2 \left(\frac{2}{3} R_{\varepsilon\varepsilon} \mathrm{x} + 2\mathrm{v}_{\varepsilon\varepsilon}\sqrt{5\mathrm{x}} + \delta^2 \|b\|^2 \right)$$

where $\mathrm{v}_{\varepsilon\varepsilon} = \delta\sqrt{\|\Sigma\|_\infty}(3 + \|b\|)$,

$$V\delta = \|A\|_\infty, \quad V^2 = \sum_{ij} a_i b_j \Sigma_{ij} \le |\!|\!|A|\!|\!|^2 \|\Sigma_\xi\|_\infty, \quad \|b\|^2 = \sum_{i:\, a_i > 0} I\!E\xi_i^2$$

Finally under assumption $\frac{2}{3} R_{\varepsilon\varepsilon} t < \mathrm{v}_{\varepsilon\varepsilon}\sqrt{5t}$ with probability $1 - 1/n$

$$\begin{aligned}
\|\Sigma_{A\xi} - \Sigma^\flat_{A\xi}\|_\infty &\le 10\sqrt{\log(np)} |\!|\!|A|\!|\!| \|A\|_\infty \|\Sigma\|_\infty (3 + \|b\|) + \|A\|_\infty^2 \|b\|^2 \\
&= 10\sqrt{\log(np)}\sqrt{2h} \|\Sigma_\xi\|_\infty (3 + \|b\|) + \|b\|^2
\end{aligned}$$

For parameter C_A from Lemma 2.5 one has to estimate a_p, σ_1 and σ_2.

$$a_p \le \frac{2}{\sqrt{p}} I\!E \max_{\tau} \left\{ \sum_t P_\tau(t) \| W_t A \widetilde{\boldsymbol{\xi}} \| \right\}$$

Since $\widetilde{\boldsymbol{\xi}} \in \mathcal{N}(0, I\!E\xi\xi^T)$ we get

$$a_p \sim \sqrt{\log n} \frac{2}{\sqrt{p}} \max_{\tau} \sum_t P_\tau(t) I\!E \| W_t A \widetilde{\boldsymbol{\xi}} \| \sim \sqrt{\log n}$$

As for σ_1 and σ_2 with $\|\gamma_t\|, \|\gamma_s\| = 1$

$$\begin{aligned} p\sigma_2^2 &= \max_{\gamma} \sum_{t,s} P_\tau(t) P_\tau(s) \gamma_t^T W_t A I\!E(\widetilde{\boldsymbol{\xi}}\widetilde{\boldsymbol{\xi}}^T) A^T W_s^T \gamma_s^T \\ &\le \max_t \| W_t A \varSigma_\xi A^T W_t^T \| \sim 1 \end{aligned}$$

and

$$\begin{aligned} p\sigma_1^2 &= \min_{\gamma} \sum_{t,s} P_\tau(t) P_\tau(s) \gamma_t^T W_t A I\!E(\widetilde{\boldsymbol{\xi}}\widetilde{\boldsymbol{\xi}}^T) A^T W_s^T \gamma_s^T \\ &\ge \lambda_{\min}(A \varSigma_\xi A^T) \sim 1 \end{aligned}$$

Subsequently

$$C_A \sim \sqrt{p \log n}$$

□

The next part of this Section evaluates the smallest parameter θ^* transition that is sufficient for change point detection in a fixed position τ and window size $2h$. Let $z_h(\alpha)$ be a quantile of $\sum_t P_\tau(t) \|\boldsymbol{\xi}_{lr}(t)\|$ such that

$$I\!P \left(\sum_t P_\tau(t) \|\boldsymbol{\xi}_{lr}(t)\| > z_h(\alpha) \right) = \alpha$$

The sufficient condition for change point detection in position τ is

$$\sum_{t=\tau-h}^{\tau+h} P_\tau(t) \sqrt{2T_h(t)} > z_h(\alpha)$$

One has to compare $z_h(\alpha)$ with $\sum_t P_\tau(t) \| D(\theta_r^* - \theta_l^*)(t) \|$. Find out the upper bound for

$z_h(\alpha)$. Consider triangle pattern example

$$P_\tau(t) = \begin{cases} 0, & t < \tau - h, \\ (t-\tau)/h + 1/2, & \tau - h \le t \le \tau, \\ (\tau - t)/h + 1/2, & \tau \le t \le \tau + h, \\ 0, & t > \tau + h \end{cases} \tag{P}$$

Theorem 4.4. Let each random vector $\boldsymbol{\xi}_{lr}(t)$ be sub-Gaussian with the second moment $\mathbb{E}\|\boldsymbol{\xi}_{lr}(t)\|^2 = p$. Assume conditions from Theorem 4.1. The sufficient condition for abrupt type change point detection of size $\varDelta$ with probability $1 - 2e^{-t}$ in position τ using triangle pattern (P) is

$$\|D_{lr}(\theta_r^* - \theta_l^*)(\tau)\| = \varDelta > \sqrt{\frac{6}{h}} z(B,t) + 21\diamondsuit(\mathtt{r}_0, \mathtt{x})$$

where matrix D_{lr} and θ_r^*, θ_l^* are defined in Theorem 4.1, $z(B,t)$ defined in Lemma 6.6 such that

$$\operatorname{tr}\{B\} = 2hp, \quad \|B\| \le 2h\|\varSigma_\xi\|$$

Proof. Set $\tau = h$. From sub-Gaussian assumption and property

$$\sum_t P_\tau(t)\|\xi_{lr}(t)\| \le \sum_{t=1}^{2h} P_\tau^2(t) \sqrt{\sum_{t=1}^{2h} \|\xi_{lr}(t)\|^2}$$

follows that (Lemma 6.6) with probability $1 - 2e^{-t}$

$$\sqrt{\sum_{t=1}^{2h} \|\xi_{lr}(t)\|^2} \le z(B,t)$$

where

$$\operatorname{tr}\{B\} = \mathbb{E}\sum_{t=1}^{2h} \|\xi_{lr}(t)\|^2 = 2hp$$

and

$$\lambda(B) \le 2h\|\varSigma_\xi\|$$

The integral sum with pattern (P) gives

$$\sum_{t=1}^{2h} P_\tau^2(t) \approx \frac{1}{6}h$$

Finally with probability $1 - 2e^{-t}$

$$z_h(2e^{-t}) \le \sqrt{\frac{h}{6}} z(B,t)$$

The abrupt type change point statistic without noise component has view

$$\sqrt{2T_h(t)} = (P_\tau(t) + 1/2)\Delta \pm 7\diamondsuit, \quad \tau - h \le t \le \tau + h$$

Involve the sufficient condition for change point discussed above

$$\frac{1}{6}\Delta h - \frac{7}{2}h\diamondsuit > \sqrt{\frac{h}{6}}z(B,t)$$

□

4.6 Covariance matrices

In the current study we are interested in a particular kind of a break – an abrupt transformation in the covariance matrix – which is motivated by applications to finance and neuroimaging. In finance the dynamics of the covariance structure of a high-dimensional process modeling return rates is crucial for a proper asset allocation in a portfolio Şerban et al. (2007); Bauwens et al. (2006); Engle et al. (1990); Mikosch et al. (2009). Analogously, break analysis in covariance structure of data in functional Magnetic Resonance Imaging is particularly important for the research on neural diseases as well as in context of brain development with emphasis on characterization of the re-configuration of the brain during learning Bassett et al. (2010); Sporns (2011); Friston (2011).

We consider the following setup. Let $X_1, ..., X_N \in I\!R^p$ denote a sample of independent zero-mean vectors. In online setting the sample size is not fixed in advance. The goal is to test the hypothesis

$$\mathbb{H}_0 = \{\forall i : \operatorname{Var} X_i = \operatorname{Var} X_{i+1}\}$$

versus the alternative suggesting the existence of a break:

$$\mathbb{H}_1 = \{\exists \tau : \operatorname{Var} X_\tau \ne \operatorname{Var} X_{\tau+1}\}$$

and localize the change-point τ as precisely as possible or (in online setting) to detect a break as soon as possible.

Now we present a formal definition of the test statistic. In order to detect a break we consider a set of window sizes $\mathfrak{N} \subset \mathbb{N}$. Denote the size of the widest window as h_+ and of the narrowest as h_-. Given a sample of length n for each window size $h \in \mathfrak{N}$ define a set of central points $t \in \{h+1, h+2, ..., n-h+1\}$. Next, for all $h \in \mathfrak{N}$ define a set of indices which belong to the window on the left side from the central point t as $\mathcal{I}_h^l(t) = \{t-h, t-h+1, ..., t-1\}$ and correspondingly for the window on the right side define $\mathcal{I}_h^r(t) = \{t, t+1, ..., t+h-1\}$. For each window size $h \in \mathfrak{N}$ and each central point t define a pair of estimators of covariance matrix as

$$\widehat{\Sigma}_h^l(t) = \frac{1}{h}\sum_{i \in \mathcal{I}_h^l(t)} X_i X_i^T, \quad \widehat{\Sigma}_h^r(t) = \frac{1}{h}\sum_{i \in \mathcal{I}_h^r(t)} X_i X_i^T$$

Let some subset of indices $\mathcal{I}_s \subseteq 1..n$ of size s (possibly, $s = n$) be chosen. Define a scaling diagonal matrix

$$S = \operatorname{diag}(\sigma_{1,1}, \sigma_{1,2}...\sigma_{p,p-1}, \sigma_{p,p})$$

where the elements $\sigma_{j,k}$ are standard deviations of corresponding elements of $X_i X_i^T$ averaged over $\mathcal{I}_s$:

$$\sigma_{j,k}^2 = \frac{1}{s} \sum_{i \in \mathcal{I}_s} \operatorname{Var}\,(X_i X_i^T)_{jk}$$

In practice the matrix S is usually unknown, hence we propose to plug-in empirical estimators $\widehat{\sigma}_{j,k}$. For each window size $h \in \mathfrak{N}$ and central point t we define a new test statistic $T_h(t)$

$$T_h(t) = \left\| \sqrt{\frac{h}{2}} S^{-1} \overline{(\widehat{\Sigma}_h^l(t) - \widehat{\Sigma}_h^r(t))} \right\|_\infty$$

Here and below we write $\overline{A}$ for a vector composed of stacked columns of matrix A and use $\|\cdot\|_\infty$ to denote the sup norm. Finally, the family of test statistics $\{T_h\}_{h \in \mathfrak{N}}$ is obtained via maximization over the central points:

$$T_h = \max_t T_h(t)$$

Remark. Generally, one can choose the diagonal matrix S arbitrarily as long as its elements are bounded. The choice does not affect Theorem 4.5. However, we prefer to bring all the elements of the covariance matrices to the same scale first, so the test focuses on a relative change. Ideally, we would like to use the $\sigma_{j,k}^2$, yet due to its unavailability we resort to their empirical estimates, whose consistency can be easily demonstrated based on sub-Gaussian assumption.

Decision rule and bootstrap calibration scheme

Our approach rejects $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if at least one of statistics T_n exceeds a corresponding threshold $x_h^\flat(\alpha)$ or formally if $\exists h \in \mathfrak{N} : T_h > x_h^\flat(\alpha)$. In order to choose thresholds $x_h^\flat(\alpha)$ the following bootstrap scheme is proposed. Define vectors $\widehat{Z}_i$ for $i \in \mathcal{I}_s$ as

$$\widehat{Z}_i = \overline{X_i X_i^T - \frac{1}{s} \sum_{i \in \mathcal{I}_s} X_i X_i^T}$$

Elements $Z_i^\flat$ for $i \in 1..n$ of bootstrap sample are proposed to be drawn with replacement from the set $\bigcup_{i \in \mathcal{I}_s} \{-\widehat{Z}_i, \widehat{Z}_i\}$. Denote the measure which $Z_i^\flat$ are distributed with respect to as $\mathbb{P}^\flat$. By construction $\mathbb{P}^\flat$ is not absolute continuous w.r.t to Lebesgue measure, which is not a problem per se, yet "high jumps" naturally complicate quantile estimation. Bringing in both $\widehat{Z}_i$ and $-\widehat{Z}_i$ reduces the "jumps".

Now we are ready to define a bootstrap counterpart $T_h^\flat(t)$ of $T_h(t)$ for all $h \in \mathfrak{N}$ and t

as

$$T_h^\flat(t) = \left\| \sqrt{\frac{h}{2}} S^{-1} \left(\frac{1}{h} \sum_{i \in \mathcal{I}_h^l(t)} Z_i^\flat - \frac{1}{h} \sum_{i \in \mathcal{I}_h^r(t)} Z_i^\flat \right) \right\|_\infty$$

The counterparts $T_h^\flat$ of T_h for all $h \in \mathfrak{N}$ are naturally defined as

$$T_h^\flat = \max_t T_h^\flat(t)$$

Now for each given $\alpha \in (0,1)$ we can define quantile functions $z_n^\flat(\alpha)$ such that

$$z_n^\flat(\alpha) = \inf \left\{ z : \mathbb{P}^\flat \left\{ T_h^\flat > z \right\} \leq \alpha \right\}$$

Next for a given significance level α we apply multiplicity correction choosing α^* as

$$\alpha^* = \sup \left\{ \alpha : \mathbb{P}^\flat \left\{ \exists h \in \mathfrak{N} : T_h^\flat > z_n^\flat(\alpha) \right\} \leq \alpha \right\}$$

and finally choose thresholds as $x_h^\flat(\alpha) = z_n^\flat(\alpha^*)$.

Remark. In most of the cases one may simply choose $\mathcal{I}_s = 1...n$ but at the same time it seems appealing to use some sub-sample which a priory does not include a break, if such information is available. On the other hand, the bootstrap justification result (Theorem 4.5) benefits from larger set $\mathcal{I}_s$.

In order to localize a change-point we have to assume that $\mathcal{I}_s \subseteq 1..\tau$. Consider the narrowest window detecting a change-point as $\widehat{h}$:

$$\widehat{h} = \min \left\{ h \in \mathfrak{N} : T_h > x_h^\flat(\alpha) \right\}$$

and the central point where this window detects a break for the first time as

$$\widehat{t} = \min \left\{ t : T_{\widehat{h}}(t) > x_{\widehat{h}}^\flat(\alpha) \right\}$$

By construction of the family of the test statistics we conclude (up to the confidence level α) that the change-point τ is localized in the interval

$$\left[\widehat{t} - \widehat{h}; \widehat{t} + \widehat{h} - 1 \right].$$

Clearly, if a non-multiscale version of the approach is employed, i.e. $|\mathfrak{N}| = \{h\}$, $h = \widehat{h}$ and precision of localization (delay of the detection in online setting) equals h.

Discuss the theoretical result demonstrating validity of the proposed bootstrap scheme i.e.

$$I\!P \left(\forall h \in \mathfrak{N} : T_h \leq x_h^\flat(\alpha) \right) \approx 1 - \alpha \tag{1}$$

Our theoretical results require the tails of the underlying distributions to be light. Specifically, we impose sub-Gaussian vector condition.

$$\exists L > 0 : \forall i \in 1..N \sup_{\substack{a \in I\!\!R^p \\ ||a||_2 \le 1}} I\!\!E\exp\left(\left(\frac{a^T X_i}{L}\right)^2\right) \le 2 \tag{sG}$$

Theorem 4.5 (Avanesov and Buzun (2016))**.** Let Assumption (sG) hold and let the dataset $X_1, X_2, ..., X_n$ be i.i.d. Allow the parameters $p, |\mathfrak{N}|, s, h_-, h_+$ grow with n. Further let $n > 2h_+ \ge 2h_-$ and $n > s$ and let the minimal window size h_- and the size s of the set $\mathcal{I}_s$ grow fast enough such that

$$\frac{|\mathfrak{N}| L^4 \log^{19}(pn)}{\min\{h_-, s\}} = o(1)$$

Then

$$\left| I\!\!P\left(\forall h \in \mathfrak{N} : T_h \le x_h^\flat(\alpha)\right) - (1-\alpha)\right| = o_p(1)$$

Proof sketch The proof consists of four straightforward steps.

1. Approximate statistics T_h by norms of a high-dimensional Gaussian vector up to the residual R_B using the high dimensional central limit theorem by Chernozhukov et al. (2017).

2. Similarly, we approximate bootstrap counterparts $T_h^\flat$ of the statistics up to the residual $R_{B^\flat}$.

3. Prove that the covariance matrix of the Gaussian vector used to approximate $T_h^\flat$ in step 2 is concentrated in the ball of radius $\varDelta_Y$ centered at its real-world counterpart involved in step 1 and employ the Gaussian comparison result provided by Chernozhukov et al. (2017) and Chernozhukov et al. (2013b).

4. Finally, obtain the bootstrap validity result combining the results of steps 1-3.

The formal proof of the theorem can be found in paper Avanesov and Buzun (2016) along with the finite-sample-size version of the result.

Proof discussion The proof of the bootstrap validity result mostly relies on the high-dimensional central limit theorems obtained by Chernozhukov et al. (2017). That paper also presents bootstrap justification results, yet does not include a comprehensive bootstrap validity statement. The theoretical treatment is complicated by the randomness of $x_h^\flat(\alpha)$. Indeed, consider Lemma 3.5 which is a straightforward combination of steps 1-3. One cannot trivially obtain result of type (1) substituting $\{x_h^\flat(\alpha)\}_{h\in\mathfrak{N}}$ from T_h due to the randomness of $x_h^\flat(\alpha)$ and dependence between $x_h^\flat(\alpha)$ and T_h. We overcome this by means of so-called “sandwiching” proof technique (see Lemma 3.5), initially used by Spokoiny and Willrich (2015). The authors had to assume normality and low dimensionality of the data. Our result is free of such limitations.

4.7 Non-parametric method

Non-parametric change point detection is encouraged by arrhythmia detection in Electrocardiogram (ECG). ECG is a one dimensional time sequence close to a periodic signal, where each period consists of 3 main parts: P wave, QRS complex and T wave. Arrhythmia corresponds to some significant perturbations of periodicity and may be one of the following types: atrial flutter, atrial fibrillation, supraventricular tachycardia, premature atrial contraction and ventricular rhythms. All this types of arrhythmia has different changes in ECG signals. But we do not distinguish between them and make only a binary classification. The formal problem statement is the following. Let X_t be the quasi-periodic signal with a period T. One has to test the hypotheses

$$\begin{gathered}\mathbb{H}_0 : \{X_t \sim I\!\!P_{f_0(t/T)},\ \forall t \in [0, n]\} \\ \mathbb{H}_1 : \{\exists \tau^* : X_t \sim I\!\!P_{f_0(t/T)}\ ; X_t \sim I\!\!P_{f_1(t)}\ \} \\ t \in [0, \tau^*] \quad \text{and} \quad t \in [\tau^*, n]\end{gathered}$$

In the notation above $I\!\!P$ represents a probability distribution, n is the dataset size, τ^* is the change point time, $f_0(t/T)$ and $f_1(t)$ are the functions parametrizing the distributions.

The major difficulty in the statistical study of the problem (1) is twofold: the dependent data and the lack of a suitable parametric model for an intricate signal, such as ECG. To address these challenges, we propose a new pipeline shown in Figure 10. In the proposed algorithm, we resort to the optimal transport (OT) approach that is capable of building a non-parametric change point statistic to test the hypotheses. We propose to apply the TDA/OT approach not to the original signal, but to a projection of the quasi-periodic function into a closed curves space (the point cloud), allowing both the periodic and the morphologic components of the original signal's waveform to be considered. Eventually, we estimate quantiles of the change point statistic with the bootstrap procedure in order to set a threshold under the null hypotheses assumption. Below we prove a theorem about the convergence of the bootstrap distribution of the statistic to the real distribution, setting a foundation stone for a plethora of possible future works on TDA/OT analysis on periodic signals.

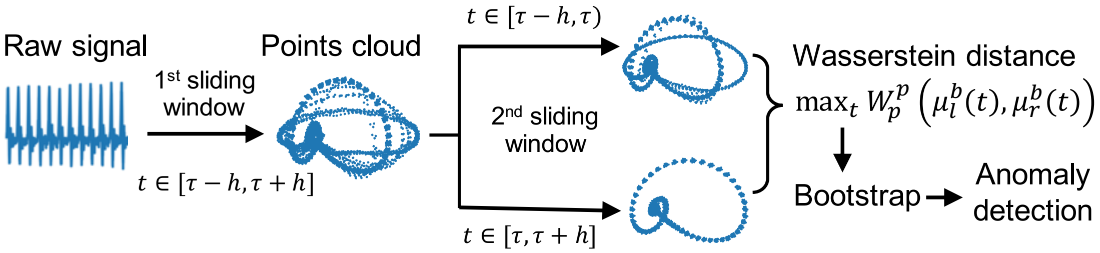


Figure 10. Pipeline of the proposed algorithm, where τ – the second sliding window center, $2h$ – the second sliding window size, W_p – Wasserstein distance, $\mu_l^b(t)$, $\mu_r^b(t)$ – Bootstrap measures in the left and the right parts of the second sliding window.

Algorithm The first step of the algorithm is mapping of the original time series into points cloud. We use method described in paper Perea (2013). Denote

$$SW(t) = \begin{bmatrix} X_t \\ X_{t+s} \\ \dots \\ X_{t+Ms} \end{bmatrix}$$

The integer M determines the window size and real number s is a step parameter. The sliding window makes an embedding of values of X_t into $I\!R^{M+1}$. Iterating by t gives a collection of points called *sliding window point cloud* for X_t. Furthermore we apply PCA to extract the most meaningful dimensions. An example with 3 PCA components for M-dimensional points cloud is presented in Figure 11.

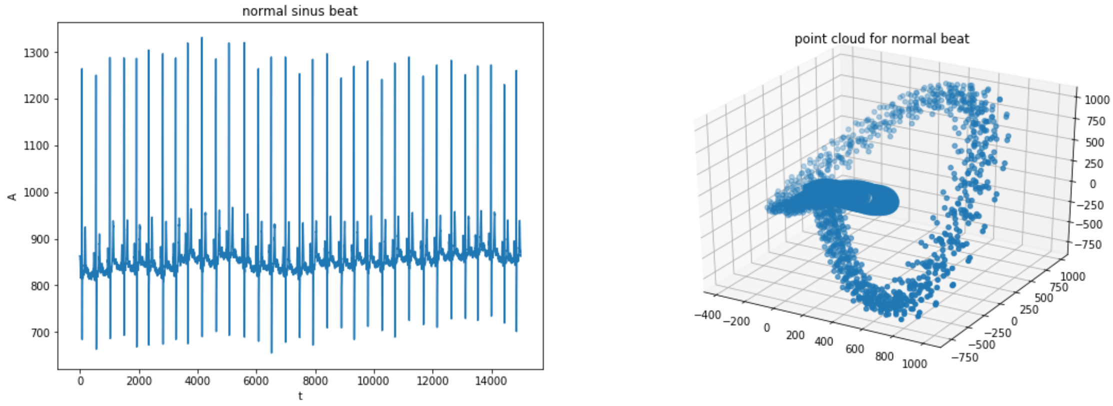


Figure 11. Example of points cloud with 3 PCA components. Left: original time series, right: projection into sliding window point cloud.

In order to find structural changes in point cloud which corresponds to structural changes in original time series we involve the second sliding window that in each step splits the points in two parts of equal size and computes Wasserstein distances on it. The size of the second sliding window equals to several curve loops. Let $X_1, \dots, X_n$ are considered points in cloud. Define the statistic for change point detection for the window size $2h$, position t and power parameter p.

$$T_h(\tau) = W_p^p \left(\sum_{t=\tau-h}^{\tau-1} \delta(X_t), \sum_{t=\tau}^{\tau+h-1} \delta(X_t) \right)$$

The empirical measures in the formula correspond to data in left and right parts of the second sliding window. An example of $T_h(\tau)$ computation for data with multiple change points is presented in Figure 12.

For the critical values calibration in this case one should use rasampling procedure that accounts dependency between data points. We use here moving blocks bootstrap for

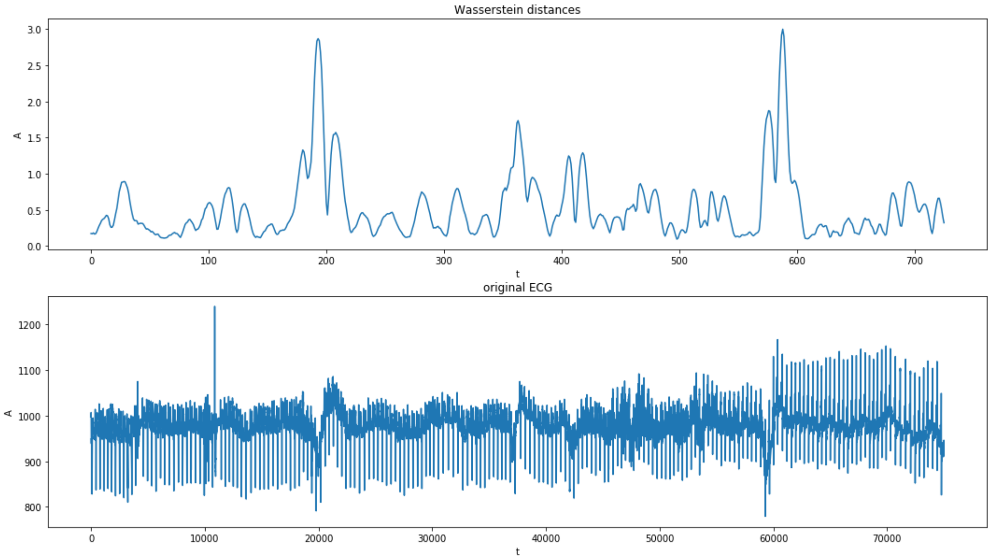


Figure 12. Original time series and corresponded Wasserstein distances between points clouds.

arrhythmia detection. As previously the bootstrap statistic looks similar to $T_h(\tau)$

$$T_h^\flat(\tau) = W_p^p \left(\sum_{t=\tau-h}^{\tau-1} \delta(X_t^\flat), \sum_{t=\tau}^{\tau+h-1} \delta(X_t^\flat) \right)$$

where $\{X_t^\flat\}_{t=1}^n$ are sampled by Moving Block Bootstrap(MBB). MBB was formulated in separate works Kunsch (1989) and Lahiri (2013) as new scheme to create pseudo-samples. The usual bootstrap forms new samples taking only random observations from the initial sample, whereas, the MBB performs this procedure only within a row of the formed blocks. We use a weighted block structure of the MBB, which generates random weights for each block and, importantly, preserves the structure of the original time series.

After the MBB resampling, we create a list of change point statistic values ($\max T_h^\flat = \max_\tau T_h^\flat(\tau)$) and set the threshold with α confidence level corresponding to the border between the normal points and the points of arrhythmia (see Figure 13). It is assumed that quantiles of $\max T_h^\flat$ are close to the quantiles of $\max T_h$ (bootstrap consistency), which we justify theoretically below.

We suppose that points located in peaks of $T_h(\tau)$ plot may correspond to arrhythmia intervals on original ECG. But to be sure we need a critical level which corresponds to some quantile of $\max T_h^\flat$. In every MBB iteration we compute $\max_\tau T_h^\flat(\tau)$. In the result, we have list of the maximum values, and after we take α-quantile we will find the board between normal points and points corresponded to arrhythmia.

Bootstrap consistency

Theorem 4.6. Let the blocks in MBB be i.i.d. and the dataset $X_1, \dots, X_n$ have bounded

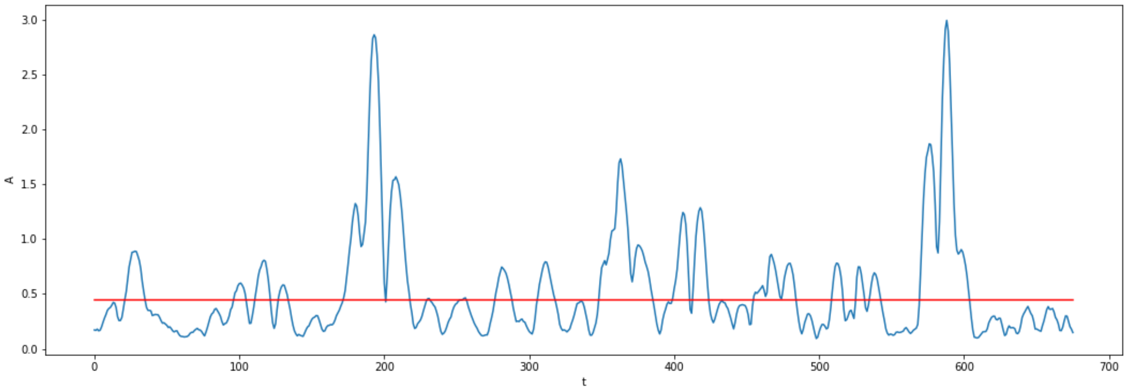


Figure 13. Example of Wasserstein distances plot with bootstrap bound corresponded to 95%-quantile of $\max T_h^\flat$.

support space. Then for a fixed τ with $h \to \infty$

$$T_h^\flat(\tau) \xrightarrow{d} T_h(\tau)$$

Proof. Bootstrap validity follows from combination of Theorem 2.7 and Theorem 2.5. Denote

$$\Phi_p = \left\{(\boldsymbol{u}, \boldsymbol{v}) : u_x + v_{x'} \le d^p\left(x, x'\right),\ (x, x') \in I\!R^d\right\}$$

There exist Gaussian vectors $Z_1, Z_2 \in \mathcal{N}(0, \Sigma_\psi)$ and generalized Fourier basis $\{\psi_i\}_{i=1}^\infty$, such that (Theorem 2.7)

$$I\!P(T_h(\tau) > x) \to I\!P\left(\max_{\boldsymbol{u},\boldsymbol{v} \in \Phi_p} \langle \boldsymbol{u}, Z_1^T \psi\rangle + \langle \boldsymbol{v}, Z_2^T \psi\rangle\right)$$

Analogically for bootstrap statistics and some Gaussian vectors $Z_1^\flat, Z_2^\flat \in \mathcal{N}(0, \Sigma_\psi^\flat)$

$$I\!P(T_h^\flat(\tau) > x) \to I\!P\left(\max_{\boldsymbol{u},\boldsymbol{v} \in \Phi_p} \langle \boldsymbol{u}, (Z_1^\flat)^T \psi\rangle + \langle \boldsymbol{v}, (Z_2^\flat)^T \psi\rangle\right)$$

By definition

$$\Sigma_\psi = I\!E \sum_i \psi\psi^T(X_i), \quad \Sigma_\psi^\flat = \sum_i \psi\psi^T(X_i)$$

$\Sigma_\psi^\flat$ converges by probability to Σ_ψ and according to Theorem 2.5 the maximum of Gaussian vectors converges to each other by distribution. □

Experiments We used the MIT-BIH arrhythmia dataset from the PhysioNet Moody and Mark (2001). The MIT-BIH Arrhythmia Dataset contains 48 half-hour excerpts of two-channel ambulatory ECG recordings, studied by the BIH Arrhythmia Laboratory between 1975 and 1979. 23 recordings were chosen at random from a set of 4000 24-hour ambulatory

ECG recordings and include most common arrhythmia types. The remaining 25 recordings include less common but clinically significant arrhythmias. Each record contains two 30-min ECG lead signal (mostly MLII lead and lead V1/V2/V4/V5) sampling the data at a frequency of 360Hz. Our algorithm proved to work without any data pre-processing or noise reduction and detected all types of arrhythmia (see results in Table 1).

Each ECG series was split to parts of different size (40,000, 80,000, and 120,000 points). If we take the indexes of the points, whose values are above the separation line calculated in the bootstrap procedure, these points in the original ECG will be the points with the arrhythmia. The PhysioNet dataset has the annotations accompanying the data; therefore, it is possible to compare the predicted labels of the points with the ground truth.

The parameters of the first sliding window have the following values $Ms = 450$, $s = 1$, $\Delta t = 2$ (Δt is step of moving window), corresponding to the typical ECG sampling parameters, such as those in the MIT-BIH dataset. The size of the second sliding window is equal to 4 curve loops, it means that the window separates the series into 2 parts with 2 curve loops in each. We chose the confidence level α=5%.

To gauge the performance of the algorithm, we use sensitivity and specificity of the prediction. To calculate them we used a hold-out test set comprising the ECG signals with the normal heart beat (160 parts) and the ECG with arrhythmias (192 parts). As a result, the specificity of 86%, and the sensitivity of 92% were obtained. We have also calculated the same metrics for the artificial data, and for all types of arrhythmia (42 time series, with arrhythmia in different parts of series). The results are the following: sensitivity 97.2% with 4.1% standard deviation; specificity 96.2% with 3.1% standard deviation. Optimal choice of prediction threshold and the size of the sliding windows define the trade-off between the high recall and the low false positive rate.

Comparison of our algorithm against several other approaches is shown in Table 1. We note that the pipeline in Figure 10 was meant to be as simple as possible, providing a robust statistical approach to predict abnormal rhythms in an unsupervised manner with high computational efficiency. Enhancing the pipeline by obvious combination with the deep learning or the hybrid model-based analysis methods is beyond the scope of this paper. Relevant to the clinical approbation, the method was tested (and correctly detected) on the short-episode arrhythmia in the long-term monitoring data stream (Figure 14).

More experiments and quality estimation are described in paper Shvetsov et al. (2020).

Table 1. Comparison of proposed approach with state-of-the-art. Definitions of sensitivity and specificity follow those in Ref. Jun et al. (2018).

Method	Sens%	Spec%	Supervision
1	**92.0 ± 4.0**	**86.0 ± 6.0**	◊
*1**	**97.2 ± 4.1**	**96.2 ±3.1**	◊
2 Truong et al. (2018)	91.6	77.0	◊
*2** Truong et al. (2018)	88.9	84.1	◊
3 Adams and MacKay (2007)	92.0	80.6	◊
*3** Adams and MacKay (2007)	85.8	88.9	◊
4 Hua et al. (2018)	70.0	98.0	△
5 Jun et al. (2018)	99.6	97.8	□
6 Alfaras et al. (2019)	84.4	99.7	□
7 Philip de Chazal et al. (2004)	75.9	77.7	□
8 Kawazoe et al. (2016)	97.0	63.0	□
9 Faganeli and Jager (2010)	98.1	85.0	□

◊ Unsupervised △ Semi-supervised □ Supervised
1: **Bootstrap** on real data, *1**: **Bootstrap** on artificial data
2: Ruptures(PELT) on Wasserstein distance data
*2**: Ruptures(PELT) on Euclidean distance data
3: BOCP on Wasserstein distance data
*3**: BOCP on Euclidean distance data
4: SVM + PCA *5*: 2D CNN *6*: Echo State Network
7: LD QRS- and time interval-based features
8: LR *9*: DT+Heart rate features

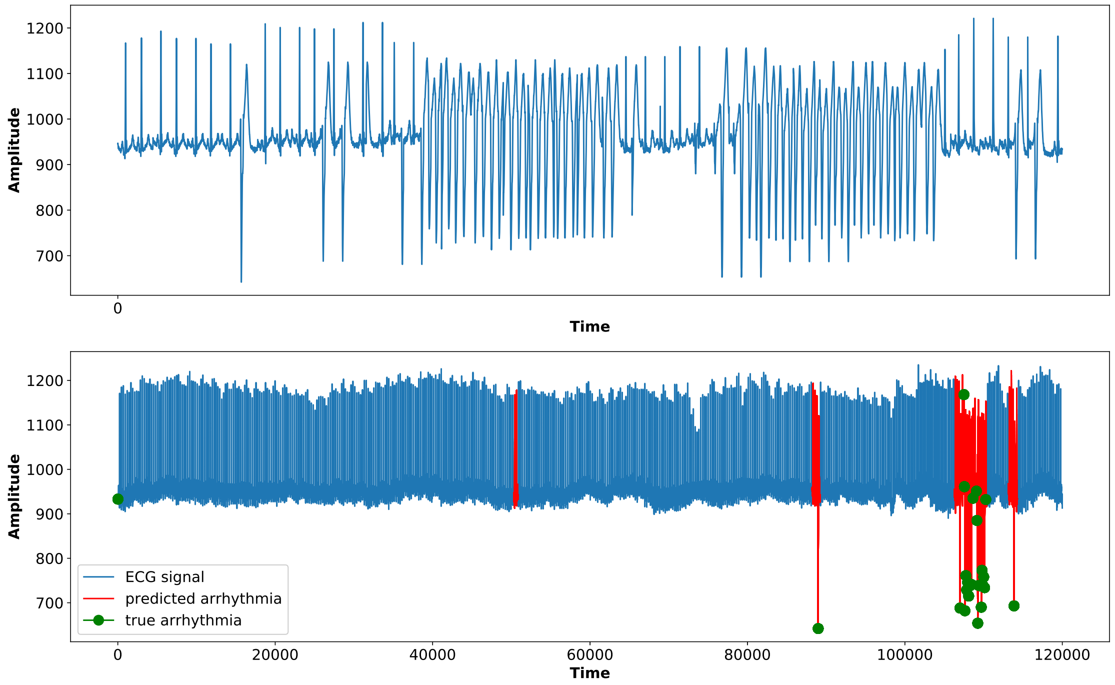


Figure 14. Detected arrhythmia example.

5 Wasserstein barycenters

Monge-Kantorovich distance or Wasserstein distance is a distance between measures. It represents a transportation cost of measure μ_1 into the other measure μ_2.

$$W_p(\mu_1, \mu_2) = \min_{\pi \in \Pi[\mu_1, \mu_2]} \left(\int \|x - y\|^p d\pi(x, y) \right)^{1/p}$$

where the condition $\pi \in \Pi[\mu_1, \mu_2]$ means that $\pi(x, y)$ has two marginal distributions: $\int_y d\pi(x, y) = d\mu_1(x)$ and $\int_x d\pi(x, y) = d\mu_2(y)$. We focus on regularized W1 distance with probabilistic space $\{\mathbb{R}^d, \mathcal{B}(\| \cdot \|_2), L^1\}$

$$\widetilde{W}_1(\mu_1, \mu_2) = \min_{\pi \in \Pi[\mu_1, \mu_2]} \int \|x - y\| d\pi(x, y) + R_\varepsilon(\pi)$$

where $R_\varepsilon(\pi)$ is a relatively small addition which improves differential properties of the distance. Namely without $R_\varepsilon(\pi)$ we can only bound the first derivative, with it we can bound also the second derivative. There is the notion of mean in Wasserstein distance, called barycenter $\widehat{\mu}$. And it is the main object in this paper. Consider a set of random measures $\{\mu_i\}_{i=1}^n$.

$$\widehat{\mu} = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^n \widetilde{W}_1(\mu, \mu_i)$$

Barycenters are center-of-mass generalization. If we look at the barycenter of a set of uniform measures it extracts the common "shape" form of these measures. If the measures are sampled from some distribution then their barycenter can be treated as an empirical approximation of the distribution mean. A simple example is a circles set with means $\{m_i \in \mathbb{R}^2\}$ and radius's $\{r_i\}$.

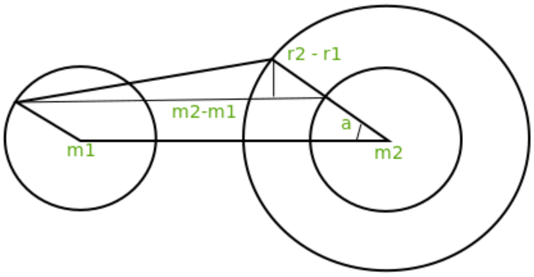


Figure 15. Illustration for W_2 distance computation between two circles (m_1, r_1) and (m_2, r_2).

$$W_2^2((m_1, r_1), (m_2, r_2))$$

$$= \frac{1}{2\pi} \int_0^{2\pi} [(m_2 - m_1) - (r_2 - r_1)\cos(a)]^2 + [(r_2 - r_1)\sin(a)]^2 da$$

$$= (m_2 - m_1)^2 + (r_2 - r_1)^2$$

Their W_2^2 barycenter is also a circle with mean $m = \frac{1}{n}\sum_{i=1}^n m_i$ and radius $r = \frac{1}{n}\sum_{i=1}^n r_i$. We refer to papers Agueh and Carlier (2011), Jeremie Bigot (2016) for an overview of the barycenters and related study.

It is well known that the center-of-mass in l_2 norm converges to a Gaussian random vector. As for the barycenter it is also expected to have some Gaussian properties. For example if the measures are Gaussian themselves or one-dimensional or circles set then the Gaussian approximation of the barycenter is proven in papers Agueh and Carlier (2011), Kroshnin et al. (2019). In circles set case the mean and radius converges to some Gaussian variables as a sum of independent observations according to Central Limit Theorem. In one-dimensional case denoting distribution functions by $F_i(x)$

$$W_2^2(\mu_1, \mu_2) = \int_0^1 |F_1^{-1}(s) - F_2^{-1}(s)|^2 ds$$

one gets

$$\widehat{F}^{-1}(s) = \frac{1}{n}\sum_{i=1}^n F_i^{-1}(s)$$

In the case of Gaussian measures with zero mean and variances $\{S_i\}$

$$W_2^2(\mu_1, \mu_2) = \operatorname{tr}\{S_1\} + \operatorname{tr}\{S_2\} - 2\operatorname{tr}\{(S_2^{1/2} S_1 S_2^{1/2})^{1/2}\}$$

and for some non-random matrix S_* (ref. Thomas Rippl (2015)) the corresponded barycenter variance is

$$\widehat{S} = \frac{1}{n}\sum_{i=1}^n (S_*^{1/2} S_i S_*^{1/2})^{1/2} + O(1/n)$$

In both last cases one deals with a mean of independent random variables that converges to a Gaussian variable (or to a Gaussian process in case of $\widehat{F}^{-1}(s)$ by Donsker's Theorem). In general case it appears to be very difficult to reveal such convergence because the barycenter doesn't have an explicit equation and it is an infinite-dimensional object. In order to handle with this difficulty we propose an approximation of the barycenter by a sum of independent variables using projection into Fourier basis and some novel results from statistical learning theory. The perspective of Fourier Analysis provides a suitable representation of the Wasserstein distance and it is already studied in the literature Steinerberger (2018). Denote a range of size p of the barycenter Fourier coefficients by

$$\widehat{\theta}_p = \mathcal{F}_p\left(\frac{d\widehat{\mu}(x)}{dx}\right)$$

The first our result states that for some non-random matrix $\breve{D}$, non-random vector θ_p^* and independent random vectors $\{\xi_i\}$

$$\left\| \breve{D}\left(\widehat{\theta}_p - \theta_p^*\right) - \sum_{i=1}^{n} \xi_i \right\| = O\left(\frac{p}{\sqrt{n}}\right)$$

Further we show that for some Gaussian vector Z

$$W_1\left(\breve{D}(\widehat{\theta}_p - \theta_p^*), Z\right) = O\left(\frac{p^{3/2}}{\sqrt{n}}\right)$$

and $\forall z$:

$$\left| P\left(\|D(\widehat{\theta}_p - \theta_p^*)\| > z\right) - P\left(\|Z\| > z\right) \right| = O\left(\frac{p}{\sqrt{n}}\right)$$

Statistical Application: The last statement allows us to obtain the confidence region of parameter $\widehat{\theta}_p$ and describe the distribution inside the region. Besides, the bootstrap procedure validity Max Sommerfeld (2016) follows from our proof as well. If one sample $\|D(\theta_p^{boot} - \widehat{\theta}_p)\|$ using bootstrap it would be close by quantiles to the random variable $\|\breve{D}(\widehat{\theta}_p - \theta_p^*)\|$, which also relates to the construction of the confidence region.

5.1 Statistical model

Consider a set of random measures (random measure is a measure-valued random element) with densities $\phi_1, \dots, \phi_n$. Let the barycenter measure $\widehat{\mu}$ has density $\widehat{\phi}$ and Fourier coefficients $\widehat{\theta} = \theta(\widehat{\phi}) \in I\!R^\infty$.

$$\widehat{\phi} = \underset{\phi}{\operatorname{argmin}} \sum_{i=1}^{n} \widetilde{W}_1(\phi, \phi_i)$$

Let Fourier basis $\{\psi_k\}_{k=1}^{\infty}$ has a Gram function of the scalar product $G(x)$, such that for any function f

$$\langle f, \psi_k \rangle_G = \int f(x)\psi_k(x)G(x)dx$$

and

$$\theta(\varphi)[k] = \int \varphi(x)\psi_k(x)dx$$

Denote Fourier coefficients of the other measures $\forall i : \theta_i = \theta(\varphi_i) \in I\!R^\infty$. Basing on Lemma 5.6 define an independent parametric model with dataset $(\theta_1, \dots, \theta_n)$ and parameter θ.

$$L(\theta) = \sum_{i=1}^{n} l(\theta - \theta_i),$$

where

$$l(\theta - \theta_i) = \max_{\eta \in \bigcap \mathcal{E}_x} \langle \eta, \theta - \theta_i \rangle - \varepsilon \eta^T (K \circ G)\eta = \widetilde{W}_1(\phi, \phi_i)$$

and $\bigcap \mathcal{E}_x$ is a Sobolev ellipsoids intersection. Each ellipsoid $\mathcal{E}_x$ has matrix $K_x = \nabla^T \psi \nabla \psi^T(x)$ such that

$$\left(\sum_{k \in \mathbb{N}_+^d} \eta_k \nabla \psi_k(x) \right)^2 = \eta^T K_x \eta$$

and

$$\bigcap \mathcal{E}_x = \left\{ \eta : \forall x : \eta^T K_x \eta \leq 1 \right\}$$

Define a positive matrix $K \circ G = \int K_x G(x) dx$ such that in case $\psi_k(x) = e^{ik^T x/T}$

$$K \circ G = \begin{pmatrix} 1/T^2 & 0 & 0 & \ldots 0 \\ 0 & \ddots & 0 & \ldots 0 \\ 0 & \ldots & k^2/T^2 & \ldots 0 \\ \ldots & \ldots & \ldots & \ldots \end{pmatrix}$$

Define for this model MLE parameter value and reference parameter value:

$$\widehat{\theta} = \operatorname*{argmin}_\theta L(\theta)$$

$$\theta^* = \operatorname*{argmin}_\theta I\!\!E L(\theta)$$

Define a local region around θ^*

$$\Omega(\mathtt{r}_0) = \{ \theta : \| D(\theta - \theta^*) \| \leq \mathtt{r}_0 \}$$

where D is a Fisher matrix of the model

$$D^2 = -\nabla^2 I\!\!E L(\theta^*)$$

Theorem 5.1. Let the random Fourier parameters of the dataset have a common density $\theta_1 \ldots \theta_n \sim q(\theta)$ and it fulfills condition

$$\int_{\theta \in \Omega(\mathtt{r}_0)} \| D^{-1} \nabla q(\theta) \| d\theta = \frac{C_Q}{\sqrt{n}}$$

Let $\widehat{\theta}, \theta^* \in I\!\!R^\infty$ be Fourier coefficients of the MLE and reference barycenter defined above, then with probability $1 - e^{-t}$

$$\left\| D(\widehat{\theta} - \theta^*) - D^{-1} \nabla L(\theta^*) \right\| \leq \diamondsuit(\mathtt{r}_0, t)$$

where $\diamondsuit(\mathtt{r}_0, t)$ is defined in Section 3.1 and has asymptotic

$$\diamondsuit(\mathtt{r}_0, t) = \frac{\sqrt{n} O(\mathtt{r}_0 C_Q + \mathtt{r}_0 \sqrt{p_D} + \sqrt{2t})}{\varepsilon \lambda_{\min}(DK \circ GD)} + o\left(\frac{1}{\sqrt{n}} \right)$$

and p_D is an ellipsoid entropy (Section 3.3) with matrix D and eigenvalues $\{\lambda_i(D)\}_{i=1}^{\infty}$

$$p_D = \sqrt{\sum_i \frac{\log^2(\lambda_i^2(D))}{\lambda_i^2(D)}}$$

and with probability $1 - e^{-t}$

$$\mathtt{r}_0 \leq 4\|D^{-1}\nabla L(\theta^*)\| \leq \frac{8\sqrt{n}(1+\sqrt{2t})}{\lambda_{\min}^{1/2}(DK \circ GD)}$$

Proof. Basing on Theorem 3.1 one has to prove Assumptions 1,2,3 from which follows

$$\left\| D(\widehat{\theta} - \theta^*) - D^{-1}\nabla L(\theta^*) \right\| \leq \{\delta(\mathtt{r}_0) + \mathfrak{z}(t)\}\mathtt{r}_0 = \diamondsuit(\mathtt{r}_0, t)$$

with probability $1 - e^{-t}$. The Assumptions 1,2,3 are proven below, where also is shown that

$$\delta(\mathtt{r}_0) = \frac{\mathtt{r}_0 n}{\varepsilon\lambda_{\min}(DK \circ GD)} \int_{\theta \in \Omega(\mathtt{r}_0)} \|D^{-1}\nabla q(\theta)\| d\theta$$

$$\mathfrak{z}(t) = E + \sqrt{2t(\mathbf{v}^2 + 2RE)} + \frac{tR}{3}$$

$$E = 6\mathbf{v}\sqrt{2p_D} + 12Rp_D$$

where

$$\mathbf{v}^2 = \frac{n}{\varepsilon^2\lambda_{\min}^2(DK \circ GD)}$$

and

$$R = \frac{1}{\varepsilon\lambda_{\min}(DK \circ GD)}$$

Setting $\mathbf{v}$ and R in the previous equations gives an asymptotic

$$\mathfrak{z}(t) = \frac{\sqrt{n}(12\sqrt{2p_D} + \sqrt{2t})}{\varepsilon\lambda_{\min}(DK \circ GD)} + O\left(\frac{1}{n}\right)$$

Lemma 5.4 provides bound

$$\|D^{-1}\nabla l\| \leq \frac{1}{\lambda_{\min}^{1/2}(DK \circ GD)}$$

From this bound and Hoefding's inequality Boucheron S. (2013) follows bound for $\|D^{-1}\nabla L(\theta^*)\|$. □

Define additional Fisher matrix corresponded to the projection into the first p elements of the parameter θ (ref. for details in Section 3.1).

$$\breve{D}^2 = D^2_{p\times p} - D^2_{p\times\infty} D^{-2}_{\infty\times\infty} D^2_{\infty\times p}$$

such that

$$D^2 = \begin{pmatrix} D^2_{p\times p} & D^2_{p\times\infty} \\ D^2_{\infty\times p} & D^2_{\infty\times\infty} \end{pmatrix}$$

and define the gradient of the projection into first p elements of the parameter θ.

$$\breve{\nabla} = \nabla_{1\ldots p} - D^2_{p\times\infty} D^{-2}_{\infty\times\infty} \nabla_{p\ldots\infty}$$

Theorem 5.2. Let $\widehat{\theta}_p, \theta^*_p \in I\!\!R^p$ are the first p Fourier coefficients of the MLE and reference barycenters, and Z is a Gaussian vector $\mathcal{N}(0, \operatorname{Var}[\breve{D}^{-1}\breve{\nabla}L(\theta^*)])$. Then, with probability $(1-e^{-t})$, W_1 and probability distance to Z are bounded as follows

$$W_1(\breve{D}(\widehat{\theta}_p - \theta^*_p), Z) \le \mu_3\, O(\log n) + \diamondsuit(\mathtt{r}_0, t)$$

and $\forall z \in I\!\!R_+$

$$|I\!\!P(\|D(\widehat{\theta}_p - \theta^*_p)\| > z) - I\!\!P(\|Z\| > z)| \le C_A(\mu_3\, O(\log^2 n) + \diamondsuit(\mathtt{r}_0, t))$$

where $\diamondsuit(\mathtt{r}_0, t)$ is defined in Theorem 5.1, $C_A = O(1/\sqrt{p})$ is anti-concentration constant defined in Theorem 2.2 and Theorem 2.7 (Götze et al. (2019)) and

$$\mu_3 \le \frac{4\sqrt{2}p}{\lambda^{1/2}_{\min}(DK \circ GD)}$$

Proof. Bind Theorems 5.1 and 2.1. Form Theorem 3.1 follows that the bound in Theorem 5.1 also holds for projection of the parameter θ:

$$\|\breve{D}(\widehat{\theta}_p - \theta^*_p) - \breve{D}^{-1}\breve{\nabla}L(\theta^*)\| \le \diamondsuit(\mathtt{r}_0, t)$$

So with probability $1 - e^{-t}$

$$W_1(\breve{D}(\widehat{\theta}_p - \theta^*_p), Z) = \min_{\pi(\widehat{\theta}, Z)} I\!\!E\|\breve{D}(\widehat{\theta}_p - \theta^*_p) - Z\|$$

$$\le W_1(\breve{D}^{-1}\breve{\nabla}L(\theta^*), Z) + \diamondsuit(\mathtt{r}_0, t)$$

Furthermore from Theorem 2.1 follows

$$W_1(\breve{D}^{-1}\breve{\nabla}L(\theta^*), Z) \le \sqrt{2}\mu_3\left(1 + \log(2\sqrt{\operatorname{tr}\{\Sigma\}}\mu_2) - \log(\mu_3)\right)$$

where $\Sigma = \mathrm{Var}[\breve{D}^{-1}\breve{\nabla}L(\theta^*)]$ and setting $X_i = \breve{D}^{-1}\breve{\nabla}l(\theta^* - \theta_i)$

$$\mu_3 = \sum_{i=1}^{n} I\!E\|\Sigma^{-1/2}(X_i - X_i')\|\|\Sigma^{-1/2}X_i\|\|X_i - X_i'\| \le 4\max\|X_i\|\sum_{i=1}^{n} I\!E X_i^T\Sigma^{-1}X_i$$

$$\sum_{i=1}^{n} I\!E X_i^T \Sigma X_i = \mathrm{tr}\left\{\Sigma^{-1}\sum_{i=1}^{n} I\!E X_i X_i^T\right\} = p$$

$$\max\|X_i\| = \|\breve{D}^{-1}\breve{\nabla}l(\theta^* - \theta_i)\| \le \|D^{-1}\nabla l(\theta^* - \theta_i)\| \le \frac{1}{\lambda_{\min}^{1/2}(DK \circ GD)}$$

Analogically one can make a consequence from Theorems 5.1 and 2.2. Let C_A is the anti-concentration constant of the distribution $I\!P(\|Z\| > z)$, then

$$|I\!P(\|\breve{D}(\widehat{\theta}_p - \theta_p^*)\| > z) - I\!P(\|Z\| > z)|$$
$$\le |I\!P(\|\breve{D}^{-1}\breve{\nabla}L(\theta^*)\| > z) - I\!P(\|Z\| > z)| + C_A\diamondsuit(\mathtt{r}_0, t)$$

and

$$|I\!P(\|\breve{D}^{-1}\breve{\nabla}L(\theta^*)\| > z) - I\!P(\|Z\| > z)| \le C_A\mu_3 O(\log^2 n)$$

As for the anti-concentration constant it can be estimated from Theorem 2.7 (Götze et al. (2019)):

$$I\!P(\|Z\|^2 \in [z, z+\Delta]) \le O\left(\frac{\Delta}{(\Lambda_{1Z}\Lambda_{2Z})^{1/2}}\right) = O\left(\frac{\Delta}{\sqrt{p}}\right),$$

where with eigenvalues of matrix Σ: $\lambda_{1Z} \ge \lambda_{2Z} \ge \dots$

$$\Lambda_{kZ}^2 = \sum_{j=k}^{\infty}\lambda_{jZ}^2, \quad k = 1, 2$$

□

We are going to show that **Assumptions 1,2,3** are fulfilled for the barycenters model defined above. Also we need to estimate $\diamondsuit(\mathtt{r}_0, t)$. Remind that we deal with Likelihood function $L(\theta) = L(\theta, \{\theta_i\}_{i=1}^n)$ where implicit random vectors $\{\theta_i\}_{i=1}^n$ is a dataset of Fourier coefficients corresponded to the random measures $\{\mu_i\}_{i=1}^n$.

Assumption 1: Set $\mathtt{r} = \|D(\theta - \theta^*)\|$, then

$$\|D^{-1}\{\nabla^2 I\!E L(\theta) - \nabla^2 I\!E L(\theta^*)\}D^{-1}\| \le \|D^{-1}\{\nabla^3 I\!E L(\theta) D^{-1}\}D^{-1}\|\mathtt{r}$$

Let $q(\theta_i)$ be distribution of each θ_i then

$$\nabla^3 I\!E_i L(\theta - \theta_i) = \sum_{i=1}^{n}\int \nabla^3 l(\theta - \theta_i)q(\theta_i)d\theta_i = -\sum_{i=1}^{n}\int \nabla^2 l(\theta - \theta_i) \times \nabla q(\theta_i)d\theta_i$$

$$\|D^{-1}\{\nabla^3 I\!E L(\theta) D^{-1}\}D^{-1}\| \le \int \|D^{-1}\nabla^2 L(\theta - \theta_x)D^{-1}\|\|D^{-1}\nabla p(\theta_x)\|d\theta_x$$

and from the consequence of Theorem 5.5 one gets

$$\|D^{-1}\nabla^2 l(\theta-\theta_i)D^{-1}\| \le \frac{1}{\varepsilon\lambda_{\min}(DK\circ GD)}$$

$$\|D^{-1}\{\nabla^2 I\!\!E L(\theta) - \nabla^2 I\!\!E L(\theta^*)\}D^{-1}\| \le \frac{\mathtt{r}n}{\varepsilon\lambda_{\min}(DK\circ GD)}\int \|D^{-1}\nabla q(\theta)\|d\theta$$

Subsequently

$$\delta(\mathtt{r}_0) = \frac{\mathtt{r}_0 n}{\varepsilon\lambda_{\min}(DK\circ GD)}\int_{\theta\in\Omega(\mathtt{r}_0)} \|D^{-1}\nabla q(\theta)\|d\theta$$

Assumption 2: From Theorem 3.2 follows that if:

$$\frac{n}{\varepsilon^2\lambda^2_{\min}(DK\circ GD)} \le \mathbf{v}^2$$

and

$$\frac{1}{\varepsilon\lambda_{\min}(DK\circ GD)} \le R$$

then

$$\mathfrak{z}(t) \le E + \sqrt{2t(\mathbf{v}^2+2RE)} + \frac{tR}{3}$$

where

$$E = 6\mathbf{v}\sqrt{2p_D} + 12Rp_D$$

and p_D is ellipsoid entropy with matrix D.
Assumption 3: Each model component $l(\theta-\theta_i)$ without regularisation is convex since

$$\begin{aligned} l(\lambda\theta_1 + (1-\lambda)\theta_2 - \theta_i) &= l(\lambda(\theta_1-\theta_i) + (1-\lambda)(\theta_2-\theta_i)) \\ &= \max_{\eta\in\bigcap\mathcal{E}_x}\langle\eta, \lambda(\theta_1-\theta_i) + (1-\lambda)(\theta_2-\theta_i))\rangle \\ &\le \max_{\eta\in\bigcap\mathcal{E}_x}\langle\eta, \lambda(\theta_1-\theta_i)\rangle + \max_{\eta\in\bigcap\mathcal{E}_x}\langle\eta, (1-\lambda)(\theta_1-\theta_i)\rangle \\ &= \lambda l(\theta_1-\theta_i) + (1-\lambda)l(\theta_2-\theta_i) \end{aligned}$$

Note that regularised l and l^2 are also convex as a composition of convex functions and the complete model L is convex ($\nabla^2 L > 0$) as a positive aggregation of convex functions. Combination of these assumptions is used in the proof of Theorem 5.1 which gives the deviation.

5.2 Support functions

Bounds for the first and second derivatives of the Likelihood of barycenters model involves additional theory from Convex analysis.

Def (*)**.** Legendre–Fenchel transform of a function $f : X \to \overline{I\!\!R}$ or the convex conjugate function calls

$$f^*(y) = \sup_{x \in X}(\langle x, y\rangle - f(x))$$

Def (s)**.** Support function for a convex body E is

$$s_E(\theta) = \sup_{\eta \in E} \theta^T \eta$$

Note that for indicator function $\delta_E(\eta)$ of a convex set E the conjugate function is support function of E

$$\delta_E^*(\theta) = s_E(\theta)$$

Def $(\oplus)$**.** Let $f_1, f_2 : E \to \overline{I\!R}$ be convex functions. The infimal convolution of them is

$$(f_1 \oplus f_2)(x) = \inf_{x_1 + x_2 = x}(f_1(x_1) + f_2(x_2))$$

Lemma 5.1. Bauschke and Combettes (2011) Let $f_1, f_2 : E \to \overline{I\!R}$ are convex lower-semi-continuous functions. Then

$$(f_1 \oplus f_2)^* = f_1^* + f_2^*$$
$$(f_1 + f_2)^* = \overline{f_1^* \oplus f_2^*}$$

Lemma 5.2. The support function of intersection $E = E_1 \cap E_2$ is infimal convolution of support functions for E_1 and E_2

$$s_E(\theta) = \inf_{\theta_1 + \theta_2 = \theta}(s_{E1}(\theta_1) + s_{E2}(\theta_2))$$

Proof. According to the previous Lemma

$$\delta_{E_1 \cap E_2}(\eta) = \delta_{E_1}(\eta) + \delta_{E_2}(\eta),$$

$$(\delta_{E_1} + \delta_{E_2})^* = \overline{\delta_{E_1}^* \oplus \delta_{E_2}^*}$$

With additional property

$$\operatorname{intdom} \delta_{E_1} \cap \operatorname{dom} \delta_{E_2} = \operatorname{int} E_1 \cap E_2 \neq \emptyset$$

one have

$$(\delta_{E_1} + \delta_{E_2})^* = \delta_{E_1}^* \oplus \delta_{E_2}^*$$

□

Lemma 5.3. Let a support function $s_E(\theta)$ be differentiable, then its gradient belongs to the border of corresponded convex set E

$$\nabla s_E(\theta) = \eta^*(\theta) \in \partial E$$

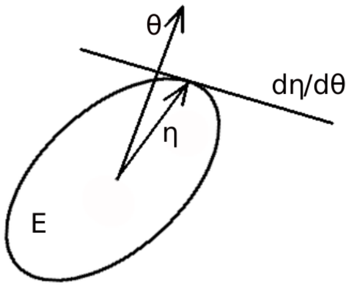


Figure 16. Optimization related to support function.

where

$$\eta^*(\theta) = \operatorname*{argmax}_{\eta \in E} \eta^T \theta$$

Proof. It follows from the convexity of E and linearity of the optimization functional.

$$\frac{\partial \eta^*(\theta)}{\partial \|\theta\|} = 0 \Rightarrow \frac{\partial \eta^*(\theta)}{\partial \theta}^T \theta = 0$$

$$\nabla s_E(\theta) = \frac{\partial \eta^*(\theta)}{\partial \theta}^T \theta + \eta^*(\theta) = \eta^*(\theta)$$

□

Lemma 5.4. Bauschke and Combettes (2011) Let $f_1, f_2 : E \to \overline{I\!R}$ be convex continuous functions. Then the subdifferential of their infimal convolution can be computed by formula

$$\partial(f_1 \oplus f_2)(x) = \bigcup_{x = x_1 + x_2} \partial f(x_1) \cap \partial f(x_2)$$

Consequence. If in addition f_1, f_2 are differentiable, then their infimal convolution is differentiable and $\exists x_1, x_2 : x = x_1 + x_2$ and

$$\nabla(f_1 \oplus f_2)(x) = \nabla f_1(x_1) = \nabla f_2(x_2)$$

Lemma 5.5. Let $f_1, \ldots, f_m : E \to \overline{I\!R}$ be convex and two times differentiable functions. There is an upper bound for the second derivative of the infimal convolution
$\forall t : \sum_{i=1}^m t_i = 1$

$$\partial \nabla^T (f_1 \oplus \ldots \oplus f_m)(x) \preceq \sum_{i=1}^m t_i^2 \nabla^2 f(x_i)$$

where $\sum_{i=1}^{m} x_i = x$.

Proof. Use notation $f = f_1 \oplus \ldots \oplus f_m$. Let

$$f(x) = \sum_i f_i(x_i)$$

According to Lemma 5.4 if all the functions are differentiable then

$$\nabla f(x) = \sum_i t_i \nabla f_i(x_i)$$

From the definition $\oplus$ also follows that

$$f(x+z) \leq \sum_i f_i(x_i + t_i z)$$

Make Tailor expansion for the left and right parts and account equality of the first derivatives.

$$z^T \partial \nabla^T f(x + \theta z) z \leq \sum_i t_i^2 z^T \nabla^2 f_i(x_i + \theta_i z) z$$

Since the direction z was chosen arbitrarily, dividing both parts of the previous equation by $\|z\|^2 \to 0$, we come to inequality

$$\partial \nabla^T f(x) \preceq \sum_i t_i^2 \nabla^2 f_i(x_i)$$

□

Remark. One can find another provement of the similar Theorem in book Bauschke and Combettes (2011) (Theorem 18.15).

Theorem 5.3. Let $f_1, \ldots, f_m : E \to \overline{I\!R}$ be convex and two times differentiable functions. There is an upper bounds for infimal convolution $f = f_1 \oplus \ldots \oplus f_m$ derivatives $\forall \gamma$ $\exists x_1, \ldots, x_m$:

$$\gamma^T \partial \nabla^T f(x) \gamma \leq \max_i \gamma^T \nabla^2 f_i(x_i) \gamma \frac{f_i(x_i)}{f(x)}$$

and

$$\gamma^T \partial \nabla^T f^2(x) \gamma \leq 2(\gamma^T \nabla f(x))^2 + 2 \max_i \gamma^T \nabla^2 f_i(x_i) \gamma f_i(x_i)$$

Proof. Choosing appropriate $\{t_i\}$ in Lemma 5.5 one get the required upper bounds. Set

$$t_i = \frac{f_i(x_i)}{\sum_{j=1}^{m} f_j(x_j)}$$

and since

$$\sum_{j=1}^{m} f_j(x_j) = f(x)$$

$$\sum_i t_i^2 \gamma^T \nabla^2 f_i(x_i)\gamma \leq \max_i t_i \gamma^T \nabla^2 f_i(x_i)\gamma = \max_i \gamma^T \nabla^2 f_i(x_i)\gamma f_i(x_i)$$

In order to prove the second formula apply this inequality in

$$\partial \nabla^T f^2 = 2\nabla f \nabla^T f + 2f \partial \nabla f$$

□

Consequence. Let $s_1, \ldots, s_m : E^* \to \overline{I\!R}$ are support functions of the bounded convex smooth sets $E_1, \ldots, E_m$. There are upper bounds for the derivatives of support function s of intersection $E_1 \cap \ldots \cap E_m$, such that $\forall i$

$$\gamma^T \partial \nabla^T s(\theta)\gamma \leq \frac{\max_i \gamma^T \partial \eta_i^* / \partial \theta_i \gamma s_i(\theta_i)}{s(\theta)}$$

$$\gamma^T \partial \nabla^T s^2(\theta)\gamma \leq 2(\gamma^T \eta_i^*)^2 + 2\max_i \gamma^T \partial \eta_i^* / \partial \theta_i \gamma s_i(\theta_i)$$

Proof. It follows from Theorem 5.3 and Lemma 5.3. □

5.3 Wasserstein distance as a support function

Def (W-dual)**.** Consider two random variables X and $Y \in \mathbb{R}^p$ with densities φ_X and φ_Y. Define Wasserstein distance in dual form between them as

$$W_1(\varphi_X, \varphi_Y) = \max_{\forall x: \|\nabla f(x)\| \leq 1} \{I\!E f(X) - I\!E f(Y)\}$$

where $\forall x : \|\nabla f(x)\| \leq 1$ means that function f is 1-Lipshits. Note that if $\pi(x, y)$ is a joint distribution with marginals φ_X and φ_Y then this definition is equivalent to the original one

$$W_1(\varphi_X, \varphi_Y) = \min_\pi I\!E\|X - Y\|$$

which follows from Kantorovich-Rubinstein duality Edwards (2011). Involve a normalized Fourier basis $\{\psi_k(x)\}_{k \in \mathbb{N}^p}$ with a scalar product Gram function $G(x)$.

Def $(W - dual - regularised)$**.** Consider two random variables X and $Y \in \mathbb{R}^p$ with densities φ_X and φ_Y. Define a penalized Wasserstein distance between them in dual form as

$$\widetilde{W}(\varphi_X, \varphi_Y) = \max_{\forall x: \|\nabla f(x)\| \leq 1} \left\{ I\!E f(X) - I\!E f(Y) - \varepsilon \int \|\nabla f(x)\|^2 G(x) dx \right\}$$

The regulariser term in this definition allows to bound the second derivative of the distance which will be shown below. Wasserstein distance in Fourier basis is a support function (ref. Def(s)). In which connection

$$f(x) = \sum_k \eta_k(f)\psi_k(x)$$

where

$$\eta_k(f) = \langle f, \psi_k\rangle_G = \int f(x)\psi_k(x)G(x)dx$$

Now we can rewrite the expectation difference as

$$Ef(X) - Ef(Y) = \langle f, \frac{\varphi_X}{G}\rangle_G - \langle f, \frac{\varphi_Y}{G}\rangle_G = \langle \eta(f), \theta(\varphi_X)\rangle - \langle \eta(f), \theta(\varphi_Y)\rangle$$

where

$$\theta_k(\varphi) = \int \varphi(x)\psi_k(x)dx$$

Define positive symmetric matrices

$$K_x = \begin{pmatrix} \nabla^T\psi_1(x) \\ \dots \\ \nabla^T\psi_k(x) \\ \dots \end{pmatrix} \begin{pmatrix} \nabla\psi_1(x) & \dots & \nabla\psi_k(x) & \dots \end{pmatrix} = (\nabla^T\psi(x))(\nabla\psi^T(x))$$

and

$$K \circ G = \int K_x G(x)dx$$

Each K_x is positive, since $\eta^T K_x \eta = \|\nabla f(x)\|^2$. Condition $\forall x : \|\nabla f(x)\| \leq 1$ is equivalent in Fourier basis to

$$\eta \in \bigcap \mathcal{E}_x = \left\{\eta : \forall x : \left(\sum_k \eta_k \nabla\psi_k(x)\right)^2 = \eta^T K_x \eta \leq 1\right\}$$

An important remark is that

$$\bigcap \mathcal{E}_x \subset \left\{\eta : \eta^T(K \circ G)\eta \leq 1\right\}$$

Finally we have come to the Wasserstein distance in Fourier basis.

Lemma 5.6. Let random vectors X and Y have densities φ_X and φ_Y with Fourier coefficients θ_X and θ_Y, then the Wasserstein distance is the support function of the convex set $\bigcap \mathcal{E}_x$ defined above, i.e.

$$W_1(\varphi_X, \varphi_Y) = \max_{\eta \in \bigcap \mathcal{E}_x} \langle \eta, \theta_X - \theta_Y\rangle$$

As for regularised case

$$\widetilde{W_1}(\varphi_X, \varphi_Y) = \max_{\eta \in \bigcap \mathcal{E}_x} \langle \eta, \theta_X - \theta_Y \rangle - \varepsilon \eta^T K \circ G \eta$$

Remind that the barycenters Likelihood consists of independent components $l_i(\theta - \theta_i)$ with random vectors $\theta_i \in I\!R^\infty$ and parameter $\theta \in I\!R^\infty$.

$$l(\theta - \theta_i) = \max_{\eta \in \bigcap \mathcal{E}_x} \langle \eta, \theta - \theta_i \rangle - \varepsilon \eta^T K \circ G \eta$$

Note that by definition the dual function of l is

$$l^*(\eta) = \delta_{\bigcap \mathcal{E}_x}(\eta) + \varepsilon \eta^T K \circ G \eta$$

Consequently from Lemma 5.1 follows that

$$l(\theta - \theta_i) = \delta^*_{\bigcap \mathcal{E}_x}(\theta - \theta_i) \oplus (\varepsilon \eta^T K \circ G \eta)^*(\theta - \theta_i)$$

$$= \max_{\eta \in \bigcap \mathcal{E}_x} \langle \eta, \theta - \theta_i \rangle \oplus \frac{1}{\varepsilon} (\theta - \theta_i)^T (K \circ G)^{-1} (\theta - \theta_i) \tag{0.1}$$

Application Theorem 5.3, taking into account $\bigcap \mathcal{E}_x \subset \Big\{ \eta : \eta^T (K \circ G) \eta \le 1 \Big\}$, provides the following bounds on the derivatives of function l.

Theorem 5.4. The gradient upper bounds of functions l and l^2 are

$$\|D^{-1} \nabla l\| \le \frac{1}{\lambda^{1/2}_{\min}(DK \circ GD)}$$

$$\|D^{-1} \nabla l^2(\theta - \theta_i)\| \le \frac{2\|(K \circ G)^{-1/2}(\theta - \theta_i)\|}{\lambda^{1/2}_{\min}(DK \circ GD)}$$

Proof. Denote

$$\eta^*(\theta) = \operatorname*{argmax}_{\eta \in \bigcap \mathcal{E}_x} \eta^T \theta$$

Use equation (0.1). By the consequence of Lemma 5.4 and Lemma 5.3 $\exists \theta_0$:

$$\nabla l(\theta - \theta_i) = \eta^*(\theta_0)$$

Since $\|(K \circ G)^{1/2} \eta^*\| \le 1$

$$\|D^{-1} \nabla l\| = \|D^{-1} \eta^*\| = \|D^{-1} (K \circ G)^{-1/2} (K \circ G)^{1/2} \eta^*\| \le \|D^{-1} (K \circ G)^{-1/2}\|$$

and from $\nabla l^2 = 2l \nabla l$ one gets

$$\|D^{-1} \nabla l^2(\theta - \theta_i)\| \le 2l(\theta - \theta_i) \|D^{-1} \nabla l\| \le 2\|(K \circ G)^{-1/2}(\theta - \theta_i)\| \|D^{-1} \nabla l\|$$

□

Theorem 5.5. The second derivative upper bounds of functions l and l^2 are

$$\|D^{-1}\partial\nabla^T l(\theta-\theta_i)D^{-1}\| \le \frac{1}{\min_x \lambda_{\min}(DK_xD)\|(K\circ G)^{-1/2}(\theta-\theta_i)\|}$$

$$\|D^{-1}\partial\nabla^T l(\theta-\theta_i)D^{-1}\| \le \frac{1}{\varepsilon\lambda_{\min}(DK\circ GD)}$$

$$\|D^{-1}\partial\nabla^T l^2 D^{-1}\| \le \frac{2}{\min_x \lambda_{\min}(DK_xD)}$$

Remark. Matrix K_x may be singular which makes the first bound non-informative. The second bound comes from the regulariser $\varepsilon\eta^T K\circ G\eta$ and has big coefficient $(1/\varepsilon)$. It is a weak part of the current theory and requires an improvement or an example which shows that this bound it tight.

Proof. Consider support function with one ellipsoid.

$$s_x(\theta) = \max_{\eta^T K_x \eta \le 1} \langle \eta, \theta\rangle = \|K_x^{-1/2}\theta\|$$

Denote $\eta^*(\theta) = \operatorname{argmax}\langle\eta,\theta\rangle$, and account that $\eta^T K_x \eta \le 1$.

$$\eta^*(\theta) = \frac{K_x^{-1}\theta}{\|K_x^{-1/2}\theta\|}$$

$$\frac{\partial\eta^*(\theta)}{\partial\theta} = \frac{K_x^{-1}\theta^T K_x^{-1}\theta - K_x^{-1}\theta\theta^T K_x^{-1}}{\left(\theta^T K_x^{-1}\theta\right)^{3/2}}$$

For some vector $\|\gamma\| = 1$ by means of property $\|a\|^2\|b\|^2 \ge (a^Tb)^2$

$$\gamma^T K_x^{-1}\gamma\theta^T K_x^{-1}\theta - \gamma^T K_x^{-1}\theta\theta^T K_x^{-1}\gamma \le \|K_x^{-1}\|\theta^T K_x^{-1}\theta$$

$$\left\|\frac{\partial\eta^*(\theta)}{\partial\theta}\right\| \le \frac{\|K_x^{-1}\|}{\left(\theta^T K_x^{-1}\theta\right)^{1/2}}$$

Apply Theorem 5.3that gives the first bound

$$\|D^{-1}\partial\nabla^T l(\theta-\theta_i)D^{-1}\| \le \max_x \left\|D^{-1}\frac{\partial\eta_x^*(\theta_x^*)}{\partial\theta}D^{-1}\right\| \frac{s_x(\theta_x^*)}{s(\theta-\theta_i)}$$

$$\le \frac{\max_x \|D^{-1}K_x^{-1}D^{-1}\|}{\|(K\circ G)^{-1/2}(\theta-\theta_i)\|}$$

The second bound for this norm follows directly from Lemma 5.5 and equation (0.1). Now consider the squared Wasserstein distance (l^2) which has a better derivative bound. From

Theorem 5.3 one gets

$$\|D^{-1}\partial\nabla^T l^2 D^{-1}\| \le 2\max_x \left\| D^{-1}\eta^*(\theta^*_x)\eta^*(\theta^*_x)^T D^{-1} + D^{-1}\frac{\partial\eta^*(\theta^*_x)}{\partial\theta}\|K_x^{-1/2}\theta^*_x\| D^{-1} \right\|$$

Note that

$$\frac{\partial\eta^*(\theta)}{\partial\theta}\|K_x^{-1/2}\theta\| = K_x^{-1} - \frac{(K_x^{-1}\theta)(K_x^{-1}\theta)^T}{\|K_x^{-1/2}\theta\|^2}$$

$$\eta^*(\theta)\eta^*(\theta)^T + \frac{\partial\eta^*(\theta)}{\partial\theta}\|K_x^{-1/2}\theta\| = K_x^{-1}$$

Finally

$$\|D^{-1}\partial\nabla^T l^2 D^{-1}\| \le 2\max_x \|D^{-1}K_x^{-1}D^{-1}\|$$

□

Remark. Wasserstein distance also may be differentiated directly. Paper Max Sommerfeld (2016) contains corresponded lemma about directional derivative. For directions h_1, h_2 it holds

$$W_1'(\mu_X, \mu_Y)(h_X, h_Y) = \max_{(u,v)\in\Phi(\mu_X,\mu_Y)} -(\langle u, h_X\rangle + \langle v, h_X\rangle)$$

where

$$\Phi = \{(u,v) : \langle u, \mu_X\rangle + \langle v, \mu_Y\rangle = W_1(\mu_X, \mu_Y),\ \forall(x,y) : u(x) + v(y) \le \|x - y\|\}$$

6 Supplementary math tools

6.1 Matrix Bernstein inequality

Lemma 6.1 (Master bound)**.** Assume that $S_1, \ldots, S_n$ are independent Hermitian matrices of the same size and $Z = \sum_{i=1}^n S_i$. Then

$$\mathbb{E}\lambda_{\max}(Z) \le \inf_{\theta>0} \frac{1}{\theta} \log \operatorname{tr} \exp\left(\sum_{i=1}^n \log \mathbb{E}e^{\theta S_i}\right)$$

$$\mathbb{P}\{\lambda_{\max}(Z) \ge z\} \le \inf_{\theta>0} e^{-\theta z} \ \operatorname{tr} \exp\left(\sum_{i=1}^n \log \mathbb{E}e^{\theta S_i}\right)$$

Proof. By the Markov inequality

$$\mathbb{P}\{\lambda_{\max}(Z) \ge z\} \le \inf_{\theta} e^{-\theta z} \mathbb{E} \exp(\theta \lambda_{\max}(Z))$$

Recall the spectral mapping theorem: for any function $f\colon \mathbb{R} \to \mathbb{R}$ and Hermitian matrix A eigenvalues of $f(A)$ are equal to eigenvalues of A. Thus

$$\exp(\theta\lambda_{\max}(Z)) = \exp(\lambda_{\max}(\theta Z)) = \lambda_{\max}\big(\exp(\theta Z)\big) \le \operatorname{tr}\ e^{\theta Z}$$

Therefore,

$$\mathbb{P}\{\lambda_{\max}(Z) \ge z\} \le \inf_{\theta} e^{-\theta z} \mathbb{E} \operatorname{tr} \exp(\theta Z)$$

and the second statement follows. To prove the first statement fix θ. Using the spectral mapping theorem one can get that

$$\mathbb{E}\lambda_{\max}(Z) = \frac{1}{\theta}\mathbb{E}\lambda_{\max}(\theta Z) = \frac{1}{\theta} \log \mathbb{E} \exp\big(\lambda_{\max}(\theta Z)\big) = \frac{1}{\theta} \log \mathbb{E}\lambda_{\max}\big(\exp(\theta Z)\big)$$

Thus we get

$$\mathbb{E}\lambda_{\max}(Z) \le \frac{1}{\theta} \log \operatorname{tr} \mathbb{E} \exp(\theta Z)$$

The final step in proving the master inequalities is to bound from above $\mathbb{E} \operatorname{tr} \exp\left(\sum_{i=1}^n S_i\right)$. To do this we use Jensen's inequality for the convex function $\operatorname{tr} \exp(H + \log(X))$ (in matrix X), where H is deterministic Hermitian matrix. For a random Hermitian matrix X one can write

$$\mathbb{E} \operatorname{tr} \exp(H + X) = \mathbb{E} \operatorname{tr} \exp(H + \log e^X) \le \operatorname{tr} \exp(H + \log \mathbb{E}e^X)$$

Convexity of function $(\operatorname{tr} \exp(H + \log(X)))$ is followed from

$$\operatorname{tr} \exp(H + \log(X)) = \max_{Y \succ 0} [\operatorname{tr}(YH) - (D(Y;X) - \operatorname{tr} X)]$$

where $D(Y;X)$ is relative entropy

$$D(Y;X) = \phi(X) - [\phi(Y) + \langle \nabla\phi(Y), X - Y\rangle] \quad \phi(X) = \operatorname{tr}(X\log X)$$

due to the partial maximum and $D(Y;X)$ are concave functions. Denote by $\mathbb{E}_i$ the conditional expectation with respect to random matrix X_i. To bound $\mathbb{E}\operatorname{tr}\exp\left(\sum_{i=1}^{n} S_i\right)$ we use the sum of independent Hermitian matrices by taking the conditional expectations with respect to i-th matrix:

$$\begin{aligned}\mathbb{E}\operatorname{tr}\exp\left(\sum_{i=1}^{n} S_i\right) &= \mathbb{E}\mathbb{E}_n \operatorname{tr}\exp\left(\sum_{i=1}^{n-1} S_i + S_n\right)\\ &\le \mathbb{E}\operatorname{tr}\exp\left(\sum_{i=1}^{n-1} S_i + \log(\mathbb{E}_n\exp(S_n))\right)\\ &\le \operatorname{tr}\exp\left(\sum_{i=1}^{n}\log\mathbb{E}\mathrm{e}^{\theta S_i}\right)\end{aligned}$$

□

The next statement was taken from Koltchinskii (2013) with the proof sketch.

Lemma 6.2 (Bernstein inequality for moment restricted matrices)**.** Suppose that

$$\forall i: \ \mathbb{E}\psi^2\left(\frac{\|S_i\|_{\mathrm{op}}}{M}\right) \le 1$$

$$\mathrm{v}^2 = \left\|\sum_{i=1}^{n}\mathbb{E}S_i^2\right\|_{\mathrm{op}}$$

$$R = M\psi^{-1}\left(\frac{2}{\delta}\frac{nM^2}{\mathrm{v}^2}\right), \quad \delta\in(0, 2/\psi(1))$$

Then under condition $zR \le (e-1)(1+\delta)\mathrm{v}^2$

$$\mathbb{P}\{\|Z\|_{\mathrm{op}} \ge z\} \le 2p\exp\left\{-\frac{z^2}{2(1+\delta)\mathrm{v}^2 + 2Rz/3}\right\}$$

If $\psi(u) = e^{u^\alpha} - 1$ then $R = M\log^{1/\alpha}(\frac{2}{\delta}\frac{nM^2}{\mathrm{v}^2} + 1)$.

Proof. According to Master bound one have to estimate $\mathbb{E}e^{\theta S}$ for S in $S_1, \ldots, S_n$. Denote a function

$$f(u) = \frac{e^u - 1 - u}{u^2}$$

Taylor expansion yields

$$\mathbb{E}e^{\theta S} \le I_p + \theta^2\mathbb{E}S^2 f(\theta\|S\|)$$

$$\log \mathbb{E} e^{\theta S} \le \theta^2 \mathbb{E} S^2 f(\theta \|S\|) \le \theta^2 f(\theta\tau) \mathbb{E} S^2 + I_p \theta^2 \mathbb{E} \|S\|^2 f(\theta \|S\|) I(\|S\| \ge \tau)$$

$$\mathbb{E} \|S\|^2 f(\theta \|S\|) I(\|S\| \ge \tau) \le M^2 \mathbb{E} \psi^2 \left(\frac{\|S\|}{M}\right) \left(\psi\left(\frac{\tau}{M}\right)\right)^{-1} \le M^2 \left(\psi\left(\frac{\tau}{M}\right)\right)^{-1}$$

$$M^2 \left(\psi\left(\frac{R}{M}\right)\right)^{-1} = \frac{\delta \mathrm{v}^2}{2n}$$

$$\log \mathbb{E} e^{\theta S} \le \theta^2 f(\theta R) \mathbb{E} S^2 + I_p \theta^2 \frac{\delta \mathrm{v}^2}{2n}$$

$$\operatorname{tr}\exp\left(\sum_{i=1}^{n} \log \mathbb{E} \exp(\theta S_i)\right) \le \operatorname{tr}\exp\left(\theta^2 f(\theta R) \mathbb{E} \sum_i S_i^2\right) \exp\left(\theta^2 \frac{\delta \mathrm{v}^2}{2}\right)$$

□

Consequence. In case $\psi(u) = e^u - 1$ with probability $1 - 2e^{-x}$

$$z \le \frac{2}{3} R \mathrm{x}_p + \mathrm{v}\sqrt{5\mathrm{x}_p}, \quad R \approx M$$

Condition for function $\psi(u) = e^u - 1$ follows from sub-Gaussian moment restriction

$$\mathbb{E} \exp\left(2 \frac{\|S_i\|_{\mathrm{op}}}{M}\right) \le 2$$

Lemma 6.3 (Deviation bound for matrix convolution with sub-Gaussian weights)**.** Let a set of symmetric $[p \times p]$ matrices $(A_1, \ldots, A_n)$ satisfy

$$\left\| \sum_i A_i^2 \right\| \le \mathrm{v}^2$$

Let ε_i be independent sub-Gaussian, $i = 1, \ldots, n$.

$$Z = \sum_{i=1}^{n} \varepsilon_i A_i$$

fulfills

$$\mathbb{P}\left(\|Z\|_{\mathrm{op}} \ge \sqrt{2\mathrm{x}_p \mathrm{v}^2} \right) \le 2\mathrm{e}^{-\mathrm{x}}$$

where $\mathtt{x}_p = \mathtt{x} + \log p$.

Proof. Apply the Master inequalities for the case

$$A_i = U_i \Lambda_i U_i^T, \quad \log \mathbb{E} e^{\varepsilon_i A_i} = \frac{1}{2} U_i \Lambda_i^2 U_i^T = \frac{1}{2} A_i^2$$

$$\operatorname{tr}\exp\left\{\sum_i \log I\!\!E e^{\theta\varepsilon_i A_i}\right\} \le p\exp\left\{\frac{1}{2}\theta^2\mathrm{v}^2\right\}$$

□

Lemma 6.4 (Deviation bound for rank one matrix convolution with sub-Gaussian weights)**.** Let vectors $u_1,\ldots,u_n$ in $I\!\!R^p$ satisfy

$$\|u_i\| \le \delta$$

for a fixed constant δ. Let ε_i be independent sub-Gaussian, $i=1,\ldots,n$. Then for each vector $b=(b_1,\ldots,b_n)^\top \in I\!\!R^n$, the matrix Z_1 with

$$Z_1 = \sum_{i=1}^n \varepsilon_i b_i u_i u_i^\top$$

fulfills

$$I\!\!P\Bigl(\|Z_1\|_{\mathrm{op}} \ge \delta^2\|b\|\sqrt{2\mathrm{x}}\Bigr) \le 2\mathrm{e}^{-\mathrm{x}}$$

Proof. As ε_i are i.i.d. standard sub-Gaussian and $I\!\!E\mathrm{e}^{a\varepsilon_i} \le \mathrm{e}^{a^2/2}$ for $|a|<1/2$, it follows from the Master inequality and property

$$\exp(\varepsilon_i u u^\top) = \frac{u_i u_i^\top}{u^2}\exp(\varepsilon_i u^2)$$

that

$$\begin{aligned} I\!\!P(\|Z_1\|_{\mathrm{op}} \ge z) &\le 2\inf_{\theta>0}\mathrm{e}^{-\theta z}\operatorname{tr}\exp\Bigl\{\sum_{i=1}^n \log I\!\!E\exp(\theta\varepsilon_i b_i u_i u_i^\top)\Bigr\} \\ &\le 2\inf_{\theta>0}\mathrm{e}^{-\theta z}\operatorname{tr}\exp\Bigl\{\sum_{i=1}^n \frac{\theta^2 b_i^2\|u_i\|^4}{2}\,\frac{u_i u_i^\top}{\|u_i\|^2}\Bigr\} \end{aligned}$$

Moreover, as $\|u_i\| \le \delta$ and $U_i = u_i u_i^\top/\|u_i\|^2$ is a rank-one projector with $\operatorname{tr} U_i = 1$, it holds

$$\operatorname{tr}\exp\Bigl\{\frac{\theta^2}{2}\sum_{i=1}^n b_i^2\|u_i\|^4 U_i\Bigr\} \le \exp\operatorname{tr}\Bigl(\frac{\theta^2\delta^4}{2}\sum_{i=1}^n b_i^2 U_i\Bigr) = \exp\frac{\theta^2\delta^4\|b\|^2}{2}$$

This implies for $z = \delta^2\|b\|\sqrt{2\mathrm{x}}$

$$I\!\!P(\|Z_1\|_{\mathrm{op}} \ge z) \le 2\inf_{\theta>0}\exp\Bigl(-\theta z + \frac{1}{2}\theta^2\delta^4\|b\|^2\Bigr) = 2\mathrm{e}^{-\mathrm{x}}$$

and the assertion follows. □

6.2 Variance deviation

Consider a sequence independent random variables $\{\varepsilon_i \varepsilon_j^T\}$, $\operatorname{cor}(\varepsilon_i, \varepsilon_j) = \Sigma_{ij}$, flatted into one vector ε. The subject of interest is upper bound for operator norm of

$$\mathcal{U} \operatorname{blockDiag}(\varepsilon\varepsilon^T)\mathcal{U}^T - I_q,$$

$$\mathcal{U} \operatorname{blockDiag}(\varepsilon\varepsilon^T)\mathcal{U}^T = \sum_{ij} \mathcal{U}_i \varepsilon_i \varepsilon_j^T \mathcal{U}_j^T, \quad \mathcal{U}\Sigma\mathcal{U}^T = I_q, \quad \left\|\mathcal{U}_i^T \mathcal{U}_j\right\|_{\mathrm{op}} \le \delta^2.$$

Analogically divide ε into mean and stochastic parts

$$\varepsilon = \mathbb{E}\varepsilon + (\varepsilon - \mathbb{E}\varepsilon) = B + \zeta.$$

Then initial term includes three parts:

$$\begin{aligned}
&\mathcal{U}\ \operatorname{blockDiag}(BB^\top)\mathcal{U}^\top \\
&+ \mathcal{U} \operatorname{blockDiag}(\zeta B^\top)\mathcal{U}^\top + \mathcal{U} \operatorname{blockDiag}(B\zeta^T)\mathcal{U}^\top \\
&+ \mathcal{U} \operatorname{blockDiag}(\zeta\zeta^T - \Sigma)\mathcal{U}^\top.
\end{aligned}$$

Estimate successively each component.

$$\begin{aligned}
\left\|\mathcal{U} \operatorname{blockDiag}(BB^\top)\mathcal{U}^\top\right\|_{\mathrm{op}} &= \sup_\gamma \sum_{i=1}^n \gamma^T \mathcal{U}_i B_i B_j^T \mathcal{U}_j^T \gamma \\
&\le \sup_\gamma \sum_{i=1}^n B_i^2 \left\|\mathcal{U}_i^T \mathcal{U}_j\right\|_{\mathrm{op}} \le \delta^2 \|B\|^2.
\end{aligned}$$

For the second and the third component one may apply Master bound 6.1, in which one have to estimate exponential moments of each element of the $\sum_i \mathcal{U}_i A_i \mathcal{U}_i^T$:

$$\operatorname{tr}\log \mathbb{E} \exp\{\mathcal{U}_i A_i \mathcal{U}_i^T\}.$$

With condition $\mathbb{E} A_i = 0$ an intuition hint is

$$\log \mathbb{E} e^{\mathcal{U}_i A_i \mathcal{U}_i^T} \le \frac{1}{2} \mathbb{E}(\mathcal{U}_i A_i \mathcal{U}_i^T)^2 + O(\left\|\mathcal{U}_i A_i \mathcal{U}_i^T\right\|_{\mathrm{op}}^3).$$

Consequently by means of Bernstein matrix inequality 6.2 one have to restrict the second moment and tail by $\log \mathbb{E} \exp\{\left\|\mathcal{U}_i A_{ij} \mathcal{U}_j^T\right\|_{\mathrm{op}}\}$.

$$\mathrm{v}^2 = \left\|\sum_{ij} \mathbb{E}(\mathcal{U}_i A_{ij} \mathcal{U}_j^T)^2\right\|_{\mathrm{op}} \le \max_i \left\|\mathcal{U}_i^T \mathcal{U}_j\right\|_{\mathrm{op}} \left\|\sum_{ij} \mathcal{U}_i \mathbb{E} A_{ij}^2 \mathcal{U}_j^T\right\|_{\mathrm{op}},$$

$$
\begin{aligned}
\mathrm{v}^2_{\zeta\zeta}/\delta^2 &= \left\| \sum_{ij} \mathcal{U}_i \mathbb{E}(\zeta_i \zeta_j^T - \Sigma_{ij})^2 \mathcal{U}_j^T \right\|_{\mathrm{op}} \\
&\le \max_{ij} \|\Sigma_{ij}\|_{\mathrm{op}} \left\| \mathbb{E}(\zeta_i^T \Sigma_{ij}^{-1} \zeta_j) \zeta_i \zeta_j^T \Sigma_{ij}^{-1} - I \right\|_{\mathrm{op}} \\
&\le \max_{ij} \|\Sigma_{ij}\|_{\mathrm{op}} \Big((\lambda^2 - 1) + \mathbb{E}(\zeta_i^T \Sigma_{ij}^{-1} \zeta_j)^2 \, \mathbb{1}(\zeta_i^T \Sigma_{ij}^{-1} \zeta_j > \lambda) \Big), \quad \lambda > 1.
\end{aligned}
$$

Upper bound for $\mathbb{E} \|\xi\|^4 \, \mathbb{1}(\|\xi\| > \sqrt{\lambda})$ from SGI with $\xi^2 = \zeta_i^T \Sigma_{ij}^{-1} \zeta_j$ leads to asymptotic $\mathrm{v}^2_{\zeta\zeta} = 9\delta^2 \max_{ij} \|\Sigma_{ij}\|_{\mathrm{op}} p$.

$$
\begin{aligned}
\mathrm{v}^2_{\zeta B}/\delta^2 &= \left\| \sum_{ij} \mathcal{U}_i \mathbb{E}(\zeta_i B_i^T)^2 \mathcal{U}_j^T \right\|_{\mathrm{op}} \\
&\le \max_{ij} \|\Sigma_{ij}\|_{\mathrm{op}} \sum_{ij} \left\| \mathcal{U}_j^T \mathcal{U}_i \right\|_{\mathrm{op}} B_i^2 \\
&\le \max_{ij} \|\Sigma_{ij}\|_{\mathrm{op}} \delta^2 \|B\|^2.
\end{aligned}
$$

Es for exponential moments for tails restriction

$$
\log \mathbb{E} \exp\{ \left\| \mathcal{U}_i A_{ij} \mathcal{U}_j^T \right\|_{\mathrm{op}} \} = \log \mathbb{E} \exp \left\{ \sup_{u_i, u_j} u_i^T A_{ij} u_j \right\}, \quad \|u_i\|^2, \|u_j\|^2 \le \delta^2.
$$

Sub-Gaussian properties for exponential moments (Lemma 6.7) lead to $\exists M_{\zeta\zeta}$, $M_{\zeta B}$:

$$
\log \mathbb{E} \exp \left\{ \sup_{u_i, u_j} \frac{2}{M_{\zeta\zeta}} u_i^T (\zeta_i \zeta_j^T - \Sigma_{ij}) u_j \right\} \le \log \mathbb{E} \exp \left\{ \frac{2\delta^2}{M_{\zeta\zeta}} (\|\zeta_i\|^2 - \lambda_{\min}(\Sigma_{ii})) \right\}
$$

$$
\le SG\{\|\xi\|^2, 2\delta^2 \|\Sigma_{ii}\|_{\mathrm{op}} / M_{\zeta\zeta}\} - \frac{2\delta^2}{M_{\zeta\zeta}} \lambda_{\min}(\Sigma_{ii}) \le \log(2),
$$

$$
\log \mathbb{E} \exp \left\{ \sup_{u_i, u_j} \frac{2}{M_{\zeta B}} u_i^T \zeta_i B_j^T u_j \right\} \le SG\{\|\xi\|, 2\delta^2 \sqrt{\|\Sigma_{ii}\|_{\mathrm{op}}} \|B_j\| / M_{\zeta B}\} \le \log(2),
$$

where $M_{\zeta\zeta} = 3\delta^2 p \|\Sigma_{ii}\|_{\mathrm{op}}$ and $M_{\zeta B} = 3\delta^2 p \sqrt{\|\Sigma_{ii}\|_{\mathrm{op}}} \|B_i\|$. Finally, as a consequence of Theorem 6.2 with probability $1 - 2e^{-x}$ and $\mathtt{x}_q = x + \log(2q)$ and $R_{**} \approx M_{**}$

$$
\left\| \mathcal{U} \, \mathrm{blockDiag}(\zeta B^\top) \mathcal{U}^\top \right\|_{\mathrm{op}} \le \frac{2}{3} R_{\zeta B} \mathtt{x}_q + 2 \mathrm{v}_{\zeta B} \sqrt{5 \mathtt{x}_q},
$$

$$
\left\| \mathcal{U} \, \mathrm{blockDiag}(\zeta \zeta^\top) \mathcal{U}^\top \right\|_{\mathrm{op}} \le \frac{2}{3} R_{\zeta\zeta} \mathtt{x}_q + 2 \mathrm{v}_{\zeta\zeta} \sqrt{5 \mathtt{x}_q}.
$$

So, the summarized error with probability $1 - 2e^{-x}$ is

$$\left\| \mathcal{U} \operatorname{blockDiag}(\varepsilon\varepsilon^T)\mathcal{U}^T - I_q \right\|_{\text{op}} \le \frac{2}{3} R_{\varepsilon\varepsilon} \mathtt{x}_q + 2 \mathrm{v}_{\varepsilon\varepsilon} \sqrt{5\mathtt{x}_q} + \delta^2 \|B\|^2 , \qquad \text{(ErrVD)}$$

where

$$R_{\varepsilon\varepsilon} \approx M_{\zeta B} + M_{\zeta\zeta} = 3\delta^2 p \max_i \left(\|\Sigma_{ii}\|_{\text{op}} + \sqrt{\|\Sigma_{ii}\|_{\text{op}}} \|B\|_\infty \right)$$

and

$$\mathrm{v}_{\varepsilon\varepsilon} = \mathrm{v}_{\zeta B} + \mathrm{v}_{\zeta\zeta} = \delta \max_i \sqrt{\|\Sigma_{ii}\|_{\text{op}}} (3\sqrt{p} + \|B\|).$$

6.3 Quadratic forms

Consider vector ξ which has restricted exponential or sub-Gaussian moments: $\exists\ \mathtt{g} > 0$

$$\log \mathbb{E} \exp(\gamma^\top \xi) \le \|\gamma\|^2/2, \qquad \gamma \in \mathbb{R}^p,\ \|\gamma\| \le \mathtt{g} \qquad \text{(SG)}$$

For ease of presentation, assume below that $\mathtt{g}$ is sufficiently large, namely, $0.3\mathtt{g} \ge \sqrt{p}$. In typical examples of an i.i.d. sample, $\mathtt{g} \asymp \sqrt{n}$. Define $\mathtt{x}_c = \mathtt{g}^2/4$.

Lemma 6.5 (Spokoiny (2016))**.** Let (SG) hold and $0.3\mathtt{g} \ge \sqrt{p}$. Then for each $\mathtt{x} > 0$

$$\mathbb{P}\bigl(\|\xi\| \ge z(p, \mathtt{x})\bigr) \le 2\mathrm{e}^{-\mathtt{x}} + 8.4\mathrm{e}^{-\mathtt{x}_c}\, \mathbb{I}(\mathtt{x} < \mathtt{x}_c)$$

where $z(p, \mathtt{x})$ is defined by

$$z(p, \mathtt{x}) = \begin{cases} (p + 2\sqrt{p\mathtt{x}} + 2\mathtt{x})^{1/2}, & \mathtt{x} \le \mathtt{x}_c \\ \mathtt{g} + 2\mathtt{g}^{-1}(\mathtt{x} - \mathtt{x}_c), & \mathtt{x} > \mathtt{x}_c \end{cases}$$

Usually the second term in previous equation can be simply ignored. Obtain similar result for quadratic form with matrix B. Define

$$\mathtt{p} = \operatorname{tr}(B), \qquad \mathrm{v}^2 = \operatorname{tr}(B^2), \qquad \lambda = \lambda_{\max}(B)$$

$$\begin{aligned} z_c^2 &= \mathtt{p} + \mathrm{v}\mathtt{g} + \lambda \mathtt{g}^2/2 \\ \mathtt{g}_c &= \frac{\sqrt{\mathtt{p}/\lambda + \mathtt{g}\mathrm{v}/\lambda + \mathtt{g}^2/2}}{1 + \mathrm{v}/(\lambda \mathtt{g})} \end{aligned}$$

Lemma 6.6 (Spokoiny (2016))**.** Let (SG) hold and $0.3\mathtt{g} \ge \sqrt{\mathtt{p}/\lambda}$. Then for each $\mathtt{x} > 0$

$$\mathbb{P}\bigl(\|B^{1/2}\xi\| \ge z(B, \mathtt{x})\bigr) \le 2\mathrm{e}^{-\mathtt{x}} + 8.4\mathrm{e}^{-\mathtt{x}_c}\, \mathbb{I}(\mathtt{x} < \mathtt{x}_c)$$

where $z(B, \mathtt{x})$ is defined by

$$z(B, \mathtt{x}) = \begin{cases} \sqrt{\mathtt{p} + 2\mathrm{v}\mathtt{x}^{1/2} + 2\lambda\mathtt{x}}, & \mathtt{x} \le \mathtt{x}_c \\ z_c + 2\lambda(\mathtt{x} - \mathtt{x}_c)/\mathtt{g}_c, & \mathtt{x} > \mathtt{x}_c \end{cases}$$

The upper quantile $z(B,\mathtt{x}) = \sqrt{\mathtt{p} + 2\mathrm{v}\mathtt{x}^{1/2} + 2\lambda\mathtt{x}}$ can be upper bounded by $\sqrt{\mathtt{p}} + \sqrt{2\lambda\mathtt{x}}$ and thus

$$z(B,\mathtt{x}) \le \begin{cases} \sqrt{\mathtt{p}} + \sqrt{2\lambda\mathtt{x}}, & \mathtt{x} \le \mathtt{x}_c \\ z_c + 2\lambda(\mathtt{x} - \mathtt{x}_c)/\mathtt{g}_c, & \mathtt{x} > \mathtt{x}_c \end{cases}$$

Lemma 6.7. Suppose (SG). For any $\mu < 1$ with $\mathtt{g}^2 > p\mu$, it holds

$$I\!\!E \exp\Bigl(\frac{\mu\|\xi\|^2}{2}\Bigr)\, \mathbb{1}\Bigl(\|\xi\| \le \mathtt{g}/\mu - \sqrt{p/\mu}\Bigr) \le 2(1-\mu)^{-p/2}$$

If $\mathtt{g}$ is sufficiently large than approximately

$$I\!\!E \exp\Bigl(\frac{\mu\|\xi\|^2}{2}\Bigr) \le (1-\mu)^{-p/2}$$

Lemma 6.8. Suppose (SG) and $\|B\|_{\mathrm{op}} = 1$. For any $\mu < 1$ with $\mathtt{g}^2/\mu \ge \mathtt{p}$, it holds

$$I\!\!E \exp(\mu\|B^{1/2}\xi\|^2/2)\, \mathbb{1}(\|B\xi\| \le \mathtt{g}/\mu - \sqrt{\mathtt{p}/\mu}) \le 2\det(I_p - \mu B)^{-1/2}$$

Proof. With $c_p(B) = (2\pi)^{-p/2}\det(B^{-1/2})$

$$\begin{aligned} & c_p(B) \int \exp\Bigl(\gamma^\top\xi - \frac{1}{2\mu}\|B^{-1/2}\gamma\|^2\Bigr)\, \mathbb{1}(\|\gamma\| \le \mathtt{g})d\gamma \\ &\quad = c_p(B)\exp\Bigl(\frac{\mu\|B^{1/2}\xi\|^2}{2}\Bigr) \int \exp\Bigl(-\frac{1}{2}\|\mu^{1/2}B^{1/2}\xi - \mu^{-1/2}B^{-1/2}\gamma\|^2\Bigr)\, \mathbb{1}(\|\gamma\| \le \mathtt{g})d\gamma \\ &\quad = \mu^{p/2}\exp\Bigl(\frac{\mu\|B^{1/2}\xi\|^2}{2}\Bigr) I\!\!P_{\boldsymbol{\xi}}(\|\mu^{-1/2}B^{1/2}\boldsymbol{\varepsilon} + B^{1/2}\xi\| \le \mathtt{g}/\mu) \end{aligned}$$

where $\boldsymbol{\varepsilon}$ denotes a standard normal vector in $I\!\!R^p$ and $I\!\!P_{\xi}$ means the conditional probability given ξ. Moreover, for any $u \in I\!\!R^p$ and $\mathtt{r} \ge \mathtt{p}^{1/2} + \|u\|$, it holds in view of $I\!\!P(\|B^{1/2}\boldsymbol{\varepsilon}\|^2 > \mathtt{p}) \le 1/2$

$$I\!\!P(\|B^{1/2}\boldsymbol{\varepsilon} - u\| \le \mathtt{r}) \ge I\!\!P(\|B^{1/2}\boldsymbol{\varepsilon}\| \le \sqrt{\mathtt{p}}) \ge 1/2$$

This implies

$$\begin{aligned} & \exp\Bigl(\mu\|B^{1/2}\xi\|^2/2\Bigr)\, \mathbb{1}(\|B\xi\| \le \mathtt{g}/\mu - \sqrt{\mathtt{p}/\mu}) \\ &\quad \le 2\mu^{-p/2}c_p(B) \int \exp\Bigl(\gamma^\top\xi - \frac{1}{2\mu}\|B^{-1/2}\gamma\|^2\Bigr)\, \mathbb{1}(\|\gamma\| \le \mathtt{g})d\gamma \end{aligned}$$

Further, by (SG)

$$\begin{aligned} & c_p(B)\mathbb{E}\int \exp\Bigl(\gamma^\top \xi - \frac{1}{2\mu}\|B^{-1/2}\gamma\|^2\Bigr)\,\mathbb{I}(\|\gamma\| \le \mathtt{g})d\gamma \\ & \le c_p(B)\int \exp\Bigl(\frac{\|\gamma\|^2}{2} - \frac{1}{2\mu}\|B^{-1/2}\gamma\|^2\Bigr)d\gamma \\ & \le \det(B^{-1/2})\det(\mu^{-1}B^{-1} - I_p)^{-1/2} = \mu^{p/2}\det(I_p - \mu B)^{-1/2} \end{aligned}$$

and the initial statement follows. □

The next object of interest is $\mathbb{E}\|\xi\|^r\,\mathbb{I}(\|\xi\| > t)$. Rather useful form of it is

$$\mathbb{E}\|\xi\|^r\,\mathbb{I}(\|\xi\| > t) = \mathbb{P}(\|\xi\| > t)t^r + r\int_t^{+\infty}\mathbb{P}(\|\xi\| > t)t^{r-1}dt$$

With $x_0 = (t - \sqrt{p})^2/2$

$$\int_t^{+\infty}\mathbb{P}(\|\xi\| > t)t^{r-1}dt = \int_{x_0}^{+\infty} 2e^{-x}(\sqrt{2x} + \sqrt{p})^{r-1}d\sqrt{2x} \le \frac{2e^{-x_0}t^{r-1}}{1 - (r-1)\log(x_0)/x_0}$$

Consequently, if $(r-1)\log(x_0)/x_0 \le 1/2$, than

$$\mathbb{E}\|\xi\|^r\,\mathbb{I}(\|\xi\| > t) \le 2e^{-(t-\sqrt{p})^2/2}\left(t^r + 2rt^{r-1}\right) \tag{SGI}$$

Analogically one is able to restrict moment with exponent part and $r\log(x_0)/x_0 + \alpha \le 1/2$

$$\mathbb{E}\|\xi\|^r e^{\alpha\|\xi\|}\,\mathbb{I}(\|\xi\| > t) \le 4e^{-(t-\sqrt{p})^2/2+\alpha t}\left(t^r + rt^{r-1}\right) \tag{SGIexp}$$

Acknowledgement

First of all, I would like to thank Prof. Vladimir Spokoiny for his support and for giving me the opportunity to work independently on an exciting topic. Furthermore, I thank Andzhey Koziuk and Valeriy Avanesov for collaboration, many fruitful discussions and development of new ideas.

This thesis has been written at three different locations: the Humboldt University of Berlin, the Weierstrass Institute for Applied Analysis and Stochastics and Skolkovo Institute of Science and Technology.

Concerning the experimental part of work, I would like to thank my colleagues from Dmitry Dylov's research group at Skoltech for discussions and software engineering, particularly in electrocardiography application so that I have addressed some vital real-world problems.

I am grateful to the members of Research Unit 1735 for funding and conferences organisation.

And certainly I thank my family for their endless support.

Bibliography

Adams, R. P. and MacKay, D. J. C. (2007). Bayesian online changepoint detection. *arXiv*, 0710.3742.

Agueh, M. and Carlier, G. (2011). Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924.

Alfaras, M., Soriano, M. C., and Ortin, S. (2019). A fast machine learning model for ecg-based heartbeat classification and arrhythmia detection. *Frontiers in Physics*, 7:103.

Avanesov, V. and Buzun, N. (2016). Change-point detection in high-dimensional covariance structure. *ArXiv:1610.03783*.

Bassett, D. S., Wymbs, N. F., Porter, M. a., Mucha, P. J., Carlson, J. M., and Grafton, S. T. (2010). Dynamic reconfiguration of human brain networks during learning. *Proceedings of the National Academy of Sciences*, 108(18):7641.

Bauschke, H. H. and Combettes, P. L. (2011). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer Publishing Company, Incorporated, 1st edition.

Bauwens, L., Laurent, S., and Rombouts, J. V. K. (2006). Multivariate GARCH models: a survey. *Journal of Applied Econometrics*, 21(1):79–109.

Bentkus, V. (2003a). A new method for approximations in probability and operator theories. *Lithuanian Mathematical Journal*, 43(4):367–388.

Bentkus, V. (2003b). On the dependence of the berry-esseen bound on dimension. *Journal of Statistical Planning and Inference*.

Biau, G., Bleakley, K., and Mason, D. M. (2016). Long signal change-point detection. *Electron. J. Statist.*, 10(2):2097–2123.

Blazek, R. and Kim, H. (2001). A novel approach to detection of denial-of-service attacks via adaptive sequential and batch-sequential change-point detection methods. In *Proc. of IEEE Workshop on Systems, Man, and Cybernetics Information Assurance*, pages 41–50. ACM Press.

Boucheron S., Lugosi G., M. P. (2013). Concentration inequalities: A nonasymptotic theory of independence. *Oxford University Press*.

Bucher, A. and Dette, H. (2013). Multiplier bootstrap of tail copulas with applications. *Bernoulli*, 19(5A):1655–1687.

Buzun N., A. V. (2017). Bootstrap for change point detection. *arXiv:1710.07285.*

Chen, J. and Gupta, A. (2012). *Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance.* Springer.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013a). Comparison and anti-concentration bounds for maxima of gaussian random vectors. *CeMMAP working papers.*

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013b). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.*, 41(6):2786–2819.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2017). Central limit theorems and bootstrap in high dimensions. *Ann. Probab.*, 45(4):2309–2352.

Şerban, M., Brockwell, A., Lehoczky, J., and Srivastava, S. (2007). Modelling the dynamic dependence structure in multivariate financial time series. *Journal of Time Series Analysis*, 28(5):763–782.

Edwards, D. (2011). On the kantorovich-rubinstein theorem. *Expositiones Mathematicae*, 29(4):387 – 398.

Engle, R. F., Ng, V. K., and Rothschild, M. (1990). Asset pricing with a factor-arch covariance structure. Empirical estimates for treasury bills. *Journal of Econometrics*, 45(1-2):213–237.

Faganeli, J. and Jager, F. (2010). Automatic classification of transient ischaemic and transient non-ischaemic heart-rate related ST segment deviation episodes in ambulatory ECG records. *Physiological Measurement*, 31(3):323–337.

Fotopoulos, S. B., Jandhyala, V. K., and Khapalova, E. (2010). Exact asymptotic distribution of change-point mle for change in the mean of gaussian sequences. *The Annals of Applied Statistics*, pages 1081–1104.

Frick, K., Munk, A., and Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):495–580.

Friston, K. J. (2011). Functional and effective connectivity: A review. *Brain Connectivity*, 1(1):13–36.

Gombay, E. (2000). Sequential change-point detection with likelihood ratios. *Statistics & probability letters*, 49(2):195–204.

Götze, F., Naumov, A., Spokoiny, V., and Ulyanov, V. (2019). Large ball probabilities, gaussian comparison and anti-concentration. *Bernoulli*, 25(4A):2538–2563.

Haccou, P., Meelis, E., and Van De Geer, S. (1987). The likelihood ratio test for the change point problem for exponentially distributed random variables. *Stochastic processes and their applications*, 27:121–139.

Horowitz, J. L. (2003). The bootstrap in econometrics. *Statist. Sci.*, 18(2):211–218.

Hua, J., Zhang, H., Liu, J., Xu, Y., and Guo, F. (2018). Direct arrhythmia classification from compressive ecg signals in wearable health monitoring system. *Journal of Circuits, Systems and Computers*, 27(06):1850088.

Jandhyala, B. and Fotopoulos, S. B. (1999). Capturing the distributional behaviour of the maximum likelihood estimator of a changepoint. *Biometrika*, 86(1):129–140.

Jeremie Bigot, Elsa Cazelles, N. P. (2016). Penalized barycenters in the wasserstein space. *arXiv:1606.01025*.

Jun, T. J., Nguyen, H. M., Kang, D., Kim, D., Kim, D., and Kim, Y.-H. (2018). Ecg arrhythmia classification using a 2-d convolutional neural network. *arXiv*, 1804.06812.

Kawazoe, H., Nakano, Y., Ochi, H., Takagi, M., Hayashi, Y., Uchimura, Y., Tokuyama, T., Watanabe, Y., Matsumura, H., Tomomori, S., Sairaku, A., Suenari, K., Awazu, A., Miwa, Y., Soejima, K., Chayama, K., and Kihara, Y. (2016). Risk stratification of ventricular fibrillation in brugada syndrome using noninvasive scoring methods. *Heart Rhythm*, 13(10):1947 – 1954. Focus Issue: Sudden Death.

Kim, H.-J. (1994). Tests for a change-point in linear regression. *Lecture Notes-Monograph Series*, pages 170–176.

Koltchinskii, V. (2013). *A remark on low rank matrix recovery and noncommutative Bernstein type inequalities*, volume Volume 9 of *Collections*, pages 213–226. Institute of Mathematical Statistics, Beachwood, Ohio, USA.

Kroshnin, A., Suvorikova, A., and Spokoiny, V. (2019). Statistical inference for bures-wasserstein barycenters. *arXiv:1901.00226*.

Kunsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The annals of Statistics*, pages 1217–1241.

Lahiri, S. N. (2013). *Resampling methods for dependent data*. Springer Science & Business Media.

Lai, T. L. (1995). Sequential changepoint detection in quality control and dynamical systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 613–658.

Liu, Y., Zou, C., and Zhang, R. (2008). Empirical likelihood ratio test for a change-point in linear regression model. *Communications in Statistics Theory and Methods*, 37(16):2551–2563.

Max Sommerfeld, A. M. (2016). Inference for empirical wasserstein distances on finite spaces. *arXiv:1610.03287v2*.

Max Sommerfeld, A. M. (2017). Inference for empirical wasserstein distances on finite spaces. *arXiv:1610.03287*.

Mikosch, T., Johansen, S., and Zivot, E. (2009). Handbook of Financial Time Series. *Time*, 468(1996):671–693.

Mikosch, T. and Starica, C. (2004). Changes of structure in financial time series and the garch model. Econometrics 0412003, EconWPA.

Moody, G. B. and Mark, R. G. (2001). The impact of the mit-bih arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):45–50.

Naumov, A., Spokoiny, V., and Ulyanov, V. (2019). Bootstrap confidence sets for spectral projectors of sample covariance. *Probability Theory and Related Fields*.

Perea, J. A. (2013). Sliding windows and persistence:an application of topological methods to signal analysis. *arXiv:1307.6188*.

Petrov, V. V. (1995). Limit theorems of probability theory: sequences of independent random variables. *Oxford, New York*.

Philip de Chazal, O'Dwyer, M., and Reilly, R. B. (2004). Automatic classification of heartbeats using ecg morphology and heartbeat interval features. *IEEE Transactions on Biomedical Engineering*, 51(7):1196–1206.

Polunchenko, A. and Tartakovsky, A. (2011). State-of-the-art in sequential change-point detection. *Methodol. Comput. Appl. Probab.*, 14:649–684.

Quandt, R. E. (1960). Tests of the hypothesis that a linear regression system obeys two separate regimes. *Journal of the American statistical Association*, 55(290):324–330.

Robert E. Gaunt, Alastair Pickett, G. R. (2015). Chi-square approximation by stein's method with application to pearson's statistic. *arXiv:1507.01707*.

Shiryaev, A. (2010). Quickest detection problems: Fifty years later. *Sequential Anal.: Design Methods and Applicat.*, 29:345–385.

Shiryaev, A. N. (1963). On optimum methods in quickest detection problems. *Theory of Probability & Its Applications*, 8(1):22–46.

Shvetsov, N., Buzun, N., and Dylov, D. V. (2020). Unsupervised non-parametric change point detection in quasi-periodic signals. *CoRR*, abs/2002.02717.

Spokoiny, V. (2009). Multiscale local change point detection with applications to value-at-risk. *Ann. of Stat.*

Spokoiny, V. (2012a). Parametric estimation. finite sample theory. *Ann. Stat.*, 40:2877–2909.

Spokoiny, V. (2012b). Penalized maximum likelihood estimation and effective dimension. *eprint arXiv:1205.0498*.

Spokoiny, V. (2016). Nonparametric estimation: parametric view.

Spokoiny, V. and Willrich, N. (2015). Bootstrap tuning in ordered model selection. *ArXiv:1507.05034*.

Spokoiny, V. and Zhilova, M. (2015). Bootstrap confidence sets under model misspecification. *Ann. Statist.*, 43(6):2653–2675.

Sporns, O. (2011). *Networks of the brain.* The MIT Press.

Srivastava, M. and Worsley, K. J. (1986). Likelihood ratio tests for a change in the multivariate normal mean. *Journal of the American Statistical Association*, 81(393):199–204.

Steinerberger, S. (2018). Wasserstein distance, fourier series and applications. *arXiv:1803.08011*.

Thomas Rippl, Axel Munk, A. S. (2015). Limit laws of the empirical wasserstein distance: Gaussian distributions. *arXiv:1507.04090v2*.

Truong, C., Oudre, L., and Vayatis, N. (2018). Selective review of offline change point detection methods. *Signal Processing*, 167.

V. Chernozhukov, D. Chetverikov, K. K. (2014). Gaussian approximation of suprema of empirical processes. *arXiv:1212.6885*.

Wang, H., Zhang, D., and Shin, K. G. (2004). Change-point monitoring for the detection of dos attacks. *Dependable and Secure Computing, IEEE Transactions on*, 1(4):193–208.

Zou, C., Liu, Y., Qin, P., and Wang, Z. (2007). Empirical likelihood ratio test for the change-point problem. *Statistics & probability letters*, 77(4):374–382.

List of publications

1. N. Shvetsov, N. Buzun, D. Dylov. Unsupervised non-parametric change point detection in electrocardiography. 32nd International Conference on Scientific and Statistical Database Management, 2020.

2. N. Buzun. Gaussian approximation for empirical barycenters. arXiv:1904.00891.

3. V. Avanesov, N. Buzun. Change-point detection in high-dimensional covariance structure. Electron. J. Statist. 12 (2018), no. 2, 3254-3294. doi:10.1214/18-EJS1484.

4. N. Buzun, V. Avanesov. Bootstrap for change point detection. arXiv:1710.07285.

5. N. Buzun, V. Spokoiny, A. Suvorikova. Multiscale parametric approach for change point detection. IITP RAS Conference, Information Technology and Systems, 2015.