Re-sampling in instrumental variables regression

Dissertation

zur Erlangung des akademischen Grades
Doktor rerum naturalium
im Fach Mathematik

eingereicht an der Mathematisch-Naturwissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin von
M.Sc. Andzhey Koziuk

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

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

Gutachter:

1. Moritz Jirak
2. Alexey Naumov
3. Vladimir Spokoiny

Tag der mündlichen Prüfung: 13.09.2019

To whom it may concern

Abstract

Instrumental variables regression in the context of a re-sampling is considered. In the work one builds a framework identifying a target of inference. It tries to generalize an idea of a non-parametric regression and motivate instrumental variables regression from a new perspective. The framework assumes a target of estimation to be formed by two factors - an environment and internal model specific structure.

Aside from the framework, the work develops a re-sampling method suited to test linear hypothesis on the target. Particular technical environment and procedure are given and explained cohesively in the introduction and in the body of the work that follows. Specifically, following the work of Spokoiny, Zilova 2015 [20], the writing justifies and applies numerically multiplier bootstrap procedure to construct non-asymptotic confidence intervals for the testing problem. The procedure and underlying statistical toolbox were chosen to account for an issue appearing in the model and overlooked by asymptotic analysis. That is weakness of instrumental variables. The issue, however, is addressed by design of the finite sample approach by Spokoiny 2014 [18] and in that sense the study contributes to econometric theory.

Moreover, in the work a set of mathematical tools crucial for the discussion were developed or in case was needed build. Among others the work covers the topics: classification of instrumental variables, general justification of finite sample approach, namely Wilks expansion, matrix concentration inequalities and a general way to regularize a probability function.

Zusammenfassung

Diese Arbeit behandelt die Instrumentalvariablenregression im Kontext der Stichprobenwiederholung. Es wird ein Rahmen geschaffen, der das Ziel der Inferenz identifiziert. Diese Abhandlung versucht, die Idee der nichtparametrischen Regression zu verallgemeinern und die Instrumentalvariablenregression von einer neuen Perspektive aus zu motivieren. Dabei wird angenommen, dass das Ziel der Schätzung von zwei Faktoren gebildet wird, einer Umgebung und einer zu einem internen Model spezifischen Struktur.

Neben diesem Rahmen entwickelt die Arbeit eine Methode der Stichprobenwiederholung, die geeignet für das Testen einer linearen Hypothese bezüglich der Schätzung des Ziels ist. Die betreffende technische Umgebung und das Verfahren werden im Zusammenhang in der Einleitung und im Hauptteil der folgenden Arbeit erklärt. Insbesondere, aufbauend auf der Arbeit von Spokoiny, Zilova 2015 [20], rechtfertigt und wendet diese Arbeit ein numerisches multiplier-bootstrap Verfahren an, um nicht asymptotische Konfidenzintervalle für den Hypothesentest zu konstruieren. Das Verfahren und das zugrunde liegende statistische Werkzeug wurden so gewählt und angepasst, um ein im Model auftretendes und von asymptotischer Analysis übersehenes Problem zu erklären, das formal als Schwachheit der Instrumentalvariablen bekannt ist. Das angesprochene Problem wird jedoch durch den endlichen Stichprobenansatz von Spokoiny 2014 [18] adressiert und leistet in diesem Sinne einen Beitrag zur ökonometrischen Theorie.

Weiterhin entwickelt diese Arbeit Werkzeuge, die entscheidend beziehungsweise notwendig für die Diskussion sind. Unter anderem werden folgende Themen angesprochen: Klassifizierung von Instrumentalvariablen, eine allgemeine Rechtfertigung für den endlichen Stichprobenansatz (Wilks Entwicklung), Konzentrationsungleichungen von Matrizen und ein allgemeiner Ansatz zur Regularisierung einer Wahrscheinlichkeitsfunktion.

Acknowledgments

I am indebted to the startling stoicism and inspiration coming from Anastasia Tcimbaluk. The work belongs to the rightful owner of my progress. I am grateful to my friend Alexandra Suvorikova, who never failed to support the development. The work owes its shape to the sharp scientific opponent and keen in the art person. The creative input was motivated by the colleagues and friends: Roland Hildebrandt, Egor Klochkov, Alexey Naumov, Alexandra Carpentier, Benjamin Stemper, Franz Besold, Arshak Minasyan, Oleksandr Zadorozny, Aleksandr Gnedko, Denis Voroshchuk, Sergei Dovgal, Maciej Kaczmarek, Denis Borovikov, Nadezda Neiland, Dmitri Ostrovsky, Maxim Panov, Nikita Zivotovsky, Nicolai Baldin, Larisa Adamian, Kirill Efimov, Timur Aslyamov, Igor Traskunov, Alexandr Tarakanov, Aleksey Khlyupin, Konstantin Sinkov, Randolf Altmeyer, Maya Zhilova, Lenka Zbonakova, Petra Burdejova and Nazar Buzun (in no particular order). The work in its entirety would have not been possible without Vladimir Spokoiny. However, the implications of the initiative belong to the future.

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 Science PhD Regulations, published in the Official Gazette of Humboldt-Universität zu Berlin no. 42/2018 on 11/07/2018.

Contents

1 Introduction

Important disclaimer is due as an entry gate to particular and every deeply technical discussion. In the work errors are inherently present, and nothing should be taken as is. Once an error is spotted there is a promise to correct it, once it is hidden it remains. It only makes sense to discuss the material.

Following the work of Spokoiny, Zilova 2015 [20], the current writing justifies and applies numerically multiplier bootstrap procedure in the problem of linear hypothesis testing on a target of inference in the regression with instrumental variables (IV). The re-sampling procedure and underlying statistical toolbox were chosen to account for an issue appearing in the model and overlooked by asymptotic analysis. The issue, however, is addressed by design of the finite sample approach by Spokoiny 2014 [18] and in that sense the study contributes to econometric theory.

Among others in the work one can find an identifying framework of an estimated in the regression target. However, it should be viewed as nothing but an attempt to motivate the model. The connection between the framework and conventionally established instrumental variables regression is not rigorous and thus presents a view on how the model appears. Specifically, under a set of assumptions one can derive the representation of the framework similar to what is called the IV regression (see the equations below [2.11-2.12]). Using the framework as a basis one states formally the hypothesis testing problem and proceeds with the analysis of the accuracy of the re-sampling procedure. It leads to the development and construction of bootstrap confidence intervals, that are further validated numerically.

Moreover, in the work a set of mathematical tools crucial for the discussion was developed or in case was needed build. The appendix, thus, can be viewed as a self-contained study about the related to the work topics. It covers classification of instrumental variables, general justification of finite sample approach, namely Wilks expansion, matrix concentration inequalities and a regularization of probability function in order to address a problem of probability measures comparison.

Outlining the major steps supporting the discussion let us mention crucial topics and their development in the work. A formalization of multiplier bootstrap procedure conclusively leads to a problem of comparison of empirically estimated and expected covariance operators - variability of an observed sample. The section 8.6 addresses the issue and matrix concentration inequalities for the operator norm of a random matrix -

$$\|S\|_{\infty} \stackrel{\text{def}}{=} \sup_{\|\boldsymbol{u}\|=1, \boldsymbol{u} \in I\!R^p} |\boldsymbol{u}^T S \boldsymbol{u}|,$$

with an additive structure $S \stackrel{\text{def}}{=} \sum_{i=1}^{n} S_i$ are considered.

The derivations generally follow techniques from Joel Tropp 2012 [22], supported by the analysis of operator functions present in the works Hansen, Pedersen [8], Effors 2008 [5] and Tropp [21]. The exposition is self-sufficient and the chapter contains required prerequisite results. The central argument in the theory builds on the concavity of the operator function

$$A \to tr\{\exp(H + \log A)\}$$

with respect to ordering on a positive-definite cone with H being fixed self-adjoint operator. This fact is due to and can be found in the paper by Lieb 1973 [11]. The derivation in appendix, however, follows more direct and short argument of Tropp 2012 [21] exploiting joint convexity of relative entropy function.

Another pivotal step in the discussion is a comparison of probability measures or non-classical Berry-Esseen inequalities. In that respect in the section 8.7 an exponential regularization procedure characterizing Kolmogorov distance in $\mathbb{R}^p$ is introduced. The tool in turn allows to study Gaussian approximation problem on the family of centered Euclidean balls (section 8.7.2). The class of the problems has been extensively studied in the literature in the context of a re-sampling justification (see [4, 14, 20]). Particular interest presents the dimensional dependence of the upper bound in the inequalities. The problem has drawn attention of many authors and considerable contributions were made by Nagaev 1976 [13], Senatov 1980 [15], Sazonov 1981 [16] and Götze 1991 [6] who demonstrated the error to be proportional to the dimension on the class of convex sets in $\mathbb{R}^p$. Finally, it was refined to $p^{\frac{1}{4}}$ by Bentkus 2005 [2] who established and holds the best known result. How and weather the dimension can be dropped is still an open problem. The development in the section was devoted to refining the existing techniques addressing the fine problem and facilitate the research on the topic via a new perspective.

On the account of the problem of measures comparison an independent from the current writing contribution was made on the problem of Gaussian comparison. Namely, in the work by Koziuk, Spokoiny 2018 [10] a characterization of difference of multivariate Gaussian measures is found on the family of centered Euclidean balls and, in particular, helps to derive an important for the development bound on the corresponding Kolmogorov distance of the test statistics. In the work the tool, however, is substituted by a more suitable and fine argument made by Götze, F. and Naumov, A. and Spokoiny, V. and Ulyanov, V. [7].

Last but not least, in the section 4.2 the problem of small modeling bias spotted in the thesis of Maya Zilova is considered and addressed by design of an assumption on a structural distributional stability of observations.

The structural outline of the work is as follows: contextual identification for a target of inference is considered and developed. Then the problem of testing of a linear hypothesis in the setting with the help of the bootstrap procedure is introduced. A brief outline of the finite sample theory is given further. The formal setting leads consequently to the problems of Gaussian comparison and approximation. Finally, the theoretical basis is verified numerically and bootstrap log-likelihood test is compared to tests from literature. In the appendix one can find formal derivations of the crucial statements.

2 Contextual identification in non-parametric regression

2.1 Motivation

Unlike non-parametric regression in the thesis a functional dependence between an input $X \in I\!R$ and output $Y \in I\!R$ in the model

$$Y = f(X) + \epsilon$$

where a random error ϵ is independent from Y, X is supposed to exist if and only if an environment identifying the function exists. Formally, the environment considered in the work is represented by the random variables

$$W^k \in I\!R$$

with $\forall k \in [1, K]$, whereas the function is structured as follows, the random error $\epsilon = Y - f(X)$ is assumed to come from an outside of the space formed by the variables $\{W^k\}_{k=\overline{1,K}}$. Formally, it is supposed to be uncorrelated with the variables W^k. Informally, it means that an input/output system is relative strictly to the environment. The idea entails the following system of the equalities

$$\begin{cases} I\!EW^1 (Y - f(X)) = 0, \\ I\!EW^2 (Y - f(X)) = 0, \\ ... \\ I\!EW^K (Y - f(X)) = 0. \end{cases} \tag{2.1}$$

Unless, however, the function comes from a narrow parametric class it is impossible to identify it uniquely based on (2.1). In most general case consider a model specific functional

$$\mathcal{L}\left(\{W^k\}_{k=\overline{1,K}}, Y, X, f\right) = const.$$

Including it in the system one comes at

$$\begin{cases} I\!EW^1 (Y - f(X)) = 0, \\ I\!EW^2 (Y - f(X)) = 0, \\ ... \\ I\!EW^K (Y - f(X)) = 0, \\ \mathcal{L}\left(\{\{W^k\}_{k=\overline{1,K}}, Y, X, f\right) = const. \end{cases} \tag{2.2}$$

A complete analysis of (2.2) with an arbitrary functional closing the system is out of the scope and complexity of the work. However, particular instance of the model leads to a view on instrumental variables regression, that is

$$\mathcal{L}\left(\{W^k\}_{k=\overline{1,K}}, Y, X, f\right) \stackrel{\text{def}}{=} \|f\|^2$$

where $\|\cdot\|$ stands for the Euclidean norm. In the next two sections one exploits effective equivalence of a Hilbert space with a linear vector space to outline specific properties of the solution.

2.2 Identification for independent identically distributed observations

Let $\mathbf{Q} \subset I\!R$ be a compact subset of a real line and random variables are coming respectively from $Y \in I\!R$, $X \in \mathbf{Q}$ and $W^k \in I\!R$ and introduce independent identically distributed observations

$$\left(Y_i, X_i, \{W_i^k\}_{k=\overline{1,K}}\right)_{i=\overline{1,n}} \in \Omega \tag{2.3}$$

from a sample set

$$\Omega \stackrel{\text{def}}{=} I\!R^{\otimes 1+K} \otimes \mathbf{Q}$$

on a probability space

$$(\Omega, \mathcal{F}(\Omega), I\!P).$$

Then assume a system of $K+1$ non-linear equations

$$\begin{cases} I\!EW_1^1 (Y_1 - f(X_1)) = 0, \\ I\!EW_1^2 (Y_1 - f(X_1)) = 0, \\ ... \\ I\!EW_1^K (Y_1 - f(X_1)) = 0, \\ \int_{\mathbf{Q}} f^2(x)dx = const. \end{cases} \tag{2.4}$$

A parametric relaxation of the system introduces a non-parametric bias. For an orthonormal functional basis

$$\{\psi_j(x) : \mathbf{Q} \to I\!R\}_{j=\overline{1,\infty}}$$

define decomposition - parametric approximation - of the function into a series of J summands

$$\widehat{f}(x) \stackrel{\text{def}}{=} \sum_{j=1}^{J} \psi_j(x)\theta_j^* \stackrel{\text{def}}{=} \boldsymbol{\Psi}(x)^T \boldsymbol{\theta}^* \tag{2.5}$$

such that

$$\theta_j^* \stackrel{\text{def}}{=} \int_{\mathbf{Q}} f(x)\psi_j(x)dx$$

and

$$\lim_{J\to\infty} \sum_{j=1}^{J} \psi_j(x)\theta_j^* = f(x).$$

Then a substitution $f(x) \to \widehat{f}(x)$ transforms (2.4) and gives

$$\begin{cases} I\!E W_1^1 \left(Y_1 - \widehat{f}(X_1)\right) = \delta_1, \\ I\!E W_1^2 \left(Y_1 - \widehat{f}(X_1)\right) = \delta_2, \\ ... \\ I\!E W_1^K \left(Y_1 - \widehat{f}(X_1)\right) = \delta_K, \\ \int_{\mathbf{Q}} \widehat{f}^2(x)dx = const, \end{cases} \tag{2.6}$$

with a bias defined as follows

$$\forall k > 0 \quad \delta_k \stackrel{\text{def}}{=} I\!E W_1^k \left(f(X_1) - \widehat{f}(X_1)\right). \tag{2.7}$$

Particular case of (2.6) under parametric assumption ($\delta_k = 0$) and with a single instrument ($K = 1$) can be seen as a popular choice of a model with instrumental variables ([1],[12]). The system is rewritten as

$$\begin{cases} I\!E W_1^1 \left(Y_1 - \widehat{f}(X_1)\right) = 0, \\ \int_{\mathbf{Q}} \widehat{f}^2(x)dx = const, \end{cases} \Rightarrow \begin{cases} \boldsymbol{\eta}_1^{*T} \boldsymbol{\theta} = I\!E W_1^1 Y_1, \\ \sum\limits_{j=1}^{J} \theta_j^2 = const \end{cases} \tag{2.8}$$

with the definition $\boldsymbol{\eta}_1^{*T} \stackrel{\text{def}}{=} \left(I\!E W_1^1 \psi_1(X_1), I\!E W_1^1 \psi_2(X_1), ..., I\!E W_1^1 \psi_J(X_1)\right)$.

Lemma 2.1. *The statements are equivalent.*

1. $\exists! \boldsymbol{\theta}^* \in I\!R^J$ *a solution to (*2.8*).*
2. $\exists! \beta > 0$ *such that* $\boldsymbol{\theta}^* = \beta \boldsymbol{\eta}_1^*$ *is a solution of (*2.8*).*

Proof. A solution to (2.8) can be represented as

$$\boldsymbol{\theta}^* = \alpha Q_\perp \boldsymbol{\eta}_\perp^* + \beta \boldsymbol{\eta}_1^*$$

for a fixed α, β and $Q_\perp \boldsymbol{\eta}_\perp^*$ such that $\boldsymbol{\eta}_\perp^{*T} \boldsymbol{\eta}_1^* = 0$ and $Q_\perp$ is a rotation of an orthogonal to $\boldsymbol{\eta}_1^*$ linear subspace in $I\!R^J$. If the vector $\boldsymbol{\theta}^*$ is unique then α must be zero otherwise there exist infinitely many distinct solutions ($Q_\perp \boldsymbol{\eta}_\perp^* \neq Q'_\perp \boldsymbol{\eta}_\perp^*$). On the other hand for $\alpha = 0$ the vector $\boldsymbol{\theta}^*$ is unique. □

The second statement helps to obtain exact form of a solution to (2.8)

$$\widehat{f}(x) = \beta \sum_{j=1}^{J} \psi_j(x) \eta_{1j}^* = \frac{I\!E W_1^1 Y_1}{\sum\limits_{j=1}^{J} \left(I\!E W_1^1 \psi_j(X_1)\right)^2} \sum_{j=1}^{J} \psi_j(x) I\!E W_1^1 \psi_j(X_1). \tag{2.9}$$

Hence, the correlation of instrumental variable W^1 with features X_1 (note $\eta^*_{1j} = I\!E W^1_1 \psi_j(X_1)$) identifies $\widehat{f}(x)$ (up to a scaling) making the choice of the variable W^1 a crucial task. An empirical relaxation to (2.8) in the literature (see [1],[12]) closely resembles the following system

$$\begin{cases} \boldsymbol{Y}_1 = \boldsymbol{Z}^T \boldsymbol{\pi} \beta + \boldsymbol{\varepsilon}_1, \\ \boldsymbol{Y}_2 = \boldsymbol{Z}^T \boldsymbol{\pi} + \boldsymbol{\varepsilon}_2, \end{cases} \tag{2.10}$$

for $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2 \in I\!R^n$, $\boldsymbol{Z} \in I\!R^{J \times n}$, $\boldsymbol{\pi} \in I\!R^J$, $\beta \in I\!R$ and

$$\begin{pmatrix} \boldsymbol{\varepsilon}_{1,i} \\ \boldsymbol{\varepsilon}_{2,i} \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} \lambda_1 & \rho \\ \rho & \lambda_2 \end{pmatrix}\right)$$

or alternatively (lemma [2.1])

$$\begin{cases} I\!E W^1_1 Y_1 = \boldsymbol{\eta}^{*T}_1 \boldsymbol{\theta}^*, \\ \|\boldsymbol{\eta}^*_1\|^2 = const \end{cases} \quad \Rightarrow \quad \begin{cases} W^1_{1,i} Y_{1,i} = W^1_{1,i} \boldsymbol{\Psi}^T(X_{1,i}) \boldsymbol{\theta} + \varepsilon_{1,i}, \\ \|W_{1,i} \boldsymbol{\Psi}(X_{1,i})\|^2 = W^1_{1,i} \boldsymbol{\Psi}^T(X_{1,i}) \boldsymbol{\theta}/\beta + \varepsilon_{2,i} \end{cases} \tag{2.11}$$

corresponding to the latter system up to a notational convention

$$W^1_{1,i} Y_{1,i} \stackrel{\text{def}}{=} \boldsymbol{Y}_{1,i}, \;\; \|W^1_{1,i} \boldsymbol{\Psi}(X_{1,i})\|^2 \stackrel{\text{def}}{=} \boldsymbol{Y}_{2,i}, \;\; W^1_{1,i} \psi_j(X_{1,i}) \stackrel{\text{def}}{=} \boldsymbol{Z}_{ji} \text{ and } \boldsymbol{\theta} \stackrel{\text{def}}{=} \beta \boldsymbol{\pi}. \tag{2.12}$$

The model was theoretically and numerically investigated in a number of papers (see [1],[12]) and in the article (see 'Numerical') is used as a numerical benchmark.

The lemma [2.1] is a special case example of a more general statement on identification in (2.6).

Lemma 2.2. *The statements are equivalent.*

1. *There exists and unique solution $\widehat{f}(x)$ to the system (2.6).*
2. *A solution to (2.6) is given by $\widehat{f}(x) = \sum\limits_{j=1}^{J} \psi_j(x) \boldsymbol{\theta}^{id}_j$ where $\boldsymbol{\theta}^{id}$ is a solution to an optimization problem*

$$\boldsymbol{\theta}^{id} = \underset{\boldsymbol{x} \in I\!R^J}{\operatorname{argmin}} \|\boldsymbol{x}\|^2 \; s.t. \; \begin{cases} \boldsymbol{\eta}^{*T}_1 \boldsymbol{x} = I\!E W^1_1 Y_1 - \delta_1, \\ \boldsymbol{\eta}^{*T}_2 \boldsymbol{x} = I\!E W^2_1 Y_1 - \delta_2, \\ ..., \\ \boldsymbol{\eta}^{*T}_K \boldsymbol{x} = I\!E W^K_1 Y_1 - \delta_K \end{cases} \tag{2.13}$$

with $\boldsymbol{\eta}^{*T}_k \stackrel{\text{def}}{=} \big(I\!E W^k_1 \psi_1(X_1), I\!E W^k_1 \psi_2(X_1), ..., I\!E W^k_1 \psi_J(X_1)\big)$.

Proof. The model (2.6) turns into

$$\begin{cases} I\!E W^1_1 \left(Y_1 - \widehat{f}(X_1)\right) = \delta_1, \\ I\!E W^2_1 \left(Y_1 - \widehat{f}(X_1)\right) = \delta_2, \\ ... \\ I\!E W^K_1 \left(Y_1 - \widehat{f}(X_1)\right) = \delta_K, \\ \int_{\mathbf{Q}} \widehat{f}^2(x) dx = const, \end{cases} \Rightarrow \begin{cases} \boldsymbol{\eta}^{*T}_1 \boldsymbol{\theta} = I\!E W^1_1 Y_1 - \delta_1, \\ \boldsymbol{\eta}^{*T}_2 \boldsymbol{\theta} = I\!E W^2_1 Y_1 - \delta_2, \\ ..., \\ \boldsymbol{\eta}^{*T}_K \boldsymbol{\theta} = I\!E W^K_1 Y_1 - \delta_K, \\ \sum\limits_{j=1}^{J} \theta^2_j = const. \end{cases} \tag{2.14}$$

A solution to (2.14) is an intersection of a J-sphere and a hyperplane $\mathbb{R}^{J-K}$. If it is unique the hyperplane is a tangent linear subspace to the J-sphere and the optimization procedure (2.13) is solved by definition of the intersection point. Conversely, if there exist a solution to the optimization problem then it is guaranteed to be unique as a solution to a convex problem with linear constraints and by definition $\widehat{f}(x)$ satisfy (2.6). □

An important identification corollary follows from the lemma [2.2].

Theorem 2.3 (Identifiability)**.** *Let $f(x) \in \mathcal{H}[\mathbf{Q}]$ and random variables $\{W^k\}_{k=\overline{1,K}}$ to be such that*

$$\lim_{J\to\infty} \delta_k = 0,$$

then $\exists!\ C_I > 0$ such that functions on a surface of the ball

$$\{\|f\|^2_{\mathbb{L}_2[\mathbf{Q}]} = C_I\}$$

contain a single solution to (2.4).

Proof. In (2.6) identifiability is equivalent to $\int_{\mathbf{Q}} f(x)\boldsymbol{\Psi}(x)dx = \boldsymbol{\theta}^{id}$ with $\|\boldsymbol{\theta}^{id}\| < \infty$ (lemma [2.2]) and the approximation converges $\lim_{J\to\infty} \widehat{f}(x) = f(x)$ in complete metric space $\mathcal{H}[\mathbf{Q}]$ to a solution of

$$\begin{cases} \mathbb{E}W_1^1\left(Y_1 - \widehat{f}(X_1)\right) = \delta_1, \\ \mathbb{E}W_1^2\left(Y_1 - \widehat{f}(X_1)\right) = \delta_2, \\ ... \\ \mathbb{E}W_1^K\left(Y_1 - \widehat{f}(X_1)\right) = \delta_K, \\ \int_{\mathbf{Q}} \widehat{f}^2(x)dx = const, \end{cases} \Rightarrow \begin{cases} \mathbb{E}W_1^1\left(Y_1 - f(X_1)\right) = 0, \\ \mathbb{E}W_1^2\left(Y_1 - f(X_1)\right) = 0, \\ ... \\ \mathbb{E}W_1^K\left(Y_1 - f(X_1)\right) = 0, \\ \int_{\mathbf{Q}} f^2(x)dx = const. \end{cases}$$

Then it inherits the equivalence from the lemma [2.1] and the ball

$$\{\|f\|^2_{\mathbb{L}_2[\mathbf{Q}]} = C_I\}$$

with $C_I \overset{\text{def}}{=} \|\boldsymbol{\theta}^{id}\|^2_{\mathbb{L}_2[\mathbf{Q}]} < \infty$, contains only a single solution.

Assume otherwise, there exists $C \neq C_I$ s.t. $\{\|f\|^2_{\mathbb{L}_2[\mathbf{Q}]} = C\}$ and $\{\|f\|^2_{\mathbb{L}_2[\mathbf{Q}]} = C_I\}$ contain unique solutions, then they must be distinct as $\{\|f\|^2_{\mathbb{L}_2[\mathbf{Q}]} = C\} \cap \{\|f\|^2_{\mathbb{L}_2[\mathbf{Q}]} = C_I\} = \emptyset$. Thus, by definition solutions to a respective parametric relaxations of (2.4) are unique and distinct for any $J > J_0$ greater than some fixed J_0 ($\delta_k^C \neq \delta_k^{C_J}$)

$$\begin{cases} \mathbb{E}W_1^1\left(Y_1 - \widehat{f}(X_1)\right) = \delta_1^C, \\ \mathbb{E}W_1^2\left(Y_1 - \widehat{f}(X_1)\right) = \delta_2^C, \\ ... \\ \mathbb{E}W_1^K\left(Y_1 - \widehat{f}(X_1)\right) = \delta_K^C, \\ \int_{\mathbf{Q}} \widehat{f}^2(x)dx = C, \end{cases} \leftrightarrow \begin{cases} \mathbb{E}W_1^1\left(Y_1 - \widehat{f}(X_1)\right) = \delta_1^{C_I}, \\ \mathbb{E}W_1^2\left(Y_1 - \widehat{f}(X_1)\right) = \delta_2^{C_I}, \\ ... \\ \mathbb{E}W_1^K\left(Y_1 - \widehat{f}(X_1)\right) = \delta_K^{C_I}, \\ \int_{\mathbf{Q}} \widehat{f}^2(x)dx = C_I. \end{cases}$$

Alternatively the lemma [2.2] states that there exist two distinct solutions to the respective optimization problem (2.13). However, in the limit $J \to \infty$ - $\delta_k^{C_I} \to 0$ and $\delta_k^C \to 0$ - optimization objectives coincide contradicting the assumption. □

Remark 2.1. *One can trace in the lemma [2.1] as well as in the theorem [2.3] that a restriction in* $\mathbb{L}_2[\mathbf{Q}]$ *norm in (2.4) enables identifiability. Otherwise an* $\mathbb{L}_q[\mathbf{Q}]$ *norm leads to an ill-posed problem.*

2.3 Identification for independent observations

Redefine

$$\left(Y_i, X_i, \{W_i^k\}_{k=\overline{1,K}}\right)_{i=\overline{1,n}} \in \Omega = I\!R \otimes \mathbf{Q} \otimes I\!R^{\otimes K} \tag{2.15}$$

on a probability space $(\Omega, \mathcal{F}(\Omega), I\!P)$. Let $\mathbf{Q} \subset I\!R$ be a compact, random variables from $Y_i \in I\!R$, $X_i \in \mathbf{Q}$, $W_i^k \in I\!R$ and let the observations identify uniquely a solution to the system

$$\forall i = \overline{1,n} \quad \begin{cases} I\!EW_i^1\left(Y_i - \widehat{f}(X_i)\right) = \delta_1, \\ I\!EW_i^2\left(Y_i - \widehat{f}(X_i)\right) = \delta_2, \\ ... \\ I\!EW_i^K\left(Y_i - \widehat{f}(X_i)\right) = \delta_K, \\ \int_{\mathbf{Q}} \widehat{f}^2(x)dx = C_I. \end{cases} \Rightarrow \forall i = \overline{1,n} \quad \begin{cases} \boldsymbol{\eta}_{1,i}^* \boldsymbol{\eta}_{1,i}^{*T} \boldsymbol{\theta} = \boldsymbol{\eta}_{1,i}^* Z_k^i \\ \boldsymbol{\eta}_{2,i}^* \boldsymbol{\eta}_{2,i}^{*T} \boldsymbol{\theta} = \boldsymbol{\eta}_{2,i}^* Z_k^i \\ ..., \\ \boldsymbol{\eta}_{K,i}^* \boldsymbol{\eta}_{K,i}^{*T} \boldsymbol{\theta} = \boldsymbol{\eta}_{K,i}^* Z_k^i \\ \sum\limits_{j=1}^{J} \theta_j^2 = C_I. \end{cases} \tag{2.16}$$

in the particular case with

$$\boldsymbol{\eta}_{k,i}^{*T} \stackrel{\text{def}}{=} \left(I\!EW_i^k \psi_1(X_i), I\!EW_i^k \psi_2(X_i), ..., I\!EW_i^k \psi_J(X_i)\right) \text{ and } Z_k^i \stackrel{\text{def}}{=} W_i^k Y_i - \delta_k.$$

Identification in non iid case complicates the fact that n is normally larger than J leading to possibly different identifiability scenarios. Distinguish them based on a rank of a matrix

$$r \stackrel{\text{def}}{=} rank\left(\sum_{i=1}^{n}\sum_{k=1}^{K} \boldsymbol{\eta}_{k,i}^* \boldsymbol{\eta}_{k,i}^{*T}\right) = rank\left(\sum_{i=1}^{n}\sum_{k=1}^{K} I\!EW_i^k \boldsymbol{\Psi}(X_i) I\!E\boldsymbol{\Psi}^T(X_i) W_i^k\right). \tag{2.17}$$

Note that the rank and, thus, a solution to [2.16] depends on a sample size n (K is assumed to be fixed). However, there is no prior knowledge of what r corresponds to the identifiable function $f(x) \in \mathcal{H}[\mathbf{Q}]$. Therefore, the discussion requires an agreement on the target of inference.

A way to reconcile uniqueness with the observed dependence is to require the function $f(x) \in \mathcal{H}[\mathbf{Q}]$ and r to be independent from n. The model (2.16) makes sense if it points consistently at a single function independently from a number of observations. Define accordingly a target function.

Definition 2.4. *Assume* $\exists N \leq \infty$ *s.t.* $\forall n \geq N$ *the rank* $r = const$*, then call a function* $\widehat{f}(x) \in \mathcal{H}[\mathbf{Q}]$ *a* ***target*** *if it solves (2.16)* $\forall n \geq N$*.*

Remark 2.2. *In the case of* $n < N$ *a bias between a solution and the target* $n > N$ *has to be considered. However, in the subsequent text it is implicitly assumed that a sample size* $n > N$*.*

Based on the convention [2.4] introduce a classification:

1. Complete model: $\forall J > 0\ \exists N \leq \infty$ s.t. $\forall n > N$ the rank $r = J$.
2. Incomplete model: $\exists J_1 > 0$ s.t $\forall J > J_1, n > 0$ the rank $r \leq J_1$.

Identification in the 'incomplete' model is equivalent to the iid case with the notational change for the number of instruments $K \leftrightarrow J_1$ and respective change of K equations with instruments to the J_1 equations from (2.16). Otherwise 'completeness' of a model allows for a direct inversion of (2.16). Generally a complete model is given without the restriction $\mathcal{F} \overset{\text{def}}{=} \{\|f\|^2_{\mathbb{L}_2[\mathbf{Q}]} = C_I\}$

$$\forall n > N:\ \forall i = \overline{1,n} \quad \begin{cases} I\!E W_i^1 \left(Y_i - \widehat{f}(X_i)\right) = \delta_1, \\ I\!E W_i^2 \left(Y_i - \widehat{f}(X_i)\right) = \delta_2, \\ ... \\ I\!E W_i^K \left(Y_i - \widehat{f}(X_i)\right) = \delta_K. \end{cases} \tag{2.18}$$

In this case a natural objective function for an inference is a quasi log-likelihood

$$L(\boldsymbol{\theta}) \overset{\text{def}}{=} -\frac{1}{2} \sum_{k=1}^{K} \sum_{i=1}^{n} \left(Z_k^i - \boldsymbol{\eta}_k^{iT} \boldsymbol{\theta}\right)^2 \tag{2.19}$$

again with

$$\boldsymbol{\eta}_k^{iT} \overset{\text{def}}{=} \left(W_i^k \psi_1(X_i), W_i^k \psi_2(X_i), ..., W_i^k \psi_J(X_i)\right)$$

and

$$Z_k^i \overset{\text{def}}{=} W_i^k Y_i - \delta_k.$$

3 Testing a linear hypothesis: bootstrap log-likelihood ratio test

Introduce an empirical relaxation of the biased (2.6)

$$\begin{cases} W_i^1 \boldsymbol{\Psi}^T(X_i)\boldsymbol{\theta} = W_i^1 Y_i - \delta_1 + \boldsymbol{\varepsilon}_{1,i}, \\ W_i^2 \boldsymbol{\Psi}^T(X_i)\boldsymbol{\theta} = W_i^2 Y_i - \delta_2 + \boldsymbol{\varepsilon}_{2,i}, \\ ... \\ W_i^K \boldsymbol{\Psi}^T(X_i)\boldsymbol{\theta} = W_i^K Y_i - \delta_K + \boldsymbol{\varepsilon}_{K,i}, \\ \|\boldsymbol{\theta}\|^2 = C_I \end{cases} \tag{3.1}$$

with centered unknown errors $\boldsymbol{\varepsilon}_{k,i}$. Courtesy of the lemma [2.2], a natural objective function is a penalized quasi log-likelihood

$$L(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} -\frac{1}{2} \sum_{k=1}^{K} \sum_{i=1}^{n} \left(Z_k^i - \boldsymbol{\eta}_k^{iT} \boldsymbol{\theta} \right)^2 - \frac{\lambda \|\boldsymbol{\theta}\|^2}{2} \tag{3.2}$$

with

$$\boldsymbol{\eta}_k^{iT} \stackrel{\text{def}}{=} \left(W_i^k \psi_1(X_i), W_i^k \psi_2(X_i), ..., W_i^k \psi_J(X_i) \right) \text{ and } Z_k^i \stackrel{\text{def}}{=} W_i^k Y_i - \delta_k.$$

Maximum likelihood estimator (MLE) and its target are given

$$\widetilde{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \underset{\boldsymbol{\theta} \in I\!R^p}{\operatorname{argmax}} L(\boldsymbol{\theta}) \text{ and } \boldsymbol{\theta}^* \stackrel{\text{def}}{=} \underset{\boldsymbol{\theta} \in I\!R^p}{\operatorname{argmax}} I\!E L(\boldsymbol{\theta}).$$

For a fixed projector $\{\Pi \in I\!R^{J \times J} : I\!R^J \to I\!R^{J_1},\ J_1 \leq J\}$ introduce a linear hypothesis and define a log-likelihood ratio test

$$\begin{aligned} \mathcal{H}_0 &: \boldsymbol{\theta}^* \in \{\Pi\boldsymbol{\theta} = 0\}, \\ \mathcal{H}_1 &: \boldsymbol{\theta}^* \in \{I\!R^p \setminus \{\Pi\boldsymbol{\theta} = 0\}\}, \\ T_{LR} &\stackrel{\text{def}}{=} \sup_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) - \sup_{\boldsymbol{\theta} \in \mathcal{H}_0} L(\boldsymbol{\theta}). \end{aligned} \tag{3.3}$$

The test weakly converges $T_{LR} \to \chi^2_{J_1}$ to chi-square distribution (theorem 4.3) and it is convenient to define a quantile as

$$z_\alpha \ : \ I\!P\left((T_{LR} - J)/\sqrt{J} < z_\alpha \right) \geq 1 - \alpha.$$

It implies that $\lim_{J\to\infty} z_\alpha = \frac{1}{2} erf^{-1}(1-\alpha) \stackrel{\text{def}}{=} \left(\frac{1}{\sqrt{\pi}} \int_0^{1-\alpha} e^{-x^2} dx \right)^{-1}$, with the notation in the formula $(\cdot)^{-1}$ for the inverse of a function. Thus, z_α weakly depends on a dimension in the sense that $\exists C < \infty$ such that $\forall J > 0$, $z_\alpha < C$.

For a set of re-sampling multipliers

$$\{u_i \sim \mathcal{N}(1,1)\}_{i=\overline{1,n}}$$

define bootstrap $L^{\flat}(\boldsymbol{\theta})$ conditional on the original data

$$L^{\flat}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}) u_i \stackrel{\text{def}}{=} \sum_{i=1}^{n} \left(\sum_{k=1}^{K} \left(-\frac{\left(Z_k^i - \boldsymbol{\eta}_k^{iT}\boldsymbol{\theta}\right)^2}{2} - \frac{\lambda \|\boldsymbol{\theta}\|^2}{2nK} \right) \right) u_i.$$

and corresponding bootstrap MLE (bMLE) and its target

$$\widetilde{\boldsymbol{\theta}}^{\flat} \stackrel{\text{def}}{=} \underset{\boldsymbol{\theta} \in I\!R^p}{\operatorname{argmax}} L^{\flat}(\boldsymbol{\theta}) \quad \text{and} \quad \widetilde{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \underset{\boldsymbol{\theta} \in I\!R^p}{\operatorname{argmax}} I\!E L^{\flat}(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta} \in I\!R^p}{\operatorname{argmax}} L(\boldsymbol{\theta}).$$

A centered hypothesis and a respective test are defined accordingly

$$\mathcal{H}_0^{\flat} : \widetilde{\boldsymbol{\theta}} \in \{\Pi(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}) = 0\},$$

$$T_{BLR} \stackrel{\text{def}}{=} \sup_{\boldsymbol{\theta}} L^{\flat}(\boldsymbol{\theta}) - \sup_{\boldsymbol{\theta} \in \mathcal{H}_0^{\flat}} L^{\flat}(\boldsymbol{\theta}). \tag{3.4}$$

And analogously $z_{\alpha}^{\flat} : I\!P^{\flat}\left((T_{BLR} - J)/\sqrt{J} < z_{\alpha}^{\flat} \right) \geq 1 - \alpha$, with the probability

$$I\!P^{\flat}(\cdot) \stackrel{\text{def}}{=} I\!P\left(\cdot \middle| \left(Y_i, X_i, \{W_i^k\}_{k=\overline{1,K}} \right)_{i=\overline{1,n}} \right)$$

relative to the aforementioned sampling and conditional on the data. The theorem [4.4] enables the same convergence in growing dimension $\lim_{J\to\infty} z_{\alpha}^{\flat} = \frac{1}{2} erf^{-1}(1-\alpha) \stackrel{\text{def}}{=} \left(\frac{1}{\sqrt{\pi}} \int_0^{1-\alpha} e^{-x^2} dx \right)^{-1}$ again with the notation in the formula $(\cdot)^{-1}$ for the inverse of a function.

Under parametric assumption - $\forall k > 0$ the non-parametric bias is zero $\delta_k = 0$ - the bootstrap log-likelihood test is empirically attainable and the quantile $z_{\alpha}^{\flat}$ is computed explicitly. On the other hand an unattainable quantile z_{α} calibrates T_{LR}. Between the two exists a direct correspondence. In the section [5] it is demonstrated that $z_{\alpha}^{\flat}$ can be effectively substituted by z_{α}.

Multiplier bootstrap procedure: (3.5)

- Sample $\{u_i \sim \mathcal{N}(1,1)\}_{i=\overline{1,n}}$ computing $z_{\alpha}^{\flat}$ satisfying $I\!P^{\flat}\left((T_{BLR} - J)/\sqrt{J} < z_{\alpha}^{\flat} \right) \geq 1 - \alpha$
- Test $\mathcal{H}_0$ against $\mathcal{H}_1$ using the inequalities

$$\mathcal{H}_0 : T_{LR} < J + z_{\alpha}^{\flat}\sqrt{J} \text{ and } \mathcal{H}_1 : T_{LR} > J + z_{\alpha}^{\flat}\sqrt{J}.$$

The idea is numerically validated in the section 6. Its theoretical justification follows immediately.

4 Finite sample theory

In most general case neither an optimization target $L(\boldsymbol{\theta})$ estimates consistently a modeled structure nor the model is justified to be characterized by an arbitrarily chosen log-likelihood function. In that sense regression with instrumental variables is known to rise concern when chosen instruments are weakly identified (see section [7]) and an inference in the problem might involve a separate testing on weakness which is then resolved separately. Therefore, a specific modeling setting can complicate an original statistical inference of testing problems.

Finite sample approach (Spokoiny 2012 [17]) is an option to construct a generic approach adjusting a modeled structure (2.3) to the log-likelihood function and in case of instrumental variables regression the approach allows to incorporate an unknown nature of instruments into the log-likelihood function.

Finite sample theory: (4.1)

- **[Identifiability]** $\sigma_k^2 \stackrel{\text{def}}{=} I\!\!E \left(Z_k^i - \boldsymbol{\eta}_k^{iT}\boldsymbol{\theta}^*\right)^2$ then $|n \sum\limits_{k=1}^{K} \left(\sigma_k^2 - 1\right) I\!\!E\boldsymbol{\eta}_k^1\boldsymbol{\eta}_k^{1T}| < \lambda$ for $\lambda > 0$
- **[Error/IV]** $\forall k$ an error $Z_k^i - \boldsymbol{\eta}_k^{iT}\boldsymbol{\theta}^*$ is independent from Z_k^i and $\boldsymbol{\eta}_k^{iT}$
- **[Design]** $\sup_j \| \sum\limits_{k=1}^{K} \boldsymbol{D}_0^{-1}\boldsymbol{\eta}_{k,j}^i\| \leq 1/2$ with $\boldsymbol{D}_0^2 = \left(n \sum\limits_{k=1}^{K} I\!\!E\boldsymbol{\eta}_k^1\boldsymbol{\eta}_k^{1T}\right) + \lambda I$
- **[Moments]** $\exists \lambda_0, C_0 < \infty$ s.t. $I\!\!E e^{\lambda_0 \epsilon_i} \leq C_0$ with $\epsilon_i \stackrel{\text{def}}{=} \sum\limits_{k=1}^{K} \left(Z_k^i - I\!\!E Z_k^i\right)$
- **[Target]** $\exists N > 0$ s.t. for a sample size $\forall n \geq N$ and any subset A of the size $|A| \geq N$ of the index set $\{1, 2, 3..., n\}$ the solution to $\sum\limits_{i \in A} \nabla I\!\!E\ell_i(\boldsymbol{\theta}) = 0$ is unique.

Remark 4.1. *The conditions validate the one from Spokoiny 2012 [17] p. 27 section 3.6 on penalized generalized linear model with the link function $g(v) : I\!R \to I\!R$ in the considered case $g(v) \stackrel{\text{def}}{=} v^2$. As for the condition 'Target' see the discussion below.*

4.1 Wilks expansion

The conditions (4.1) give a ground to statistical analysis of a quasi log-likelihood. An objective function assumes concentration of an estimation $\widetilde{\boldsymbol{\theta}}$ around the parameter $\boldsymbol{\theta}^*$. Thus, the log-likelihood behavior dominantly depend on a local approximation in the vicinity of the target. Based on the conditions (4.1) one can derive formally the Wilks expansion (Spokoiny 2012 [17]) for the quasi log-likelihood $L(\boldsymbol{\theta})$.

Theorem 4.1. *Suppose conditions (4.1) are fulfilled. Define a score vector*

$$\boldsymbol{\xi} \stackrel{\text{def}}{=} (\Delta\!\!\!E L(\boldsymbol{\theta}^*))^{-1/2} \nabla L(\boldsymbol{\theta}^*).$$

then it holds with a universal constant $C > 0$

$$\left| \sqrt{2L(\widetilde{\boldsymbol{\theta}}) - 2L(\boldsymbol{\theta}^*)} - \|\boldsymbol{\xi}\| \right| \leq C\,(J + x)\,/\sqrt{Kn}$$

at least with the probability $1 - 5e^{-x}$.

Bootstrap analogue of the Wilks expansion also follows. It was claimed in theorem B.4, section B.2 in Spokoiny, Zhilova 2015 [20].

Theorem 4.2. *Suppose conditions (4.1) are fulfilled. Define a bootstrap score vector*

$$\boldsymbol{\xi}^\flat \stackrel{\text{def}}{=} (\Delta\!\!\!E L(\boldsymbol{\theta}^*))^{-1/2} \nabla \left(L^\flat(\boldsymbol{\theta}^*) - L(\boldsymbol{\theta}^*) \right),$$

then it holds with a universal constant $C > 0$

$$\left| \sqrt{2L^\flat(\widetilde{\boldsymbol{\theta}}^\flat) - 2L(\widetilde{\boldsymbol{\theta}})} - \|\boldsymbol{\xi}^\flat\| \right| \leq C\,(J + x)\,/\sqrt{Kn}$$

at least with the probability $1 - 5e^{-x}$.

Moreover, the log-likelihood statistic follows the same local approximation in the context of hypothesis testing and the T_{LR} satisfies (see appendix - section (8.5)).

Theorem 4.3. *Assume conditions (4.1) are satisfied then with a universal constant $C > 0$*

$$\left| \sqrt{2T_{LR}} - \|\boldsymbol{\xi}^s\| \right| \leq C\,(J + x)\,/\sqrt{Kn}$$

with probability $\geq 1 - Ce^{-x}$. The score vector is defined respectively

$$\boldsymbol{\xi}^s \stackrel{\text{def}}{=} D_0^{-1/2} \left(\nabla_{\Pi\boldsymbol{\theta}} L(\boldsymbol{\theta}^*) - (I - \Pi)\,\Delta\!\!\!E L(\boldsymbol{\theta}^*) \Pi^T \left((I - \Pi)\,\Delta\!\!\!E L(\boldsymbol{\theta}^*)\,(I - \Pi)^T \right)^{-1} \nabla_{(I-\Pi)\boldsymbol{\theta}} L(\boldsymbol{\theta}^*) \right),$$

and Fisher information matrix

$$D_0^2 \stackrel{\text{def}}{=} -\Pi\,\Delta\!\!\!E L(\boldsymbol{\theta}^*)\Pi^T + (I - \Pi)\,\Delta\!\!\!E L(\boldsymbol{\theta}^*)\Pi^T \left((I - \Pi)\,\Delta\!\!\!E L(\boldsymbol{\theta}^*)\,(I - \Pi)^T \right)^{-1} \Pi\,\Delta\!\!\!E L(\boldsymbol{\theta}^*)\,(I - \Pi)^T.$$

Similar statement can be proven in the bootstrap world.

Theorem 4.4. *Assume conditions (4.1) are fulfilled then with probability $\geq 1 - Ce^{-x}$ holds*

$$\left| \sqrt{2T_{BLR}} - \|\boldsymbol{\xi}^s_\flat\| \right| \leq C\,(J + x)\,/\sqrt{Kn},$$

with a universal constant $C > 0$, where a score vector is given

$$\boldsymbol{\xi}^s_\flat \stackrel{\text{def}}{=} D_0^{-1/2} \left(\nabla_{\Pi\boldsymbol{\theta}} L^\flat(\boldsymbol{\theta}^*) - (I - \Pi)\,\Delta\!\!\!E L(\boldsymbol{\theta}^*) \Pi^T \left((I - \Pi)\,\Delta\!\!\!E L(\boldsymbol{\theta}^*)\,(I - \Pi)^T \right)^{-1} \nabla_{(I-\Pi)\boldsymbol{\theta}} L^\flat(\boldsymbol{\theta}^*) \right).$$

The theorem is effectively the same for $L(\boldsymbol{\theta})$ as the re-sampling procedure replicates sufficient for the statement assumptions of a quasi log-likelihood (shown in section 8.3 Appendix).

4.2 Small Modelling Bias

In view of the re-sampling justification a separate discussion deserves a small modeling bias from Spokoiny, Zhilova 2015 [20]. The condition appears from the general way to prove the re-sampling procedure. Namely, for a small error term $\delta > 0$ it is claimed

$$\sup_t |I\!P\left(T_{LR} < t\right) - I\!P\left(T_{BLR} < t\right)| \leq \delta + \|H_0^{-1} B_0^2 H_0^{-1}\|_{op}$$

with the matrices

$$H_0^2 = \sum_{i=1}^n I\!E \nabla \ell_i(\boldsymbol{\theta}^*) \nabla^T \ell_i(\boldsymbol{\theta}^*) \text{ and } B_0^2 = \sum_{i=1}^n \nabla I\!E \ell_i(\boldsymbol{\theta}^*) \nabla^T I\!E \ell_i(\boldsymbol{\theta}^*),$$

where the term $\|H_0^{-1} B_0^2 H_0^{-1}\|_{op}$ is assumed to be of the error order essentially meaning that the deterministic bias is small. However, the assumption

$$\|H_0^{-1} B_0^2 H_0^{-1}\|_{op} \approx \delta$$

appears in the current development only in the form of the condition 'Target' in (4.1). The substitution is possible because of the next lemma.

Theorem 4.5. *Assume that the condition 'Target' (4.1) holds, then* $\|H_0^{-1} B_0^2 H_0^{-1}\|_{op} = 0$.

Proof. By definition of a target of estimation

$$\sum_{i=1}^N \nabla I\!E \ell_i(\boldsymbol{\theta}_0^*) = 0, \text{ and } \nabla I\!E \ell_j(\boldsymbol{\theta}_1^*) + \sum_{i=1}^N \nabla I\!E \ell_i(\boldsymbol{\theta}_1^*) = 0.$$

The condition 'Target' implies that $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0^* = \boldsymbol{\theta}_1^*$. Meaning, that any particular choice of the term $\nabla I\!E \ell_j(\boldsymbol{\theta}^*)$ with the index $j \in \{1, 2, 3..., n\}$ is also zero - $\sum_{i=1}^N \nabla I\!E \ell_i(\boldsymbol{\theta}_0^*) = \sum_{i=1}^N \nabla I\!E \ell_i(\boldsymbol{\theta}_1^*)$. Thus, $B_0^2 = 0$ and the statement follows. □

5 Gaussian comparison and approximation

There are two founding results justifying the re-sampling (3.5) in the thesis. The first - Gaussian comparison - is due to Götze, F. and Naumov, A. and Spokoiny, V. and Ulyanov, V. [7] and was adapted to the needs and notations in the study.

Theorem 5.1. *Assume centered Gaussian vectors* $\boldsymbol{\xi}_0 \sim \mathcal{N}(0, \Sigma_0)$ *and* $\boldsymbol{\xi}_1 \sim \mathcal{N}(0, \Sigma_1)$ *then it holds*

$$\sup_t |I\!P(\|\boldsymbol{\xi}_1\| < t) - I\!P(\|\boldsymbol{\xi}_0\| < t)| \leq \sup_{j=\{0,1\}} C\sqrt{Tr\Sigma_j}\|I - \Sigma_0^{-1}\Sigma_1\|_{op}$$

with a universal constant $C < \infty$*, where* $\|\cdot\|_{op}$ *stands for the operator norm of a matrix.*

The second - Gaussian approximation - has been developed independently in the appendix (section [8.7]). The framework is as follows: define the vectors with an additive structure

$$\boldsymbol{\xi}_1 \stackrel{\text{def}}{=} \sum_{i=1}^n \boldsymbol{\xi}_{1,i}, \text{ and } \boldsymbol{\xi}_0 \stackrel{\text{def}}{=} \sum_{i=1}^n \boldsymbol{\xi}_{0,i}$$

such that $\boldsymbol{\xi}_{1,i_0}$ and $\boldsymbol{\xi}_{0,i_1}$ are independent and sub-Gaussian, that is for $s \in \{0,1\}$ and $\forall \boldsymbol{\gamma} \in I\!R^p$

$$I\!E e^{\boldsymbol{\gamma}^T \boldsymbol{\xi}_{s,i_s}} \leq e^{\frac{\boldsymbol{\gamma}^T \Sigma \boldsymbol{\gamma}}{2}}$$

and with respective covariance matrix $I\!E\boldsymbol{\xi}_1\boldsymbol{\xi}_1^T = I\!E\boldsymbol{\xi}_0\boldsymbol{\xi}_0^T = \Sigma$. Then an adapted version of the theorem [8.27] from the appendix holds.

Theorem 5.2. *Assume the framework above, then it holds*

$$\sup_t |I\!P(\|\boldsymbol{\xi}_1\| < t) - I\!P(\|\boldsymbol{\xi}_0\| < t)| \leq C\left(Tr\Sigma\right)^{3/2}/\sqrt{n}\left(1 + O\left(n^{-1/2}\right)\right)$$

with the universal constant of the order $r_0^3\lambda_{max}(\Sigma)$ *with* r_0 *defined in the Lemma (*8.23*).*

The same statement holds true for the weaker - compared to sub-Gaussian vectors above - condition 'Moments' on the random vectors from the set of conditions enabling finite sample theory in the respective section above (4.1).

Finally, the critical value z_α and the empirical $z_\alpha^\flat$ are compared with a help of matrix concentration inequalities from the section (8.6). The essence of the re-sampling is to translate the closeness of z_α and $z_\alpha^\flat$ into the closeness of the covariance matrices $I\!E\boldsymbol{\xi}^s\boldsymbol{\xi}^{sT} \sim I\!E\boldsymbol{\xi}_\flat^s\boldsymbol{\xi}_\flat^{sT}$ through the Wilks expansion (theorems [4.3,4.4]) and Gaussian comparison result and approximate unknown $\boldsymbol{\xi}_s, \boldsymbol{\xi}_\flat^s$ by the respective Gaussian counterparts. It all amounts to the central statement in the work.

Theorem 5.3. *The parametric model (2.6) in the introduction - $\delta_k = 0$ - under the assumption (4.1) enables*

$$\left| I\!P\left(\left(T_{LR} - J\right)/\sqrt{J} > z_{\alpha}^{\flat} \right) - \alpha \right| \leq C_0 \frac{J^{3/2}}{\sqrt{Kn}} + C_1 \sqrt{\frac{J \log J}{Kn}}$$

with a dominating probability and universal constants $C_0, C_1 < \infty$.

Remark 5.1. *Note that the critical value $z_{\alpha}^{\flat}$ depends on experimental data at hand and is fixed when the expectation is taken with respect to the data generating T_{LR} statistics.*

6 Numerical: conditional and bootstrap log-likelihood ratio tests

Calibrate BLR test on a model from Andrews, Moreira and Stock [1]. In the paper the authors proposed conditional likelihood ratio test (CLR - T_{CLR}) used here as a benchmark. The simulated model reads as

$$\boldsymbol{Y}_1 = \boldsymbol{Z}^T\boldsymbol{\pi}\beta + \boldsymbol{\varepsilon}_1, \tag{6.1}$$

$$\boldsymbol{Y}_2 = \boldsymbol{Z}^T\boldsymbol{\pi} + \boldsymbol{\varepsilon}_2, \tag{6.2}$$

where $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2 \in I\!R^n$, $\boldsymbol{Z} \in I\!R^{J\times n}$, $\boldsymbol{\pi} \in I\!R^J$ and $\beta \in I\!R$ with a matrix $\boldsymbol{Z}_{i,j} \overset{\text{def}}{=} \cos\left(\frac{2\pi ij}{n}\right)$, $\beta^* = 1$ and $\boldsymbol{\pi}_i^* \sim i$ (see section 1). And the hypothesis

$$H_0 : \ \beta^* = \beta_0 \ \text{ against } \ H_1 : \ \beta^* \neq \beta_0$$

on a value of a structural parameter β. For the hypothesis Moreira [12] and later Andrews, Moreira and Stock [1] construct a CLR test based on the two vectors

$$\boldsymbol{S} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-\frac{1}{2}}\boldsymbol{Z}^T\boldsymbol{Y}\boldsymbol{b}(\boldsymbol{b}^T\Omega\boldsymbol{b})^{-\frac{1}{2}}$$

and

$$\boldsymbol{T} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-\frac{1}{2}}\boldsymbol{Z}^T\boldsymbol{Y}\boldsymbol{a}(\boldsymbol{a}^T\Omega^{-1}\boldsymbol{a})^{-\frac{1}{2}}$$

with the notations $\boldsymbol{Y} \overset{\text{def}}{=} [\boldsymbol{Y}_1, \boldsymbol{Y}_2]$, $\boldsymbol{a}^T \overset{\text{def}}{=} (\beta_0, 1)$ and $\boldsymbol{b}^T \overset{\text{def}}{=} (1, -\beta_0)$. $\boldsymbol{S}$ and $\boldsymbol{T}$ are independent and together present sufficient statistics for the model (6.1) with only $\boldsymbol{T}$ depending on instruments' identification, thus conditioning on $\boldsymbol{T}$ and CLR test. Log-likelihood ratio statistics in (6.1) is represented as (see Moreira 2003 [12]) -

$$T_{LR} = \boldsymbol{S}^T\boldsymbol{S} - \boldsymbol{T}^T\boldsymbol{T} + \sqrt{(\boldsymbol{S}^T\boldsymbol{S} - \boldsymbol{T}^T\boldsymbol{T})^2 + 4(\boldsymbol{S}^T\boldsymbol{T})^2}.$$

Additionally Lagrange multiplier and Anderson-Rubin tests are given by

$$T_{LM} = \frac{(\boldsymbol{S}^T\boldsymbol{T})^2}{\boldsymbol{T}^T\boldsymbol{T}},$$

$$T_{AR} = \frac{\boldsymbol{S}^T\boldsymbol{S}}{J}$$

The latter two are known to perform reasonably except for the weakly identified case.
First, correctly specified model is generated for the sample of $n = 200$ and with weak instruments ($\boldsymbol{\pi}^{*T}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{\pi}^* = \frac{C}{n}$). In this case powers of T_{BLR}, T_{CLR} and true T_{LR} tests are drawn on the figure (8.1). To be consistent T_{BLR} is also compared to T_{LM} and T_{AR}. The comparison is given on the figure (8.2) and the data in the case is aggregated in the table (1).
Moreover an important step is to check how robust T_{BLR} to a misspecification of the model. Three special examples are simulated:

1. $\varepsilon_1, \varepsilon_2 \sim Laplace(0, 1)$,
2. $\varepsilon_{1i}, \varepsilon_{2i} \sim \mathcal{N}(0, \frac{5i}{n}\Omega)$,
3. $\varepsilon_{1i}, \varepsilon_{2i} \sim \mathcal{N}(0, (2 + 1.5\sin(6\pi i/n))\Omega)$.

Experiment (1) can be found on the figure (8.3), (8.4) and in the table (2). Numerical study of the experiment (2) with misspecified heteroskedastic error is given on the figure (8.5) and is collected in the table (3). The last experiment is shown on the figure (8.6) and in the table (4).

Remark 6.1. *All the figures and tables are collected in the end of the work.*

7 Strength of instrumental variables

On practice one wants to distinguish instruments based on its strength. For the clarity of exposition the section considers a simplified log-likelihood (2.19) identifying complete model with the Fisher information matrix

$$\boldsymbol{D}_0^2 = -\Delta\!\!\!\mathbb{E} L(\boldsymbol{\theta}^*) = \sum_{i=1}^n \sum_{k=1}^K \mathbb{E}\boldsymbol{\eta}_{ki}^* \boldsymbol{\eta}_{ki}^{*T} = \sum_{i=1}^n \sum_{k=1}^K \mathbb{E} W_i^k \boldsymbol{\Psi}(X_i) \boldsymbol{\Psi}^T(X_i) W_i^k.$$

Weak instrumental variables introduce an unavoidable lower bound on estimation error (lemma [7.1], see the proof in the appendix (8.1)).

Lemma 7.1. *Let conditions (4.1) hold then*

$$\exists N > 0,\ s.t.\ \forall n > N\ \ \mathbb{E}\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2 \geq \frac{C_J}{\sup_{\|\boldsymbol{u}\|=1} \sum\limits_{i=1}^n \sum\limits_{k=1}^K \mathbb{E}\left(\boldsymbol{u}^T \boldsymbol{\Psi}(X_i) W_i^k\right)^2},$$

with a factor $C_J > 0$ depending on dimensionality J.

In view of a hypothesis testing it amounts to an indifference region of a test (see the section 'Numerical').

Classification of Instrumental Variables:

1. Weak instruments

$$\sup_{\|\boldsymbol{u}\|=1} \sum_{i=1}^n \sum_{k=1}^K \mathbb{E}\left(\boldsymbol{u}^T \boldsymbol{\Psi}(X_i) W_i^k\right)^2 \sim K/C \ \text{ and } \ \mathbb{E}\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2 \geq \frac{CC_J}{K}$$

2. Semi-strong instruments with $0 < \alpha < 1$

$$\sup_{\|\boldsymbol{u}\|=1} \sum_{i=1}^n \sum_{k=1}^K \mathbb{E}\left(\boldsymbol{u}^T \boldsymbol{\Psi}(X_i) W_i^k\right)^2 \sim Kn^\alpha/C \ \text{ and } \ \mathbb{E}\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2 \geq \frac{CC_J}{Kn^{1-\alpha}}$$

3. Strong instruments

$$\sup_{\|\boldsymbol{u}\|=1} \sum_{i=1}^n \sum_{k=1}^K \mathbb{E}\left(\boldsymbol{u}^T \boldsymbol{\Psi}(X_i) W_i^k\right)^2 \sim Kn/C \ \text{ and } \ \mathbb{E}\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2 \geq \frac{CC_J}{Kn}$$

Weak instruments effectively cancel an analysis based on a limiting distribution of a test statistics. Therefore, an IV regression requires a treatment under the finite sample assumption.

8 Appendix

8.1 Classification of instrumental variables

Lemma 8.1. *Let conditions (4.1) hold then*

$$\exists N > 0,\ s.t.\ \forall n > N\ \ I\!E\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2 \geq \frac{C_J}{\sup_{\|\boldsymbol{u}\|=1} \sum_{i=1}^{n} \sum_{k=1}^{K} I\!E \left(\boldsymbol{u}^T \boldsymbol{\Psi}(X_i) W_i^k\right)^2},$$

with a factor $C_J > 0$ depending on dimensionality J.

Proof. Fisher expansion (Spokoiny [17]) on the set of dominating probability $I\!P(\Upsilon) > 1 - Ce^{-x}$ is written as

$$\|\boldsymbol{D}_0 \left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right) - \boldsymbol{\xi}\| \leq C(J + x)/\sqrt{n}.$$

with the matrix $\boldsymbol{D}_0^2 = \sum_{i=1}^{n} \sum_{k=1}^{K} I\!E W_i^k \boldsymbol{\Psi}(X_i) \boldsymbol{\Psi}^T(X_i) W_i^k$. Introduce also an inequality

$$\|\boldsymbol{\xi}\|^2 \leq \left(\|\boldsymbol{\xi} - \boldsymbol{D}_0 \left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)\| + \|\boldsymbol{D}_0 \left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)\|\right)^2 \leq 2\|\boldsymbol{\xi} - \boldsymbol{D}_0 \left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)\|^2 + 2\|\boldsymbol{D}_0 \left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)\|^2.$$

It gives

$$I\!E\|\boldsymbol{D}_0(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2 \geq I\!E_\Upsilon\|\boldsymbol{D}_0(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2 \geq \frac{1}{2} I\!E_\Upsilon\|\boldsymbol{\xi}\|^2 - C(J + x)^2/n$$

and the inquired statement follows with $N > 0$ s.t. $\inf_N\{\frac{1}{2} I\!E_\Upsilon\|\boldsymbol{\xi}\|^2 - C(J+x)^2/N > 0\}$ and a constant $C_J \stackrel{\text{def}}{=} \frac{1}{2} I\!E_\Upsilon\|\boldsymbol{\xi}\|^2 - C(J + x)^2/N$. □

8.2 Non-parametric bias

The bias term - $\boldsymbol{b}_J \stackrel{\text{def}}{=} \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|$ - between parametric and non-parametric functions from the model in chapter 2 is quantified in the lemma.

Lemma 8.2. *Assume that basis functions ψ_j follow -*

$$\overset{(s_2)}{j} \leq j^{s_2} \psi_j$$

with some positive constants s_1, s_2, s_3. Let $f(x)$ be s.t. $f \in \mathcal{S}^s$ where

$$\mathcal{S}^s \stackrel{\text{def}}{=} \{f : ||D^s g|| \leq C_f\},$$

with the notation $D^s(\cdot) \stackrel{\text{def}}{=} \frac{\partial^{s_1}}{\partial x} \frac{\partial^{s_2}}{\partial w} \frac{\partial^{s_3}}{\partial z}(\cdot)$ for $s_{i=1:3} \leq s$. Then bias satisfies

$$\boldsymbol{b}_J = \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \leq C_f J^{-s}.$$

***Proof* :**
Straightforwardly using smoothness of functions from a Sobolev class it can be argued for $s < \infty$ that
-

$$J^s\|\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\| = J^s\|\sum_j \theta_j^*\psi_j - \sum_{j\le J}\theta_j^*\psi_j\| \le \|\sum_j \theta_j^* j^s\psi_j\| \le \|D^s f\| \le C_f$$

and the result follows.
End of Proof

8.3 Re-sampled quasi log-likelihood

A basis for the statistical investigation of a re-sampled log-likelihood builds on the probabilistic equivalence with an original quasi log-likelihood. In the section one also uses notations from Spokoiny 2012 [17].

An analogue to $(\boldsymbol{\mathcal{ED}_0})$ condition for re-sampled log-likelihood - will be referred to as $(\boldsymbol{\mathcal{EDB}_0})$ - readily follows from normality of re-sampling weights $\{u_i\}_{i=\overline{1,n}}$.

Lemma 8.3. *Suppose that conditions (*4.1*) are justified, then there exist a positive symmetric matrix* V_0 *and constants* $\nu_0 \ge 1$ *and* $g \ge 0$ *such that* $\mathrm{Var}\left(\nabla\zeta(\boldsymbol{\theta}^*)\right) \le V_0^2$ *and*

$$\forall\|\boldsymbol{\gamma}\| = 1 \quad \log I\!\!E^\flat \exp\left(\lambda\frac{\boldsymbol{\gamma}^T\nabla\zeta^\flat(\boldsymbol{\theta}^*)}{\|V_0\boldsymbol{\gamma}\|}\right) \le \frac{\nu_0^2\lambda^2}{2},\ |\lambda| \le g$$

with probability $\ge 1 - e^{-x}$.

Proof. Define a vector

$$\boldsymbol{s}_i \stackrel{\mathrm{def}}{=} \frac{\nabla\ell_i(\boldsymbol{\theta}^*)}{\|V_0\boldsymbol{\gamma}\|},$$

then using normality of re-sampling weights u_i rewrite

$$\log I\!\!E^\flat \exp\left(\lambda\frac{\boldsymbol{\gamma}^T\nabla\zeta^\flat(\boldsymbol{\theta}^*)}{\|V_0\boldsymbol{\gamma}\|}\right) = \log I\!\!E^\flat \exp\left(\frac{\lambda}{\|V_0\boldsymbol{\gamma}\|}\left(\boldsymbol{\gamma}^T\sum_{i=1}^n \nabla\ell(Y_i,\boldsymbol{\theta}^*)(u_i-1)\right)\right) =$$

$$= \log I\!\!E^\flat \exp\left(\sum_{i=1}^n \lambda\gamma^T\boldsymbol{s}_i(u_i-1)\right) \le \frac{\nu_0^2\lambda^2}{2}\sum_{i=1}^n\left(\gamma^T\boldsymbol{s}_i\right)^2 \le \frac{\nu_0'^2\lambda^2}{2},$$

where $\nu_0' \stackrel{\mathrm{def}}{=} \sqrt{\nu_0^2 + C\delta}$ for some positive constant $C > 0$ and small δ. The last inequality is derived using $\sum\limits_{i=1}^n\left(\gamma^T\boldsymbol{s}_i\right)^2 \le 1 + C\delta$ from definition of V_0 and matrix concentration inequality (thm [8.20]). □

Re-sampling analogue to the condition $(\boldsymbol{\mathcal{ED}_2})$ (Spokoiny 2012 [17]) also follows.

Lemma 8.4. *Let conditions (4.1) hold true then there exist a positive value* $\omega_1(r) \stackrel{\text{def}}{=} \sqrt{4\nu_0^2\omega^2 x + \frac{2C_\delta^2(r)}{n}}$ *and for each* $r \geq 0$*, a constant* $g(r) \geq 0$ *such that it holds for any* $\boldsymbol{v} \in \Upsilon(r)$

$$\log I\!\!E^\flat \exp\left(\frac{\lambda}{\omega_1(r)} \frac{\boldsymbol{\gamma}_1^T \nabla^2 \zeta^\flat(\boldsymbol{\theta}^*)\boldsymbol{\gamma}_2}{\|\boldsymbol{D}_0\boldsymbol{\gamma}_1\| \, \|\boldsymbol{D}_0\boldsymbol{\gamma}_2\|}\right) \leq \frac{\nu_0^2\lambda^2}{2}, \ |\lambda| \leq g(r).$$

Proof. Here it is convenient to reformulate conditions $(\boldsymbol{\mathcal{L}})$ and $(\boldsymbol{\mathcal{ED}_2})$. Bound on deterministic covariance structure can be rewritten as

$$\|\boldsymbol{D}_0^{-1}(\boldsymbol{D}^2(\boldsymbol{\theta}) - \boldsymbol{D}_0^2)\boldsymbol{D}_0^{-1}\| = \|\boldsymbol{D}_0^{-1}(-\sum_{i=1}^n \nabla^2 I\!\!E\ell(Y_i, \boldsymbol{\theta}) - \boldsymbol{D}_0^2)\boldsymbol{D}_0^{-1}\| =$$

$$= \|\sum_{i=1}^n (\boldsymbol{D}_0^{-1}\nabla^2 I\!\!E\ell(Y_i, \boldsymbol{\theta})\boldsymbol{D}_0^{-1} + \frac{I_p}{n})\| = n\|\boldsymbol{D}_0^{-1}\nabla^2 I\!\!E\ell(Y_i, \boldsymbol{\theta})\boldsymbol{D}_0^{-1} + \frac{I_p}{n}\| \leq \delta(r),$$

and it follows

$$\|\boldsymbol{D}_0^{-1}\nabla^2 I\!\!E\ell(Y_i, \boldsymbol{\theta})\boldsymbol{D}_0^{-1}\| \leq \frac{C_\delta(r)}{n}.$$

Next, rewrite $(\boldsymbol{\mathcal{ED}_2})$ mostly in the same fashion, so that it is capable to quantify $\boldsymbol{D}_0^{-1}\nabla^2\zeta_i(\boldsymbol{\theta})\boldsymbol{D}_0^{-1}$. It follows

$$\log I\!\!E \exp\{\frac{\lambda}{\omega} \frac{\boldsymbol{\gamma}_1^T \nabla^2 \zeta(\boldsymbol{\theta})\boldsymbol{\gamma}_2}{\|\boldsymbol{D}_0\boldsymbol{\gamma}_1\| \, \|\boldsymbol{D}_0\boldsymbol{\gamma}_2\|}\} = \log I\!\!E \exp\{\frac{\lambda}{\omega} \sum_{i=1}^n \frac{\boldsymbol{\gamma}_1^T \nabla^2 \zeta_i(\boldsymbol{\theta})\boldsymbol{\gamma}_2}{\|\boldsymbol{D}_0\boldsymbol{\gamma}_1\| \, \|\boldsymbol{D}_0\boldsymbol{\gamma}_2\|}\} =$$

$$= n \log I\!\!E \exp\{\frac{\lambda}{\omega} \frac{\boldsymbol{\gamma}_1^T \nabla^2 \zeta_i(\boldsymbol{\theta})\boldsymbol{\gamma}_2}{\|\boldsymbol{D}_0\boldsymbol{\gamma}_1\| \, \|\boldsymbol{D}_0\boldsymbol{\gamma}_2\|}\},$$

where $\zeta_i(\boldsymbol{v}) = \ell(Y_i, \boldsymbol{\theta}) - I\!\!E\ell(Y_i, \boldsymbol{\theta})$. This means that component-wise $(\boldsymbol{\mathcal{ED}_2})$ condition holds true, namely that

$$\sup_{\gamma_1, \gamma_2 \in I\!R^p} \log I\!\!E \exp\{\frac{\lambda}{\omega} \frac{\boldsymbol{\gamma}_1^T \nabla^2 \zeta_i(\boldsymbol{\theta})\boldsymbol{\gamma}_2}{\|\boldsymbol{D}_0\boldsymbol{\gamma}_1\| \, \|\boldsymbol{D}_0\boldsymbol{\gamma}_2\|}\} \leq \frac{\nu_0^2\lambda^2}{2n}, \ |\lambda| \leq g(r).$$

The two constitute the substance of the proof. Define complementary variables $s_i \stackrel{\text{def}}{=} \frac{\boldsymbol{\gamma}_1^T \nabla^2 \ell(Y_i, \boldsymbol{\theta})\boldsymbol{\gamma}_2}{\|\boldsymbol{D}_0\boldsymbol{\gamma}_1\| \, \|\boldsymbol{D}_0\boldsymbol{\gamma}_2\|}$ and rewrite

$$\log I\!\!E^\flat \exp\{\frac{\lambda}{\omega_1(r)} \frac{\boldsymbol{\gamma}_1^T \nabla^2 \zeta^\flat(\boldsymbol{\theta})\boldsymbol{\gamma}_2}{\|\boldsymbol{D}_0\boldsymbol{\gamma}_1\| \, \|\boldsymbol{D}_0\boldsymbol{\gamma}_2\|}\} = \log I\!\!E^\flat \exp\{\sum_{i=1}^n \frac{\lambda}{\omega_1(r)} s_i(u_i - 1)\} \leq \frac{\nu_0^2\lambda^2}{2\omega_1^2(r)} \sum_{i=1}^n s_i^2.$$

To claim the statement it is sufficient to limit sum $\sum_{i=1}^n s_i^2$. Once again rewrite this sum using mentioned above $(\boldsymbol{\mathcal{L}})$ -

$$\sum_{i=1}^n s_i^2 = \sum_{i=1}^n \left(\frac{\boldsymbol{\gamma}_1^T \nabla^2 \zeta_i(\boldsymbol{\theta})\boldsymbol{\gamma}_2}{\|\boldsymbol{D}_0\boldsymbol{\gamma}_1\| \, \|\boldsymbol{D}_0\boldsymbol{\gamma}_2\|} + \frac{\boldsymbol{\gamma}_1^T \nabla^2 I\!\!E\ell(Y_i, \boldsymbol{\theta})\boldsymbol{\gamma}_2}{\|\boldsymbol{D}_0\boldsymbol{\gamma}_1\| \, \|\boldsymbol{D}_0\boldsymbol{\gamma}_2\|}\right)^2 \leq$$

$$\leq 2\sum_{i=1}^n \left(\frac{\boldsymbol{\gamma}_1^T \nabla^2 \zeta_i(\boldsymbol{v})\boldsymbol{\gamma}_2}{\|\boldsymbol{D}_0\boldsymbol{\gamma}_1\| \, \|\boldsymbol{D}_0\boldsymbol{\gamma}_2\|}\right)^2 + 2\sum_{i=1}^n \left(\frac{\boldsymbol{\gamma}_1^T \nabla^2 I\!\!E\ell(Y_i, \boldsymbol{\theta})\boldsymbol{\gamma}_2}{\|\boldsymbol{D}_0\boldsymbol{\gamma}_1\| \, \|\boldsymbol{D}_0\boldsymbol{\gamma}_2\|}\right)^2 \leq$$

$$\leq 2\sum_{i=1}^{n}\left(\frac{\gamma_1^T\nabla^2\zeta_i(\boldsymbol{\theta})\gamma_2}{\|\boldsymbol{D}_0\gamma_1\|\,\|\boldsymbol{D}_0\gamma_2\|}\right)^2+\frac{2C_\delta^2(r)}{n}$$

The left term in the sum is bounded under $(\boldsymbol{\mathcal{ED}_2})$ and Markov exponential inequality

$$I\!P\left(\frac{\gamma_1^T\nabla^2\zeta_i(\boldsymbol{\theta})\gamma_2}{\|\boldsymbol{D}_0\gamma_1\|\,\|\boldsymbol{D}_0\gamma_2\|}\leq t\right)\leq I\!E\exp\{\frac{\lambda'\gamma_1^T\nabla^2\zeta_i(\boldsymbol{\theta})\gamma_2}{\omega\|\boldsymbol{D}_0\gamma_1\|\,\|\boldsymbol{D}_0\gamma_2\|}-\frac{\lambda' t}{\omega}\}\leq$$

$$\leq\exp\{\frac{\nu_0^2\lambda'^2}{2n}-\frac{\lambda' t}{\omega}\}\leq\exp\{-\frac{t^2 n}{2\nu_0^2\omega^2}\},$$

and

$$I\!P\left(\frac{\gamma_1^T\nabla^2\zeta_i(\boldsymbol{\theta})\gamma_2}{\|\boldsymbol{D}_0\gamma_1\|\,\|\boldsymbol{D}_0\gamma_2\|}\leq\nu_0\omega\sqrt{\frac{2x}{n}}\right)\leq e^{-x}.$$

Therefore it holds

$$\sum_{i=1}^{n}s_i^2\leq 4\nu_0^2\omega^2 x+\frac{2C_\delta^2(r)}{n},$$

and now we can see that controlling $\omega_1(r)$ in the way -

$$\omega_1(r)\stackrel{\text{def}}{=}\sqrt{4\nu_0^2\omega^2 x+\frac{2C_\delta^2(r)}{n}},$$

justifies inquired in the theorem inequality. □

The lemma in turn helps to bound a stochastic part of re-sampled log-likelihood. The demonstrated equivalence allows to translate statements for log-likelihood into the re-sampled counterpart.

A result requiring only $(\boldsymbol{\mathcal{ED}_0})$ is the deviation bound on $\|\boldsymbol{\xi}\|$ (Spokoiny Zhilova 2013 [19]). In the work of Spokoiny [17] it has been proven.

Theorem 8.5. *Let $(\boldsymbol{\mathcal{ED}_0})$ is fulfilled, then for $g\geq\sqrt{2tr(\boldsymbol{D}_0^{-1}V_0^2\boldsymbol{D}_0^{-1})}$, where $V_0^2\geq\operatorname{Var}\nabla\zeta(\boldsymbol{\theta}^*)$ it holds:*

$$I\!P(\|\boldsymbol{\xi}\|^2\geq\mathfrak{z}^2(x,\boldsymbol{D}_0^{-1}V_0^2\boldsymbol{D}_0^{-1}))\leq 2e^{-x}+8.4e^{-x_c},$$

for function $\mathfrak{z}^2(x,\boldsymbol{D}_0^{-1}V_0^2\boldsymbol{D}_0^{-1})$ and small positive constant x_c (thm 8.6).

Let us claim the same for $\|\boldsymbol{\xi}^\flat\|$ using the lemma [8.3].

Theorem 8.6. *Let $(\boldsymbol{\mathcal{EDB}_0})$ is fulfilled, then for $g\geq\sqrt{1+\frac{C\delta}{g}}\sqrt{2tr(\boldsymbol{D}_0^{-1}V_0^2\boldsymbol{D}_0^{-1})}$, where $V_0^2\geq\operatorname{Var}\{\nabla\zeta(\boldsymbol{\theta}^*)\}$ it holds with dominating probability:*

$$I\!P^\flat(\|\boldsymbol{\xi}^\flat\|^2\geq\mathfrak{z}^2(x,\boldsymbol{D}_0^{-1}V_0^2\boldsymbol{D}_0^{-1}))\leq 2e^{-x}+8.4e^{-x_{c_1}},$$

for function $\mathfrak{z}^2(x,\boldsymbol{D}_0^{-1}V_0^2\boldsymbol{D}_0^{-1})$ and small positive constant x_{c_1}, specified below.

The function $\mathfrak{z}(x, X)$, where $x \in I\!R$ and $X \in I\!R^{p\times p}$, is given by the following formula

$$\mathfrak{z}^2(x, X) \stackrel{\text{def}}{=} \begin{cases} tr(X^2) + \sqrt{8tr(X^4)x},\ x \le \frac{\sqrt{2tr(X^4)}}{18\lambda_{max}(X^2)} \\ tr(X^2) + 6x\lambda_{max}(X^2),\ \frac{\sqrt{2tr(X^4)}}{18\lambda_{max}(X^2)} \le x \le x_c \\ |z_c + 2(x - x_c)/g_c|^2\lambda_{max}(X^2),\ x \ge x_c, \end{cases}$$

where in tern numerical constants x_c, z_c, g_c are defined as follows

$$2x_c \stackrel{\text{def}}{=} 2z_c^2/3 + \log\det(I_p - 2X^2/3\lambda_{max}(X^2))$$
$$z_c^2 \stackrel{\text{def}}{=} (9g^2/4 - 3tr(X^2)/2)/\lambda_{max}(X^2)$$
$$g_c \stackrel{\text{def}}{=} \sqrt{g^2 - 2tr(X^2)/3}/\sqrt{\lambda_{max}(X^2)}.$$

This technical result is used extensively for the proof of squared-root Wilks result.

Another key result is that $(\boldsymbol{\mathcal{ED}_2})$ condition justifies a bound on stochastic part of log-likelihood function. The fact formally is stated in the next theorem.

Theorem 8.7. *Let $(\boldsymbol{\mathcal{ED}_2})$ and $(\boldsymbol{\mathcal{I}})$ hold then $\forall \boldsymbol{v} \in I\!R^p$ following inequality is fulfilled*

$$\|\boldsymbol{D}_0^{-1}\nabla\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\| \le 6\nu_0\omega\mathfrak{Z}(x)r.$$

Also $\forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in I\!R^p$ holds

$$\|\boldsymbol{D}_0^{-1}\nabla\zeta(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\| \le 12\nu_0\omega\mathfrak{Z}(x)r,$$

where $\mathfrak{Z}(x)$ is defined as

$$\mathfrak{Z}(x) \stackrel{\text{def}}{=} \begin{cases} \mathbb{H}_1 + \sqrt{2x} + g^{-1}(g^{-2}x + 1)\mathbb{H}_2, & \\ \sqrt{\mathbb{H}_2 + 2x}, & if\ \mathbb{H}_2 + 2x \le g^2, \\ g^{-1}x + \frac{1}{2}(g^{-1}\mathbb{H}_2 + g), & if\ \mathbb{H}_2 + 2x \ge g^2. \end{cases}$$

Here $\mathbb{H}_2 = 4p$ and $\mathbb{H}_1 = 2p^{\frac{1}{2}}$; see theorem A.15 in [17].

Let us provide a proof of that statement.

Proof. Consider quantity $\|\boldsymbol{D}_0^{-1}\nabla\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\|$ and rewrite it as $\|\boldsymbol{D}_0^{-1}\nabla^2\zeta(\boldsymbol{\theta}')\boldsymbol{D}_0\boldsymbol{D}_0^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|$, then introducing vector $Y(s) \stackrel{\text{def}}{=} \boldsymbol{D}_0^{-1}\nabla\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$, where $s \stackrel{\text{def}}{=} \boldsymbol{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ we can see that $\nabla_s Y(s) = \boldsymbol{D}_0^{-1}\nabla^2\zeta(\boldsymbol{\theta}')\boldsymbol{D}_0^{-1}$ and from $(\boldsymbol{\mathcal{ED}_2})$, which holds for $\nabla_s Y(s)$, we have for stochastic process $Y(s)$ by an argument from Spokoiny [17]

$$\sup_{\boldsymbol{\theta}\in\Upsilon(r_0)} \|\boldsymbol{D}_0^{-1}\nabla\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\| \le 6\nu_0\mathfrak{Z}(x)\omega r,$$

which is generally drawn from empirical processes theory. Furthermore, one can use triangle inequality to generalize result

$$\|\boldsymbol{D}_0^{-1}\nabla\zeta(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\| \le \|\boldsymbol{D}_0^{-1}\nabla\zeta(\boldsymbol{\theta}_1, \boldsymbol{\theta}^*)\| + \|\boldsymbol{D}_0^{-1}\nabla\zeta(\boldsymbol{\theta}_2, \boldsymbol{\theta}^*)\| \le 12\nu_0\mathfrak{Z}(x)\omega r,$$

and finalize the proof. □

Once again it is obviously translated using ($\boldsymbol{\mathcal{EDB}_2}$), justified by the lemma 8.4. Therefore, formally one comes at the theorem.

Theorem 8.8. *Let ($\boldsymbol{\mathcal{EDB}_2}$) hold true then $\forall \boldsymbol{\theta} \in \mathbb{R}^p$ following inequality is fulfilled*

$$\|\boldsymbol{D}_0^{-1} \nabla \zeta^{\flat}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\| \leq 6\nu_0 \omega_1(r) \mathfrak{Z}(x) r,$$

and $\forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^p$ also holds

$$\|\boldsymbol{D}_0^{-1} \nabla \zeta^{\flat}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\| \leq 12\nu_0 \omega_1(r) \mathfrak{Z}(x) r.$$

8.4 Concentration of MLE and bMLE

This is technical part of the paper and thus the full version of theorems without unnecessary simplifications is presented. An important result in the section (4) is formulated by the theorem below.

Theorem 8.9. *Let conditions ($\boldsymbol{\mathcal{L}_0}$), ($\boldsymbol{\mathcal{L}}$), ($\boldsymbol{\mathcal{ED}_0}$), ($\boldsymbol{\mathcal{ED}_2}$), ($\boldsymbol{\mathcal{I}}$), (($\boldsymbol{\mathcal{EB}}$)), ($\boldsymbol{\mathcal{SMB}}$) and ($\boldsymbol{\mathcal{IB}}$) hold true, then for r_0 such that following inequalities are fulfilled simultaneously*

$$\begin{cases} b(r) r \geq 2\mathfrak{z}(x, B) \vee 4\sqrt{tr(\sum\limits_{i=1}^{n} \mathbb{E}\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T)} + 24\nu_0 \omega \mathfrak{Z}(x + \log \frac{2r}{r_0}), \\ b(r) r \geq 3\mathfrak{z}(x, B) + 12\nu_0 \mathfrak{Z}(x + \log \frac{2r}{r_0})(\omega + \omega_1(r)), \end{cases}$$

where $B \stackrel{\text{def}}{=} \boldsymbol{D}_0^{-1} \operatorname{Var}\{\nabla L(\boldsymbol{v}^*)\} \boldsymbol{D}_0^{-1}$ following inequalities are fulfilled

1. $\mathbb{P}(\widetilde{\boldsymbol{\theta}} \notin \Upsilon(r_0)) \leq C_1 e^{-x}$,
2. $\mathbb{P}^*(\widetilde{\boldsymbol{\theta}}^{\flat} \notin \Upsilon(r_0)) \leq C_2 e^{-x}$.

Up to constants and quantities smaller than $\sqrt{\frac{p}{n}}$ the concentration radii follows $r_0 \sim \sqrt{p + x}$

We will utilize uniform version of local deviation bound on stochastic processes $\nabla \zeta(\boldsymbol{\theta})$ and $\nabla \zeta^{\flat}(\boldsymbol{\theta})$ from theorems 8.7 and 8.8 and also bounds on $\|\boldsymbol{\xi}\|$ outlined in previous section to prove this result.

Proof. Let us list the key facts needed in the proof in an informal fashion to get an idea of the background required.

1. ($\boldsymbol{\mathcal{L}}$) condition to bound deterministic part of log-likelihood function
   $-2\mathbb{E}L(\boldsymbol{v}, \boldsymbol{v}^*) \geq b(r) r^2$
2. Uniform bound on stochastic processes $\nabla \zeta(\boldsymbol{v})$ and $\nabla \zeta^{\flat}(\boldsymbol{v})$
   $|\zeta(\boldsymbol{v}, \boldsymbol{v}^*) - (\boldsymbol{v} - \boldsymbol{v}^*) \nabla \zeta(\boldsymbol{v}^*)| \leq \rho(x, r) r$
   $|\zeta^{\flat}(\boldsymbol{v}, \boldsymbol{v}^*) - (\boldsymbol{v} - \boldsymbol{v}^*) \nabla \zeta^{\flat}(\boldsymbol{v}^*)| \leq \rho_1(x, r) r$

3. Deviation bound on $\|\boldsymbol{\xi}\|$ and $\|\boldsymbol{\xi}^\flat\|$
   $\|\boldsymbol{\xi}\| \geq \mathfrak{z}(x, B)$
   $\|\boldsymbol{\xi}^\flat\| \geq \mathfrak{z}(x, B)$

These are sufficient to prove results number one and two in the theorem. Let us divide the proof in parts accordingly to the results provided in the statement.

1.*Real world concentration of MLE*

Notice that an inequality $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \geq 0$ always hold and thus by definition binds MLE $\widetilde{\boldsymbol{\theta}}$ structurally to $\boldsymbol{\theta}^*$. So, if one justifies that there exist minimum r_0 such that for $r \geq r_0$ the property breaks than one can claim that $\widetilde{\boldsymbol{\theta}}$ concentrates within $\Upsilon(r_0)$. Therefore, one need to have a uniform bound on log-likelihood function. Spokoiny [17] has proven that with dominating probability -

$$|\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}^*)\zeta(\boldsymbol{\theta}^*)| \leq \rho(x, r)r,$$

where $\rho(x, r) \stackrel{\text{def}}{=} 6\nu_0\mathfrak{Z}(x + \log\frac{2r}{r_0})\omega$. Local analogue of which is to be proved in the next section. Then using theorem (8.5) and condition $(\boldsymbol{\mathcal{L}})$ it is possible to see that r_0 satisfies

$$b(r)r \geq 2\mathfrak{z}(x, B) + 2\rho(x, r),$$

then $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is most probably $(\geq 1 - 3e^{-x})$ less then zero.

2.*Bootstrap world concentration of bMLE*

Interestingly in the bootstrap world one needs to extend the set where $\widetilde{\boldsymbol{\theta}}^\flat$ concentrates. However, the key idea of the proof remains.

By definition $Ł^\flat(\widetilde{\boldsymbol{v}}^\flat, \boldsymbol{v}_\flat^*)$ is positive. A uniform bound on $\zeta^\flat(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ over $\mathbb{R}^p \setminus \Upsilon(r_0)$ translates as

$$|\zeta^\flat(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}) - (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})\nabla\zeta^\flat(\widetilde{\boldsymbol{\theta}})| \leq \rho_1(x, r)r,$$

where $\rho_1(x, r) \stackrel{\text{def}}{=} 6\nu_0\mathfrak{Z}(x + \log\frac{2r}{r_0})\omega_1(r)$. Rewriting it one has

$$|Ł^\flat(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})\nabla\zeta^\flat(\widetilde{\boldsymbol{\theta}})| \leq \rho_1(x, r)r,$$

and the deviation bound on $\|\boldsymbol{\xi}^\flat\|$, from theorem 8.6, and part one of the proof enable with probability $\geq 1 - 3e^{-x}$ an inequality

$$|L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)| \leq \rho(r, x)r + r\mathfrak{z}(x, B) - \frac{r^2 b(r)}{2}.$$

And $Ł^\flat(\widetilde{\boldsymbol{\theta}}^\flat, \widetilde{\boldsymbol{\theta}})$ is negative for $r_0^\flat$ satisfying inequality

$$b(r)r \geq 12\nu_0\mathfrak{Z}(x + \log\frac{2r}{r_0})(\omega + \omega_1(r)) + 3\mathfrak{z}(x, B).$$

□

8.5 Square root Wilks expansion

Theorem 8.10. *Let conditions (4.1) to be fulfilled, then with probability* $\geq 1 - Ce^{-x}$ *holds*

$$\left| \sqrt{2T_{LR}} - \|\boldsymbol{\xi}^s\| \right| \leq 7 \diamondsuit(r_0, x),$$

where $\diamondsuit(r, x)$ *is given by*

$$\diamondsuit(r, x) \stackrel{\text{def}}{=} (\delta(r) + 6\nu_0 \omega \mathfrak{Z}(x))r.$$

Proof. Compared to the body of the work redefine

$$\boldsymbol{v} \leftrightarrow \boldsymbol{\theta}, \;\; \boldsymbol{\theta} \to \Pi \boldsymbol{v}, \;\; \boldsymbol{\eta} \leftrightarrow (I - \Pi)\, \boldsymbol{v}$$

In the proof one relies on local linear approximation of the quasi log-likelihood following with dominating probability from bound on stochastic component (theorem 8.7) and $(\boldsymbol{\mathcal{L}_0})$. For a quadratic form of parameters $\boldsymbol{v}'$ and $\boldsymbol{v}'_1$: $\mathbb{L}(\boldsymbol{v}', \boldsymbol{v}'_1) = (\boldsymbol{v}' - \boldsymbol{v}'_1)^T \nabla L(\boldsymbol{v}'_1) - \frac{\|\boldsymbol{D}'_0(\boldsymbol{v}' - \boldsymbol{v}'_1)\|^2}{2}$ introduce residual on the set $\Upsilon(r_0)$

$$\alpha(\boldsymbol{v}'_1, \boldsymbol{v}'_2) = L(\boldsymbol{v}'_1, \boldsymbol{v}'_2) - \mathbb{L}(\boldsymbol{v}'_1, \boldsymbol{v}'_2).$$

Then from the inequality

$$\|\boldsymbol{D}'^{-1}_0 \nabla \!\!\!/ L(\boldsymbol{v}', \boldsymbol{v}'^{,*}) + \boldsymbol{D}'_0(\boldsymbol{v}' - \boldsymbol{v}'^{,*})\| \leq \delta(r_0) r_0,$$

directly following from $(\boldsymbol{\mathcal{L}_0})$and theorem 8.7 one concludes for $\boldsymbol{v}' \in \Upsilon(r_0)$

$$\|\boldsymbol{D}'^{-1}_0 \nabla \alpha(\boldsymbol{v}', \boldsymbol{v}'^{,*})\| \leq \diamondsuit(r_0, x),$$

with the notation $\diamondsuit(r) = (\delta(r) + 6\nu_0 \mathfrak{Z}(x)\omega)r$. Triangle inequality for $\boldsymbol{v}'_1, \boldsymbol{v}'_2 \in \Upsilon(r_0)$ gives

$$\|\boldsymbol{D}'^{-1}_0 \nabla \alpha(\boldsymbol{v}'_1, \boldsymbol{v}'_2)\| \leq 2 \diamondsuit(r_0, x).$$

and it is evident that

$$|\sqrt{2L(\boldsymbol{v}'_1, \boldsymbol{v}'_2)} - \sqrt{-2\mathbb{L}(\boldsymbol{v}'_1, \boldsymbol{v}'_2)}| \sqrt{-2\mathbb{L}(\boldsymbol{v}'_1, \boldsymbol{v}'_2)} \leq 4 \|\mathcal{D}_0(\boldsymbol{v}'_1 - \boldsymbol{v}'_2)\| \diamondsuit(r_0, x),$$

fro points $\boldsymbol{v}'_1, \boldsymbol{v}'_2$ such that $L(\boldsymbol{v}'_1, \boldsymbol{v}'_2) \geq 0$. Moving forward consider transformation matrices

$$K \stackrel{\text{def}}{=} \begin{pmatrix} 1 & -\boldsymbol{D}^{-1}_{\boldsymbol{\theta}} \boldsymbol{D}_{\boldsymbol{\theta},\boldsymbol{\eta}} \\ -\boldsymbol{D}^{-1}_{\boldsymbol{\eta}} \boldsymbol{D}_{\boldsymbol{\eta},\boldsymbol{\theta}} & 1 \end{pmatrix}$$

and

$$K_1 \stackrel{\text{def}}{=} \begin{pmatrix} 1 & -\boldsymbol{D}_{\boldsymbol{\theta},\boldsymbol{\eta}} \boldsymbol{D}^{-1}_{\boldsymbol{\eta}} \\ \boldsymbol{D}_{\boldsymbol{\eta},\boldsymbol{\theta}} \boldsymbol{D}^{-1}_{\boldsymbol{\theta}} & 1 \end{pmatrix},$$

then it can be seen that

$$\widehat{\boldsymbol{D}} \stackrel{\text{def}}{=} \begin{pmatrix} \widehat{\boldsymbol{D}}_0 & 0 \\ 0 & \widehat{\boldsymbol{D}}_1 \end{pmatrix} = \boldsymbol{D}'_0 K,$$

and furthermore

$$\boldsymbol{D}_0'^{-1} = K\widehat{\boldsymbol{D}}^{-1} = \widehat{\boldsymbol{D}}^{-1}K_1.$$

The transformation helps to get rid of non-diagonal entries of matrix $\boldsymbol{D}_0'$ and shape the form of the score $\boldsymbol{\xi}^s$.

Using proven above inequality under the alternative one has

$$\left|\sqrt{2T_{LR}} - \|\widehat{\boldsymbol{D}}(\widetilde{\boldsymbol{v}}'' - \boldsymbol{v}_1'') + \boldsymbol{b}\|\right| \leq 4\diamondsuit(r_0, x),$$

where $\widetilde{\boldsymbol{v}}''$ and $\boldsymbol{v}_1''$ are such that $\boldsymbol{v}' \stackrel{\text{def}}{=} K\boldsymbol{v}''$. The fact that norm of truncated score vector is less then norm of full vector and local expansion for $\boldsymbol{D}_0^{-1}\nabla\alpha$ yields

$$\|\widehat{\boldsymbol{D}}_0 \begin{pmatrix} \widetilde{\boldsymbol{\theta}}'' - \boldsymbol{\theta}_1'' \\ 0 \end{pmatrix}\| \leq 2\diamondsuit(r_0, x),$$

and

$$\|\widehat{\boldsymbol{D}}_0 \begin{pmatrix} 0 \\ \widetilde{\boldsymbol{\eta}}'' - \boldsymbol{\eta}_1'' \end{pmatrix} - \boldsymbol{\xi}_{\mathcal{H}}\| \leq \diamondsuit(r_0, x).$$

Combining these three suffice the announced statement. □

In the bootstrap world an almost complete analogue of the theorem is attainable. It is evident that it takes place since we show that exactly similar conditions as in real world are replicated in the bootstrap world.

Theorem 8.11. *Let conditions (4.1) hold then with probability* $\geq 1 - Ce^{-x}$

$$\left|\sqrt{2T_{BLR}} - \|\boldsymbol{\xi}_\flat^s\|\right| \leq 7\diamondsuit^\flat(r_0, x),$$

where $\diamondsuit^\flat(r, x)$ *is given by*

$$\diamondsuit^\flat(r, x) \stackrel{\text{def}}{=} \diamondsuit(r, x) + 6\nu_0\omega_1(r)\mathfrak{Z}(x)r.$$

Let us specify the proof of this fact.

Proof :

The underlying in the previous proof result - local linear approximation of a gradient - is sufficient. We are aiming thus at establishing -

$$\|\boldsymbol{D}_0^{-1}\nabla\alpha^\flat(\boldsymbol{v}, \boldsymbol{v}^*)\| \leq \diamondsuit^\flat(r_0, x).$$

It is easy to note that

$$\|\boldsymbol{D}_0^{-1}\nabla\alpha^\flat(\boldsymbol{v}, \boldsymbol{v}^*)\| \leq \|\boldsymbol{D}_0^{-1}\nabla\zeta^\flat(\boldsymbol{v}, \boldsymbol{v}^*)\| + \|\boldsymbol{D}_0^{-1}\nabla\alpha(\boldsymbol{v}, \boldsymbol{v}^*)\| \leq 6\nu_0\omega_1(r)r\mathfrak{Z}(x) + \diamondsuit(r, x),$$

which follows from the theorem 8.6 and the previous proof. Therefore, for bootstrap world square root Wilks result is true with the same notations and with a minor change for $\boldsymbol{b}^\flat \equiv 0$ since the hypothesis is exact and $\diamondsuit \to \diamondsuit^\flat$.

End of Proof

8.6 Matrix Inequalities

8.6.1 Concavity theorem of Leib

The gound for the development builds on the Leib's concavity theorem.

Theorem 8.12. *(Lieb, 1973) For the fixed self-adjoint matrix H function*

$$A \to tr\{\exp(H + \log A)\}$$

is concave with respect to positive-definite cone.

The proof of the required corollary from the Leib's concavity theorem requires several supporting lemmas. Generalization of the Jensen inequality for operator functions is important, however the core constructive point in the proof is operator convexity of entropy function. The observation allows to infer that relative entropy as a perspective of entropy is jointly convex. In view of the fact subsequent text contains slightly abused notation for relative entropy so that it equals exactly to the perspective.

Lemma 8.13. *(Lowner-Heinz) Define operator function (Entropy) - $\phi_e(X) \stackrel{\text{def}}{=} X \log X$ and define (relative entropy) - $\phi_e(X, Y) \stackrel{\text{def}}{=} X \log X - X \log Y$. Where $X \in \mathbb{R}^{p \times p}$ lie in Hilbert space of positive definite operators $\mathbb{H}_p^{++}$. Then*

1. *$\phi_e(X)$ - operator convex. Namely for any positive definite X_1, X_2 and $\lambda \in (0, 1)$*

$$\phi_e(\lambda X_1 + (1 - \lambda)X_2) \leq \lambda \phi_e(X_1) + (1 - \lambda)\phi_e(X_2)$$

2. *$\phi_e(X, Y)$ - jointly operator convex. For any positive definite X_1, Y_1, X_2, Y_2 and $\lambda \in (0, 1)$ holds*

$$\phi_e(\lambda X_1 + (1 - \lambda)Y_1, \lambda X_2 + (1 - \lambda)Y_2) \leq \lambda \phi_e(X_1, Y_1) + (1 - \lambda)\phi_e(X_2, Y_2)$$

Generalization of the lemma can be found under the name - Lowner-Heinz theorem. Below is an adopted proof of the required statement.

Proof **:**

Let us demonstrate that inverse function $f : \mathbb{R}^{++} \to \mathbb{R}^{++}$ s.t. $f(t) = t^{-1}$ is operator convex function. It is evident from definition for any invertible matrix $A \in \mathbb{R}^{p \times p}$ that

$$\lambda_i^{A^{-1}} = 1/\lambda_i^A, \quad i = \overline{1, p}$$

where λ_i^A is i-th eigenvalue of matrix A. And, therefore, also

$$\lambda_i^{(I+A)^{-1}} = 1/\lambda_i^{I+A} = 1/(1 + \frac{1}{\lambda^{A^{-1}}}), \quad i = \overline{1, p}$$

which will be useful next. Also it is worth notion that in view of continuity only middle point convexity needs to be shown for function $f(t) = t^{-1}$ being convex.
Therefore, convexity is implied by the inequality

$$\frac{1}{2}X_1^{-1} + \frac{1}{2}X_2^{-1} - \left(\frac{X_1 + X_2}{2}\right)^{-1} \succ 0$$

with respect to positive definite cone. Using the fact that matrix $C \stackrel{\text{def}}{=} X_1^{-1/2} X_2 X_1^{-1/2} \succ 0$ is positive definite helps to rearrange terms to get -

$$\frac{1}{2}I + \frac{1}{2}C^{-1} - \left(\frac{I + C}{2}\right)^{-1} \succ 0.$$

Multiplying from both sides left hand side of inequality with unit vectors from orthogonal basis of eigenvectors matrix C and using relations for eigenvalues above the matrix inequality is reduced to the p inequalities on real line

$$\frac{1 + \lambda_i^{C^{-1}}}{2} - \left(\frac{2}{1 + \frac{1}{\lambda_i^{C^{-1}}}}\right) \geq 0, \quad i = \overline{1, p}$$

which obviously hold representing difference between arithmetic and harmonic means. Therefore $f(t) = t^{-1}$ is operator convex.
Next step is to demonstrate that entropy function can be represented as a weighted sum of functions t^{-1}. For that purpose introduce an integral representation of power of a matrix X. It can be seen that

$$X^p = c_p \int_0^\infty t^p (\frac{1}{t} - \frac{1}{t + X}) dt,$$

for $p \in (0, 1)$ and c_p is a constant depending only on p. Also multiplying by X we get

$$X^p = c_p \int_0^\infty t^{p-1} (\frac{X}{t} + \frac{1}{t + X} - I) dt,$$

which now converges in the interval $p \in (1, 2)$. And adding here

$$X \log X \stackrel{\text{def}}{=} \lim_{p \to 1} \frac{X^p - X}{p - 1},$$

is sufficient to see that entropy is operator convex function. It follows a representation which is convex as a sum with positive coefficients of a convex functions.
Now it is left to demonstrate that relative entropy as was defined $\phi_e(X, Y) = X \log X - X \log Y$ is jointly convex function. Joint convexity can be seen via Hansen-Pedersen inequality [8] and relative entropy being perspective of $\phi_e(X)$ -

$$\phi_e(X, Y) = \phi_e(XY^{-1})Y.$$

Hansen-Pedersen inequality states

$$\phi_e(A^T X_1 A + B^T X_2 B) \leq A^T \phi_e(X_1) A + B^T \phi_e(X_2) B$$

for A, B s.t. $A^T A + B^T B = I$. Then for $X = \lambda X_1 + (1-\lambda) X_2$ and $Y = \lambda Y_1 + (1-\lambda) Y_2$ and matrices $A = \lambda^{1/2} Y^{-1/2} Y_1^{1/2}$ and $B = (1-\lambda)^{1/2} Y^{-1/2} Y_2^{1/2}$ we receive

$$\phi_e(X,Y) = \phi_e(A^T \frac{X_1}{Y_1} A + B^T \frac{X_2}{Y_2} B) Y \leq A^T \phi_e(\frac{X_1}{Y_1}) A Y + B^T \phi_e(\frac{X_2}{Y_2}) B Y \leq \lambda \phi_e(X_1, Y_2) + (1-\lambda) \phi_e(X_2, Y_2),$$

which ends the proof.
End of Proof

Following article by Tropp 2012 [21] let us rely on geometric properties of $\phi_e(X)$. Quantifying the approach let us use Bregman operator divergence for entropy function and try to built its affine approximation which is in turn by lemma 8.13 gives inequality

$$D_{\phi_e}(X,Y) \stackrel{\text{def}}{=} \phi_e(X) - \phi_e(Y) - (\nabla \phi(Y), X - Y) \geq 0.$$

Above Bregman divergence was defined - $D_{\phi_e}(X,Y)$, and it is easy to see that $D_{\phi_e}(X,Y) = 0$ iff $X = Y$. Therefore,the following lemma can be concluded.

Lemma 8.14. *(Variational Formula for Trace) Let Y be a positive definite matrix, then*

$$trY = \sup_{X>0} tr(X \log Y - X \log X + X)$$

Informally argument is presented above and one can skip the rigorous formal proof below.
Proof :
Obviously from $D_{\phi_e}(X,Y) \geq 0$ follows inequality for trace of $trD_{\phi_e}(X,Y) \geq 0$. Therefore,

$$trY \geq tr(X \log Y - X \log X + X).$$

But equality holds iff $X = Y$, then we conclude the statement of the lemma.
End of Proof
Operator concavity also helps to derive the following lemma.

Lemma 8.15. *Function $\sup_{X>0} g(X,Y)$ is concave if $g(X,Y)$ is jointly concave.*

Proof :
First, suggest existence of X_1, X_2 and Y_1, Y_2 s.t. they provide a partial maximum to function $g(X,Y)$. Namely define them as

$$X_1, Y_1 : \quad \sup_X g(X, Y_1) = g(X_1, Y_1),$$

$$X_2, Y_2 : \quad \sup_X g(X, Y_2) = g(X_2, Y_2).$$

Then observe that the set of inequalities hold

$$\sup_X g(X, \lambda Y_1 + (1-\lambda)Y_2) \leq g(\lambda X_1 + (1-\lambda)X_2, \lambda Y_1 + (1-\lambda)Y_2) \leq$$

$$\leq \lambda g(X_1, Y_1) + (1-\lambda)g(X_2, Y_2) = \lambda \sup_X g(X, Y_1) + (1-\lambda) \sup_X g(X, Y_2)$$

where joint operator convexity of g was used along with the definition of points (X_1, Y_1) and (X_2, Y_2).

End of Proof

Now we are in position to provide the proof of the theorem 8.12 of Lieb, 1973.
Proof :
(Theorem 8.12).
Let us start with variational formula from Lemma 8.14 for trace function saying

$$trY = \sup_{X>0} tr(-\phi(X, Y) + X).$$

Also from Lemma 8.13 we know that $\phi(X, Y)$ is jointly convex and, therefore the trace of it is also jointly convex and, thus, supremum of the trace function is convex according to Lemma 8.15. Now to demonstrate final result it suffice to substitute Y with matrix $\exp\{H + \log A\}$ in variational formula for trace giving

$$tr \exp\{H + \log A\} = sup_{X>0} tr\{-\phi(X, A) - XH + X\}$$

and finally providing the advertised statement.
End of Proof

8.6.2 Master Bound

Compared to the use of Golden-Thompson inequality

$$tr \exp\{X + Y\} \leq tr \exp\{X\} \exp\{Y\},$$

suitable for iid case one can follow theorem 8.12 and improve upper bounds on tail distribution of operator norm of the random matrix. This improvement is inherent to the study of independent but not identically distributed random variables.
Obvious corollary from theorem 8.12 can be useful further applications.

Corollary 8.16. *For any probability measure $\mathbb{P}$ and set of independent random matrices $\{S_i, i = \overline{1, n}\}$ holds*

$$\mathbb{E} tr \exp\{\sum_{i=1}^{n} S_i\} \leq tr \exp\{\sum_{i=1}^{n} \log \mathbb{E}_i \exp S_i\}$$

Proof :
Product structure of probability measure composed from independent marginal parts - $I\!P \stackrel{\text{def}}{=} \prod_{i=1}^{n} I\!P^i$ allows to write

$$I\!Etr\exp\{\sum_{i=1}^{n} S_i\} = I\!E_1 I\!E_2 ... I\!E_n tr\exp\{\sum_{i=1}^{n-1} S_i + \log\exp S_i\}.$$

Using theorem 8.12 helps to arrive at

$$I\!E_1 I\!E_2 ... I\!E_n tr\exp\{\sum_{i=1}^{n-1} S_i + \log\exp S_i\} \leq I\!E_1 I\!E_2 ... I\!E_{n-1} tr\exp\{\sum_{i=1}^{n-1} S_i + \log I\!E_n\exp S_i\}.$$

Iterating n-times the inequality accounting for the independence of $\{S_i\}$ finally relates

$$I\!Etr\exp\{\sum_{i=1}^{n} S_i\} \leq tr\exp\{\sum_{i=1}^{n}\log I\!E_i\exp S_i\}$$

End of Proof

This result can be easily combined with Markov exponential inequality to receive a bound on operator norm's tail probability.

Theorem 8.17. *(Master Bound) Suppose* $\{S_i \in I\!R^{p\times p}, i = \overline{1,n}\}$ *are independent and let us denote* $S = \sum_{i=1}^{n} S_i$. *Then following bound hold*

$$I\!P(\|S\|_\infty \geq t) \leq 2\inf_{\theta>0} e^{-\theta t} tr\exp(\sum_{i=1}^{n}\log I\!E_i\exp\theta S_i),$$

for $\theta > 0$ *and* $\|S\|_\infty = \sup_{\|\boldsymbol{u}\|_2=1} |\boldsymbol{u}^T S\boldsymbol{u}|$.

Proof :
The theorem follows directly from corollary 8.16 and Markov exponential inequality. Write

$$I\!P(\|S\|_\infty \geq z) = I\!P(\lambda_{max}(S) \vee \lambda_{max}(-S)) \leq$$

$$\leq I\!P(\lambda_{max}(S)) + I\!P(\lambda_{max}(-S)).$$

It will be evident from the subsequent derivation that it is enough to control one of the probabilities. Also the spectral mapping theorem allows to state $\forall i$

$$\exp\{\theta\lambda_{max}(S)\} = \lambda_{max}\exp\{\theta S\}$$

and combined with a trivial inequality $\lambda_{max}\exp\{S\} \leq tr\exp\{S\}$ gives

$$I\!P(\lambda_{max}(S) \geq t) = I\!P(\exp\{\theta\lambda_{max}(S)\} \geq \exp\{\theta t\}) = I\!P(\lambda_{max}\exp\{\theta(S)\} \geq \exp\{\theta t\}) \leq$$

$$\leq I\!P(tr\exp\{\theta(S)\} \geq \exp\{\theta t\}) \leq e^{-\theta t} I\!Etr\exp\{\theta S\}.$$

Now, applying corollary 8.16 to the sum $S = \sum_i S_i$ of independent matrices we achieve desired result

$$I\!P(||S||_\infty \geq t) \leq 2\inf_{\theta>0} e^{-\theta t} I\!Etr\exp\{\theta S\} \leq 2\inf_{\theta>0} e^{-\theta t} tr\{\exp(\sum_{i=1}^{n}\log I\!E_i\exp\theta S_i)\}.$$

End of Proof

The subject of next two chapters - where we derive Bernstein inequality for two types of conditions on matrices S_i - is to bound exponential moment of each independent matrix S_i, amounting to the bound on $\sum_{i=1}^{n}\log I\!E_i\exp(\theta S_i)$.

8.6.3 Bernstein inequality for uniformly bounded matrices.

The matrix version of Bernstein type inequality requires supporting lemma for uniformly bounded matrices S_i in a sense that $\|S_i\|_\infty \leq R$ for some positive and universal constant R.

Lemma 8.18. *Suppose that random matrices S_i for $i = \overline{1,n}$ are such that for some positive number R we can found $\|S_i\|_\infty \leq R$ then it holds*

$$\begin{cases} \log I\!E_i\exp\{\theta S_i\} \leq I\!E_i S_i^2\psi_2(\theta R)/R^2 & \text{if } \forall\theta > 0, \\ \log I\!E_i\exp\{\theta S_i\} \leq \frac{\theta^2 I\!E_i S_i^2}{2(1-\frac{R\theta}{3})} & \text{if } 0 < \theta < \frac{3}{R}, \end{cases}$$

where we denote by $\psi_2(u) \stackrel{\text{def}}{=} e^{u^2} - 1$.

Proof :
The proof is classic and relies on the following series of inequalities. Let us decompose the expectation of exponent

$$I\!E_i\exp\{\theta S_i\} = I\!E_i\left[I_p + \theta S_i + \theta^2 S_i^2\left(\frac{I_p}{2!} + \frac{\theta S_i}{3!} + \frac{\theta^2 S_i^2}{4!} + ...\right)\right] \leq$$

$$\leq I\!E_i\left[I_p + \theta S_i + \theta^2 S_i^2\left(\frac{1}{2!} + \frac{\theta\|S_i\|_\infty}{3!} + \frac{\theta^2\|S_i\|_\infty^2}{4!} + ...\right)\right] \leq$$

$$\leq I_p + \theta^2 I\!E_i S_i^2\left[\frac{\exp\{\theta\|S_i\|_\infty\} - 1 - \theta\|S_i\|_\infty}{\theta^2\|S_i\|_\infty^2}\right].$$

To proceed further it is suffice to denote that function - $\left[\frac{\exp\{u\}-1-u}{u^2}\right]$ - is non-decreasing in its argument and, therefore, last inequality can be substituted with a bound - $I_p + I\!E_i S_i^2 \left[\frac{\exp\{\theta R\}-1-\theta R}{R^2}\right]$. Making also contribution here from inequalities $e^x - x \leq e^{x^2}$ and $1 + x \leq e^x$ we arrive at

$$I\!E_i \exp\{\theta S_i\} \leq I_p + I\!E_i S_i^2 \frac{\psi_2(\theta R)}{R^2} \leq \exp\{\frac{\psi_2(\theta R) I\!E_i S_i^2}{R^2}\}.$$

This concludes the first part of our statement. However, it is useful sometimes to have more convenient expression to work with. In the fashion of sub-exponential random variables it is nice to derive result with leading term proportional to θ^2 in the right hand side of inequalities. This can be easily seen if we estimate series - $\left(\frac{1}{2!} + \frac{\theta R}{3!} + \frac{\theta^2 R^2}{4!} + ...\right)$ - using inequality $k! \geq 23^{k-2}$. Explicitly we have for $\theta \leq \frac{3}{R}$

$$\left(\frac{1}{2!} + \frac{\theta R}{3!} + \frac{\theta^2 R^2}{4!} + ...\right) \leq \frac{1}{2}\left(\sum_{k=2}^{\infty} \frac{(\theta R)^{k-2}}{3^{k-2}}\right) = \frac{1}{2(1-\theta R/3)},$$

finally justifying the second part of the lemma

$$I\!E_i \exp\{\theta S_i\} \leq I_p + \theta^2 I\!E_i S_i^2 \left(\frac{1}{2!} + \frac{\theta \|S_i\|_\infty}{3!} + \frac{\theta^2 \|S_i\|_\infty^2}{4!} + ...\right) \leq$$

$$\leq I_p + \frac{\theta^2 I\!E_i S_i^2}{2(1-\theta R/3)} \leq \exp \frac{\theta^2 I\!E_i S_i^2}{2(1-\theta R/3)}.$$

End of Proof

Matrix Bernstein inequality is easy step now to accomplish. All essential tools to provide concentration bound for the norm of random matrix was derived above. In essence one have to have two facts - first is the lemma 8.18 and second is master bound from previous section (theorem 8.17). Those two are sufficient to justify Bernstein inequality for matrices.

Theorem 8.19. *Suppose that random matrix* $S = \sum_{i=1}^{n} S_i$ *is s.t* $\forall i$ *there exist positive number* R *bounding above* $\|S_i\|_\infty \leq R$. *Also denote* $\sigma^2 \stackrel{\text{def}}{=} \|\sum_{i=1}^{n} I\!E_i S_i^2\|_\infty$ *and* $\psi_2(u) = e^{u^2} - 1$ *then it holds for* $\theta_{opt} \stackrel{\text{def}}{=} \frac{4\sigma^2 \psi_2(\theta_{opt} R)}{R^2 t}$

$$I\!P(\|S\|_\infty \geq t) \leq 2p \exp\{-\frac{4\sigma^2 \psi_2(\theta_{opt} R)}{R^2}\} = 2p \exp\{-\theta_{opt} t\},$$

which incurs
a. for $t < \frac{4\psi_2(R)\sigma^2}{R^2} \stackrel{\text{def}}{=} t_{max}^2$

$$I\!P(\|S\|_\infty \geq t) \leq 2p \exp\{-\frac{R^2 t^2}{4\psi_2(R)\sigma^2}\} = 2p \exp\{-(t/t_{max})^2\},$$

b.Bernstein inequality

$$I\!P(\|S\|_\infty \geq t) \leq 2p \exp\{-\frac{t^2}{2\sigma^2(1 + Rt/3\sigma^2)}\}$$

Proof **:**
Straightforwardly apply master bound and lemma 8.18 to get -

$$I\!P(\|S\|_\infty \geq t) \leq 2 \inf_{\theta>0} e^{-\theta t} tr \exp(\sum_{i=1}^{n} \log I\!E_i \exp \theta S_i) \leq 2 \inf_{\theta>0} e^{-\theta t} tr \exp\left(\sum_{i=1}^{n} I\!E_i S_i^2 \psi_2(\theta R)/R^2\right) \leq$$

$$\leq 2p \inf_{\theta>0} \exp\left(-\theta t + \sigma^2 \psi_2(\theta R)/R^2\right).$$

Analogously for the second case in lemma 8.18 for $0 < \theta < \frac{3}{R}$

$$I\!P(\|S\|_\infty \geq t) \leq 2p \inf_{\theta>0} \exp\left(-\theta t + \frac{\theta^2 \sigma^2}{2(1-\theta R/3)}\right).$$

And the most unwieldy thing here is to optimize over θ. Let us first deal with upper inequality above, namely we try to choose θ in a way to receive almost Gaussian like type of behavior for tails. For that we introduce $\alpha \stackrel{\text{def}}{=} \frac{\theta}{t} - \frac{\psi_2(\theta R)\sigma^2}{t^2 R^2}$. It is evident that if lover bound on $\inf_{\theta>0} \alpha(\theta)$ is established then an upper-bound on right hand side of the first inequality will follow

$$\exp\left(-\theta t + \sigma^2 \psi_2(\theta R)/R^2\right) = \exp\{-\alpha t^2\} \leq \exp\{-\inf_{\theta>0} \alpha(\theta) t^2\}.$$

To proceed we rearrange alpha in the following way

$$\alpha = -\left(\sqrt{\frac{\sigma^2 \psi_2(\theta R)}{\theta^2 R^2}} \frac{\theta}{t} - \sqrt{\frac{\theta^2 R^2}{4\sigma^2 \psi_2(\theta R)}}\right)^2 + \frac{\theta^2 R^2}{4\sigma^2 \psi_2(\theta R)},$$

and now choose $\theta_{opt} \stackrel{\text{def}}{=} \frac{4\sigma^2 \psi_2(\theta_{opt} R)}{R^2 t}$ to approximate optimal α. Then we have $\alpha(\theta_{opt}) = \frac{4\sigma^2 \psi_2(\theta_{opt} R)}{R^2 t^2}$ and finally tail behavior -

$$I\!P(\|S\|_\infty \geq t) \leq 2p \exp\{-\frac{4\sigma^2 \psi_2(\theta_{opt} R)}{R^2}\} = 2p \exp\{-\theta_{opt} t\}.$$

Now we can analyze in more details the last formula. For example, in the case $\theta_{opt} < 1$ it is easily seen that $\psi_2(\theta_{opt} R) < \theta_{opt}^2 \psi_2(R)$ and, therefore, $\theta_{opt} > \frac{R^2 t}{4\psi_2(R)\sigma^2}$, which in view of $\psi_2(\theta_{opt} R) \geq \theta_{opt}^2 R^2$ recovers Gaussian tail behavior

$$I\!P(\|S\|_\infty \geq t) \leq 2p \exp\{-\frac{R^2 t^2}{2\psi_2(R)\sigma^2}\}.$$

This is useful illustration that if $R \to 0$ then obviously one gets more Gaussian like tail behavior

Also Bernstein inequality can be recovered in a classical form. Following below statements are usually can be seen as an argument to the proof of Bernstein like inequality and were used previously in the proof of lemma 8.18. In words using Taylor decomposition with inequality $k! \geq 23^{k-2}$ yield estimate for all $0 < \theta < \frac{3}{R}$

$$\psi_2(\theta_{opt} R) \leq \frac{\theta_{opt}^2 R^2}{2(1 - 2\theta_{opt} R/3)}.$$

Once again from definition of optimal point we can see that $\theta_{opt} \geq \frac{t(1-2\theta_{opt}R/3)}{2\sigma^2}$ and $\theta_{opt} \geq \frac{t}{2\sigma^2(1+Rt/3\sigma^2)}$. It can be easily verified for new point $\theta^1_{opt} \stackrel{\text{def}}{=} \frac{t}{2\sigma^2(1+Rt/3\sigma^2)}$ that $\theta^1_{opt}R < 3$ and, therefore, we receive identical to classical Bernstein result

$$I\!P(\|S\|_\infty \geq t) \leq 2p \exp\{-\frac{t^2}{2\sigma^2(1+tR/3\sigma^2)}\}.$$

This finalizes the proof of the theorem. It is left to establish only Bernstein type inequality in a conventional way. For that purpose let us use the second part of lemma 8.18 which yields inequality

$$I\!P(\|S\|_\infty \geq t) \leq 2p \inf_{\theta>0} \exp\left(-\theta t + \frac{\theta^2\sigma^2}{2(1-\theta R/3)}\right).$$

Instead of optimization choosing $\theta = \frac{t}{\sigma^2(1+\frac{tR}{3\sigma^2})}$ we arrive at

$$I\!P(\|S\|_\infty \geq t) \leq 2p \exp\left(-\frac{t^2}{2\sigma^2(1+\frac{tR}{3\sigma^2})}\right)$$

and finalize the second part of the theorem.
End of Proof

8.6.4 Bernstein ineqaulity for sub-gaussian matrices

To develop the theory in the section let us explore a bound analogous to the previous, however, requiring only sub-Gaussian tail behavior of a norm of the random matrix $S = \sum\limits_{i=1}^{n} S_i$. Analogous result can be found in the work by Koltchinskii [9].

Define for convex function $\psi_\alpha(u) \stackrel{\text{def}}{=} e^{u^\alpha} - 1$ (see van der Vaart and Wellner [23]) and operator norm $\|S_i\|_{op}$ a moment

$$\|S_i\|^{\psi_\alpha}_\infty \stackrel{\text{def}}{=} I\!E_i \exp\{\|S_i\|^\alpha_{op}\} - 1.$$

If we bound this distance we will get Gaussian like behavior for tails and thus can complement our earlier discussion with more soft bound for the tail probability. I essence we can state

Theorem 8.20. *Suppose that random matrix $S = \sum\limits_{i=1}^{n} S_i \in I\!R^{p\times p}$ is s.t $\forall i$ there exist two positive numbers $C_n > \theta$ and $C_p > 0$ for which*

$$\|\theta S_i\|^{\psi_1}_\infty \leq C_p.$$

And choose R and δ to satisfy

$$\frac{\delta\psi_2(3)\psi_1(3)}{R^3} = \frac{1}{\sigma^2}\sum_{i=1}^{n} \|6S_i/R\|^{\psi_1}_\infty.$$

Then Bernstein matrix inequality holds again

$$I\!P(\|S\|_{op} \geq t) \leq 2p \exp\{-\frac{t^2}{2\sigma^2(1+\delta)(1+Rt/3\sigma^2)}\},$$

where $\sigma^2 \stackrel{\text{def}}{=} \|\sum_i I\!E_i S_i^2\|_{op}$.

Proof :
Let us start with the bound for exponential moments analogous to the ones in lemma 8.18. One can see for some positive constant R

$$I\!E_i \exp\{\theta S_i\} \leq I_p + \frac{I\!E_i S_i^2 \psi_2(\theta R)}{R^2} + I\!E_i S_i^2 \frac{\psi_1(\theta\|S_i\|_\infty)}{\|S_i\|_\infty^2} \mathbf{1}(\|S_i\|_\infty > R).$$

The derivation remains the same as in the theorem 8.19 if the term is bounded

$$I\!E_i S_i^2 \frac{\psi_1(\theta\|S_i\|_\infty)}{\|S_i\|_\infty^2} \mathbf{1}(\|S_i\|_\infty > R)$$

with the goal to establish R as small as possible such that it further sharpens bound on the quadratic term above according to the results from theorem 8.19. However, it is also important to keep reminder term with indicator small or at least proportional to the quadratic one, which naturally requires larger values of R. Resolving this trade off one comes at an optimal value R.

Proceed with substitution of indicator function with smooth approximation

$$\mathbf{1}(\|S_i\|_\infty > R) \leq \frac{\psi_1(\theta\|S_i\|_\infty) R}{\psi_1(\theta R)\|S_i\|_\infty},$$

where it was used that $\psi_1(u)/u$ is non-decreasing function. Thus, it leads to

$$I\!E_i S_i^2 \frac{\psi_1(\theta\|S_i\|_\infty)}{\|S_i\|_\infty^2} \mathbf{1}(\|S_i\|_\infty > R) \leq I\!E_i S_i^2 \frac{\psi_1^2(\theta\|S_i\|_\infty) R}{\psi_1(\theta R)\|S_i\|_\infty^3} \leq$$

$$\leq \frac{R}{\psi_1(\theta R)} I\!E_i \frac{\psi_1^2(\theta\|S_i\|_\infty)}{\|S_i\|_\infty} I_p$$

And to be consistent sum over i of these terms needs to resemble quadratic one in the inequality above

$$\frac{\delta \psi_2(\theta R)\psi_1(\theta R)}{R^3} = \frac{1}{\|\sum_{i=1}^n I\!E_i S_i^2\|_{op}} \sum_{i=1}^n I\!E_i \frac{\psi_1^2(\theta\|S_i\|_\infty)}{\|S_i\|_\infty}.$$

Observe here that the function to the left is increasing and for sufficiently large values of R and sufficiently small δ equality can be always satisfied. However, one additionally need to bound right hand side to demonstrate that such an R exists. It can be done by following rough estimate for $\theta < C_n/2$

$$I\!E_i \frac{\psi_1^2(\theta\|S_i\|_\infty)}{\|S_i\|_\infty} \leq I\!E_i \psi_1(2\theta\|S_i\|_\infty) \leq C_p.$$

Although it is rough it provides enough evidence to justify existence of R. As to what value it equals exactly needs to be addressed implicitly via equality above. Since solution exists one can establish

$$\sum_{i=1}^{n} \log I\!E_i \exp\{\theta S_i\} \leq (1+\delta)\frac{\| \sum\limits_{i=1}^{n} I\!E_i S_i^2 \|\psi_2(\theta R)}{R^2}$$

and the first result follows from theorem 8.19.
Let us dwell here finally on the constants R and δ. From the proof of theorem 8.19 we have $\theta R^* < 3$. Then the definition of R above helps tp built an upper estimate R' on it given by

$$\frac{\delta\psi_2(3)\psi_1(3)}{R'^3} = \frac{1}{\sigma^2}\sum_{i=1}^{n} I\!E_i\psi_1(6\|S_i\|_\infty/R').$$

and finalize the proof of theorem.
End of Proof

Apply this result to specific case when matrices S_i are built based on sub-exponential random vectors $\boldsymbol{x}_i \in I\!R^p$, for which we know that

$$I\!E_i \exp(\boldsymbol{\gamma}\boldsymbol{x}_i) \leq \exp\{||\boldsymbol{\gamma}||_2^2/2n\},$$

holds for any $i = \overline{1,n}$ and $\boldsymbol{\gamma} \in I\!R^p$. Namely, define matrix S_i as

$$S_i \stackrel{\text{def}}{=} \boldsymbol{x}_i\boldsymbol{x}_i^T - I\!E_i\boldsymbol{x}_i\boldsymbol{x}_i^T.$$

we can draw from definition following inequality

$$||S_i||_\infty \leq ||\boldsymbol{x}_i||_2^2 + ||I\!E_i\boldsymbol{x}_i\boldsymbol{x}_i^T||_\infty.$$

In view of this note one can establish next corollary.

Corollary 8.21. *For matrices $S_i \stackrel{\text{def}}{=} \boldsymbol{x}_i\boldsymbol{x}_i^T - I\!E_i\boldsymbol{x}_i\boldsymbol{x}_i^T$, vectors $\boldsymbol{x}_i$, for which exponential moment condition above holds with $n > 2$, the constants from theorem* 8.20 *are*

$$R = \frac{12p}{n}$$

and there exist $0 < \alpha < 0.012$ such that

$$\delta \leq \alpha\frac{p^3}{n^2\sigma^2}$$

Proof :
For this technical proof one need to upper bound θS_i. And then applying definition of R and δ from theorem 8.20 leads to the result. Notice that

$$\|\theta S_i\|_\infty^{\psi_1} \stackrel{\text{def}}{=} I\!E_i \exp\{\theta\|S_i\|_\infty\} - 1 \leq e^{\theta\|I\!E_i\boldsymbol{x}_i\boldsymbol{x}_i^T\|_\infty} I\!E_i \exp\{\theta\|\boldsymbol{x}_i\|_2^2\} - 1.$$

The expectation on right hand side of inequality can be explicitly calculated using exponential moment condition for $\boldsymbol{x}_i$. It is evident that such an integral converges for $\theta < n/2$ and explicit calculation then gives

$$I\!E_i \exp\{\theta\|\boldsymbol{x}_i\|_2^2\} = \frac{1}{\sqrt[p]{2\pi}} I\!E_i \int_{I\!R^p} \exp\{\sqrt{2\theta}\boldsymbol{x}_i\boldsymbol{\gamma} - \frac{\|\boldsymbol{\gamma}\|^2}{2}\}d\boldsymbol{\gamma} \leq$$

$$\leq \frac{1}{\sqrt[p]{2\pi}} \int_{I\!R^p} \exp\{\frac{(2\theta/n)\|\boldsymbol{\gamma}\|^2}{2} - \frac{\|\boldsymbol{\gamma}\|^2}{2}\}d\boldsymbol{\gamma}$$

which easily gives us

$$I\!E_i \exp\{\theta\|\boldsymbol{x}_i\|_2^2\} \leq (1 - 2\theta/n)^{-p/2}.$$

and adapting it to θS_i yields

$$\|\theta S_i\|_\infty^{\psi_1} \leq e^{\theta\|I\!E_i\boldsymbol{x}_i\boldsymbol{x}_i^T\|_\infty}(1 - 2\theta/n)^{-p/2} - 1.$$

Now choose $R = 12p/n$. Knowing that $\theta R < 3$ we can check that here $2\theta < n/2$ as required for a norm to be finite. And using theorem 8.20 δ is given by the formula

$$\delta = \frac{(12p)^3}{\psi_2(3)\psi_1(3)n^3\sigma^2} \sum_{i=1}^n \|nS_i/2p\|_{\psi_1} \leq$$

$$\leq \frac{(12p)^3(e^{n\|I\!E_i\boldsymbol{x}_i\boldsymbol{x}_i^T\|_\infty/4p+1/2}\sqrt{p/(p-1)} - 1)}{\psi_2(3)\psi_1(3)n^2\sigma^2} =$$

$$= 0.012\frac{p^3(e^{n\|I\!E_i\boldsymbol{x}_i\boldsymbol{x}_i^T\|_\infty/4p+1/2}\sqrt{p/(p-1)} - 1)}{n^2\sigma^2}.$$

If we further note $\|I\!E_i\boldsymbol{x}_i\boldsymbol{x}_i^T\|_\infty \leq C/n$, then using assumption on n the order of δ is shown to be

$$\delta \leq \alpha\frac{p^3}{n^3\sigma^2} \sim \frac{p^2}{n^2}$$

for a positive constant satisfying $\alpha < 0.012$.

End of Proof

This example of an empirical covariance matrix demonstrates sharp - in view of small parameter $\frac{p^3}{n}$ - bounds on constants R and δ, however not optimal ones.

8.7 Gaussian approximation

8.7.1 Smooth representation of Kolmogorov distance.

Introduce a smooth indicator function

$$f(x) = I\!\!I(x > 0) - \frac{1}{2}\mathbf{sign}(x)e^{-|x|}$$

and define a regular difference

$$g_\alpha(t) \stackrel{\text{def}}{=} I\!E f\left(\alpha\|\boldsymbol{x}_0\|^2 - \alpha t\right) - I\!E f\left(\alpha\|\boldsymbol{x}_1\|^2 - \alpha t\right).$$

One aims at studying the limiting object

$$g_\infty(t) \stackrel{\text{def}}{=} I\!E\, \mathbb{1}\left(\|\boldsymbol{x}_0\|^2 < t\right) - I\!E\, \mathbb{1}\left(\|\boldsymbol{x}_1\|^2 < t\right) = I\!P\left(\|\boldsymbol{x}_0\| < t\right) - I\!P\left(\|\boldsymbol{x}_1\| < t\right)$$

the difference between multivariate probabilities. The smoothing function on the other hand allows for a structural characterization of the relation between g_α and g_∞.

Lemma 8.22. *Assume that $g_\alpha(t)$ has smooth second derivative. Then it satisfies an ODE*

$$g_\alpha(t) = g_\infty(t) + \frac{g''_\alpha(t)}{\alpha^2}.$$

Moreover, an ordering holds

$$\forall \alpha > 0 \quad \sup_t |g_\alpha(t)| \leq \sup_t |g_\infty(t)|.$$

Proof. The kernel function f admits an ODE representation

$$\mathbb{L}_{\boldsymbol{x}}\left(f(\alpha\|\boldsymbol{x}\|^2 - \alpha t)\right) = \mathbb{L}_{\boldsymbol{x}}\left(\mathbb{1}\left(\|\boldsymbol{x}\|^2 > t\right)\right) + \frac{1}{\alpha^2}\left(\mathbb{L}_{\boldsymbol{x}}\left(f(\alpha\|\boldsymbol{x}\|^2 - \alpha t)\right)\right)''_t$$

with a linear integral operator $\mathbb{L}_{\boldsymbol{x}}(\cdot)$ and an inequality

$$\sup_t \left|\mathbb{L}_{\boldsymbol{x}}\left(f(\alpha\|\boldsymbol{x}\|^2 - \alpha t)\right)\right| \leq \sup_t \left|\mathbb{L}_{\boldsymbol{x}}\left(\mathbb{1}\left(\|\boldsymbol{x}\|^2 > t\right)\right)\right|$$

follows from the characterization of extreme points - second derivative in maximum is negative and positive in minimum. The same applies for the difference $g_\alpha(t)$. □

A natural candidate for the investigation of an underlying structure of the problem is Fourier analysis as the ODE in the lemma [8.22] resembles an oscillator with a complex α. Thus, define a spectrum of $g_\alpha(t)$ and $g_\infty(t)$ as follows

$$G_\infty(\omega) = \mathcal{F}\left(g_\infty(t)\right) = \int_{-\infty}^{\infty} g_\infty(t) e^{-i\omega t} dt$$

$$G_\alpha(\omega) = \mathcal{F}\left(g_\alpha(t)\right) = \int_{-\infty}^{\infty} g_\alpha(t) e^{-i\omega t} dt$$

respectively and with the convention for ω being a frequency scaled by 2π. Additionally we analytically extend the spectra on $\omega \in \mathbb{C}$ which is the derivation crucial in the inversion step and we elaborate on that later (see lemma 8.23).

Easy to notice that in the Fourier world the connection between $G_\infty(\omega)$ and $G_\alpha(\omega)$ is straightforward and given by

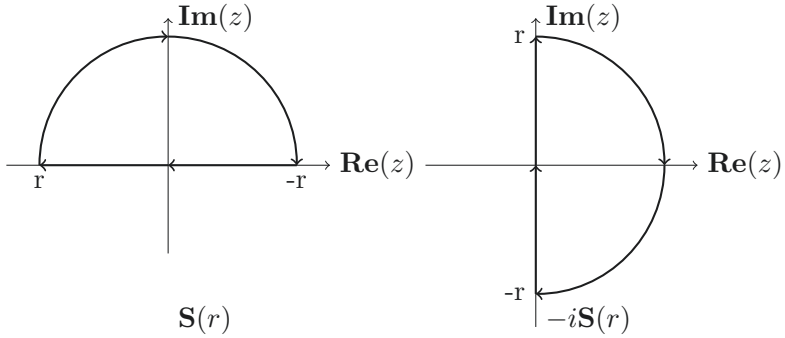
$$G_\alpha(\omega) = G_\infty(\omega) - \frac{\omega^2}{\alpha^2} G_\alpha(\omega)$$

which yields

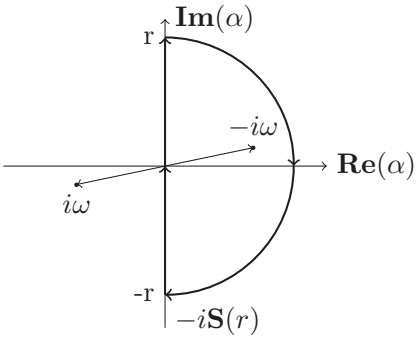
$$G_\alpha(\omega) = \frac{\alpha^2 G_\infty(\omega)}{(\alpha - i\omega)(\alpha + i\omega)}. \tag{8.1}$$

The central observation for further analysis is that α can be taken as a complex number $\alpha \in \mathbb{C}$ in the ODE leaving equation 8.1 intact.

Introduce supplementary clockwise oriented contours $\mathbf{S}(r), -i\mathbf{S}(r)$ in complex plane.



One option to find the closed-form connection between $G_\alpha(\omega)$ and $G_\infty(\omega)$ independent from α is to integrate $G_\alpha(\omega)$ over $\alpha \in -i\mathbf{S}(r)$. The step gains an additional smoothness as we will see below. After inspecting the poles of $G_\alpha(\omega)$ on the picture



it is obvious in view of the Cauchy's residue theorem to conclude for the convolution

$$\frac{1}{i\pi}\int_{-i\mathbf{S}(r_0)} (-\alpha)^{k-1} G_\alpha(\omega) d\alpha = \tag{8.2}$$

$$= \frac{1}{i\pi}\int_{-i\mathbf{S}(r_0)} \frac{(-\alpha)^{k+1} G_\infty(\omega)}{(\alpha + i\omega)(\alpha - i\omega)} d\alpha = \begin{cases} (i\omega)^k G_\infty(\omega) & \omega \in \mathbf{S}(r_0), \\ (-i\omega)^k G_\infty(\omega) & -\omega \in \mathbf{S}(r_0), \\ 0 & \textbf{else}. \end{cases}$$

where we also multiplied the spectrum by additional $(-\alpha)^{k-1}$ to generalize and expand on the idea later (see corollary 8.26).

The formula 8.2 gives clear explanation how initial function $g_\infty(t)$ can be regularized through the $g_\alpha(t)$. The answer above suggests that convolution of our 'kernels' $g_\alpha(t)$ is equivalent to the differentiation. For now the connection is settled in the Fourier world and one need to translate the result back into the initial objects. For the purpose let us rewrite the Fourier inversion formula as an integration in a complex plane.

Lemma 8.23. *Assume continuous p.d.f. of $\|\boldsymbol{x}_0\|^2$ and $\|\boldsymbol{x}_1\|^2$, then the functions $g_\alpha(t)$ and $g_\infty(t)$ can be represented as*

$$g_\alpha(t) = \frac{1}{2\pi}\int_{\mathbf{S}(r_0)} G_\alpha(\omega)e^{i\omega t}d\omega$$

and

$$g_\infty(t) = \frac{1}{2\pi}\int_{\mathbf{S}(r_0)} G_\infty(\omega)e^{i\omega t}d\omega$$

for $t > 0$ and r_0 s.t. $\mathbf{S}(r_0)$ covers all the poles of a spectrum $G_\infty(\omega)$.

Proof. Let us compute explicitly $G_\infty(\omega)$ to proceed -

$$G_\infty(\omega) = \int_{-\infty}^{\infty} g_\infty(t)e^{-i\omega t}dt = I\!\!E\left[\mathcal{F}\left(\mathbb{1}(\boldsymbol{x}_0 \in \mathcal{B}_t)\right) - \mathcal{F}\left(\mathbb{1}(\boldsymbol{x}_1 \in \mathcal{B}_t)\right)\right]$$

$$= \frac{I\!\!E e^{-i\omega\|\boldsymbol{x}_0\|^2} - I\!\!E e^{-i\omega\|\boldsymbol{x}_1\|^2}}{i\sqrt{2\pi}\omega} + \sqrt{\frac{\pi}{2}}I\!\!E e^{-i\omega\|\boldsymbol{x}_0\|^2}\delta(\omega) - I\!\!E e^{-i\omega\|\boldsymbol{x}_1\|^2}\delta(\omega)$$

$$= \frac{I\!\!E e^{-i\omega\|\boldsymbol{x}_0\|^2} - I\!\!E e^{-i\omega\|\boldsymbol{\xi}_1\|^2}}{i\omega\sqrt{2\pi}}.$$

On the other hand the contour $\mathbf{S}(r)$ can be seen as a sum of the real-line and semicircle parts - $\mathbf{S}(r) = [-r, r] \cup \mathbf{Arc}(r)$ - where the latter conforms the limit

$$\lim_{r\to\infty}\int_{\mathbf{Arc}(r)} G_\infty(\omega)e^{i\omega t}d\omega \le \lim_{r\to\infty}\pi r \sup_{\omega\in\mathbf{Arc}(r)}|G_\infty(\omega)| \le$$

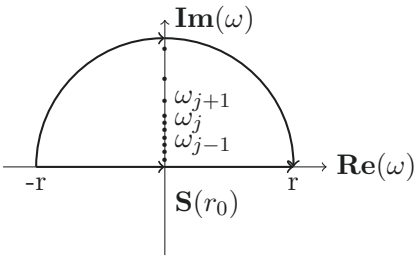
$$\le \lim_{r\to\infty}\sup_{\omega\in\mathbf{Arc}(r)}\left|I\!\!E e^{-i\omega\|\boldsymbol{x}_0\|^2} - I\!\!E e^{-i\omega\|\boldsymbol{x}_1\|^2}\right| = 0.$$

Therefore, the inverse is given as an integral over $\mathbf{S}(\infty)$

$$g_\infty(t) \stackrel{\text{def}}{=} \frac{1}{2\pi}\int_{-\infty}^{\infty} G_\infty(\omega)e^{i\omega t}d\omega + \frac{1}{2\pi}\int_{\mathbf{Arc}(\infty)} G_\infty(\omega)e^{i\omega t}d\omega$$

$$= \frac{1}{2\pi}\int_{\mathbf{S}(\infty)} G_\infty(\omega)e^{i\omega t}d\omega.$$

Defining now the critical points of $G_\infty(\omega), G_\alpha(\omega)$ as $\{\omega_{j=\overline{1,n}}\}$ and $\{\omega_{j=\overline{1,n}}, -i\alpha, i\alpha\}$ respectively (see the equation 8.1) we see that by the assumption of the lemma they are covered by the $\mathbf{S}(r_0)$.



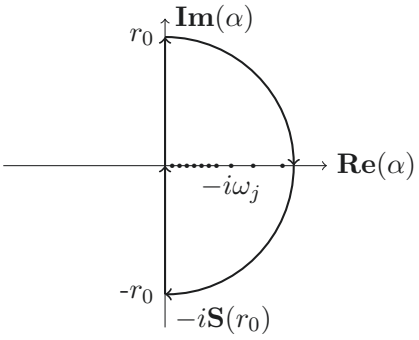
Therefore, the Cauchy's residue theorem puts the equivalence

$$g_\infty(t) = \frac{1}{2\pi}\int_{\mathbf{S}(\infty)} G_\infty(\omega)e^{i\omega t}d\omega = \frac{1}{2\pi}\int_{\mathbf{S}(r_0)} G_\infty(\omega)e^{i\omega t}d\omega$$

and completes the argument. □

Remark 8.1. *Note that in the proof above we conclude from positiveness of $\|\boldsymbol{x}_0\|^2, \|\boldsymbol{x}_1\|^2$ that all the poles but for the $i\alpha$ lie above the real line.*

With these two facts - the inversion above and convolution over α - one comes to the concluding step of the section. From the α perspective the pole structure of the $G_\infty(\omega)$ looks like it is drawn on the picture below.



Where by the definition of r_0 the convolution over α preserves the pole structure of $g_\infty(t)$. Thus, the inversion from the lemma [8.23] allows to relate explicitly the function $g_\infty(t)$ and a part of the equation 8.2 where $\omega \in \mathbf{S}(r_0)$. Merging the two one can state the theorem.

Theorem 8.24. *Assume k-continuous c.d.f. of $\|\boldsymbol{x}_0\|^2$ and $\|\boldsymbol{x}_1\|^2$, then it holds*

$$g_\infty^{(k)}(t) = (-1)^k 2i\int_{-i\mathbf{S}(r_0)} \alpha^{k-1} g_\alpha(t)d\alpha.$$

for r_0 s.t. $\mathbf{S}(r_0)$ covers all the poles of the spectrum $G_\infty(\omega)$.

***Proof* :**
Justified by the lemma [8.23] and using the equation [8.2] integrate over α

$$\int_{-i\mathbf{S}(r_0)} (-\alpha)^{k-1} g_\alpha(t) d\alpha \overset{L\,8.23}{=} \frac{1}{2\pi} \int_{-i\mathbf{S}(r_0)} (-\alpha)^{k-1} \int_{\mathbf{S}(r_0)} G_\alpha(\omega) e^{i\omega t} d\omega d\alpha$$

$$= \frac{1}{2\pi} \int_{\mathbf{S}(r_0)} i\pi \left(i\omega\right)^k G_\infty(\omega) e^{i\omega t} d\omega = \frac{i}{2} \left(\int_{\mathbf{S}(r_0)} G_\infty(\omega) e^{i\omega t} d\omega \right)_t^{(k)} \overset{L\,8.23}{=} -\frac{g_\infty^{(k)}(t)}{2i}.$$

The answer concludes the proof.
End of Proof
The theorem in turn amounts to the corollaries.

Corollary 8.25. *Introduce a function*

$$h_\alpha(\boldsymbol{x}, t) = \max\{\|\boldsymbol{x}\|^2 - t, 0\} + \frac{1}{2\alpha} e^{-\alpha \left| \|\boldsymbol{x}\|^2 - t \right|}$$

and an integral operator

$$\mathcal{A}\left(\cdot\right) \overset{\text{def}}{=} 2i \int_{-i\mathbf{S}(r_0)} I\!E\left(\cdot\right) d\alpha,$$

then under assumptions in the theorem [8.24] we have for $k = 1$

$$I\!P\left(\|\boldsymbol{x}_0\|^2 < t\right) - I\!P\left(\|\boldsymbol{x}_1\|^2 < t\right) = \mathcal{A}\left(h_\alpha(\boldsymbol{x}_1, t) - h_\alpha(\boldsymbol{x}_0, t)\right).$$

Corollary 8.26. *Introduce a function*

$$h_\alpha(\boldsymbol{x}_1, \boldsymbol{x}_0, t) = \int_0^t \left[f(\alpha \|\boldsymbol{x}_1\|^2 - \alpha x) - f(\alpha \|\boldsymbol{x}_0\|^2 - \alpha x) \right] dx$$

and an integral operator

$$\mathcal{B}\left(\cdot\right) \overset{\text{def}}{=} 2i \int_{-i\mathbf{S}(r_0)} I\!E\left(\cdot\right) \alpha d\alpha,$$

then under assumptions in the theorem [8.24] we have for densities of $\|\boldsymbol{x}_1\|^2$ *and* $\|\boldsymbol{x}_0\|^2$

$$\rho_{\|\boldsymbol{x}_1\|^2}(t) - \rho_{\|\boldsymbol{x}_0\|^2}(t) = \mathcal{B} h_\alpha(\boldsymbol{x}_1, \boldsymbol{x}_0, t).$$

In essence the theorem 8.24 and corollary 8.25 allow for the direct application of a simple Taylor expansion to the function h_α. Namely the statements claim that one can differentiate under the operator $\mathcal{A}$. It is a subject of the next chapter to explore the use case of Gaussian approximation.

8.7.2 GAR on Euclidean balls.

The road-map of the following application case of the section above is to use Taylor decomposition up to the third term. Aside from the fact the proof is technical and presents no specific interest except for the outcome, which is comparable to the work of Betnkus 2005 [3].

Classic Lindenberg construction entails the following framework -

- Define vectors $\boldsymbol{\xi}_1 \overset{\text{def}}{=} \sum_{i=1}^{n} \boldsymbol{\xi}_{1,i}$ and $\boldsymbol{\xi}_2 \overset{\text{def}}{=} \sum_{i=1}^{n} \boldsymbol{\xi}_{2,i}$ such that
  1. $\boldsymbol{\xi}_{1,i}$ and $\boldsymbol{\xi}_{2,i}$ are independent mutually and over $i = \overline{1,n}$
  2. $I\!E \boldsymbol{\xi}_{1,i} \boldsymbol{\xi}_{1,i}^T = I\!E \boldsymbol{\xi}_{2,i} \boldsymbol{\xi}_{2,i}^T = \Sigma/n$
- and the chain
$$\boldsymbol{\xi}_{/i} \overset{\text{def}}{=} \sum_{j=0}^{i-1} \boldsymbol{\xi}_{1,j} + \sum_{j=i+1}^{n+1} \boldsymbol{\xi}_{2,j}$$
under a convention $\boldsymbol{\xi}_{1,0} = \boldsymbol{\xi}_{2,n+1} = 0$.

With the introduced notations one can claim the theorem.

Theorem 8.27. *Assume continuous measures and the framework above, then it holds*

$$\sup_t \left| I\!P \left(\|\boldsymbol{\xi}_1\| < t \right) - I\!P \left(\|\boldsymbol{\xi}_2\| < t \right) \right| \leq C \left(Tr\Sigma \right)^{3/2} / \sqrt{n} \left(1 + O\left(n^{-1/2} \right) \right)$$

with the universal constant of the order $C \sim r_0 \lambda_{max}(\Sigma)$ *and* r_0 *defined in the Lemma (*8.23*).*

Proof **:**
Let us start the proof forming a chain

$$g_\infty(t) = \sum_{i=1}^{n} I\!E_i g_\infty^i(t) =$$

$$= \sum_{i=1}^{n} I\!E_i \left[I\!P_{/i} \left(\|\boldsymbol{\xi}_{/i} + \boldsymbol{\xi}_{1,i}\|^2 > t \right) - I\!P_{/i} \left(\|\boldsymbol{\xi}_{/i} + \boldsymbol{\xi}_{2,i}\|^2 > t \right) \right],$$

where the latter chaining sum is called the Lindenberg device. The main objective of the device is to exploit independence of $\boldsymbol{\xi}_{/i}$, $\boldsymbol{\xi}_{1,i}$ and $\boldsymbol{\xi}_{2,i}$ amounting to the cancellation of first and second terms in Taylor expansion.

However to exploit Taylor decomposition we first define smooth counterpart of the Lindenberg summand $g_\infty^i(t)$ as per the corollary 8.25

$$h_\alpha^i(\boldsymbol{\xi}_{1,i}, \boldsymbol{\xi}_{2,i}, t) \overset{\text{def}}{=} I\!E_{/i} \left[\int_0^t f\left(\alpha \|\boldsymbol{\xi}_{/i} + \boldsymbol{\xi}_{1,i}\|^2 - \alpha x \right) dx - \int_0^t f\left(\alpha \|\boldsymbol{\xi}_{/i} + \boldsymbol{\xi}_{2,i}\|^2 - \alpha x \right) dx \right].$$

From the theorem 8.24, the contour integral over $|\alpha| < r_0$ recovers the limiting difference

$$I\!P\left(\|\boldsymbol{\xi}_{/i}+\boldsymbol{\xi}_{1,i}\|^2>t\right)-I\!P\left(\|\boldsymbol{\xi}_{/i}+\boldsymbol{\xi}_{2,i}\|^2>t\right)=\mathcal{A}h^i_\alpha(\boldsymbol{\xi}_{1,i},\boldsymbol{\xi}_{2,i},t)$$

$$=2i\int_{i\mathbf{S}(r_0)}I\!E_i h^i_\alpha(\boldsymbol{\xi}_{1,i},\boldsymbol{\xi}_{2,i},t)d\alpha$$

for some fixed positive constant r_0. On the other hand assuming continuous pdf the function

$$f_\alpha(\boldsymbol{x},t)\stackrel{\text{def}}{=}I\!E_{/i}\int_0^t f\left(\alpha\|\boldsymbol{\xi}_{/i}+\boldsymbol{x}\|^2-\alpha x\right)dx$$

is tree times continuously differentiable with respect to $\boldsymbol{x}$ and admits Taylor decomposition up to the third term.

Computing derivatives one has

- $\frac{\partial f_\alpha(0,t)}{\partial \boldsymbol{x}}=2\alpha I\!E_{/i}f\left(\alpha\|\boldsymbol{\xi}_{/i}\|^2-\alpha t\right)\boldsymbol{\xi}_{/i}$
- $\frac{\partial^2 f_\alpha(0,t)}{\partial \boldsymbol{x}\partial \boldsymbol{x}}=4\alpha^2 I\!E_{/i}f'\left(\alpha\|\boldsymbol{\xi}_{/i}\|^2-\alpha t\right)\boldsymbol{\xi}_{/i}\boldsymbol{\xi}_{/i}^T+2\alpha I\!E_{/i}f\left(\alpha\|\boldsymbol{\xi}_{/i}\|^2-\alpha t\right)I_p$
- $\frac{\partial^3 f_\alpha(\boldsymbol{a},t)}{\partial \boldsymbol{x}\partial \boldsymbol{x}\partial \boldsymbol{x}}=8\alpha^3 I\!E_{/i}f''\left(\alpha\|\boldsymbol{\xi}_{/i}+\boldsymbol{a}\|^2-\alpha t\right)\left(\boldsymbol{\xi}_{/i}+\boldsymbol{a}\right)\otimes\left(\boldsymbol{\xi}_{/i}+\boldsymbol{a}\right)\otimes\left(\boldsymbol{\xi}_{/i}+\boldsymbol{a}\right)$

$$+8\alpha^2 I\!E_{/i}f'\left(\alpha\|\boldsymbol{\xi}_{/i}+\boldsymbol{a}\|^2-\alpha t\right)\left(\boldsymbol{\xi}_{/i}+\boldsymbol{a}\right)\otimes I_p$$

and, therefore, we obtain for the zero, first and second order approximation of the difference

$$I\!E_i\left[f_\alpha(\boldsymbol{\xi}_{1,i},t)-f_\alpha(\boldsymbol{\xi}_{2,i},t)\right]$$

the following equalities

- $I\!E_i\left[f_\alpha(0,t)-f_\alpha(0,t)\right]=0$
- $I\!E_i\left[\frac{\partial^T f_\alpha(0,t)}{\partial \boldsymbol{\xi}_{1,i}}\boldsymbol{\xi}_{1,i}-\frac{\partial^T f_\alpha(0,t)}{\partial \boldsymbol{\xi}_{2,i}}\boldsymbol{\xi}_{2,i}\right]=0$
- $I\!E_i\left[\boldsymbol{\xi}_{1,i}^T\frac{\partial^2 f_\alpha(0,t)}{\partial \boldsymbol{\xi}_{1,i}\partial \boldsymbol{\xi}_{1,i}}\boldsymbol{\xi}_{1,i}-\boldsymbol{\xi}_{2,i}^T\frac{\partial^2 f_\alpha(0,t)}{\partial \boldsymbol{\xi}_{2,i}\partial \boldsymbol{\xi}_{2,i}}\boldsymbol{\xi}_{2,i}\right]=0$

by the designed independence of the $\boldsymbol{\xi}_{/i},\boldsymbol{\xi}_{1,i}$ and $\boldsymbol{\xi}_{2,i}$. Thus, Taylor expansion of the function $I\!E_i h^i_\alpha(\boldsymbol{\xi}_{1,i},\boldsymbol{\xi}_{2,i},t)$ reads as

$$I\!E_i h^i_\alpha(\boldsymbol{\xi}_{1,i},\boldsymbol{\xi}_{2,i},t)=\frac{1}{6}I\!E_i\left[\frac{\partial^3 f_\alpha(\boldsymbol{a}_i,t)}{\partial \boldsymbol{x}\partial \boldsymbol{x}\partial \boldsymbol{x}}\cdot\boldsymbol{\xi}_{1,i}\otimes\boldsymbol{\xi}_{1,i}\otimes\boldsymbol{\xi}_{1,i}-\frac{\partial^3 f_\alpha(\boldsymbol{b}_i,t)}{\partial \boldsymbol{x}\partial \boldsymbol{x}\partial \boldsymbol{x}}\cdot\boldsymbol{\xi}_{2,i}\otimes\boldsymbol{\xi}_{2,i}\otimes\boldsymbol{\xi}_{2,i}\right]$$

for vectors $\boldsymbol{a}_i,\boldsymbol{b}_i$ such that $\|\boldsymbol{a}_i\|\le\|\boldsymbol{\xi}_{1,i}\|$ and $\|\boldsymbol{b}_i\|\le\|\boldsymbol{\xi}_{2,i}\|$ respectively as is suggested by Taylor decomposition with the remainder in Lagrange's form. The difference is given

$$I\!E_i h^i_\alpha(\boldsymbol{\xi}_{1,i},\boldsymbol{\xi}_{2,i},t)=\frac{1}{6}I\!E_i\left(\frac{\partial^3 f_\alpha(\boldsymbol{a}_i,t)}{\partial \boldsymbol{x}\partial \boldsymbol{x}\partial \boldsymbol{x}}\boldsymbol{\gamma}_i^1\otimes\boldsymbol{\gamma}_i^1\otimes\boldsymbol{\gamma}_i^1\right)\|\boldsymbol{\xi}_{1,i}\|^3+\frac{1}{6}I\!E_i\left(\frac{\partial^3 f_\alpha(\boldsymbol{b}_i,t)}{\partial \boldsymbol{x}\partial \boldsymbol{x}\partial \boldsymbol{x}}\boldsymbol{\gamma}_i^2\otimes\boldsymbol{\gamma}_i^2\otimes\boldsymbol{\gamma}_i^2\right)\|\boldsymbol{\xi}_{2,i}\|^3$$

additionally we know for a vector $\boldsymbol{a}_i$ such that $\|\boldsymbol{a}_i\| \leq \|\boldsymbol{\xi}_{1,i}\|$ and a fixed $\boldsymbol{\gamma}$ such that $\|\boldsymbol{\gamma}\| = 1$

$$\begin{aligned}\left|\frac{\partial^3 f_\alpha(\boldsymbol{a}_i,t)}{\partial \boldsymbol{x}\partial \boldsymbol{x}\partial \boldsymbol{x}}\boldsymbol{\gamma}\otimes\boldsymbol{\gamma}\otimes\boldsymbol{\gamma}\right| &\leq 4\alpha^3 I\!\!E_{/i}\left|\boldsymbol{\gamma}^T\boldsymbol{\xi}_{/i}+\boldsymbol{\gamma}^T\boldsymbol{a}_i\right|^3+4\alpha^2 I\!\!E_{/i}\left|\boldsymbol{\gamma}^T\boldsymbol{\xi}_{/i}+\boldsymbol{\gamma}^T\boldsymbol{a}_i\right| \\ &\leq 4\alpha^3 I\!\!E_{/i}\left|\boldsymbol{\gamma}^T\boldsymbol{\xi}_{/i}+\|\boldsymbol{\xi}_{1,i}\|\right|^3+4\alpha^2 I\!\!E_{/i}\left|\boldsymbol{\gamma}^T\boldsymbol{\xi}_{/i}+\|\boldsymbol{\xi}_{1,i}\|\right| \\ &\leq 4r_0^3\left(\lambda_{max}\left(\Sigma\right)+\|\boldsymbol{\xi}_{1,i}\|\right)^3+4r_0^2\lambda_{max}\left(\Sigma\right)+4r_0^2\|\boldsymbol{\xi}_{1,i}\|\end{aligned}$$

by $|f'| \leq 0.5$ and $|f''| \leq 0.5$ and thus

$$\int_{i\mathbf{S}(r_0)} I\!\!E_i h_\alpha^i(\boldsymbol{\xi}_{1,i},\boldsymbol{\xi}_{2,i},t)d\alpha \leq C(I\!\!E_i\|\boldsymbol{\xi}_{1,i}\|^3+I\!\!E_i\|\boldsymbol{\xi}_{2,i}\|^3)\left(1+O\left(n^{-1/2}\right)\right)$$

Therefore, using the bound on the moments from lemma 8.28 in the appendix and summing over Lindenberg chain, we come at the inquired statement

$$\sup_t\left|I\!\!P(\|\boldsymbol{\xi}_1\|^2>t)-I\!\!P(\|\boldsymbol{\xi}_2\|^2>t)\right| \leq C\left(Tr\Sigma\right)^{3/2}/\sqrt{n}\left(1+O\left(n^{-1/2}\right)\right).$$

End of Proof

Following the general scheme of a proof of the Berry-Esseen bound we are bound to work with the moments of a random vector, therefore in need of the technical lemma.

Lemma 8.28. *Under the assumptions in the framework above holds*

$$I\!\!E_i\|\boldsymbol{\xi}_i\|^3 \leq 14\left(Tr\Sigma/n\right)^{3/2}$$

Proof :
Generally for - $I\!\!E_i\|\boldsymbol{\xi}_i\|^k$ we can write

$$I\!\!E_i\|\boldsymbol{\xi}_i\|^k = \frac{k}{2}\int_0^\infty I\!\!P(\|\boldsymbol{\xi}_i\|^2>s)s^{(k-2)/2}ds$$

Theorem 3.1 or 4.1 from Spokoiny, Zhilova [19] on the sharp deviation bound for a sub-Gaussian vector $\boldsymbol{\xi}_i$ suggests for $s>0$

$$I\!\!P(\|\boldsymbol{\xi}_i\|^2>p+s) \leq 2e^{-\frac{s^2}{6.6p}\vee\frac{s}{6.6}}.$$

and additionally for a $s\in[-p,0]$ we know

$$I\!\!P(\|\boldsymbol{\xi}_i\|^2>p+s) \leq 1.$$

We also note that under dimension we understand $p=Tr\Sigma/n$ in the derivation. Since it was mentioned in the introduction that we work with an appropriately scaled random vectors.

Using following change of variables $s = s' + \frac{p}{n}$ we come at the inequality

$$I\!\!E_i \|\boldsymbol{\xi}_i\|^k = \frac{k}{2} \int_{-p}^{\infty} I\!\!P(\|\boldsymbol{\xi}_i\|^2 > p + s')(s' + p)^{(k-2)/2} ds' \leq$$

$$\leq k \int_0^{\infty} e^{-\frac{s^2}{6.6p}} (s' + p)^{(k-2)/2} ds' + \frac{k}{2} \int_{-p}^{0} (s' + p)^{(k-2)/2} ds' \leq$$

$$\leq k p^{k/2} \int_0^{\infty} e^{-\frac{ps^2}{6.6}} (s' + p)^{(k-2)/2} ds' + k \left(p\right)^{k/2} \leq$$

$$\leq k p^{k/2} \int_0^{\infty} e^{-\frac{s^2}{6.6}} (s + 1)^{(k-2)/2} ds + k p^{k/2}$$

and explicit calculation for $k = 3$ yields the result

$$I\!\!E_i \|\boldsymbol{\xi}_i\|^3 \leq 14 p^{3/2}.$$

End of Proof

8.8 Log-likelihood multiplier re-sampling

Theorem 8.29. *The parametric model (2.6) in the introduction - $\delta_k = 0$ - under the assumption (4.1) enables*

$$\left| I\!\!P\left(\left(T_{LR} - J\right) / \sqrt{J} > z_{\alpha}^{\flat} \right) - \alpha \right| \leq C_0 \frac{J^{3/2}}{\sqrt{Kn}} + C_1 \sqrt{\frac{J \log J + x}{Kn}}$$

with a dominating probability $> 1 - C_2 e^{-x}$ and universal constants $C_0, C_1 < \infty$.

Proof. Using respective Wilks expansions let us reduce the log-likelihood ratio statistics T_{LR}, T_{BLR} to the norms of score vectors $\|\boldsymbol{\xi}^s\|, \|\boldsymbol{\xi}_{\flat}^s\|$ - sub-exponential random vectors based on the finite sample theory assumptions (section [4]). One has from the theorems [4.3,4.4]

$$\left| \sqrt{2T_{LR}} - \|\boldsymbol{\xi}^s\| \right| \leq C \left(J + x\right) / \sqrt{Kn},$$

$$\left| \sqrt{2T_{BLR}} - \|\boldsymbol{\xi}_{\flat}^s\| \right| \leq C \left(J + x\right) / \sqrt{Kn}.$$

Both score vectors are reduced to the respective Gaussian counterparts

$$\widetilde{\boldsymbol{\xi}}_{\flat}^s \sim \mathcal{N}\left(0, \frac{1}{n} \sum_i \boldsymbol{\xi}_i^s \boldsymbol{\xi}_i^{sT}\right), \;\; \widetilde{\boldsymbol{\xi}}^s \sim \mathcal{N}\left(0, I\!\!E \boldsymbol{\xi}^s \boldsymbol{\xi}^{sT}\right).$$

In view of Gaussian approximation result (theorem 5.2) one can state

$$\sup_t \left| I\!\!P\left(\|\boldsymbol{\xi}^s\| < t\right) - I\!\!P\left(\|\widetilde{\boldsymbol{\xi}}^s\| < t\right) \right| \leq C \frac{J^{3/2}}{\sqrt{Kn}}$$

with a universal constant $C < \infty$, and analogously the theorem implies

$$\sup_t \left| I\!P\left(\|\boldsymbol{\xi}^s_\flat\| < t\right) - I\!P\left(\|\widetilde{\boldsymbol{\xi}}^s_\flat\| < t\right)\right| \leq C\frac{J^{3/2}}{\sqrt{Kn}}$$

with the universal constant $C < \infty$. In turn, Gaussian comparison result [5.1] and Bernstein matrix inequality allow to derive

$$\sup_t \left| I\!P\left(\|\widetilde{\boldsymbol{\xi}}^s_\flat\| < t\right) - I\!P\left(\|\widetilde{\boldsymbol{\xi}}^s\| < t\right)\right| \leq C_1\sqrt{J}\|I - \left(I\!E\boldsymbol{\xi}^s\boldsymbol{\xi}^{sT}\right)^{-1/2}\frac{1}{n}\sum_i \boldsymbol{\xi}^s_i\boldsymbol{\xi}^{sT}_i\left(I\!E\boldsymbol{\xi}^s\boldsymbol{\xi}^{sT}\right)^{-1/2}\|_{op}$$

$$\overset{\text{thm8.20}}{\leq} C_1\sqrt{\frac{J\log J + x}{Kn}}$$

with an exponentially large probability $1 - C_2e^{-x}$.

Finally, let us use the anti-concentration result (theorem 2.7) from Götze, F. and Naumov, A. and Spokoiny, V. and Ulyanov, V. [7], stating for a Gaussian vector $\boldsymbol{x} \sim \mathcal{N}(0, \Sigma)$

$$I\!P\left(t < \|\boldsymbol{x}\| < t + \epsilon\right) \leq \frac{C\epsilon}{\|\Sigma\|_{Fr}}.$$

It allows to translate Wilks expansions into the probabilistic language. Assembling all the statements in a cohesive structure one comes at

$$\sup_t \left| I\!P\left(\left(T_{LR} - J\right)/\sqrt{J} < t\right) - I\!P\left(\left(T_{BLR} - J\right)/\sqrt{J} < t\right)\right|$$

$$\overset{\text{Wilks+AC+GAR}}{\leq} \sup_t \left| I\!P\left(\left(\|\widetilde{\boldsymbol{\xi}}^s\| - J\right)/\sqrt{J} < t\right) - I\!P\left(\left(\|\widetilde{\boldsymbol{\xi}}^s_\flat\| - J\right)/\sqrt{J} < t\right)\right| + 2C\frac{J+x}{\sqrt{JKn}} + 2C\frac{J^{3/2}}{\sqrt{Kn}}$$

$$\overset{\text{GComp}}{\leq} C_1\sqrt{\frac{J\log J + x}{Kn}} + 2C\frac{J+x}{\sqrt{JKn}} + 2C\frac{J^{3/2}}{\sqrt{Kn}},$$

which helps to infer straightforwardly the statement of the theorem. □

Bibliography

[1] Donald W.K. Andrews, Marcelo J. Moreira, and James H. Stock. Optimal two-sided invariant similar tests for instrumental variables regression. *Econometrica*, 74(3):715–752, 2006.

[2] Vidmantas Bentkus. On the dependence of the berry–esseen bound on dimension. *Journal of Statistical Planning and Inference*, 113(2):385–402, 2003.

[3] Vidmantas Bentkus. A lyapunov-type bound in rd. *Theory of Probability & Its Applications*, 49(2):311–323, 2005.

[4] Victor Chernozhukov, Denis Chetverikov, Kengo Kato, et al. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819, 2013.

[5] Edward G Effros. A matrix convexity approach to some celebrated quantum inequalities. *Proceedings of the National Academy of Sciences*, 106(4):1006–1008, 2009.

[6] F Gotze. On the rate of convergence in the multivariate clt. *The Annals of Probability*, pages 724–739, 1991.

[7] F. Götze, A. Naumov, V. Spokoiny, and V. Ulyanov. Large ball probability, Gaussian comparison and anti-concentration. *ArXiv e-prints*, August 2017.

[8] Frank Hansen and Gert K Pedersen. Jensen's operator inequality. *Bulletin of the London Mathematical Society*, 35(4):553–564, 2003.

[9] Vladimir Koltchinskii et al. A remark on low rank matrix recovery and noncommutative bernstein type inequalities. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pages 213–226. Institute of Mathematical Statistics, 2013.

[10] A. Koziuk and V. Spokoiny. Toolbox: Gaussian comparison on Eucledian balls. *ArXiv e-prints*, April 2018.

[11] Elliott H Lieb. Convex trace functions and the wigner-yanase-dyson conjecture. *Advances in Mathematics*, 11(3):267–288, 1973.

[12] Marcelo J. Moreira. A conditional likelihood ratio test for structural models. *Econometrica*, 71(4):1027–1048, 2003.

[13] S Nagaev. An estimate of the remainder term in the multidimensional central limit theorem. In *Proceedings of the Third JapanUSSR Symposium on Probability Theory*, pages 419–438. Springer, 1976.

[14] A. Naumov, V. Spokoiny, and V. Ulyanov. Bootstrap confidence sets for spectral projectors of sample covariance. *ArXiv e-prints*, 2017.

[15] Vladimir Vasil'evich Senatov. Several uniform estimates of the rate of convergence in the multi-dimensional central limit theorem. *Teoriya Veroyatnostei i ee Primeneniya*, 25(4):757–770, 1980.

[16] VV Senatov. Uniform estimates of the rate of convergence in the multi-dimensional central limit theorem. *Theory of Probability & Its Applications*, 25(4):745–759, 1981.

[17] Vladimir Spokoiny. Penalized maximum likelihood estimation and effective dimension. *arXiv preprint arXiv:1205.0498*, 2012.

[18] Vladimir Spokoiny. Bernstein - von mises theorem for growing parameter dimension. *arXiv preprint arXiv:1302.3430*, 2013.

[19] Vladimir Spokoiny and Mayya Zhilova. Sharp deviation bounds for quadratic forms. *Mathematical Methods of Statistics*, 22(2):100–113, 2013.

[20] Vladimir Spokoiny, Mayya Zhilova, et al. Bootstrap confidence sets under model misspecification. *The Annals of Statistics*, 43(6):2653–2675, 2015.

[21] Joel Tropp. From joint convexity of quantum relative entropy to a concavity theorem of lieb. *Proceedings of the American Mathematical Society*, 140(5):1757–1760, 2012.

[22] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[23] Aad W Van Der Vaart and Jon A Wellner. *Weak Convergence*. Springer, 1996.

Figure 8.1: The empirical power of T_{LR}, T_{BLR} and T_{CLR} with weak instruments.

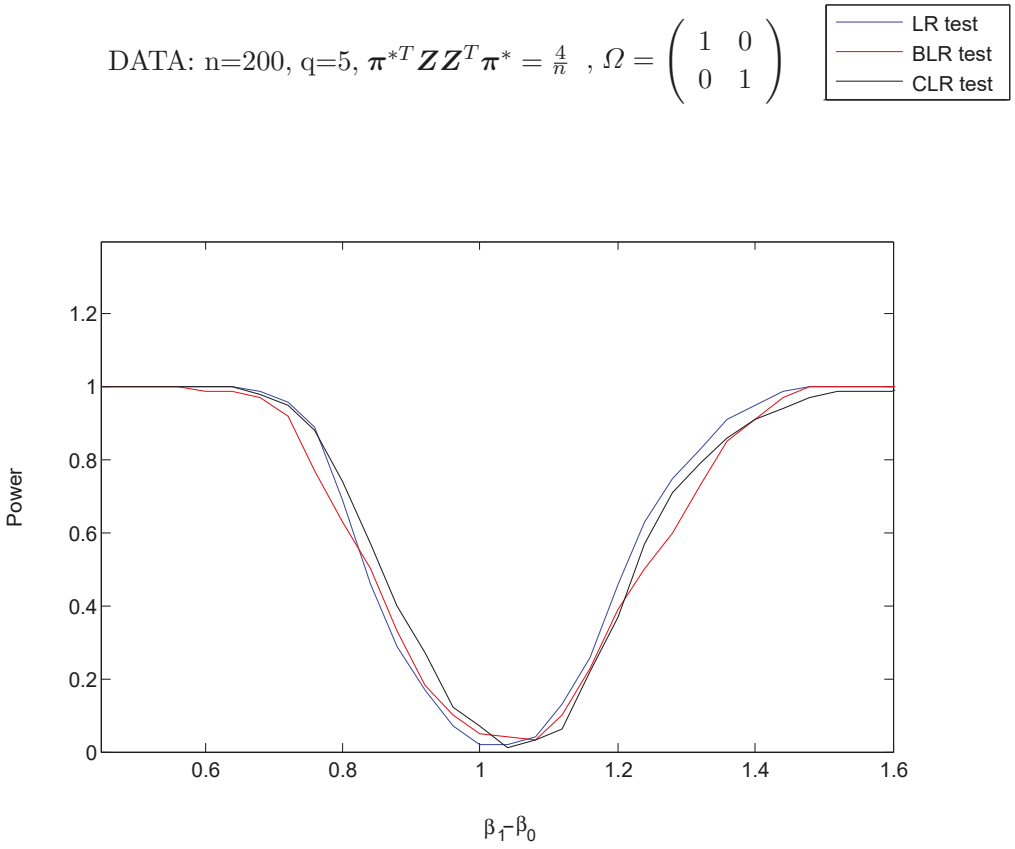


Figure 8.2: The empirical power of T_{LR}, T_{AR} and T_{LM} with weak instruments.

DATA: n=200, q=5, $\boldsymbol{\pi}^{*T}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{\pi}^* = \frac{4}{n}$, $\Omega \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

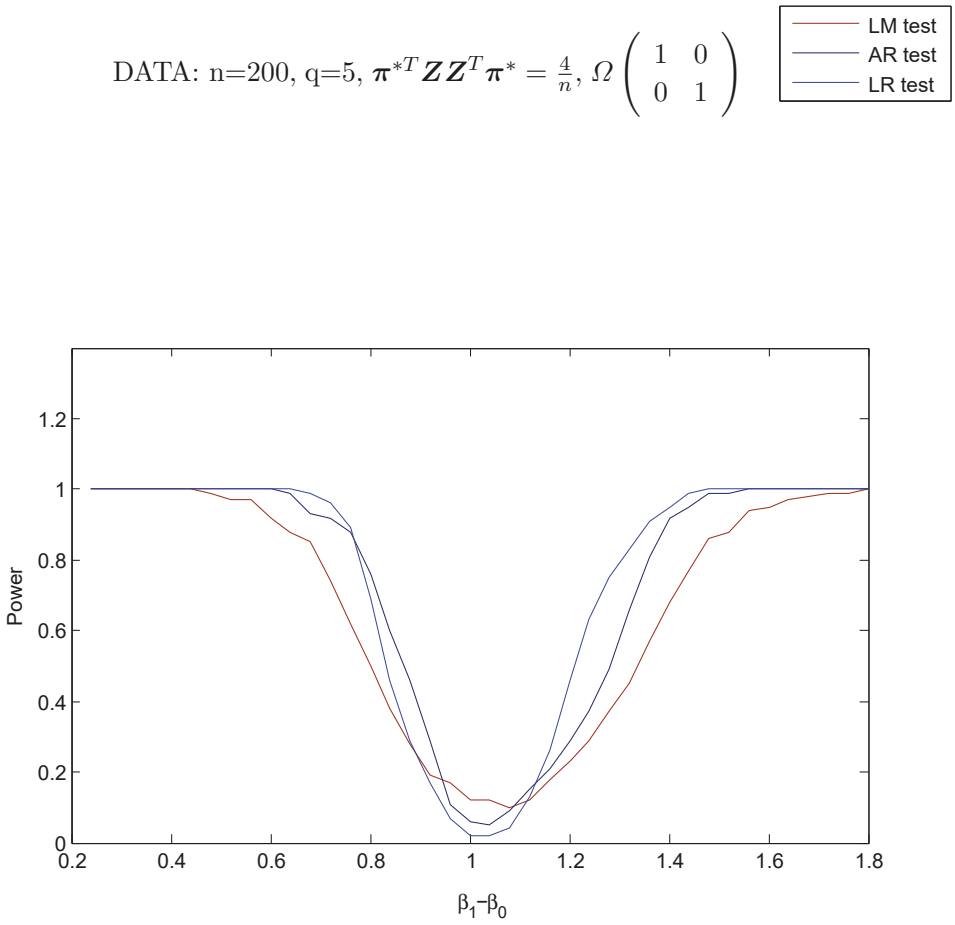


Table 1: **Power weak instrumental variables.**

DATA: n=200, q=5, $\boldsymbol{\pi}^{*T}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{\pi}^* = \frac{4}{n}$, $\Omega \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

$\beta_1 - \beta_0$	**LR**	**BLR**	**CLR**	**AR**	**LM**
0.48	1	1	1	0.99	1
0.56	1	1	1	0.97	1
0.64	1	0.99	1	0.88	0.99
0.72	0.96	0.92	0.95	0.74	0.92
0.8	0.69	0.63	0.74	0.5	0.76
0.88	0.29	0.33	0.4	0.28	0.46
0.96	0.07	0.1	0.12	0.17	0.11
1.04	0.02	0.04	0.01	0.12	0.05
1.12	0.13	0.1	0.06	0.12	0.15
1.2	0.46	0.39	0.37	0.23	0.29
1.28	0.75	0.6	0.71	0.37	0.49
1.36	0.91	0.85	0.86	0.57	0.81
1.44	0.99	0.97	0.94	0.77	0.95
1.52	1	1	0.99	0.88	0.99
1.6	1	1	0.99	0.95	1
1.68	1	1	1	0.98	1
1.76	1	1	1	0.99	1

Figure 8.3: The empirical power of T_{LR}, T_{BLR} and T_{CLR} with weak instruments and Laplace errors.

DATA: n=200, q=5, $\boldsymbol{\pi}^{*T}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{\pi}^* = \frac{2.56}{n}$

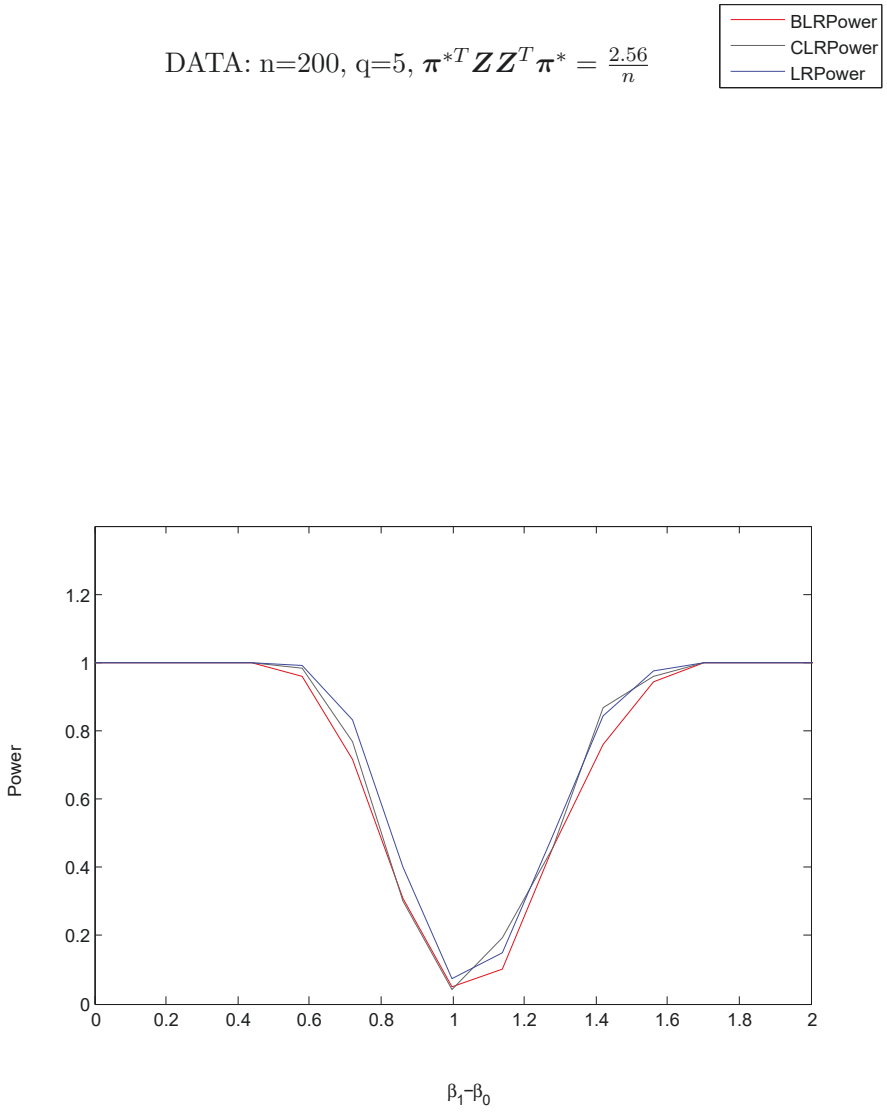


Figure 8.4: The empirical power of T_{LR}, T_{AR} and T_{LM} with weak instruments and Laplace errors.

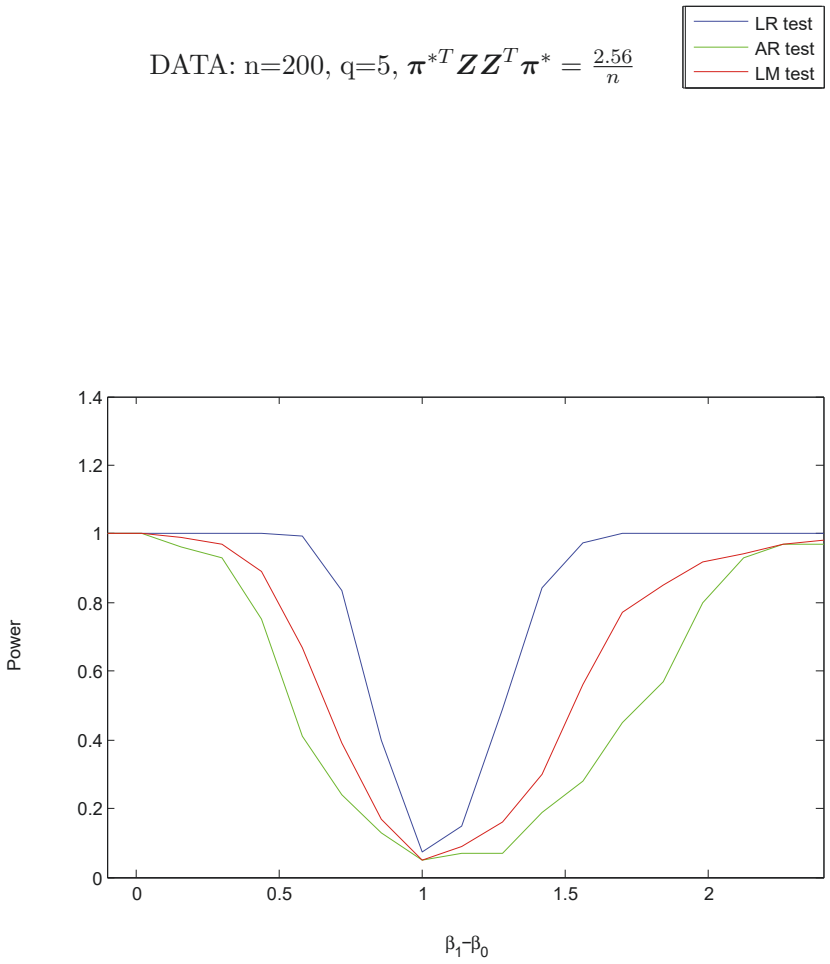


Table 2: **Power - Weak instrumental variables and Laplace noise.**

DATA: n=200, q=5, $\boldsymbol{\pi}^{*T}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{\pi}^* = \frac{2.56}{n}$

$\beta_1 - \beta_0$	**LR**	**BLR**	**CLR**	**AR**	**LM**
0.02	1	1	1	1	1
0.16	1	1	1	0.96	0.99
0.3	1	1	1	0.93	0.97
0.44	1	1	1	0.75	0.89
0.58	0.99167	0.95833	0.98333	0.41	0.67
0.72	0.83333	0.71667	0.76667	0.24	0.39
0.86	0.4	0.30833	0.3	0.13	0.17
1	0.075	0.05	0.041667	0.05	0.05
1.14	0.15	0.1	0.19167	0.07	0.09
1.28	0.49167	0.45833	0.45833	0.07	0.16
1.42	0.84167	0.75833	0.86667	0.19	0.3
1.56	0.975	0.94167	0.95833	0.28	0.56
1.7	1	1	1	0.45	0.77
1.84	1	1	1	0.57	0.85
1.98	1	1	1	0.8	0.92
2.12	1	1	1	0.93	0.94
2.26	1	1	1	0.97	0.97

Figure 8.5: The empirical power of T_{LR}, T_{BLR} and T_{CLR} with weak instruments and heteroskedastic errors.

DATA: n=200, q=5, $\boldsymbol{\pi}^{*T}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{\pi}^* = \frac{2.56}{n}$

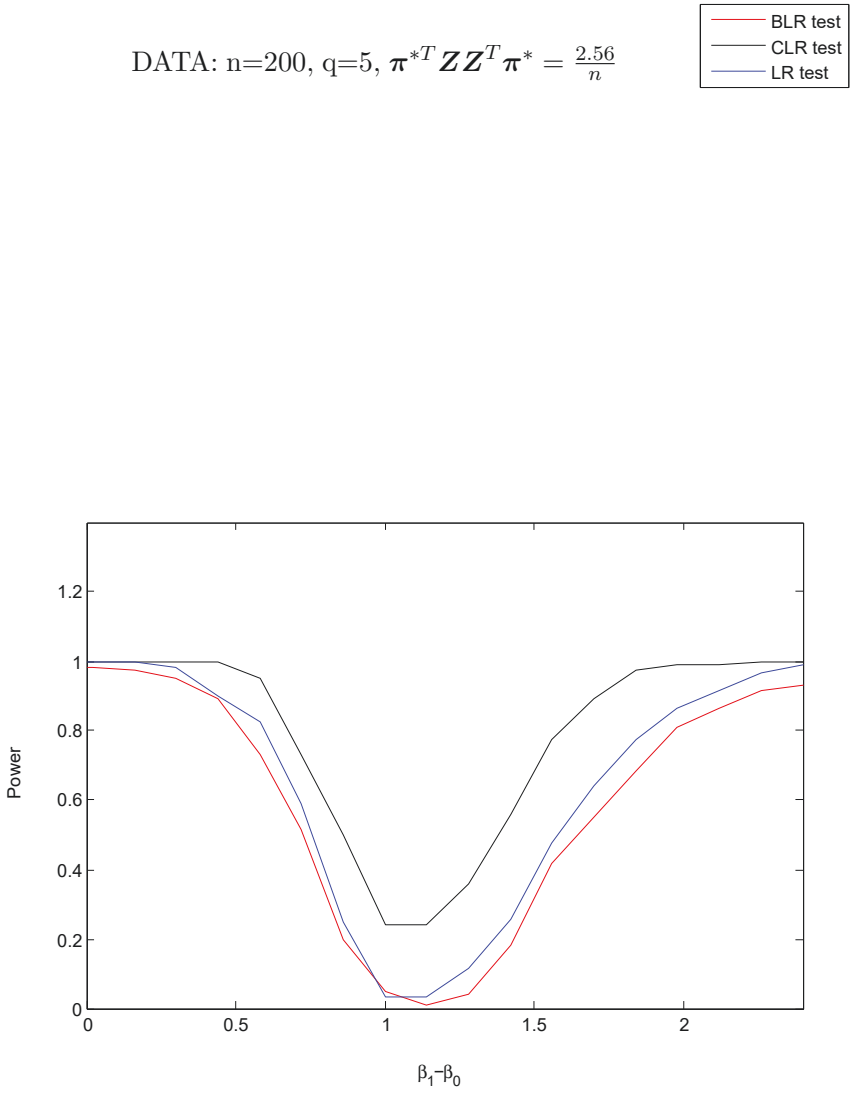


Table 3: **Power - Weak instrumental variables and heteroskedastic noise.**

DATA: n=200, q=5, $\boldsymbol{\pi}^{*T}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{\pi}^* = \frac{2.56}{n}$

$\beta_1 - \beta_0$	**LR**	**BLR**	**CLR**	**AR**	**LM**
-0.26	1	0.99167	1	1	1
-0.12	1	0.98333	1	1	1
0.02	1	0.98333	1	1	1
0.16	1	0.975	1	1	1
0.3	0.98333	0.95	1	1	1
0.44	0.9	0.89167	1	0.98333	0.99167
0.58	0.825	0.73333	0.95	0.94167	0.93333
0.72	0.59167	0.51667	0.73333	0.80833	0.8
0.86	0.25	0.2	0.5	0.7	0.50833
1	0.033333	0.05	0.24167	0.5	0.24167
1.14	0.033333	0.0083333	0.24167	0.50833	0.19167
1.28	0.11667	0.041667	0.35833	0.63333	0.35833
1.42	0.25833	0.18333	0.55833	0.73333	0.59167
1.56	0.475	0.41667	0.775	0.825	0.775
1.7	0.64167	0.55	0.89167	0.9	0.89167
1.84	0.775	0.68333	0.975	0.95833	0.95
1.98	0.86667	0.80833	0.99167	0.975	0.975
2.12	0.91667	0.86667	0.99167	1	0.99167
2.26	0.96667	0.91667	1	1	0.99167
2.4	0.99167	0.93333	1	1	0.99167

Figure 8.6: The empirical power of T_{LR}, T_{BLR} and T_{CLR} with weak instruments and heteroskedastic (periodic) errors - case 3.

DATA: n=200, q=5, $\boldsymbol{\pi}^{*T}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{\pi}^* = \frac{2.56}{n}$

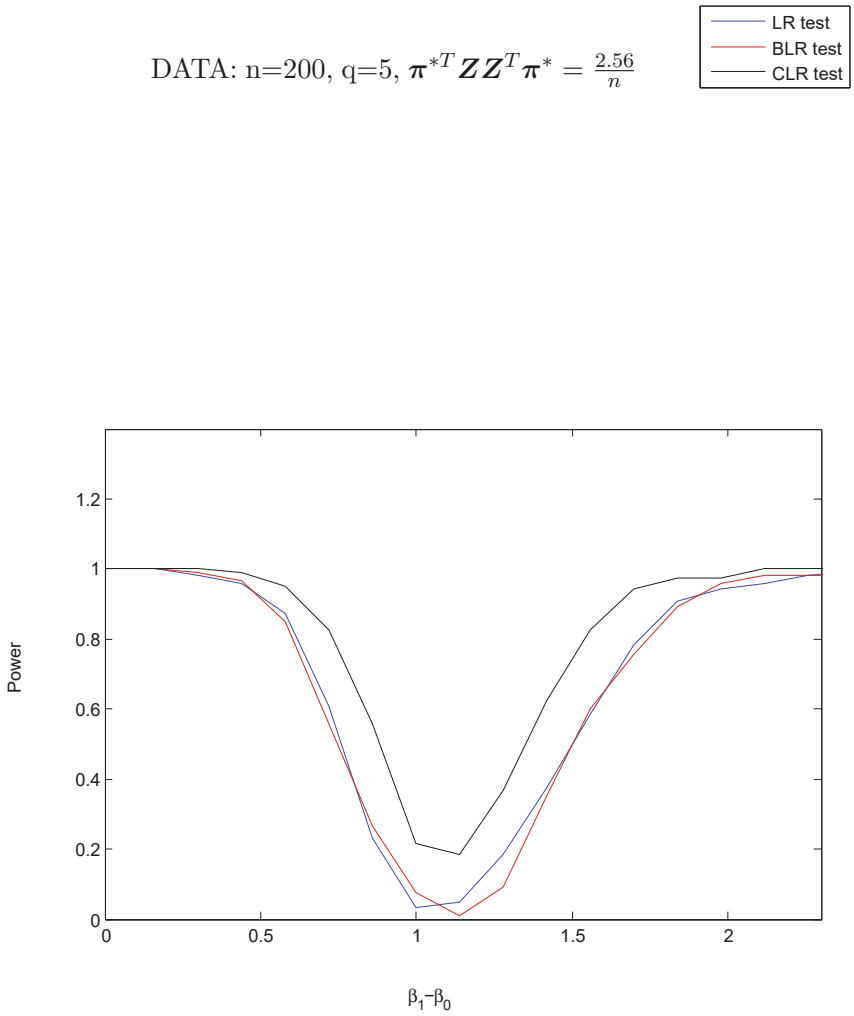


Table 4: **Power - Weak instrumental variables and Heteroskedastic periodic noise.**

DATA: n=200, q=5, $\boldsymbol{\pi}^{*T}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{\pi}^* = \frac{2.56}{n}$

$\beta_1 - \beta_0$	**LR**	**BLR**	**CLR**	**AR**	**LM**
0.16	1	1	1	1	1
0.3	0.98333	0.99167	1	0.99167	1
0.44	0.95833	0.96667	0.99167	0.96667	1
0.58	0.875	0.85	0.95	0.9	0.98333
0.72	0.60833	0.55833	0.825	0.8	0.84167
0.86	0.23333	0.26667	0.55833	0.55	0.49167
1	0.033333	0.075	0.21667	0.35	0.175
1.14	0.05	0.0083333	0.18333	0.35	0.16667
1.28	0.18333	0.091667	0.36667	0.525	0.325
1.42	0.375	0.35	0.625	0.68333	0.58333
1.56	0.58333	0.6	0.825	0.81667	0.825
1.7	0.78333	0.75833	0.94167	0.91667	0.925
1.84	0.90833	0.89167	0.975	0.95833	0.96667
1.98	0.94167	0.95833	0.975	0.99167	0.99167
2.12	0.95833	0.98333	1	1	0.99167
2.26	0.98333	0.98333	1	1	0.99167
2.4	0.98333	0.99167	1	1	0.99167