Random Dimension Low Sample Size Asymptotics

Gerd Christoph[1] and Vladimir V. Ulyanov[2,3](✉)

[1] Department of Mathematics, University of Magdeburg, Magdeburg, Germany
gerd.christoph@ovgu.de

[2] Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Moscow, Russian Federation
vulyanov@cs.msu.ru

[3] Faculty of Computer Science, HSE University, 109028 Moscow, Russian Federation

Abstract. A first investigation of high-dimensional low-sample-size (HDLSS) asymptotics, Hall, Marron and Neeman (2005) discovered a surprisingly rigid geometric structure. A sample of size k taken from the standard m-dimensional normal distribution is for large m close to the vertices of the k-dimensional simplex in m-dimensional vector space. It follows from the analysis of three geometric statistics: the length of an observation, the distance between any two independent observations and the angle between these vectors. We generalize and refine the results constructing the second order Chebyshev-Edgeworth expansions under assumption that the data dimension is random and different scaling factors are chosen.

Keywords: HDLSS data · Chebyshev-Edgeworth expansions · Random dimension · Student's t-distribution · Laplace approximation

1 Three Geometric Statistics of Gaussian Vectors

We continue to study properties of high-dimensional Gaussian random vectors. In our earlier papers Christoph, Prokhorov and Ulyanov [8] and Bobkov, Naumov and Ulyanov [5] two-sided bounds were constructed for a probability density function of the distance of a Gaussian random element Y with zero mean from a point a in a Hilbert space $\mathbb{H}$. We get new results for basic geometric statistics connected with high-dimensional random normal vectors.

Let $\mathbf{X}_1 = (X_{1,1}, ..., X_{1,m})^T$,..., $\mathbf{X}_k = (X_{k,1}..., X_{k,m})^T$ be a random sample.

In a high-dimension low-sample-size (HDLSS) data it is assumed that dimension m tends to infinity and sample size k is fixed.

One of the first investigation of HDLSS data was done in Hall, Marron and Neeman (2005) [14]. It became the basis of research in high-dimensional mathematical statistics. See a recent survey on HDLSS asymptotics and its applications in Aoshima et al. [1]. Further development see e.g. in Fujikoshi, Ulyanov

A. N. Shiryaev et al. (Eds.): ICSM-5 2020, PROMS 371, pp. 215–228, 2021.
https://doi.org/10.1007/978-3-030-83266-7_16

and Shimizu [12] when both m and k may tend to infinity. This is an important framework of the current data analysis called *Big data*. In [14] it was discovered a surprisingly rigid geometric structure. A sample of size k taken from the standard m-dimensional normal distribution is close for large m to the vertices of the k-dimensional simplex in $\mathbb{R}^m$. It follows from the analysis of three geometric statistics:

the **length** $||\mathbf{X}_i||_m$ of an observation,

the **distance** $||\mathbf{X}_i - \mathbf{X}_j||_m$ between any two independent observations, and the **angle** $\theta_m = \text{ang}(\mathbf{X}_i, \mathbf{X}_j)$ between these vectors.

We generalize and refine the results constructing the second order Chebyshev-Edgeworth expansions under assumption that the data dimension is random and different scaling factors are chosen.

In case of $\dim \mathbb{H} < \infty$ we consider a sample of size k when the dimension of the observations is a random variable N_n with values in $\mathbb{N}_+ = \{1, 2, \ldots\}$.

The present work continues our investigations in Christoph and Ulyanov [9] on these three geometric statistics of Gaussian vectors with randomly distributed dimension N_n which depends on parameter $n \in \mathbb{N}_+$ and $N_n \to \infty$ in probability as $n \to \infty$. Let the vectors $\mathbf{X}_1, ..., \mathbf{X}_k$ and $N_1, N_2, ...$ be defined on one and the same probability space and it is assumed that they are independent. If $T_m := T_m(\mathbf{X}_1, ..., \mathbf{X}_k)$ is some statistic of the vectors $\mathbf{X}_1, ..., \mathbf{X}_k$ with *non-random dimension* $m \in \mathbb{N}_+$ then the random variable $T_{N_n} = T_{N_n}(\omega)$ is defined as:

$$T_{N_n}(\omega) := T_{N_n(\omega)}\left(\mathbf{X}_1(\omega), ..., \mathbf{X}_k(\omega)\right), \quad \omega \in \Omega \quad \text{and} \quad n \in \mathbb{N}_+.$$

Therefore, the statistics T_{N_n} based on statistics T_m are constructed from the sample $\{\mathbf{X}_1, ..., \mathbf{X}_k\}$, where these vectors have the dimension N_n.

In [9], the distribution function of the normalized angle $\theta_m = \text{ang}(\mathbf{X}_i, \mathbf{X}_j)$ was approximated by a second order Chebyshev-Edgeworth expansion with a bound $\leq C m^{-2}$ for all $m \in \mathbb{N}_+$. Furthermore, the fixed dimension m of the Gaussian vectors was substituted by a random number N_n and expansions for statistics θ_{N_n} were proved.

A natural question arises whether similar results hold for the length $||\mathbf{X}_i||_{N_n}$ and the distance $||\mathbf{X}_i - \mathbf{X}_j||_{N_n}$ of Gaussian vectors with random dimension N_n.

Two cases of random dimensions (or random sample sizes) N_n are considered as e.g. in Bening, Galieva and Korolev [2], Christoph, Monakhov and Ulyanov [7] and Christoph and Ulyanov [9]:

i) The random dimension $N_n = N_n(r) \in \mathbb{N}_+$ has negative binomial distribution displaced by 1 with probability of success $1/n$, positive parameter $r > 0$ and probabilities

$$\mathbb{P}(N_n(r) = j) = \frac{\Gamma(j + r - 1)}{\Gamma(j)\,\Gamma(r)} \left(\frac{1}{n}\right)^r \left(1 - \frac{1}{n}\right)^{j-1}, \ j \in \mathbb{N}_+. \tag{1}$$

ii) The random dimension $N_n = N_n(s) \in \mathbb{N}_+$ is discrete Pareto-like distributed with parameters $n \in \mathbb{N}_+$, $s > 0$ and distribution function

$$\mathbb{P}(N_n(s) \le k) = \left(\frac{k}{s+k}\right)^n \quad \text{where} \quad N_n(s) = \max_{1\le j\le n} Y_j(s), \tag{2}$$

and $Y(s), Y_1(s), Y_2(s), ...,$ are independent discrete Pareto II distributed random variables with the common distribution

$$\mathbb{P}\big(Y(s) \le k\big) = \frac{k}{s+k} \quad \text{and} \quad \mathbb{P}(Y(s) = k) = \frac{s}{(s+k)\,(s+k-1)}, \; k \in \mathbb{N}_+. \tag{3}$$

The discrete $Y(s)$ on integers is the discretized continuous Pareto II (Lomax) random variable, see Buddana and Kozubowski [6].

Both cases of random dimensions of the Gaussian vectors are also interesting because $\mathbb{E}N_n(r) = r(n-1)+1 < \infty$ and $\mathbb{E}N_n(s) = \infty$, which has an influence on the normalization factors.

The rest of the paper is organized as follows: In Sect. 2, Chebyshev-Edgeworth expansions are proved for the geometric statistics of Gaussian vectors with fixed dimension m. Section 3 presents the transfer theorem for results with fixed sample size (in our case the dimension of the vectors) m to those with random sample size N_n. The main results are given in Sects. 4 and 5 when the random sample size is negative binomial $N_n(r)$ or discrete Pareto-like $N_n(s)$ distributed, respectively. In Sect. 6 the main results are proved.

2 Approximation for Geometric Statistics of m-Dimensional Normal Vectors

Let $\mathbf{X}_i = (X_{i,1}, ..., X_{i,m})^T$,..., $\mathbf{X}_j = (X_{j,1}..., X_{j,m})^T$ be m-dimensional vectors chosen from a sample $\{\mathbf{X}_1, ..., \mathbf{X}_k\}$ of normal distribution $\mathcal{N}(\mathbf{0}_m, \mathrm{I}_m)$ with mean vectors $\mathbb{E}\mathbf{X}_k = \mathbf{0}_m$ and covariance matrix I_m for $1 \le i < j \le k \le m$.

The **length** of the vector $\mathbf{X}_j$ is defined by the Euclidean distance $||\cdot||_m$:

$$||\mathbf{X}_i||_m = S_m^{1/2} \quad \text{with} \quad S_m = \sum\nolimits_{k=1}^m X_{i,k}^2\,. \tag{4}$$

and similarly the **distance** $||\mathbf{X}_i - \mathbf{X}_j||_m$ between any two independent vectors

$$||\mathbf{X}_i - \mathbf{X}_i||_m = \sum\nolimits_{k=1}^m (X_{i,k} - X_{j,k})^2\,. \tag{5}$$

The distribution of distance $||\mathbf{X}_i - \mathbf{X}_j||_m$ is closely linked to the distribution of length $||\mathbf{X}_i||_m$, since $(X_{i,k} - X_{j,k})/\sqrt{2}$ has also standard normal distribution $\Phi(x)$. Therefore

$$\mathbb{P}(||\mathbf{X}_i - \mathbf{X}_j||_m/\sqrt{2} \le x) = \mathbb{P}(||\mathbf{X}_i||_m \le x). \tag{6}$$

The **angle** $\theta_m = \text{ang}(\mathbf{X}_i, \mathbf{X}_j)$ between these two independent vectors with vertex at the origin and the **sample correlation coefficient** $R_m(\mathbf{X}_i, \mathbf{X}_j)$ are connected by:

$$\cos\theta_m = \frac{||\mathbf{X}_i||_m^2 + ||\mathbf{X}_j||_m^2 - ||\mathbf{X}_i - \mathbf{X}_j||_m^2}{2\,||\mathbf{X}_i||_m\,||\mathbf{X}_j||_m} = R_m(\mathbf{X}_i, \mathbf{X}_j) = R_m. \tag{7}$$

Hall, Marron and Neeman [14] showed

- for the length $||\mathbf{X}_i||_m = \sqrt{m} + \mathcal{O}_p(1)$,
- for the distance $||\mathbf{X}_i - \mathbf{X}_j||_m = \sqrt{2m} + \mathcal{O}_p(1)$ with $i \neq j$ and
- for the $\theta_m =$ angle $\text{ang}(\mathbf{X}_i, \mathbf{X}_j) = \frac{1}{2}\pi + \mathcal{O}_p(m^{-1/2})$ with $i \neq j$,

where $1 \leq i < j \leq k \leq m$ and $\mathcal{O}_p$ refers to the stochastic boundedness.

The length of the vector $\mathbf{X}_i$ drawn from an m-dimensional normal distribution $\mathcal{N}(\mathbf{0}, \mathrm{I}_m)$ is defined in (4) as $||\mathbf{X}_i||_m = S_m^{1/2}$, where the statistics S_m as a sum of the squares of m independent standard normal random variables has **chi-square distribution** with m degrees of freedom and

$$V_m = \frac{S_m - m}{\sqrt{2\,m}} \tag{8}$$

is asymptotically standard normally distributed. With the two-term Chebyshev-Edgeworth expansions in the central limit theorem for the distribution function of V_m, the following inequality results for all $m \in \mathbb{N}$

$$\left| P\Big(V_m \leq x\Big) - \Phi(x) - \varphi(x)\left(\frac{\lambda_3\,H_2(x)}{6\sqrt{m}} + \frac{\lambda_3^2\,H_5(x)}{72\,m} + \frac{\lambda_4\,H_3(x)}{24\,m}\right)\right| \leq \frac{C}{m^{3/2}}$$

where $H_2(x) = x^2 - 1$, $H_3(x) = x^3 - 3x$, $H_5(x) = x^5 - 10x^3 + 15x$ are the Chebyshev-Hermite polynomials, skewness $\lambda_3 = \sqrt{8}$ and excess kurtosis $\lambda_4 = 12$ of S_1, see Petrov [19, Sec. 5.7, Theorem 5.18].

Then $S_m = m(1 + \sqrt{2/m}\,V_m)$ and Tayor expansion of $(1+u)^{1/2}$ lead to

$$||\mathbf{X}_i||_m = S_m^{1/2} = \sqrt{m}\left(1 + \frac{1}{\sqrt{2\,m}}\,V_m - \frac{1}{4\,m}\,V_m^2 + \frac{\sqrt{2}}{8\,m^{3/2}}V_m^3 + ...\right) \tag{9}$$

Define the statistics

$$Z_m = \sqrt{2}\Big(\frac{||\mathbf{X}_i||_m}{\sqrt{m}} - 1\Big) \quad \text{and} \quad Z_m^* = \sqrt{2}\Big(\frac{||\mathbf{X}_i - \mathbf{X}_j||_m}{\sqrt{2\,m}} - 1\Big), \tag{10}$$

then (6) results in

$$P\Big(\sqrt{m}\,Z_m \leq x\Big) = P\Big(\sqrt{m}\,Z_m^* \leq x\Big). \tag{11}$$

It follows from (9) that the statistic $T_1 = \sqrt{m}Z_m$ holds

$$T_1 = \sqrt{m}Z_m = V_m - \frac{\sqrt{2}}{4\sqrt{m}}\,V_m^2 + \frac{\sqrt{1}}{4\,m}\,V_m^3 + ... \tag{12}$$

Following the sketch of the proof in Kawaguchi, Ulyanov and Fujikoshi [16, Theorem 1] (The coefficients in the polynomial $l_2(x)$ are incorrect.) and calculating the characteristic function $f_{T_1}(t)$, we obtain

$$f_{T_1}(t) = \mathbb{E}\left[\mathrm{e}^{\mathrm{i}tV_m}\left(1-\frac{\sqrt{2}(\mathrm{i}t)}{4\sqrt{m}}V_m^2+\frac{(\mathrm{i}t)}{4\,m}V_m^3+\frac{(\mathrm{i}t)^2}{16\,m}V_m^4+\mathcal{O}_p(m^{-3/2})\right)\right]$$
$$= \mathrm{e}^{-t^2/2}\left(1-\frac{\sqrt{2}((\mathrm{i}t)^3+3(\mathrm{i}t))}{12\sqrt{m}}+\frac{(\mathrm{i}t)^6-6(\mathrm{i}t)^4-9(\mathrm{i}t)^2}{144\,m})\right)+\mathcal{O}(m^{-3/2}). \quad (13)$$

This results in the related expansion of the corresponding distribution function:

Proposition 1. *Let $\mathbf{X}_i$ be a vector drawn from an m-dimensional normal distribution $\mathcal{N}(\mathbf{0}_m, \mathrm{I}_m)$. Then with the asymptotic expansion for the distribution of normalized length $Z_m = \sqrt{2}\Big(\dfrac{||\mathbf{X}_i||_m}{\sqrt{m}} - 1\Big)$ we obtain the following inequality for all $m \in \mathbb{N}$:*

$$\left|P\Big(\sqrt{m}\,Z_m \le x\Big) - \Phi(x) - \varphi(x)\left(\frac{x^2-4}{6\sqrt{2\,m}} + \frac{x^5-16x^3+24x}{144\,m}\right)\right| \le \frac{C}{m^{3/2}}. \quad (14)$$

Corollary 1. *Let $\mathbf{X}_i$ and $\mathbf{X}_j$, $i \ne j$ be independent random vectors with an m-dimensional normal distribution $\mathcal{N}(\mathbf{0}_m, \mathrm{I}_m)$. Due to (11), distribution function of the normalized distance $Z_m^* = \sqrt{2}\Big(\dfrac{||\mathbf{X}_i - \mathbf{X}_j||_m}{\sqrt{2\,m}} - 1\Big)$ has the same asymptotic expansion as the distribution of normalized length Z_m and inequality (14) with replacing Z_m by Z_m^*.*

Second order Chebyshev-Edgeworth expansion of the angle $\theta_m = \mathrm{ang}(\mathbf{X}_i, \mathbf{X}_j)$ between independent vectors $\mathbf{X}_i$ and $\mathbf{X}_j$ with vertex at the origin and the corresponding sample correlation coefficient $R_m(\mathbf{X}_i, \mathbf{X}_j)$ with computable error bounds of approximation are shown in Christoph and Ulyanov [9, Section 2], using results of Konishi [17, Sect. 4], Johnson, Kotz and Balakrishnan [15, Chap. 32], Christoph, Ulyanov and Fujikoshi [11]:

$$\sup_x\left|P\Big(\sqrt{m}\;R_m \le x\Big) - \Phi(x) - \frac{x^3-5x}{4\,m}\,\varphi(x)\right| \le \frac{B_1}{m^2} \quad (15)$$

and

$$\sup_x\left|P\left(\sqrt{m}(\theta_m - \frac{\pi}{2}) \le x\right) - \Phi(x) - \frac{x^3-15x}{12\,m}\,\varphi(x)\right| \le \frac{B_2}{m^2}. \quad (16)$$

The estimates (15) and (16) were used in Christoph and Ulyanov [9] to obtain second order approximations the statistics R_{N_n} and $\Theta_{N_n} = \theta_{N_n} - \pi/2$ when the non-random dimension m of the vectors is replaced be a random dimension N_n, where the random dimension $N_n \to \infty$ in probability when the parameter $n \to \infty$.

Analogous results for the statistics $||\mathbf{X}_i||_m$ and $||\mathbf{X}_i - \mathbf{X}_j||_m$ are proven in Sects. 4 and 5 below, when the non-random dimension m is replaced be a random dimension N_n.

3 Auxiliary Proposition

In this section, expansions for the distribution function of statistics T_{N_n} obtained from samples with random sample size (here with random dimension N_n of the considered vectors $\mathbf{X}_i$) are obtained. These depend directly on the expansions concerning statistics T_m based on non-random samples size m and expansions regarding the random sample size N_n.

First we formulate the conditions determining expansions for the statistic T_m with $\mathbb{E}T_m = 0$ and the normalized random dimension N_n:

Assumption A: *Given $\gamma \in \{-1/2, 0, 1/2\}$, $a > 1$, $C_1 > 0$ and differentiable functions $f_1(x), f_2(x)$ with bounded derivatives $f_1'(x), f_2'(x)$ such that*

$$\sup_x \left|\mathbb{P}\big(m^\gamma T_m \le x\big) - \Phi(x) - \frac{f_1(x)}{\sqrt{m}} - \frac{f_2(x)}{m}\right| \le \frac{C_1}{m^a} \quad \text{for all} \quad m \in \mathbb{N}. \quad (17)$$

Remark 1. Statistics satisfying Assumption A are shown in (14), (15) and (16).

Assumption B: *Given constants $b > 0$ and $C_2 > 0$, real numbers g_n with $0 < g_n \uparrow \infty$ if $n \to \infty$, a distribution function $H(y)$ with $H(0+) = 0$ and a function $h_2(y)$of bounded variation that*

$$\sup_{y \ge 0} \left|\mathbb{P}\left(\frac{N_n}{g_n} \le y\right) - H(y) - \frac{h_2(y)\, \mathbb{I}_{\{b>1\}}(b)}{n}\right| \le \frac{C_2}{n^b} \quad \text{for all} \quad n \ge 1. \quad (18)$$

where $\mathbb{I}_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$ *defines the indicator function of a set $A \subset \mathbb{R}$.*

Remark 2. The random dimensions $N_n(r)$ and $N_n(s)$ given in (1) and (2), respectively, fulfill Assumption B as shown in [9, Propositions 1 and 2], see (29) and (39) below.

Proposition 2. *Let $\gamma \in \{1/2,\, 0,\, -1/2\}$ and both Assumption A and B as well as the following requirements on $H(.)$ and $h_2(.)$ are fulfilled*

$$\left.\begin{array}{lll} i: & H(1/g_n) \le c_1\, g_n^{-b} & \text{for } b > 0, \\ ii: & \int_0^{1/g_n} y^{-1/2} dH(y) \le c_2\, g_n^{-b+1/2} & \text{for } b > 1/2, \\ iii: & \int_0^{1/g_n} y^{-1} dH(y) \le c_3\, g_n^{-b+1} & \text{for } b > 1, \end{array}\right\} \quad (19)$$

$$\left.\begin{array}{ll} i: \; h_2(0) = 0, \quad \text{and} \quad |h_2(1/g_n)| \le c_4\, n\, g_n^{-b} & \text{for } b > 1, \\ ii: \int_0^{1/g_n} y^{-1} |h_2(y)| dy \le c_5\, n\, g_n^{-b} & \text{for } b > 1, \end{array}\right\} \quad (20)$$

where b is the convergence rate in (18). Then for all $n \ge 1$ is valid:

$$\sup_{x \in \mathbb{R}} \left|\mathbb{P}\Big(g_n^\gamma T_{N_n} \le x\Big) - G_{n,2}(x)\right| \le C_1\, \mathbb{E}\left(N_n^{-a}\right) + (C_3 D_n + C_4)\, n^{-b} + I_n, \quad (21)$$

where with $a > 1, b > 0, f_1(z), f_2(z), h_2(y)$ are given in (17) and (18)

$$G_{n,2}(x) = \begin{cases} \int\limits_0^\infty \Phi(x\,y^\gamma)\mathrm{d}H(y), & 0 < b \le 1/2, \\ \int\limits_0^\infty \Big(\Phi(xy^\gamma) + \dfrac{f_1(x\,y^\gamma)}{\sqrt{g_n y}}\Big)\mathrm{d}H(y) =: G_{n,1}(x), & 1/2 < b \le 1, \\ G_{n,1}(x) + \int\limits_0^\infty \dfrac{f_2(x\,y^\gamma)}{g_n y}\mathrm{d}H(y) + \int\limits_0^\infty \dfrac{\Phi(x\,y^\gamma)}{n}\mathrm{d}h_2(y), & b > 1, \end{cases} \tag{22}$$

$$D_n = \sup_x \int_{1/g_n}^\infty \left| \frac{\partial}{\partial y}\left(\Phi(xy^\gamma) + \frac{f_1(xy^\gamma)}{\sqrt{g_n y}} + \frac{f_2(xy^\gamma)}{y g_n}\right)\right| \mathrm{d}y, \tag{23}$$

$$I_n = \sup\nolimits_x \left(|I_1(x,n)| + |I_2(x,n)|\right), \tag{24}$$

$$I_1(x,n) = \int_{1/g_n}^\infty \Big(\frac{f_1(x\,y^\gamma)\;\mathbb{I}_{(0,1/2]}(b)}{\sqrt{g_n y}} + \frac{f_2(x\,y^\gamma)}{g_n\,y}\Big)\mathrm{d}H(y), \qquad b \le 1, \tag{25}$$

and

$$I_2(x,n) = \int_{1/g_n}^\infty \Big(\frac{f_1(x\,y^\gamma)}{n\,\sqrt{g_n y}} + \frac{f_2(x\,y^\gamma)}{n\,g_n y}\Big)\mathrm{d}h_2(y), \qquad b > 1. \tag{26}$$

The constants C_1, C_3, C_4 are independent of n.

Proof. The proof is based on the statement in [2, Theorem 3.1] for $\gamma \ge 0$. Since in Theorems 1 and 2 in the present paper as well as in Christoph and Ulyanov [9, Theorems 1 and 2] the case $\gamma = -1/2$ is also considered, therefore the proof was adapted to $\gamma \in \{1/2,\, 0,\, -1/2\}$ in [9]. The conditions (19) and (20) guarantee integration range $(0, \infty)$ of the integrals in (22). The approximation function $G_{n,2}(x)$ in (22) is now a polynomial in $g_n^{-1/2}$ and $n^{-1/2}$. Present Proposition 2 differs from Theorems 1 and 2 in [9] only by the term $f_1(xy^\gamma)\,(g_n y)^{-1/2}$ and the added condition (19ii) to estimate this term. Therefore here the details are omitted. □

Remark 3. The domain $[1/g_n, \infty)$ of integration depends on g_n in (23), (25) and (26). Some of the integrals in (25) and (26) could tend to infinity with $1/g_n \to 0$ as $n \to \infty$ and thus worsen the convergence rates of the corresponding terms. See (47) in Sect. 6.

In the next two sections we consider the statistics Z_m and Z_m^* defined in (10) and the cases when the random dimension N_n is given in either (1) or (2). We use Proposition 2 when the limit distributions of scaled statistics Z_{N_n} are scale mixtures $G_\gamma(x) = \int_0^\infty \Phi(x\,y^\gamma)\mathrm{d}H(y)$ with $\gamma \in \{1/2,\, 0,\, -1/2\}$ that can be expressed in terms of the well-known distributions. We obtain non-asymptotic results for the statistics Z_{N_n} and $Z_{N_n}^*$, using second order approximations the statistics Z_m and Z_m^* given in (14) as well as for the random sample size N_n. In both cases the jumps of the distribution function of the random sample size N_n only affect the function $h_2(y)$ in formula (18).

4 The Random Dimension $N_n(r)$ is Negative Binomial Distributed

The negative binomial distributed dimension $N_n(r)$ has probability mass function (1)) and $g_n = \mathbb{E}(N_n(r)) = r\,(n-1)+1$. Schluter and Trede [21] (Sect. 2.1) underline the advantage of this distribution compared to the Poisson distribution for counting processes. They showed in a general unifying framework

$$\lim_{n\to\infty} \sup_y |\mathbb{P}(N_n(r)/g_n \le y) - G_{r,r}(y)| = 0, \tag{27}$$

where $G_{r,r}(y)$ is the Gamma distribution function with the identical shape and scale parameters $r>0$ and density

$$g_{r,r}(y) = \frac{r^r}{\Gamma(r)}\; y^{r-1}\, e^{-ry}\; \mathbb{I}_{(0\,,\,\infty)}(y) \quad \text{for all} \quad y\in\mathbb{R}. \tag{28}$$

Statement (27) was proved earlier in Bening and Korolev [3, Lemma 2.2].

In [9, Proposition 1] the following inequality was proved for $r>0$:

$$\sup_{y\ge 0}\left|\mathbb{P}\left(\frac{N_n(r)}{g_n}\le y\right) - G_{r,r}(y) - \frac{h_{2;r}(y)\,\mathbb{I}_{\{r>1\}}(r)}{n}\right| \le \frac{C_2(r)}{n^{\min\{r,2\}}}, \tag{29}$$

where $\quad h_{2;r}(y) = \frac{1}{2\,r}\, g_{r,r}(y)\,\big((y-1)(2-r) + 2Q_1\big(g_n\, y\big)\big) \quad$ for $\quad r>1,$

$$Q_1(y) = 1/2 - (y-[y]) \quad \text{and} \quad [y] \quad \text{is the integer part of a value } y. \tag{30}$$

Both Bening, Galieva and Korolev [2] and Gavrilenko, Zubov and Korolev [13] showed the rate of convergence in (29) for $r\le 1$. In Christoph, Monakhov and Ulyanov [7, Theorem 1] the Chebyshev-Edgeworth expansion (29) for $r>1$ is proved.

Remark 4. The random dimension $N_n(r)$ satisfies Assumption 2 of the Transfer Propositions 2 with $g_n = \mathbb{E}N_n(r)$, $H(y) = G_{r,r}(y)$, $h_2(y) = h_{2;r}(y)$ and $b=2$.

In (21), negative moment $\mathbb{E}(N_n(r))^{-a}$ is required where m^{-a} is rate of convergence of Chebyshev-Edgeworth expansion for T_m in (17). Negative moments $\mathbb{E}(N_n(r))^{-a}$ fulfill the estimate:

$$\mathbb{E}\big(N_n(r)\big)^{-a} \le C(a,r)\begin{cases} n^{-\min\{r,\,a\}}, r\ne a \\ \ln(n)\, n^{-a}, r=a \end{cases} \quad \text{for all} \quad r>0 \quad \text{and} \quad a>0. \tag{31}$$

For $r=a$ the factor $\ln n$ cannot be removed. In Christoph, Ulyanov and Bening [10, Corollary 4.2] leading terms for the negative moments of $\mathbb{E}\big(N_n(r)\big)^{-p}$ were derived for any $p>0$ that lead to (31).

The expansions of the length of the vector Z_m in (14) as well as of the sample correlation coefficient R_n in (15) and the angle θ_m in (16) have as limit distribution the standard normal distribution $\Phi(x)$. Therefore, with $g_n = \mathbb{E}N_n(r)$ and $\gamma \in \{1/2, 0, -1/2\}$, limit distributions for

$$\mathbb{P}\Big(g_n^\gamma (N_n(r))^{1/2-\gamma} Z_{N_n(r)} \le x\Big) \quad \text{are} \quad G_\gamma(x,r) = \int_0^\infty \Phi(x\, y^\gamma) \mathrm{d}G_{r,r}(y).$$

These scale mixtures distributions $G_\gamma(x,r)$ are calculated in Christoph and Ulyanov [9, Theorems 3–5]. We apply Proposition 2 to the statistics

$$T_{N_n(r)} = N_n(r)^{1/2-\gamma}\, Z_{N_n(r)} \quad \text{with the normalizing factor} \quad g_n^\gamma = \mathbb{E}(N_n(r))^\gamma.$$

The limit distributions are:

- for $\gamma = 1/2$ and $r > 0$ the **Student's t-distribution** $S_{2\,r}(x)$ with density

$$s_{2\,r}(x) = \frac{\Gamma(r+1/2)}{\sqrt{2\,r\pi}\,\Gamma(r)} \left(1 + \frac{x^2}{2\,r}\right)^{-(r+1/2)}, \quad x \in \mathbb{R}, \tag{32}$$

- for $\gamma = 0$ the **normal law** $\Phi(x)$,
- for $\gamma = -1/2$ and $r = 2$ the **generalized Laplace distributions** $L_2(x)$ with density $l_2(x)$:

$$L_2(x) = \frac{1}{2} + \frac{1}{2}\,\mathrm{sign}(x)\,(1 - (1+|x|)\,e^{-2\,|x|}) \quad \text{and} \quad l_2(x) = \left(\frac{1}{2} + |x|\right) e^{-2\,|x|}.$$

  For arbitrary $r > 0$ Macdonald functions $K_{r-1/2}(x)$ occur in the density $l_r(x)$, which can be calculated in closed form for integer values of r.

The standard Laplace density with variance 1 is $l_1(x) = \dfrac{1}{\sqrt{2}}\,\mathrm{e}^{-\sqrt{2}\,|x|}$.

Theorem 1. *Let Z_m and $N_n(r)$ with $r > 0$ be defined by (10) and (1), respectively. Suppose that (14) is satisfied for Z_m and (29) for $N_n(r)$. Then the following statements hold for all $n \in \mathbb{N}_+$:*

(i) **Student's t approximation** *using scaling factor $\sqrt{\mathbb{E}N_n(r)}$ by $Z_{N_n(r)}$*

$$\sup_x \left|\mathbb{P}\left(\sqrt{g_n}\, Z_{N_n(r)} \le x\right) - S_{2r;n}(x)\right| \le C_r \begin{cases} n^{-\min\{r,3/2\}}, & r \ne 3/2, \\ \ln(n)\, n^{-3/2}, & r = 3/2, \end{cases} \tag{33}$$

where

$$\begin{aligned} S_{2r;n}(x) = S_{2r}(x) + s_{2r}(x) \Bigg(& \frac{\sqrt{2}\,((2r-5)x^2 - 8r)}{12\,(2r-1)\sqrt{g_n}}\, \mathbb{I}_{\{r>1/2\}}(r) \\ & + \frac{96r^2 x + (-64r^2 + 128r)x^3 + (4r^2 - 32r + 39)x^5}{(x^2+2r)(2r-1)\, g_n} \mathbb{I}_{\{r>1\}}(r)\Bigg), \end{aligned} \tag{34}$$

(ii) **Normal approximation** *with random scaling factor* $N_n(r)$ *by* $Z_{N_n(r)}$

$$\sup_x \left|\mathbb{P}(\sqrt{N_n(r)}\, Z_{N_n(r)} \le x) - \Phi_{n\,,\,2}(x)\right| \le C_r \begin{cases} n^{-\min\{r,3/2\}}, & r \ne 3/2, \\ \ln(n)\, n^{-3/2}, & r = 3/2, \end{cases} \tag{35}$$

where

$$\Phi_{n\,,\,2}(x) = \Phi(x) + \frac{\sqrt{2}\, r\, \Gamma(r-1/2)}{12\, \Gamma(r)\, \sqrt{g_n}}\,(x^2-4)\varphi(x)\, \mathbb{I}_{\{r>1/2\}}(r) + \frac{x^5 - 16\,x^3 + 24\,x}{144\, g_n} \left(\frac{r}{r-1}\, \mathbb{I}_{\{r>1\}}(r) \,+\, \ln n\, \mathbb{I}_{\{r=1\}}(r)\right). \tag{36}$$

(iii) **Generalized Laplace approximation** *if* $r = 2$ *with mixed scaling factor* $g_n^{-1/2}\, N_n(2)$ *by* $Z_{N_n(2)}$

$$\sup_x \left|\mathbb{P}\left(g_n^{-1/2}\, N_n(2)\, Z_{N_n(2)} \,\le\, x\right) - L_{n;2}(x)\right| \le C_2\, n^{-3/2} \tag{37}$$

where

$$L_{n;2}(x) = L_2(x) - \frac{1}{3\,\sqrt{g_n}} \left(\frac{1}{\sqrt{2}} + \sqrt{2}|x| - x^2\right) \mathrm{e}^{-2|x|} + \frac{1}{33\, g_n}\,(12\,\sqrt{2}\,x - 15|x|\,x + 2\,x^3)\,\mathrm{e}^{-2|x|}\,. \tag{38}$$

5 The Random Dimension $N_n(s)$ is Discrete Pareto-Like Distributed

The Pareto-like distributed dimension $N_n(s)$ has probability mass function (2) and $\mathbb{E}(N_n(s)) = \infty$. Hence $g_n = n$ is chosen as normalizing sequence for $N_n(s)$.

Bening and Korolev [4, Sect. 4.3] showed that for integer $s \ge 1$

$$\lim_{n\to\infty} \sup_{y>0} |\mathbb{P}(N_n(s) \le n\, y) - H_s(y)| = 0.$$

where $H_s(y) = \mathrm{e}^{-s/y}\mathbb{I}_{(0\,,\,\infty)}(y)$ is the continuous distribution function of the inverse exponential $W(s) = 1/V(s)$ with exponentially distributed $V(s)$ having rate parameter $s > 0$. As $\mathbb{P}(N_n(s) \le y)$, so $H_s(y)$ is heavy tailed with shape parameter 1 and $\mathbb{E}W(s) = \infty$.

Lyamin [18] proved a bound $|\mathbb{P}(N_n(s) \le n\, y) - H_s(y)| \le C/n$ and $C < 0.37$ for integer $s \ge 1$.

In [9, Proposition 2] the following results are presented for $s > 0$:

$$\sup_{y>0} \left|\mathbb{P}\left(\frac{N_n(s)}{n} \le y\right) - H_s(y) - \frac{h_{2;s}(y)}{n}\right| \le \frac{C_3(s)}{n^2}, \quad \text{for all} \quad n \in \mathbb{N}_+, \tag{39}$$

with $H_s(y) = \mathrm{e}^{-s/y}$ and $h_{2;s}(y) = s\,\mathrm{e}^{-s/y}\left(s-1+2Q_1(n\,y)\right)/\left(2\,y^2\right)$ for $y > 0$, where $Q_1(y)$ is defined in (30). Moreover

$$\mathbb{E}\big(N_n(s)\big)^{-p} \le C(p)\, n^{-\min\{p,2\}}, \tag{40}$$

where for $0 < p \le 2$ the order of the bound is optimal.

The Chebyshev-Edgeworth expansion (39) is proved in Christoph, Monakhov and Ulyanov [7, Theorem 4]. The leading terms for the negative moments $\mathbb{E}\big(N_n(s)\big)^{-p}$ were derived in Christoph, Ulyanov and Bening [10, Corollary 5.2] that lead to (40).

Remark 5. The random dimension $N_n(s)$ satisfies Assumption 2 of the Transfer Propositions 2 with $H_s(y) = \mathrm{e}^{-s/y}$, $h_2(y) = h_{2;s}(y)$, $g_n = n$ and $b = 2$.

With $g_n = n$ and $\gamma \in \{1/2,\, 0,\, -1/2\}$, the limit distributions for

$$\mathbb{P}\left(n^{\gamma} N_n(s)^{1/2-\gamma} Z_{N_n(s)} \le x\right) \quad \text{are now} \quad G_\gamma(x,s) = \int_0^\infty \varPhi(x\, y^\gamma) \mathrm{d}H_s(y).$$

These scale mixtures distributions $G_\gamma(x,s)$ are calculated in Christoph and Ulyanov [9, Theorems 6–8]. We apply Proposition 2 to statistics

$$T_{N_n(s)} = N_n(s)^{1/2-\gamma}\, Z_{N_n(s)} \quad \text{with the normalizing factor } n^\gamma.$$

The limit distributions are:

- for $\gamma = 1/2$ **Laplace distributions** $L_{1/\sqrt{s}}(x)$ with density

$$l_{1/\sqrt{s}}(x) = \sqrt{s/2}\, \mathrm{e}^{-\sqrt{2\,s}|x|},$$

- for $\gamma = 0$ the **standard normal law** $\varPhi(x)$ and
- for $\gamma = -1/2$ the **scaled Student's t-distribution** $S_2^*(x; \sqrt{s})$ with density

$$s_2^*(x;\sqrt{s}) = \frac{1}{2\sqrt{2\,s}}\Big(1 + \frac{x^2}{2\,s}\Big)^{-3/2}.$$

Theorem 2. *Let Z_m and $N_n(s)$ with $s > 0$ be defined by (10) and (2), respectively. Suppose that (14) is satisfied for Z_m and (39) for $N_n(s)$. Then the following statements hold for all $n \in \mathbb{N}_+$:*

(i) **Laplace approximation** *with non-random scaling factor n^γ by $Z_{N_n(s)}$:*

$$\sup_x \left|\mathbb{P}\left(\sqrt{n}\, Z_{N_n(s)} \le x\right) - L_{1/\sqrt{s};n}(x)\right| \le C_s\, n^{-3/2} \tag{41}$$

where

$$L_{1/\sqrt{s};n}(x) = L_{1/\sqrt{s}}(x) + l_{1/\sqrt{s}}(x)\left(\frac{\sqrt{2}}{12\, s\sqrt{n}}\,(sx^2 - 2\,(1+\sqrt{2\,s}\,|x|)\right.$$
$$\left. + \frac{s}{72\, n}\left(\frac{x^3\,|x|}{\sqrt{2\,s}} - \frac{8\,x^2}{s} + \frac{6\,x}{s^2}\,(1+\sqrt{2\,s}\,|x|)\right)\right) \tag{42}$$

,

(ii) **Normal approximation** *with random scaling factor* $\sqrt{N_n(s)}$ *by* $Z_{N_n(r)}$*:*

$$\sup_x \left| \mathbb{P}\left(\sqrt{N_n(s)}\, Z_{N_n(s)} \le x \right) - \Phi_{n,2}(x) \right| \le C_s\, n^{-3/2}, \tag{43}$$

where

$$\Phi_{n,2}(x) = \Phi(x) + \varphi(x) \left(\frac{\sqrt{2\,\pi}(x^2-4)}{24\,\sqrt{n}} + \frac{x^5 - 16x^3 + 24x}{144\,s\,n} \right) \tag{44}$$

(iii) **Scaled Student's t-distribution** *with mixed scaling factor by* $Z_{N_n(s)}$

$$\sup_x \left| \mathbb{P}\left(n^{-1/2}\, N_n(s)\, Z_{N_n(s)} \le x \right) - S^*_{n;2}(x) \right| \le C_s\, n^{-3/2}, \tag{45}$$

where

$$\begin{aligned} S^*_{n;2}(x;\sqrt{s}) = S^*_2(x;\sqrt{s}) + s^*_2(x;\sqrt{s}) \Bigg(&-\frac{\sqrt{2}\,(x^2+8\,s)}{12(2\,s+x^2)\,\sqrt{n}} \\ &+ \frac{1}{144\,n}\left(\frac{105x^5}{(2\,s+x^2)^3} + \frac{240x^3}{(2\,s+x^2)^2} + \frac{72x}{2\,s+x^2} \right) \Bigg). \end{aligned} \tag{46}$$

6 Proofs of Main Results

Proof. The proofs of Theorems 1 and 2 are based on Proposition 2. The structure of the functions f_1, f_2 and $h_2$in Assumptions A and B is similar to the structure of the corresponding functions in Conditions 1 and 2 in [9]. Therefore, the estimates of the term D_n and of the integrals $I_1(x,n)$ and $I_2(x,n)$ in (23), (25) and (24) as well as the validity of (19) and (20) in Proposition 2 when $H(y)$ is $G_{r,r}(y)$ or $H_s(y)$ can be shown analogously to the proofs for Lemmas 1, 2 or 4 in [9]. In Remark 3 above it was pointed out that the integrals in (25) and (25) can degrade the convergence rate. Let $r<1$. With $|f_2(x\,y^\gamma| \le c^*$ we get

$$\int_{1/g_n}^{\infty} \frac{|f_2(x\,y^\gamma)|}{g_n\,y} \mathrm{d}G_{r,r}(y) \le \frac{c^* r^r}{\Gamma(r)\,g_n} \int_{1/g_n}^{\infty} y^{r-2} \mathrm{d}y \le \frac{c^* r^r}{(1-r)\Gamma(r)}\, g_n^{-r}. \tag{47}$$

The additional term $f_1(xy^\gamma)\,(g_n y)^{-1/2}$ in (17) in Assumption A is to be estimated with condition (19ii).

Moreover, the bounds for $\mathbb{E}(N_n)^{-3/2}$ follow from (31) and (40), since $a=3/2$ in Assumption A, considering the approximation (14).

The integrals in (22) in Proposition 2 are still to be calculated. Similar integrals are calculated in great detail in the proofs of Theorems 3–8 in [9]. To obtain (34), we compute the integrals with Formula 2.3.3.1 in Prudnikov et al. [20]

$$M_\alpha(x) = \frac{r^r}{\Gamma(r)\,\sqrt{2\pi}} \int_0^\infty y^{\alpha-1} \mathrm{e}^{-(r+x^2/2)y} dy = \frac{\Gamma(\alpha)\,r^{r-\alpha}}{\Gamma(r)\,\sqrt{2\,\pi}} \left(1+x^2/(2r)\right)^{-\alpha} \tag{48}$$

for $\alpha = r-1/2,\ r\ +\ 1/2,\ r\ +\ 3/2$ and $p = r + x^2/2$.

Lemma 2 in [9] and $\int_0^\infty y^{-1} dG_{r,r}(y) = r/(r-1)$ for $r > 1$ lead to (36).

To show (38) we use Formula 2.3.16.2 in [20] with $n = 0, 1$ and Formula 2.3.16.3 in [20] with $n = 1, 2$ and $p = 2$ and $q = x^2/2$.

To obtain (42), we calculate the integrals again with Formula 2.3.16.3 in [20], with $p = x^2/2 > 0$, $q = s > 0$, $n = 0, 1, 2$.

Lemma 4 in [9] and $\int_0^\infty y^{-a-1} \mathrm{e}^{-s/y} dy = s^{-a} \Gamma(a)$ for $a = 3/2$, 2 lead to (44).

Finally, in $\int_0^\infty f_k(x/y^\gamma) y^{-2-k/2} \mathrm{e}^{-s/y} \mathrm{d}y$ we use the substitution $s/y = u$ to obtain, with (48), the terms in (46). $\square$

Acknowledgements. Theorem 1 has been obtained under support of the Ministry of Education and Science of the Russian Federation as part of the program of the Moscow Center for Fundamental and Applied Mathematics under the agreement № 075-15-2019-1621. Theorem 2 was proved within the framework of the HSE University Basic Research Program.

References

1. Aoshima, M., Shen, D., Shen, H., Yata, K., Zhou, Y.-H., Marron, J.S.: A survey of high dimension low sample size asymptotics. Aust. N. Z. J. Stat. **60**(1), 4–19 (2018). https://doi.org/10.1111/anzs.12212
2. Bening, V.E., Galieva, N.K., Korolev, V.Y.: Asymptotic expansions for the distribution functions of statistics constructed from samples with random sizes [in Russian]. Inf. Appl. IPI RAN **7**(2), 75–83 (2013)
3. Bening, V.E., Korolev, V.Y.: On the use of Student's distribution in problems of probability theory and mathematical statistics. Theory Probab. Appl. **49**(3), 377–391 (2005)
4. Bening, V.E., Korolev, V.Y.: Some statistical problems related to the Laplace distribution [in Russian]. Inf. Appl. IPI RAN **2**(2), 19–34 (2008)
5. Bobkov, S.G., Naumov, A.A., Ulyanov V.V.: Two-sided inequalities for the density function's maximum of weighted sum of chi-square variables. arXiv:2012.10747v1 (2020). https://arxiv.org/pdf/2012.10747.pdf
6. Buddana, A., Kozubowski, T.J.: Discrete Pareto distributions. Econ. Qual. Control **29**(2), 143–156 (2014)
7. Christoph, G., Monakhov, M.M., Ulyanov, V.V.: Second-order Chebyshev-Edgeworth and Cornish-Fisher expansions for distributions of statistics constructed with respect to samples of random size. J. Math. Sci. (N.Y.) **244**(5), 811–839 (2020). Translated from Zapiski Nauchnykh Seminarov POMI, 466, Veroyatnost i Statistika. 26, 167–207 (2017)
8. Christoph, G., Prokhorov, Yu., Ulyanov, V.: On distribution of quadratic forms in Gaussian random variables. Theory Prob. Appl. **40**(2), 250–260 (1996)
9. Christoph, G., Ulyanov, V.V.: Second order expansions for high-dimension low-sample-size data statistics in random setting. Mathematics **8**(7), 1151 (2020)
10. Christoph, G., Ulyanov, V.V., Bening, V.E.: Second order expansions for sample median with random sample size. arXiv:1905.07765v2 (2020). https://arxiv.org/pdf/1905.07765.pdf

11. Christoph, G., Ulyanov, V.V., Fujikoshi, Y.: Accurate approximation of correlation coefficients by short Edgeworth-Chebyshev expansion and its statistical applications. In: Shiryaev, A.N., Varadhan, S.R.S., Presman, E.L. (eds.) Prokhorov and Contemporary Probability Theory. In Honor of Yuri V. Prokhorov. Springer Proceedings in Mathematics & Statistics, vol. 33, pp. 239–260. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-33549-5_13
12. Fujikoshi, Y., Ulyanov, V.V., Shimizu, R.: Multivariate Statistics. High-Dimensional and Large-Sample Approximations. Wiley Series in Probability and Statistics. Wiley, Hoboken (2010)
13. Gavrilenko, S.V., Zubov, V.N., Korolev, V.Y.: The rate of convergence of the distributions of regular statistics constructed from samples with negatively binomially distributed random sizes to the Student distribution. J. Math. Sci. (N.Y.) **220**(6), 701–713 (2017)
14. Hall, P., Marron, J.S., Neeman, A.: Geometric representation of high dimension, low sample size data. J. R. Stat. Soc. Ser. **67**, 427–444 (2005)
15. Johnson, N.L., Kotz, S., Balakrishnan, N.: Continuous Univariate Distributions, vol. 2, 2nd edn. Wiley, New York (1995)
16. Kawaguchi, Y., Ulyanov, V.V., Fujikoshi, Y.: Asymptotic distributions of basic statistics in geometric representation for high-dimensional data and their error bounds (Russian). Inf. Appl. **4**, 12–17 (2010)
17. Konishi, S.: Asymptotic expansions for the distributions of functions of a correlation matrix. J. Multivar. Anal. **9**, 259–266 (1979)
18. Lyamin, O.O.: On the rate of convergence of the distributions of certain statistics to the Laplace distribution. Mosc. Univ. Comput. Math. Cybern. **34**(3), 126–134 (2010)
19. Petrov, V.V.: Limit Theorems of Probability Theory. Sequences of Independent Random Variables. Clarendon Press, Oxford (1995)
20. Prudnikov, A.P., Brychkov, Y.A., Marichev, O.I.: Integrals and Series, Volume 1: Elementary Functions, 3rd edn. Gordon & Breach Science Publishers, New York (1992)
21. Schluter, C., Trede, M.: Weak convergence to the student and Laplace distributions. J. Appl. Probab. **53**(1), 121–129 (2016)