Entropy **2011**, *13*, 805-819; doi:10.3390/e13040805

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Large-Sample Asymptotic Approximations for the Sampling and Posterior Distributions of Differential Entropy for Multivariate Normal Distributions

Guillaume Marrelec [1,2,3,⋆] **and Habib Benali** [1,2,3]

[1] Inserm, U678, Paris, F-75013, France; E-Mail: benali@imed.jussieu.fr

[2] UPMC Univ Paris 06, UMR_S U678, Paris, F-75013, France

[3] Inserm, Université de Montréal, and UPMC Univ Paris 06, LINeM, Montréal, QC, H3W 1W5, Canada

⋆ Author to whom correspondence should be addressed; E-Mail: marrelec@imed.jussieu.fr.

Received: 15 February 2011; in revised form: 29 March 2011 / Accepted: 31 March 2011 / Published: 6 April 2011

Abstract: In the present paper, we propose a large sample asymptotic approximation for the sampling and posterior distributions of differential entropy when the sample is composed of independent and identically distributed realization of a multivariate normal distribution.

Keywords: differential entropy; large sample; asymptotic approximation; multivariate normal distribution; sampling distribution; posterior distribution; mutual information; multiinformation; total correlation; multivariate constraint

1. Introduction

Entropy has been an active topic of research for over 50 years and much has been published about this measure in various contexts. In statistics, recent developments have investigated how to estimate entropy from data, either in a parametric [1–3] or nonparametric framework [4,5], as well as the reliability and convergence properties of these estimators [6,7].

By contrast, relatively little is known about the statistical distribution of entropy, even in the simple case of a multivariate normal distribution. For instance, the differential entropy $H(X)$ of a

D-dimensional random variable X that is normally distributed with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is given by

$$H(X) = h(\boldsymbol{\Sigma}) = \frac{D}{2}\left[1 + \ln(2\pi)\right] + \frac{1}{2}\ln|\boldsymbol{\Sigma}| \tag{1}$$

If $(\boldsymbol{x}_n)_{n=1,\ldots,N}$ are N independent and identically distributed realizations of X and $\boldsymbol{S}$ the corresponding sum of square, then the sample differential entropy $h(\boldsymbol{S}/N)$ is used as the so-called plug-in estimator for $H(X)$. However, $h(\boldsymbol{S}/N)$ is also a random variable whose sampling distribution could be studied. Ahmed *et al.* provided the exact expression for the mean and variance of this variable [1]. Similarly, in a Bayesian framework, given $h(\boldsymbol{S}/N)$, what are the probable values of $h(\boldsymbol{\Sigma})$? We are not aware of any study in this direction for multivariate normal distributions (but see, e.g., [8,9] for the posterior moments of entropy in the case of multinomial distributions). In the present paper, we provide an asymptotic approximation for both the sampling distribution of $h(\boldsymbol{S}/N)$ and, in a Bayesian framework, the posterior distribution of $h(\boldsymbol{\Sigma})$ given $h(\boldsymbol{S}/N)$. To this aim, we first calculate the moments of $|\boldsymbol{S}|/|\nu\boldsymbol{\Sigma}|$ in the same condition as above. We then use this result to provide a closed form expression for the cumulant-generating function of $U = -\ln(|\boldsymbol{S}|/|\nu\boldsymbol{\Sigma}|)$, from which we derive closed form expressions for the cumulants, together with asymptotic expansions when $\nu \to \infty$. Using the characteristic function of U, we then provide an asymptotic normal approximation for the distribution of this variable. We finally apply these result to the sample and posterior entropy of multivariate normal distributions.

2. General Result

Assume that $\boldsymbol{S}$ is distributed according to a Wishart distribution with $\nu \geq D$ degrees of freedom and scale matrix $\boldsymbol{\Sigma}$, *i.e.*, [10] (Chapter 7)

$$\mathrm{p}(\boldsymbol{S}|\boldsymbol{\Sigma},\nu) = \frac{1}{Z_D(\nu)}|\boldsymbol{\Sigma}|^{-\frac{\nu}{2}}|\boldsymbol{S}|^{\frac{\nu-D-1}{2}}\exp\left[-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S})\right]$$

where $Z_D(\nu)$ is the normalizing constant,

$$Z_D(\nu) = 2^{\frac{\nu D}{2}}\pi^{\frac{D(D-1)}{4}}\prod_{d=1}^{D}\Gamma\left(\frac{\nu+1-d}{2}\right) \tag{2}$$

Direct calculation show that we have, for $t \in \mathbb{R}$,

$$\begin{aligned}
\mathrm{E}\left[\left(\frac{|\boldsymbol{S}|}{|\nu\boldsymbol{\Sigma}|}\right)^t\right] &= \int\left(\frac{|\boldsymbol{S}|}{|\nu\boldsymbol{\Sigma}|}\right)^t\cdot\frac{1}{Z_D(\nu)}|\boldsymbol{\Sigma}|^{-\frac{\nu}{2}}|\boldsymbol{S}|^{\frac{\nu-D-1}{2}}\exp\left[-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S})\right]\mathrm{d}\boldsymbol{S} \\
&= \frac{Z_D(\nu+2t)}{Z_D(\nu)}\nu^{-Dt}\int\frac{1}{Z_D(\nu+2t)}|\boldsymbol{\Sigma}|^{-\frac{\nu+2t}{2}}|\boldsymbol{S}|^{\frac{(\nu+2t)-D-1}{2}}\exp\left[-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S})\right]\mathrm{d}\boldsymbol{S} \\
&= \frac{Z_D(\nu+2t)}{Z_D(\nu)}\nu^{-Dt} \qquad (3)
\end{aligned}$$

provided that the integral sums to one, *i.e.*, $\nu + 2t \geq D$ or, equivalently, $t \geq (D-\nu)/2$.

2.1. *Cumulant-Generating Function, Cumulants, and Central Moments of U*

Cumulant-generating function Let U be the function defined in the introduction, *i.e.*,

$$U = -\ln\frac{|\boldsymbol{S}|}{|\nu\boldsymbol{\Sigma}|} \tag{4}$$

and $g_U(t) = \ln \mathrm{E}\left[e^{tU}\right]$ its cumulant-generating function. $g_U(t)$ is the log of the quantity calculated in Equation (3)

$$g_U(t) = Dt \ln \nu + \ln Z_D(\nu - 2t) - \ln Z_D(\nu) \tag{5}$$

$\ln Z_D(\nu)$ and $\ln Z_D(\nu - 2t)$ can be expressed using Equation (2), leading to

$$g_U(t) = Dt \ln \frac{\nu}{2} + \sum_{d=1}^{D} \ln \Gamma \left(\frac{\nu - 2t + 1 - d}{2} \right) - \sum_{d=1}^{D} \ln \Gamma \left(\frac{\nu + 1 - d}{2} \right) \tag{6}$$

Cumulants By construction, the nth cumulant of U is given by $\kappa_n = g_U^{(n)}(0)$. In the present case, $g_U^{(n)}(t)$ can be obtained by direct derivation, yielding for the cumulants

$$\kappa_1 = g'_U(0) = D \ln \frac{\nu}{2} - \sum_{d=1}^{D} \psi \left(\frac{\nu + 1 - d}{2} \right) \tag{7}$$

and

$$\kappa_n = g_U^{(n)}(0) = (-1)^n \sum_{d=1}^{D} \psi^{(n-1)} \left(\frac{\nu + 1 - d}{2} \right) \tag{8}$$

for $n \geq 2$, where ψ is the digamma function, *i.e.*, $\psi(t) = d[\ln \Gamma(t)]/dt$, and $\psi^{(n)}$ its nth derivative [11] (pp. 258–260). For any $n \geq 1$, κ_n is always strictly positive. It is an increasing function of D and a decreasing function of ν. It tends to 0 when ν tends to infinity. For a proof of these properties, see the appendix.

Central moments Cumulants and central moments are related as follows: If we denote by μ, σ^2, γ and γ_2 the mean, variance, skewness and excess kurtosis of U, respectively, we have $\mu = \kappa_1$, $\sigma^2 = \kappa_2$, $\gamma_1 = \kappa_3/\kappa_2^{3/2}$, and $\gamma_2 = \kappa_4/\kappa_2^2$. Note that, by definition, μ is equal to the expression of Equation (7) and σ^2 to that of Equation (8) with $n = 2$.

2.2. *Asymptotic Expansion*

When ν is large, ψ can be approximated using the following asymptotic expansion [11] (p. 260)

$$\psi(z) = \ln z - \frac{1}{2z} - \frac{1}{12z^2} + O\left(\frac{1}{z^3}\right)$$

where $O(1/z^n)$ refers to Landau notation and stands for any function $f(z)$ for which there exists z_0 so that $z^n f(z)$ is bounded for $z \geq z_0$. This leads to

$$\begin{aligned}
\psi\left(\frac{\nu+1-d}{2}\right) &= \ln\left(\frac{\nu+1-d}{2}\right) - \frac{1}{\nu+1-d} - \frac{1}{3(\nu+1-d)^2} + O\left(\frac{1}{\nu^3}\right) \\
&= \ln\frac{\nu}{2} + \ln\left(1 + \frac{1-d}{\nu}\right) - \frac{1}{\nu\left(1+\frac{1-d}{\nu}\right)} - \frac{1}{3\nu^2\left(1+\frac{1-d}{\nu}\right)^2} + O\left(\frac{1}{\nu^3}\right) \\
&= \ln\frac{\nu}{2} + \left[\frac{1-d}{\nu} - \frac{1}{2}\left(\frac{1-d}{\nu}\right)^2\right] - \frac{1}{\nu}\left(1 - \frac{1-d}{\nu}\right) - \frac{1}{3\nu^2} + O\left(\frac{1}{\nu^3}\right) \\
&= \ln\frac{\nu}{2} - \frac{d}{\nu} + \frac{1-3d^2}{6\nu^2} + O\left(\frac{1}{\nu^3}\right)
\end{aligned}$$

Incorporating this expansion in Equation (7) yields for the first cumulant κ_1 or, equivalently, the mean μ

$$\kappa_1 = \mu = \frac{D(D+1)}{2\nu} + \frac{2D^3 + 3D^2 - D}{12\nu^2} + O\left(\frac{1}{\nu^3}\right) \tag{9}$$

For the cumulants and central moments of order 2 and up, we use the following approximation of $\psi^{(n)}$ [11] (p. 260)

$$\psi^{(n)}(z) = (-1)^{n-1}\left[\frac{(n-1)!}{z^n} + \frac{n!}{2z^{n+1}} + O\left(\frac{1}{z^{n+2}}\right)\right] \tag{10}$$

Each term in the sum of Equation (8) can therefore be approximated as

$$\begin{aligned}
\psi^{(n-1)}\left(\frac{\nu+1-d}{2}\right) &= (-1)^{n-2}\left[\frac{2^{n-1}(n-2)!}{\nu^{n-1}\left(1+\frac{1-d}{\nu}\right)^{n-1}} + \frac{2^{n-1}(n-1)!}{\nu^n\left(1+\frac{1-d}{\nu}\right)^n} + O\left(\frac{1}{\nu^{n+1}}\right)\right] \\
&= (-1)^{n-2}\left[\frac{2^{n-1}(n-2)!}{\nu^{n-1}}\left(1 - \frac{(n-1)(1-d)}{\nu}\right) + \frac{2^{n-1}(n-1)!}{\nu^n} + O\left(\frac{1}{\nu^{n+1}}\right)\right] \\
&= (-1)^{n-2}\left[\frac{2^{n-1}(n-2)!}{\nu^{n-1}} + \frac{2^{n-1}(n-1)!d}{\nu^n} + O\left(\frac{1}{\nu^{n+1}}\right)\right]
\end{aligned}$$

leading to an approximation of $\kappa_n = g_U^{(n)}(0)$ of the form

$$\kappa_n = \frac{2^{n-1}D(n-2)!}{\nu^{n-1}} + \frac{2^{n-1}D(D+1)(n-1)!}{2\nu^n} + O\left(\frac{1}{\nu^{n+1}}\right) \tag{11}$$

Taking n equal to 2, 3, and 4 respectively yields for the cumulants of order 2, 3, and 4

$$\kappa_2 = \frac{2D}{\nu} + \frac{D(D+1)}{\nu^2} + O\left(\frac{1}{\nu^3}\right) \tag{12}$$

$$\kappa_3 = \frac{4D}{\nu^2} + \frac{4D(D+1)}{\nu^3} + O\left(\frac{1}{\nu^4}\right) \tag{13}$$

$$\kappa_4 = \frac{16D}{\nu^3} + \frac{24D(D+1)}{\nu^4} + O\left(\frac{1}{\nu^5}\right) \tag{14}$$

We can now provide asymptotic approximations for the corresponding central moments. The variance $\sigma^2 = \kappa_2$ is given by Equation (12). Approximation for the skewness $\gamma_1 = \kappa_3/\kappa_2^{3/2}$ can be obtained from Equations (12) and (13) as

$$\begin{aligned}
\gamma_1 &= \frac{4D}{\nu^2}\left[1 + \frac{D+1}{\nu} + O\left(\frac{1}{\nu^2}\right)\right]\left(\frac{2D}{\nu}\right)^{-\frac{3}{2}}\left[1 + \frac{D+1}{2\nu} + O\left(\frac{1}{\nu^2}\right)\right]^{-\frac{3}{2}} \\
&= \sqrt{\frac{2}{D\nu}}\left[1 + \frac{D+1}{4\nu} + O\left(\frac{1}{\nu^2}\right)\right]
\end{aligned}$$

γ_1 being asymptotically positive, the distribution is skewed on the right. Finally, the approximation for $\gamma_2 = \kappa_4/\kappa_2^2$ can be expressed as

$$\begin{aligned}
\gamma_2 &= \frac{16D}{\nu^3}\left[1 + \frac{3(D+1)}{2\nu} + O\left(\frac{1}{\nu^2}\right)\right]\left(\frac{2D}{\nu}\right)^{-2}\left[1 + \frac{D+1}{2\nu} + O\left(\frac{1}{\nu^2}\right)\right]^{-2} \\
&= \frac{4}{D\nu}\left(1 + \frac{D+1}{2\nu}\right) + O\left(\frac{1}{\nu^3}\right)
\end{aligned}$$

which is asymptotically positive, corresponding to a leptokurtic distribution.

2.3. Asymptotic Distribution of U

We now use the previous results to prove that U is asymptotically normally distributed with mean $D(D+1)/2\nu$ and variance $2D/\nu$. To this aim, set

$$V_\nu = \frac{U - \frac{a}{\nu}}{\frac{b}{\sqrt{\nu}}} \tag{15}$$

with $a = D(D+1)/2$ and $b = \sqrt{2D}$. The logarithm of the characteristic function of V_ν reads

$$\begin{aligned}
\ln \phi_{V_\nu}(t) &= \ln \mathrm{E}\left\{ \exp\left[it \left(\frac{U - \frac{a}{\nu}}{\frac{b}{\sqrt{\nu}}} \right) \right] \right\} \\
&= -\frac{ita}{b\sqrt{\nu}} + \ln \mathrm{E}\left\{ \exp\left[\left(\frac{it\sqrt{\nu}}{b} \right) U \right] \right\} \\
&= -\frac{ita}{b\sqrt{\nu}} + \ln \phi_U\left(\frac{it\sqrt{\nu}}{b} \right) \\
&= \ln \phi_U\left(\frac{it\sqrt{\nu}}{b} \right) + O\left(\frac{1}{\sqrt{\nu}} \right)
\end{aligned}$$

where $\phi_U(t)$ is the characteristic function of U. We proved Equation (3) as an analytic identity for $t \in \mathbb{R}$. This expression will, however, be valid in the range where $Z_D(\nu + 2t)$ is analytic. We can thus obtain an expression for $\phi_U(it\sqrt{\nu}/b)$ by replacing t by $-it\sqrt{\nu}/b$ in Equation (3), leading to

$$\begin{aligned}
\ln \phi_U\left(\frac{it\sqrt{\nu}}{b} \right) &= \ln \left[\frac{Z_D\left(\nu - \frac{2it\sqrt{\nu}}{b} \right)}{Z_D(\nu)} \right] + \frac{itD\sqrt{\nu}\ln\nu}{b} \\
&= \ln \left[\frac{2^{\frac{\left(\nu - \frac{2it\sqrt{\nu}}{b}\right)D}{2}} \pi^{\frac{D(D-1)}{2}} \prod_{d=1}^{D} \Gamma\left(\frac{\nu - \frac{2it\sqrt{\nu}}{b} + 1 - d}{2} \right)}{2^{\frac{\nu D}{2}} \pi^{\frac{D(D-1)}{2}} \prod_{d=1}^{D} \Gamma\left(\frac{\nu+1-d}{2} \right)} \right] + \frac{itD\sqrt{\nu}\ln\nu}{b} \\
&= \frac{itD\sqrt{\nu}}{b} \ln\frac{\nu}{2} + \sum_{d=1}^{D} \ln \left[\frac{\Gamma\left(\frac{\nu - \frac{2it\sqrt{\nu}}{b} + 1 - d}{2} \right)}{\Gamma\left(\frac{\nu+1-d}{2} \right)} \right]
\end{aligned} \tag{16}$$

We then use Stirling's approximation [11] (p. 257)

$$\ln\Gamma(z) = \left(z - \frac{1}{2} \right) \ln z - z + \frac{1}{2}\ln 2\pi + O\left(\frac{1}{z} \right)$$

to approximate each term of the sum in the second term of the right-hand side of Equation (16) when ν is large, yielding

$$\begin{aligned}
\ln\left[\frac{\Gamma\left(\frac{\nu-\frac{2it\sqrt{\nu}}{b}+1-d}{2}\right)}{\Gamma\left(\frac{\nu+1-d}{2}\right)}\right] &= \frac{\nu-\frac{2it\sqrt{\nu}}{b}-d}{2}\ln\left(\frac{\nu-\frac{2it\sqrt{\nu}}{b}+1-d}{2}\right)-\frac{\nu-\frac{2it\sqrt{\nu}}{b}+1-d}{2} \\
&\quad -\frac{\nu-d}{2}\ln\left(\frac{\nu+1-d}{2}\right)+\frac{\nu+1-d}{2}+O\left(\frac{1}{\sqrt{\nu}}\right) \\
&= \frac{\nu-\frac{2it\sqrt{\nu}}{b}-d}{2}\left[\ln\frac{\nu}{2}+\ln\left(1-\frac{2it}{b\sqrt{\nu}}+\frac{1-d}{\nu}\right)\right]+\frac{it\sqrt{\nu}}{b} \\
&\quad -\frac{\nu-d}{\nu}\left[\ln\frac{\nu}{2}+\ln\left(1+\frac{1-d}{\nu}\right)\right]+O\left(\frac{1}{\sqrt{\nu}}\right) \\
&= -\frac{it\sqrt{\nu}}{b}\ln\frac{\nu}{2}+\frac{it\sqrt{\nu}}{b} \\
&\quad +\frac{\nu-2it\frac{\sqrt{\nu}}{b}-d}{2}\left[-\frac{2it}{b\sqrt{\nu}}+\frac{1-d}{\nu}+\frac{2t^2}{b^2\nu}+O\left(\frac{1}{\nu^{3/2}}\right)\right] \\
&\quad -\frac{\nu-d}{2}\left[\frac{1-d}{\nu}+O\left(\frac{1}{\nu^{3/2}}\right)\right]+O\left(\frac{1}{\sqrt{\nu}}\right) \\
&= -\frac{it\sqrt{\nu}}{b}\ln\frac{\nu}{2}-\frac{t^2}{b^2}+O\left(\frac{1}{\sqrt{\nu}}\right)
\end{aligned}$$

We consequently have for the characteristic moment of V_ν

$$\begin{aligned}
\ln\phi_{V_\nu}(t) &= \ln\phi_U\left(it\frac{\sqrt{\nu}}{b}\right)+O\left(\frac{1}{\sqrt{\nu}}\right) \\
&= -\frac{Dt^2}{b^2}+O\left(\frac{1}{\sqrt{\nu}}\right) \\
&= -\frac{t^2}{2}+O\left(\frac{1}{\sqrt{\nu}}\right)
\end{aligned}$$

As ν tends towards infinity, $\phi_{V_\nu}(t)$ achieves pointwise convergence toward $e^{-t^2/2}$, which is continuous in $t=0$. According to Lévi's continuity theorem, V_ν therefore converges in distribution to the standard normal distribution,

$$V_\nu=\frac{U-\frac{D(D+1)}{2\nu}}{\sqrt{\frac{2D}{\nu}}}\overset{\nu\to\infty}{\sim}\mathcal{N}(0,1)$$

3. Application to Differential Entropy

We can use the results of the previous section to obtain the exact and asymptotic cumulants of the sample and posterior entropy when the data are multivariate normal.

3.1. Sampling Distribution

The differential entropy $H(X)$ of a D-dimensional random variable X that is normally distributed with (known) mean $\boldsymbol{\mu}$ and (unknown) covariance matrix $\boldsymbol{\Sigma}$ is given by Equation (1). Let $(\boldsymbol{x}_n)_{n=1,\ldots,N}$ be N independent and identically distributed realizations of X. Set $\boldsymbol{S}$ the sum of square, *i.e.*,

$$\boldsymbol{S} = \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^{\mathrm{t}} \tag{17}$$

$\boldsymbol{S}$ follows a Wishart distribution with $\nu = N$ degrees of freedom and scale matrix $\boldsymbol{\Sigma}$ [12] (Th. 7.2.2). Define the sample differential entropy corresponding to the N realizations as $h(\boldsymbol{S}/N)$. Using the fact that $|\boldsymbol{S}/N|/|\boldsymbol{\Sigma}| = |\boldsymbol{S}|/|N\boldsymbol{\Sigma}|$, we obtain that $h(\boldsymbol{S}/N) - h(\boldsymbol{\Sigma}) = -U/2$, where U was defined in Equation (4). The mean and variance of $h(\boldsymbol{S}/N) = h(\boldsymbol{\Sigma}) - U/2$ can therefore be expressed as functions of the corresponding central moments of U, *i.e.*, $\mu = \kappa_1$ [Equations (7) and (9)] and $\sigma^2 = \kappa_2$ [Equations (6) and (12)], leading to the following closed form expressions and approximations

$$\begin{aligned}
\mathrm{E}[h(\boldsymbol{S}/N)|N, \boldsymbol{\Sigma}] &= h(\boldsymbol{\Sigma}) - \frac{\mu}{2} \\
&= h(\boldsymbol{\Sigma}) - \frac{D}{2}\ln\frac{N}{2} + \frac{1}{2}\sum_{d=1}^{D}\psi\left(\frac{N+1-d}{2}\right) && (18) \\
&= h(\boldsymbol{\Sigma}) - \frac{D(D+1)}{4N} - \frac{2D^3+3D^2-D}{24N^2} + O\left(\frac{1}{N^3}\right) && (19)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}[h(\boldsymbol{S}/N)|\nu, \boldsymbol{\Sigma}] &= \frac{\sigma^2}{4} \\
&= \frac{1}{4}\sum_{d=1}^{D}\psi'\left(\frac{N+1-d}{2}\right) && (20) \\
&= \frac{D}{2N} + \frac{D(D+1)}{4N^2} + O\left(\frac{1}{N^3}\right) && (21)
\end{aligned}$$

Furthermore, use of Section 2.3 shows that, given N and $\boldsymbol{\Sigma}$, $h(\boldsymbol{S}/N)$ is asymptotically normally distributed with mean $-D(D+1)/4N$ and variance $D/2N$. If $\boldsymbol{\mu}$ is unknown, we replace $\boldsymbol{\mu}$ by the sample mean $\boldsymbol{m}$ in Equation (17). $\boldsymbol{S}$ is then still Wishart distributed with scale matrix $\boldsymbol{\Sigma}$ but $\nu = N-1$ degrees of freedom [12] (Cor. 7.2.2). The exact expectation and variance of $h[\boldsymbol{S}/(N-1)]$ are therefore given by Equations (18) and (20), respectively where N is replaced by $N-1$. Performing asymptotic expansion of this expression leads to

$$\mathrm{E}\left\{h\left[\boldsymbol{S}/(N-1)\right]|N, \boldsymbol{\Sigma}\right\} = h(\boldsymbol{\Sigma}) - \frac{D(D+1)}{4N} - \frac{2D^3+9D^2+5D}{24N^2} + O\left(\frac{1}{N^3}\right)$$

and

$$\mathrm{Var}\left\{h\left[\boldsymbol{S}/(N-1)\right]|\nu, \boldsymbol{\Sigma}\right\} = \frac{D}{2N} + \frac{D(D+3)}{4N^2} + O\left(\frac{1}{N^3}\right)$$

Furthermore, since the first-order approximation is the same for $h[\boldsymbol{S}/(N-1)]$ for $h(\boldsymbol{S}/N)$, both quantities have the same asymptotic distribution.

3.2. Posterior Distribution

With the same assumptions as above, and assuming a non-informative Jeffreys prior for Σ, *i.e.*,

$$\mathrm{p}(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{D+1}{2}}$$

the posterior distribution for $\boldsymbol{\Sigma}$ given the N realizations of X is inverse Wishart with $n = N - 1$ degrees of freedom and scale matrix $\boldsymbol{S}^{-1}$ [13]. This implies that $\boldsymbol{\Upsilon} = \boldsymbol{\Sigma}^{-1}$, the concentration matrix, is Wishart distributed with n degrees of freedom and scale matrix $\boldsymbol{S}^{-1}$. Results of Section 3.1 therefore apply to $h(\boldsymbol{\Upsilon}/n) - h(\boldsymbol{S}^{-1})$. But, since for any matrix $\boldsymbol{A}$, $\ln|\boldsymbol{A}^{-1}| = \ln|\boldsymbol{A}|^{-1} = -\ln|\boldsymbol{A}|$, we have that $h(\boldsymbol{\Upsilon}/n) - h(\boldsymbol{S}^{-1})$ is equal to $h(\boldsymbol{S}) - h(n\boldsymbol{\Sigma})$ or, equivalently, to $h(\boldsymbol{S}/n) - h(\boldsymbol{\Sigma})$. As a consequence,

$$\begin{aligned} \mathrm{E}[h(\boldsymbol{\Sigma})|N, \boldsymbol{S}] &= h(\boldsymbol{S}/n) + \frac{D}{2}\ln\frac{\nu}{2} - \frac{1}{2}\sum_{d=1}^{D}\psi\left(\frac{N-d}{2}\right) && (22)\\ &= h(\boldsymbol{S}/n) + \frac{D(D+1)}{4N} + \frac{2D^3 + 9D^2 + 5D}{24N^2} + O\left(\frac{1}{N^3}\right) && (23) \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}[h(\boldsymbol{\Sigma})|n, \boldsymbol{S}] &= \frac{1}{4}\sum_{d=1}^{D}\psi'\left(\frac{N-d}{2}\right)\\ &= \frac{D}{2N} + \frac{D(D+3)}{4N^2} + O\left(\frac{1}{N^3}\right) \end{aligned}$$

Also, $h(\boldsymbol{\Sigma})$ is asymptotically normally distributed with mean $D(D+1)/4N$ and variance $D/2N$.

4. Application to Mutual Information and Multiinformation

Similar results can also be derived about the first cumulant of mutual information and multiinformation, its generalization to more than two variables. The mutual information between two sets of variables X_1 (of dimension D_1) and X_2 (of dimension D_2) is defined as

$$I(X_1, X_2) = H(X_1) + H(X_2) - H(X_1, X_2)$$

For multivariate normal variables, we have

$$I(X_1, X_2) = i(\boldsymbol{\Sigma}) = h(\boldsymbol{\Sigma}_1) + h(\boldsymbol{\Sigma}_2) - h(\boldsymbol{\Sigma}) \qquad (24)$$

where $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are the two block diagonal elements of $\boldsymbol{\Sigma}$ and where h was defined in Equation (1).

4.1. Sampling Mean

Define the sample mutual information as $i(\boldsymbol{S}/N)$. Using Equation (24), direct calculation shows that we have

$$\mathrm{E}[i(\boldsymbol{S}/N)|N, \boldsymbol{\Sigma}] = \mathrm{E}[h(\boldsymbol{S}_1/N)|N, \boldsymbol{\Sigma}] + \mathrm{E}[h(\boldsymbol{S}_2/N)|N, \boldsymbol{\Sigma}] - \mathrm{E}[h(\boldsymbol{S}/N)|N, \boldsymbol{\Sigma}]$$

An asymptotic approximation for $\mathrm{E}[h(\boldsymbol{S}/N)|N, \boldsymbol{\Sigma}]$ can be obtained by direct use of Equation (19). For $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$, we proceed as follows. If $\boldsymbol{S}$ is Wishart distributed with N degrees of freedom and scale matrix $\boldsymbol{\Sigma}$, then $\boldsymbol{S}_j$ ($j \in \{1,2\}$) is also Wishart distributed with N degrees of freedom and scale matrix $\boldsymbol{\Sigma}_j$ [12] (Th. 7.3.4). Equation (19) can therefore be applied to matrix $\boldsymbol{S}_j$ with the proper scale matrix, yielding

$$\begin{aligned} \mathrm{E}[h(\boldsymbol{S}_j/N)|N,\boldsymbol{\Sigma}] &= \mathrm{E}[h(\boldsymbol{S}_j/N)|N,\boldsymbol{\Sigma}_j] \\ &= h(\boldsymbol{\Sigma}_j) - \frac{D_j(D_j+1)}{4N} - \frac{2D_j^3+3D_j^2-D_j}{24N^2} + O\left(\frac{1}{N^3}\right) \end{aligned}$$

$\mathrm{E}[i(\boldsymbol{S}/N)|N,\boldsymbol{\Sigma}]$ consequently reads

$$\mathrm{E}[i(\boldsymbol{S}/N)|N,\boldsymbol{\Sigma}] = i(\boldsymbol{\Sigma}) + \frac{D_1D_2}{2N}\left[1+\frac{D_1+D_2+1}{2N}\right] + O\left(\frac{1}{N^3}\right)$$

A similar result can be obtained for the generalization of i to K sets of variables X_k (of size D_k) as a measure called total correlation [14], multivariate constraint [15], δ [16], or multiinformation [17]. In that case, we have

$$\mathrm{E}[i(\boldsymbol{S}/N)|N,\boldsymbol{\Sigma}] = i(\boldsymbol{\Sigma}) + \frac{\sum_{i<j} D_iD_j}{2N} + \frac{\sum_{i\neq j} D_iD_j\left(D_i+\sum_{k\neq i,j} D_k+1\right)}{4N^2} + O\left(\frac{1}{N^3}\right)$$

and, in the particular case where each X_k is one-dimensional (*i.e.*, $D_k = 1$),

$$\mathrm{E}[i(\boldsymbol{S}/N)|N,\boldsymbol{\Sigma}] = i(\boldsymbol{\Sigma}) + \frac{D(D-1)}{4N} + \frac{2D^3+3D^2-5D}{24N^2} + O\left(\frac{1}{N^3}\right)$$

4.2. Posterior Mean

A similar argument can be applied to the Bayesian posterior mean of i. Using Equation (24) again, we have

$$\mathrm{E}[i(\boldsymbol{\Sigma})|N,\boldsymbol{S}] = \mathrm{E}[h(\boldsymbol{\Sigma}_1)|N,\boldsymbol{S}] + \mathrm{E}[h(\boldsymbol{\Sigma}_2)|N,\boldsymbol{S}] - \mathrm{E}[h(\boldsymbol{\Sigma})|N,\boldsymbol{S}]$$

An asymptotic approximation for $\mathrm{E}[h(\boldsymbol{\Sigma})|N,\boldsymbol{S}]$ can be obtained by direct use of Equation (23). Now, if $\boldsymbol{\Sigma}$ is inverse Wishart distributed with n degrees of freedom and scale matrix $\boldsymbol{S}$, then $\boldsymbol{\Sigma}_j$ ($j \in \{1,2\}$) is also inverse Wishart distributed with $n - D_k$ ($k \in \{1,2\}$, $k \neq j$) degrees of freedom and scale matrix $\boldsymbol{S}_j$ [18]. Application of Equation (23) with the proper degrees of freedom and scale matrix leads to

$$\begin{aligned} \mathrm{E}[h(\boldsymbol{\Sigma}_j)|N,\boldsymbol{S}] &= h[\boldsymbol{S}_j/(n-D_k)] + \frac{D_j(D_j+1)}{4(N-D_k)} + O\left(\frac{1}{N^2}\right) \\ &= h(\boldsymbol{S}_j/n) - \frac{D_j}{2}\ln\left(1-\frac{D_k}{N}\right) + \frac{D_j(D_j+1)}{4N} + O\left(\frac{1}{N^2}\right) \\ &= h[\boldsymbol{S}_j/n] + \frac{D_1D_2}{2N} + \frac{D_j(D_j+1)}{4N} + O\left(\frac{1}{N^2}\right) \end{aligned}$$

where we only retained the expansion terms of order up to $1/N$ for the sake of simplicity. $\mathrm{E}[i(\boldsymbol{\Sigma})|N,\boldsymbol{S}]$ consequently reads

$$\mathrm{E}[i(\boldsymbol{\Sigma})|N,\boldsymbol{S}] = i(\boldsymbol{S}/n) + \frac{D_1D_2}{2N} + O\left(\frac{1}{N^2}\right)$$

For posterior multiinformation, we have

$$\mathrm{E}[i(\boldsymbol{\Sigma})|N,\boldsymbol{S}] = i(\boldsymbol{S}/n) + \frac{\sum_{i<j} D_iD_j}{2N} + O\left(\frac{1}{N^2}\right)$$

and, in the particular case where each X_k is one-dimensional (*i.e.*, $D_k = 1$),

$$\mathrm{E}[i(\mathbf{\Sigma})|N, \boldsymbol{S}] = i(\boldsymbol{S}/n) + \frac{D(D-1)}{4N} + O\left(\frac{1}{N^2}\right)$$

5. Simulation Study

We conducted the following computations for $D \in \{2, 5, 10\}$. To assess the accuracy of the asymptotic expansion of the cumulants of sample entropy, we calculated the error made by the first and second central moments (*i.e.*, the mean and variance of the distribution) compared to the exact values as a function of ν. As a way of comparison, we computed the same quantities for 500 different homogeneous positive definite matrices $\mathbf{\Sigma}$ (*i.e.*, with all non-diagonal elements equal to the same value ρ, generated uniformly); for each value of $\mathbf{\Sigma}$ and ν, we generated 1,000 samples from $\boldsymbol{S} \sim \mathrm{Wishart}(\nu, \mathbf{\Sigma})$, computed the corresponding values of sample entropy, and approximated the moments by the corresponding sampling moments. The results are reported in Figure 1.

6. Discussion

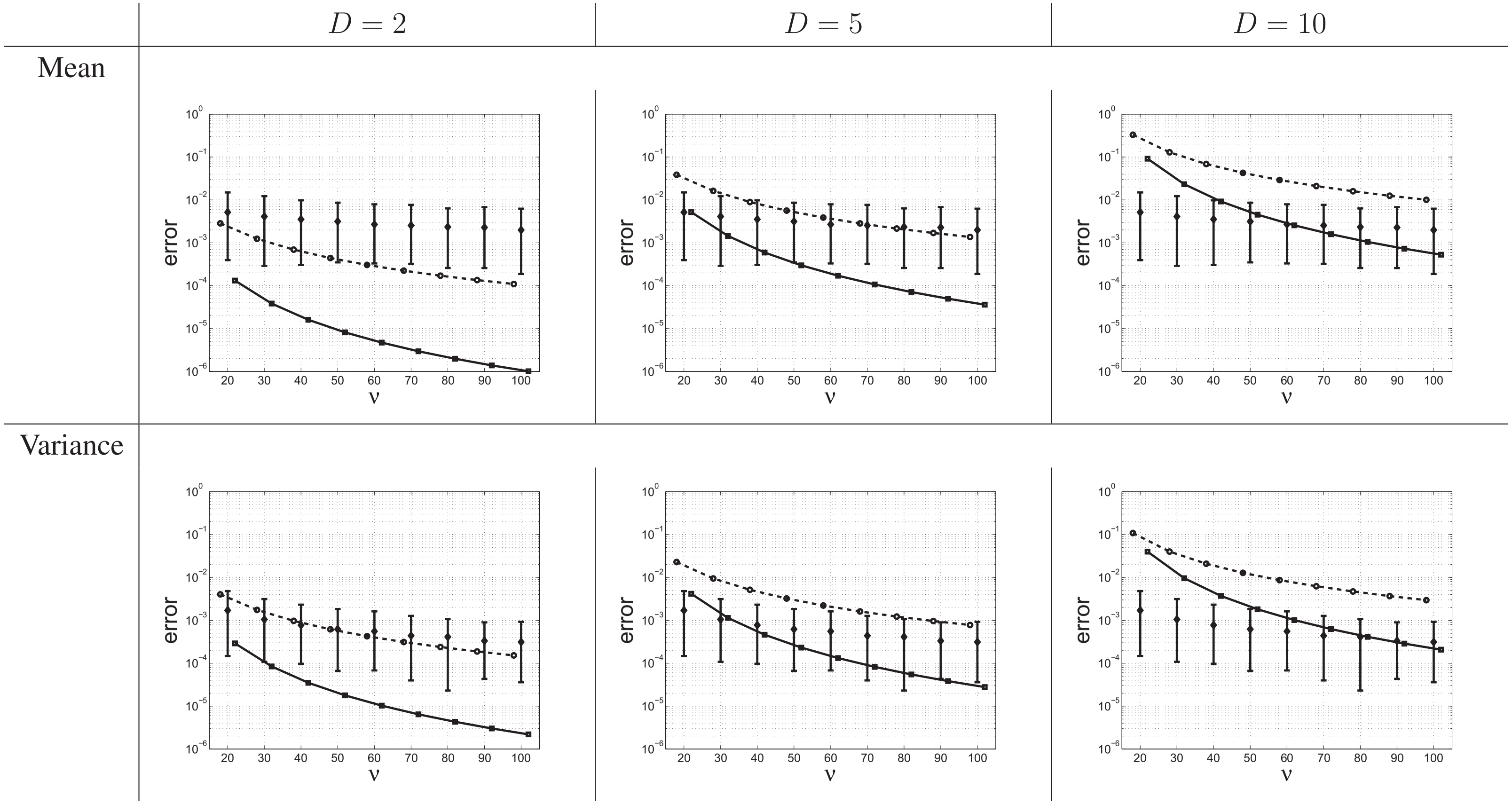
In this work, we calculated both the moments of $|\boldsymbol{S}|/|\nu\mathbf{\Sigma}|$ and the cumulant-generating function of $U = -\ln(|\boldsymbol{S}|/|\nu\mathbf{\Sigma}|)$ when $\boldsymbol{S}$ is Wishart distributed with ν degrees of freedom and scale matrix $\mathbf{\Sigma}$. From there, we provided an asymptotic approximation of the first four central moments of U. We also proved that U is asymptotically normally distributed. We then demonstrated the quality of the normal approximation compared to simulations. We finally applied these results to the multivariate normal distribution to provide asymptotic approximations of the sample and posterior distributions of differential entropy, as well as an asymptotic approximation of the sample and posterior mean of multiinformation.

Interestingly, the moments of $|\boldsymbol{S}|/|\nu\mathbf{\Sigma}|$ and, as a way of consequence, the cumulant-generating function of U depends on the distribution that $\boldsymbol{S}$ follows only through the matrix dimension D and the degree of freedom ν, but not through $\mathbf{\Sigma}$. This means that the exact distribution of U is also independent from that parameter and could possibly be tabulated as a function of the two integer parameters.

As mentioned in the introduction, the sample differential entropy defined in Equation (1) is equal to the plug-in estimator for differential entropy. The present work provides a quantification in the case of multivariate normal samples for the well-known negative bias for this estimator [7]. Obviously, Equation (18) confirms that, to correct from this bias, one must take the uniformly minimum variance unbiased (UMVU) estimator [1].

The posterior derivation that we presented here is a particular case of the Bayesian posterior estimate obtained by [3] with, in our case, the prior distribution for $\mathbf{\Sigma}$ taken as Jeffreys prior (*i.e.*, $q = -1$ and $\boldsymbol{B} = \mathbf{0}$ with their notations). While the same analysis as in [3] could have been performed, it would essentially lead to the same result, since we only consider the asymptotic case, where the sample is large and the prior distribution is supposed to have very little influence—provided that it does not contradict the data. The present study also shows an interesting feature of Bayesian estimation with respect to the above-mentioned negative bias. As the sample differential entropy tends to underestimate $H(\mathbf{\Sigma})$ by a factor of $m/2$, if one takes the posterior mean as the Bayesian estimate of $H(\mathbf{\Sigma})$, then the negative bias is corrected by the opposite factor.

Figure 1. Error on the mean (top row) and variance (bottom row) of sample entropy for various values of D and ν when using the first-order approximation (circles), the second-order approximation (squares), or the sampling scheme (diamonds). The error was calculated as the absolute value of the difference between the approximation and the true value. For the sampling scheme are represented the median as well as the symmetrical 90% probability interval of the error. Scale on y axis is logarithmic.



We were also able to obtain an asymptotic approximation of the sampling and posterior expectations of mutual information and multiinformation. Contrary to the general argument developed by [7], we proved that, for multivariate normal distributions, the negative bias for differential entropy does entail a positive bias for mutual information. This result is in agreement with the fact that, under the null hypothesis of $\boldsymbol{\Sigma}$ diagonal matrix, corresponding to $i(\boldsymbol{\Sigma}) = 0$, $\nu i(\boldsymbol{S}/\nu)$ is asymptotically chi square distributed with $\sum_{i<j} D_i D_j/2$ degrees of freedom and, hence, has an expectation equal to that value [19] (pp. 306–307). Surprisingly, and unlike what was said for entropy, the positive bias of the sample multiinformation was not corrected by the Bayesian approach. A naive correction of minus the positive bias could lead to negative values, which is impossible by construction of multiinformation. Note that, using the present results alone, we were not able to obtain an asymptotic approximation for the variance of the same measures.

In the present paper, we used loose versions of the inequalities proposed in [20] to prove the monotonicity and sign of the cumulants of U (see Section 2.1 and Appendix). Note that, using the same inequalities, it seems that it would also be possible to obtain lower and upper bounds for these quantities, instead of asymptotic approximations. These bounds would be useful complements to the approximations provided in the present manuscript.

Acknowledgements

The authors are grateful to Pierre Bellec for helpful discussions.

References

1. Ahmed, N.A.; Gokhale, D.V. Entropy expressions and their estimators for multivariate distributions. *IEEE Trans. Inform. Theory* **1989**, *35*, 688–692.
2. Misra, N.; Singh, H.; Demchuk, E. Estimation of the entropy of a multivariate normal distribution. *J. Multivariate Anal.* **2005**, *92*, 324–342.
3. Gupta, M.; Srivastava, S. Parametric Bayesian estimation od differential entropy and relative entropy. *Entropy* **2010**, *12*, 818–843.
4. Beirlant, J.; Dudewicz, E.J.; Györfi, L.; van der Meulen, E.C. Nonparametric entropy estimation: An overview. *Int. J. Math. Stastist. Sci.* **1997**, *6*, 17–39.
5. Strong, S.P.; Koberle, R.; de Ruyter van Steveninck, R.R.; Bialek, W. Entropy and information in neural spike trains. *Phys. Rev. Lett.* **1998**, *80*, 197–200.
6. Antos, A.; Kontoyiannis, I. Convergence properties of functional estimates for discrete distributions. *Random Struct. Algor.* **2001**, *19*, 163–193.
7. Paninski, L. Estimation of entropy and mutual information. *Neural Comput.* **2003**, *15*, 1191–1253.
8. Wolpert, D.H.; Wolf, D.R. Estimating functions of probability distributions from a finite set of samples. *Phys. Rev. E* **1995**, *52*, 6841–6854.
9. Wolpert, D.H.; Wolf, D.R. Erratum: Estimating functions of probability distributions from a finite set of samples. *Phys. Rev. E* **1996**, *54*, 6973.
10. Anderson, T.W. *An Introduction to Multivariate Statistical Analysis*; John Wiley and Sons: New York, NY, USA, 1958.

11. Abramowitz, M.; Stegun, I.A. *Handbook of Mathematical Functions*; Applied Mathematics Series 55; National Bureau of Standards: Washington, DC, USA, 1972.
12. Anderson, T.W. *An Introduction to Multivariate Statistical Analysis*, 3rd ed.; Series in Probability and Mathematical Statistics; John Wiley and Sons: New York, NY, USA, 2003.
13. Gelman, A.; Carlin, J.B.; Stern, H.S.; Rubin, D.B. *Bayesian Data Analysis*; Texts in Statistical Science; Chapman & Hall: London, UK, 1998.
14. Watanabe, S. Information theoretical analysis of multivariate correlation. *IBM J. Res. Dev.* **1960**, *4*, 66–82.
15. Garner, W.R. *Uncertainty and Structure as Psychological Concepts*; John Wiley & Sons: New York, NY, USA, 1962.
16. Joe, H. Relative entropy measures of multivariate dependence. *J. Am. Statist. Assoc.* **1989**, *84*, 157–164.
17. Studený, M.; Vejnarová, J. The multiinformation function as a tool for measuring stochastic dependence. In *Proceedings of the NATO Advanced Study Institute on Learning in Graphical Models*; Jordan, M.I., Ed.; MIT Press: Cambridge, MA, USA, 1998; pp. 261–298.
18. Press, S.J. *Applied Multivariate Analysis. Using Bayesian and Frequentist Methods of Inference*, 2nd ed.; Dover: Mineola, NY, USA, 2005.
19. Kullback, S. *Information Theory and Statistics*; Dover: Mineola, NY, USA, 1968.
20. Chen, C.P. Inequalities for the polygamma functions with application. *Gener. Math.* **2005**, *13*, 65–72.

Appendix

Results Regarding the Cumulants

The proofs differ for κ_1 and κ_n, $n \geq 2$.

1. Results for κ_1

For $\nu \geq D > 0$, set $f_D(\nu) = \kappa_1$ as defined in Equation (7).

Result 1: $f_D(\nu)$ is a decreasing function of ν. Derivation of $f_D(\nu)$ with respect to ν leads to

$$f'_D(\nu) = \sum_{d=1}^{D} \left[\frac{1}{\nu} - \frac{1}{2}\psi'\left(\frac{\nu+1-d}{2}\right) \right] \tag{25}$$

We use the following inequality [20]

$$\psi'(x) > \frac{1}{x} + \frac{1}{2x^2}$$

This implies that

$$\frac{1}{\nu} - \frac{1}{2}\psi'\left(\frac{\nu+1-d}{2}\right) < \frac{1}{\nu} - \frac{1}{\nu+1-d} - \frac{1}{(\nu+1-d)^2}$$

For $1 \leq d \leq \nu$, we have $1/\nu \leq 1/(\nu+1-d)$. Consequently, each term in the sum of Equation (25) is strictly negative, and so is $f'_D(\nu)$. $f_D(\nu)$ is therefore a strictly decreasing function of ν.

Result 2: $f_D(\nu)$ is an increasing function of D. We have

$$f_{D+1}(\nu) = f_D(\nu) + \ln\frac{\nu}{2} - \psi\left(\frac{\nu+1-D}{2}\right)$$

Using the following inequality [20]

$$\psi(u) < \ln u - \frac{1}{2u} < \ln u$$

we obtain that

$$\psi\left(\frac{\nu+1-D}{2}\right) < \ln\left(\frac{\nu+1-D}{2}\right)$$

leading to

$$\ln\frac{\nu}{2} - \psi\left(\frac{\nu+1-D}{2}\right) > -\ln\left(1+\frac{1-D}{\nu}\right)$$

Since $\ln(1+x) < x$, we have

$$\ln\frac{\nu}{2} - \psi\left(\frac{\nu+1-D}{2}\right) > \frac{D-1}{\nu}$$

and, therefore, $f_{D+1}(\nu) > f_D(\nu)$.

Result 3: $f_D(\nu)$ is positive. $f_D(\nu)$ is the sum of terms that are strictly positive (cf previous paragraph); it is thus strictly positive.

Result 4: $f_D(\nu)$ tends to infinity as D increases. From the proof of Result 2, we have

$$f_D(\nu) > \sum_{d=1}^{D} \frac{d-1}{\nu} = \frac{D(D-1)}{2\nu}$$

which tends to infinity when D tends to infinity.

Result 5: $f_D(\nu)$ tends to 0 as ν increases. We use the following inequality [20]

$$\ln x - \frac{1}{2x} - \frac{1}{12x^2} < \psi(u)$$

This implies that

$$\psi\left(\frac{\nu+1-d}{2}\right) > \ln\left(\frac{\nu+1-d}{2}\right) - \frac{1}{\nu+1-d} - \frac{1}{3(\nu+1-d)^2}$$

leading to

$$\ln\frac{\nu}{2} - \psi\left(\frac{\nu+1-d}{2}\right) < -\ln\left(1+\frac{1-d}{\nu}\right) + \frac{1}{\nu+1-d} + \frac{1}{3(\nu+1-d)^2}$$

Since $x - x^2/2 < \ln(1+x)$, we have

$$\ln\frac{\nu}{2} - \psi\left(\frac{\nu+1-d}{2}\right) < \frac{d-1}{\nu} + \frac{1}{\nu+1-d} + \frac{1}{3(\nu+1-d)^2} < \frac{d-1}{\nu} + \frac{1}{\nu-(D-1)} + \frac{1}{3[\nu-(D-1)]^2}$$

Summing over d yields

$$f_D(\nu) < \frac{D(D-1)}{2\nu} + \frac{D}{\nu-(D-1)} + \frac{D}{3[\nu-(D-1)]^2}$$

which tends to 0 when ν increases.

2. Results for κ_n, $n \geq 2$

Define $f_D(\nu) = \kappa_n$ as in Equation (6), $(-1)^{n+1}\psi^{(n)}$ is completely monotonic. As a consequence, κ_n is a decreasing function of ν. We also use the following inequality [20]

$$\frac{(n-1)!}{x^n} < (-1)^{n+1}\psi^{(n)}(x) < \frac{(n-1)!}{x^n} + \frac{n!}{2x^{n+1}} + \frac{B_2\Gamma(n+2)}{2x^{n+2}}$$

This implies that $(-1)^{n+1}\psi^{(n)}(x)$ is strictly positive and, as a consequence, that κ_n is an increasing function of D. It also implies that κ_n tends to 0 as ν tends to infinity.