entropy

MDPI

Communication

Geometric Insights into the Multivariate Gaussian Distribution and Its Entropy and Mutual Information

Dah-Jing Jwo [1,*], Ta-Shun Cho [2] and Amita Biswal [1]

[1] Department of Communications, Navigation and Control Engineering, National Taiwan Ocean University, 2 Peining Rd., Keelung 202301, Taiwan; amitabiswal1988@gmail.com
[2] Department of Business Administration, Asia University, 500 Liufeng Road, Wufeng, Taichung 41354, Taiwan; cho2022@asia.edu.tw
* Correspondence: djjwo@mail.ntou.edu.tw

Abstract: In this paper, we provide geometric insights with visualization into the multivariate Gaussian distribution and its entropy and mutual information. In order to develop the multivariate Gaussian distribution with entropy and mutual information, several significant methodologies are presented through the discussion, supported by illustrations, both technically and statistically. The paper examines broad measurements of structure for the Gaussian distributions, which show that they can be described in terms of the information theory between the given covariance matrix and correlated random variables (in terms of relative entropy). The content obtained allows readers to better perceive concepts, comprehend techniques, and properly execute software programs for future study on the topic's science and implementations. It also helps readers grasp the themes' fundamental concepts to study the application of multivariate sets of data in Gaussian distribution. The simulation results also convey the behavior of different elliptical interpretations based on the multivariate Gaussian distribution with entropy for real-world applications in our daily lives, including information coding, nonlinear signal detection, etc. Involving the relative entropy and mutual information as well as the potential correlated covariance analysis, a wide range of information is addressed, including basic application concerns as well as clinical diagnostics to detect the multi-disease effects.

Keywords: multivariate Gaussians; correlated random variables; visualization; entropy; relative entropy; mutual information

Citation: Jwo, D.-J.; Cho, T.-S.; Biswal, A. Geometric Insights into the Multivariate Gaussian Distribution and Its Entropy and Mutual Information. *Entropy* **2023**, *25*, 1177. https://doi.org/10.3390/e25081177

Academic Editor: Ashwin Vaidya

Received: 6 July 2023
Revised: 31 July 2023
Accepted: 4 August 2023
Published: 7 August 2023

1. Introduction

Understanding the ways knowledge concerning an external variable or the reciprocal information of its parts can assist in characterizing and inferring the underlying mechanics and function of the system. This goal has driven the development of several techniques for dissecting the elements of a set of variables' combined entropy or for dissecting the contributions of a set of variables to the mutual information about the variable of interest. In actuality, this association and its modifications exist for any input signal and the widest range of Gaussian pathways, comprising discrete-time and continuous-time pathways in scalar or vector forms.

In a more general way, mutual information and mean-square error (MSE) are the fundamental concepts of information theory and estimation theory, respectively. In contrast to the minimum MSE (MMSE), which determines how precisely each input sample can be restored using the channel's outcomes, the input-output mutual information is an estimation of whether the information can be consistently delivered over a channel given a specific input signal. An inactive functioning characterization for mutual information is provided by the substantial relevance of mutual information to estimation and filtering. Therefore, the significance of identity is not only obvious, but the link is also fascinating and merits an in-depth explanation [1–3]. Relations between the MMSE of the approximation

of the output given the input and the localized actions of the mutual information at diminishing signal-to-noise ratio (SNR) are presented in [4]. The authors of [5] give an idea about the probabilistic ratios of geometric characteristics of signal detection in Gaussian noise. Furthermore, whether in a continuous-time [5–7] or discrete-time [8] context, the likelihood ratio is difficult in the relationship between observation and estimation.

Considering the specific instance of parametric computation (or Gaussian inputs), correlations relating to causal and non-causal estimation errors have been investigated in [9,10], involving the limit on the loss owing to the causality restriction. Knowing how data pertaining to an external parameter, or inversely related data within its parts, distributes across the parts of a multivariate system can assist in categorizing and determining the fundamental mechanics and functionality of the structure. The mechanism served as the impetus for the development of various techniques for decomposing the various elements of a set of parameters' joint entropy [11–19] or for deconvoluting the additions of a set of elements to the mutual information about a target variable [13]. The mutual information techniques can be used to examine a variety of intricate systems, including those in the physical distinction domain, such as gene networks [17] or brain coding [20], as well as those in the social domain, such as selection agents [21] and community behavior [22]. It can also be used to analyze artificial agents [23]. Furthermore, some new proposals deviated more significantly from the original framework by incorporating novel principles such as the consideration of the presence of harmful elements associated with errors and the use of joint entropy subdivisions in place of mutual information [24,25].

In the multivariate scenario, the challenges of breaking down mutual information into redundancy and complementary sections have nevertheless been significantly increased. The maximum entropy framework allows for a more straightforward generalization of the efficiency measurements to the multivariate case [26,27]. The novel redundancy determines that were initially developed are only defined for the bivariate situation or allow negative components [28], whereas measurements of coordination are more readily extended to the multivariate case, especially when using the maximum entropy architecture [29,30]. By either utilizing the associations between lattices formed by various numbers of parameters or utilizing the multiple interactions between redundant lattices and information loss lattices, for which collaborative efforts are more actually defined, the study in [31,32] established two analogous techniques for constructing multivariate redundant metrics. Information theory variables have a benefit compared to more known test results measurements in that they may be employed when numerous ailments are being considered as well as when a test of diagnosis can produce several or continuous findings [33].

Although there are some valuable references detailing entropy-related topics, both discrete and continuous, they may not be easily accessible to some readers from the existing publications. Therefore, in this present study, we propose an extension of the bivariate Gaussian distribution technique to calculate multivariate redundant metrics inside the maximum entropy context. The importance of the maximum entropy approach in the multivariate scenario, where it offers constraints for the actual redundancy, unique information, and efficiency terms under logical presumptions shared by additional criteria, acts as the motivation for this particular focus [26,34]. The maximum entropy measurements, specifically, offer a lower limit for the actual cooperation and redundant terms and a higher limit for the actual specific information if it is presumed that a bivariate non-negative disintegration exists and that redundancy can be calculated from the bivariate distributions of the desired outcome with every source. Furthermore, if these bivariate distributions are consistent with possibly having little interaction under the previous hypotheses, then the maximum entropy decomposition returns not only boundaries but also the precise actual terms. Here, in the proposed framework, we also demonstrated that, under similar presumptions, the maximum entropy reduction also plays this dominant role in the multivariate situation [35]. This paper intends to convey the important issues and inspire new applications of information theory to a number of areas, such as information coding, nonlinear signal detection, and clinical diagnostic testing.

The remainder of this paper is organized as follows. A brief review of the geometry of the Gaussian distribution is reviewed in Section 2. The three consecutive sections deal with various important topics on information entropy with illustrative examples, with an emphasis on visualization of the information and discussion. In Section 3, continuous entropy/differential entropy are presented. In Section 4, the relative entropy (Kullback–Leibler divergence) is presented. Mutual information is presented in Section 5. Conclusions are given in Section 6.

2. Geometry of the Gaussian Distribution

In this section, the background relations of the Gaussian distribution from different parametric points of view will be discussed. The exploratory objective of the fundamental analysis is to identify "the framework" in multivariate datasets. Ordinary least-squares regression and principal component analysis (PCA), respectively, analyze the measurements for dependency (the predicted connection between particular components) and rigidity (the degree of prominence of the probability density function (pdf) around a low-dimensional axis) for bivariate Gaussian distributions. Mutual information, an established measure of dependency, is not an accurate indicator of rigidity since it is not invariant with an opposite rotation of the parameters. For bivariate Gaussian distributions, a suitable rotating invariant compactness measure is constructed and demonstrated to reduce the corresponding PCA measure.

2.1. Standard Parametric Representation of an Ellipse

For the uncorrelated data, which has zero covariance, the ellipse is not rotated and the axis is aligned. The radii of the ellipse in both directions are the variances. Geometrically, a not-rotated ellipse at point (0, 0) and radii a and b for the x_1- and x_2-direction is described by:

$$\left(\frac{x_1}{a}\right)^2 + \left(\frac{x_2}{b}\right)^2 = 1. \tag{1}$$

The general probability density function for the multivariate Gaussian is given by the following:

$$f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\mu}, \sum) = \frac{1}{\left(\sqrt{2\pi}\right)^n \left|\sum\right|^{1/2}} e^{\{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^T \sum^{-1}(\mathbf{X}-\boldsymbol{\mu})\}}, \tag{2}$$

where $\boldsymbol{\mu} = E[\mathbf{X}]$ and $\sum = \mathrm{Cov}(\mathbf{X}) = E[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T]$ is a symmetric, positive semi-definite matrix. If $\sum$ is the identity matrix, then the Mahalanobis distance reduces to the standard Euclidean distance between $\mathbf{X}$ and $\boldsymbol{\mu}$.

For bivariate Gaussian distributions, the mean and covariance matrix are given by the following:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}; \sum = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \tag{3}$$

where the linear correlation coefficient $|\rho| \leqslant 1$.

Variance measures the variation of a single random variable, whereas covariance is a measure of the two random variables varying together. With the covariance, we can calculate the entries of the covariance matrix, which is a square matrix. In addition, the covariance matrix is symmetric. The diagonal entries of the covariance matrix are the variances; however, the other entries are the covariances. Due to this reason, the covariance matrix is often called the variance-covariance matrix.

2.2. The Confidence Ellipse

A typical way to visualize two-dimensional Gaussian-distributed data is by plotting a confidence ellipse. The distance $d_M = (\mathbf{X}-\boldsymbol{\mu})^{\mathrm{T}} \sum^{-1} (\mathbf{X}-\boldsymbol{\mu})$ is a constant value referred to

as the Mahalanobis distance, which is a random variable distributed by the chi-squared distribution, denoted as χ^2_k.

$$\mathrm{P}[(\mathbf{X}-\boldsymbol{\mu})^{\mathrm{T}}\sum^{-1}(\mathbf{X}-\boldsymbol{\mu}) \leqslant \chi^2_k(\alpha)] = 1-\alpha, \tag{4}$$

where k is the number of degrees of freedom and α is the given probability related to the confidence ellipse. For example, if $\alpha = 0.95$, 95% confidence ellipse is defined. Extension from Equation (1): the radius in each direction is the standard deviation σ_1 and σ_2 parametrized by a scale factor s, known as the Mahalanobis radius of the ellipsoid:

$$\left(\frac{x_1}{\sigma_1}\right)^2 + \left(\frac{x_2}{\sigma_2}\right)^2 = s. \tag{5}$$

The goal is to determine the scale s such that confidence p is met. Since the data are multivariate Gaussian-distributed, the left-hand side of the equation is the sum of squares of Gaussian-distributed samples, which follows a χ^2 distribution. A χ^2 distribution is defined by the degrees of freedom, and since we have two dimensions, the number of degrees of freedom is two. Now, we have calculated the probability with the sum, and therefore s has a certain value under a χ^2 distribution.

This ellipse with a probability contour defines the region of a minimum area (or volume in the multivariate case) containing a given probability under the Gaussian assumption. The equation can be solved using a χ^2 table or simply using the relationship $s = -2\ln(1-p)$. The confidence interval can be evaluated through the following:

$$p = 1 - exp(-0.5s). \tag{6}$$

for $s = 1$ we have $p = 1 - exp(-0.5) \approx 0.3935$. Furthermore, typical values include $s = 2.279$, $s = 4.605$, $s = 5.991$, and $s = 9.210$ for $p = 0.68$, $p = 0.9$, $p = 0.95$, and $p = 0.99$, respectively. The ellipse can then be drawn with radii $\sigma_1\sqrt{s}$ and $\sigma_2\sqrt{s}$. Figure 1 shows the relationship between the confidence interval and the scale factor s.

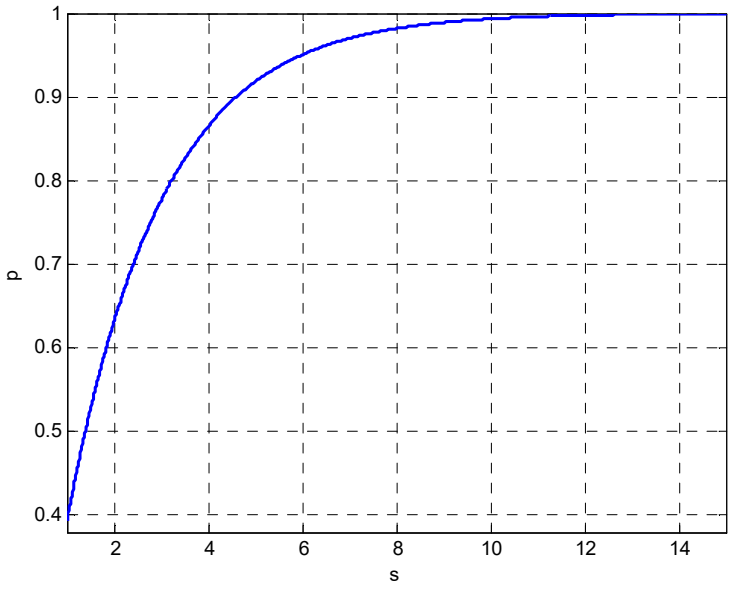


Figure 1. Relationship between the confidence interval and the scale factor s.

The Mahalanobis distance accounts for the variance of each variable and the covariance between variables.

$$\begin{aligned}
&(\mathbf{X}-\boldsymbol{\mu})^T \sum{}^{-1} (\mathbf{X}-\boldsymbol{\mu}) \\
&= \begin{bmatrix} x_1-\mu_1 & x_2-\mu_2 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix} \\
&= \begin{bmatrix} x_1-\mu_1 & x_2-\mu_2 \end{bmatrix} \frac{\begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix}}{\sigma_1^2\sigma_2^2(1-\rho^2)} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix} \\
&= \frac{1}{1-\rho^2}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} \right)
\end{aligned} \tag{7}$$

Geometrically, it does this by transforming the data into standardized, uncorrelated data and computing the ordinary Euclidean distance for the transformed data. In this way, the Mahalanobis distance is like a univariate z-score: it provides a way to measure distances that takes into account the scale of the data.

In the general case, covariances σ_{12} and σ_{21} are not zero, and therefore the ellipse-coordinate system is not axis-aligned. In such a case, instead of using the variance as a spread indicator, we use the eigenvalues of the covariance matrix. The eigenvalues represent the spread in the direction of the eigenvectors, which are the variances under a rotated coordinate system. By definition, a covariance matrix is positive and definite; therefore, all eigenvalues are positive and can be seen as a linear transformation of the data. The actual radii of the ellipse are $\sqrt{\lambda_1}$ and $\sqrt{\lambda_2}$ for the two eigenvalues λ_1 and λ_2 of the scaled covariance matrix $s \cdot \sum$.

Based on Equations (2) and (7), the bivariate Gaussian distributions can be represented as follows:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2}\left(\frac{1}{1-\rho^2}\right)\left\{\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right\}}, \tag{8}$$

and the level surface of $f(x_1, x_2)$ are concentric ellipses:

$$\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} = c, \tag{9}$$

where c is the Mahalanobis distance possessing the following properties:

- It accounts for the fact that the variances in each direction are different;
- It accounts for the covariance between variables;
- It reduces to the familiar Euclidean distance for uncorrelated variables with unit variance.

The length of the ellipse axes is a function of the given probability of the chi-squared distribution with 2 degrees of freedom $\chi_2^2(\alpha)$, the eigenvalues $\boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 & \lambda_2 \end{bmatrix}^T$ and the linear correlation coefficient ρ. If $\alpha = 0.95$, 95% confidence ellipse is defined by:

$$\begin{bmatrix} x_1-\mu_1 & x_2-\mu_2 \end{bmatrix} \sum{}^{-1} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix} \leqslant \chi_2^2(0.05) \tag{10}$$

where

$$\sum{}^{-1} = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)} \begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix}, \tag{11}$$

as $\sum$ denotes a symmetric matrix, the eigenvectors of $\sum$ is linearly independent (or orthogonal).

2.3. Similarity Transform

The simplest similarity transformation method for eigenvalue computation is the Jacobi method, which deals with the standard eigenproblems. In the multivariate Gaussian distribution, the covariance matrix $\sum$ can be expressed in terms of eigenvectors:

$$\sum = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}} = \begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_2\end{bmatrix}\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}\begin{bmatrix}\mathbf{u}_1^T\\ \mathbf{u}_2^T\end{bmatrix}, \tag{12}$$

where $\mathbf{U} = \begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_2\end{bmatrix}$ are the eigenvectors of $\sum$ and $\mathbf{\Lambda}$ is the diagonal matrix of the eigenvalues $\boldsymbol{\lambda} = \begin{bmatrix}\lambda_1 & \lambda_2\end{bmatrix}^T$

$$\mathbf{\Lambda} = \begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix},$$

replacing $\sum$ by $\sum^{-1} = \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^{-1}$, the square of the difference can be written as:

$$\begin{bmatrix}x_1-\mu_1 & x_2-\mu_2\end{bmatrix}\mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^{-1}\begin{bmatrix}x_1-\mu_1\\ x_2-\mu_2\end{bmatrix} \leqslant \chi_2^2(0.05), \tag{13}$$

as $\mathbf{U}^T = \mathbf{U}^{-1}$. Denoting

$$\begin{bmatrix}y_1\\ y_2\end{bmatrix} = \mathbf{U}^{-1}\begin{bmatrix}x_1-\mu_1\\ x_2-\mu_2\end{bmatrix}, \tag{14}$$

the square of the difference can then be expressed as:

$$\begin{bmatrix}y_1 & y_2\end{bmatrix}\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}\begin{bmatrix}y_1\\ y_2\end{bmatrix} \leqslant \chi_2^2(0.05). \tag{15}$$

If the above equation is further evaluated, the resulting equation is the equation of an ellipse aligned with the axis y_1 and y_2 in the new coordinate system:

$$\frac{y_1^2}{\chi_2^2(0\ .05)\lambda_1} + \frac{y_2^2}{\chi_2^2(0\ .05)\lambda_2} \leqslant 1, \tag{16}$$

the axes of the ellipse are defined by y_1 axis with a length $2\sqrt{\lambda_1\chi_2^2(0\ .05)}$ and y_2 axis with a length $2\sqrt{\lambda_2\chi_2^2(0\ .05)}$.

When $\rho = 0$, the eigenvectors are equal to $\lambda_1 = \sigma_1$ and $\lambda_2 = \sigma_2$. Additionally, $\mathbf{U}$ matrix whose elements are the eigenvectors, of $\sum$ becomes an identity matrix. The final equation of an ellipse is then defined by:

$$\frac{(x_1-\mu_1)^2}{\chi_2^2(0\ .05)\lambda_1} + \frac{(x_2-\mu_2)^2}{\chi_2^2(0\ .05)\lambda_2} \leqslant 1. \tag{17}$$

It is clear from the equation given above that the axes of the ellipse are parallel to the coordinate axes. The lengths of the axes of the ellipse are then defined as $2\sqrt{\sigma_{11}\chi_2^2(0\ .05)}$ and $2\sqrt{\sigma_{22}\chi_2^2(0\ .05)}$.

The covariance matrix can be presented by its eigenvectors and eigenvalues: $\sum\mathbf{U} = \mathbf{U}\mathbf{\Lambda}$, where $\mathbf{U}$ is the matrix whose columns are the eigenvectors of $\sum$ and $\mathbf{\Lambda}$ is the diagonal matrix with diagonal elements given by the eigenvalues of $\sum$. Transformation is performed based on the three steps involving scaling, rotation, and translation:

1. Scaling

The covariance matrix can be written as $\sum = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1} = \mathbf{U}\mathbf{S}\mathbf{S}\mathbf{U}^{-1}$, where $\mathbf{S}$ is a diagonal scaling matrix $\mathbf{S} = \mathbf{\Lambda}^{1/2} = \mathbf{S}^T$;

2. Rotation

U is generalized from the normalized eigenvectors of the covariance matrix $\sum$.

$$\mathbf{U} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}, \tag{18}$$

it can be noted that **U** is an orthogonal matrix $\mathbf{U}^{-1} = \mathbf{U}^T$ and $|\mathbf{U}| = 1$. Here, we have calculated the matrix with rotation and scaling $\mathbf{T} = \mathbf{US}$ and $\mathbf{T}^T = (\mathbf{US})^T = \mathbf{S}^T\mathbf{U}^T = \mathbf{SU}^{-1}$. Thus, the covariance matrix can be written as $\sum = \mathbf{TT}^T$ and $\mathbf{U}^T \sum \mathbf{U} = \mathbf{\Lambda}$ with diagonal eigenvalues λ_i. Since $\mathbf{T} = \mathbf{US}$, we have $\mathbf{Y} = \mathbf{TX} = \mathbf{USX} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{X}$.

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} u_{1x} & u_{2x} \\ u_{1y} & u_{2y} \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1}\cos(t) \\ \sqrt{\lambda_2}\sin(t) \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1}\cos(t) \\ \sqrt{\lambda_2}\sin(t) \end{bmatrix} \tag{19}$$

The similarity transform is applied to obtain the relationship between $\mathbf{X}^T \sum^{-1} \mathbf{X} = \mathbf{Y}^T\mathbf{U}^T \sum^{-1} \mathbf{UY} = \mathbf{Y}^T\mathbf{\Lambda}^{-1}\mathbf{Y}$, and the pdf of **Y** vector, which can be found by considering the below expression:

$$f_{\mathbf{Y}}(\mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sqrt{\lambda_i}} e^{-\frac{1}{2}\frac{y_i^2}{\lambda_i}}, \tag{20}$$

the ellipse in the transformed frame can be represented as:

$$\frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} = c, \tag{21}$$

where the eigenvectors are equal to $\lambda_1 = \sigma_1^2$ and $\lambda_2 = \sigma_2^2$;

3. Translation

$$x_1(t) = \sqrt{\lambda_1}\cos(\theta)\Big)\cos(t) - \sqrt{\lambda_2}\sin(\theta)\Big)\sin(t) + \mu_1, \tag{22}$$

$$x_2(t) = \sqrt{\lambda_1}\sin(\theta)\Big)\cos(t) + \sqrt{\lambda_2}\cos(\theta)\Big)\sin(t) + \mu_2, \tag{23}$$

the eigenvalues $\boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 & \lambda_2 \end{bmatrix}^T$ can be calculated from:

$$\lambda_1 = \frac{1}{2}\left[\sigma_1^2 + \sigma_2^2 + \sqrt{\left(\sigma_1^2 - \sigma_2^2\right)^2 + 4\rho^2\sigma_1^2\sigma_2^2}\right];\ \lambda_2 = \frac{1}{2}\left[\sigma_1^2 + \sigma_2^2 - \sqrt{\left(\sigma_1^2 - \sigma_2^2\right)^2 + 4\rho^2\sigma_1^2\sigma_2^2}\right],$$

and thus

$$\left|\sum\right| = \lambda_1 \cdot \lambda_2 = \sigma_1^2\sigma_2^2(1 - \rho^2). \tag{24}$$

From another point of view, the covariance matrix can be calculated as:

$$\begin{aligned} \sum = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T &= \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}, \\ &= \begin{bmatrix} \lambda_1\cos(\theta) & -\lambda_2\sin(\theta) \\ \lambda_1\sin(\theta) & \lambda_2\cos(\theta) \end{bmatrix} \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}, \\ &= \begin{bmatrix} \lambda_1\cos^2(\theta) + \lambda_2\sin^2(\theta) & (\lambda_1 - \lambda_2)(\sin(\theta) - \cos(\theta)) \\ syms & \lambda_1\sin^2(\theta) + \lambda_2\cos^2(\theta) \end{bmatrix}. \end{aligned} \tag{25}$$

Calculation for the determinant of the above covariance matrix gives the same result, and the inverse is:

$$\begin{aligned}\sum^{-1} &= \frac{1}{\lambda_1 \cdot \lambda_2}\begin{bmatrix} \lambda_1 \sin^2(\theta) + \lambda_2 \cos^2(\theta) & (\lambda_2 - \lambda_1)(\sin(\theta) - \cos(\theta)) \\ syms & \lambda_1 \cos^2(\theta) + \lambda_2 \sin^2(\theta) \end{bmatrix}, \\ &= \begin{bmatrix} \frac{\sin^2(\theta)}{\lambda_2} + \frac{\cos^2(\theta)}{\lambda_1} & \sin(\theta)\cos(\theta)\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right) \\ syms & \frac{\sin^2(\theta)}{\lambda_1} + \frac{\sin^2(\theta)}{\lambda_2} \end{bmatrix}. \end{aligned} \tag{26}$$

2.4. Simulation with a Given Variance-Covariance Matrix

With the given data $\mathbf{X} \sim N(\boldsymbol{\mu}, \sum)$, an ellipse represents the confidence p, which can be plotted by calculating the radii, its center, and the rotation. Here, θ (by which $\mathbf{U}$ can be obtained) and $\mathbf{S}$ for generating the covariance matrix $\sum$, from which ρ can be derived. The inclination angle is calculated by:

$$\theta = \begin{cases} 0 & if\ \sigma_{12} = 0\ and\ \sigma_1^2 \geqslant \sigma_2^2 \\ \pi/2 & if\ \sigma_{12} = 0\ and\ \sigma_1^2 < \sigma_2^2, \\ \tan^{-1}(\lambda_1 - \sigma_1^2, \sigma_{12}) & else \end{cases} \tag{27}$$

which can be used in calculations with the values of $\mathbf{U}$

$$\mathbf{U} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}, \tag{28}$$

and the covariance can be evaluated by: $\sum = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T = \mathbf{USSU}^{\mathrm{T}}$ if $\mathbf{S}$ is specified. On the other hand, with the correlation coefficient ρ and variances for generating the covariance matrix $\sum$, θ can be obtained.

To generate the sampling points that meet the specified correlation, the following procedure can be followed. Given two random variables X_1 and X_2, their linear combination is $Y = \alpha X_1 + \beta X_2$. For the generation of correlated random variables, if we have two Gaussian, uncorrelated random variables X_1, X_2 then we can create two correlated random variables using the formula:

$$Y = \rho X_1 + \sqrt{1 - \rho^2} X_2, \tag{29}$$

and then Y will have a correlation ρ with X_1:

$$\rho = \sigma_{12}/(\sigma_1 \sigma_2).$$

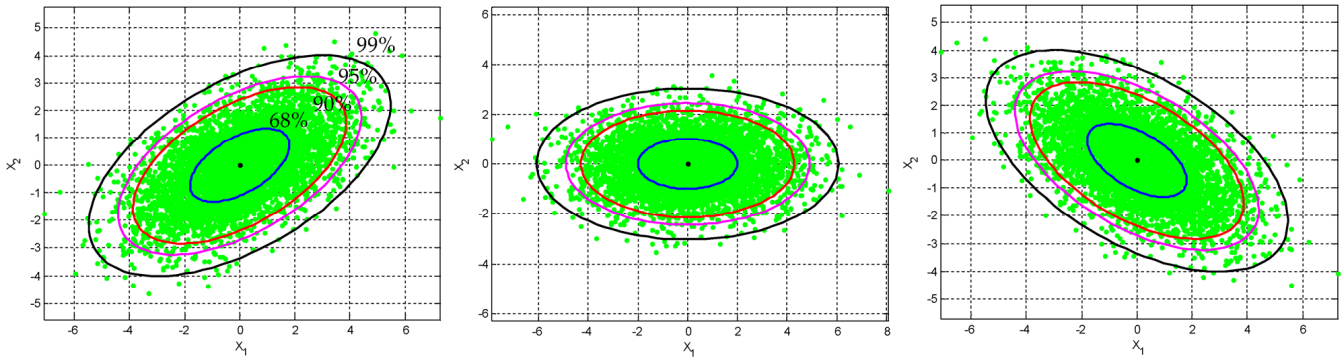
Based on the relationship: $X = AZ + \mu$, $Z \sim N(0, 1)$, the following equation can be employed to generate the sampling points for the scatter plots using the MATLAB software:

$$\mathbf{X} = \mathbf{A} * randn(2, K) + \boldsymbol{\mu} * ones(1, K), \tag{30}$$

where the Cholesky decomposition of $\sum$ has a lower triangular matrix for $\mathbf{A}$, $\sum = \mathbf{AA}^T$ and $\boldsymbol{\mu}$ is the vector of mean values.

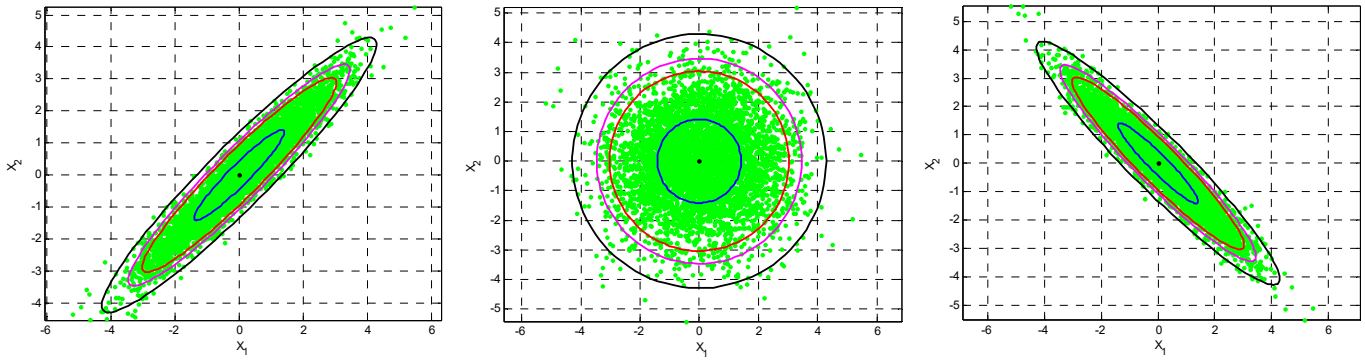
When $\rho = 0$, the axes of the ellipse are parallel to the original coordinate system, and when $\rho \neq 0$, the axes of the ellipse are aligned with the rotated axes in the transformed coordinate system. Figures 2 and 3 show the ellipses for various levels of confidence. The plots provide the idea of confidence (error) ellipses with different confidence levels (i.e., 68%, $s = 2.279$; 90%, $s = 4.605$; 95%, $s = 5.991$; and 99%, $s = 9.210$) from inner to outer ellipses, respectively, by considering the cases where the random variables are: (1) positively correlated $\rho > 0$, (2) negatively correlated $\rho < 0$, and (3) independent $\rho = 0$. More specifically, in Figure 2, the position of the ellipse with various correlation coefficients given by the angle of inclination is specified θ to obtain ρ, $\rho = \sigma_{12}/(\sigma_1 \sigma_2)$: (a) $\theta = 30^\circ$, $\rho \approx 0.55$; (b) $\theta = 0^\circ$, $\rho = 0$; and (c) $\theta = 150^\circ$, $\rho \approx -0.55$, respectively. On the other hand, in Figure 3, the position of the ellipse with various values of the correlation constant given

the angle of inclination is specified ρ to obtain θ: (a) $\rho = 0.95^\circ$, $\theta = 45^\circ$; (b) $\rho = 0$, $\theta = 0^\circ$; and (c) $\rho = -0.95^\circ$, $\theta = 135^\circ$, respectively. The rotation angle is measured $0 \leqslant \theta \leqslant 180^\circ$ with respect to the positive axis. When $\rho > 0$, the angle is in the first quadrant and when $\rho < 0$, the angle is in the second quadrant.



(**a**) (**b**) (**c**)

Figure 2. The position of the ellipse with various correlation coefficients given by the angle of inclination is specified θ to obtain ρ, $\rho = \sigma_{12}/(\sigma_1\sigma_2)$: (**a**) $\theta = 30^\circ$, $\rho \approx 0.55$; (**b**) $\theta = 0^\circ$,$\rho = 0$; and (**c**) $\theta = 150^\circ$, $\rho \approx -0.55$, respectively.



(**a**) (**b**) (**c**)

Figure 3. The position of the ellipse for various values of the correlation constant with the angle of inclination ρ is specified to obtain θ: (**a**) $\rho = 0.95$, $\theta = 45^\circ$; (**b**) $\rho = 0$, $\theta = 0^\circ$; and (**c**) $\rho = -0.95$, $\theta = 135^\circ$, respectively.

In the following, two case studies involve more illustrations:

(1) Equal variances for two random variables with nonzero ρ:

Case 1: fixed correlation coefficient. As an example, when $\rho = 0.5$, and the variances $\sigma_1 = \sigma_2 = \sigma$ range from $2 \sim 5$, as shown in Figure 4. As can be seen, the contours and the scatter plots are ellipses instead of circles.

$$\sum = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} 4^2 & 0.5(4)(2) \\ 0.5(4)(2) & 2^2 \end{bmatrix} = \begin{bmatrix} 4^2 & 4 \\ 4 & 2^2 \end{bmatrix}$$

Subplot (a) in Figure 5 shows the ellipses for $\rho = 0.5$ with varying variances. Here and in subsequent illustrations, 95% confidence levels are shown;

Case 2: increasing the correlation coefficient ρ from zero correlation. With fixed variance $\sigma_1 = \sigma_2 = \sigma$, the contour will initially be a circle when $\rho = 0$ and then an ellipse as ρ increases when $\rho \neq 0$. Subplot (b) in Figure 5 provides the contours with scatter plots for $\rho = 0, 0.5, 0.9, 0.99$, respectively, when $\sigma_1 = \sigma_2 = 2$. The eccentricity of the ellipses increases with the increase of ρ.

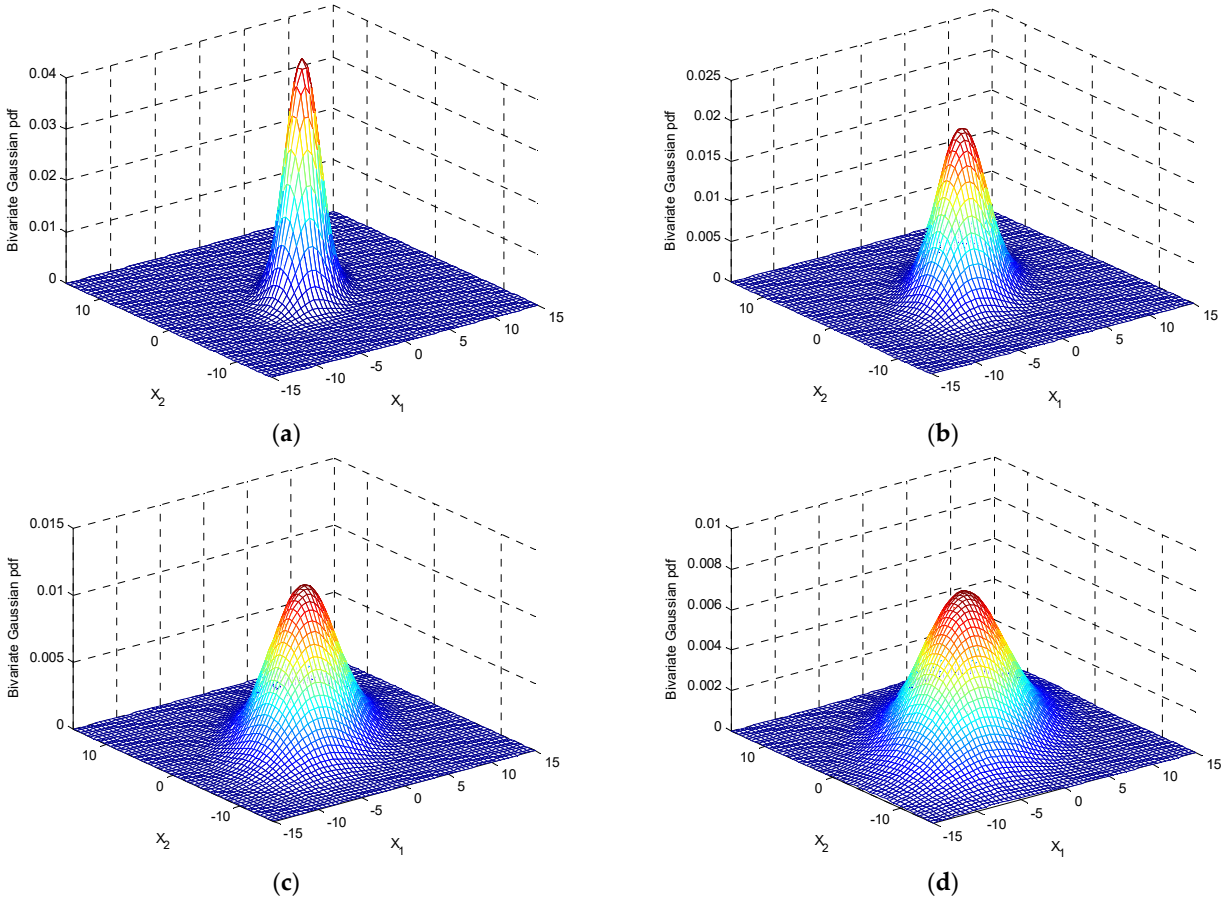


Figure 4. The contours and the scatter plots of ellipses for equal variances $\sigma_1 = \sigma_2 = \sigma$ with a fixed $\rho = 0.5$: (**a**) $\sigma = 2$ (**b**) $\sigma = 3$ (**c**) $\sigma = 4$ (**d**) $\sigma = 5$.

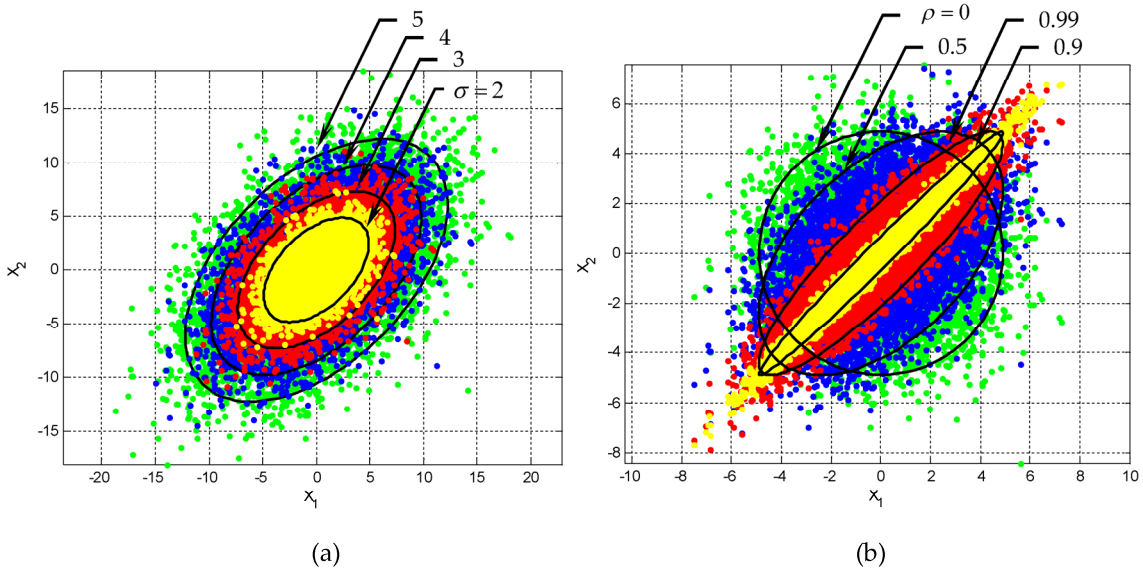


Figure 5. Ellipses for (**a**) $\rho = 0.5$ with varying variances $\sigma_1 = \sigma_2 = \sigma = 2 \sim 5$; and (**b**) equal variances $\sigma_1 = \sigma_2 = 2$ with varying $\rho = 0; 0.5; 0.9; 0.99$.

(2) Unequal variances for two random variables, $\sigma_1 \neq \sigma_2$ with fixed correlation coefficient $\rho = 0.5$.

Case 1: $\sigma_1 > \sigma_2$. The variations of three-dimensional surfaces and ellipses are presented in Figures 6 and 7a with the increase of σ_1/σ_2, where $\sigma_1 = 2 \sim 5$ and $\sigma_2 = 2$.

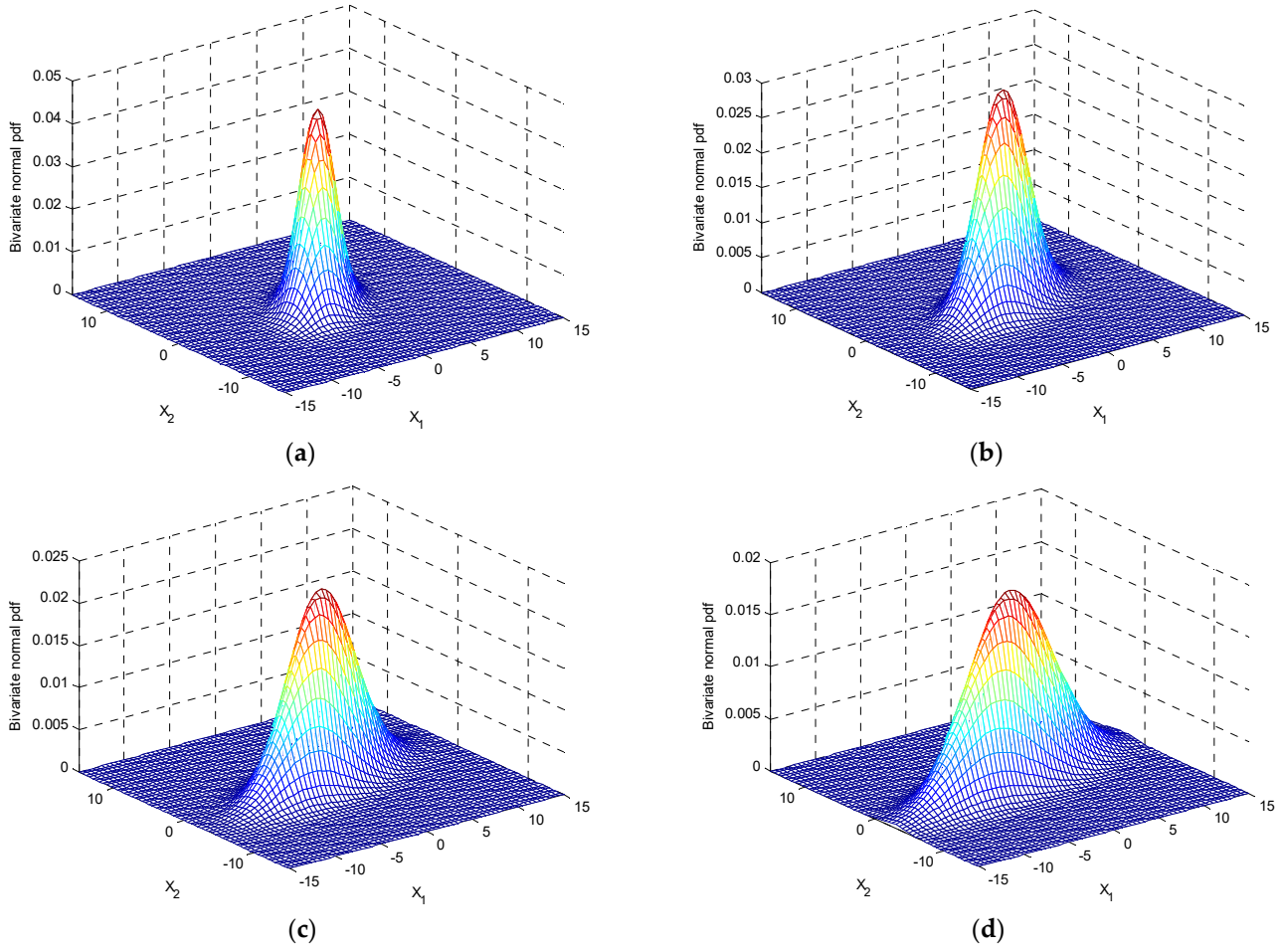


Figure 6. The variations of surface plots in three-dimensional with the increase in σ_1/σ_2 for a fixed $\rho = 0.5$ where $\sigma_2 = 2$: (**a**) $\sigma_1 = 2$; (**b**) $\sigma_1 = 3$; (**c**) $\sigma_1 = 4$; and (**d**) $\sigma_1 = 5$.

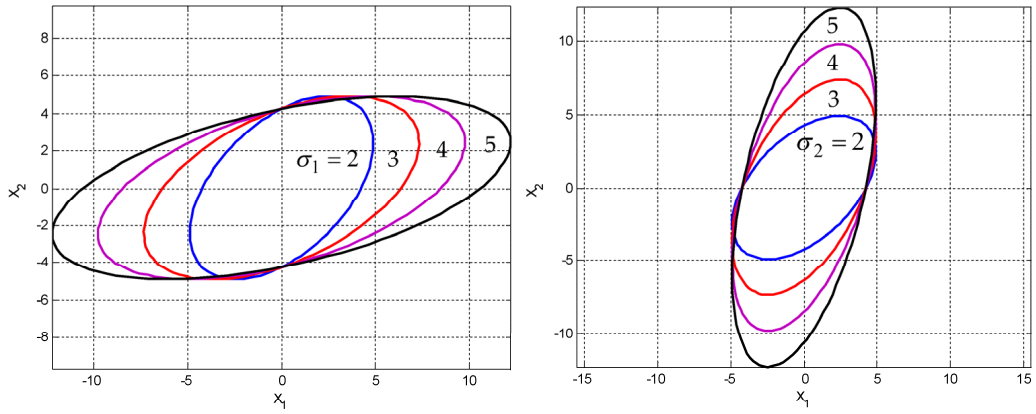


Figure 7. Ellipses for a fixed correlation coefficient when $\sigma_1 \neq \sigma_2$ for a fixed $\rho = 0.5$: (**a**) $\sigma_1 > \sigma_2$, σ_1/σ_2 increases where $\sigma_1 = 2 \sim 5$ and $\sigma_2 = 2$; and (**b**) $\sigma_2 > \sigma_1$, σ_2/σ_1 increases where $\sigma_2 = 2 \sim 5$ and $\sigma_1 = 2$.

Case 2: $\sigma_2 > \sigma_1$. The variation of the ellipses is presented in Figure 7b with the increase of σ_2/σ_1, where $\sigma_2 = 2 \sim 5$ and $\sigma_1 = 2$. Figure 8 shows the variation of inclination angle as a function of σ_1 and σ_2, for $\rho = 0$ and $\rho = 0.5$ for providing further insights on the variation of inclination angle θ with respect to σ_1 and σ_2.

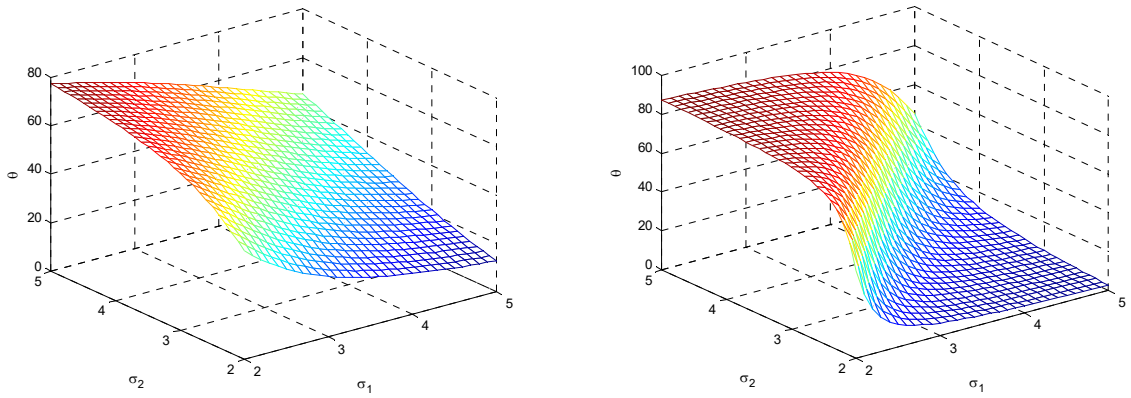


Figure 8. The variation for inclination angle with a function of σ_1 and σ_2, for (**a**) $\rho = 0.5$; and (**b**) $\rho = 0$.

(3) Variation of the ellipses for the various positive and negative correlations. For a given variance, when ρ is specified, the eigenvalues and the inclination angle are obtained accordingly. Figure 9 presents results for the cases of $\sigma_1 > \sigma_2$ ($\sigma_1 = 4, \sigma_2 = 2$ in this example) and $\sigma_2 > \sigma_1$ ($\sigma_1 = 2, \sigma_2 = 4$ in this example) with various correlation coefficients (namely, positive, zero, and negative), including $\rho = 0, 0.5, 0.9, 0.99$ and $\rho = 0, -0.5, -0.9, -0.99$. In the figure, $\sigma_1 = 4, \sigma_2 = 2$ are applied for the top plots, while $\sigma_1 = 2, \sigma_2 = 4$ are applied for the bottom plots. On the other hand, $\rho = 0, 0.5, 0.9, 0.99$ are applied for the left plots, while $\rho = 0, -0.5, -0.9, -0.99$ are applied for the right plots. Furthermore, Figure 10 provides a comparison of the ellipses for various σ_1 and σ_2 for the following cases: (i) $\sigma_1 = 2, \sigma_2 = 4$; (ii) $\sigma_1 = 4, \sigma_2 = 2$; (iii) $\sigma_1 = \sigma_2 = 2$; and (iv) $\sigma_1 = \sigma_2 = 4$, while $\rho = 0.5$.

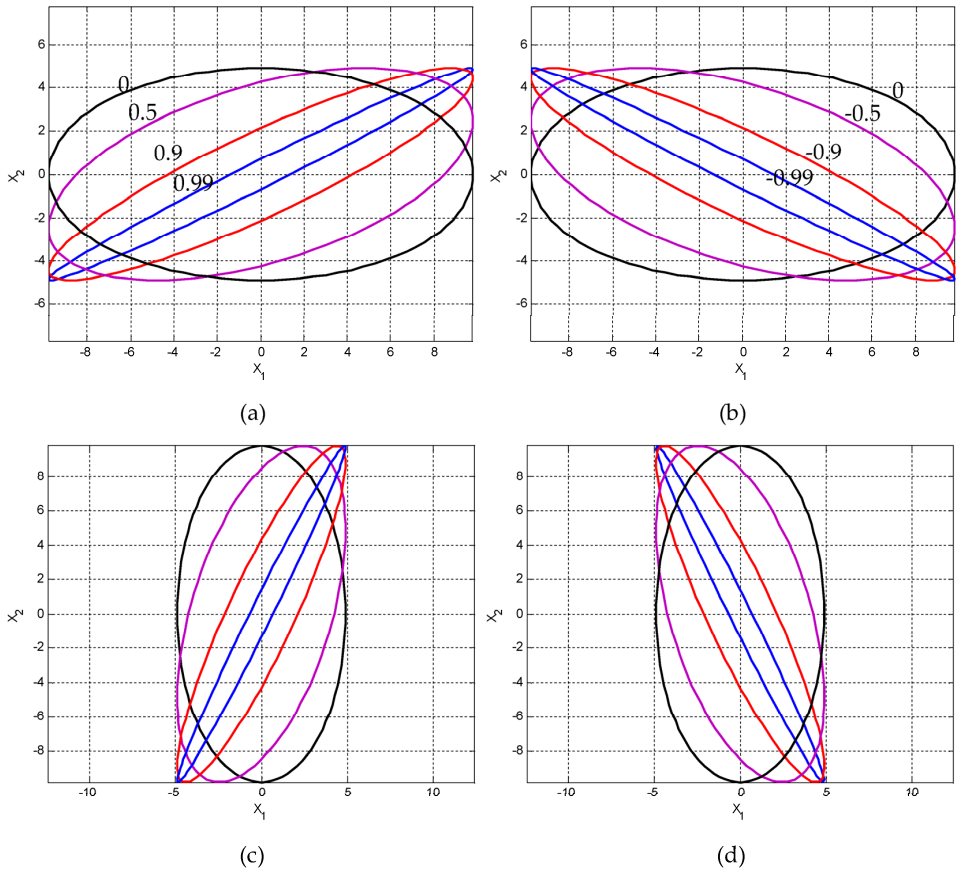


Figure 9. ($\sigma_1 = 4, \sigma_2 = 2$) with (**a**) $\rho = 0, 0.5, 0.9, 0.99$; (**b**) $\rho = 0, -0.5, -0.9, -0.99$ as compared to $\sigma_2 > \sigma_1$ ($\sigma_1 = 2, \sigma_2 = 4$) with (**c**) $\rho = 0, 0.5, 0.9, 0.99$; and (**d**) $\rho = 0, -0.5, -0.9, -0.99$.

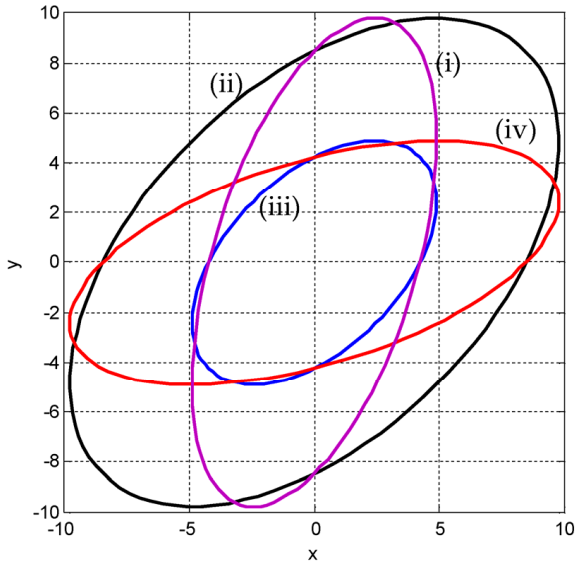


Figure 10. Comparison of the ellipses for various (i) $\sigma_1 = 2, \sigma_2 = 4$; (ii) $\sigma_1 = 4, \sigma_2 = 2$; (iii) $\sigma_1 = \sigma_2 = 2$; and (iv) $\sigma_1 = \sigma_2 = 4$, while $\rho = 0.5$.

3. Continuous Entropy/Differential Entropy

Differential entropy (also referred to as continuous entropy) is a concept in information theory that began as an attempt by Claude Shannon to extend the idea of (Shannon) entropy, a measure of the average surprise of a random variable, to continuous probability distributions. Unfortunately, Shannon did not derive this formula and rather just assumed it was the correct continuous analog of discrete entropy, but it is not [1]. The actual continuous version of discrete entropy is the limiting density of discrete points (LDDP). Differential entropy (described here) is commonly encountered in the literature, but it is a limiting case of the LDDP and one that loses its fundamental association with discrete entropy.

In the following discussion, differential entropy and relative entropy are measured in bits, which are used in the definition. Instead, if ln is used, it is then measured in nats, and the only difference in the expression is the $\log_2 e$ factor.

3.1. Entropy of a Univariate Gaussian Distribution

If we have a continuous random variable X with a probability density function (pdf) $f_X(x)$, the differential entropy of X in bits is expressed as:

$$h(X) = -E[\log_2 f_X(x)] = -\int f_X(x) \log_2 f_X(x) dx, \tag{31}$$

let X be a Gaussian random variable $X \sim N(\mu, \sigma^2)$

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

The differential entropy for this univariate Guassian distribution can be evaluated as follows:

$$\begin{aligned} & h(X) \\ &= -E[\log_2 f_X(x)] \\ &= -\int f_X(x) \log_2 f_X(x) dx \\ &= -\int f_X(x) \log_2 \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right) dx \\ &= \tfrac{1}{2} \log_2 (2\pi e \sigma^2) \end{aligned} . \tag{32}$$

Figure 11 shows the differential entropy as a function σ^2 for the univariate Gaussian variable, which is concave downward and grows first very fast and then much more slowly at high values of σ^2.

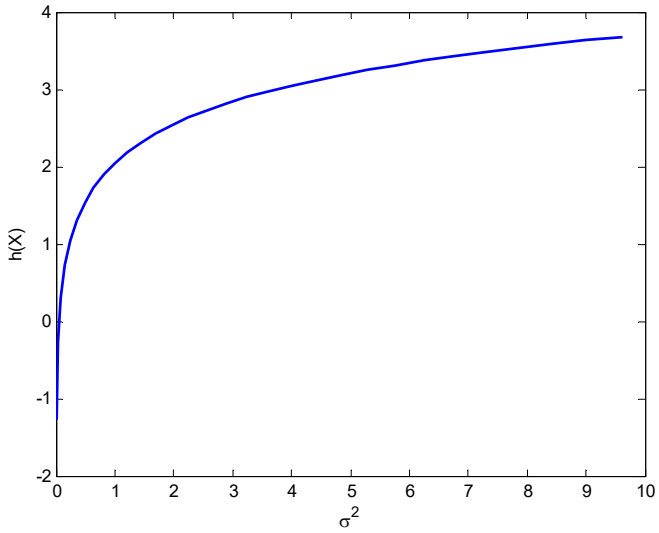


Figure 11. The differential entropy as a function of σ^2 for a univariate Gaussian variable.

3.2. Entropy of a Multivariate Gaussian Distribution

Let $\mathbf{X}$ follow a *multivariate* Gaussian distribution $\mathbf{X} \sim N(\mu, \sum)$, as given by Equation (2), then the differential entropy of $\mathbf{X}$ in nats is:

$$h(\mathbf{X}) = -E[\log_2 f_{\mathbf{X}}(\mathbf{x})] = -\int f_{\mathbf{X}}(\mathbf{x}) \log_2 f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}, \tag{33}$$

and the differential entropy is given by Appendix B:

$$h(\mathbf{X}) = \frac{1}{2} \log_2((2\pi e)^n |\sum|). \tag{34}$$

The above calculation involves the evaluation of expectations of the Mahalanobis distance as (Appendix C):

$$E[(\mathbf{x} - \boldsymbol{\mu})^T \sum^{-1} (\mathbf{x} - \boldsymbol{\mu})] = n. \tag{35}$$

For a fixed variance, the normal distribution is the pdf that maximizes entropy. Let $\mathbf{X} = \begin{bmatrix} X_1 & X_2 \end{bmatrix}^{\mathrm{T}}$ be a 2D Gaussian vector, and the entropy of $\mathbf{X}$ can be calculated to be:

$$h(\mathbf{X}) = h(X_1, X_2) = \frac{1}{2} \log_2\left((2\pi e)^2 |\sum|\right) = \log_2(2\pi e \sigma_1 \sigma_2 \sqrt{1-\rho^2}), \tag{36}$$

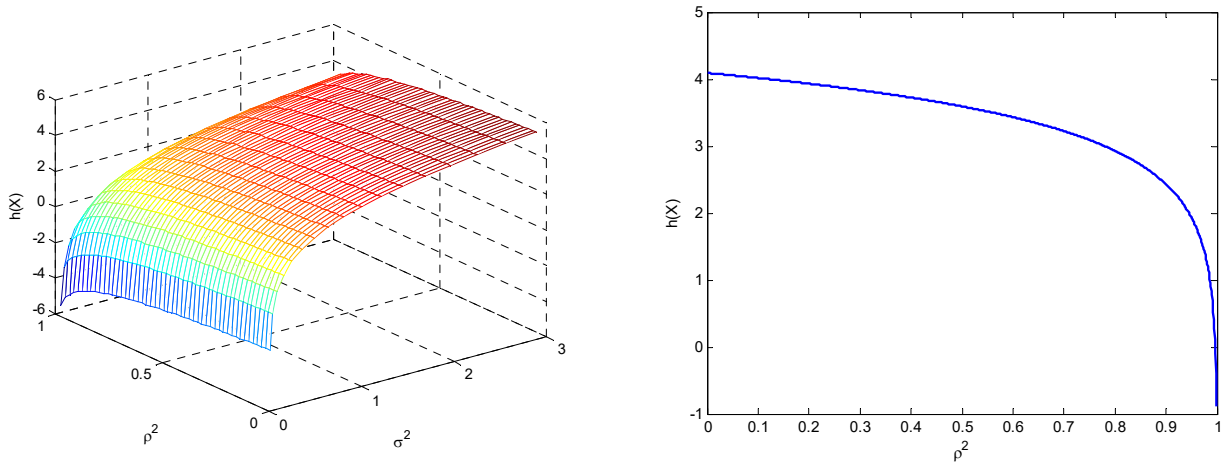
with a covariance matrix:

$$\sum = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

If $\sigma_1 = \sigma_2 = \sigma$, this becomes:

$$h(X_1, X_2) = \log_2(2\pi e \sigma^2 \sqrt{1-\rho^2}), \tag{37}$$

which is a function of ρ^2 concave downward and grows first very fast and then much more slowly for high ρ^2 values, as shown in Figure 12.



(**a**) (**b**)

Figure 12. The variation in differential entropy for the bivariate Gaussian distribution (**a**) as a function of ρ^2 and σ^2, and (**b**) as a function of ρ^2 when $\sigma_1 = \sigma_2 = 1$.

3.3. The Differential Entropy in the Transformed Frame

The differential entropy is invariant to a translation (change in the mean of the pdf):

$$h(X + a) = h(X),$$

and

$$h(bX) = h(X) + \log_2 |b|.$$

For a random variable vector, the differential entropy in the transformed frame remains the same as the one in the original frame. It can be shown in general that:

$$h(\mathbf{Y}) = h(\mathbf{UX}) = h(\mathbf{X}) + \log_2 |\mathbf{U}| = h(\mathbf{X}). \tag{38}$$

For the case of a multivariate Gaussian distribution, we have:

$$h(\mathbf{X}) = \frac{1}{2}\log_2\Big((2\pi e)^n |\sum|\Big) = \frac{n}{2}\log_2(2\pi e) + \frac{1}{2}\log_2|\sum| = \frac{n}{2}\log_2(2\pi e) + \sum_{i=1}^{n}\frac{1}{2}\log_2\lambda_i$$

It is known that the determinant of the covariance matrix is equal to the product of its eigenvalues:

$$|\sum| = \prod_{i=1}^{n}\lambda_i.$$

For the case of a bivariate Gaussian distribution, $n = 2$, we have:

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \prod_{i=1}^{2}\frac{1}{\sqrt{2\pi}\sqrt{\lambda_i}}e^{-\frac{1}{2}\frac{y_i^2}{\lambda_i}} \\ &= \frac{1}{\sqrt{2\pi}\sqrt{\lambda_1}}e^{-\frac{1}{2}\frac{y_1^2}{\lambda_1}} \cdot \frac{1}{\sqrt{2\pi}\sqrt{\lambda_2}}e^{-\frac{1}{2}\frac{y_2^2}{\lambda_2}} \\ &= \frac{1}{2\pi\sqrt{\lambda_1\lambda_2}}e^{-\frac{1}{2}(\frac{y_1^2}{\lambda_1}+\frac{y_2^2}{\lambda_2})} \end{aligned} \tag{39}$$

It can be shown that the entropy in the transformed frame is given by:

$$h(\mathbf{Y}) = \frac{2}{2}\log_2(2\pi e) + \sum_{i=1}^{2}\log_2(\lambda_i) = \log_2(2\pi e) + \log_2(\lambda_1 \cdot \lambda_2).$$

Detailed derivations are provided in Appendix D. As discussed, the determinant of the covariance matrix is equal to the product of its eigenvalues:

$$\begin{aligned} |\sum| &= \lambda_1 \cdot \lambda_2 \\ &= \left(\frac{1}{2}\left[\sigma_1^2 + \sigma_2^2 + \sqrt{\left(\sigma_1^2 - \sigma_2^2\right)^2 + 4\sigma_1^2\sigma_2^2\rho^2}\right]\right)\left(\frac{1}{2}\left[\sigma_1^2 + \sigma_2^2 - \sqrt{\left(\sigma_1^2 - \sigma_2^2\right)^2 + 4\sigma_1^2\sigma_2^2\rho^2}\right]\right), \\ &= \sigma_1^2\sigma_2^2(1-\rho^2) \end{aligned} \tag{40}$$

and thus, the entropy can be presented as:

$$h(Y_1, Y_2) = \frac{1}{2}\log_2(2\pi e)^2|\sum| = \frac{1}{2}\log_2(2\pi e)^2\sigma_1^2\sigma_2^2(1-\rho^2)) = \log_2(2\pi e\sigma_1\sigma_2\sqrt{1-\rho^2}). \tag{41}$$

The result confirms the statement that the differential entropy remains unchanged in the transformed frame.

4. Relative Entropy (Kullback–Leibler Divergence)

In this section, various important issues regarding relative entropy (Kullback–Leibler divergence) are discussed. Despite the aforementioned flaws, there is a possibility of information theory in the continuous case. A key result is that the definitions for relative entropy and mutual information follow naturally from the discrete case and retain their usefulness.

The relative entropy is a type of statistical distance that provides a measure of probability distribution $f_{\mathbf{X}}$, is different from a second reference probability distribution $g_{\mathbf{X}}$, denoted as:

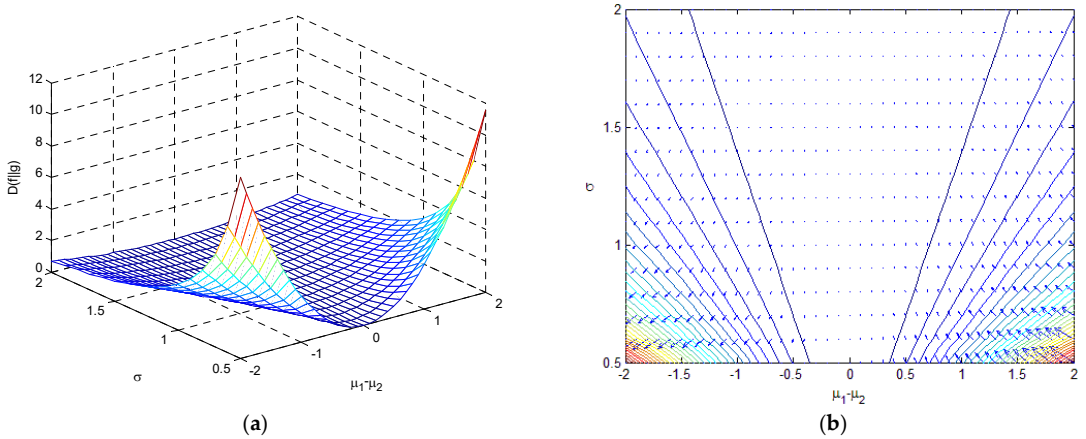
$$D_{KL}(f\|g) = \int f_{\mathbf{X}}(\mathbf{x})\log_2\frac{f_{\mathbf{X}}(\mathbf{x})}{g_{\mathbf{X}}(\mathbf{x})}d\mathbf{x}. \tag{42}$$

A detailed derivation is provided in Appendix E. The relative entropy between two Gaussian distributions with different mean and variance is given by:

$$D_{KL}(f\|g) = \frac{1}{2}\left[\ln\left(\frac{\sigma_2^2}{\sigma_1^2}\right) + \frac{\sigma_1^2}{\sigma_2^2} + \left(\frac{\mu_1-\mu_2}{\sigma_2}\right)^2 - 1\right]\cdot\log_2 e. \tag{43}$$

It is worth noting that the relative entropy measured in bits where $\log_2$ is used in the definition. However, if ln is used, then it would be measured in nats. The only difference in the expression is the $\log_2 e$ factor. Several conditions are discussed with examples of the characteristics of relative entropy:

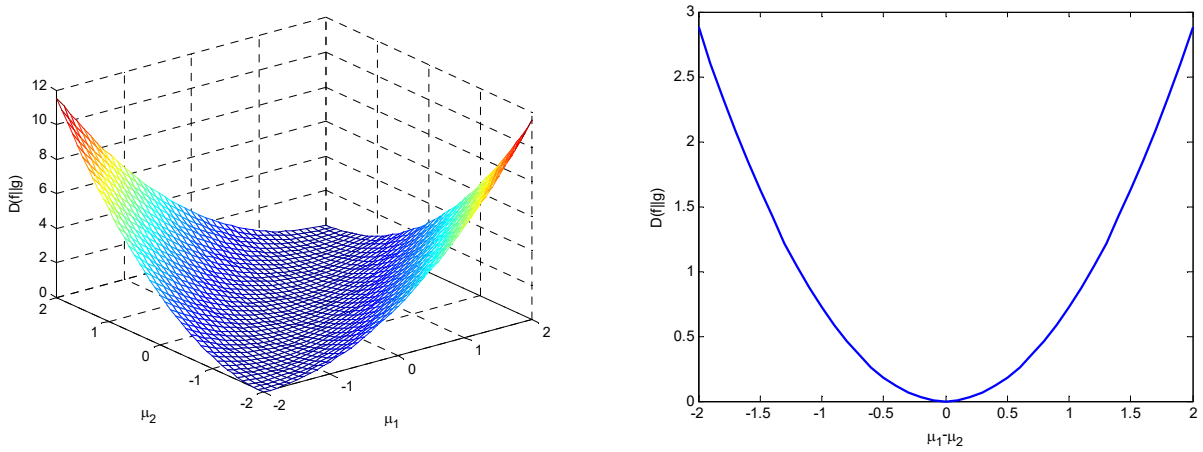
(1) If $\sigma_1 = \sigma_2 = \sigma$, $D_{KL}(f\|g) = \frac{1}{2}\left(\frac{\mu_1-\mu_2}{\sigma}\right)^2\log_2 e$, which is 0 when $\mu_1 = \mu_2$. Figure 13 shows the relative entropy as a function of σ and $\mu_1 - \mu_2$ when $\sigma_1 = \sigma_2 = \sigma$, where a three-dimensional surface and a contour with an entropy gradient are provided.



(**a**) (**b**)

Figure 13. The variation in relative entropy as a function of σ and $\mu_1 - \mu_2$ when $\sigma_1 = \sigma_2 = \sigma$ for (**a**) a three-dimensional surface; and (**b**) a contour with an entropy gradient.

(2) If $\sigma_1 = \sigma_2 = 1$, $D_{KL}(f\|g) = \frac{1}{2}(\mu_1 - \mu_2)^2 \cdot \log_2 e$, which is an even function with a minimum value of 0 when $\mu_1 = \mu_2$. Figure 14 illustrates the variations of relative entropy as a function of μ_1 and μ_2 and as a function of $\mu_1 - \mu_2$.

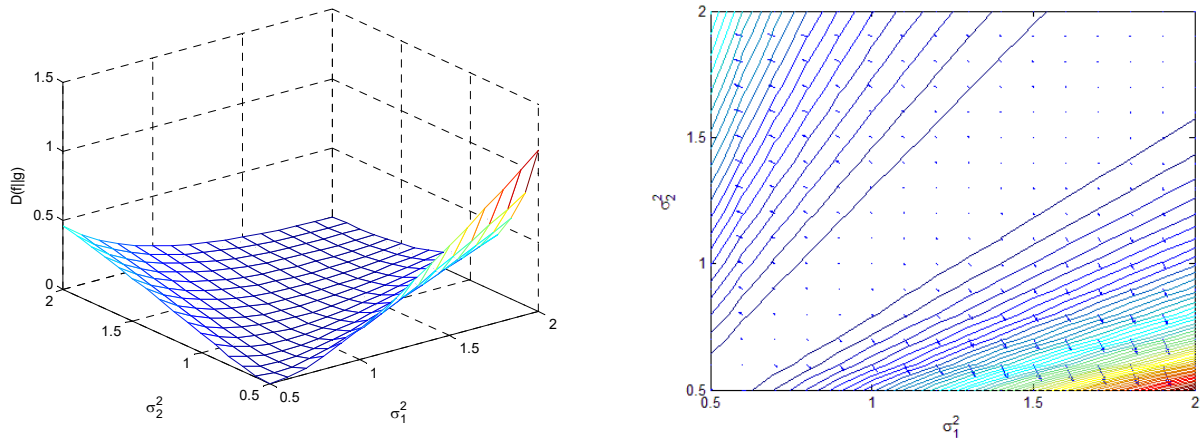


(**a**) (**b**)

Figure 14. The variations in relative entropy with $\sigma_1 = \sigma_2 = 1$ for (**a**) a three-dimensional surface as a function of μ_1 and μ_2; and (**b**) as a function of $\mu_1 - \mu_2$.

- If $\mu_2 = 0$, $D_{KL}(f\|g) = \frac{1}{2}\mu_1^2 \log_2 e$, it is a function of μ_1 concave upward.
- If $\mu_1 = 0$, $D_{KL}(f\|g) = \frac{1}{2}\mu_2^2 \log_2 e$, it is a function of μ_2 concave upward.

(3) If $\mu_1 = \mu_2$, $D_{KL}(f\|g) = \frac{1}{2}\left[\ln\left(\frac{\sigma_2^2}{\sigma_1^2}\right) + \frac{\sigma_1^2}{\sigma_2^2} - 1\right] \cdot \log_2 e$. Figure 15 demonstrates relative entropy as a function of σ_1 and σ_2 when $\mu_1 = \mu_2$, where a three-dimensional surface and the contour with an entropy gradient are plotted.



(**a**) (**b**)

Figure 15. The variation in relative entropy as a function of σ_1 and σ_2 with $\mu_1 = \mu_2$ for (**a**) at the three-dimensional surface; and (**b**) contour with an entropy gradient.

When $\sigma_2 = 1$, $D_{KL}(f\|g) = \frac{1}{2}\left[\ln\left(\frac{1}{\sigma_1^2}\right) + \sigma_1^2 - 1\right] \cdot \log_2 e$.

When $\sigma_1 = 1$, $D_{KL}(f\|g) = \frac{1}{2}\left[\ln(\sigma_2^2) + \frac{1}{\sigma_2^2} - 1\right] \cdot \log_2 e$.

Figure 16 illustrates the variations of relative entropy as a function of the variance when the other variance is unity under the condition $\mu_1 = \mu_2$.

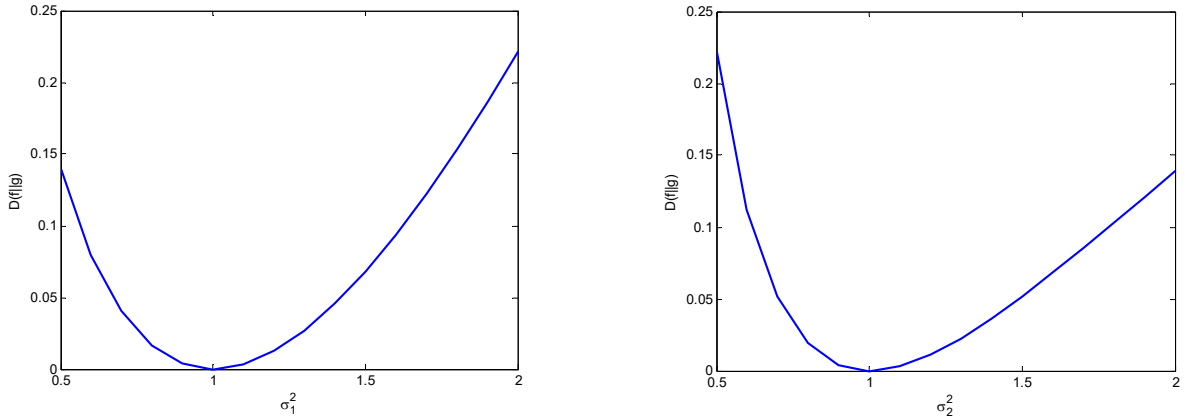


Figure 16. Variations of relative entropy as a function of (**a**) σ_1 when fixed $\sigma_2 = 1$ and (**b**) σ_2 when fixed $\sigma_1 = 1$, respectively ($\mu_1 = \mu_2$).

A sensitivity analysis of the relative entropy due to changes in variance and mean is carried out. The gradient of $D_{KL}(f\|g)$ given by:

$$\frac{\partial D_{KL}(\sigma_1, \sigma_2, \mu_1, \mu_2)}{\partial \mathbf{x}} = \left[\frac{\partial D_{KL}}{\partial \sigma_1} \quad \frac{\partial D_{KL}}{\partial \sigma_2} \quad \frac{\partial D_{KL}}{\partial \mu_1} \quad \frac{\partial D_{KL}}{\partial \mu_2}\right],$$

can be calculated with the partial derivatives where the chain rule is involved. Based on the relationship $\frac{d}{dx}\ln x = \frac{1}{x}$, we have:

$$\frac{\partial}{\partial \sigma_1}\left[\ln\left(\frac{\sigma_2^2}{\sigma_1^2}\right)\right] = \frac{\sigma_1^2}{\sigma_2^2}\cdot(-2)\sigma_2^2\sigma_1^{-3} = -\frac{2}{\sigma_1},$$

and the following expressions are obtained:

(1) $\frac{\partial D_{KL}}{\partial \sigma_1} = \frac{\partial}{\partial \sigma_1}\left[\ln\left(\frac{\sigma_2^2}{\sigma_1^2}\right) + \frac{\sigma_1^2}{\sigma_2^2}\right]\cdot\frac{1}{2}\log_2 e = \left(\frac{\sigma_1}{\sigma_2^2} - \frac{1}{\sigma_1}\right)\cdot\log_2 e,$

(2) $\frac{\partial D_{KL}}{\partial \sigma_2} = \frac{\partial}{\partial \sigma_2}\left[\ln\left(\frac{\sigma_2^2}{\sigma_1^2}\right) + \frac{\sigma_1^2}{\sigma_2^2} + \frac{(\mu_1-\mu_2)^2}{\sigma_2^2}\right]\cdot\frac{1}{2}\log_2 e = \left[\frac{1}{\sigma_2} - \frac{\sigma_1^2}{\sigma_2^3} - \frac{(\mu_1-\mu_2)^2}{\sigma_2^3}\right]\cdot\log_2 e,$

(3) $\frac{\partial D_{KL}}{\partial \mu_1} = \frac{\partial}{\partial \mu_1}\left(\frac{\mu_1-\mu_2}{\sigma_2}\right)^2\cdot\frac{1}{2}\log_2 e = \left(\frac{\mu_1-\mu_2}{\sigma_2^2}\right)\cdot\log_2 e,$

(4) $\frac{\partial D_{KL}}{\partial \mu_2} = \frac{\partial}{\partial \mu_2}\left(\frac{\mu_1-\mu_2}{\sigma_2}\right)^2\cdot\frac{1}{2}\log_2 e = \left(\frac{\mu_2-\mu_1}{\sigma_2^2}\right)\cdot\log_2 e.$

The optimal condition for each of the above cases can be:

$\frac{\partial D_{KL}}{\partial \sigma_1} = \frac{\sigma_1}{\sigma_2^2} - \frac{1}{\sigma_1} = 0$ when $\sigma_1^2 = \sigma_2^2$,

$\frac{\partial D_{KL}}{\partial \sigma_2} = \frac{1}{\sigma_2} - \frac{\sigma_1^2}{\sigma_2^3} - \frac{(\mu_1-\mu_2)^2}{\sigma_2^3} = 0$ when $\sigma_2^2 = \sigma_1^2 + (\mu_1-\mu_2)^2$,

$\left(\frac{\mu_1-\mu_2}{\sigma_2^2}\right)\cdot\log_2 e = 0$ when $\mu_1 = \mu_2$,

$\left(\frac{\mu_2-\mu_1}{\sigma_2^2}\right)\cdot\log_2 e = 0$ when $\mu_1 = \mu_2$.

5. Mutual Information

Mutual information is one of many quantities that measures one random variable and tells us about another. It is a dimensionless quantity with (generally) units of bits and can be thought of as the reduction in uncertainty about one random variable given knowledge

of another. The mutual information $I(X;Y)$ between two variables with joint pdf $f_{XY}(x,y)$ is given by:

$$I(X;Y) = E\left[\log \frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}\right] = \int\int f_{XY}(x,y) \log \frac{f_{XY}(x,y)}{f_X(x)f_Y(y)} dxdy. \tag{44}$$

The mutual information between the random variables X and Y has the following relationships:

$$I(X;Y) = I(Y;X), \tag{45}$$

where

$$I(X;Y) = h(X) - h(X|Y) \geqslant 0, \tag{46}$$

and

$$I(Y;X) = h(Y) - h(Y|X) \geqslant 0, \tag{47}$$

implying that $h(X) \geqslant h(X|Y)$ and $h(Y) \geqslant h(Y|X)$. The mutual information of a random variable with itself is self-information, which is entropy. High mutual information indicates a large reduction in uncertainty; low mutual information indicates a small reduction; and zero mutual information between two random variables, $I(X;Y) = 0$, meaning that the variables are independent. In such a case, $h(X) = h(X|Y)$ and $h(Y) = h(Y|X)$.

Let's consider the mutual information between the correlated Gaussian variables X and Y given by:

$$\begin{array}{l} I(X;Y) = h(X) + h(Y) - h(X,Y) \\ = \frac{1}{2}\log_2(2\pi e)\sigma_x^2 + \frac{1}{2}\log_2(2\pi e)\sigma_y^2 - \frac{1}{2}\log_2(2\pi e)^2\sigma_x^2\sigma_y^2(1-\rho^2) \\ = -\frac{1}{2}\log_2(1-\rho^2) \end{array} . \tag{48}$$

Figure 17 presents the mutual information versus ρ^2, where it grows first much slower and then very fast for high values of ρ^2. If $\rho = \pm 1$, the random variables X and Y are perfectly correlated, the mutual information is infinite. It can be seen that $I(X;Y) = 0$ for $\rho = 0$ and that $I(X;Y) \to \infty$ for $\rho \to \pm 1$.

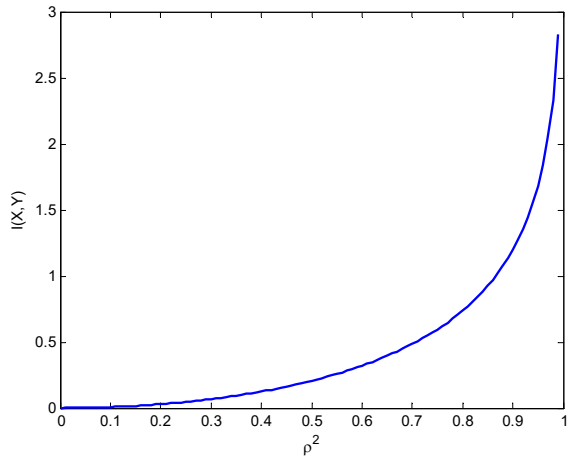


Figure 17. Mutual information versus ρ^2 between the correlated Gaussian variables.

On the other hand, considering the additive white Gaussian noise (AWGN) channel, shown in Figure 18, the mutual information is given by:

$$I(X;Y) = h(Y) - h(Y|X) = \frac{1}{2}\log_2\left(\frac{2\pi e(\sigma_x^2 + \sigma_n^2)}{2\pi e\sigma_n^2}\right) = \frac{1}{2}\log_2\left(1 + \frac{\sigma_x^2}{\sigma_n^2}\right), \tag{49}$$

where $h(Y|X) = h(N) = h(X,Y) - h(X)$, and

$$h(\mathrm{Y}) = \frac{1}{2}\log_2\Big(2\pi e(\sigma_x^2 + \sigma_n^2)\Big);\ h(Y|X) = h(N) = \frac{1}{2}\log_2(2\pi e)\sigma_n^2$$

Mutual information for the additive white Gaussian noise (AWGN) channel is shown in Figure 19, including the three-dimensional surface as a function of σ_x^2 and σ_n^2, and also in terms of the signal-to-noise ratio SNR $= \sigma_x^2/\sigma_n^2$. It can be seen that the mutual information grows first very fast and then much more slowly for high values of the signal-to-noise ratio.

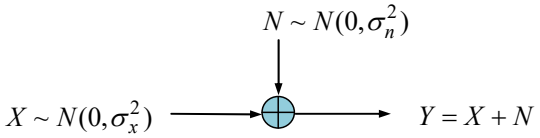


Figure 18. Schematic illustration of the additive white Gaussian noise (AWGN) channel.

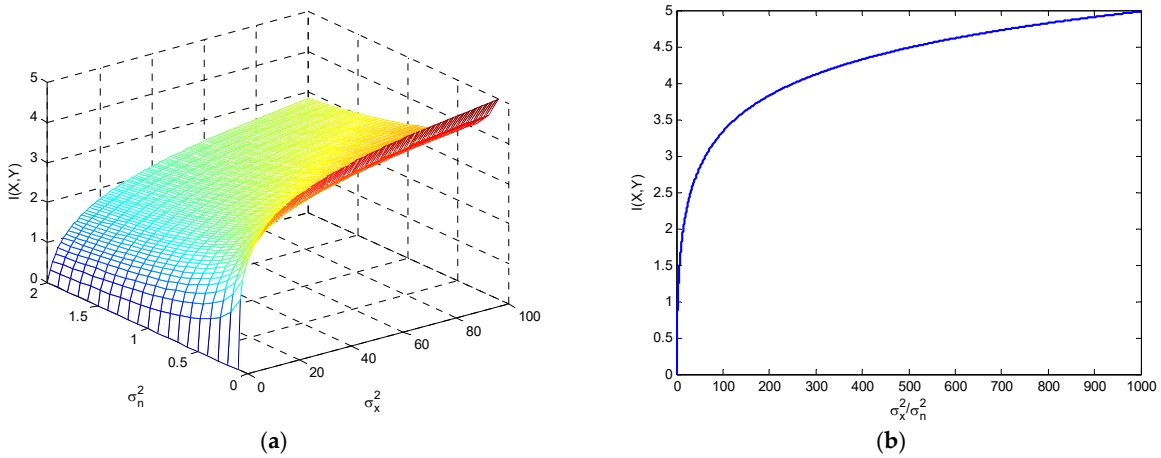


Figure 19. The mutual information with the additive white Gaussian noise (AWGN) channel for (**a**) the three-dimensional surface as a function of σ_x^2 and σ_n^2; and (**b**) in terms of the signal-to-noise ratio.

6. Conclusions

This paper intends to serve the readers as a supplement note for the multivariate Gaussian distribution with its entropy, relative entropy, and mutual information. The illustrative examples are discussed to provide further insights into the geometric interpretation and visualization, enabling the readers to correctly interpret the theory for future design. The fundamental objective is to study the application of multivariate sets of data to a Gaussian distribution. This paper examines broad measurements of structure for Gaussian distributions, which show that they can be described in terms of the information theory between the given covariance matrix and correlated random variables (in terms of relative entropy). To develop the multivariate Gaussian distribution with entropy and mutual information, several significant methodologies are presented through the discussion supported by illustrations, both technically and statistically. The content obtained allows readers to better perceive concepts, comprehend techniques, and properly execute software programs for future study on the topic's science and implementations. It also helps readers grasp the themes' fundamental concepts. Involving the relative entropy and mutual information as well as the potential correlated covariance analysis based on differential equations, a wide range of information is addressed, from basic to application concerns. Moreover, the proposed techniques of multivariate Gaussian distribution and mutual information are intended to inspire new applications of information theory to a number of areas, including information coding, nonlinear signal detection, and clinical diagnostic testing, particularly when data from improved testing equipment becomes accessible.

Author Contributions: Conceptualization, D.-J.J.; methodology, D.-J.J.; software, D.-J.J.; validation, D.-J.J. and T.-S.C.; writing—original draft preparation, D.-J.J. and T.-S.C.; writing—review and editing, D.-J.J., T.-S.C. and A.B.; supervision, D.-J.J. All authors have read and agreed to the published version of the manuscript.

Funding: The author gratefully acknowledges the support of the National Science and Technology Council, Taiwan, under grant number NSTC 111-2221-E-019-047.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Derivation of the Differential Entropy for the Univariate Gaussian Distribution

$$
\begin{aligned}
& h(X) \\
&= -E[\log_2 f_X(x)] \\
&= -\int f_X(x) \log_2 f_X(x) dx \\
&= -\int f_X(x) \log_2 \left(\tfrac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right) dx \\
&= -\int f_X(x) \left(\log_2 (2\pi\sigma^2)^{-\frac{1}{2}} + \log_2 e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right) dx \\
&= -\int f_X(x) \left[\left(-\tfrac{1}{2} \log_2(2\pi\sigma^2) - \tfrac{(x-\mu)^2}{2\sigma^2} \log_2 e \right) \right] dx \\
&= \tfrac{1}{2} \log_2(2\pi\sigma^2) \textstyle\int_{-\infty}^{\infty} f_X(x) dx + \tfrac{\log_2 e}{2\sigma^2} \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x) dx \\
&= \tfrac{1}{2} \log_2(2\pi\sigma^2) + \tfrac{\sigma^2}{2\sigma^2} \log_2 e \\
&= \tfrac{1}{2} \log_2(2\pi e\sigma^2)
\end{aligned}
$$

Appendix B. Derivation of the Differential Entropy for the Multivariate Gaussian Distribution

$$
\begin{aligned}
& h(\mathbf{X}) \\
&= -E\left[\log_2 f_{\mathbf{X}}(\mathbf{x})\right] \\
&= -E\left[\log_2 \left(\tfrac{1}{\sqrt{(2\pi)^n |\sum|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \sum^{-1} (\mathbf{x}-\boldsymbol{\mu})} \right) \right] \\
&= -E\left[-\tfrac{n}{2} \log_2(2\pi) - \tfrac{1}{2} \log_2 |\sum| - \log_2 e^{\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \sum^{-1} (\mathbf{x}-\boldsymbol{\mu})} \right] \\
&= -\int f_{\mathbf{X}}(\mathbf{x}) \left[\left(-\tfrac{1}{2} \log_2 (2\pi)^n |\sum|) - \tfrac{\log_2 e}{2} ((\mathbf{x}-\boldsymbol{\mu})^T \sum^{-1} (\mathbf{x}-\boldsymbol{\mu})) \right) \right] d\mathbf{x} \\
&= \tfrac{1}{2} \log_2((2\pi)^n |\sum|) + \tfrac{n}{2} \log_2 e \\
&= \tfrac{n}{2} \log_2(2\pi) + \tfrac{1}{2} \log_2 |\sum| + \log_2 e^{\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \sum^{-1} (\mathbf{x}-\boldsymbol{\mu})} \\
&= \tfrac{n}{2} \log_2(2\pi) + \tfrac{1}{2} \log_2 |\sum| + \log_2 e^{\frac{n}{2}} \\
&= \tfrac{n}{2} \log_2(2\pi) + \tfrac{1}{2} \log_2 |\sum| + \tfrac{n}{2} \log_2 e \\
&= \tfrac{1}{2} \log_2 \left((2\pi e)^n |\sum| \right)
\end{aligned}
$$

The calculation involves the evaluation of expectations of the Mahalanobis distance.

Appendix C. Evaluation of Expectations of the Mahalanobis Distance

$E[(\mathbf{x}-\boldsymbol{\mu})^T \sum^{-1}(\mathbf{x}-\boldsymbol{\mu})] = n$

$$\begin{aligned}
&E[(\mathbf{x}-\boldsymbol{\mu})^T \sum^{-1}(\mathbf{x}-\boldsymbol{\mu})] \\
&= E[tr((\mathbf{x}-\boldsymbol{\mu})^T \sum^{-1}(\mathbf{x}-\boldsymbol{\mu}))] \\
&= E[tr(\sum^{-1}(\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu}))] \\
&= tr(\sum^{-1} E[(\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu})]) \\
&= tr(\sum^{-1}\sum) \\
&= tr(I_n) \\
&= n
\end{aligned}$$

A special case for $n = 1$

$$\begin{aligned}
&E[(\mathbf{x}-\boldsymbol{\mu})^T \sum^{-1}(\mathbf{x}-\boldsymbol{\mu})] \\
&= E\left[\tfrac{(x-\mu)^2}{\sigma^2}\right] \\
&= \int f_X(x)\left(\tfrac{(x-\mu)^2}{\sigma^2}\right)dx \\
&= \tfrac{1}{\sigma^2}\int (x-\mu)^2 f_X(x)dx \\
&= 1
\end{aligned}$$

Appendix D. Derivation of the Differential Entropy in the Transformed Frame

$$\begin{aligned}
&h(\mathbf{Y}) \\
&= -E\left[\log_2\left(\prod_{i=1}^{n}\tfrac{1}{\sqrt{2\pi}\sqrt{\lambda_i}}e^{-\frac{1}{2}\frac{y_i^2}{\lambda_i}}\right)\right] \\
&= -E\left[\sum_{i=1}^{n}\log_2\left(\tfrac{1}{\sqrt{2\pi}\sqrt{\lambda_i}}e^{-\frac{1}{2}\frac{y_i^2}{\lambda_i}}\right)\right] \\
&= -\sum_{i=1}^{n}E\left[\log_2\left(\tfrac{1}{\sqrt{2\pi}\sqrt{\lambda_i}}e^{-\frac{1}{2}\frac{y_i^2}{\lambda_i}}\right)\right] \\
&= -\sum_{i=1}^{n}E\left[\log_2(\tfrac{1}{\sqrt{2\pi}\sqrt{\lambda_i}}) + \log_2 e^{-\frac{1}{2}\frac{y_i^2}{\lambda_i}}\right] \\
&= -\sum_{i=1}^{n}\left[\int f_Y(y_i)\left(\log_2(\tfrac{1}{\sqrt{2\pi}\sqrt{\lambda_i}}) + \log_2 e^{-\frac{1}{2}\frac{y_i^2}{\lambda_i}}\right)dy_i\right] \\
&= -\sum_{i=1}^{n}\left[\int f_Y(y_i)\left(-\tfrac{1}{2}\log_2(2\pi\lambda_i) - \tfrac{1}{2}\tfrac{y_i^2}{\lambda_i}\log_2 e\right)dy_i\right] \\
&= \sum_{i=1}^{n}\left[\tfrac{1}{2}\log_2(2\pi\lambda_i) + \tfrac{1}{2}\log_2 e\right] \\
&= \sum_{i=1}^{n}\left[\tfrac{1}{2}\log_2(2\pi) + \tfrac{1}{2}\log_2(\lambda_i) + \tfrac{1}{2}\log_2 e\right] \\
&= \tfrac{n}{2}\log_2(2\pi e) + \tfrac{1}{2}\sum_{i=1}^{n}\log_2(\lambda_i)
\end{aligned}$$

The eigenvalues λ_i are the diagonal elements of the covariance matrix, namely variances, in the transformed frame. When $\rho = 0$, the eigenvectors are equal to $\lambda_i = \sigma_i^2$.

Appendix E. Derivation of the Kullback–Leibler Divergence between Two Normal Distributions

$$
\begin{aligned}
&D_{KL}(f\|g)\\
&=\int f_X(x)\log_2\frac{f_X(x)}{g_X(x)}dx\\
&=\int f_X(x)\log_2\frac{\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2}}{\frac{1}{\sqrt{2\pi}\sigma_2}e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2}}dx\\
&=\int f_X(x)\log_2\left(\frac{\sigma_2}{\sigma_1}\right)dx+\int f_X(x)\log_2\left[\exp\left(-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2+\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2\right)\right]dx\\
&=\log_2\left(\frac{\sigma_2}{\sigma_1}\right)-\frac{\log_2 e}{2\sigma_1^2}\int f_X(x)(x-\mu_1)^2dx+\frac{\log_2 e}{2\sigma_2^2}\int f_X(x)(x-\mu_2)^2dx\\
&=\log_2\left(\frac{\sigma_2}{\sigma_1}\right)-\frac{\log_2 e}{2}+\frac{\log_2 e}{2\sigma_2^2}\int f_X(x)(x-\mu_1+\mu_1-\mu_2)^2dx\\
&=\log_2\left(\frac{\sigma_2}{\sigma_1}\right)-\frac{\log_2 e}{2}+\frac{\log_2 e}{2\sigma_2^2}\int f_X(x)\Big((x-\mu_1)^2+(\mu_1-\mu_2)^2+2(x-\mu_1)(\mu_1-\mu_2)\Big)dx\\
&=\frac{1}{2}\log_2\left(\frac{\sigma_2^2}{\sigma_1^2}\right)-\frac{\log_2 e}{2}+\frac{\log_2 e}{2\sigma_2^2}\left[\sigma_1^2+(\mu_1-\mu_2)^2\right]\\
&=\frac{1}{2}\left[\ln\left(\frac{\sigma_2^2}{\sigma_1^2}\right)+\frac{\sigma_1^2}{\sigma_2^2}+\left(\frac{\mu_1-\mu_2}{\sigma_2}\right)^2-1\right]\cdot\log_2 e
\end{aligned}
$$

where equality $\log_2(\cdot)=\log_2 e\cdot\ln(\cdot)$ was used.

References

1. Verdú, S. On channel capacity per unit cost. *IEEE Trans. Inf. Theory* **1990**, *36*, 1019–1030. [CrossRef]
2. Lapidoth, A.; Shamai, S. Fading channels: How perfect need perfect side information be? *IEEE Trans. Inf. Theory* **2002**, *48*, 1118–1134. [CrossRef]
3. Verdú, S. Spectral efficiency in the wideband regime. *IEEE Trans. Inf. Theory* **2002**, *48*, 1319–1343. [CrossRef]
4. Prelov, V.; Verdú, S. Second-order asymptotics of mutual information. *IEEE Trans. Inf. Theory* **2004**, *50*, 1567–1580. [CrossRef]
5. Kailath, T. A general likelihood-ratio formula for random signals in Gaussian noise. *IEEE Trans. Inf. Theory* **1969**, *IT-15*, 350–361. [CrossRef]
6. Kailath, T. A note on least squares estimates from likelihood ratios. *Inf. Control* **1968**, *13*, 534–540. [CrossRef]
7. Kailath, T. A further note on a general likelihood formula for random signals in Gaussian noise. *IEEE Trans. Inf. Theory* **1970**, *IT-16*, 393–396. [CrossRef]
8. Jaffer, A.G.; Gupta, S.C. On relations between detection and estimation of discrete time processes. *Inf. Control* **1972**, *20*, 46–54. [CrossRef]
9. Duncan, T.E. On the calculation of mutual information. *SIAM J. Appl. Math.* **1970**, *19*, 215–220. [CrossRef]
10. Kadota, T.T.; Zakai, M.; Ziv, J. Mutual information of the white Gaussian channel with and without feedback. *IEEE Trans. Inf. Theory* **1971**, *17*, 368–371. [CrossRef]
11. Amari, S.I. *Information Geometry and Its Applications*; Springer: Berlin/Heidelberg, Germany, 2016; Volume 194.
12. Schneidman, E.; Still, S.; Berry, M.J.; Bialek, W. Network information and connected correlations. *Phys. Rev. Lett.* **2003**, *91*, 238701. [CrossRef]
13. Timme, N.; Alford, W.; Flecker, B.; Beggs, J.M. Synergy, redundancy, and multivariate information measures: An experimentalist's perspective. *J. Comput. Neurosci.* **2014**, *36*, 119–140. [CrossRef]
14. Ahmed, N.A.; Gokhale, D.V. Entropy expressions and their estimators for multivariate distributions. *IEEE Trans. Inform. Theory* **1989**, *35*, 688–692. [CrossRef]
15. Misra, N.; Singh, H.; Demchuk, E. Estimation of the entropy of a multivariate normal distribution. *J. Multivar. Anal.* **2005**, *92*, 324–342. [CrossRef]
16. Arellano-Valle, R.B.; Contreras-Reyes, J.E.; Genton, M.G. Shannon entropy and mutual information for multivariate skew-elliptical distributions. *Scand. J. Stat.* **2013**, *40*, 42–62. [CrossRef]
17. Liang, K.C.; Wang, X. Gene regulatory network reconstruction using conditional mutual information. *EURASIP J. Bioinform. Syst. Biol.* **2008**, *2008*, 253894. [CrossRef] [PubMed]
18. Novais, R.G.; Wanke, P.; Antunes, J.; Tan, Y. Portfolio optimization with a mean-entropy-mutual information model. *Entropy* **2022**, *24*, 369. [CrossRef]
19. Verdú, S. Error exponents and α-mutual information. *Entropy* **2021**, *23*, 199. [CrossRef] [PubMed]
20. Panzeri, S.; Magri, C.; Logothetis, N.K. On the use of information theory for the analysis of the relationship between neural and imaging signals. *Magn. Reson. Imaging* **2008**, *26*, 1015–1025. [CrossRef]

21. Katz, Y.; Tunstrøm, K.; Ioannou, C.C.; Huepe, C.; Couzin, I.D. Inferring the structure and dynamics of interactions in schooling fish. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 18720–18725. [CrossRef] [PubMed]
22. Cutsuridis, V.; Hussain, A.; Taylor, J.G. (Eds.) *Perception-Action Cycle: Models, Architectures, and Hardware*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2011.
23. Ay, N.; Bernigau, H.; Der, R.; Prokopenko, M. Information-driven self-organization: The dynamical system approach to autonomous robot behavior. *Theory Biosci.* **2012**, *131*, 161–179. [CrossRef] [PubMed]
24. Rosas, F.; Ntranos, V.; Ellison, C.J.; Pollin, S.; Verhelst, M. Understanding interdependency through complex information sharing. *Entropy* **2016**, *18*, 38. [CrossRef]
25. Ince, R.A. The Partial Entropy Decomposition: Decomposing multivariate entropy and mutual information via pointwise common surprisal. *arXiv* **2017**, arXiv:1702.01591.
26. Harder, M.; Salge, C.; Polani, D. Bivariate measure of redundant information. *Phys. Rev. E* **2013**, *87*, 012130. [CrossRef]
27. Rauh, J.; Banerjee, P.K.; Olbrich, E.; Jost, J.; Bertschinger, N. On extractable shared information. *Entropy* **2017**, *19*, 328. [CrossRef]
28. Ince, R.A. Measuring multivariate redundant information with pointwise common change in surprisal. *Entropy* **2017**, *19*, 318. [CrossRef]
29. Perrone, P.; Ay, N. Hierarchical quantification of synergy in channels. *Front. Robot. AI* **2016**, *2*, 35. [CrossRef]
30. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.; Ay, N. Quantifying unique information. *Entropy* **2014**, *16*, 2161–2183. [CrossRef]
31. Chicharro, D.; Panzeri, S. Synergy and redundancy in dual decompositions of mutual information gain and information loss. *Entropy* **2017**, *19*, 71. [CrossRef]
32. Michalowicz, J.V.; Nichols, J.M.; Bucholtz, F. Calculation of differential entropy for a mixed Gaussian distribution. *Entropy* **2008**, *10*, 200. [CrossRef]
33. Benish, W.A. A review of the application of information theory to clinical diagnostic testing. *Entropy* **2020**, *22*, 97. [CrossRef]
34. Cadirci, M.S.; Evans, D.; Leonenko, N.; Makogin, V. Entropy-based test for generalised Gaussian distributions. *Comput. Stat. Data Anal.* **2022**, *173*, 107502. [CrossRef]
35. Goethe, M.; Fita, I.; Rubi, J.M. Testing the mutual information expansion of entropy with multivariate Gaussian distributions. *J. Chem. Phys.* **2017**, *147*, 224102. [CrossRef]