Communication

A Geometric Interpretation of 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; djjwo@mail.ntou.edu.tw (D.-J.J.), amitabiswal1988@gmail.com (A.B.)
[2] Department of Business Administration, Asia University, 500 Liufeng Road, Wufeng, Taichung 41354, Taiwan, cho2022@asia.edu.tw (T.-S.C.)
* Correspondence: djjwo@mail.ntou.edu.tw

Abstract: The fundamental objective is to study the application of multivariate sets of data in Gaussian distribution. This paper examines broad measurements of structure for both Gaussian and non-Gaussian distributions, which shows that they can be described in terms of the information-theoretic between the given covariance matrix and correlated random variables (in terms of relative entropy). 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 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, including basic to application concerns.

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

1. Introduction

Understanding the ways knowledge concerning an external variable, or the reciprocal information of its parts, is distributed across the parts of a multivariate system can assist characterize and infer 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 are the fundamental concepts of information theory and estimating theory, respectively. In contrast to the 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 estimate 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 diminish-

ing SNR are presented in [4]. [6] gives the 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 setting [8] context, the likelihood ratio is difficult in the relationship between observation and estimation [9].

Considering the specific instance of parametric computation (or Gaussian inputs), correlations relating to causal and non-causal estimation errors have been investigated in [10, 11], involving the limit on the loss owing to the causality restriction is specified. Knowing how data pertaining of an external parameter, or inversely related data within its parts, distributes across the parts of a multivariate system can assist categorize and determining the fundamental mechanics and functionality of the structure. This goal served as the impetus for the development of various techniques for decomposing the various elements of a set of parameters' joint entropy [12, 13] or to deconvolute the additions of a set of elements to the mutual information about a target variable [14]. These techniques can be used to examine a variety of intricate systems, including those in the physical distinctions domain, such as gene networks [15] or brain coding [16], as well as those in the social domain, such as selection agents [17] and community behavior [18]. They can also be used to analyze artificial agents [19]. Additionally, some new proposals diverge more significantly from the original framework, either through the adoption of novel principles, the consideration of the presence of detrimental elements linked to erroneous, or the implementation of joint entropy subdivisions in place of mutual information [20, 21].

In the multivariate scenario, the challenges of breaking down mutual information into redundancy and complimentary sections have nevertheless been significantly increased. The novel redundancy determines that were initially developed are only defined for the bivariate situation [24, 25], or allow negative components [26], whereas measurements of coordination are more readily extended to the multivariate case, especially when using the maximum entropy architecture [22, 23]. 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 [27] established two analogous techniques for constructing multivariate redundant metrics. The maximum entropy framework allows for a more straightforward generalization of the efficiency measurements to the multivariate case [24, 25].

In the 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 bivariate 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 [24]. 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, we demonstrate that, under similar presumptions, the maximum entropy reduction also plays this dominant role in the multivariate situation.

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 consecutive three sections deal with various important topics on information entropy with illustrative examples with emphasis on visualization of the information and discussion. In Section 3, continuous entropy/differential entropy is 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1909.v1

2. Geometry of the Gaussian Distribution

In this section, the background relations on Gaussian distribution for different parametric point of view has been discussed. The exploratory analysis's fundamental objective is to identify "the framework" in multivariate datasets. Ordinary least-squares regression and principal component analysis (PCA), respectively, offer typical 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.

The Gaussian pdf (a) does not have a framework in either of the above-described definitions and represents the independent variables without any settling around a lower-dimensional region. The Gaussian pdf (b), on the other hand, has greater variance along one axis over another. Despite being independent, their combined pdf is small. Although the variables are associated and therefore likewise characterized by dependency, the Gaussian pdf (c) is equally focused around one dimension as is (b).

2.1. Standard Parametric Represenatation of an Ellipse

If the data is uncorrelated and therefore has zero covariance, the ellipse is not rotated and axis aligned. The radii of the ellipse in both directions are then 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

$$\left(\frac{x}{a}\right)^2+\left(\frac{y}{b}\right)^2=1 \tag{1}$$

Figure 1 represents the construction of single points of an ellipse is due to de La Hire. It is based on the standard parametric representation.

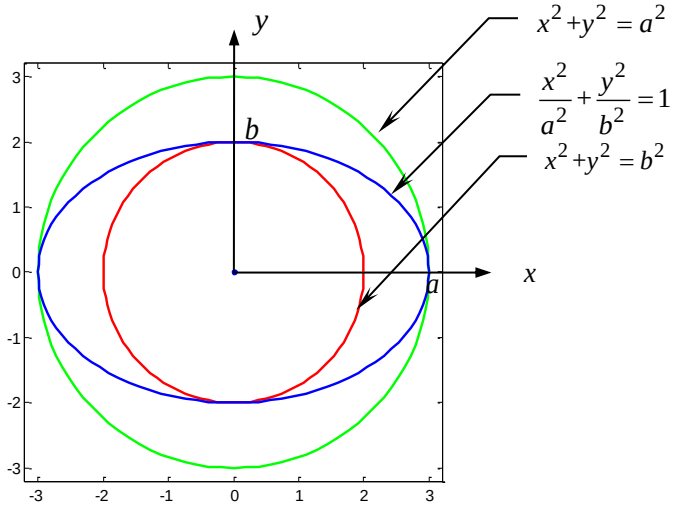


Figure 1. Standard parametric represenatation of ellipse followed by de La Hire's point construction.

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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1909.v1

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

For bivariate Gaussian distributions with zero mean, the pdf can be expressed as

$$f_{\mathbf{X}}(\mathbf{x}\,|\,\boldsymbol{\mu},\Sigma) = \frac{1}{2\pi\sqrt{|\Sigma|}} e^{\left\{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu})\right\}} \tag{3}$$

and mean and covariance matrix are given by

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix};\ \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \tag{4}$$

respectively, where the linear correlation coefficient $|\rho| \leq 1$.

Variance measures the variation of a single random variable, whereas covariance is a measure of how much two random variables vary together. With the covariance we can calculate 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 cause, the covariance matrix is often called as the variance-covariance matrix.

2.2. The Confidence Ellipse

A typical way to visualize two-dimensional Gaussian distributed data is plotting a confidence ellipse. The distance $d_M = (\mathbf{X}-\boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{X}-\boldsymbol{\mu})$ is a constant value referred to as the Mahalanobis distance, which is a random variable distributed by the chi-squared distribution, denoted as χ_k^2

$$\mathrm{P}[(\mathbf{X}-\boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{X}-\boldsymbol{\mu}) \leq \chi_k^2(\alpha)] = 1-\alpha \tag{5}$$

where k is the number of degree 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:

$$\left(\frac{x_1}{\sigma_1}\right)^2 + \left(\frac{x_2}{\sigma_2}\right)^2 = s \tag{6}$$

The goal must be to determine the scale *s* such that confidence *p* is met. Since the data is 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 also two. We now want to know the probability that the sum and therefore *s* has a certain value under a χ^2 distribution.

This ellipse, also a probability contour, defines the region of a minimum area (or volume in multivariate case) containing a given probability under the Gaussian assumption. This equation can be solved using a χ^2 table or simply using the relation $s = -2\ln(1-p)$. The confidence interval can be evaluated through $p = 1 - exp(-0.5s)$. For

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1909.v1

$s=1$, 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 $\sigma_2\sqrt{s}$.

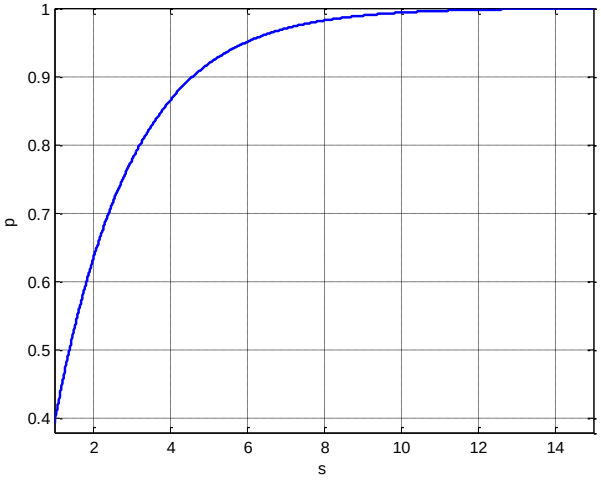


Figure 2. Relation of the confidence interval and the scale factor s.

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

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

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 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 definite therefore all eigenvalues are positive and can be seen as a linear transformation to 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 $s\cdot\Sigma$.

Based on Equations (3) and (7), the bivariate Gaussian distributions can be represented as

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1909.v1

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

Level surface of $f(x_1,x_2)$ are concentric ellipses

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

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 are 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

$$\begin{bmatrix}x_1-\mu_1 & x_2-\mu_2\end{bmatrix}\Sigma^{-1}\begin{bmatrix}x_1-\mu_1\\ x_2-\mu_2\end{bmatrix}\le\chi_2^2(0.05) \quad (10)$$

where

$$\Sigma^{-1}=\frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)}\begin{bmatrix}\sigma_2^2 & -\rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_1^2\end{bmatrix}$$

As Σ denotes a symmetric matrix, the eigenvectors of Σ 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 Σ can be expressed in terms of eigenvectors

$$\Sigma=\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{-1}=\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}=\begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_2\end{bmatrix}\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}\begin{bmatrix}\mathbf{u}_1^T\\ \mathbf{u}_2^T\end{bmatrix} \quad (11)$$

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

$$\boldsymbol{\Lambda}=\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix} \quad (12)$$

Replacing Σ by $\Sigma^{-1}=\mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^{-1}$, the square of the difference can be written as:

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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1909.v1

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

the square of the difference can then be expressed as:

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

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.

$$\frac{y_1^2}{\chi_2^2(0.05)\lambda_1} + \frac{y_2^2}{\chi_2^2(0.05)\lambda_2} \leq 1 \quad (16)$$

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 ρ=0, the eigenvectors are equal to λ_1=σ_1 and λ_2=σ_2. Also, $\mathbf{U}$ matrix whose elements are the eigenvectors of $\sum$ becomes an identity matrix. The final equation of an ellipse is then defined by

$$\frac{(x_1 - \mu_1)^2}{\chi_2^2(0.05)\lambda_1} + \frac{(x_2 - \mu_2)^2}{\chi_2^2(0.05)\lambda_2} \leq 1 \quad (17)$$

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 performated based on the three steps involving scaling, rotation, and translation.

1. Scaling

The coariance 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

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

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

Note that $\mathbf{U}$ is an orthogonal matrix $\mathbf{U}^{-1} = \mathbf{U}^T$, and $|\mathbf{U}| = 1$. Define the matrix with rorartion and scaling $\mathbf{T} = \mathbf{U}\mathbf{S}$, $\mathbf{T}^T = (\mathbf{U}\mathbf{S})^T = \mathbf{S}^T\mathbf{U}^T = \mathbf{S}\mathbf{U}^{-1}$. The covariance matrix can thus be written as $\sum = \mathbf{T}\mathbf{T}^T$ and $\mathbf{U}^T \sum \mathbf{U} = \mathbf{\Lambda}$ being diagonal with eigenvalues λ_i. Since $\mathbf{T} = \mathbf{U}\mathbf{S}$, we have $\mathbf{Y} = \mathbf{T}\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{X} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{X}$.

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1909.v1

The similarity transform is applied to obtain the relation $\mathbf{X}^T \Sigma^{-1} \mathbf{X}=\mathbf{Y}^T \mathbf{U}^T \Sigma^{-1} \mathbf{U}\mathbf{Y} = \mathbf{Y}^T \mathbf{\Lambda}^{-1}\mathbf{Y}$, and the pdf of $\mathbf{Y}$ vector can be found to be

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

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

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

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

The eigenvalues $\boldsymbol{\lambda}=\begin{bmatrix}\lambda_1 & \lambda_2\end{bmatrix}^T$ can be calculated through

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

and thus

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

From other view point for calculation of covariance matrix

$$\Sigma=\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T=\begin{bmatrix}\cos(\theta) & -\sin(\theta)\\ \sin(\theta) & \cos(\theta)\end{bmatrix}\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}\begin{bmatrix}\cos(\theta) & \sin(\theta)\\ -\sin(\theta) & \cos(\theta)\end{bmatrix}$$

$$=\begin{bmatrix}\lambda_1\cos(\theta) & -\lambda_2\sin(\theta)\\ \lambda_1\sin(\theta) & \lambda_2\cos(\theta)\end{bmatrix}\begin{bmatrix}\cos(\theta) & \sin(\theta)\\ -\sin(\theta) & \cos(\theta)\end{bmatrix}$$

$$=\begin{bmatrix}\lambda_1\cos^2(\theta)+\lambda_2\sin^2(\theta) & (\lambda_1-\lambda_2)(\sin(\theta)-\cos(\theta))\\ syms & \lambda_1\sin^2(\theta)+\lambda_2\cos^2(\theta)\end{bmatrix} \tag{25}$$

Claculation for the determinant of covariance matrix above gives the same result and the inverse is

$$\Sigma^{-1}=\frac{1}{\lambda_1\cdot\lambda_2}\begin{bmatrix}\lambda_1\sin^2(\theta)+\lambda_2\cos^2(\theta) & (\lambda_2-\lambda_1)(\sin(\theta)-\cos(\theta))\\ syms & \lambda_1\cos^2(\theta)+\lambda_2\sin^2(\theta)\end{bmatrix}$$

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

2.4. Simulation with a Given Variance-covariance Matrix

Given the data $\mathbf{X}\sim N(\boldsymbol{\mu},\Sigma)$, an ellipse representing the confidence p can be plotted by calculating the radii of the ellipse, its center and rotation. Specify θ (by

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1909.v1

which $\mathbf{U}$ can be obtained) and $\mathbf{S}$ for generating the covariance matrix $\sum$, thus ρ can be derived. The inclination angle is calculated through:

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

which can be used in calculation 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: $\sum = \mathbf{U}\Lambda\mathbf{U}^T = \mathbf{USSU}^T$ if $\mathbf{S}$ is specified. On the other way, given the correlation coefficient ρ and variances for generating the covariance matrix $\sum$, thus 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 $Y = \alpha X_1 + \beta X_2$. As for the generation of correlated random variables, if we have two Gaussian, uncorrelated random variables X_1 , X_2 then we can create 2 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 relation: $X = AZ + \mu$, $Z \sim N(0,1)$, 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 vectors of mean values .

When ρ=0, the axes of the ellipse are parallel to the original coordinate system and when $\rho \neq 0$, axes of the ellipse are aligned with the rotated axes in the transformed coordinate system. Figures 3 and 4 display ellipses drawn for various levels of confidences. The plots provide illustration of confidence (error) ellipses with different confidence levels (i.e. 68%, $s = 2.279$; 90%, $s = 4.605$; 95%, $s = 5.991$; 99%, $s = 9.210$ from inner to outer ellipses), 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 3, the position of ellipse with various correlation coefficient given by the angel of inclination, specify θ to obtain ρ, $\rho = \sigma_{12} / (\sigma_1 \sigma_2)$: (a) $\theta = 30^\circ$, $\rho \approx 0.55$; (b) $\theta = 0^\circ$, $\rho = 0$; (c) $\theta = 150^\circ$, $\rho \approx -0.55$, respectively. On the other hand, in Figure 4, the position of ellipse with various values of correlation constant given the angel of inclination, specify ρ to obtain θ: (a) $\rho = 0.95^\circ$, $\theta = 45^\circ$; (b) $\rho = 0$, $\theta = 0^\circ$; (c) $\rho = -0.95^\circ$, $\theta = 135^\circ$, respectively. Rotation angle is measrued $0 \leq \theta \leq 180^\circ$ with respect to the positive axis. When $\rho > 0$, the angle is in the first quadrant and $\rho < 0$, the angle is in the second quadrant.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1909.v1

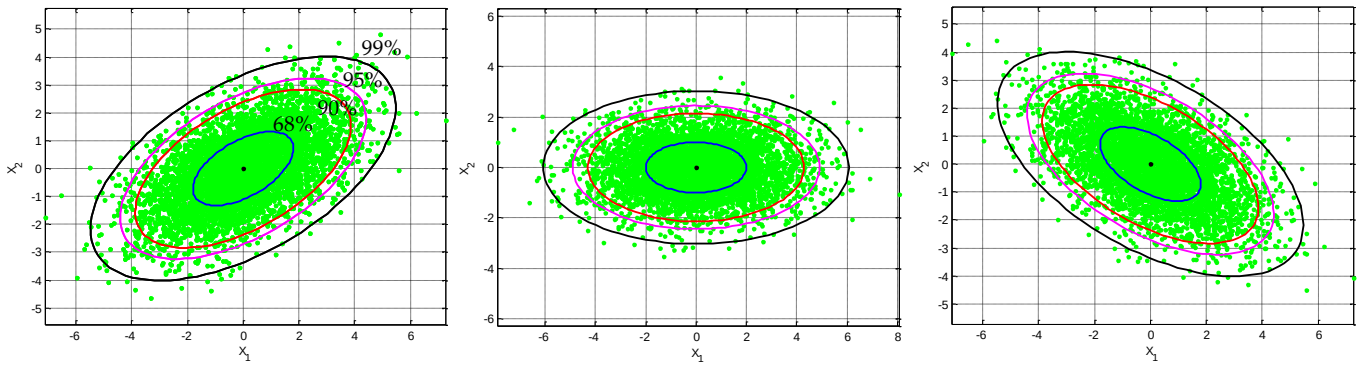


Figure 3. The position of ellipse with various correlation coefficient given by the angel of inclination, specify θ to obtain ρ, $\rho=\sigma_{12}/(\sigma_1\sigma_2)$: (a) $\theta=30^\circ$, $\rho\approx 0.55$; (b) $\theta=0^\circ$, $\rho=0$; (c) $\theta=150^\circ$, $\rho\approx -0.55$, respectively.

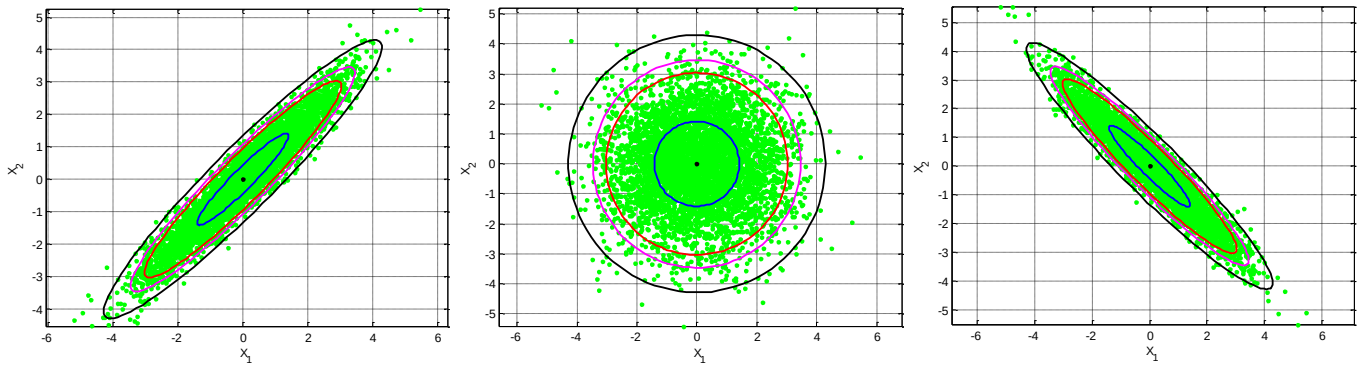


Figure 4. The position of ellipse with various values of correlation constant given the angel of inclination, specify ρ to obtain θ: (a) $\rho=0.95$, $\theta=45^\circ$; (b) $\rho=0$, $\theta=0^\circ$; (c) $\rho=-0.95$, $\theta=135^\circ$, respectively.

In the following, two scenarios cases invloving more illustrations will be visited.

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

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

$$\Sigma=\begin{bmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{21} & \sigma_2^2\end{bmatrix}=\begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}=\begin{bmatrix}4^2 & 0.5(4)(2)\\ 0.5(4)(2) & 2^2\end{bmatrix}=\begin{bmatrix}4^2 & 4\\ 4 & 2^2\end{bmatrix} \quad (326)$$

Subplot (a) in Figure 6 shows the ellipses for $\rho=0.5$ with varying variances. In the present and subsequent illustartions, 95% confidence levels are shown.

Case 2: Increasing 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 $\rho\neq 0$. Sbuplot (b) in Figure 6 provides the contours with scatter plots for $\rho=0,0.5,0.9,0.99$, respectively when $\sigma_1=\sigma_2=2$. The eccestricity of the ellipses increases with the increase of ρ.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1909.v1

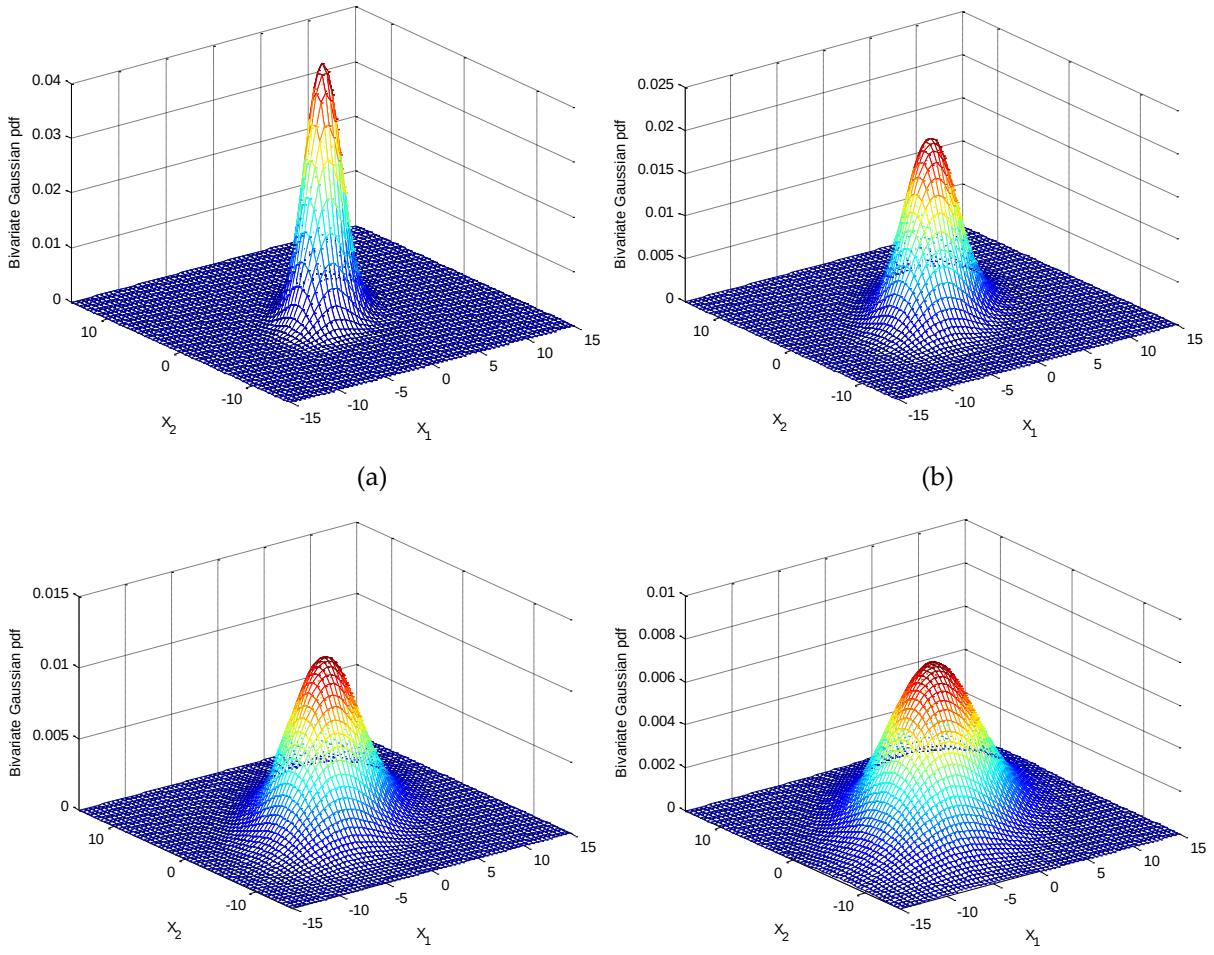


Figure 5. Equal variances $\sigma_1 = \sigma_2 = \sigma$ for a fixed $\rho = 0.5$: (a) $\sigma = 2$ (b) $\sigma = 3$ (c) $\sigma = 4$ (d) $\sigma = 5$.

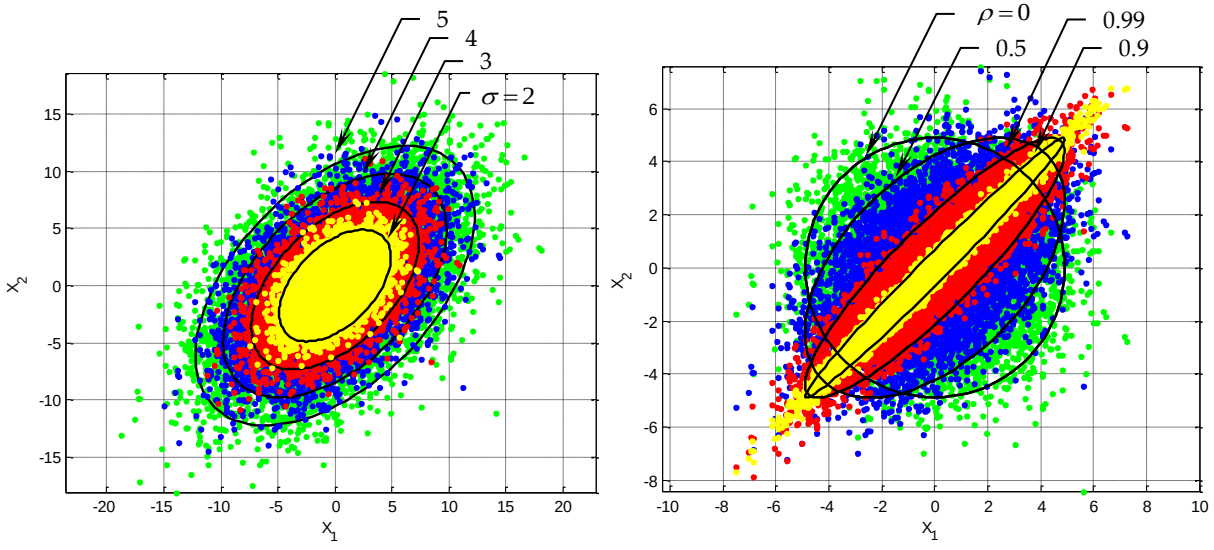


Figure 6. Ellipses for (a) $\rho = 0.5$ with varying variances $\sigma_1 = \sigma_2 = \sigma = 2 \sim 5$; (b) equal variances $\sigma_1 = \sigma_2 = 2$ with varying $\rho = 0; 0.5; 0.9; 0.99$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1909.v1

(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 variation of three dimensional surfaces and ellipses are presented in Figure 7 and Figure 8 (a) with the increase of σ_1 / σ_2, where $\sigma_1 = 2 \sim 5$ and $\sigma_2 = 2$.

Case 2: $\sigma_2 > \sigma_1$. The variation of the ellipses are presented in Figure 8 (b) with the increase of σ_2 / σ_1, where $\sigma_2 = 2 \sim 5$ and $\sigma_1 = 2$. Figure 9 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.

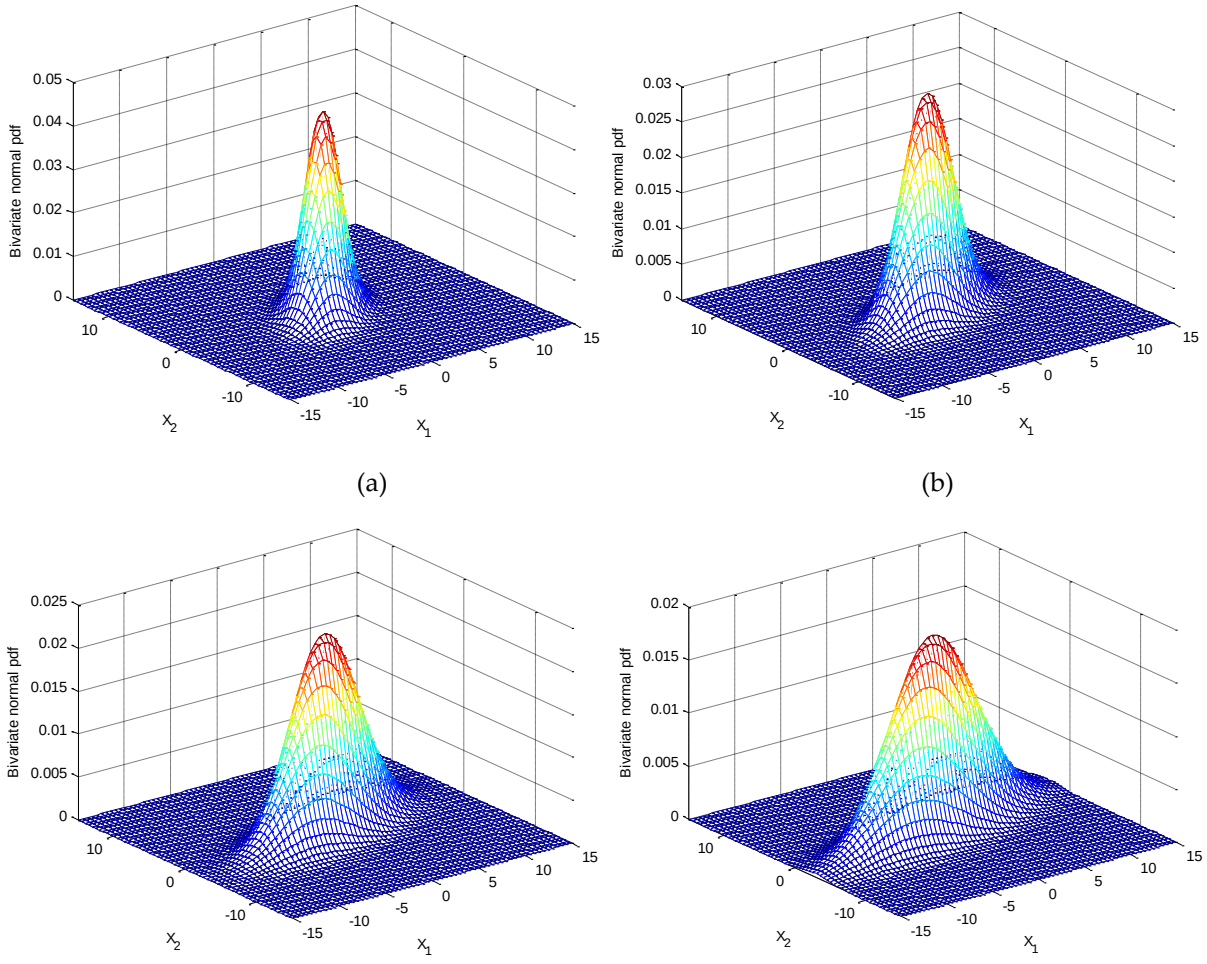
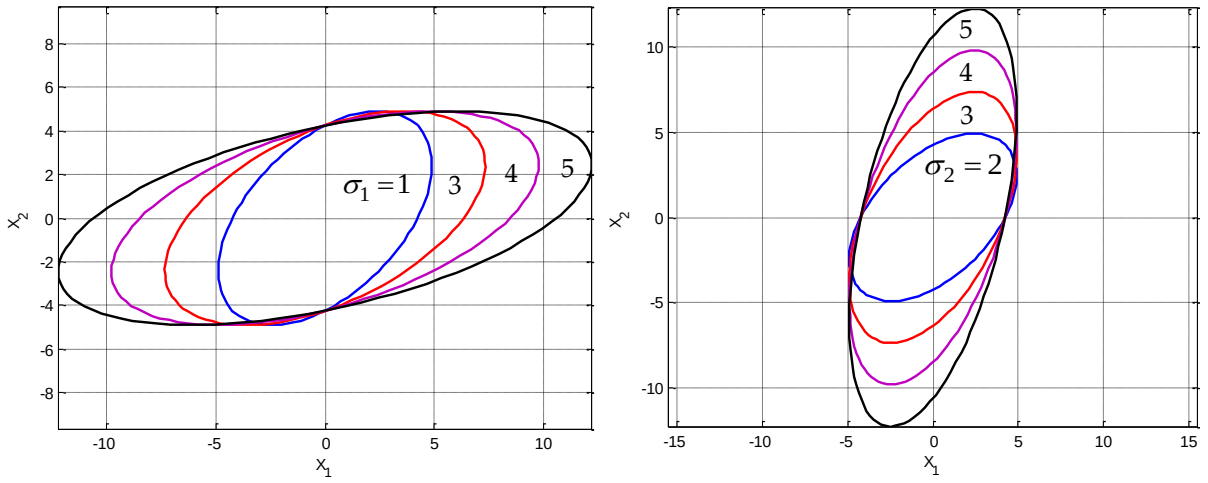


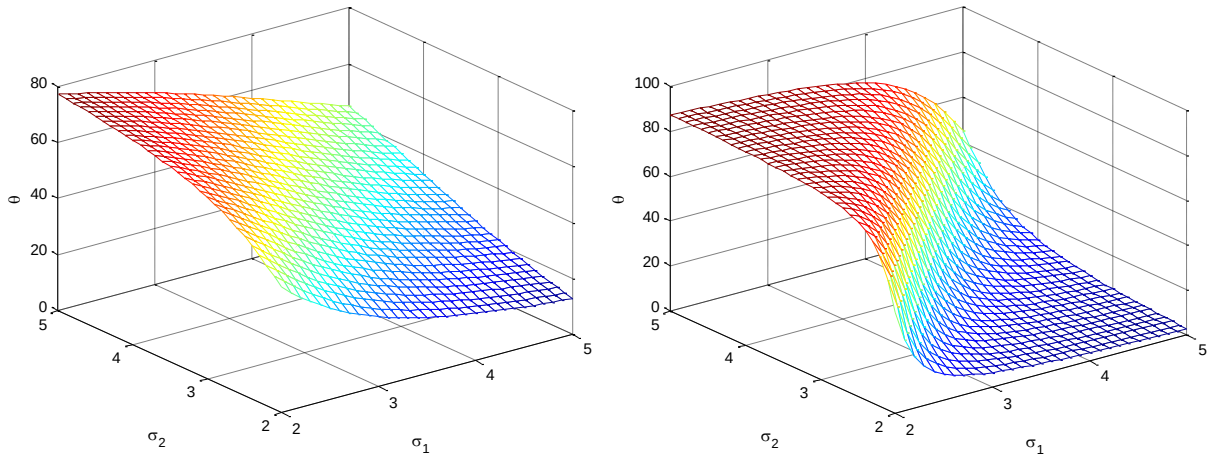
Figure 7. $\sigma_1 \neq \sigma_2$, $\sigma_1 > \sigma_2$, σ_1 / σ_2 increases $\sigma_1 = 2 \sim 5$, $\sigma_2 = 2$ for a fixed $\rho = 0.5$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1909.v1



(a) (b)

Figure 8. 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 $\sigma_2 = 2$; (b) $\sigma_2 > \sigma_1$, σ_2 / σ_1 increases where $\sigma_2 = 2 \sim 5$ and $\sigma_1 = 2$.



(a) (b)

Figure 9. Variation of inclination angle as a function of σ_1 and σ_2, for (a) $\rho = 0.5$; (b) $\rho = 0$.

(3) Variation of the ellipses for the various positive and negaive correlation. For a given variance, when ρ is specified, thus the eigenvalues and the inclination angle are obtained accordingly. Figure 10 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 left plots. Furthermore, Figure 11 provides comparsion 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) $\sigma_1 = \sigma_2 = 2$; (iv) $\sigma_1 = \sigma_2 = 4$, while fived $\rho = 0.5$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1909.v1

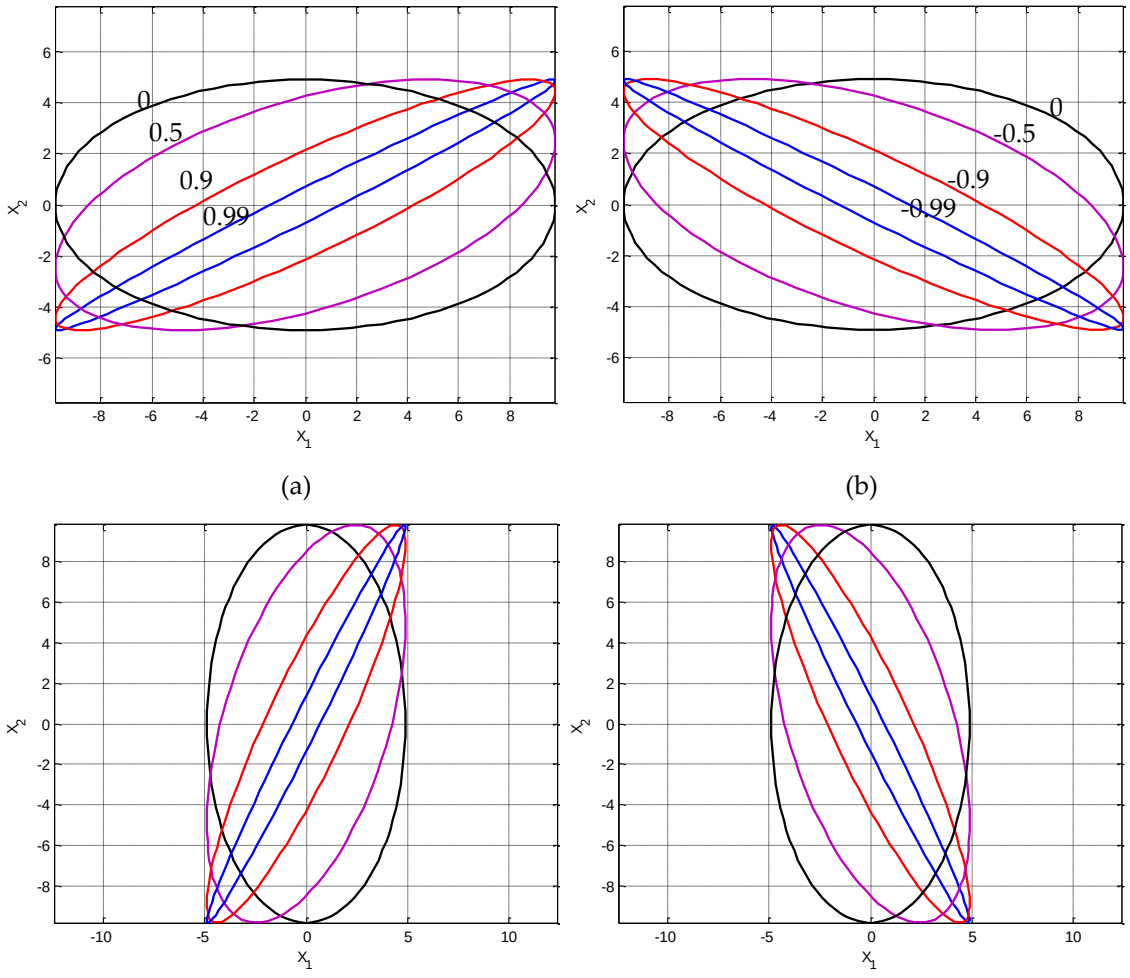


Figure 10. $\sigma_1 > \sigma_2$ ($\sigma_1 = 4$, $\sigma_2 = 2$) with (a) $\rho = 0, 0.5, 0.9, 0.99$; (b) $\rho = 0, -0.5, -0.9, -0.99$ as compared to $\sigma_2 > \sigma_1$ ($\sigma_1 = 2$, $\sigma_2 = 4$) with (c) $\rho = 0, 0.5, 0.9, 0.99$ (d) $\rho = 0, -0.5, -0.9, -0.99$.

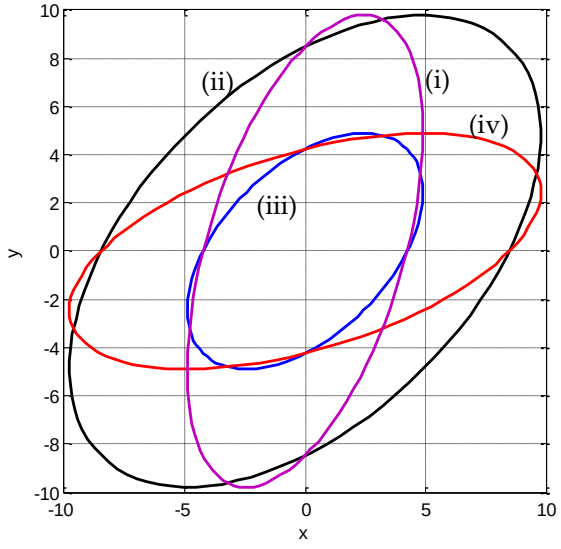


Figure 11. Comparsion of the ellipses for various (i) $\sigma_1 = 2$, $\sigma_2 = 4$; (ii) $\sigma_1 = 4$, $\sigma_2 = 2$; (iii) $\sigma_1 = \sigma_2 = 2$; (iv) $\sigma_1 = \sigma_2 = 4$, while fived $\rho = 0.5$.

 doi:10.20944/preprints202305.1909.v1

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 average surprisal 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]: [181–218]. 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 is 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

$$h(X) = -E[\log_2 f_X(x)] = -\int f_X(x) \log_2 f_X(x) dx \tag{30}$$

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

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

The differential entropy for this univariate Guassian distribution can be evaluated

$$\begin{aligned} & h(X) \\ & = -E[\log_2 f_X(x)] \\ & = -\int f_X(x) \log_2 f_X(x) dx \\ & = -\int f_X(x) \log_2 \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right) dx \\ & = \frac{1}{2} \log_2 (2\pi e \sigma^2) \end{aligned} \tag{35}$$

Figure 12 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 slower at high values of σ^2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1909.v1

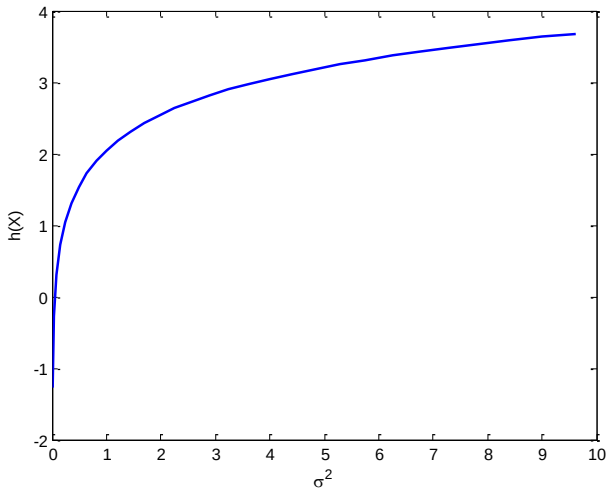


Figure 12. The differential entropy as a function σ^2 for a univariate Gaussian variable.

3.2. Entropy of a Multivariate Gaussian Distribution

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

$$h(\mathbf{X}) = -E[\log_2 f_{\mathbf{X}}(\mathbf{x})] = -\int f_{\mathbf{X}}(\mathbf{x}) \log_2 f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \tag{36}$$

and the differential entropy is given by (Appendix B)

$$h(\mathbf{X}) = \frac{1}{2} \log_2((2\pi e)^n |\Sigma|) \tag{37}$$

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

$$E[(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})] = n \tag{38}$$

For a fixed variance, the normal distribution is the pdf that maximizes entropy. Let $\mathbf{X} = [X_1 \quad X_2]^{\mathsf{T}}$ be a 2D Gaussian vector, the entropy of $\mathbf{X}$ can be calculated to be

$$h(\mathbf{X}) = h(X_1, X_2) = \frac{1}{2} \log_2 \left((2\pi e)^2 |\Sigma| \right) = \log_2 (2\pi e \sigma_1 \sigma_2 \sqrt{1-\rho^2}) \tag{39}$$

with covaraince matrix

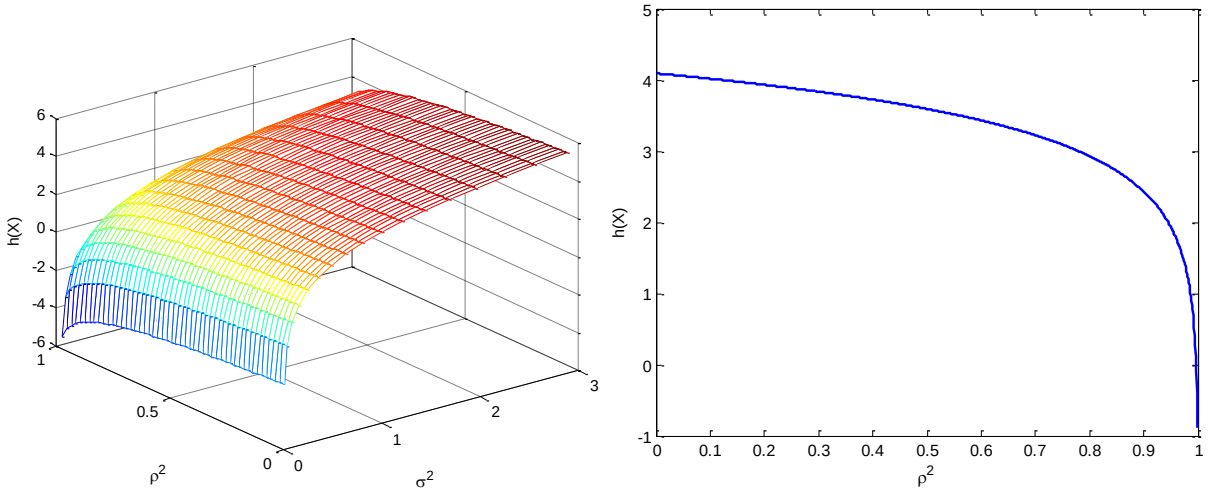
$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

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

$$h(X_1, X_2) = \log_2 (2\pi e \sigma^2 \sqrt{1-\rho^2}) \tag{40}$$

which is a function of ρ^2 concave downward, and grows first very fast and then much slower for high ρ^2 values, shown as in Figure 13.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1909.v1



(a) (b)

Figure 13. Differential entropy for the bivariate Gaussian distribution (a) as function of ρ^2 and σ^2, (b) as 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

$$h(\mathbf{Y}) = h(\mathbf{UX}) = h(\mathbf{X}) + \log_2 |\mathbf{U}| = h(\mathbf{X}) \quad (41)$$

For the case of multivariate Gaussian distribution, we have

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

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

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

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

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \prod_{i=1}^{2} \frac{1}{\sqrt{2\pi}\sqrt{\lambda_i}} e^{-\frac{1}{2}\frac{\mathbf{y}_i^2}{\lambda_i}} \\ &= \frac{1}{\sqrt{2\pi}\sqrt{\lambda_1}} e^{-\frac{1}{2}\frac{\mathbf{y}_1^2}{\lambda_1}} \cdot \frac{1}{\sqrt{2\pi}\sqrt{\lambda_2}} e^{-\frac{1}{2}\frac{\mathbf{y}_2^2}{\lambda_2}} \\ &= \frac{1}{2\pi\sqrt{\lambda_1\lambda_2}} e^{-\frac{1}{2}\left(\frac{\mathbf{y}_1^2}{\lambda_1} + \frac{\mathbf{y}_2^2}{\lambda_2}\right)} \end{aligned} \quad (42)$$

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1909.v1

$$h(\mathbf{Y}) = \frac{2}{2}\log_2(2\pi e) + \sum_{i=1}^{2}\log_2(\lambda_i) = \log_2(2\pi e) + \log_2(\lambda_1 \cdot \lambda_2) \tag{460}$$

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

$$\begin{aligned} |\Sigma| &= \lambda_1 \cdot \lambda_2 \\ &= \left(\frac{1}{2}\left[\sigma_1^2 + \sigma_2^2 + \sqrt{(\sigma_1^2 - \sigma_2^2)^2 + 4\sigma_1^2\sigma_2^2\rho^2}\right]\right)\left(\frac{1}{2}\left[\sigma_1^2 + \sigma_2^2 - \sqrt{(\sigma_1^2 - \sigma_2^2)^2 + 4\sigma_1^2\sigma_2^2\rho^2}\right]\right) \\ &= \sigma_1^2\sigma_2^2(1-\rho^2) \end{aligned} \tag{43}$$

and thus the entropy can be presented as

$$h(Y_1, Y_2) = \frac{1}{2}\log_2(2\pi e)^2|\Sigma| = \frac{1}{2}\log_2(2\pi e)^2\sigma_1^2\sigma_2^2(1-\rho^2)) = \log_2(2\pi e\sigma_1\sigma_2\sqrt{1-\rho^2}) \tag{44}$$

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 the relative entropy (Kullback-Leibler divergence will be delivered. Despite the aforementioned flaws, there is a possibility of information theory in the continuous case. A key result is that 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 how one probability distribution $f_{\mathbf{X}}$ is different from a second, reference probability distribution $g_{\mathbf{X}}$, denoted as

$$D_{KL}(f \parallel g) = \int f_{\mathbf{X}}(\mathbf{x})\log_2\frac{f_{\mathbf{X}}(\mathbf{x})}{g_{\mathbf{X}}(\mathbf{x})}d\mathbf{x} \tag{45}$$

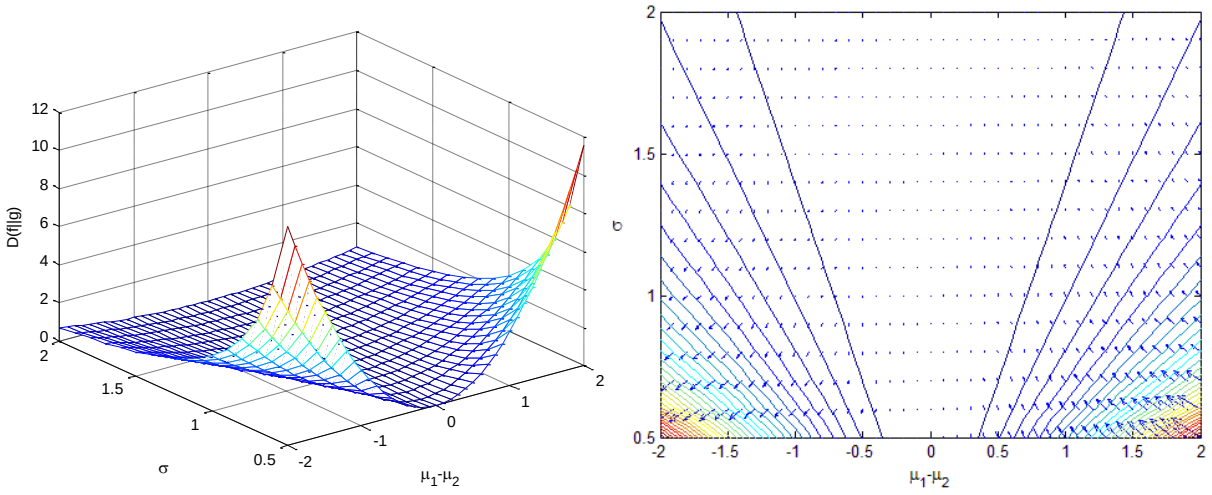
Detailed derivation is provided in Appendix E. The relative entropy between two Gaussian distributions with different means and variances are given by

$$D_{KL}(f \parallel g) = \frac{1}{2}\left[\ln\left(\frac{\sigma_2^2}{\sigma_1^2}\right) + \frac{\sigma_1^2}{\sigma_2^2} + \left(\frac{\mu_1 - \mu_2}{\sigma_2}\right)^2 - 1\right]\cdot\log_2 e \tag{46}$$

Notice that the relative entropy here is measured in bits where $\log_2$ is used in the definition. In stead, if ln is used, it would be measured in nats. The only difference in the expression is the $\log_2 e$ factor. Several conditions are discussed.

(1) If $\sigma_1 = \sigma_2 = \sigma$, $D_{KL}(f \parallel g) = \frac{1}{2}\left(\frac{\mu_1 - \mu_2}{\sigma}\right)^2\log_2 e$, which is 0 when $\mu_1 = \mu_2$.

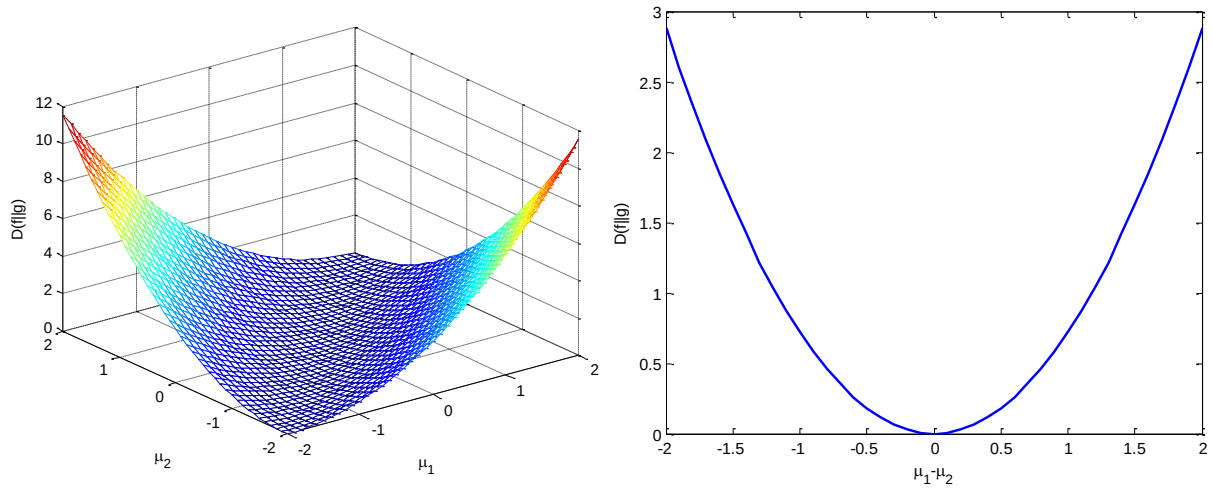
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1909.v1



(a) (b)

Figure 14. Relative entropyas a function of σ and $\mu_1-\mu_2$ when $\sigma_1=\sigma_2=\sigma$: (a) three dimensional surface; (b) contour with 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 a even function with a minimum value of 0 when $\mu_1=\mu_2$.



(a) (b)

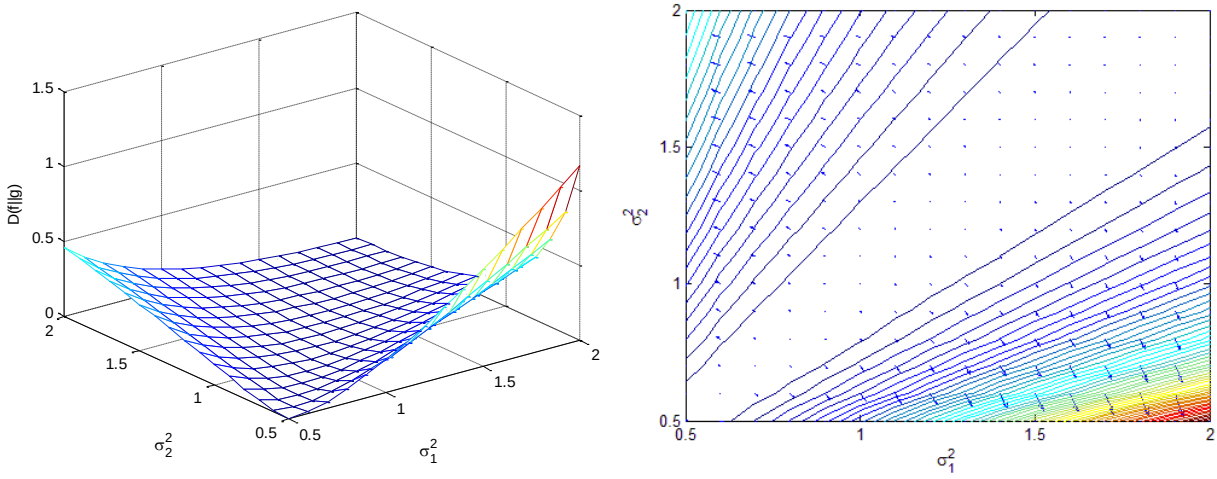
Figure 15. Variations of relative entropy when $\sigma_1=\sigma_2=1$: (a) three dimensional surface as a function of μ_1 and μ_2 (b) as a function of $\mu_1-\mu_2$.

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1909.v1

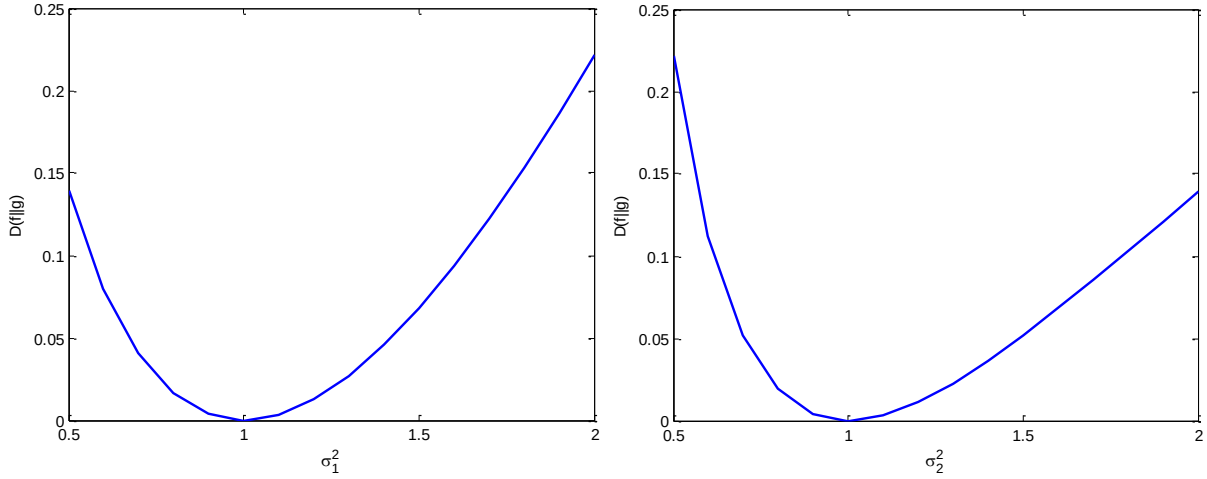


(b) (b)

Figure 17. Relative entropy as a function of σ_1 and σ_2 when $\mu_1 = \mu_2$: (a) the three dimensional surface; (b) contour with entropy gradient.

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

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



(a) (b)

Figure 18. 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 ($\mu_1 = \mu_2$).

Sensitivity analysis of the relative entropy due to change of variances and means. The gradient of $D_{KL}(f \parallel g)$ 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 where the calculation deals with partial derivatives where the cahin rule is involved. Based on the relation $\frac{d}{dx}\ln x = \frac{1}{x}$, we have

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1909.v1

$$\frac{\partial}{\partial\sigma_1}\left[\ln\left(\frac{\sigma_2^2}{\sigma_1^2}\right)\right]=\frac{\sigma_1^2}{\sigma_2^2}\cdot(-2)\sigma_2^2\sigma_1^{-3}=-\frac{2}{\sigma_1} \tag{513}$$

and the following derivatives are obtained. 514

(1) $$\frac{\partial D_{KL}}{\partial\sigma_1}=\frac{\partial}{\partial\sigma_1}\left[\ln\left(\frac{\sigma_2^2}{\sigma_1^2}\right)+\frac{\sigma_1^2}{\sigma_2^2}\right]\cdot\frac{1}{2}\log_2 e=\left(\frac{\sigma_1}{\sigma_2^2}-\frac{1}{\sigma_1}\right)\cdot\log_2 e \tag{515}$$

(2) $$\frac{\partial D_{KL}}{\partial\sigma_2}=\frac{\partial}{\partial\sigma_2}\left[\ln\left(\frac{\sigma_2^2}{\sigma_1^2}\right)+\frac{\sigma_1^2}{\sigma_2^2}+\frac{(\mu_1-\mu_2)^2}{\sigma_2^2}\right]\cdot\frac{1}{2}\log_2 e=\left[\frac{1}{\sigma_2}-\frac{\sigma_1^2}{\sigma_2^3}-\frac{(\mu_1-\mu_2)^2}{\sigma_2^3}\right]\cdot\log_2 e \tag{516}$$

(3) $$\frac{\partial D_{KL}}{\partial\mu_1}=\frac{\partial}{\partial\mu_1}\left(\frac{\mu_1-\mu_2}{\sigma_2}\right)^2\cdot\frac{1}{2}\log_2 e=\left(\frac{\mu_1-\mu_2}{\sigma_2^2}\right)\cdot\log_2 e \tag{517}$$

(4) $$\frac{\partial D_{KL}}{\partial\mu_2}=\frac{\partial}{\partial\mu_2}\left(\frac{\mu_1-\mu_2}{\sigma_2}\right)^2\cdot\frac{1}{2}\log_2 e=\left(\frac{\mu_2-\mu_1}{\sigma_2^2}\right)\cdot\log_2 e \tag{518}$$

For optimality for each of the above cases, we have 519

$$\frac{\partial D_{KL}}{\partial\sigma_1}=\frac{\sigma_1}{\sigma_2^2}-\frac{1}{\sigma_1}=0 \quad \text{when} \quad \sigma_1^2=\sigma_2^2 \tag{520}$$

$$\frac{\partial D_{KL}}{\partial\sigma_2}=\frac{1}{\sigma_2}-\frac{\sigma_1^2}{\sigma_2^3}-\frac{(\mu_1-\mu_2)^2}{\sigma_2^3}=0 \quad \text{when} \quad \sigma_2^2=\sigma_1^2+(\mu_1-\mu_2)^2 \tag{521}$$

$$\left(\frac{\mu_1-\mu_2}{\sigma_2^2}\right)\cdot\log_2 e=0 \quad \text{when} \quad \mu_1=\mu_2 \tag{522}$$

$$\left(\frac{\mu_2-\mu_1}{\sigma_2^2}\right)\cdot\log_2 e=0 \quad \text{when} \quad \mu_1=\mu_2 \tag{523}$$

5. Mutual Information 524

Mutual information is one of many quantities that measures how much one 525
random variables tells us about another. It is a dimensionless quantity with (generally) 526
units of bits, and can be thought of as the reduction in uncertainty about one random 527
variable given knowledge of another. The mutual information $I(X;Y)$ between two 528
variables with joint pdf $f_{XY}(x,y)$ is given by 529

$$I(X;Y)=E\left[\log\frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}\right]=\int\int f_{XY}(x,y)\log\frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}dxdy \tag{47}$$ 530

The mutual information between the random variables X and Y has the following 531
relation 532

$$I(X;Y)=I(Y;X) \tag{48}$$ 533

where 534

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1909.v1

$$I(X;Y) = h(X) - h(X \mid Y) \geq 0 \tag{49}$$

and

$$I(Y;X) = h(Y) - h(Y \mid X) \geq 0 \tag{50}$$

implying that $h(X) \geq h(X \mid Y)$ and $h(Y) \geq h(Y \mid X)$. The mutual information of a random variable with itself is the self information, which is the 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 case, $h(X) = h(X \mid Y)$ and $h(Y) = h(Y \mid X)$.

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

$$\begin{aligned} I(X;Y) &= h(X) + h(Y) - h(X,Y) \\ &= \frac{1}{2}\log_2(2\pi e)\sigma_x^2 + \frac{1}{2}\log_2(2\pi e)\sigma_y^2 - \frac{1}{2}\log_2(2\pi e)^2\sigma_x^2\sigma_y^2(1-\rho^2) \\ &= -\frac{1}{2}\log_2(1-\rho^2) \end{aligned} \tag{51}$$

Figure 19 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 mutually information is infinite. It can be ceen that $I(X;Y) = 0$ for $\rho = 0$ and that $I(X;Y) \to \infty$ for $\rho \to \pm 1$.

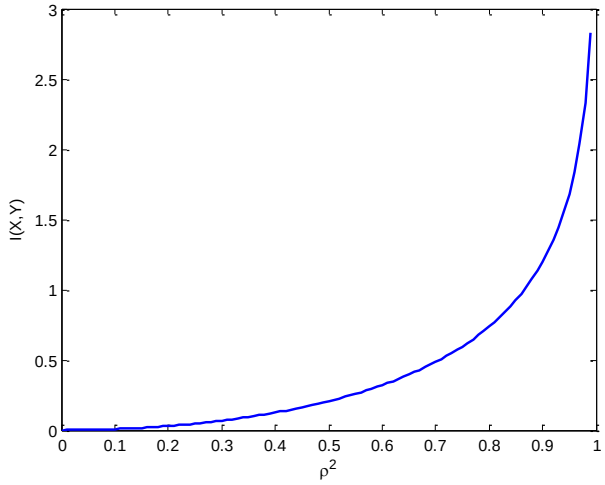


Figure 19. Mutual information versus ρ^2 between the correlated Gaussian variables.

On the other hand, consider the additive white Gaussian noise (AWGN) channel shown as in Figure 20, the mutual information is given by

$$I(X;Y) = h(Y) - h(Y \mid X) = \frac{1}{2}\log_2\left(\frac{2\pi e(\sigma_x^2 + \sigma_n^2)}{2\pi e\sigma_n^2}\right) = \frac{1}{2}\log_2\left(1 + \frac{\sigma_x^2}{\sigma_n^2}\right) \tag{52}$$

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1909.v1

$$h(\mathrm{Y})=\frac{1}{2}\log_2\left(2\pi e(\sigma_x^2+\sigma_n^2)\right);\quad h(Y\mid X)=h(N)=\frac{1}{2}\log_2(2\pi e)\sigma_n^2$$

Mutual information for the additive white Gaussian noise (AWGN) channel is shown in Figure 21, including the three-dimensional surface as a function of σ_x^2 and σ_n^2;, and also in terms of the the signal-to-noise-ratio $\mathrm{SNR}=\sigma_x^2/\sigma_n^2$. It can be seen that The mutual information grows first very fast and then much slower for high values of the signal-to-noise ratio.

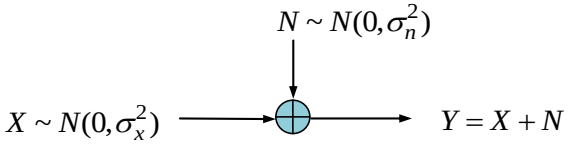


Figure 20. Schematic illustration of the additive white Gaussian noise (AWGN) channel.

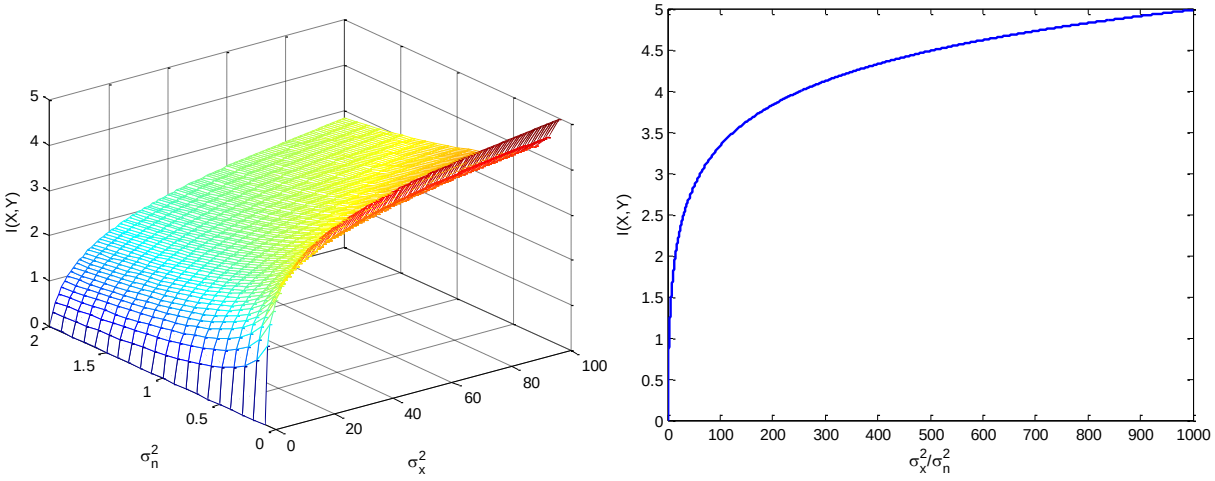


Figure 21. Mutual information for the additive white Gaussian noise (AWGN) channel: (a) the three-dimensional surface as a function of σ_x^2 and σ_n^2 ; (b) in terms of the the signal-to-noise-ratio.

6. Conclusions

This paper intends to serve to the readers as a supplement note on the geometric interpretation of the multivariate Gaussian distribution and its entropy, relative entropy, and mutual information. The illustrative examples are employed to provide further insights into the geometric interpretation of the multivariate Gaussian distribution and its entropy and mutual information, enabling the readers to correctly interpret the theory for future design. The fundamental objective is to study the application of multivariate sets of data in Gaussian distribution. This paper examines broad measurements of structure for Gaussian distributions, which shows that they can be described in terms of the information-theoretic between the given covariance matrix and correlated random variables (in terms of relative entropy). To develop the multivariate Gaussian distribution with the 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 sci-

 doi:10.20944/preprints202305.1909.v1

ence 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, including basic to application concerns.

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.

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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1909.v1

Appendix B. Derivation of the differential entropy for the multivariate Gaussian distribution

$$
\begin{aligned}
& h(\mathbf{X}) \\
&= -E\left[\log_2 f_{\mathbf{X}}(\mathbf{x})\right] \\
&= -E\left[\log_2\left(\frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}\right)\right] \\
&= -E\left[-\frac{n}{2}\log_2(2\pi) - \frac{1}{2}\log_2|\Sigma| - \log_2 e^{\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}\right] \\
&= -\int f_{\mathbf{X}}(\mathbf{x})\left[\left(-\frac{1}{2}\log_2(2\pi)^n|\Sigma|) - \frac{\log_2 e}{2}((\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)\right] d\mathbf{x} \\
&= \frac{1}{2}\log_2((2\pi)^n|\Sigma|) + \frac{n}{2}\log_2 e \\
&= \frac{n}{2}\log_2(2\pi) + \frac{1}{2}\log_2|\Sigma| + \log_2 e^{\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})} \\
&= \frac{n}{2}\log_2(2\pi) + \frac{1}{2}\log_2|\Sigma| + \log_2 e^{\frac{n}{2}} \\
&= \frac{n}{2}\log_2(2\pi) + \frac{1}{2}\log_2|\Sigma| + \frac{n}{2}\log_2 e \\
&= \frac{1}{2}\log_2\left((2\pi e)^n|\Sigma|\right)
\end{aligned}
\tag{612}
$$

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

Appendix C. Evaluation of expectation of the Mahalanobis distance $E[(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})] = n$

$$
\begin{aligned}
& E[(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})] \\
&= E[tr((\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}))] \\
&= E[tr(\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})^T (\mathbf{x}-\boldsymbol{\mu}))] \\
&= tr(\Sigma^{-1} E[(\mathbf{x}-\boldsymbol{\mu})^T (\mathbf{x}-\boldsymbol{\mu})]) \\
&= tr(\Sigma^{-1}\Sigma) \\
&= tr(I_n) \\
&= n
\end{aligned}
\tag{617}
$$

A special case for $n=1$

$$
\begin{aligned}
& E[(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})] \\
&= E\left[\frac{(x-\mu)^2}{\sigma^2}\right] \\
&= \int f_X(x)\left(\frac{(x-\mu)^2}{\sigma^2}\right) dx \\
&= \frac{1}{\sigma^2}\int (x-\mu)^2 f_X(x) dx \\
&= 1
\end{aligned}
\tag{619}
$$

Appendix D. Derivation of the differential entropy in the transformed frame

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

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

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

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

References

1. Verdú, S. (**1990**). On channel capacity per unit cost, *IEEE Trans. Inf. Theory, vol. 36*, no. 5, pp. 1019–1030.
2. Lapidoth, A. and Shamai (Shitz), S. (**2002**). Fading channels: How perfect need perfect side information be? *IEEE Trans. Inf. Theory, vol. 48, no. 5*, pp. 1118–1134.
3. Verdú, S. (**2002**). Spectral efficiency in the wideband regime, *IEEE Trans. Inf. Theory, vol. 48*, no. 6, pp. 1319–1343.
4. Prelov, V. and Verdú, S. (**2004**). Second-order asymptotics of mutual information, *IEEE Trans. Inf. Theory, vol. 50*, no. 8, pp. 1567–1580.
5. Kailath, T. (**1968**). A note on least squares estimates from likelihood ratios, *Inf. Contr., vol. 13*, pp. 534–540.
6. Kailath, T. (**1969**). A general likelihood-ratio formula for random signals in Gaussian noise, *IEEE Trans. Inf. Theory, vol. IT-15, no. 2*, pp. 350–361.
7. Kailath, T. (**1970**). A further note on a general likelihood formula for random signals in Gaussian noise, *IEEE Trans. Inf. Theory, vol. IT-16, no. 4*, pp. 393–396.
8. Jaffer A. G. and Gupta S. C. (**1972**). On relations between detection and estimation of discrete time processes, *Inf. Contr., vol. 20*, pp. 46–54.
9. Jwo, D. J., Biswal, A. (**2023**). Implementation and Performance Analysis of Kalman Filters with Consistency Validation. *Mathematics, 11*, 521.
10. Duncan, T. E. (**1970**). On the calculation of mutual information, *SIAM J. Applied Mathematics, vol. 19*, pp. 215–220.

11. Kadota, T. T., Zakai, M. and Ziv, J. (**1971**). Mutual information of the white Gaussian channel with and without feedback, *IEEE Trans. Inf. Theory*, *vol. IT-17, no. 4*, pp. 368–371.
12. Amari, S. I. (**2016**). Information geometry and its applications, *Vol. 194. Springer*.
13. Schneidman, E., Still, S., Berry, M. J. and Bialek, W. (**2003**). Network information and connected correlations. *Physical review letters*, *91(23)*, p.238701.
14. Timme, N., Alford, W., Flecker, B. and Beggs, J. M. (**2014**). Synergy, redundancy, and multivariate information measures: an experimentalist's perspective. *Journal of computational neuroscience, 36*, pp.119-140.
15. Liang, K. C. and Wang, X. (**2008**). Gene regulatory network reconstruction using conditional mutual information. *EURASIP Journal on Bioinformatics and Systems Biology*, *2008*, pp.1-14.
16. Panzeri, S., Magri, C. and Logothetis, N. K. (**2008**). On the use of information theory for the analysis of the relationship between neural and imaging signals. *Magnetic resonance imaging, 26(7),* pp.1015-1025.
17. Katz, Y., Tunstrøm, K., Ioannou, C. C., Huepe, C. and Couzin, I. D. (**2011**). Inferring the structure and dynamics of interactions in schooling fish. *Proceedings of the National Academy of Sciences, 108(46)*, pp.18720-18725.
18. Cutsuridis, V., Hussain, A. and Taylor, J. G. eds. (**2011**). Perception-action cycle: Models, architectures, and hardware. *Springer Science & Business Media.*
19. Ay, N., Bernigau, H., Der, R. and Prokopenko, M. (**2012**). Information-driven self-organization: the dynamical system approach to autonomous robot behavior. *Theory in Biosciences, 131*, pp.161-179.
20. Rosas, F., Ntranos, V., Ellison, C. J., Pollin, S. and Verhelst, M. (**2016**). Understanding interdependency through complex information sharing. *Entropy, 18(2),* p.38.
21. Ince, R. A. (**2017**). The Partial Entropy Decomposition: Decomposing multivariate entropy and mutual information via pointwise common surprisal. *arXiv preprint arXiv:1702*.01591.
22. Perrone, P. and Ay, N. (**2016**). Hierarchical quantification of synergy in channels. *Frontiers in Robotics and AI, 2*, p.35.
23. Bertschinger, N., Rauh, J., Olbrich, E., Jost, J. and Ay, N. (**2014**). Quantifying unique information. *Entropy, 16(4),* pp.2161-2183.
24. Harder, M., Salge, C. and Polani, D. (**2013**). Bivariate measure of redundant information. *Physical Review E, 87(1)*, p.012130.
25. Rauh, J., Banerjee, P. K., Olbrich, E., Jost, J. and Bertschinger, N. (**2017**). On extractable shared information. *Entropy, 19(7),* p.328.
26. Ince, R. A. (**2017**). Measuring multivariate redundant information with pointwise common change in surprisal. *Entropy, 19(7),* p.318.
27. Chicharro, D. and Panzeri, S. (2017). Synergy and redundancy in dual decompositions of mutual information gain and information loss. *Entropy, 19(2)*, p.71.