III.2 Gaussiana Bidimensional

May 13, 2019

1 Distribución Gaussiana Bidimensional

Consideremos dos variables aleatorias, X e Y, agrupadas en un vector aleatorio $\mathbf{V} = (X, Y)^{\mathrm{T}}$. Su **vector de medias** es

$$\boldsymbol{\mu}_{\boldsymbol{V}} \equiv E(\mathbf{V}) = \left(\begin{array}{c} E(X) \\ E(Y) \end{array} \right) \equiv \left(\begin{array}{c} \mu_X \\ \mu_Y \end{array} \right)$$

Y su **matriz de covarianza** es

$$\boldsymbol{\Sigma}_{\boldsymbol{V}} \equiv \mathbf{C_{XY}} = E\left((\mathbf{V} - \boldsymbol{\mu}_{\boldsymbol{V}})(\mathbf{V} - \boldsymbol{\mu}_{\boldsymbol{V}})^{\mathrm{T}} \right) = \left[\begin{array}{cc} \sigma_X^2 & \rho_{XY}\sigma_X\sigma_Y \\ \rho_{XY}\sigma_X\sigma_Y & \sigma_Y^2 \end{array} \right]$$

donde:

$$\mathbf{C_{XY}} = \begin{pmatrix} E((X-\mu_X)^2) & E((X-\mu_X)(Y-\mu_Y)) \\ E((X-\mu_X)(Y-\mu_Y)) & E((Y-\mu_Y)^2) \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & C_{XY} \\ C_{XY} & \sigma_Y^2 \end{pmatrix}$$

1.1 Función de densidad conjuntamente Gaussiana

El vector aleatorio $\mathbf{V} = (X, Y)^{\mathrm{T}}$ es Normal o Gaussiano o, alternativamente, las variables aleatorias X e Y son conjuntamente Gaussianas si la función de densidad de probabilidad conjunta es ($N = 2$):

$$f_{\mathbf{V}}(\mathbf{v}) \equiv f_{XY}(x, y) = \frac{1}{\sqrt{(2\pi)^N |\boldsymbol{\Sigma}_{\boldsymbol{V}}|}} \exp\left(-\frac{1}{2} (\mathbf{v} - \boldsymbol{\mu}_{\boldsymbol{V}})^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{V}}{}^{-1} (\mathbf{v} - \boldsymbol{\mu}_{\boldsymbol{V}}) \right)$$

Esta expresión es general para vectores aleatorios con N componentes aleatorias. En el caso bidimensional:

$$|\boldsymbol{\Sigma}_{\boldsymbol{V}}| = \sigma_X^2 \sigma_Y^2 \left(1 - \rho_{XY}^2\right)$$

$$\boldsymbol{\Sigma}_{\boldsymbol{V}}{}^{-1} = \frac{1}{\sigma_X^2 \sigma_Y^2 \left(1 - \rho_{XY}^2\right)} \left[\begin{array}{cc} \sigma_Y^2 & -\rho_{XY}\sigma_X\sigma_Y \\ -\rho_{XY}\sigma_X\sigma_Y & \sigma_X^2 \end{array} \right]$$

Sustituyendo y desarrollando, la **función de densidad de probabilidad conjunta** es:

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{XY}^2}} \exp\left(-\frac{1}{2\left(1-\rho_{XY}^2\right)} \left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho_{XY}(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} \right] \right)$$

Las **funciones de densidad de probabilidad marginales** son:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}}$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_Y} e^{-\frac{(y-\mu_X)^2}{2\sigma_Y^2}}$$

Adviértase que la caracterización marginal de X e Y no permite, en general, alcanzar conclusiones sobre la caracterización conjunta. En particular, si X e Y son marginalmente normales, ello ni siquiera supone que (X, Y) sean conjuntamente normales.

Sin embargo, si las variables aleatorias X e Y son **independientes** puede obtenerse la función de densidad conjunta a partir de las marginales:

$$f_{XY}(x, y) = f_X(x) f_Y(y) = \frac{1}{2\pi\sigma_X\sigma_Y} \exp\left(-\frac{1}{2}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right)$$

Adviértase que si las variables aleatorias X e Y están incorreladas, $\rho_{XY} = 0$ y la función de densidad conjunta es igual al caso en que las variables aleatorias sean independientes. Por tanto, **si las variables aleatorias son conjuntamente normales la independencia no solo implica incorrelación, sino que ambas propiedades son equivalentes**.

Las *isolíneas* de la función de densidad conjunta corresponden a los puntos del plano (X, Y) en los que dicha función es constante. En general son elipses y definen la denominada **distancia de Mahalanobis**, d_M del punto $\mathbf{v} = (x, y)^{\mathrm{T}}$ al valor medio $\boldsymbol{\mu_V} = (\mu_X, \mu_Y)^{\mathrm{T}}$:

$$d_M^2(\mathbf{v}; \boldsymbol{\mu_V}) = (\mathbf{v} - \boldsymbol{\mu_V})^{\mathrm{T}} \boldsymbol{\Sigma_V}^{-1} (\mathbf{v} - \boldsymbol{\mu_V})$$

Esta expresión es general para N variables aleatorias. Desarrollando para el caso de dos:

$$d_M^2\left((x, y); (\mu_X, \mu_Y)\right) = \frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho\,(x-\mu_X)\,(y-\mu_Y)}{\sigma_X\sigma_Y}$$

1.1.1 Vector aleatorio Normal estándar

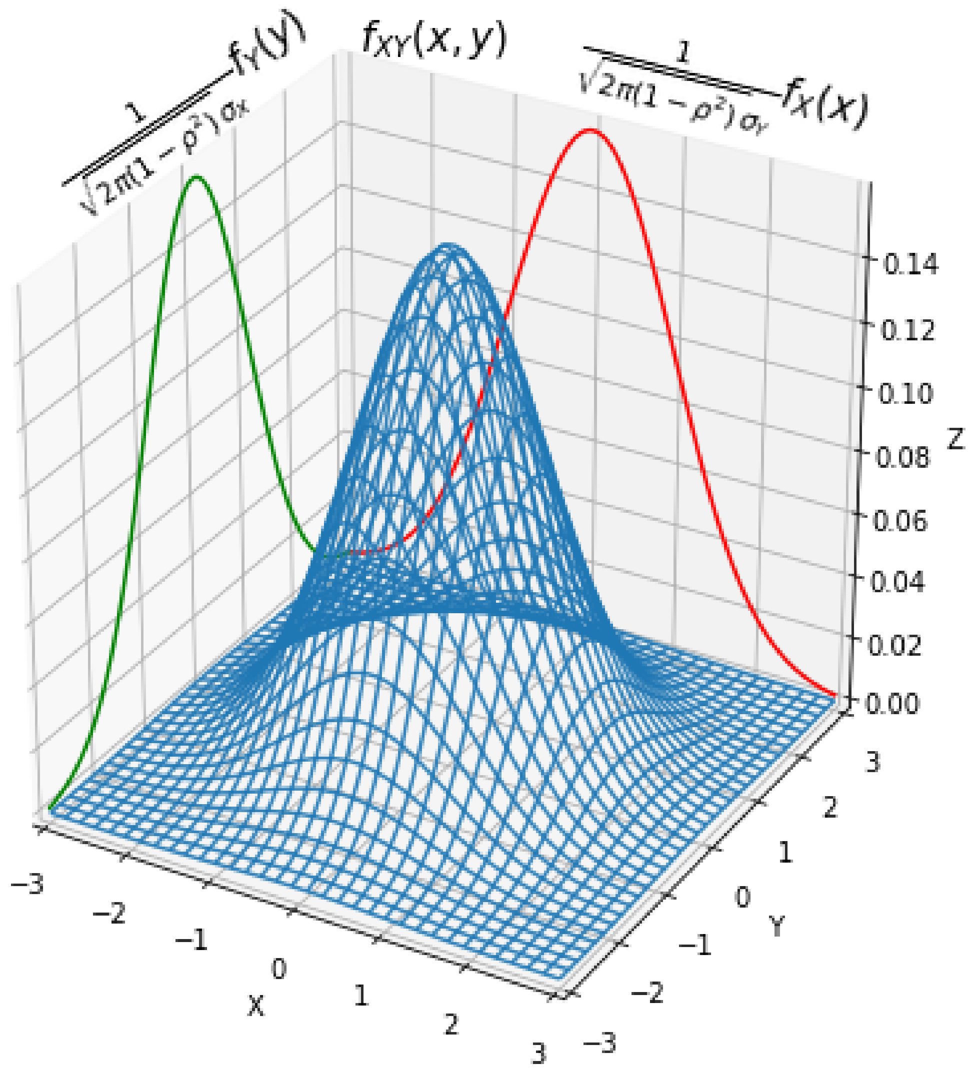
Es el caso en el que:

- las variables aleatorias son independientes y, por tanto, incorreladas,
- las medias son nulas y
- las desviaciones típicas son la unidad.

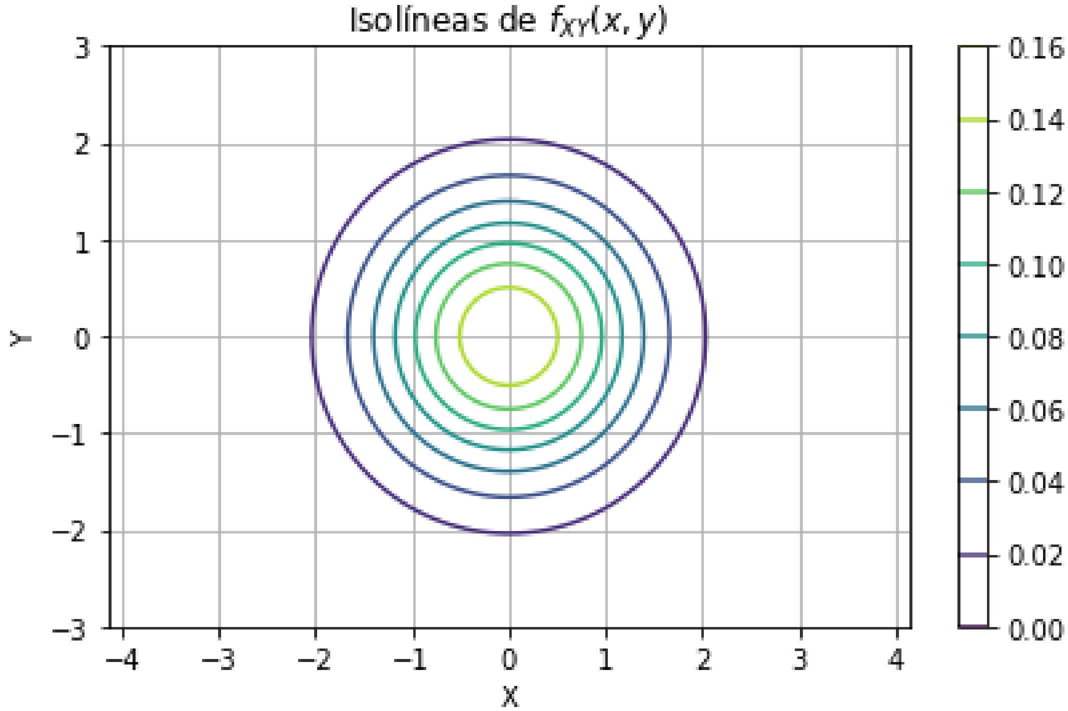
Ello supone que **la matriz de covarianza es la matriz identidad**:

$$f_{XY}(x, y) = \frac{1}{2\pi} e^{-\frac{1}{2}\left(x^2+y^2\right)}$$

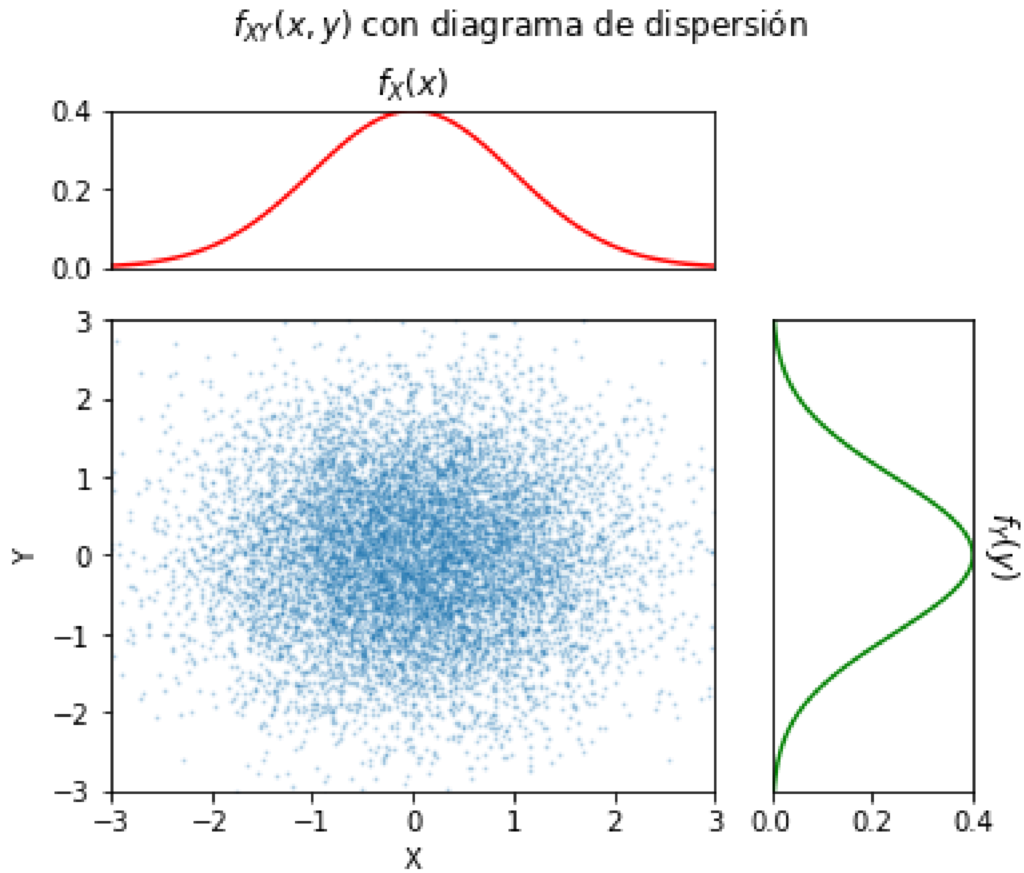
En el gráfico que sigue, se muestran las funciones de densidad marginales escaladas para que las modas coincidan con la moda de la densidad conjunta.



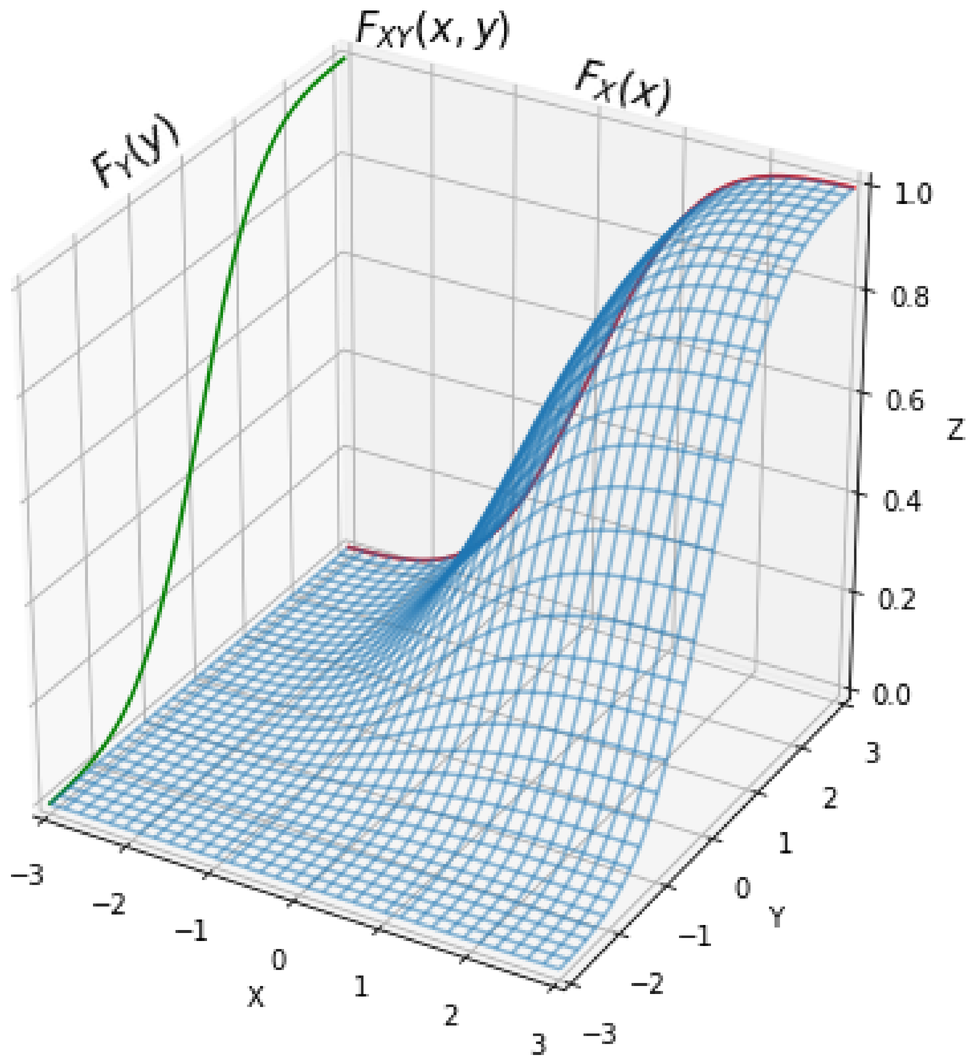
En este casi las isolíneas son circunferencias, centradas en la media, y la distancia de Mahalanobis corresponde a la distancia Euclídea habitual:



Puede hacerse una representación intuitiva muestreando la función de densidad conjunta y mostrando el diagrama de dispersión resultante, junto con las densidades marginales:



La **función de distribución de probabilidad** es:



1.1.2 Independencia con desviaciones típicas diferentes

Es este caso:

- las variables aleatorias son independientes y, por tanto, incorreladas,
- las medias son nulas y
- las desviaciones típicas son diferentes, σ_X y σ_Y.

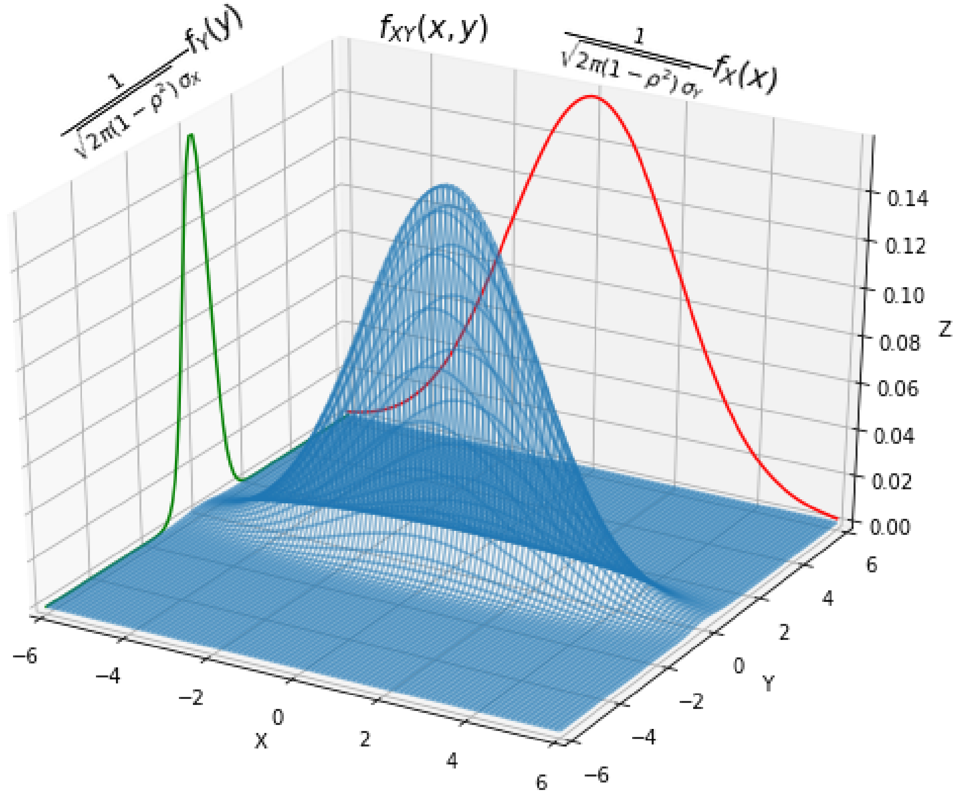
Ello supone que **la matriz de covarianza es diagonal** con matriz inversa:

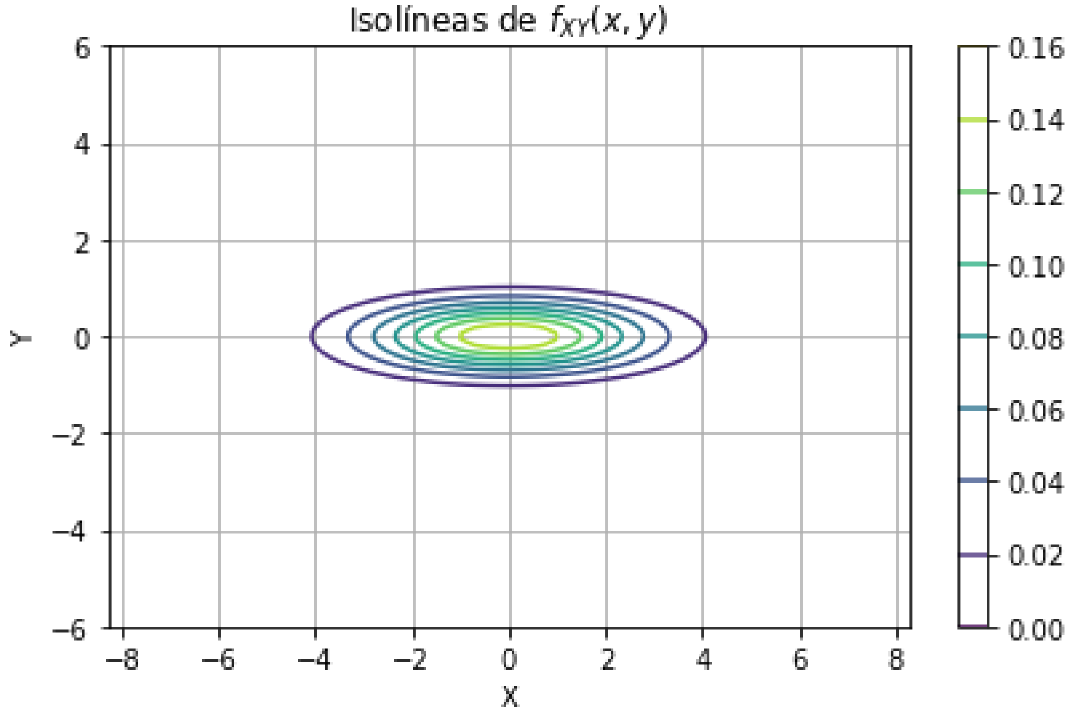
$$\mathbf{\Sigma_V} \equiv \mathbf{C_{XY}} = \begin{pmatrix} \sigma_X^2 & 0 \\ 0 & \sigma_Y^2 \end{pmatrix} \qquad \mathbf{\Sigma_V}^{-1} = \begin{bmatrix} \frac{1}{\sigma_X^2} & 0 \\ 0 & \frac{1}{\sigma_Y^2} \end{bmatrix}$$

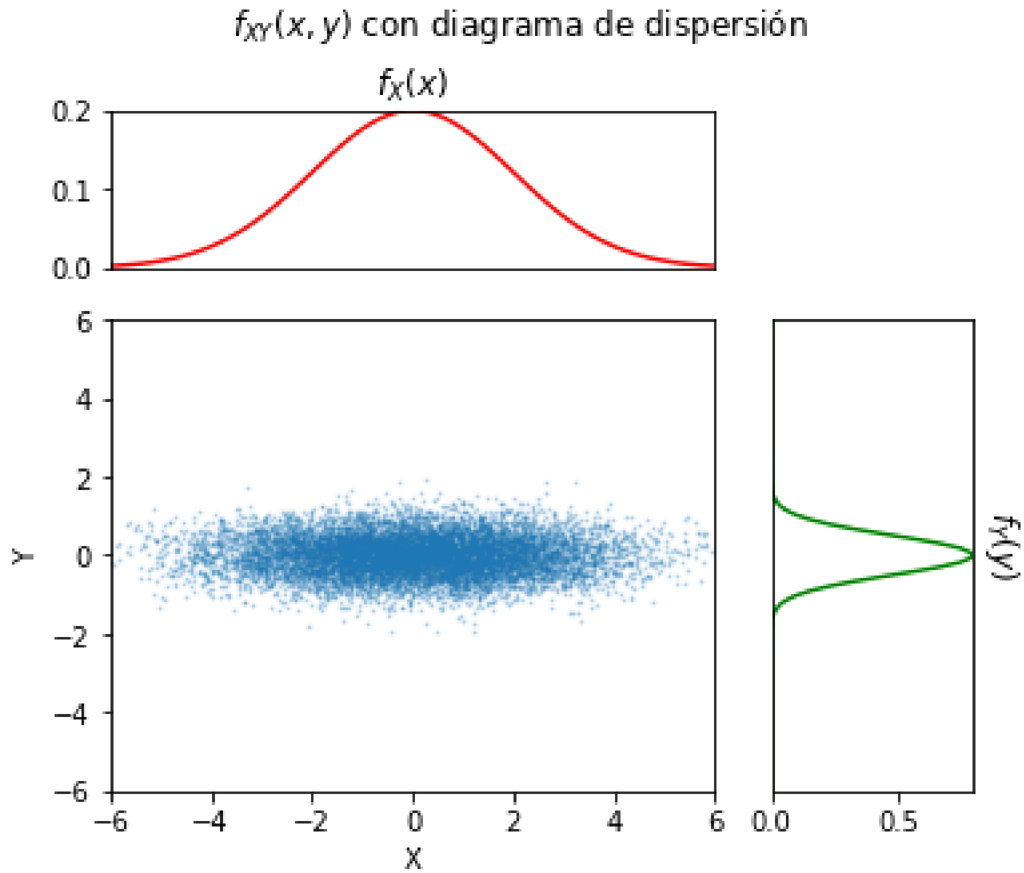
La función de densidad conjunta resulta:

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y} e^{-\frac{1}{2}\left[(\frac{x}{\sigma_X})^2 + (\frac{y}{\sigma_Y})^2\right]}$$

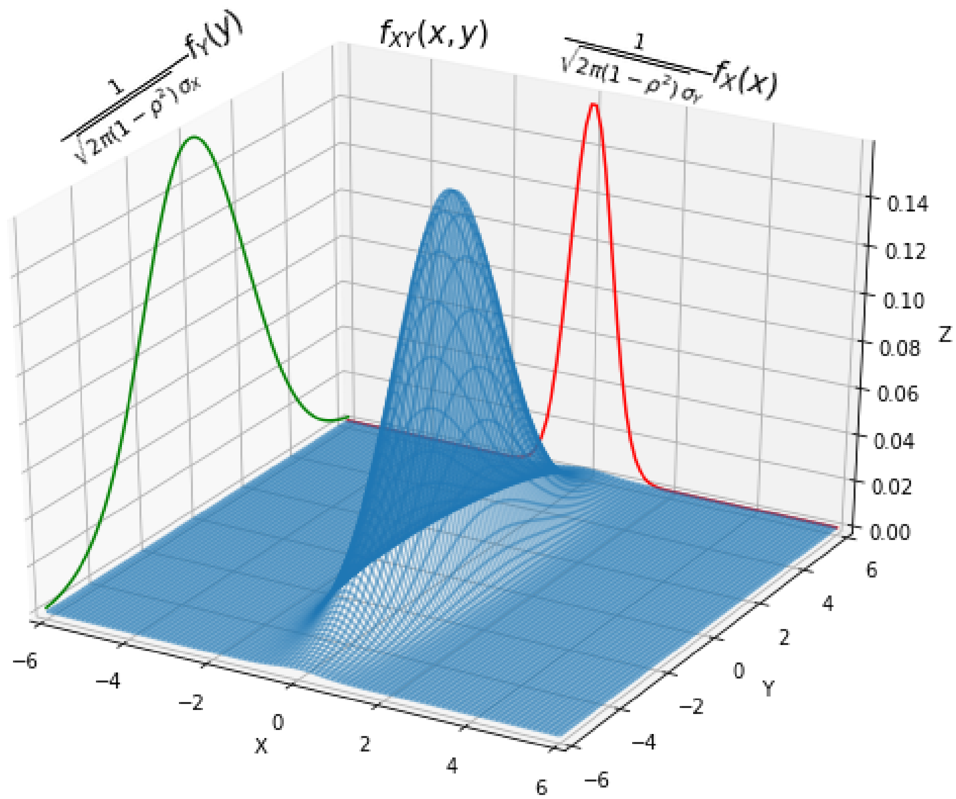
$\sigma_X = 2, \sigma_Y = 1/2$

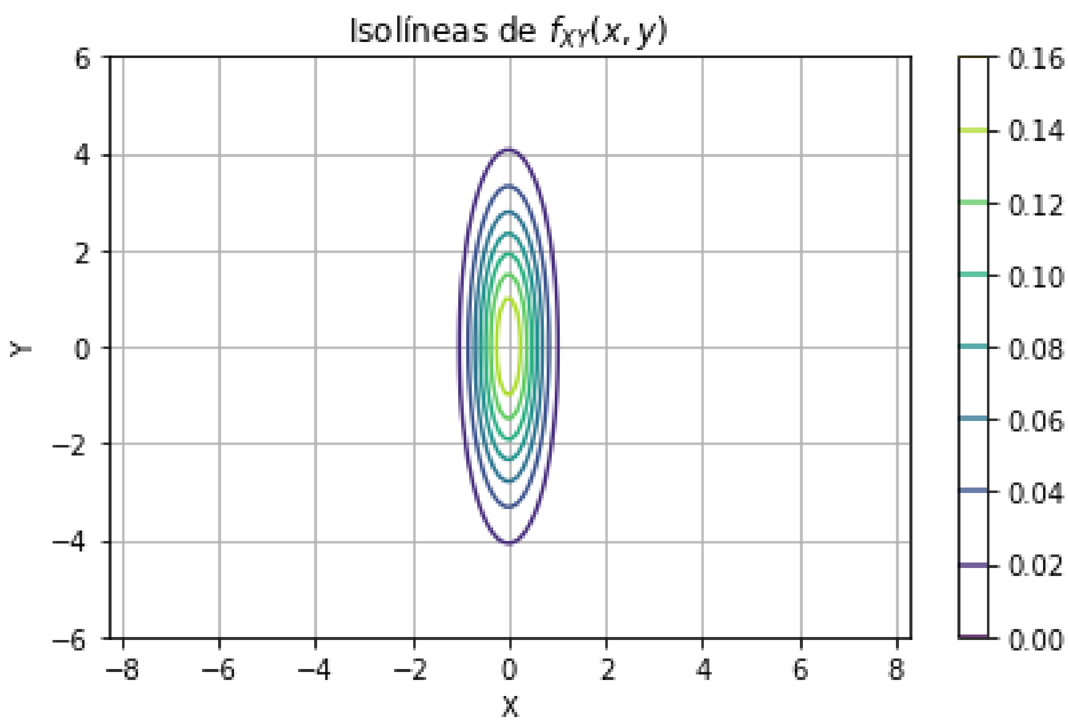


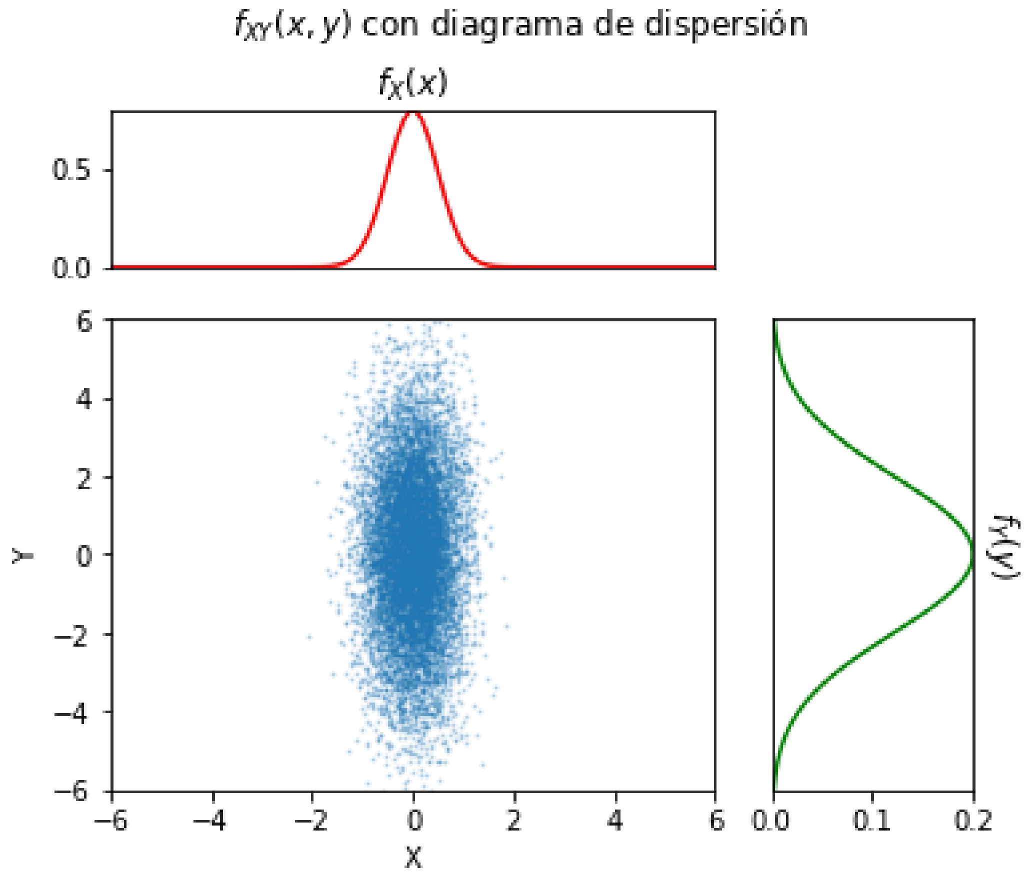

Isolíneas de $f_{XY}(x, y)$
Y
X
6
4
2
0
−2
−4
−6
−8
8
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00



$\sigma_X = 1/2$, $\sigma_Y = 2$

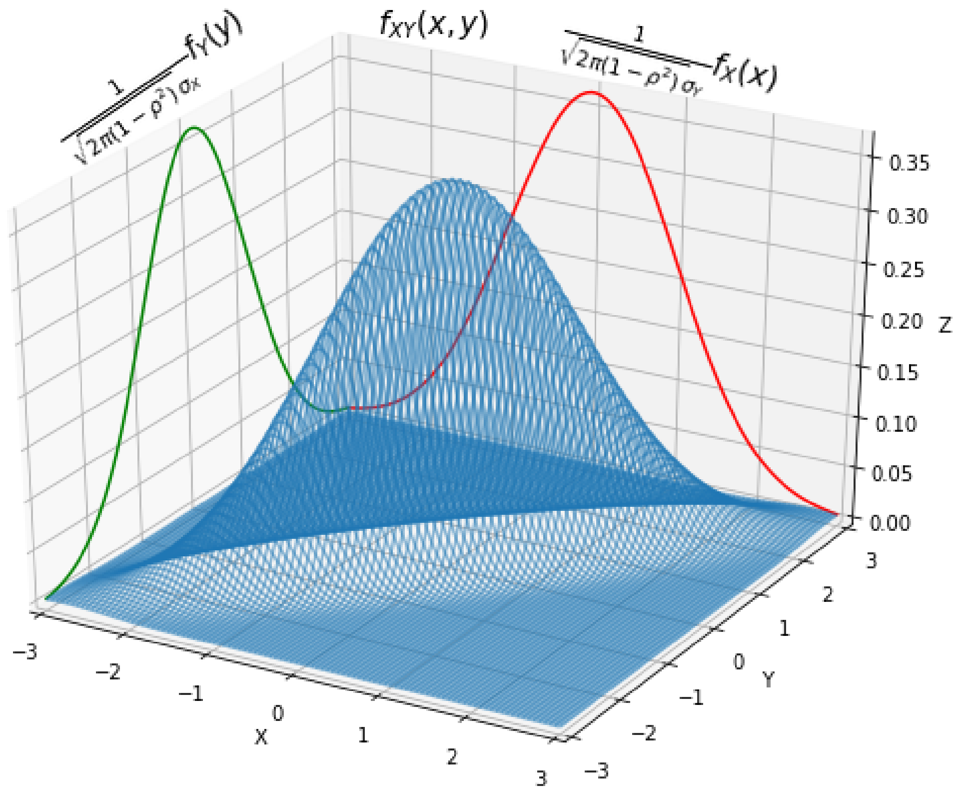

$f_{XY}(x, y)$
$\frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma_X} f_Y(y)$
$\frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma_Y} f_X(x)$
X
Y
Z


Isolíneas de $f_{XY}(x, y)$
X
Y

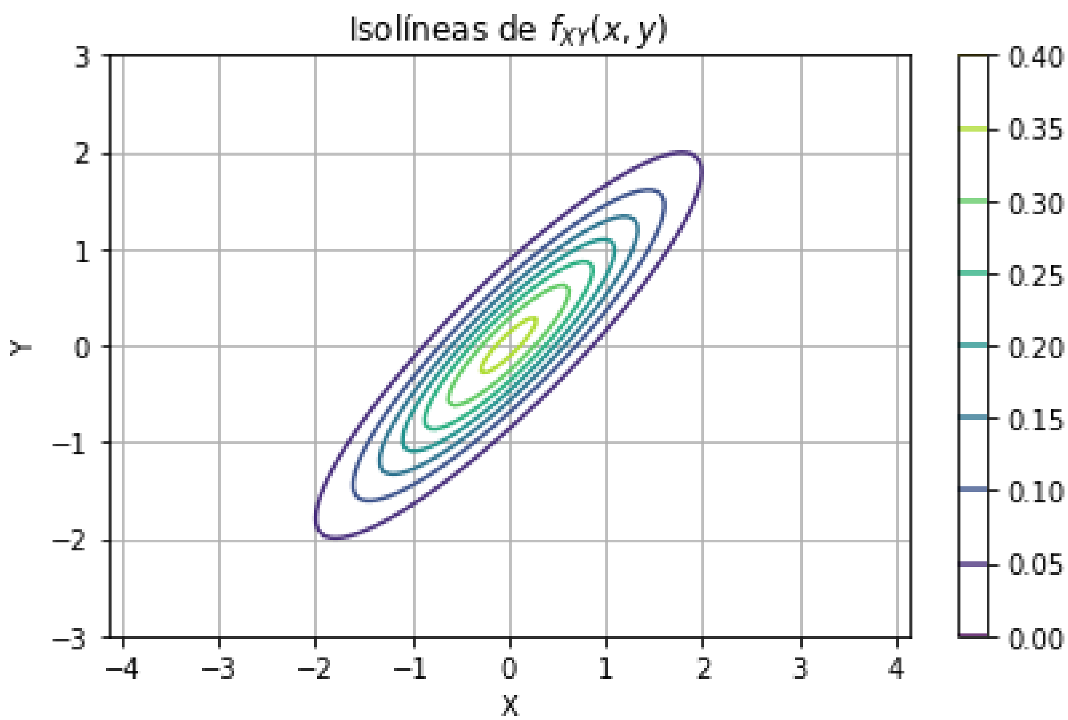


1.1.3 Dos variables aleatorias correladas

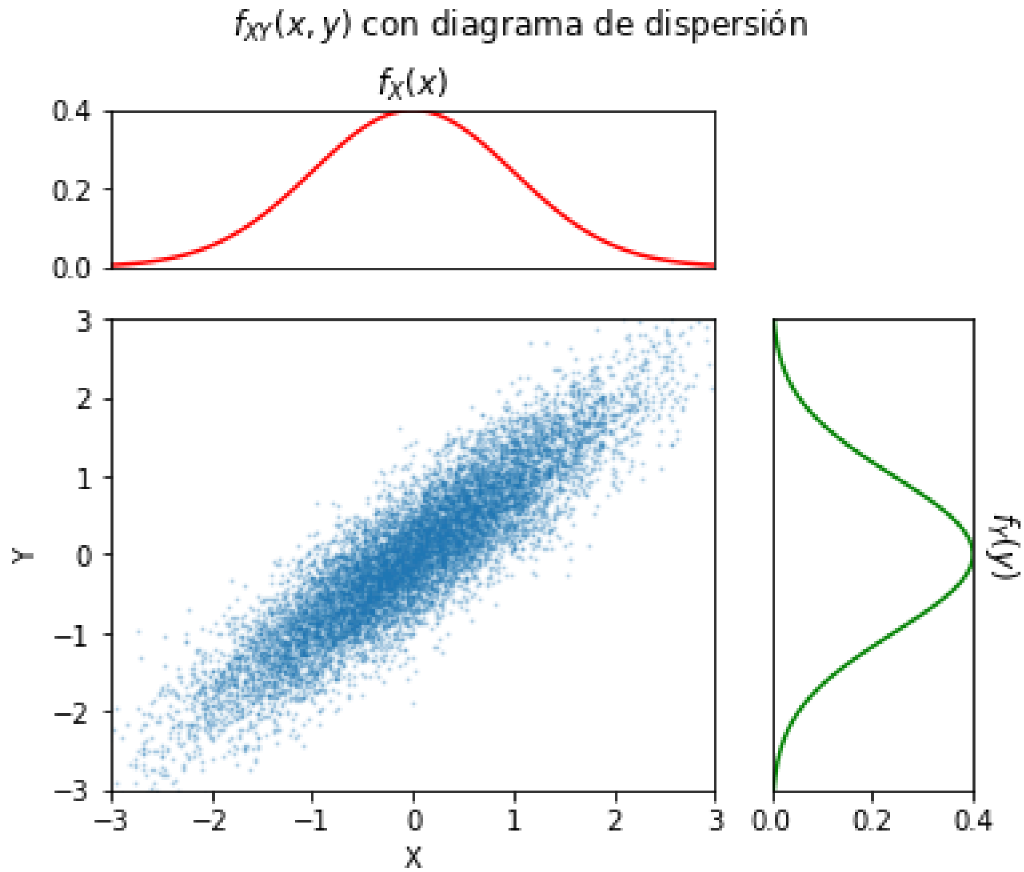
Consideremos, a modo de ejemplo, que ambas funciones de densidad marginal sean normales estándar, pero que tengan una correlación $\rho_{XY} = 0.9$.



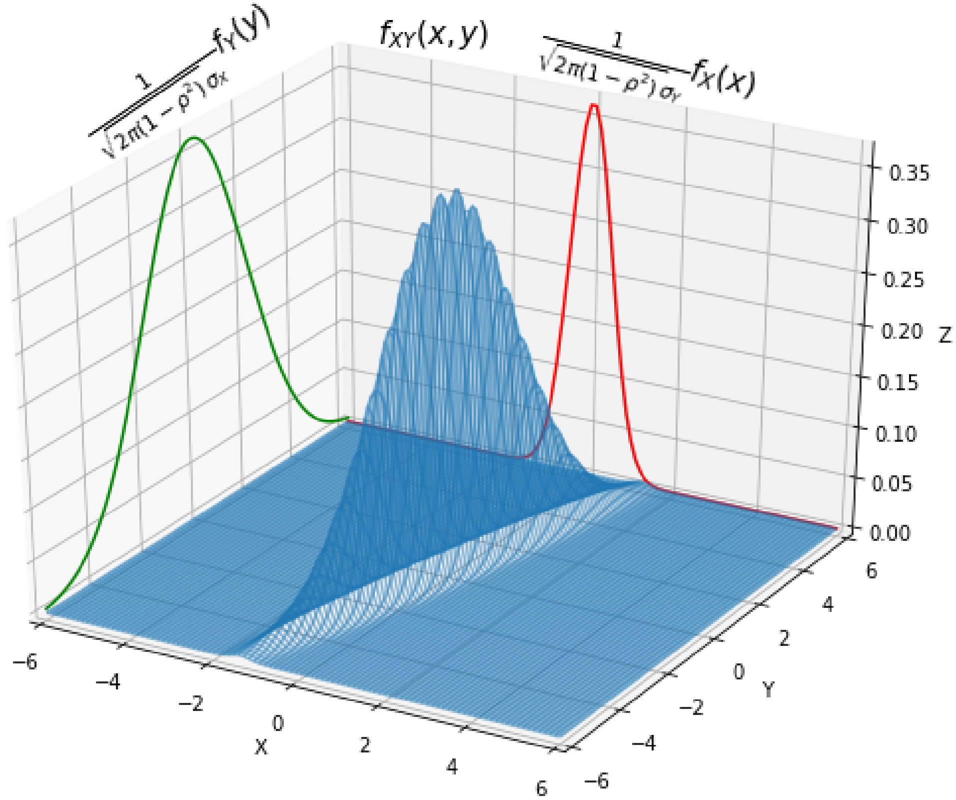
$f_Y(y)$
$\frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma_X}$
$f_{XY}(x,y)$
$\frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma_Y}$
$f_X(x)$
Z
X
Y



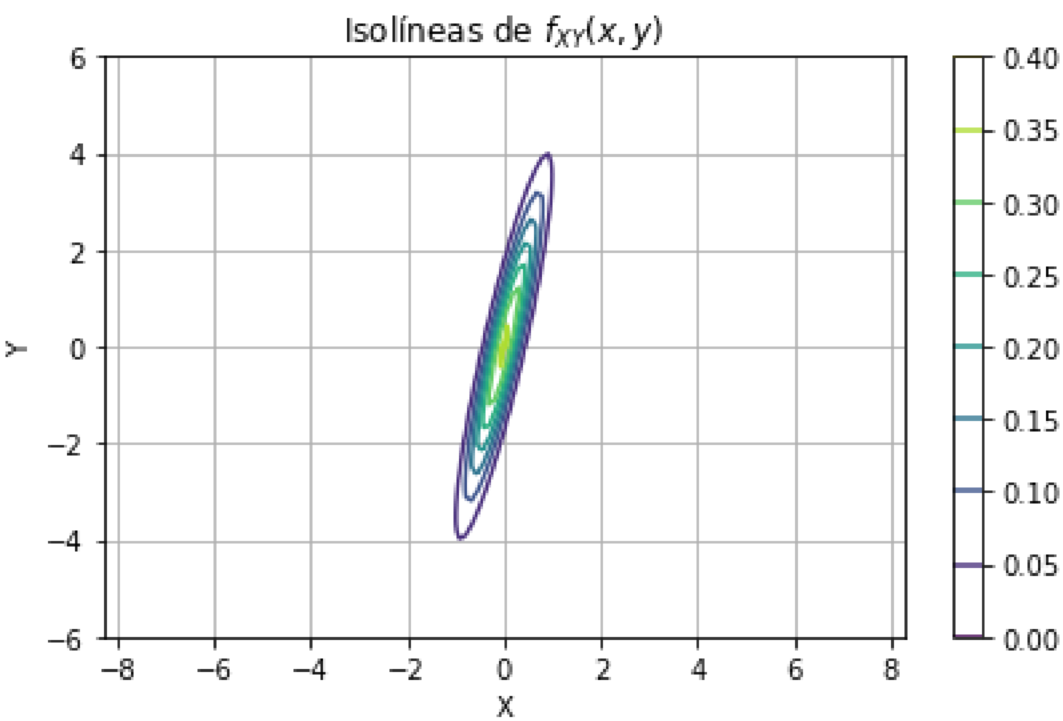
Isolíneas de $f_{XY}(x,y)$
X
Y



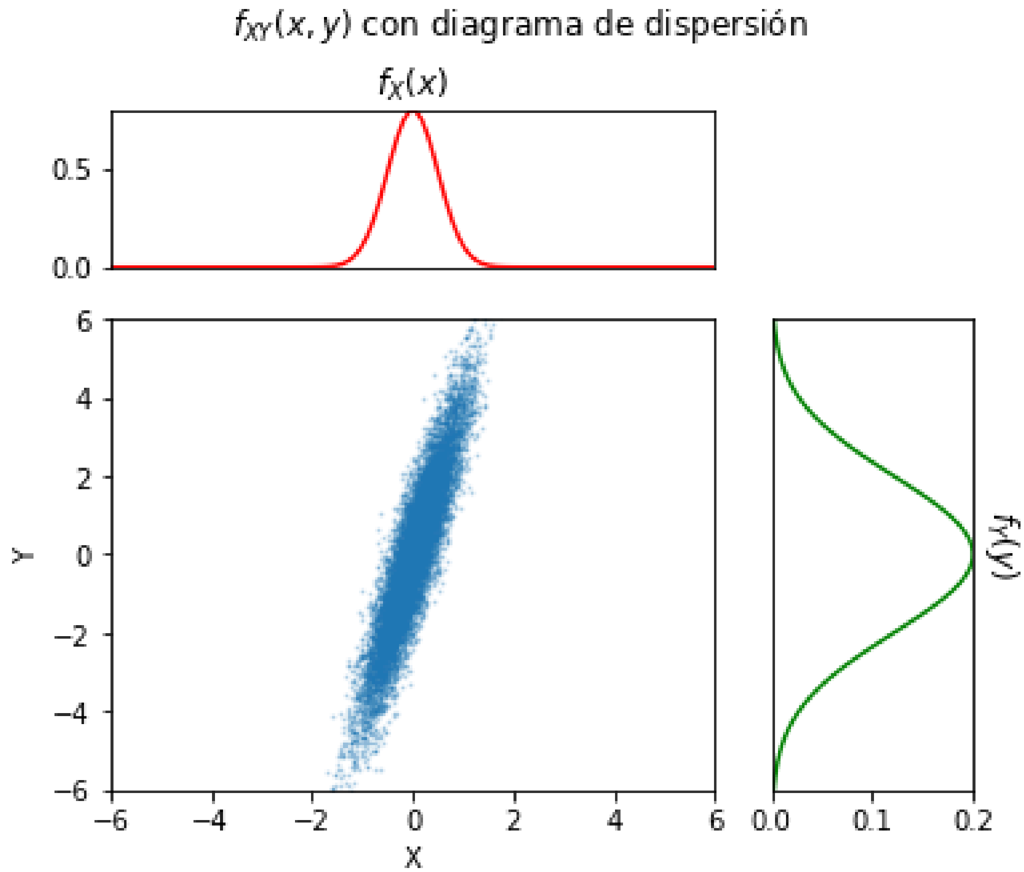
$\sigma_X = 1/2$, $\sigma_Y = 2$, $\rho_{XY} = 0.9$



$f_{XY}(x,y)$
$\frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma_X} f_Y(y)$
$\frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma_Y} f_X(x)$
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Z
X
Y
−6
−4
−2
0
2
4
6



Isolíneas de $f_{XY}(x,y)$
Y
X
6
4
2
0
−2
−4
−6
−8
8
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00



1.1.4 Transformación lineal de variables aleatorias conjuntamente Gaussianas

Si $\mathbf{V} = (X,Y)^{\mathrm{T}}$ es un vector conjuntamente Gaussiano ya sabemos que las distribuciones marginales de sus componentes también lo son. Además:

La variable aleatoria $W = aX + bY + c = \mathbf{k}^T\mathbf{V} + c$ es Gaussiana, donde:

$$E(W) \equiv \mu_W = \mathbf{k}^T\boldsymbol{\mu}_{\boldsymbol{V}} + c$$

$$Var(W) \equiv \sigma_W^2 = \mathbf{k}^{\mathrm{T}}E\left((\mathbf{V}-\boldsymbol{\mu}_{\boldsymbol{V}})(\mathbf{V}-\boldsymbol{\mu}_{\boldsymbol{V}})^{\mathrm{T}}\right)\mathbf{k} = \mathbf{k}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{V}}\mathbf{k}$$

Adviértase que si $c = 0$ y $\mathbf{k}$ es ortonormal ($\|\mathbf{k}\| = 1$), la transformación representa una proyección ortogonal de la distribución sobre la dirección definida por $\mathbf{k}$. Son casos particulares interesantes:

- $a = 1, b = c = 0$, que proporciona la distribución marginal sobre el eje X
- $a = c = 0, b = 1$, que proporciona la distribución marginal sobre el eje Y

Las variables aleatorias W, Z, definidas como sigue, son conjuntamente Gaussianas

$$W = aX + bY + c = \mathbf{k_1}^T \mathbf{V} + c \tag{1}$$
$$Z = dX + eY + f = \mathbf{k_2}^T \mathbf{V} + f \tag{2}$$

Que puede expresarse matricialmente como sigue:

$$\begin{pmatrix} W \\ Z \end{pmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} + \begin{pmatrix} c \\ f \end{pmatrix}$$

O de forma más compacta:

$$\mathbf{U} = \mathbf{K}^{\mathrm{T}}\mathbf{V} + \mathbf{p}$$

donde $\mathbf{U} = (W, Z)^{\mathrm{T}}$, $\mathbf{p} = (c, f)^{\mathrm{T}}$, $\mathbf{K} = [\mathbf{k_1}\ \mathbf{k_2}]$.

El vector transformado $\mathbf{U} = (W, Z)^{\mathrm{T}}$ tiene el siguiente vector de medias y matriz de covarianza:

$$E(\mathbf{U}) \equiv \boldsymbol{\mu}_U = \mathbf{K}^{\mathrm{T}}\boldsymbol{\mu}_V + \mathbf{p}$$

$$\mathbf{C_{WZ}} \equiv \boldsymbol{\Sigma}_U = \mathbf{K}^{\mathrm{T}} E\left((\mathbf{V} - \boldsymbol{\mu}_V)(\mathbf{V} - \boldsymbol{\mu}_V)^{\mathrm{T}} \right) \mathbf{K} = \mathbf{K}^{\mathrm{T}}\boldsymbol{\Sigma}_V \mathbf{K}$$

En general, esta transformación representa una **transformación afín** del plano (X, Y). Un caso particular de gran importancia es cuando se trata de un **movimiento**, esto es de la combinación de una **traslación** y de una **rotación**, en cuyo caso la matriz $\mathbf{K} \equiv \mathbf{R}$ es **ortogonal** $\left(\mathbf{R}\mathbf{R}^{\mathrm{T}} = \mathbf{I} \implies \mathbf{R}^{-1} = \mathbf{R}\right)$ y su determinante es uno ($|\mathbf{R}| = 1$).

Como las matrices de covarianza son siempre simétricas semidefinidas positivas, pueden ser diagonalizadas en una base ortonormal de autovectores $\boldsymbol{\Phi} = [\hat{\mathbf{e_1}}\ \hat{\mathbf{e_2}}]$, $\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} = \mathbf{I}$, siendo los autovalores correspondientes $\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$. Por tanto:

$$\boldsymbol{\Sigma}_V \hat{\mathbf{e}}_\mathbf{i} = \lambda_i \hat{\mathbf{e}}_\mathbf{i} \implies \boldsymbol{\Sigma}_V \boldsymbol{\Phi} = \boldsymbol{\Phi}\boldsymbol{\Lambda} = \boldsymbol{\Phi}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Lambda}^{1/2}$$

$$\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Sigma}_V \boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2} = \mathbf{I}$$

De ello se deduce que, si $\mathbf{K} = \boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2}$ la matriz de covarianza resultante de la transformación lineal es la identidad. Podemos extender al caso de dos variables aleatorias la tipificación de una única variable aleatoria como sigue:

$$\mathbf{U} = \mathbf{K}^{\mathrm{T}}(\mathbf{V} - \boldsymbol{\mu}_V) = \left(\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2}\right)^{\mathrm{T}} (\mathbf{V} - \boldsymbol{\mu}_V)$$

Resultando $E(\mathbf{U}) \equiv \boldsymbol{\mu}_U = \mathbf{0}$ y $\mathbf{C_{WZ}} \equiv \boldsymbol{\Sigma}_U = \mathbf{I}$.

Esta transformación suelle llamarse **blanqueado** y se extiende fácilmente al caso de N variables aleatorias conjuntas.

1.2 Distribuciones condicionadas

Consideremos dos variables aleatorias, X e Y, conjuntamente Gaussianas y que conocemos su función de densidad conjunta $f_{XY}(x,y)$ o, lo que es equivalente, que conocemos sus valores medios y su matriz de covarianza.

Ya hemos visto que, en tal caso, tanto las distribuciones marginales son Gaussianas, así como que cualquier transformación lineal resulta conjuntamente Gaussiana.

Además, la función de densidad condicionada de una variable aleatoria por la otra, también resulta Gaussiana:

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} = N(\mu_{X|Y}, \sigma_{X|Y})$$

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} = N(\mu_{Y|X}, \sigma_{Y|X})$$

Que quedan totalmente determinadas por el valor medio y la desviación típica de una variable aleatoria condicionada por la otra.

Sin más que aplicar la definición y operar:

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho_{XY}^2}} \exp\left[-\frac{1}{2}\left(\frac{y-\left[\mu_Y+\rho_{XY}\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right]}{\sigma_Y\sqrt{1-\rho_{XY}^2}}\right)^2\right]$$

donde:

$$\mu_{Y|X}(x) \equiv E(Y|X) = \mu_Y + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(x-\mu_X)$$

$$\sigma_{Y|X}^2 \equiv Var(Y|X) = \sigma_Y^2(1-\rho_{XY}^2)$$

$f_{X|Y}(x|y)$ y sus momentos se obtiene sin más que intecambiar ambas variables.

1.2.1 Ejemplo de estimación con variables conjuntamente Gaussianas

Consideremos que dos variables aleatorias (X,Y) son conjuntamente Gaussianas, y que conocemos la función de densidad de las observaciones Y dadas las causas X, $f_{Y|X}(y|x)$ o verosimilitud $L(x)$ de la causa una vez hecha la observación. La Gaussianidad implica que es suficiente con conocer la media y la desviación típica condicionadas. La media condicionada tiene necesariamente una relación lineal con las causas x como se vio en la expresión general:

$$L(x) \equiv f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi}\sigma_{Y|X}} e^{-\frac{[y-(ax+b)]^2}{2\sigma_{Y|X}^2}}$$

donde $\mu_{Y|X}(x) = ax+b$. En nuestro ejemplo vamos a considerar

$$\mu_{Y|X}(x) = 3.6x - 0.8$$

$$\sigma_{Y|X}^2 = 0.76$$

Supongamos para nuestro ejemplo que observamos una medida $Y = 1.5$. Nuestro objetivo es estimar X a partir de la observación.

En ausencia de información adicional hemos de aplicar el **principio de máxima verosimilitud**, esto es, una vez observada $Y = 1.5$, qué valor de x maximiza $L(x)$. Como sabemos, la moda de la Gaussiana coeincide con su valor medio. La simetría del exponente permite concluir que el estimador que maximiza la verosimilitud se obtiene cuando $ax + b = y$. Por tanto:

$$\hat{X} = \frac{y - b}{a} = \frac{1.5 + 0.8}{3.6} \approx 0.639$$

El estimador puede mejorarse si conocemos la **distribución *a priori*** de las causas X. De nuevo, la Gaussianidad implica que es suficiente con conocer la media y la desviación típica a priori. Consideremos pra el ejemplo que la media es $\mu_X = 0.5$ y que la desviación típica es $\sigma_Y = 0.5$.

El Teorema de Bayes junto con el de la probabilidad total nos permiten obtener la **función de densidad *a posteriori***. Sin embargo, la Gaussianidad conjunta proporciona expresiones cerradas tanto para la media como para la desviación típica a posteriori, lo que simplifica enormemente el problema.

Recapitulemos. Conocemos:

- Verosimilitud:
  - $\mu_{Y|X}(x) = ax + b = 3.6x - 0.8$
  - $\sigma^2_{Y|X} = 0.76$
- Distribución *a priori*:
  - $\mu_X = 0.5$
  - $\sigma_X = 0.5$

Necesitamos obtener μ_Y, σ_Y y ρXY y, con ello, tenemos inmediatamente tanto la caracterización conjunta como la caracterización *a posteriori*

$$ax + b = \mu_Y + \rho_{XY} \frac{\sigma_Y}{\sigma_X} (x - \mu_X)$$

Identificando coeficientes, obtenemos dos ecuaciones:

$$a = \rho_{XY} \frac{\sigma_Y}{\sigma_X} \implies \rho_{XY}\sigma_Y = a\sigma_X = 3.6\frac{1}{2} = 1.8 \tag{3}$$

$$b = \mu_Y - \rho_{XY} \frac{\sigma_Y}{\sigma_X} \mu_X \implies \mu_Y = b + a\mu_X = -0.8 + 3.6\frac{1}{2} = 1.0 \tag{4}$$

A la que añadimos la siguiente, para obtener tres ecuaciones para tres incógnitas

$$\sigma^2_{Y|X} = \sigma^2_Y(1 - \rho^2_{XY}) \implies \sigma^2_Y = \sigma^2_{Y|X} + \sigma^2_Y \rho^2_{XY} = \sigma^2_{Y|X} + (a\sigma_X)^2 = 0.76 + (1.8)^2 = 4.0$$

Por tanto: $\mu_Y = 1.0$, $\sigma_Y = 2$, $\rho_{XY} = 0.9$ (adviértase que ρ_{XY} lleva el signo de a).

Ya tenemos toda la información para obtener la **función de densidad *a posteriori***, $f_{X|Y}(x|y)$, que queda complematmente especificada por la media y la desviación típica *a posteriori* gracias a la Gaussianidad:

$$\mu_{X|Y}(y) \equiv E(X|Y) = \mu_X + \rho_{XY}\frac{\sigma_X}{\sigma_Y}(y - \mu_Y) = 0.5 + 0.225(y-1)$$

$$\sigma^2_{X|Y} \equiv Var(X|Y) = \sigma^2_X(1-\rho^2_{XY}) = 0.0475 \implies \sigma_{X|Y} \approx 0.218 < \sigma_X$$

Pueden obtenerse expresiones cerradas en función de los momentos de la verosimilitud y de la densidad *a priori*.

Podemos ahora utilizar el **principio Máximo *a Posteriori*** que hace corresponder el estimador $\hat{X}$ a la media condicionada *a posteriori*:

$$\hat{X} = \mu_{X|Y}(y) \equiv E(X|Y) = 0.5 + 0.9\frac{0.5}{2}(1.5-1) = 0.6125$$

Puede demostrarse que el estimador **Máximo *a Posteriori*** es **óptimo** en el sentido de minimizar el error cuadrático medio $E((\hat{X}-X)^2)$.