Introduction to Probability and Statistics (Continued)

Prof. Nicholas Zabaras
Center for Informatics and Computational Science
https://cics.nd.edu/
University of Notre Dame
Notre Dame, Indiana, USA

Email: nzabaras@gmail.com
URL: https://www.zabaras.com/

August 28, 2018
Contents

- Covariance, Uncorrelated Random Variables, Multivariate Random Variables, Independence Vs Uncorrelated Random Variables

- Marginal and Conditional Densities, Conditional Expectation

- The multivariate Gaussian, Multivariate Student t distribution

- Transformations of random variables

- Dirichlet distribution
References

• Following closely Chris Bishops’ PRML book, Chapter 2

• Kevin Murphy’s, Machine Learning: A probabilistic perspective, Chapter 2


Consider two random variables \( X, Y : \Omega \rightarrow \mathbb{R} \).

The joint probability distribution is defined as:

\[
P\{ X \in A, Y \in B \} = P\{ X^{-1}(A) \cap Y^{-1}(B) \} = \int_{A \times B} p(x, y) \, dx \, dy
\]

Two random variables are independent if

\[
p(x, y) = p(x) p(y)
\]

The covariance of \( X \) and \( Y \) is defined as:

\[
\text{cov}(X, Y) = \mathbb{E}\left[ (X - \mathbb{E}[X])(Y - \mathbb{E}[Y]) \right]
\]

It is straightforward to verify that:

\[
\text{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y]
\]
Correlation, Center Normalized Random Variables

- Consider two random variables $X, Y : \Omega \rightarrow \mathbb{R}$.

- The correlation coefficient of $X$ and $Y$ is defined as:

$$corrc(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

where the standard deviations of $X$ and $Y$ are

$$\sigma_X = \sqrt{\text{cov}(X)}, \sigma_Y = \sqrt{\text{cov}(Y)}$$

- The center normalized random variables are defined as:

$$\begin{align*}
\tilde{X} &= \frac{X - \mathbb{E}[X]}{\sigma_X} \\
\tilde{Y} &= \frac{Y - \mathbb{E}[Y]}{\sigma_Y}
\end{align*}$$

- It is straightforward to verify that:

$$\mathbb{E}[\tilde{X}] = \mathbb{E}[\tilde{Y}] = 0 \quad \text{var}[\tilde{X}] = \text{var}[\tilde{Y}] = 1$$
Variance and Covariance

- **Variance**
  \[
  \text{var}[f] = \mathbb{E}\left[ (f(X) - \mathbb{E}[f(X)])^2 \right] = \mathbb{E}[f(X)^2] - \mathbb{E}[f(X)]^2
  \]

- **Covariance**
  \[
  \text{cov}[X,Y] = \mathbb{E}_{X,Y}\left[ (X - \mathbb{E}[X])(Y - \mathbb{E}[Y]) \right] = \mathbb{E}_{X,Y}[XY] - \mathbb{E}[X]\mathbb{E}[Y]
  \]

It expresses the extent to which \(X\) and \(Y\) vary (linearly) together.

- If \(X\) and \(Y\) are independent, \(p(X,Y) = p(X)p(Y)\), their covariance vanishes.
Consider two random variables $X, Y : \Omega \rightarrow \mathbb{R}$.

We say that $X$ and $Y$ are uncorrelated when:

\[ \text{cov}(X, Y) = 0 \Rightarrow \mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y] \]

If $X$ and $Y$ are independent, then they are uncorrelated:

\[ \text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[(X - \mathbb{E}[X])] \mathbb{E}[(Y - \mathbb{E}[Y])] = 0 \]

The opposite is not true: Uncorrelated random variables are not independent. **Independency affects the whole density, not just the expectation.**

$X$ and $Y$ are orthogonal if

\[ \mathbb{E}[XY] = 0 \]

In the last case, the following holds:

\[ \mathbb{E}[(X + Y)^2] = \mathbb{E}[X^2] + \mathbb{E}[Y^2] \]
Consider

\[ X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} : \Omega \to \mathbb{R}^n \]

where each component \( X_i \) is an \( \mathbb{R} \) - valued variable.

\( X \) is defined by the joint probability density of its components

\[ p_X : \mathbb{R}^n \to \mathbb{R}^+ \]

Define the cumulative distribution function is defined as:

\[ F(x_1, x_2, \ldots, x_n) = \Pr\left[ X_1 < x_1, X_2 < x_2, \ldots, X_n < x_n \right] \in [0, 1] \]

Then the probability density function of \( X \) is defined as

\[ p(x_1, x_2, \ldots, x_n) = \frac{\partial^n F(x)}{\partial x_1 \partial x_2 \ldots \partial x_n} \text{ and } \int p(x)dx = 1 \]
Consider

\[ X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} : \Omega \rightarrow \mathbb{R}^n \]

where each component \( X_i \) is an \( \mathbb{R} \)-valued variable.

The expectation is defined as

\[
\mathbb{E}[X] = \int_{\mathbb{R}^n} xp(x)dx \in \mathbb{R}^n, \quad \text{or} \quad \mathbb{E}[X_i] = \int_{\mathbb{R}^n} x_i p(x)dx = \int_{\mathbb{R}} x_i p(x_i)dx_i \in \mathbb{R}, \quad i = 1, 2, \ldots, n
\]
Consider

\[ X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} : \Omega \to \mathbb{R}^n \]

The covariance matrix is:

\[
\text{cov}[X] = \int_{\mathbb{R}^n} (x - \mathbb{E}[X])(x - \mathbb{E}[X])^T p(x)dx \in \mathbb{R}^{n \times n}
\]

or equivalently:

\[
\text{cov}[X]_{ij} = \int_{\mathbb{R}^n} (x_i - \mathbb{E}[X_i])(x_j - \mathbb{E}[X_j]) p(x)dx \in \mathbb{R}, 1 \leq i, j \leq n.
\]

The covariance matrix is symmetric and positive semi-definite, i.e.

\[
\forall v \in \mathbb{R}^n, v \neq 0, \quad v^T \text{cov}[X]v = \int_{\mathbb{R}^n} [v^T(x - \bar{x})][(x - \bar{x})^Tv] p(x)dx = \int_{\mathbb{R}^n} [v^T(x - \bar{x})]^2 p(x)dx \geq 0.
\]

Note that the diagonal of the covariance matrix gives the variances of the individual components:

\[
\text{cov}[X]_{ii} = \int_{\mathbb{R}^n} (x_i - \mathbb{E}[X_i])^2 p(x)dx = \int_{\mathbb{R}^n} (x_i - \mathbb{E}[X_i])^2 \int_{\mathbb{R}^{n-1}} p(x_i, x_{-i})dx_{-i}dx_i = \int_{\mathbb{R}^n} (x_i - \mathbb{E}[X_i])^2 p(x_i)dx_i = \text{var}[X_i]
\]
The covariance matrix of a vector $X$ can be written explicitly:

$$\text{cov}[X] = \begin{pmatrix}
\text{var}[X_1] & \text{cov}[X_1, X_2] & \ldots & \text{cov}[X_1, X_d] \\
\text{cov}[X_2, X_1] & \text{var}[X_2] & \ldots \\
\text{cov}[X_d, X_1] & \text{cov}[X_d, X_2] & \ldots & \text{var}[X_d]
\end{pmatrix}$$

A normalized version of this is the correlation matrix (all elements between $[-1,1]$ (diagonal elements $= 1$))

$$R = \begin{pmatrix}
\text{corr}[X_1, X_1] & \text{corr}[X_1, X_2] & \ldots & \text{corr}[X_1, X_d] \\
\text{corr}[X_2, X_1] & \text{corr}[X_2, X_2] & \ldots \\
\text{corr}[X_d, X_1] & \text{corr}[X_d, X_2] & \ldots & \text{corr}[X_d, X_d]
\end{pmatrix}$$

$$\text{corr}[X,Y] = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X) \cdot \text{var}(Y)}}$$
Consider two scalar random variables $X$ and $Y$. We can write the following:

$$0 \leq \text{var} \left[ \frac{X}{\sigma_X} + \frac{Y}{\sigma_Y} \right] = \mathbb{E} \left[ \left( \frac{X}{\sigma_X} + \frac{Y}{\sigma_Y} \right)^2 \right] - \left( \mathbb{E} \left[ \frac{X}{\sigma_X} + \frac{Y}{\sigma_Y} \right] \right)^2$$

$$= \frac{\text{var}[X]}{\sigma_X^2} + \frac{\text{var}[Y]}{\sigma_Y^2} + 2 \frac{\text{cov}[X,Y]}{\sigma_X \sigma_Y} = 1 + 1 + 2 \text{Corr}[X,Y]$$

$$\Rightarrow \text{Corr}[X,Y] \geq -1$$

Similarly starting with

$$0 \leq \text{var} \left[ \frac{X}{\sigma_X} - \frac{Y}{\sigma_Y} \right] \Rightarrow \text{Corr}[X,Y] \leq 1$$

$$\text{corr}[X,Y] = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$$
The covariance of the vector random variables is:

$$\text{cov}[ X, Y ] = \mathbb{E}_{x,y} \left[ ( X - \mathbb{E}[X] ) ( Y - \mathbb{E}[Y] )^T \right] = \mathbb{E}_{x,y} \left[ XY^T \right] - \mathbb{E}[X] \mathbb{E}[Y^T]$$

The covariance between the components of a vector:

$$\text{cov}[ X, X ] = \mathbb{E}_{x,x} \left[ XX^T \right] - \mathbb{E}[X] \mathbb{E}[X^T]$$
Correlation as a Degree of Linearity

- It can be shown that
  
  \[
  \text{If } Y = aX + b, a > 0, \text{ then } : \text{corr}[X, Y] = +1
  \]
  
  \[
  \text{If } Y = aX + b, a < 0, \text{ then } : \text{corr}[X, Y] = -1
  \]

- The regression coefficient is \( a = \frac{\text{cov}[X, Y]}{\text{var}[X]} \).

- Think of the correlation coefficient as a degree of linearity.

- If \( X \) and \( Y \) are independent, \( p(X, Y) = p(X)p(Y) \), then \( \text{cov}[X, Y] = 0 \), and hence \( \text{corr}[X, Y] = 0 \) so they are uncorrelated.

- The converse is not true: uncorrelated does not imply independence.
Independent vs Uncorrelated

Note that:

\[
\text{var}[X + Y] = \mathbb{E}\left[\left( X + Y \right)^2 \right] - \left( \mathbb{E}[X] + \mathbb{E}[Y] \right)^2 = \\
= \mathbb{E}\left[ X^2 \right] - \left( \mathbb{E}[X] \right)^2 + \mathbb{E}\left[ Y^2 \right] - \left( \mathbb{E}[Y] \right)^2 + 2\left( \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y] \right) = \\
= \text{var}[X] + \text{var}[Y] + 2 \text{cov}[X, Y]
\]

From the above equation, we note that if \( X, Y \) are independent then:

\[
\text{var}[X + Y] = \text{var}[X] + \text{var}[Y]
\]

but note that the linearity of expectation is valid even when the variables are not independent:

\[
\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]
\]
Uncorrelated does not imply independent.

For example, let $X \sim \mathcal{U}(-1,1)$ and $Y = X^2$. Clearly $Y$ is dependent on $X$, yet one can show that $\text{corr} [X, Y] = 0$.

$$
\mathbb{E}[X] = \frac{-1+1}{2} = 0, \quad \text{var}[X] = \frac{(1-(-1))^2}{12} = \frac{1}{3}
$$

$$
\mathbb{E}[Y] = \mathbb{E}[X^2] = \text{var}[X] + (\mathbb{E}[X])^2 = \frac{1}{3} + 0^2 = \frac{1}{3}
$$

$$
\mathbb{E}[XY] = \int x^3 p(x)dx = \int_{-1}^{1} x^3 \frac{1}{2} dx = 0
$$

$$
\text{Corr}[X, Y] = \frac{\text{cov}[X, Y]}{\sigma_X \sigma_Y} = \frac{\mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y]}{\sigma_X \sigma_Y} = \frac{0 - 0 \times 1/3}{\sigma_X \sigma_Y} = 0
$$
The Figure given next shows several data sets where there is clear dependence between $X$ and $Y$, and yet the correlation coefficient is 0.

A more general measure of dependence between random variables is mutual information.

The mutual information is zero if and only if the variables are truly independent.
Several sets of \((x, y)\) points, with the correlation coefficient of \(x\) and \(y\) for each set.

The correlation reflects the \textit{noisiness and direction of a linear relationship} (top row), but not the slope of that relationship (middle), nor nonlinear relationships (bottom).

The figure in the center has a slope of 0 but the correlation coefficient is undefined because \(\text{var}[Y] = 0\).
Consider two random variables \( X, Y : \Omega \rightarrow \mathbb{R} \) with joint probability density \( p(x, y) \).

The probability density of \( X \) when \( Y \) can take any value is defined as:

\[
p(x) = \int_{\mathbb{R}} p(x, y) dy
\]

Similarly:

\[
p(y) = \int_{\mathbb{R}} p(x, y) dx.
\]
Consider two random variables $X, Y : \Omega \to \mathbb{R}$ with joint density $p(x, y)$.

The probability density of $X$ assuming that $Y = y$ is defined as

$$p(x \mid y) = \frac{p(x, y)}{p(y)}, \quad p(y) \neq 0$$

One can show this by noting the following:

$$P(a \leq X \leq b \mid y - \varepsilon \leq Y \leq y + \varepsilon) = \int_{y-\varepsilon}^{y+\varepsilon} \int_{a}^{b} p(x, y) dx dy \approx \int_{a}^{b} 2\varepsilon p(x, y) dx = \int_{a}^{b} \frac{p(x, y)}{p(y|Y=y)} dx$$

From this we derive the following important identity:

$$p(x, y) = p(x \mid y)p(y) = p(y \mid x)p(x).$$

Bayes’ rule in terms of densities now follows as:

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}$$
Conditional Expectations

- Consider two random variables $X, Y : \Omega \to \mathbb{R}$.

- We define the conditional expectation as: $\mathbb{E}[X \mid y] = \int xp(x \mid y)dx$

- The expectation of $X$ via conditional expectation can be computed as:

$$
\mathbb{E}[X] = \int xp(x)dx = \int x \left( \int \frac{p(x, y)dy}{p(x \mid y)p(y)} \right)dx \Rightarrow
$$

$$
\mathbb{E}[X] = \int \left( \int xp(x \mid y)dx \right)p(y)dy = \int \mathbb{E}[X \mid y]p(y)dy \Rightarrow
$$

$$
\mathbb{E}[X] = \int \mathbb{E}[X \mid y]p(y)dy
$$
Suppose $y = f(x) = Ax + b$. You can show that:

$$
\mathbb{E}[y] = A\mathbb{E}[x] + b
$$

$$
\text{cov}[y] = A\text{cov}[x]A^T
$$

For a scalar-valued function $y = f(x) = a^T x + b$:

$$
\mathbb{E}[y] = a^T\mathbb{E}[x] + b
$$

$$
\text{var}[y] = a^T\text{cov}[x]a
$$
A random variable $X \in \mathbb{R}$ is Gaussian or normally distributed $X \sim \mathcal{N}(x_0, \sigma^2)$ if:

$$P\{X \leq t\} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{t} \exp\left(-\frac{1}{2\sigma^2}(x-x_0)^2\right) dx$$

A **multivariate** $X \in \mathbb{R}^D$ is Gaussian if its probability density is

$$p(x) = \left(\frac{1}{(2\pi)^D \det \Sigma}\right)^{1/2} \exp\left(-\frac{1}{2}(x-x_0)^T \Sigma^{-1} (x-x_0)\right)$$

where $x_0 \in \mathbb{R}^D$, $\Sigma \in \mathbb{R}^{D \times D}$ is symmetric positive definite (covariance matrix).

The symmetry property of the covariance matrix does not affect the value of $(x-x_0)^T \Sigma^{-1} (x-x_0)$. **However, for symmetric covariance matrices we only need to describe** $D(D+1)/2$ **elements rather than** $D^2$.

It is invariant under linear transformations, i.e. for $A, B \in \mathbb{M}^{M \times D}$, $c \in \mathbb{R}^M$

$$X_1 \sim \mathcal{N}(\mu_1, \Sigma_1), \ X_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$$

and $X_1, X_2$ independent $\Rightarrow$

$$AX_1 + BX_2 + c \sim \mathcal{N}(A\mu_1 + B\mu_2 + c, A\Sigma_1 A^T + B\Sigma_2 B^T)$$
Conditional and Marginal Probability Densities

Ellipsoids: equiprobability curves of $p(x, y)$.

$p(x|y=2)$

$p(x)$

Conditional

Marginal

Link [here](#) for a MatLab program to generate these figures.
A probability density transforms differently from functions.

Let \( x = g(y) \).

\[
p_y(y) = p_x(g(y)) \left| \frac{dx}{dy} \right| = p_x(g(y)) |g'(y)| = p_x(g(y))sg'(y), \quad s \in \{-1,1\}
\]

This is easily derived by taking observations in the interval \((x, x + dx)\) to be transformed to observations in \((y, y + dy)\), i.e.

\[
p_y(y)dy = p_x(x)dx
\]
For example consider the Gamma distribution

\[
\text{Gamma}(x \mid a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-xb}
\]

Let us compute the density of \( Y = 1/X \).

\[
p_y(y) = p_x(g(y)) \left| \frac{dx}{dy} \right| = \frac{b^a}{\Gamma(a)} y^{-(a-1)} e^{-\frac{b}{y}} \left| -\frac{1}{y^2} \right| = \frac{b^a}{\Gamma(a)} y^{-(a+1)} e^{-\frac{b}{y}},
\]

where \( \frac{dx}{dy} = -\frac{1}{y^2} \)

This is the Inverse Gamma distribution

\[
\text{InvGamma}(y \mid a, b) = \frac{b^a}{\Gamma(a)} y^{-(a+1)} e^{-\frac{b}{y}}
\]
If \( f \) is an invertible mapping, we can define the pdf of the transformed variables using the Jacobian of the inverse mapping \( y \to x \):

\[
p_y(y) = p_x(x) \left| \det \frac{\partial x}{\partial y} \right|, \quad \frac{\partial y}{\partial x} = \begin{pmatrix}
\frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\
\vdots & \ddots & \vdots \\
\frac{\partial y_n}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_n}
\end{pmatrix}
\]

As an example, it is trivial to show that transforming a density from Cartesian coordinates \( x = (x, y) \) to polar coordinates \( y = (r, \theta) \), where \( x = r \cos \theta \) and \( y = r \sin \theta \), gives:

\[
p_{r,\theta}(r, \theta) = p_{x,y}(r \cos \theta, r \sin \theta) r
\]

\[
p_{r,\theta}(r, \theta) drd\theta = p_{x,y}(r \cos \theta, r \sin \theta) r drd\theta
\]
Transformation of Probability Density

- Recall \( p_y(y) = p_x(g(y)) \frac{dx}{dy} = p_x(g(y))sg'(y), \ s \in \{-1, 1\} \)

- Using this Eq., note that modes of densities depend on the choice of variables (see 2\(^{nd}\) term on the rhs below):
  \[ p'_y(y) = sp'_x(x) \left\{ g'(y) \right\}^2 + sp_x(g(y))g''(y) \]

- Consider \( X \sim \mathcal{N}(6,1) \) and the following
  \[ x = g(y) = \ln \frac{y}{1-y} + 5, \ y = g^{-1}(x) = \frac{1}{1 + e^{-x+5}} \]

- Transforming \( p_x(x) \) as a function gives the same mode for \( p_x(g(y)) \). The actual mode of \( p_y(y) \) is shifted.

- The histogram of \( p_y(y) \) is obtained as:
  \[ y^{(s)} = g^{-1}(x^{(s)}), \ \text{where} : x^{(s)} \sim p_x(x) \]
Multivariate Student’s $\mathcal{I}$ Distribution

\[ p(x \mid \mu, a, b) = \int_0^\infty \mathcal{N}(x \mid \mu, \tau^{-1}) \Gamma((a,b) \mid \tau) d\tau \]

- If we return to the derivation of the univariate Student’s $\mathcal{I}$ distribution and substitute $\nu = 2a$, $\lambda = \frac{a}{b}$, $\eta = \tau b / a$, and use

\[ \Gamma(a, b) = \frac{b^a}{\Gamma(a)} \tau^{a-1} e^{-b\tau} \]

we can write the Student’s $\mathcal{I}$ distribution as:

\[ \mathcal{I}(x \mid \mu, \lambda, \nu) = \int_0^\infty \mathcal{N}(x \mid \mu, (\eta\lambda)^{-1}) \Gamma(\eta \mid \nu / 2, \nu / 2) d\eta \]

- This form is useful in providing generalization to a multivariate Student’s $\mathcal{I}$

\[ \mathcal{I}(x \mid \mu, \Lambda, \nu) = \int_0^\infty \mathcal{N}(x \mid \mu, (\eta\Lambda)^{-1}) \Gamma(\eta \mid \nu / 2, \nu / 2) d\eta \]

*Use change of variables for distributions, also $d\tau = \lambda d\eta$, and notice that the extra $\lambda$-terms that appear cancel out.

Statistical Computing, University of Notre Dame, Notre Dame, IN, USA (Fall 2018, N. Zabaras)
Multivariate Student's $T$ Distribution

$$\mathcal{T}(x \mid \mu, \Lambda, \nu) = \int_0^\infty \mathcal{N}(x \mid \mu, (\eta\Lambda)^{-1}) \operatorname{Gamma}(\eta \mid \nu/2, \nu/2) \, d\eta$$

- This integral can be computed analytically as:

$$\mathcal{T}(x \mid \mu, \Lambda, \nu) = \frac{\Gamma\left(\frac{D + \nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left|\Lambda\right|^{D/2} \left(\pi\nu\right)^{D/2} \left[1 + \frac{\Delta^2}{\nu}\right]^{-\nu/2 - D/2}$$

$$\Delta^2 = (x - \mu)^T \Lambda (x - \mu) \ (\text{Mahalanobis Distance})$$

- One can derive the above form of the distribution by substitution in the Eq. on the top.

$$\mathcal{T}(x \mid \mu, \Lambda, \nu) = \frac{(\nu/2)^{D/2}}{\Gamma(\nu/2)} \frac{|\Lambda|^{D/2}}{(2\pi)^{D/2}} \int_0^\infty \eta^{D/2} \eta^{\nu - 1} e^{-\nu\eta/2} e^{-\eta\Delta^2/2} \, d\eta$$

Use $\tau = \eta\left(\nu/2 + \Delta^2/2\right)$

$$= \frac{(\nu/2)^{D/2}}{\Gamma(\nu/2)} \frac{|\Lambda|^{D/2}}{(2\pi)^{D/2}} \int_0^\infty \tau^{D/2 + \nu/2} e^{-\tau} \, d\tau = \frac{\Gamma(\nu/2 + d/2)}{\Gamma(\nu/2)} \frac{|\Lambda|^{D/2}}{(\pi\nu)^{D/2}} \left(1 + \frac{\Delta^2}{\nu}\right)^{-D/2 - \nu/2}$$

- Normalization proof is immediate from the normalization of the normal & Gamma distributions.
Some useful results of the multivariate Student's $T$ are given below:

\[ \mathbb{E}[x] = \mu \quad \text{if } \nu > 1, \quad \text{cov}[x] = \frac{\nu}{\nu - 2} \Lambda^{-1} \quad \text{if } \nu > 2, \quad \text{mode}[x] = \mu \]

One can show easily the expression for the mean by using $x = z + \mu$:

\[ \mathbb{E}[x] = \int_{-\infty}^{+\infty} \frac{\nu}{\nu - 2} \Lambda^{-1} \left[ 1 + \frac{\Delta^2}{\nu} \right]^{-\nu/2-D/2} (z + \mu) dz \]

The 1\textsuperscript{st} term drops out since $\mathcal{S}(z | 0, \Lambda, \nu)$ is even. The 2\textsuperscript{nd} term gives $\mu$ from the normalization of the distribution.

The covariance is computed as:

\[
\text{cov}[x] = \int_{\eta=0}^{+\infty} \left[ \int_{x} \mathcal{N}\left( x | \mu, (\eta\Lambda)^{-1} \right)(x - \mu)(x - \mu)^T \ dx \right] \text{Gamma}(\eta | \nu/2, \nu/2) \ d\eta = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \int_{\eta=0}^{+\infty} \eta^{\nu/2-1} e^{-\nu\eta/2} d\eta
\]

\[ = \Lambda^{-1} \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)^{(\nu/2)^{\nu/2}} (\nu/2)^{\nu/2-1}} = \frac{(\nu/2)\Gamma(\nu/2-1)}{\Gamma(\nu/2)} \Lambda^{-1} = \frac{\nu/2}{\nu/2-1} \Lambda^{-1} = \frac{\nu}{\nu-2} \Lambda^{-1} \]
### Multivariate Student's $\mathcal{T}$ Distribution

$$
\mathcal{T}(x | \mu, \Lambda, \nu) = \frac{\Gamma\left(\frac{\nu + D}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\nu^{D/2}} \left| \Lambda^{1/2} \right| \left(1 + \frac{\Lambda^2}{\nu} \right)^{-\nu/2-D/2}
$$

- Differentiation with respect to $x$ also shows the mode being $\boldsymbol{\mu}$:

  $$
  \mathbb{E}[x] = \mu \quad \text{if } \nu > 1, \quad \text{cov}[x] = \frac{\nu}{\nu-2} \Lambda^{-1} \quad \text{if } \nu > 2, \quad \text{mode}[x] = \mu
  $$

- The Student’s $\mathcal{T}$ has fatter tails than a Gaussian. *The smaller $\nu$ is the fatter the tails.*

- For $\nu \to \infty$, the distribution approaches a Gaussian. Indeed note that:

  $$
  \left[1 + \frac{\Lambda^2}{\nu}\right]^{-\nu/2-D/2} \approx \exp\left(-\left(\frac{\nu}{2} + \frac{D}{2}\right) \ln\left[1 + \frac{\Lambda^2}{\nu}\right] \right) \approx \exp\left(-\frac{\nu}{2} \left(\frac{\Lambda^2}{\nu} - \frac{1}{2} \left(\frac{\Lambda^2}{\nu}\right)^2\right)\right) = \exp\left(-\frac{\Lambda^2}{2} + O(\nu^{-1})\right)
  $$

- The distribution can also be written in terms of $\Sigma = \Lambda^{-1}$ (scale matrix – not the covariance) or $\mathbf{V} = \nu \Sigma$. 

---

Statistical Computing, University of Notre Dame, Notre Dame, IN, USA (Fall 2018, N. Zabaras)
We introduce the Dirichlet distribution as a family of “conjugate priors” (to be formally introduced in a follow up lecture) for the parameters $\mu_k$ in the multinomial distribution.

The Dirichlet distribution $\mathcal{D}(\alpha)$, is a family of continuous multivariate probability distributions parametrized by the vector $\alpha$ of positive reals.

It is the multivariate generalization of the Beta distribution.
Its probability density function returns the belief that the probabilities of $K$ rival events are $\mu_k$ given that each event has been observed $\alpha_k - 1$ times:

$$p(\mu | \alpha) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1},$$

$$0 \leq \mu_k \leq 1,$$

$$\sum_{k=1}^{K} \mu_k = 1$$

The distribution over the space of $\mu_k$ is $K - 1$ dimensional due to the last constraint above.
The Dirichlet distribution of order $K \geq 2$ with parameters $\alpha_1, \ldots, \alpha_K > 0$ has a PDF with respect to Lebesgue measure on $\mathbb{R}^{K-1}$ given by

$$p(\mu | \alpha) = \frac{1}{\text{Beta}(\alpha)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

for all $\mu_1, \ldots, \mu_{K-1} > 0$ satisfying $\mu_1 + \ldots + \mu_{K-1} < 1$, where $\mu_K$ is an abbreviation for $1 - \mu_1 - \cdots - \mu_{K-1}$. The normalizing constant is the multinomial Beta function:

$$\text{Beta}(\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}, \quad \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K)^T$$

The Dirichlet distribution over $(\mu_1, \mu_2, \mu_3)$ is confined on a plane as shown.
We write the Dirichlet distribution as:

\[ p(\mu | \alpha) = K(\alpha) \prod_{k=1}^{K} \mu_k^{\alpha_k-1}, \quad K(\alpha) = \frac{\Gamma(a_0)}{\Gamma(a_1) \cdots \Gamma(a_K)}, \quad a_0 = a_1 + \ldots + a_K \]

Note the following useful relation:

\[ \frac{\partial}{\partial \alpha_j} \prod_{k=1}^{K} \mu_k^{\alpha_k-1} = \frac{\partial}{\partial \alpha_j} \prod_{k=1}^{K} e^{(\alpha_k-1) \ln \mu_k} = \ln \mu_j \prod_{k=1}^{K} \mu_k^{\alpha_k-1} \]

From this we can derive an interesting expression for \( \mathbb{E}[\ln \mu_j] \):

\[
\mathbb{E}[\ln \mu_j] = K(\alpha) \int_{0}^{1} \ldots \int_{0}^{1} \ln \mu_j \prod_{k=1}^{K} \mu_k^{\alpha_k-1} d\mu_1 \ldots d\mu_K = K(\alpha) \int_{0}^{1} \ldots \int_{0}^{1} \frac{\partial}{\partial \alpha_j} \prod_{k=1}^{K} \mu_k^{\alpha_k-1} d\mu_1 \ldots d\mu_K = \\
K(\alpha) \frac{\partial}{\partial \alpha_j} \int_{0}^{1} \ldots \int_{0}^{1} \prod_{k=1}^{K} \mu_k^{\alpha_k-1} d\mu_1 \ldots d\mu_K = K(\alpha) \frac{\partial}{\partial \alpha_j} \left( \frac{1}{K(\alpha)} \right) = - \frac{\partial \ln K(\alpha)}{\partial \alpha_j} = \frac{\partial \ln \Gamma(\alpha_j)}{\partial \alpha_j} - \frac{\partial \ln \Gamma(\alpha_0)}{\partial \alpha_0}
\]

where \( \Psi(\alpha) = d \ln \Gamma(\alpha) / d\alpha \) is the digamma function.

\[
\mathbb{E}[\ln \mu_j] = \Psi(\alpha_j) - \Psi(\alpha_0), \quad \Psi(\alpha_j) = \frac{\partial \ln \Gamma(\alpha_j)}{\partial \alpha_j}, \quad \Psi(\alpha_0) = \frac{\partial \ln \Gamma(\alpha_0)}{\partial \alpha_0}
\]
To show the normalization, we use induction. The case for $M = 2$ was shown earlier for the Beta distribution.

Assume that the Dirichlet normalization formula is valid for $M - 1$ terms. We will show the formula for $M$ terms:

$$p_M (\mu_1, \ldots, \mu_{M-1}) = C_M \prod_{k=1}^{M-1} \mu_k^{\alpha_k-1} \left(1 - \sum_{j=1}^{M-1} \mu_j \right)^{\alpha_M-1}$$

Let us integrate out $\mu_{M-1}$:

$$p_{M-1} (\mu_1, \ldots, \mu_{M-2}) = C_M \int_0^{1-\sum_{j=1}^{M-2} \mu_j} \left(\prod_{k=1}^{M-2} \mu_k^{\alpha_k-1}\right) \mu_{M-1}^{\alpha_{M-1}-1} \left(1 - \sum_{j=1}^{M-2} \mu_j - \mu_{M-1} \right)^{\alpha_{M-1}-1} d\mu_{M-1} =$$

$$= C_M \left(\prod_{k=1}^{M-2} \mu_k^{\alpha_k-1}\right) \int_0^{1-\sum_{j=1}^{M-2} \mu_j} t^{\alpha_{M-1}-1} \left(1 - \sum_{j=1}^{M-2} \mu_j \right)^{\alpha_{M-1}-1+\alpha_M-1+1} (1-t)^{\alpha_M-1} dt$$
Dirichlet Distribution: Normalization

\[
p_{M-1}(\mu_1, \ldots, \mu_{M-2}) = C_M \left( \prod_{k=1}^{M-2} \mu_k^{\alpha_k-1} \right) \left( 1 - \sum_{j=1}^{M-2} \mu_j \right)^{\alpha_{M-1} + \alpha_M-1} \int_0^1 t^{\alpha_{M-1}-1} (1-t)^{\alpha_M-1} dt = \\
= C_M \left( \prod_{k=1}^{M-2} \mu_k^{\alpha_k-1} \right) \left( 1 - \sum_{j=1}^{M-2} \mu_j \right)^{\alpha_{M-1} + \alpha_M-1} \frac{\Gamma(\alpha_{M-1})\Gamma(\alpha_M)}{\Gamma(\alpha_{M-1} + \alpha_M)}
\]

The last step above comes from the normalization of Beta.

What we have above is an \((M - 1)\) term Dirichlet distribution with coefficients \(\alpha_1, \ldots, \alpha_{M-2}, \alpha_{M-1} + \alpha_M\). Since we assumed that the normalization formula is valid for \((M - 1)\) terms, we must have:

\[
1 = C_M \frac{\Gamma(\alpha_1) \ldots \Gamma(\alpha_{M-2}) \Gamma(\alpha_{M-1} + \alpha_M)}{\Gamma(\alpha_1 + \ldots + \alpha_M)} \frac{\Gamma(\alpha_{M-1}) \Gamma(\alpha_M)}{\Gamma(\alpha_{M-1} + \alpha_M)} \Rightarrow \\
C_M = \frac{\Gamma(\alpha_1 + \ldots + \alpha_M)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_{M-2}) \Gamma(\alpha_{M-1}) \Gamma(\alpha_M)}
\]
Using the multinomial as a “likelihood” and the Dirichlet as “the conjugate prior”, we arrive at the following “posterior”

\[
p(\mu | \mathcal{D}) \propto p(\mathcal{D} | \mu) \ p(\mu) \implies p(\mu | \mathcal{D}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1}
\]

which is a Dirichlet distribution \( \text{Dir}(\mu | \alpha_1 + m_1, \ldots, \alpha_K + m_K) \).

The normalization factor is computed easily from the normalization factor of the Dirichlet as:

\[
p(\mu | \mathcal{D}) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k + N\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k + m_k)} \left(\prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1}\right)
\]

\( \alpha_k \) can be interpreted as “the effective number of prior observations of \( x_k = 1 \)”. 

Dirichlet Distribution
Dirichlet Distribution

Examples of Dirichlet distribution over \((\mu_1, \mu_2, \mu_3)\) which can be plotted in 2D since \(\mu_3 = 1 - \mu_1 - \mu_2\).

- **Uniform** at \((1/3, 1/3, 1/3)\)
- **Narrow centered** at \((1/3, 1/3, 1/3)\)
- **Broad centered** at \((1/3, 1/3, 1/3)\)

\[ a_0 = a_1 + \ldots + a_K \]

controls how peaked the distribution is.

The \(a_K\) control the location of the peak.
Dirichlet Distribution

The Dirichlet distribution over \((\mu_1, \mu_2, \mu_3)\) where the horizontal axes are \(\mu_1\) and \(\mu_2\) and the vertical axis is the density.

\[
\{\alpha_k\} = 0.1
\]

\[
\{\alpha_k\} = 10
\]

\[
\{\alpha_k\} = 1
\]

MatLab Code
The Dirichlet distribution over \((\mu_1, \mu_2, \mu_3)\) where the horizontal axes are \(\mu_1\) and \(\mu_2\) and the vertical axis is the density.

\[\{\alpha_K\} = \{2, 2, 2\}\]

If \(a_k < 1/3\) for all \(k\), we obtain spikes at the corners.

\[\{\alpha_K\} = \{0.1, 0.1, 0.1\}\]

\[\{\alpha_K\} = \{10, 10, 10\}\]

Run `visDirichletGui` & `dirichlet3dPlot` from PMTK.
Dirichlet Distribution

Samples from 5-dimensional symmetric Dirichlet distribution.

\[ \{ \alpha_K \} = \{ 5, 5, \ldots, 5 \} \]

Run `dirichletHistogramDemo` from `PMTK`
In closing, we have the following properties (you only need the normalization \( \int \prod_{k=1}^{K} \mu_k^{a_k-1} d\mu = \frac{\Gamma(a_1) \cdots \Gamma(a_K)}{\Gamma(a_0)} \), \( a_0 = a_1 + \cdots + a_K \) of the Dirichlet distribution and the property \( \Gamma(x+1) = x\Gamma(x) \) to prove them):

- \( \mathbb{E}[\mu_k] = \frac{\alpha_k}{\alpha_0} \), \( \text{mode}[\mu_k] = \frac{\alpha_k - 1}{\alpha_0 - 1} \), \( \text{var}[\mu_k] = \frac{\alpha_k (\alpha_0 - \alpha_k)}{\alpha_0^2 (\alpha_0 + 1)} \), \( \text{cov}[\mu_j, \mu_l] = -\frac{\alpha_j \alpha_l}{\alpha_0^2 (\alpha_0 + 1)} \) (\( j \neq l \))

where: \( \alpha_0 = \sum_{k=1}^{K} \alpha_k \)

- Often we use:

\[ \alpha_k = \alpha / K \]

In this case:

\[ \mathbb{E}[\mu_k] = \frac{1}{K}, \quad \text{var}[\mu_k] = \frac{K - 1}{K^2 (\alpha + 1)} \]

Increasing \( \alpha \) increases the precision of the distribution.