# 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 Expectaction
- 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 probablistic perspective, Chapter 2
- Jaynes, E. T. (2003). Probability Theory: The Logic of Science. Cambridge University Press.
- Bertsekas, D. and J. Tsitsiklis (2008). Introduction to Probability. Athena Scientific. 2nd Edition
- Wasserman, L. (2004). All of statistics. A Concise Course in Statistical Inference. Springer.

# Covariance

- Consider two random variables $X, Y : \Omega \to \mathbb{R}$.
- The joint probability distribution is defined as:

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

- Two random variables are independent if

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

- The covariance of *X* and *Y* is defined as:

$$\operatorname{cov}(X,Y) = \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)\big(Y - \mathbb{E}[Y]\big)\Big]$$

- It is straight forward to verify that: $\operatorname{cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$

# Correlation, Center Normalized Random Variables

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

- 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:

$$\begin{cases} \tilde{X} = \dfrac{X - \mathbb{E}[X]}{\sigma_X} \\ \tilde{Y} = \dfrac{Y - \mathbb{E}[Y]}{\sigma_Y} \end{cases}$$

- It is straight forward to verify that:

$$\mathbb{E}\left[\tilde{X}\right] = \mathbb{E}\left[\tilde{Y}\right] = 0 \qquad \text{var}\left[\tilde{X}\right] = \text{var}\left[\tilde{Y}\right] = 1$$

# Variance and Covariance

- **Variance**

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

- **Covariance**

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

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.

# Uncorrelated Random Variables

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

- 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:

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

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:

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

# Multivariate Random Variables

- Consider $$X = \begin{bmatrix} X_1 \\ X_2 \\ .. \\ X_n \end{bmatrix} : \Omega \to \mathbb{R}^n$$

  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:

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

- *Then the probability density function of X* is defined as

$$p(x_1, x_2, ..., x_n) = \frac{\partial^n F(\boldsymbol{x})}{\partial x_1 \partial x_2 ... \partial x_n} \; and \; \int p(\boldsymbol{x}) dx = 1$$

# Multivariate Random Variables

- Consider

$$X = \begin{bmatrix} X_1 \\ X_2 \\ .. \\ X_n \end{bmatrix} : \Omega \to \mathbb{R}^n$$

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

- The expectation is defined as

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

# Covariance Matrix

- Consider $$X = \begin{bmatrix} X_1 \\ X_2 \\ .. \\ X_n \end{bmatrix} : \Omega \to \mathbb{R}^n$$

- The covariance matrix is: $$\text{cov}[X] = \int_{\mathbb{R}^n} \left(x - \mathbb{E}[X]\right)\left(x - \mathbb{E}[X]\right)^T p(x)dx \in \mathbb{R}^{n \times n}$$

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

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

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

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

$$\text{cov}[X]_{ii} = \int_{\mathbb{R}^n} (x_i - \mathbb{E}[X_i])^2 p(x)dx = \int_{\mathbb{R}^n} (x_i - \mathbb{E}[X_i])^2 \int_{\mathbb{R}^{n-1}} p(x_i, x_i')dx_i' dx_i = \int_{\mathbb{R}^n} (x_i - \mathbb{E}[X_i])^2 p(x_i)dx_i = \text{var}[X_i]$$

# Covariance Matrix

- The covariance matrix of a vector $\boldsymbol{X}$ can be written explicitly:

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

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

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

$$corr[X, Y] = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}}$$

# Correlation Coefficient Between -1 and 1

- Consider two scalar random variables $X$ and $Y$. We can write the following:

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

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

$$\Rightarrow Corr[X,Y] \ge -1$$

- Similarly starting with

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

$$corr[X,Y] = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}}$$

# Variance and Covariance

- The covariance of the vector random variables is:

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

- The covariance between the components of a vector:

$$\text{cov}\left[\boldsymbol{X},\boldsymbol{X}\right]=\mathbb{E}_{\boldsymbol{X},\boldsymbol{X}}\left[\boldsymbol{X}\boldsymbol{X}^{T}\right]-\mathbb{E}\left[\boldsymbol{X}\right]\mathbb{E}\left[\boldsymbol{X}^{T}\right]$$

# Correlation as a Degree of Linearity

- It can be shown that

$$If\ Y = aX + b, a > 0, then: corr\left[X, Y\right] = +1$$

$$If\ Y = aX + b, a < 0, then: corr\left[X, Y\right] = -1$$

- The regression coefficient is $a = cov[X, Y] / var[X]$.
- *Think of the correlation coefficient as a degree of linearity.*
- If $X$ and $Y$ are independent, $p(X, Y) = p(X)p(Y)$, then $cov[X, Y] = 0$, and hence $corr[X, Y] = 0$ so they are uncorrelated.
- The converse is not true: uncorrelated does not imply independence.

# Independent vs Uncorrelated

- Note that:

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

$$= \mathbb{E}\left[X^2\right] - \left(\mathbb{E}[X]\right)^2 + \mathbb{E}\left[Y^2\right] - \left(\mathbb{E}[Y]\right)^2 + 2\left(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\right) =$$

$$= \text{var}[X] + \text{var}[Y] + 2\,\text{cov}[X,Y]$$

- 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:

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

# Correlation and Dependence

- 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 $corr\,[X,Y\,] = 0$.

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

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

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

$$Corr[X,Y] = \frac{\mathrm{cov}[X,Y]}{\sigma_X \sigma_Y} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sigma_X \sigma_Y} = \frac{0 - 0 \times 1/3}{\sigma_X \sigma_Y} = 0$$

# Mutual Information

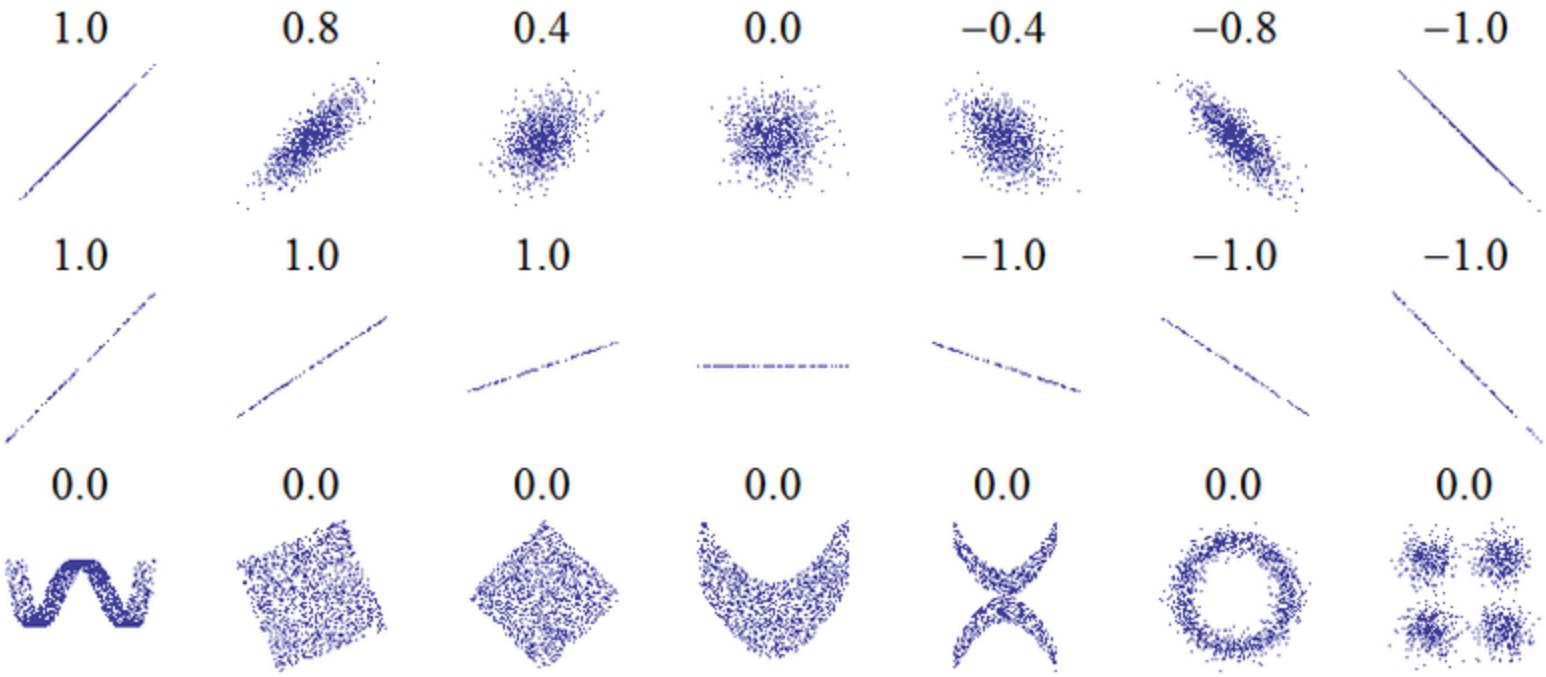
- The Figure given next shows several data sets where there is clear dependence between $X$ and $Y$, and yet the correlation coefficient is $0$.

- 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.

# Uncorrelated Random Variables

- Several sets of $(x, y)$ points, with the correlation coefficient of $x$ and $y$ for each set.
- The correlation reflects the *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 $var[Y] = 0$.



# Marginal Density

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

$$p(y) = \int_{\mathbb{R}} p(x, y) dx.$$

# Conditional Probability Density

- 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

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

- One can show this by noting the following:

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

- 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: $\mathbb{E}[X \mid y] = \int_{\mathbb{R}} x p(x \mid y) dx$

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

$$\mathbb{E}[X] = \int x p(x) dx = \int x \left( \int \underbrace{p(x, y)}_{p(x|y)\,p(y)} dy \right) dx \Rightarrow$$

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

$$\boxed{\mathbb{E}[X] = \int \mathbb{E}[X \mid y] p(y) dy}$$

# Linear Transformations

- Suppose $\boldsymbol{y} = f(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$. You can show that:

$$\mathbb{E}[\boldsymbol{y}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{x}] + \boldsymbol{b}$$
$$\text{cov}[\boldsymbol{y}] = \boldsymbol{A}\,\text{cov}[\boldsymbol{x}]\boldsymbol{A}^T$$

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

$$\mathbb{E}[y] = \boldsymbol{a}^T\mathbb{E}[\boldsymbol{x}] + b$$
$$\text{var}[y] = \boldsymbol{a}^T\,\text{cov}[\boldsymbol{x}]\boldsymbol{a}$$

# Multivariate Gaussian

- 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

$$p(\boldsymbol{x}) = \left(\frac{1}{(2\pi)^D \det \Sigma}\right)^{1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{x}_0)\right)$$

where $\boldsymbol{x}_0 \in \mathbb{R}^D, \Sigma \in \mathbb{R}^{D \times D}$ is *symmetric positive definite (covariance matrix).*

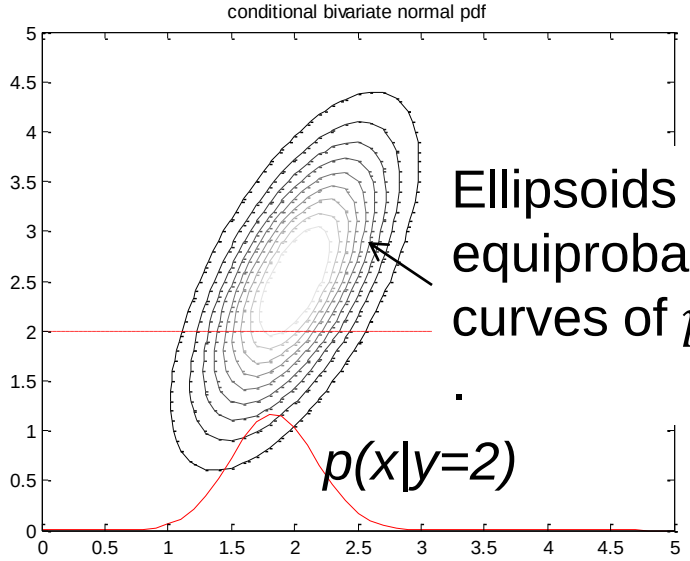
- The symmetry property of the covariance matrix does not affect the value of $(\boldsymbol{x} - \boldsymbol{x}_0)^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{x}_0)$. *However, for symmetric covariance matrices we only need to describe $D(D+1)/2$ elements rather than $D^2$.*

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

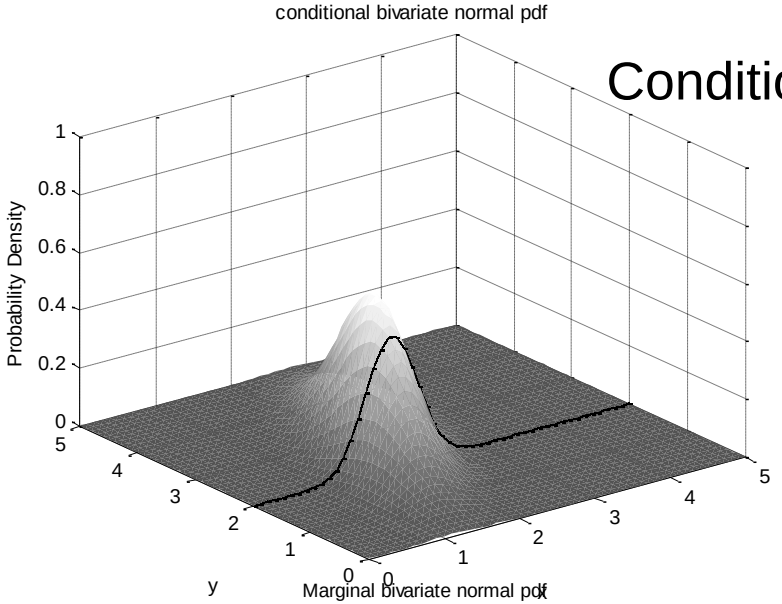
$$\boldsymbol{X}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1),\ \boldsymbol{X}_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$
$$\textit{and } \boldsymbol{X}_1, \boldsymbol{X}_2 \textit{ independent} \Rightarrow$$

$$\boldsymbol{A}\boldsymbol{X}_1 + \boldsymbol{B}\boldsymbol{X}_2 + \boldsymbol{c} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu}_1 + \boldsymbol{B}\boldsymbol{\mu}_2 + \boldsymbol{c}, \boldsymbol{A}\boldsymbol{\Sigma}_1\boldsymbol{A}^T + \boldsymbol{B}\boldsymbol{\Sigma}_2\boldsymbol{B}^T)$$

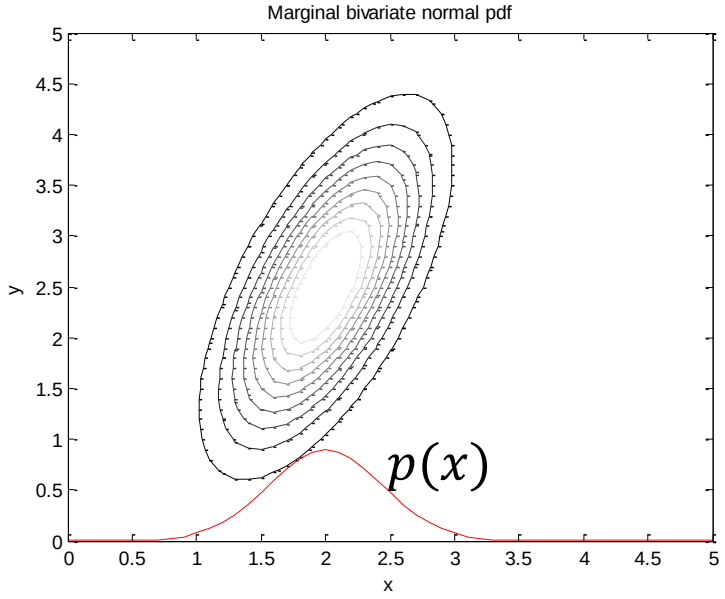
# Conditional and Marginal Probability Densities

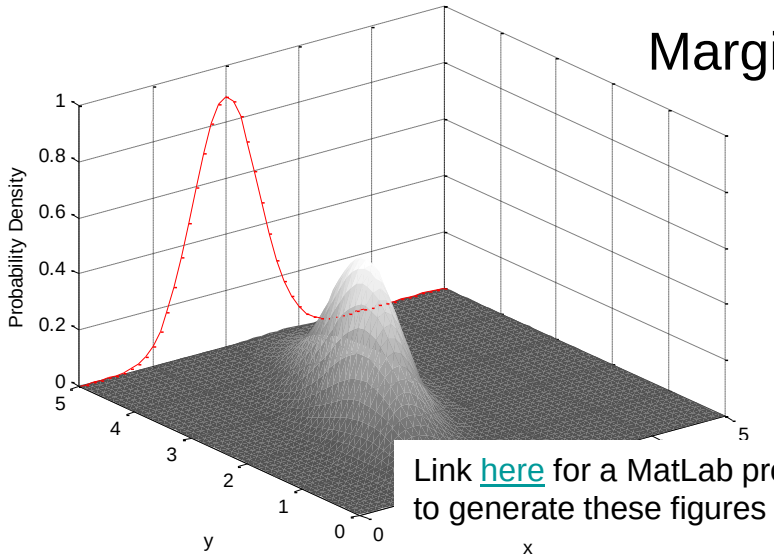



Ellipsoids : equiprobability curves of $p(x, y)$ .




Conditional




Marginal

Link here for a MatLab program to generate these figures

# Transformation of Probability Density

- A probability density transforms differently from functions.

- Let $x = g(y)$.

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

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

$$p_y(y)dy = p_x(x)dx$$

# Transformation of Probability Density

- For example consider the Gamma distribution

$$\mathcal{Gamma}(x \mid a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-xb}$$

- 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

$$\mathcal{InvGamma}(y \mid a,b) = \frac{b^a}{\Gamma(a)} y^{-(a+1)} e^{-\frac{b}{y}}$$

# Multivariate Change of Variables

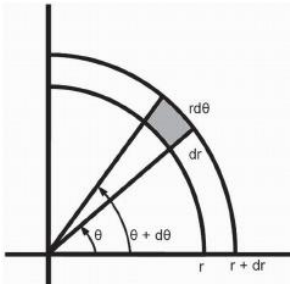
- If $f$ is an invertible mapping, we can define the pdf of the transformed variables using the Jacobian of the inverse mapping $\boldsymbol{y} \rightarrow \boldsymbol{x}$:

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

- As an example, it is trivial to show that transforming a density from Cartesian coordinates $\boldsymbol{x} = (x, y)$ to polar coordinates $\boldsymbol{y} = (r, \theta)$, where $x = r\cos\theta$ and $y = r\sin\theta, \left| \frac{\partial(x,y)}{\partial(r,\theta)} \right| = r$, gives:

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

$$p_{r,\theta}(r,\theta) dr d\theta = p_{x,y}(r\cos\theta, r\sin\theta) r dr d\theta$$

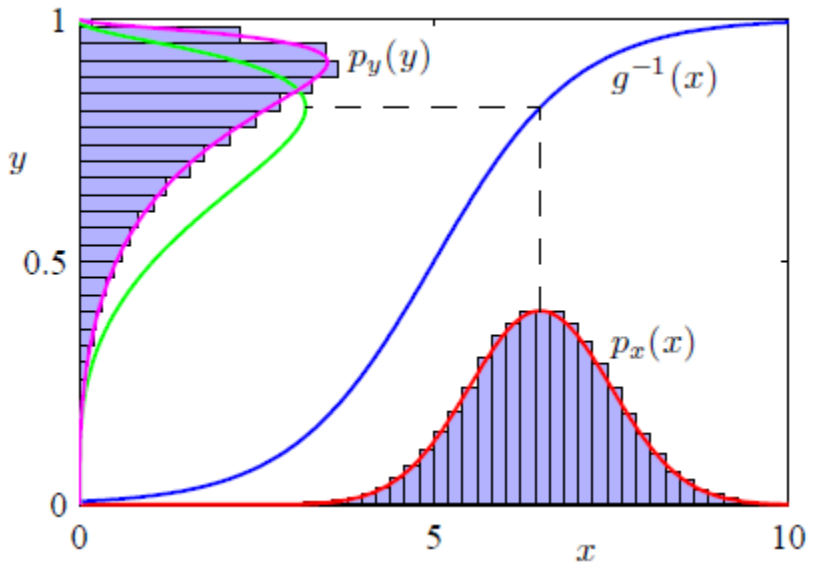


# Transformation of Probability Density

- Recall $p_y(y) = p_x(g(y))\left|\frac{dx}{dy}\right| = p_x(g(y))sg'(y),\ s \in \{-1,1\}$
- Using this Eq., note that modes of densities depend on the choice of variables (see 2nd term on the rhs below):

$$p_y'(y) = sp_x'(x)\{g'(y)\}^2 + sp_x(g(y))g''(y)$$

- 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:

$$y^{(s)} = g^{-1}(x^{(s)}),\ where: x^{(s)} \sim p_x(x)$$

# Multivariate Student's $\mathcal{T}$ Distribution

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

- If we return to the derivation of the univariate Student's $\mathcal{T}$ distribution and substitute $\upsilon = 2a, \lambda = \frac{a}{b}, \eta = \tau b / a$ , and use

$$\mathcal{Gamma}\left(\tau \mid a, b\right) = \frac{b^a}{\Gamma(a)} \tau^{a-1} e^{-b\tau}$$

we can write the Student's $\mathcal{T}$ distribution as:*

$$\mathcal{T}(x \mid \mu, \lambda, \upsilon) = \int_0^\infty \mathcal{N}\left(x \mid \mu, (\eta\lambda)^{-1}\right) \mathcal{Gamma}\left(\eta \mid \upsilon/2, \upsilon/2\right) d\eta$$

- This form is useful in *providing generalization to a multivatiate Student's $\mathcal{T}$*

$$\mathcal{T}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}, \upsilon) = \int_0^\infty \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right) \mathcal{Gamma}\left(\eta \mid \upsilon/2, \upsilon/2\right) d\eta$$

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

# Multivariate Student's $\mathcal{T}$ Distribution

$$\mathcal{T}(\boldsymbol{x}\,|\,\boldsymbol{\mu},\Lambda,\upsilon)=\int_0^{\infty}\mathcal{N}\left(\boldsymbol{x}\,|\,\boldsymbol{\mu},(\eta\Lambda)^{-1}\right)\mathcal{Gamma}\left(\eta\,|\,\upsilon/2,\upsilon/2\right)d\eta$$

- This integral can be computed analytically as:

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

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

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

$$\mathcal{T}(\boldsymbol{x}\,|\,\boldsymbol{\mu},\Lambda,\upsilon)=\frac{(\upsilon/2)^{\upsilon/2}}{\Gamma(\upsilon/2)}\frac{|\Lambda|^{1/2}}{(2\pi)^{D/2}}\int_0^{\infty}\eta^{D/2}\eta^{\upsilon/2-1}e^{-\upsilon\eta/2}e^{-\eta\Delta^2/2}d\eta \qquad \textit{Use } \tau=\eta\left(\upsilon/2+\Delta^2/2\right)$$

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

- Normalization proof is immediate from the normalization of the normal & Gamma distributions.

# Multivariate Student's $\mathcal{T}$ Distribution

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

- Some useful results of the multivatiate Student's $\mathcal{T}$ are given below:

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

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

$$\mathbb{E}[\boldsymbol{x}]=\int_{-\infty}^{+\infty}\frac{\Gamma(\frac{D}{2}+\frac{\upsilon}{2})}{\Gamma(\frac{\upsilon}{2})}\frac{|\Lambda|^{1/2}}{(\pi\upsilon)^{D/2}}\left[1+\frac{\Delta^2}{\upsilon}\right]^{-\upsilon/2-D/2}(\boldsymbol{z}+\boldsymbol{\mu})\,d\boldsymbol{z}$$

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

- The covariance is computed as:

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

$$=\Lambda^{-1}\frac{(\upsilon/2)^{\upsilon/2}}{\Gamma(\upsilon/2)}\frac{\Gamma(\upsilon/2-1)}{(\upsilon/2)^{\upsilon/2}(\upsilon/2)^{\upsilon/2-1}}=\frac{(\upsilon/2)\Gamma(\upsilon/2-1)}{\Gamma(\upsilon/2)}\Lambda^{-1}=\frac{\upsilon/2}{\upsilon/2-1}\Lambda^{-1}=\frac{\upsilon}{\upsilon-2}\Lambda^{-1}$$

# Multivariate Student's $\mathcal{T}$ Distribution

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

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

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

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

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

$$\left[1+\frac{\Delta^2}{\upsilon}\right]^{-\upsilon/2-D/2}=\exp\left(-\left(\frac{\upsilon}{2}+\frac{D}{2}\right)\ln\left[1+\frac{\Delta^2}{\upsilon}\right]\right)\overset{\upsilon\rightarrow\infty}{=}\exp\left(-\frac{\upsilon}{2}\left(\frac{\Delta^2}{\upsilon}-\frac{1}{2}\left(\frac{\Delta^2}{\upsilon}\right)^2\right)\right)=\exp\left(-\frac{\Delta^2}{2}+O\left(\upsilon^{-1}\right)\right)$$

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

# Dirichlet Distribution

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

- **It is the multivariate generalization of the $\mathcal{B}eta$ distribution.**

# Dirichlet 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:

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

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

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

- The distribution over the space of $\mu_k$ is $K - 1$ dimensional due to the last constraint above.

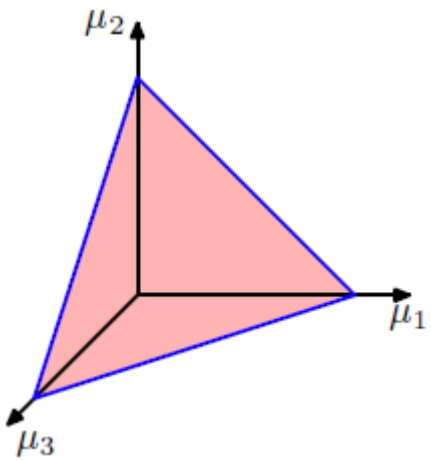
# Dirichlet Distribution

- 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

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

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*:

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




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

# Dirichlet Distribution

- We write the Dirichlet distribution as:

$$p(\boldsymbol{\mu}\,|\,\boldsymbol{\alpha}) = K(\boldsymbol{\alpha})\prod_{k=1}^{K}\mu_k^{\alpha_k-1},\; K(\boldsymbol{\alpha}) = \frac{\Gamma(a_0)}{\Gamma(a_1)\ldots\Gamma(a_K)},\; a_0 = a_1+\ldots+a_K$$

- 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 $\mathbb{E}\left[\ln\mu_j\right]$

$$\mathbb{E}\left[\ln\mu_j\right] = K(\boldsymbol{\alpha})\int_0^1\!\!..\int_0^1\ln\mu_j\prod_{k=1}^{K}\mu_k^{\alpha_k-1}d\mu_1....d\mu_K = K(\boldsymbol{\alpha})\int_0^1\!\!..\int_0^1\frac{\partial}{\partial\alpha_j}\prod_{k=1}^{K}\mu_k^{\alpha_k-1}d\mu_1....d\mu_K =$$

$$K(\boldsymbol{\alpha})\frac{\partial}{\partial\alpha_j}\int_0^1\!\!..\int_0^1\prod_{k=1}^{K}\mu_k^{\alpha_k-1}d\mu_1....d\mu_K = K(\boldsymbol{\alpha})\frac{\partial}{\partial\alpha_j}\frac{1}{K(\boldsymbol{\alpha})} = -\frac{\partial\ln K(\boldsymbol{\alpha})}{\partial\alpha_j} = \underbrace{\frac{\partial\ln\Gamma(\alpha_j)}{\partial\alpha_j}}_{\Psi(\alpha_j)} - \underbrace{\frac{\partial\ln\Gamma(\alpha_0)}{\partial\alpha_0}}_{\Psi(\alpha_0)}$$

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

$$\mathbb{E}\left[\ln\mu_j\right] = \Psi(\alpha_j) - \Psi(\alpha_0),\; \Psi(\alpha_j) = \frac{\partial\ln\Gamma(\alpha_j)}{\partial\alpha_j},\; \Psi(\alpha_0) = \frac{\partial\ln\Gamma(\alpha_0)}{\partial\alpha_0}$$

# Dirichlet Distribution: Normalization

- 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:

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

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

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

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

# Dirichlet Distribution: Normalization

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

$$= \underbrace{C_M \left(\prod_{k=1}^{M-2}\mu_k^{\alpha_k-1}\right)\left(1-\sum_{j=1}^{M-2}\mu_j\right)^{\alpha_{M-1}+\alpha_M-1}}_{Dirichlet\,(M-1)} \frac{\Gamma(\alpha_{M-1})\Gamma(\alpha_M)}{\Gamma(\alpha_{M-1}+\alpha_M)}$$

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

$$1 = C_M \frac{\Gamma(\alpha_1)...\Gamma(\alpha_{M-2})\Gamma(\alpha_{M-1}+\alpha_M)}{\Gamma(\alpha_1+...+\alpha_M)} \frac{\Gamma(\alpha_{M-1})\Gamma(\alpha_M)}{\Gamma(\alpha_{M-1}+\alpha_M)} \Rightarrow$$

$$C_M = \frac{\Gamma(\alpha_1+...+\alpha_M)}{\Gamma(\alpha_1)...\Gamma(\alpha_{M-2})\Gamma(\alpha_{M-1})\Gamma(\alpha_M)}$$

# Dirichlet Distribution

- Using the multinomial as a "likelihood" and the Dirichlet as "the conjugate prior", we arrive at the following "posterior"

$$p(\boldsymbol{\mu}\,|\,\mathcal{D}) \propto \underbrace{p(\mathcal{D}\,|\,\boldsymbol{\mu})}_{\text{Multinomial}}\; p(\boldsymbol{\mu}) \Rightarrow p(\boldsymbol{\mu}\,|\,\mathcal{D}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1}$$

  which is a Dirichlet distribution $\mathcal{D}ir(\boldsymbol{\mu}\,|\,\alpha_1 + m_1, \ldots, \alpha_K + m_K)$.

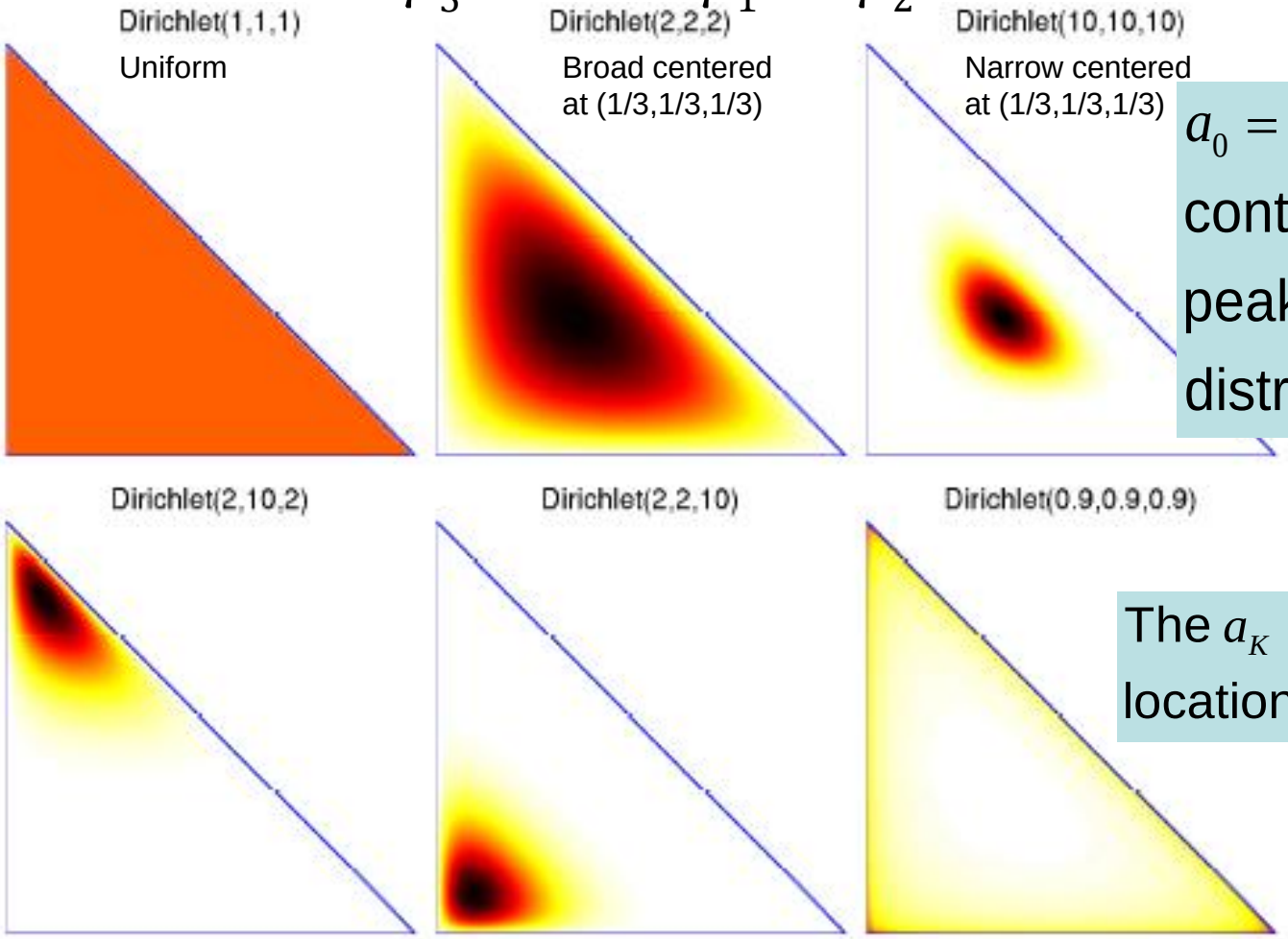
- The normalization factor is computed easily from the normalization factor of the Dirichlet as:

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

- $a_k$ can be interpreted as "the effective number of prior observations of $x_k = 1$".

# Dirichlet Distribution

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



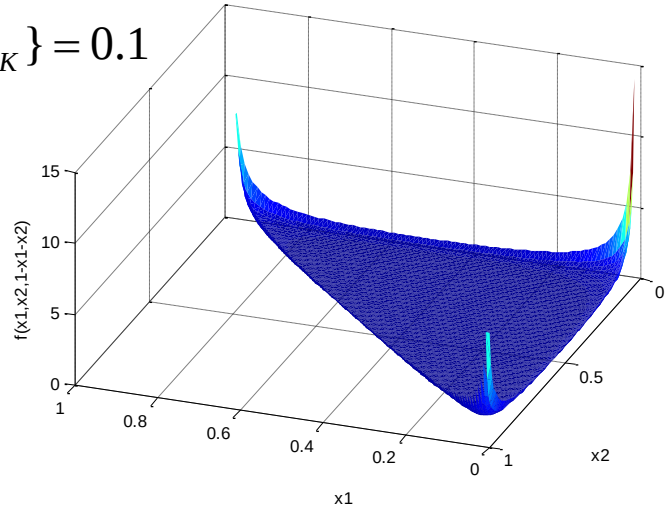

$a_0 = a_1 + \ldots + a_K$ controls how peaked the distribution is

The $a_K$ control the location of the peak

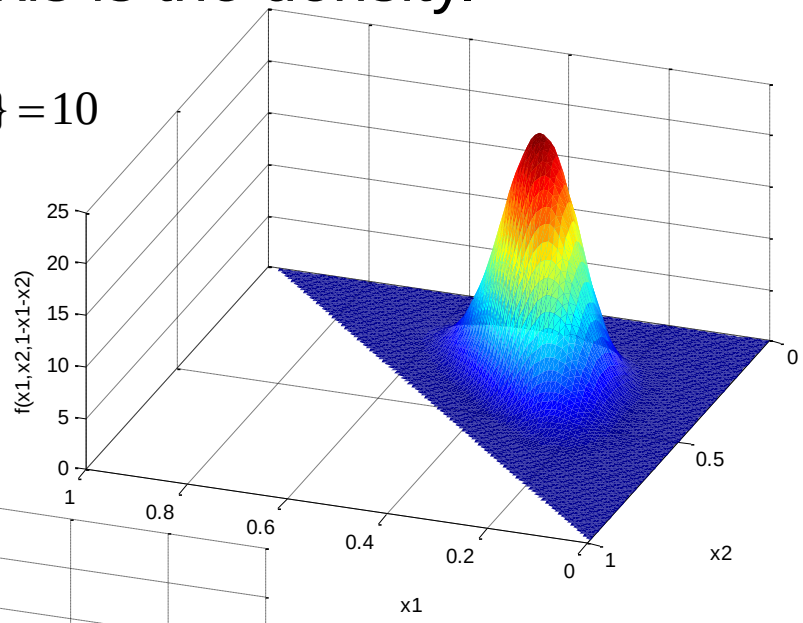
# Dirichlet Distribution

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.

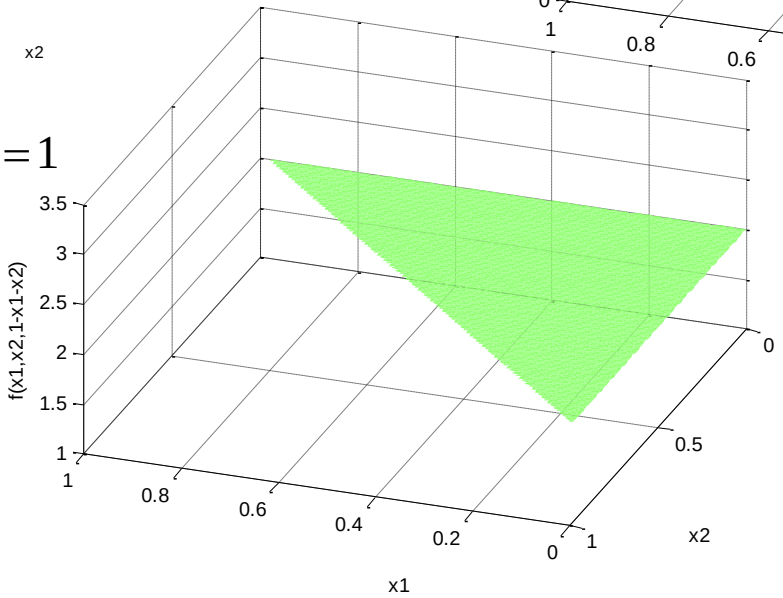
$\{\alpha_K\} = 0.1$




$\{\alpha_K\} = 10$



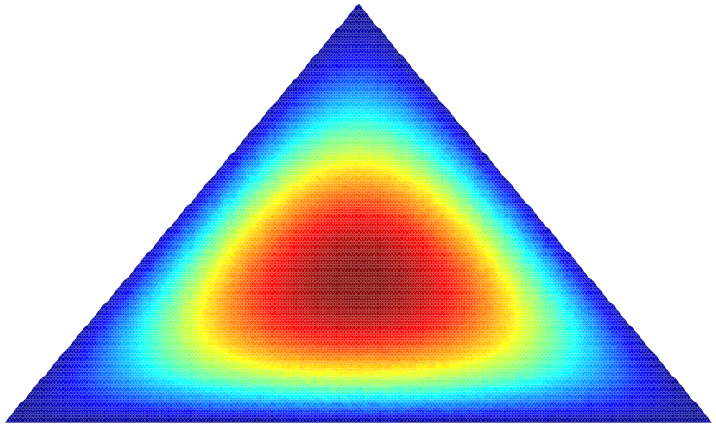

$\{\alpha_K\} = 1$



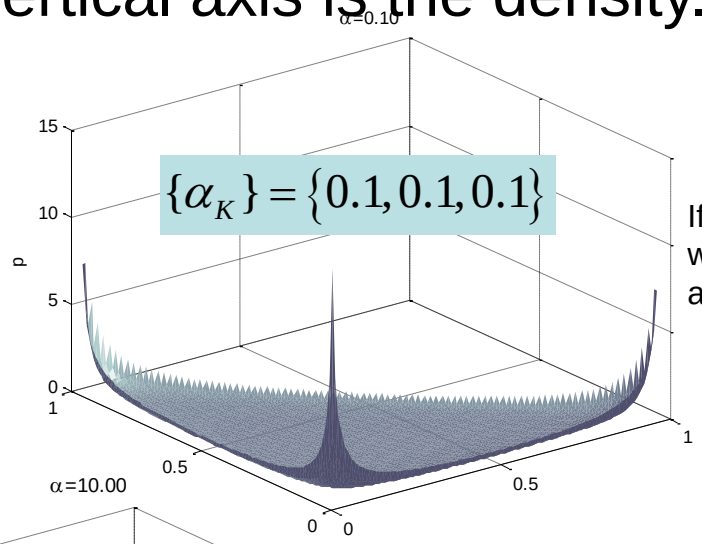

MatLab Code

# 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.

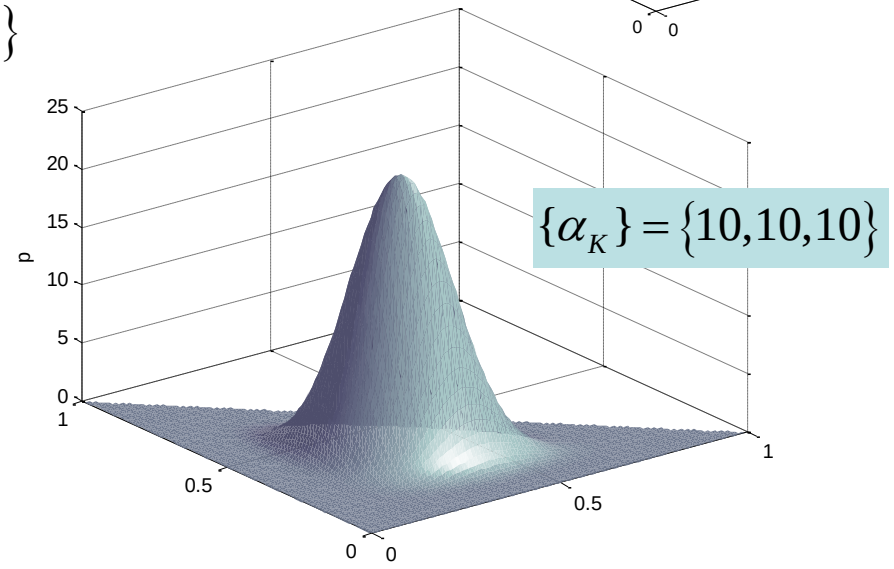


$$\{\alpha_K\} = \{2, 2, 2\}$$




$$\{\alpha_K\} = \{0.1, 0.1, 0.1\}$$

If $a_k$<1/3 for all k, we obtain spikes at the corners

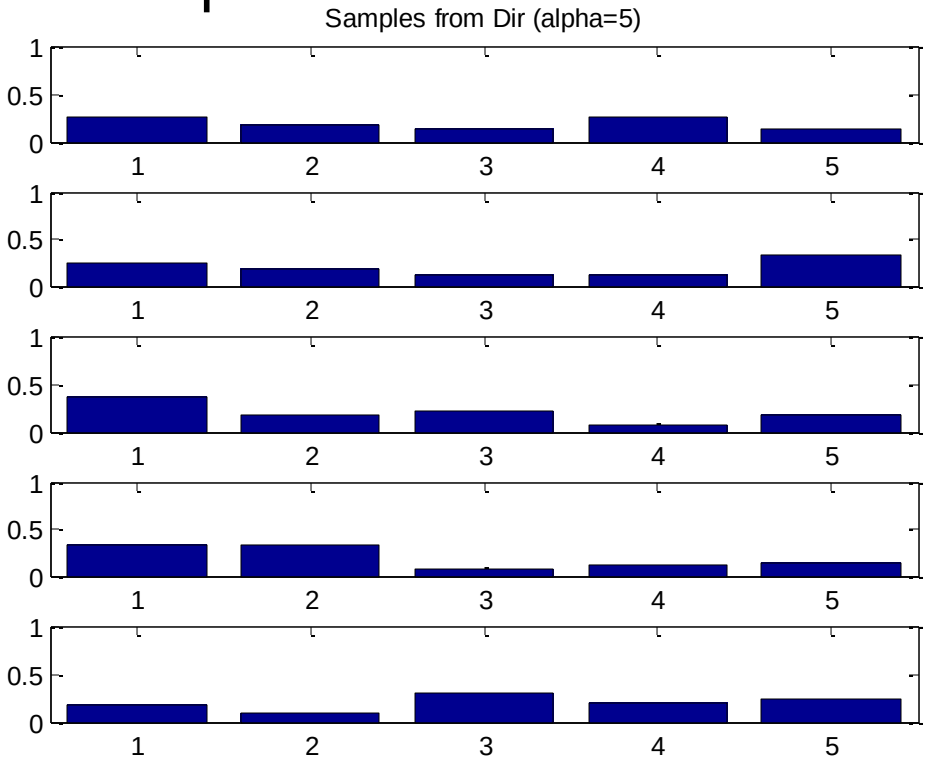



$$\{\alpha_K\} = \{10, 10, 10\}$$

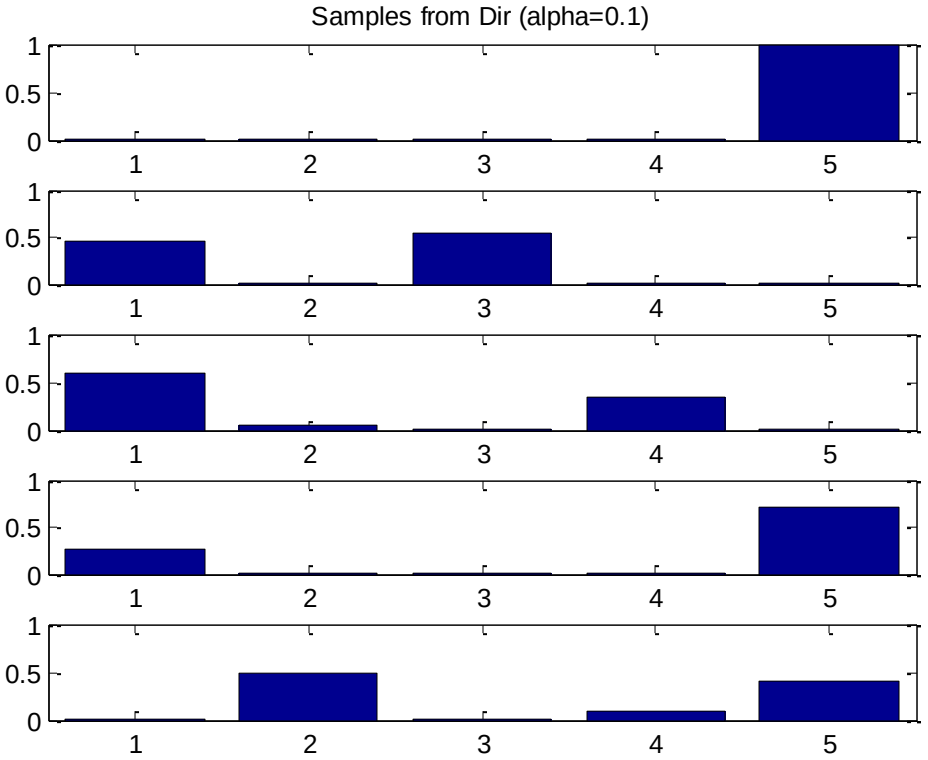
Run *visDirichletGui* & dirichlet3dPlot from PMTK

# Dirichlet Distribution

Samples from 5-dimensional symmetric Dirichlet distribution.




$$\{\alpha_K\} = \{5, 5, ..., 5\}$$




$$\{\alpha_K\} = \{0.1, 0.1, ..., 0.1\}$$

Run dirichletHistogramDemo
from PMTK

# Dirichlet Distribution

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

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

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

- Often we use:

$$\alpha_k = \alpha / K$$

In this case:

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

*Increasing $\alpha$ increases the precision of the distribution.*