Adding Discrete RVs

$\{X, Y - \text{Independent}\}$

$\{Z = X + Y\}$

ex. $X \sim \text{Bernoulli}(p_x)$ and $Y \sim \text{Bernoulli}(p_y)$.

$1 - p_x$

$0 - 1 - p_x$

$$\begin{cases} P(Z=0) = P(X=0) \cdot P(Y=0) = (1-p_x) \quad P(X=0) \cdot P(Y=0 \mid X=0) \\ P(Z=1) = P(X=0)\, P(Y=1) + P(X=1) \cdot P(Y=0) \\ P(Z=2) = P(X=1)\, P(Y=1) \end{cases}$$

Q. X, Y be general RVs with ranges $R(X)$, $R(Y)$

$$P(Z=z) = \left(\sum_{x \in R(X)} \sum_{y \in R(Y)} \{ I(x+y=z) \} P(X=x)\, P(Y=y) \right)$$

$$= \left\{ \sum_{x \in R(X)} P(X=x) \cdot P(Y=z-x) \right\}$$

What about continuous RVs ?

Discrete

pmf

$$P(Z=z) = \sum_{x \in R(X)} P(X=x)\, P(Y=z-x)$$

pmf

$\int$ pdf

$Z = f(X,Y)$

$$\int_{-\infty}^{\infty} p_X(x)\, p_Y(f(x,y)=z)\, dx$$

analogy

(say $R(X) = R(Y) = \mathbb{R}$)

Continuous

PDF

convolution

$$p_Z(z) = \left(\int_{-\infty}^{\infty} p_X(x)\, p_Y(z-x)\, dx \right)$$

Is this formula correct?

* Is this type of discrete $\longrightarrow$ continuous analogy always valid?

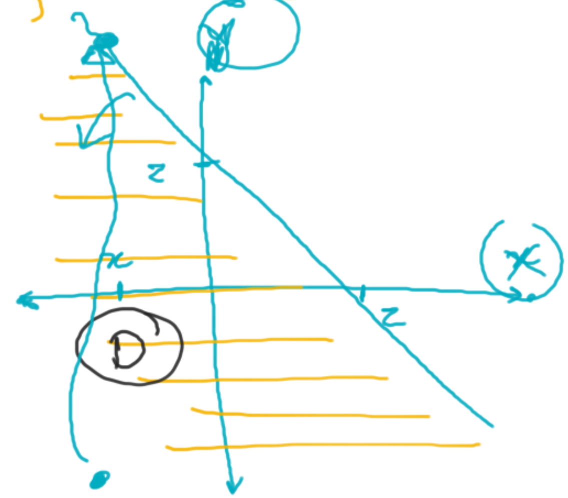
NO

1 - yes

yes / NO

Rigorous Argument for the convolution formula

{ $F_X(\cdot)$ → CDF of X and $p_X(\cdot)$ → PDF of X } → convention



$Z = X + Y$. We know $p_X(\cdot)$ & $p_Y(\cdot)$

$$F_Z(z) = P(Z \le z) = \iint_D p_X(x)\, p_Y(y)\, d(x,y)$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} p_X(x)\, p_Y(y)\, dy\, dx$$

$$F_Z(z) = \int_{-\infty}^{\infty} p_X(x) \left(\int_{-\infty}^{z-x} p_Y(y)\, dy \right) dx$$

$$= \boxed{\int_{-\infty}^{\infty} p_X(x)\, p_Y(z-x)\, dx}$$

$$p_Z(z) = \frac{d}{dz} F_Z(z)$$

$$= \int_{-\infty}^{\infty} p_X(x) \frac{d}{dz} \int_{-\infty}^{z-x} p_Y(y)\, dy\, dx$$

$$X \longrightarrow \text{Gaussian}\ (\mu_X, \sigma_X^2)$$

$$Y \qquad\qquad\qquad\qquad \mu_Y, \sigma_Y^2$$

What if $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$?

Z

$$p_Z(z) = \left\{ \int_{-\infty}^{\infty} p_X(x)\, p_Y(z-x)\, dx \right\}$$

$z - x - \mu_Y$

$$= \eta \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\mu_X)^2}{2\sigma_X^2} - \frac{((z-x)-\mu_Y)^2}{2\sigma_Y^2} \right) dx$$

$$= \eta \int_{-\infty}^{\infty} \exp\left[-\left(ax^2 + bz^2 + 2cx + 2dz + 2exz \right) \right] dx$$

$$= \eta\, e^{(-bz^2 - 2dz)} \int_{-\infty}^{\infty} \exp\left(-a\left(\underbrace{x^2 + 2x(cz+d)}_{+(cz+d)^2}\right) \right) dx$$

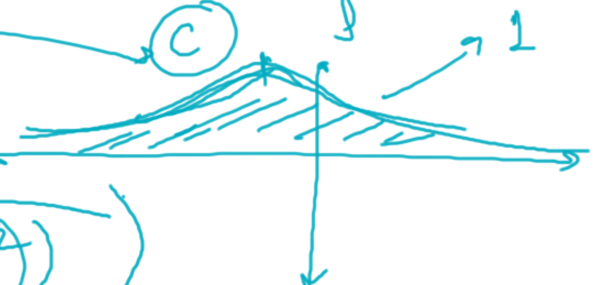
$+a(cz+d)^2$

$$\left\{ = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sqrt{2\pi}\sigma_X\sigma_Y} \exp\left[-\frac{x^2(\sigma_X^2+\sigma_Y^2) - 2x(\sigma_X^2(z-\mu_Y)+\sigma_Y^2\mu_X) + \sigma_X^2(z^2+\mu_Y^2-2z\mu_Y)+\sigma_Y^2\mu_X^2}{2\sigma_Y^2\sigma_X^2}\right] dx \right\}$$

η_z

$$= \eta \cdot \exp(-az^2 - bz) \cdot \left\{ \int_{-\infty}^{\infty} \exp\left(-c\,(x - ez - f)^2\right) dx \right\}$$

η'



$$= \eta' \exp(-az^2 - bz + \dots)$$

$$\mathcal{N}(\mu_z, \sigma_z^2)$$

Gaussian

So, $Z \sim N(\mu_z, \sigma_z^2)$ for some μ_z, σ_z

$$\underline{\underline{\mu_z}} = E(Z) = E(X+Y) = E(X) + E(Y) = \mu_x + \mu_z$$

$$\underline{\underline{\sigma_z^2}} = Var(Z) = Var(X+Y) = \sigma_x^2 + \sigma_y^2$$

$$Var(X) + Var(Y) + \cancel{2cov}(X,Y)$$

$\underline{\underline{23^{rd}}}$

What if $Z = Y - X$? $\;\rightarrow\; \{ Y = Z + X \} \longrightarrow Z \sim$

$$p_Z(z) = \int_{-\infty}^{\infty} p_X(x)\, p_Y(z+x)\, dx$$

$$= \left\{ \eta \int_{-\infty}^{\infty} \exp\left(- \frac{(x-\mu_X)^2}{2\sigma_X^2} - \frac{(x+z-\mu_Y)^2}{2\sigma_Y^2} \right) dx \right\}$$

change variable to $(-x)$

$$-\mu_Y + \mu_X = \eta \int_{-\infty}^{\infty} \exp\left(- \frac{(x-(-\mu_X))^2}{2\sigma_X^2} - \frac{(-x+z-\mu_Z)^2}{2\sigma_Y^2} \right) dx$$

$z = Y - X$

$$\sim \left\{ N\left(\mu_Y - \mu_X,\ \sigma_X^2 + \sigma_Y^2 \right) \right\} \longrightarrow$$

* What about product?

$1 - \qquad 2 \sim$

ex. $X \sim 1 + \text{Bernoulli}(p_X)$ $\qquad$ $Y \sim 1 + \text{Bernoulli}(p_Y)$ $\quad$ & $\quad$ $Z = XY$

$$\begin{cases} P(Z=1) = \\ P(Z=2) = \\ P(Z=4) = \end{cases}$$

General:

$$\boxed{P(Z=z) = \sum_{x \in R(X)} P(X=x)\, P\left(Y = \frac{z}{x}\right)}$$

Intuitive $p_Z(z)$ when X, Y are continuous in $(0, \infty)$?

$$p_Z(z) = \int_0^\infty p_X(x)\, p_Y(z/x)\, dx$$

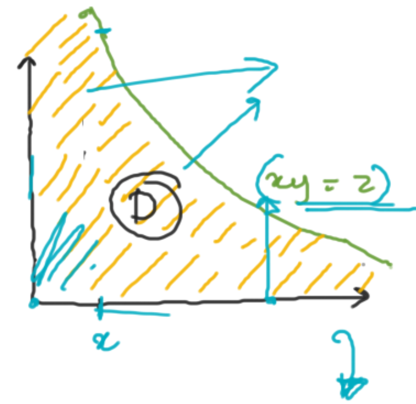
f'

$$= \int_0^\infty \frac{p_X(x)\, p_Y(y)}{\left(\left| \frac{\partial f}{\partial y}(x,y) \right| \right)}\, dx$$

Intuitively!

Go by CDF method

$(0,\infty)$ $Z = XY$



$$F_Z(z) = P(Z \leq z) = \iint_D p_X(x)\, p_Y(y)\, d(x,y)$$

$$= \int_{0^+}^{\infty} \int_{0^+}^{z/x} p_X(x)\, p_Y(y)\, dy\, dx$$

$$F_Z(z) = \int_0^{\infty} p_X(x) \left[\frac{d}{dz} \int_0^{z/x} p_Y(y)\, dy \right] dx$$

$$\frac{d}{dz}(z/x)$$

$$f_Z(z) = \frac{d}{dz} F_Z(z) = \int_0^{\infty} \frac{p_X(x)\, p_Y(z/x)}{x}\, dx$$

$$\frac{\partial f}{\partial y} = \frac{\partial xy}{\partial y}$$

x

WHY DID THE ANALOGY FAIL ?

$$\boxed{p_X(x) \neq P(X = x)}$$

PDF $\not\longleftrightarrow$ PMF

CS-215 Tutorial

Harshit Varma

IIT Bombay

August 24, 2021

Binomial Distribution

- Notation: $B(n, p)$
- Parameters:
  - $n \in \mathbb{N}$: number of independent trials
  - $p \in [0, 1]$: probability of success in each Bernoulli trial
- PMF: $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$
- Discrete probability distribution of the number of successes observed in a sequence of Bernoulli trials
- Thus, a binomial random variable X can be written as the sum of n iid Bernoulli random variables with parameter p

$$X = X_1 + \ldots + X_n \quad \left(\{X_i \sim \text{Bernoulli}(p)\}_{i=1}^{n} \text{ are iid} \right)$$

Sum of Binomials

Consider $Z = X + Y$ where $X \sim B(n_1, p)$ and $Y \sim B(n_2, p)$ are independent.
Thus,

$$P(X = k) = \binom{n_1}{k} p^k (1-p)^{n_1 - k}$$
$$P(Y = k) = \binom{n_2}{k} p^k (1-p)^{n_2 - k}$$

$P(Z = k) = \sum_{i=0}^{k} P(X = i, Y = k - i)$

Sum of Binomials

$$P(Z = k) = \sum_{i=0}^{k} P(X = i, Y = k - i)$$

$$P(Z = k) = \sum_{i=0}^{k} P(X = i) \cdot P(Y = k - i) \quad (X, Y \text{ are independent})$$

$$P(Z = k) = \sum_{i=0}^{k} \binom{n_1}{i} p^i (1 - p)^{n_1 - i} \cdot \binom{n_2}{k - i} p^{k-i} (1 - p)^{n_2 - k + i}$$

$$P(Z = k) = p^k (1 - p)^{n_1 + n_2 - k} \sum_{i=0}^{k} \binom{n_1}{i} \cdot \binom{n_2}{k - i}$$

Sum of Binomials

$$\sum_{i=0}^{k} \binom{n_1}{i} \cdot \binom{n_2}{k-i} = \binom{n_1+n_2}{k} \quad \text{(Vandermonde's Identity)}$$

Proof:

Number of ways of choosing k items from a collection of n_1 items of type A and n_2 items of type B

Sum of Binomials

$$
\begin{aligned}
P(Z = k) &= p^k (1-p)^{n_1+n_2-k} \sum_{i=0}^{k} \binom{n_1}{i} \cdot \binom{n_2}{k-i} \\
&= \binom{n_1+n_2}{k} p^k (1-p)^{n_1+n_2-k}
\end{aligned}
$$

Thus, $Z \sim B(n_1 + n_2, p)$

Sum of Binomials: Another Proof

As $X \sim B(n_1, p)$, $X = \beta_1 + \beta_2 + \ldots + \beta_{n_1 - 1} + \beta_{n_1}$ where $\{\beta_i\}_{i=1}^{n_1}$ are n_1 iid Bernoulli random variables with parameter p

Similarly, $Y = \alpha_1 + \alpha_2 + \ldots + \alpha_{n_2 - 1} + \alpha_{n_2}$ where $\{\alpha_j\}_{j=1}^{n_2}$ are n_2 iid Bernoulli random variables with parameter p

Thus, $Z = X + Y = \beta_1 + \ldots + \beta_{n_1} + \alpha_1 + \ldots + \alpha_{n_2}$

As X, Y are independent, β_i and α_j are independent for all i, j
They are also identically distributed

Thus, $\beta_1, \ldots, \beta_{n_1}, \alpha_1, \ldots, \alpha_{n_2}$ is a sequence of $n_1 + n_2$ iid Bernoulli random variables with parameter p

Thus, $Z \sim B(n_1 + n_2, p)$

Sum of Binomials

What happens when $X \sim B(n_1, p_1)$ and $Y \sim B(n_2, p_2)$ are independent, but $p_1 \neq p_2$?

$$
\begin{aligned}
P(Z = k) &= \sum_{i=0}^{k} \binom{n_1}{i} p_1^i (1 - p_1)^{n_1 - i} \cdot \binom{n_2}{k-i} p_2^{k-i} (1 - p_2)^{n_2 - k + i} \\
&= p_2^k (1 - p_1)^{n_1} (1 - p_2)^{n_2 - k} \sum_{i=0}^{k} \binom{n_1}{i} \binom{n_2}{k-i} \left(\frac{p_1}{p_2} \cdot \frac{1 - p_2}{1 - p_1} \right)^i \\
&= p_2^k (1 - p_1)^{n_1} (1 - p_2)^{n_2 - k} \sum_{i=0}^{k} \binom{n_1}{i} \binom{n_2}{k-i} \gamma^i
\end{aligned}
$$

In this case, Z is not binomially distributed

But we can show that $Var(Z) \leq Var(W)$, where $W \sim B\left(n_1 + n_2, \frac{p_1 + p_2}{2}\right)$ (Binomial sum variance inequality)

Summary

- If $X \sim B(n_1, p)$ and $Y \sim B(n_2, p)$ are independent, then $Z = X + Y \sim B(n_1 + n_2, p)$
- Doesn't hold when:
  - X, Y are dependent
  - $X \sim B(n_1, p_1)$, $Y \sim B(n_2, p_2)$ and $p_1 \neq p_2$

Multivariate Gaussians

Sums and Conditionals

Dhruv Arora

October 12, 2021

Topics Covered

- Sum of multivariate Gaussian random variables
- Conditional probability with multivariate Gaussians

Sum of Gaussian RVs

Question

Given two random variables X, Y *with distributions* $X \sim \mathcal{N}(\mu_X, \Sigma_X)$ *and* $Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$*.*
Can you determine the distribution of $Z = X + Y$*?*

No!

What do you need ?

Relationship between X, Y

Sum of independent Gaussian RVs

Question

Given two independent random variables X, Y with distributions $X \sim \mathcal{N}(\mu_X, \Sigma_X)$ and $Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$. Can you determine the distribution of $Z = X + Y$?

Yes!

How ?

Sum of independent Gaussian RVs

Recall from the tutorial on univariate Gaussians

$$f_Z(z) = \int_{\mathbb{R}} f_X(x) f_Y(z - x) dx$$

This can be extended to multivariate situations as :

$$f_Z(z) = \int_{\mathbb{R}^d} f_X(x) f_Y(z - x) dx$$

if $X, Y \in \mathbb{R}^d$

Sum of independent Gaussian RVs

Do you need so much calculation though? Or is there a simpler way ?

Can you atleast get the mean and variance of Z without any calculation?

Ofcourse!

$$\mu_Z = E[Z] = E[X] + E[Y] = \mu_X + \mu_Y$$

$$Cov(Z_i, Z_j) = Cov(X_i + Y_i, X_j + Y_j) = Cov(X_i, X_j) + Cov(Y_i, Y_j) \text{ [why?]}$$

And thus, elementwise,

$$\Sigma_Z = \Sigma_X + \Sigma_Y$$

Sum of independent Gaussian RVs

But how do you argue that Z is Gaussian?

By definition!

Recall from lectures that X is said to be distributed according to a multivariate Gaussian distribution if

$$X = A_X w + \mu_X$$

where $A \in \mathbb{R}^{d \times n}(n \geq d)$ and each w_i is i.i.d standard normal RV.

Sum of independent Gaussian RVs

So, $X = A_X w_1 + \mu_X$ and $Y = A_Y w_2 + \mu_Y$ ($A_X \in \mathbb{R}^{d \times n_1}$ and $A_Y \in \mathbb{R}^{d \times n_2}$) where all elements of $w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ are iid standard normals.

$$Z = A_X w_1 + A_Y w_2 + (\mu_X + \mu_Y)$$

Define $A_Z = \begin{bmatrix} A_X & A_Y \end{bmatrix}$ and convince yourself that

$$Z = A_Z w + \mu_Z$$

And since both $n_1, n_2 \geq d$, $n_1 + n_2 \geq d$.

Conditionals on multivariate Gaussians

Question

Let $X \sim \mathcal{N}(\mu, \Sigma)$. If I sample X and tell you the values of some of it's dimensions, can you get a probability distribution on the rest?

w.l.o.g assume I give you the last few entries (why w.l.o.g?)

i.e.

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

and I'm telling you $X_2 = \alpha$

Conditionals on multivariate Gaussians

You need $P(X_1 = x|X_2 = \alpha)$. Which is

$$P(X_1 = x|X_2 = \alpha) = \frac{P(X_1 = x, X_2 = \alpha)}{P(X_2 = \alpha)}$$

The $P(X_2 = \alpha)$ is called marginal distribution (covered in next part by Harshit)

But for our purposes, it is a constant independent of x.

And

$$P(X_1 = x, X_2 = \alpha) = P(X = \begin{bmatrix} x \\ \alpha \end{bmatrix})$$

Conditionals on multivariate Gaussians

Which we know...

$$P(X = \begin{bmatrix} x \\ \alpha \end{bmatrix}) \propto \exp - \frac{\begin{bmatrix} x - \mu_1 \\ \alpha - \mu_2 \end{bmatrix}^T \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} x - \mu_1 \\ \alpha - \mu_2 \end{bmatrix}}{2}$$

For now, call

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}$$

we will deal with getting Λ later.

Conditionals on multivariate Gaussians

Can you already see that this is a Gaussian distribution?

Focus only on the exponent! Replace $x - \mu_1$ by y and $\alpha - \mu_2$ by β for good measure.

Convince yourself that

$$\begin{bmatrix} y \\ \beta \end{bmatrix}^T \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix} \begin{bmatrix} y \\ \beta \end{bmatrix} = y^T \Lambda_{11} y + y^T \Lambda_{12} \beta + \beta^T \Lambda_{21} y + \beta^T \Lambda_{22} \beta$$

Which is the same as

$$y^T \Lambda_{11} y + y^T (\Lambda_{12} + \Lambda_{21}^T) \beta + C$$

Conditionals on multivariate Gaussians

By coefficient comparison in

$$y^T\Lambda_{11}y + y^T(\Lambda_{12} + \Lambda_{21}^T)\beta + C = (y - \mu_*)^T\Sigma_*^{-1}(y - \mu_*)$$

$$\Sigma_*^{-1} = \Lambda_{11}$$

$$-2\Sigma_*^{-1}\mu_* = (\Lambda_{12} + \Lambda_{21}^T)\beta$$

Conditionals on multivariate Gaussians

In conclusion,

$$X_1 - \mu_1 = Y \sim \mathcal{N}(\mu_*, \Sigma_*)$$

And therefore,

$$X_1 \sim \mathcal{N}(\mu_1 + \mu_*, \Sigma_*)$$

where

$$\Sigma_* = \Lambda_{11}^{-1}$$

and

$$\mu_* = -\frac{\Sigma_*(\Lambda_{12} + \Lambda_{21}^T)(\alpha - \mu_2)}{2}$$

Getting Λ

Just for completeness (i.e., don't memorize this), this is how you get Λ :

$$M^{-1} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (M/D)^{-1} & 0 \\ 0 & (M/A)^{-1} \end{bmatrix} \begin{bmatrix} I & -BD^{-1} \\ -CA^{-1} & I \end{bmatrix}$$

where

$$M/A = D - CA^{-1}B \text{ and } M/D = A - BD^{-1}C$$

CS-215 Tutorial

Harshit Varma

IIT Bombay

October 12, 2021

Multivariate Gaussian

- Notation: $\mathcal{N}(\mu, C)$
- Parameters:
  - $\mu \in \mathbb{R}^d$: mean
  - $C \in \mathbb{R}^{d\times d}$: covariance matrix
- PDF: $p(x) = \frac{1}{\sqrt{(2\pi)^d \det(C)}} \exp\left(-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)\right)$

Product of Gaussian Densities

Given $X_1 \sim \mathcal{N}(\mu_1, C_1)$ and $X_2 \sim \mathcal{N}(\mu_2, C_2)$

$$p_1(x) = \frac{1}{\sqrt{(2\pi)^d \det(C_1)}} \exp\left(-\frac{1}{2}(x - \mu_1)^T C_1^{-1}(x - \mu_1)\right)$$
$$p_2(x) = \frac{1}{\sqrt{(2\pi)^d \det(C_2)}} \exp\left(-\frac{1}{2}(x - \mu_2)^T C_2^{-1}(x - \mu_2)\right)$$

What will be the PDF of the random variable associated with $p(x) \propto p_1(x)p_2(x) = kp_1(x)p_2(x)$? ($k$ is the normalizing constant)

Product of Gaussian Densities

For univariate case?

$$\frac{1}{\sigma^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}$$

$$\mu = \sigma^2 \left(\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2} \right)$$

(Seen in Quiz-1)

Product of Gaussian Densities

For multivariate

$$
\begin{aligned}
\log p(x) &\propto \log p_1(x) + \log p_2(x) \\
\log p_1(x) &= -\frac{1}{2}\log\left((2\pi)^d \det(C_1)\right) - \frac{1}{2}(x-\mu_1)^T C_1^{-1}(x-\mu_1) \\
&\propto -\frac{1}{2}(x-\mu_1)^T C_1^{-1}(x-\mu_1) \\
&\propto -\frac{1}{2}\left(x^T C_1^{-1} x - \mu_1^T C_1^{-1} x - x^T C_1^{-1}\mu_1 + \mu_1^T C_1^{-1}\mu_1\right) \\
&\propto -\frac{1}{2}\left(x^T C_1^{-1} x - \mu_1^T C_1^{-1} x - x^T C_1^{-1}\mu_1\right) \\
&\propto -\frac{1}{2}\left(x^T C_1^{-1} x - 2\mu_1^T C_1^{-1} x\right)
\end{aligned}
$$

Note that $\mu_1^T C_1^{-1} x$ is a scalar and C_1 is symmetric, thus

$$\mu_1^T C_1^{-1} x = (\mu_1^T C_1^{-1} x)^T = x^T (C_1^{-1})^T \mu_1 = x^T C_1^{-1} \mu_1$$

Product of Gaussian Densities

$$\begin{aligned}\log p(x) &\propto \log p_1(x) + \log p_2(x)\\ &\propto -\frac{1}{2}\left(x^T(C_1^{-1}+C_2^{-1})x - 2(\mu_1^T C_1^{-1} + \mu_2^T C_2^{-1})x\right)\\ &\propto -\frac{1}{2}\left(x^T(C_1^{-1}+C_2^{-1})x - 2(\mu_1^T C_1^{-1} + \mu_2^T C_2^{-1})I_d x\right)\end{aligned}$$

Now, as C_1, C_2 are positive definite (pd), C_1^{-1}, C_2^{-1} and $C_1^{-1} + C_2^{-1}$ are also pd, and thus invertible

$$C^{-1} = C_1^{-1} + C_2^{-1}$$

$$CC^{-1} = C^T C^{-1} = I_d$$

$$\begin{aligned}\log p(x) &\propto -\frac{1}{2}\left(x^T C^{-1} x - 2(\mu_1^T C_1^{-1} + \mu_2^T C_2^{-1})(C^T C^{-1})x\right)\\ &\propto -\frac{1}{2}\left(x^T C^{-1} x - 2(C(C_1^{-1}\mu_1 + C_2^{-1}\mu_2))^T C^{-1} x\right)\\ &\propto -\frac{1}{2}\left(x^T C^{-1} x - 2\mu^T C^{-1} x\right)\end{aligned}$$

Product of Gaussian Densities

Thus, $p(x)$ corresponds to a Gaussian with mean μ and covariance C

$$
\begin{aligned}
C^{-1} &= C_1^{-1} + C_2^{-1} \\
\mu &= C(C_1^{-1}\mu_1 + C_2^{-1}\mu_2)
\end{aligned}
$$

(Similar in form to the univariate case)

Marginals

If $X = [X_1, \ldots, X_d]^T \sim \mathcal{N}(\mu, C)$, then what will be the distribution of any subset Y of $\{X_i\}_{i=1}^{d}$?

Let the size of the subset be s
Let v be a list of indices of size s such that $X_{v_i} \in Y \;\; \forall i \in [1, \ldots, s]$
Let $B \in \{0, 1\}^{s \times d}$ be a selection matrix, with $B_{i,v_i} = 1 \;\; \forall i \in [1, \ldots, s]$ and all other entries 0.
Then, $Y = BX$

Marginals

Example:
Let $X = [X_1, X_2, X_3, X_4]^T \sim \mathcal{N}(\mu, C)$ and $Y = [X_1, X_3]^T$
$d = 4, s = 2$ and $v = [1, 3]$
Thus, $B_{1,1} = 1, B_{2,3} = 1$

$$B = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Note that $Y = BX$

Marginals

Recall that $X \sim \mathcal{N}(\mu, C) \iff \exists\ \mu \in \mathbb{R}^d, A \in \mathbb{R}^{d \times m}$ $(d \leq m)$ such that $X = AW + \mu$, where $C = AA^T$ and W is a random vector of length m with all it's components being iid $\sim \mathcal{N}(0, 1)$.

Thus, $Y = BX = (BA)W + (B\mu)$
As $BA \in \mathbb{R}^{s \times m}$ and $B\mu \in \mathbb{R}^s$,
Y is also a multivariate Gaussian random variable with mean $B\mu$ and covariance $(BA)(BA)^T = BAA^TB^T = BCB^T$

$B\mu$ just contains the the relevant elements of μ
B selects the relevant rows from C and B^T selects the relevant columns

Example (contd.)

Let $\mu = [\mu_1, \mu_2, \mu_3, \mu_4]^T$ and $C = \begin{bmatrix} c_{11} & c_{12} & c_{13} & c_{14} \\ c_{21} & c_{22} & c_{23} & c_{24} \\ c_{31} & c_{32} & c_{33} & c_{34} \\ c_{41} & c_{42} & c_{43} & c_{44} \end{bmatrix}$

Then, $B\mu = [\mu_1, \mu_3]^T$

$$BC = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} c_{11} & c_{12} & c_{13} & c_{14} \\ c_{21} & c_{22} & c_{23} & c_{24} \\ c_{31} & c_{32} & c_{33} & c_{34} \\ c_{41} & c_{42} & c_{43} & c_{44} \end{bmatrix} \quad \text{(selects the relevant rows)}$$

$$= \begin{bmatrix} c_{11} & c_{12} & c_{13} & c_{14} \\ c_{31} & c_{32} & c_{33} & c_{34} \end{bmatrix}$$

Example (contd.)

$$BCB^T = \begin{bmatrix} c_{11} & c_{12} & c_{13} & c_{14} \\ c_{31} & c_{32} & c_{33} & c_{34} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{(selects the relevant columns)}$$

$$= \begin{bmatrix} c_{11} & c_{13} \\ c_{31} & c_{33} \end{bmatrix}$$

Marginals

Note that whenever C is pd, $D = BCB^T = (BA)(BA)^T$ is also pd

Proof that D is psd
Check that BCB^T is symmetric.
Now, consider any $v \neq 0 \in \mathbb{R}^s$
$v^T Dv = v^T(BA)(BA)^T v = ((BA)^T v)^T (BA)^T v = ||(BA)^T v||^2 \geq 0$
By definition, D is psd.
Now, to show D is pd, only need to show that D is invertible.

Proof that D is invertible
As all rows of B are independent and $s < d$, B has full row rank.
As $C = AA^T$ is invertible and $d \leq m$, A also has full row rank.
Thus, $s < m$ and BA has full row rank $\implies (BA)(BA)^T$ is invertible.

Thus, D is pd

In the previous slide, we have used the following lemmas
(a) For $A \in \mathbb{R}^{d \times m}$, AA^T is invertible $\iff$ A has full row rank
(b) Given A, B have full row rank, BA will also have full row rank

Proof (a)

($\Rightarrow$) Let A not have a full row rank $\implies$ A^T doesn't have full column rank, so for a non-zero $x \in \mathbb{R}^d$, $A^T x = 0$. But if this is the case, $AA^T x = 0 \implies AA^T$ is not invertible, a contradiction.

($\Leftarrow$) Let AA^T not be invertible, i.e., for a non-zero x, $AA^T x = 0 \implies x^T AA^T x = 0 \implies ||A^T x||^2 = 0 \implies A^T x = 0 \implies A^T$ doesn't have full column rank, and thus A doesn't have a full row rank, a contradiction.

Proof (b)

Try it yourself

CS-215 Tutorial

Harshit Varma

IIT Bombay

October 26, 2021

Categorical Distribution

- Generalization of the Bernoulli distribution to multiple categories, thus also called the 'Multinoulli' distribution
- Notation: $X \sim \text{Cat}\left(K, \{p_k\}_{k=1}^{K}\right)$, X is a K-dim. random vector
- Parameters:
  - $K > 0$: number of categories
  - $p_k \geq 0$: probability of the k^{th} category
  - $\sum_{k=1}^{K} p_k = 1$
- X_k models whether category k was observed or not
- Support: $\boldsymbol{x} \in \{0,1\}^K$ such that $\sum_{k=1}^{K} x_k = 1$ (i.e., the set of 'one-hot' encoded categories)
- PMF:

$$P(X = \boldsymbol{x}) = \prod_{k} p_k^{x_k}$$

- Example: outcome of tossing a dice

Categorical Distribution

Properties

- For all k, X_k can be interpreted as a Bernoulli RV with parameter p_k
- $E[X_k] = p_k, Var(X_k) = p_k(1 - p_k)$
- $Cov(X_i, X_j) = -p_i p_j$ for $i \neq j$
- For $K = 2$, $\text{Cat}\left(2, \{p, 1 - p\}\right)$ gives the Bernoulli distribution

Multinomial Distribution

- Generalizes the Binomial distribution to multiple categories
- Notation: $X \sim \text{Mult}\left(N, K, \{p_k\}_{k=1}^{K}\right)$, X is a K-dim. random vector
- Parameters:
  - $N > 0$: number of independent trials
  - $K > 0$: number of categories
  - $p_k \geq 0$: probability of the k^{th} category, $\sum_{k=1}^{K} p_k = 1$
- X_k models the number of times category k is observed in the N trials
- Support: $\boldsymbol{x} \in \mathbb{Z}_{\geq 0}^{K}$ such that $\sum_{k=1}^{K} x_k = N$
- PMF:

$$P(X = \boldsymbol{x}) = \frac{N!}{\prod_k (x_k!)} \cdot \prod_k p_k^{x_k} = \frac{\Gamma(N+1)}{\prod_k \Gamma(x_k+1)} \cdot \prod_k p_k^{x_k}$$

- Example: number of times each side of a dice appears in N throws

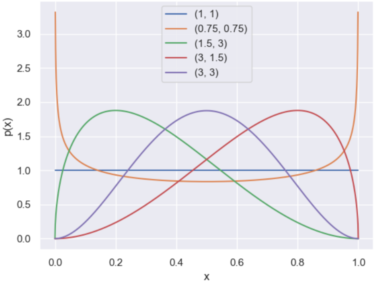
Multinomial Distribution

Properties

- Each trial has a Categorical distribution, thus a Multinomial RV can be written as a sum of N iid Categorical RVs $\sim \text{Cat}(K, \{p_k\}_{k=1}^{K})$
- For all k, X_k can be written as $\sum_{n=1}^{N} \beta_{k,n}$ where $\{\beta_{k,n}\}_{n=1}^{N}$ are iid Bernoulli random variables with parameter p_k, thus $X_k \sim \text{Bin}(N, p_k)$
- $E[X_k] = np_k, Var(X_k) = np_k(1 - p_k)$
- $Cov(X_i, X_j) = -np_ip_j$ for $i \neq j$ (trials are independent, X_i, X_j aren't)
- For $K = 2$, $\text{Mult}\left(N, 2, \{p, 1-p\}\right)$ gives the Binomial distribution
- For $N = 1$, $\text{Mult}\left(1, K, \{p_k\}_{k=1}^{K}\right)$ gives the Categorical Distribution

Recall the Beta Distribution

- Notation: $X \sim \text{Beta}(\alpha_1, \alpha_2)$
- Parameters:
  - $\alpha_1, \alpha_2 > 0$: shape parameters
- Support: $x \in (0, 1)$



- PDF:

$$p(x) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \cdot x^{\alpha_1 - 1}(1 - x)^{\alpha_2 - 1}$$

- Serves as a conjugate prior for the binomial and bernoulli distributions

Dirichlet Distribution

Multivariate generalization of the Beta Distribution, also known as the Multivariate Beta Distribution

- Notation: $X \sim \mathrm{Dir}(\boldsymbol{\alpha})$, X is a K-dim. random vector
- Parameters:
  - $\{\alpha_k > 0\}_{k=1}^{K}$: called the concentration parameters
- Support: $\boldsymbol{x} \in \mathbb{R}^K$ such that $\sum_{k=1}^{K} x_k = 1$ and $\forall k \, x_k \in (0, 1)$
  Can be interpreted as the set of all K-dimensional probability vectors, thus Dirichlet is also sometimes called a "distribution over distributions"
- PDF:

$$p(\boldsymbol{x}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \cdot \prod_k x_k^{\alpha_k - 1}$$

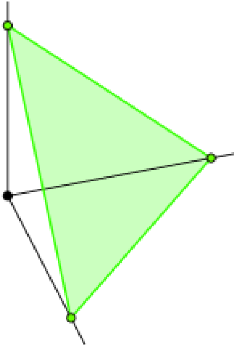
Dirichlet Distribution

Properties

- $X_k \sim \text{Beta}(\alpha_k, s - \alpha_k), s = \sum_k \alpha_k$
- $E[X_k] = \frac{\alpha_k}{s}$
- $Var(X_k) = \frac{\alpha_k(s-\alpha_k)}{s^2(s+1)}$
- $K = 2$ gives the Beta distribution

Visualizing the Support

Recall the support was $\boldsymbol{x} \in \mathbb{R}^K$ such that $\sum_{k=1}^{K} x_k = 1$ and $\forall k \, x_k \in (0, 1)$
Formally called the Open Standard $(K-1)$-simplex
For $K = 3$,



Visualizing the PDF (code for generating the plots)

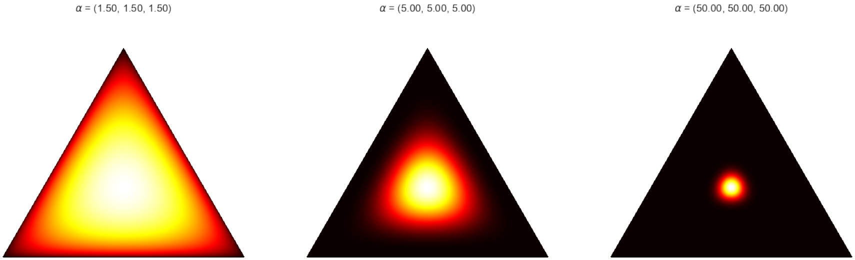


Figure: Effect of $\sum_k \alpha_k$

Dirichlet Distribution

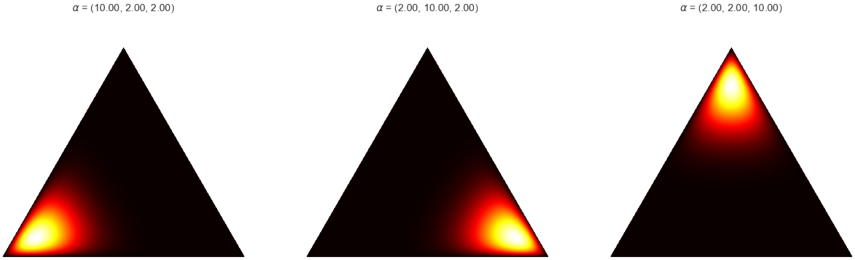


Figure: Effect of individual α_ks, keeping the sum fixed

Dirichlet Distribution

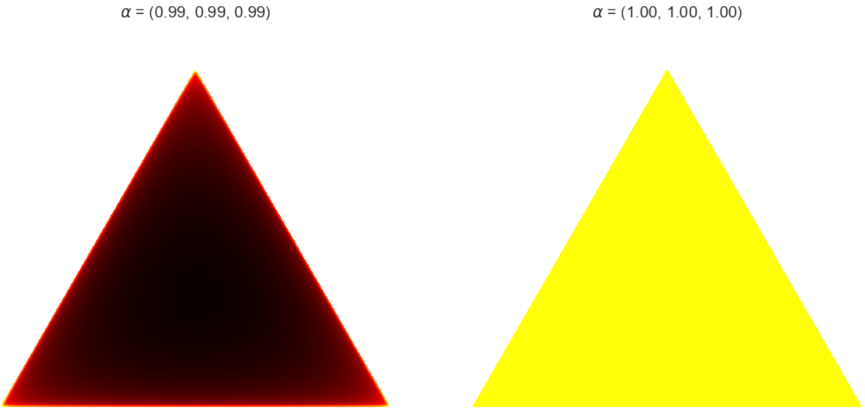


Figure: $\alpha_k < 1$ and $\alpha_k = 1$

Bayesian Inference

Multinomial Dirichlet Interaction

Dhruv Arora

October 27, 2021

The Dirichlet Distribution

Definition (Standard Simplex)

The standard simplex in n dimensions ($\mathcal{S}^n$) is the $n-1$ dimensional set of points

$$\left\{ x \in \mathbb{R}^n_{0+} \middle| \sum_{i=1}^{n} x_i = 1 \right\}$$

Definition (Dirichlet Distribution)

A dirichlet distribution is defined over $\mathcal{S}^n$ with parameter $\alpha \in \mathbb{R}^n_+$ as

$$Dir(X;\alpha) = \frac{\Gamma(\sum_{k=1}^{n} \alpha_k)}{\prod_{k=1}^{n} \Gamma(\alpha_k)} \prod_{k=1}^{n} x_k^{\alpha_k - 1} \propto \prod_{k=1}^{n} x_k^{\alpha_k - 1}$$

Conjugate Prior for Multinomial Distribution

Definition (Conjugate Prior)

A family of distribution D_1 is a conjugate prior for the likelihood family D_2 if the posterior probability obtained using $d_1 \in D_1$ and $d_2 \in D_2$ is $\in D_1$. It helps in mathematical simplification while using Bayesian estimation.

Does the dirichlet distribution look like a conjugate prior for a very common distribution?

Recall the multinomial distribution ($p \in \mathcal{S}^n$)

$$Multinom(X; p) = \frac{(\sum_{i=1}^{n} x_i)!}{\prod_{i=1}^{n} x_i!} \prod_{i=1}^{n} p_i^{x_i} \propto \prod_{i=1}^{n} p_i^{x_i}$$

Conjugate Prior for Multinomial Distribution

Let $X \sim Multinom(p)$ where p is unknown. Surely $\sum_{i=1}^{n} p_i = 1$. We model p using a dirichlet prior with parameter α. Then :

$$P(p|X) \propto P(X|p) \cdot P(p)$$

$$P(p|X) \propto \prod_{i=1}^{n} p_i^{x_i} \cdot \prod_{i=1}^{n} p_i^{\alpha_i - 1} = \prod_{i=1}^{n} p_i^{\alpha_i + x_i - 1}$$

$$P(p|X) = Dir(p; \alpha + X)$$

Conjugate Prior for Multinomial Distribution

Therefore dirichlet distribution is a conjugate prior for multinomial distribution.

It also requires little overhead after making observations to update the belief.

This is a desirable property for modelling.

Mean and Mode of a Dirichlet Distribution

How would you find the mode of $Dir(X; \alpha)$ (for a fixed α)?

You want

$$argmax_{X \in \mathcal{S}^n} \prod_{i=1}^{n} x_i^{\alpha_i - 1}$$

That is :

$$argmax_{x_1, x_2, \ldots x_n} \prod_{i=1}^{n} x_i^{\alpha_i - 1}$$

constrained to

$$\sum_{i=1}^{n} x_i = 1$$

Use Langrange multipliers! (Recall MA111)

Mean and Mode of a Dirichlet Distribution

That is, set

$$\nabla_x \prod_{i=1}^{n} x_i^{\alpha_i - 1} = \lambda \nabla_x \sum_{i=1}^{n} x_i$$

You get

$$\frac{\alpha_k - 1}{x_k} \prod_{i=1}^{n} x_i^{\alpha_i - 1} = \lambda \implies x_k = \lambda' \cdot (\alpha_k - 1)$$

for each $k \in 1, \ldots, n$ and of course, $\sum_{i=1}^{n} x_i = 1$

Convince yourself that the solution is

$$x_k = \frac{\alpha_k - 1}{\sum_{i=1}^{n} (\alpha_i - 1)}$$

Mean and Mode of a Dirichlet Distribution

What about the mean?

$$E[X] = \int_{X \in \mathcal{S}^n} X \cdot Dir(X; \alpha) dX$$

$$E[x_k] = C \int_{X \in \mathcal{S}^n} x_k \cdot \prod_{i=1}^{n} x_i^{\alpha_i - 1} dx = C \int_{X \in \mathcal{S}^n} \prod_{i=1}^{n} x_i^{\alpha_i + I(i==k) - 1} dx$$

But this is a dirichlet distribution itself (without the constant factor). Thus,

$$E[x_k] = C \frac{\prod_{i=1}^{n} \Gamma(\alpha_i + I(i == k))}{\Gamma(\sum_{i=1}^{n} \alpha_i + 1)}$$

Mean and Mode of a Dirichlet Distribution

Where

$$C = \frac{\Gamma(\sum_{i=1}^{n} \alpha_i)}{\prod_{i=1}^{n} \Gamma(\alpha_i)}$$

Convince yourself that

$$E[x_k] = \frac{\Gamma(\sum_{i=1}^{n} \alpha_i)}{\Gamma(\sum_{i=1}^{n} \alpha_i + 1)} \cdot \frac{\Gamma(\alpha_k + 1)}{\Gamma(\alpha_k)}$$

A well known property :

$$\Gamma(x+1) = x \cdot \Gamma(x)$$

$$E[x_k] = \frac{\alpha_k}{\sum_{i=1}^{n} \alpha_i}$$

Dirichlet + Multinomial : MAP, Posterior Mean, MLE

Let $X \sim Multinom(p)$ such that $X \in \mathbb{R}^n$ and $\sum_{i=1}^{n} x_i = N$

What is the likelihood L of obtaining X?

$$L \propto \prod_{i=1}^{n} p_i^{x_i}$$

Therefore, MLE estimate (as discussed above) is :

$$p_i^{MLE} = \frac{x_i}{N}$$

Dirichlet + Multinomial : MAP, Posterior Mean, MLE

What if we assume a $Dir(p; \alpha)$ prior ?

As we've already seen, the posterior distribution has parameter $\alpha + X$

The MAP estimate (mode) therefore is :

$$p_i^{MAP} = \frac{x_i + (\alpha_i - 1)}{N + \sum_{i=1}^{n}(\alpha_i - 1)}$$

The Posterior mean therefore is :

$$p_i^{PM} = \frac{x_i + \alpha_i}{N + \sum_{i=1}^{n} \alpha_i}$$

Do MAP and PM estimates converge to MLE

Does PM converge to MLE?

$$\lim_{N\to\infty} p_i^{MLE} - p_i^{PM} = \lim_{N\to\infty} \frac{x_i}{N} - \frac{x_i + a}{N + b}$$

$$= \lim_{N\to\infty} (\frac{x_i}{N} - \frac{x_i}{N + b}) - \lim_{N\to\infty} \frac{a}{N + b}$$

$$= \lim_{N\to\infty} \frac{x_i}{N} - \frac{x_i}{N + b}$$

Do MAP and PM estimates converge to MLE

$$\forall N \in \mathbb{N}\ \frac{x_i}{N} - \frac{x_i}{N+b} \geq 0 \implies \lim_{N\to\infty} \frac{x_i}{N} - \frac{x_i}{N+b} \geq 0$$

And

$$\frac{x_i}{N} - \frac{x_i}{N+b} = \frac{x_i \cdot b}{N(N+b)} \leq \frac{b}{N+b} \ \text{(why?)}$$

Therefore,

$$\lim_{N\to\infty} \frac{x_i}{N} - \frac{x_i}{N+b} \leq \lim_{N\to\infty} \frac{b}{N+b} = 0$$

Therefore p_i^{PM} and p_i^{MAP} both converge to p_i^{MLE}

But what does this achieve?

Recall coin flipping!

If you never observe tails for 100 trials do you assume you never would?

This smoothing is what dirichlet distribution achieves in the multinomial case.

In some sense

$$\texttt{Dirichlet} : \text{Beta} :: \text{Multinomial} : \text{Binomial}$$

Multinomial and Binomial processes are very common in statistical modelling

Therefore, so is the use of Beta and Dirichlet priors

Extra : Sampling a Dirichlet Distribution

Can you come up with a method to sample from an n dimensional Dirichlet distribution with parameter α given only a $[0, 1]$ uniform random generator?

Maybe you can! I'm not telling you how. HF