Lecture 4

Multivariate probabilities

Covariance

$$\mathrm{Cov}[X,Y] = E[(X-E(X))(Y-E(Y))] =$$

$$= E[XY] - E[X]E[Y].$$

Correlation

$$\rho = \mathrm{corr}[X,Y] = \frac{\mathrm{Cov}[X,Y]}{\sqrt{V[X]\,V[Y]}} = \frac{\mathrm{Cov}[X,Y]}{\sigma_X \sigma_Y}.$$

Consider $0 \le V\left[\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right] = V\left[\frac{X}{\sigma_X}\right] + V\left[\frac{Y}{\sigma_Y}\right] +$

$$+ 2\,\mathrm{Cov}\left[\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right] = \frac{V[X]}{\sigma_X^2} + \frac{V[Y]}{\sigma_Y^2} +$$

$$+ 2\,\frac{\mathrm{Cov}[X,Y]}{\sigma_X \sigma_Y} = 2(1+\rho), \text{ or}$$

$$\rho \ge -1$$

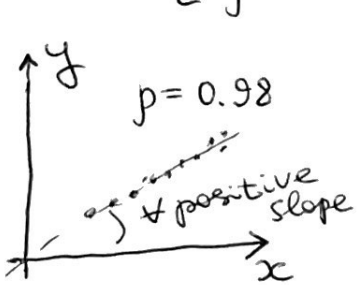
Likewise, $0 \le V\left[\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right] = 2(1-\rho)$, or

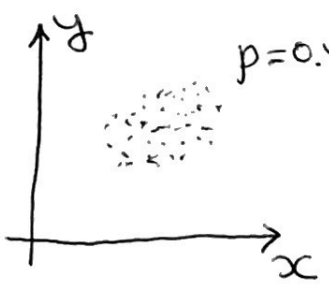
$$\rho \le 1.$$

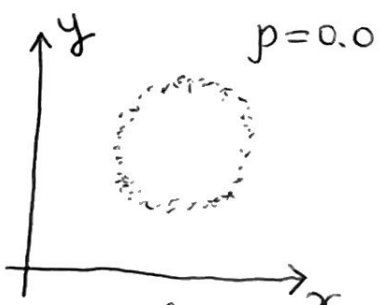
Thus, $\underline{-1 \le \rho \le 1}$

Note that $\rho = 1$ iff $Y = aX + b$, $a > 0$

[$\rho = -1$ iff $Y = aX + b$, $a < 0$]





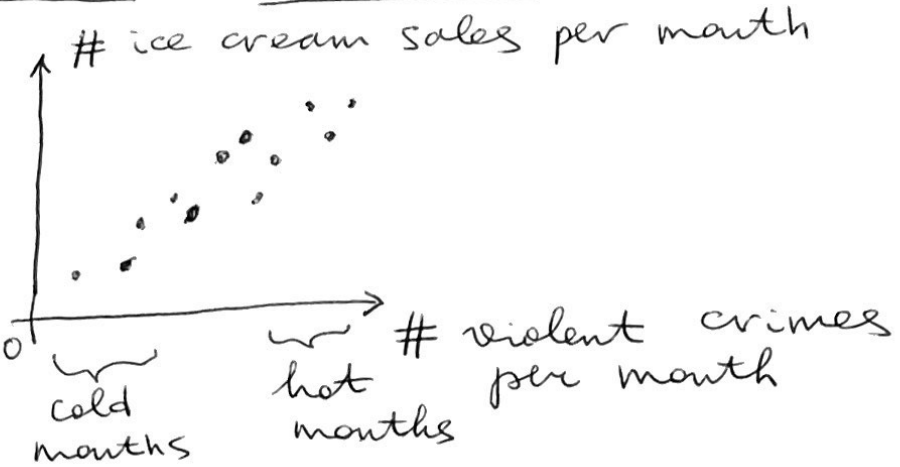


uncorrelated ⇏ independent

indep. ⇒ uncorrel.

Clust

Correlation ≠> causation

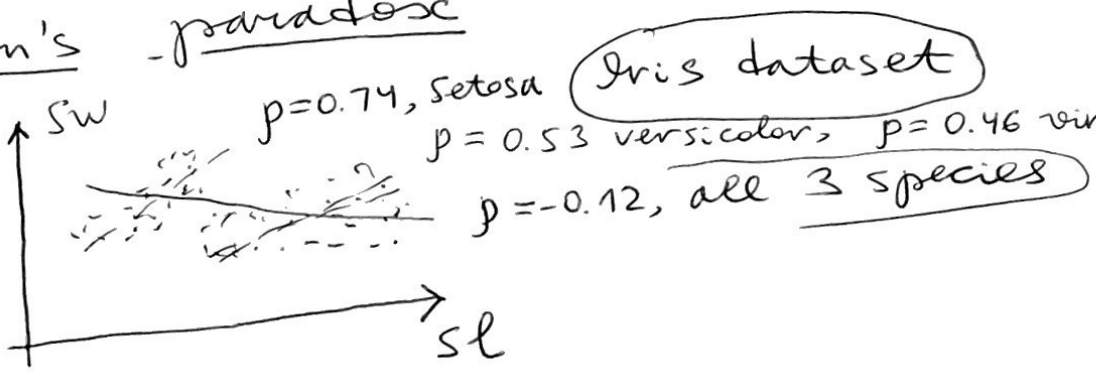


Hidden common cause: weather

storks ~ birth rates in some areas

Hidden common cause: wealth (?)

Simpson's paradox



Multi-D gaussian distribution

$$\mathcal{N}(\vec{y} \mid \vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(\vec{y}-\vec{\mu})^T \times \Sigma^{-1}(\vec{y}-\vec{\mu})}$$

$\vec{y} = (y_1, \ldots, y_D)$

$\Sigma_{D \times D} = \mathrm{Cov}[\vec{y}]$ is the covariance matrix:

$$\Sigma = E[(\vec{y} - \underbrace{\vec{\mu}}_{\uparrow E[\vec{y}]})(\vec{y} - \vec{\mu})^T], \text{ or}$$

$$\Sigma = E[\vec{y}\vec{y}^T] - \vec{\mu}\cdot\vec{\mu}^T.$$

$D=2$: $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$ (ρ: corr'n coeffs.)

Isotropic $\Sigma = \sigma^2 \underbrace{\mathbb{I}_D}_{\text{unit matrix}}$

Marginals and conditionals

Consider $\vec{y} = (\vec{y}_1, \vec{y}_2)$ ($\vec{y}$: $2D$; $\vec{y}_1$: D; $\vec{y}_2$: D)

assume that $\vec{y}$ is gaussian distributed:

$$\mathcal{N}(\vec{y} \mid \vec{\mu}, \Sigma) \Rightarrow \vec{\mu} = \begin{pmatrix} \vec{\mu}_1 \\ \vec{\mu}_2 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \quad \underbrace{\Lambda}_{\text{precision matrix}} = \Sigma^{-1} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}$$

Marginals: $\begin{cases} p(\vec{y}_1) = \mathcal{N}(\vec{y}_1 \mid \vec{\mu}_1, \Sigma_{11}), \\ p(\vec{y}_2) = \mathcal{N}(\vec{y}_2 \mid \vec{\mu}_2, \Sigma_{22}). \end{cases}$

Moreover,

$$p(\vec{y}_1 | \vec{y}_2) = \mathcal{N}(\vec{y}_1 | \vec{\mu}_{1|2}, \Sigma_{1|2}),$$

where

$$\begin{cases} \vec{\mu}_{1|2} = \vec{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\vec{y}_2 - \vec{\mu}_2) = \\ \quad = \vec{\mu}_1 - \Lambda_{11}^{-1}\Lambda_{12}(\vec{y}_2 - \vec{\mu}_2), \\ \Sigma_{1|2} = \Lambda_{11}^{-1} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{cases}$$

2D example $\quad \Sigma = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \Rightarrow \Lambda = \Sigma^{-1} = \frac{1}{\sigma^2(1-\rho^2)} \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}$

$$\Sigma = \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix} \qquad \begin{cases} p(y_1) = \mathcal{N}(y_1 | \mu_1, \sigma^2), \\ p(y_2) = \mathcal{N}(y_2 | \mu_2, \sigma^2) \end{cases}$$

$$p(y_1 | y_2) = ?$$

$$\begin{cases} \mu_{1|2} = \mu_1 + \rho\sigma^2 \frac{1}{\sigma^2}(y_2 - \mu_2), \\ \sigma_{1|2} = \sigma^2 - \rho\sigma^2 \frac{1}{\sigma^2}\rho\sigma^2 = \\ \quad = \sigma^2[1 - \rho^2]. \end{cases}$$

Thus, $p(y_1 | y_2) = \mathcal{N}(y_1 | \mu_{1|2}, \sigma_{1|2}^2) = \mathcal{N}(y_1 | \mu_1 + \rho(y_2 - \mu_2), \sigma^2(1-\rho^2))$.

If $\mu_1 = \mu_2 = 0$, $\sigma = 1$, $\rho = 0.8$:

$$E[y_1 | y_2 = 1] = 0.8\,(1 - 0) = 0.8$$

if y_2 increases by 1.0 (from the mean), y_1 increases by 0.8 on average since $\rho = 0.8$

Next, $V[y_1|y_2=1] = 1 - 0.8^2 = 0.36 < V[y_1] = 1$.

Variance went down b/c observing a correlated variable y_2 provides info about y_1.

Note that if $\rho = 0$,

$$p(y_1|y_2) = \mathcal{N}(y_1|\mu_1, \sigma^2) = p(y_1).$$

Observing y_2 does not help since y_1 & y_2 are uncorrelated.

Linear gaussian systems

$\vec{z} \in R^L$ hidden variables

$\vec{y} \in R^D$ noisy observables

Then we can assume:

$$\begin{cases} p(\vec{z}) = \mathcal{N}(\vec{z}|\vec{\mu}_z, \Sigma_z) \\ p(\vec{y}|\vec{z}) = \mathcal{N}(\vec{y}|W_{D\times L}\vec{z} + \vec{b}_D, \Sigma_y) \end{cases}$$

Thus, the posterior is

$$p(\vec{z}|\vec{y}) = \mathcal{N}(\vec{z}|\vec{\mu}_{z|y}, \Sigma_{z|y}), \text{ where}$$

$$\begin{cases} \Sigma_{z|y}^{-1} = \Sigma_z^{-1} + W^T \Sigma_y^{-1} W, \\ \vec{\mu}_{z|y} = \Sigma_{z|y}[W^T \Sigma_y^{-1}(\vec{y} - \vec{b}) + \Sigma_z^{-1}\vec{\mu}_z] \end{cases}$$

Moreover,

$$p(\vec{y}) = \int d\vec{z}\, p(\vec{y}|\vec{z})\, p(\vec{z}) = \mathcal{N}(\vec{y} \mid W\vec{\mu}_z + \vec{b},\ \Sigma_y + W\Sigma_z W^T).$$

Ex. Bayesian inference of $\vec{z}$ from

$\vec{y} = (y_1 \dots y_N) \Rightarrow W = \begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}_N$, $\vec{b} = \vec{0}_N$

λ_0 = precision

Assume $p(\vec{z}) = \mathcal{N}(\vec{z} \mid \mu_0, \underbrace{\lambda_0^{-1}}_{=\sigma_0^2})$,

↑ prior

$$p(y_i|z) = \mathcal{N}(y_i \mid z, \underbrace{\lambda^{-1}}_{=\sigma^2}). \qquad i = 1, \dots, N$$

↑ each y_i is a noisy measurement of z

$p(\vec{y}|z) = \prod_i p(y_i|z) = \mathcal{N}(\vec{y} \mid \vec{z}, \lambda^{-1}\mathbb{I}_N)$ $\left(\vec{z} = \begin{pmatrix}z\\\vdots\\z\end{pmatrix}\right)$

Then $\mu_{z|y} = \Sigma_{z|y}\left[N\lambda\,\bar{y} + \lambda_0\mu_0\right]$ (≡)

$\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}_N z + \vec{0}_N$; W

$\underbrace{1\dots1}\ \lambda\begin{pmatrix}1&0\\0&1\\&&\ddots\\&&&1\end{pmatrix}\begin{pmatrix}y_1\\\vdots\\y_N\end{pmatrix} = \lambda\sum_{j=1}^{N} y_j = N\lambda\bar{y}$; $\bar{y}$ = average

1D

$\Sigma_y^{-1} = \lambda\mathbb{I}_N$; $\Sigma_z^{-1} = \lambda_0$

$$\Sigma_{z|y}^{-1} = \lambda_0 + N\lambda \equiv \lambda_N$$ ← posterior precision, sum of λ_0 & $N\lambda$

average of y's

(≡) $\dfrac{N\lambda}{N\lambda+\lambda_0}\bar{y} + \dfrac{\lambda_0}{N\lambda+\lambda_0}\mu_0$.

weighted average of $\bar{y}$ & μ_0

If $N=1$, $\bar{y} \to y$ (single datapoint) and

$$\mu_1 = \frac{\lambda_0}{\lambda+\lambda_0}\mu_0 + \frac{\lambda}{\lambda+\lambda_0}y = \quad (\lambda = \sigma^{-2},\ \lambda_0 = \sigma_0^{-2})$$

$$= \frac{1}{\frac{\sigma_0^2}{\sigma^2}+1}\mu_0 + \frac{1}{1+\frac{\sigma^2}{\sigma_0^2}}y = \frac{\sigma^2}{\sigma^2+\sigma_0^2}\mu_0 + \frac{\sigma_0^2}{\sigma^2+\sigma_0^2}y \quad (\text{II})$$

$$(=)\ \mu_0 + (y-\mu_0)\frac{\sigma_0^2}{\sigma^2+\sigma_0^2} \ (\Leftrightarrow)$$

↑ prior ↑ correction due to data

$$(\Leftrightarrow)\ y - (y-\mu_0)\frac{\sigma^2}{\sigma^2+\sigma_0^2}.$$

↑ data ↑ adjustment towards the mean; 'shrinkage'

$$\begin{cases} \sigma_0 \ll \sigma : \mu_1 \simeq \mu_0, & \text{'prior wins'} \\ \sigma_0 \gg \sigma : \mu_1 \simeq y, & \text{'data wins'} \end{cases}$$