

Lecture 03: Multivariate Probability

Introduction to Machine Learning

Sajjad Amini

Department of Electrical Engineering
Sharif University of Technology

Review!

Contents

- 1 Important Notation Definition
- 2 Basic Definitions
- 3 Sample Distributions
- 4 Linear Gaussian Systems
- 5 Mixture Models

Section 2

Basic Definitions

Basic Definitions

- Suppose two random variables X and Y . The Covariance is defined as:

$$\text{Cov}[X, Y] \triangleq E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

- Assume $\mathbb{X} = [X_1, X_2, \dots, X_D]^T$ is a D-dimensional random vector, then its covariance matrix is defined as:

$$\text{Cov}[\mathbb{X}] \triangleq E[(\mathbb{X} - E[\mathbb{X}])(\mathbb{X} - E[\mathbb{X}])^T] = \Sigma$$

Variance terms.

$$= \begin{bmatrix} \text{Cov}[X_1, X_1] & \text{Cov}[X_1, X_2] & \cdots & \text{Cov}[X_1, X_D] \\ \text{Cov}[X_2, X_1] & \text{Cov}[X_2, X_2] & \cdots & \text{Cov}[X_2, X_D] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[X_D, X_1] & \text{Cov}[X_D, X_2] & \cdots & \text{Cov}[X_D, X_D] \end{bmatrix}$$

Covariance

some thing

- Cross-covariance: $\text{Cov}[\mathbb{X}, \mathbb{Y}] = E[(\mathbb{X} - E[\mathbb{X}])(\mathbb{Y} - E[\mathbb{Y}])^T]$
- we mech. $\text{Cov}([\mathbb{X}][\mathbb{Y}])$ has more info.*

- $E[\mathbb{X}\mathbb{X}^T] = \Sigma + \mu\mu^T$, $\mu \triangleq E[\mathbb{X}]$

- $\text{Cov}[A\mathbb{X} + b] = A \text{Cov}[\mathbb{X}] A^T$ *no change of basis!*

Basic Definitions

Correlation

- Suppose two random variables X and Y . The Correlation that measure the level of **Linear** relation between two variables is defined as:

$$\rho \triangleq \text{Cor}[X, Y] \triangleq \frac{\text{Cov}[X, Y]}{\sqrt{\text{V}[X] \text{V}[Y]}}$$

- If \mathbb{X} is a D-dimensional random vector, its correlation matrix is defined as:

$$\text{Cor}[\mathbb{X}] \triangleq \begin{bmatrix} \text{Cor}[X_1, X_1] = 1 & \text{Cor}[X_1, X_2] & \cdots & \text{Cor}[X_1, X_D] \\ \text{Cor}[X_2, X_1] & \text{Cor}[X_2, X_2] = 1 & \cdots & \text{Cor}[X_2, X_D] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cor}[X_D, X_1] & \text{Cor}[X_D, X_2] & \cdots & \text{Cor}[X_D, X_D] = 1 \end{bmatrix}$$

Properties

- One can show that $-1 \leq \rho \leq 1$
- $|\text{Cor}[X, Y]| = 1$ iff $Y = aX + b$

Correlation and Nonlinear Dependencies [1]

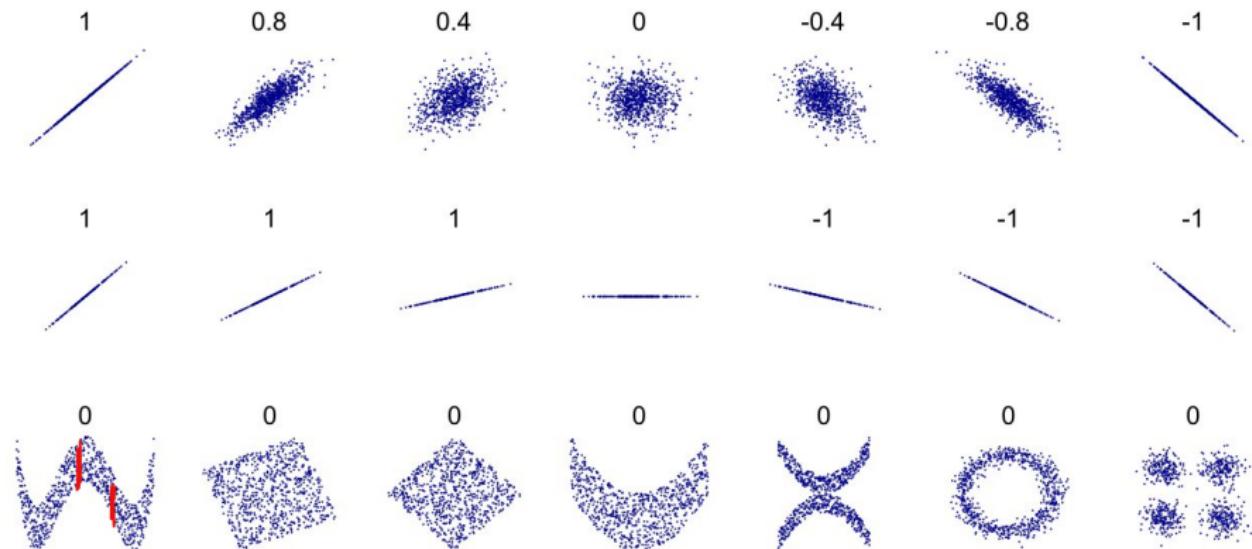


Figure: Visual interpretation of conditional probability

Uncorrelatedness vs. Independence

$$\begin{aligned}\text{Cov}[X, Y] &= E[XY] - E[X]E[Y] = E[X]E[Y] - E[X]E[Y] = 0 \\ \Rightarrow \text{Cor}[X, Y] &= \frac{\text{Cov}[X, Y]}{\sqrt{V[X]V[Y]}} = 0\end{aligned}$$

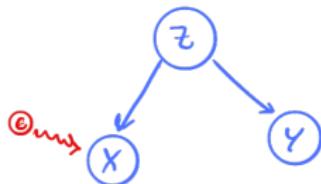
Suppose: $\begin{cases} X \sim U(-1, 1) \\ Y = X^2 \end{cases}$ Then: $\begin{cases} \text{Cor}[X, Y] = 0 \text{ (Uncorrelated)} \\ X \not\perp\!\!\!\perp Y \end{cases}$

Independence iff $P(X, Y) = P(X)P(Y)$ $\rightsquigarrow E[XY] = E[X]E[Y]$
for those who know information theory: $I(X; Y) = 0$

Correlatedness vs. Causation

Suppose: $\begin{cases} X \sim U(-1, 1) \\ Y = X^2 \end{cases}$ Then: $\begin{cases} \text{Cor}[X, Y] = 0 \text{ (Uncorrelated)} \\ X \text{ clearly causes } Y. \end{cases}$

$\begin{cases} Z \sim U(-1, 1) \\ X = Z^2 \\ Y = Z^2 \end{cases}$ Then: $\begin{cases} \text{Cor}[X, Y] = 1 \text{ (Correlated)} \\ X \text{ and } Y \text{ don't have causal effect on each other.} \end{cases}$



Spurious Correlation [2]

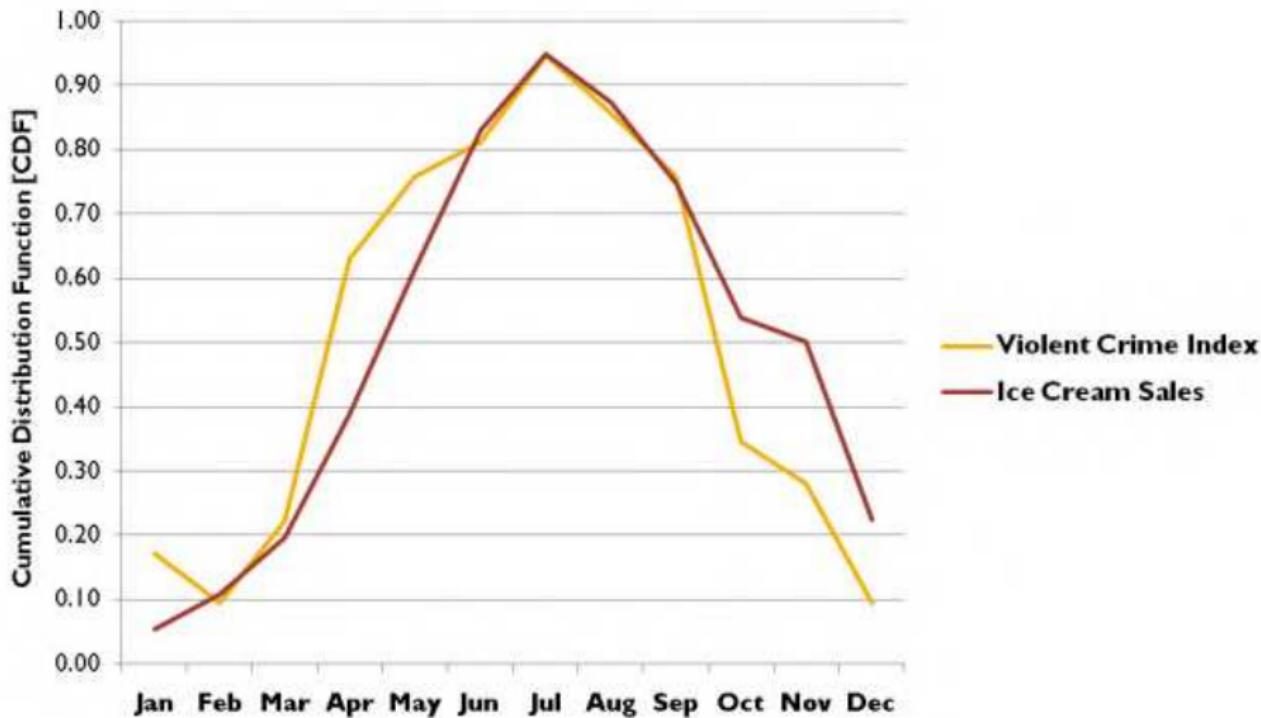


Figure: Violent Crime Index vs Ice Cream Sales

Section 3

Sample Distributions

The Multivariate Gaussian (Normal) Distribution (MVN)

Random vector \mathbb{Y} is said to be multivariate normally distributed if every linear combination of its components has a univariate normal distribution.

The PDF for MVN with dimension D is defined as:

$$\mathcal{N}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left[-\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right]$$

where:

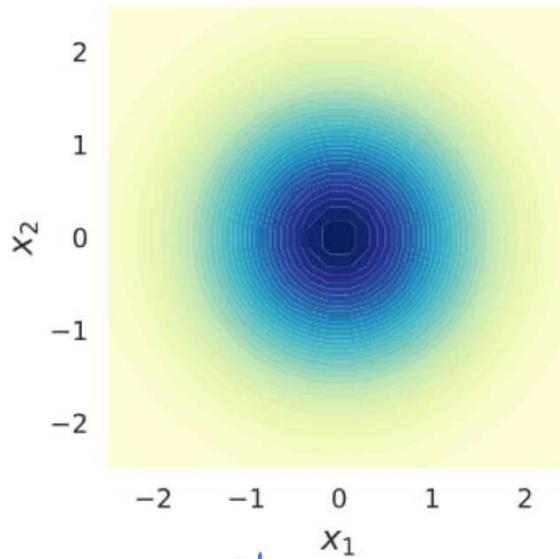
$$\boldsymbol{\mu} = \text{E}[\mathbb{Y}] \in \mathbb{R}^D$$

$$\boldsymbol{\Sigma} = \text{Cov}[\mathbb{Y}] \in \mathbb{R}^{D \times D}$$

$$p(x_1, x_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_1^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_2 - x_1)^2}{2}}$$

Bivariate normal distributions

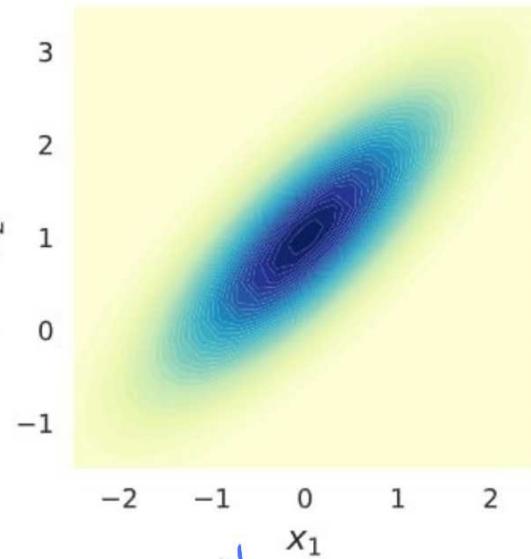
Independent variables



$$X_1, X_2 \stackrel{iid}{\sim} N(0, 1)$$

$$\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

Correlated variables

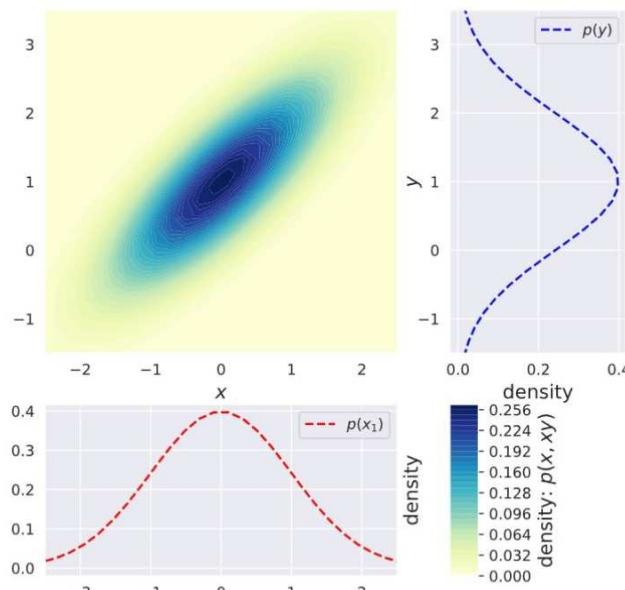


$$Z_1, Z_2 \stackrel{iid}{\sim} N(0, 1), X_1 = Z_1, X_2 = 0.8Z_1 + 0.6Z_2 + 1$$

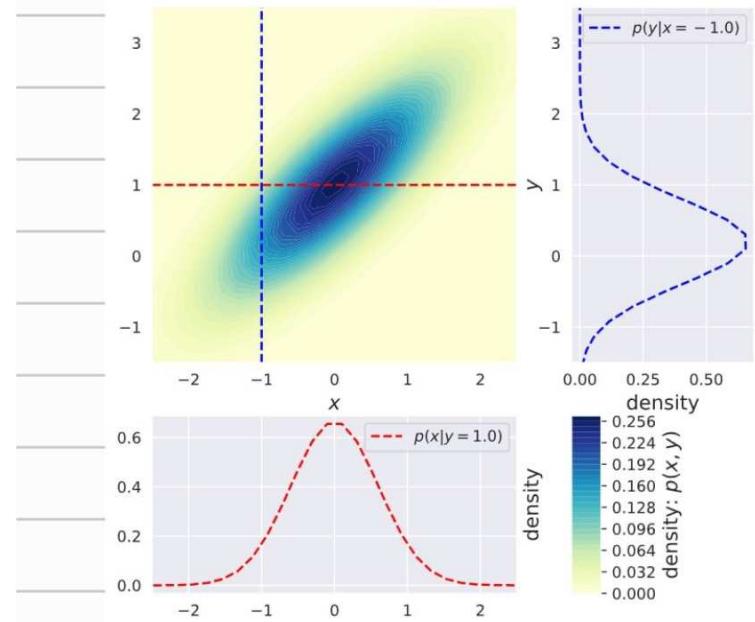
$$\mathcal{N}\left(\begin{bmatrix} 1 \\ 0.8 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}\right)$$

how to generate?

Marginal distributions



Conditional distributions



$$X \sim \mathcal{N}(0, 1)$$

$$\text{let } Y = \begin{cases} X & |x| < 1 \\ -x & |x| \geq 1 \end{cases} \Rightarrow Y \sim \mathcal{N}(0, 1)$$

good rule of thumb:

$\Rightarrow (X, Y)$ not MVN

$$X \text{ MVN} \text{ iff } \exists L, \mu : Z \sim \mathcal{N}(0, I), X = \mu + LZ$$

MVN Covariance Matrix Properties

Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is symmetric iff $\mathbf{A} = \mathbf{A}^T$ (We usually show this by $\mathbf{A} \in \mathbb{S}^n$)

Suppose $\mathbf{A} \in \mathbb{S}^n$. Then $\forall \mathbf{v} \in \mathbb{R}^n \setminus \{0\}$:

$$\mathbf{A} \text{ is positive definite (PD), denoted } \mathbf{A} \succ 0 \iff \mathbf{v}^T \mathbf{A} \mathbf{v} > 0$$

$$\mathbf{A} \text{ is positive semidefinite (PSD), denoted } \mathbf{A} \succeq 0 \iff \mathbf{v}^T \mathbf{A} \mathbf{v} \geq 0$$

$$\mathbf{A} \text{ is negative definite (ND), denoted } \mathbf{A} \prec 0 \iff \mathbf{v}^T \mathbf{A} \mathbf{v} < 0$$

$$\mathbf{A} \text{ is negative semidefinite (NSD), denoted } \mathbf{A} \preceq 0 \iff \mathbf{v}^T \mathbf{A} \mathbf{v} \leq 0$$

\mathbf{A} is indefinite iff it is none of the above.

MVN Covariance Matrix Properties

Covariance Matrix PSD

Assume Σ to be the covariance matrix of \mathbb{X} D-dimensional random vector.
Then:

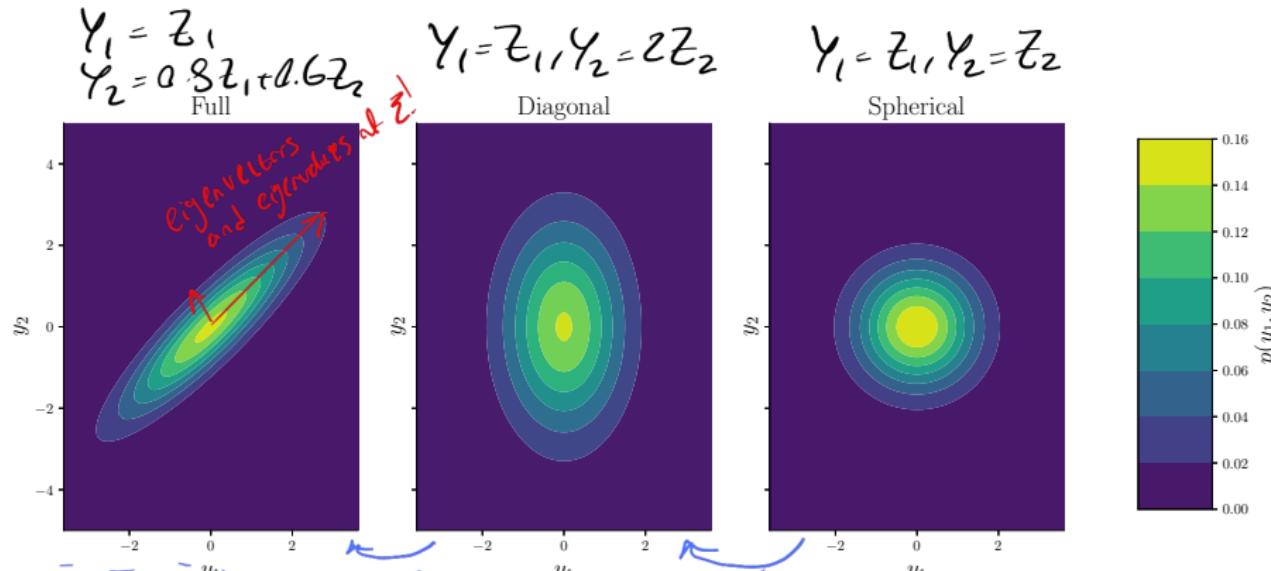
- $\Sigma \in \mathbb{S}^D$ based on definition.
- $\Sigma \succeq 0$ (PSD) because:

$$\mathbf{v}^T \Sigma \mathbf{v} = V[\mathbf{v}^T \mathbb{X}] \geq 0, \quad \forall \mathbf{v} \in \mathbb{R}^D$$

- If \mathbb{X} is distributed normally, then $\Sigma \succ 0$ (PD) because:

$$\exists \mathbf{v} \neq \mathbf{0} : \mathbf{v}^T \Sigma \mathbf{v} = 0 \rightarrow V[\mathbf{v}^T \mathbb{X}] = 0 \rightarrow \mathbf{v}^T \mathbb{X} \text{ is not normally distributed}$$

Bivariate Normal (D=2)



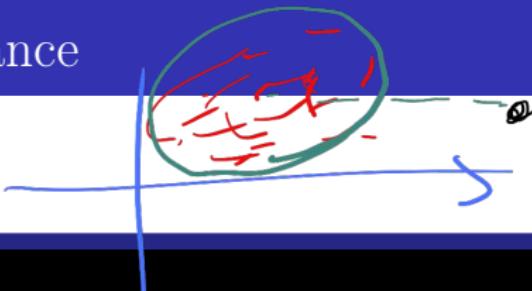
$$-\frac{1}{5\sqrt{2}}z_1 + \frac{9}{5\sqrt{2}}z_2 = x_1$$

$$\frac{1}{5\sqrt{2}}z_1 + \frac{9}{5\sqrt{2}}z_2 = x_2$$

Figure: Level set of constant probability density

$z_1, z_2 \stackrel{\text{iid}}{\sim} N(0, 1)$: source of uncertainty

Mahalanobis Distance



Mahalanobis Distance

Mahalanobis Distance (Δ) is a metric to calculate the distance between point y and distribution p with mean μ and covariance matrix Σ and is defined as:

$$\Delta^2 \triangleq (y - \mu)^T \Sigma^{-1} (y - \mu)$$

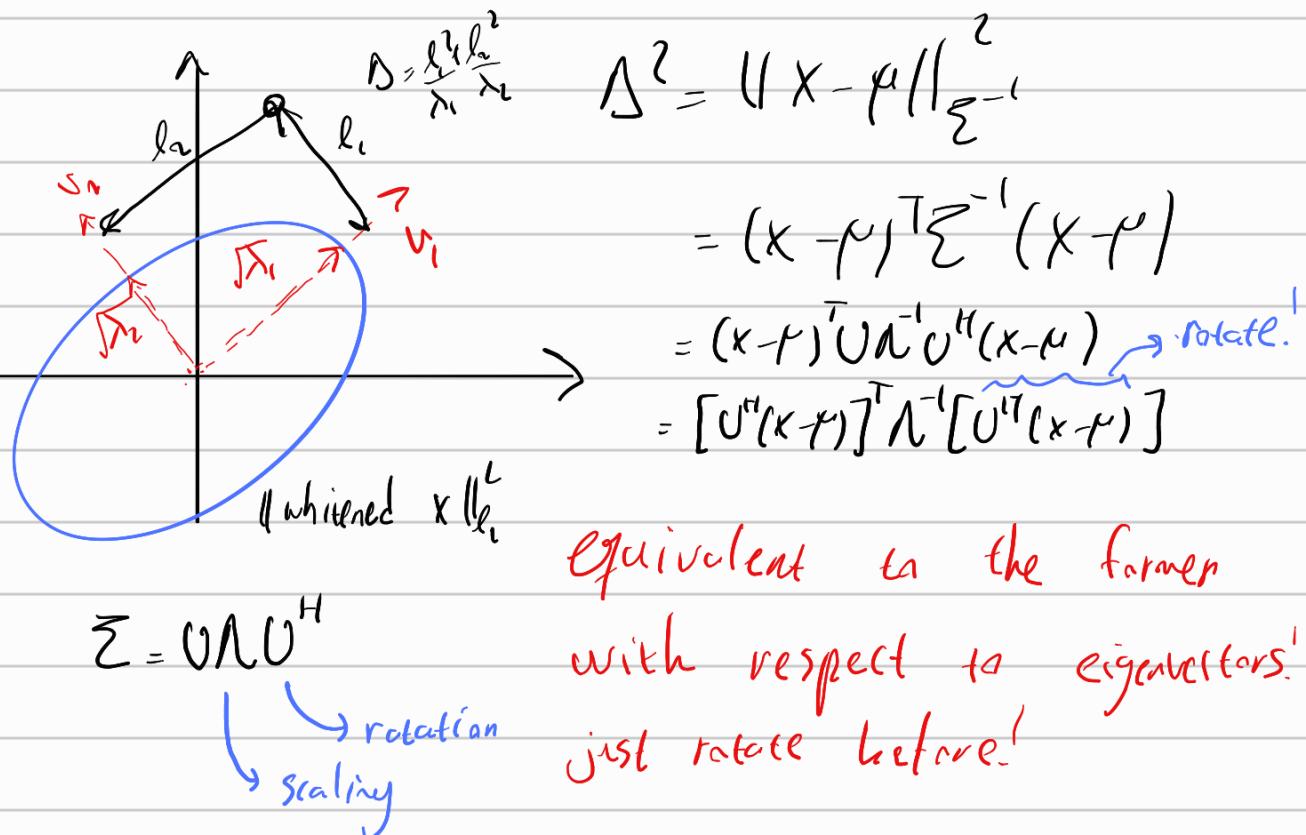
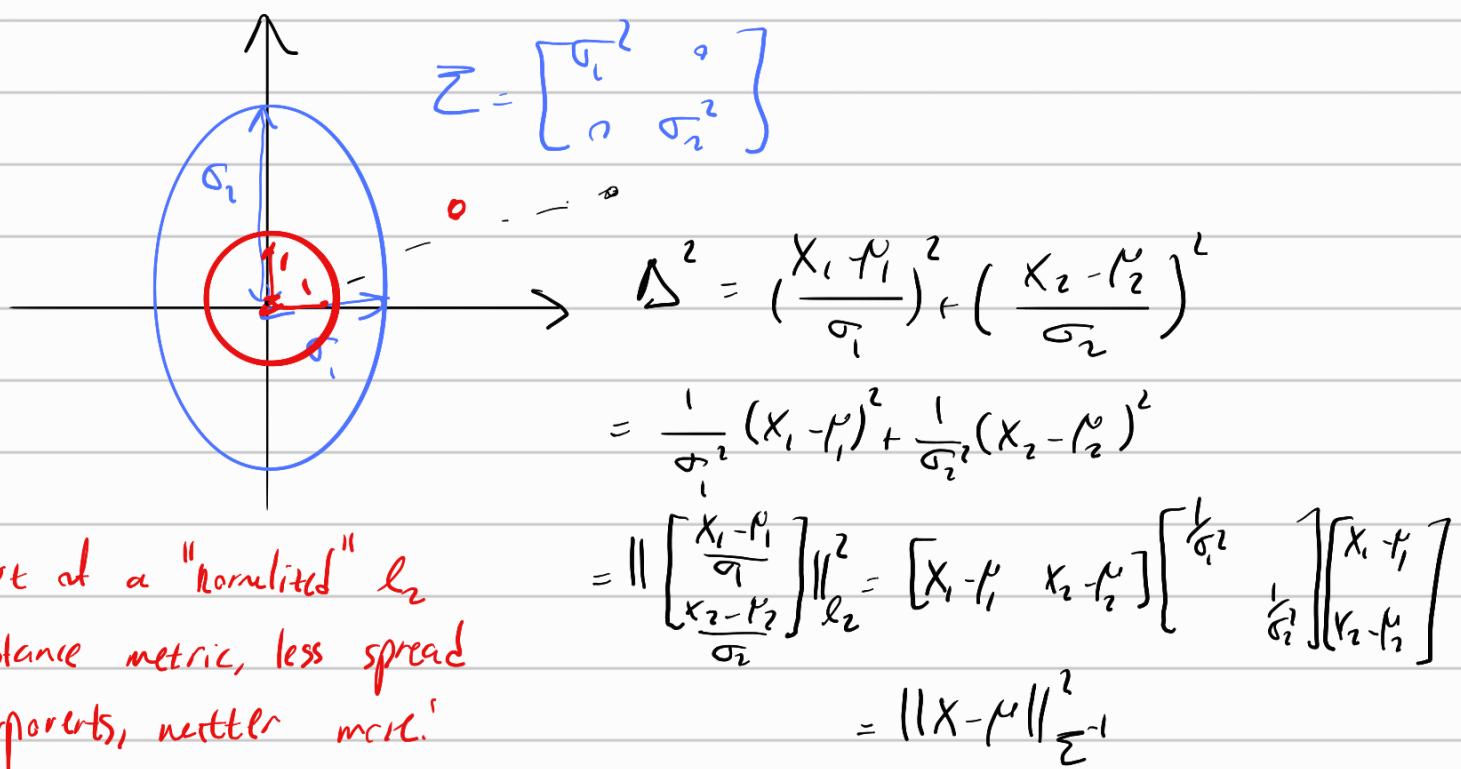
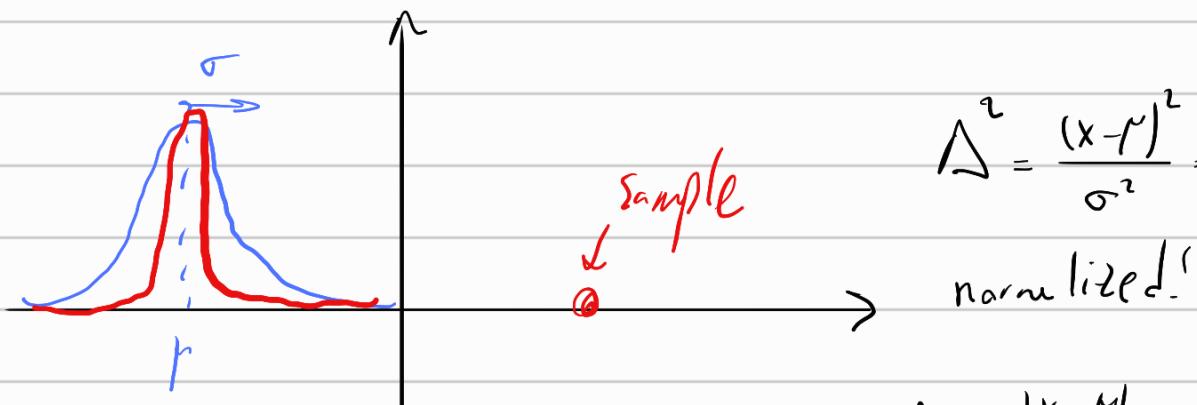
*Actually used for
Euclidean
models!*

The log probability of MVN at a specific point y is given by:

optimise

$$\log p(y|\mu, \Sigma) = -\frac{1}{2} \underbrace{(y - \mu)^T \Sigma^{-1} (y - \mu)}_{\Delta^2} + \text{constant}$$

MLE, MAP



Inference for MVN

Marginals and Conditionals of an MVN

Suppose $\mathbb{Y} = (\mathbb{Y}_1, \mathbb{Y}_2)$ where \mathbb{Y}_1 and \mathbb{Y}_2 have D_1 and D_2 dimension, respectively (thus \mathbb{Y} is (D_1+D_2) -dimensional). Assume \mathbb{Y} to be Gaussian with following parameters:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}, \quad \boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{bmatrix}$$

where $\boldsymbol{\mu}_1 \in \mathbb{R}^{D_1}$, $\boldsymbol{\mu}_2 \in \mathbb{R}^{D_2}$, $\boldsymbol{\Sigma}_{ij} \in \mathbb{R}^{D_i \times D_j}$ and $\boldsymbol{\Lambda}_{ij} \in \mathbb{R}^{D_i \times D_j}$. Then the marginals and conditionals are given by:

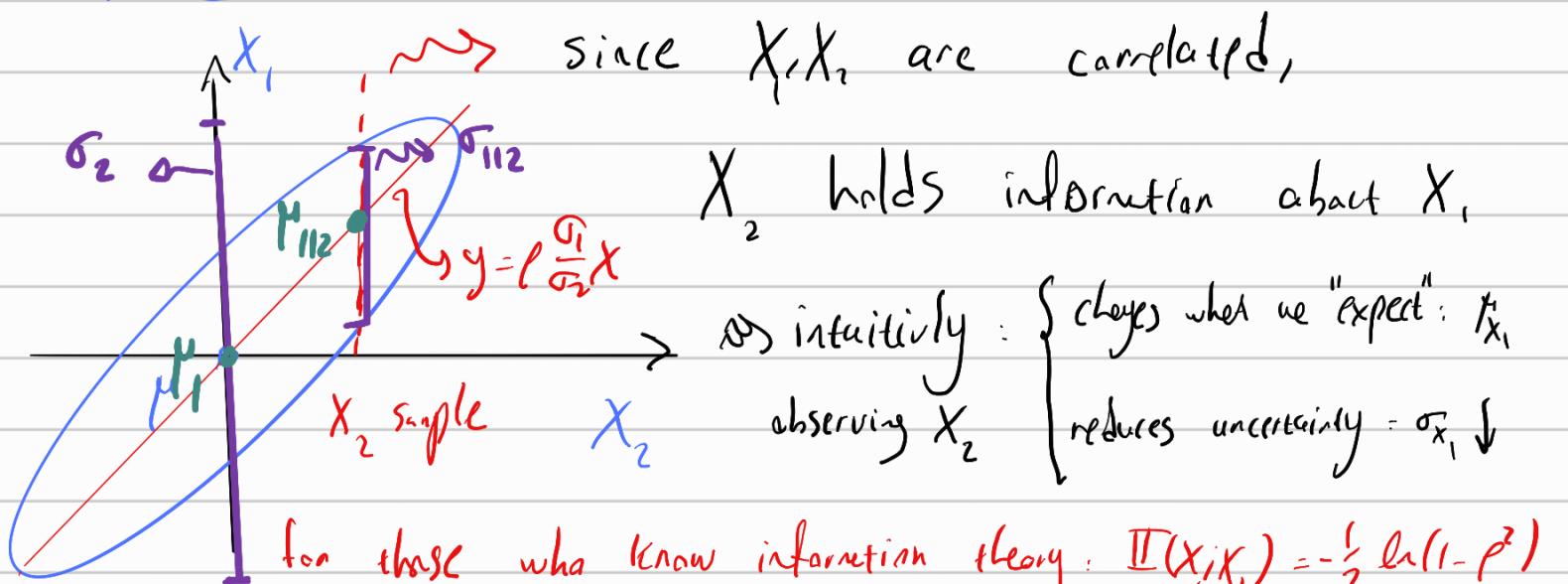
$$\begin{aligned} p(\mathbf{y}_1) &= \mathcal{N}(\mathbf{y}_1 | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \\ p(\mathbf{y}_2) &= \mathcal{N}(\mathbf{y}_2 | \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}) \\ p(\mathbf{y}_1 | \mathbf{y}_2) &= \mathcal{N}(\mathbf{y}_1 | \boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}) \end{aligned}$$

where:

inferred *observed*

$$\begin{aligned} \boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2) \quad (\text{Affine function of observed vector } \mathbf{y}_2) \\ \boldsymbol{\Sigma}_{1|2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \quad (\text{Independent of observed vector } \mathbf{y}_2) \end{aligned}$$

2D case?



Since for gaussians: variance = uncertainty (entropy)
 Reduces variance bits of information we gain.

$$\begin{aligned}\mu_{1|2} &= \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (y_2 - \mu_2) = \mu_1 + \rho \sigma_1 \sigma_2 \cdot \frac{1}{\sigma_2^2} (y_2 - \mu_2) \\ &= \mu_1 + \rho \cdot \frac{\sigma_1}{\sigma_2} (y_2 - \mu_2) \rightsquigarrow \text{linear dependence!}\end{aligned}$$

$$\begin{aligned}\Sigma_{1|2} &= \underbrace{\Sigma_1}_{\text{Schur complement}} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} = \sigma_1^2 - \rho \sigma_1 \sigma_2 \cdot \frac{1}{\sigma_2^2} \cdot \rho \sigma_1 \sigma_2 \\ &= \sigma_1^2 (1 - \rho^2) \rightsquigarrow \text{reduced var!} \\ &\quad \text{looks like } I(X_1; X_2)!!\end{aligned}$$

for MNN, just use block inverses,

Some analogy!

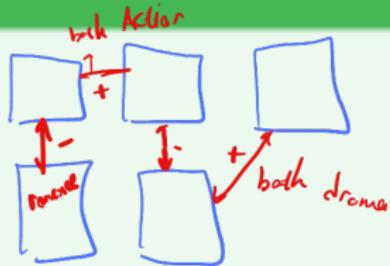
Using MVN Marginals

Imputing Missing Values

Consider the following scenario:

- Select D movies
- Ask N people to give them scores ($\mathbb{Y} \in \mathbb{R}^D$)
- Some people have not scored all movies.
- You know that the scoring vector comes from $\mathcal{N}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$

How to fill missing scores by MVN marginals?



Solution

We can fill person n scoring vector as:

- could be MAP*
- Compute $p(\mathbf{y}_{n,h} | \mathbf{y}_{n,v}, \boldsymbol{\theta})$ where:
$$\begin{cases} \boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}) : \text{Parameters} \\ h : \text{missing (hidden) score indices} \\ v : \text{submitted (visible) score indices} \end{cases}$$
 - Impute missing values by:
$$\begin{cases} \bar{\mathbf{y}}_{n,h} = E[\mathbb{Y}_{n,h} | \mathbf{y}_{n,v}, \boldsymbol{\theta}] : \text{Posterior mean} \\ \text{Posterior sampling} \end{cases} \Rightarrow \text{MLE}$$

Imputing Missing Values

How to estimate $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$? *Solution:* By using *Expectation Maximization*.

Section 4

Linear Gaussian Systems

Linear Gaussian Systems (LGS)

Given: $\mathbb{Z} \sim \mathcal{N}(\mu_z, \Sigma_z)$

Given: $\mathbb{Y} \sim \mathcal{N}(\mathbf{W}\mathbb{Z} + \mathbf{b}, \Sigma_y)$

Assume the following items:

- $\mathbb{Z} \in \mathbb{R}^L$: Unknown vector
- $\mathbb{Y} \in \mathbb{R}^D$: Noisy measurements
- The following distributions hold:
 - $p(\mathbb{z}) = \mathcal{N}(\mathbb{z} | \mu_z, \Sigma_z)$
 - $p(\mathbb{y} | \mathbb{z}) = \mathcal{N}(\mathbb{y} | \mathbf{W}\mathbb{z} + \mathbf{b}, \Sigma_y)$, $\mathbf{W} \in \mathbb{R}^{D \times L}$, $\mathbf{b} \in \mathbb{R}^D$

then:

- Joint distribution $p(\mathbb{z}, \mathbb{y}) = p(\mathbb{z})p(\mathbb{y} | \mathbb{z})$ is a $L + D$ dimensional Gaussian with the following parameters:

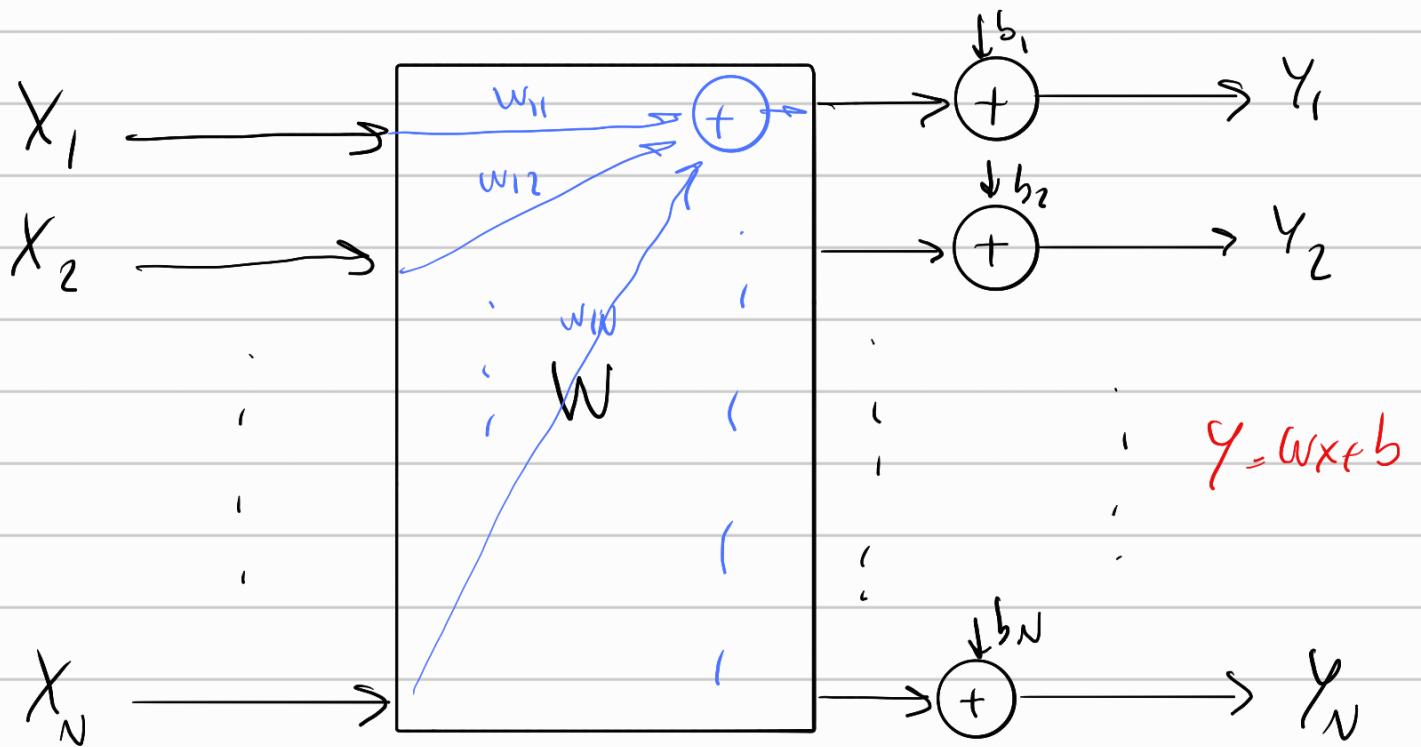
$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_z \\ \mathbf{W}\boldsymbol{\mu}_z + \mathbf{b} \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_z & \boldsymbol{\Sigma}_z \mathbf{W}^T \\ \mathbf{W} \boldsymbol{\Sigma}_z & \boldsymbol{\Sigma}_y + \mathbf{W} \boldsymbol{\Sigma}_z \mathbf{W}^T \end{bmatrix},$$

- Using Bayes rule, the posterior $p(\mathbb{z} | \mathbb{y})$ is also L dimensional Gaussian with the following parameters:

$$\boldsymbol{\Sigma}_{z|y}^{-1} = \boldsymbol{\Sigma}_z^{-1} + \mathbf{W}^T \boldsymbol{\Sigma}_y^{-1} \mathbf{W}$$

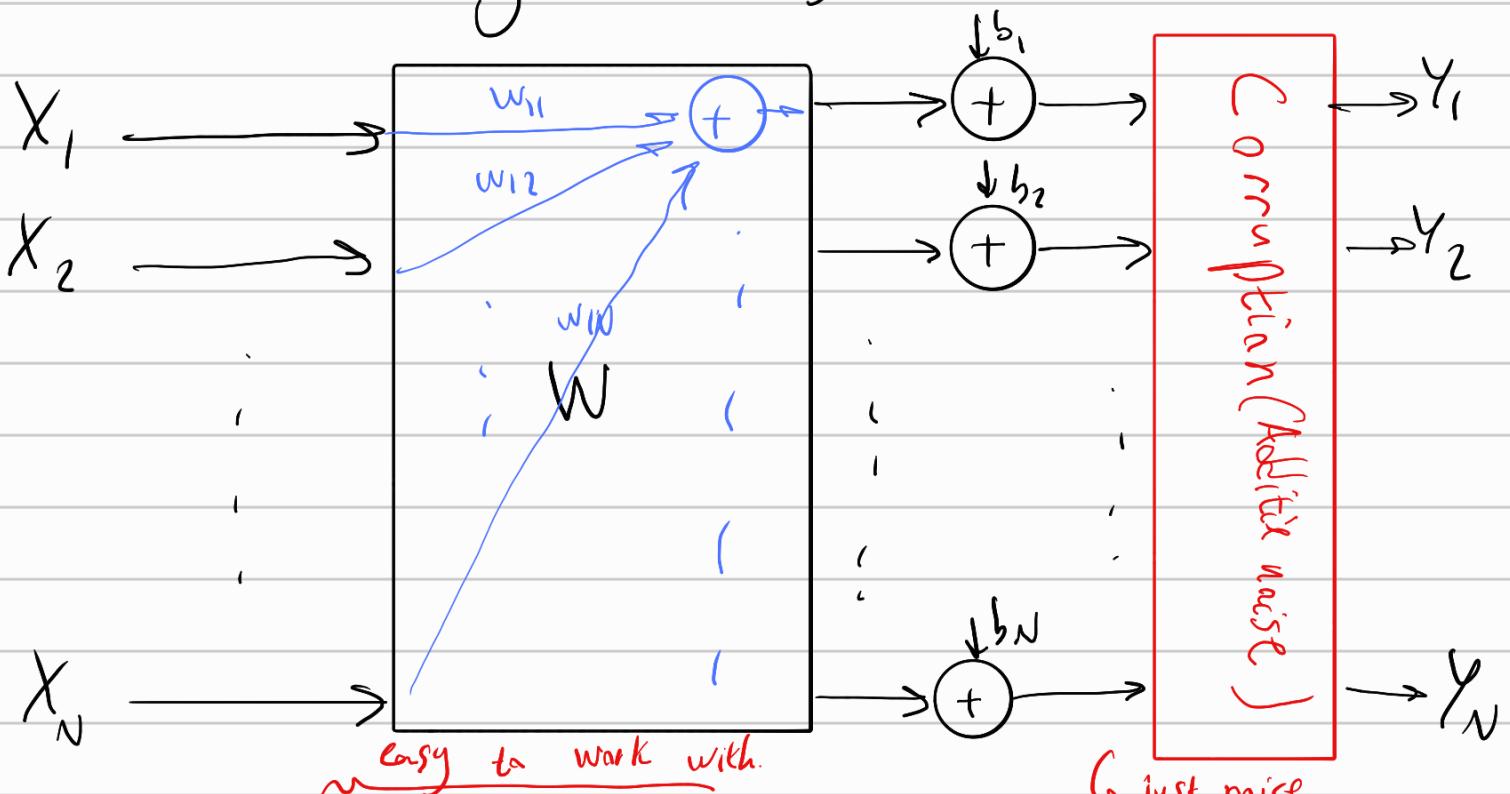
$$\boldsymbol{\mu}_{z|y} = \boldsymbol{\Sigma}_{z|y} \left[\mathbf{W}^T \boldsymbol{\Sigma}_y^{-1} (\mathbb{y} - \mathbf{b}) + \boldsymbol{\Sigma}_z^{-1} \boldsymbol{\mu}_z \right]$$

Linear (Affine) system.



e.g. FFT, DSP, DCT(JPEG), ...

Noisy Linear system.



$$p(y|z) = N(y|Wz+b, \Sigma_z)$$

↳ just noise uncertainty quantification error

equivalent to $\epsilon \sim N(0, \Sigma_\epsilon)$, $y = Wz + b + \epsilon$

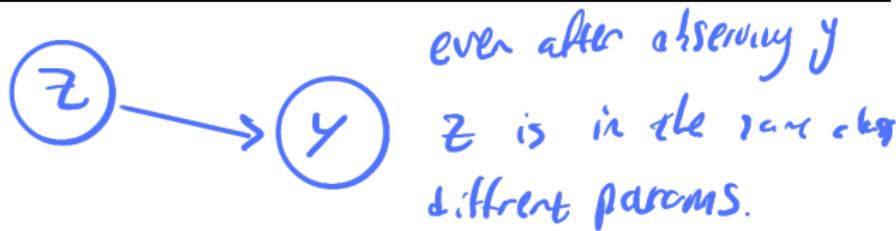
Conjugate Priors

Conjugate Priors

Assume \mathcal{F} as a family of distribution functions (e.g. Gaussian). We say that a prior $p(\mathbf{z}) \in \mathcal{F}$ is a conjugate prior for a likelihood function $p(\mathbf{y}|\mathbf{z})$ if the posterior is in the same family of distribution, i.e., $p(\mathbf{z}|\mathbf{y}) \in \mathcal{F}$.

Conjugate Priors

Based on slide 22, Gaussian prior is a conjugate prior for the Gaussian likelihood.



Linear Gaussian System

Suppose:

- *Prior*: We want to estimate unknown quantity Z where $p(z) = \mathcal{N}(z|\mu_0, \lambda_0^{-1})$
- *Likelihood*: We have N independent noisy measurements y_i distributed as

$$p(y_i|z) = \mathcal{N}(y_i|z, \lambda_y^{-1}) \quad \text{MLE?} \quad Z = \frac{y_1 + \dots + y_N}{N} = \bar{y}$$

compute the posterior $p(z|y_1, \dots, y_N)$.

Goal:

We start by defining $\mathbb{Y} = (y_1, \dots, y_N)$. Then we can easily show that the problem is linear Gaussian system with $\mathbf{W} = \mathbf{1}_N$ and $\Sigma_y^{-1} = \text{diag}(\lambda_y \mathbf{I})$. Thus:

$$p(z|\mathbf{y}) = \mathcal{N}(z|\mu_N, \lambda_N^{-1})$$

where:

$$\begin{aligned}\Sigma_{z|\mathbf{y}}^{-1} &= \Sigma_z^{-1} + \mathbf{W}^T \Sigma_y^{-1} \mathbf{W} \Rightarrow \lambda_{z|\mathbf{y}} = \lambda_0 + \mathbf{1}^T \text{diag}(\lambda_y \mathbf{I}) \mathbf{1} = \lambda_0 + N\lambda_y \\ \mu_{z|\mathbf{y}} &= \Sigma_{z|\mathbf{y}} \left[\mathbf{W}^T \Sigma_y^{-1} (\mathbf{y} - \mathbf{b}) + \Sigma_z^{-1} \boldsymbol{\mu}_z \right] \Rightarrow \mu_{z|\mathbf{y}} = \lambda_{z|\mathbf{y}}^{-1} \left[\mathbf{1}^T \text{diag}(\lambda_y \mathbf{I})(\mathbf{y} - \mathbf{0}) + \lambda_0 \boldsymbol{\mu}_0 \right] \\ \Rightarrow \mu_{z|\mathbf{y}} &= \frac{N\lambda_y \bar{y} + \lambda_0 \mu_0}{\lambda_{z|\mathbf{y}}} = \frac{N\lambda_y}{N\lambda_y + \lambda_0} \bar{y} + \frac{\lambda_0}{N\lambda_y + \lambda_0} \mu_0\end{aligned}$$

prior scale \uparrow \bar{y}

$$\text{MMSE} = \text{MAP} = P_{z|y}$$

this is a simple case:

$$P(z|y) = \frac{P(y|z) \cdot P(z)}{P(y)} \propto P(y|z) \cdot P(z) = P(y_1|z) \cdots P(y_N|z) P(z)$$

$$= P(z) \prod_{i=1}^N P(y_i|z)$$

$$P(z) = \frac{\lambda}{\sqrt{2\pi}} \exp\left(-\frac{1}{2\lambda} (\lambda(z - \mu))^2\right)$$

$$P(y_i|z) = \frac{\sqrt{\lambda_y}}{\sqrt{2\pi}} \exp\left(-\frac{1}{2\lambda_y} (\lambda_y(y_i - z))^2\right)$$

$$P(z) \prod_{i=1}^N P(y_i|z) \propto \exp\left(-\frac{\lambda}{2\lambda} (z - \mu)^2\right) \cdot \prod_{i=1}^N \exp\left(-\frac{\lambda_y}{2\lambda_y} (y_i - z)^2\right)$$

$$= \exp\left(-\frac{1}{2\lambda} [\lambda_z(z - \mu)^2 + \lambda_y \sum_{i=1}^N (y_i - z)^2]\right)$$

$$\text{MAP} : \frac{\partial}{\partial z} = 0 \Rightarrow \lambda_z(z - \mu) - \lambda_y \sum (y_i - z) = 0$$

$$\lambda_z(z - \mu) - N\lambda_y \bar{y} + N\lambda_y z_0 = 0$$

$$z(\lambda_z + N\lambda_y) = N\lambda_y \bar{y} + \lambda_z \mu$$

$$z = \frac{N\lambda_y \bar{y} + \lambda_z \mu}{\lambda_z + N\lambda_y}$$

Linear Gaussian System

All plots: assuming $\bar{y}=0$

LGS system with $N = 1, \lambda_y = 1.0$

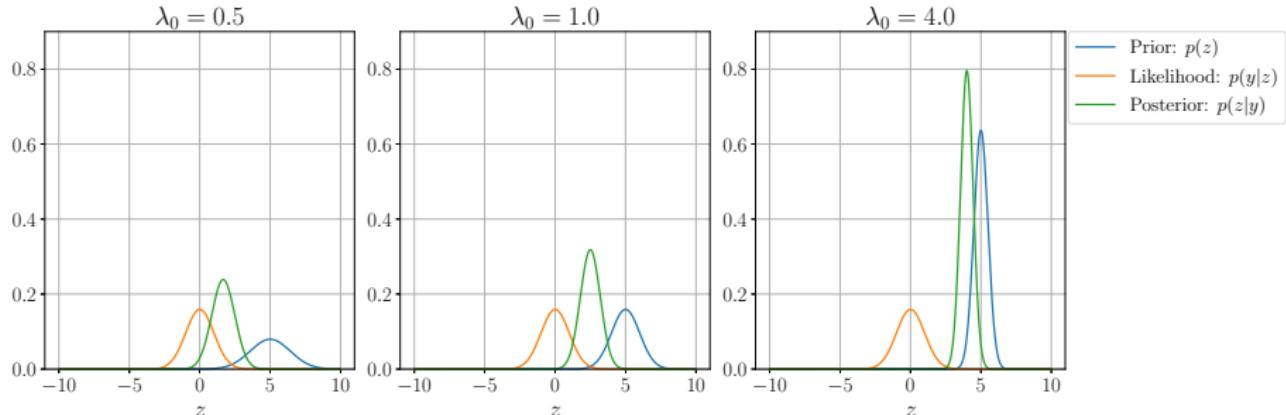


Figure: Prior precision (λ_0) effect

Linear Gaussian System

$$\bar{y} = 0$$

LGS system with $N = 1, \lambda_0 = 1.0$

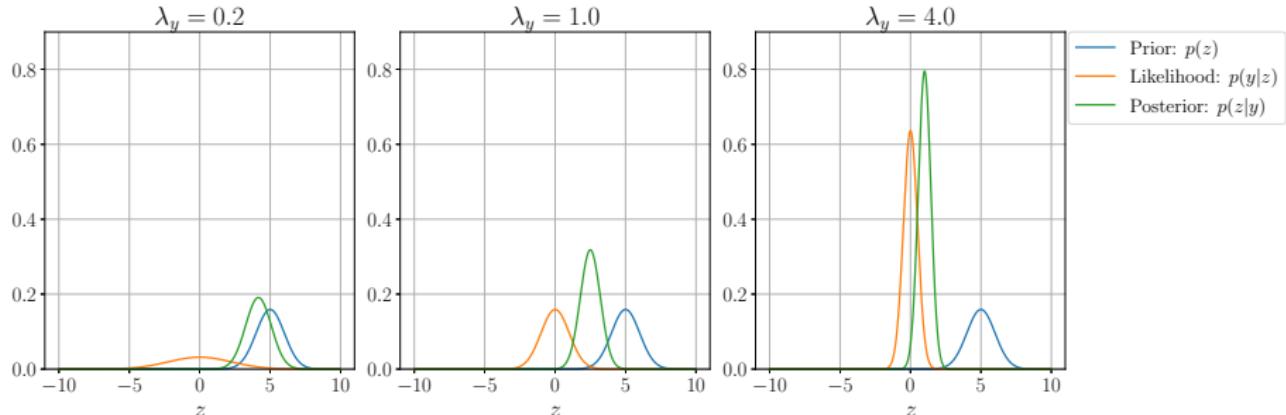


Figure: Likelihood precision (λ_y) effect

Linear Gaussian System

LGS system with $\lambda_0 = 1.0, \lambda_y = 1.0$

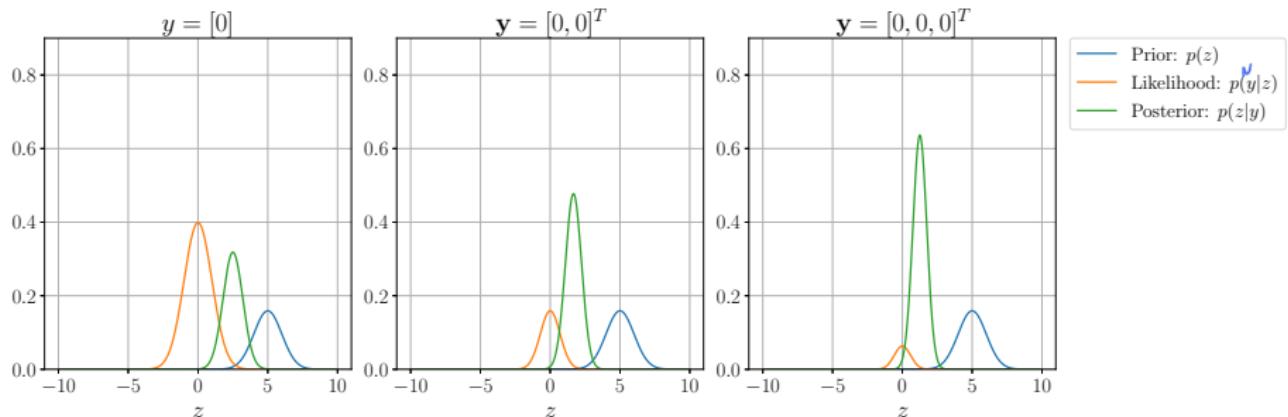


Figure: Number of measurements (N) effect

Linear Gaussian System

Suppose:

- *Prior*: We want to estimate unknown vector \mathbb{Z} where $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mu_0, \Sigma_0)$
- *Likelihood*: We have 2 sensors and 1 measurements of each sensor, denoted \mathbb{Y}_1 and \mathbb{Y}_2 , distributes as $\mathcal{N}(\mathbf{y}_i|\mathbf{z}, \Sigma_i)$ (Σ_i demonstrates the reliability for i -th sensor).

compute the posterior $p(\mathbf{z}|\mathbf{y}_1, \mathbf{y}_2)$.

We start by defining $\mathbb{Y} = (\mathbb{Y}_1, \mathbb{Y}_2)$. Then we can easily show that the problem is linear Gaussian system with $\mathbf{W} = [\mathbf{I}; \mathbf{I}]$ and $\Sigma_y = \begin{bmatrix} \Sigma_1 & \mathbf{0} \\ \mathbf{0} & \Sigma_2 \end{bmatrix}$. Thus the posterior $p(\mathbf{z}|\mathbf{y}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{z|y}, \Sigma_{z|y})$ where $\boldsymbol{\mu}_{z|y}$ and $\Sigma_{z|y}$ can be calculated using formulas in Slide 22.

Sensor Fusion

Suppose the sensor fusion example in Slide 28, with the following parameters:

$$\boldsymbol{\mu}_0 = [0; 0], \Sigma_0 = 1000\mathbf{I}, \Sigma_1 = \Sigma_2 = 0.01\mathbf{I}$$

and assume $\mathbf{y}_1 = (0, -1)$ and $\mathbf{y}_2 = (1, 0)$. Visualize the measurements and posterior $p(\mathbf{z}|\mathbf{y})$.

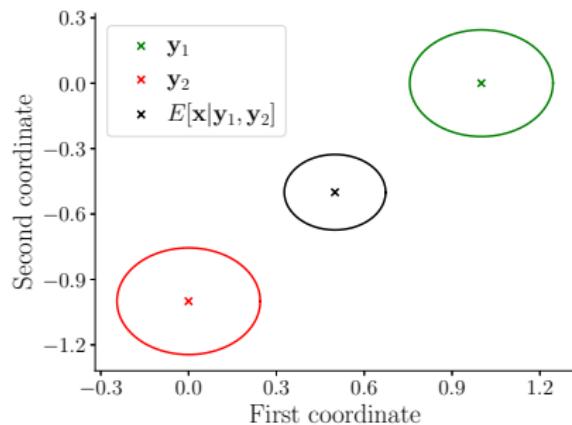


Figure: Sensor fusion result

Sensor Fusion

Suppose the sensor fusion example in Slide 28, with the following parameters:

$$\boldsymbol{\mu}_0 = [0; 0], \quad \boldsymbol{\Sigma}_0 = 1000\mathbf{I}, \quad \boldsymbol{\Sigma}_1 = 0.01\mathbf{I}, \quad \boldsymbol{\Sigma}_2 = 0.05\mathbf{I}$$

and assume $\mathbf{y}_1 = (0, -1)$ and $\mathbf{y}_2 = (1, 0)$. Visualize the measurements and posterior $p(\mathbf{z}|\mathbf{y})$.

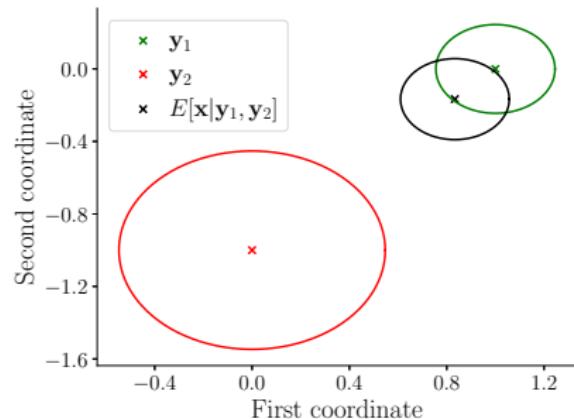


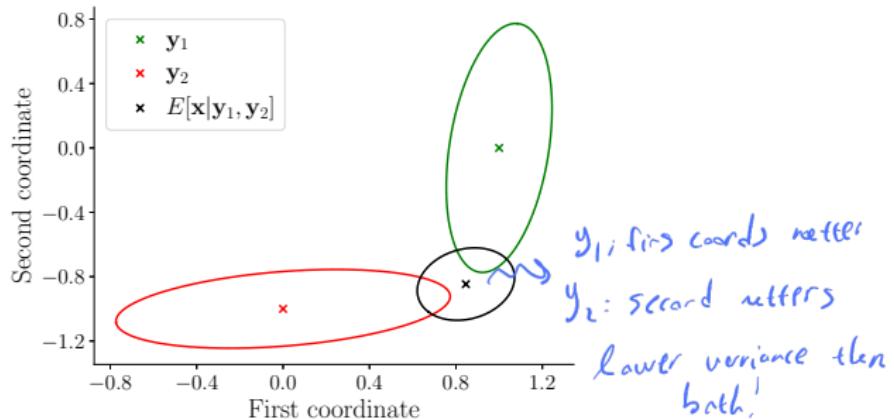
Figure: Sensor fusion result

Sensor Fusion

Suppose the sensor fusion example in Slide 28, with the following parameters:

$$\mu_0 = [0; 0], \Sigma_0 = 1000I, \Sigma_1 = 0.01 \begin{bmatrix} 10 & 1 \\ 1 & 1 \end{bmatrix}, \Sigma_2 = 0.01 \begin{bmatrix} 1 & 1 \\ 1 & 10 \end{bmatrix}$$

and assume $y_1 = (0, -1)$ and $y_2 = (1, 0)$. Visualize the measurements and posterior $p(z|y)$.



Gaussian: simple! ↗ Mixture

general distribution: hard! ↘

Section 5

Mixture Models

$$f(x) = \alpha(x_1 + x_2 e^{x_3}) \quad x \in [0, 1]$$

Mixture Models

Mixture Model

One way to create more complex probability models is to take a convex combination of simple distributions. This is called a mixture model. This has the form $p(\mathbf{y}|\boldsymbol{\theta}) = \sum_{k=1}^K \pi_k p_k(\mathbf{y})$ where:

- p_k is the k -th mixture component
- $\{\pi_k\}_{k=1}^K$ are mixture weights with the following constraints:
 - $0 \leq \pi_k \leq 1, k = 1, \dots, K$
 - $\sum_{k=1}^K \pi_k = 1$

Sampling from Mixture Models

Suppose latent variable Z to be a categorical RV and distributed as $p(z|\boldsymbol{\theta}) = \text{Cat}(z|\boldsymbol{\pi})$ and conditional $p(\mathbf{y}|z = k, \boldsymbol{\theta}) = p_k(\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta}_k)$. We can interpret mixture models as follows:

- We sample a specific component.
- We generate \mathbf{y} using sampled value of z .

Using the above procedure, we have:

basically, assume some z exists.
 $p(\mathbf{y}|z)$: simple } $\Rightarrow p(\mathbf{y})$: complex
 $p(z)$: simple }

$$p(\mathbf{y}|\boldsymbol{\theta}) = \sum_{k=1}^K p(z = k|\boldsymbol{\theta})p(\mathbf{y}|z = k, \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k p(\mathbf{y}|\boldsymbol{\theta}_k)$$

Gaussian Mixture Model

Gaussian Mixture Model

Gaussian Mixture Model (GMM) or Mixture of Gaussian (MoG) is defined as:

$$p(\mathbf{y}|\boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

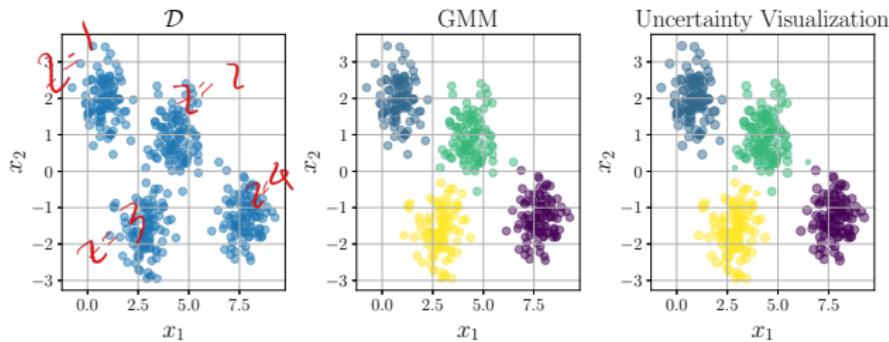
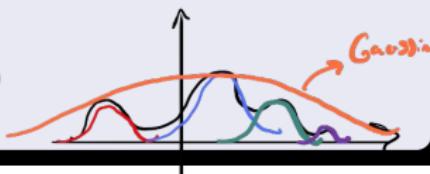


Figure: Sample GMM distribution and its application for clustering

Unsupervised clustering

References I

-  “Pearson correlation coefficient,”
https://en.wikipedia.org/wiki/Pearson_correlation_coefficient.
-  “The logic of causal conclusions: How we know that fire burns, fertilizer helps plants grow, and vaccines prevent disease,”
<http://icbseverywhere.com/blog/2014/10/the-logic-of-causal-conclusions/>.