# Probabilistic Inference and Learning
# Lecture 06
# Gaussian Probability Distributions

Philipp Hennig

04 May 2021

Eberhard Karls Universität Tübingen

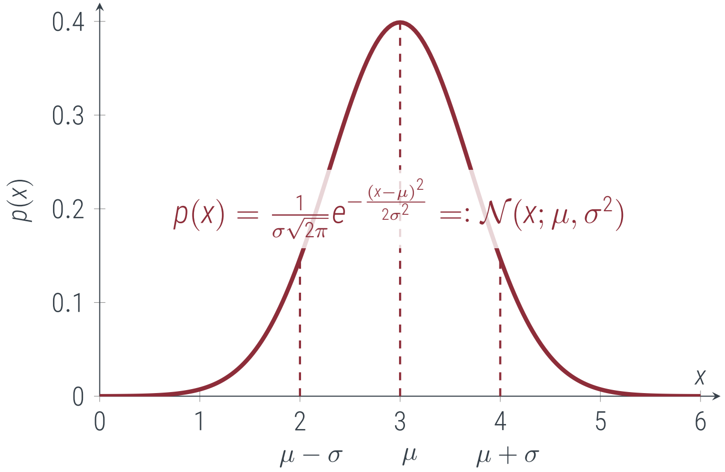
Faculty of Science
Department of Computer Science
Chair for the Methods of Machine Learning

| # | date | content | Ex | # | date | content | Ex |
|---|---|---|---|---|---|---|---|
| 1 | 20.04. | Introduction | 1 | 14 | 09.06. | Logistic Regression | 8 |
| 2 | 21.04. | Reasoning under Uncertainty | | 15 | 15.06. | Exponential Families | |
| 3 | 27.04. | Continuous Variables | 2 | 16 | 16.06. | Graphical Models | 9 |
| 4 | 28.04. | Monte Carlo | | 17 | 22.06. | Factor Graphs | |
| 5 | 04.05. | Markov Chain Monte Carlo | 3 | 18 | 23.06. | The Sum-Product Algorithm | 10 |
| **6** | **05.05.** | **Gaussian Distributions** | | 19 | 29.06. | Example: Topic Models | |
| 7 | 11.05. | Parametric Regression | 4 | 20 | 30.06. | Mixture Models | 11 |
| 8 | 12.05. | Understanding Deep Learning | | 21 | 06.07. | EM | |
| 9 | 18.05. | Gaussian Processes | 5 | 22 | 07.07. | Variational Inference | 12 |
| 10 | 19.05. | An Example for GP Regression | | 23 | 13.07. | Example: Topic Models | |
| 11 | 25.05. | Understanding Kernels | 6 | 24 | 14.07. | Example: Inferring Topics | 13 |
| 12 | 26.05. | Gauss-Markov Models | | 25 | 20.07. | Example: Kernel Topic Models | |
| 13 | 08.06. | GP Classification | 7 | 26 | 21.07. | Revision | |


DEUTSCHE BUNDESBANK
Banknote
10
ZEHN DEUTSCHE MARK
1777–1855 Carl Friedr. Gauß
$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
GN4480100S8

# The (univariate) Gaussian distribution

an exponentiated square




$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} =: \mathcal{N}(x; \mu, \sigma^2)$$

- $\mu$ the **mean** of $x$
- $\sigma^2$ the **variance** of $x$
- $\sigma$ the **standard deviation** of $x$

## Definition

$$\mathcal{N}(x;\mu,\sigma^2) =: \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \text{with } \mu,\sigma \in \mathbb{R}$$

will be called the **Gaussian** or **normal distribution** of $x$. We call $x$ the **argument** or **variable**, $\mu, \sigma^2$ the **parameters**. We write $x \sim \mathcal{N}(\mu,\sigma^2)$ to say that the variable $x$ is distributed with pdf $\mathcal{N}(x;\mu,\sigma^2)$.

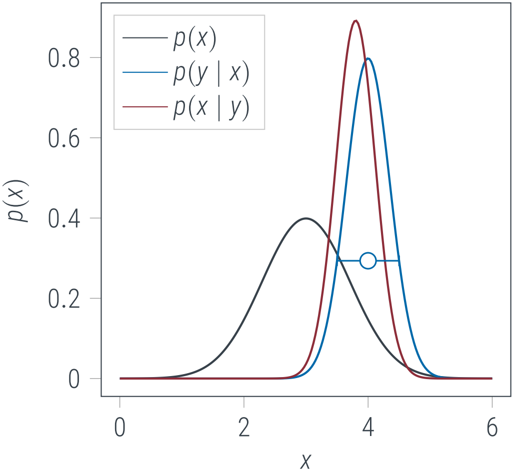
- $\int \mathcal{N}(x;\mu,\sigma^2)\,dx = 1$ and $\mathcal{N}(x;\mu,\sigma^2) > 0\ \forall x \in \mathbb{R}$. So $\mathcal{N}$ is the density of a probability measure.
- Symmetry in $x$ and $\mu$: $\mathcal{N}(x;\mu,\sigma^2) = \mathcal{N}(\mu;x,\sigma^2)$
- An **exponential of a quadratic polynomial** of the **natural parameters** $(a,\eta,\tau)$ :

$$\mathcal{N}(x;\mu,\sigma^2) = \exp\left(a + \eta x - \frac{1}{2}\tau x^2\right) \quad \text{with} \quad \tau = \sigma^{-2}\ (\text{“precision”}), \eta = \sigma^{-2}\mu$$

$$a = -\frac{1}{2}\left(\log(2\pi) - \log\lambda^2 + \lambda^2\eta^2\right)$$

# Gaussian Inference

The Gaussian is its own **conjugate prior**.

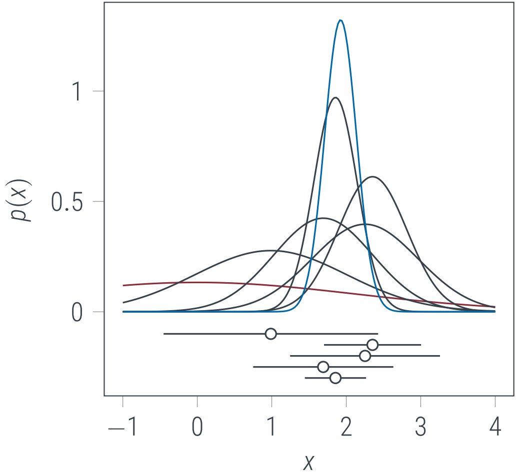



Let

$$p(x) = \mathcal{N}(x; \mu, \sigma^2)$$
$$p(y \mid x) = \mathcal{N}(y; x, \nu^2)$$

Then

$$p(x \mid y) = \frac{p(x)p(y \mid x)}{\int p(x)p(y \mid x)\, dx}$$
$$= \mathcal{N}(x; m, s^2), \text{ with}$$
$$s^2 := \frac{1}{\sigma^{-2} + \nu^{-2}}$$
$$m := \frac{\sigma^{-2}\mu + \nu^{-2} y}{\sigma^{-2} + \nu^{-2}}$$




$$p(x) = \mathcal{N}(x; \mu, \sigma^2)$$

$$p(\boldsymbol{y} \mid x) = \prod_{i=1}^{N} \mathcal{N}(y_i; x, \nu_i^2)$$

$$p(x \mid y) = \frac{p(x)p(\boldsymbol{y} \mid x)}{\int p(x)p(\boldsymbol{y} \mid x)\, dx}$$

$$= \mathcal{N}(x; m, s^2), \text{ with}$$

$$s^{-2} := \sigma^{-2} + \sum_{i=1}^{N} \nu_i^{-2}$$

$$s^{-2}m := \sigma^{-2}\mu + \sum_{i=1}^{N} \nu_i^{-2} y_i$$

If $\sigma^{-2} \to 0$, $\nu_i = \nu \; \forall i$, then $m$ is the **arithmetic mean**.

# The Method of Least Squares

The Gaussian distribution is the unique choice yielding a mean that is the mean of measurements.

[image: C.A. Jensen, 1840]

so wird allgemein sein müssen $\varphi'(M-p)+\varphi'(M'-p)+\varphi'(M''-p)+$ etc. $=0$, wenn für $p$ der Werth $\frac{1}{\mu}(M+M'+M''+\text{etc.})$ substituirt wird, welches positive Ganzes nun auch durch $\mu$ ausdrückt sein mag. Setzt man daher voraus $M' = M'' =$ etc. $= M-\mu N$, so wird allgemein, d. h. für jeden ganzen positiven Werth für $\mu$, sein $\varphi'(\mu-1)N=(1-\mu)\varphi'(-N)$, woraus man leicht sicht, dass allgemein $\frac{\varphi'\Delta}{\Delta}$ eine constante Grösse sein müsse, welche ich mit $k$ bezeichnen will. Hieraus wird $\log\varphi\Delta = \frac{1}{2}k\Delta\Delta +$ Const., oder wenn man die Basis der hyperbolischen Logarithmen mit $e$ bezeichnet und die Constante $= \log\varkappa$ setzt,

$$\varphi\Delta = \varkappa e^{\frac{1}{2}k\Delta\Delta}.$$

Ferner sieht man leicht ein, dass $k$ nothwendig negativ sein müsse, damit $\Omega$ in der That ein Grösstes werden könne, weshalb wir setzen $\frac{1}{2}k = -hh$; und da vermittelst des eleganten, zuerst von Laplace*) gefundenen Theorems das Integral $\int e^{-hh\Delta\Delta}\mathrm{d}\Delta$, von $\Delta = -\infty$ bis zu $\Delta = +\infty$, wird $= \frac{\sqrt{\pi}}{h}$ (wobei $\pi$ den halben Kreisumfang für den Radius $= 1$ bezeichnet), so wird unsere Function werden:

$$\varphi\Delta = \frac{h}{\sqrt{\pi}}e^{-hh\Delta\Delta}.$$

**178.**

Die so eben ermittelte Function kann zwar nicht in aller Strenge die Wahrscheinlichkeiten der Fehler ausdrücken; denn da die möglichen Fehler
(213) stets in gewisse Grenzen eingezwängt sind, so müsste die Wahrscheinlichkeit grösserer Fehler immer $= 0$ herauskommen, während unsere Formel stets einen begrenzten Werth darstellt. Dennoch aber ist dieser Mangel, an welchem jede analytische Function ihrer Natur nach laboriren muss, für jeden praktischen

*) In v. Zach „Monatliche Correspondenz“ Band 21, S. 280 äussert Gauss: „Dass Euler schon das Theorem gefunden hat, woraus der schöne, von mir Laplace beigelegte Lehrsatz sehr leicht abgeleitet werden kann, fiel mir selbst schon früher ein, als aber die Stelle S. 212 schon abgedruckt war; ich wollte es aber nicht unter die Errata setzen, weil Laplace wenigstens das obige Theorem doch erst in der dort gebrauchten Form aufgestellt hat.“ *Anmerkung des Uebersetzers.*

# The **Multivariate** Gaussian distribution

An exponentiated quadratic form

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

## Definition (multivariate Gaussian distribution)

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\intercal \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \qquad \boldsymbol{x}, \boldsymbol{\mu} \in \mathbb{R}^n, \Sigma \in \mathbb{R}^{n \times n}, \text{spd.}$$

$\Sigma$ must be **symmetric positive definite**.

# The **Multivariate** Gaussian distribution

An exponentiated quadratic form

## Definition (multivariate Gaussian distribution)

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\intercal \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right) \qquad \boldsymbol{x}, \boldsymbol{\mu} \in \mathbb{R}^n, \Sigma \in \mathbb{R}^{n \times n}, \text{spd.}$$

$\Sigma$ must be **symmetric positive definite**.

## Definition (symmetric positive definite matrix)

A matrix $A \in \mathbb{R}^{n \times n}$ is called **symmetric positive (semi-) definite** if $A = A^\intercal$, and

$$v^\intercal A v \geq 0 \; \forall v \in \mathbb{R}^n.$$

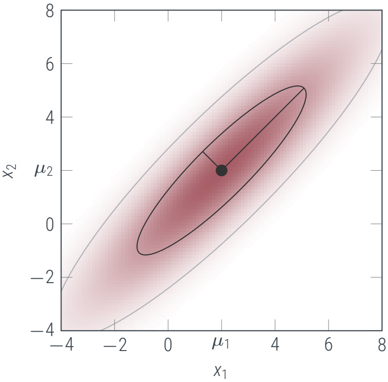
Equivalent statement: All eigenvalues of the symmetric matrix $A$ are non-negative.

# The **Multivariate** Gaussian distribution

Equiprobability lines are ellipsoids



$$\mathcal{N}(x; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\intercal \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \qquad \boldsymbol{x}, \boldsymbol{\mu} \in \mathbb{R}^n, \Sigma \in \mathbb{R}^{n\times n}, \text{spd.}$$




- $\int \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = 1$ and $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) > 0 \ \forall x \in \mathbb{R}^n$.
- Symmetry in $\boldsymbol{x}$ and $\boldsymbol{\mu}$: $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \mathcal{N}(\boldsymbol{\mu}; \boldsymbol{x}, \Sigma)$
- An exponential of a quadratic polynomial:

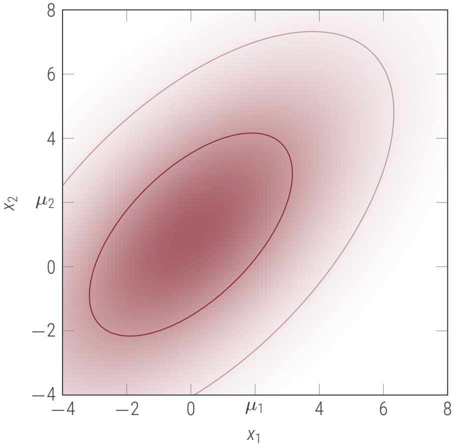
$$\mathcal{N}(x; \mu, \Sigma) = \exp\left(a + \eta^\intercal \boldsymbol{x} - \frac{1}{2}x^\intercal \Lambda x\right) \qquad (1)$$

$$= \exp\left(a + \eta^\intercal \boldsymbol{x} - \frac{1}{2}\operatorname{tr}(xx^\intercal \Lambda)\right) \qquad (2)$$

with the **natural parameters** $\Lambda = \Sigma^{-1}$ (precision matrix), $\eta = \Lambda\mu$, and the **sufficient statistics** $x, xx^\intercal$.

# Products of Gaussians are Gaussians

## Closure under Multiplication




**To multiply Gaussians, add the natural parameters**

$$\mathcal{N}(x; a, A)\mathcal{N}(x; b, B) = \mathcal{N}(x; c, C)Z$$

$$C = (A^{-1} + B^{-1})^{-1}$$

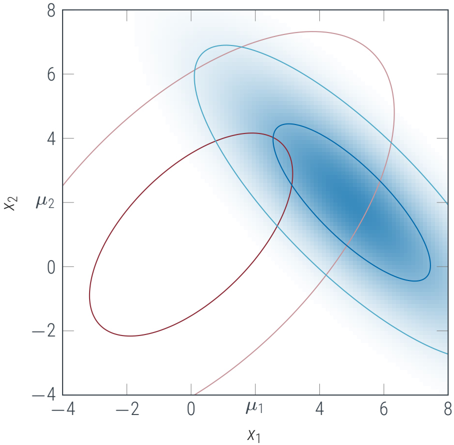
$$c = C(A^{-1}a + B^{-1}b)$$

$$Z = \mathcal{N}(a; b, A + B)$$

Note similarity to univariate case.

# Products of Gaussians are Gaussians

Closure under Multiplication




**To multiply Gaussians, add the natural parameters**

$$\mathcal{N}(x; a, A)\mathcal{N}(x; b, B) = \mathcal{N}(x; c, C)Z$$

$$C = (A^{-1} + B^{-1})^{-1}$$

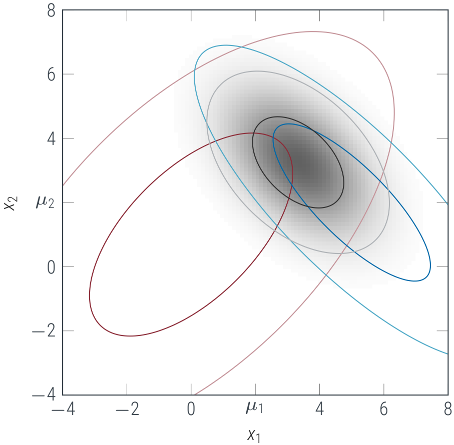
$$c = C(A^{-1}a + B^{-1}b)$$

$$Z = \mathcal{N}(a; b, A + B)$$

Note similarity to univariate case.

# Products of Gaussians are Gaussians

## Closure under Multiplication




**To multiply Gaussians, add the natural parameters**

$$\mathcal{N}(x; a, A)\mathcal{N}(x; b, B) = \mathcal{N}(x; c, C)Z$$

$$C = (A^{-1} + B^{-1})^{-1}$$

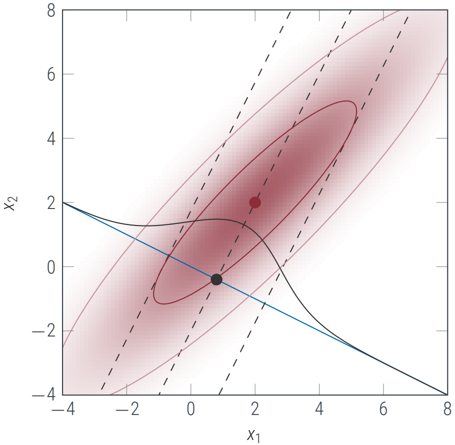
$$c = C(A^{-1}a + B^{-1}b)$$

$$Z = \mathcal{N}(a; b, A + B)$$

Note similarity to univariate case.

# Linear Projections of Gaussians are Gaussians

Closure under linear maps




**To linearly project a Gaussian variable, project the parameters**

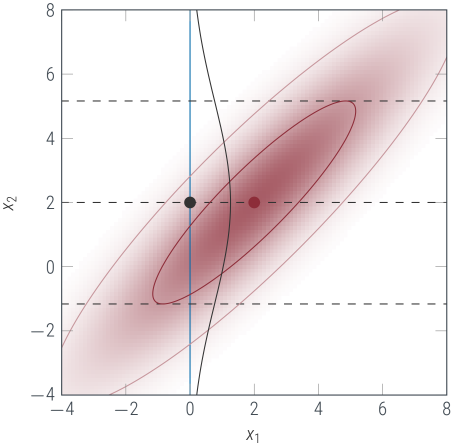
$$p(z) = \mathcal{N}(z; \mu, \Sigma)$$

$$\Rightarrow \quad p(Az) = \mathcal{N}(Az, A\mu, A\Sigma A^\intercal)$$

# Marginals of Gaussians are Gaussians

Closure under marginalization






$$p(z) = \mathcal{N}(z; \mu, \Sigma) \quad \Rightarrow \quad p(Az) = \mathcal{N}(Az, A\mu, A\Sigma A^\intercal)$$

$$\text{choose } A = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

$$\int \mathcal{N}\left[\begin{pmatrix} x \\ y \end{pmatrix}; \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}\right] dy = \mathcal{N}(x; \mu_x, \Sigma_{xx})$$

- this is the sum rule

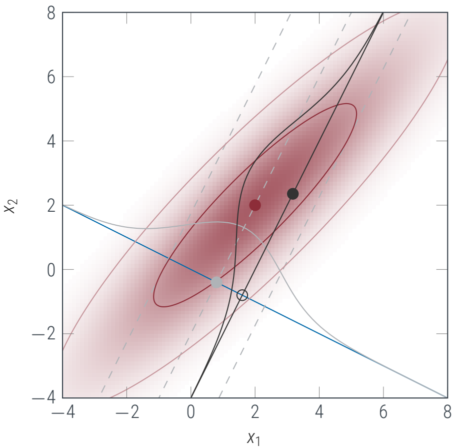
$$\int p(x,y)\,dy = \int p(y \mid x)p(x)\,dy = p(x)$$

- so every finite-dim Gaussian is a marginal of infinitely many more

# Cuts through Gaussians are Gaussians

## Closure under conditioning






$$p(x \mid Ax = y) = \frac{p(x,y)}{p(y)}$$

$$= \mathcal{N}\left(x; \mu + \Sigma A^\intercal (A\Sigma A^\intercal)^{-1}(y - A\mu), \Sigma - \Sigma A^\intercal (A\Sigma A^\intercal)^{-1} A\Sigma\right)$$

- this is the product rule
- so Gaussians are closed under the rules of probability

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

## Theorem

$$
\begin{aligned}
\text{If } p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \Sigma) \\
\text{and } p(\mathbf{y} \mid \mathbf{x}) &= \mathcal{N}(\mathbf{y}; A\mathbf{x} + \mathbf{b}, \Lambda), \\
\text{then } p(\mathbf{y}) &= \mathcal{N}(\mathbf{y}; A\boldsymbol{\mu} + \mathbf{b}, \Lambda + A\Sigma A^\intercal) \\
\text{and } p(\mathbf{x} \mid \mathbf{y}) &= \mathcal{N}(\mathbf{x}; \boldsymbol{\mu} + \underbrace{\Sigma A^\intercal (A\Sigma A^\intercal + \Lambda)^{-1}}_{\text{gain}} \underbrace{(\mathbf{y} - (A\boldsymbol{\mu} + \mathbf{b}))}_{\text{residual}}, \Sigma - \Sigma A^\intercal (\underbrace{A\Sigma A^\intercal + \Lambda}_{\text{Gram matrix}})^{-1} A\Sigma) \\
&= \mathcal{N}(\mathbf{x}; (\underbrace{\Sigma^{-1} + A^\intercal \Lambda^{-1} A}_{\text{precision matrix}})^{-1} (A^\intercal \Lambda^{-1} (\mathbf{y} - \mathbf{b}) + \Sigma^{-1} \boldsymbol{\mu}), (\underbrace{\Sigma^{-1} + A^\intercal \Lambda^{-1} A}_{\text{precision matrix}})^{-1})
\end{aligned}
$$

# The Core Insight for All of This

Gaussian inference is linear algebra at its core

[image: Konrad Jacobs]

$$A = \begin{bmatrix} P & Q \\ R & S \end{bmatrix} \qquad M := (S - RP^{-1}Q)^{-1}$$

$$A^{-1} = \begin{bmatrix} P^{-1} + P^{-1}QMRP^{-1} & -P^{-1}QM \\ -MRP^{-1} & M \end{bmatrix}$$

$$(Z + UWV^\intercal)^{-1} = Z^{-1} - Z^{-1}U(W^{-1} + V^\intercal Z^{-1}U)^{-1}V^\intercal Z^{-1}$$

$$|Z + UWV^\intercal| = |Z| \cdot |W| \cdot |W^{-1} + V^\intercal Z^{-1}U|$$

Issai Schur (1875–1941)

# Gaussians provide the **linear algebra of inference**

if all joints are Gaussian and all observations are linear, all posteriors are Gaussian



- products of Gaussians are Gaussians

$$\begin{aligned}\mathcal{N}(x; a, A)\mathcal{N}(x; b, B)\\ = \mathcal{N}(x; c, C)\mathcal{N}(a; b, A + B)\end{aligned}$$

$$C := (A^{-1} + B^{-1})^{-1} \qquad c := C(A^{-1}a + B^{-1}b)$$

- linear projections of Gaussians are Gaussians

$$\begin{aligned} p(z) &= \mathcal{N}(z; \mu, \Sigma)\\ \Rightarrow \quad p(Az) &= \mathcal{N}(Az, A\mu, A\Sigma A^\intercal)\end{aligned}$$

- marginals of Gaussians are Gaussians

$$\int \mathcal{N}\left[\begin{pmatrix} x \\ y \end{pmatrix}; \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}\right] dy = \mathcal{N}(x; \mu_x, \Sigma_{xx})$$

- (linear) conditionals of Gaussians are Gaussians

$$\begin{aligned} p(x \mid y) &= \frac{p(x, y)}{p(y)}\\ &= \mathcal{N}\left(x; \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y), \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\right)\end{aligned}$$

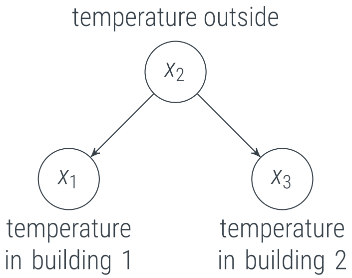
**Bayesian inference becomes linear algebra**

If $p(x) = \mathcal{N}(x; \mu, \Sigma)$ and $p(y \mid x) = \mathcal{N}(y; A^\intercal x + b, \Lambda)$, then

$$p(B^\intercal x + c \mid y) = \mathcal{N}[B^\intercal x + c; B^\intercal \mu + c + B^\intercal \Sigma A(A^\intercal \Sigma A + \Lambda)^{-1}(y - A^\intercal \mu - b), B^\intercal \Sigma B - B^\intercal \Sigma A(A^\intercal \Sigma A + \Lambda)^{-1} A^\intercal \Sigma B]$$

# Example 1: Conditional Independence, Marginal Correlation

[DJC MacKay, *The humble Gaussian distribution*, 2006]

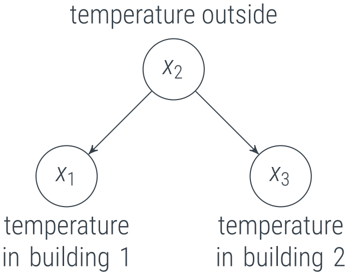



$$x_2 = \nu_2 \qquad p(\nu_2) = \mathcal{N}(\nu_2; \mu_2, \sigma_2^2)$$
$$x_1 = w_1 x_2 + \nu_1 \qquad p(\nu_1) = \mathcal{N}(\nu_1; \mu_1, \sigma_1^2)$$
$$x_3 = w_3 x_2 + \nu_3 \qquad p(\nu_3) = \mathcal{N}(\nu_3; \mu_3, \sigma_3^2)$$

# Example 1: Conditional Independence, Marginal Correlation



[DJC MacKay, *The humble Gaussian distribution*, 2006]




$$x_2 = \nu_2 \qquad p(\nu_2) = \mathcal{N}(\nu_2; \mu_2, \sigma_2^2)$$

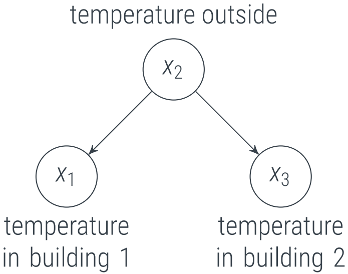
$$x_1 = w_1 x_2 + \nu_1 \qquad p(\nu_1) = \mathcal{N}(\nu_1; \mu_1, \sigma_1^2)$$

$$x_3 = w_3 x_2 + \nu_3 \qquad p(\nu_3) = \mathcal{N}(\nu_3; \mu_3, \sigma_3^2)$$

$$p(\boldsymbol{\nu}) = \mathcal{N}(\boldsymbol{\nu}; \boldsymbol{\mu}, \operatorname{diag}(\boldsymbol{\sigma}^2))$$

$$A = \begin{bmatrix} 1 & w_1 & 0 \\ 0 & 1 & 0 \\ 0 & w_3 & 1 \end{bmatrix} \qquad \Longrightarrow$$

$$p(\boldsymbol{x} = A\nu) = \mathcal{N}\left(\boldsymbol{x}; \underbrace{A\boldsymbol{\mu}}_{=:m}, \underbrace{\begin{bmatrix} w_1\sigma_2^2 + \sigma_1^2 & w_1\sigma_2^2 & w_1 w_3 \sigma_2^2 \\ & \sigma_2^2 & w_3\sigma_2^2 \\ & & w_3^2\sigma_2^2 + \sigma_3^2 \end{bmatrix}}_{=:\Sigma}\right)$$




$$x_2 = \nu_2 \qquad p(\nu_2) = \mathcal{N}(\nu_2; \mu_2, \sigma_2^2)$$
$$x_1 = w_1 x_2 + \nu_1 \qquad p(\nu_1) = \mathcal{N}(\nu_1; \mu_1, \sigma_1^2)$$
$$x_3 = w_3 x_2 + \nu_3 \qquad p(\nu_3) = \mathcal{N}(\nu_3; \mu_3, \sigma_3^2)$$

$$p(\boldsymbol{\nu}) = \mathcal{N}(\boldsymbol{\nu}; \boldsymbol{\mu}, \operatorname{diag}(\boldsymbol{\sigma}^2))$$

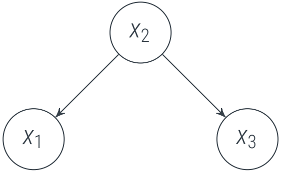
$$A = \begin{bmatrix} 1 & w_1 & 0 \\ 0 & 1 & 0 \\ 0 & w_3 & 1 \end{bmatrix} \qquad \Longrightarrow$$

$$p(\boldsymbol{x} = A\nu) = \mathcal{N}\left(\boldsymbol{x}; \underbrace{A\boldsymbol{\mu}}_{=:m}, \underbrace{\begin{bmatrix} w_1\sigma_2^2 + \sigma_1^2 & w_1\sigma_2^2 & w_1 w_3 \sigma_2^2 \\ & \sigma_2^2 & w_3\sigma_2^2 \\ & & w_3^2\sigma_2^2 + \sigma_3^2 \end{bmatrix}}_{=:\Sigma}\right)$$

From graph: $x_1 \perp\!\!\!\perp x_3 \mid x_2$. Where can we see this in the pdf?

# Example 1: Conditional Independence, Marginal Correlation

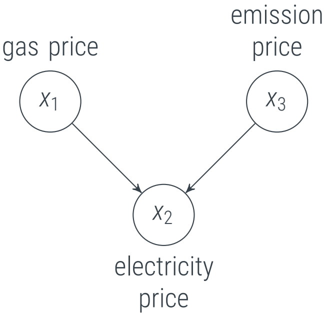
A zero in the precision matrix means independence conditional on everything else [DJC MacKay, *The humble Gaussian distribution*, 2006]




$$x_2 = \nu_2 \qquad p(\nu_2) = \mathcal{N}(\nu_2; \mu_2, \sigma_2^2)$$
$$x_1 = w_1 x_2 + \nu_1 \qquad p(\nu_1) = \mathcal{N}(\nu_1; \mu_1, \sigma_1^2)$$
$$x_3 = w_3 x_2 + \nu_3 \qquad p(\nu_3) = \mathcal{N}(\nu_3; \mu_3, \sigma_3^2)$$

to simplify exposition, set $\boldsymbol{\mu} = 0$.

$$p(x_1, x_2, x_3) = p(x_2) \cdot p(x_1 \mid x_2) \cdot p(x_3 \mid x_2)$$
$$= \frac{1}{Z_1 Z_2 Z_3} \exp\left(-\frac{1}{2}\left(\frac{x_2^2}{\sigma_2^2} + \frac{(x_1 - w_1 x_2)^2}{\sigma_1^2} + \frac{(x_3 - w_3 x_2)^2}{\sigma_3^2}\right)\right)$$
$$= \frac{1}{Z_1 Z_2 Z_3} \exp\left(-\frac{1}{2}\left(x_2^2\left(\frac{1}{\sigma_2^2} + \frac{w_1^2}{\sigma_1^2} + \frac{w_3^2}{\sigma_3^2}\right) + x_1^2 \frac{1}{\sigma_1^2} - 2x_1 x_2 \frac{w_1}{\sigma_1^2} + x_3^2 \frac{1}{\sigma_3^2} - 2x_3 x_2 \frac{w_3}{\sigma_3^2}\right)\right)$$
$$= \frac{1}{Z_1 Z_2 Z_3} \exp\left(-\frac{1}{2}\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_1^2} & -\frac{w_1}{\sigma_1^2} & 0 \\ -\frac{w_1}{\sigma_1^2} & \left(\frac{1}{\sigma_2^2} + \frac{w_1^2}{\sigma_1^2} + \frac{w_3^2}{\sigma_3^2}\right) & -\frac{w_3}{\sigma_3^2} \\ 0 & -\frac{w_3}{\sigma_3^2} & \frac{1}{\sigma_3^2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\right)$$




$$x_1 = \nu_1 \qquad p(\nu_1) = \mathcal{N}(\nu_1; \mu_1, \sigma_1^2)$$

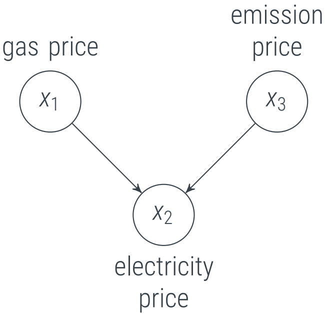
$$x_3 = \nu_3 \qquad p(\nu_3) = \mathcal{N}(\nu_3; \mu_3, \sigma_3^2)$$

$$x_2 = w_1 x_1 + w_3 x_3 + \nu_2 \qquad p(\nu_2) = \mathcal{N}(\nu_2; \mu_2, \sigma_2^2)$$

# Example 2: Explaining away

Bayesian Inference with Gaussians

[DJC MacKay, *The humble Gaussian distribution*, 2006]




$$p(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{m}, \underbrace{\begin{bmatrix} \sigma_1^2 & w_1\sigma_1^2 & 0 \\ & \sigma_2^2 + w_1^2\sigma_1^2 + w_3^2\sigma_3^2 & w_3\sigma_3^2 \\ & & \sigma_3^2 \end{bmatrix}}_{\Sigma}\right)$$

$$p(x_1, x_3) = \mathcal{N}\left(\begin{bmatrix} x_1 \\ x_3 \end{bmatrix}; \begin{bmatrix} \mu_1 \\ \mu_3 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_3^2 \end{bmatrix}\right)$$

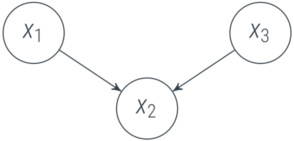
$$x_1 = \nu_1 \qquad p(\nu_1) = \mathcal{N}(\nu_1; \mu_1, \sigma_1^2)$$

$$x_3 = \nu_3 \qquad p(\nu_3) = \mathcal{N}(\nu_3; \mu_3, \sigma_3^2)$$

$$x_2 = w_1 x_1 + w_3 x_3 + \nu_2 \qquad p(\nu_2) = \mathcal{N}(\nu_2; \mu_2, \sigma_2^2)$$

# Example 2: Explaining away

a $\pm$ value in the precision matrix implies $\mp$ correlation conditional on everything else   [DJC MacKay, *The humble Gaussian distribution*, 2006]




$$x_1 = \nu_1 \qquad p(\nu_1) = \mathcal{N}(\nu_1; \mu_1, \sigma_1^2)$$

$$x_3 = \nu_3 \qquad p(\nu_3) = \mathcal{N}(\nu_3; \mu_3, \sigma_3^2)$$

$$x_2 = w_1 x_1 + w_3 x_3 + \nu_2 \qquad p(\nu_2) = \mathcal{N}(\nu_2; \mu_2, \sigma_2^2)$$

$$p(x_1, x_2, x_3) = p(x_1) \cdot p(x_3) \cdot p(x_2 \mid x_1, x_3)$$

$$= \frac{1}{Z_1 \cdot Z_2 \cdot Z_3} \exp\left(-\frac{1}{2}\left(\frac{x_1}{\sigma_1^2} + \frac{x_3}{\sigma_3^2} + \frac{x_2 - w_1 x_1 - w_3 x_3}{\sigma_2^2}\right)\right)$$

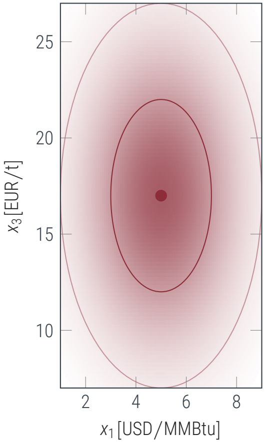
$$= \frac{1}{Z_1 \cdot Z_2 \cdot Z_3} \exp\left(-\frac{1}{2}\left(x_1^2\left(\frac{1}{\sigma_1^2} + \frac{w_1^2}{\sigma_2^2}\right) + x_2^2 \frac{1}{\sigma_2^2} - 2x_1 x_2 \frac{w_1}{\sigma_2^2} + x_3^2\left(\frac{1}{\sigma_3^2} + \frac{w_3^2}{\sigma_2^2}\right) - 2x_2 x_3 \frac{w_3}{\sigma_2^2} + 2x_3 x_1 \frac{w_1 w_3}{\sigma_2^2}\right)\right)$$

$$= \frac{1}{Z_1 \cdot Z_2 \cdot Z_3} \exp\left(-\frac{1}{2}\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} \left(\frac{1}{2\sigma_1^2} + \frac{w_1^2}{\sigma_2^2}\right) & -\frac{w_1}{\sigma_2^2} & \frac{w_1 w_3}{\sigma_2^2} \\ -\frac{w_1}{\sigma_2^2} & \frac{1}{\sigma_2^2} & -\frac{w_3}{\sigma_2^2} \\ \frac{w_1 w_3}{\sigma_2^2} & -\frac{w_3}{\sigma_2^2} & \left(\frac{1}{2\sigma_3^2} + \frac{w_3^2}{\sigma_2^2}\right) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\right)$$

# Example 2: Explaining away

Bayesian Inference with Gaussians

[DJC MacKay, *The humble Gaussian distribution*, 2006]

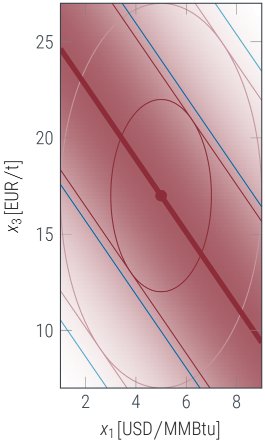



$$p(x_1, x_3) = \mathcal{N}\left(\begin{bmatrix} x_1 \\ x_3 \end{bmatrix}; \begin{bmatrix} \mu_1 \\ \mu_3 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_3^2 \end{bmatrix}\right)$$

# Example 2: Explaining away

Bayesian Inference with Gaussians

[DJC MacKay, *The humble Gaussian distribution*, 2006]



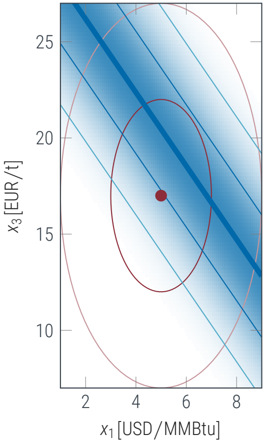

$$p(x_1, x_3) = \mathcal{N}\left(\begin{bmatrix} x_1 \\ x_3 \end{bmatrix}; \begin{bmatrix} \mu_1 \\ \mu_3 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_3^2 \end{bmatrix}\right)$$

$$p(x_2) = \mathcal{N}\left(x_2; w_1\mu_1 + w_3\mu_3 + \mu_2, \sigma_2^2 + w_1^2\sigma_1^2 + w_3^2\sigma_3^2\right)$$

# Example 2: Explaining away

Bayesian Inference with Gaussians

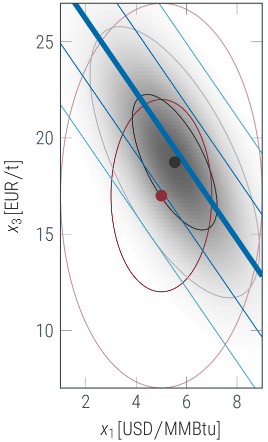
[DJC MacKay, *The humble Gaussian distribution*, 2006]




$$p(x_1, x_3) = \mathcal{N}\left(\begin{bmatrix} x_1 \\ x_3 \end{bmatrix}; \begin{bmatrix} \mu_1 \\ \mu_3 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_3^2 \end{bmatrix}\right)$$

$$p(x_2) = \mathcal{N}\left(x_2; w_1\mu_1 + w_3\mu_3 + \mu_2, \sigma_2^2 + w_1^2\sigma_1^2 + w_3^2\sigma_3^2\right)$$

$$p(x_2 \mid x_1, x_3) = \mathcal{N}(x_2; w_1 x_1 + w_3 x_3 + \mu_2, \sigma_2^2)$$




$$p(x_1, x_3) = \mathcal{N}\left(\begin{bmatrix} x_1 \\ x_3 \end{bmatrix}; \begin{bmatrix} \mu_1 \\ \mu_3 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_3^2 \end{bmatrix}\right)$$

$$p(x_2) = \mathcal{N}\left(x_2; w_1\mu_1 + w_3\mu_3 + \mu_2, \sigma_2^2 + w_1^2\sigma_1^2 + w_3^2\sigma_3^2\right)$$

$$p(x_2 \mid x_1, x_3) = \mathcal{N}(x_2; w_1 x_1 + w_3 x_3 + \mu_2, \sigma_2^2)$$

$$p(x_1, x_3 \mid x_2) = \mathcal{N}\left(\boldsymbol{x}_{1,3}; \boldsymbol{\mu}_{1,3} - \Sigma_{1,3}\boldsymbol{w}^\intercal \frac{x_2 - \boldsymbol{w}\boldsymbol{\mu}_{1,3} - \mu_2}{\boldsymbol{w}\Sigma_{1,3}\boldsymbol{w}^\intercal + \sigma_2^2},\right.$$

$$\left.\Sigma_{1,3} - \Sigma_{1,3}\boldsymbol{w}^\intercal \frac{1}{\boldsymbol{w}\Sigma_{1,3}\boldsymbol{w}^\intercal + \sigma_2^2}\boldsymbol{w}\Sigma_{1,3}\right)$$

$$= \mathcal{N}\left(\begin{bmatrix} x_1 \\ x_3 \end{bmatrix}; \begin{bmatrix} \mu_1 \\ \mu_3 \end{bmatrix} - \begin{bmatrix} w_1\sigma_1^2 \\ w_3\sigma_3^2 \end{bmatrix} \frac{x_2 - w_1\mu_1 - w_3\mu_3 - \mu_2}{w_1^2\sigma_1^2 + w_3^2\sigma_3^2 + \sigma_2^2},\right.$$

$$\left.\begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_3^2 \end{bmatrix} - \begin{bmatrix} w_1\sigma_1^2 \\ w_3\sigma_3^2 \end{bmatrix} \frac{1}{w_1^2\sigma_1^2 + w_3^2\sigma_3^2 + \sigma_2^2} \begin{bmatrix} w_1\sigma_1^2 & w_3\sigma_3^2 \end{bmatrix}\right)$$

$$\mathcal{N}(x;\boldsymbol{\mu},\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\boldsymbol{\mu})^\intercal \Sigma^{-1}(x-\boldsymbol{\mu})\right)$$

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Today:

- **Gaussian distributions** provide the **linear algebra of inference**.
  - products of Gaussians are Gaussians
  - linear maps of Gaussian variables are Gaussian variables
  - marginals of Gaussians are Gaussians
  - linear conditionals of Gaussians are Gaussians

If all variables in a generative model are **linearly** related, and the distributions of the parent variables are Gaussian, then all conditionals, joints and marginals are Gaussian, with means and covariances computable by linear algebra operations.

- A zero off-diagonal element in the **covariance** matrix implies independence if all other variables are integrated out
- A zero off-diagonal element in the **precision** matrix implies independence conditional on all other variables

$$[\Sigma]_{ij} = 0 \quad \Rightarrow \quad p(x_i, x_j) = \mathcal{N}(x_i; [\boldsymbol{\mu}]_i, [\Sigma]_{ii}) \cdot \mathcal{N}(x_j; [\boldsymbol{\mu}]_j, [\Sigma]_{jj})$$

$$[\Sigma^{-1}]_{ij} = 0 \quad \Rightarrow \quad p(x_i, x_j \mid x_{\neq i,j}) = \mathcal{N}(x_i \mid x_{\neq i,j}) \cdot \mathcal{N}(x_j \mid x_{\neq i,j})$$

# The Toolbox

## Framework:

$$\int p(x_1, x_2)\, dx_2 = p(x_1) \qquad p(x_1, x_2) = p(x_1 \mid x_2) p(x_2) \qquad p(x \mid y) = \frac{p(y \mid x) p(x)}{p(y)}$$

## Modelling:

- Directed Graphical Models
- Gaussian Distributions
-
-
-
-

## Computation:

- Monte Carlo
- Linear algebra / Gaussian inference
-
-
-