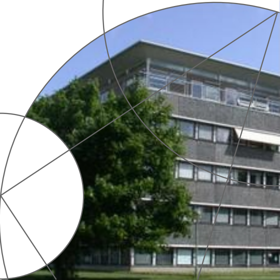
Faculty of Science

Gaussian distributions and graphical models

Niels Richard Hansen
Department of Mathematical Sciences



The missing definition

The book is missing a definition:

Definition (Belongs to Section 2.1.7)

If X and Y are two continuous random variables we define their **covariance** as

$$\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y).$$

Their **correlation** is

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{V(X)V(Y)}}.$$

Proposition 2.5 implies that if $X \perp Y$ then $\mathrm{Cov}(X, Y) = 0$.

The following rules for computing with covariances apply:

$$\begin{aligned} \mathrm{Cov}(X, Y) &= \mathrm{Cov}(Y, X) \\ \mathrm{Cov}(cX + bZ + a, Y) &= c\mathrm{Cov}(X, Y) + b\mathrm{Cov}(Z, Y) \\ \mathrm{V}(X + Y) &= \mathrm{V}(X) + \mathrm{V}(Y) + 2\mathrm{Cov}(X, Y) \end{aligned}$$

The multivariate Gaussian distribution

Let $\mu \in \mathbb{R}^n$ be a vector and let Σ be an $n \times n$ matrix. We call Σ **positive definite** if it is symmetric, $\Sigma^T = \Sigma$, and

$$x^T \Sigma x > 0$$

for all $x \neq 0$.

Definition

The multivariate Gaussian distribution parametrized by μ and a positive definite Σ is the distribution with density

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} \det(\Sigma)^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)}$$

The multivariate Gaussian distribution

Let $\mu \in \mathbb{R}^n$ be a vector and let Σ be an $n \times n$ matrix. We call Σ **positive definite** if it is symmetric, $\Sigma^T = \Sigma$, and

$$x^T \Sigma x > 0$$

for all $x \neq 0$.

Definition

The multivariate Gaussian distribution parametrized by μ and a positive definite Σ is the distribution with density

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} \det(\Sigma)^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)}$$

That Σ is positive definite is equivalent to (cf. Cholesky decomposition)

$$J = \Sigma^{-1} = C^T C$$

for a full rank $n \times n$ matrix C, and from the last slide on Monday 1/3 this shows that p is a probability density on $\mathbb{R}^n$.

Gaussian distributions for $n = 2$

For $n = 2$ with

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}; \begin{pmatrix} \sigma_1^2 & \gamma \\ \gamma & \sigma_2^2 \end{pmatrix}\right)$$

We can write out the density as

$$p(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2 - \gamma^2}} \exp\left(-\frac{(x_1-\mu_1)^2\sigma_2^2 + (x_2-\mu_2)^2\sigma_1^2 - 2(x_1-\mu_1)(x_2-\mu_2)\gamma}{2(\sigma_1^2\sigma_2^2 - \gamma^2)}\right)$$

Gaussian distributions for $n = 2$

The case $\gamma = 0$:

$$p(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} \exp\left(-\frac{x_1^2\sigma_2^2 + x_2^2\sigma_1^2}{2\sigma_1^2\sigma_2^2}\right)$$

Mean, variance and covariance in the bivariate Gaussian distribution:

See also `https://www2.stat.duke.edu/courses/Spring12/sta104.1/Lectures/Lec22.pdf`

Marginalization

If $(\mathbf{X}, \mathbf{Y})$ is multivariate Gaussian we write their parameters in block form:

$$\mu = \begin{pmatrix} \mu_{\mathbf{X}} \\ \mu_{\mathbf{Y}} \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_{\mathbf{XX}} & \Sigma_{\mathbf{XY}} \\ \Sigma_{\mathbf{YX}} & \Sigma_{\mathbf{YY}} \end{pmatrix}.$$

Lemma (7.1, p. 250)

The marginal distributions of $\mathbf{X}$ *and* $\mathbf{Y}$ *are*

$$\mathbf{X} \sim \mathcal{N}(\mu_{\mathbf{X}}, \Sigma_{\mathbf{XX}}) \quad \mathbf{Y} \sim \mathcal{N}(\mu_{\mathbf{Y}}, \Sigma_{\mathbf{YY}}).$$

Proof:

Marginalization

A special case of the above result is to consider the distribution of just two variables, X_i and X_j, from $\mathbf{X}$ with $\mathbf{X} \sim \mathcal{N}(\mu; \Sigma)$:

$$\begin{pmatrix} X_i \\ X_j \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_i \\ \mu_j \end{pmatrix}; \begin{pmatrix} \Sigma_{i,i} & \Sigma_{i,j} \\ \Sigma_{j,i} & \Sigma_{j,j} \end{pmatrix} \right)$$

Marginalization

A special case of the above result is to consider the distribution of just two variables, X_i and X_j, from $\mathbf{X}$ with $\mathbf{X} \sim \mathcal{N}(\mu; \Sigma)$:

$$\begin{pmatrix} X_i \\ X_j \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_i \\ \mu_j \end{pmatrix}; \begin{pmatrix} \Sigma_{i,i} & \Sigma_{i,j} \\ \Sigma_{j,i} & \Sigma_{j,j} \end{pmatrix} \right)$$

Thus

$$\begin{aligned} E(X_i) &= \mu_i \\ E(X_j) &= \mu_j \\ V(X_i) &= \Sigma_{i,i} \\ V(X_j) &= \Sigma_{j,j} \\ Cov(X_i, X_j) &= \Sigma_{i,j} = \Sigma_{j,i}, \end{aligned}$$

Marginalization

A special case of the above result is to consider the distribution of just two variables, X_i and X_j, from $\mathbf{X}$ with $\mathbf{X} \sim \mathcal{N}(\mu; \Sigma)$:

$$\begin{pmatrix} X_i \\ X_j \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_i \\ \mu_j \end{pmatrix}; \begin{pmatrix} \Sigma_{i,i} & \Sigma_{i,j} \\ \Sigma_{j,i} & \Sigma_{j,j} \end{pmatrix}\right)$$

Thus

$$\begin{aligned} E(X_i) &= \mu_i \\ E(X_j) &= \mu_j \\ V(X_i) &= \Sigma_{i,i} \\ V(X_j) &= \Sigma_{j,j} \\ Cov(X_i, X_j) &= \Sigma_{i,j} = \Sigma_{j,i}, \end{aligned}$$

and $X_i \perp X_j$ if and only if $\Sigma_{i,j} = 0$ (Theorem 7.1).

Gaussian Markov random fields

A completely different place to start is by factors of the form

$$\varphi_{ij}(x_i, x_j) = e^{-J_{i,j} x_i x_j}$$

for $i \neq j$ and

$$\varphi_{ii}(x_i, x_i) = e^{-\frac{1}{2} J_{i,i} x_i^2 + h_i x_i}.$$

Gaussian Markov random fields

A completely different place to start is by factors of the form

$$\varphi_{ij}(x_i, x_j) = e^{-J_{i,j}x_i x_j}$$

for $i \neq j$ and

$$\varphi_{ii}(x_i, x_i) = e^{-\frac{1}{2}J_{i,i}x_i^2 + h_i x_i}.$$

Then

$$p(\mathbf{x}) \propto \prod_{\{i,j\}:i\neq j} e^{-J_{i,j}x_i x_j} \prod_i e^{-\frac{1}{2}J_{i,i}x_i^2 + h_i x_i}$$

factorizes over the Markov network structure with i—j if and only if $J_{i,j} \neq 0$ for $i \neq j$.

Gaussian Markov random fields

A completely different place to start is by factors of the form

$$\varphi_{ij}(x_i, x_j) = e^{-J_{i,j}x_i x_j}$$

for $i \neq j$ and

$$\varphi_{ii}(x_i, x_i) = e^{-\frac{1}{2}J_{i,i}x_i^2 + h_i x_i}.$$

Then

$$\begin{aligned} p(\mathbf{x}) &\propto \prod_{\{i,j\}:i\neq j} e^{-J_{i,j}x_i x_j} \prod_i e^{-\frac{1}{2}J_{i,i}x_i^2 + h_i x_i} \\ &= \exp\left(-\frac{1}{2}\left(\sum_{i,j:i\neq j} J_{i,j}x_i x_j + \sum_i J_{i,i}x_i^2\right) + \sum_i h_i x_i\right) \end{aligned}$$

factorizes over the Markov network structure with i—j if and only if $J_{i,j} \neq 0$ for $i \neq j$.

Gaussian Markov random fields

A completely different place to start is by factors of the form

$$\varphi_{ij}(x_i, x_j) = e^{-J_{i,j} x_i x_j}$$

for $i \neq j$ and

$$\varphi_{ii}(x_i, x_i) = e^{-\frac{1}{2} J_{i,i} x_i^2 + h_i x_i}.$$

Then

$$\begin{aligned} p(\mathbf{x}) &\propto \prod_{\{i,j\}: i \neq j} e^{-J_{i,j} x_i x_j} \prod_i e^{-\frac{1}{2} J_{i,i} x_i^2 + h_i x_i} \\ &= \exp\left(-\frac{1}{2}\left(\sum_{i,j: i \neq j} J_{i,j} x_i x_j + \sum_i J_{i,i} x_i^2\right) + \sum_i h_i x_i\right) \\ &= \exp\left(-\frac{1}{2}\mathbf{x}^T J \mathbf{x} + \mathbf{h}^T \mathbf{x}\right) \end{aligned}$$

factorizes over the Markov network structure with i—j if and only if $J_{i,j} \neq 0$ for $i \neq j$.

Gaussian Markov random fields

For the factorization to define a probability density, the partition function has to be finite;

$$Z(J, \mathbf{h}) = \int \exp\left(-\frac{1}{2}\mathbf{x}^T J \mathbf{x} + \mathbf{h}^T \mathbf{x}\right) \, \mathrm{d}\mathbf{x} < \infty.$$

- The markov random field defined by J and $\mathbf{h}$ has pairwise potentials
- The pairwise energies (log-potentials) are the quadratic functions

  $$\epsilon_{i,j}(x_i, x_j) = -J_{i,j} x_i x_j$$

  and the node log-potentials are the quadratic functions

  $$\epsilon_i(x_i) = -\frac{1}{2} J_{i,i} x_i^2 + h_i x_i$$

- All quadratic log-potentials are of this form for a symmetric J and a vector $\mathbf{h}$

Connecting the dots ...

Starting with a Gaussian distribution,

$$
\begin{aligned}
\log p(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu) - \frac{1}{2}\log\left((2\pi)^n \det(\Sigma)\right) \\
&=
\end{aligned}
$$

Connecting the dots ...

Starting with a Gaussian distribution,

$$\begin{aligned}\log p(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu) - \frac{1}{2}\log\left((2\pi)^n \det(\Sigma)\right)\\ &= \; -\frac{1}{2}\mathbf{x}^T J\mathbf{x} + \mathbf{h}^T\mathbf{x} - \frac{1}{2}\mu^T J\mu + \frac{1}{2}\log\left((2\pi)^{-n}\det(J)\right)\end{aligned}$$

where $J = \Sigma^{-1}$ and $\mathbf{h} = J\mu = \Sigma^{-1}\mu$.

Proposition (not explicit in book)

If J^{-1} is **positive definite** *the pairwise factorization gives a well defined probability distribution, which is a Gaussian distribution with parameters $\mu = J^{-1}\mathbf{h}$ and $\Sigma = J^{-1}$. Moreover,*

$$Z(J,\mathbf{h}) = \frac{(2\pi)^{n/2}}{\det(J)^{1/2}} e^{\frac{1}{2}\mathbf{h}^T J^{-1}\mathbf{h}}.$$

Positive definite J

It holds that $\Sigma = J^{-1}$ is positive definite if and only if J is positive definite, but otherwise there is no simple way to determine if the local pairwise interactions $J_{i,j}$ give a valid Gaussian Markov random field.

Proposition (7.2, p. 256)

If J is **diagonally dominant**, *that is*

$$\sum_{j \neq i} |J_{i,j}| < J_{i,i}$$

for all $i = 1, \ldots, n$, then J gives a valid Gaussian Markov random field.

Conditional independence

It follows from Theorem 4.2 that:

Theorem (7.2, p. 251)

Let $\mathbf{X}$ *follow a multivariate Gaussian distribution with* $J = \Sigma^{-1}$. *Then* $X_i \perp X_j \mid \mathbf{X} \backslash \{X_i, X_j\}$ *if and only if* $J_{ij} = 0$.

Conditioning

If $(\mathbf{X}, \mathbf{Y})$ is multivariate Gaussian we write their **information parameters** in block form:

$$\mathbf{h} = \begin{pmatrix} \mathbf{h}_{\mathbf{X}} \\ \mathbf{h}_{\mathbf{Y}} \end{pmatrix} \quad J = \begin{pmatrix} J_{\mathbf{XX}} & J_{\mathbf{XY}} \\ J_{\mathbf{YX}} & J_{\mathbf{YY}} \end{pmatrix}.$$

Proposition (not explicit in book)

The **conditional** *distribution of* $\mathbf{Y}$ *given* $\mathbf{X} = \mathbf{x}$ *is Gaussian with information parameters*

$$\mathbf{h}_{\mathbf{x}} = \mathbf{h}_{\mathbf{Y}} - J_{\mathbf{YX}}\mathbf{x} \quad \textit{and} \quad J = J_{\mathbf{YY}}.$$

Proof:

Block matrix inversion

$$\begin{pmatrix} \Sigma_{\mathbf{XX}} & \Sigma_{\mathbf{XY}} \\ \Sigma_{\mathbf{YX}} & \Sigma_{\mathbf{YY}} \end{pmatrix}^{-1} = \begin{pmatrix} (\Sigma_{\mathbf{XX}})^{-1} + (\Sigma_{\mathbf{XX}})^{-1}\Sigma_{\mathbf{XY}} J_{\mathbf{YY}} \Sigma_{\mathbf{YX}} (\Sigma_{\mathbf{XX}})^{-1} & -(\Sigma_{\mathbf{XX}})^{-1}\Sigma_{\mathbf{XY}} J_{\mathbf{YY}} \\ -J_{\mathbf{YY}} \Sigma_{\mathbf{YX}} (\Sigma_{\mathbf{XX}})^{-1} & J_{\mathbf{YY}} \end{pmatrix}$$

where

$$J_{\mathbf{YY}} = (\Sigma_{\mathbf{YY}} - \Sigma_{\mathbf{YX}}(\Sigma_{\mathbf{XX}})^{-1}\Sigma_{\mathbf{XY}})^{-1}.$$

The matrix $(J_{\mathbf{YY}})^{-1}$ is known as the **Schur complement** of $\Sigma_{\mathbf{XX}}$.

https://en.wikipedia.org/wiki/Invertible_matrix#Blockwise_inversion

Conditioning

If $(\mathbf{X}, \mathbf{Y})$ is multivariate Gaussian we write their parameters in block form:

$$\mu = \begin{pmatrix} \mu_{\mathbf{X}} \\ \mu_{\mathbf{Y}} \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_{\mathbf{XX}} & \Sigma_{\mathbf{XY}} \\ \Sigma_{\mathbf{YX}} & \Sigma_{\mathbf{YY}} \end{pmatrix}.$$

Theorem (7.4, p. 253, generalized)

The conditional distribution of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ is a Gaussian distribution with mean

$$\mu_{\mathbf{x}} = \mu_{\mathbf{Y}} + \Sigma_{\mathbf{YX}}(\Sigma_{\mathbf{XX}})^{-1}(\mathbf{x} - \mu_{\mathbf{X}})$$

and covariance matrix

$$\Sigma = \Sigma_{\mathbf{YY}} - \Sigma_{\mathbf{YX}}(\Sigma_{\mathbf{XX}})^{-1}\Sigma_{\mathbf{XY}}$$

Proof:

Summary

For the multivariate Gaussian distribution we have the following properties:

- All marginal distributions are Gaussian
- All conditional distributions are Gaussian

Summary

For the multivariate Gaussian distribution we have the following properties:

- All marginal distributions are Gaussian
- All conditional distributions are Gaussian

But different parametrizations serve the two operations differently:

- In the (μ, Σ)-parametrization it is straightforward to compute parameters for marginal distributions
- In the $(\mathbf{h}, J)$-parametrization it is straightforward to compute parameters for conditional distributions

Summary

For the multivariate Gaussian distribution we have the following properties:

- All marginal distributions are Gaussian
- All conditional distributions are Gaussian

But different parametrizations serve the two operations differently:

- In the (μ, Σ)-parametrization it is straightforward to compute parameters for marginal distributions
- In the $(\mathbf{h}, J)$-parametrization it is straightforward to compute parameters for conditional distributions

Going between the parametrizations,

$$\mathbf{h} = \Sigma^{-1}\mu,\ J = \Sigma^{-1} \quad \Leftrightarrow \quad \mu = J^{-1}\mathbf{h},\ \Sigma = J^{-1},$$

requires matrix inversion, with standard algorithms of $O(n^3)$ run time.