CS189/CS289A
Introduction to Machine Learning
Lecture 8: More on the Multivariate Normal Distribution

Peter Bartlett

February 12, 2015

Outline

Outline

- Review: Diagonal covariance matrices.

Outline

- Review: Diagonal covariance matrices.
- Non-diagonal covariance and diagonalization.

Outline

- Review: Diagonal covariance matrices.
- Non-diagonal covariance and diagonalization.
- Properties of multivariate Gaussians.

Outline

- Review: Diagonal covariance matrices.
- Non-diagonal covariance and diagonalization.
- Properties of multivariate Gaussians.
- Estimating Gaussians.

Outline

- **Review: Diagonal covariance matrices.**
- Non-diagonal covariance and diagonalization.
- Properties of multivariate Gaussians.
- Estimating Gaussians.

Review: Covariance matrices

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

Review: Covariance matrices

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \qquad \mathbb{E}X = \begin{bmatrix} \mathbb{E}X_1 \\ \mathbb{E}X_2 \\ \mathbb{E}X_3 \end{bmatrix}$$

Review: Covariance matrices

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \qquad \mathbb{E}X = \begin{bmatrix} \mathbb{E}X_1 \\ \mathbb{E}X_2 \\ \mathbb{E}X_3 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}$$

Review: Covariance matrices

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \qquad \mathbb{E}X = \begin{bmatrix} \mathbb{E}X_1 \\ \mathbb{E}X_2 \\ \mathbb{E}X_3 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \mu,$$

Review: Covariance matrices

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \qquad \mathbb{E}X = \begin{bmatrix} \mathbb{E}X_1 \\ \mathbb{E}X_2 \\ \mathbb{E}X_3 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \mu,$$

$$\mathrm{Cov}(X_i, X_j) = \mathbb{E}(X_i - \mu_i)(X_j - \mu_j),$$

Review: Covariance matrices

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \qquad \mathbb{E}X = \begin{bmatrix} \mathbb{E}X_1 \\ \mathbb{E}X_2 \\ \mathbb{E}X_3 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \mu,$$

$$\mathrm{Cov}(X_i, X_j) = \mathbb{E}(X_i - \mu_i)(X_j - \mu_j),$$

$$\Sigma = \mathbb{E}(X - \mu)(X - \mu)'$$

Review: Covariance matrices

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \qquad \mathbb{E}X = \begin{bmatrix} \mathbb{E}X_1 \\ \mathbb{E}X_2 \\ \mathbb{E}X_3 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \mu,$$

$$\mathrm{Cov}(X_i, X_j) = \mathbb{E}(X_i - \mu_i)(X_j - \mu_j),$$

$$\begin{aligned} \Sigma &= \mathbb{E}(X - \mu)(X - \mu)' \\ &= \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \mathrm{Cov}(X_1, X_3) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \mathrm{Cov}(X_2, X_3) \\ \mathrm{Cov}(X_3, X_1) & \mathrm{Cov}(X_3, X_2) & \mathrm{Var}(X_3) \end{bmatrix}. \end{aligned}$$

Review: Covariance matrices

Properties of Σ

Review: Covariance matrices

Properties of Σ

- Symmetric: $\Sigma_{i,j} = \Sigma_{j,i}$.

Review: Covariance matrices

Properties of Σ

- Symmetric: $\Sigma_{i,j} = \Sigma_{j,i}$.
- Non-negative diagonal entries: $\Sigma_{i,i} \geq 0$.

Review: Covariance matrices

Properties of Σ

- Symmetric: $\Sigma_{i,j} = \Sigma_{j,i}$.
- Non-negative diagonal entries: $\Sigma_{i,i} \geq 0$.
- Positive semidefinite: for all $v \in \mathbb{R}^d$, $v'\Sigma v \geq 0$.

Review: Diagonal covariance

Consider $X \sim N(\mu, \Sigma)$ where Σ is diagonal:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_d^2 \end{bmatrix}.$$

Review: Diagonal covariance

Consider $X \sim N(\mu, \Sigma)$ where Σ is diagonal:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_d^2 \end{bmatrix}.$$

The components of X are uncorrelated:

$$\mathrm{Cov}(X_i, X_j) = 0 \qquad \text{for} \qquad i \neq j.$$

Review: Diagonal covariance

Consider $X \sim N(\mu, \Sigma)$ where Σ is diagonal:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_d^2 \end{bmatrix}.$$

The components of X are uncorrelated:

$$\mathrm{Cov}(X_i, X_j) = 0 \qquad \text{for} \qquad i \neq j.$$

Consider the super-level sets of the pdf:

$$\mathcal{E}_r = \left\{ x \in \mathbb{R}^d : (x-\mu)'\Sigma^{-1}(x-\mu) \leq r^2 \right\}$$

Review: Diagonal covariance

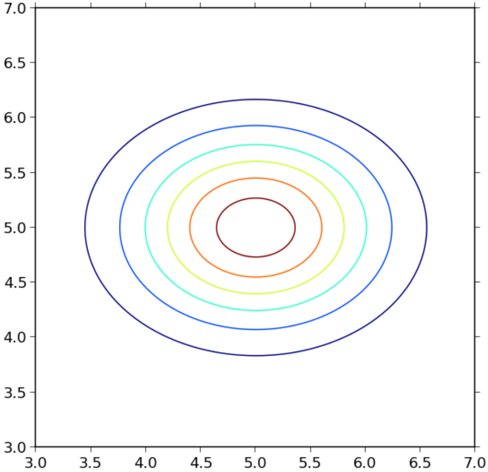
The super-level sets of the pdf,

$$\mathcal{E}_r = \left\{ x \in \mathbb{R}^d : (x-\mu)'\Sigma^{-1}(x-\mu) \le r^2 \right\}$$
$$= \left\{ x \in \mathbb{R}^d : \sum_{i=1}^{d} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \le r^2 \right\},$$

correspond to *axis-aligned ellipsoids* in $\mathbb{R}^d$, of length $2r\sigma_i$ in the x_i direction.
The volume of $\mathcal{E}_r$ is proportional to $\prod_{i=1}^d \sigma_i = \sqrt{|\Sigma|}$.
Hence, for a fixed r, $\Pr(\mathcal{E}_r)$ does not depend on μ and Σ (but does depend on d).

Review: Diagonal covariance

$$\mu = \begin{bmatrix} 5.0 \\ 5.0 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 0.8 & 0.0 \\ 0.0 & 0.6 \end{bmatrix}$$



Outline

- Review: Diagonal covariance matrices.
- **Non-diagonal covariance and diagonalization.**
- Properties of multivariate Gaussians.
- Estimating Gaussians.

Non-diagonal covariance

What if Σ is not diagonal?

Non-diagonal covariance

What if Σ is not diagonal?
Suppose we start with a diagonal Σ (so the contours of the pdf are axis-aligned ellipsoids)

Non-diagonal covariance

What if Σ is not diagonal?

Suppose we start with a diagonal Σ (so the contours of the pdf are axis-aligned ellipsoids), and then consider the transformed vector

$$Y = \mu_X + Q(X - \mu_X),$$

where Q is a matrix that corresponds to a rotation about the origin.

Non-diagonal covariance

What if Σ is not diagonal?
Suppose we start with a diagonal Σ (so the contours of the pdf are axis-aligned ellipsoids), and then consider the transformed vector

$$Y = \mu_X + Q(X - \mu_X),$$

where Q is a matrix that corresponds to a rotation about the origin.

$$\text{e.g.} \qquad Q = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

Non-diagonal covariance

This leaves the mean unchanged:

$$\mathbb{E}Y = \mu_X + \mathbb{E}\left[Q(X - \mu_X)\right] = \mu_X + Q(\mathbb{E}\left[X\right] - \mu_X) = \mu_X.$$

Non-diagonal covariance

This leaves the mean unchanged:

$$\mathbb{E}Y = \mu_X + \mathbb{E}\left[Q(X - \mu_X)\right] = \mu_X + Q(\mathbb{E}\left[X\right] - \mu_X) = \mu_X.$$

But the covariance is now transformed

Non-diagonal covariance

This leaves the mean unchanged:

$$\mathbb{E}Y = \mu_X + \mathbb{E}\left[Q(X - \mu_X)\right] = \mu_X + Q(\mathbb{E}\left[X\right] - \mu_X) = \mu_X.$$

But the covariance is now transformed:

$$\Sigma_Y = \mathbb{E}(Y - \mu_X)(Y - \mu_X)'$$

Non-diagonal covariance

This leaves the mean unchanged:

$$\mathbb{E}Y = \mu_X + \mathbb{E}\left[Q(X - \mu_X)\right] = \mu_X + Q(\mathbb{E}\left[X\right] - \mu_X) = \mu_X.$$

But the covariance is now transformed:

$$\begin{aligned}\Sigma_Y &= \mathbb{E}(Y - \mu_X)(Y - \mu_X)' \\ &= \mathbb{E}\left[Q(X - \mu_X)(X - \mu_X)'Q'\right]\end{aligned}$$

Non-diagonal covariance

This leaves the mean unchanged:

$$\mathbb{E}Y = \mu_X + \mathbb{E}\left[Q(X - \mu_X)\right] = \mu_X + Q(\mathbb{E}\left[X\right] - \mu_X) = \mu_X.$$

But the covariance is now transformed:

$$\begin{aligned}
\Sigma_Y &= \mathbb{E}(Y - \mu_X)(Y - \mu_X)' \\
&= \mathbb{E}\left[Q(X - \mu_X)(X - \mu_X)'Q'\right] \\
&= Q\mathbb{E}\left[(X - \mu_X)(X - \mu_X)'\right] Q'
\end{aligned}$$

Non-diagonal covariance

This leaves the mean unchanged:

$$\mathbb{E}Y = \mu_X + \mathbb{E}\left[Q(X - \mu_X)\right] = \mu_X + Q(\mathbb{E}\left[X\right] - \mu_X) = \mu_X.$$

But the covariance is now transformed:

$$\begin{aligned}
\Sigma_Y &= \mathbb{E}(Y - \mu_X)(Y - \mu_X)' \\
&= \mathbb{E}\left[Q(X - \mu_X)(X - \mu_X)'Q'\right] \\
&= Q\mathbb{E}\left[(X - \mu_X)(X - \mu_X)'\right]Q' \\
&= Q\Sigma_X Q'.
\end{aligned}$$

Non-diagonal covariance

This leaves the mean unchanged:

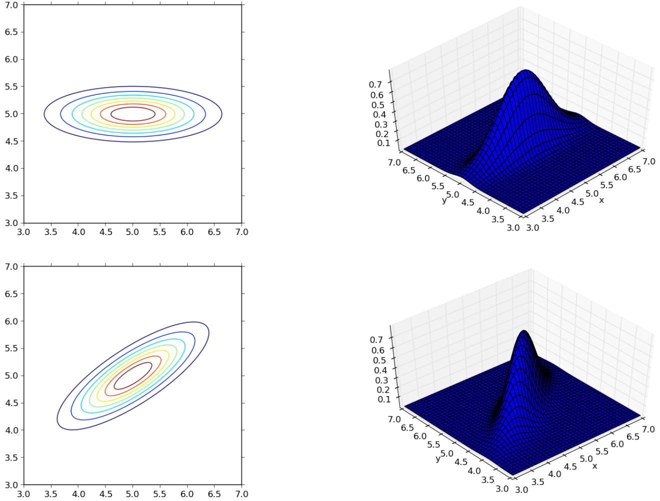
$$\mathbb{E}Y = \mu_X + \mathbb{E}\left[Q(X - \mu_X)\right] = \mu_X + Q(\mathbb{E}\left[X\right] - \mu_X) = \mu_X.$$

But the covariance is now transformed:

$$\begin{aligned}\Sigma_Y &= \mathbb{E}(Y - \mu_X)(Y - \mu_X)' \\ &= \mathbb{E}\left[Q(X - \mu_X)(X - \mu_X)'Q'\right] \\ &= Q\mathbb{E}\left[(X - \mu_X)(X - \mu_X)'\right]Q' \\ &= Q\Sigma_X Q'.\end{aligned}$$

It is no longer diagonal.

Non-diagonal covariance



Non-diagonal covariance

We can always view a covariance matrix in this way: as a rotated version of a diagonal covariance.

Non-diagonal covariance

We can always view a covariance matrix in this way: as a rotated version of a diagonal covariance.

Eigenvalues and eigenvectors

For a square matrix A, λ is an *eigenvalue* and x is an *eigenvector* if

$$Ax = \lambda x.$$

Non-diagonal covariance

We can always view a covariance matrix in this way: as a rotated version of a diagonal covariance.

Eigenvalues and eigenvectors

For a square matrix A, λ is an *eigenvalue* and x is an *eigenvector* if

$$Ax = \lambda x.$$

That is, the transformation A corresponds to a rescaling in the x direction.

Non-diagonal covariance

We can always view a covariance matrix in this way: as a rotated version of a diagonal covariance.

Eigenvalues and eigenvectors

For a square matrix A, λ is an *eigenvalue* and x is an *eigenvector* if

$$Ax = \lambda x.$$

That is, the transformation A corresponds to a rescaling in the x direction.

(Eigenvalues are roots of the *characteristic polynomial*, $\det(A - \lambda I)$.)

Non-diagonal covariance

Spectral Theorem

Non-diagonal covariance

Spectral Theorem

For a symmetric real matrix $A \in \mathbb{R}^{n \times n}$, we can find n orthonormal eigenvectors of A $(v_1, \ldots, v_n)$, and the eigenvalues $(\lambda_1, \ldots, \lambda_n)$ are real.

Non-diagonal covariance

Spectral Theorem

For a symmetric real matrix $A \in \mathbb{R}^{n \times n}$, we can find n orthonormal eigenvectors of A $(v_1, \ldots, v_n)$, and the eigenvalues $(\lambda_1, \ldots, \lambda_n)$ are real. Hence, we can write

$$AU = U\Lambda, \qquad A = U\Lambda U',$$

(this is called *diagonalization*), where

$$U = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}, \qquad \Lambda = \operatorname{diag}\left(\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \cdots \\ \lambda_n \end{bmatrix}\right).$$

Non-diagonal covariance

Spectral Theorem

For a symmetric real matrix $A \in \mathbb{R}^{n \times n}$, we can find n orthonormal eigenvectors of A $(v_1, \ldots, v_n)$, and the eigenvalues $(\lambda_1, \ldots, \lambda_n)$ are real. Hence, we can write

$$AU = U\Lambda, \qquad A = U\Lambda U',$$

(this is called *diagonalization*), where

$$U = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}, \qquad \Lambda = \mathrm{diag}\left(\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \cdots \\ \lambda_n \end{bmatrix}\right).$$

(Why "hence"? $U'U = I$, so $UU' = I$.)

Non-diagonal covariance

A covariance matrice Σ is symmetric, so we can write

$$\Sigma = U \Lambda U'.$$

Non-diagonal covariance

A covariance matrice Σ is symmetric, so we can write

$$\Sigma = U \Lambda U'.$$

It is also positive semidefinite, so we must have all of the $\lambda_i \geq 0$

Non-diagonal covariance

A covariance matrice Σ is symmetric, so we can write

$$\Sigma = U \Lambda U'.$$

It is also positive semidefinite, so we must have all of the $\lambda_i \geq 0$:

$$v_i' \Sigma v_i = v_i' U \Lambda U' v_i = \lambda_i \geq 0.$$

Non-diagonal covariance

A covariance matrice Σ is symmetric, so we can write

$$\Sigma = U \Lambda U'.$$

It is also positive semidefinite, so we must have all of the $\lambda_i \geq 0$:

$$v_i' \Sigma v_i = v_i' U \Lambda U' v_i = \lambda_i \geq 0.$$

Also, the *precision matrix* can be written

$$\Sigma^{-1} = U \Lambda^{-1} U'.$$

(Check!)

Non-diagonal covariance

Gaussian pdf

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right)$$

Non-diagonal covariance

Gaussian pdf

$$
\begin{aligned}
p(x) &= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right) \\
&= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)'U\Lambda^{-1}U'(x-\mu)\right).
\end{aligned}
$$

Non-diagonal covariance

Gaussian pdf

$$
\begin{aligned}
p(x) &= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right) \\
&= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)'U\Lambda^{-1}U'(x-\mu)\right).
\end{aligned}
$$

$U'(x-\mu)$ computes the representation of $x-\mu$ in the basis of the eigenvectors of Σ:

Non-diagonal covariance

Gaussian pdf

$$
\begin{aligned}
p(x) &= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right) \\
&= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)'U\Lambda^{-1}U'(x-\mu)\right).
\end{aligned}
$$

$U'(x-\mu)$ computes the representation of $x-\mu$ in the basis of the eigenvectors of Σ:

$$
U'(x-\mu) = \begin{bmatrix} v_1' \\ v_2' \\ \vdots \\ v_d' \end{bmatrix} (x-\mu)
$$

Non-diagonal covariance

Gaussian pdf

$$\begin{aligned} p(x) &= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right) \\ &= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)'U\Lambda^{-1}U'(x-\mu)\right). \end{aligned}$$

$U'(x-\mu)$ computes the representation of $x-\mu$ in the basis of the eigenvectors of Σ:

$$U'(x-\mu) = \begin{bmatrix} v_1' \\ v_2' \\ \vdots \\ v_d' \end{bmatrix} (x-\mu) = \begin{bmatrix} v_1'(x-\mu) \\ v_2'(x-\mu) \\ \vdots \\ v_d'(x-\mu) \end{bmatrix}.$$

Non-diagonal covariance

Gaussian pdf

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right)$$
$$= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)'U\Lambda^{-1}U'(x-\mu)\right).$$

$U'(x-\mu)$ computes the representation of $x-\mu$ in the basis of the eigenvectors of Σ:

$$U'(x-\mu) = \begin{bmatrix} v_1' \\ v_2' \\ \vdots \\ v_d' \end{bmatrix} (x-\mu) = \begin{bmatrix} v_1'(x-\mu) \\ v_2'(x-\mu) \\ \vdots \\ v_d'(x-\mu) \end{bmatrix}.$$

In this new basis, the covariance is the diagonal matrix Λ^{-1}.

Non-diagonal covariance

The level sets are ellipsoids with axes aligned with the eigenvectors, and the eigenvalues determine the scaling of the ellipsoids:

Non-diagonal covariance

The level sets are ellipsoids with axes aligned with the eigenvectors, and the eigenvalues determine the scaling of the ellipsoids:

Level sets with non-diagonal covariance

$$\mathcal{E}_r = \left\{ x \in \mathbb{R}^d : (x - \mu)' \Sigma^{-1} (x - \mu) \leq r^2 \right\}$$

Non-diagonal covariance

The level sets are ellipsoids with axes aligned with the eigenvectors, and the eigenvalues determine the scaling of the ellipsoids:

Level sets with non-diagonal covariance

$$\mathcal{E}_r = \left\{ x \in \mathbb{R}^d : (x-\mu)'\Sigma^{-1}(x-\mu) \leq r^2 \right\}$$

The $\mathcal{E}_r$ correspond to eigenvector-aligned ellipsoids in $\mathbb{R}^d$:

$$(x-\mu)'\Sigma^{-1}(x-\mu)$$

Non-diagonal covariance

The level sets are ellipsoids with axes aligned with the eigenvectors, and the eigenvalues determine the scaling of the ellipsoids:

Level sets with non-diagonal covariance

$$\mathcal{E}_r = \left\{ x \in \mathbb{R}^d : (x-\mu)'\Sigma^{-1}(x-\mu) \leq r^2 \right\}$$

The $\mathcal{E}_r$ correspond to eigenvector-aligned ellipsoids in $\mathbb{R}^d$:

$$(x-\mu)'\Sigma^{-1}(x-\mu) = (x-\mu)'U\Lambda^{-1}U'(x-\mu)$$

Non-diagonal covariance

The level sets are ellipsoids with axes aligned with the eigenvectors, and the eigenvalues determine the scaling of the ellipsoids:

Level sets with non-diagonal covariance

$$\mathcal{E}_r = \left\{ x \in \mathbb{R}^d : (x-\mu)'\Sigma^{-1}(x-\mu) \leq r^2 \right\}$$

The $\mathcal{E}_r$ correspond to eigenvector-aligned ellipsoids in $\mathbb{R}^d$:

$$(x-\mu)'\Sigma^{-1}(x-\mu) = (x-\mu)'U\Lambda^{-1}U'(x-\mu) = \sum_{i=1}^{d} \frac{(v_i'(x-\mu))^2}{\lambda_i^2}.$$

Non-diagonal covariance

The level sets are ellipsoids with axes aligned with the eigenvectors, and the eigenvalues determine the scaling of the ellipsoids:

Level sets with non-diagonal covariance

$$\mathcal{E}_r = \left\{ x \in \mathbb{R}^d : (x-\mu)'\Sigma^{-1}(x-\mu) \le r^2 \right\}$$

The $\mathcal{E}_r$ correspond to eigenvector-aligned ellipsoids in $\mathbb{R}^d$:

$$(x-\mu)'\Sigma^{-1}(x-\mu) = (x-\mu)'U\Lambda^{-1}U'(x-\mu) = \sum_{i=1}^{d} \frac{(v_i'(x-\mu))^2}{\lambda_i^2}.$$

For instance, if $v_2'(x-\mu) = \cdots = v_d'(x-\mu) = 0$, then $x \in \mathcal{E}_r$ when

Non-diagonal covariance

The level sets are ellipsoids with axes aligned with the eigenvectors, and the eigenvalues determine the scaling of the ellipsoids:

Level sets with non-diagonal covariance

$$\mathcal{E}_r = \left\{ x \in \mathbb{R}^d : (x-\mu)'\Sigma^{-1}(x-\mu) \leq r^2 \right\}$$

The $\mathcal{E}_r$ correspond to eigenvector-aligned ellipsoids in $\mathbb{R}^d$:

$$(x-\mu)'\Sigma^{-1}(x-\mu) = (x-\mu)'U\Lambda^{-1}U'(x-\mu) = \sum_{i=1}^{d} \frac{(v_i'(x-\mu))^2}{\lambda_i^2}.$$

For instance, if $v_2'(x-\mu) = \cdots = v_d'(x-\mu) = 0$, then $x \in \mathcal{E}_r$ when

$$(v_1(x-\mu))^2 \leq r^2\lambda_1^2$$

Non-diagonal covariance

The level sets are ellipsoids with axes aligned with the eigenvectors, and the eigenvalues determine the scaling of the ellipsoids:

Level sets with non-diagonal covariance

$$\mathcal{E}_r = \left\{ x \in \mathbb{R}^d : (x-\mu)'\Sigma^{-1}(x-\mu) \le r^2 \right\}$$

The $\mathcal{E}_r$ correspond to eigenvector-aligned ellipsoids in $\mathbb{R}^d$:

$$(x-\mu)'\Sigma^{-1}(x-\mu) = (x-\mu)'U\Lambda^{-1}U'(x-\mu) = \sum_{i=1}^{d} \frac{(v_i'(x-\mu))^2}{\lambda_i^2}.$$

For instance, if $v_2'(x-\mu) = \cdots = v_d'(x-\mu) = 0$, then $x \in \mathcal{E}_r$ when

$$(v_1(x-\mu))^2 \le r^2\lambda_1^2 \qquad \Leftrightarrow \qquad |v_1(x-\mu)| \le r\lambda_1.$$

Non-diagonal covariance

The level sets are ellipsoids with axes aligned with the eigenvectors, and the eigenvalues determine the scaling of the ellipsoids:

Level sets with non-diagonal covariance

$$\mathcal{E}_r = \left\{ x \in \mathbb{R}^d : (x-\mu)'\Sigma^{-1}(x-\mu) \le r^2 \right\}$$

The $\mathcal{E}_r$ correspond to eigenvector-aligned ellipsoids in $\mathbb{R}^d$:

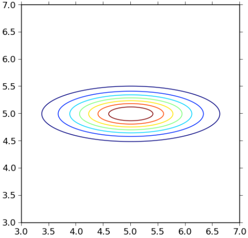
$$(x-\mu)'\Sigma^{-1}(x-\mu) = (x-\mu)'U\Lambda^{-1}U'(x-\mu) = \sum_{i=1}^{d} \frac{(v_i'(x-\mu))^2}{\lambda_i^2}.$$

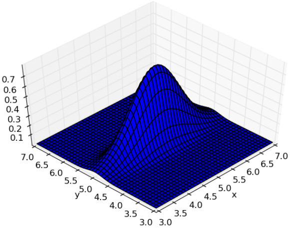
For instance, if $v_2'(x-\mu) = \cdots = v_d'(x-\mu) = 0$, then $x \in \mathcal{E}_r$ when

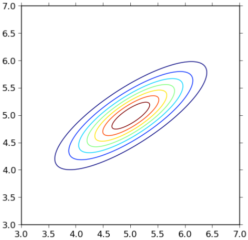
$$(v_1(x-\mu))^2 \le r^2\lambda_1^2 \qquad \Leftrightarrow \qquad |v_1(x-\mu)| \le r\lambda_1.$$

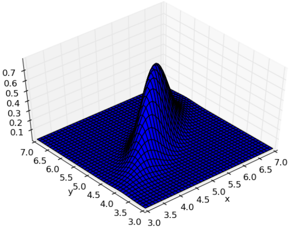
This is the same as the diagonal case, but with a new coordinate system, defined by the eigenvectors.

Non-diagonal covariance









Non-diagonal covariance

The volume of the ellipsoid $\mathcal{E}_r$ is proportional to

$$\prod_{i=1}^{d} \lambda_i = \sqrt{|\Lambda|}$$

Non-diagonal covariance

The volume of the ellipsoid $\mathcal{E}_r$ is proportional to

$$\prod_{i=1}^{d} \lambda_i = \sqrt{|\Lambda|} = \sqrt{|U| \cdot |\Lambda| \cdot |U'|}$$

Non-diagonal covariance

The volume of the ellipsoid $\mathcal{E}_r$ is proportional to

$$\prod_{i=1}^{d} \lambda_i = \sqrt{|\Lambda|} = \sqrt{|U| \cdot |\Lambda| \cdot |U'|} = \sqrt{|U \Lambda U'|}$$

Non-diagonal covariance

The volume of the ellipsoid $\mathcal{E}_r$ is proportional to

$$\prod_{i=1}^{d} \lambda_i = \sqrt{|\Lambda|} = \sqrt{|U| \cdot |\Lambda| \cdot |U'|} = \sqrt{|U \Lambda U'|} = \sqrt{|\Sigma|}.$$

Non-diagonal covariance

The volume of the ellipsoid $\mathcal{E}_r$ is proportional to

$$\prod_{i=1}^{d} \lambda_i = \sqrt{|\Lambda|} = \sqrt{|U| \cdot |\Lambda| \cdot |U'|} = \sqrt{|U \Lambda U'|} = \sqrt{|\Sigma|}.$$

(As in the diagonal case.)

Outline

- Review: Diagonal covariance matrices.
- Non-diagonal covariance and diagonalization.
- **Properties of multivariate Gaussians.**
- Estimating Gaussians.

Properties of multivariate Gaussians

Properties of multivariate Gaussians

Probability measure

$$\int_{\mathbb{R}^d} p(x; \mu, \Sigma)\, dx = 1.$$

Properties of multivariate Gaussians

Probability measure

$$\int_{\mathbb{R}^d} p(x; \mu, \Sigma)\, dx = 1.$$

Expectation

$$\mathbb{E}[X] = \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_d] \end{bmatrix} = \int_{\mathbb{R}^d} x p(x; \mu, \Sigma)\, dx = \mu.$$

Properties of multivariate Gaussians

Probability measure

$$\int_{\mathbb{R}^d} p(x; \mu, \Sigma)\, dx = 1.$$

Expectation

$$\mathbb{E}[X] = \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_d] \end{bmatrix} = \int_{\mathbb{R}^d} x p(x; \mu, \Sigma)\, dx = \mu.$$

Covariance

$$\mathrm{Cov}[X] = \mathbb{E}\left[(X - \mu)(X - \mu)'\right] = \int_{\mathbb{R}^d} (x - \mu)(x - \mu)' p(x; \mu, \Sigma)\, dx = \Sigma.$$

Properties of multivariate Gaussians

Sums of independent Gaussians

For independent d-dimensional Gaussians,
$X \sim \mathcal{N}(\mu_X, \Sigma_X)$ and $Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$,

Properties of multivariate Gaussians

Sums of independent Gaussians

For independent d-dimensional Gaussians,
$X \sim \mathcal{N}(\mu_X, \Sigma_X)$ and $Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$,
$X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \Sigma_X + \Sigma_Y)$.

- NB: Adding the random variables, not the densities!
  This corresponds to *convolving* the densities.
  The family of Gaussian densities is closed under convolutions.

Properties of multivariate Gaussians

Sums of independent Gaussians

For independent d-dimensional Gaussians,
$X \sim \mathcal{N}(\mu_X, \Sigma_X)$ and $Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$,
$X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \Sigma_X + \Sigma_Y)$.

- NB: Adding the random variables, not the densities!
  This corresponds to *convolving* the densities.
  The family of Gaussian densities is closed under convolutions.
- Since $X + Y$ is Gaussian, we only need its mean and covariance.

Properties of multivariate Gaussians

Sum of two scalar standard normals: completing the square

$$
\begin{aligned}
p_{X+Y}(z) &= \int p_X(x) p_Y(z-x)\, dx = (p_X * p_Y)(z) \\
&= \frac{1}{2\pi} \int \exp\left(-\frac{1}{2} \left(2x^2 - 2xz + z^2 \right) \right) dx \\
&= \frac{1}{2\pi} \int \exp\left(-\frac{1}{2} \left(2(x - z/2)^2 - z^2/2 + z^2 \right) \right) dx \\
&= \frac{\exp(-z^2/4)}{\sqrt{2\pi \cdot 2}} \frac{1}{\sqrt{2\pi/2}} \int \exp\left(-\frac{1}{2} \left(\frac{(x - z/2)^2}{1/2} \right) \right) dx \\
&= \frac{\exp(-z^2/4)}{\sqrt{2\pi \cdot 2}} \qquad \text{i.e., } \mathcal{N}(0, 2).
\end{aligned}
$$

Properties of multivariate Gaussians

Properties of multivariate Gaussians

- $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

Properties of multivariate Gaussians

- $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.
- For covariance, let's consider the mean-zero case (which has the same covariance):

$$\mathrm{Cov}(X+Y) = \mathbb{E}[(X+Y)(X+Y)']$$

Properties of multivariate Gaussians

- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.
- For covariance, let's consider the mean-zero case (which has the same covariance):

$$\begin{aligned}\mathrm{Cov}(X + Y) &= \mathbb{E}[(X + Y)(X + Y)'] \\ &= \mathbb{E}[XX' + XY' + YX' + YY']\end{aligned}$$

Properties of multivariate Gaussians

- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.
- For covariance, let's consider the mean-zero case (which has the same covariance):

$$\begin{aligned}
\mathrm{Cov}(X + Y) &= \mathbb{E}[(X + Y)(X + Y)'] \\
&= \mathbb{E}[XX' + XY' + YX' + YY'] \\
&= \mathbb{E}[XX'] + \mathbb{E}[XY'] + \mathbb{E}[YX'] + \mathbb{E}[YY']
\end{aligned}$$

Properties of multivariate Gaussians

- $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.
- For covariance, let's consider the mean-zero case (which has the same covariance):

$$\begin{aligned}
\mathrm{Cov}(X+Y) &= \mathbb{E}[(X+Y)(X+Y)'] \\
&= \mathbb{E}[XX' + XY' + YX' + YY'] \\
&= \mathbb{E}[XX'] + \mathbb{E}[XY'] + \mathbb{E}[YX'] + \mathbb{E}[YY'] \\
&= \mathbb{E}[XX'] + 0 + 0 + \mathbb{E}[YY']
\end{aligned}$$

Properties of multivariate Gaussians

- $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.
- For covariance, let's consider the mean-zero case (which has the same covariance):

$$\begin{aligned}
\mathrm{Cov}(X+Y) &= \mathbb{E}[(X+Y)(X+Y)'] \\
&= \mathbb{E}[XX' + XY' + YX' + YY'] \\
&= \mathbb{E}[XX'] + \mathbb{E}[XY'] + \mathbb{E}[YX'] + \mathbb{E}[YY'] \\
&= \mathbb{E}[XX'] + 0 + 0 + \mathbb{E}[YY'] \\
&= \Sigma_X + \Sigma_Y.
\end{aligned}$$

Properties of multivariate Gaussians

Sums of independent Gaussians

For independent d-dimensional Gaussians,
$X \sim \mathcal{N}(\mu_X, \Sigma_X)$ and $Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$,
$X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \Sigma_X + \Sigma_Y)$.

- We need X and Y independent.

Properties of multivariate Gaussians

Sums of independent Gaussians

For independent d-dimensional Gaussians,
$X \sim \mathcal{N}(\mu_X, \Sigma_X)$ and $Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$,
$X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \Sigma_X + \Sigma_Y)$.

- We need X and Y independent. For example: $X \sim \mathcal{N}(0, 1)$,

$$Y = \begin{cases} -X & \text{if } |X| < 1, \\ X & \text{if } |X| > 1. \end{cases}$$

Properties of multivariate Gaussians

Sums of independent Gaussians

For independent d-dimensional Gaussians,
$X \sim \mathcal{N}(\mu_X, \Sigma_X)$ and $Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$,
$X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \Sigma_X + \Sigma_Y)$.

- We need X and Y independent. For example: $X \sim \mathcal{N}(0, 1)$,

$$Y = \begin{cases} -X & \text{if } |X| < 1, \\ X & \text{if } |X| > 1. \end{cases}$$

Then $X \sim \mathcal{N}(0, 1)$ and $Y \sim \mathcal{N}(0, 1)$.

Properties of multivariate Gaussians

Sums of independent Gaussians

For independent d-dimensional Gaussians,
$X \sim \mathcal{N}(\mu_X, \Sigma_X)$ and $Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$,
$X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \Sigma_X + \Sigma_Y)$.

- We need X and Y independent. For example: $X \sim \mathcal{N}(0,1)$,

$$Y = \begin{cases} -X & \text{if } |X| < 1, \\ X & \text{if } |X| > 1. \end{cases}$$

Then $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,1)$.
But (X, Y) is not jointly Gaussian. (Check!)

Properties of multivariate Gaussians

Sums of independent Gaussians

For independent d-dimensional Gaussians,
$X \sim \mathcal{N}(\mu_X, \Sigma_X)$ and $Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$,
$X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \Sigma_X + \Sigma_Y)$.

- We need X and Y independent. For example: $X \sim \mathcal{N}(0, 1)$,

$$Y = \begin{cases} -X & \text{if } |X| < 1, \\ X & \text{if } |X| > 1. \end{cases}$$

Then $X \sim \mathcal{N}(0, 1)$ and $Y \sim \mathcal{N}(0, 1)$.
But (X, Y) is not jointly Gaussian. (Check!)
And $X + Y$ is not Gaussian. (Check!)

Properties of multivariate Gaussians

Sums of independent Gaussians

For independent d-dimensional Gaussians,
$X \sim \mathcal{N}(\mu_X, \Sigma_X)$ and $Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$,
$X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \Sigma_X + \Sigma_Y)$.

- We need X and Y independent. For example: $X \sim \mathcal{N}(0, 1)$,

$$Y = \begin{cases} -X & \text{if } |X| < 1, \\ X & \text{if } |X| > 1. \end{cases}$$

Then $X \sim \mathcal{N}(0, 1)$ and $Y \sim \mathcal{N}(0, 1)$.
But (X, Y) is not jointly Gaussian. (Check!)
And $X + Y$ is not Gaussian. (Check!)
And, replacing the threshold 1 with a suitable constant, we can ensure that X, Y are uncorrelated but dependent. (Check!)

Properties of multivariate Gaussians

Marginals

Given a d-dimensional Gaussian $X \sim \mathcal{N}(\mu, \Sigma)$, write

$$X = \begin{bmatrix} Y \\ Z \end{bmatrix}, \qquad \mu = \begin{bmatrix} \mu_Y \\ \mu_Z \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZY} & \Sigma_{ZZ} \end{bmatrix},$$

where $Y \in \mathbb{R}^m$ and $Z \in \mathbb{R}^{d-m}$.

Properties of multivariate Gaussians

Marginals

Given a d-dimensional Gaussian $X \sim \mathcal{N}(\mu, \Sigma)$, write

$$X = \begin{bmatrix} Y \\ Z \end{bmatrix}, \qquad \mu = \begin{bmatrix} \mu_Y \\ \mu_Z \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZY} & \Sigma_{ZZ} \end{bmatrix},$$

where $Y \in \mathbb{R}^m$ and $Z \in \mathbb{R}^{d-m}$. Then $Y \sim \mathcal{N}(\mu_Y, \Sigma_{YY})$.

- The mean and covariance of Y are immediate from the definitions.

Properties of multivariate Gaussians

Marginals

Given a d-dimensional Gaussian $X \sim \mathcal{N}(\mu, \Sigma)$, write

$$X = \begin{bmatrix} Y \\ Z \end{bmatrix}, \qquad \mu = \begin{bmatrix} \mu_Y \\ \mu_Z \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZY} & \Sigma_{ZZ} \end{bmatrix},$$

where $Y \in \mathbb{R}^m$ and $Z \in \mathbb{R}^{d-m}$. Then $Y \sim \mathcal{N}(\mu_Y, \Sigma_{YY})$.

- The mean and covariance of Y are immediate from the definitions.
- To show that Y has a Gaussian distribution, we write its density as an integral (integrating out the Z variables), *complete the squares* to compute the integral, and recognize the answer as a normal density. (And necessarily it's a $\mathcal{N}(\mu_Y, \Sigma_{YY})$ density.)

Properties of multivariate Gaussians

Affine transformations

Given a d-dimensional Gaussian $X \sim \mathcal{N}(\mu, \Sigma^2)$, matrix $A \in \mathbb{R}^{m \times d}$ and vector $b \in \mathbb{R}^m$, define

$$Y = AX + b.$$

Properties of multivariate Gaussians

Affine transformations

Given a d-dimensional Gaussian $X \sim \mathcal{N}(\mu, \Sigma^2)$, matrix $A \in \mathbb{R}^{m \times d}$ and vector $b \in \mathbb{R}^m$, define

$$Y = AX + b.$$

Then $Y \sim \mathcal{N}(A\mu + b, A\Sigma A')$.

Properties of multivariate Gaussians

Affine transformations

Given a d-dimensional Gaussian $X \sim \mathcal{N}(\mu, \Sigma^2)$, matrix $A \in \mathbb{R}^{m \times d}$ and vector $b \in \mathbb{R}^m$, define

$$Y = AX + b.$$

Then $Y \sim \mathcal{N}(A\mu + b, A\Sigma A')$.

- Why is Y Gaussian?
  e.g., $m = d$ and A has full rank. Then Y is a shifted, rescaled version of X, which is Gaussian.

Properties of multivariate Gaussians

Affine transformations

Given a d-dimensional Gaussian $X \sim \mathcal{N}(\mu, \Sigma^2)$, matrix $A \in \mathbb{R}^{m \times d}$ and vector $b \in \mathbb{R}^m$, define

$$Y = AX + b.$$

Then $Y \sim \mathcal{N}(A\mu + b, A\Sigma A')$.

- Why is Y Gaussian?
  e.g., $m = d$ and A has full rank. Then Y is a shifted, rescaled version of X, which is Gaussian.
- $\mathbb{E}Y = A\mu + b$.

Properties of multivariate Gaussians

Affine transformations

Given a d-dimensional Gaussian $X \sim \mathcal{N}(\mu, \Sigma^2)$, matrix $A \in \mathbb{R}^{m \times d}$ and vector $b \in \mathbb{R}^m$, define

$$Y = AX + b.$$

Then $Y \sim \mathcal{N}(A\mu + b, A\Sigma A')$.

- Why is Y Gaussian?
  e.g., $m = d$ and A has full rank. Then Y is a shifted, rescaled version of X, which is Gaussian.
- $\mathbb{E}Y = A\mu + b$.
- $\mathrm{Cov}(Y) = \mathbb{E}\left[(Y - \mathbb{E}Y)(Y - \mathbb{E}Y)'\right] = \mathbb{E}\left[(A(X - \mu))(A(X - \mu))'\right] = A\mathbb{E}[(X - \mu)(X - \mu)']A' = A\Sigma A'$.

Properties of multivariate Gaussians

Example

Given a d-dimensional standard normal $X \sim \mathcal{N}(0, I)$, we have

$$Y = AX + \mu$$

Properties of multivariate Gaussians

Example

Given a d-dimensional standard normal $X \sim \mathcal{N}(0, I)$, we have

$$Y = AX + \mu \sim \mathcal{N}(\mu, AA').$$

Properties of multivariate Gaussians

Example

Given a d-dimensional standard normal $X \sim \mathcal{N}(0, I)$, we have

$$Y = AX + \mu \sim \mathcal{N}(\mu, AA').$$

Example

Given a d-dimensional normal $X \sim \mathcal{N}(\mu, \Sigma)$ with Σ positive definite,

$$Y = \Sigma^{-1/2}(X - \mu)$$

Properties of multivariate Gaussians

Example

Given a d-dimensional standard normal $X \sim \mathcal{N}(0, I)$, we have

$$Y = AX + \mu \sim \mathcal{N}(\mu, AA').$$

Example

Given a d-dimensional normal $X \sim \mathcal{N}(\mu, \Sigma)$ with Σ positive definite,

$$Y = \Sigma^{-1/2}(X - \mu) \sim \mathcal{N}(0, I).$$

Properties of multivariate Gaussians

Example

Given a d-dimensional standard normal $X \sim \mathcal{N}(0, I)$, we have

$$Y = AX + \mu \sim \mathcal{N}(\mu, AA').$$

Example

Given a d-dimensional normal $X \sim \mathcal{N}(\mu, \Sigma)$ with Σ positive definite,

$$Y = \Sigma^{-1/2}(X - \mu) \sim \mathcal{N}(0, I).$$

This is called *whitening* X.

Properties of multivariate Gaussians

Example

Given a d-dimensional standard normal $X \sim \mathcal{N}(0, I)$, we have

$$Y = AX + \mu \sim \mathcal{N}(\mu, AA').$$

Example

Given a d-dimensional normal $X \sim \mathcal{N}(\mu, \Sigma)$ with Σ positive definite,

$$Y = \Sigma^{-1/2}(X - \mu) \sim \mathcal{N}(0, I).$$

This is called *whitening* X.

What is $\Sigma^{-1/2}$? We can write $\Sigma = U\Lambda U'$ for U with orthonormal columns and Λ diagonal. Then defining $\Sigma^{-1/2} = U\Lambda^{-1/2}U'$, we have

$$\Sigma^{-1/2}\Sigma^{-1/2} = U\Lambda^{-1/2}U'U\Lambda^{-1/2}U' = U\Lambda^{-1}U' = \Sigma^{-1}.$$

Outline

- Review: Diagonal covariance matrices.
- Non-diagonal covariance and diagonalization.
- Properties of multivariate Gaussians.
- **Estimating Gaussians.**

Maximum likelihood

Gaussian Estimation

- Work with the log likelihood:

$$\ell(\mu, \sigma^2) = \log\left(\prod_{i=1}^{n}\left(\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right)\right)$$

Maximum likelihood

Gaussian Estimation

- Work with the log likelihood:

$$\ell(\mu, \sigma^2) = \log\left(\prod_{i=1}^{n}\left(\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right)\right)$$

$$= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Maximum likelihood

Gaussian Estimation

- Work with the log likelihood:

$$\ell(\mu, \sigma^2) = \log\left(\prod_{i=1}^{n}\left(\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right)\right)$$

$$= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Set $\nabla\ell(\theta) = 0$ and solve.

Maximum likelihood

Gaussian maximum likelihood estimation

$\nabla \ell(\theta) = 0$ for

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i, \qquad \frac{n}{2\sigma^2} = \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2.$$

Other estimators for Gaussian parameters

Gaussian Estimation

Other estimators for Gaussian parameters

Gaussian Estimation

- We can also use penalized maximum likelihood estimators:

$$\ell(\mu, \sigma^2) - \text{penalty}(\mu, \sigma^2)$$
$$= \log \left(\prod_{i=1}^{n} \left(\frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right) \right) \right) - \text{penalty}(\mu, \sigma^2)$$

Other estimators for Gaussian parameters

Gaussian Estimation

- We can also use penalized maximum likelihood estimators:

$$\ell(\mu, \sigma^2) - \text{penalty}(\mu, \sigma^2)$$
$$= \log\left(\prod_{i=1}^{n}\left(\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right)\right) - \text{penalty}(\mu, \sigma^2)$$

- And Bayesian estimators:

$$\begin{aligned} \text{prior distribution:} \quad & \pi(\theta) = 1 \\ \text{posterior distribution:} \quad & p(\theta | X_1 = x_1) \propto \underbrace{p(X_1 = x_1 | \theta)}_{\text{likelihood}} \underbrace{\pi(\theta)}_{\text{prior}} \\ & = \frac{p(X_1 = x_1|\theta)\pi(\theta)}{\int p(X_1 = x_1 | q)\, d\pi(q)} \end{aligned}$$

Other estimators for Gaussian parameters

Gaussian Estimation

- Penalized maximum likelihood estimators and Bayesian estimators are particularly effective in the high-dimensional setting, when the number of parameters is large compared to the amount of data.

Maximum likelihood

Multivariate Gaussian Estimation

- Log likelihood:

$$\ell(\mu, \Sigma) = \log\left(\prod_{i=1}^{n}\left(\frac{1}{(2\pi)^{d/2}\left|\Sigma\right|^{1/2}}\exp\left(-\frac{1}{2}(x_i-\mu)'\Sigma^{-1}(x_i-\mu)\right)\right)\right)$$

Maximum likelihood

Multivariate Gaussian Estimation

- Log likelihood:

$$\begin{aligned}\ell(\mu, \Sigma) &= \log\left(\prod_{i=1}^{n}\left(\frac{1}{(2\pi)^{d/2}\,|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x_i-\mu)'\Sigma^{-1}(x_i-\mu)\right)\right)\right)\\ &= -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)'\Sigma^{-1}(x_i-\mu).\end{aligned}$$

Maximum likelihood

Multivariate Gaussian Estimation

- Log likelihood:

$$\ell(\mu, \Sigma) = \log \left(\prod_{i=1}^{n} \left(\frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x_i - \mu)'\Sigma^{-1}(x_i - \mu) \right) \right) \right)$$

$$= -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)'\Sigma^{-1}(x_i - \mu).$$

Set $\nabla \ell(\theta) = 0$ and solve.

Maximum likelihood

Gaussian maximum likelihood estimation

$\nabla_\mu \ell(\mu, \Sigma) = 0$ for

$$\sum_{i=1}^{n} \Sigma^{-1}(\mu - x_i) = 0 \qquad \text{check: } \nabla_\mu^2 \ell = -\frac{n}{2}\Sigma^{-1}$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i. \qquad \text{(sample mean)}$$

And it's possible to show that the maximum of $\ell(\mu, \Sigma)$ occurs at

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})(x_i - \hat{\mu})'. \qquad \text{(sample covariance)}$$

Outline

- Review: Diagonal covariance matrices.
- Non-diagonal covariance and diagonalization.
- Properties of multivariate Gaussians.
- Estimating Gaussians.