CS281: Advanced ML **September 11, 2017**

Lecture 3: Multivariate Normal Distributions

Lecturer: Sasha Rush *Scribes: Christopher Mosch, Lindsey Brown, Ryan Lapcevic*

3.1 Examples

Multivariate gaussians are used for modeling in various applications, where knowing mean and variance is useful:

- radar: mean and variance of approaching objects (like invading aliens)
- weather forecasting: predicting the position of a hurricane, where the uncertainty in the storm's position increases for timepoints farther away
- tracking the likely outcome of a sports game: last year's superbowl is an example of a failure of modeling with multivariate gaussians as the Patriots still won after a large Falcons' lead

3.2 Review: Eigendecomposition

Let $\mathbf{\Sigma}$ be a square, symmetric matrix. Then its eigendecomposition is given by $\mathbf{\Sigma} = \boldsymbol{U}^T \mathbf{\Lambda} \boldsymbol{U}$, where $\boldsymbol{U}$ is an orthogonal matrix and $\mathbf{\Lambda}$ is a diagonal matrix. In the special case that $\mathbf{\Sigma}$ is positive semidefinite (as is the case for covariance matrices), denoted $\mathbf{\Sigma} \succeq 0$, all its eigenvalues are nonnegative, $\Lambda_{ii} \geq 0$, and we can decompose its inverse as $\mathbf{\Sigma}^{-1} = \boldsymbol{U}^T \mathbf{\Lambda}^{-1} \boldsymbol{U}$, where $\Lambda_{ii}^{-1} = 1/\Lambda_{ii}$.

3.3 Multivariate Normal Distributions (MVNs)

Let X be a D-dimensional MVN random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$, denoted $X \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$. Then the pdf of X is

$$p(x) = (2\pi)^{-D/2} |\mathbf{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(x-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(x-\boldsymbol{\mu})\right],$$

where for many problems we focus on the quadratic form $(x-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(x-\boldsymbol{\mu})$ (which geometrically can be thought of as distance) and ignore the normalization factor $(2\pi)^{-D/2} |\mathbf{\Sigma}|^{-1/2}$. Note that MVNs are identifiable using just their first and second moments.

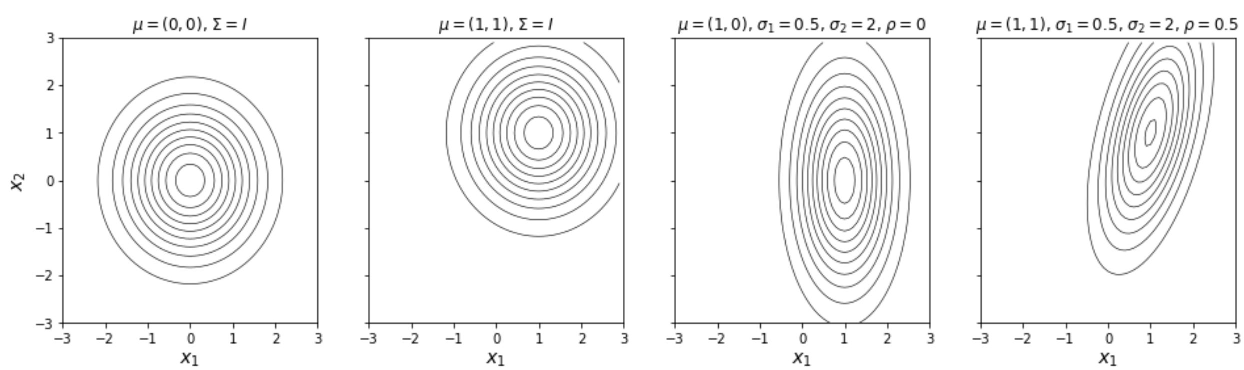

Figure 3.1: Contours of a bivariate Normal for various $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ (in the figure, ρ denotes the off-diagonal elements of $\mathbf{\Sigma}$, given by the covariance of x_1 and x_2)

Note that we can decompose $\mathbf{\Sigma}$ as

$$\begin{aligned}
\mathbf{\Sigma} &= (x-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(x-\boldsymbol{\mu}) \\
&= (x-\boldsymbol{\mu})^T \left(\boldsymbol{U}^T \mathbf{\Lambda}^{-1} \boldsymbol{U}\right)(x-\boldsymbol{\mu}) \\
&= (x-\boldsymbol{\mu})^T \left(\sum_d \frac{1}{\lambda_d} U_d U^T\right)(x-\boldsymbol{\mu}) \\
&= \sum_d \frac{1}{\lambda_d} (x-\boldsymbol{\mu})^T U_d U_d^T (x-\boldsymbol{\mu}),
\end{aligned}$$

where $(x-\boldsymbol{\mu})^T U_d$ can be interpreted as the projection of $(x-\boldsymbol{\mu})$ onto U_d (which can each be thought of as univariate gaussians), the eigenvector corresponding to the eigenvalue λ_d. Since $\mathbf{\Sigma}$ is the weighted sum of the dot product of such projections (with weights being given by $1/\lambda_d$, which can be thought of as the scale $1/\sigma^2$), we can describe the MVN as tiling of univariates.

3.3.1 Manipulating MVNs: Stretches, Rotations, and Shifts

Let $x \sim \mathcal{N}(0, \boldsymbol{I})$ and $y = \boldsymbol{A}x + \boldsymbol{b}$. We want to consider two ways of obtaining the complete distribution of y.

- 'Overkill': We can perform a change of variables[1]. Here, we have $x = \boldsymbol{A}^{-1}$ and $|dx/dy| = |A^{-1}|$, leading to

$$\begin{aligned}
p(y) &= \mathcal{N}\left(\boldsymbol{A}^{-1}(y-\boldsymbol{b})|0, \boldsymbol{I}\right)|\boldsymbol{A}^{-1}| \\
&= \frac{1}{z}\exp\left[(\boldsymbol{A}^{-1}(y-\boldsymbol{b}))^T(\boldsymbol{A}^{-1}(y-\boldsymbol{b}))\right] \\
&= \frac{1}{z}\exp\left[(y-\boldsymbol{b})^T(\boldsymbol{A}^{-1})^T(\boldsymbol{A}^{-1})(y-\boldsymbol{b})\right] \\
&= \mathcal{N}(y|\boldsymbol{b}, \boldsymbol{A}\boldsymbol{A}^T),
\end{aligned}$$

  where z is the normalizing constant.

- Using the properties of MVN, we know that y is also MVN, so is completely specified by its mean and covariance matrix which can easily be derived,

$$\mathbb{E}(y) = \mathbb{E}(\boldsymbol{A}x+\boldsymbol{b}) = \boldsymbol{A}\mathbb{E}(x)+\boldsymbol{b} \qquad \mathrm{cov}(y) = \boldsymbol{A}\boldsymbol{A}^T.$$

Thus, we can generate MVN from $\mathcal{N}(0, \boldsymbol{I})$ via the transformation $y = \boldsymbol{A}x + \boldsymbol{b}$, where we set $\boldsymbol{A} = \boldsymbol{U}\mathbf{\Lambda}^{1/2}$, leading to $\mathbf{\Sigma}_Y = \boldsymbol{U}^T\mathbf{\Lambda}\boldsymbol{U}$. Then shifts are represented by $\boldsymbol{b}$, stretches by $\mathbf{\Lambda}$, and rotations by $\boldsymbol{U}$.

3.3.2 Detour: MVN in High-Dimensions ($D \gg 0$)

Let x be a D-dimensional random vector, distributed as $\mathcal{N}(0, I/D)$, where I is the identity. The expected length of x is given by

$$\mathbb{E}\left(\|x\|^2\right) = \mathbb{E}\left(\sum_d x_d^2\right) = D\sigma_d^2 = 1,$$

[1] A change of variables can be done in the following way: Let $y = f(x)$ and assume f is invertible so that $x = f^{-1}(y)$. Then $p(y) = p(x)|dx/dy|$. This is a technique which will be used often in this course.

which means that x is expected to be on the boundary of a unit sphere centered at the origin. Moreover, the variance of the length is

$$\text{var}\left(\|x\|^2\right) = D \cdot \left(\mathbb{E}(x^4) - \mathbb{E}\left(x^2\right)^2\right) = D \cdot (3\sigma^4 - \sigma^4) = 2D/D^2 = 2/D$$

Thus, it is not only expected that x lies on the boundary but as D increases most of its realizations will in fact fall on the boundary[2]. A good illustration can be found at `http://www.inference.vc/high-dimensional-gaussian-distributions-are-soap-bubble/`.

3.3.3 Key Formulas for MVN: Marginalization and Conditioning

Let $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Note that Σ is written in block matrix form, rather than scalar entries. It turns out that the marginals, X_1 and X_2, are also MVN, and their mean and covariance matrice are given by $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_{22}$ respectively. A sketch of the proof is provided below.

$$p(x_1) = \int_{x_2} N(x|\boldsymbol{\mu}, \boldsymbol{\Sigma}) dx_2,$$

which can be written as

$$0.5 \int_{x_2} \exp\left[(x_1 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_{11}^{-1}(x_1 - \mu_1) + 2(x_1 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_{12}^{-1}(x_2 - \mu_2) + (x_2 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_{22}^{-1}(x_2 - \mu_2)\right] dx_2.$$

Note that this equals

$$p(x_1) \int_{x_2} p(x_2|x_1) dx_2,$$

implying that $X_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$.
While the marginals have a simple form, the conditionals are more complicated. (For a complete derivation, which requires matrix inversion lemmas, refer to Murphy 4.3.1 and 4.3.4.3.) It can be shown that $X_1|X_2 \sim \mathcal{N}\left(\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}\right)$ with

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(x_2 - \boldsymbol{\mu}_2), \quad \boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$

3.3.4 Information Form

An alternative parametrization, called information form, uses the precision matrix (inverse variance) $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$. Suppose $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ are called the **moment parameters** of the distribution. Sometimes it is useful to use the **canonical parameters** or **natural parameters**, defined as

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}, \ \boldsymbol{\xi} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$$

We can convert back to the moment parameters using

$$\boldsymbol{\mu} = \boldsymbol{\Lambda}^{-1}\boldsymbol{\xi}, \ \boldsymbol{\Sigma} = \boldsymbol{\Lambda}^{-1}$$

Using the canonical parameters, we can write the MVN in **information form** (i.e., in exponential family form, see Murphy Section 9.2):

$$\mathcal{N}_c(\boldsymbol{x}|\boldsymbol{\xi}, \boldsymbol{\Lambda}) = (2\pi)^{-D/2}|\boldsymbol{\Lambda}|^{\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\boldsymbol{x}^\intercal \boldsymbol{\Lambda} \boldsymbol{x} + \boldsymbol{\xi}^\intercal \boldsymbol{\Lambda}^{-1} \boldsymbol{\xi} - 2\boldsymbol{x}^\intercal \boldsymbol{\xi}\right)\right]$$

[2] It is left as an exercise to show that this formula holds. Hint: Use the fact that we assumed no covariance.

where we use the notation $\mathcal{N}_c()$ to distinguish from the moment parameterization $\mathcal{N}()$. Partitioning Λ as

$$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{\Lambda}_{11} & \mathbf{\Lambda}_{12} \\ \mathbf{\Lambda}_{21} & \mathbf{\Lambda}_{22} \end{pmatrix},$$

we can derive the marginalization and conditioning formulas in information form:

$$\begin{aligned} p(\boldsymbol{x}_2) &= \mathcal{N}_c\left(\boldsymbol{x}_2 | \boldsymbol{\xi}_2 - \mathbf{\Lambda}_{21}\mathbf{\Lambda}_{11}^{-1}\boldsymbol{\xi}_1, \mathbf{\Lambda}_{22} - \mathbf{\Lambda}_{21}\mathbf{\Lambda}_{11}^{-1}\mathbf{\Lambda}_{12}\right) \\ p(\boldsymbol{x}_1|\boldsymbol{x}_2) &= \mathcal{N}_c\left(\boldsymbol{x}_1 | \boldsymbol{\xi}_1 - \mathbf{\Lambda}_{12}\boldsymbol{x}_2, \mathbf{\Lambda}_{11}\right). \end{aligned}$$

The covariance matrices of the conditional distributions have a simple form. For example, the covariance matrix of $\boldsymbol{x}_1$ given $\boldsymbol{x}_2$ is $\mathbf{\Lambda}_{1|2} = \mathbf{\Lambda}_{11}$. However, the simplicity of the conditional precision comes at the cost of marginalization (which was simple when using $\mathbf{\Sigma}$) becoming a more complicated expression (see Murphy subsection 4.3 for more details).

Exercise 3.1. See hint 2.

Proof. x is a D-dimensional random vector, distributed as $\mathcal{N}(0, I/D)$, where I is the identity. So for each dimension $d \in \{1, 2, \ldots, D\}$, we have $x_d \sim \mathcal{N}(0, 1/D)$ independently. Thus the expected length of x is

$$\mathbb{E}\left(\|x\|^2\right) = \mathbb{E}\left(\sum_d x_d^2\right) = \sum_d \mathbb{E}\left(x_d^2\right) = \sum_d 1/D = D \cdot 1/D = 1$$

which means that x is expected to be on the boundary of a unit sphere centered at the origin.
Moreover, we have

$$\begin{aligned} \mathbb{E}(\|x\|^4) =& \mathbb{E}\left(\left(\sum_d x_d^2\right)^2\right) = \mathbb{E}\left(\sum_d x_d^4 + \sum_d \sum_{d':d'\neq d} x_d^2 x_{d'}^2\right) \\ =& \sum_d \mathbb{E}\left(x_d^4\right) + \sum_d \sum_{d':d'\neq d} \mathbb{E}\left(x_d^2 x_{d'}^2\right) \\ =& \sum_d 3/D^2 + \sum_d \sum_{d':d'\neq d} \mathbb{E}\left(x_d^2\right)\mathbb{E}\left(x_{d'}^2\right) \qquad \text{since } x_d \perp\!\!\!\perp x_{d'} \\ =& 3/D + \sum_d \sum_{d':d'\neq d} 1/D^2 \\ =& 3/D + D(D-1) \cdot 1/D^2 \\ =& 3/D + (D-1)/D \\ =& 2/D + 1 \end{aligned}$$

Therefore, the variance of the length is

$$\mathrm{var}\left(\|x\|^2\right) = \mathbb{E}\left(\|x\|^4\right) - \left(\mathbb{E}(\|x\|^2)\right)^2 = 2/D + 1 - 1 = 2/D \to 0 \quad \text{as } D \to \infty$$

Thus, it is not only expected that x lies on the boundary but as D increases most of its realizations will in fact fall on the boundary. $\square$