CS189Hw2

Trustin Nguyen

October 28, 2024

Multivariate Gaussians: A review

Multivariate Gaussian distributions crop up everywhere in machine learning, from priors on model parameters to assumptions on noise distributions. Being able to manipulate multivariate Gaussians also becomes important for analyzing correlations in data and preprocessing it for better regression and classification. We want to make sure to first cover the MVG fundamentals here.

Note that the probability density function of a non-degenerate (i.e. the covariance matrix is positive definite and, thus, invertible) multivariate Gaussian RV with mean vector, $\mu \in \mathbb{R}^2$, and covariance matrix, $\Sigma \in \mathbb{R}^{2\times 2}$, is:

$$f(z) = \frac{1}{\sqrt{(2\pi)^2|\Sigma|}} \exp\left(-\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)\right)$$

(a) Consider a two dimensional, zero mean random variable $Z = \begin{bmatrix} Z_1 & Z_2 \end{bmatrix}^T \in \mathbb{R}^2$. In order for the random variable to be jointly Gaussian, a necessary and sufficient condition which we call the *first characterization* is that

- Z_1 and Z_2 are each marginally Gaussian, and
- $Z_1 \mid Z_2 = z$ is Gaussian, and $Z_2 \mid Z_1 = z$ is Gaussian.

A *second characterization* of a jointly Gaussian zero mean RV $Z \in \mathbb{R}^2$ is that it can be written as $Z = AX$, where $X \in \mathbb{R}^2$ is a collection of i.i.d. standard normal RVs and $A \in \mathbb{R}^{2\times 2}$ is a matrix.

Let X_1 and X_2 be i.i.d. standard normal RVs. Let U denote a binary random variable uniformly that is equal to 1 with probability $\frac{1}{2}$ and -1 with probability $\frac{1}{2}$, independent of everything else.

For each of the below subproblems, complete the following *two* steps: (1) Using one of the characterizations given above, determine whether the RVs are jointly Gaussian. If using the second characterization, clearly specify the A matrix. (2) Calculate the covariance matrix of Z (regardless of whether the RVs are jointly Gaussian or not).

(i.) $Z_1 = X_1$ and $Z_2 = X_2$

Answer. I will be using the first characterization. To show that Z_1 and Z_2 are marginally Gaussian, we need to show that:

$$p_{Z_2}(z) = \int_{-\infty}^{\infty} p(Z_1 = z', Z_2 = z)\, dz'$$

is Gaussian. Since the RVs are independent:

$$\begin{aligned} p_{Z_2}(z) &= \int_{-\infty}^{\infty} p(Z_1 = z')p(Z_2 = z)\, dz' \\ &= p(Z_2 = z)\int_{-\infty}^{\infty} p(Z_1 = z')\, dz' \\ &= p(Z_2 = z) \end{aligned}$$

so p_{Z_2} is Gaussian. The other way is symmetric.

To see that $Z_1 \mid Z_2 = z$ is Gaussian, we have:

$$p(Z_1 = z_1 \mid Z_2 = z) = \frac{p(Z_1 = z_1, Z_2 = z)}{p(Z_2 = z)}$$

Since they are independent:

$$p(Z_1 = z_1 \mid Z_2 = z) = \frac{p(Z_1 = z_1)p(Z_2 = z)}{p(Z_2 = z)} = p(Z_1 = z_1)$$

So the distribution of $Z_1 \mid Z_2 = z$ is the same as that of Z_1, and is therefore Gaussian. The other case is symmetric.

The covariance matrix of Z is the variance of Z_1 along the diagonal because the distributions are independent.

(ii.) $Z_1 = X_1$ and $Z_2 = X_1 + 2X_2$. If using the first characterization, assume that you already know $(Z_1 \mid Z_2 = z)$ is Gaussian.

Answer. It is jointly Gaussian because

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

by the second classification. From lecture, the covariance matrix is given by AA^T.

(iii.) $Z_1 = X_1$ and $Z_2 = -X_1$.

Answer. Yes these are joint Gaussian because $Z = \begin{bmatrix} 1 & 0 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$. Same covariance matrix in the previous question: AA^T.

(iv.) $Z_1 = X_1$ and $Z_2 = UX_1$.

Answer. We will use the first characterization here. To show marginal Gaussian:

$$p_{X_1}(z) = p(X_1 = z) = \int_{-\infty}^{\infty} p(X_1 = z, UX_1 = z')\, dz'$$

must be Gaussian. We know that it has non-zero values when $z' = \pm z$:

$$p_{X_1}(z) = \sum_{z'=\pm z} p(X_1 = z, UX_1 = z')$$

We can expand:

$$p(X_1 = z, UX_1 = z) + p(X_1 = z, UX_1 = -z) = p(UX_1 = z)p(U = 1) + p(X_1 = z)p(U = -1)$$

we know that:

$$p(U = \pm 1) = 1/2$$

So this turns out to just $p(X_1 = z)$. So $p_{X_1}(z)$ is gaussian.

Now for the other marginal:

$$p_{UX_1}(z) = p(UX_1 = z) = \int_{-\infty}^{\infty} p(X_1 = z', UX_1 = z)\, dz'$$

Again, nonzero when $X_1 = \pm z$:

$$\sum_{z'=\pm z} p(X_1 = z', UX_1 = z) = p(X_1 = z)p(U = 1) + p(X_1 = -z)p(U = -1)$$

which is

$$\frac{1}{2}(p(X_1 = z) + p(X_1 = -z))$$

Since X_1 is standard normal, this becomes $p(X_1 = z)$ which is normal.

Now on to the conditional probabilities. We have

$$p(X_1 = x \mid UX_1 = x') = \frac{p(X_1 = x, UX_1 = x')}{p(UX_1 = x')}$$

expand:

$$\frac{p(X_1 = x, UX_1 = x') + p(X_1 = x, UX_1 = -x')}{p(X_1 = x')(p(U = -1) + p(U = 1))}$$

which is

$$\frac{p(X_1 = x)p(X_1 = x')}{p(X_1 = x')} = p(X_1 = x)$$

which is gaussian.

For the other conditional:

$$p(UX_1 = x \mid X_1 = x') = \frac{p(UX_1 = x, X_1 = x')}{p(X_1 = x')}$$

expand:

$$\frac{p(X_1 = x')(p(UX_1 = -x) + p(UX_1 = x))}{p(X_1 = x')} = p(UX_1 = x) + p(UX_1 = -x)$$

so this is gaussian also.

Now for the covariance,

$$Cov(UX, X) = \mathbb{E}[(UX - \mathbb{E}(UX))(X)]$$

or

$$Cov(UX, X) = \mathbb{E}[UX^2 - X] = -\mathbb{E}[X] = 0$$

So the matrix is

$$\begin{bmatrix} 1 & 0 \\ 0 & Var(UX) \end{bmatrix}$$

and

$$Var(UX) = \mathbb{E}[U^2X^2] - \mathbb{E}[UX]^2 = \mathbb{E}[U^2X^2] = \mathbb{E}[U^2]\mathbb{E}[X^2] = \mathbb{E}[X^2] = 1$$

(b) Show that two Gaussian random variables can be uncorrelated, but not independent. On the other hand, show that two uncorrelated, jointly Gaussian RVs are independent.

Answer. In part (d) of last example, we computed a covariance matrix of $\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$ but UX, X are not independent. So we see that non correlation does not mean independence.

If two Gaussians are uncorrelated, then from the joint pdf:

$$p_{Z_1,Z_2}(z) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left\{-\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)\right\}$$

Σ is a diagonal matrix, so the exponent part factors into

$$Var(z_1)\cdot(z_1-\mu_1)^2 + Var(z_2)\cdot(z_2-\mu_2)^2$$

The determinant in the denominator also factors into: $Var(z_1)Var(z_2)$. So the joint pdf factors into their marginal distributions:

$$p_{Z_1,Z_2}(z) = \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left\{-\frac{1}{2}\left(\frac{z_1-\mu_1}{\sigma_1}\right)^2\right\}\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left\{-\frac{1}{2}\left(\frac{z_2-\mu_2}{\sigma_2}\right)^2\right\}$$

(c) With the setup in (a), let $Z = VX$, where $V \in \mathbb{R}^{2\times 2}$, and $Z, X \in \mathbb{R}^2$. What is the covariance matrix Σ_Z? If X is not a multivariate Gaussian but has the identity matrix $I \in \mathbb{R}^{2\times 2}$ as its covariance matrix, is your computed Σ_Z still the covariance of Z?

Answer. Recall from the previous homework that if $Z = AX + b$, then

$$Var(Z) = A\Sigma A^T$$

where Σ is the covariance matrix of X. Then the covariance matrix of Z is

$$V\Sigma V^T$$

Here, Σ is the identity because X is standard normal. This also implies that if X has an identity matrix as the covariance matrix, the computed covariance does not change.

(d) Given a jointly Gaussian zero mean RV $Z = \begin{bmatrix}Z_1 & Z_2\end{bmatrix}^T \in \mathbb{R}^2$ with covariance matrix $\Sigma_Z = \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{12} & \Sigma_{22}\end{bmatrix}$, derive the conditional distribution of $(Z_1 \mid Z_2 = z)$.

Hint: The following identity may be useful

$$\begin{bmatrix}a & b\\ b & c\end{bmatrix}^{-1} = \begin{bmatrix}1 & 0\\ -\frac{b}{c} & 1\end{bmatrix}\begin{bmatrix}\left(a-\frac{b^2}{c}\right)^{-1} & 0\\ 0 & \frac{1}{c}\end{bmatrix}\begin{bmatrix}1 & -\frac{b}{c}\\ 0 & 1\end{bmatrix}$$

Answer. The joint pdf is:

$$\frac{1}{\sqrt{2\pi|\Sigma|}}\exp\left\{-\frac{1}{2}(Z-\mu)^T\Sigma^{-1}(Z-\mu)\right\}$$

And from Bayes's rule:

$$p_{Z_1}(Z_1 = z_1, Z_2 = z) = p(Z_1 \mid Z_2 = z)p(Z_2 = z)$$

What we want to solve for is $p(Z_1 \mid Z_2 = z)$ on the RHS. So the LHS is:

$$\text{LHS} = \frac{1}{\sqrt{2\pi\left|\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{12} & \Sigma_{22}\end{bmatrix}\right|}}\exp\left\{-\frac{1}{2}\begin{bmatrix}z_1 & z\end{bmatrix}\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{12} & \Sigma_{22}\end{bmatrix}^{-1}\begin{bmatrix}z_1\\ z\end{bmatrix}\right\}$$

$$= \frac{1}{\sqrt{2\pi(\Sigma_{11}\Sigma_{22}-\Sigma_{12}^2)}}\exp\left\{-\frac{1}{2}\begin{bmatrix}z_1 & z\end{bmatrix}\begin{bmatrix}1 & 0\\ -\frac{\Sigma_{12}}{\Sigma_{22}} & 1\end{bmatrix}\begin{bmatrix}\frac{\Sigma_{22}}{\Sigma_{11}\Sigma_{22}-\Sigma_{12}^2} & 0\\ 0 & \frac{1}{\Sigma_{22}}\end{bmatrix}\begin{bmatrix}1 & -\frac{\Sigma_{12}}{\Sigma_{22}}\\ 0 & 1\end{bmatrix}\begin{bmatrix}z_1\\ z\end{bmatrix}\right\}$$

Recall that we still had the $p(Z_2 = z)$ term on the RHS which is:

$$p(Z_2 = z) = \frac{1}{\sqrt{2\pi\Sigma_{22}}} \exp\left\{-\frac{1}{2}\begin{bmatrix} z_1 & z \end{bmatrix}\begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{\Sigma_{22}} \end{bmatrix}\begin{bmatrix} z_1 \\ z \end{bmatrix}\right\}$$

So when we take $p_{Z_1}(Z_1 = z_1, Z_2 = z)/p(Z_2 = z)$, we would end up subtracting the exponents in the exp() term:

$$\begin{bmatrix} z_1 & z \end{bmatrix}\begin{bmatrix} 1 & 0 \\ -\frac{\Sigma_{12}}{\Sigma_{22}} & 1 \end{bmatrix}\begin{bmatrix} \frac{\Sigma_{22}}{\Sigma_{11}\Sigma_{22}-\Sigma_{12}^2} & 0 \\ 0 & \frac{1}{\Sigma_{22}} \end{bmatrix}\begin{bmatrix} 1 & -\frac{\Sigma_{12}}{\Sigma_{22}} \\ 0 & 1 \end{bmatrix}\begin{bmatrix} z_1 \\ z \end{bmatrix} - \begin{bmatrix} z_1 & z \end{bmatrix}\begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{\Sigma_{22}} \end{bmatrix}\begin{bmatrix} z_1 \\ z \end{bmatrix}$$

and notice that:

$$\begin{bmatrix} z_1 & z \end{bmatrix}\begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{\Sigma_{22}} \end{bmatrix}\begin{bmatrix} z_1 \\ z \end{bmatrix} = \begin{bmatrix} z_1 & z \end{bmatrix}\begin{bmatrix} 1 & 0 \\ -\frac{\Sigma_{12}}{\Sigma_{22}} & 1 \end{bmatrix}\begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{\Sigma_{22}} \end{bmatrix}\begin{bmatrix} 1 & -\frac{\Sigma_{12}}{\Sigma_{22}} \\ 0 & 1 \end{bmatrix}$$

So subtracting yields:

$$\begin{bmatrix} z_1 & z \end{bmatrix}\begin{bmatrix} 1 & 0 \\ -\frac{\Sigma_{12}}{\Sigma_{22}} & 1 \end{bmatrix}\begin{bmatrix} \frac{\Sigma_{22}}{\Sigma_{11}\Sigma_{22}-\Sigma_{12}^2} & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 1 & -\frac{\Sigma_{12}}{\Sigma_{22}} \\ 0 & 1 \end{bmatrix}\begin{bmatrix} z_1 \\ z \end{bmatrix}$$

We can simplify the matrix in the middle down to:

$$\begin{bmatrix} \frac{\Sigma_{22}}{|\Sigma|} & -\frac{\Sigma_{12}}{|\Sigma|} \\ -\frac{\Sigma_{12}}{|\Sigma|} & \frac{\Sigma_{12}^2}{\Sigma_{22}|\Sigma|} \end{bmatrix}$$

The rest is just algebraic manipulation.

Projections and Linear Regression

We are given $X \in \mathbb{R}^{n \times d}$ where $n > d$ and $\mathrm{rank}(X) = d$. We are also given a vector $y \in \mathbb{R}^n$. Define the orthogonal projection of y onto $\mathrm{range}(X)$ as $P_{\mathrm{range}(X)}(y)$.

Background on orthogonal projections: For any finite-dimensional subspace W (here, $\mathrm{range}(X)$) of a vector space V (here, $\mathbb{R}^n$), any vector $v \in V$ can be decomposed as

$$v = w + u,\ w \in W,\ u \in W^\perp$$

where $W^\perp$ is the orthogonal complement of W. Furthermore, this decomposition is unique: if $v = w' + u'$ where $w' \in W, u' \in W^\perp$, then $w' = w$ and $u' = u$. These two facts allow us to define P_W, the orthogonal projection operator onto W. Given a vector v with decomposition $v = w + u$, we define

$$P_W(v) = w.$$

It can also be shown using these two facts that P_W is linear.

(a) Prove that $P_{\mathrm{range}(X)}(y) = \mathrm{argmin}_{w \in \mathrm{range}(X)} \|y - w\|_2^2$

Answer. We know that $y = x + w$ for $x \in \mathrm{range}(X)$ and $w \in X^\perp$. So the equation becomes:

$$\operatorname*{argmin}_{x' \in \mathrm{range}(X)} \|x + w - x'\|_2^2$$

The norm is:

$$(x + w - x') \cdot (x + w - x')$$

a dot product:

$$\begin{aligned}((x - x') + w) \cdot ((x - x') + w) &= (x - x') \cdot (x - x') + 2(w \cdot (x - x')) + w \cdot w \\ &= \|x - x'\| + \|w\|\end{aligned}$$

We want the x' which minimizes this. So $x' = x = P_{\mathrm{range}(X)}(y)$.

(b) An orthogonal projections is a linear transformation. That is, $P_{\mathrm{range}(X)}(y) = Py$ for some projection matrix P

Specifically, given $1 \leqslant d \leqslant n$, a matrix $P \in \mathbb{R}^{n \times n}$ is said to be a rank-d orthogonal projections matrix if

- $\mathrm{rank}(P) = d$
- $P = P^T$
- $P^2 = P$

Prove that P is a rank-d projection matrix if an only if there exists a $U \in \mathbb{R}^{n \times d}$ such that $P = UU^T$ and $U^TU = I$.

Answer. ($\rightarrow$) From the three statements above, $P = P^TP$, which means that P is symmetric, there is an eigenvalue decomposition. The eigenvalues are 1 for elements in the basis of X and 0 in the subspace W/X. So:

$$P = U'IU'^T$$

but U' has d non-zero vectors in its column space. So $P = UU^T$, $U^TU = I$ because U consists of orthonormal vectors.

($\leftarrow$) Since $U^TU = I$, U is full rank. We know this because the dot products not along the diagonal are 0, so there cant be any linear dependence. This means that the rank of P is also d because:

$$U^TU = I \implies UU^TU = I \implies PU = U$$

and therefore, the rank of P cannot be any less. It also cannot be any more because $P = UU^T$ and is restricted by $\mathrm{rank}(U)$. The other two properties of an orthogonal projection matrix can be shown by algebraic manipulation.

(c) The Singular Value Decomposition theorem states that we can write any matrix X as

$$X = \sum_{i=1}^{\min(n,d)} \sigma_i u_i v_i^T = \sum_{i:\sigma_i>0} \sigma_i u_i v_i^T$$

where $\sigma_i \geqslant 0$, and $\{u_i\}_{i=1}^n$ and $\{v_i\}_{i=1}^d$ are orthonormal bases for $\mathbb{R}^n$ and $\mathbb{R}^d$ respectively. Some of the singular values σ_i may equal 0, indicating that the associated left and right singular vectors u_i and v_i do not contribute to the sum, but sometimes it is still convenient to include them in the SVD so we have complete orthonormal bases for $\mathbb{R}^n$ and $\mathbb{R}^d$ to work with. Show that

(i) $\{u_i : \sigma_i > 0\}$ is an orthonormal basis for the column space of X.

Answer.

(ii) Similarly, $\{v_i : \sigma_i > 0\}$ is an orthonormal basis for the row space of X.

(d)

(e)

(f)

Some MLEs

For this question, assume you observe n (data point, label) pairs $(x_i, y_i)_{i=1}^n$, with $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ for all $i = 1, \ldots, n$. We denote X as the data matrix containing all the data points and y as the label vector containing all the labels:

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \in \mathbb{R}^{n \times d}, \qquad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$$

(a) Ignoring y for now, suppose we model the data points as coming from a d-dimensional Gaussian with diagonal covariance:

$$\forall i = 1, \ldots, n,\ x_i \sim N(\mu, \Sigma);\ \Sigma = \begin{bmatrix} \sigma_1^2 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_d^2 \end{bmatrix}$$

If we consider $\mu \in \mathbb{R}^d$ and $(\sigma_1^2, \ldots, \sigma_d^2)$, where each $\sigma_i^2 > 0$, to be unknown, the parameter space here is $2d$-dimensional. When we refer to Σ as a parameter, we are referring to the d-tuple $(\sigma_1^2, \ldots, \sigma_d^2)$, but inside a linear algebraic expression, Σ denotes the diagonal matrix $\mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$.

Solve the following problems:

(i) Prove that log-likelihood $l(\mu, \Sigma) = \log p(X \mid \mu, \Sigma)$ is equal to

$$-\frac{n}{2}\left(d \log 2\pi - \sum_{j=1}^{d} \log \frac{1}{\sigma_j^2}\right) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^T \Sigma^{-1}(x_i - pm)$$

(ii)

(iii)

(iv)

(b)

(c)

(d)