STA561: Probabilistic machine learning

Gaussian Models (9/9/13)

Lecturer: Barbara Engelhardt *Scribes: Xi He, Jiangwei Pan, Ali Razeen, Animesh Srivastava*

1 Multivariate Normal Distribution

The multivariate normal distribution (MVN), also known as **multivariate gaussian**, is a generalization of the one-dimensional normal distribution to higher dimensions. The probability density function (pdf) of an MVN for a random vector $x \in \mathbb{R}^d$ as follows:

$$\mathcal{N}(x|\mu, \Sigma) \triangleq \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right] \tag{1}$$

where $\mu = \mathbb{E}[x] \in \mathbb{R}^d$ is the mean vector, and $\Sigma = cov[x]$ is $d \times d$ symmetric positive definite matrix, known as the covariance matrix. Σ^{-1} is known as the precision matrix.

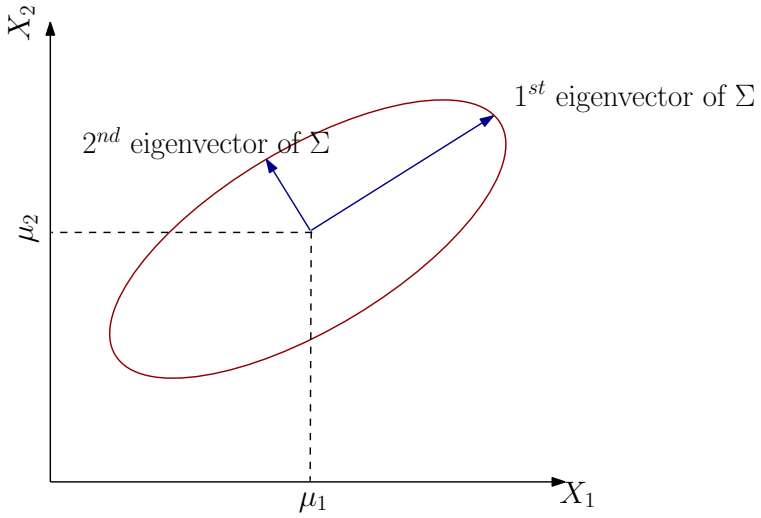

Figure 1: 2 dimensional Gaussian density.

Fig. 1 shows a 2-dimensional Gaussian density. The random vectors span two dimensions and are denoted in the plot by X_1 (x-axis) and X_2 (y-axis). The means of X_1 and X_2 are μ_1 and μ_2 respectively. The density at μ is highest, and as the random vector moves away from μ, the density goes down. All of the points on the red contour (level set) have the same density. The first and second eigenvectors of the covariance matrix are orthogonal to each other as shown in the Fig. 1. The first eigenvalue is the direction of maximum variance in the MVN; the second eigenvector is orthogonal to the first.

The expression inside the exponent can be rewritten as: $\sqrt{(x-\mu)^T\Sigma^{-1}(x-\mu)}$. The Mahalanobis distance between two vectors x_1 and x_2 is equivalent to the MVN, calculating the pdf of one of the two points, with

the other as the mean:

$$md(x_1, x_2) = \sqrt{(x_1 - x_2)^T \Sigma^{-1} (x_1 - x_2)}. \tag{2}$$

Note that this distance is symmetric.

2 MLE for MVN

In order to determine the MLE for a MVN, we need some basic results from linear algebra. Recall the following definitions.

- The *trace* of a matrix $A \in \mathbb{R}^{d \times d}$ is defined as the sum of A's diagonal elements, i.e., $tr(A) = \sum_i^d A_{ii}$.
- The *determinant* of a matrix $A \in \mathbb{R}^{d \times d}$ is defined as the product of its eigenvalues. A positive definite matrix A has positive eigenvalues, so the determinant will always be positive.
- Symmetric positive definite matrices (as we will consider for our covariance matrices) are defined as having eigenvalues that are strictly positive.
- Trace has the *cyclic permutation* property:

$$tr(ABC) = tr(CAB) = tr(BCA)$$

Given vectors a, b and matrices A, B, C, we have the following facts:

- $\frac{\partial b^T a}{\partial a} = b$
- $\frac{\partial (a^T A a)}{\partial a} = (A + A^T)a$. (Note that if A is symmetric, this equals to $2Aa$)
- $\frac{\partial tr(BA)}{\partial A} = B^T$
- $\frac{\partial \log |A|}{\partial A} = A^{-T} \triangleq (A^{-1})^T$
- *Trace trick:* $a^T A a = tr(a^T A a) = tr(a a^T A) = tr(A a a^T)$

We will be using these aforementioned facts in deriving the MLEs, $\hat{\mu}_{MLE}$ and $\hat{\Sigma}_{MLE}$, for a MVN given a data set $\mathcal{D} = \{x_1, x_2, x_3, ..., x_n\}$, where $x_i \in \mathbb{R}^d$ is a sample vector from the MVN. The log likelihood of the data set $\mathcal{D}$ given MVN parameters μ, Σ can be written as

$$\begin{aligned} \mathcal{L}(\mu, \Sigma; \mathcal{D}) &= \log \prod_{i=1}^{n} p(x_i | \mu, \Sigma) && (3) \\ &= \frac{n}{2} \log |\Sigma^{-1}| - \frac{1}{2} \sum_{i=1}^{n} tr[(x_i - \mu)(x_i - \mu)^T \Sigma^{-1}] && (4) \end{aligned}$$

Set the partial derivative with respect to μ to 0,

$$\frac{\partial \mathcal{L}(\mu, \Sigma; \mathcal{D})}{\partial \mu} = -\frac{1}{2} \sum_{i=1}^{n} -2\Sigma^{-1}(x_i - \mu)] = 0$$

we get MLE of μ as follows,

$$\begin{aligned} & \sum_{i=1}^{n}(x_i - \mu) = 0 \\ \Rightarrow \quad & \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \mu = 0 \\ \Rightarrow \quad & \mu_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i \end{aligned} \tag{5}$$

This means that the MLE of μ for MVN is just the empirical mean of the samples.

Similarly, setting the partial derivative of the log likelihood (Equation 4) with respect to Σ^{-1} to 0,

$$\frac{\partial \mathcal{L}(\mu, \Sigma; \mathcal{D})}{\partial \Sigma^{-1}} = \frac{n}{2}\Sigma - \frac{1}{2}(\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T) = 0$$

we get MLE of Σ as follows,

$$\Sigma_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T \tag{6}$$

This expression is just the empirical covariance of the data, centered on μ.

3 The MVN is in the Exponential Family

We have already seen that if $x \sim \mathcal{N}(\mu, \Sigma)$ then $E[x] = \mu$ and $cov[x] = \Sigma$. These are also called the *mean* or the *moment* parameters of the distribution. We can also express the MVN in exponential family form in terms of the *natural parameters* as

$$\Lambda = \Sigma^{-1}, \ \eta = \Sigma^{-1}\mu. \tag{7}$$

Similarly, we can convert the natural parameters back to moment parameters as

$$\Sigma = \Lambda^{-1}, \ \mu = \Lambda^{-1}\eta. \tag{8}$$

Note that the natural parameter covariance matrix is the precision matrix. Also note that the relationship between the mean parameters and the natural parameters is an invertible relationship, so the MLE for the natural parameters can be converted into the MLE for the mean parameters (and vice versa). This enables us to work in the most mathematically convenient space, and convert afterwards between parameterizations.

We can rewrite the MVN density, in Eqn (1), in exponential family form using the natural parameters as follows:

$$\begin{aligned} P(x|\eta, \Lambda) &= (2\pi)^{-d/2}|\Lambda|^{1/2}\exp\Big[-\frac{1}{2}(x^T\Lambda x + \eta^T\Lambda\eta - 2x^T\eta)\Big] \\ &= (2\pi)^{-d/2}|\Lambda|^{1/2}\exp\Big[\eta^T x - \frac{1}{2}x^T\Lambda x - \frac{1}{2}\eta^T\Lambda\eta\Big] \\ &= (2\pi)^{-d/2}|\Lambda|^{1/2}\exp\Big[\eta^T x - \frac{1}{2}tr(\Lambda x x^T) - \frac{1}{2}\eta^T\Lambda\eta\Big] \end{aligned} \tag{9}$$

Recall the exponential family form:

$$P(X|\eta) = h(X)\exp\{\eta^T T(X) - A(\eta)\}, \tag{10}$$

where η in Eqn (10) denotes the natural parameter vector, and T(x) is the vector of sufficient statistics for the MVN.

We can see then that Eqn (9) is in exponential family form, and we read out the sufficient statistics of MVN:

$$T(x) = \begin{bmatrix} x \\ xx^T \end{bmatrix},$$

and the natural parameters are

$$\tilde{\eta} = \begin{bmatrix} \eta \\ -\frac{1}{2}\Lambda \end{bmatrix}$$

and the log partition function is

$$A(\eta, \Lambda) = \frac{1}{2}\eta^T \Lambda \eta = \frac{1}{2}\Lambda \eta \eta^T.$$

Thus, the sufficient statistics for a MVN are the empirical mean and the empirical covariance.

4 Marginals and Conditionals for an MVN

Let's consider an example where $d = 2$. If it is simpler, let $X = (x_1, x_2)$ where x_1 and x_2 are scalar. However, if we consider x_1 and x_2 to be a split of the MVN data in dimension $d > 2$, where each x_1 and x_2 is a vector, all of this subsequent section goes through naturally. Suppose that x_1 and x_2 are jointly Gaussian:

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad , \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \quad \text{and hence } \Lambda = \Sigma^{-1} = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}$$

Can we find the marginal and conditional distributions in this space? Recall that

$$P(x_1) = \int_{x_2} \mathcal{N}(x_1, x_2 | \mu, \Sigma) dx_2.$$

Using this model we can derive the following distributions with Eqn. (8):

- Marginal: $P(x_1) = \mathcal{N}(x_1 | \mu_1, \Sigma_{11})$, where $\mu_1 = \eta_2 - \Lambda_{21}\Lambda_{11}^{-1}\eta_1$ and $\Sigma_{11} = \Lambda_{22} - \Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12}$.
- Marginal (equivalent): $P(x_2) = \mathcal{N}(x_2 | \mu_2, \Sigma_{22})$, where $\mu_2 = \eta_1 - \Lambda_{12}\Lambda_{22}^{-1}\eta_2$ and $\Sigma_{22} = \Lambda_{11} - \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}$.
- Conditional distribution: $P(x_1 | x_2) = \mathcal{N}(x_1 | \mu_{1|2}, \Sigma_{1|2})$, where $\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$, $\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, or more concisely (natural parameters), $\mu_{1|2} = \eta_1 - \Lambda_{12}x_2$, $\Sigma_{1|2} = \Lambda_{11}$.

The converse conditional distribution, $p(x_2|x_1)$ is written out equivalently (swapping the 1, 2 indices). These formulas are derived using the Schur complement of a matrix and the matrix inversion lemma. Note that conditional probabilities are straightforward to consider in the natural parameter space, where marginal probabilities are much simpler in the mean parameter space.

The marginal distribution in the mean parameter space is a simple projection of a (for example) 2D MVN cloud onto each of the univariate Gaussian distributions i one dimension. The conditional distribution is a similar projection, but considering only a slice of the space at the conditional random variable. When the off-diagonal elements of the covariance matrix are 0, the conditional distribution is identical to the marginal distribution (as the two univariate Gaussians are independent).

5 Conjugate prior

The conjugate prior for the mean term μ of a multivariate normal distribution is a multivariate normal distribution:

$$p(\mu|X) \propto p(\mu)p(X|\mu), \tag{11}$$

where $p(\mu)$ is a multivariate normal distribution, $\mu \sim \mathcal{N}(\mu_0, \Sigma_0)$. The implication of this prior is that the mean term has a Gaussian distribution across the space that it might lie in: generally large values of Σ_0 are preferable unless we have good prior information about the mean term (e.g., that it will be right around zero).

The conjugate prior for the covariance matrix Σ of a multivariate normal distribution is the inverse Wishart distribution:

$$p(\Sigma|X) \propto p(\Sigma)p(X|\Sigma), \tag{12}$$

where $p(\Sigma)$ is an inverse Wishart distribution $\Sigma \sim \mathcal{IW}(\nu, \Psi)$. The inverse Wishart is a PDF for positive definite matrices.