

Matrix calculus and MLE for the multivariate Normal (9/24/07)

Lecturer: Daniel Ting

Scribes:

Suppose we want to find the MLE under a multivariate Normal distribution with unknown mean  $\mu$  and covariance  $\Sigma$ .

Recall the density of a  $d$ -dimensional multivariate  $Normal(\mu, \Sigma)$ .

$$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-(x-\mu)^T \Sigma^{-1} (x-\mu)/2}$$

So given data  $\mathcal{D} = x_i : i = 1 \dots N \in \mathcal{R}^d$  we get a log-likelihood

$$l(\mu, \Sigma|\mathcal{D}) = -\frac{Nd}{2} \log(2\pi) - \frac{N}{2} \log|\Sigma| - \sum_{i=1}^N (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)/2$$

We want to take the derivatives wrt the parameters  $\mu$  and  $\Sigma$  and set them to 0 to find the MLE, but how? This is a basic tutorial on matrix calculus. (Might be useful if you decide to take EE227a convex opt too)

## 1 Vector calculus

### 1.1 Simple linear function warmup

We start with a simple linear function. Let  $A$  be an  $m \times n$  matrix

$$f(x) = Ax$$

This is a function from  $\mathbb{R}^n$  to  $\mathbb{R}^m$  which means the derivative is an  $m \times n$  matrix. (Write out the matrix of partial derivatives)

The derivative is easy.

$$\frac{dAx}{dx} = A$$

Note that this falls out immediately from the definition of derivative as the linear operator that's a "good approximation" ie. the derivative  $L_x$  of  $f$  at  $x$  is the linear operator such that

$$\frac{f(x+h) - f(x) - L_x h}{\|h\|} = o(h)$$

In this case  $f(x) = Ax$  and trying  $L_x = A$  gives

$$\frac{A(x+h) - Ax - Ah}{\|h\|} = 0 = o(h)$$

so  $L_x = A$  is the derivative. This is super intuitive since the best linear “approximation” to a linear operator  $A$  is just itself.

## 1.2 Quadratic forms

### 1.2.1 Basic quadratic form

Now consider something slightly more complicated. Let  $A$  be a symmetric  $n \times n$  matrix

$$f(x) = x^T Ax$$

Now we have a function from  $\mathbb{R}^n \rightarrow \mathbb{R}$  and the derivative should be an  $1 \times n$  matrix (it also maps  $\mathbb{R}^n \rightarrow \mathbb{R}$ ).

We can also use first principles in calculating the partial derivatives.

Do some calculations by taking

$$\frac{(x + he_i)^T A(x + he_i) - x^T Ax}{h} = x^T A^T he_i + x^T Ahe_i + h^2 e_i^T Ae_i$$

Since  $A$  is symmetric, we get

$$\frac{dx^T Ax}{dx} = x^T A^T + x^T A = x^T A + x^T A = 2x^T A$$

(We could also have just used the product rule to derive this)

### 1.2.2 Product rule

Assume  $y$  is  $m$ -dim,  $x$  is  $n$ -dim  $B$  is  $m \times m$  symmetric.

$A$  is an  $m \times n$  matrix.

Consider

$$f(x) = (y - Ax)^T B(y - Ax)$$

Again we have a function from  $\mathbb{R}^n \rightarrow \mathbb{R}$  and the derivative should be an  $1 \times n$  matrix.

We have two methods for calculating the derivative already, write out the analytic expression and take the partial derivatives or use first principles and manipulate some vectors scaled to have norm  $h$ .

We can also try to take a shortcut and use the product rule.

$$\frac{\partial YZ}{\partial X} = \frac{\partial Y}{\partial X} Z + Y \frac{\partial Z}{\partial X}$$

Note that since matrix multiplication does not commute, you cannot switch the order of  $Y$  and  $Z$ .

We also need another trick. For a vector  $v$

$$\frac{dx^T v}{dx} = \frac{dv^T x}{dx} = v^T$$

We do this so we can get the  $x$  on the right side and go back to our simple linear operator warmup case. This works since  $x^T v$  is a scalar and the transpose of a scalar is just itself. The derivative is  $v^T$  from our warmup.

$$\begin{aligned} & \frac{d(y - Ax)^T B(y - Ax)}{dx} \\ = & \frac{d(y - Ax)^T}{dx} B(y - Ax) + (y - Ax)^T \frac{dB(y - Ax)}{dx} \\ = & (y - Ax)^T B^T \frac{d(y - Ax)}{dx} (y - Ax) + (y - Ax)^T \frac{dB(y - Ax)}{dx} \\ = & (y - Ax)^T B^T (-A) + (y - Ax)^T B(-A) \\ = & -2(y - Ax)^T B A \end{aligned}$$

Note that  $B(y - Ax)$  is a vector which is how the transpose trick comes in. We use our assumption that  $B$  is symmetric for the last step to group the two terms together.

## 2 Trace

The trace of a square matrix  $A$  is the sum of the diagonals

$$tr(A) = \sum_i a_{ii}$$

Trace is often important because it “sort of” commutes. It is invariant under cyclic permutations

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

or in the simple case

$$tr(AB) = tr(BA)$$

In these eqns,  $A, B, C$  need not be square matrices, as long as their dimensions are conformal and the product  $ABC$  is a square matrix (these automatically imply that  $BCA$  and  $CAB$  are square).

(do the calc for  $tr(AB) = tr(BA)$  by calculating the terms on the diagonal)

This applies to ML estimation for the normal since

$$(x_i - \mu)^T \Sigma^{-1} (x_i - \mu) = tr((x_i - \mu)^T \Sigma^{-1} (x_i - \mu)) = tr(\Sigma^{-1} (x_i - \mu)(x_i - \mu)^T)$$

(take sum over  $i$  and group terms by pulling in the sum. Note that we can do this since trace is a linear function of the matrix entries)

## 2.1 Derivatives

Let  $A$  be  $m \times n$ ,  $B$  be  $n \times m$ .

We consider the matrix derivative

$$\frac{dtr(AB)}{dA}$$

Instead of a vector, we are now taking the derivatives wrt to some matrix  $A$ . What we are doing is taking the partial derivatives wrt each matrix entry.

(\*The more mathematical interpretation is to view an  $m \times n$  matrix as an  $mn$ -dim vector. That way the mathematical definition of derivative as a linear operator in a vector space still applies. We leave out the details of this interpretation.)

For a function  $\mathbb{R}^{m \times n} \rightarrow \mathbb{R}$  this means that we should get a derivative that is an  $m \times n$  matrix.

(Write out the terms on the diagonal of  $AB$  and take the derivative wrt  $b_{ij}$  to get derivative wrt  $a_{ij}$  is  $b_{ji}$ )

$$(AB)_{ii} = \sum_j a_{ij} b_{ji}$$

$$\frac{\partial}{\partial a_{ij}} tr(AB) = \frac{\partial}{\partial a_{ij}} \sum_k (AB)_{kk} = b_{ji}$$

so

$$\frac{dtr(AB)}{dA} = B^T$$

Also, since  $tr(AB) = tr(BA)$ , we have  $\frac{dtr(AB)}{dB} = A^T$

## 3 Determinant

In the normal log-likelihood, we find that we need to calculate the derivative of

$$\log|\Sigma|$$

Instead of  $\Sigma$  we use some arbitrary square matrix  $A$  with positive determinant.

Applying the chain rule gives us

$$\frac{\partial \log|A|}{\partial a_{ij}} = \frac{1}{|A|} \frac{\partial |A|}{\partial a_{ij}}$$

Recall the cofactor expansion of the determinant of a matrix.

$$|A| = \sum_j (-1)^{i+j} a_{ij} M_{ij}$$

(Describe the minors of a matrix: Delete the  $i$ th row,  $j$ th column and then calculate the determinant)

so the partial derivatives are given by

$$\frac{\partial |A|}{\partial a_{ij}} = (-1)^{i+j} M_{ij}$$

Thus the derivative  $\frac{d|A|}{dA}$  is the matrix of cofactors.

Recall Cramer's rule for the inverse of a matrix in terms of the determinant of a matrix and the minors.

$$A^{-1} = \frac{1}{|A|} \tilde{A}$$

where  $\tilde{A}_{ij} = (-1)^{i+j} M_{ji}$ , the adjugate matrix. The adjugate matrix is the transpose of matrix of cofactors.

This gives us

$$\frac{d|A|}{dA} = \tilde{A}^T$$

and

$$\frac{d \log |A|}{dA} = A^{-T}$$

For symmetric matrices (such as covariance matrices), the inverse is also symmetric ( $A^{-1}A = I = I^T = (AA^{-1})^T = (A^{-1})^T A^T = (A^{-1})^T A$  so  $A^{-1} = (A^{-1})^T$ ), so we have

$$\frac{d \log |A|}{dA} = A^{-1}$$

## 4 MLE

We now put this all together

$$l(\mu, \Sigma | \mathcal{D}) = -\frac{Nd}{2} \log(2\pi) - \frac{N}{2} \log |\Sigma| + \sum_{i=1}^N (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) / 2$$

Take the derivative wrt  $\mu$

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^N (x_i - \mu)^T \Sigma^{-1}$$

(Take  $B = \Sigma^{-1}$ ,  $A = I$  from our prev calculation.)

setting to 0, multiplying both sides by  $\Sigma$  and we get our MLE for  $\mu$

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N x_i$$

#### 4.1 MLE for $\Sigma$

We first use the fact that  $|A^{-1}| = |A|^{-1}$  and the trace trick.

$$l = \frac{N}{2} \log |\Sigma^{-1}| - \sum_{i=1}^N \text{tr} (\Sigma^{-1} (x_i - \mu)(x_i - \mu)^T) / 2 + C$$

Taking derivatives wrt to  $\Sigma^{-1}$  we get

$$\frac{\partial l}{\partial \Sigma^{-1}} = \frac{N}{2} \Sigma - \sum_{i=1}^N (x_i - \mu)(x_i - \mu)^T / 2$$

(Setting to 0, using our MLE  $\hat{\mu}$  and doing some algebra gives)

$$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{\mu})(x_i - \hat{\mu})^T$$