ECE7252
Statistical Learning for Signal Processing

Matrix Algebra for Multivariate Gaussian

Chin-Hui Lee

School of Electrical and Computer Engineering

Georgia Institute of Technology

Atlanta, GA 30332, USA

chl@ece.gatech.edu

Outline

- What is multivariate Gaussian?
- Parameterizations
- Mathematical Preparation
- Joint distributions, Marginalization and conditioning
- Maximum likelihood estimation

What is Multivariate Gaussian?

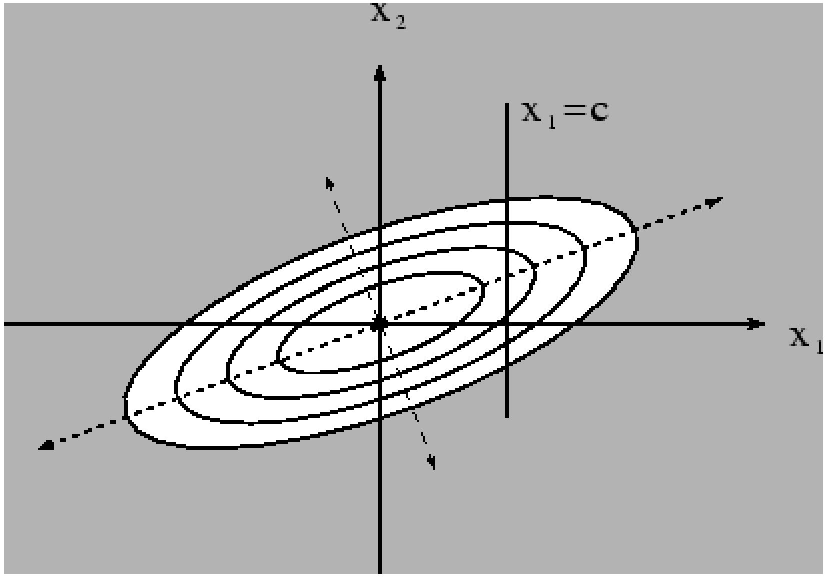
$$p(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\}$$

Where *x* is a *n*1* vector, Σ is an *n*n*, symmetric matrix

$$\Sigma^{-1} = \begin{pmatrix} \langle x_1 \rangle^2 & \langle x_1, x_2 \rangle & \cdots \\ \langle x_1, x_2 \rangle & \langle x_2 \rangle^2 & \vdots \\ \vdots & \cdots & \ddots \end{pmatrix}$$

Geometrical Interpretation

- This is a ellipse with the coordinate x1 and x2



Thus we can easily image that when *n* increases the ellipse became higher dimension ellipsoids

Parameterization

Another type of parameterization, putting it into the form of exponential family:

$$\mu = E(x)$$

$$\Sigma = E(x-\mu)(x-\mu)^T$$

$$p(x|\eta,\Lambda) = \exp\{a + \eta^T x - \frac{1}{2} x^T \Lambda x\}$$

$$\Lambda = \Sigma^{-1}$$

$$\eta = \Sigma^{-1}\mu$$

$$a = -\frac{1}{2}(n\log(2\pi) - \log|\Lambda| + \eta^T \Lambda \eta)$$

Mathematical Preparation

- In order to get the marginalization and conditioning of the partitioned multivariate Gaussian distribution, we need the theory of block diagonalization of a partitioned matrix
- In order to do maximum likelihood estimation, we need the knowledge of the traces of the covariance matrix

Partitioned Matrices

- Consider a general partitioned matrix

$$M = \begin{bmatrix} E & F \\ G & H \end{bmatrix}$$

To zero out the upper-right-hand and lower-left-hand corner of M, we can pre-multiply and post-multiply matrices in the following form

$$\begin{bmatrix} I & -FH \\ 0 & I \end{bmatrix} \begin{bmatrix} E & F \\ G & H \end{bmatrix} \begin{bmatrix} I & 0 \\ -H^{-1}G & I \end{bmatrix} = \begin{bmatrix} E-FH^{-1}G & 0 \\ 0 & H \end{bmatrix}$$

Partitioned Matrices (Continued)

- Define the Schur complement of Matrix M with respect to H , denote M/H as the term $E-FH^{-1}G$

Since

$$(XYZ)^{-1} = Z^{-1}Y^{-1}X^{-1} = W^{-1}$$

$$Y^{-1} = ZW^{-1}X$$

So

$$\begin{bmatrix} E & F \\ G & H \end{bmatrix}^{-1} = \begin{bmatrix} I & 0 \\ -H^{-1}G & I \end{bmatrix} \begin{bmatrix} (M/H)^{-1} & 0 \\ 0 & H^{-1} \end{bmatrix} \begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix}$$

$$= \begin{bmatrix} (M/H)^{-1} & -(M/H)^{-1}FH^{-1} \\ -H^{-1}G(M/H)^{-1} & H^{-1} + H^{-1}G(M/H)^{-1}FH^{-1} \end{bmatrix}$$

Partitioned Matrices

- Note that we could alternatively have decomposed the matrix *m* in terms of *E* and *M/E*, yielding the following for the inverse

$$\begin{bmatrix} I & 0 \\ -GE^{-1} & I \end{bmatrix}\begin{bmatrix} E & F \\ G & H \end{bmatrix}\begin{bmatrix} I & -E^{-1}F \\ 0 & I \end{bmatrix}=\begin{bmatrix} E & F \\ 0 & -GE^{-1}F-H \end{bmatrix}\begin{bmatrix} I & -E^{-1}F \\ 0 & I \end{bmatrix}=\begin{bmatrix} E & 0 \\ 0 & -GE^{-1}F-H \end{bmatrix}$$

$$\begin{bmatrix} E & F \\ G & H \end{bmatrix}^{-1}=\begin{bmatrix} I & -E^{-1}F \\ 0 & I \end{bmatrix}\begin{bmatrix} E^{-1} & 0 \\ 0 & (M/E)^{-1} \end{bmatrix}\begin{bmatrix} I & 0 \\ -GE^{-1} & I \end{bmatrix}$$

$$=\begin{bmatrix} E^{-1} & -E^{-1}F(M/E)^{-1} \\ 0 & (M/E)^{-1} \end{bmatrix}\begin{bmatrix} I & 0 \\ -GE^{-1} & I \end{bmatrix}$$

$$=\begin{bmatrix} E^{-1}+E^{-1}F(M/E)^{-1}GE^{-1} & -E^{-1}F(M/E)^{-1} \\ -(M/E)^{-1}GE^{-1} & (M/E)^{-1} \end{bmatrix}$$

Partitioned Matrices (Continued)

- Thus we get

$$(E-FH^{-1}G)^{-1}=E^{-1}+E^{-1}F(H-GE^{-1}F)^{-1}GE^{-1}$$

$$(E-FH^{-1}G)^{-1}FH^{-1}=E^{-1}F(H-GE^{-1}F)^{-1}$$

- At the same time we get the conclusion

$|M| = |M/H||H|$

Theory of Traces

- Define

$$tr[A] \Box \sum_i a_{ii} = \sum_i \lambda_i$$

It has the following properties:

tr*[ABC]* = tr*[CAB]* = tr*[BCA]*

$$x^T Ax = tr[x^T Ax] = tr[xx^T A]$$

Theory of Traces (continued)

$$\frac{\delta}{\delta a_{ij}} tr[AB] = \frac{\delta}{\delta a_{ij}} \sum_k \sum_l a_{kl} b_{lk} = b_{ji}$$

so

$$\frac{\delta}{\delta A} tr[BA] = B^T$$

$$\frac{\delta}{\delta A} x^T A x = \frac{\delta}{\delta A} tr[xx^T A] = [xx^T]^T = xx^T$$

Theory of Traces (continued)

We want to show that

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

Since

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

Recall

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

This is equivalent to prove

$$\frac{\delta}{\delta a_{ij}}|A| = \tilde{A}$$

Noting that

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

Joint Distributions, Marginalization & Conditioning

We partition the n by 1 vector *x* into *p* by 1 and *q* by 1, which *n* = *p* + *q*

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

$$p(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{(p+q)/2} \left|\Sigma\right|^{1/2}} \exp\{-\frac{1}{2}\begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^T \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}\}$$

Marginalization and Conditioning

$$\exp\{-\frac{1}{2}\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}^T \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}\}$$

$$= \exp\{\frac{1}{2}\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}^T \begin{bmatrix} I & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I \end{bmatrix} \begin{bmatrix} (\Sigma/\Sigma_{22})^{-1} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}$$

$$\begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}{}^{-1} \\ 0 & I \end{bmatrix} \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}\}$$

$$= \exp\left\{-\frac{1}{2}(x_1-\mu_1-\Sigma_{12}\Sigma_{22}{}^{-1}(x_2-\mu_2))^T(\Sigma/\Sigma_{22})^{-1}(x_1-\mu_1-\Sigma_{12}\Sigma_{22}{}^{-1}(x_2-\mu_2))\right\}$$

$$\square\exp\left\{\frac{1}{2}(x_2-\mu_2)^T\Sigma_{22}{}^{-1}(x_2-\mu_2)\right\}$$

Normalization Factor

$$\frac{1}{(2\pi)^{(p+q)/2}\left|\Sigma\right|^{1/2}} = \frac{1}{(2\pi)^{(p+q)/2}(|\Sigma/\Sigma_{22}||\Sigma_{22}|)^{1/2}}$$

$$= \left(\frac{1}{(2\pi)^{p/2}(|\Sigma/\Sigma_{22}|)^{1/2}}\right)\left(\frac{1}{(2\pi)^{q/2}(|\Sigma_{22}|)^{1/2}}\right)$$

Marginalization and Conditioning

Thus

$$p(x_2) = \left(\frac{1}{(2\pi)^{q/2} (|\Sigma_{22}|)^{1/2}} \right) \exp\left\{ \frac{1}{2} (x_2 - \mu_2)^T \Sigma_{22}{}^{-1} (x_2 - \mu_2) \right\}$$

$$p(x_1 | x_2) = \left(\frac{1}{(2\pi)^{p/2} (|\Sigma / \Sigma_{22}|)^{1/2}} \right) \square$$

$$\exp\left\{ -\frac{1}{2} (x_1 - \mu_1 - \Sigma_{12} \Sigma_{22}{}^{-1} (x_2 - \mu_2))^T (\Sigma / \Sigma_{22})^{-1} (x_1 - \mu_1 - \Sigma_{12} \Sigma_{22}{}^{-1} (x_2 - \mu_2)) \right\}$$

Marginalization and Conditioning (Cont)

Marginalization

$$\mu_2^{\ m} = \mu_2$$
$$\Sigma_2^{\ m} = \Sigma_{22}$$

Conditioning

$$\mu^{c}_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{\ -1}(x_2 - \mu_2)$$
$$\Sigma^{c}_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{\ -1}\Sigma_{21}$$

In Another Form

Marginalization

$$\eta_2^{\ m} = \eta_2 - \Lambda_{21}\Lambda_{11}^{\ -1}\eta_1$$

$$\Lambda_2^{\ m} = \Lambda_{22} - \Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12}$$

Conditioning

$$\eta^{c}_{\ 1|2} = \eta_1 - \Lambda_{12}x_2$$

$$\Lambda^{c}_{\ 1|2} = \Lambda_{11}$$

Maximum Likelihood Estimation

Likelihood function expression:

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

Taking derivative with respect to µ

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

Setting to zero

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

Estimating Σ

We need to take the derivative with respect to Σ

$$l(\Sigma \mid D) = -\frac{N}{2}\log|\Sigma| - \frac{1}{2}\sum_n (x-\mu)^T \Sigma^{-1}(x-\mu)$$

$$= \frac{N}{2}\log|\Sigma^{-1}| - \frac{1}{2}\sum_n tr[(x-\mu)^T \Sigma^{-1}(x-\mu)]$$

$$= \frac{N}{2}\log|\Sigma^{-1}| - \frac{1}{2}\sum_n tr[(x-\mu)(x-\mu)^T \Sigma^{-1}]$$

According to the property of traces

$$\frac{\delta l}{\delta \Sigma^{-1}} = \frac{N}{2}\Sigma - \frac{1}{2}\sum_n (x_n-\mu)(x_n-\mu)^T$$

Estimating Σ (Continued)

Thus the maximum likelihood estimator is

$$\hat{\Sigma}_{ML} = \frac{1}{N}\sum_{n}(x_n - \mu)(x_n - \mu)^T$$

The maximum likelihood estimator of canonical parameters are

$$\hat{\Lambda} = \hat{\Sigma}_{ML}^{-1}$$

$$\hat{\eta} = \hat{\Sigma}_{ML}^{-1}\,\hat{\mu}_{ML}$$