CS281A/Stat241A: Statistical Learning Theory

Multivariate Gaussian, Factor Analysis, and EM Algorithm (10/28/04)

Lecturer: Michael I. Jordan *Scribes: Albert C. To*

An important operation involving multivariate Gaussian distribution is the block matrix inversion. First, define the Schur complement of the matrix M with respect to H as

$$M/H = E - FH^{-1}G \tag{1}$$

where

$$M = \begin{pmatrix} E & F \\ G & H \end{pmatrix} \tag{2}$$

It can be verified that

$$\begin{pmatrix} I & -FH^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} E & F \\ G & H \end{pmatrix} \begin{pmatrix} I & 0 \\ -H^{-1}G & I \end{pmatrix} = \begin{pmatrix} E - FH^{-1}G & 0 \\ 0 & H \end{pmatrix} \tag{3}$$

Then,

$$M^{-1} = \begin{pmatrix} I & 0 \\ -H^{-1}G & I \end{pmatrix} \begin{pmatrix} (M/H)^{-1} & 0 \\ G & H^{-1} \end{pmatrix} \begin{pmatrix} I & -FH^{-1} \\ 0 & I \end{pmatrix} \tag{4}$$

$$= \begin{pmatrix} (M/H)^{-1} & -(M/H)^{-1}FH^{-1} \\ -HG(M/H)^{-1} & H^{-1} + H^{-1}G(M/H)^{-1}FH^{-1} \end{pmatrix} \tag{5}$$

From the above expression, it can be verified that

$$|M| = |M/H||H| \tag{6}$$

and also,

$$(M/E)^{-1} = (H - GE^{-1}F)^{-1} = H^{-1} + H^{-1}G(E - FH^{-1}G)^{-1}FH^{-1} \tag{7}$$

Matrix inversion lemma:

Let

$$X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \tag{8}$$

Then,

$$\exp\left\{-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right\} \tag{9}$$

$$= \exp\left\{-\frac{1}{2}\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}^T \begin{pmatrix} I & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I \end{pmatrix} \begin{pmatrix} (\Sigma/\Sigma_{22})^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix} \begin{pmatrix} I & -\Sigma_{21}\Sigma_{22}^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}\right\} \tag{10}$$

$$= \exp\left\{-\frac{1}{2}(x_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2))^T(\Sigma/\Sigma_{22})^{-1}(x_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2))\right\}$$
$$\cdot \exp\left\{-\frac{1}{2}(x_2-\mu_2-)^T\Sigma_{22}^{-1}(x_2-\mu_2)\right\} \tag{11}$$

The normalization can be split into two factors:

$$\frac{1}{(2\pi)^{(p+q)/2}|\Sigma|^{1/2}} = \frac{1}{(2\pi)^{p/2}|\Sigma/\Sigma_{22}|^{1/2}} \frac{1}{(2\pi)^{q/2}|\Sigma_{22}|^{1/2}} \tag{12}$$

Conditional distribution:

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \tag{13}$$
$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{14}$$

Marginalization:

$$\mu_2^m = \mu_2 \tag{15}$$
$$\Sigma_{22}^m = \Sigma_{22} \tag{16}$$

Canonical Parameters:

$$\Lambda = \Sigma^{-1} \tag{17}$$
$$\eta = \Sigma^{-1}\mu \tag{18}$$

Conditioning:

$$\eta_{1|2}^c = \eta_1 - \Lambda_{12}x_2 \tag{19}$$
$$\Lambda_{1|2}^c = \Lambda_{11} \tag{20}$$

Marginalization:

$$\eta_2^m = \eta_2 - \Lambda_{21}\Lambda_{11}^{-1}\eta_1 \tag{21}$$
$$\Lambda_2^m = \Lambda_{22} - \Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12} \tag{22}$$

Factor Analysis:

$$x_n \sim N(0, I) \tag{23}$$
$$y_n \sim N(\mu + \Lambda x_n, \Psi) \tag{24}$$

Write

$$y_n = \mu + \Lambda x_n + \epsilon_n; \quad \epsilon_n \sim N(0, \Psi) \tag{25}$$

Maximum Likelihood (ML) estimation:

Let

$$\mathcal{D} = \{y_n\}_{n=1}^N; \quad \theta = (\mu, \Lambda, \Psi) \tag{26}$$

$$\ell(\theta|\mathcal{D}) = \sum_{n=1}^N \log p(y_n|\theta) \tag{27}$$
$$= -\frac{N}{2}\log|\Lambda\Lambda^T + \Psi| - \frac{1}{2}\left\{\sum_{n=1}^N (y_n - \mu)^T(\Lambda\Lambda^T + \Psi)^{-1}(y_n - \mu)\right\} \tag{28}$$

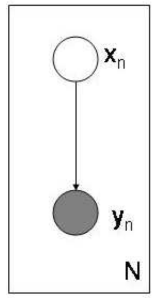

Figure 1: Graphical model for multinomial density estimation with Dirichlet prior with hyper-parameter α.

$$\begin{aligned} E(y_n) &= \mu + \Lambda E(x_n) + E(\epsilon_n) = \mu && (29)\\ \mathrm{Var(y_n)} &= E(\Lambda x_n + \epsilon_n)(\Lambda x_n + \epsilon_n)^T && (30)\\ &= E(\Lambda x_n x_n^T \Lambda^T + \epsilon_n x_n^T \Lambda^T + \Lambda x_n \epsilon^T + \epsilon_n \epsilon_n^T) && (31)\\ &= \Lambda\Lambda^T && (32) \end{aligned}$$

Important formulas:

$$\begin{aligned} E(Y) &= E(E(Y|X)) && (33)\\ \mathrm{Var(Y)} &= E(\mathrm{Var(Y|X)}) + \mathrm{Var(E(Y|X))} && (34) \end{aligned}$$

So,

$$\begin{aligned} E(y_n|x_n) &= \mu + \Lambda x_n && (35)\\ \mathrm{Cov(x_n, y_n)} &= E[x_n(\mu + \Lambda x_n + \epsilon_n - \mu)^T] = \Lambda^T && (36) \end{aligned}$$

Thus, the overall mean and covariance matrix is given by

$$\begin{aligned} \mu &= \begin{pmatrix} 0 \\ \mu \end{pmatrix} && (37)\\ \Sigma &= \begin{pmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Psi \end{pmatrix} && (38) \end{aligned}$$

EM algorithm:

Write complete log likelihood (using the trace trick):

$$\ell(\theta|\mathcal{D}) = -\frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_{n=1}^{N} x_n^T x_n - \frac{1}{2}\left\{\sum_{n=1}^{N}(y_n - \Lambda x_n)^T \Psi^{-1}(y_n - \Lambda x_n)\right\} \qquad (39)$$

$$= \quad -\frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_{n=1}^{N}\mathrm{tr}(\mathrm{x_n x_n^T}) - \frac{1}{2}\left\{\sum_{\mathrm{n=1}}^{\mathrm{N}}\mathrm{tr}[(\mathrm{y_n} - \Lambda \mathrm{x_n})(\mathrm{y_n} - \Lambda \mathrm{x_n})^{\mathrm{T}}\Psi^{-1}]\right\} \tag{40}$$

Use the following formulas to obtain the EM algorithm:

$$\frac{\partial}{\partial A}\log|A| = A^{-T} \tag{41}$$

$$\frac{\partial}{\partial A}\mathrm{tr}(\mathrm{BA}) = \mathrm{B^T} \tag{42}$$

$$\frac{\partial}{\partial A}\mathrm{tr}(\mathrm{BA^TCA}) = 2\mathrm{CAB} \tag{43}$$