Properties of Multivariate Gaussians

2/23/11
Jason L. Pacheco
CS195F

Conditional Gaussian Distributions:

If two RV's are jointly Gaussian then the distribution of one conditional on the other is also Gaussian.

Consider X_a and X_b jointly Gaussian

$$\underbrace{\begin{pmatrix} X_a \\ X_b \end{pmatrix}}_{X} \sim N\left(\underbrace{\begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}}_{\mu}, \underbrace{\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ab}^T & \Sigma_{bb} \end{pmatrix}}_{\Sigma}\right)$$

Sometimes it's useful to parameterize by $\Lambda = \Sigma^{-1}$ the <u>precision</u> <u>matrix</u>.

$$\Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ab}^T & \Lambda_{bb} \end{pmatrix}$$

NOTE: Λ is also symmetric, but $\Lambda_{aa} \neq \Sigma_{aa}^{-1}$.

Now we solve for the conditional,

$$p(X_a \mid X_b) = N(X_a \mid \mu_{a|b}, \Sigma_{a|b})$$

COMPLETING THE SQUARE

To do this we use a common technique. The quadratic form of a Gaussian expands as,

$$-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) = \cdots$$

$$= -\frac{1}{2}x^T \underset{\substack{\downarrow \\ \text{precision}}}{\Sigma^{-1}} x + x^T \underbrace{\Sigma^{-1}\mu}_{\text{precision} \times \text{mean}} - \underbrace{\frac{1}{2}\mu^T\Sigma^{-1}\mu}_{\text{const.}}$$

So if we can alter into this form we can "see" the precision directly, then solve for mean μ.

In our example we expand the quadratic form as,

$$-\frac{1}{2}(x-\mu)^T\Lambda(x-\mu) = \cdots$$

$$= -\frac{1}{2}(x_a-\mu_a)^T\Lambda_{aa}(x_a-\mu_a) - \frac{1}{2}(x_a-\mu_a)^T\Lambda_{ab}(x_b-\mu_b)$$
$$-\frac{1}{2}(x_b-\mu_b)^T\Lambda_{ba}(x_a-\mu_a) - \frac{1}{2}(x_b-\mu_b)^T\Lambda_{bb}(x_b-\mu_b)$$

$$= -\frac{1}{2}x_a^T\Lambda_{aa}x_a + x_a^T\Lambda_{aa}\mu_a - \frac{1}{2}x_a^T\Lambda_{ab}x_b + \frac{1}{2}x_a^T\Lambda_{ab}\mu_b$$
$$-\frac{1}{2}x_b^T\Lambda_{ba}x_a + \frac{1}{2}\mu_b^T\Lambda_{ba}x_a + \text{const}$$

Since $\Lambda_{ab} = \Lambda_{ba}^T$ we have,

$$= -\frac{1}{2}x_a^T\Lambda_{aa}x_a + x_a^T\Lambda_{aa}\mu_a - \cancel{\tfrac{1}{2}}\, x_a^T\Lambda_{ab}x_b + x_a^T\Lambda_{ab}\mu_b$$
$$+ \text{const}$$

Collecting terms we have,

$$-\frac{1}{2}(x-\mu)^T \Lambda (x-\mu) = -\frac{1}{2} x_a^T \Lambda_{aa} x_a + x_a^T \left(\Lambda_{aa} \mu_a - \Lambda_{ab}(x_b - \mu_b) \right)$$
$$+ \text{const}$$

So,

$$\Sigma_{a|b} = \Lambda_{aa}^{-1}$$

$$\Sigma_{a|b}^{-1} \mu_{a|b} = \left(\Lambda_{aa} \mu_a - \Lambda_{ab} (x_b - \mu_b) \right)$$

And multiplying both sides by $\Sigma_{a|b}$ the men is,

$$\mu_{a|b} = \Sigma_{a|b} \left\{ \Lambda_{ab} \mu_a - \Lambda_{ab} (x_b - \mu_b) \right\}$$

$$= \mu_a - \Lambda_{aa}^{-1} \Lambda_{ab} (x_b - \mu_b) \quad \Leftarrow \text{Linear in } x_b$$

So, if $p(x_a, x_b)$ is Gaussian $N(\mu, \Sigma)$ then the

Conditional $p(x_a | x_b)$ is Gaussian $N(\mu_{a|b}, \Sigma_{a|b})$

MARGINAL GAUSSIANS

What about the marginal?

$$p(x_a) = \int p(x_a, x_b)\, dx_b$$

This is also GAUSSIAN!

$$p(x_a) = N(x_a ; \mu_a, \Sigma_{aa})$$

BAYE'S THEOREM FOR GAUSSIAN RV's:

So far we know if x_a and x_b are jointly Gaussian, then the marginal $p(x_a)$ is Gaussian and Conditional $p(x_a|x_b)$ is Gaussian.

Furthermore, the conditional mean is a linear function of x_b. Generalizing assume we have

$$p(x) = N(x \mid \mu, \Lambda^{-1}) \qquad \text{(think prior)}$$

$$p(y|x) = N(y \mid Ax + b, L^{-1}) \qquad \text{(think likelihood)}$$

Multiplying and completing the square we look at the quadratic form,

$$-\tfrac{1}{2}(x-\mu)^T\Lambda(x-\mu) - \tfrac{1}{2}(y-Ax-b)^T L (y-Ax-b)$$

$$= -\tfrac{1}{2}x^T\Lambda x + x^T\Lambda\mu - \tfrac{1}{2}y^TLy + \tfrac{1}{2}y^TLAx$$

$$+\tfrac{1}{2}y^TLb + \tfrac{1}{2}x^TA^TLy - \tfrac{1}{2}x^TA^TLAx - \tfrac{1}{2}x^TA^TLb$$

$$+\tfrac{1}{2}b^TLy - \tfrac{1}{2}b^TLAx + \text{const}$$

$$= -\tfrac{1}{2}x^T(\Lambda + A^TLA)x - \tfrac{1}{2}y^TLy + \tfrac{1}{2}y^TLAx$$

$$+\tfrac{1}{2}x^TA^TLy + x^T\Lambda\mu - x^TA^TLb + y^TLb$$

Collecting terms we have,

$$-\frac{1}{2}\begin{pmatrix} x \\ y \end{pmatrix}^T \begin{pmatrix} \Lambda + A^T L A & -A^T L \\ -LA & L \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} x \\ y \end{pmatrix}^T \begin{pmatrix} \Lambda \mu - A^T L b \\ L b \end{pmatrix}$$

Which is the form of the full joint. We "see" the invers precision matrix

Using our previous results we can compute,

$$p(y) = N(y \mid A\mu + b, L^{-1} + A\Lambda^{-1}A^T)$$
$$p(x|y) = N(x \mid \Sigma\{A^T L(y-b) + \Lambda\mu\}, \Sigma)$$

where,

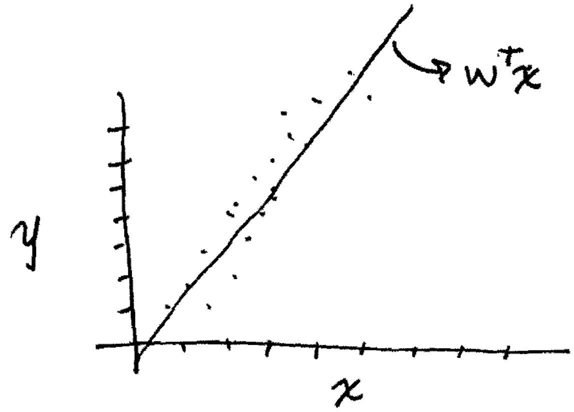
$$\Sigma = (\Lambda + A^T L A)^{-1}$$

These are general formulas which can be used!

Linear Regression

Assum we have data

$$D \in \{(y_1, x_1), \ldots, (y_N, x_N)\}$$



Where,

$$y_i \in \mathbb{R} \quad , \quad x_i \in \mathbb{R}^M$$

We want to fit a line w/ coefficients $w \in \mathbb{R}^M$

$$y_i = w^T x_i + \varepsilon \quad , \quad \varepsilon \sim N(0, \sigma^2)$$

This is equivalent to the likelihood,

$$p(y_i \mid x_i, w, \sigma^2) = N(y_i \mid w^T x_i, \sigma^2)$$

Maximum Likelihood $\hat{w}_{ML}$

The negative log-likelihood is,

$$\ell(w) = -\sum_{i=1}^{N} \log p(y_i \mid x_i, w, \sigma^2)$$

$$= \sum_{i=1}^{N} \left\{ \frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \sigma^2 + \frac{1}{2\sigma^2} (y_i - w^T x_i)^2 \right\}$$

$$= \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - w^T x_i)^2 + \text{Const}$$

$$\hat{w}_{ML} = \arg\min_w \ell(w) = \arg\min_w \underbrace{\frac{1}{N}}_{\text{Doesn't affect optimum}} \sum_{i=1}^{N} (y_i - w^T x_i)^2 \quad (*)$$

~~So Maximum Likelihood~~

LEAST SQUARES:

So ML is equivalent to MMSE for linear regression. Furthermore, objective is <u>convex</u>.

Let,

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1M} \\ x_{21} & x_{22} & \cdots & x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NM} \end{pmatrix} \in \mathbb{R}^{N \times M}$$

We rewrite (*) as,

$$\hat{w}_{MMSE} = \underset{w}{\operatorname{argmin}} \frac{1}{N} (\vec{y} - \mathbf{X}\vec{w})^T (\vec{y} - \mathbf{X}\vec{w})$$

$$= \underset{w}{\operatorname{argmin}} \frac{1}{N} y^T y - \frac{2}{N} w^T (\mathbf{X}^T y) + \frac{1}{N} w^T (\mathbf{X}^T \mathbf{X}) w$$

$$= \underset{w}{\operatorname{argmin}} \frac{1}{2N} w^T (\mathbf{X}^T \mathbf{X}) w - \frac{1}{N} w^T (\mathbf{X}^T y) \qquad (**)$$

<u>ASIDE</u>

Derivative of scalar product: $\nabla_x a^T x = a$

Derivative of quadratic form: $\nabla_x x^T A x = (A + A^T) x$

<u>RETURN</u>

The objective is convex which means unique optimum.

We take gradient of (**) w.r.t. $\vec{w}$ and solve,

$$\nabla_w \hat{W}_{MMSE} = 0 = \frac{1}{2N}\left(\mathbf{X}^T\mathbf{X} + \mathbf{X}^T\mathbf{X}\right)w - \frac{1}{N}\mathbf{X}^T y$$

$$0 = \frac{1}{N}\left(\mathbf{X}^T\mathbf{X}w - \mathbf{X}^T y\right)$$

$$\mathbf{X}^T y = \mathbf{X}^T\mathbf{X}w$$

$$\hat{W}_{MMSE} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T y ,$$

(NORMAL EQUATIONS)