Multivariate Gaussian

The univariate Gaussian distribution can be written as

$$N(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

where μ is the mean and σ^2 is the variance

For a D-dim vector $\vec{x}$ the mvn-Gaussian distribution ~~takes the form~~ the following 3 defs are equiv.

Definition 1)

$$f_X(\vec{x}) = N(\vec{x} \mid \vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1}(x-\mu)\right\}$$

which is equivalent to both

Definition 2) the moment generating function of X is

$$M_X(\vec{t}) \equiv \mathbb{E}\left[e^{\vec{t}' X}\right] = \exp\left\{\vec{\mu}'\vec{t} + \frac{1}{2}\vec{t}'\Sigma\vec{t}\right\}$$

Definition 3) X has the same distribution as

$$AZ + \mu$$

where $Z = [z_1, \ldots, z_k]$ is a sample from $N(0, I)$ and $A_{n \times k}$ satisfied $AA' = \Sigma$

Here $\vec{\mu}$ is a vector, Σ is a PSD matrix and for $\vec{x}$ all it can be written that

$$\vec{x} \sim N_n(\vec{x} \mid \vec{\mu}, \Sigma)$$

Thm: Def's 1, 2, and 3 are equiv. for Σ pos. def. Def's 2 & 3 are equivalent for Σ pos. semi. def.

Proof that Def 3 $\Rightarrow$ Def 2

For $z_i \sim N(0,1)$

$$M_{z_i}(t_i) = \mathbb{E}\left[e^{t_i z_i}\right] = \int_{-\infty}^{\infty} e^{z_i t_i} \frac{e^{-z_i^2/2}}{\sqrt{2\pi}} \, dz_i$$

$$= \int_{-\infty}^{\infty} e^{z_i t_i} \frac{e^{-(z_i - t_i)^2/2}}{\sqrt{2\pi}} e^{-z_i t_i} e^{t_i^2/2} \, dz_i$$

$$= e^{t_i^2/2} \int_{-\infty}^{\infty} \frac{e^{-(z_i - t_i)^2/2}}{\sqrt{2\pi}} \, dz_i$$

$$= e^{t_i^2/2} \frac{\sqrt{2\pi}}{\sqrt{2\pi}} = e^{t_i^2/2}$$

If $\vec{z} = [z_1, \ldots, z_k]$ is a random sample from $N(0,1)$ then

$$M_{\vec{z}}(\vec{t}) = \mathbb{E}\left[e^{\vec{z}'\vec{t}}\right] = \mathbb{E}\left[e^{\sum_i z_i t_i}\right]$$

$$= \mathbb{E}\left[\prod_{i=1}^{k} e^{z_i t_i}\right] \quad \text{because } z_i \sim iid\ N(0,1)$$

$$= \prod_{i=1}^{k} \mathbb{E}\left[e^{z_i t_i}\right]$$

$$= \prod_{i=1}^{k} M_{z_i}(t_i)$$

$$= \exp\left\{\sum_{i=1}^{k} t_i^2/2\right\}$$

$$= \exp\left\{\vec{t}'\vec{t}/2\right\}$$

Which gives us the M.G.F. for $\vec{z} \sim N(0, I)$

If $\vec{X} = A\vec{Z} + \vec{\mu}$ then

$$M_{\vec{X}}(\vec{t}) = E[\exp\{\vec{X}'\vec{t}\}]$$

$$= E[\exp\{(A\vec{Z}+\vec{\mu})'\vec{t}]$$

$$= \exp(\vec{\mu}'\vec{t}) \cdot E[\exp\{(A\vec{Z})'\vec{t}\}]$$

$$= \exp(\vec{\mu}'\vec{t}) \cdot E[\exp\{\cancel{(A\vec{Z})'}\vec{Z}'A'\vec{t}\}]$$

$$= \exp(\vec{\mu}'\vec{t}) \cdot M_{\vec{Z}}(A'\vec{t})$$

$$= \exp(\vec{\mu}\vec{t}) \cdot \exp\{\tfrac{1}{2}(A'\vec{t})'(A'\vec{t})\}$$

$$= \exp(\vec{\mu}\vec{t}) \cdot \exp\{\tfrac{1}{2} t'AA't\}$$

$$= \exp(\vec{\mu}\vec{t} + \tfrac{1}{2}\vec{t}'\Sigma\vec{t})$$

- Note, if A is invertable then $A^{-1}(\vec{X}-\mu) \sim N(\vec{0}, \mathbb{I})$ which can be useful for hypothesis testing, data pre-conditioning, etc.

- If two ~~RVs~~ distributions have the same MGF then they are identical at all points.

Proof that Def 2 $\Rightarrow$ Def 3

Since $\Sigma \geq 0$ (and $\Sigma = \Sigma'$) there is an orthogonal matrix $C^{n \times n}$ such that
$C' \Sigma C = \Lambda$ where Λ is diagonal with non-negative elements. So

$$\begin{aligned} \Sigma &= C \Lambda C' \\ &= C \Lambda^{1/2} \Lambda^{1/2} C' \\ &= (C \Lambda^{1/2})(C \Lambda^{1/2})' \\ &= A A' \end{aligned}$$

But we known that the moment generating function of $A\vec{Z} + \vec{\mu}$ is (by the previous proof)

$$\exp\{\vec{\mu}'\vec{t} + \tfrac{1}{2}\vec{t}'\Sigma\vec{t}\}$$

i.e. the same as $\vec{X}$. Because the m.g.f. uniquely determines the distribution, $\vec{X}$ has the same distribution as $A\vec{Z} + \vec{\mu}$

∎

Proof of Def 3 $\Rightarrow$ Def 1: (for pos. def. Σ)

Because Σ is pos. def. there is a non-singular $A_{n \times n}$ s.t. $AA' = \Sigma$. Let $\vec{X} = A\vec{z} + \mu$ where $\vec{z} \sim N(0, \mathbb{1})$. The density of $\vec{z}$ is

$$f_{\vec{z}}(\vec{z}) = \prod_{i=1}^{D} (2\pi)^{-1/2} \exp\left\{-\frac{1}{2} z_i^2\right\} = (2\pi)^{-D/2} \exp\left\{-\frac{1}{2}\vec{z}'\vec{z}\right\}$$

By the change of variable rule (transformation of distribution) the density function of X is

$$f_{\vec{X}}(x) = f_{\vec{z}}(\vec{z}(\vec{x}))\,|J|$$ determinant of the jacobian

$$\vec{z}(\vec{x}) = A^{-1}(\vec{y} - \mu) \quad , \quad \frac{\partial z(\vec{x})}{\partial \vec{y}} = \frac{\partial A^{-1}\vec{y}}{\partial \vec{y}} = A^{-1}$$

$$\Rightarrow J = A^{-1}$$

$$\Rightarrow f_{\vec{X}}(\vec{x}) = f_{\vec{z}}(z(\vec{x}))\,|A^{-1}| = f_{\vec{z}}(z(\vec{x}))\,|A|^{-1}$$

Therefore

$$f_{\vec{X}}(\vec{x}) = f_{\vec{z}}(A^{-1}(\vec{x} - \mu))\,|A|^{-1}$$

$$= (2\pi)^{-D/2} |A|^{-1} \exp\left\{-\frac{1}{2}[A^{-1}(\vec{x} - \vec{\mu})]'[A^{-1}(x - \mu)]\right\}$$

$$= (2\pi)^{-D/2} |AA'|^{-1/2} \exp\left\{-\frac{1}{2}(\vec{x} - \vec{\mu})'(A^{-1})'A^{-1}(x - \mu)\right\}$$

$$= (2\pi)^{-D/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\vec{x} - \vec{\mu})'(AA')^{-1}(\vec{x} - \vec{\mu})\right\}$$

$$= (2\pi)^{-D/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\vec{x} - \vec{\mu})'\Sigma^{-1}(\vec{x} - \vec{\mu})\right\}$$

Note $|A|^{-1}$
$= |A|^{-1/2}|A|^{-1/2}$
$= |A|^{-1/2}|A'|^{-1/2}$
$= (|A||A'|)^{-1/2}$
$= |AA'|^{-1/2}$

Pf that Def 1 $\Rightarrow$ Def 2 left to reader.

Properties of MVN distributions

1) If $X \sim N(\vec{x} | \vec{\mu}, \Sigma)$

$$\mathbb{E}[\vec{x}] = \vec{\mu}, \quad \mathrm{cov}[\vec{x}] = \Sigma$$

Pf. We know $\vec{x}$ has the same distribution as $A\vec{z} + \vec{\mu}$ so

$$\mathbb{E}[\vec{x}] = \mathbb{E}[A\vec{z} + \vec{\mu}] = \mathbb{E}[A\vec{z}] + \vec{\mu} = \vec{\mu}$$

$$\begin{aligned}
\mathrm{Cov}[\vec{x}] &= \mathbb{E}\left[(\vec{x} - \mathbb{E}[\vec{x}])(\vec{x} - \mathbb{E}[\vec{x}])'\right] \\
&= \mathbb{E}\left[(\vec{x} - \vec{\mu})(\vec{x} - \vec{\mu})'\right] \\
&= \mathbb{E}\left[(A\vec{z})(A\vec{z})'\right] \\
&= \mathbb{E}\left[A\vec{z}\vec{z}'A'\right] \\
&= A\,\mathrm{Cov}(\vec{z})\,A' \\
&= AIA' = AA' = \Sigma
\end{aligned}$$

2) If ~~$\vec{z} \sim N(0, I_{D\times D})$ $\vec{z} = \{z_1, \dots, z$~~
$\vec{z} = (z_1, \dots, z_D)$, $z_i \sim_{iid} N(0, \sigma^2)$ then

$\vec{z} \sim N_D(\vec{0}_D, I_{D\times D}\sigma^2)$ (matrix normal)

$$\begin{aligned}
p(\vec{z}) &= \prod_{i=1}^{D} (2\pi)^{-\frac{1}{2}} \sigma^{-1} \exp\left\{-\frac{1}{2\sigma^2}(z)^2\right\} \\
&= (2\pi)^{-D/2} \sigma^{-D} \exp\left\{-\frac{1}{2}\sum_{i=1}^{D} z_i^2/\sigma^2\right\} \\
&= (2\pi)^{-D/2} |I\sigma^2|^{-1/2} \exp\left\{-\frac{1}{2}\left(\vec{z}'(I\sigma^{-2})\vec{z}\right)\right\}
\end{aligned}$$

Linear transformations of MVN Vectors

1. If $\vec{y} \sim N_n(\vec{\mu}, \Sigma)$ and $C_{p\times n}$ is a constant matrix of rank p then $C\vec{y} \sim N_p(C\vec{\mu}, C\Sigma C')$

Proof: By def 3, $\vec{y} = A\vec{z} + \vec{\mu}$ where $AA' = \Sigma$. So

$$C\vec{y} = C(A\vec{z} + \vec{\mu})$$
$$= CA\vec{z} + C\vec{\mu}$$

but by def. 3

$$CA\vec{z} + C\vec{\mu} \sim N(C\vec{\mu}, CA(CA)')$$

$$= N(C\vec{\mu}, CAA'C')$$

$$= N(C\vec{\mu}, C\Sigma C')$$

2. $\vec{y}$ is MVN iff $\vec{a}'\vec{Y}$ is normally distributed for all non-zero constant vectors $\vec{a}$

Pf. If $\vec{y} \sim N_n(\vec{\mu}, \Sigma)$ then by the previous claim

$$\vec{a}'\vec{Y} \sim N(\vec{a}'\vec{\mu}, \vec{a}'\Sigma\vec{a})$$

Conversely, assume $x = \vec{a}'\vec{Y}$ is univariate normal for all nonzero $\vec{a}$, Then since $E[x] = \vec{a}'\vec{\mu}$ and $\mathrm{cov}(x) = \sigma^2\{x\} = \mathrm{var}\{x\} = \mathrm{Cov}(a'\vec{Y}) = \vec{a}'\Sigma\vec{a}$.

The univariate normal M.G.F. is

$$E[\exp(xt)] = M_x(t) = \exp\{(a'\mu)t + \tfrac{1}{2}(a'\Sigma a)t^2\}$$

for all t. Setting $t = 1$ we have

$$E[\exp(\vec{a}'\vec{Y})] = \exp\{(\vec{a}'\vec{\mu}) + \tfrac{1}{2}(\vec{a}'\Sigma\vec{a})\} = M(a)$$

which is the M.G.F. of $N_n(\vec{\mu}, \Sigma)$ and therefore $\vec{y} \sim N_n(\vec{\mu}, \Sigma)$.

Conditional Gaussian Distributions

Important features: if two sets of var's are jointly Gaussian then 1) the cond. dist. of one set given the other is Gaussian and 2) the marginal dist of each set is also Gaussian.

Suppose $x \in \mathbb{R}^D$ and $x \sim \mathcal{N}(x \mid \mu, \Sigma)$. let x_a be (w.l.o.g.) the first M components of x and x_b the remaining $D-M$ components.

ie.

$$x = \begin{pmatrix} x_a \\ x_b \end{pmatrix}$$

let

$$\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}$$

and let

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$$

$\leftarrow$ note Σ_{aa}, Σ_{bb} symmetric and $\Sigma_{ab} = \Sigma_{ba}^T$

be a covariance matrix

The precision matrix $\Lambda \equiv \Sigma^{-1}$ is sometimes easier to work with.

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

since Λ is also symmetric the same stuff holds

To start lets find $p(x_a \mid x_b)$

By the product rule we know that we can get $p(x_a \mid x_b)$ simply by fixing x_b in the joint $p(x_a, x_b)$ and renormalizing by inspection, ie.

$$p(x_a \mid x_b) \propto p(x_a, x_b)$$

$$= \frac{p(x_a, x_b)}{\int p(x_a, x_b)\, d x_a} = \frac{p(x_a, x_b)}{p(x_a)}$$

But we will proceed by inspection

$$\ln p(x_a | x_b, \mu, \Sigma) \propto -\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) + \text{const}$$

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

$$-\frac{1}{2}(x_b-\mu_b)^T \Lambda_{ba}(x_a-\mu_a) - \frac{1}{2}(x_b-\mu_b)^T \Lambda_{bb}(x_b-\mu_b) + \text{const}$$

but by inspection we can see that, as a function of ~~μ_a and~~ x_a this is again a quadratic form

Reminder: Completing the square

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

If we collect the terms quadratic in x_a we can immediately identify the ~~mean~~ cov Σ of $x_a | x_b$ ie,

$$-\frac{1}{2}x^T \Sigma^{-1} x = -\frac{1}{2} x_a^T \Lambda_{aa} x_a$$

$$\Rightarrow \Sigma^{-1} = \Lambda_{aa}$$

to identify the mean we have to collect the linear terms in x_a

$$x_a^T \{\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b-\mu_b)\}$$

$$\Rightarrow x_a^T \Lambda_{aa}\mu = x_a^T \{\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b-\mu_b)\}$$

$$\Rightarrow \mu = \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b-\mu_b)$$

yielding

$$x_a | x_b \sim N(x_a | \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b-\mu_b), \Lambda_{aa}^{-1})$$

If we wish to have everything in terms of the covariance matrix instead of the precision matrix, we need the following matrix inversion identity ~~(cat.~~

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1}+D^{-1}CMBD^{-1} \end{pmatrix}$$

where $M = (A - BD^{-1}C)^{-1}$ (called the Schur complement)

Since $\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$

and using the inversion identity we have

$$\Lambda_{aa} = (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}$$
$$\Lambda_{ab} = -(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}\Sigma_{ab}\Sigma_{bb}^{-1}$$

which gives us (after some algebra)

$$\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b)$$
$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$$

Note that $\mu_{a|b}$ is a linear function of x_b and the covariance is independent of x_a. This is an example of a linear gaussian model.

Marginal Gaussian Distributions

We know now that if $p(x_a, x_b)$ is Gaussian then $p(x_a | X_b)$ is also Gaussian, what about $p(x_a)$?

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

Using the same tricks (completing squares, recognizing quadratic forms, etc.) it can be arrived at that

$$\mathbb{E}[x_a] = \mu_a$$
$$\mathrm{cov}[x_a] = \Sigma_{aa}$$

i.e. the marginal distribution of x_a arises from simply ignoring the parts of the distribution corresponding to x_b.

In Summary

Given a joint Gaussian dist $\mathcal{N}(\vec{x}|\vec{\mu}, \Sigma)$ with $\Lambda \equiv \Sigma^{-1}$ and

$$\vec{x} = \begin{pmatrix} \vec{x}_a \\ \vec{x}_b \end{pmatrix}, \quad \vec{\mu} = \begin{pmatrix} \vec{\mu}_a \\ \vec{\mu}_b \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \qquad \Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$$

Condition

$$p(\vec{x}_a | \vec{x}_b) = \mathcal{N}(\vec{x}_a | \vec{\mu}_{a|b}, \Lambda_{aa}^{-1})$$
$$\vec{\mu}_{a|b} = \vec{\mu}_a - \Lambda_{aa}^{-1}\Lambda_{ab}(\vec{x}_b - \vec{\mu}_b)$$

Marginal

$$p(\vec{x}_a) = \mathcal{N}(\vec{x}_a | \vec{\mu}_a, \Sigma_{aa})$$

Bayes Thm for Gaussian Variables

Take

$$p(\vec{x}) = \mathcal{N}(\vec{x} \mid \vec{\mu}, \Lambda^{-1})$$

and

$$p(\vec{y} \mid \vec{x}) = \mathcal{N}(\vec{y} \mid A\vec{x} + \vec{b}, L^{-1})$$

Λ, L - precision matrices

μ, A, b - params governing means

We want $p(\vec{y})$ and $p(\vec{x} \mid \vec{y})$

To do this we define a R.V.

$$\vec{z} = \begin{pmatrix} \vec{x} \\ \vec{y} \end{pmatrix}$$

stacking vectors

and construct the joint distribution of this R.V. by considering the log of the joint

$$\ln p(\vec{z}) = \ln p(\vec{x}) + \ln p(\vec{y} \mid \vec{x})$$

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

$$-\frac{1}{2}(y-Ax-b)^T L (y-Ax-b) + \text{const}$$

terms indep of x & y

In this we complete the square by first collecting quadratic terms in x & y

$$-\frac{1}{2} x^T (\Lambda + A^T L A) x - \frac{1}{2} y^T L y + \frac{1}{2} x^T A^T L y + \frac{1}{2} y^T L A x$$

and note that this equals

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

where

$$R = \begin{pmatrix} \Lambda + A^T L A & -A^T L \\ -LA & L \end{pmatrix}$$

which remembering $-\frac{1}{2}(z-\mu_z)^T \mathrm{cov}[z]^{-1}(z-\mu_z)$

$$= -\frac{1}{2} z^T \mathrm{cov}[z]^{-1} z + z^T \mathrm{cov}[z]^{-1} \mu_z$$

$\Rightarrow \quad \mathrm{cov}[z]^{-1} = R \quad , \quad \mathrm{cov}[z] = R^{-1}$

$\Rightarrow \quad \mathrm{cov}[z] = \begin{pmatrix} \Lambda^{-1} & \Lambda^{-1}A^T \\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^T \end{pmatrix}$ verify? HW

from partitioned inverse eqn.

Now that we have the covariance matrix for z we can find the mean of the distribution (complete square). For this we need the linear terms from the lnjoint

$$x^T \Lambda \mu - y^T L b \mp x^T A^T L b = \begin{pmatrix} x \\ y \end{pmatrix}^T \begin{pmatrix} \Lambda\mu \mp A^T L b \\ Lb \end{pmatrix}$$

$$\Rightarrow \mathbb{E}[z] = \mu_z = R^{-1}\begin{pmatrix} \Lambda\mu \mp A^T L b \\ Lb \end{pmatrix}$$

which after some algebra yields

$$\vec{z} \sim \mathcal{N}\left(\begin{pmatrix} \mu \\ A\mu + b \end{pmatrix}, \begin{pmatrix} \Lambda^{-1} & \Lambda^{-1}A^T \\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^T \end{pmatrix}\right)$$

but, since we have analytic forms for the conditionals and marginals of Gaussian distributions we essentially have our answers.

By inspection

$$\mathbb{E}[Y] = A\mu + b$$
$$\text{cov}[Y] = L^{-1} + A\Lambda^{-1}A^T$$

← And now to get $p(\vec{x}\,|\,\vec{Y})$ we use conditional ~~And~~, remembering

$$p(x_a|x_b) = \mathcal{N}(x_a|\mu_{a|b}, \Lambda_{aa}^{-1})$$
$$\mu_{a|b} = \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b - \mu_b)$$

$$\cancel{\Rightarrow P(\vec{Y}|\vec{x}) = \mathcal{N}(\vec{Y}}$$

$$\Rightarrow P(\vec{x}|\vec{Y}) = \mathcal{N}(\vec{x}\,|\,\mu_{x|Y}, \Lambda_{xx}^{-1})$$
$$\mu_{x|Y} = \mu_x - \Lambda_{xx}^{-1}\Lambda_{xY}(Y - \mu_Y)$$

where
$$\Lambda_{xx}^{-1} = (\Lambda + A^TLA)^{-1}$$
$$\Lambda_{xY} = -A^TL$$
$$\mu_Y = A\mu + b$$
$$\mu_x = \mu$$
$$\mu_{x|Y} = \mu + (\Lambda + A^TLA)^{-1}A^TL(Y - A\mu - b)$$

$$\Rightarrow P(\vec{x}|\vec{Y}) = \mathcal{N}(\vec{x}\,|\,\mathbb{E}[x|Y], \text{cov}[x|Y])$$

where
$$\mathbb{E}[x|Y] = (\Lambda + A^TLA)^{-1}(A^TL(Y-b) + \Lambda\mu)$$
$$\text{cov}[x|Y] = (\Lambda + A^TLA)^{-1}$$

Summarizing, if
$$p(x) = \mathcal{N}(x|\mu, \Lambda^{-1}) \quad \text{and}$$
$$p(y|x) = \mathcal{N}(y|Ax + b, L^{-1})$$
then
$$p(y) = \mathcal{N}(y|A\mu + b, L^{-1} + A\Lambda^{-1}A^T)$$
$$p(x|y) = \mathcal{N}(x|\Sigma\{A^TL(y-b) + \Lambda\mu\}, \Sigma)$$

where
$$\Sigma = (\Lambda + A^TLA)^{-1}$$

Maximum likelihood for the Gaussian

Given a dataset $X = (x_1, \dots, x_n)^T$ where $x_i \overset{iid}{\sim} N(\mu, \Sigma)$ by assumption we can estimate μ and Σ from the data using maximum likelihood

$$\ln p(X|\mu, \Sigma) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^T\Sigma^{-1}(x_n-\mu)$$

to find an ML estimator for μ we take the partial w.r.t. μ

$$\frac{\partial}{\partial\mu}\ln p(X|\mu,\Sigma) = -\frac{1}{2}\sum_{n=1}^{N}\frac{\partial}{\partial\mu}\left[(x_n-\mu)^T\Sigma^{-1}(x_n-\mu)\right]$$

$$= -\frac{1}{2}\sum_{n=1}^{N} -2\,\Sigma^{-1}(x_n-\mu)$$

now we set this equal to zero and solve

$$\sum_{n=1}^{N}(x_n-\mu) = 0 \implies \boxed{\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N}x_n}$$

To solve for the ML estimator for Σ we need the following ~~two~~ facts

$$\boxed{\frac{\partial \ln|X|}{\partial X} = (X^{-1})^T = (X^T)^{-1}}$$

and

$$\boxed{\frac{\partial(a^T X b)}{\partial X} = b a^T}$$

and

$$\boxed{\frac{\partial\, a^T X^{-1} b}{\partial X} = -X^{-T} a b^T X^{-T}}$$

$$\frac{\partial}{\partial \Sigma} \ln p(X|\mu, \Sigma) = -\frac{N}{2} \frac{\partial}{\partial \Sigma} \ln|\Sigma|$$

$$-\frac{1}{2} \sum_{n=1}^{N} \frac{\partial}{\partial \Sigma} (x_n - \mu)^T \Sigma^{-1} (x_n - \mu)$$

$$= -\frac{N}{2} (\Sigma^{-1})^T + \frac{1}{2} \sum_{n=1}^{N} \Sigma (x_n - \mu)(x_n - \mu)^T \Sigma$$

Now we set this expression equal to zero and solve

$$\frac{N}{\cancel{2}} \Sigma^{-1} = \frac{1}{\cancel{2}} \sum_{n=1}^{N} \Sigma^{-1} (x_n - \mu)(x_n - \mu)^T \Sigma^{-1}$$

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

But we don't know μ. We <u>do</u> know that μ_{ML} optimizes the same.