Maximum Likelihood and the Multivariate Gaussian

Let $\vec{\mu}, \vec{x} \in \mathbb{R}^D$, $\Sigma \in PSD^{D \times D}$ then let

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

be the "multivariate normal density function" of RV $\vec{x}$ given vector mean $\vec{\mu}$ and covariance matrix Σ.

An important thing to know about MVN distributed vectors is that

$\vec{x}$ has the same distribution as

$A\vec{z} + \vec{\mu}$ where

$\vec{z} = [z_1, \ldots z_D]$, $z_d \sim N(0, I)$

and A satisfies $AA^T = \Sigma$ (Cholesky)

Certainly we can note that

$$\mathbb{E}[A\vec{z} + \vec{\mu}] = A\,\mathbb{E}[\vec{z}] + \vec{\mu} = A \cdot \vec{0} + \vec{\mu} = \vec{\mu}$$

and

$$Cov[A\vec{z} + \vec{\mu}] = Cov[A\vec{z}] = A\, Cov[\vec{z}]\, A^T = AA^T = \Sigma$$

But higher-order moments could appear.

We can "derive" the MVN PDF starting from —

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

By the change of variable rule, let

$$\vec{x} = A\vec{z} + \vec{\mu} \quad \Rightarrow A^{-1}(\vec{x} - \vec{\mu}) = \vec{z}$$

where $A A^T = \Sigma$

The multivariate change of variable rule says,

$$p_x(\vec{x}) = p_z(\vec{z}) \left| \begin{matrix} \partial z_1/\partial x_1 & \partial z_1/\partial x_2 & \cdots & \partial z_1/\partial x_D \\ \partial z_2/\partial x_1 & & & \\ \vdots & & \ddots & \\ \partial z_D/\partial x_1 & & \cdots & \partial z_D/\partial x_D \end{matrix} \right|$$

$$g(\vec{x}) = f(\vec{z}) \left| \det\left(\frac{\partial \vec{z}}{\partial \vec{x}}\right) \right|$$

$\vec{x} = r(\vec{z})$

$r: z \to x$

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

$$A^{-1} = \frac{\partial \vec{z}}{\partial \vec{x}}$$

and

$$\begin{bmatrix} a^{-1}_{11} & a^{-1}_{12} \\ a^{-1}_{D1} & a^{-1}_{D2} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_D \end{bmatrix} = \begin{bmatrix} z_1 \\ \vdots \\ z_D \end{bmatrix}$$

we have

$$p_x(\vec{x}) = p_z(\vec{z}) \left| A^{-1} \right| \qquad \text{note } |A^{-1}| = |A|^{-1}$$

$$= p_z\left(A^{-1}(\vec{x} - \mu)\right) |A|^{-1}$$

$$p_{\vec{x}}(\vec{x}) = p_{\vec{z}}\left(A^{-1}(\vec{x}-\vec{\mu})\right)\left|A\right|^{-1}$$

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

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

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

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

Where we have used

$$\left|A\right|^{-1} = \left|A\right|^{-\frac{1}{2}}\left|A\right|^{-\frac{1}{2}} = \left|A\right|^{-\frac{1}{2}}\left|A^{T}\right|^{-\frac{1}{2}} = \left(\left|A\right|\left|A^{T}\right|\right)^{-\frac{1}{2}}$$
$$= \left|AA^{T}\right|^{-\frac{1}{2}}$$

and $$(A^{-1})^{T}A^{-1} = (A^{T})^{-1}A^{-1} = (AA^{T})^{-1}$$

with individual steps from
Sam Roweis' Matrix cheat sheets
or the "matrix cookbook"
notably $|A| = |A^{T}|$, $(A^{-1})^{T} = (A^{T})^{-1}$
$(AB)^{-1} = B^{-1}A^{-1}$ and $|AB| = |A||B|$

So we can represent a multivariate Gaussian as an affine transformation of a vector of individual $N(0,1)$ distributed R.V.'s

But, given a collection of N vectors $\vec{x}_1, \ldots \vec{x}_N$ which we assume to be iid $N(\mu, \Sigma)$ how do we get μ & Σ?

M.L. For MV Gaussians

Let $X = \{\vec{x}_1 \dots \vec{x}_N\}$

$$\mathcal{L}(X; \{\mu, \Sigma\}) = \prod_{n=1}^{N} 2\pi^{-D/2} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(x_n - \mu^T)\Sigma^{-1}(x_n - \mu)\right\}$$

like before we will estimate μ_{ML} & Σ_{ML} by taking derivatives of log likelihoods and setting the — equal to zero.

We will need the following, also from the Matrix cookbook.

with W symmetric

$$\frac{\partial}{\partial s}(x-s)^T W (x-s) = -2W(x-s)$$

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

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

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

First ML Est $\hat{\mu}$ for μ:

$$\frac{\partial}{\partial \vec{\mu}} \ln \mathcal{L}(X; \{\mu, \Sigma\})$$

$$= \sum_{n=1}^{N} \frac{\partial}{\partial \vec{\mu}} \left(-\frac{1}{2}(\vec{x}_n - \vec{\mu})^T \Sigma^{-1} (\vec{x}_n - \vec{\mu})\right\} = 0$$

$$= \sum_{n=1}^{N} \Sigma^{-1} (\vec{x}_n - \vec{\mu}) = 0$$

$$\Rightarrow \sum_{n=1}^{N} \vec{x}_n - N\vec{\mu} = 0$$

$$\Rightarrow \hat{\vec{\mu}} = \frac{1}{N} \sum_{n=1}^{N} \vec{x}_n$$

Note

$$\mathbb{E}[\hat{\vec{\mu}}] = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[\vec{x}_n] = \frac{N}{N}\vec{\mu} \Rightarrow \text{Bias}(\hat{\vec{\mu}}) = 0$$

Now $\hat{\Sigma}_{ML} = \arg\max_{\Sigma} \log \mathcal{L}(X; \hat{\mu}, \Sigma)$

$$\frac{\partial}{\partial \Sigma} \log \mathcal{L}(X; \hat{\mu}, \Sigma)$$

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

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

Setting this expression equal to zero and solving yields

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

Σ symmetric so $(\Sigma^{-1})^T = (\Sigma^T)^{-1} = (\Sigma)^{-1}$

note $\hat{\mu}$'s

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

So now, given data, we can analytically compute the ML parameters of a multivariate Gaussian.

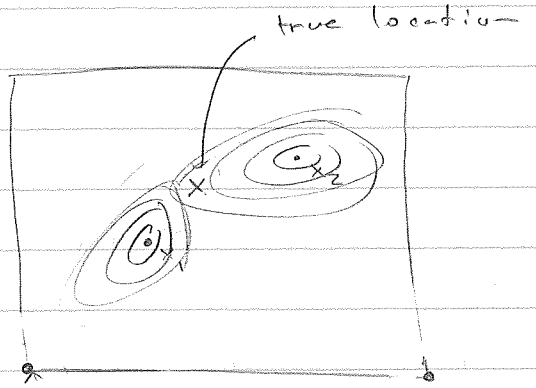
The maximum likelihood principle has other "uses".

A Note: the ML estimator, particularly for exponential families, has nice properties, particularly in the limit. It can be shown, for instance, that the sampling distribution of most ML estimator tends asymptotically to a normal distribution.

Extra reading: Fisher Information Matrix.

Example: ML For Sensor Fusion

Assume sensors are unbiased but produce "noisy" measurements of 2D location.



Assume 2 sensors that make independent measurements of an object's location in 2d space.

. From the sensors we have $x_1, x_2 \in \mathbb{R}^2$ the measurements

From the sensor manufacturer we know the MVN observation noise variance for each sensor

$$x_i - x_{true} \sim N(0, \Sigma_i) \iff x_{true} \sim N(x_i, \Sigma_i) \iff x_i \sim N(x_{true}, \Sigma_i)$$

Let's say we believe both sensors equally and that they are independent. Then

$$p(x_1, x_2 \mid x_{true}) = p(x_1 \mid x_{true}) p(x_2 \mid x_{true}) = N(x_1; x_{true}, \Sigma_1) N(x_2; x_{true}, \Sigma_2)$$

The ML priciple says that

$$\hat{x}_{true} = \arg\max_{x_{true}} \log N(x_1; x_{true}, \Sigma_1) + \log N(x_2; x_{true}, \Sigma_2)$$

Setting up as usual

$$\frac{\partial}{\partial x_{true}}\Big((x_1 - x_{true})^T \Sigma_1^{-1} (x_1 - x_{true}) + (x_2 - x_{true})^T \Sigma_2^{-1} (x_2 - x_{true}) \Big) = 0$$

$$-2\Sigma_1^{-1}(x_1 - x_{true}) - 2\Sigma_2^{-1}(x_2 - x_{true}) = 0$$

$$\Sigma_1^{-1} x_1 + \Sigma_2^{-1} x_2 = \left(\Sigma_1^{-1} + \Sigma_2^{-1}\right) x_{true}$$

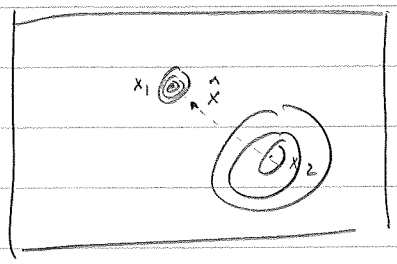
$$\hat{x}_{true} = \left(\Sigma_1^{-1} + \Sigma_2^{-1}\right)^{-1} \left(\Sigma_1^{-1} x_1 + \Sigma_2^{-1} x_2\right)$$

In multiple dimensions this is a little difficult to see, but, let's assume that the derivation holds in 1-D, then

$$\hat{x}_{true} = \left(\frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}\right)\left(\frac{1}{\sigma_1^2} x_1 + \frac{1}{\sigma_2^2} x_2\right)$$

which is just a weighted average with higher weight on the smallest variance.

Graphically (back in 2D)



ML & Linear Regression (is not what you think it is)

If you think linear regression is about drawing a line through points, think again! (OK, it's that too)

Linear regression as you know it-ish:

$n \in \{1, \dots N\}$

$y_n \in \mathbb{R}$ $\quad w \in \mathbb{R}$ $\quad Y = \{y_1, \dots, y_N\}$

$x_n \in \mathbb{R}$ $\quad b \in \mathbb{R}$ $\quad X = \{x_1, \dots x_N\}$

$$y_n | x_n \sim \mathcal{N}(w x_n + b, \sigma^2)$$

$$\Rightarrow \mathcal{L}(X; w, b) = \prod_{n=1}^{N} \mathcal{N}(y_n; w x_n + b, \sigma^2)$$

$$\underset{w}{\text{argmax}} \log \prod \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}(y_n - w x_n + b)/\sigma^2\right\}$$

= take deriv's w.r.t. w, b set $= 0$, = gross!

Use matrix linear regression with

$$\vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \qquad \vec{X} = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{bmatrix} \qquad \vec{w} = \begin{bmatrix} w \\ b \end{bmatrix}$$

with likelihood

$$\vec{y} - \vec{X}\vec{w} \sim \mathcal{N}(\vec{0}, I)$$

and

$$\underset{w}{\text{argmax}} \log \mathcal{L}(Y; \vec{w})$$

$$= w \text{ s.t. } \frac{\partial}{\partial \vec{w}} \log \mathcal{N}(\vec{y} - \vec{X}w; I) = \vec{0}$$

Plugging in to MVN PDF and taking logs we see

$$0 = \frac{\partial}{\partial \vec{w}} (\vec{y} - \vec{X}\vec{w})^T I (\vec{y} - \vec{X}\vec{w})$$

(least squares: $\sum_n (y_n - x_n \vec{w})^2 = \sum_n (y_n - (wx_n + b))^2$)

$$= \frac{\partial}{\partial \vec{w}} -2(\vec{X}\vec{w})^T \vec{y} + (\vec{X}\vec{w})^T (\vec{X}\vec{w})$$

$$= \frac{\partial}{\partial \vec{w}} -2\vec{w}^T \vec{X}^T \vec{y} + \vec{w}^T \vec{X}^T \vec{X} \vec{w}$$

From matrix cook book

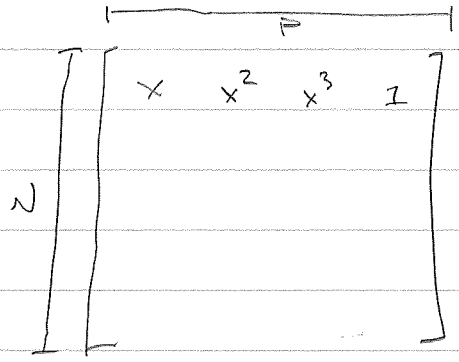
$$= -2\vec{X}^T \vec{y} + (\vec{X}^T \vec{X} + (\vec{X}^T \vec{X})^T)\vec{w}$$

$$= -\vec{X}^T \vec{y} + \vec{X}^T \vec{X} \vec{w}$$

$$\Rightarrow \vec{w}_{ML} = (\vec{X}^T \vec{X})^{-1} \vec{X}^T \vec{y} = \vec{w}_{\text{Least squares}}$$

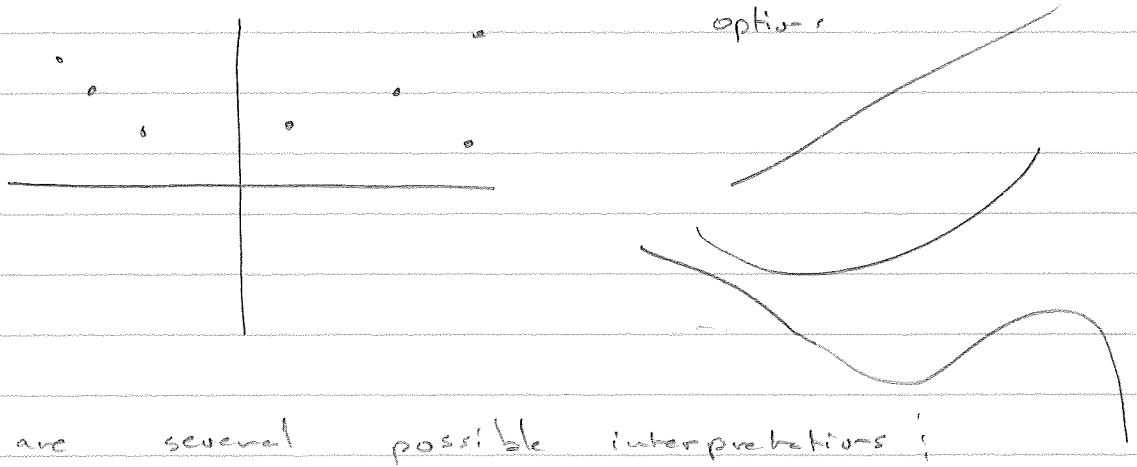
OK, so, we have that the linear regression estimator (linear in the parameters) is the same as the least squares estimator.

We can ask, when can this estimator be computed. Let's look at X



$X^T X \in \mathbb{R}^{P \times P}$ and $(X^T X)^{-1}$ requires $X^T X$ nonsingular, ie. $\text{rank}(X^T X) \geq p$.

You might ask, "When could the design matrix (X) be rank deficient?" Let's consider the following linear regression



There are several possible interpretations of said data

Let's say I prefer $y = ax^3 + bx^2 + cx + d$ for some unknown $\vec{\omega} = [a\ b\ c\ d]^T$.

Note that nothing from the above need change except the definitions.

$$X = \begin{bmatrix} x_1^3 & x_1^2 & x_1 & 1 \\ x_2^3 & x_2^2 & x_2 & 1 \\ & & & \end{bmatrix} \qquad \vec{\omega} = \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}$$

It is entirely feasible then that $p > n$ and X^TX becomes rank deficient.