Wishart distributions

- **Definition:** Let G be a $n \times p$ random matrix. Each column is a random vector, they are iid and follow the n-variate normal distribution

$$G = [\vec{g}^1, \dots, \vec{g}^p], \quad \vec{g}^i \sim \mathcal{N}_n(0, V)$$

Define $S = GG^T$. It results in a $n \times n$ matrix and it follows a new probability measure, called Wishart of dimension n

$$S = GG^T \sim W_n(V, p)$$

We can think it in terms of σ-algebra. Let $S_n(\mathbb{R})$ be the vector space of real symmetric matrices $n \times n$. We know the Frobenius norm

$$\|A\|_F = \sqrt{tr(A^TA)} = \sqrt{\sum_{i,j} a_{i,j}^2}, \quad \forall A = [a_{ij}] \in S_n(\mathbb{R})$$

Consider the topology τ^F induced by this norm
Now, consider the induced topology τ_+^F onto $S_n^{++}(\mathbb{R})$ the positive-definite matrices
Consider the Borel σ-algebra in $(S_n^{++}(\mathbb{R}), \tau_+^F)$, say $\mathfrak{I}_+$
In this case: $(S_n^{++}(\mathbb{R}), \mathfrak{I}_+, W_n(V,p))$ is a measure space

We can think that, $\forall A \in \Sigma_{\mathbb{R}^{n \times n}}$,

$$W_n(V, n)(S^{-1}(A)) = \mu(A)$$

Lebesgue measure in $\mathbb{R}^{n \times n}$

- **Properties:**
  - S is symmetric (obvious, because $S = G.G^T = (G.G^T)^T = S^T$)
  - Expectation: $\mathbb{E}[S] = pV$

- Inversibility: The rank of S is equal to the rank of G. As a condition for the inversion of a matrix is it have maximum rank (ie, non-linear dependences between rows and columns and, consequently, $\det \neq 0$), we need that $p \geq n$.
- χ^2-distribution: When $n = 1$ and $V = 1$, G will be just a random vector which each entrance will follow $\mathcal{N}(0,1)$. Consequently, $S = \sum_{i=1}^{p} (\vec{g}^{\,i})^2$, $\vec{g}^{\,i} \sim \mathcal{N}(0,1)$. By definition, this random variable follows the χ^2_p-distribution with p freedom parameters

$$S \sim W_1(1,p) \equiv \chi^2_p$$

- Assymptotic behavior: when $p \longrightarrow \infty$, S concentrates itself around pV. It's a consequence of the Law of Large Numbers

- We can define $W_n(V,p)$ in terms of its probability density. Let S be a $n \times n$ symmetric positive semi-definite random matrix. Let V be a $n \times n$ symmetric positive definite matrix. For $p \geq n$, $W_n(V,p)$ is a Wishart distribution if its probability density is

$$\rho_S([x]) = \frac{\det([x])^{\frac{p-n-1}{2}}}{2^{np/2} \det(V)^{p/2} \Gamma_n\left(\frac{p}{2}\right)} e^{-\frac{1}{2} tr(V^{-1}[x])}$$

↳ We supposed S semi-definite because it was positive semi-definite in our construction

$$\vec{z}^T S \vec{z} = \vec{z}^T G G^T \vec{z} = (G^T\vec{z})^T (G^T\vec{z}) = \|G^T\vec{z}\|^2 \geq 0$$

$[G]^T_{n \times n} \cdot [\vec{z}]_{n \times 1} = [G^T\vec{z}]_{n \times 1} \Rightarrow [G^T\vec{z}]^T_{1 \times n} [G^T\vec{z}]_{n \times 1} \in \mathbb{R}$, then $\|\cdot\|^2$ here is Euclidean

Multivariate gamma function

$$\Gamma_n\left(\frac{p}{2}\right) = \pi^{\frac{n(n-1)}{4}} \prod_{j=1}^{n} \Gamma\left(\frac{p}{2} - \frac{j-1}{2}\right)$$

- If $[x]$ is positive semi-definite, ie, if one of its eigenvalues is zero, then $\det([x]) = 0$ and, consequently, $\rho_S([x]) = 0$.
  - We can consider just the samples (matrices $n \times n$) which are positive definite

- Theorem (Maintenance by conjugation): Let $S \sim W_n(V,p)$ and let C be a matrix $q \times n$ such that $rk(C) = q$. Then,

$$\begin{cases} [C]_{q\times n}[S]_{n\times n}[C]^T_{n\times q} = [C]_{q\times n}[SC^T]_{n\times q} = [CSC^T]_{q\times q} \\ CSC^T \sim W_q(CVC^T, p) \end{cases}$$

- **Corollary (χ^2-distributions):** Now, consider $q=1 \Rightarrow C$ is a n-vector, ie, $C=(S_1 \dots S_n)^T$. Let's notate C as $\vec{z}$

$$\frac{\vec{z} S \vec{z}^T}{\vec{z} V \vec{z}^T} \sim W_1(\vec{z} V \vec{z}^T, n)$$

Naturality of the Wishart distributions

Let $\vec{X}_1, \dots, \vec{X}_n$ p-random vectors i.i.d. following the n-variate normal distribution: $\vec{X}_i \sim \mathcal{N}_p(\vec{\mu}, \Sigma)$

Notice we could work with the random matrix $M=[X_i]_{n\times n}$, but I would like to construct a a process analogous to the Gauss derivation for the gaussian distributions

Consider the MLF for $\vec{X}_1, \dots, \vec{X}_n$ below, where we wrote $\theta = (\vec{\mu}, \Sigma)$:

$$L(\theta | \vec{x}_1, \dots, \vec{x}_n) = \prod_{i=1}^{n} \rho_{\vec{X}_i}(\vec{x}_i | \theta)$$

$$= \left(\frac{1}{(2\pi)^{p/2} \det(\Sigma)^{1/2}} \right)^n \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (\vec{x}_i - \vec{\mu})^T \Sigma^{-1} (\vec{x}_i - \vec{\mu}) \right)$$

Let's compute the log-likehood function

$$\ell(\theta | \vec{x}_1, \dots, \vec{x}_n) = -\frac{np}{2} \ln(2\pi) - \frac{n}{2} \ln(\det \Sigma) - \frac{1}{2} \sum_{i=1}^{n} (\vec{x}_i - \vec{\mu})^T \Sigma^{-1} (\vec{x}_i - \vec{\mu})$$

Fixing Σ, let's derivate it in order to obtain a estimation for $\vec{\mu}$:

$$\frac{\partial}{\partial \vec{\mu}} \ell(\theta|\vec{x}_1,\dots,\vec{x}_n) = \frac{\partial}{\partial \vec{\mu}}\left(-\frac{1}{2}\sum_{i=1}^{n}(\vec{x}_i-\vec{\mu})^T\Sigma^{-1}(\vec{x}_i-\vec{\mu})\right)$$

The vector derivative is more subtle

$$\frac{\partial}{\partial \vec{x}}(A\vec{x}+\vec{b})^T C(A\vec{x}+\vec{b}) = (A\vec{x}+\vec{b})C^TA + (A\vec{x}+\vec{b})^T CA$$

Here, $A=-I$, $\vec{b}=\vec{x}_i$, $\vec{x}=\vec{\mu}$, $C=\Sigma^{-1}=(\Sigma^{-1})^T$

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

$$= \sum_{i=1}^{n}(\vec{x}_i-\vec{\mu})^T\Sigma^{-1}$$

$$= \sum_{i=1}\vec{x}_i^T\cdot\Sigma^{-1} - n\vec{\mu}^T\Sigma^{-1}$$

$\downarrow \sum \vec{x}_i^T =: \vec{x}_M^T$

$$= (\vec{x}_M^T - n\vec{\mu})^T\Sigma^{-1}$$

Applying the MLE method, we have

$$(\vec{x}_M^T - n\hat{\vec{\mu}})^T\Sigma^{-1} = 0 \Rightarrow \vec{x}_M = n\hat{\vec{\mu}} \Rightarrow \boxed{\hat{\vec{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\vec{x}_i}$$

Now, let's do the same for Σ:

$$\frac{\partial}{\partial \vec{\mu}} \ell(\theta|\vec{x}_1,\dots,\vec{x}_n) = -\frac{n}{2}\frac{\partial}{\partial\Sigma}\left(\ln(\det\Sigma)\right) - \frac{1}{2}\frac{\partial}{\partial\Sigma}\left(\sum_{i=1}^{n}(\vec{x}_i-\vec{\mu})^T\Sigma^{-1}(\vec{x}_i-\vec{\mu})\right)$$

$$\frac{\partial}{\partial A}\ln(\det A) = A^{-1}$$

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

$$a^Tb = tr(ba^T) \Rightarrow tr(a^TBa) = Baa^T$$

$$= -\frac{n}{2}\Sigma^{-1} - \frac{1}{2}\frac{\partial}{\partial\Sigma}\left(\sum_{i=1}^{n}tr\left[\Sigma^{-1}(\vec{x}_i-\vec{\mu})(\vec{x}_i-\vec{\mu})^T\right]\right)$$

Trace is linear

$$= -\frac{n}{2}\Sigma^{-1} - \frac{1}{2}\frac{\partial}{\partial \Sigma}\left(tr\left[\Sigma^{-1}\underbrace{\sum_{i=1}^{n}(\vec{x}_i-\vec{\mu})(\vec{x}_i-\vec{\mu})^T}_{S}\right]\right)$$

$$= -\frac{n}{2}\Sigma^{-1} - \frac{1}{2}\frac{\partial}{\partial \Sigma} tr(\Sigma^{-1}S)$$

$$\Big\downarrow \quad \frac{\partial}{\partial X} tr(X^{-1}A) = -X^{-1}AX^{-1}$$

$$= -\frac{n}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}S\Sigma^{-1}$$

Applying the MLE method, we have

$$\frac{\partial}{\partial \Sigma}\ell(\theta|\vec{x}_1,\dots,\vec{x}_n) = 0 \iff -\frac{n}{2}\hat{\Sigma}^{-1} + \frac{1}{2}\hat{\Sigma}^{-1}S\hat{\Sigma}^{-1} = 0$$

$$\iff -n\hat{\Sigma}^{-1} + \hat{\Sigma}^{-1}S\hat{\Sigma}^{-1} = 0$$

$$\iff \hat{\Sigma}^{-1} = \frac{1}{n}\hat{\Sigma}^{-1}S\hat{\Sigma}^{-1}$$

$$\iff \hat{\Sigma} = \frac{\hat{\Sigma}\hat{\Sigma}^{-1}S\hat{\Sigma}^{-1}\hat{\Sigma}}{n}$$

$$\iff \boxed{\hat{\Sigma} = \frac{S}{n}, \quad S = \sum_{i=1}^{n}(\vec{x}_i-\vec{\mu})(\vec{x}_i-\vec{\mu})^T}$$

Notice that, fixed $\vec{\mu}$, then $(\vec{x}_i-\vec{\mu})$ follows a normal distribution. Consequently,

$$S \sim W_p(\Sigma, n)$$

So, here, Wishart distributions appears naturally from the maximum likelihood estimation for Σ.

We need to be careful about $\vec{\mu}$.

Consider we know a "prior" mean $\mu = \mathbb{E}[X_1] = \mathbb{E}[X_2] = \dots = \mathbb{E}[X_n]$. Then,

$(\vec{x}_i - \vec{\mu}) \sim \mathcal{N}_p(0, \Sigma)$ and $S \sim W_p(\Sigma, n)$. Notice that, in this case, $\vec{\mu}$ is known previously.

Now, consider $(x_i - \hat{\vec{\mu}})$, where $\hat{\vec{\mu}}$ is the estimator used in the ML method, for example. Consequently, the random variables $(\vec{x}_1 - \hat{\vec{\mu}}), \dots (\vec{x}_n - \hat{\vec{\mu}})$ are not independent (each one depends on the sum over all of them).

In this last case, I can write $X_1 = n\hat{\vec{\mu}} - \sum_{i=2}^{n} X_i$ and X_1 is completely dependent on the others X_i. Then, I don't have anymore n levels of freedom to choose X_i, but $n-1$. In this case, S follows the called Wishart central distribution with $n-1$ levels of freedom

$$S \sim W_p(\Sigma, n-1)$$

Inverse Wishart distribution

Definition: Let $\varsigma \in S_n^{++}(\mathbb{R})$. ς follows the inverse Wishart distribution $W_p^{-1}(\Psi, n)$, for $\Psi \in S_n^{++}(\mathbb{R})$ and $p \geq n$, if $\varsigma^{-1} \sim W_p(\Psi^{-1}, n)$

$\varsigma \sim W_n^{-1}(\Psi, p)$

$\varsigma \in S_n^{++}(\mathbb{R})$

$\Psi \in S_n^{++}(\mathbb{R})$

$p \geq n$ levels of freedom

$\varsigma = G^{-1}(G^T)^{-1}$ $\begin{cases} G \sim \mathcal{N}_n(0, V) \\ G \in Mat_{n \times p}(\mathbb{R}) \end{cases}$

$S \sim W_n(\Psi^{-1}, p)$

$S = \varsigma^{-1} \in S_n^{++}(\mathbb{R})$

$\Psi^{-1} \in S_n^{++}(\mathbb{R})$

$p \geq n$ levels of freedom

$S = G^T G$ $\begin{cases} G \sim \mathcal{N}_n(0, V) \\ G \in Mat_{n \times p}(\mathbb{R}) \end{cases}$

- Its density is

$$\rho_S([x]) = \frac{\det(\Psi)^{p/2} \det([x])^{-\frac{(p+n+1)}{2}}}{2^{\frac{pn}{2}} \Gamma_n\left(\frac{p}{2}\right)} e^{-\frac{1}{2} tr(\Psi [x]^{-1})}$$