Maximum likelihood estimation of the parameters

of the multivariate normal distribution and

the distribution of these estimators (the Wishart distribution)

Definition 1 *The $p \times p$ random matrix $\mathbf{W}$ is a (centered) Wishart-matrix if it is of the form $\mathbf{W} = \mathbf{X}\mathbf{X}^T$, where the column vectors of the $p \times n$ random matrix $\mathbf{X}$ are i.i.d. $\mathcal{N}_p(\mathbf{0}, \mathbf{C})$ random vectors. In other words, the joint distribution of the entries of $\mathbf{W}$ is called Wishart-distribution with parameters p (dimension), n (degrees of freedom), and $\mathbf{C}$ (positive definite covariance matrix).*

This matrix is named after Wishart (1934) and it is the first random matrix of the history. We use the notation: $\mathbf{W} \sim \mathcal{W}_p(n, \mathbf{C})$.

We remark the following:

- It is easy to see that $\mathbf{W}$ is symmetric, positive semidefinite (Gram-matrix). Because of its symmetry, $\mathbf{W}$ follows, in fact, a $p(p+1)/2$-dimensional distribution.
- Denoting the column vectors of $\mathbf{X}$ by $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$, $\quad \mathbf{W} = \sum_{k=1}^n \mathbf{X}_k \mathbf{X}_k^T$.
- If $\mathbf{C} > 0$ and $n > p$, then $\mathbf{W} > 0$ (positive definite) with probability 1.
- The $\mathcal{W}_p(n, \mathbf{I}_p)$-distribution is called *standard Wishart-distribution.* In the case of $p = 1$, it is the $\chi^2(n)$-distribution.

Proposition 1 (Standardization) *Let $\mathbf{C} > 0$ be symmetric, positive definite. Then $\mathbf{W} \sim \mathcal{W}_p(n, \mathbf{C})$ holds if and only if $\mathbf{C}^{-1/2}\mathbf{W}\mathbf{C}^{-1/2} \sim \mathcal{W}_p(n, \mathbf{I}_p)$.*

The proof is the easy consequence of the fact, learned in Lesson 1, that $\mathbf{X}_i \in \mathcal{N}_p(\mathbf{0}, \mathbf{C})$ if and only if $\mathbf{Y}_i \in \mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$.

Proposition 2 (Additivity) *If $\mathbf{W}_1 \sim \mathcal{W}_p(n, \mathbf{C})$ and $\mathbf{W}_2 \sim \mathcal{W}_p(m, \mathbf{C})$ are independent, then $\mathbf{W}_1 + \mathbf{W}_2 \sim \mathcal{W}_p(n+m, \mathbf{C})$.*

Theorem 1 (Lukács) *Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \sim \mathcal{N}_p(\mathbf{m}, \mathbf{C})$ be i.i.d. sample, further*

$$\bar{\mathbf{X}} := \frac{1}{n}\sum_{k=1}^n \mathbf{X}_k \quad \textit{and} \quad \mathbf{S} := \sum_{k=1}^n (\mathbf{X}_k - \bar{\mathbf{X}})(\mathbf{X}_k - \bar{\mathbf{X}})^T.$$

Then

1. $\bar{\mathbf{X}} \sim \mathcal{N}_p(\mathbf{m}, \frac{1}{n}\mathbf{C})$,
2. $\mathbf{S} \sim \mathcal{W}_p(n-1, \mathbf{C})$,
3. $\bar{\mathbf{X}}$ *and* $\mathbf{S}$ *are (stochastically) independent. (Equivalently, $\bar{\mathbf{X}}$ is independent of both the empirical- and corrected empirical covariance matrices of the sample.)*

Proof: The first statement follows from the previously learned facts: $\bar{\mathbf{X}}$ is linear combination and we proved that its covariance matrix is $\frac{1}{n}\mathbf{C}$.

To prove the other two statements, let us transform the vectors $\mathbf{X}_k$'s with an $n \times n$ orthogonal matrix $\mathbf{V}$ such that it contains entries equal to $1/\sqrt{n}$ all along its last line. This implies that the sum of the entries in its first $n-1$ rows are all zeros. The transformation is the following:

$$\mathbf{Y}_i = \sum_{k=1}^{n} v_{ik}\mathbf{X}_k, \quad (i = 1, \ldots, n),$$

where v_{ik}'s are the entries of $\mathbf{V}$. Equivalently, with matrix notation: Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ and $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ be $p \times n$ matrices. Then

$$\mathbf{Y}^T = \mathbf{V}\mathbf{X}^T \qquad \text{or} \qquad \mathbf{Y} = \mathbf{X}\mathbf{V}^T.$$

The so defined random vectors $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ also have multivariate normal distribution and they are independent. Indeed, because of the multivariate normality it suffices to check that their cross-covariance matrices are zeros:

$$\begin{aligned}
\mathrm{Cov}(\mathbf{Y}_i, \mathbf{Y}_j) &= \mathbb{E}\left[(\mathbf{Y}_i - \mathbb{E}\mathbf{Y}_i)(\mathbf{Y}_j - \mathbb{E}\mathbf{Y}_j)^T\right] = \\
&= \mathbb{E}\left[\sum_{k=1}^{n} v_{ik}(\mathbf{X}_k - \mathbb{E}\mathbf{X}_k)\sum_{l=1}^{n} v_{jl}(\mathbf{X}_l - \mathbb{E}\mathbf{X}_l)^T\right] = \\
&= \sum_{k=1}^{n}\sum_{l=1}^{n} v_{ik}v_{jl}\mathbb{E}\left[(\mathbf{X}_k - \mathbb{E}\mathbf{X}_k)(\mathbf{X}_l - \mathbb{E}\mathbf{X}_l)^T\right] = \sum_{k=1}^{n}\sum_{l=1}^{n} v_{ik}v_{jl}\delta_{kl}\mathrm{Var}\mathbf{X}_k = \\
&= \sum_{k=1}^{n} v_{ik}v_{jk}\mathrm{Var}\mathbf{X}_k = \mathbf{C}\sum_{k=1}^{n} v_{ik}v_{jk} = \delta_{ij}\mathbf{C}.
\end{aligned}$$

This implies that the cross-covariance matrix of $\mathbf{Y}_i$ and $\mathbf{Y}_j$ is the zero matrix if $i \neq j$; therefore, they are independent of each other. Further, in the $i = j$ case we obtain that the covariance matrix of any $\mathbf{Y}_i$ is $\mathbf{C}$.

Their expectation vectors are:

$$\mathbb{E}\mathbf{Y}_i = \sum_{j=1}^{n} v_{ij}\mathbb{E}\mathbf{X}_j = \mathbf{m}\sum_{j=1}^{n} v_{ij} = \mathbf{0}$$

for $i = 1, \ldots, n-1$, and $\mathbb{E}\mathbf{Y}_n = \sqrt{n}\mathbf{m}$, since $\mathbf{Y}_n = \sqrt{n}\bar{\mathbf{X}}$.

Summarizing, $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n-1} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{C})$, $\mathbf{Y}_n \sim \mathcal{N}_p(\sqrt{n}\mathbf{m}, \mathbf{C})$, and they are completely independent.

We will also need that

$$\sum_{k=1}^{n} \mathbf{Y}_k\mathbf{Y}_k^T = \mathbf{Y}\mathbf{Y}^T = \mathbf{X}\mathbf{V}^T\mathbf{V}\mathbf{X}^T = \mathbf{X}\mathbf{X}^T = \sum_{k=1}^{n} \mathbf{X}_k\mathbf{X}_k^T,$$

where we used the above transformations and the fact that $\mathbf{V}$ being orthogonal, $\mathbf{V}^T\mathbf{V} = \mathbf{I}_n$.

Finally, we will write $\mathbf{S}$ in terms of the first $n-1$ $\mathbf{Y}_i$'s. In view of the multidimensional Steiner equality:

$$\mathbf{S} = \sum_{k=1}^{n}(\mathbf{X}_k - \bar{\mathbf{X}})(\mathbf{X}_k - \bar{\mathbf{X}})^T = \sum_{k=1}^{n}\mathbf{X}_k\mathbf{X}_k^T - n\bar{\mathbf{X}}\bar{\mathbf{X}}^T =$$
$$= \sum_{k=1}^{n}\mathbf{Y}_k\mathbf{Y}_k^T - n\left(\frac{\mathbf{Y}_n}{\sqrt{n}}\right)\left(\frac{\mathbf{Y}_n}{\sqrt{n}}\right)^T = \sum_{k=1}^{n-1}\mathbf{Y}_k\mathbf{Y}_k^T.$$

In this way, $\mathbf{S}$ is the dyadic sum of $n-1$ independent $\mathcal{N}_p(\mathbf{0},\mathbf{C})$ random vectors. Hence, by the definition of the Wishart-distribution, $\mathbf{S} \sim \mathcal{W}_p(n-1,\mathbf{C})$. Since $\mathbf{Y}_n$, and therefore $\bar{\mathbf{X}}$, is independent of the first $n-1$ $\mathbf{Y}_i$'s, it is also independent of $\mathbf{S}$. This finishes the proof.

Theorem 2 (ML-estimation of the multivariate normal parameters) *Based on the $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \sim \mathcal{N}_p(\mathbf{m},\mathbf{C})$ i.i.d. sample, the maximum likelihood (ML) estimators of the parameters are:*

$$\hat{\mathbf{m}} = \bar{\mathbf{X}}, \qquad \hat{\mathbf{C}} = \frac{1}{n}\mathbf{S}.$$

Proof: Consider the likelihood function written in the following form (see the preceding lesson):

$$L_{\mathbf{m},\mathbf{C}}(\mathbf{X}_1,\ldots,\mathbf{X}_n) = \frac{1}{(2\pi)^{np/2}|\mathbf{C}|^{n/2}}e^{-\frac{1}{2}\sum_{k=1}^{n}(\mathbf{X}_k-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{X}_k-\mathbf{m})}$$
$$= \frac{1}{(2\pi)^{np/2}|\mathbf{C}|^{n/2}}e^{-\frac{1}{2}\mathrm{tr}(\mathbf{C}^{-1}\mathbf{S})}e^{-\frac{1}{2}n(\bar{\mathbf{X}}-\mathbf{m})^T\mathbf{C}^{-1}(\bar{\mathbf{X}}-\mathbf{m})}.$$

The ln function being strictly increasing, to find the place of the maximum, it suffices to maximize the so-called log-likelihood function with respect to the parameters:

$$l_{\mathbf{m},\mathbf{C}}(\mathbf{X}_1,\ldots,\mathbf{X}_n) := \ln L_{\mathbf{m},\mathbf{C}}(\mathbf{X}_1,\ldots,\mathbf{X}_n)$$
$$= c - \frac{n}{2}\ln|\mathbf{C}| - \frac{1}{2}\mathrm{tr}(\mathbf{C}^{-1}\mathbf{S}) - \frac{1}{2}n(\bar{\mathbf{X}}-\mathbf{m})^T\mathbf{C}^{-1}(\bar{\mathbf{X}}-\mathbf{m}),$$

where the constant c does not depend on the parameters.

Observe that the quadratic form in the last term is nonnegative (since $\mathbf{C}$, and hence, $\mathbf{C}^{-1}$ is positive definite), and it is zero if and only if $\bar{\mathbf{X}} - \mathbf{m} = \mathbf{0}$, irrespective of the actual value of $\mathbf{C}$. Therefore, the ML-estimator of $\mathbf{m}$ is $\hat{\mathbf{m}} = \bar{\mathbf{X}}$, and substituting it for $\mathbf{m}$, the last term vanishes. Then we maximize the remaining part

$$l_{\hat{\mathbf{m}},\mathbf{C}}(\mathbf{X}_1,\ldots,\mathbf{X}_n) = c - \frac{n}{2}\ln|\mathbf{C}| - \frac{1}{2}\mathrm{tr}(\mathbf{C}^{-1}\mathbf{S})$$

of the log-likelihood function with respect to $\mathbf{C}$, or equivalently, with respect to $\mathbf{C}^{-1}$. That is, we maximize the function

$$g(\mathbf{C}^{-1}) = c + \frac{n}{2}\ln|\mathbf{C}^{-1}| - \frac{1}{2}\mathrm{tr}(\mathbf{C}^{-1}\mathbf{S})$$

with respect to $\mathbf{C}^{-1}$. For this purpose, we take its derivative with respect to $\mathbf{C}^{-1}$, where the result of the differentiation is a matrix, the ij-th entry of which

is the derivative of the scalar function $g(\mathbf{C}^{-1})$ with respect to the ij-th entry of $\mathbf{C}^{-1}$. (In fact, this is the gradient, the components of which vector are stored is matrix form.)

We will use the following rules of differentiating the matrix–scalar functions determinant and trace with respect to the matrix itself. If $\mathbf{A}$ and $\mathbf{S}$ are symmetric matrices, then

$$\frac{\partial|\mathbf{A}|}{\partial\mathbf{A}} = adj(\mathbf{A})$$

the adjoint matrix (of the signed minors) of $\mathbf{A}$, and

$$\frac{\partial \mathrm{tr}(\mathbf{AS})}{\partial\mathbf{A}} = \mathbf{S}.$$

Using these rules, we get that

$$\frac{\partial g(\mathbf{C}^{-1})}{\partial\mathbf{C}^{-1}} = \frac{n}{2}\frac{1}{|\mathbf{C}^{-1}|}\frac{\partial|\mathbf{C}^{-1}|}{\partial\mathbf{C}^{-1}} - \frac{1}{2}\frac{\partial \mathrm{tr}(\mathbf{C}^{-1}\mathbf{S})}{\partial\mathbf{C}^{-1}} = \frac{n}{2}\mathbf{C} - \frac{1}{2}\mathbf{S}.$$

If we make it equal to the $p\times p$ zero matrix, then we get that the ML-estimator of $\mathbf{C}$ is $\frac{1}{n}\mathbf{S}$. Since the log-likelihood function is continuously differentiable with respect to the parameters, and it tends to $-\infty$ when we approach the boundary of its domain, its only stationary point obtained above should be the place of its only local, and hence, global maximum. This finishes the proof.

Only for the Multivariate Statistics course

Theorem 3 *The density of the standard Wishart matrix* $\mathbf{W}\sim\mathcal{W}_p(\mathbf{0},\mathbf{I}_p)$ *and that of its eigenvalues is*

$$c_{np}|\mathbf{W}|^{\frac{n-p-1}{2}}e^{-\frac{1}{2}\mathrm{tr}\mathbf{W}} \quad \textit{and} \quad \kappa_{np}(\prod_{j=1}^{p}\lambda_j)^{\frac{n-p-1}{2}}e^{-\frac{1}{2}\sum_{j=1}^{p}\lambda_j}\prod_{j\neq k}|\lambda_j-\lambda_k|,$$

where the normalizing constants c_{np} *and* κ_{np} *only depend on* p *and* n $(n>p)$.

Using transformation formulas and Proposition 1, we obtain the density of a general Wishart-matrix:

Theorem 4 *The density of the Wishart-matrix* $\mathbf{W}\sim\mathcal{W}_p(\mathbf{0},\mathbf{C})$ *is*

$$c_{np}|\mathbf{W}|^{\frac{n-p-1}{2}}|\mathbf{C}|^{-\frac{n}{2}}e^{-\frac{1}{2}\mathrm{tr}\mathbb{C}^{-1}\mathbf{W}}.$$