Parameter estimation in multivariate models

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be i.i.d. sample from the $\mathbf{P}_{\underline{\theta}}$ distribution, where $\underline{\theta} \in \Theta$ and $\Theta \subset \mathbb{R}^k$ is the parameter space. The unknown parameter $\underline{\theta}$ is estimated by means of a $\mathbf{T} = \mathbf{T}(\mathbf{X}_1, \ldots, \mathbf{X}_n) = \mathbf{T}(\mathbf{X}) \in \mathbb{R}^k$ statistic, which depends on the sample condensed into the $p \times n$ matrix $\mathbf{X}$ column-wise. Here $\mathbf{X}^T$ is the *data matrix*.

For example, if our sample is from the $\mathcal{N}_p(\mathbf{m}, \mathbf{C})$ distribution, and p is given, then $\underline{\theta} = (\mathbf{m}, \mathbf{C})$ and $k = p + p(p+1)/2$. It is estimated with the $\mathbf{T} = (\bar{\mathbf{X}}, \hat{\mathbf{C}})$ or $(\bar{\mathbf{X}}, \hat{\mathbf{C}}^*)$ statistics.

Definition 1 *The statistic* $\mathbf{T}$ *is an unbiased estimator of* $\underline{\theta}$ *if* $\mathbb{E}_{\underline{\theta}}(\mathbf{T}) = \underline{\theta}$, $\forall \underline{\theta} \in \Theta$.

Definition 2 *The sequence of statistics* $\mathbf{T}_n = \mathbf{T}(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ *is asymptotically unbiased estimator of* $\underline{\theta}$ *if* $\lim_{n\to\infty} \mathbb{E}_{\underline{\theta}}(\mathbf{T}_n) = \underline{\theta}$, $\forall \underline{\theta} \in \Theta$.

For example, if our sample is from the $\mathcal{N}_p(\mathbf{m}, \mathbf{C})$ distribution, and p is given, then $\bar{\mathbf{X}}$ is an unbiased estimator of $\mathbf{m}$, whereas $\hat{\mathbf{C}}$ is asymptotically unbiased, and $\hat{\mathbf{C}}^*$ is unbiased estimator of $\mathbf{C}$.

Definition 3 *Let* $\mathbf{T}_1$ *and* $\mathbf{T}_2$ *be two unbiased estimators of the parameter* $\underline{\theta}$*, based on the same sample. We say that* $\mathbf{T}_1$ *is at least as efficient as* $\mathbf{T}_2$ *if for their covariance matrices*

$$\mathrm{Var}_{\underline{\theta}}(\mathbf{T}_1) \leq \mathrm{Var}_{\underline{\theta}}(\mathbf{T}_2) \qquad \forall \underline{\theta} \in \Theta$$

holds, which means that $\mathrm{Var}_{\underline{\theta}}(\mathbf{T}_2) - \mathrm{Var}_{\underline{\theta}}(\mathbf{T}_1) \geq 0$ *(positive semidefinite).*

An unbiased estimator is efficient if it is at least as efficient as any other unbiased estimator.

- Efficient estimator does not always exist, but if yes, then it is unique with probability 1.
- If the covariance matrix of an unbiased estimator attains the Cramér–Rao information (matrix) limit (see the forthcoming definition), then it is the efficient estimator.
- Even if the information limit cannot be attained with any unbiased estimator, there may exist an efficient estimator. As a consequence of the Rao–Blackwell–Kolmogorov theorem, an unbiased estimator is efficient if it is also a sufficient and complete statistic.

Definition 4 *The statistic* $\mathbf{T}$ *is sufficient for* $\underline{\theta}$ *if the distribution of the sample, conditioned on the given value of* $\mathbf{T}$*, dos not depend on* $\underline{\theta}$ *any more.*

In fact, a sufficient statistic contains all the information that can be retrieved from the sample for the parameter. We usually find sufficient statistics with the help of the following theorem.

Theorem 1 (Neyman–Fisher factorization) *The statistic* $\mathbf{T}$ *is sufficient for* $\underline{\theta}$ *if and only if the likelihood function (joint p.m.f. or p.d.f. of the sample) can be factorized as*

$$L_{\underline{\theta}}(\mathbf{X}) = g_{\underline{\theta}}(\mathbf{T}(\mathbf{X})) \cdot h(\mathbf{X}), \quad \forall \underline{\theta} \in \Theta.$$

Definition 5 *The statistic* $\mathbf{T}$ *is complete if* $\mathbb{E}_{\underline{\theta}}[g(\mathbf{T})] = \mathbf{0}$ *(*$\forall \underline{\theta} \in \Theta$*) implies that* $g = \mathbf{0}$ *with probability 1. (Here* $g : \mathbb{R}^k \to \mathbb{R}^k$ *is a measurable function.)*

A complete and sufficient statistic is also minimal sufficient (it is a function of any other sufficient statistic).

Definition 6 *The sequence of statistics* $\mathbf{T}_n = \mathbf{T}(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ *is a strongly consistent estimator of* $\underline{\theta}$ *if* $\mathbf{T}_n \to \underline{\theta}$ *as* $n \to \infty$ *almost surely (with probability 1),* $\forall \underline{\theta} \in \Theta$.

For example, if our sample is from the $\mathcal{N}_p(\mathbf{m}, \mathbf{C})$ distribution, and p is given, then in view of the Strong Law of Large Numbers, $\bar{\mathbf{X}}$ is a strongly consistent estimator of $\mathbf{m}$, whereas $\hat{\mathbf{C}}$ and $\hat{\mathbf{C}}^*$ are both strongly consistent estimators of $\mathbf{C}$.

Now, the multivariate counterpart of the Cramér–Rao inequality will be formulated. First we generalize the notion of the Fisher-information.

Definition 7 *The Fisher-information matrix of the* n*-element* p*-dimensional sample* $\mathbf{X}_1, \ldots, \mathbf{X}_n$*, taken from the* $\mathbf{P}_{\underline{\theta}}$ *distribution* ($\underline{\theta} \in \Theta \subset \mathbb{R}^k$) *is*

$$\mathbf{Inf}_n(\underline{\theta}) = \mathrm{Var}_{\underline{\theta}} \left(\nabla_{\underline{\theta}} \ln L_{\underline{\theta}}(\mathbf{X}_1, \ldots, \mathbf{X}_n) \right)$$

where $\nabla_{\underline{\theta}}$ *denotes the* k*-dimensional gradient vector, and* $\mathbf{Inf}_n(\underline{\theta})$ *is a* $k \times k$ *positive semidefinite matrix.*

Proposition 1 *Under some regularity conditions, e.g., the support of the p.m.f. or p.d.f. does not depend on the parameter, which always holds in exponential families (and the multivariate normal distribution belongs here),*

$$\mathbb{E}_{\underline{\theta}} \left(\nabla_{\underline{\theta}} \ln f_{\underline{\theta}}(\mathbf{X}_1) \right) = \mathbf{0}$$

and so,

$$\mathbf{Inf}_n(\underline{\theta}) = \mathbb{E}_{\underline{\theta}} \left[\left(\nabla_{\underline{\theta}} \ln L_{\underline{\theta}}(\mathbf{X}_1, \ldots, \mathbf{X}_n) \right) \left(\nabla_{\underline{\theta}} \ln L_{\underline{\theta}}(\mathbf{X}_1, \ldots, X_n) \right)^T \right].$$

Further,

$$\mathbf{Inf}_n(\underline{\theta}) = n\mathbf{Inf}_1(\underline{\theta}),$$

where

$$\begin{aligned} \mathbf{Inf}_1(\underline{\theta}) &= \mathrm{Var}_{\underline{\theta}} \left(\nabla_{\underline{\theta}} \ln f_{\underline{\theta}}(\mathbf{X}_1) \right) \\ &= \mathbb{E}_{\underline{\theta}} \left[\left(\nabla_{\underline{\theta}} \ln f_{\underline{\theta}}(\mathbf{X}_1) \right) \left(\nabla_{\underline{\theta}} \ln f_{\underline{\theta}}(\mathbf{X}_1) \right)^T \right] \end{aligned}$$

with f *denoting the p.d.f. of* $\mathbf{X}_1$ *(or of any sample entry) if it comes from an absolutely continuous distribution (e.g., from a p-variate normal distribution), otherwise the p.m.f. of a discrete distribution is to be used.*

Theorem 2 (Cramér–Rao inequality) *Let* $\mathbf{T}$ *be an unbiased estimator of* $\underline{\theta}$ *based on the i.i.d. sample* $\mathbf{X}_1, \ldots, \mathbf{X}_n$*, the covariance matrix of which exists. Then under some regularity conditions (imposed on the distribution itself),*

$$\mathrm{Var}_{\underline{\theta}}(\mathbf{T}) \geq \mathbf{Inf}_n^{-1}(\underline{\theta}) = \frac{1}{n}\mathbf{Inf}_1^{-1}(\underline{\theta}), \qquad \forall \underline{\theta} \in \Theta$$

where the inequality again means that the difference of the left- and right-hand side matrices is positive semidefinite.

Note that equality holds (the information limit is attained) only if this difference is the zero matrix. Applying the theorem for a multivariate normal sample with $\mathbf{T} = (\bar{\mathbf{X}}, \hat{\mathbf{C}}^*)$, the equality cannot be attained (even in the $p = 1$, $k = 2$ case). However, $\mathbf{T}$ is an efficient estimator, as a consequence of the Rao–Blackwell–Kolmogorov theorem. In fact, its covariance matrix asymptotically attains the information bound, akin to the covariance matrix of the ML-estimator $(\bar{\mathbf{X}}, \hat{\mathbf{C}})$ to be introduced in the next lesson.

Finally, let us find a sufficient statistic based on an i.i.d. $\mathbf{X}_1, \dots, \mathbf{X}_n \sim \mathcal{N}_p(\mathbf{m}, \mathbf{C})$ sample. We will apply the Neyman–Fisher factorization theorem for the likelihood function (joint density of the sample):

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

Here, in the exponent we used the multivariate Steiner equality and the fact that the tr operator is cyclically commutative:

$$\sum_{i=1}^{n}(\mathbf{X}_i - \mathbf{m})^T\mathbf{C}^{-1}(\mathbf{X}_i - \mathbf{m}) = \sum_{i=1}^{n}\mathrm{tr}[(\mathbf{X}_i - \mathbf{m})^T\mathbf{C}^{-1}(\mathbf{X}_i - \mathbf{m})] =$$
$$= \sum_{i=1}^{n}\mathrm{tr}[\mathbf{C}^{-1}(\mathbf{X}_i - \mathbf{m})(\mathbf{X}_i - \mathbf{m})^T] = \mathrm{tr}[\mathbf{C}^{-1}\sum_{i=1}^{n}(\mathbf{X}_i - \mathbf{m})(\mathbf{X}_i - \mathbf{m})^T] =$$
$$= \mathrm{tr}\mathbf{C}^{-1}\left[\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T + n(\bar{\mathbf{X}} - \mathbf{m})(\bar{\mathbf{X}} - \mathbf{m})^T\right] =$$
$$= \mathrm{tr}\mathbf{C}^{-1}\mathbf{S} + n\mathrm{tr}[(\bar{\mathbf{X}} - \mathbf{m})^T\mathbf{C}^{-1}(\bar{\mathbf{X}} - \mathbf{m})] = \mathrm{tr}\mathbf{C}^{-1}\mathbf{S} + n(\bar{\mathbf{X}} - \mathbf{m})^T\mathbf{C}^{-1}(\bar{\mathbf{X}} - \mathbf{m}).$$

As the likelihood function depends on the sample only through the statistics $\bar{\mathbf{X}}$ and $\mathbf{S}$, they will provide the sufficient statistics (the other factor is 1). Equivalently, the pair $(\bar{\mathbf{X}}, \hat{\mathbf{C}})$ or $(\bar{\mathbf{X}}, \hat{\mathbf{C}}^*)$ is also sufficient.

Only for the Multivariate Statistics course: c.d.f. of the multivariate normal distribution

Numerical approximations for

$$F(x_1,\dots,x_p)=\int_{-\infty}^{x_1}\dots\int_{-\infty}^{x_p} f(t_1,\dots,t_p)\,dt_1,\dots dt_p.$$

1. *Monte Carlo method.* Approximate the probability

$$F(x_1,\dots,x_p)=\mathbb{P}(X_1<x_1,\dots,X_p<x_p)$$

with the corresponding relative frequency based on an $\mathbf{X}_1,\dots,\mathbf{X}_n \sim \mathcal{N}_p(\mathbf{m},\mathbb{C})$ i.i.d. sample. How to do it?

2. *Expansion by means of Hermite polynomials*

Definition 8 *The correlation matrix $\mathbb{R}$ corresponding to the covariance matrix $\mathbf{C}$ is $\mathbb{R}=\mathbf{D}^{-1/2}\mathbf{C}\mathbf{D}^{-1/2}$ corr where the diagonal matrix $\mathbf{D}$ contains the positive diagonal entries of the covariance matrix $\mathbf{C}$ in its main diagonal.*

Definition 9 *The l^{th} orthogonal Hermite polynomial:*

$$H_l(x)=(-1)^l e^{x^2/2}\frac{d^l}{dx^l}e^{-x^2/2},\quad (l=0,1,2,\dots)$$

Proposition 2 *Let $\mathbf{X}=(X_1,\dots,X_p)\sim\mathcal{N}_p(\mathbf{0},\mathbb{R})$, and suppose that the eigenvalues of the matrix $\mathbf{R}-\mathbf{I}$ are less then 1 in absolute value (or equivalently, the spectrum of the correlation matrix $\mathbf{R}$ is in the (0,2) interval). Then*

$$\mathbb{P}(X_1\geq x_1,\dots,X_p\geq x_p)=\prod_{m=1}^{p}(1-\Phi(x_m))+$$

$$+\prod_{m=1}^{p}\phi(x_m)\cdot\sum_{k=1}^{\infty}\sum_{k_{ij}}\left(\prod_{i=1}^{p-1}\prod_{j=i+1}^{p}\frac{r_{ij}^{k_{ij}}}{k_{ij}!}\right)\prod_{q=1}^{p}H_{l_q-1}(x_q),$$

where the summation if for k_{ij}'s such that

$$k_{ij}\geq 0 \text{ integer},\quad \sum_{i=1}^{p-1}\sum_{j=i+1}^{p}k_{ij}=k,$$

r_{ij}'s are entries of $\mathbb{R}$, further $l_q=\sum_{i=1}^{q-1}k_{iq}+\sum_{j=q+1}^{p}k_{qj}$, H_l is the l^{th} Hermite polynomial, and $H_{-1}(x):=1-\Phi(x)$. This series is absolutely and uniformly convergent over $\mathbb{R}^p$.