A Historical Perspective on the Schützenberger-van Trees Inequality: A Posterior Uncertainty Principle

Olivier Rioul[0000−0002−8681−8916]

LTCI, Télécom Paris, Institut Polytechnique de Paris, France
olivier.rioul@telecom-paris.fr
https://perso.telecom-paristech.fr/rioul/

Abstract. The Bayesian Cramér-Rao Bound (BCRB) is generally attributed to Van Trees who published it in 1968. According to Stigler's law of eponymy, no scientific discovery is named after its first discoverer. This is the case not only for the Cramér-Rao bound itself—due in particular to the French mathematicians Fréchet and Darmois—but also for the van Trees inequality: The French physician, geneticist, epidemiologist and mathematician Marcel-Paul (Marco) Schützenberger, in a paper of just fifteen lines written in 1956—more than a decade before van Trees—had not only derived the BCRB but, as a close examination of his proof shows, used a very original approach based on the Weyl-Heisenberg uncertainty principle on the square root of the posterior distribution. This work reviews and extends Schützenberger's approach to Fisher information matrices, which opens up new perspectives.

1 An Overview of the (So-Called) Cramér-Rao Inequality

The well-known Cramér-Rao bound (CRB) allows one to easily evaluate the best possible parametric estimation performance in terms of quadratic risk. It is a fact known today by researchers in the field—at least in France [2]—that the bound was established for a real parameter $\theta \in \mathbb{R}$ by the French mathematician Maurice Fréchet for i.i.d. observations in his 1943 seminal paper [11]. As Fréchet recalls on the first page of his article,

> « *Le contenu de ce mémoire a formé une partie de notre cours de statistique mathématique à l'Institut Henri Poincaré pendant l'hiver 1939-1940.* »
> [The contents of this memoir formed part of our mathematical statistics course at the Institut Henri Poincaré during the winter of 1939-1940.]

The extension to non-i.i.d. observations and to the vector case of several parameters $\theta \in \mathbb{R}^d$ was soon made by his colleague Georges Darmois in 1945 [4]. Contrary to what is sometimes asserted, their student Daniel Dugué had not, it seems, already established this inequality in his 1937 thesis [7]. Fréchet cites the early 1898 work of Pearson and Filon [14], as well as later works of Edgeworth [8] and Fisher [9], but as he points out, these authors only proved an approximate bound for large values of n under an asymptotic normality assumption. He also cites Doob [6]:

> « *qui semble avoir obtenu le premier la formule non asymptotique et sans l'hypothèse de normalité faite par ses prédécesseurs* » [who seems to have been the first to obtain the non-asymptotic formula without the normality assumption made by his predecessors]

yet the relationship between the inequalities that Doob derived on Fisher informations and the CRB is not clear. The very same bound (in the scalar and vector cases) was eventually independently established by C. R. Rao [15] in 1945 and by Cramér in his excellent book [3, Chap. 32] in 1946, and is now known as the "Cramér-Rao bound".

The bound can be written as follows. For an unbiased estimator $\hat{\theta}(x)$ of parameter $\theta \in \mathbb{R}^d$, computed from observations $x = (x_1, x_2, \ldots, x_n)$, and under some well-known regularity conditions[1], the *quadratic risk* $\mathbf{R}_\theta$ (which equals the estimator's covariance) is lower bounded by the inverse of the *Fisher information matrix* (FIM):

$$\mathbf{R}_\theta = \mathrm{Cov}(\hat{\theta}(X)) = \mathbb{E}\big\{(\hat{\theta}(X) - \theta)(\hat{\theta}(X) - \theta)^t\big\} \geq \frac{1}{\mathbf{J}_\theta}. \tag{1}$$

Here "$\geq$" is the Loewner order on the set of symmetric matrices, defined by $A \geq B \iff A - B$ is positive semi-definite, and the FIM $\mathbf{J}_\theta$ is defined as the covariance matrix of the score (gradient of log-likelihood):

$$\mathbf{J}_\theta = \mathrm{Cov}(\nabla \log p_\theta(X)) = \mathbb{E}\big\{\nabla \log p_\theta(X) \nabla^t \log p_\theta(X)\big\}, \tag{2}$$

where $p_\theta(x) = p_\theta(x_1, x_2, \ldots, x_n)$ describes the parametric model of the joint distribution of the n observations and ∇ denotes the gradient w.r.t. θ. A classical calculation gives

$$\mathbf{J}_\theta = -\,\mathbb{E}\big\{\nabla\nabla^t \log p_\theta(X)\big\}, \tag{3}$$

where $\nabla\nabla^t$ is the Hessian operator w.r.t. θ. The latter identity shows that for independent and identically distributed (i.i.d.) observations we simply have $\mathbf{J}_\theta = n\mathbf{J}_{\theta,1}$ where $\mathbf{J}_{\theta,1}$ is the Fisher information for a single observation ($n = 1$).

A simple proof of (1) is based on the *matrix Cauchy-Schwarz inequality* (see Subsection A.1):

$$\mathrm{Cov}\hat{\theta}(X) \geq \mathrm{Cov}\{\hat{\theta}(X), \nabla\log p_\theta(X)\} \cdot \mathrm{Cov}(\nabla \log p_\theta(X))^{-1} \cdot \mathrm{Cov}\{\nabla\log p_\theta(X), \hat{\theta}(X)\} \tag{4}$$

where the inter-covariance $\mathrm{Cov}\{\hat{\theta}(X), \nabla\log p_\theta(X)\} = \mathrm{Cov}\{\nabla\log p_\theta(X), \hat{\theta}(X)\}^t$ simply reduces to $\mathbb{E}\big(\hat{\theta}(X)\frac{\nabla^t p_\theta(X)}{p_\theta(X)}\big) = \nabla^t\,\mathbb{E}(\hat{\theta}(X)) = \nabla^t\theta = I$ since the estimator

[1] Essentially, that the parametric model of the joint distribution of the n observations, $p_\theta(x) = p_\theta(x_1, x_2, \ldots, x_n)$, has support that does not depend on θ, is twice differentiable, and uniformly integrable as well as its first and second derivatives on its support.

is unbiased. Fréchet [11] and Cramér [3] also extended this inequality to the case of a *biased* estimator, with (possibly non-zero) bias $B_{\hat\theta}(\theta) = \mathbb{E}(\hat\theta(X)) - \theta$, in which case $\nabla^t \mathbb{E}(\hat\theta(X)) = \nabla^t(\theta + B_{\hat\theta}(\theta))$. This gives the following modified CRB with the gradient of the bias:

$$\mathbf{R}_\theta = \mathbf{K}_\theta + B_{\hat\theta}(\theta)B_{\hat\theta}(\theta)^t \geq (I + \nabla^t B_{\hat\theta}(\theta)) \cdot \mathbf{J}_\theta^{-1} \cdot \bigl(I + \nabla^t B_{\hat\theta}(\theta)\bigr)^t + B_{\hat\theta}(\theta)B_{\hat\theta}(\theta)^t. \quad (5)$$

The interest of such an improved bound is limited since it generally depends on the estimator itself *via* its bias. Subsequently, many other versions of the Cramér-Rao bound have been discussed and have allowed to take into account different regularity conditions from the classical framework [1], constraints in the parameter vector [13], a periodicity constraint inherent to certain estimation problems [17], and more recently the geometric structure of the parameters [21]. The Cramér-Rao bound has have found numerous applications in engineering problems, sometimes at the limit of abuse [23].

2 An Overview of the (So-Called) van Trees Inequality

One of the most important extensions of the Cramér-Rao bound is the *Bayesian Cramér-Rao bound* (BCRB) in a Bayesian context where the parameter of interest $\theta \in \mathbb{R}^d$ is assumed random and follows a known prior distribution $p(\theta)$. The model of the data distribution $p(x|\theta)$ depends on the variable θ and the quadratic risk matrix is no longer defined for a fixed "true" value of θ, but averaged over the prior distribution:

$$\mathbf{R} \triangleq \mathbb{E}_{x,\theta}\bigl\{(\hat\theta(X) - \theta)(\hat\theta(X) - \theta)^t\bigr\}, \quad (6)$$

where the expectation is now over the joint distribution $p(x, \theta) = p(x|\theta)p(\theta)$.

The BCRB is almost always attributed to van Trees who proved it in his reknown textbook [24] published in 1968:

$$\mathbf{R} \geq \mathbf{J}^{-1} = (\tilde{\mathbf{J}} + \mathbb{E}_\theta \mathbf{J}_\theta)^{-1} \quad (7)$$

where $\mathbf{J}$ is the *joint FIM*:

$$\mathbf{J} = \mathbb{E}_{x,\theta}\bigl\{\nabla \log p(X, \theta) \nabla^t \log p(X, \theta)\bigr\}. \quad (8)$$

A calculation identical to that which proves (3) from (2) gives (with appropriate regularity and decay assumptions on the prior):

$$\mathbf{J} = -\mathbb{E}_{x,\theta}\{\nabla\nabla^t \log p(X, \theta)\}. \quad (9)$$

Since $p(x, \theta) = p(\theta)p(x|\theta)$ under the logarithm, we have the relation $\mathbf{J} = \tilde{\mathbf{J}} + \mathbb{E}_\theta \mathbf{J}_\theta$ where

$$\tilde{\mathbf{J}} = \mathbb{E}_\theta\bigl\{\nabla \log p(\theta) \nabla^t \log p(\theta)\bigr\} = -\mathbb{E}_\theta\{\nabla\nabla^t \log p(\theta)\} \quad (10)$$

is the *prior FIM*, and where $\mathbf{J}_\theta$ is the classical Fisher information given by (2)-(3) for $p_\theta(x) = p(x|\theta)$, which equals $\mathbf{J}_\theta = n\mathbf{J}_{\theta,1}$ in the case of i.i.d. observations (conditionally to θ). We see in particular that the influence of the *a priori* eventually disappears for a very large number of observations since $\mathbf{J} = \tilde{\mathbf{J}} + n\,\mathbb{E}_\theta \mathbf{J}_{\theta,1} \sim \mathbb{E}_\theta\,\mathbf{J}_\theta$ when $n \to \infty$, which reduces to the classical bound when $\mathbf{J}_\theta$ does not depend on θ.

Van Trees' proof (which is detailed only in the scalar case $d = 1$) is directly inspired by the Cauchy-Schwarz inequality (4) used in the classical case, but applied to the joint distribution $p(x, \theta)$ in place of $p_\theta(x) = p(x|\theta)$. In the general case $d \geq 1$ this reads (see Subsection A.1)

$$\mathbf{R} \triangleq \mathbf{R}_{\hat{\theta}(X)-\theta} \geq \mathbf{R}_{\hat{\theta}(X)-\theta,\nabla \log p(X,\theta)} \cdot \mathbf{R}^{-1}_{\nabla \log p(X,\theta)} \cdot \mathbf{R}_{\nabla \log p(X,\theta),\hat{\theta}(X)-\theta} \tag{11}$$

where the cross-covariance matrix equals

$$\begin{aligned}\mathbf{R}_{\hat{\theta}(X)-\theta,\nabla \log p(X,\theta)} &= \iint (\hat{\theta}(x) - \theta) \nabla^t p(x, \theta) \\ &= \iint (\nabla^t \theta) p(x, \theta) + \int \nabla \int (\hat{\theta}(x) - \theta) p(x, \theta) \\ &= I + \int \nabla \{p(\theta) B_\theta(\theta)\} = I\end{aligned} \tag{12}$$

under the following assumption on the bias:

$$\lim_{|\theta| \to \infty} p(\theta) B_{\hat{\theta}}(\theta) = 0, \tag{13}$$

a crucial assumption under which van Trees proves the inequality (7). Moreover, the equality condition in the Cauchy-Schwarz inequality (11) implies that the BCRB is achieved when the posterior distribution $p(\theta|x)$ is Gaussian. All these results obtained by van Trees have become classical today, and reproduced verbatim in most textbooks. But few know that he was preceded by more than a decade by Marcel-Paul Schützenberger.

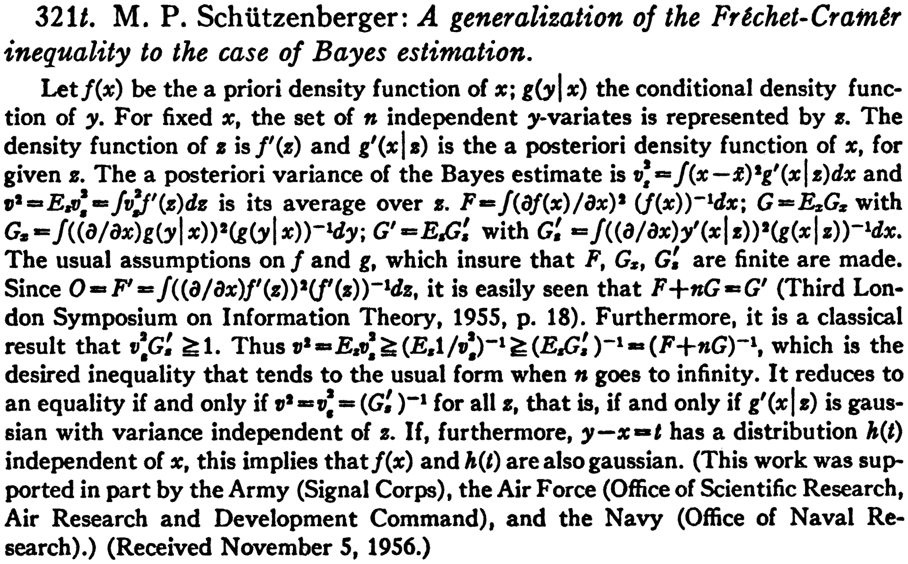
3 Marcel-Paul Schützenberger's 1956 Contribution

It was only in 2007, during the publication of a collection of articles on Bayesian bounds [25], that van Trees mentioned that his BCRB bound had been derived independently by "Shutzenberger" (sic) and commented:

> "*This derivation is a model of economy (1/3 of a page) but does not appear to have been noticed by either the engineering or statistical communities.*"

Figure 1 shows the third of a page in question, a small paragraph of about fifteen lines. It is actually a simple announcement in the AMS bulletin which was later republished in more detail (and in French) in [19].

The author of this note is in fact Marcel-Paul Schützenberger, French physician, geneticist, epidemiologist and mathematician, a colorful character [16] who defended his thesis in 1953 under the direction of Georges Darmois, after having



321*t*. M. P. Schützenberger: ***A generalization of the Frĕchet-Cramĕr inequality to the case of Bayes estimation.***

Let $f(x)$ be the a priori density function of x; $g(y|x)$ the conditional density function of y. For fixed x, the set of n independent y-variates is represented by z. The density function of z is $f'(z)$ and $g'(x|z)$ is the a posteriori density function of x, for given z. The a posteriori variance of the Bayes estimate is $v_z^2=\int(x-\bar{x})^2g'(x|z)dx$ and $v^2=E_zv_z^2=\int v_z^2f'(z)dz$ is its average over z. $F=\int(\partial f(x)/\partial x)^2\,(f(x))^{-1}dx$; $G=E_xG_x$ with $G_x=\int((\partial/\partial x)g(y|x))^2(g(y|x))^{-1}dy$; $G'=E_zG_z'$ with $G_z'=\int((\partial/\partial x)y'(x|z))^2(g(x|z))^{-1}dx$. The usual assumptions on f and g, which insure that F, G_x, G_z' are finite are made. Since $O=F'=\int((\partial/\partial x)f'(z))^2(f'(z))^{-1}dz$, it is easily seen that $F+nG=G'$ (Third London Symposium on Information Theory, 1955, p. 18). Furthermore, it is a classical result that $v_z^2G_z'\geqq 1$. Thus $v^2=E_zv_z^2\geqq(E_z1/v_z^2)^{-1}\geqq(E_zG_z')^{-1}=(F+nG)^{-1}$, which is the desired inequality that tends to the usual form when n goes to infinity. It reduces to an equality if and only if $v^2=v_z^2=(G_z')^{-1}$ for all z, that is, if and only if $g'(x|z)$ is gaussian with variance independent of z. If, furthermore, $y-x=t$ has a distribution $h(t)$ independent of x, this implies that $f(x)$ and $h(t)$ are also gaussian. (This work was supported in part by the Army (Signal Corps), the Air Force (Office of Scientific Research, Air Research and Development Command), and the Navy (Office of Naval Research).) (Received November 5, 1956.)

Fig. 1: The entirety of the publication [18] written in 1956.

completed his medical thesis entitled *Contribution to the study of sex at birth.* Already in his 1953 thesis, he established deep connections between statistics (Fisher information in particular) and information theory (Shannon's mutual information). He also discovered the famous *Pinsker inequality* with optimal first and second-order constants, 7 years before Pinsker himself and 17 years before the precise republication of this inequality by Kullback—see [16] for more details.

Schützenberger was certainly aware of the work of Maurice Fréchet, his thesis jury president, on the so-called CRB which he actually called the *Fréchet-Cramér inequality.* Invited by Claude Shannon to MIT during the 1956-57 academic year, he wrote his article on the BCRB before November 1956. It is likely that his inspiration came in particular from discussions with David Slepian, who was working at Bell Laboratories at the time on estimation problems [20] and was well acquainted with the CRB via Cramér's book [3]. In fact, Slepian is mentioned in a footnote by Schützenberger [19] as having "independently obtained" the BCRB, although Slepian's work has apparently not been published.

A second researcher is also mentioned in the same footnote by Schützenberger [19], as having independently obtained the BCRB: "Mr J. Dard (article to appear in *Annals of Mathematical Statistics*)". As it seems, this is no other than John J. Gart, whose article was indeed published in this journal in 1959 [12]. This article is also included at the end of van Trees' collection [25] without much comment. Gart's proof simply rewrites the inequality of the classical case (4) by also averaging over the *a priori* $p(\theta)$, to obtain a Bayesian version of (5) where the derivative of the bias $B_{\hat{\theta}}(\theta)$ appears but where the *a priori* information $\tilde{J}$ has disappeared. Unfortunately, Gart only considers the diagonal terms in the FIM without taking into account the inter-parameter correlations.

Schützenberger's work on the BCRB is particularly interesting. It is one of his very last in the field of statistics. He never returned to this field afterwards, focusing from his time at MIT on the theory of codes, formal languages, automata, word combinatorics, etc.—all domains of theoretical computer science and combinatorics for which he is best known today.

Albeit fallen into oblivion, Schützenberger's proof of the BCRB is surprisingly original. The original publication deals only with the scalar case $\theta \in \mathbb{R}$, but can be easily extended to the vector case $\theta \in \mathbb{R}^d$ as shown below.

First of all, a careful decryption of his article (Fig. 1) shows that the emphasis is placed on the posterior $p(\theta|x)$ rather than on the prior, and on the following *posterior FIM*:

$$\tilde{\mathbf{J}}(x) = \mathbb{E}_{\theta|x}\bigl\{\nabla \log p(\theta|x)\nabla^t \log p(\theta|x)\bigr\} = -\,\mathbb{E}_{\theta|x}\bigl\{\nabla\nabla^t \log p(\theta|x)\bigr\} \qquad (14)$$

which hardly seems to be mentioned anywhere in the literature and satisfies the relation $\mathbf{J} = \mathbb{E}_x\,\tilde{\mathbf{J}}(x)$, as can be seen by expanding $p(x,\theta) = p(x)p(\theta|x)$ in $\mathbf{J}$—the unconditional data distribution $p(x)$ disappears in the differentiation since it does not depend on θ.

The introduction of the posterior is indeed natural since the optimal estimator, which minimizes the quadratic risk, is precisely given by the mean of the posterior distribuition $\hat{\theta}^*(x) = \mathbb{E}(\theta|x)$. Schützenberger, therefore, begins his derivation by showing that it suffices to prove the BCRB (7) on the minimal risk

$$\min \mathbf{R}_{\theta-\mathbb{E}(\theta|x)} = \mathbb{E}_x \operatorname{Cov}(\theta|x) \qquad (15)$$

where $\operatorname{Cov}(\theta|x) = \mathbb{E}_{\theta|x}\{(\theta - \mathbb{E}(\theta|x))(\theta - \mathbb{E}(\theta|x))^t\}$ is the covariance matrix of the *posterior.* The crucial step in Schützenberger's proof is the following inequality that resembles the BCRB, but for a given observation vector:

$$\operatorname{Cov}(\theta|x) \geq \tilde{\mathbf{J}}(x)^{-1} \qquad (16)$$

Schützenberger (in the scalar case $d = 1$) simply says that this is a "classical result" (cf. Figure 1) in [18]. However, he specifies in [19] that it is "Weyl's inequality". Although this is not obvious, one can indeed see it as the Weyl-Heisenberg inequality which constitutes the famous *uncertainty principle* in quantum mechanics, which Hermann Weyl proves in his 1928 book on quantum mechanics [26, App. 1]. The proof of this inequality (as well as the equality case) is classical in the scalar case $d = 1$ and carried out in Subsection A.2 in the general case of dimension d, where it is equivalent to the matrix inequality

$$\mathbf{R}_{t\cdot f} \geq \frac{1}{4}\mathbf{R}_{\nabla f}^{-1} \qquad (17)$$

for any function $f \in L^2(\mathbb{R}^d)$ such that $tf(t) \in L^2$ and $\nabla f \in L^2$.

Although he does not write it explicitly, the key point of Schützenberger's proof consists in applying Weyl's inequality (17) to the function $f(\theta) = \sqrt{p(\theta|x)}$ for fixed x. After making a change of variable $\theta \leftarrow \bigl(\theta - \mathbb{E}(\theta|x)\bigr)$, this reads

$$\operatorname{Cov}(\theta|x) \geq \frac{1}{4}\mathbf{R}^{-1}_{\nabla\sqrt{p(\theta|x)}}. \qquad (18)$$

Now since $\nabla\sqrt{p(\theta|x)} = \frac{1}{2\sqrt{p(\theta|x)}}\nabla p(\theta|x)$, we have

$$\begin{aligned}\mathbf{R}_{\nabla\sqrt{p(\theta|x)}} &= \mathbb{E}_{\theta|x}\,\frac{1}{4p(\theta|x)}\nabla p(\theta|x)\nabla^t p(\theta|x)\\ &= \frac{1}{4}\,\mathbb{E}_{\theta|x}\bigl\{\nabla\log p(\theta|x)\nabla^t\log p(\theta|x)\bigr\} = \frac{1}{4}\tilde{\mathbf{J}}(x),\end{aligned} \tag{19}$$

which gives (16) for any fixed data x. Finally, taking the expectation over the unconditional law $p(x)$ yields the BCRB:

$$\mathbf{R} \geq \mathbb{E}_x\,\mathrm{Cov}(\theta|x) \geq \mathbb{E}_x\bigl(\tilde{\mathbf{J}}(x)^{-1}\bigr) \geq \bigl(\mathbb{E}_x\,\tilde{\mathbf{J}}(x)\bigr)^{-1} = \mathbf{J}^{-1} \tag{20}$$

where the last inequality comes from the operator convexity of the function $A \mapsto A^{-1}$ (see Subsection A.3). Like van Trees a decade later, Schützenberger writes this bound in the form (7) with $\mathbf{J} = \tilde{\mathbf{J}} + n\,\mathbb{E}_\theta\,\mathbf{J}_{\theta,1}$ since he assumes i.i.d. observations [18,19] (see Figure 1).

4 Conclusion and Perspectives

It is unfortunate that Schützenberger's pioneering work has been forgotten for so long. Admittedly, his abstract in [18] (Figure 1) is not easy to decipher and the barely longer version [19] certainly had the drawback of being written in French. Nevertheless, Schützenberger's idea of a resemblance or equivalence between the CRB and the uncertainty principle has reappeared several times independently since then. As early as 1972, one speaks of the (non-Bayesian) CRB as "resembling that of Heisenberg" [10, p. 198] without further precision. In 1991, Dembo *et al.* [5] show an equivalence between the classical (non-Bayesian) CRB and the uncertainty principle for a location parameter $p(x|\theta) = p(x-\theta)$. It seems that this kind of equivalence is *not* possible *in general* for the non-Bayesian CRB.

Schützenberger's proof also has several advantages over the usual approach, due to the fact that it is equivalent to an uncertainty inequality on the posterior: First, unlike van Trees (or Gart), it does not assume conditions on the bias, such as (13). Indeed, Weyl's inequality is automatically verified as soon as the quantities involved in the BCRB are finite. Second, the equality condition becomes obvious since it is that of the Weyl-Heisenberg inequality, namely that the posterior must be Gaussian. Finally, it opens new perspectives based on improvements or variants of uncertainty inequalities.

A Appendix

A.1 Matrix Cauchy-Schwarz Inequality

Somewhat surprisingly, the generalization of the classical Cauchy-Schwarz inequality to matrices does not seem to be well known in the literature, except perhaps in econometrics where Tripathi [22] proved it in the particular case where the dominating measure μ below is a probability measure. As Tripathi noticed,

> *"Although this inequality looks astonishingly familiar, I have been unable to discover any references to it in the literature".*

For completeness we give a simplified derivation in the more general case of any dominating measure.

Consider d-dimensional functions $f : \mathbb{R}^m \to \mathbb{C}^d$, where $f = (f_i)_i$ is written by convention as a column vector, and whose components f_i are square integrable w.r.t. some measure μ (e.g., a probability measure): $\int \|f\|^2 \,\mathrm{d}\mu < +\infty$ (we write $f \in L^2(\mu)$). We let

$$\mathbf{R}_{f,g} \triangleq \int f g^\dagger \,\mathrm{d}\mu \tag{21}$$

be the $d \times d'$ "*cross-correlation*" matrix between $f : \mathbb{R}^m \to \mathbb{C}^d$, $g : \mathbb{R}^m \to \mathbb{C}^{d'}$, where $\dagger$ denotes the conjugate transpose. We also let $\mathbf{R}_f \triangleq \mathbf{R}_{f,f} = \int f f^\dagger \,\mathrm{d}\mu$ be the "*auto-correlation*" matrix of f. Some obvious properties are: $\mathbf{R}_f = \mathbf{R}_{-f} = \mathbf{R}_{if} \geq 0$ (positive semi-definite), $\mathbf{R}_{g,f} = \mathbf{R}_{f,g}^\dagger$ and $\mathbf{R}_f = \mathbf{R}_f^\dagger$ (Hermitian symmetry), $\mathbf{R}_{f+g} = \mathbf{R}_f + \mathbf{R}_g + \mathbf{R}_{f,g} + \mathbf{R}_{g,f}$ (sesquilinearity), and under linear transformations A, B, $\mathbf{R}_{Af,Bg} = A\mathbf{R}_{f,g}B^\dagger$ and $\mathbf{R}_{Af} = A\mathbf{R}_f A^\dagger$.

If $\mathbf{R}_g > 0$ (positive definite), the matrix *Cauchy-Schwarz inequality* reads

$$\boxed{\mathbf{R}_f \geq \mathbf{R}_{f,g}\mathbf{R}_g^{-1}\mathbf{R}_{g,f}}. \tag{22}$$

where $\geq$ denotes the Loewner order ($A \geq B$ iff $A - B \geq 0$).

Proof. Let $h = f - \mathbf{R}_{f,g}\mathbf{R}_g^{-1}g$ and expand $\mathbf{R}_h \geq 0$ using sesquilinearity: $\mathbf{R}_h = \mathbf{R}_f + \mathbf{R}_{f,g}\mathbf{R}_g^{-1}\mathbf{R}_g\mathbf{R}_g^{-1}\mathbf{R}_{g,f} - 2\mathbf{R}_{f,g}\mathbf{R}_g^{-1}\mathbf{R}_{g,f} = \mathbf{R}_f - \mathbf{R}_{f,g}\mathbf{R}_g^{-1}\mathbf{R}_{g,f}$. □

It is easily seen that equality holds in (22) iff $f = Ag$ for some linear transformation A. The case $d = d' = 1$ reduces to the classical (scalar) Cauchy-Schwarz inequality $|\int f g^* \,\mathrm{d}\mu|^2 \leq \int |f|^2 \,\mathrm{d}\mu \int |g|^2 \,\mathrm{d}\mu$.

For any $f \in L^1(\mu) \cap L^2(\mu)$, define $\mu_f = \int f \,\mathrm{d}\mu$ and similarly for g. Their $d \times d'$ *cross-covariance* matrix is defined as

$$\mathbf{K}_{f,g} = \mathrm{Cov}(f,g) \triangleq \mathbf{R}_{f-\mu_f,g-\mu_g} = \mathbf{R}_{f,g} - \mu_f\mu_g^\dagger. \tag{23}$$

We also let $\mathbf{K}_f = \mathrm{Cov}(f) \triangleq \mathbf{K}_{f,f} = \mathbf{R}_{f,g} - \mu_f\mu_f^\dagger$ be the "*(auto)-covariance*" matrix of f. Applied to $f - \mu_f$ and $g - \mu_g$, the matrix *Cauchy-Schwarz inequality* reads

$$\mathbf{R}_f \geq \boxed{\mathbf{K}_f \geq \mathbf{K}_{f,g}\mathbf{K}_g^{-1}\mathbf{K}_{g,f}}. \tag{24}$$

A.2 Matrix Weyl-Heisenberg Inequality

Let $\mu = \mathrm{d}x$ be the Lebesgue measure, let f, g be as in the preceding Subsection, and let $\hat{f}, \hat{g}$ be their respective (componentwise) Fourier transforms, e.g., $\hat{f} = (\hat{f}_j)_j$ where $\hat{f}_j(\nu) = \int_{\mathbb{R}^m} f_j(t)e^{-2i\pi\nu\cdot t}\,\mathrm{d}t$ (equality in the L^2-sense). By Parseval identities $\langle f_j \mid g_k \rangle = \langle \hat{f}_j \mid \hat{g}_k \rangle$, one has the matrix *Parseval-Plancherel identity*

$$\boxed{\mathbf{R}_{f,g} = \mathbf{R}_{\hat{f},\hat{g}}}. \tag{25}$$

In particular let $d = 1$, $f = g : \mathbb{R}^m \to \mathbb{C}$ with Fourier transform $\hat{f}$, then $\nabla f \in L^2$ has Fourier transform $2i\pi\nu\hat{f}(\nu)$, and replacing we have

$$\mathbf{R}_{\nabla f} = 4\pi^2 \mathbf{R}_{\nu \hat{f}}. \tag{26}$$

Assume for simplicity that f takes real values and $\int f^2(t)\,\mathrm{d}t = 1$. The cross-correlation $\mathbf{R}_{t\cdot f,\nabla f} = \int t f(t) \nabla^t f(t)\,\mathrm{d}t = \frac{1}{2}\int t \nabla^t f^2(t)\,\mathrm{d}t = -\frac{1}{2} I$ by integration by parts. Applying (22) gives the matrix *Weyl-Heisenberg inequality*

$$\boxed{\mathbf{R}_{t\cdot f} \geq \frac{1}{16\pi^2} \mathbf{R}^{-1}_{\nu\cdot\hat{f}}}. \tag{27}$$

The general case where f takes complex values is proved similarly. One easily sees that the equality case corresponds to the case where $f(t)$ is a (multivariate) Gaussian by solving a first-order differential equation. A classical argument allows one to replace t by $t - \mu_t$ and ν by $\nu - \mu_\nu$, where $\mu_t = \int t |f(t)|^2\,\mathrm{d}t$ and $\mu_\nu = \int \nu |\hat{f}(\nu)|^2\,\mathrm{d}\nu$. The matrix inequality (27) does not seem to be well-known; it appears in [5, Thm. 19]. The case $m = 1$ reduces to the classical (scalar) Weyl-Heisenberg inequality $\sigma_t^2 \geq \frac{1}{16\pi^2}\sigma_\nu^{-1}$, that is, $\sigma_t \sigma_\nu \geq \frac{1}{4\pi}$.

A.3 Operator Convexity of the Inverse

For completeness we give a simple proof that $\mathbb{E}(A^{-1}) \geq \big(\mathbb{E}(A)\big)^{-1}$.

Proof. Let $B = \mathbb{E}(A)$ and expand $0 \leq B^{-1}(B - A)A^{-1}(B - A)B^{-1} = A^{-1} - 2B^{-1} + B^{-1}AB^{-1}$. Taking expectations gives $\mathbb{E}(A^{-1}) - B^{-1} \geq 0$. □

Acknowledgement The author wishes to thank Alexandre Renaux for his expertise on the BCRB, and Valérie Girardin for pointing out reference [22] during the review process.

References

1. Bar-Shalom, Y., Osborne, R.W., Willett, P., Daum, F.E.: Cramér-Rao-Leibniz lower bound — A new estimation bound for finite support measurement noise. In: 53rd IEEE Conf. Decision Control. pp. 2609–2614. L.A., CA, USA (2014)
2. Barbaresco, F.: Le cours de Maurice Fréchet à l'IHP pendant l'hiver 1939, borne de Fréchet-Darmois, densités distinguées et géométrie de l'information. In: Conférence internationale Maurice Fréchet: Les mathématiques, l'abstrait et le concret. Institut Henri Poincaré, Paris, France (Oct, 9–11 2023)
3. Cramér, H.: Mathematical Methods of Statistics, vol. 9. Princeton Univ. Press (Sep 1946)
4. Darmois, G.: Sur les limites de la dispersion de certaines estimations. Revue Inst. Int. Stat. **13**(1/4), 9–15 (1945)
5. Dembo, A., Cover, T.M., Thomas, J.A.: Information theoretic inequalities. IEEE Transactions on Information Theory **37**(6), 1501–1518 (Nov 1991)

6. Doob, J.L.: Statistical estimation. Transactions of the American Mathematical Society **39**(3), 410–421 (May 1936)
7. Dugué, D.: Application des propriétés de la limite au sens du calcul des probabilités à l'étude des diverses questions d'estimation. Journal de l'Ecole Polytechnique **3**(4), 305–372 (1937)
8. Edgeworth, F.Y.: On the probable errors of frequency-constants. Journal of the Royal Statistical Society **71 & 72**, 381–397, 499–512, 651–678, 81–90 (1908–1909)
9. Fisher, R.A.: On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London. Series A **CCXXII**, 309–368 (Apr 1922)
10. Fourgeaud, C., Fuchs, A.: Statistique. Dunod (1972)
11. Fréchet, M.: Sur l'extension de certaines évaluations statistiques au cas de petits échantillons. Revue Inst. Int. Stat. **11**(3/4), 182–205 (1943)
12. Gart, J.J.: An extension of the Cramér-Rao inequality. Ann. Math. Statist. **30**, 367–380 (Jun 1959)
13. Gorman, J.D., Hero, A.O.: Lower bounds for parametric estimation with constraints. IEEE Transactions on Information Theory **36**(6), 1285–1301 (Nov 1990)
14. Pearson, K., Filon, L.N.G.: Mathematical contributions to the theory of evolution.—IV. On the probable errors of frequency constants and on the influence of random selection on variation and correlation. Philosophical Transactions of the Royal Society of London. Series A **191**, 229–311 (Dec 1898)
15. Rao, C.R.: Information and accuracy attainable in the estimation of statistical parameters. Bulletin of the Calcutta Mathematical Society **37**, 81–91 (1945)
16. Rioul, O.: A historical perspective on Schützenberger-Pinsker inequalities (extended version). Information Geometry **7**, S737–S779 (2024)
17. Routtenberg, T., Tabrikian, J.: Non-bayesian periodic Cramér-Rao bound. IEEE Transactions on Signal Processing **61**(4), 1019–1032 (Feb 2013)
18. Schützenberger, M.P.: A generalization of the Fréchet-Cramér inequality to the case of Bayes estimation. Bulletin of the American Mathematical Society **63**, 142 (1957)
19. Schützenberger, M.P.: A propos de l'inégalité de Fréchet-Cramer. Publ. Inst. Statist. Univ. Paris **7**, 3–6 (1958)
20. Slepian, D.S.: Estimation of signal parameters in the presence of noise. Trans. IRE Professional Group Inf. Theory **3**(3), 68–69 (1954)
21. Smith, S.T.: Covariance, subspace, and intrinsic Cramér-Rao bounds. IEEE Trans.Signal Proc. **53**(5), 1610–1630 (May 2005)
22. Tripathi, G.: A matrix extension of the Cauchy-Schwarz inequality. Economics Letters **63**, 1–3 (1999)
23. Vallisneri, M.: Use and abuse of the Fisher information matrix in the assessment of gravitational-wave parameter-estimation prospects. Phys. Rev. D **77** (Jun 2008)
24. van Trees, H.L.: Detection, Estimation and Modulation Theory, vol. 1. John Wiley & Sons (1968)
25. van Trees, H.L., Bell, K.L. (eds.): Bayesian Bounds for Parameter Estimation and Nonlinear Filtering/Tracking. IEEE Press, John Wiley & Sons (2007)
26. Weyl, H.: Gruppentheorie und Quantenmechanik. S.Hirzel Verlag (1928)