A Proof of the Fisher Information Inequality via a Data Processing Argument *

Ram Zamir

Dept. of Elect. Eng. - Systems, Tel Aviv University, ISRAEL e-mail: zamir@eng.tau.ac.il

November 23, 1997

Abstract

The Fisher information $J(X)$ of a random variable X under a translation parameter appears in information theory in the classical proof of the Entropy-Power Inequality (EPI). It enters the proof of the EPI via the De-Bruijn identity, where it measures the variation of the differential entropy under a Gaussian perturbation, and via the convolution inequality $J(X+Y)^{-1} \geq J(X)^{-1} + J(Y)^{-1}$ (for independent X and Y), known as the Fisher Information Inequality (FII). The FII is proved in the literature directly, in a rather involved way. We give an alternative derivation of the FII, as a simple consequence of a "data-processing inequality" for the Cramer-Rao lower bound on parameter estimation.

Key Words: Fisher information, data processing inequality, Cramer-Rao bound, non-Gaussian noise, linear modeling, pre-filtering, entropy-power inequality.

*This research was supported in part by the Wolfson Research Awards administered by the Israel Academy of Science and Humanities. This work was presented in part in the Information Theory Workshop on Multiple Access and Queuing, April 1995, St. Louis, Missouri.

I. Introduction

The data processing inequality (or the data processing theorem) is used in information theory for proving the converse channel-coding theorem [4, sec. 5.3,5.4], [6, sec. 2.8, 8.9]. This inequality asserts that if the random variables $W - X - Y$ form a Markov chain in this order, then the mutual informations between them satisfy

$$I(W;Y) \leq I(W;X). \tag{1}$$

In the special case where Y is given by a deterministic function ϕ of X, (1) becomes

$$I(W;\phi(X)) \leq I(W;X), \tag{2}$$

with equality if $W - \phi(X) - X$ form a Markov chain, e.g., if $\phi(\cdot)$ is an invertible function. The proof of (1) follows straightforwardly from the chain rule and the positivity of the mutual information [6].

The name "data processing inequality" apparently came from the analogy to the problem of optimal filtering. Suppose that W, X, Y are real variables. In analogy with (1) and (2), it is clear and easy to verify that the conditional variance, i.e., the mean squared error of the conditional mean estimator of W, satisfies the data processing inequalities

$$\mathrm{VAR}(W|Y) \geq \mathrm{VAR}(W|X) \quad \text{and} \quad \mathrm{VAR}(W|\phi(X)) \geq \mathrm{VAR}(W|X) \tag{3}$$

where $\mathrm{VAR}(W|X) \stackrel{\Delta}{=} E[W - E(W|X)]^2$.

When the estimated quantity is a parameter $\boldsymbol{\theta}$ (i.e., not a random variable), it is impossible to use the conditional variance as a measure for the goodness of the optimal estimator. Instead, it is common to use the Fisher Information matrix (FI) of the measurement $\mathbf{X}$ relative to the parameter vector $\boldsymbol{\theta}$, defined as [10, 4, 6]

$$\mathbf{J}(\mathbf{X};\boldsymbol{\theta}) \stackrel{\Delta}{=} \mathrm{COV}\Big\{\frac{\partial}{\partial\boldsymbol{\theta}} \ln\Big(f_{\boldsymbol{\theta}}(\mathbf{X})\Big)\Big\} = \int \frac{1}{f_{\boldsymbol{\theta}}(\mathbf{x})}\Big(\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial\boldsymbol{\theta}}\Big) \cdot \Big(\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial\boldsymbol{\theta}}\Big)^t d\mathbf{x} \tag{4}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$, the set $\{f_{\boldsymbol{\theta}}(\mathbf{x})\}$ is a family of densities of $\mathbf{X}$ parameterized by $\boldsymbol{\theta}$, $\frac{\partial}{\partial \boldsymbol{\theta}}$ denotes the gradient (i.e., a column vector of partial derivatives) with respect to the parameters $\theta_1, \ldots \theta_m$, $\ln(\cdot)$ denotes the natural logarithm, and $\mathrm{COV}\{\cdot\}$ denotes the $m \times m$ covariance matrix calculated relative to the distribution of $\mathbf{X}$. Here $\mathbf{X}$ may either be a single measurement or a vector of n measurements. The importance of the matrix $\mathbf{J}(\mathbf{X}; \boldsymbol{\theta})$ follows from the Cramer-Rao Bound (CRB), [10, 4, 6], saying that for any un-biased estimator $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{x})$ (i.e., estimator for which $E\{\hat{\boldsymbol{\theta}}(\mathbf{X})\} = \boldsymbol{\theta}$) the error vector $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ satisfies

$$\mathrm{COV}\Big\{\hat{\boldsymbol{\theta}}(\mathbf{X})\Big\} \geq \mathbf{J}(\mathbf{X}; \boldsymbol{\theta})^{-1} \tag{5}$$

where throughout the paper an inequality between (nonnegative definite) matrices means that the difference matrix is nonnegative definite. As it turns out (see Lemma 3 below), the notion of data processing extends easily to the FI; if $\boldsymbol{\theta} - \mathbf{X} - \mathbf{Y}$ satisfy a chain relation of the form $f(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) = f_{\boldsymbol{\theta}}(\mathbf{x}) f(\mathbf{y}|\mathbf{x})$ (i.e., the conditional distribution of Y given X is independent of $\boldsymbol{\theta}$), then we have the data processing inequality

$$\mathbf{J}(\mathbf{Y}; \boldsymbol{\theta}) \leq \mathbf{J}(\mathbf{X}; \boldsymbol{\theta}), \tag{6}$$

whose deterministic version (in analogy with (2)) is

$$\mathbf{J}(\phi(\mathbf{X}); \boldsymbol{\theta}) \leq \mathbf{J}(\mathbf{X}; \boldsymbol{\theta}). \tag{7}$$

Equality in (7) holds if $\phi(X)$ is a *sufficient statistic* relative to the family $\{f_{\boldsymbol{\theta}}(x)\}$, i.e., $\boldsymbol{\theta} - \phi(\mathbf{X}) - \mathbf{X}$ form a chain [6, sec. 2.10][1].

In the context of information theoretic inequalities, e.g., in the derivation of the Entropy Power Inequality (EPI), there appears a special form of the FI matrix, namely the FI of a random vector with respect to a *translation parameter*

$$\mathbf{J}(\mathbf{N}) \triangleq \mathbf{J}(\boldsymbol{\theta} + \mathbf{N}; \boldsymbol{\theta}) = \mathrm{COV}\Big\{\frac{\partial}{\partial \mathbf{N}} \ln f(\mathbf{N})\Big\} = \int \frac{1}{f(\mathbf{n})} \left(\frac{\partial f(\mathbf{n})}{\partial \mathbf{n}}\right) \cdot \left(\frac{\partial f(\mathbf{n})}{\partial \mathbf{n}}\right)^t d\mathbf{n} \tag{8}$$

[1] An alternative ("Bayesian") way to express the equality condition is that $\boldsymbol{\Theta} - \phi(\mathbf{X}) - \mathbf{X}$ form a Markov chain for any distribution on the parameter θ.

where $f(\mathbf{n})$ is the density function of the vector $\mathbf{N}$ (which is independent of $\boldsymbol{\theta}$), and $\mathbf{J}(\mathbf{N})$ is a square matrix whose dimension equals that of $\mathbf{N}$; see [3, 8, 4, 6]. Unlike the general case (4), this form of the FI is a function of the density of the random vector alone, and not of its parameterization[2].

The FI under translation (8) exhibits some well known properties [1, 7], e.g.,

$$\mathbf{J}(A\mathbf{N}) = A^{-t}\mathbf{J}(\mathbf{N})A^{-1} \tag{9}$$

for any non-singular square matrix A, and

$$\mathbf{J}(\mathbf{N}) \geq \mathrm{COV}(\mathbf{N})^{-1} \tag{10}$$

with equality iff $\mathbf{N}$ is Gaussian. Another property which is of particular interest for us is a convolution inequality, called the *Fisher Information Inequality* (FII). Let N_1 and N_2 be statistically independent random variables. Then

$$J(N_1 + N_2)^{-1} \geq J(N_1)^{-1} + J(N_2)^{-1}, \tag{11}$$

with equality iff N_1 and N_2 are Gaussian. Vector, matrix and "convex" versions of (11) exist in the literature [8, 7], some of which will be mentioned in the sequel. The FII (11), together with the De-Bruijn identity[3], consist the key tools in the classical proof of the EPI [3, 4, 8]. Both FII and EPI relate to the tendency towards Gaussianity of the sum of independent random variables [1, 16].

Existing proofs of the FII (11) [3, 4] involve a direct calculation of the convolution of the densities of $\mathbf{N}_1$ and $\mathbf{N}_2$ and application of the Cauchy-Schwartz inequality, and they are rather technical. In this paper we show that the FII follows from the Fisher information data processing inequality given in (7). We derive the FII by applying the data processing inequality to a suitable linear model

[2]In some references the FI of $\mathbf{N}$ is defined as $K(\mathbf{N}) = \mathrm{trace}\{\mathbf{J}(\mathbf{N})\}$.

[3]The scalar form of the De-Bruijn identity is $\frac{d}{dt}h(X + \sqrt{t}Z) = \frac{1}{2}J(X + \sqrt{t}Z)$, where Z is a standard normal variable.

relating the measurements and the parameters. This model provides an interesting interpretation to the difference between the two sides of inequality (11): $J(N_1+N_2)^{-1} - J(N_1)^{-1} - J(N_2)^{-1}$ amounts to the *loss in the CRB after optimal linear estimation.* If N_1 and N_2 are Gaussian, linear estimation is globally optimum, and the CRB loss is zero. However, if N_1 and N_2 are not Gaussian, non-invertible linear operation may increase the CRB.

In our proof we consider a generalized form of the FII (11), namely, the matrix form of the FII, which was presented originally in [16, 15]. The new derivation of the FII (11) is given in Section II. Some additional properties of the matrix form of the FII are given in Section III. In Section IV we use the matrix-FII to analyze the loss in FI (or in CRB) due to pre-filtering in a certain linear model for parameter estimation. This part of the work appeared originally in [14].

II. Derivation of Results

In this section we prove a matrix form of the FII using the Fisher information data processing inequality (7). For completeness we give also a proof for ineq. (7), for which we could not find a reference in the literature. We then show that the matrix form of the FII implies the form in (11).

The derivation follows a sequence of lemmas. Below we assume that $\{f_{\boldsymbol{\theta}}(x,y)\}$ is a family of density functions parameterized by $\boldsymbol{\theta}$, where the first and second derivatives of $f_{\boldsymbol{\theta}}(x,y)$ with respect to $\boldsymbol{\theta}$ exist and are absolutely integrable (see [10, pp. 66, eq. (c)]).

Lemma 1 (Chain rule for the FI matrix)

$$\mathbf{J}(X,Y;\boldsymbol{\theta}) = \mathbf{J}(X;\boldsymbol{\theta}) + \mathbf{J}(Y;\boldsymbol{\theta}|X) \tag{12}$$

where

$$\mathbf{J}(Y;\boldsymbol{\theta}|X) = E_X\Big\{\mathbf{J}(Y;\boldsymbol{\theta}|X=x)\Big\} = E_{XY}\left\{\left(\frac{\partial \ln f_{\boldsymbol{\theta}}(Y|X)}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ln f_{\boldsymbol{\theta}}(Y|X)}{\partial \boldsymbol{\theta}}\right)^t\right\}. \tag{13}$$

Note that the argument of the first expectation in (13), $\mathbf{J}(Y;\boldsymbol{\theta}|X=x)$, is the FI of Y relative to $\boldsymbol{\theta}$, calculated with respect to the conditional density of Y given a specific value $X=x$.

Proof: We prove assuming θ is scalar. The generalization to a vector $\boldsymbol{\theta}$ is straightforward. Since $J(X,Y;\theta) = E\{[\partial \ln f_\theta(X,Y)/\partial\theta]^2\}$, and $\ln f_\theta(x,y) = \ln f_\theta(x) + \ln f_\theta(y|x)$, we see that $J(X,Y;\theta)$ is given by the sum in the right hand side of (12), plus a cross term which is twice

$$E\left\{\frac{\partial \ln f_\theta(X)}{\partial\theta} \cdot \frac{\partial \ln f_\theta(Y|X)}{\partial\theta}\right\} = E_X\left\{\frac{\partial \ln f_\theta(X)}{\partial\theta} \cdot E_Y\left\{\frac{\partial \ln f_\theta(Y|X)}{\partial\theta}\middle| X\right\}\right\}, \tag{14}$$

where the right hand side follows by iterating the expectation. This cross term is zero since the inner expectation in the right hand side of (14) is zero for each value of X (see e.g. [10, pp. 67, eq. (190)])[4]. □

Lemma 2 **(Data refinement inequality)**

$$\mathbf{J}(X,Y;\boldsymbol{\theta}) \geq \mathbf{J}(X;\boldsymbol{\theta}) \tag{15}$$

with equality if X is a sufficient statistic relative to the family $\{f_{\boldsymbol{\theta}}(x,y)\}$, i.e., if $\boldsymbol{\theta} - X - (X,Y)$ form a chain.

Proof: The inequality follows from Lemma 1 by the non-negativity of the FI $\mathbf{J}(Y;\boldsymbol{\theta}|X=x)$. The equality condition implies that $\mathbf{J}(Y;\boldsymbol{\theta}|X=x) = 0$ for each x. □

Lemma 3 **(Data processing inequality)**

$$\mathbf{J}(X;\boldsymbol{\theta}) \geq \mathbf{J}(\phi(X);\boldsymbol{\theta}) \tag{16}$$

with equality if $\phi(X)$ is a sufficient statistic relative to the family $\{f_{\boldsymbol{\theta}}(x)\}$, e.g., if $\phi(\cdot)$ is an invertible function.

Proof: By Lemma 2 we have $\mathbf{J}(\phi(X);\boldsymbol{\theta}) \leq \mathbf{J}(\phi(X),X;\boldsymbol{\theta}) = \mathbf{J}(X;\boldsymbol{\theta})$, where the second equality follows from the chain rule (12) since $\phi(X)$ is deterministic given X thus $\mathbf{J}(\phi(X);\boldsymbol{\theta}|X) = 0$. The inequality becomes equality if $\mathbf{J}(X;\boldsymbol{\theta}|\phi(X)) = 0$, i.e., if $\phi(X)$ is a sufficient statistic. □

[4]As pointed out by A. Yeredor, the proof follows even more easily from the equivalent formula for FI $J(X,Y;\theta) = -E\{\partial^2 \ln f_\theta(X,Y)/\partial\theta^2\}$.

Lemma 4 (Parameter transformation) *Let $\{f_{\boldsymbol{\phi}}(\mathbf{x})\}$ be a family of densities parameterized by $\boldsymbol{\phi}$, and suppose that the vector of parameters $\boldsymbol{\phi} \in \mathcal{R}^n$ is a function of the vector $\boldsymbol{\theta} \in \mathcal{R}^m$. Then*

$$\mathbf{J}(X;\boldsymbol{\theta}) = \left(\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}\right)^t \cdot \mathbf{J}(X;\boldsymbol{\phi}(\boldsymbol{\theta})) \cdot \left(\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}\right) \tag{17}$$

where $\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}} = \left\{\frac{\partial \phi_i}{\partial \theta_j}\right\}$ $i = 1 \ldots n, j = 1 \ldots m$ denotes the matrix of partial derivatives of the vector function $\boldsymbol{\phi}(\boldsymbol{\theta})$. In the linear case $\boldsymbol{\phi} = Q^t \boldsymbol{\theta}$, where Q is some $m \times n$ matrix, we have

$$\mathbf{J}(X;\boldsymbol{\theta}) = Q \cdot \mathbf{J}(X;\phi = Q^t \boldsymbol{\theta}) \cdot Q^t. \tag{18}$$

Proof: By the chain rule for derivatives, we have

$$\frac{\partial \ln f_{\boldsymbol{\phi}(\boldsymbol{\theta})}(x)}{\partial \boldsymbol{\theta}} = \left(\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}\right)^t \frac{\partial \ln f_{\boldsymbol{\phi}}(x)}{\partial \boldsymbol{\phi}}$$

which when substituted in the definition of $J(X;\boldsymbol{\theta})$ (4) results in (17). □

We now define a linear model relating the estimated parameters $\boldsymbol{\theta}$ with their measurements $\mathbf{X}$ and $\mathbf{Y}$. Let the random vector $\mathbf{N} \in \mathcal{R}^n$ have some joint density $f_{\mathbf{N}}$, and let Q and P be some $m_\theta \times n$ and $m_y \times n$ matrices, respectively, where $m_\theta \leq m_y \leq n$. Let $\boldsymbol{\theta} \in \mathcal{R}^{m_\theta}$ be a vector of parameters, and consider the linear model

$$\mathbf{X} = Q^t \cdot \boldsymbol{\theta} + \mathbf{N} \quad \text{and} \quad \mathbf{Y} = P \cdot \mathbf{X}, \tag{19}$$

where $\mathbf{X} \in \mathcal{R}^n$ and $\mathbf{Y} \in \mathcal{R}^{m_y}$.

Lemma 5 *For the model in (19)*

$$\mathbf{J}(\mathbf{X};\boldsymbol{\theta}) = Q\mathbf{J}(\mathbf{N})Q^t \quad \textit{and} \quad \mathbf{J}(\mathbf{Y};\boldsymbol{\theta}) = QP^t\mathbf{J}(P\mathbf{N})PQ^t \tag{20}$$

where $\mathbf{J}(\mathbf{N})$ and $\mathbf{J}(P\mathbf{N})$ are the $n \times n$ and the $m_y \times m_y$ FI matrices under a translation parameter (8) of the vectors $\mathbf{N}$ and $P\mathbf{N}$, respectively.

Proof: Let $\phi = Q^t\boldsymbol{\theta}$. Using Lemma 4 we have $\mathbf{J}(\mathbf{X};\theta) = \mathbf{J}(\phi + \mathbf{N};\boldsymbol{\theta}) = Q\mathbf{J}(\phi + \mathbf{N};\phi)Q^t = Q\mathbf{J}(\mathbf{N})Q^t$, where the last equality follows from the definition of FI under a translation parameter (8). The expression for $\mathbf{J}(\mathbf{Y})$ follows similarly. □

Lemma 6 *For the model in (19)*

$$\mathbf{J}(\mathbf{X};\boldsymbol{\theta}) \geq \mathbf{J}(\mathbf{Y};\boldsymbol{\theta}). \tag{21}$$

Proof: The inequality follows from the data processing inequality (16) since $\mathbf{Y} = P\mathbf{X}$. $\square$

Lemma 5 and Lemma 6 above imply three inequalities regarding the effect of linear transformation on the FI under translation:

Corollary 1 **A)** *For any $m \times n$ matrix A and random vector $\mathbf{N} \in \mathcal{R}^n$*

$$A^t \mathbf{J}(A\mathbf{N}) A \leq \mathbf{J}(\mathbf{N}). \tag{22}$$

B) *For any $m \times n$ matrix Λ with orthonormal rows, i.e., $\Lambda\Lambda^t = I_{m\times m} \triangleq$ the $m \times m$ identity matrix, and any $\mathbf{N}$*

$$\mathbf{J}(\Lambda\mathbf{N}) \leq \Lambda \mathbf{J}(\mathbf{N}) \Lambda^t. \tag{23}$$

C) *For any $m \times n$ matrix A with a full row rank, and random vector $\mathbf{N} \in \mathcal{R}^n$ with non-singular FI matrix $\mathbf{J}(\mathbf{N})$,*

$$\mathbf{J}(A\mathbf{N}) \leq (A\mathbf{J}(\mathbf{N})^{-1}A^t)^{-1}, \tag{24}$$

with equality if $m = n$ or if $\mathbf{N}$ is Gaussian.

Remark: ineq. (24) was shown in [16, 13] for the case where the components of $\mathbf{N}$ are statistically independent.

Proof:

A) Assume $m_\theta = 1$. Combining (20) and (21) implies $Q \cdot [\mathbf{J}(\mathbf{N}) - P^t\mathbf{J}(\mathbf{N})P]Q^t \geq 0$ for any $1 \times m_\theta$ vector Q. This implies that the inner difference is a nonnegative definite matrix.

B) Substitute $m_\theta = m_y$ and $Q = P = \Lambda$ in (20), use $\Lambda\Lambda^t = I_{m\times m}$, and substitute in (21).

C) Substitute $m_\theta = m_y$, $P = A$, and $Q = (A\mathbf{J}(\mathbf{N})^{-1}A^t)^{-1}A\mathbf{J}(\mathbf{N})^{-1}$ in (20), and substitute in (21) to obtain the inequality. The sufficient conditions for equality follow from (9) and (10), noting that $\mathrm{COV}(AN) = A\mathrm{COV}(N)A^t$. See Proposition 3 in the next section for a necessary and sufficient condition for $N_1 \ldots N_n$ statistically independent.

□

In the model above the vector $\mathbf{N}$ was arbitrary. We now specialize to a vector with *independent components* and give our main result.

Theorem 1 (FII) *Let N_1 and N_2 be statistically independent random variables. Then*

$$J(N_1+N_2) \leq \left(J(N_1)^{-1}+J(N_2)^{-1}\right)^{-1} \tag{25}$$

with equality if N_1 and N_2 are Gaussian. The inverse of the left and right hand sides of (25) are equal to the CRBs when θ is estimated from X_1+X_2 and (X_1,X_2), respectively, where $X_i=\alpha_i\theta+N_i$ and $\alpha_i = 1-\frac{J(N_i)}{J(N_1)+J(N_2)}$ $\quad i=1,2$.

Proof: Since N_1 and N_2 are independent, $\mathbf{J}([N_1,N_2])$ is a 2×2 diagonal matrix whose diagonal elements are $J(N_1)$ and $J(N_2)$. Hence the inequality follows by substituting $m=1$, $n=2$ and $A=[1,1]$ in (24). Equality for N_1 and N_2 Gaussian follows from the equality condition of (10).

The interpretation of $J(N_1+N_2)^{-1}$ and $J(N_1)^{-1}+J(N_2)^{-1}$ as CRBs can be seen by substituting $P=[1,1]$ and $Q=[\alpha_1,\alpha_2]$ in the linear model (19), and then using (20) to calculate the corresponding CRBs. □

We thus proved that the FII follows from the Fisher information data processing inequality. Furthermore, the FII corresponds to the loss in CRB due to "filtering" in a certain linear additive-noise model for parameter estimation. This loss is due to the non-Gaussianity of the noise and vanishes if the noise is Gaussian. A certain drawback in this alternative derivation of the FII is that the necessity of the equality condition does not follow easily, and requires some additional effort. See Proposition 3 in the next Section.

III. Additional Results Regarding the Matrix Form of the FII

For completeness, we review below additional results regarding the matrix form of the FII, some of which appeared elsewhere. We start with a convex-matrix form of the FII (11) which was presented in [15, 12].

Proposition 1 (Convex-matrix form of FII) *Let* $\mathbf{N} = (N_1, \ldots, N_n)$ *be a vector with statistically independent components, and* $\Lambda = \{\lambda_{i,j}\}$ *be an* $m \times n$ *matrix with orthonormal rows, i.e.,* $\Lambda\Lambda^t = I_{m\times m}$. *Then,*

$$trace\Big\{\mathbf{J}(\Lambda\mathbf{N})\Big\} \leq \sum_{i=1}^{m}\sum_{j=1}^{n}\lambda_{i,j}^2 J(N_j). \tag{26}$$

In particular, if $J(N_1) = \ldots = J(N_n) = J_N$, the inequality becomes $\frac{1}{m}\mathrm{trace}\Big\{\mathbf{J}(\Lambda\mathbf{N})\Big\} \leq J_N$, since the orthonormality of the rows implies $\sum_{j=1}^{n}\lambda_{i,j}^2 = 1$ for $i = 1\ldots m$.

Proof: Since $\mathbf{N}$ is an independent vector, the matrix $\mathbf{J}(\mathbf{N})$ is diagonal $\mathbf{J}(\mathbf{N}) = \mathrm{diag}[J(N_1), \ldots, J(N_n)]$. Thus the i,ith component of the matrix $\Lambda\mathbf{J}(\mathbf{N})\Lambda^t$ in the right hand side of (23) is given by $\sum_{j=1}^{n}\lambda_{ij}^2 J(N_j)$. Taking the trace of both sides of (23) results in (26). $\square$

We turn to a convex-matrix version of the Entropy-Power Inequality (EPI) which also appears in [15, 12].

Proposition 2 (Convex-matrix-EPI) *Let* $\mathbf{N}$ *and* Λ *be as in Proposition 1 above. Then*

$$h(\Lambda\mathbf{N}) \geq \sum_{i=1}^{m}\sum_{j=1}^{n}\lambda_{i,j}^2 h(N_j), \tag{27}$$

where $h(\cdot)$ *denotes (joint) differential entropy.*

Proof: The proof is based on the convex-matrix form of the FII in Proposition 1 and the De-Bruijn identity, and follows the line of the proof of Theorem 7 in [8, pp. 1509]. $\square$

From (27) it is easy to obtain the matrix form of the EPI presented in [16]; see also [15, 12].

We end this section with a special case of ineq. (24), regarding a linear transformation of an *independent* vector, for which we can prove a necessary and sufficient condition. The proof can be found in [13].

Proposition 3 (Necessary and Sufficient Condition for Equality in the Matrix-FII for an Independent Vector) *Let* A *be an* $m \times n$ *matrix,* $m \leq n$. *Let* $\mathcal{I}_R(A)$ *denote the set of indices* $j \in \{1\ldots n\}$ *for which* n_j *is uniquely determined by* $A\mathbf{n}$, *and let* $\mathcal{I}_0(A)$ *denote the set of indices of the all-zero columns of* A. *Assume that* $\mathbf{N} \in \mathcal{R}^n$ *is a random vector with independent components, each having finite FI. Then, inequality (24),* $\mathbf{J}(A\mathbf{N})^{-1} \geq A\mathbf{J}(\mathbf{N})^{-1}A^t$, *holds with equality*

if and only if N_j *is Gaussian for all* $j \notin \mathcal{I}_R(A) \cup \mathcal{I}_0(A)$.

The necessary condition of Proposition 3 asserts that the *joint* FI can be used as a "contrast" (or "objective") function for blind de-convolution / signal separation [9, 5].

Clearly, if $m < n$ and A does not have all-zero columns, then there is at least one index j not in $\mathcal{I}_R(A)$. We thus have the following corollary of Proposition 3.

Corollary 2 *If the $m \times n$ matrix A does not have all-zero columns, and $N_1 \dots N_n$ are i.i.d. random variables with $J(N_i) = J_N$, then $J(A\mathbf{N})^{-1} \geq J_N^{-1} AA^t$, with equality if and only if $m = n$ or the N_i's are identical Gaussians.*

IV. Parameter Estimation from Pre-filtered Measurements

Some insight into the problem of parameter estimation in the presence of *non-Gaussian* measurement noise can be gained by investigating the properties of the FI in the model $\mathbf{X} = Q^t\boldsymbol{\theta} + \mathbf{N}$ and $\mathbf{Y} = P\mathbf{X}$; $\boldsymbol{\theta} \in \mathcal{R}^{m_\theta}$; $\mathbf{X}, \mathbf{N} \in \mathcal{R}^n$; $\mathbf{Y} \in \mathcal{R}^{m_y}$ given in (19). In many practical situations a large number of noisy measurements $X_1 \dots X_n$ is taken to estimate a small number of parameters $\theta_1 \dots \theta_m$. For example, the linear relation $\mathbf{X} = Q^t\boldsymbol{\theta} + \mathbf{N}$ may represent a radar application, where $X_1 \dots X_n$ are the outputs of an n-element phased array which observes the targets $\theta_1 \dots \theta_m$, where $n \gg m$; here the entry q_{ij} of the matrix Q is the (complex) gain of the jth element towards the ith target [11]. Another possible application is that of AR parameter estimation, where a long "training" sequence $X_1 \dots X_n$ is modeled concisely by $X_j = \theta_1 X_{j-1} + \ldots + \theta_m X_{j-m} + N_j, \quad j = 1 \dots n$; here $q_{ij} = X_{j-i}$.

In the applications above $\mathbf{Y} = P\mathbf{X}$ represents a prefiltered version of the measurements, from which we wish to estimate $\boldsymbol{\theta}$. The matrix P thus represents a *non invertible* linear operation. For instance, we may perform such a prefiltering operation to reduce the dimensionality and hence the complexity of the (possibly non linear) estimator. See e.g. [11, 2] and the references therein. A case of interest would be when the matrix P is a projection matrix (i.e. $PP^t = I_{m_y \times m_y}$) whose rows span the row space of Q (i.e., $Q = GP$ for some $m \times m$ matrix G). Furthermore, it makes sense to assume that the noise samples $N_1 \dots N_n$ are i.i.d., and that the n columns of Q have a fixed, say, unit norm, i.e., $\sum_{i=1}^{m_\theta} q_{ij}^2 = 1 \ \forall j$. (The latter condition corresponds to a fixed total power gain per each element in the radar phased array application above).

Under these quite natural assumptions, one may wonder how does the quality of the estimation vary with n and with the noise properties, and how much do we loose (do we?) for applying the projection operation P prior to estimation.

In order to isolate the effect of the SNR on the performance, we introduce a noise gain parameter α, and consider the model

$$\mathbf{X} = Q^t\boldsymbol{\theta} + \alpha\mathbf{N} \quad \text{and} \quad \mathbf{Y} = P\mathbf{X}.$$

Using Lemma 5 and Lemma 6, and the identity $QP^tPQ^t = QQ^t$ implied by the model above, it is easy to prove that for any $m_\theta \le m_y \le n$,

$$\mathbf{J}(\mathbf{X};\boldsymbol{\theta}) = \frac{J_N}{\alpha^2}QQ^t \quad \text{and} \quad \mathbf{J}(\mathbf{Y};\boldsymbol{\theta}) = QP^t\mathbf{J}(P\mathbf{N})PQ^t \le \frac{J_N}{\alpha^2}QQ^t \tag{28}$$

where J_N is the FI of the (i.i.d.) noise samples. Furthermore, since $\sum_{i=1}^{m_\theta} q_{ij}^2 = 1$ for $j = 1 \dots n$, it follows that

$$n\frac{\sigma_N^{-2}}{\alpha^2} \le \operatorname{trace}\{\mathbf{J}(\mathbf{Y};\boldsymbol{\theta})\} \le \operatorname{trace}\{\mathbf{J}(\mathbf{X};\boldsymbol{\theta})\} = n\frac{J_N}{\alpha^2} \tag{29}$$

where σ_N^2 is the variance of the noise, and the lower bound follows from (10). Equality holds in both inequalities in (29) if N is Gaussian (in which case $J_N = \sigma_N^{-2}$). Furthermore, for any noise distribution the right inequality holds with equality if $n = m_y = m_\theta$. Otherwise the inequalities are strict; see Corollary 2.

One simple implication of (29) is that without prefiltering the total FI increases *linearly with the number of the measurements.* The same is true even after projection by P if the measurement noise is Gaussian. However, for non-Gaussian noise projection causes a loss of FI whenever $m_y < n$. Thus, from the FI / CRB point of view, the optimal estimator cannot be decomposed into projection followed by non-linear processing.

This phenomena can be explained by the tendency towards Gaussianity of the sum of independent random variables. Projection, which is a non invertible linear transformation, makes the residual noise *more Gaussian* and thus less favorable for estimation. A similar phenomena causes *increase of entropy* after non invertible filtering [9, 16].

In [14] we suggested another way to interpret (29), namely, as an accuracy-quantity tradeoff relation. Notice that $1/\alpha^2$ represents the accuracy (or the resolution) of the measurements. Thus without prefiltering, keeping the quantity / accuracy product n/α^2 fixed keeps the FI constant. The same is true for a Gaussian noise even *after* (appropriate) prefiltering, but not true when the noise is not Gaussian. Thus, if prefiltering (projection) is used prior to estimation in the presence of a non-Gaussian noise, *it is better to take few accurate measurements than many noisy ones.*

Acknowledgments

My joint work with Meir Feder on information theoretic inequalities formed the basis for this paper. I also thank Toby Berger, Jean-Francois Cardoso, Hagit Messer, Nadav Shulman, Yossef Steinberg, Tony Weiss and Arie Yeredor for helpful discussions.

References

[1] A.R. Barron. Entropy and the central limit theorem. *The Annals of Probability*, 14, No. 1:336–342, 1986.

[2] A. Bartov and H. Messer. Lower bound on the achievable DSP performance for localizing step-like continuous signals in noise. *IEEE Trans. Signal Processing*, accepted Aug. 1997.

[3] N. M. Blachman. The convolution inequality for entropy powers. *IEEE Trans. Information Theory*, IT-11:267–271, 1965.

[4] R. E. Blahut. *Principles and Practice of Information Theory.* Addison Wesley, Reading, MA, 1987.

[5] J.-F. Cardoso. Blind signal separation: A review. *Proceedings of the IEEE*, Submitted 1997.

[6] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, New York, 1991.

[7] A. Dembo. Information inequalities and uncertainty principles. Technical Report 75, Department of Statistics, Stanford University, July 1990.

[8] A. Dembo, T.M.Cover, and J.A.Thomas. Information theoretic inequalities. *IEEE Trans. Information Theory*, IT-37:1501–1518, Nov. 1991.

[9] D. Donoho. On minimum entropy deconvolution. *Applied Time Series Analysis II*, pages 565–608, Academic Press, NY, 1981.

[10] H. L. Van Trees. *Detection Estimation and Modulation theory (Part I).* Wiley, New York, 1968.

[11] A. J. Weiss and B. Friedlander. Preprocessing for direction finding with minimal variance degradation. *IEEE Trans. Signal Processing*, pages 1478–1485, June 1994.

[12] R. Zamir. *Universal Encoding of Signals by Entropy Coded Dithered Quantization.* PhD thesis, Tel-Aviv University, 1994.

[13] R. Zamir. A necessary and sufficient condition for equality in the matrix Fisher-information-inequality. Technical report, Tel-Aviv University, Dept. of EE-Systems, 1997.

[14] R. Zamir. It is better to take few accurate measurements than many noisy ones. In *Proc. of IT Workshop on Info. Theory, Multiple Access and Queueing, St.Louis, Missouri*, page 35, April 1995.

[15] R. Zamir and M. Feder. A generalization of information theoretic inequalities to linear transformations of independent vector. *Proceedings of the Sixth Joint Swedish-Russian International Workshop on Information Theory*, pages 254–258, Molle, Sweden, Aug. 1993.

[16] R. Zamir and M. Feder. A generalization of the Entropy Power Inequality with applications. *IEEE Trans. Information Theory*, IT-39:1723–1727, Sept. 1993.

Footnotes

1. An alternative ("Bayesian") way to express the equality condition is that $\Theta - \phi(X) - X$ form a Markov chain for any distribution of the parameter θ.

2. In some references the FI of $\mathbf{N}$ is defined as $K(\mathbf{N}) = \text{trace}\{\mathbf{J}(\mathbf{N})\}$.

3. The scalar form of the De-Bruijn identity is $\frac{d}{dt}h(X + \sqrt{t}Z) = \frac{1}{2}J(X + \sqrt{t}Z)$, where Z is a standard normal variable.

4. As pointed out by A. Yeredor, the proof follows even more easily from the equivalent formula for FI $J(X, Y; \theta) = -E\{\partial^2 \ln f_\theta(X, Y)/\partial\theta^2\}$.