Tech. Report EE-S-97-11: A Necessary and Sufficient Condition for Equality in the Matrix Fisher-Information-Inequality *

Ram Zamir
Dept. of Elect. Eng. - Systems, Tel Aviv University, Tel Aviv 69978, ISRAEL
zamir@eng.tau.ac.il

November 23, 1997

Abstract

A necessary and sufficient condition for equality in the matrix form of the Fisher-Information inequality is derived.

Key Words: Information theoretic inequalities, Fisher-information, non-Gaussian noise, linear filtering.

I. The Fisher Information Inequality

In this technical report we consider the matrix form of the Fisher Information Inequality (FII), which was introduced in [5]. In particular, we provide a necessary and sufficient condition for equality in the matrix-FII.

We first introduce a special form of Fisher Information (FI), called the *FI with respect to a translation parameter*, which does not involve explicitly a parameter. Let $f(\underline{x})$ be the density of

*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 Queueing, April 1995, St. Louis, Missouri.

some random vector $\underline{X}$. The Fisher Information of $\underline{X}$ with respect to a translation parameter is defined as (see e.g. [2, 3])

$$K(\underline{X}) = \mathrm{COV}\left\{\nabla_{\underline{x}} \ln\left(f(\underline{X})\right)\right\} = E\left\{\left(\frac{\nabla_{\underline{x}} f(\underline{X})}{f(\underline{X})}\right)\left(\frac{\nabla_{\underline{x}} f(\underline{X})}{f(\underline{X})}\right)^t\right\};, \tag{1}$$

where $\nabla_{\underline{x}}$ is the gradient vector with respect to $\underline{x}$. If the components of $\underline{X}$ are independent, then the matrix $K(\underline{X})$ is diagonal, with the scalar FI's of the components of $\underline{X}$ on its diagonal. The Fisher Information (1) satisfies the scaling property $K(\alpha\underline{X}) = 1/\alpha^2 K(\underline{X})$, or more generally

$$K(A\underline{X}) = (A^{-1})^t K(\underline{X}) A^{-1} \tag{2}$$

for any non singular square matrix A. A lower bound on $K(\underline{X})$ is provided by the inverse covariance

$$K(\underline{X}) \geq \mathrm{COV}(\underline{X})^{-1}\ , \tag{3}$$

with equality iff $\underline{X}$ is a Gaussian vector. Note that for a Gaussian vector $\underline{X}^*$ we have

$$K(A\underline{X}^*) = \mathrm{COV}(A\underline{X}^*)^{-1} = \left(AK(\underline{X}^*)^{-1}A^t\right)^{-1} \tag{4}$$

for *any* matrix A (not necessarily square).

The special form of Fisher Information defined in (1) satisfies another interesting inequality, called *the Fisher Information Inequality* (FII); see e.g. [2, pp. 494-497] and [3]. We next propose a matrix form of the FII, which was shown recently in [6] and [5]. Given the $m \times n$ matrix A, define $\mathcal{I}_R(A) \subseteq \{1 \ldots n\}$ to be the subset of indices j such that x_j is uniquely determined by $A\underline{x}$. Define $\mathcal{I}_0(A) \subseteq \{1 \ldots n\}$ to be the subset of indices of the all-zero columns of A. If $j \in \mathcal{I}_R(A)$ we say that "x_j is extractable", while if $j \in \mathcal{I}_0(A)$ we say that "x_j is irrelevant". Of course both sets may be empty, while if A is a non-singular square matrix $\mathcal{I}_R(A) = \{1 \ldots n\}$. Finally, given the random vector $\underline{X}$, define $\mathcal{I}_G(\underline{X}) \subseteq \{1 \ldots n\}$ to be the subset of indices j such that X_j is Gaussian.

Theorem 1 ("Matrix FII" - Zamir Feder [6]) *Suppose the components of the random vector $\underline{X} = X_1 \dots X_n$ are statistically independent. Then, for any $m \times n$ matrix A, $m \leq n$, having a full row-rank (i.e., Rank $A = m$),*

$$K(A\underline{X}) \leq K(A\underline{X}^*) = \left(AK(\underline{X})^{-1}A^t\right)^{-1} \tag{5}$$

where $\underline{X}^*$ is a Gaussian vector with independent components such that the variance of X_j^* equals $K(X_j)^{-1}$ for $j = 1 \dots n$ (i.e., $X_1^* \dots X_n^*$ have the same scalar Fisher Informations as $X_1 \dots X_n$). The matrix inequality (5) is in the sense that the difference matrix is non-negative definite. Equality holds if and only if

$$\mathcal{I}_R(A) \bigcup \mathcal{I}_0(A) \bigcup \mathcal{I}_G(\underline{X}) = \{1 \dots n\} \ , \tag{6}$$

i.e., if every X_j is either "extractable" or "irrelevant" or Gaussian.

In particular, equality in (5) holds if $X_1 \dots X_n$ are all Gaussian random variables, in which case

$$K(A\underline{X}) = \mathrm{COV}(A\underline{X})^{-1} = \left(A\mathrm{COV}(\underline{X})A^t\right)^{-1} = \left(AK(\underline{X})^{-1}A^t\right)^{-1} \ .$$

or if $m = n$ (in which case A is invertible so (5) coincides with (2)).

This theorem appeared originally in [5], but without the exact condition (6) for equality.

II. Proof of the Matrix FII

In this appendix we refine the proof of the matrix-FII given in [5], and examine in detail the necessary and sufficient conditions for the inequality to become equality. The entire derivation is similar to that in [4] and [1], where the basic FII was shown.

II..1 A Lemma

We first introduce and prove Lemma 1, which we use later to prove Theorem 1.

Lemma 1 *Let $X_1 \dots X_n$ be independent random variables having differentiable densities, and let $\underline{Y} = A\underline{X}$, where A is an $m \times n$ matrix, $m \leq n$, having a full row-rank (i.e., Rank $A = m$). Let $f_{\underline{x}}(\cdot)$ and $f_{\underline{y}}(\cdot)$ denote the densities of $\underline{X}$ and $\underline{Y}$, respectively, and let $\nabla f_{\underline{x}}$ and $\nabla f_{\underline{y}}$ denote their gradient vectors. Then,*

$$A^t \cdot \frac{\nabla f_{\underline{y}}(\underline{Y})}{f_{\underline{y}}(\underline{Y})} = E\left\{ \frac{\nabla f_{\underline{x}}(\underline{X})}{f_{\underline{x}}(\underline{X})} \,\middle|\, \underline{Y} \right\} \qquad \textit{with probability } 1, \tag{7}$$

where $E\{\cdot|\underline{Y}\}$ denotes conditional expectation given $\underline{Y}$.

Proof: We consider three cases. Assume first $m = n$, in which case A is invertible (since by assumption A has a full row-rank). Thus $f_{\underline{y}}(\underline{y}) = f_{\underline{x}}(A^{-1}\underline{y})/|A|$, where $|A|$ is the absolute determinant of A, which implies $\nabla f_{\underline{y}}(\underline{y}) = (A^{-1})^t \cdot \nabla f_{\underline{x}}(A^{-1}\underline{y})/|A|$, and we have the identity

$$A^t \cdot \frac{\nabla f_{\underline{y}}(\underline{y})}{f_{\underline{y}}(\underline{y})} = \frac{\nabla f_{\underline{x}}(A^{-1}\underline{y})}{f_{\underline{x}}(A^{-1}\underline{y})} \; . \tag{8}$$

Since $\underline{Y} = A\underline{X}$ with probability 1, we have from (8)

$$A^t \cdot \frac{\nabla f_{\underline{y}}(\underline{Y})}{f_{\underline{y}}(\underline{Y})} = \frac{\nabla f_{\underline{x}}(\underline{X})}{f_{\underline{x}}(\underline{X})} \qquad \text{with probability 1,} \tag{9}$$

which implies (7)[1].

Consider now arbitrary $m \le n$. The matrix identity (7) amounts to the set of equalities

$$\underline{b}_j^t \cdot \frac{\nabla f_{\underline{y}}(\underline{Y})}{f_{\underline{y}}(\underline{Y})} = E\left\{ \left. \frac{f'_{x_j}(X_j)}{f_{x_j}(X_j)} \right| \underline{Y} \right\} \qquad j = 1 \ldots n \; , \tag{10}$$

where $\underline{b}_j$ is the j-th column of A, and the conditional expectation is taken over X_j given $\underline{Y} = A\underline{X}$. We need to verify (10) for each j. Define A^{-j} as the matrix A after removing its j-th column, and define $\underline{U}_j = A^{-j}\underline{X} = \underline{Y} - X_j \cdot \underline{b}_j$. Since the components of $\underline{X}$ are statistically independent, X_j and $\underline{U}_j$ are also statistically independent.

Assume first that $\underline{b}_j$ is linearly independent of all the other columns of A, implying that A^{-j} has a deficient rank. (This case includes the case $m = n$ discussed above.) Under this assumption the m-dimensional random vector $\underline{U}_j$ belongs with probability 1 to an $m-1$ dimensional linear space, so it may be expressed as a linear transformation of an $m-1$ dimensional random vector $(V_2, \ldots, V_m)$. Thus, we may find an invertible $m \times m$ matrix Ω such that $\underline{Y} = \Omega\underline{V}$, where $(\underline{V} = V_1, \ldots, V_n)$, the first column of Ω is $\underline{b}_j$, $V_1 = X_j$, and $V_2 \ldots V_m$ are jointly statistically independent of $V_1 = X_j$. Using the identity (8), we then get

$$\Omega^t \frac{\nabla f_{\underline{y}}(\underline{Y})}{f_{\underline{y}}(\underline{Y})} = \frac{\nabla f_{\underline{v}}(\underline{V})}{f_{\underline{v}}(\underline{V})} \qquad \text{with probability 1.} \tag{11}$$

[1] In this case (7) holds pointwise, i.e., even without taking the (conditional) expectation.

The first component of the vector in the left hand side of (11) is $\underline{b}_j^t \cdot \frac{\nabla f_{\underline{y}}(\underline{Y})}{f_{\underline{y}}(\underline{Y})}$, while the first component of the vector in the right hand side of (11) is

$$\frac{\frac{\partial}{\partial V_1} f_{\underline{v}}(\underline{V})}{f_{\underline{v}}(\underline{V})} = \frac{\frac{\partial}{\partial X_j} f(X_j) f(V_2, \ldots, V_m)}{f(X_j) f(V_2, \ldots, V_m)} = \frac{f'(X_j)}{f(X_j)} \ , \tag{12}$$

and (10) follows for the j-th component (again pointwise, like in the previous case).

Lastly, assume that $\underline{b}_j$ depends linearly on the other columns of A. In this case A^{-j} has a full row-rank so $\underline{U}_j$ must have a density. Since $\underline{Y}$ is the independent sum of $\underline{U}_j$ and $X_j \underline{b}_j$, we can write $f_{\underline{y}}$ as the convolution

$$f_{\underline{y}}(\underline{y}) = \int f_{x_j}(x) f_{\underline{u}_j}(\underline{y} - x\underline{b}_j) dx \ . \tag{13}$$

Note that

$$\frac{d}{dx} f_{\underline{u}_j}(\underline{y} - x\underline{b}_j) = -\underline{b}_j^t \cdot \nabla f_{\underline{u}_j}(\underline{y} - x\underline{b}_j) \ . \tag{14}$$

Taking the gradient of $f_{\underline{y}}$ with respect to $\underline{y}$ by changing the order of differentiation and integration in (13), and using (14), we obtain

$$\underline{b}_j^t \cdot \frac{\nabla f_{\underline{y}}(\underline{y})}{f_{\underline{y}}(\underline{y})} = -\frac{1}{f_{\underline{y}}(\underline{y})} \int f_{x_j}(x) \frac{d}{dx} f_{\underline{u}_j}(\underline{y} - x\underline{b}_j) dx \ . \tag{15}$$

Finally, integrating (15) in parts, and noting that $f_{x_j}(-\infty|\underline{y}) = f_{x_j}(+\infty|\underline{y}) = 0$, we obtain

$$\begin{aligned} \underline{b}_j^t \cdot \frac{\nabla f_{\underline{y}}(\underline{y})}{f_{\underline{y}}(\underline{y})} = \frac{1}{f_{\underline{y}}(\underline{y})} \int f'_{x_j}(x) f_{\underline{u}_j}(\underline{y} - x\underline{b}_j) dx \ &= \int \frac{f_{x_j}(x) f_{\underline{u}_j}(\underline{y} - x\underline{b}_j)}{f_{\underline{y}}(\underline{y})} \frac{f'_{x_j}(x)}{f_{x_j}(x)} dx \\ &= \int f_{x_j|\underline{y}}(x, \underline{y}) \frac{f'_{x_j}(x)}{f_{x_j}(x)} dx \ , \end{aligned} \tag{16}$$

which shows (10), and the proof of the lemma is completed. □

II..2 Proof of the Inequality

Define the $m \times n$ matrix

$$T = (AK(\underline{X})^{-1} A^t)^{-1} AK(\underline{X})^{-1} \ , \tag{17}$$

and set $\underline{Y} = A\underline{X}$. Multiplying (7) by T, and noting that $TA^t = I_m$, we get

$$\frac{\nabla f_{\underline{y}}(\underline{Y})}{f_{\underline{y}}(\underline{Y})} = E\left\{T \cdot \frac{\nabla f_{\underline{x}}(\underline{X})}{f_{\underline{x}}(\underline{X})} \;\middle|\; \underline{Y}\right\} \quad \text{with probability 1.} \tag{18}$$

Using the Cauchy-Schwartz inequality $EWEW^t \leq EWW^t$, it follows from (18) that

$$\begin{aligned} \left(\frac{\nabla f(\underline{Y})}{f(\underline{Y})}\right)\left(\frac{\nabla f(\underline{Y})}{f(\underline{Y})}\right)^t &= E\left\{T\frac{\nabla f(\underline{X})}{f(\underline{X})}\middle|\underline{Y}\right\} E\left\{\frac{\nabla f(\underline{X})^t}{f(\underline{X})}T^t\middle|\underline{Y}\right\} \\ &\leq E\left\{T\left(\frac{\nabla f(\underline{X})}{f(\underline{X})}\right)\left(\frac{\nabla f(\underline{X})}{f(\underline{X})}\right)^t T^t\middle|\underline{Y}\right\}, \end{aligned} \tag{19}$$

with equality if and only if

$$\frac{\nabla f_{\underline{y}}(\underline{Y})}{f_{\underline{y}}(\underline{Y})} = T \cdot \frac{\nabla f_{\underline{x}}(\underline{X})}{f_{\underline{x}}(\underline{X})}, \quad \text{with probability 1.} \tag{20}$$

Using definition (1) and averaging (19) over $\underline{Y}$ gives $K(\underline{Y}) \leq TK(\underline{X})T^t$, which by the definition of T above is the desired inequality (5).

II..3 Proof of the Equality Condition

We turn to show that (6) is a sufficient and necessary condition for equality in (5). To clarify the proof we divide it into a few steps.

Equivalence principle: Let Q and F be arbitrary $m \times m$ and $n \times n$ non singular matrices, respectively. Define $B = QAF$, and $\underline{\widetilde{X}} = F^{-1}\underline{X}$. By the scaling property of Fisher information with respect to invertible linear transformation (2), we have $K(B\underline{\widetilde{X}}) = (Q^{-1})^t K(A\underline{X})Q^{-1}$, and $(BK(\underline{\widetilde{X}})^{-1}B^t)^{-1} = (Q^{-1})^t(AK(\underline{X})^{-1}A^t)^{-1}Q^{-1}$. Since Q is invertible and in light of (5), these identities mean that equality in the FII for $A\underline{X}$ implies equality in the FII for $B\underline{\widetilde{X}}$, and vice versa. Assume in addition that the matrix F results from row (or column) permutations of some non singular diagonal matrix. Then for some permutation π of the index set $\{1, \ldots, n\}$, we have $\mathcal{I}_R(B) = \pi\mathcal{I}_R(A)$, $\mathcal{I}_0(B) = \pi\mathcal{I}_0(A)$, and $\mathcal{I}_G(\underline{\widetilde{X}}) = \pi\mathcal{I}_G(\underline{X})$, where the subsets $\mathcal{I}_R(\cdot)$, $\mathcal{I}_0(\cdot)$ and $\mathcal{I}_G(\cdot)$ are defined before Theorem 1. Thus, condition (6) holds for $A\underline{X}$ if and only if it holds for $B\underline{\widetilde{X}}$.

We conclude that without loss of generality we may check conditions for equality in the matrix FII with respect to $B\underline{\widetilde{X}}$ instead of $A\underline{X}$.

Canonical form I: Let $|\mathcal{I}_R(A)| = n_R$, $|\mathcal{I}_0(A)| = n_0$ and $n_g = n - n_R - n_0$. We can rearrange the the columns of A and correspondingly permute the vector $\underline{X}$ so that $\mathcal{I}_R(A) = \{n_g+1, \ldots, n_g+n_R\}$ and $\mathcal{I}_0(A) = \{n_g + n_R + 1, \ldots, n\}$. This rearrangement is equivalent to multiplying the matrix A from the right by some $n \times n$ permutation matrix F'. Also, by the definition of the subset $\mathcal{I}_R(A)$ it follows that there exists an $m \times m$ matrix R such that $(RA\underline{x})_j = x_j$ for $j \in \mathcal{I}_R(A)$. Thus, by the equivalence principle above (setting $Q = R$ and $F = F'$), we may assume without loss of generality that the $m \times n$ matrix A has the following form

$$\begin{pmatrix} P & | & 0 & | & \\ - & | & - & | & 0 \\ 0 & | & I_{n_R} & | & \end{pmatrix}, \tag{21}$$

where I_{n_R} is the $n_R \times n_R$ identity matrix, and P is an $(m - n_R) \times n_g$ matrix whose subsets $\mathcal{I}_R(P)$ and $\mathcal{I}_0(P)$ are *empty.* Note that the matrix R is also responsible for zeroing the block above the submatrix I_{n_R} via row operations.

Sufficiency: We are now in a position to show that condition (6) is indeed sufficient for equality in the matrix FII. Define $(X_1, \ldots, X_{n_g})^t = \underline{X}_G$, and $(X_{n_g+1}, \ldots, X_{n_g+n_R})^t = \underline{X}_R$. By the assumed canonical form (21) of the matrix A, we have $A\underline{X} = ((P\underline{X}_G)^t, X_{n_g+1}, \ldots, X_{n_g+n_R})^t$. Thus, since $\underline{X}_G$ is independent of $\underline{X}_R$, the Fisher information matrix of $A\underline{X}$ has a block diagonal form $K(A\underline{X}) = \mathrm{diag}[K(P\underline{X}_G), K(\underline{X}_R)]$. On the other hand

$$(AK(\underline{X})^{-1}A^t)^{-1} = \mathrm{diag}[PK(\underline{X}_G)^{-1}P^t, K(\underline{X}_R)^{-1}]^{-1} \,. \tag{22}$$

But since by condition (6) the subvector $\underline{X}_G$ is Gaussian, we have from (4) $K(P\underline{X}_G) = (PK(\underline{X}_G)^{-1}P^t)^{-1}$, so $K(A\underline{X})$ and $(AK(\underline{X})^{-1}A^t)^{-1}$ are equal as desired.

Canonical form II: Define $\underline{\widetilde{X}} = K(\underline{X})^{1/2} \cdot \underline{X}$. By (2), $K(\underline{\widetilde{X}}) = I_n$. Since A has a full row rank, there exists an $m \times m$ matrix G such that the matrix $G \cdot A \cdot K(\underline{X})^{-1/2}$ is orthonormal. Following the equivalence principle above (setting $Q = G$ and $F = K(\underline{X})^{-1/2}$), we can assume without loss of generality that the matrix A is orthonormal, i.e., $AA^t = I_m$, and that $K(\underline{X}) = I_n$.

Necessity: We are now in a position to prove that condition (6) is necessary for equality in the matrix FII. In view of the canonical form (21), we need to show that if equality holds, the components of $\underline{X}_G = (X_1, \ldots, X_{n_g})^t$ must be Gaussian and (since $K(\underline{X}) = I_n$) have unit variance. Assuming $AA^t = I_m$ and $K(\underline{X}) = I_n$, the matrix T defined in (17) is equal to A. Condition (20) for equality in the matrix-FII becomes

$$A \cdot \frac{\nabla f_{\underline{x}}}{f_{\underline{x}}}(\underline{X}) = \frac{\nabla f_{\underline{y}}}{f_{\underline{y}}}(\underline{Y}) \qquad \text{with probability 1,} \tag{23}$$

where for convenience we regard here $\frac{\nabla f}{f}$ as a single function. In particular, in view of (21) it is necessary that

$$P \cdot \frac{\nabla f_{\underline{x}_G}}{f_{\underline{x}_G}}(\underline{X}_G) = \frac{\nabla f_{\underline{y}_G}}{f_{\underline{y}_G}}(P\underline{X}_G) \qquad \text{with probability 1,} \tag{24}$$

where $f_{\underline{x}_G}$ and $f_{\underline{y}_G}$ are the density functions of $\underline{X}_G$ and $P\underline{X}_G$, respectively. Note that $\frac{\nabla f_{\underline{x}_G}}{f_{\underline{x}_G}}(\underline{x}) = (\frac{f'_{x_1}}{f_{x_1}}(x_1), \ldots, \frac{f'_{x n_g}}{f_{x n_g}}(x_{n_g}))^t$ and that $\frac{\nabla f_{\underline{y}_G}}{f_{\underline{y}_G}}(\cdot)$ is an m-dimensional function of an m-vector. Condition (24) means that for every vector $(x_1, \ldots, x_{n_g})$ in the support of $f_{\underline{x}_G}$,

$$P \cdot \begin{pmatrix} \mathrm{func}_1(x_1) \\ \vdots \\ \mathrm{func}_{n_g}(x_{n_g}) \end{pmatrix} = \mathrm{func}_0 \left\{ P \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_{n_g} \end{pmatrix} \right\} \tag{25}$$

for some functions $\mathrm{func}_0(\cdot), \ldots, \mathrm{func}_{n_g}(\cdot)$. But since the subsets $\mathcal{I}_R(P)$ and $\mathcal{I}_0(P)$ are empty, this is possible only if all these functions are <u>linear</u>, and satisfy

$$\mathrm{func}_j(x) = \lambda x \qquad \text{for } j = 0 \ldots n_g \text{ for some constant } \lambda. \tag{26}$$

Thus

$$\frac{f'_{x_j}}{f_{x_j}}(x) = \frac{d \ln f_{x_j}(x)}{dx} = \lambda x \qquad \text{for } j = 1 \ldots n_g. \tag{27}$$

Equation (27) implies that $X_1 \ldots X_{n_g}$ are Gaussian with equal variances, proving the necessity of condition (6). This also completes the proof of the Theorem. □

References

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

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

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

[4] A. J. Stam. Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Inform. Control*, 2:101–112, June 1959.

[5] 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.

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