It Is Better to Take Few Accurate Measurements rather than Many Noisy Ones *

Ram Zamir

330 E&TC Building, Cornell University, Ithaca, NY 14853 . e-mail: zamir@ee.cornell.edu

April 1995

Abstract

Linear pre-filtering (projection) of the measurement space is often used in parameter estimation to reduce the dimensionality, and hence the complexity, of the (generally non-linear) processor. We examine the tradeoff between the number and the accuracy of the measurements, as reflected by the Fisher Information after the prefilter. We observe the following phenomena. Taking twice as much but half as accurate measurements does not preserve the Fisher information after the prefiletr, unless the measurement noise is Gaussian. Thus, when the processor dimension is fixed and the noise is not Gaussian, it is better to take few accurate measurements rather than many noisy ones.

Key Words: Cramer-Rao Bound, non-Gaussian noise, linear modeling, pre-filtering.

*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 the Information Theory Workshop on Multiple Access and Queueing, April 1995, St. Louis, Missouri.

I. Introduction

In many situations a large number of noisy measurements is used to estimate a small number of parameters. Examples may be found in array processing, in linear modeling (e.g., equalization for multi-path channels, or linear prediction for speech coding), and in digital-to-analog conversion of oversampled data. In certain applications, the number of measurements is used to trade for their accuracy in order to achieve a desirable estimation error. Suppose the "cost" of a measurement is some function of its accuracy. The system designer may wish, then, to find an optimal operation point in terms of the amount and the accuracy of the measurements.

In the case where the measurement noise is Gaussian, the solution to this problem is very simple and intuitive, due to the linear structure of the optimal estimator. However, in non-Gaussian noise cases the situation is less clear, since often in practice linear pre-filtering precedes the (possibly non linear) estimator in order to reduce its dimensionality and hence its complexity [8].

In this paper we are concerned with the effect of linear pre-filtering, and with the question of the optimal tradeoff between the amount and the accuracy of measurements in parameter estimation. Section II presents the linear additive non-Gaussian noise model which we analyze. In Section III we present our main result which is an explicit upper bound on the Fisher information matrix (or via the Cramer Rao bound, a lower bound on the mean squared estimation error). The important property of this bound is that it provides insight to the quantity-quality tradeoff discussed above, while making a distinction between the Gaussian and the non-Gaussian noise cases. In Section IV we prove our main result, using a matrix form of the Fisher Information Inequality (FII) which was recently presented in [10]. The last section provides an example that illustrates the quantity-quality tradeoff in a case where the Cramer Rao bound is tight. A detailed proof of the matrix-FII is given in the appendix.

We note that the degradation of the *marginal* Fisher information due to linear projection has already been observed before (see, e.g., [5] and the literature on blind deconvolution). In this respect, our work extends the scope to the multi parameter case.

II. The Statistical Model

Suppose we need to estimate a vector of unknown (nonrandom) real parameters $\underline{\theta} = \theta_1 \ldots \theta_m$ by observing

$$\underline{Y} = H \cdot \underline{\theta} + \alpha \cdot \underline{N} \ , \tag{1}$$

where H is a known $n \times m$ matrix, α is a known scalar parameter, and $\underline{N} = N_1 \ldots N_n$ is a vector of independent noises. Each measurement, $Y_j, j = 1 \ldots n$, is thus some linear combination of the parameters, corrupted by its own independent noise. Assume that $n \geq m$ and that Rank $H = m$, so that for noiseless measurements equation (1) is invertible.

In order to give a physical insight to this problem, one may interpret $\underline{Y}$ as the output of an array of n sensors in the presence of m targets. By this interpretation, the m entries in the j-th row of H are the gains of the j-th sensor with respect to the targets $\theta_1 \ldots \theta_m$, the m columns of H are the virtual beams of the array towards the m targets, and the parameter α controls the overall signal-to-noise ratio, or the accuracy, of the array.

The processor we consider in this paper has the special structure shown in Figure 1. First, the n measurements are projected onto an $\tilde{m}$-space, $m \leq \tilde{m} \leq n$, by means of linear transformation. Then a (possibly non-linear) processor is applied to obtain an unbiased estimate $\hat{\underline{\theta}}$ of the parameters. The input of the non-linear processor is thus

$$\widetilde{\underline{Y}} = P \cdot (H\underline{\theta} + \alpha\underline{N}) \ , \tag{2}$$

where P is an $\tilde{m} \times n$ orthonormal matrix, i.e., $PP^t = I_{\tilde{m}}$, where $I_{\tilde{m}}$ is the $\tilde{m} \times \tilde{m}$ identity matrix. Note that any more general pre-filter can be rewritten in this form by suitably modifying P and H. We further assume that Rank $PH = m$, i.e., the rows of P span the columns space of H,

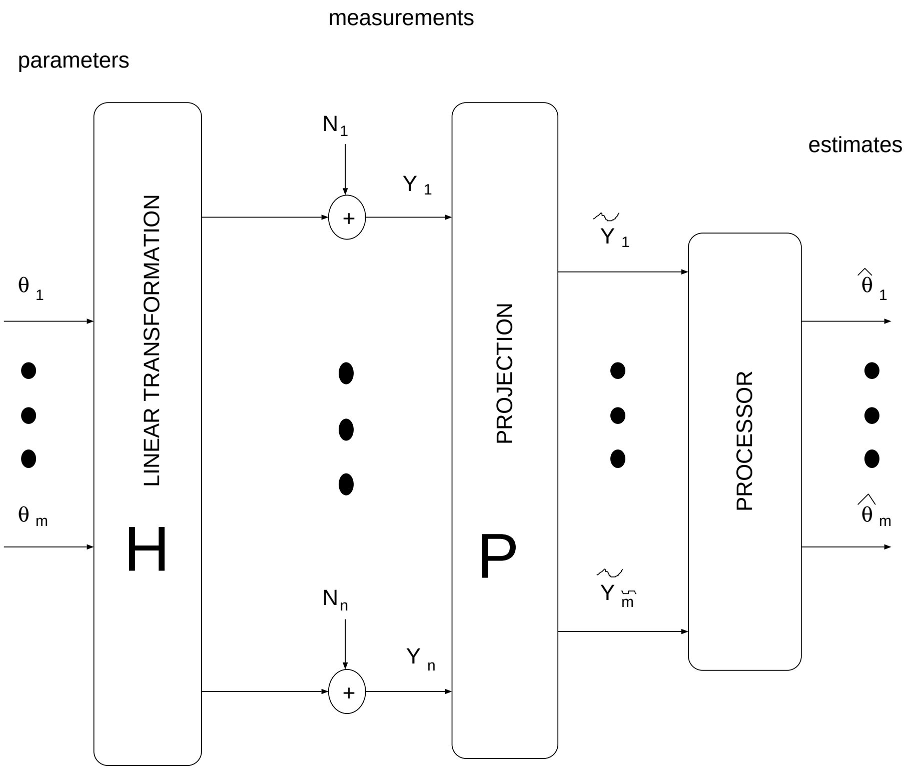

Figure 1: The statistical model of the measurements and the structure of the estimator.

so that (2) is invertible when the measurements are noiseless. The two extreme situations of this processor are $\tilde{m} = n$ (e.g., when $P = I_n$), where the non-linear processor has direct access to all the n measurements, and $\tilde{m} = m$, where the dimension of the measurements space is reduced to the number of parameters, i.e., to its minimal possible value.

We are interested in investigating the quality of the estimate as a function of the number of measurements, n, and the dimensionality of the estimator, $\tilde{m}$, when the number of parameters, m, is fixed. For that we need to impose two additional "physical" constraints. First, we assume that

$$N_1 \ldots N_n \text{ are i.i.d. random variables.} \tag{3}$$

Second, we assume that all rows of H have a unit square norm, i.e.,

$$\sum_{j=1}^{m} h_{i,j}^2 = 1 \ , \quad \text{for } i = 1 \ldots n \ . \tag{4}$$

Continuing our interpretation above, these two conditions means that the accuracy and the total gain of each one of the sensors are the same. Furthermore, enlarging the array is equivalent simply to adding rows of unit norm to H and extending respectively the i.i.d. noise vector $\underline{N}$.

III. The Main Result

The processor wishes to minimize the estimation error covariance matrix

$$\mathrm{COV}(\hat{\underline{\theta}} - \underline{\theta}) = E\left\{(\hat{\underline{\theta}} - \underline{\theta}) \cdot (\hat{\underline{\theta}} - \underline{\theta})^t\right\} \ . \tag{5}$$

A useful tool for assessing the performance limit of any unbiased estimator[1] which has access to $\underline{\widetilde{Y}}$, is the Cramer-Rao Lower Bound (CRB), [7], which states that

$$\mathrm{COV}(\hat{\underline{\theta}} - \underline{\theta}) \geq J(\underline{\theta})^{-1} \ , \tag{6}$$

where the inequality between the two matrices means that the difference matrix is nonnegative definite. The $m \times m$ matrix $J(\underline{\theta})$ is the Fisher Information matrix

$$J(\underline{\theta}) = \mathrm{COV}\left\{\nabla_{\underline{\theta}} \ln\left(f_{\underline{\tilde{y}}}(\underline{\widetilde{Y}} \ ; \ \underline{\theta})\right)\right\} \ , \tag{7}$$

where $\mathrm{COV}(\cdot)$ denotes covariance matrix as in (5), $\ln(\cdot)$ is the natural logarithm, $f_{\underline{\tilde{y}}}(\underline{\widetilde{Y}} \ ; \ \underline{\theta})$ is the probability density function of $\underline{\widetilde{Y}}$ for some parameter vector $\underline{\theta}$, and

$$\nabla_{\underline{\theta}} = \left(\frac{\partial}{\partial\theta_1}, \ldots, \frac{\partial}{\partial\theta_m}\right)$$

is the gradient vector with respect to the parameters. Note that due to the simple linear additive noise model (2), $J(\underline{\theta})$ is in our case independent of the value of $\underline{\theta}$.

For a Gaussian noise $N \sim \mathcal{N}(0, \sigma_N^2)$, the maximum likelihood (ML) estimate

$$\hat{\underline{\theta}}_{ML}(\underline{\widetilde{Y}}) = \arg\max_{\underline{\theta}} f_{\underline{\tilde{y}}}(\underline{\widetilde{Y}}; \underline{\theta}) \tag{8}$$

[1] An unbiased estimator satisfies $E\{\hat{\underline{\theta}}\} = \underline{\theta}$, for all $\underline{\theta}$ in some neighborhood of the true parameter.

is unbiased and *efficient*, i.e., it satisfies the CRB (6) with equality; see [7, p. 68]. More specifically, in the Gaussian case $\underline{\tilde{Y}}$ is a multivariate normal vector, with mean $PH \cdot \underline{\theta}$, and covariance $\alpha^2\sigma_N^2 \cdot I_{\tilde{m}}$, and so

$$J(\underline{\theta}) = \frac{H^t H}{\alpha^2 \sigma_N^2} \tag{9}$$

(see Section IV). Also, since we may always write

$$PH = \tilde{P}^t G \ ,$$

where $\tilde{P}$ is an $m \times \tilde{m}$ orthonormal matrix (i.e., $\tilde{P}\tilde{P}^t = I_m$) and G is a non-singular $m \times m$ matrix, the ML estimate (8) may be written in the Gaussian case in a linear form, as

$$\hat{\underline{\theta}}_{ML} = G^{-1}\tilde{P} \cdot \underline{\tilde{Y}} \ . \tag{10}$$

Thus, the ML estimate has a Gaussian distribution, with mean $\underline{\theta}$ and covariance

$$\text{COV}(\hat{\underline{\theta}}_{ML} - \underline{\theta}) = J(\underline{\theta})^{-1} = \alpha^2\sigma_N^2 \cdot (H^t H)^{-1} \ . \tag{11}$$

As for the non-Gaussian noise case, let $\sigma_N^2 = \text{VAR}(N)$ denote the variance of N, and let, [3, pp. 494-497],

$$J_N = \text{VAR}\left\{\frac{d}{dN}\ln(f(N))\right\} = \int \frac{1}{f(n)}\left(\frac{df(n)}{dn}\right)^2 dn \tag{12}$$

denote the (scalar) *Fisher Information with respect to a translation parameter* of the measurement noise N_i, where $f(n)$ is the density function of N_i (which is identical for $N_1 \ldots N_n$). Note that $J_N \geq \sigma_N^{-2}$, with equality if N is Gaussian [1].

Theorem 1 *Assume the model defined in (1), (2) and (3). Then, for any $m \leq \tilde{m} \leq n$,*

$$\frac{\sigma_N^{-2}}{\alpha^2} \cdot H^t H \leq J(\underline{\theta}) \leq \frac{J_N}{\alpha^2} \cdot H^t H \tag{13}$$

where the inequality between the matrices is in the sense of (6). The upper bound is tight if $\tilde{m} = n$, and both inequalities are tight (for any $\tilde{m}$) if the noises are Gaussian. Furthermore, if we also assume (4), we have

$$n \cdot \frac{\sigma_N^{-2}}{\alpha^2} \leq \mathit{trace}\{J(\underline{\theta})\} \leq n \cdot \frac{J_N}{\alpha^2} \ , \tag{14}$$

where trace$\{\cdot\}$ denotes diagonal element sum, and equality holds under the same conditions as in (13).

The proof is given in Section IV. The main contribution of Theorem 1 is in the upper bound on the Fisher Information matrix, which is based on a matrix form of the Fisher Information Inequality (FII) given in Section IV. The matrix-FII was stated originally in [10] and proved partially in [9]. For completeness we provide its full proof in the appendix.

Compare the upper and lower bounds in (13) with the Fisher information in the Gaussian noise case (9). We observe that non Gaussian measurements provide us with more information than Gaussian measurements having the same variance, but with *less* information than Gaussian measurements having the same Fisher information.

We now make some additional interesting observations from the second part of Theorem 1, which bounds the Fisher Information sum[2]. Consider first the case $\tilde{m} = n$ in which by Theorem 1 the upper bound in (14) is tight, i.e.,

$$\text{trace}\left\{J(\underline{\theta})\right\} = J_N \cdot \frac{n}{\alpha^2} \ . \tag{15}$$

Recall that α^2 is a measure for the accuracy of the measurements. The relation (15) implies that the Fisher Information sum is constant as long as the ratio n/α^2 is kept constant. In other words, when the non linear processor has direct access to all the n measurements there is *a simple tradeoff between the number of measurements and the accuracy of each measurement, which keeps the Fisher Information sum fixed.* Consider now the case where $\tilde{m} < n$. It follows from Theorem 1 that for non Gaussian noise the Fisher Information sum is in this case *less* than $J_N \cdot \frac{n}{\alpha^2}$. Namely, when the measurement space is projected into a smaller sub-space prior to the non-linear processor the Fisher Information sum decreases (unless the measurement noises are Gaussian). Thus, if we have the choice of taking as many measurements as we want while keeping $n/\alpha^2 =$ constant, but in the same time the dimension of the processor input must be kept as low as possible, i.e., $\tilde{m} = m$, $\forall n$,

[2] The diagonal elements of the Fisher Information matrix may not be uniform (in terms of our array processing interpretation, the array may have strong beams pointing to some of the targets, and weak beams pointing to others), and so it is more meaningful to consider their sum or their average.

then *it is better to take few accurate measurements rather than many noisy ones.*

Theorem 1 also provides insight into the structure of the optimal estimator. In the Gaussian case the optimal unbiased estimator (10) is linear, composed of projection of the measurements vector on the parameters space and linear transformation. In the non-Gaussian case Theorem 1 tells us that the optimal estimator *cannot be decomposed into projection followed by some (non linear) operation.* This may be explained by the fact that projection, which is a non invertible linear transformation, makes the additive noise *more Gaussian* (see [10]) and thus less favorable for the estimator. This phenomena is illustrated in the example in the last section.

The second part of Theorem 1 is stated with respect to the Fisher Information sum rather than in terms of the CRB (6) (i.e., the inverse Fisher Information matrix) which directly lower bounds the estimation errors $E(\hat{\theta}_i - \theta_i)^2, i = 1 \ldots m$. Nevertheless, if the columns of H are orthogonal and have the same square norm, then by (4),

$$H^t H = \frac{n}{m} \cdot I_m \,. \tag{16}$$

In this case each of the diagonal elements of the inverse Fisher Information matrix satisfies

$$\frac{m}{n} \cdot \alpha^2 \sigma_N^2 \geq J(\underline{\theta})^{-1}_{i,i} \geq \frac{m}{n} \cdot \frac{\alpha^2}{J_N} \,, \quad i = 1 \ldots m \,. \tag{17}$$

Equality in the lower bound in (17) holds if $\tilde{m} = n$, and both inequalities are tight (for any $\tilde{m}$) if the noises are Gaussian.

In the array processing interpretation, when (16) is satisfied we say that the array has "orthogonal beams having equal gains". Paradoxically, in order to satisfy (16) in dynamic situations (e.g., in order to ensure high resolution between moving targets), an array should have much more sensors than targets. Thus, in light of (17) and contrary to our conclusion following (15), it might be desirable to have $n \gg m$.

IV. Derivation of Results

We first introduce a special form of Fisher Information which does not involve explicitly a parameter, and is useful for our purpose. Let $f(\underline{x})$ be the density of some random vector $\underline{X}$. The *Fisher Information of $\underline{X}$ with respect to a translation parameter* is defined as

$$K(\underline{X}) = \mathrm{COV}\left\{\nabla_{\underline{x}} \ln\left(f(\underline{X})\right)\right\} = E\left\{\left(\frac{\nabla_{\underline{x}} f(\underline{X})}{f(\underline{X})}\right)\left(\frac{\nabla_{\underline{x}} f(\underline{X})}{f(\underline{X})}\right)^t\right\}; , \tag{18}$$

where $\nabla_{\underline{x}}$ is the gradient vector with respect to $\underline{x}$. Note that the quantity J_N defined in (12) is the scalar form of (18). 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 (18) satisfies the scaling property $K(\alpha\underline{X}) = 1/\alpha^2 K(\underline{X})$, or more generally

$$K(A\underline{X}) = (A^{-1})^t K(\underline{X}) A^{-1} \tag{19}$$

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

$$K(\underline{X}) \geq \mathrm{COV}(\underline{X})^{-1} , \tag{20}$$

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

$$K(A\underline{X}^*) = \mathrm{COV}(A\underline{X}^*)^{-1} = \left(AK(\underline{X}^*)^{-1}A^t\right)^{-1} \tag{21}$$

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

The special form of Fisher Information defined in (18) satisfies another interesting inequality, called *the Fisher Information Inequality* (FII); see e.g. [3, pp. 494-497] and [4]. We next propose a matrix form of the FII, which was shown recently in [10] and [9]. Given the $m \times n$ matrix A, define $\mathcal{I}_R(A) \subseteq \{1 \dots n\}$ to be the subset of indices j such that x_j is uniquely determined by $A\underline{x}$, and define $\mathcal{I}_0(A) \subseteq \{1 \dots n\}$ 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 $\mathcal{I}_R(A) = \{1 \dots n\}$. Finally, given the random vector $\underline{X}$, define $\mathcal{I}_G(\underline{X}) \subseteq \{1 \dots n\}$ to be the subset of indices j such that X_j is Gaussian.

Theorem 2 ("Matrix FII" - Zamir Feder [10]) *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 $A = m$),*

$$K(A\underline{X}) \leq K(A\underline{X}^*) = \left(AK(\underline{X})^{-1}A^t\right)^{-1} \tag{22}$$

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 (12) as $X_1 \dots X_n$). The matrix inequality (22) is in the sense of (6). Equality holds if and only if

$$\mathcal{I}_R(A) \bigcup \mathcal{I}_0(A) \bigcup \mathcal{I}_G(\underline{X}) = \{1 \dots n\} \, , \tag{23}$$

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

The proof which originally appeared in [9] without the exact condition (23) for equality, is given for completeness in the appendix. In particular, equality in (22) holds if $X_1 \dots X_n$ are all Gaussian random variables, or if $m = n$ (in which case A is invertible so (22) coincides with (19)). We may turn now to prove Theorem 1.

Proof of Theorem 1: Following the simple linear additive noise model (2), we obtain

$$\nabla_{\underline{\theta}} \ln \left(f_{\tilde{\underline{y}}}(\underline{x} \, ; \, \underline{\theta})\right) = (PH)^t \cdot \nabla_{\underline{x}} \ln \left(f_{\alpha P \underline{N}}(\underline{x} - PH\underline{\theta})\right) \, , \tag{24}$$

where $f_{\alpha P \underline{N}}(\cdot)$ is the density of the random vector $\alpha P \underline{N}$ (which exists since P has a full row rank). Using definitions (7) and (18), and the scaling property of $K(\cdot)$, we then obtain

$$J(\underline{\theta}) = \frac{1}{\alpha^2}(PH)^t \cdot K(P\underline{N}) \cdot (PH) \, . \tag{25}$$

Note that (25) holds for any joint probability density function of the measurement noise vector.

To show the lower bound in (13), we use (20) to obtain

$$J(\underline{\theta}) \geq \frac{1}{\alpha^2}(PH)^t \cdot \mathrm{COV}(P\underline{N})^{-1} \cdot (PH) \, . \tag{26}$$

Since $N_1 \ldots N_n$ are i.i.d., and since $PP^t = I_{\tilde{m}}$, we have $\mathrm{COV}(P\underline{N}) = \sigma_N^2 \cdot I_{\tilde{m}}$, so

$$J(\underline{\theta}) \geq \frac{H^t P^t P H}{\alpha^2 \sigma_N^2} = \frac{H^t H}{\alpha^2 \sigma_N^2} \ , \tag{27}$$

where the second equality above follows since the (orthonormal) rows of P span the columns space of H (see the discussion following (2)), so $P^t P$ is virtually a unit matrix with respect to the column space of H.

As for the upper bound in (13), we use the fact that $N_1 \ldots N_n$ are independent, and apply the matrix form of the FII (22), to obtain from (25)

$$J(\underline{\theta}) \leq \frac{1}{\alpha^2} \cdot H^t P^t \left(P K(\underline{N})^{-1} P^t \right)^{-1} P H = \frac{J_N}{\alpha^2} \cdot H^t H \ , \tag{28}$$

where for the second equality we substituted $K(\underline{N}) = J_N \cdot I_n$ and $PP^t = I_{\tilde{m}}$, and we used the same argument used in (27) by which $H^t P^t P H = H^t H$.

In the special case $\tilde{m} = n$ we have $P^{-1} = P^t$, so from (19), $K(P\underline{N}) = J_N \cdot I_n$. Substituting in (25) we get $J(\underline{\theta}) = J_N/\alpha^2 \cdot H^t H$, which implies that the upper bound is tight in this case. In the Gaussian case, $J_N = \sigma_N^{-2}$, resulting equality.

To obtain (14), we take the trace of (13) (which preserves the inequality), and we utilize assumption (4) which implies $\mathrm{trace}\{H^t H\} = \sum_{i,j} h_{i,j}^2 = n$. □

V. Example

To illustrate the phenomena predicted by Theorem 1, we present in this section a simple example of estimating a parameter corrupted by a “very un-Gaussian” noise. We show the loss due to pre-projection of the measurements by examining the Fisher Information and the actual minimum achievable MSE in this example. We assume a doubly modal Gaussian noise, distributed as

$$N \sim \frac{1}{2} \cdot \mathcal{N}(-\Delta, \delta^2) + \frac{1}{2} \cdot \mathcal{N}(\Delta, \delta^2) \ , \tag{29}$$

i.e., N has a density $f_N(x) = \frac{1}{2\delta}\{\phi(x/\delta+\Delta)+\phi(x/\delta-\Delta)\}$, where $\phi(x) = \frac{1}{\sqrt{2\pi}}\exp(-x^2/2)$ denotes the standard normal distribution. To make N indeed "very un-Gaussian", assume that $\Delta \gg \delta$. We consider estimating a single parameter θ (i.e., $m = 1$), given the measurement vector

$$Y_i = \theta + \alpha N_i \ , \quad i = 1 \ldots n \ ,$$

where the N_i's are independent random variables, identically distributed according to (29), and for convenience we set $\alpha = 1$.

Our aim is to compare the performance of the estimator $\hat{\theta}(Y_1 \ldots Y_n)$ which has access to all the measurements, with that of the estimator $\hat{\theta}(\widetilde{Y})$ which receives only a scalar projection $\widetilde{Y} = \sum_i p_i Y_i$ of the measurements. In terms of the definitions in Section II, these are the two extreme cases $\tilde{m} = n$ and $\tilde{m} = 1$. As a natural choice for the p_i's in the second case we take a symmetric projection, i.e., $p_1 \ldots p_n = 1/\sqrt{n} \ldots 1/\sqrt{n}$.

We first apply Theorem 1 to this example. The variance of N is given by $VAR(N) = \Delta^2 + \delta^2$. By the assumption $\Delta \gg \delta$, the Fisher Information of N with respect to a translation parameter (as defined in (12)) is given approximately[3] by

$$J_N = \int \frac{(f'_N(x))^2}{f_N(x)}dx \approx \frac{1}{2}\left\{\int \frac{(\frac{1}{\delta^2}\phi'(x/\delta+\Delta))^2}{\frac{1}{\delta}\phi(x/\delta+\Delta)}dx + \int \frac{(\frac{1}{\delta^2}\phi'(x/\delta-\Delta))^2}{\frac{1}{\delta}\phi(x/\delta-\Delta)}dx\right\} = \frac{1}{\delta^2} \ , \tag{30}$$

where $\phi'(x) = -\frac{x}{\sqrt{2\pi}}\exp(-x^2/2)$. Substituting in Theorem 1, we get that the Fisher Information $J(\theta, \tilde{m})$, for $\tilde{m} = 1$ and $\tilde{m} = n$, satisfies

$$\frac{n}{\Delta^2+\delta^2} \leq J(\theta, 1) < J(\theta, n) = \frac{n}{\delta^2} \ . \tag{31}$$

Next, we examine directly the minimum mean squared error in unbiased estimation of θ in both cases discussed above, and show that it coincides with the CRB. Consider first the case $\tilde{m} = n$. Due to the doubly modal nature of f_N, and since $\Delta \gg \delta$, most of the measurements will be concentrated

[3] More precisely, $J_N \to 1/\delta^2$ as $\Delta \to \infty$.

in two accumulations with width of about 4δ each [4]. We thus can divide the measurements into two distinct groups, where $\{Y_i^+\}_{i=1}^{n+}$ are those in the upper accumulation, and $\{Y_i^-\}_{i=1}^{n-}$ are those in the lower one. With high probability each accumulation is associated with only one (Gaussian) mode of N, and $n^+ + n^- = n$. We claim that

$$\hat{\theta}(Y_1 \dots Y_n) = \frac{1}{n}\left\{\sum_{i=1}^{n^+}(Y_i^+ - \Delta) + \sum_{i=1}^{n^-}(Y_i^- + \Delta)\right\} \tag{32}$$

is efficient, i.e., it achieves the CRB $J(\theta, n)^{-1} = \frac{\delta^2}{n}$ given in (31). To see that, note that with high probability the estimator (32) is distributed as $\mathcal{N}(\theta, \delta^2/n)$. Thus, $\hat{\theta}$ is an unbiased estimator having MSE of δ^2/n, as we claim[5].

Consider now the case $\tilde{m} = 1$. For n large enough the Central Limit Theorem holds, and the distribution of $\widetilde{Y} = \frac{1}{\sqrt{n}}\sum_i Y_i$ is approximately normal, with mean $\sqrt{n}\theta$ and variance $(\Delta^2 + \delta^2)$. Thus, asymptotically the CRB is tight, given by $J(\theta, 1)^{-1} = (\Delta^2 + \delta^2)/n$, and it is actually achieved by the estimator $\hat{\theta}(\widetilde{Y}) = \widetilde{Y}/\sqrt{n}$. We thus conclude that for large n,

$$\frac{\min_{\hat{\theta}} E(\hat{\theta}(\underline{Y}) - \theta)^2}{\min_{\hat{\theta}} E(\hat{\theta}(\widetilde{Y}) - \theta)^2} \approx \frac{J(\theta, n)^{-1}}{J(\theta, 1)^{-1}} \approx \frac{\delta^2}{\Delta^2 + \delta^2}\ , \tag{33}$$

where the minimizations above are taken with respect to all possible unbiased estimators. This demonstrates the loss due to pre-projection of the measurements for large n.

Acknowledgments

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

[4] More precisely, on the average 99% of the measurements will be within two intervals of width 4δ each, and only 0.1% of them will be outside two intervals of width 6δ each.

[5] Following the tightness of the CRB, $\hat{\theta}$ of (32) is also the *maximum likelihood* estimator in this case, [7, p. 68].

References

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

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

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

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

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

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

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

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

[9] 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.

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