Empirical non-parametric estimation of the Fisher Information

Visar Berisha and Alfred O. Hero

Abstract

The Fisher information matrix (FIM) is a foundational concept in statistical signal processing. The FIM depends on the probability distribution, assumed to belong to a smooth parametric family. Traditional approaches to estimating the FIM require estimating the probability distribution, or its parameters, along with its gradient or Hessian. However, in many practical situations the probability distribution of the data is not known. Here we propose a method of estimating the Fisher information directly from the data that does not require knowledge of the underlying probability distribution. The method is based on non-parametric estimation of an f-divergence over a local neighborhood of the parameter space and a relation between curvature of the f-divergence and the FIM. Thus we obtain an empirical estimator of the FIM that does not require density estimation and is asymptotically consistent. We empirically evaluate the validity of our approach using two experiments.

Index Terms

Fisher information, f-divergence, information geometry, non-parametric estimation

I. Introduction

The Fisher information matrix (FIM) is an important quantity in signal processing and statistical estimation. It can be used to benchmark the performance of an estimator (via the Cramer-Rao Lower Bound (CRLB)) [1], design an optimal experiment to maximize the trace or determinant

V. Berisha is with the School of Electrical, Computer, and Energy Engineering and the Department of Speech and Hearing Science, Arizona State University, e-mail: visar@asu.edu. This author gratefully acknowledges partial support of this work by ONR grant N000141410722 and NIH-NIDCD grant 1R21DC012558.

A. Hero is with the Department of Electrical Engineering and Computer Science, University of Michigan, e-mail: hero@eecs.umich.edu. This author gratefully acknowledges partial support of this work by ARO grant W911NF-11-1-0391 and NSF grant CCF-1217880.

of the FIM [2], or develop super-efficient estimators [3]. The FIM can be computed as the covariance matrix of the gradient of the log likelihood function (the score function). However, in some cases the statistical model may not be known or the covariance computation may be intractable to manipulate. In some cases one has access to a generative model, a black box that can generate multiple realizations from an distribution for various settings of its parameters, e.g., when one can generate data from a controlled experiment with accessible and tunable operating parameters. In such cases it makes sense to try and compute the FIM empirically from multiple perturbed experiments. This is the approach that we take here.

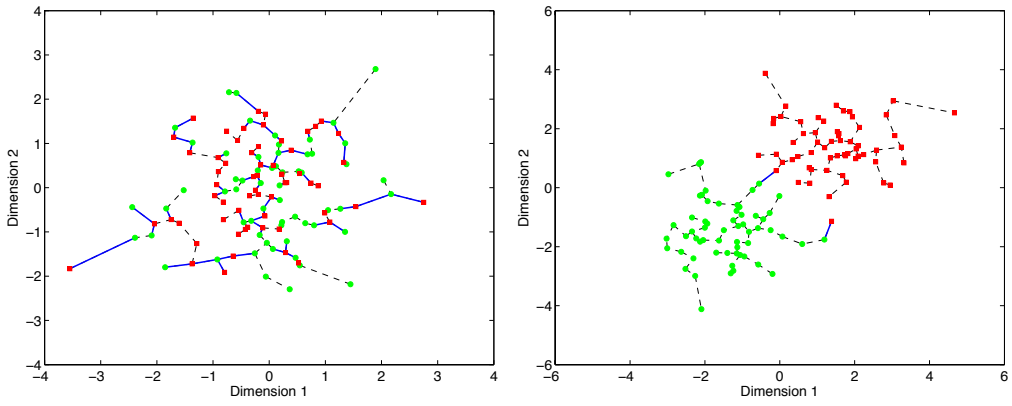
The first approach that might come to mind would be to perform non-parametric density estimation for each of the perturbed experiments and then compute a numerical gradient of the log density followed by sample averaging. Another, related, approach would be to estimate an f-divergence measure between the estimated densities and use the relation between the f-divergence and the FIM through its second variational term in the Taylor expansion. These density estimate-and-plug strategies are not the most natural approaches. Indeed it seems ill-advised to approach the problem of estimation of a finite dimensional quantity, e.g. the elements of the FIM, by first generating an estimate of an infinite dimensional quantity, e.g., the density function. Here we introduce a more direct way of estimating the FIM that uses the estimates an f-divergence measure directly from the data, and does not require density estimation. We empirically evaluate this procedure on two experiments: (1) estimating the FIM from multidimensional data for which the truth is known in closed form and (2) using the FIM as a proxy for the difficulty of an estimation problem using data from a cochlear implant signal processing chain.

II. The Henze-Penrose Divergence

For a parameter $\alpha \in [0, 1]$ consider the following affinity measure between two probability density functions p and q with domain $\mathbf{R}^d$:

$$A(p, q) = 2\alpha(1 - \alpha) \int \frac{p(x)q(x)}{\alpha p(x) + (1 - \alpha)q(x)} dx \tag{1}$$

This measure satisfies the property that $A(p, q) = 1$ if $p = q$ (a.e.) and is locally monotonic decreasing in expected value of the difference $\Delta^2 = (1 - \frac{q}{p})^2$. It has the remarkable property that it can be estimated directly without estimation or plug-in of the densities p and q based on an extension of a multi-variate two sample test. Consider the following problem: given sample realizations from p and q, denoted by X_p and X_q, find an estimator for the null hypothesis H_0 :



(a) (b)

Fig. 1. The HP divergence estimator: The minimum spanning tree showing edges connecting points from different classes for (a) overlapping distributions and (b) separable distributions.

$p(x) = q(x)$. An underexplored, non-parametric estimator for this scenario is the multivariate runs test proposed by Friedman and Rafsky (FR) [4] based on constructing the the minimal spanning tree over X_p and X_q. This is a multivariate generalization of the Wald$-$Wolfowitz test for the two sample problem [5].

In the one-dimensional case, observations from both distributions are ranked in ascending order. Each observation is then replaced by a binary variable corresponding to the class to which it belongs and the total number of runs is the test statistic for accepting or rejecting H_0. For the multivariate case, the "SORT" is replaced with a minimum spanning tree on the data. Given a data set $X \in \mathbb{R}^{N \times d}$, the spanning tree on X is defined as a connected graph, $\mathcal{G}(E, X)$, with vertex set X, edge weights given by Euclidean distances between samples from X, and no cycles. The length $\mathcal{L}(G)$ of the spanning tree is the sum of its edge lengths. The minimal spanning tree is defined as the spanning tree with minimum length.

Given the two data sets, $X_p \in \mathbb{R}^{N_p \times d}$ and $X_q \in \mathbb{R}^{N_q \times d}$, let $\mathcal{G}(E, X_p \cup X_q)$ denote the MST associated with the concatenated data set. We define $\mathcal{C}(X_p, X_q)$ as the number of edges of $\mathcal{G}(E, X_p \cup X_q)$ connecting a data point from p to a data point from q. Henze and Penrose proved the following theorem related to this test statistic [6]:

Theorem 1. *As $N_p \to \infty$ and $N_q \to \infty$ in a linked manner such that $\frac{N_p}{N_p+N_q} \to \alpha$,*

$$\frac{\mathcal{C}(X_p, X_q)}{N_p + N_q} \to 2\alpha(1-\alpha) \int \frac{p(x)q(x)}{\alpha p(x) + (1-\alpha)q(x)} dx$$

almost surely.

In Fig. 5 and 5 we show two examples of this statistical test. We plot samples from two distributions and evaluate the value of $\mathcal{C}$. In Fig. 5, both data sets are drawn from the same distribution, $\mathcal{N}([0,0]^{\mathrm{T}}, \mathbf{I})$. In Fig. 5, we plot data drawn from $\mathcal{N}([-1.5,-1.5]^{\mathrm{T}}, \mathbf{I})$ and $\mathcal{N}([1.5,1.5]^{\mathrm{T}}, \mathbf{I})$. $\mathbf{I}$ is the identity matrix. The dotted line in each figure represents the MST associated with that data. The blue lines represent the edges of the MST connecting points from different sets. The blue line count, normalized by the total number of points is the statistic proposed by Friedman and Rafsky in [4]: $\frac{\mathcal{C}(X_p,X_q)}{N_p+N_q}$. It is clear from the figures that this value is much smaller for the data in Fig. 5 than the data in Fig. 5. As Theorem 1 suggests, in the limit, this statistic converges to the integral used in the affinity measure in 1 - a ratio of the distributions product and sum.

Based on the affinity measure in 1, we can define the Henze-Penrose (HP) probability distance measure as $D_{\mathrm{HP}}(p,q) = 1 - 2A(p,q)$. It is easy to show that this divergence measure belongs to the class of f-divergences or Ali-Silvey distances [7]. Intuitively, an f-divergence is an average of the ratio of two distributions, weighted by some function $f(t)$: $D_f(p,q) = \int f(\frac{p(x)}{q(x)})q(x)dx$. Many common divergences used in statistical signal processing fall in this category, including the KL-divergence, the Hellinger distance, the total variation distance, among others. We refer the reader to [7] for a detailed discussion of this family of probability distance measures. It is easy to show that for D_{HP}, the corresponding function $f(t)$ is

$$f(t) = \frac{(t-1)^2}{t+1}. \tag{2}$$

Combining the results of Theorem 1 with the definition of D_{HP}, we identify an f-divergence that can be directly estimated from data without ever having to explicitly model the data. We use this fact to identify a non-parametric estimator of the Fisher Information directly from data.

III. FIM Estimation

Let the densities p and q be parameterized by a multidimensional parameter θ and consider the case where q is a directional perturbation of p about some particular value of θ:

$$q = p_{\theta+u} \tag{3}$$

Here u is a d-dimensional vector that is assumed to be small, e.g., $||u||^2$. Amari showed that *any* f-divergence induces a unique information monotonic Riemannian metric, given by the Fisher information matrix (see Theorem 5 in [8]), $F_\theta = E_\theta[\nabla p_\theta \nabla p_\theta^T]$. Using a Taylor expansion he showed that any f-divergence measure is related to the FIM through the asymptotic relation

$$D_{\mathrm{HP}}(p_\theta, p_{\theta+u}) = \alpha(1-\alpha)u^T F_\theta u + o(||u||^2). \tag{4}$$

Combining this with the relationship shown in Theorem 1, we identify an estimator of the Fisher information based on

$$Q_{pq} = \frac{1}{\alpha(1-\alpha)} D_{\mathrm{HP}}(p,q) \approx u^T F_\theta u \tag{5}$$

where we can estimate $D_{\mathrm{HP}}(p,q)$ using the FR runs statistic computed using the datasets from $p = p_\theta$ and $q = p_{\theta+u}$.

The above is the key relation that allows one to estimate F_θ directly from the data. In particular, assume that the experimenter can generate a data set $X_{01}, \ldots X_{0N_f}$ from the reference density p_θ and M other data sets $X_{k1}, \ldots X_{kN_g}$, $k = 1, \ldots, M$, from perturbed densities $p_{\theta+u_1}, \ldots, p_{\theta+u_M}$. Assume that M is at least equal to $d(d+1)/2$, where d is the dimension of the parameter space. Let $Q = [Q_1, \ldots, Q_M]^T$ be the M different values of Q_{pq} that are computed from the FR runs statistic applied to the pairs of datasets $\{X_{0i}\}_{i=1}^{N_0} \cup \{X_{ki}\}_{i=1}^{N_k}$, $k = 1, \ldots, M$. Then, a least squares estimator of the distinct (lower triangular) part of the FIM is easily constructed. Define the $d(d+1)/2$ element vector of direction component pairs $[u_{11}^2, \ldots, u_{1d}^2, 2u_{11}u_{12}, \ldots, 2u_{1(d-1)}u_{1d}]$, and likewise for the other $M-1$ directions $u_2, \ldots, u_M$. Concatenate all of these direction component pair vectors in the $M \times d(d+1)/2$ matrix U. Define the $d(d+1)/2$ element FIM vector:

$$F = [F_{11}, \ldots, F_{dd}, F_{12}, \ldots, F_{d(d-1)}]^{\mathrm{T}} \tag{6}$$

A simple least squares estimator for the FIM is given by:

$$\hat{F} = (U^T U)^{-1} U^T Q. \tag{7}$$

This approach, however, does not ensure that the resulting estimate is positive semidefinite (PSD). We propose an alternate estimator that further constrains the problem by ensuring the resulting matrix is PSD.

For a given parameter set, θ, we can estimate the diagonal values of the corresponding FIM by repeating the procedure described above. Specifically, we can perturb the individual components

of θ independently and iteratively reconstruct the diagonal FIM values, $\hat{F}_d$, using the least squares procedure. This now provides an additional set of constraints with which to further regularize the full FIM reconstruction. To estimate the full FIM, we propose the following semi-definite program

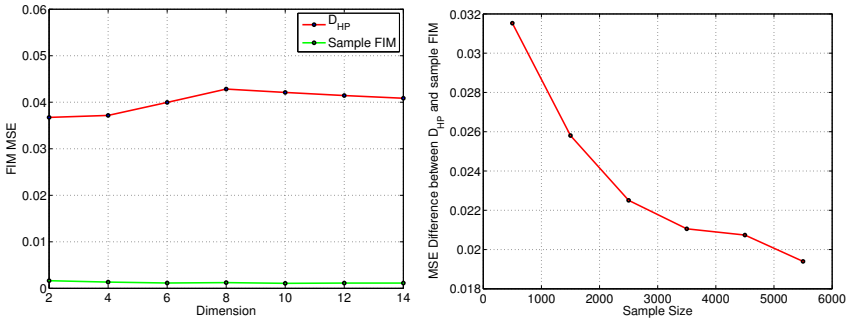
$$\begin{aligned} \underset{F}{\text{minimize}} \quad & ||UF - Q||^2 \\ \text{subject to} \quad & F_i = \hat{F}_{d,i},\ i \in 1 \dots d, \\ & \text{mat}(F) \succeq 0. \end{aligned}$$

This formulation aims to minimize the L2 error with respect to the vectorized FIM, subject to the constraint that the diagonal values are the same as those estimated individually and that the FIM (denoted by $\text{mat}(F)$) is PSD.

IV. Results

We validate the proposed procedure. In the first experiment, we estimate the mean square error (MSE) of the FIM estimator for different data dimensions. We compare it against the sample FIM, known in closed form for the selected model. In the second experiment, we empirically analyze a complicated signal processing chain for which it becomes impossible to derive an analytical relation between parameters and observations. We also analyze a cochlear implant (CI) simulator similar to the one in [9]. Through the relationship between the CRLB and the FIM, we empirically estimate the minimum variance of the best unbiased estimator of the signal energy at different frequency components in the input speech signal (parameters, θ), given only current levels of the different electrodes. This is exactly the problem that CI patients must solve - given electrode stimulation, resolve the input speech.

Empirical FIM Estimation: The objective of this experiment is to compare estimates of the FIM with the closed form solution for spherically-symmetric Gaussian data at different dimensions. We start with the d-dimensional distribution, $f(x) = \mathcal{N}(\mathbf{0}_{d\times 1}, \sigma^2\mathbf{I}_{d\times d})$, and perturb f about the mean, $f_u(x) = \mathcal{N}(\mathbf{0}_{d\times 1} + \mathbf{u}, \sigma^2\mathbf{I}_{d\times d})$, where $\mathbf{u} \sim \mathcal{N}(0, \sigma_u^2)$. For this experiment, the closed form estimate of the FIM is known (given by the identity matrix, $\mathbf{I}_{d\times d}$, since $\sigma^2 = 1$). We estimate the FIM from empirical data using the algorithm proposed here and evaluate the FIM MSE. In particular, we draw $N = 1000$ points from both $f(x)$ and $f_u(x)$ and estimate the HP divergence for multiple values of u ($\sigma_u^2 = 0.25$). From this set of parameters, we use the least squares estimator in 3 to obtain an estimate of the FIM. We do this for different problem



(a) (b)

Fig. 2. (a) MSE of the FIM estimate for the D_{HP}-based estimator (b) The MSE difference between the two estimators in (a).

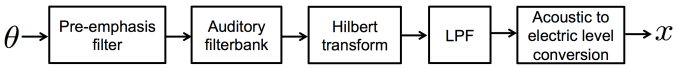


Fig. 3. The cochlear implant signal processing chain.

dimensions ranging from 2 to 14. In Fig. 5, we plot the FIM MSE (averaged over 25 Monte Carlo simulations) as a function of dimension. We compare the MSE of this estimator with the MSE of the sample FIM estimated from the sample covariance matrix. In Fig. 5 we plot the MSE difference between the D_{HP}-based estimator and the sample estimator as the sample size increases. As the figure shows, the sample FIM starts with a lower MSE, however the difference between the two estimators decreases as the sample size increases. It is important to note that the sample FIM assumes full knowledge of the analytical form of the FIM. The estimator here makes no such assumption and is directly estimated from the data.

Estimating the CRLB in a Complex Signal Processing System: Beyond estimating the Fisher Information itself, we can use the HP divergence to characterize the difficulty of an estimation problem. It is well known that the FIM can be used to benchmark the performance of an estimator (via the CRLB). If $\mathbf{T}(x)$ is an unbiased estimator of θ, then the inverse FIM lower bounds the covariance matrix of the estimator by $\mathrm{cov}_\theta(\mathbf{T}(x)) \geq F_\theta^{-1}$. The CRLB is known in closed form for certain distributions; however, for complex signal processing systems, it becomes impossible to characterize the CRLB for a specific parameter unless simplifying assumptions are made. We consider a signal processing simulator for a 16-electrode cochlear implant device. Cochlear implants (CI) provide hearing in patients with damage to sensory hair cells by stimulating the 16 electrodes with appropriate energy. In a CI, a microphone picks up audio from the environment,

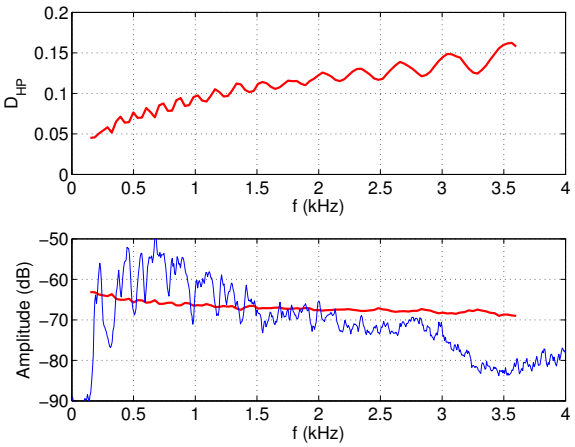


Fig. 4. (Top) D_{HP} as a function of frequency (Bottom) The average speech spectrum and the empirically-estimated CRLB

a speech processor decomposes the speech in 16 sub-bands and identifies the appropriate amount of current to stimulate each of the 16 electrodes, and stimulates the electrode array. We have implemented a CI simulator in Matlab, similar to the Simulink simulator by Loizou in [9] (see the block diagram in Fig. 3). This speech processor includes linear (filtering) and non-linear (rectification, intensity compression) operations. The input of the processor is the spectrum of the speech signal (θ) and the output of the simulator (x) represents the current level applied to each of the 16 output electrodes. Here, we are interested in characterizing the lower bound on the variance of the estimator for the amplitude spectrum of the speech signal when all we can observe is the electrode current level. This will serve as a proxy for how well a patient with a cochlear implant can resolve the amplitude of any given frequency component in a speech signal.

We evaluate the algorithm on the well-known TIMIT corpus - a phonemically balanced database that spans multiple speakers and multiple regional dialects in the US [10]. We sample 100 sentences from this database, downsample the data to 8 kHz, normalize the average energy of each signal to a fixed value, and perform the experiment as described in section 3. We first process each speech signal through the cochlear implant simulator. The simulator analyzes each input speech signal using a 10ms frame size (with 80% overlap) and estimates the energy associated with each of the 16 electrodes per frame. If the total number of frames is N, this results in a data matrix of $N \times 16$. We denote this matrix by X_θ. For a given frequency component, we perturb this parameter by a small value u by simply adding a sinusoid of appropriate energy to

the original speech. We then process the resulting audio through the CI simulator to generate a new data matrix, $X_{\theta+u}$. We estimate $\mathcal{C}(X_\theta, X_{\theta+u})$ using the statistical test from Theorem 1. This is done multiple times and the FIM is estimated using the procedure described in section 3. In this simulation, rather than estimating the FIM in a single step, we restrict ourselves to the diagonal components of the FIM. We sweep through the frequency spectrum by iteratively perturbing the spectrum of the original speech signal and generating a new data set for each perturbation (through the CI simulator). We use the relationship between the HP divergence and the FIM to iteratively estimate the diagonal components of this matrix one at a time. For every speech signal, at every frequency we perform 10 perturbations with u drawn from a normal distribution of mean 3.5 and variance 1. This is done for all selected signals in the corpus.

In Fig. 4 (top), we plot the average HP divergence as a function of frequency from 150 Hz to 3.6 kHz. As the figure shows, there is a general increasing trend in the plot. This makes sense intuitively as the average speech spectrum contains more signal energy at low frequencies. Therefore, for the same perturbation, D_{HP} will be greater for higher frequencies than lower frequencies. The many peaks and troughs of the divergence measure are also consistent with the filter bank used in the simulator. In a CI simulator, one of the first steps is a 16-channel auditory filter bank. As expected, the troughs of this signal match exactly with the center frequencies of this filter bank. In Fig. 4 (bottom), we plot the square root of the estimated CRLB (the standard deviation associated with the optimal unbiased estimator of each amplitude component) on top of the frequency spectrum. The spectrum represents an average of the speech spectrum using 150 ms frames and a Hamming window. As the figure shows, the amplitude and the standard deviation of the estimator are often of the same order. Indeed, for a number of frequencies, the standard deviation is greater than the mean indicating that, at those frequencies, even the best estimator will have difficulty identifying statistically significant differences between the true signal energy and no energy. This is consistent with the noisy and poor quality speech associated with real cochlear implants [13]. Performing this analysis for voiced and unvoiced speech segments independently reveals that the CRLB almost always dominates the signal energy for unvoiced; however, this is not the case for voiced speech segments. This is consistent with results from the literature, where subjective tests show that individuals with cochlear implants have higher intelligibility rates for vowels than for consonants [11].

In addition to calculating the diagonal CRLB for the TIMIT database, we also compare the

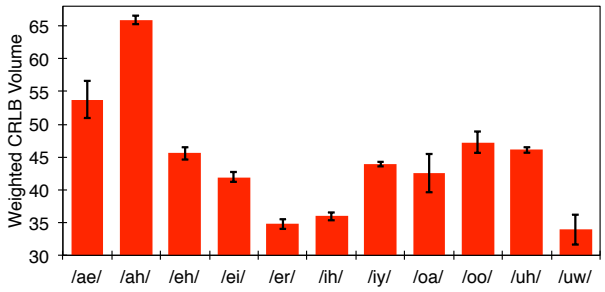


Fig. 5. The CRLB volume for each of the 11 vowels in [12].

difficulty of estimating specific vowels by calculating the full CRLB matrix for each vowel. We use the vowels from male speaker "m01" in the Hillenbrand data set [12]. As before, we downsample the data to 8 kHz, normalize the average energy of each signal to a fixed value, and perform the experiment as described in section 3. For each vowel, each component of u is drawn from a normal distribution of mean 3.5 and variance 1. In order to manage computational complexity, rather than estimating the full FIM associated with all input frequencies, we elect to reconstruct the 16×16 sub-matrix associated with the center frequencies of the auditory filter bank in Fig. 3. We justify this by noting that the CI model is most sensitive at these center frequencies (see peaks in Fig. 4 (top)). As a result, this sub matrix acts as a best-case estimation bound.

We estimate F_θ using the SDP-based estimator in section 3 and we invert to calculate the CRLB, $C_\theta = F_\theta^{-1}$. Prior to inverting, we use the Bayesian-based diagonal loading procedure by Haff for regularization [14]. For each of the vowels in the analysis, we compute the volume of the CRLB matrix, weighted by a perceptual weighting filter giving more weight to bands that contain more signal energy. The weighting matrix, W, is a diagonal matrix with the components given by the signal energy at the location of the 16 frequencies associated with θ. Perceptual weighting of frequency bands by signal energy is common in many speech applications [15], [16]. The volume of the weighted CRLB matrix is given by $\mathrm{Vol} = \log(\det(VDWV^{\mathrm{T}}))$, where $C_\theta = VDV^{\mathrm{T}}$. This estimate of the volume serves as a proxy for the uncertainty associated with estimating the spectrum of the speech signal from only the current at the electrodes. As such, it is expected that vowels with a small volume are more intelligible than vowels with a large volume by patients with cochlear implants. In Fig. 5 we plot the weighted CRLB volume for

the vowels in the Hillenbrand dataset.

A similar behavioral study was conducted by Dorman et. al in [17]. Using the same dataset, the authors encode the vowels with a simulated CI and conduct an intelligibility assessment task where they ask participants to correctly identify the encoded vowels. The authors show that there are only two vowels that prove to be difficult for the listeners, /ah/ as in "hod" (43% correct classification rate) and /ae/ as in "had" (68% correct classification rate) - see Table II in [17]. Interestingly, these are the two vowels with the highest CRLB volume in Fig. 5. Furthermore, correlating the average CRLB volumes in Fig. 5 with the intelligibility values in Table II in [17] reveals a correlation coefficient of -0.801. This suggests that there exists a strong inverse relationship between this volume and the intelligibility scores for the corresponding vowels. In short, vowels with a small CRLB volume are easier to identify than those with a large CRLB volume.

V. Conclusion

In this paper we have identified a new data-driven approach for estimating the Fisher information by exploiting the relationship between the FIM and the family of f-divergence measures. The estimator relies on multiple perturbations of a signal model followed by a semidefinite program that reconstructs the FIM. We evaluated the algorithm by evaluating its ability to reconstruct the FIM for a signal model where we have a closed form solution. In addition, we showed that we can predict the difficulty of resolving different speech samples processed through a cochlear implant by comparing intelligibility values estimated through behavioral experiments with the volume of the CRLB. Future work will analyze the asymptotic properties of this estimator and discretization bias. This would allow us to better understand the approximation error as a function of the underlying density and could serve as a starting point for developing theory around an optimal value of σ_u^2 given the sample size N and some properties of the data distribution.

References

[1] H. Cramer, *A contribution to the theory of statistical estimation,* Skand. Aktuaries Tidskrift, vol. 29, pp. 458463, 1946.

[2] V. V. Federov, *Theory of optimal experiments,* Academic Press, Orlando, 1972.

[3] L. Le Cam, *On some asymptotic properties of maximum likelihood estimates and related Bayes estimates,* University of California Publications in Statistics 1, pp. 277-330, 1953.

[4] J. Friedman and L. Rafsky, *Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-Sample Tests.* Annals of Statistics, 1979.

[5] A. Wald and J. Wolfowitz, *On a test whether two samples are from the same population.* Annals Math. Stat., 1940.

[6] N. Henze and M. Penrose, *On the Multivariate Runs Test.* Annals of Statistics, 1999.

[7] I. Csiszr and P. Shields, *Information Theory and Statistics: A Tutorial,* Foundations and Trends in Communications and Information Theory, pp. 417528, 2004.

[8] S. I. Amari and A. Cichocki. *Information geometry of divergence functions,* Bulletin of the Polish Academy of Sciences: Technical Sciences, pp. 183-195, 2010.

[9] Mathworks, *Cochlear Implant Speech Processor,* DSP Systems Toolbox Examples, 2014.

[10] William M. Fisher, George R. Doddington, and Kathleen M. Goudie-Marshall, *The DARPA Speech Recognition Research Database: Specifications and Status*, Proceedings of DARPA Workshop on Speech Recognition, pp.93–99, 1986.

[11] F. G. Zeng and J. J. Galvin, *Amplitude mapping and phoneme recognition in cochlear implant listeners,* Ear and hearing, pp. 60-74, 1999.

[12] James Hillenbrand, Laura A. Getty, Michael J. Clark, and Kimberlee Wheeler, *Acoustic characteristics of American English vowels*, Journal of the Acoustical Society of America, 1995.

[13] P. Loizou, *Mimicking the Human Ear,* IEEE Signal Processing Magazine, Vol. 15, No. 5, pp. 101-130, 1998.

[14] L.R. Haff, *Empirical Bayes estimation of the multivariate normal covariance matrix,* Ann. Statist., pp. 586597, 1980.

[15] R. Udrea, V. Nicolae, and C. Silviu, *An improved spectral subtraction method for speech enhancement using a perceptual weighting filter,* Digital Signal Processing, pp. 581-587, 2008.

[16] M. Mauc and W. Navarro, *Analysis-by-synthesis speech coding method with truncation of the impulse response of a perceptual weighting filter.* U.S. Patent No. 5,963,898, 1999.

[17] P. Loizou, M. Dorman, and V. Powell, *The recognition of vowels produced by men, women, boys, and girls by cochlear implant patients using a six-channel CIS processor,* The Journal of the Acoustical Society of America, pp. 1141-1149,1998.

[18] J. Costa and A. Hero, *Geodesic entropic graphs for dimension and entropy estimation in manifold learning.* IEEE Trans. on Sig. Proc. , vol.52, no.8, pp.2210,2221, Aug. 2004.

[19] A. Hero, M. Bing, O. Michel, and J. Gorman, *Applications of entropic spanning graphs.* IEEE Trans. on Sig. Proc., vol.19, no.5, pp.85,95, Sep 2002.