Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

A note on estimation using quantized data

Alessandro Chiuso [*]

[*] *Dipartimento di Tecnica e Gestione dei Sistemi Industriali,*
Università di Padova (sede di Vicenza), stradella San Nicola, 3 -
36100 Vicenza, Italy. E-mail: `chiuso@dei.unipd.it`

Abstract: In this paper we discuss estimation of parameters from quantized data. Extending some recent results appeared in the literature we show that, under a regularity assumption on the parametric model describing the data, the Maximum Likelihood estimator can be found, asymptotically, in closed form in two steps. The first is a non-linear (and invertible) mapping of the observed relative frequencies which makes the dependence on the parameters linear; the second is a linear estimator. Some simulations which demonstrate the results are included.

Keywords: Estimation, Identification, Quantization, Communication Constraints, Sensor Networks

1. INTRODUCTION

Traditionally control/estimation and communication were considered and tackled as separate problems, the control/estimation algorithms assuming “perfect” communication was available. This is no longer possible when it comes to designing control and communication strategies with severe performance requirements and/or bandwidth limitation. This is particularly evident for Wireless Sensor Networks (WSN) in which cheap sensors might produce very coarsely quantized measurements (one or few bits) and very restrictive energy-consumption requirements might impose severe limitations to the data communication rate.

Essentially for these reasons the recent years have witnessed an increasing interest in the interplay between sensing, communication, estimation and control.

In this paper we shall be concerned with implications of quantization in parameter estimation and system identification. The literature on this topic is vast but rather scattered in the fields of Signal-Information Processing, Communications, Control and Statistics making it impossible, in this short note, to exhaustively survey the literature. In particular we shall limit our interest to the “large sample” case, studying simple algorithms which allow to perform estimation using only, possibly coarsely, quantized observations. We shall also be interested in algorithms which, even though non *distributed* in nature, are prone to *distributed implementations*, i.e. can be implemented in a network only through local communications.

This problem, with particular reference to WSN, has been recently studied in some papers by Ribeiro and Giannakis (2006a,b) and solved via Maximum Likelihood (ML) estimation. However, besides a few simple exceptions, the estimators proposed in Ribeiro and Giannakis (2006a,b) require performing a non-linear search over parameter space. In fact computing the ML estimator requires finding the minimum of a (non-linear) function of the parameters. This has two drawbacks: the first regards computational complexity and the second the fact that this approach requires this optimization to be centralized (i.e. all data have to be available to a central processing unit).

Parameter estimation from quantized measurements includes, of course, also system identification when input-output data have to be transmitted over a bandlimited communication channel. With respect to the latter, the recent papers Wang et al. (2006b,a); Wang and Yin (2007) demonstrate that, provided the system is excited by a periodic input [1], system identification using quantized observations boils down to estimating a constant corrupted by white noise from quantized measurements, thus motivating Example 1 below. Note that Example 2 below, which could not be tackled with the techniques in Wang and Yin (2007), is also important in System Identification since, as already discussed for the binary case in Wang et al. (2006b), assuming perfect knowledge of the noise distribution is often unrealistic [2].

We have to remind that, besides the engineering community, also statisticians have for long time been interested in *categorical data analysis*, in which measurements are “categories” or “classes” and the parameters to be estimated describe the (parametric) class of probabilistic models describing the data (e.g. parametric multinomial models, see Fisher (1928); Birch (1961); Cox (1984); Zacks (1971) for some early references and also Section 2 for more details). It is also worth recalling that the very recent and thorough paper by Morales et al. (2006) discusses asymptotic efficiency of minimum ϕ-divergence estimators of (continuous) parameters from quantized observations. ML estimators are a special case of minimum ϕ-divergence

[⋆] This work has been supported in part by the national project *New techniques and applications of identification and adaptive control* funded by MIUR

[1] See e.g. Wang and Yin (2007) Section 10.

[2] The reader has just to keep in mind that the problems stated in Examples 1 and 2 are not oversimplified cases but find important applications in System Identification. In this paper we shall not enter into the question of how, from the estimation of a constant, one recovers the parameters of a linear systems. We refer the reader to Wang et al. (2006b,a); Wang and Yin (2007).

978-1-1234-7890-2/08/$20.00 © 2008 IFAC

10.3182/20080706-5-KR-1001.3270

and hence more classical results found in the statistics literature mentioned above are re-captured under this general framework.

The framework discussed in this paper is strongly related (and in fact extends) the Quasi-Convex Combination Estimator (QCCE) discussed in Wang et al. (2006a); Wang and Yin (2007). In particular, with respect to Wang and Yin (2007), we improve along the following directions: (i) we allow for partially unknown (parametrized) noise distribution functions; in particular note that the simulation setup reported in Section 5, referring Example 2, is not covered by the results in Wang and Yin (2007) since the noise variance is not assumed to be known. We state a *sufficient* condition under which the two-stages estimator is asymptotically ML, hence (ii) proving asymptotic efficiency using standard tools in asymptotic statistics for the whole class of estimators (regardless of the specific parametric model); instead in Wang and Yin (2007) asymptotic efficiency of the QCCE estimator was shown by direct comparison with the Cramér-Rao lower bound. We remind the reader that the proof in Wang and Yin (2007) was limited to the case in which θ is a scalar location parameter and the noise distribution is completely known while here the parametric model is much more general.

As an illustrating example, using the technique of this paper we also tackle two other problems discussed in Ribeiro and Giannakis (2006b), in particular a scalar parameter estimation problem with *unknown* noise variance (Example 2) and a vector parameter estimation with unknown noise variance (Example 3); in the paper Ribeiro and Giannakis (2006b) instead an iterative method, based on minimization of the likelihood, was proposed.

We stress that we work under the assumption that the quantization is fixed; of course the asymptotic results can be utilized to optimize the partition as to minimize estimation error variance (maximizing information); this has already been done, to some extent, e.g. in Venkitasubramaniam et al. (2007); Ribeiro and Giannakis (2006a); Wang et al. (2006a) and *will not* be addressed here. Also a relevant question when dealing with quantization is when and whether the information contained in quantized data converges to the non-quantized case as the number of partitions grows. This problem has been studied by several authors; we refer the reader to the papers Vajda (2002); Liese et al. (2006) and references therein for an up-to-date account of the results available.

Besides the result per se, which, we should say, is rather straightforward, the purpose of this paper is also to provide a link between recent papers published in the areas of Signal Processing (Ribeiro and Giannakis (2006a,b)), System Identification (Wang et al. (2006a); Wang and Yin (2007)), Information Theory (Liese et al. (2006); Vajda (2002); Morales et al. (2006)) and more classical results in Statistics (Birch (1961); Cox (1984); Fisher (1928); Rao (1958)); to the author's opinion this relation has been partially overlooked in recent works.

The structure of the paper is as follows: in Section 2 we shall state the problem while in Section 3 we derive, under suitable conditions, a closed-form estimator which is asymptotically maximum likelihood (ML). Section 4 applies the result of the paper to three examples which have been recently studied in the literature Wang and Yin (2007); Ribeiro and Giannakis (2006a,b) while Section 5 contains some experimental results. In Section 6 conclusions are drawn.

1.1 Notation

Boldface lowercase letters (e.g. $\mathbf{x}$) denotes random variables (rv's); x shall be the sample value of $\mathbf{x}$. Given a sequence of rv's $\mathbf{x}_N$, converging to zero in probability, $\sqrt{N}\mathbf{x}_N \xrightarrow{\mathcal{L}} \mathbf{x}$ denotes convergence in law (see e.g. Ferguson (1996); van der Vaart (1998)). The variance of the limiting distribution is called asymptotic variance.

Given two sequences of rv's $\mathbf{x}_N$ and $\mathbf{v}_N$, we shall say that $\mathbf{x}_N = o_P(\mathbf{v}_N)$ if, $\forall \delta > 0$,

$$\lim_{N\to\infty} P[\mid \mathbf{x}_N/\mathbf{v}_N \mid > \delta] = 0.$$

When two sequences of rv's ($\mathbf{v}_N$ and $\mathbf{w}_N$) differ only up to $o_P(1/\sqrt{N})$ terms (which we shall denote as $\mathbf{v}_N \doteq \mathbf{w}_N$), then (multiplied by $\sqrt{N}$) they have the same asymptotic distribution (see e.g. Ferguson (1996); van der Vaart (1998)); we shall also say that $\mathbf{v}_N$ and $\mathbf{w}_N$ are *asymptotically equivalent.*

2. STATEMENT OF THE PROBLEM

Let us consider a *partition* of $\mathbb{R}^n$, i.e. a collection of sets $\{A_i\}_{i=1,..,k}$, $A_i \subseteq \mathbb{R}^n$ such that $\bigcup_{i=1}^k A_i = \mathbb{R}^n$, $A_i \bigcap A_j = \emptyset$ for $i \neq j$ and let $Q_i : \mathbb{R}^n \to \{0,1\}$ denote the indicator function of the set A_i.Assume we are given N independent identically distributed (i.i.d.) quantized measurements $\mathbf{z}(j)$, $j = 1,..,N$ where $\mathbf{z}_i(j)$, the i-th component of $\mathbf{z}(j)$, is given by

$$\mathbf{z}_i(j) = Q_i(\mathbf{y}(j)), \quad i = 1,..,k \quad j = 1,..,N. \tag{2.1}$$

We shall be concerned with the problem of estimating a (vector) constant $\theta \in \Theta \subseteq \mathbb{R}^s$ which parametrizes the (common) distribution function $F_{\mathbf{y}}(y;\theta) := P[\mathbf{y} \leq y;\theta]$ from measurements $\mathbf{z}(j)$, $j = 1,..,N$. Whenever needed we shall make the assumption that the $\mathbf{y}(j)$'s are absolutely continuous w.r.t. the Lebesgue measure and denote with $f_{\mathbf{y}}(y;\theta)$ the density function.

A simple yet important example (see Wang et al. (2006a); Wang and Yin (2007); Ribeiro and Giannakis (2006a)) is the estimation of a constant from quantized noisy measurements; this is formalized in Example 1 below. The distribution function of $\mathbf{y}(i)$ is in this case given by $F_{\mathbf{y}}(y;\theta) = F_{\mathbf{w}}(y-\phi)_{|\phi=\theta}$, i.e. θ is a (scalar) location parameter.

Note that, our setup encompasses also the case in which the distribution function $F_{\mathbf{w}}(w;\eta) := P[\mathbf{w} \leq w;\eta]$ is only known up to a (nuisance) vector parameter η (see Example 2 below). In this case the distribution function of $\mathbf{y}_i$ would be of the form

$$F_{\mathbf{y}}(y;\theta) = F_{\mathbf{w}}(y-\phi;\eta) \qquad \theta := \left[\, \phi^\top \; \eta^\top \,\right]^\top$$

Let us denote with e_i the k-dimensional vector with all entries equal to zero except for the i-th entry which is equal to 1. Since the sets A_i are mutually disjoint the variables $\mathbf{z}(j)$ take values in the set $\{e_1,..,e_k\}$ with probabilities

$$p_i(\theta) := P[\mathbf{z}(j) = e_i] = P[\mathbf{y}(j) \in A_i] = \int_{A_i} f_{\mathbf{y}}(y;\theta)\, dy \tag{2.2}$$

Let us define $\mathbf{n}_i := \sum_{j=1}^{N} \mathbf{z}_i(j)$, i.e. $\mathbf{n}_i$ is the number of times the variables $\mathbf{y}(j)$ take values in A_i.

The joint probability distribution function of the observation vector $\{\mathbf{z}(1),..,\mathbf{z}(N)\}$ is hence given by

$$P[\mathbf{z}(1)=z(1),..,\mathbf{z}(N)=z(N);\theta] = N!\prod_{i=1}^{k}\frac{p_i(\theta)^{n_i}}{n_i!}. \tag{2.3}$$

Note that the vector $\mathbf{n} := \{\mathbf{n}_1,..,\mathbf{n}_k\}$ follows a multinomial distribution with parameters $p(\theta) := \{p_1(\theta),..,p_k(\theta)\}$. We recall that the multinomial distribution is of the exponential type with sufficient statistic (for the parameter θ) given by[3] $\underline{\mathbf{n}} := \{\mathbf{n}_1,..,\mathbf{n}_{k-1}\}$[4]. It is convenient at this point to define the relative frequencies $\hat{\mathbf{p}}_{i,N} := \mathbf{n}_i/N$. Of course also $\underline{\hat{\mathbf{p}}}_N := \{\hat{\mathbf{p}}_{1,N},..,\hat{\mathbf{p}}_{k-1,N}\}$ is a sufficient statistic for θ. We shall also denote $\underline{p}(\theta) := \{p_1(\theta),..,p_{k-1}(\theta)\}$ and define the cumulative probabilities

$$P_i(\theta) := \sum_{l=1}^{i} p_l(\theta) \quad i=1,..,k. \tag{2.4}$$

From sufficiency of $\underline{\hat{\mathbf{p}}}_N$, w.l.o.g., it is possible to restrict the class of estimators $\hat{\theta}(\mathbf{z}(1),..,\mathbf{z}(N))$ of θ to be of form $\hat{\theta}(\underline{\hat{\mathbf{p}}}_N) = \hat{\theta}(\hat{\mathbf{p}}_{1,N},..,\hat{\mathbf{p}}_{k-1,N})$, i.e. only functions of the sufficient statistic.

It follows that the problem is reduced to that of estimating the parameter θ in a multinomial model. This problem has been studied thoroughly in the statistical community, see for instance Birch (1961); Cox (1984); Rao (1958). In fact, the well-known Pearson-Fisher theorem[5] (sometimes referred to also as Birch's theorem, see Cox (1984)) states that, under suitable regularity assumptions which we shall also assume throughout (see Birch (1961); Cox (1984) for details), the ML estimator exists and is unique for large samples. It is asymptotically normal and its asymptotic variance, which can be easily computed, reaches the Cramér-Rao lower bound.

3. TWO STAGES ASYMPTOTIC MAXIMUM LIKELIHOOD ESTIMATION

In this section, similarly to Wang et al. (2006a); Wang and Yin (2007), we shall be concerned with the problem of finding, when possible, the ML estimator in a simple (possibly closed) form.

In order to do so we shall restrict to the asymptotic case in which $N \to \infty$. In particular our aim is to derive an estimator $\hat{\theta}_N$ which is asymptotically equivalent to the ML estimator, i.e. $\hat{\theta}_N \dot{=} \hat{\theta}_{ML}(\hat{\mathbf{p}}_{1,N},..,\hat{\mathbf{p}}_{k-1,N})$. Under this assumption $\hat{\theta}_N$ shall inherit the nice properties of ML, e.g. asymptotic normality and asymptotic efficiency.

We shall first derive general sufficient conditions under which this is possible; we shall then see how this methodology applies to a number of interesting examples.

Assumption 1. Let us assume that we can find $k-1$ functions $g_i : \mathbb{R}^{k-1} \to \mathbb{R}$, $i=1,..,k-1$, $k > s$, and vectors $s_i \in R^s$ such that

$$g_i(p_1(\theta),..,p_{k-1}(\theta)) = s_i^\top \theta \tag{3.1}$$

so that the matrix $S := [s_1, s_2,..,s_{k-1}]^\top$ has full rank s and $g := [g_1,..,g_{k-1}]^\top : \mathbb{R}^{k-1} \to \mathbb{R}^{k-1}$ is invertible and continuously differentiable. We shall denote with $G(\theta)$ the jacobian at $\underline{p}(\theta)$, i.e. $[G(\theta)]_{ij} := \frac{\partial g_i}{\partial p_j}_{|\underline{p}=\underline{p}(\theta)}$. From the invertibility of g the matrix $G(\theta)$ is non-singular.

Remark 3.1 Note that Assumption 1 is equivalent to assuming that $\underline{p}(\theta)$ is the restriction of a continuously differentiable and invertible function $\underline{t} : \mathbb{R}^{k-1} \to \mathbb{R}^{k-1}$ to an s-dimensional subspace of $\mathbb{R}^{k-1}$. In fact $\underline{t} = \underline{g}^{-1}$ ◇

Remark 3.2 Invertibility of $g(\underline{p})$ ensures that also $\underline{\hat{\mathbf{q}}}_N := g(\underline{\hat{\mathbf{p}}}_N)$ is a sufficient statistic for θ. ◇

Example 1. As a first illustration we shall consider the problem addressed in Wang et al. (2006a); Wang and Yin (2007). It is assumed that θ is a scalar parameter and that $\mathbf{y}(j)$ are $\mathbf{y}(j) = \theta + \mathbf{w}(j)$ where $\mathbf{w}(j)$ are i.i.d. with known distribution function $F_{\mathbf{w}}(w)$. Consider the partitions $A_i := [c_{i-1}, c_i]$ with $c_0 = -\infty$ and $c_k = +\infty$. It is immediate to recognize that

$$F_{\mathbf{w}}(c_i - \theta) = F_{\mathbf{y}}(c_i;\theta) = P_i(\theta).$$

It follows that

$$F(c_i - \theta) = P_i(\theta) = [\underbrace{1,..,1}_{i}\ \underbrace{0,..,0}_{k-i-1}]\begin{bmatrix} p_1(\theta) \\ \vdots \\ p_{k-1}(\theta)\end{bmatrix}.$$

Therefore, assuming it is possible to compute the inverse function of F, we have $c_i - F^{-1}(P_i(\theta)) = \theta$.

This satisfies Assumption 1 with

$$g_i(p_1(\theta),..,p_{k-1}(\theta)) := c_i - F^{-1}(P_i(\theta)) \quad i=1,..,k-1$$

and $s_i = 1$. Of course the matrix $S := [s_1,..,s_{k-1}]^\top = [1,1,..,1]^\top$ has full rank equal to 1 (i.e. the number of parameters). Clearly each g_i is invertible as a function of $P_i(\theta) = \sum_{l=1}^{i} p_l(\theta)$ and hence also $g := [g_1,..,g_{k-1}]$ is invertible. Differentiability of the functions g_i follows easily from the absolute continuity of $\mathbf{w}$.

Example 2. As a second example we consider a modification of Example 1 in which the noise variance is unknown. This generalizes the problem considered in Ribeiro and Giannakis (2006b), Section III.C, where the number of thresholds per node is allowed to be larger than 2, similarly to what is done in Ribeiro and Giannakis (2006b), Section IV.B for the case of completely unknown noise probability distribution function. We assume $\mathbf{y}(j) = \phi + \sigma\mathbf{w}(j)$ where $\mathbf{w}(j)$ are i.i.d. with known distribution function $F_{\mathbf{w}}(w)$ and the parameter vector to be estimated is $\theta := [\phi\ \sigma]^\top$.

We assume the partitions A_i are the same as in Example 1 and also that $F_{\mathbf{w}}$ is invertible. Then,

$$F_{\mathbf{w}}\left(\frac{c_i - \phi}{\sigma}\right) = F_{\mathbf{y}}(c_i;\theta) = P_i(\theta).$$

For simplicity of exposition assume $k-1$ is even and consider all the pairs $(i, i+1)$ with $i = 2j-1$, $j = 1,..,\frac{k-1}{2}$. Then, for any pair of indices $(2j-1, 2j)$, it is straightforward to see that

[3] It is sufficient to consider only the first $k-1$ since trivially $\mathbf{n}_k = N - \sum_{i=1}^{k-1}\mathbf{n}_i$.

[4] The underlined symbol denotes the vector with the last component removed

[5] The Pearson-Fisher theorem states also that, asymptotically, the goodness-of-fit statistics on the sample proportions is χ^2 distributed with $k-s-1$ degrees of freedom and independent of $\sqrt{N}\left(\hat{\theta}_N^{ML} - \theta\right)$, see Fisher (1928).

$$\begin{aligned} g_{2j-1}(\underline{p}) &:= \frac{F_{\mathbf{w}}^{-1}(P_{2j-1})\,c_{2j} - F_{\mathbf{w}}^{-1}(P_{2j})\,c_{2j-1}}{F_{\mathbf{w}}^{-1}(P_{2j-1}) - F_{\mathbf{w}}^{-1}(P_{2j})} = [1\ 0]\,\theta \\ g_{2j}(\underline{p}) &:= \frac{c_{2j-1} - c_{2j}}{F_{\mathbf{w}}^{-1}(P_{2j-1}) - F_{\mathbf{w}}^{-1}(P_{2j})} = [0\ 1]\,\theta. \end{aligned} \tag{3.2}$$

Note that, for $j = 1, .., \frac{k-1}{2}$, the functions $[g_{2j-1},\, g_{2j}]^\top : \mathbb{R}^2 \to \mathbb{R}^2$ are invertible as functions of $P_{2j-1} = \sum_{l=1}^{2j-1} p_l(\theta)$ and $P_{2j} = \sum_{l=1}^{2j} p_l(\theta)$ and hence also $g := [g_1,\, g_2, .., g_{k-1}]^\top$ is invertible. Differentiability follows from absolute continuity of $\mathbf{w}$.

It is now sufficient to observe that $s_i^\top = [1\ 0]$ for $i$odd and $s_i^\top = [0\ 1]$ for $i$even so that $S^\top := [\, s_1\ s_2\ \ldots\ s_{k-1}\,]$ has rank $2 = \dim(\theta)$.

Essentially Assumption 1 guarantees that the sufficient statistic $\underline{\hat{\mathbf{p}}}_N$ can be made into a linear function of θ using a "sufficiently regular" function g. This will allow us to show that the transformed statistic $\underline{\hat{\mathbf{q}}}_N = g(\underline{\hat{\mathbf{p}}}_N)$, which is still sufficient for θ, can be thought of as the output of a linear model:

$$\underline{\hat{\mathbf{q}}}_N = S\theta + \mathbf{v}_N \tag{3.3}$$

Remark 3.3 Note also that the papers Wang et al. (2006b), Ribeiro and Giannakis (2006b) discuss the case of (partially) unknown noise distribution but deal only with *binary* (two classes) quantizers. It is apparent from Assumption 1 that using a binary quantizer (i.e. $k = 2$) it is not possible to find an invertible mapping g so that S has full column rank (S would have one row and, at least, two columns). In fact, in Wang et al. (2006b) the authors needed to modify the thresholds and also act multiplicatively on the input in order to be able to estimate the noise distribution (see eg. Wang et al. (2006b) Section 5, Example 4). Similarly, in Ribeiro and Giannakis (2006b), Section III.B, the binary sensors are divided in two groups (say A and B); the quantizers in group A have a different threshold from those of group B. Alternatively, in Ribeiro and Giannakis (2006b), Section III.C quantizers with 2 thresholds (i.e. 3 classes) are considered. Our result generalizes this situation to an arbitrary number of thresholds, as already stated in Example 2. ◇

The asymptotic distribution of $\mathbf{v}_N$ is given by the following proposition:

Proposition 3.1. The noise term $\mathbf{v}_N$ in (3.3) is asymptotically normal with asymptotic covariance matrix $\Sigma_{\mathbf{v}}(\theta) := G(\theta)\Sigma_{\mathbf{p}}(\theta)G^\top(\theta)$ where

$$\Sigma_{\mathbf{p}}(\theta) := \left[\operatorname{diag}\left(\underline{p}\right) - \underline{p}\underline{p}^\top\right]_{|\underline{p}=\underline{p}(\theta)},$$

i.e. $\sqrt{N}\mathbf{v}_N \to^{\mathcal{L}} \mathcal{N}\left(0, \Sigma_{\mathbf{v}}(\theta)\right)$

Proof. The proof can be found in the extended version of this paper available at *www.dei.unipd.it/∼chiuso.* □

We are now ready to state the main result of the paper. From the linear model (3.3) and Proposition 3.1 it follows that, asymptotically, the transformed statistic $\underline{\hat{\mathbf{q}}}_N$ is normal. The (asymptotic) ML estimator is then the weighted least squares solution (Markov estimator)

$$\hat{\theta}_N = \left(S^\top \Sigma_{\mathbf{v}}^{-1}(\theta) S\right)^{-1} S^\top \Sigma_{\mathbf{v}}^{-1}(\theta)\underline{\hat{\mathbf{q}}}_N.$$

However, the reader may argue, the asymptotic covariance matrix of $\underline{\hat{\mathbf{q}}}_N$, $\Sigma_{\mathbf{v}}(\theta) = G(\theta)\Sigma_{\mathbf{p}}(\theta)G^\top(\theta)$, depends on the true (but unknown) parameter θ. It is a standard fact that, provided a $\sqrt{N}-$consistent estimator $\hat{\theta}$ of θ is available, one can use $\Sigma_{\mathbf{v}}(\hat{\theta})$ in lieu of $\Sigma_{\mathbf{v}}(\theta)$ without altering the asymptotic properties of $\hat{\theta}_N$ [6] .

This we state in the form of a theorem.

Theorem 3.2. Under Assumption 1 and given a consistent estimator [7] $\hat{\theta}$ of θ, the weighted least squares estimator

$$\hat{\theta}_N^{WLS} := \left(S^\top \Sigma_{\mathbf{v}}^{-1}(\hat{\theta}) S\right)^{-1} S^\top \Sigma_{\mathbf{v}}^{-1}(\hat{\theta})\underline{\hat{\mathbf{q}}}_N \tag{3.4}$$

is, asymptotically, a ML estimator. It satisfies:

$$\sqrt{N}\left(\hat{\theta}_N^{WLS} - \theta\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \Sigma_\theta\right) \tag{3.5}$$

where the asymptotic covariance matrix Σ_θ is given by

$$\Sigma_\theta := \left(S^\top \Sigma_{\mathbf{v}}^{-1}(\theta) S\right)^{-1} \tag{3.6}$$

In particular this result shows that, under Assumption 1, the estimator (3.4) is asymptotically efficient.

4. APPLICATIONS

In this section we shall revisit some special cases encountered in the literature using Theorem 3.2.

First of all we consider the problem of estimating a constant from noisy and quantized measurements as discussed in Wang and Yin (2007); Ribeiro and Giannakis (2006a). This problem has already been described in Example 1.

We have seen that we can take the functions g_i to be of the form

$$g_i(p_1(\theta), .., p_{k-1}(\theta)) := c_i - F^{-1}\left(P_i(\theta)\right) \quad i = 1, .., k-1$$

and $s_i = 1$, so that $S := [s_1, .., s_{k-1}]^\top = [1, .., 1]^\top = \mathbb{1}^\top$.

The noise variance is $\Sigma_{\mathbf{v}}(\theta) = G(\theta)\Sigma_{\mathbf{p}}(\theta)G^\top(\theta)$ where

$$G(\theta) = -\begin{bmatrix} f_{\mathbf{w}}^{-1}(P_1) & 0 & \ldots & 0 \\ f_{\mathbf{w}}^{-1}(P_2) & f_{\mathbf{w}}^{-1}(P_2) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ f_{\mathbf{w}}^{-1}(P_{k-1}) & f_{\mathbf{w}}^{-1}(P_{k-1}) & \ldots & f_{\mathbf{w}}^{-1}(P_{k-1}) \end{bmatrix}$$

evaluated at $P_i = \sum_{l=1}^{i} p_l(\theta)$ and

$$\Sigma_{\mathbf{p}}(\theta) = \begin{bmatrix} p_1 - p_1^2 & -p_1p_2 & \ldots & -p_1p_{k-1} \\ -p_1p_2 & p_2 - p_2^2 & \ldots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ -p_1p_{k-1} & -p_2p_{k-1} & \ldots & p_{k-1} - p_{k-1}^2 \end{bmatrix}$$

evaluated at $\underline{p} := \underline{p}(\theta)$. The asymptotic variance is:

$$\Sigma_\theta := \operatorname{AsVar}\left\{\sqrt{N}\left(\hat{\theta}_N^{WLS} - \theta\right)\right\} = \left(\mathbb{1}^\top \Sigma_{\mathbf{v}}^{-1}(\theta)\mathbb{1}\right)^{-1}. \tag{4.1}$$

We shall now show that the expression (4.1) is indeed the Cramér-Rao bound found in Wang and Yin (2007); Ribeiro and Giannakis (2006a).

[6] Note that here continuity of $G(\theta)$ is needed to ensure convergence. This is the reason why we assumed g to be continuously differentiable.

[7] A consistent estimator is, for instance, given by $\hat{\theta} := (S^\top S)^{-1}S^\top \hat{\mathbf{q}}_N$.

First of all let us define $h_i(\theta) := f_{\mathbf{w}}(P_i)_{|P_i=\sum_{l=1}^{i} p_l(\theta)}$; note that

$$\begin{aligned} G(\theta) &= \mathrm{diag}\{h_1^{-1}(\theta),..,h_{k-1}^{-1}(\theta)\} \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix} \qquad (4.2) \\ &:= H^{-1}(\theta)J \end{aligned}$$

where the last equation defines H and J. For convenience of notation we shall also define $\underline{h}(\theta) := [h_1(\theta),..,h_{k-1}(\theta)]^\top$.

Note now that

$$\begin{aligned} \mathbb{1}^\top \Sigma_{\mathbf{v}}^{-1}(\theta)\mathbb{1} &= \mathbb{1}^\top G^{-\top}(\theta)\Sigma_{\mathbf{p}}^{-1}(\theta)G^{-1}(\theta)\mathbb{1} \\ &= \mathbb{1}^\top H(\theta)J^{-\top}\Sigma_{\mathbf{p}}^{-1}(\theta)J^{-1}H(\theta)\mathbb{1} \qquad (4.3) \\ &= \underline{h}^\top(\theta)J^{-\top}\Sigma_{\mathbf{p}}^{-1}(\theta)J^{-1}\underline{h}(\theta) \end{aligned}$$

It is now simple to check that

$$J^{-\top} = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & -1 \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix}.$$

Let us now define $\underline{\tilde{h}}(\theta) := J^{-1}\underline{h}(\theta)$ so that equation (4.3) becomes:

$$\mathbb{1}^\top \Sigma_{\mathbf{v}}^{-1}(\theta)\mathbb{1} = \underline{\tilde{h}}^\top(\theta)\Sigma_{\mathbf{p}}^{-1}(\theta)\underline{\tilde{h}}(\theta) \qquad (4.4)$$

It is immediate to recognize that the components of the vector $\underline{\tilde{h}}(\theta)$ are exactly $\tilde{h}_i$ defined in Wang and Yin (2007), Section 6. Therefore, using Lemma 7 in Wang and Yin (2007), and (4.1),

$$\begin{aligned} \mathrm{AsVar}\left\{\sqrt{N}\left(\hat{\theta}_N^{WLS} - \theta\right)\right\} &= \underline{\tilde{h}}^\top(\theta)\Sigma_{\mathbf{p}}^{-1}(\theta)\underline{\tilde{h}}(\theta) \\ &= \left(\sum_{i=1}^{k} \frac{\tilde{h}_i^2(\theta)}{p_i(\theta)}\right)^{-1} \qquad (4.5) \end{aligned}$$

where the last term on the right hand side is the Cramér-Rao lower bound (see Wang and Yin (2007), Lemma 9 and also Ribeiro and Giannakis (2006a), formula (43)).

Remark 4.4 We would like to stress that this latter result follows here from Theorem 3.2, which guarantees that $\hat{\theta}_N^{WLS}$ is asymptotically a ML estimator and hence asymptotically efficient; the proof is based on simple and standard facts in asymptotic statistics rather that (tedious) direct manipulations as in[8] Wang and Yin (2007). ◇

We now move to Example 2, which was studied in Ribeiro and Giannakis (2006b) in the particular case $k = 3$. In that paper, besides a few special cases in which the ML estimator could be found in closed form, the authors resort to gradient descent methods to find the solution to the likelihood equations. One of the special cases in which a closed form exist (see Section III.C in Ribeiro and Giannakis (2006b)) is precisely when $k = 3$, i.e. 2 thresholds. Under these circumstances the sufficient statistic has dimension 2 and is invertible as a function of the parameters; therefore, using the invariance principle (see e.g. Zacks (1971)), the ML estimator follows from equations (3.2) for $j = 1$. Using the technique of this paper the asymptotic ML estimator can be found in closed form, for an arbitrary number of thresholds (i.e. quantization levels). For reasons of space we shall not report explicitly the expressions for the jacobian $G(\theta)$ and refer the reader to Section 5 for some experimental results concerning this Example.

As a last example we also consider a minor modification of the one presented in Ribeiro and Giannakis (2006b), Section VI.B, which involves estimation of a vector parameter.

Example 3. Assume one has to measure a vector flow $v := (v_1, v_2)^\top$. Each sensor measures the flow normal to its surface, identified by the normal vector $\underline{u} := (u_1, u_2)^\top$. We assume that there are J sensors (with normal vectors $\underline{u}_j$, $j = 1,..,J$) and each sensor performs N_j independent measurements[9]

$$\mathbf{y}_{ij} = v^\top \underline{u}_j + \mathbf{w}_{ij} \quad j = 1,..,J, \;\; i = 1,..,N_j$$

We also assume that the measurement noises are i.i.d., Gaussian with zero mean and unknown variance σ^2.

The vector of parameters to be estimated is, therefore, $\theta := (v_1, v_2, \sigma)$. Assume also, w.l.o.g., that all sensors quantize their measurements with k levels. For each sensor $j = 1,..,J$ let us define with $\hat{\underline{\mathbf{p}}}_{j,N_j}$ the vectors of relative frequencies and, as in Example 2, assume $k-1$ is even. It is straightforward to see that $\{\hat{\underline{\mathbf{p}}}_{j,N_j}\}$, $j = 1,..,J$ is a sufficient statistic for θ. Also, using the same function g as in Example 2 and defining

$$\hat{\underline{\mathbf{q}}}_{j,N_j} := g(\hat{\underline{\mathbf{p}}}_{j,N_j}) = S_j\theta + \mathbf{v}_{j,N_j} \quad j = 1,..,J$$

also $\{\hat{\underline{\mathbf{q}}}_{j,N_j}\}_{j=1,..,J}$ is a sufficient statistic. Let us define $\hat{\underline{\mathbf{q}}}_{\underline{N}} := [\sqrt{N_1}\hat{\underline{\mathbf{q}}}_{1,N_1}^\top, ..., \sqrt{N_J}\hat{\underline{\mathbf{q}}}_{J,N_J}^\top]$ and similarly $\hat{\underline{\mathbf{v}}}_{\underline{N}}$.

The matrices S_j depend on the normal vectors $\underline{u}_j$. It is also a simple check to verify that in the linear model

$$\hat{\underline{\mathbf{q}}}_{\underline{N}} S\theta + \hat{\underline{\mathbf{v}}}_{\underline{N}}$$

the matrix $S := [\sqrt{N_1}S_1^\top, .., \sqrt{N_J}S_J^\top]^\top$ is of full column rank provided one takes measurements along at least two independent directions u_j's. Using the same arguments as in Proposition 3.1 it is easy to see that, asymptotically in $\underline{N} := (N_1, .., N_J)$[10], the error term $\mathbf{v}_{\underline{N}} := [\sqrt{N_1}\mathbf{v}_{1,N_1}^\top, .., \sqrt{N_J}\mathbf{v}_{J,N_J}^\top]^\top$ is normally distributed. Since the sensors are independent, the asymptotic covariance matrix $\Sigma_{\mathbf{v}}(\theta)$ is block diagonal where each diagonal block has the same form as that found in Example 2. Hence the asymptotic ML estimator can be found as in Theorem 3.2 provided $\hat{\underline{\mathbf{q}}}_N$ is replaced with $\hat{\underline{\mathbf{q}}}_{\underline{N}}$.

5. SIMULATIONS

We consider Example 2 above with $\phi = 0.3$ and $\sigma = \sqrt{2}$. We consider two scenarios in which quantization is performed using respectively $k = 5$ and $k = 7$ regions (4 and 6 thresholds). In particular we have chosen, somewhat

[8] The estimator $\hat{\theta}_N^{WLS}$ was called QCCE - Quasi-Convex Combination Estimator - in Wang and Yin (2007)

[9] Analogously one could consider a similar setup in which there is a network of $\sum_{j=1}^{J} N_j$; there are J groups of sensors, the $j-th$ group is composed of N_j elements and all sensors in the group measure the flow along the direction $\underline{u}_j$.

[10] I.e. as $\min(\underline{N}) \to \infty$.

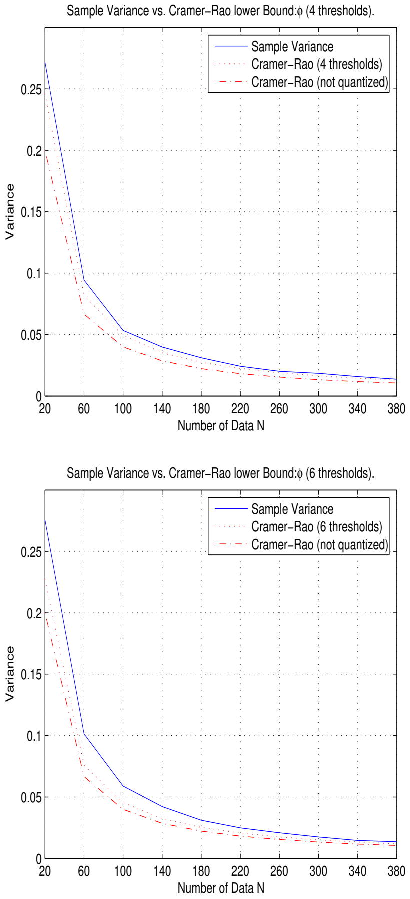

Fig. 1. Estimation of ϕ: sample variance (5000 Monte Carlo runs) vs. Cramér-Rao lower bound.

arbitrarily $[c_1\ c_2\ c_3\ c_4] = [-1\ \ -0.25\ 0.25\ 1]$ for $k = 5$ and $[c_1\ c_2\ c_3\ c_4\ c_5\ c_6] = [-1.5\ \ -1\ \ -0.5\ 0.5\ 1\ 1.5]$ for $k = 7$.

We regard σ as a nuisance parameter and hence we show only the results for the estimators of ϕ. We report (i) the sample variance (solid) from 5000 Monte Carlo experiments, (ii) the Cramér-Rao lower bound for the specified number of thresholds (dotted, red) and (iii) the Cramér-Rao lower bound for the non-quantized data (i.e. σ^2/N).

Note that the Cramér-Rao bound with 6 thresholds is essentially indistinguishable from the Cramér-Rao bound for non-quantized data (see figure 1) and very little loss is observed is 4 thresholds are used. This is in line with the considerations reported in Ribeiro and Giannakis (2006a,b). It should also be stressed that, for small sample size ($N < 200$, see figure 1) the sample variance of the estimator with 6 thresholds is larger that the estimator with 4 thresholds. This is an effect which of course the asymptotic theory does not predict.

6. CONCLUSIONS

We have considered estimation of parameters from quantized measurements. We have seen that under a regularity assumption on the probabilistic model, the ML estimator can be found, asymptotically, from a linear least squares problem on a transformed statistics, i.e. in closed form. Our setup provides a common framework which includes, but is not limited to, several important problems considered recently in the literature Ribeiro and Giannakis (2006a,b); Wang et al. (2006a); Wang and Yin (2007).

REFERENCES

M.W. Birch. A new proof of the Pearson-Fisher theorem. *The Annals of Mathematical Statistics*, 35(2):817–824, 1961.

C. Cox. An elementary introduction to maximum likelihood estimation for multinomial models: Birch's theorem and the delta method. *The American Statistician*, 38(4):283–287, 1984.

T. Ferguson. *A Course in Large Sample Theory.* Chapman and Hall, 1996.

R.A. Fisher. On a property connecting χ^2 measure of discrepancy with the method of maximum likelyhood. In *Atti Congresso Int. Mat.*, pages 95–100, Bologna, Italy, 1928. Reprinted in *Contributions to Mathematical Statistics* by R.A. Fisher (1950), Wiley, New York.

F. Liese, D. Morales, and I. Vajda. Asymptotically sufficient partitions and quantizations. *IEEE Trans. on Information Theory*, 52(12):5599–5606, 2006.

D. Morales, L. Pardo, and I. Vajda. On efficient estimation in continuous models based on finitely quantized observations. *Communications in Statistics: Theory and Methods*, 35(9):1629 – 1653, 2006.

C.R. Rao. Maximum likelihood estimation for the multinomial model with infinite number of cells. *Sankhyā*, 20: 211–218, 1958.

A. Ribeiro and G.B. Giannakis. Bandwidth-constrained distributed estimation for wireless sensor networks - part i: Gaussian case. *IEEE Transactions on Signal Processing*, 54(3):1131–1143, 2006a.

A. Ribeiro and G.B. Giannakis. Bandwidth-constrained distributed estimation for wireless sensor networks - part ii: Unknown probabilistic density function. *IEEE Transactions on Signal Processing*, 54(7):2784–2796, 2006b.

I. Vajda. On convergence of information contained in quantized observations. *IEEE Trans. on Information Theory*, 48(8):2163–2172, 2002.

A.W. van der Vaart. *Asymptotic Statistics.* Cambridge University Press, 1998.

P. Venkitasubramaniam, L. Tong, and A. Swami. Score-function quantization for distributed estimation. *IEEE Transactions on Signal Processing*, 55(7):3596–3604, 2007.

L.Y. Wang and G.G. Yin. Asymptotically efficient parameter estimation using quantized output observations. *Automatica*, 43:1178–1191, 2007.

L.Y. Wang, G.G. Yin, and J. Zhang. System identification using quantized data. In *Proc. of IFAC Symposium on System Identification*, pages 255–260, Newcastle, Australia, 2006a.

L.Y. Wang, G.G. Yin, and J.F. Zhang. Joint identification of plant rational models and noise distribution functions using binary-valued observations. *Automatica*, 42:533–547, 2006b.

S. Zacks. *The Theory of Statistical Inference.* Wiley Series in Probability and Mathematical Statistics. Wiley, 1971.