Multiterminal Estimation - Extensions and a Geometric Interpretation.

Rebecka Jörnsten and Bin Yu
Department of Statistics, UC Berkeley
`rebecka,binyu@stat.berkeley.edu`

Abstract

In multiterminal estimation the basic theoretical question is to prove the existence of encoding and decoding schemes that can achieve a certain rate of compression, while resulting in a particular statistical estimation efficiency. This is by comparison a much less studied field than multiterminal source coding. Essentially, only two approaches have been reported. Zhang and Berger [3] established an upper bound on the asymptotic estimation efficiency under certain rate-compatibility constraints, and a test channel constraint referred to as the *solvability condition.* Han and Amari [2] tightened the upper bound under weaker constraints on both the rates and the test channel distributions. However, their bound is in most cases prohibitively complex to compute. Here we unify the two approaches. We are able to construct an upper bound that is asymptotically equal to Han and Amari's bound, under the same rate compatibility conditions. Our bound is valid under weaker constraints on the test channels than those of Zhang and Berger. Moreover, the bound is easily computed for most source distributions. We also present a new geometric interpretation of the upper bound on asymptotic estimation efficiency.

1 Introduction

Assume that the sources X^n and Y^n are i.i.d. according to $p_\theta(x, y)$, where θ is a (possibly vector valued) parameter. We assume the existence of an estimator $\tilde{\theta}(X^n, Y^n)$ which is asymptotically unbiased. We assume the estimator has an asymptotic (co)variance index $V(\theta)$ where

$$\lim_{n\to\infty} nV(\tilde{\theta}(X^n, Y^n)) = V(\theta),$$

a quantity that depends on the true value of θ. If we restrict the transmission rate of source X to R_1 bits, and the rate of source Y to R_2 bits, how much estimation efficiency of the parameter θ can we hope to retain? We encode the data strings with encoding functions f and g respectively and form an estimator $\hat{\theta}(f(X^n), g(Y^n))$, where we place the following rate constraint on the encoding functions;

$$\frac{1}{n}\log(||f||) \leq R_1,\ \frac{1}{n}\log(||g||) \leq R_2,$$

where $||.||$ denotes the cardinality. We assume that the estimator $\hat{\theta}(f(X^n), g(Y^n))$ is asymptotically unbiased and that there exists a (co)variance index

$$\lim_{n\to\infty} nV(\hat{\theta}(f(X^n), g(Y^n))) = V(\theta|R_1, R_2).$$

The compression of the data sources leads to a loss of information about the parameter θ such that

$$V(\theta|R_1, R_2) \geq V(\theta).$$

However, if this loss of estimation efficiency is minor compared to $V(\theta)$ itself, we can conclude that compression does not seriously affect estimation.

2 The approaches of Zhang and Berger and Han and Amari

To prove the existence of encoding functions f and g and estimators $\hat{\theta}$ that achieve a certain covariance index $V(\theta|R_1, R_2)$ for given rate constraints, two approaches exist, by Zhang and Berger (1988) and Han and Amari (1995,1998), respectively. They are similar with respect to the information theoretic coding arguments used, but widely different in the approach to establishing the achievable covariance index of a given code. Han and Amari provide an upper bound on the covariance index, which is tight if the *optimal* coding function f and g are given. The bound is difficult to compute for even quite simple data source distributions. Zhang and Berger give an upper bound which exceeds, or equals the bound of Han and Amari. They place stronger constraints on the encoding functions. Even so, their bound is simple to compute and applies to continuous data sources.

The existence of encoding functions f and g that provide 'good' estimates for a parameter θ, are proven via universal coding arguments. For simplicity we will restrict the discussion to discrete data sources. Recall that the data *type* is used to denote the relative frequencies of each letter outcome in a data string. Thus if data source X is distributed over alphabet $\{1, 2, ..., M\}$ the type of the data string X^n is

$$t(X^n) = (\sum_{i=1}^{n} 1_{\{X_i=1\}}, \sum_{i=1}^{n} 1_{\{X_i=2\}}, \cdots, \sum_{i=1}^{n} 1_{\{X_i=M\}})/n.$$

We can introduce joint types for X^n, Y^n, and conditional types in a similar fashion. The standard approach in information theory to proving existence of effective codes is to introduce auxiliary variables, which form a collection of codewords for each data source. These auxiliary variables or codewords U and V are generated according to the "test channel" distributions defined by $p_\theta(u|x)$ and $p_\theta(v|y)$. The auxiliary variables and the data sources thus form a Markov chain;

$$U \to X \to Y \to V.$$

The test channel distributions depend on θ only through the marginal distributions $p_\theta(x)$ and $p_\theta(y)$ respectively. Since the true value of θ is unknown we approximate the test channels by $p_{t_X}(u|x)$ and $p_{t_Y}(v|y)$, where t_X, t_Y are the marginal types of the data sequences X^n and Y^n. We construct a large set of such codewords for each data type and use a random mapping assignment to members of the codebooks. This is the first step of encoding. It can be shown that the rate constraints map into restrictions on the test channel distributions. Zhang and Berger (1988) proved the existence of codes f, g under the rate constraints imposed by the random codebook mapping with exponentially decaying encoding error probability;

$$R_1 \geq I(U; X), \; R_2 \geq I(V; Y).$$

Han and Amari (1995, 1998) showed that these rate-compatibility constraints can be weakened by adding a second step of encoding. The regular encoding argument ($X \to U$, $Y \to V$) is followed by the binning of codewords U and V, and minimum-entropy decoding. The resulting constraints on the test channels can be expressed through the following inequalities;

$$R_1 \geq I(U;X|V),\ R_2 \geq I(V;Y|U),\ R_1 + R_2 \geq I(U,V;X,Y).$$

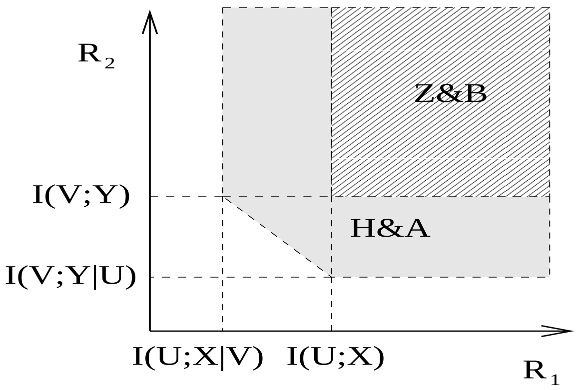

Figure 1: Rate compatibility conditions for the methods of Zhang and Berger (ZB), Han and Amari (HA).

Given encoding functions f and g we can construct an estimator $\hat{\theta}(f(X^n), g(Y^n))$. Here the encoding functions are either the result of the first encoding step (Zhang and Berger), or the first and second step and minimum entropy decoding (Han and Amari). We want to determine the covariance index of the estimator $\hat{\theta}(f,g)$. All the information about the parameter θ can be deduced from the observed (and decoded) data type t_{UXYV}. Han and Amari form the maximum likelihood estimate based on the distribution $p_\theta(t_{UXYV})$. By construction, the test channels place constraints on the data type t_{UXYV}. Han and Amari formulate these constraints through a projection operator H on the space of data types (marginals and joint). This generalized MLE is elegant and provides an upper bound bound on estimation efficiency of estimators $\hat{\theta}(f(X^n), g(Y^n))$ given a rate constraint (R_1, R_2), i.e.;

$$V(\theta|R_1,R_2) \leq V_{HA}(\theta|R_1,R_2) = (F_\theta(H))^{-1} + O(n^{-1/2}),$$

where F_θ is the Fisher information with respect to $p_\theta(t_{UXYV})$, a function of the projection matrix H. However, forming the projection operator H is no small feat, even for such limited and simple cases as binary data sources. Computing the bound on estimation efficiency using the techniques of Han and Amari ($V_{HA}(\theta|R_1,R_2)$), is therefore prohibitively complex for larger source alphabets, and moreover it is unclear whether this approach can be extended to continuous data sources, even in an abstract form.

Zhang and Berger (1988) construct their bounds on estimation efficiency by computing the *ensemble* mean and variance over the randomly generated codebooks (first step of encoding). Their argument is only valid for additive estimators, i.e. estimators such that

$$\tilde{\theta}(X^n, Y^n) = \frac{1}{n}\sum_{i=1}^{n} \tilde{\theta}(X_i, Y_i)$$

holds. Moreover, they assume that an additive estimator based on the encoded data can be obtained as the solution to the linear equation system

$$\sum_{u,v} p_\theta(u|x)p_\theta(v|y)\hat{\theta}(u,v) = \tilde{\theta}(x,y), \ \forall x,y.$$

This puts a very limiting constraint on the test channel distributions, but the additivity of the estimator $\hat{\theta}(f(X^n), g(Y^n))$ follows. For such *additive* estimators, a repeated random coding argument ensures the existence of encoder functions and estimators, whose means and variances come arbitrarily close to the ensemble quantities. Zhang and Berger thus avoid the construction of the data type distribution. In fact, computation of the efficiency bounds only requires moments under distribution $p_\theta(UXYV)$ (in contrast to $p_\theta(t_{UXYV})$ for V_{HA}). Zhang and Berger's upper bound on the asymptotic efficiency equals

$$V(\theta|R_1,R_2) \leq V_{ZB}(\theta|R_1,R_2) =$$

$$= V(\hat{\theta}(U,V)) + E[E(\tilde{\theta}|X)]^2 + E[E(\tilde{\theta}|Y)]^2 - E[E(\hat{\theta}|UX]^2 - E[E(\hat{\theta}|VY]^2 + O(n^{-1/2}).$$

3 Extending Zhang and Berger's approach

The limitation of Han and Amari's approach lies in the complexity of the distribution $p_\theta(t_{UXYV})$. The solvability condition, and the one-step encoding limits the approach of Zhang and Berger. In order to construct computable efficiency bounds we choose to extend the approach of Zhang and Berger. We place the following constraints on the test channel distributions. Assume an exponential family source distribution $p_\theta(x,y)$. Restrict the test channels $p_\theta(u|x)$ and $p_\theta(v|y)$ to map to another exponential family distribution $p_\theta(u,v)$ such that $I_\theta(U,X) > 0, I_\theta(V,Y) > 0$ holds. Refer to the canonical parameters of $p_\theta(u,v)$ as η. Note that η is not equal to η', the canonical parameters of $p_\theta(x,y)$. However, a reparameterization of $p_\theta(x,y)$ with parameters η, δ is in general possible. In e.g. the multinomial case η may correspond to linear combinations of outcome probabilities. Assume the existence of functionals $h : \eta \to \theta$ (where both η and θ may be vector valued) such that

$$|\partial^4_{i^q, j^{4-q}} h(\eta_1, .., \eta_i, .., \eta_j, ..)| < \infty, \ \forall i,j, \ q = 0, .., 4.$$

This restriction on the test channel distributions is stronger than Han and Amari who only need assume the existence of bounded and continuous first order derivatives of $p_\theta(u|x)$ with respect to $p_\theta(x)$ (and similarly for y). However, it is weaker than solvability condition, and can easily be verified in practice.

We compute the *decoder ensemble moments* of estimators $\hat{\eta}$, which are additive and asymptotically efficient. Using the decoder ensemble mean allows us to extend Zhang and Berger's rate compatibility region (Fig. 1) to that of Han and Amari. The binning of codewords and minimum entropy decoding is an additional random step over which to average. Under the rate compatibility conditions

$$R_1 \geq I(U;X|V), \ R_2 \geq I(V;Y|U), \ R_1 + R_2 \geq I(U,V;X,Y)$$

the encoder and decoder ensemble moments are asymptotically equal. For estimates of the canonical parameters we can thus establish a bound, which asymptotically coincides with the bound of Han and Amari since the ML estimates are indeed additive for η. For now, let η be the parameter of interest. The asymptotic efficiency bound equals

$$V(\eta|R_1, R_2) \leq V_{eZB}(\eta|R_1, R_2) = \tag{1}$$

$$= V(\hat{\eta}(U,V)) - E[E(\hat{\eta}(U,V)|X) - E(\hat{\eta}(U,V)|UX)]^2 - E[E(\hat{\eta}(U,V)|Y) - E(\hat{\eta}(U,V)|VY)]^2.$$

The test channel constraints on the data type distribution, which in Han and Amari's work entered through the projection operator H, is now featured in the second and third term of V_{eZB}. The first term is the variance of the estimator based on encoded information when these constraints are ignored, i.e. under the marginal distribution $p_\eta(U,V)$. The second and third term reduces this quantity by, as we see below, the variances over the constructed test channels.

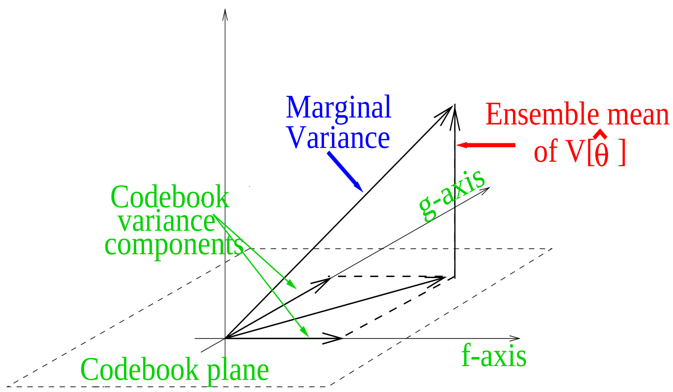

Figure 2: Geometric interpretation of the asymptotic estimation efficiency bound.

We can rewrite V_{eZB} as follows,

$$V_{eZB} = E_{f,g}[V(\hat{\eta}(f,g)|f,g)].$$

$E_{f,g}$ is an operator that corresponds to the ensemble mean over the decoded codebooks, f and g. We can identify the terms in equation (1) by those of a variance decomposition,

$$V_{eZB} = E_{f,g}[V(\hat{\eta}(f,g)|f,g)] = V(\hat{\eta}(f,g)) - V_{f,g}[E(\hat{\eta}(f,g)|f,g)] \tag{2}$$

The conditional expectation $E(\cdot|f,g)$ is a projection operator onto the *orthogonal* codebook components, since in the multiterminal setup the codebooks f and g are generated independently. Thus, the second and third term of equation (1) equal

$$V_{f,g}[E(\hat{\eta}(f,g)|f,g)] = V_f[E(\hat{\eta}(f,g)|f,g)_f] + V_g[E(\hat{\eta}(f,g)|f,g)_g],$$

where the subscripts f and g denote the orthogonal components. A geometric illustration is shown in Figure 2. The quantity of interest is the ensemble mean of the variance of the estimator, $E_{f,g}[V(\hat{\eta}(f,g)|f,g)]$. Obviously, the marginal variance $V(\hat{\eta}(f,g))$ is an overestimate of this quantity since the codebook construction is known (i.e. determined by 0-rate quantities). Thus, this overestimate is corrected by removing the portion of the variance of $\hat{\eta}$ that we control, i.e. the variance over the constructed codebooks. With independent codebook components we get the expression in equation (2).

With the restriction placed on the test channels through the function h we can now construct a bound for an estimator $\hat{\theta}$ using the bound for $\hat{\eta}$. By a delta-method argument we can show that

$$V(\theta|R_1,R_2) \leq VeZB(\theta|R_1,R_2) = \sum_{i:\partial_i h(\eta)\neq 0} |\partial_i h(\eta)|^2 [V(\hat{\eta}_i(f,g)) - V_{f,g}[E(\hat{\eta}_i(f,g)|f,g)] +$$

$$+ \sum_{i,j:\partial_{i,j} h(\eta)\neq 0} (\partial_{i,j} h(\eta))[Cov(\hat{\eta}_i(f,g),\hat{\eta}_j(f,g)) - Cov_{f,g}[E(\hat{\eta}_i(f,g)|f,g), E(\hat{\eta}_j(f,g)|f,g)] + O(n^{-1/2}).$$

The construction of the covariance bounds proceeds in a similar fashion to the above and is left out to conserve space.

We conclude with a simple example. Assume a bivariate Gaussian model: $p_\theta(x^n,y^n)$, $\theta = (\sigma_x^2,\sigma_y^2,\rho)$. We use the test channels $p(u|x) \sim N(x,\sigma_n^2)$ and $p(v|y) \sim N(y,\sigma_m^2)$ which maps between exponential families. The V_{eZB} bound still applies if we discretize (U,X,Y,V) to $(\tilde{U},\tilde{X},\tilde{Y},\tilde{V})$, with the number discretization levels growing with the sample size n at rate $O(n^\alpha)$, where $\alpha \in (\frac{1}{2},1)$. With α in this range the marginal types can still be transmitted at 0-rate, and the resulting summary statistics differ from those of the continuous distributions by no more than $O(n^{-1})$. Let ρ be the parameter of interest. It is easy to verify that the condition on h applies for this parameter and the given test channels. We compute the V_{eZB} bound as

$$nV(\hat{\rho}(h(f(X^n),g(Y^n))) \leq (1-\rho^2)^2 + (\frac{1}{2^{2R_1}-1} + \frac{1}{2^{2R_2}-1}) + (\frac{1}{2^{2R_1}-1}\frac{1}{2^{2R_2}-1}) +$$

$$-\rho^2(\frac{1}{2^{2R_1}-1} + \frac{1}{2^{2R_2}-1} + \frac{1}{2^{2R_2}(2^{2R_1}-1)} + \frac{1}{2^{2R_1}(2^{2R_2}-1)})$$

Zhang and Berger's bound is given by

$$nV(\hat{\rho}(f(X^n),g(Y^n)) \leq (1+\rho^2) + (\frac{1}{2^{2R_1}-1} + \frac{1}{2^{2R_2}-1}) + (\frac{1}{2^{2R_1}-1}\frac{1}{2^{2R_2}-1}) +$$

$$-\rho^2(\frac{1}{2^{2R_1}-1} + \frac{1}{2^{2R_2}-1}),$$

which is obviously larger than V_{eZB}. Their bound also uses a suboptimal full-data estimator as the baseline for comparison, which is reflected in the first term $(1+\rho^2)$. The interpretation of the bound is in this example particularly simple. It corresponds to estimating the parameter ρ from noisy Gaussian data when the signal-to-noise ratios, or equivalently the noise variances, are known.

References

[1] Amari, S. and Han, T. (1998). Statistical Inference Under Multiterminal Data compression. *IEEE Trans. Inform. Theory*, **44(6)** 2300-2323.

[2] Han, T. and Amari, S-I. (1995). Parameter Estimation with Multiterminal Data Compression. *IEEE Trans. Inform. Theory*, **41(6)**, 1802-1833.

[3] Zhang, Z. and Berger, T. (1988). Estimation via Compressed Information. *IEEE Trans. Inform. Theory*, **34(2)**, 198-211.