Information Inequalities and the Central Limit Theorem

Andrew Barron

Yale University, Department of Statistics

Presentation at Boston University, March 1, 2007

Outline

- Entropy and the Central Limit Problem*
- Entropy Power Inequality (EPI)
- Monotonicity of Entropy and new subset sum EPI**
- Variance Drop Lemma**
- Projection and Fisher Information
- Rates of Convergence in the CLT***

* Andrew Barron, Annals of Probability 1986

** Mokshay Madiman and Andrew Barron, ISIT 2006 and IEEE IT Submitted

*** Oliver Johnson and Andrew Barron, PTRF 2004

Entropy Basics

- For a mean zero random variable X with density $f(x)$ and finite variance $\sigma^2 = 1$,

  the differential entropy is $H(X) = E[\log \frac{1}{f(X)}]$

  the entropy power of X is $e^{2H(X)}/2\pi e$

- For a Normal$(0, \sigma^2)$ random variable Z, with density function ϕ,

  the differential entropy is $H(Z) = (1/2)\log(2\pi e \sigma^2)$

  the entropy power of Z is σ^2

- The relative entropy is $D(f||\phi) = \int f(x) \log \frac{f(x)}{\phi(x)} dx$

  it is non-negative: $D(f||\phi) \geq 0$ with equality iff $f = \phi$

  it is larger than $(1/2)||f - \phi||_1^2$

Maximum entropy property

Boltzmann, Jaynes, Shannon

Let Z be a normal random variable with the same mean and variance as a random variable X, then $H(X) \leq H(Z)$ with equality iff X is normal

The relative entropy quantifies the entropy gap

$$H(Z) - H(X) = D(f||\phi)$$

Maximum entropy property

Boltzmann, Jaynes, Shannon

Let Z be a normal random variable with the same mean and variance as a random variable X, then $H(X) \leq H(Z)$ with equality iff X is normal.

The relative entropy quantifies the entropy gap. Indeed, this is Kullback's proof of the maximum entropy property

$$\begin{aligned} H(Z) - H(X) &= \int \phi(x) \log \frac{1}{\phi(x)} dx - \int f(x) \log \frac{1}{f(x)} dx \\ &= \int f(x) \log \frac{1}{\phi(x)} dx - \int f(x) \log \frac{1}{f(x)} dx \\ &= \int f(x) \log \frac{f(x)}{\phi(x)} dx \\ &= D(f||\phi) \\ &\geq 0 \end{aligned}$$

Here $\log \frac{1}{\phi(x)} = \frac{x^2}{2\sigma^2} \log e + \frac{1}{2} \log 2\pi\sigma^2$ is quadratic in x, so both f and ϕ give it the same expectation, which is $\frac{1}{2} \log 2\pi e \sigma^2$.

Fisher Information Basics

- For a mean zero random variable X with differentiable density $f(x)$ and finite variance $\sigma^2 = 1$,

  the score function is $score(X) = \frac{d}{dx} \log f(x)$

  the Fisher information is $I(X) = E[score^2(X)]$.

- For a Normal$(0, \sigma^2)$ random variable Z, with density function ϕ,

  the score function is linear $score(Z) = -Z/\sigma^2$

  the Fisher information is $I(Z) = 1/\sigma^2$

- The relative Fisher information is $J(f||\phi) = \int f(x) \left(\frac{d}{dx} \log \frac{f(x)}{\phi(x)} \right)^2 dx$

  it is non-negative

  it is larger than $D(f||\phi)$

- Minimum Fisher info property (Cramer-Rao ineq): $I(X) \geq 1/\sigma^2$

  equality iff Normal

- The information gap satisfies: $I(X) - I(Z) = J(f||\phi)$

The Central Limit Problem

For independent identically distributed random variables $X_1, X_2, \ldots, X_n$, with $E[X] = 0$ and $VAR[X] = \sigma^2 = 1$, consider the standardized sum

$$\frac{X_1 + X_2 + \ldots + X_n}{\sqrt{n}}.$$

Let its density function be f_n and its distribution function F_n.

Let the standard normal density be ϕ and its distribution function Φ.

Natural questions:

- In what sense do we have convergence to the normal?
- Do we come closer to the normal with each step?
- Can we give clean bounds on the "distance" from the normal and a corresponding rate of convergence?

Convergence

- **In distribution:** $F_n(x) \to \Phi(x)$

  Classical via Fourier methods or expansions of expectations of smooth functions.

  Linnick 59, Brown 82 via info measures applied to smoothed distributions.

- **In density:** $f_n(x) \to \phi(x)$

  Prohorov 52 showed $||f_n - \phi||_1 \to 0$ iff f_n exists eventually.

  Kolmogorov & Gnedenko 54 $||f_n - \phi||_\infty \to 0$ iff f_n bounded eventually.

- **In Shannon Information:** $H(\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i) \to H(Z)$

  Barron 86 shows $D(f_n||\phi) \to 0$ iff it is eventually finite.

- **In Fisher Information:** $I(\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i) \to 1/\sigma^2$

  Johnson & Barron 04 shows $J(f_n||\phi) \to 0$ iff it is eventually finite.

Original Entropy Power Inequality

Shannon 48, Stam 59: For independent random variables with densities,

$$e^{2H(X_1+X_2)} \geq e^{2H(X_1)} + e^{2H(X_2)}$$

where equality holds if and only if the X_i are normal.

Also

$$e^{2H(X_1+\ldots+X_n)} \geq \sum_{j=1}^{n} e^{2H(X_j)}$$

Original Entropy Power Inequality

Shannon 48, Stam 59: For independent random variables with densities,

$$e^{2H(X_1+X_2)} \geq e^{2H(X_1)} + e^{2H(X_2)}$$

where equality holds if and only if the X_i are normal.

Central Limit Theorem Implication

For X_i i.i.d., let $H_n = H\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\right)$

- nH_n is superadditive

$$H_{n_1+n_2} \geq \frac{n_1}{n_1+n_2} H_{n_1} + \frac{n_2}{n_1+n_2} H_{n_2}$$

- monotonicity for doubling sample size

$$H_{2n} \geq H_n$$

- The superadditivity of nH_n and the monotonicity for the powers of two subsequence are key in the proof of entropy convergence [Barron '86]

Leave-one-out Entropy Power Inequality

Artstein, Ball, Barthe and Naor 2004 (ABBN): For independent X_i

$$e^{2H(X_1+\ldots+X_n)} \geq \frac{1}{n-1}\sum_{i=1}^{n} e^{2H\left(\sum_{j\neq i} X_j\right)}$$

Remarks

- This strengthens the original EPI of Shannon and Stam.
- ABBN's proof is elaborate.
- Our proof (Madiman & Barron 2006) uses familiar and simple tools and proves a more general result, that we present.
- The leave-one-out EPI implies in the iid case that entropy is increasing:

$$H_n \geq H_{n-1}$$

- A related proof of monotonicity is developed contemporaneously in Tulino & Verdú 2006.
- Combining with Barron 1986 the monotonicity implies

$$H_n \nearrow H(\text{Normal}) \quad \text{and} \quad D_n = \int f_n \log \frac{f_n}{\phi} \searrow 0$$

New Entropy Power Inequality

Subset-sum EPI (Madiman and Barron)

For any collection $\mathcal{S}$ of subsets s of indices $\{1, 2, \ldots, n\}$,

$$e^{2H(X_1+\ldots+X_n)} \geq \frac{1}{r(\mathcal{S})} \sum_{s \in \mathcal{S}} e^{2H(\text{sum}_s)}$$

where $\text{sum}_s = \sum_{j \in s} X_j$ is the subset-sum

$r(\mathcal{S})$ is the *prevalence*, the maximum number of subsets in $\mathcal{S}$ in which any index i can appear

Examples

- $\mathcal{S}$=singletons, $r(\mathcal{S}) = 1$, original EPI
- $\mathcal{S}$=leave-one-out sets, $r(\mathcal{S}) = n{-}1$, ABBN's EPI
- $\mathcal{S}$=sets of size m, $r(\mathcal{S}) = \binom{n-1}{m-1}$, leave $n{-}m$ out EPI
- $\mathcal{S}$=sets of m consecutive indices, $r(\mathcal{S}) = m$

New Entropy Power Inequality

Subset-sum EPI

For any collection $\mathcal{S}$ of subsets s of indices $\{1, 2, \ldots, n\}$,

$$e^{2H(X_1+\ldots+X_n)} \geq \frac{1}{r(\mathcal{S})} \sum_{s \in \mathcal{S}} e^{2H(\mathsf{sum}_s)}$$

Discriminating and balanced collections $\mathcal{S}$

- *Discriminating* if for any i, j, there is a set in $\mathcal{S}$ containing i but not j
- *Balanced* if each index i appears in the same number $r(\mathcal{S})$ of sets in $\mathcal{S}$

Equality in the Subset-sum EPI

For discriminating and balanced $\mathcal{S}$, equality holds in the subset-sum EPI if and only if the X_i are normal

In this case, it becomes $\sum_{i=1}^{n} a_i = \frac{1}{r(\mathcal{S})} \sum_{s \in \mathcal{S}} \sum_{i \in s} a_i$ with $a_i = \mathsf{Var}(X_i)$

New Entropy Power Inequality

Subset-sum EPI

For any collection $\mathcal{S}$ of subsets s of indices $\{1, 2, \ldots, n\}$,

$$e^{2H(X_1+\ldots+X_n)} \geq \frac{1}{r(\mathcal{S})} \sum_{s\in\mathcal{S}} e^{2H(\mathrm{sum}_s)}$$

CLT Implication

Let X_i be independent, but not necessarily identically distributed.

The entropy of variance-standardized sums increases "on average":

$$H\left(\frac{\mathrm{sum}_{\mathrm{total}}}{\sigma_{\mathrm{total}}}\right) \geq \sum_{s\in\mathcal{S}} \lambda_s \, H\left(\frac{\mathrm{sum}_s}{\sigma_s}\right)$$

where

- $\sigma^2_{\mathrm{total}}$ is the variance of $\mathrm{sum}_{\mathrm{total}} = \sum_{i=1}^n X_i$ and σ_s^2 is the variance of $\mathrm{sum}_s = \sum_{j\in s} X_j$
- The weights $\lambda_s = \frac{\sigma_s^2}{r(\mathcal{S})\sigma^2_{\mathrm{total}}}$ are proportional to σ_s^2
- The weights add to 1 for balanced collections $\mathcal{S}$

New Fisher Information Inequality

For independent X_1, $X_2, \ldots, X_n$ with differentiable densities,

$$\frac{1}{I(\text{sum}_{\text{total}})} \geq \frac{1}{r(\mathcal{S})} \sum_{s \in \mathcal{S}} \frac{1}{I(\text{sum}_s)}$$

Remarks

- This extends Fisher information inequalities of Stam and ABBN
- Recall from Stam '59

$$\frac{1}{I(X_1 + \ldots + X_n)} \geq \frac{1}{I(X_1)} + \ldots + \frac{1}{I(X_n)}$$

- For discriminating and balanced $\mathcal{S}$, equality holds iff the X_i are normal

New Fisher Information Inequality

For independent X_1, $X_2, \ldots, X_n$ with differentiable densities,

$$\frac{1}{I(\text{sum}_{\text{total}})} \geq \frac{1}{r(\mathcal{S})} \sum_{s \in \mathcal{S}} \frac{1}{I(\text{sum}_s)}$$

CLT Implication

- For i.i.d. X_i , let $I_n = I\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i\right)$

  The Fisher information I_n is a decreasing sequence:

  $$I_n \leq I_{n-1} \qquad \text{[ABBN '04]}$$

  Combining with Johnson and Barron '04 implies $I_n \searrow I(\text{Normal})$ and

  $$J(f_n || \phi) \searrow 0$$

- For i.n.i.d. X_i, the Fisher info. of standardized sums decreases on average

$$I\left(\frac{\text{sum}_{\text{total}}}{\sigma_{\text{total}}}\right) \leq \sum_{s \in \mathcal{S}} \lambda_s \, I\left(\frac{\text{sum}_s}{\sigma_s}\right)$$

The Link between H and I

Definitions

- Shannon entropy: $H(X) = E\left[\log \frac{1}{f(X)}\right]$
- Score function: $\text{score}(X) = \frac{\partial}{\partial x} \log f(X)$
- Fisher information: $I(X) = E\left[\text{score}^2(X)\right]$

Relationship

For a standard normal Z independent of X,

- Differential version:

$$\frac{d}{dt} H(X + \sqrt{t}Z) = \frac{1}{2} I(X + \sqrt{t}Z)$$ [de Bruijn, see Stam '59]

- Integrated version:

$$H(X) = \frac{1}{2}\log(2\pi e) - \frac{1}{2}\int_0^\infty \left[I(X + \sqrt{t}Z) - \frac{1}{1+t} \right] dt$$ [Barron '86]

The Projection Tool

For each subset s,

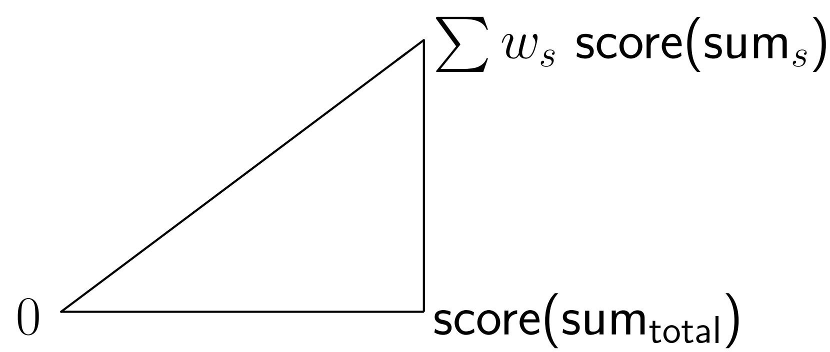
$$\text{score}(\text{sum}_{\text{total}}) = E\left[\text{score}(\text{sum}_s) \,\middle|\, \text{sum}_{\text{total}}\right]$$

Hence, for weights w_s that sum to 1,

$$\text{score}(\text{sum}_{\text{total}}) = E\left[\sum_{s\in\mathcal{S}} w_s \ \text{score}(\text{sum}_s) \,\middle|\, \text{sum}_{\text{total}}\right]$$

Pythagorean inequality

The Fisher info. of the sum is the mean squared length of the projection



$$I(\text{sum}_{\text{total}}) \le E\left[\sum_{s\in\mathcal{S}} w_s \ \text{score}(\text{sum}_s)\right]^2$$

The Heart of the Matter

Recall the Pythagorean inequality

$$I(\mathsf{sum}_{\mathsf{total}}) \leq E\left[\sum_{s\in\mathcal{S}} w_s \ \mathsf{score}(\mathsf{sum}_s)\right]^2$$

and apply the variance drop lemma to get

$$I(\mathsf{sum}_{\mathsf{total}}) \leq r(\mathcal{S}) \sum_{s\in\mathcal{S}} w_s^2 I(\mathsf{sum}_s)$$

The Variance Drop Lemma

Let X_1, $X_2, \ldots, X_n$ be independent. Let $\underline{X}_s = (X_i : i \in s)$ and $g_s(\underline{X}_s)$ be some mean-zero function of $\underline{X}_s$. Then sums of such functions

$$g(X_1, X_2, \ldots, X_n) = \sum_{s \in \mathcal{S}} g_s(\underline{X}_s)$$

have the variance bound

$$Eg^2 \leq r(\mathcal{S}) \sum_{s \in \mathcal{S}} Eg_s^2(\underline{X}_s)$$

The Variance Drop Lemma

Let X_1, $X_2, \ldots, X_n$ be independent. Let $\underline{X}_s = (X_i : i \in s)$ and $g_s(\underline{X}_s)$ be some mean-zero function of $\underline{X}_s$. Then sums of such functions

$$g(X_1, X_2, \ldots, X_n) = \sum_{s \in \mathcal{S}} g_s(\underline{X}_s)$$

have the variance bound

$$Eg^2 \le r(\mathcal{S}) \sum_{s \in \mathcal{S}} Eg_s^2(\underline{X}_s)$$

Remarks

- Note that $r(\mathcal{S}) \le |\mathcal{S}|$, hence the "variance drop"
- Examples:

  $\mathcal{S}$=singletons has $r = 1$: additivity of variance with independent summands

  $\mathcal{S}$=leave-one-out sets has $r = n-1$ as in the study of the jackknife and U-statistics
- Proof is based on ANOVA decomposition [Hoeffding '48, Efron and Stein '81]
- Introduced in leave-one-out case to info. inequality analysis by ABBN '04

Optimized Form for I

We have, for all weights w_s that sum to 1,

$$I(\mathsf{sum}_{\mathsf{total}}) \le r(\mathcal{S}) \sum_{s \in \mathcal{S}} w_s^2 I(\mathsf{sum}_s)$$

Optimizing over w yields the new Fisher information inequality

$$\frac{1}{I(\mathsf{sum}_{\mathsf{total}})} \ge \frac{1}{r(\mathcal{S})} \sum_{s \in \mathcal{S}} \frac{1}{I(\mathsf{sum}_s)}$$

Optimized Form for H

We have (again)

$$I(\mathsf{sum}_{\mathsf{total}}) \leq r(\mathcal{S}) \sum_{s \in \mathcal{S}} w_s^2 I(\mathsf{sum}_s)$$

Equivalently,

$$I(\mathsf{sum}_{\mathsf{total}}) \leq \sum_{s \in \mathcal{S}} w_s I\left(\frac{\mathsf{sum}_s}{\sqrt{r(\mathcal{S}) w_s}}\right)$$

Adding independent normals and integrating,

$$H(\mathsf{sum}_{\mathsf{total}}) \geq \sum_{s \in \mathcal{S}} w_s H\left(\frac{\mathsf{sum}_s}{\sqrt{r(\mathcal{S}) w_s}}\right)$$

Optimizing over w yields the new Entropy Power Inequality

$$e^{2H(\mathsf{sum}_{\mathsf{total}})} \geq \frac{1}{r(\mathcal{S})} \sum_{s \in \mathcal{S}} e^{2H(\mathsf{sum}_s)}$$

Fisher information and M.M.S.E. Estimation

Model: $Y = X + Z$
where $Z \sim N(0,1)$ and X is to be estimated

- Optimal estimate: $\hat{X} = E[X|Y]$

  Fact: score$(Y) = \hat{X} - Y$

  Note: $X - \hat{X}$ and $\hat{X} - Y$ are orthogonal, and sum to $-Z$

  Hence: $I(Y) = E(\hat{X} - Y)^2 = 1 - E(X - \hat{X})^2$
  $= 1 -$ Minimal M.S.E.

  From L.D. Brown '70's [c.f. the text of Lehmann and Casella '98]

- Thus derivative of entropy can be expressed equivalently in terms of either $I(Y)$ or minimal M.S.E.

- Guo, Shamai and Verdú, 2005 use the minimal M.S.E. interpretation to give a related proof of the EPI and Tulino and Verdú 2006 use this M.S.E. interpretation to give a related proof of monotonicity in the CLT

Recap: Subset-sum EPI

For any collection $\mathcal{S}$ of subsets s of indices $\{1, 2, \ldots, n\}$,

$$e^{2H(\mathsf{sum}_{\mathsf{total}})} \geq \frac{1}{r(\mathcal{S})} \sum_{s \in \mathcal{S}} e^{2H(\mathsf{sum}_s)}$$

- Generalizes original EPI and ABBN's EPI
- Simple proof using familiar tools
- Equality holds for normal random variables

Comment on CLT rate bounds

For iid X_i let

$$J_n = J(f_n||\phi)$$

and

$$D_n = D(f_n||\phi)$$

Suppose the distribution of the X_i has a finite Poincaré constant R.

Using the pythagorean identity for score projection, Johnson & Barron '04 show:

$$J_n \leq \frac{2R}{n} J_1$$

$$D_n \leq \frac{2R}{n} D_1$$

- Implies a $1/\sqrt{n}$ rate of convergence in distribution, known to hold for random variables with non-zero finite third moment.
- Our finite Poincaré assumption implies finite moments of all orders.
- Do similar bounds on information distance hold assuming only finite initial information distance and finite third moment?

Summary

Two ingredients

- score of sum $=$ projection of scores of subset-sums
- variance drop lemma

yield the conclusions

- existing Fisher information and entropy power inequalities
- new such inequalities for arbitrary collections of subset-sums
- monotonicity of I and H in central limit theorems

refinements using the pythagorean identity for the score projection yield

- convergence in information to the Normal
- order $1/n$ bounds on information distance from the Normal

○ — ○ — ○