The Monotonicity of Information in the Central Limit Theorem and Entropy Power Inequalities

Andrew Barron and Mokshay Madiman

Original Entropy Power Inequality

For independent random variables with densities,

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

[Shannon '48, Stam '59]

- When X has density $f(x)$,

  the differential entropy is $H(X) = E[-\log f(X)]$

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

- Also

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

- Equality holds if and only if the X_i are normal

ABBN's Entropy Power Inequality

Leave-one-out Inequality for independent X_i

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

[Artstein, Ball, Barthe and Naor '04 (ABBN)]

Remarks

- This strengthens the original EPI of Shannon and Stam
- ABBN's proof is elaborate; ours uses familiar tools and proves a more general result

CLT Implications

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

- Original EPI implies nH_n is superadditive and

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

- ABBN's EPI implies that entropy is an increasing sequence:

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

- Combining with Barron '86 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

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 $\{1, 2, \ldots, n\}$,

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

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 $a_i = \text{Var}(X_i)$

New Entropy Power Inequality

Subset-sum EPI

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

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

CLT Implication

Let X_i be independent, but not necessarily identically distributed.

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

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

where

- σ^2_{total} is the variance of $\text{sum}_{\text{total}} = \sum_{i=1}^n X_i$ and σ_s^2 is the variance of $\text{sum}_s = \sum_{j \in s} X_j$
- The weights $\lambda_s = \frac{\sigma_s^2}{r(\mathcal{S})\sigma^2_{\text{total}}}$ 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 $I_n \searrow I(\text{Normal})$

- 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,

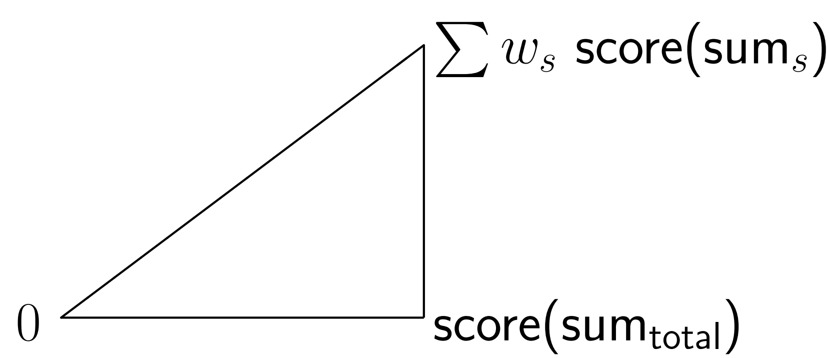
$$\text{score}(\text{sum}_{\text{total}}) = E\left[\text{score}(\text{sum}_s) \,\middle|\, \text{sum}_{\text{total}}\right]$$

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

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

and apply the variance drop lemma to get

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

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,

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

Optimizing over w yields the new Fisher information inequality

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

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, Verdú, Tulino '05-'06 use the minimal M.S.E. interpretation to give a related proof of the original and leave-one-out EPI's, and associated monotonicity

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

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

○ — ○ — ○