Celebrations for
three influential scholars in Information Theory and Statistics:

- Tom Cover: On the occasion of his 70th birthday
  Coverfest.stanford.edu
  Elements of Information Theory Workshop
  Stanford University, May 16, 2008
- Imre Csiszár: On the occasion of his 70th birthday
  www.renyi.hu/∼infocom
  Information and Communication Conference
  Renyi Institute, Budapest, August 25-28, 2008
- Jorma Rissanen: On the occasion of his 75th birthday
  Festschrift at www.cs.tut.fi/∼tabus/
  presented at the *IEEE Information Theory Workshop*
  Porto, Portugal, May 8, 2008: THIS MORNING!

MDL, Penalized Likelihood and Statistical Risk

Andrew Barron
Department of Statistics
Yale University

Coauthors: Jonathan Li, Cong Huang, Xi Luo

May 8, 2008
ITW - Porto, Portugal

On the Occasion of the Festschrift for Jorma Rissanen

Outline

Outline

Shannon Codes

- Kraft-McMillan characterization:
  Uniquely decodeable codelengths

$$L(\underline{x}), \quad \underline{x} \in \underline{\mathcal{X}}, \qquad \qquad \sum_{\underline{x}} 2^{-L(\underline{x})} \leq 1$$

$$L(\underline{x}) = \log 1/p(\underline{x}) \qquad \qquad p(\underline{x}) = 2^{-L(\underline{x})}$$

- Operational meaning of probability:
  A probability distribution p is given by a choice of code

Codelength Comparison

- Targets p^* are possible distributions
- Compare codelength $\log 1/p(\underline{x})$ to alternatives $\log 1/p^*(\underline{x})$
- Redundancy or regret

$$\left[\log 1/p(\underline{x}) - \log 1/p^*(\underline{x}) \right]$$

- Expected redundancy

$$D(P^*_{\underline{X}} \| P_{\underline{X}}) = E_{P^*}\left[\log \frac{p^*(\underline{X})}{p(\underline{X})} \right]$$

Universal Codes

- MODELS

  Family of coding strategies $\Leftrightarrow$ Family of prob. distributions

$$\{L_{\theta,m}(\underline{x}) : \theta \in \Theta_m\} \Leftrightarrow \{p_{\theta,m}(\underline{x}) : \theta \in \Theta_m\}$$

  Model index $m \in \mathcal{M}$

- Universal codes $\Leftrightarrow$ Universal probabilities $q_m(\underline{x})$

$$L_m(\underline{x}) = \log 1/q_m(\underline{x})$$

- Redundancy: $\left[\ \log 1/q_m(\underline{x}) - \log 1/p_{\theta,m}(\underline{x})\ \right]$

  Want it small either uniformly in $\underline{x}, \theta$ or in expectation

Statistical Aim

- Training data $\underline{x}$ $\Rightarrow$ estimator $\hat{p} = p_{\hat{\theta},\hat{m}}$
- Subsequent data $\underline{x}'$
- Want $\log 1/\hat{p}(\underline{x}')$ to compare favorably to $\log 1/p_{\theta,m}(\underline{x}')$
- Likewise for p^* close to but not necessarily in the families

Loss

- Kullback Information-divergence:

$$D(P^*_{\underline{X}'} \| P_{\underline{X}'}) = E\left[\, \log p^*(\underline{X}')/p(\underline{X}')\,\right]$$

- Bhattacharyya, Hellinger, Chernoff, Rényi divergence:

$$d(P^*_{\underline{X}'}, P_{\underline{X}'}) = 2 \log 1/E[p(\underline{X}')/p^*(\underline{X}')]^{1/2}$$

- Product model case: $p(\underline{x}') = \prod_{i=1}^n p(x_i')$

$$D(P^*_{\underline{X}'} \| P_{\underline{X}'}) = n\, D(P^* \| P)$$

Likewise $$d(P^*_{\underline{X}'}, P_{\underline{X}'}) = n\, d(P^*, P)$$

Loss

- Relationship:

$$d(P^*, P) \leq D(P^* \| P)$$

- and, if the log density ratio is not more than B, then

$$D(P^* \| P) \leq C_B \, d(P^*, P)$$

with $C_B \leq 2 + B$

Outline

MDL

- Universal coding brought into statistical play
- Minimum Description Length Principle:

  The shortest code for data gives the best statistical model

MDL: Two-stage Version

- Two-stage codelength (parametric case):

$$L(\underline{x}) = \min_{m} \min_{\theta \in \Theta_m} \left[\log 1/p_{\theta,m}(\underline{x}) + L(\theta, m) \right]$$

bits for $\underline{x}$ given θ, m $+$ bits for θ, m

Corresponding statistical estimator $\hat{p} = p_{\hat{\theta},\hat{m}}$

- Two-stage codelength (function case):

$$L(\underline{x}) = \min_{f \in \mathcal{F}} \left[\log 1/p_f(\underline{x}) + L(f) \right]$$

Corresponding statistical estimator $\hat{p} = p_{\hat{f}}$

MDL: Two-stage Version

- Two-stage codelength (parametric case):

$$L(\underline{x}) = \min_{m} \min_{\theta \in \Theta_m} \left[\log 1/p_{\theta,m}(\underline{x}) + L(\theta, m) \right]$$

bits for $\underline{x}$ given θ, m $+$ bits for θ, m

Corresponding statistical estimator $\hat{p} = p_{\hat{\theta},\hat{m}}$

- Typically $L(\theta, m)$ is of order

$$\frac{\dim(\Theta_m)}{2} \log n \; + \; L(m)$$

MDL: Mixture and Predictive Versions

- Codelength based on a selection of mixture models

$$L(\underline{x}) = \min_m \left[\log \frac{1}{\int_{\Theta_m} p_m(\underline{x}|\theta) w_m(\theta) d\theta} \quad + \quad L(m) \right]$$

describe $\underline{x}$ given m $\quad + \quad$ describe m

average case optimal

- Corresponding statistical estimators are $\hat{m}$ and

$$\hat{p}(\underline{x}') = p_m(\underline{x}'|\underline{x}) = \frac{\int p_m(\underline{x}'|\theta) p_m(\underline{x}|\theta) w_m(\theta) d\theta}{\int p_m(\underline{x}|\theta) w_m(\theta) d\theta}$$

which is a predictive distribution

MDL: Predictive Version

- Codelength based on predictive distributions

$$L(\underline{x}) = \log \frac{1}{p(x_1)} + \log \frac{1}{p(x_2|x_1)} + \ldots \log \frac{1}{p(x_n|x_1, \ldots, x_{n-1})}$$

- Corresponding statistical estimator at $x' = x_{n+1}$

$$\hat{p}(x') = p(x_{n+1}|x_1, \ldots, x_n)$$

MDL: Two-stage Code Redundancy

- Expected codelength minus target at $p^* = p_{f^*}$

$$\text{Redundancy} = E\left[\min_{f \in \mathcal{F}} \left\{ \log \frac{1}{p_f(\underline{x})} + L(f) \right\} - \log \frac{1}{p_{f^*}(\underline{x})} \right]$$

Redundancy and Resolvability

- Redundancy $= E \min_{f \in \mathcal{F}} \left[\log \frac{p_{f^*}(\underline{x})}{p_f(\underline{x})} + L(f) \right]$
- Resolvability $= \min_{f \in \mathcal{F}} E \left[\log \frac{p_{f^*}(\underline{x})}{p_f(\underline{x})} + L(f) \right]$ $= \min_{f \in \mathcal{F}} \left[D(P_{\underline{X}|f^*} \| P_{\underline{X}|f}) + L(f) \right]$
- Ideal tradeoff of Kullback approximation error & complexity
- Population analogue of the two-stage code MDL criterion
- Divide by n to express as a rate. In the i.i.d. case

$$R_n(f^*) = \min_{f \in \mathcal{F}} \left[D(f^* \| f) + \frac{L(f)}{n} \right]$$

Risk of Estimator based on Two-stage Code

- Estimator $\hat{f}$ is the choice achieving the minimization

$$\min_{f \in \mathcal{F}} \left\{ \log \frac{1}{p_f(\underline{x})} + \mathcal{L}(f) \right\}$$

- Codelengths for f are $\mathcal{L}(f) = 2L(f)$ with $\sum_{f \in \mathcal{F}} 2^{-L(f)} \leq 1$.

- Total loss $d_n(f^*, \hat{f})$ with $d_n(f^*, f) = d(P_{\underline{X}'|f^*}, P_{X'|f})$

$$\text{Risk} = E[d_n(f^*, \hat{f})]$$

- Info-Thy bound on risk: Barron (1985), Barron and Cover (1991), Jonathan Li (1999)

Risk $\leq$ Redundancy $\leq$ Resolvability

Risk of Estimator based on Two-stage Code

- Estimator $\hat{f}$ achieves $\min_{f \in \mathcal{F}} \{\log 1/p_f(\underline{x}) + \mathcal{L}(f)\}$
- Codelengths require $\sum_{f \in \mathcal{F}} 2^{-L(f)} \leq 1$.
- *Risk* $\leq$ Resolvability
- Specialize to i.i.d. case:

$$Ed(f^*, \hat{f}) \leq \min_{f \in \mathcal{F}} \left[D(f^* \| f) + \frac{L(f)}{n} \right]$$

- As $n \nearrow$, tolerate more complex f if needed to get near f^*
- Rate is $1/n$, or close to that rate if f^* is simple
- Drawback: Code interpretation entails countable $\mathcal{F}$

Key to Risk Analysis

- log likelihood-ratio discrepancy at training $\underline{x}$ and future $\underline{x}'$

$$\left[\log \frac{p_{f^*}(\underline{x})}{p_f(\underline{x})} - d_n(f^*, f)\right]$$

- Proof shows, for $L(f)$ satisfying Kraft, that

$$\min_{f \in \mathcal{F}} \left\{ \left[\log \frac{p_{f^*}(\underline{x})}{p_f(\underline{x})} - d_n(f^*, f)\right] + \mathcal{L}(f) \right\}$$

has expectation ≥ 0. From which the risk bound follows.

Outline

Information-theoretically Valid Penalty

- Penalized Likelihood

$$\hat{f} = \mathrm{argmin}_{f \in \mathcal{F}} \left\{ \log \frac{1}{p_f(\underline{x})} + Pen_n(f) \right\}$$

- Possibly uncountable $\mathcal{F}$
- Task: determine a condition on $Pen_n(f)$ such that the risk is captured by the population analogue

$$Ed_n(f^*, \hat{f}) \leq \inf_{f \in \mathcal{F}} \left\{ E \log \frac{p_{f^*}(\underline{X})}{p_f(\underline{X})} + Pen_n(f) \right\}$$

Info-thy Penalty: Link Uncountable & Countable

- Suppose for uncountable $\mathcal{F}$ and penalty $Pen_n(f), f \in \mathcal{F}$ there is a countable $\tilde{\mathcal{F}}$ and $\mathcal{L}_n(\tilde{f})$ satisfying Kraft, such that, for all $\underline{x}$, f^*,

$$\min_{f \in \mathcal{F}} \left\{ \left[\log \frac{p_{f^*}(\underline{x})}{p_f(\underline{x})} - d_n(f^*, f) \right] + Pen_n(f) \right\}$$

$$\geq \min_{\tilde{f} \in \tilde{\mathcal{F}}} \left\{ \left[\log \frac{p_{f^*}(\underline{x})}{p_{\tilde{f}}(\underline{x})} - d_n(f^*, \tilde{f}) \right] + \mathcal{L}_n(\tilde{f}) \right\}$$

- Proof of the risk conclusion:
  The second expression has expectation ≥ 0,
  so the first expression does too.

Variable Complexity, Variable Distortion Cover (Code)

- Equivalently: $Pen_n(f)$ is a valid penalty if for all $\underline{x}$,

$$Pen_n(f) \geq \min_{\tilde{f} \in \tilde{\mathcal{F}}} \left[\mathcal{L}(\tilde{f}) + \Delta_n(\tilde{f}, f) \right]$$

  where the distortion $\Delta_n(\tilde{f}, f)$ is the difference in the discrepancies at $\tilde{f}$ and f

- Equivalently: For each f in $\mathcal{F}$ there is a representer $\tilde{f}$ in $\tilde{\mathcal{F}}$ with complexity $L(\tilde{f})$, distortion $\Delta_n(\tilde{f}, f)$ and

$$Pen_n(f) \geq \mathcal{L}(\tilde{f}) + \Delta_n(\tilde{f}, f)$$

Linear Span of a Dictionary

- $\mathcal{G}$ is a dictionary of candidate basis functions
- Wavelets, splines, polynomials, trigonometric terms, sigmoids, explanatory variables and their interactions

- Candidate functions in the linear span

$$f(x) = f_\theta(x) = \sum_{g \in \mathcal{G}} \theta_g \, g(x)$$

- ℓ_1 norm of coefficients

$$\|\theta\|_1 = \sum_g |\theta_g|$$

Example Models

- Regression

$$p_f(y|x) = \text{Normal}(f(x), \sigma^2)$$

- Logistic regression with $y \in \{0, 1\}$

$$p_f(y|x) = \text{Logistic}(f(x)) \quad \text{for } y = 1$$

- Log-density estimation

$$p_f(x) = \frac{p_0(x) \exp\{f(x)\}}{c_f}$$

ℓ_1 Penalty

- $pen_n(f_\theta) = \lambda_n \|\theta\|_1$ where $f_\theta(x) = \sum_{g \in \mathcal{G}} \theta_g \, g(x)$
- Popular penalty: Chen & Donoho (96) Basis Pursuit; Tibshirani (96) LASSO; Efron et al (04) LARS; Precursors: Jones (92), B.(90,93,94) greedy algorithm and analysis of combined ℓ_1 and ℓ_0 penalty
- Risk analysis: specify valid λ_n for risk $\leq$ resolvability
- Computation analysis: bound accuracy of new ℓ_1-penalized greedy pursuit algorithm

ℓ_1 penalty is valid for λ_n of order $1/\sqrt{n}$

- Example: ℓ_1 penalized log-density estimation, i.i.d. case

$$\hat{\theta} = \mathrm{argmin}_\theta \left\{ \frac{1}{n} \log \frac{1}{p_{f_\theta}(\underline{x})} + \lambda_n \|\theta\|_1 \right\}$$

- Risk bound:

$$Ed(f^*, f_{\hat{\theta}}) \leq \inf_\theta \left\{ D(f^* \| f_\theta) + \lambda_n \|\theta\|_1 \right\}$$

- Valid for

$$\lambda_n \geq \sqrt{\frac{H}{n}} \quad \text{with } H = \log Card(\mathcal{G})$$

- For infinite $\mathcal{G}$ use metric entropy in place of H
- Results for regression shown in a companion paper

ℓ_1 penalty is valid for λ_n of order $1/\sqrt{n}$

- Example: ℓ_1 penalized log-density estimation, i.i.d. case

$$\hat{\theta} = \text{argmin}_\theta \left\{ \frac{1}{n} \log \frac{1}{p_{f_\theta}(\underline{x})} + \lambda_n \|\theta\|_1 \right\}$$

- Risk bound:

$$Ed(f^*, f_{\hat{\theta}}) \leq \inf_\theta \left\{ D(f^* \| f_\theta) + \lambda_n \|\theta\|_1 \right\}$$

- True with

$$\lambda_n = \sqrt{\frac{H}{n}} \quad \text{with } H = \log Card(\mathcal{G})$$

- Risk of order λ_n when the target has finite ℓ_1 norm

Comment on proof

- Shannon-like demonstration of the existence of the variable complexity cover property
- Inspiration from technique originating with Lee Jones (92)
- Representer $\tilde{f}$ of f_θ of the form

$$\tilde{f}(x) = \frac{v}{m} \sum_{k=1}^{m} g_k(x)$$

- $g_1, \ldots g_m$ picked at random from $\mathcal{G}$, independently, where g arises with probability proportional to $|\theta_g|$
- May pick them in greedy fashion as in demonstration of fast computation properties
- In the paper in the Festschrift with summary in the ITW proceedings

Outline

Summary

- Handle penalized likelihoods with continuous domains for f
- Information-theoretically valid penalties:
  Penality exceed complexity plus distortion of optimized representors of f
- Yields statistical risk controlled by resolvability
- ℓ_0 penalty $\frac{dim}{2}\log n$ classically analyzed
- ℓ_1 penalty $\lambda_n \|\theta\|_1$ analyzed here: valid for $\lambda_n \geq \sqrt{H/n}$.