Celebrations for
three influential scholars in Information Theory and Statistics:

- 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
- 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
  Rényi Institute, Budapest, August 25-28, 2008; NOW!

MDL Procedures with ℓ_1 Penalty and their Statistical Risk

Andrew Barron
Department of Statistics
Yale University

Collaborators: Jonathan Li, Cong Huang, Xi Luo

August 19, 2008
Budapest, Hungary

Outline

Outline

Shannon, Kraft, McMillan

- Characterization of uniquely decodeable codelengths

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

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

- Operational meaning of probability:

  A distribution is given by a choice of code

Codelength Comparison

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

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

- Expected redundancy

$$D(P_{\underline{X}} \| Q_{\underline{X}}) = E_P \left[\log \frac{p(\underline{X})}{q(\underline{X})} \right]$$

Universal Codes

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

$$\{L_\theta(\underline{x}) : \theta \in \Theta\} \Leftrightarrow \{p_\theta(\underline{x}) : \theta \in \Theta\}$$

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

$$L(\underline{x}) = \log 1/q(\underline{x})$$

- Redundancy: $\left[\ \log 1/q(\underline{x}) - \log 1/p_\theta(\underline{x})\ \right]$
  Want it small either uniformly in $\underline{x}, \theta$ or in expectation

Statistical Aim

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

Loss

- Kullback Information-divergence:

$$D(P_{\underline{X}'} \| Q_{\underline{X}'}) = E\left[\, \log p(\underline{X}')/q(\underline{X}')\,\right]$$

- Bhattacharyya, Hellinger, Chernoff, Rényi divergence:

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

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

$$D(P_{\underline{X}'} \| Q_{\underline{X}'}) = n\, D(P \| Q)$$

Likewise $$d(P_{\underline{X}'}, Q_{\underline{X}'}) = n\, d(P, Q)$$

Loss

- Relationship:

$$d(P, Q) \leq D(P \| Q)$$

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

$$D(P \| Q) \leq C_B \, d(P, Q)$$

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:

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

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

- The corresponding statistical estimator is $\hat{p} = p_{\hat{\theta}}$

- Typically in d-dimensional families $L(\theta)$ is of order

$$\frac{d}{2} \log n$$

MDL: Mixture Versions

- Codelength based on a mixture model

$$L(\underline{x}) = \log \frac{1}{\int p(\underline{x}|\theta)w(\theta)d\theta}$$

average case optimal and pointwise optimal for a.e. θ

- Codelength approximation (Barron 1985, Clarke and Barron 1990,1994)

$$\log \frac{1}{p(\underline{x}|\hat{\theta})} + \frac{d}{2}\log\frac{n}{2\pi} + \log\frac{|\hat{I}(\hat{\theta})|^{1/2}}{w(\hat{\theta})}$$

where $\hat{I}(\hat{\theta})$ is the empirical Fisher Information at the MLE

MDL: Two-stage Code Redundancy

- Expected codelength minus target at p_{θ^*}

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

- Redundancy approx in smooth families

$$\frac{d}{2}\log\frac{n}{2\pi} + \log\frac{|I(\theta^*)|^{1/2}}{w(\theta^*)}$$

Redundancy and Resolvability

- Redundancy $= E \min_{\theta \in \Theta} \left[\log \frac{p_{\theta^*}(\underline{x})}{p_\theta(\underline{x})} + L(\theta)) \right]$
- Resolvability $= \min_{\theta \in \Theta} E \left[\log \frac{p_{\theta^*}(\underline{x})}{p_\theta(\underline{x})} + L(\theta) \right]$
  $= \min_{\theta \in \Theta} \left[D(P_{\underline{X}|\theta^*} \| P_{\underline{X}|\theta}) + L(\theta) \right]$
- 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

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

Risk of Estimator based on Two-stage Code

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

$$\min_{\theta \in \Theta} \left\{ \log \frac{1}{p_\theta(\underline{x})} + \mathcal{L}(\theta) \right\}$$

- Codelengths for θ are $\mathcal{L}(\theta) = 2L(\theta)$ with $\sum_{\theta \in \Theta} 2^{-L(\theta)} \le 1$.
- Total loss $d_n(\theta^*, \hat{\theta})$ with $d_n(\theta^*, \theta) = d(P_{\underline{X}'|\theta^*}, P_{\underline{X}'|\theta})$

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

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

$$\textit{Risk} \le \text{Redundancy} \le \text{Resolvability}$$

Risk of Estimator based on Two-stage Code

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

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

- As $n \nearrow$, tolerate more complex $P_{X|\theta}$ if needed to get near $P_{X|\theta^*}$
- Rate is $1/n$, or close to that rate if the target is simple
- Drawback: Code interpretation entails countable Θ

Key to Risk Analysis

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

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

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

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

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

Outline

Information-theoretically Valid Penalties

- Penalized Likelihood

$$\min_{\theta \in \Theta} \left\{ \log \frac{1}{p_\theta(\underline{x})} + Pen(\theta) \right\}$$

- Possibly uncountable Θ
- Yields data compression interpretation if there exists a countable $\tilde{\Theta}$ and L satisfying Kraft such that the above is not less than

$$\min_{\tilde{\theta} \in \tilde{\Theta}} \left\{ \log \frac{1}{p_{\tilde{\theta}}(\underline{x})} + L(\tilde{\theta}) \right\}$$

Data-Compression Valid Penalties

- Equivalently, $Pen(\theta)$ is valid for penalized likelihood with uncountable Θ to have a data compression interpretation if there is such a countable $\tilde{\Theta}$ and Kraft summable $L(\tilde{\theta})$, such that, for every θ in Θ, there is a representor $\tilde{\theta}$ in $\tilde{\Theta}$ such that

$$Pen(\theta) \;\geq\; L(\tilde{\theta}) + \log \frac{p_\theta(\underline{x})}{p_{\tilde{\theta}}(\underline{x})}$$

- This is the link between uncountable and countable cases

Statistical-Risk Valid Penalties

- Penalized Likelihood

$$\hat{\theta} = \mathrm{argmin}_{\theta \in \tilde{\Theta}} \left\{ \log \frac{1}{p_\theta(\underline{x})} + Pen(\theta) \right\}$$

- Again: possibly uncountable Θ
- Task: determine a condition on $Pen(\theta)$ such that the risk is captured by the population analogue

$$E d_n(\theta^*, \hat{\theta}) \leq \inf_{\theta \in \Theta} \left\{ E \log \frac{p_{\theta^*}(\underline{X})}{p_\theta(\underline{X})} + Pen(\theta) \right\}$$

Statistical-Risk Valid Penalty

- For an uncountable Θ and a penalty $Pen(\theta)$, $\theta \in \Theta$, suppose there is a countable $\tilde{\Theta}$ and $\mathcal{L}(\tilde{\theta}) = 2L(\tilde{\theta})$ where $L(\tilde{\theta})$ satisfies Kraft, such that, for all $\underline{x}, \theta^*$,

$$\min_{\theta \in \Theta} \left\{ \left[\log \frac{p_{\theta^*}(\underline{x})}{p_{\theta}(\underline{x})} - d_n(\theta^*, \theta) \right] + Pen(\theta) \right\}$$

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

- Proof of the risk conclusion:
  The second expression has expectation ≥ 0,
  so the first expression does too.
- This condition and result is obtained with J. Li and X. Luo (in Rissanen Festschrift 2008)

Variable Complexity, Variable Distortion Cover

- Equivalent statement of the condition: $Pen(\theta)$ is a valid penalty if for each θ in Θ there is a representor $\tilde{\Theta}$ in $\tilde{\Theta}$ with complexity $L(\tilde{\theta})$, distortion $\Delta_n(\tilde{\theta}, \theta)$ and

$$Pen(\theta) \;\geq\; \mathcal{L}(\tilde{\theta}) + \Delta_n(\tilde{\theta}, \theta)$$

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

$$\Delta_n(\tilde{\theta}, \theta) = \log \frac{p_\theta(\underline{x})}{p_{\tilde{\theta}}(\underline{x})} + d_n(\theta, \theta^*) - d_n(\tilde{\theta}, \theta^*)$$

A Setting for Regression and log-density Estimation: Linear Span of a Dictionary

- $\mathcal{G}$ is a dictionary of candidate basis functions
  E.g. wavelets, splines, polynomials, trigonometric terms, sigmoids, explanatory variables and their interactions
- Candidate functions in the linear span

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

- weighted ℓ_1 norm of coefficients

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

- weights $a_g = \|g\|_n$ where $\|g\|_n^2 = \frac{1}{n} \sum_{i=1}^n g^2(x_i)$

Example Models

- Regression (focus of current presentation)

$$p_\theta(y|x) = \text{Normal}(f_\theta(x), \sigma^2)$$

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

$$p_\theta(y|x) = \text{Logistic}(f_\theta(x)) \quad \text{for } y = 1$$

- Log-density estimation (focus of Festschrift paper)

$$p_\theta(x) = \frac{p_0(x) \exp\{f_\theta(x)\}}{c_f}$$

- Gaussian graphical models

ℓ_1 Penalty

- $pen(\theta) = \lambda\|\theta\|_1$ where θ are coeff of $f_\theta(x) = \sum_{g\in\mathcal{G}} \theta_g\, g(x)$
- 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
- We want to avoid cross-validation in choice of λ
- Data-compression: specify valid λ for coding interpretation
- Risk analysis: specify valid λ for risk $\leq$ resolvability
- Computation analysis: bounds accuracy of ℓ_1-penalized greedy pursuit algorithm

Outline

Regression with ℓ_1 penalty

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

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

- Regression with Gaussian model, fixed σ^2

$$\min_\theta \left\{ \frac{1}{2\sigma^2} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_\theta(x_i))^2 + \frac{1}{2} \log 2\pi\sigma^2 + \frac{\lambda_n}{\sigma} \|\theta\|_1 \right\}$$

- Valid for

$$\lambda_n \geq \sqrt{\frac{2 \log(2M_{\mathcal{G}})}{n}} \quad \text{with } M_{\mathcal{G}} = Card(\mathcal{G})$$

Adaptive risk bound

- For log density estimation with suitable λ_n

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

- For regression with fixed design points x_i, fixed σ, and $\lambda_n = \sqrt{\frac{2\log(2M)}{n}}$,

$$\frac{E\|f^* - f_{\hat{\theta}}\|_n^2}{4\sigma^2} \leq \inf_{\theta} \left\{ \frac{\|f^* - f_\theta\|_n^2}{2\sigma^2} + \frac{\lambda_n}{\sigma} \|\theta\|_1 \right\}$$

Adaptive risk bound specialized to regression

- Again for fixed design and $\lambda_n = \sqrt{\frac{2\log 2M}{n}}$, multiplying through by $4\sigma^2$,

$$E\|f^* - f_{\hat{\theta}}\|_n^2 \leq \inf_{\theta} \left\{ 2\|f^* - f_\theta\|_n^2 + 4\sigma\lambda_n\|\theta\|_1 \right\}$$

- In particular for all targets $f^* = f_{\theta^*}$ with finite $\|\theta^*\|$ the risk bound $4\sigma\lambda_n\|\theta^*\|$ is of order $\sqrt{\frac{\log M}{n}}$
- Details in Barron, Luo (proceedings Workshop on Information Theory Methods in Science & Eng. 2008), Tampere, Finland: last week

Comments on proof

- Likelihood discrepancy plus complexity

$$\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - f_{\tilde{\theta}}(x_i))^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - f_{\theta}(x_i))^2 + K\log(2M)$$

- Representor $f_{\tilde{\theta}}$ of f_θ of following form, with v near $\|\theta\|_1$

$$f_{\tilde{\theta}}(x) = \frac{v}{K}\sum_{k=1}^{K} g_k(x)/\|g_k\|$$

- $g_1, \ldots g_K$ picked at random from $\mathcal{G}$, independently, where g arises with probability proportional to $|\theta_g|a_g$
- Shows exists representor with like. discrep. $+$ complexity

$$\frac{nv^2}{2K} + K\log(2M)$$

Comments on proof

- Optimizing it yields penalty proportional to ℓ_1 norm
- Penalty $\lambda\|\theta\|_1$ is valid for both data compression and statistical risk requirements for $\lambda \geq \lambda_n$ where

$$\lambda_n = \sqrt{\frac{2\log(2M)}{n}}$$

- Especially useful for very large dictionaries
- Improvement for small dictionaries gets rid of log factor: $\log(2M)$ may be replaced by $\log(2e\max\{\frac{M}{\sqrt{n}}, 1\})$

Comments on proof

- Existence of respresentor shown by random draw is a Shannon-like demonstration of the variable cover (code)
- Similar approximation in analysis of greedy computation of ℓ_1 penalized least squares

Aside: Random design case

- May allow X_i random with distribution P_X
- Presuming functions in the library $\mathcal{G}$ are uniformly bounded
- Analogous risk bounds hold (in submitted paper by Huang, Cheang, and Barron)

$$E\|f^* - f_{\hat{\theta}}\|^2 \leq \inf_{\theta} \left\{ 2\|f^* - f_\theta\|^2 + c\lambda_n \|\theta\|_1 \right\}$$

- Allows for libraries $\mathcal{G}$ of infinite cardinality, replacing the $\log M_{\mathcal{G}}$ with a metric entropy of $\mathcal{G}$
- If the linear span of $\mathcal{G}$ is dense in L_2 then for all L_2 functions f^* the approximation error $\|f^* - f_\theta\|^2$ can be arranged to go to zero as the size of $v = \|\theta^*\|_1$ increases.

Aside: Simultaneous Minimax Rate Optimality

- The resolvability bound tends to zero as $\lambda_n = \sqrt{\frac{2 \log M_{\mathcal{G}}}{n}}$ gets small

$$E\|f^* - f_{\hat{\theta}}\|^2 \leq \inf_{\theta} \left\{ 2\|f^* - f_\theta\|^2 + c\lambda_n \|\theta\|_1 \right\}$$

- Current work with Cong Huang shows for a broad class of approximate rates, the risk rate obtained from this resolvability bound is minimax optimal for the class of target functions $\{f^* : inf_{\theta:\|\theta\|_1 = v} \|f^* - f_\theta\|^2 \leq A_v\}$, simultaneously for all such classes

Fixed σ versus unknown σ

- MDL with ℓ_1 penalty for each possible σ. Recall

$$\min_{\theta}\left\{\frac{1}{2\sigma^2}\frac{1}{n}\sum_{i=1}^{n}(Y_i - f_\theta(x_i))^2 + \frac{1}{2}\log 2\pi\sigma^2 + \frac{\lambda_n}{\sigma}\|\theta\|_1\right\}$$

- Provides a family of fits indexed by σ.
- For unknown σ suggest optimization over σ as well as θ

$$\min_{\theta,\sigma}\left\{\frac{1}{2\sigma^2}\frac{1}{n}\sum_{i=1}^{n}(Y_i - f_\theta(x_i))^2 + \frac{1}{2}\log 2\pi\sigma^2 + \frac{\lambda_n}{\sigma}\|\theta\|_1\right\}$$

- Slight modification of this does indeed satisfy our condition for an information-theoretically valid penalty and risk bound (details in the WITMSE 2008 proceedings)

Best σ

- Best σ for each θ solves the quadratic equation

$$\sigma^2 = \sigma \lambda_n \|\theta\|_1 + \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - f_\theta(x_i)\right)^2$$

- By the quadratic formula the solution is

$$\sigma = \frac{1}{2}\lambda_n \|\theta\|_1 + \sqrt{\left[\frac{1}{2}\lambda_n \|\theta\|_1\right]^2 + \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - f_\theta(x_i)\right)^2}$$

Outline

Summary

- Handle penalized likelihoods with continuous domains for θ

- Information-theoretically valid penalties: Penalty exceed complexity plus distortion of optimized representor of θ
- Yields MDL interpretation and statistical risk controlled by resolvability
- ℓ_0 penalty $\frac{dim}{2} \log n$ classically analyzed
- ℓ_1 penalty $\sigma \lambda_n \|\theta\|_1$ analyzed here: valid in regression for

$$\lambda_n \geq \sqrt{2(log 2M)/n}$$

- Can handle fixed or unknown σ