Overview of Recent Developments in Penalized Likelihood and MDL

Andrew R. Barron

Yale University

Collaborators
Penal. Like: Sabyasachi Chatterjee, Cong Huang, Jonathan Li, Xi Luo
NML vs Bayes: Teemu Roos, Kazuho Watanabe

Workshop on Information Theory Methods in Science & Engineering
WITMSE, 5 July 2014, Wakiki, Hawaii
Outline

1. Some Principles of Information Theory and Statistics
   - Data Compression
   - Minimum Description Length Principle
   - Two-stage Codes, Mixture Codes, Normalized Max Like.

2. Penalized Likelihood
   - Redundancy, Resolvability, and Risk
   - Risk-Valid Penalties
   - Penalized Likelihood Analysis for Continuous Parameters
   - $\ell_0$ and $\ell_1$ Penalties are Codelength-Valid and Risk-Valid

3. Summary
Outline

1. Some Principles of Information Theory and Statistics
   - Data Compression
   - Minimum Description Length Principle
   - Two-stage Codes, Mixture Codes, Normalized Max Like.

2. Penalized Likelihood
   - Redundancy, Resolvability, and Risk
   - Risk-Valid Penalties
   - Penalized Likelihood Analysis for Continuous Parameters
   - $\ell_0$ and $\ell_1$ Penalties are Codelength-Valid and Risk-Valid

3. Summary
Characterization of uniquely decodeable codelengths

\[ L(x), \quad x \in \mathcal{X}, \quad \sum_{x} 2^{-L(x)} \leq 1 \]

\[ L(x) = \log \frac{1}{q(x)} \quad q(x) = 2^{-L(x)} \]

Operational meaning of probability:

A distribution is given by a choice of code
Code-length Comparison

- Targets $p$ are possible distributions
- Compare code-length $\log \frac{1}{q(x)}$ to targets $\log \frac{1}{p(x)}$
- Redundancy or regret

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

- Expected redundancy

$$D(P_X \parallel Q_X) = E_P \left[ \log \frac{p(X)}{q(X)} \right]$$
Universal Codes

- MODELS
  Family of coding strategies $\Leftrightarrow$ Family of prob. distributions
  Indexed by parameters or functions:

  \[
  \{ L_\theta(x) : \theta \in \Theta \} \Leftrightarrow \{ p_\theta(x) : \theta \in \Theta \}
  \]

  \[
  \{ L_f(x) : f \in F \} \Leftrightarrow \{ p_f(x) : f \in F \}
  \]

- Universal codes $\Leftrightarrow$ Universal probabilities $q(x)$

  \[
  L(x) = \log 1 / q(x)
  \]

- Redundancy:

  \[
  \left[ \log 1 / q(x) - \log 1 / p_\theta(x) \right]
  \]

  Want it small either uniformly in $x, \theta$ or in expectation
Statistical Aim

- Training data $x \Rightarrow$ estimator $\hat{p} = p_{\hat{\theta}}$

- Subsequent data $x'$

- Want $\log \frac{1}{\hat{p}(x')}$ to compare favorably to $\log \frac{1}{p(x')}$

- For targets $p$ close to or in the families
Loss

- Kullback Information-divergence:
  \[ D(P_{X'} \| Q_{X'}) = E \left[ \log p(X')/q(X') \right] \]

- Bhattacharyya, Hellinger, Chernoff, Rényi divergence:
  \[ d(P_{X'}, Q_{X'}) = 2 \log 1/E[q(X')/p(X')]^{1/2} \]

- Product model case: \( p(x') = \prod_{i=1}^{n} p(x'_i) \)
  \[ D(P_{X'} \| Q_{X'}) = n D(P \| Q) \]
  Likewise \[ d(P_{X'}, Q_{X'}) = n d(P, Q) \]
Minimum Description Length (MDL) Principle

- 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:
  \[
  L(x) = \min_{\theta \in \Theta} \left[ \log 1/p_{\theta}(x) + L(\theta) \right]
  \]
  bits for \(x\) given \(\theta\) + bits for \(\theta\)

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

- Typically in \(d\)-dimensional families \(L(\theta)\) is of the form
  \[
  \frac{d}{2} \log n + C_d(\theta)
  \]
MDL: Mixture Versions

- Code length based on a Bayes mixture

\[ L(x) = \log \frac{1}{q(x)} \]

where

\[ q(x) = \int p(x|\theta)w(\theta)d\theta \text{ or } \sum_\theta p(x|\theta)w(\theta) \]

minimax optimal with least favorable \( w \) (capacity achieving)

- Code length approximation (Barron 1985, Clarke and Barron 1990,1994)

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

where \( \hat{I}(\hat{\theta}) \) is the empirical Fisher Information at the MLE
MDL: Mixture Versions

- Codelength based on a Bayes mixture
  \[ q(x) = \int p(x|\theta)w(\theta)d\theta \]

- Corresponding statistical estimator is the predictive distrib
  \[ \hat{p}(x') = q(x'|x) = \frac{\int p(x'|\theta)p(x|\theta)w(\theta)d\theta}{\int p(x|\theta)w(\theta)d\theta} \]

- It has a clean relative entropy risk bound
  \[ ED(P||\hat{P}) \leq \text{Resolvability} \]
  \[ = \text{Kullback Approx Error} + \log 1/(\text{Posterior prob of neigh.}) \]
NML Version

- Code-length via normalized maximum likelihood (NML)
  \[ NML(x) = \frac{\max_\theta p(x|\theta)}{C_{Shtarkov}} \]

- Minimax optimal for pointwise regret

- It has the same code-length approximation as Bayes with asympt. least favorable prior \((\text{Takeuchi} \ & \ B. \ 1998, \ B., \ Rissanen, \ Yu \ 1998)\)

- Is NML exactly Bayes in finite samples? \((B., \ Roos, \ Watanabe \ 2014)\)
NML Version

- Codelength via normalized maximum likelihood (NML)

\[
NML(x) = \frac{\max_\theta p(x|\theta)}{C_{Shtarkov}}
\]

- Minimax optimal for pointwise regret

- It has the same codelength approximation as Bayes with asympt. least favorable prior (Takeuchi & B. 1998, B., Rissanen, Yu 1998)

- Is NML exactly Bayes in finite samples? (B., Roos, Watanabe 2014)
NML Version

- Codelength via normalized maximum likelihood (NML)
  \[
  NML(x) = \frac{\max_{\theta} p(x|\theta)}{C_{\text{Shtarkov}}}
  \]
- minimax optimal for pointwise regret
- It has the same codelength approximation as Bayes with asympt. least favorable prior (Takeuchi & B. 1998, B., Rissanen, Yu 1998)
- Is NML exactly Bayes in finite samples? (B., Roos, Watanabe 2014)
NML Version

- Codelength via normalized maximum likelihood (NML)
  \[ NML(x) = \frac{\max_{\theta} p(x|\theta)}{C_{Shtarkov}} \]

- Minimax optimal for pointwise regret

- It has the same codelength approximation as Bayes with asympt. least favorable prior (Takeuchi & B. 1998, B., Rissanen, Yu 1998)

- Is NML exactly Bayes in finite samples? (B., Roos, Watanabe 2014)
  Yes, with sufficiently many linearly independent \( p(\cdot|\theta) \), [though in esoteric cases some weights are negative]:
  \[ NML(x) = \sum_{\theta} p(x|\theta) w(\theta) \]
NML Version

- Codelength via normalized maximum likelihood (NML)
  \[ NML(x) = \frac{\max_{\theta} p(x|\theta)}{C_{Shtarkov}} \]

- Minimax optimal for pointwise regret

- It has the same codelength approximation as Bayes with asymptotic least favorable prior (Takeuchi & B. 1998, B., Rissanen, Yu 1998)

- NML is exactly Bayes in finite samples (B., Roos, Watanabe 2014)
  \[ NML(x) = \sum_{\theta} p(x|\theta) w(\theta) \]

  Allows fast computation of marginals and predictive distribution.
Outline

1. Some Principles of Information Theory and Statistics
   - Data Compression
   - Minimum Description Length Principle
   - Two-stage Codes, Mixture Codes, Normalized Max Like.

2. Penalized Likelihood
   - Redundancy, Resolvability, and Risk
   - Risk-Valid Penalties
   - Penalized Likelihood Analysis for Continuous Parameters
   - $\ell_0$ and $\ell_1$ Penalties are Codelength-Valid and Risk-Valid

3. Summary
Penalized Likelihood

Penalized likelihood estimators

\[ \hat{\theta} = \text{argmin}_\theta \left\{ \log \frac{1}{p(x|\theta)} + \text{pen}(\theta) \right\} \]

where the penalty \( \text{pen}(\theta) \) arises from

- prior density: \( \text{pen}(\theta) = \log \frac{1}{w(\theta)} \)
- code length: \( \text{pen}(\theta) = L(\theta) \)
- dimensionality: \( \text{pen}(\theta) = \lambda_n \dim(\Theta) \)
- sparsity control: \( \text{pen}(\theta) = \lambda_n \|\theta\|_1 \)
- roughness penalty: \( \text{pen}(\theta) = \text{norm of derivative} \)
- maximum likelihood: \( \text{pen}(\theta) = \text{constant} \)
Penalized Likelihood

Penalized likelihood estimators

$$\hat{\theta} = \arg\min_{\theta} \{\log \frac{1}{p(x|\theta)} + \text{pen}(\theta)\}$$

where the penalty $\text{pen}(\theta)$ arises from

- prior density: $\text{pen}(\theta) = \log \frac{1}{w(\theta)}$
- codelength: $\text{pen}(\theta) = L(\theta)$
- dimensionality: $\text{pen}(\theta) = \lambda_n \text{dim}(\Theta)$
- sparsity control: $\text{pen}(\theta) = \lambda_n \|\theta\|_1$
- roughness penalty: $\text{pen}(\theta) = \text{norm of derivative}$
- maximum likelihood: $\text{pen}(\theta) = \text{constant}$

MDL analysis reveals which size penalties are valid for good prediction and compression properties.
Two-stage Code Redundancy

- Two-stage codes: the start of penalized likelihood analysis
- Expected codelength minus target at $p_{\theta^*}$

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

- Redundancy approx in smooth families of dimension $d$

$$\frac{d}{2} \log n + C_d(\theta)$$
Redundancy and Resolvability

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

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

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

- Code lengths for $\theta$ are $\mathcal{L}(\theta) = 2L(\theta)$ with $\sum_{\theta \in \Theta} 2^{-L(\theta)} \leq 1$.

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

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

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

  \[\text{Risk} \leq \text{Redundancy} \leq \text{Resolvability}\]
Risk of Estimator based on Two-stage Code

- Estimator $\hat{\theta}$ is the choice achieving the minimization
  \[
  \min_{\theta \in \Theta} \left\{ \log \frac{1}{p_\theta(x)} + L(\theta) \right\}
  \]
- Codelengths for $\theta$ are $L(\theta) = 2L(\theta)$ with $\sum_{\theta \in \Theta} 2^{-L(\theta)} \leq 1$.
- Total loss $d_n(\theta^*, \hat{\theta})$ with $d_n(\theta^*, \theta) = d(P_{X'|\theta^*}, P_{X'|\theta})$
  \[
  \text{Risk} = E[d_n(\theta^*, \hat{\theta})]
  \]
- Info-Thy bound on risk: (Barron 1985, Barron and Cover 1991, Jonathan Li 1999)
  \[
  \text{Risk} \leq \text{Redundancy} \leq \text{Resolvability}
  \]
- Drawback: Two-part code interpretation needs countable $\Theta$
Key to Risk Analysis (in the countable $\Theta$ case)

- log likelihood-ratio discrepancy at training $x$ and future $x'$

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

- Proof shows, for $L(\theta) = Lcal(\theta)/2$ satisfying Kraft, that

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

has expectation $\geq 0$. From which the risk bound follows.
Penalized Likelihood

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

Possibly uncountable \( \Theta \)

It is still a codelength if there exists a countable \( \tilde{\Theta} \) and \( L \) satisfying Kraft such that the above is not less than

\[
\min_{\tilde{\theta} \in \tilde{\Theta}} \left\{ \log \frac{1}{p_{\tilde{\theta}}(x)} + L(\tilde{\theta}) \right\}
\]
Equivalently, \( Pen(\theta) \) is valid for penalized likelihood to be a code length if there is such a countable \( \tilde{\Theta} \) and Kraft summable \( L(\tilde{\theta}) \), such that, for every \( \theta \) in \( \Theta \), there is a representor \( \tilde{\theta} \) in \( \tilde{\Theta} \) such that

\[
Pen(\theta) \geq L(\tilde{\theta}) + \log \frac{p_\theta(x)}{p_{\tilde{\theta}}(x)}
\]

This is the link between uncountable and countable cases.
Statistical-Risk Valid Penalties

- Penalized Likelihood
  \[ \hat{\theta} = \arg\min_{\theta \in \Theta} \left\{ \log \frac{1}{p_\theta(x)} + Pen(\theta) \right\} \]

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

\[ Ed_n(\theta^*, \hat{\theta}) \leq \inf_{\theta \in \Theta} \left\{ E \log \frac{p_{\theta^*}(X)}{p_\theta(X)} + Pen(\theta) \right\} \]
Statistical-Risk Valid Penalty

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

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

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

Proof of the risk conclusion:
The second expression has expectation $\geq 0$, 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:** \( \text{Pen}(\theta) \) is a valid penalty if for each \( \theta \) in \( \Theta \) there is a representor \( \tilde{\theta} \) in \( \tilde{\Theta} \) with complexity \( L(\tilde{\theta}) \), distortion \( \Delta_n(\tilde{\theta}, \theta) \) and

\[
\text{Pen}(\theta) \geq L(\tilde{\theta}) + \Delta_n(\tilde{\theta}, \theta)
\]

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

\[
\Delta_n(\tilde{\theta}, \theta) = \log \frac{p_\theta(x)}{p_{\tilde{\theta}}(x)} + d_n(\theta, \theta^*) - d_n(\tilde{\theta}, \theta^*)
\]
A Setting for Regression and log-density Estimation: Linear Span of a Dictionary

- $G$ 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
  \[ f_\theta(x) = \sum_{g \in G} \theta_g g(x) \]

- weighted $\ell_1$ norm of coefficients
  \[ \|\theta\|_1 = \sum_g a_g |\theta_g| \]

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

Barron Penalties and MDL

Penalized Likelihood Analysis for Continuous Parameters
$\ell_0$ and $\ell_1$ Penalties are Codelength-Valid and Risk-Valid
### Example Models

- **Regression** (focus of current presentation)
  
  \[ p_\theta(y \mid x) = \text{Normal}(f_\theta(x), \sigma^2) \]

- Logistic regression with \( y \in \{0, 1\} \)
  
  \[ p_\theta(y \mid x) = \text{Logistic}(f_\theta(x)) \quad \text{for } y = 1 \]

- Log-density estimation (focus of Festschrift paper)
  
  \[ p_\theta(x) = \frac{p_0(x) \exp\{f_\theta(x)\}}{c_f} \]

- Gaussian graphical models
\begin{itemize}
    \item $pen(\theta) = \lambda \|\theta\|_1$ where $\theta$ are coeff of $f_\theta(x) = \sum_{g \in G} \theta g(x)$
\end{itemize}
Regression with $\ell_1$ penalty

- $\ell_1$ penalized log-density estimation, i.i.d. case

$$\hat{\theta} = \arg\min_{\theta} \left\{ \frac{1}{n} \log \frac{1}{p_{f_\theta}(x)} + \lambda_n ||\theta||_1 \right\}$$

- Regression with Gaussian model, fixed $\sigma^2$

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

- Valid for

$$\lambda_n \geq \sqrt{\frac{2 \log(2M_G)}{n}}$$

with $M_G = \text{Card}(\mathcal{G})$
Adaptive risk bound

- For log density estimation with suitable $\lambda_n$

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

- For regression with fixed design points $x_i$, fixed $\sigma$, and

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

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

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

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

- 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
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 \( f_{\theta} \) 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 $\ell_1$ 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 representor shown by random draw is a Shannon-like demonstration of the variable cover (code)
- Similar approximation in analysis of greedy computation of $\ell_1$ penalized least squares
Fixed $\sigma$ versus unknown $\sigma$

- MDL with $\ell_1$ penalty for each possible $\sigma$. 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 $\sigma$.
- For unknown $\sigma$ suggest optimization over $\sigma$ as well as $\theta$

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

- 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 $\sigma$

- Best $\sigma$ for each $\theta$ solves the quadratic equation

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

- By the quadratic formula the solution is

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

1. Some Principles of Information Theory and Statistics
   - Data Compression
   - Minimum Description Length Principle
   - Two-stage Codes, Mixture Codes, Normalized Max Like.

2. Penalized Likelihood
   - Redundancy, Resolvability, and Risk
   - Risk-Valid Penalties
   - Penalized Likelihood Analysis for Continuous Parameters
   - $\ell_0$ and $\ell_1$ Penalties are Codelength-Valid and Risk-Valid

3. Summary
Summary

- We equated forms of MDL including NML and Bayes
- We related MDL and penalized likelihood
- Allow penalized likelihoods with continuous domains for $\theta$
- Information-theoretically valid penalties exceed complexity plus distortion of optimized representor of $\theta$
- Yields MDL interpretation and stat. risk $\leq$ resolvability
- $\ell_1$ penalty $\lambda_n \|\theta\|_1$ valid in regression and related problems for $\lambda_n \geq \sqrt{2(\log 2M)/n}$
- $\ell_0$ penalty $\frac{d}{2} \log n + C_d(\theta)$ is a classical codelength penalty.
- The next talk by Sabyasachi Chatterjee shows how to avoid the growth of $C_d(\theta)$ with the size of $\|\theta\|$.
- He shows that $\frac{d}{2} \log n + constant_d$ is conditionally codelength valid and risk valid.
Summary

- We equated forms of MDL including NML and Bayes
- We related MDL and penalized likelihood
- Allow penalized likelihoods with continuous domains for $\theta$
- Information-theoretically valid penalties exceed complexity plus distortion of optimized representor of $\theta$
- Yields MDL interpretation and stat. risk $\leq$ resolvability
- $\ell_1$ penalty $\lambda_n \|\theta\|_1$ valid in regression and related problems for $\lambda_n \geq \sqrt{2(\log 2M)/n}$
- $\ell_0$ penalty $\frac{d}{2} \log n + C_d(\theta)$ is a classical codelength penalty.
- The next talk by Sabyasachi Chatterjee shows how to avoid the growth of $C_d(\theta)$ with the size of $\|\theta\|$.
- He shows that $\frac{d}{2} \log n + constant_d$ is conditionally codelength valid and risk valid.
Summary

- We equated forms of MDL including NML and Bayes
- We related MDL and penalized likelihood
- Allow penalized likelihoods with continuous domains for $\theta$
- Information-theoretically valid penalties exceed complexity plus distortion of optimized representor of $\theta$
- Yields MDL interpretation and stat. risk $\leq$ resolvability
- $\ell_1$ penalty $\lambda_n \|\theta\|_1$ valid in regression and related problems for $\lambda_n \geq \sqrt{2(\log 2M)/n}$
- $\ell_0$ penalty $\frac{d}{2} \log n + C_d(\theta)$ is a classical codelength penalty.
- The next talk by Sabyasachi Chatterjee shows how to avoid the growth of $C_d(\theta)$ with the size of $\|\theta\|$.
- He shows that $\frac{d}{2} \log n + constant_d$ is conditionally codelength valid and risk valid.
We equated forms of MDL including NML and Bayes
We related MDL and penalized likelihood
Allow penalized likelihoods with continuous domains for $\theta$
Information-theoretically valid penalties exceed complexity plus distortion of optimized representor of $\theta$
Yields MDL interpretation and stat. risk $\leq$ resolvability
$\ell_1$ penalty $\lambda_n \|\theta\|_1$ valid in regression and related problems for $\lambda_n \geq \sqrt{2(\log 2M)/n}$
$\ell_0$ penalty $\frac{d}{2} \log n + C_d(\theta)$ is a classical codelength penalty.
The next talk by Sabyasachi Chatterjee shows how to avoid the growth of $C_d(\theta)$ with the size of $\|\theta\|$.
He shows that $\frac{d}{2} \log n + constant_d$ is conditionally codelength valid and risk valid.
Summary

- We equated forms of MDL including NML and Bayes
- We related forms of MDL and penalized likelihood
- Allow penalized likelihoods with continuous domains for $\theta$
- Information-theoretically valid penalties exceed complexity plus distortion of optimized representor of $\theta$
- Yields MDL interpretation and stat. risk $\leq$ resolvability
- $\ell_1$ penalty $\lambda_n \|\theta\|_1$ valid in regression and related problems for $\lambda_n \geq \sqrt{2(\log 2M)/n}$
- $\ell_0$ penalty $\frac{d}{2} \log n + C_d(\theta)$ is a classical codelength penalty.
- The next talk by Sabyasachi Chatterjee shows how to avoid the growth of $C_d(\theta)$ with the size of $\|\theta\|$.
- He shows that $\frac{d}{2} \log n + \text{constant}_d$ is conditionally codelength valid and risk valid.