Modern MDL Meets Data Mining
Insight, Theory, and Practice
—Part III—
Stochastic Complexity

Kenji Yamanishi
Graduate School of Information Science and Technology, the University of Tokyo
August 4th 2019
KDD Tutorial
Part III. Stochastic Complexity

3.1. Stochastic Complexity and NML Codelength

3.2. G-function and Fourier Transformation

3.3. Latent Variable Model Selection
   3.3.1. Latent Stochastic Complexity
   3.3.2. Decomposed NML Codelength
   3.3.3. Applications to a Variety of Models

3.4. High-dimensional Sparse Model Selection
   3.4.1. High Dimensional Penalty Selection
   3.4.2. High Dimensional Minimax Prediction
3.1. Stochastic Complexity and NML Codelength
What is Information?

- Case 1: Data distribution is known

Data sequence: \( \mathbf{x} = x_1, \ldots, x_n \)

Probability mass (density) function: 
\[
p(\mathbf{x})
\]

Shannon information: 
\[
I(\mathbf{x}) = - \log p(\mathbf{x})
\]
Characterization of Shannon Information

Theorem 3.1.1 (Shannon’s Source Coding Theorem)

\[ \forall \mathcal{L} : \text{prefix code-length function,} \]

\[ E_p[\mathcal{L}(x)] \geq E_p[-\log p(x)] = H_n(p) \]

Entropy

Note:

Prefix code-length function \( \mathcal{L} \) is defined as \( \mathcal{L}(x) \geq 0 \) and \( \sum_x 2^{-\mathcal{L}(x)} \leq 1 \)\n
Kraft’s Inequality

Shannon information gives optimal code length when the true distribution is known in advance.
Case 2: Data distribution is unknown.

Data seq.: \( \mathbf{x} = x_1, \ldots, x_n \)

Probabilistic model class.: 

\[
\mathcal{P}_M = \{ p(\mathbf{x}; \theta) : \theta \in \Theta_M \} \\
M: \text{model}
\]

Normalized Maximum Likelihood (NML) Distribution

\[
p_{\text{NML}}(\mathbf{x}; M) = \frac{\max_{\theta} p(\mathbf{x}; \theta)}{\sum_y \max_{\theta} p(y; \theta)}
\]

\( \left( \text{Note: } \sum_{\mathbf{x}} \max_{\theta} p(\mathbf{x}; \theta) > 1 \right) \)

Stochastic Complexity of \( \mathbf{x} \) relative to \( P_M \)

\( = \text{NML codelength} \)

\[
I(\mathbf{x}^n; M) = - \log p_{\text{NML}}(\mathbf{x}; M)
= - \log \max_{\theta} p(\mathbf{x}; \theta) + \log C_n(M)
\]

Parametric Complexity of \( P_M \)

\[
C_n(M) = \sum_{\mathbf{x}} \max_{\theta} p(\mathbf{x}; \theta)
\]
Theorem 3.1.2  (Minimax optimality of NML codelength)
[Shtarkov  Probl. Inf. Trans. 87]

NML codelength achieves the minimum of the risk

$$I(x; M) = \arg\min_{\hat{M}} \max_{x} \left\{ \mathcal{L}(x) - \left( -\log \max_{\theta} p(x; \theta) \right) \right\}$$

Shtarkov’s minimax risk  baseline
How to calculate Parametric Complexity

Theorem 3.1.3 （Asymptotic formula for parametric complexity）[Rissanen IEEE IT1996]

Under the condition of central limit theorem:
\( \sqrt{n}(\hat{\theta}(\mathbf{x}) - \theta) \sim \mathcal{N}(0, I^{-1}(\theta)), \)
(\( \hat{\theta} = \arg\max_{\theta} p(\mathbf{x}; \theta) \) : maximum likelihood estimator), it holds:

\[
\log C_n(M) = \log \sum_{\mathbf{x}} \max_{\theta} p(\mathbf{x}; \theta) \\
= \frac{k}{2} \log \frac{n}{2\pi} + \log \int |I(\theta)|^{1/2} d\theta
\]

\( \lim_{n \to \infty} o(1) = 0 \) uniformly over \( \mathbf{x}. \)

\( k: \# \text{parameters, } n: \text{data length}, \)
\( I(\theta) = \lim_{n \to \infty} E_p[-\frac{1}{n} \frac{\partial^2 \log p(\mathbf{x}; \theta)}{\partial \theta \partial \theta^T}] \) (Fisher information matrix)
Example 3.1.1 (Multinomial Distribution)

\[ X = \{0, 1, \ldots, K\} \]

\[ p(X = i; \theta) = \theta_i \ (i = 0, \ldots, K), \]

\[ \Theta_K = \{\theta = (\theta_0, \ldots, \theta_K) : \sum_{i=0}^{K} \theta_i = 1, \ \theta_i \geq 0\} \]

\[ x = x_1, \ldots, x_n: \text{ data sequence} \]

\[ n_i: \# \text{ occurrences of } X = i \]

\[ \hat{\theta} = \left( \frac{n_0}{n}, \ldots, \frac{n_K}{n} \right): \text{ m.l.e. of } \theta \]

\[ I(\theta) = \prod_{i=0}^{K} \theta_i^{-1}: \text{ Fisher inf.} \]

Stochastic Complexity

\[
I(x^n; K) = - \log p(x; \hat{\theta}) + \frac{K}{2} \log \frac{n}{2\pi} + \log \int \sqrt{|I(\theta)|} \, d\theta
\]

\[
= n H \left( \frac{n_0}{n}, \ldots, \frac{n_K}{n} \right) + \frac{K}{2} \log \frac{n}{2\pi} + \log \frac{\pi^{K+1}}{\Gamma \left( \frac{K+1}{2} \right)}
\]

where \( H(z_0, \ldots, z_K) = - \sum_{i=0}^{K} z_i \log z_i. \)
Example 3.1.2 (1-dimensional Gaussian distribution)

\[ p(X; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp \left\{ -\frac{(X - \mu)^2}{2\sigma^2} \right\} \]

Parameter Space \( \tau = \sigma^2, \ \Theta = \{ (\mu, \tau) : \mu \in (-\infty, +\infty), \tau > 0 \} \)

Fisher Information

\[ I(\mu, \tau) = \begin{pmatrix} 1/\tau & 0 \\ 0 & 1/2\tau^2 \end{pmatrix}, \quad |I(\mu, \tau)| = \frac{1}{2\tau^3} \]

m.l.e.

\[ \hat{\mu} = \frac{1}{n} \sum_{t=1}^{n} x_t, \quad \hat{\tau} = \frac{1}{n} \sum_{t=1}^{n} (x_t - \hat{\mu})^2 \]

s, r: smallest integers such that \( \hat{\mu} \leq 2^s, \ \hat{\tau} \geq 2^{-2r} \)

Restricted Parameter Space \( \tilde{\Theta} = \{ (\mu, \tau) : \mu \leq 2^s, \ \tau \geq 2^{-2r} \} \)

\[ \int_{(\mu, \tau) \in \tilde{\Theta}} \sqrt{|I(\mu, \tau)|} d\mu d\tau = 2^{s+r+1/2} \]

Stochastic complexity

\[ \frac{n}{2} \log(2\pi e \hat{\tau}) + \log \frac{n}{2\pi} + \frac{1}{2} + s + r + \log^* s + \log^* r. \]

\[ \log^* x = \log c + \log x + \log \log x + \ldots (c = 2.865): \text{codelength of an integer } x \]
Characterization of Stochastic Complexity

Theorem 3.1.4 (Rissanen’s lower bound) [Rissanen 1989]

Under the assumption of the central limit theorem:
\[ \sqrt{n}(\hat{\theta}(x^n) - \theta) \sim \mathcal{N}(0, I^{-1}(\theta)), \]
except points in \( \Theta_0 \) such that \( \text{vol}(\Theta_0) \to 0 \) \( (n \to \infty) \),
\( \forall \epsilon > 0, \forall L: \) prefix codelength function, it holds:

\[
E_\theta[L(x)] \geq E_\theta[-\log p(x; \theta)] + \frac{k - \epsilon}{2} \log n
\]
\[= E_\theta[I(x : M)] + o(\log n). \]

Stochastic complexity gives optimal codelength when the true distribution is unknown.
Sequential NML Code length

Sequential NML (SNML)
= Cumulative code length for sequential NML coding

\[ \mathbf{x} = x_1, \ldots, x_n \quad : \text{data sequence} \]

\[ \tilde{I}(\mathbf{x}; M) = \sum_{t=1}^{n} \left( - \log \frac{p(x_t; \hat{\theta}(x_t \cdot x^{t-1}))}{\sum_X p(X; \hat{\theta}(X \cdot x^{t-1}))} \right) \]

Theorem 3.1.5 (Property of SNML)
For model classes under some regular conditions

\[ I(\mathbf{x}; M) = \tilde{I}(\mathbf{x}; M) + o(\log n) \]

E.g. Regression model \[ \text{[Rissanen IEEE IT2000]} \]
\[ \text{[Roos, Myllymaki, Rissanen MVA 2009]} \]
Example 3.1.3. (Auto-regression model)

\[
p(x_t|x_{t-1}; \theta) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp \left\{ -\frac{1}{2\sigma^2} \left( x_t - \sum_{i=1}^{k} a^{(i)} x_{t-i} \right)^2 \right\}
\]

\[
\bar{x}_t = (x_{t-1}, \ldots, x_{t-k})^T, \quad \theta^T = (a^T, \sigma^2) = (a^{(1)}, \ldots, a^{(k)}, \sigma^2)
\]

\[
\hat{a}_t = \arg\min_{a \in \mathbb{R}^k} \sum_{j=1}^{t} (x_j - a^T \bar{x}_j)^2 = V_t \sum_{j=1}^{t} \bar{x}_j x_j = \hat{a}_{t-1} + \frac{V_{t-1} \bar{x}_t (x_t - \bar{x}_t \hat{a}_{t-1})}{1 + c_t}.
\]

where

\[
V_t \overset{\text{def}}{=} \left( \sum_{j=1}^{t} \bar{x}_j \bar{x}_j^T \right)^{-1} = V_{t-1} - \frac{V_{t-1} \bar{x}_t \bar{x}_t^T V_{t-1}}{1 + c_t}, \quad c_t \overset{\text{def}}{=} \bar{x}_t^T V_{t-1} \bar{x}_t.
\]

When \( \sigma \) is fixed, then SNNL dist. becomes

\[
p_{\text{SNML}}(x_t|x^{t-1}) = \frac{p(x_t|x^{t-1}; \hat{a}_t)}{K(x^{t-1})} = \frac{1}{\sqrt{2\pi(1 + c_t)^2}\sigma^2} \exp \left( -\frac{(y_t - \hat{a}_t^T \bar{x}_t)^2}{2(1 + c_t)^2\sigma^2} \right).
\]
MDL Criterion (1/2)

$\mathcal{M} = \{M\}$: model class

NML distribution for fixed $M$:

$$\hat{p}(\boldsymbol{x}; M) = \frac{\max_{\theta} p(\boldsymbol{x} : \theta, M)}{C_n(M)}, \quad C_n(M) = \sum_{\boldsymbol{x}} \max_{\theta} p(\boldsymbol{x}; \theta, M)$$

Stochastic complexity of $\boldsymbol{x}$ relative to $\mathcal{M}$:

$$I(\boldsymbol{x}; \mathcal{M}) = -\log \frac{\max_{M \in \mathcal{M}} \{\hat{p}(\boldsymbol{x}; M)\}}{C}$$

$$= \min_{M \in \mathcal{M}} \left\{ -\max_{\theta} \log p(\boldsymbol{x}; \theta, M) + \log C_n(M) \right\} + \log C$$

MDL (Minimum Description Length) criterion

where $C = \sum_{\boldsymbol{x}} \max_{M \in \mathcal{M}} \{\hat{p}(\boldsymbol{x}; M)\}$ → does not depend on $M$. 
MDL Criterion (2/2)

\[ \mathcal{M} = \{ M \}: \text{ model class} \]  

Under the condition of central limit theorem:
\[
\sqrt{n}(\hat{\theta}(\mathbf{x}) - \theta) \sim \mathcal{N}(0, 1/I(\theta)),
\]

\[ \mathbf{x} = x_1, \ldots, x_n: \text{ given} \]

\[
I(\mathbf{x}; M) + \ell(M)
\]
\[
= -\log \max_{\theta} p(\mathbf{x}; \theta) + \frac{k}{2} \log \frac{n}{2\pi} + \log \int \sqrt{|I(\theta)|} d\theta + \ell(M)
\]
\[ \implies \min \text{ w.r.t. } M. \]

\( k \): #parameters, \( n \): data length, \( I(\theta) = \lim_{n \to \infty} E_p \left[ -\frac{1}{n} \frac{\partial^2 \log p(\mathbf{x}; \theta)}{\partial \theta \partial \theta^T} \right] \)

\( \ell(M) \): Complexity of \( M \) s.t. \( \sum_{M \in \mathcal{M}} \exp(-\ell(M)) \leq 1. \)
Example 3.1.4. (Histogram Density)

\( X = [0, 1] \ldots K \) disjoint cells

\[ X = \bigcup_{i=0}^{K} C_i, \quad C_i = \left[ \frac{i}{K+1}, \frac{i+1}{K+1} \right) \quad (i = 0, \ldots, K - 1), \quad C_K = \left[ \frac{K}{K+1}, 1 \right]. \]

\( \theta_i = \text{Prob}(X \in C_i) \)

Class of histogram densities with equal length cells:

\[
\mathcal{P}_{\text{HIS}} = \left\{ p(X; \theta) : \theta = (\theta_0, \ldots, \theta_k), \theta_i \geq 0, \sum_{i=0}^{k} \theta_i = 1 \right\}
\]

if \( X \in C_i \), then \( p(X; \theta) = (K+1)\theta_i \) \((i = 0, \ldots, K)\).

\[
I(x^n; K) = \min_{\theta} \left\{ -\log p(x^n; \theta) \right\} + \frac{K}{2} \log \frac{n}{2\pi} + \log \int \sqrt{I(\theta)} \, d\theta + \mathcal{L}(K)
\]

\[
= -\sum_{i=0}^{K} n_i \log \frac{n_i}{n} - n \log(K+1) + \frac{K}{2} \log \frac{n}{2\pi} + \log \frac{\pi^{K+1}}{\Gamma \left( \frac{K+1}{2} \right)} + \log^* K
\]

\( \Rightarrow \) \( \min \) w.r.t. \( K \) (optimal \( K \))
Optimality of MDL Estimation

**Theorem 3.1.6 (Optimality of MDL estimator) [Rissanen 2012]**

\[ \bar{\theta}, \bar{M}: \text{ any given estimators} \]

Define a **general normalized distribution** associated with \( \bar{\theta} \) and \( \bar{M} \) by

\[
\bar{p}(\mathbf{x}) = \frac{p(\mathbf{x}; \bar{\theta}, \bar{M})}{\sum_{\mathbf{y}} p(\mathbf{y}; \bar{\theta}, \bar{M})},
\]

then it holds:

\[
\min_{\bar{\theta}, \bar{M}} \max_{\theta, M} D(p_{\theta, M} \| \bar{p}) = \log \hat{C}_n(\hat{M}) + \log \hat{C}_n
\]

**Kullback-Leibler divergence**

\[
D(f \| g) = E_f[\log(f(x)/g(x))]
\]

minimum is achieved by \( \bar{\theta} = \hat{\theta}(\text{m.l.e.}) \), \( \bar{M} = \hat{M} \) **(MDL estimator)**.

\[
\hat{C}_n(M) \equiv \sum_{\mathbf{x}} \max_{\theta} p(\mathbf{x}; \theta, M)
\]

\[
\hat{C}_n \equiv \sum_{\mathbf{x}} \max_{M} \left\{ \max_{\theta} p(\mathbf{x}; \theta, M) / C_n(M) \right\}
\]
Why is the MDL?

- Optimal solution to Shtarkov minimax risk
- Attaining Rissanen’s lower bound
- Consistency  [Rissanen  Auto. Control 1978]  
  [Barron 1989]
- Index of Resolvability  [Barron and Cover IEEE IT1991]
- Rapid convergence rate with PAC learning  
  [Yamanishi  Mach. Learning 1992]  
  [Rissanen Yu, Learn. and Geo. 1996]  
  [Chatterjee and Barron ISIT2014]  
  [Kawakita, Takeuchi, ICML2016]
- Estimation optimality  [Rissanen 2012]
- A huge number of case studies
3.2. G-Function and Fourier Transformation
Computational Problem in Parametric Complexity

Parametric Complexity

\[ C_n(M) = \sum_{x} \max_{\theta} p(x; \theta) \]

It is hard to calculate for general model M. Beyond Rissanen’s asymptotic formulae for small data

1) Calculate it non-asymptotically.
2) Calculate it exactly and efficiently.

A) g-function
B) Fourier transformation
C) Combinatorial Methods
g-Function

Density decomposition

\[ p(\mathbf{x}; \theta) = p(\mathbf{x} | \hat{\theta}(\mathbf{x})) g(\hat{\theta}(\mathbf{x}); \theta) \]

where

\[ \hat{\theta}(\mathbf{x}) \overset{\text{def}}{=} \arg\max_{\theta} p(\mathbf{x}; \theta): \text{m.l.e.} \]

\[ g(\bar{\theta}; \theta) \overset{\text{def}}{=} \sum_{\mathbf{x}: \hat{\theta}(\mathbf{x}) = \bar{\theta}} p(\mathbf{x}; \theta) \]

\( g(\bar{\theta}; \theta) \) is a probability density function of \( \bar{\theta} \) (g-function).

\[ \int g(\bar{\theta}; \theta) d\bar{\theta} = \int d\bar{\theta} \left( \sum_{\mathbf{y}: \hat{\theta}(\mathbf{y}) = \bar{\theta}} p(\mathbf{y}; \theta) \right) = 1. \]
Calculation of Parametric Complexity via g-function

\[ C_n(M) = \sum_x p(x; \hat{\theta}(x)) \]

\[ = \int d\hat{\theta} \sum_{y: \hat{\theta}(y)=\hat{\theta}} p(y; \hat{\theta}) \]

\[ = \int g(\hat{\theta}; \hat{\theta}) d\hat{\theta}. \]

Variable transformation
Example 3.2.1 (g-function for exponential distributions)

\[ \mathcal{P}_{\text{Exp}} \overset{\text{def}}{=} \{ p(X; \theta) = \theta \exp(-\theta X) : \theta \in \mathbb{R} \} \]

\[ x = x_1, \ldots, x_n: \text{data sequence} \quad \hat{\theta}(x^n) = \frac{n}{\sum_{i=1}^{n} x_i} : \text{m.l.e.} \]

\[
p(x; \theta) = \exp \left\{ -\theta \sum_{i=1}^{n} x_i + n \log \theta \right\}
\]

\[
= \theta^n \exp \left\{ -\frac{n\theta}{\hat{\theta}} \right\}
\]

\[
= f(x | \hat{\theta}(x)) g(\hat{\theta}(x); \theta).
\]

\[ \Rightarrow \quad \text{Gamma dist. of } n/\hat{\theta}(x^n) \text{ with shape } n \text{ and scale } 1/\theta. \]

\[
g(\hat{\theta}(x); \theta) = \frac{\theta^n n^n}{\Gamma(n)\hat{\theta}(x)^{n+1}} \exp \left\{ -\theta \cdot \frac{n}{\hat{\theta}(x)} \right\}
\]
Fix $\hat{\theta}(x^n) = \hat{\theta}$,
\[ g(\hat{\theta}; \hat{\theta}) = \frac{n^n}{e^n(n - 1)!} \cdot \frac{1}{\hat{\theta}}. \]

Since $\int g(\hat{\theta}; \hat{\theta}) d\hat{\theta}$ diverges, restrict
\[ Y(\theta_{\min}, \theta_{\max}) \overset{\text{def}}{=} \{ \hat{\theta} : \theta_{\min} \leq \hat{\theta} \leq \theta_{\max} \} \]

where $\theta_{\min}, \theta_{\max}$ are given constants.

Then
\[
C_n = \int_{Y(\theta_{\min}, \theta_{\max})} g(\hat{\theta}; \hat{\theta}) d\hat{\theta} \\
= \frac{n^n}{e^n(n - 1)!} \int_{\theta_{\min}}^{\theta_{\max}} \frac{1}{\hat{\theta}} d\hat{\theta} \\
= \frac{n^n}{e^n(n - 1)!} \log \frac{\theta_{\max}}{\theta_{\min}}.
\]

$\Rightarrow$ Extended into general exponential family
[Hirai and Yamanishi IEEE IT2013]
Fourier Transformation Method (1/2)

Theorem 3.2.1 (Parametric complexity based on Fourier transformation) [Suzuki and Yamanishi ISIT 2018]

\[ C_n(M) = \sum_{x} \max_{\theta} p(x; \theta) = \int d\theta h(\theta), \]

where for \( \xi \in \mathbb{R}^k \) (\( k \): \# parameters),

\[ h(\theta) \overset{\text{def}}{=} \frac{1}{(2\pi)^k} \int d\xi \sum_{x} p(x; \theta) \exp(i\xi^T(\hat{\theta}(x) - \theta)), \]

where \( \hat{\theta}(x) = \arg\max_{\theta} p(x; \theta). \)
Proof Sketch.

\( \tilde{p}(x; \xi) \): Fourier transformation of \( p(x; \theta) \)

\[
\tilde{p}(x; \xi) = \frac{1}{(2\pi)^{k/2}} \int d\theta \exp(-i\xi^T \theta)p(x; \theta),
\]

\[
p(x; \theta) = \frac{1}{(2\pi)^{k/2}} \int d\xi \exp(i\xi^T \theta)\tilde{p}(x; \xi).
\]

\[
\sum_x \max_{\theta} p(x; \theta) = \sum_x \frac{1}{(2\pi)^{k/2}} \int d\xi \exp(i\xi^T \hat{\theta}(x))\tilde{p}(x; \xi)
\]

\[
= \frac{1}{(2\pi)^{k/2}} \int d\xi \sum_x \int d\theta \exp(i\xi^T (\hat{\theta}(x) - \theta))p(x; \theta)
\]

\[
= \frac{1}{(2\pi)^{k/2}} \int d\theta \int d\xi \sum_x p(x; \theta) \exp(i\xi^T (\hat{\theta}(x) - \theta))
\]

\[
= \int d\theta h(\theta).
\]
Exponential Family

\[ p(x; \eta) = m(x) \exp(\eta^T t(x)) / Z(\eta). \]

\[ \tau(\eta) = \int dx \cdot t(x)p(x; \eta) \]

Theorem 3.2.2 (Parametric complexity of Exponential Family with Fourier method) [Suzuki and Yamanishi ISIT 2018]

\[ \mathbf{x} = x_1, \ldots, x_n \]

\[ \int d\mathbf{x} \max_\tau p(\mathbf{x}; \eta(\tau)) = \frac{1}{(2\pi)^k} \int d\tau \int d\xi \exp(-\xi^T \tau) \left( \frac{Z(\eta(\tau)) + i\xi/n}{Z(\eta(\tau))} \right)^n \]
## Parametric Complexity of Exponential Family Computed with Fourier Method

[Suzuki and Yamanishi ISIT 2018]

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Density function $f(x^N; \theta)$</th>
<th>Sufficient statistics $u_k(x)$</th>
<th>Canonical parameter $\eta$</th>
<th>Partition function $Z(\eta)$</th>
<th>Parametric complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal dist. with known variance $\nu$</td>
<td>$f(x^N; \mu)$</td>
<td>$x$</td>
<td>$\eta = \frac{\mu}{\nu} \in (-\infty, +\infty)$</td>
<td>$v^{\frac{1}{2}} \exp \left( \frac{1}{2} v \eta^2 \right)$</td>
<td>$\int_{-\infty}^{+\infty} d\mu \frac{w(\mu)}{\sqrt{2\pi N}}$</td>
</tr>
<tr>
<td>Normal dist. with known mean $\mu$</td>
<td>$f(x^N; \nu)$</td>
<td>$(x - \mu)^2$</td>
<td>$\eta = -\frac{1}{2\nu} \in (-\infty, 0)$</td>
<td>$\frac{1}{\sqrt{-\eta}}$</td>
<td>$(\frac{1}{2} N)^{\frac{1}{2}} N \exp(-\frac{1}{2} N) \Gamma\left(\frac{1}{2} N\right)$ \times $\int_{0}^{+\infty} d\nu \frac{w(\nu)}{\nu}$</td>
</tr>
<tr>
<td>Laplace dist. with known mean $\mu$</td>
<td>$f(x^N; b)$</td>
<td>$</td>
<td>x - \mu</td>
<td>$</td>
<td>$\eta = -\frac{1}{b} \in (-\infty, 0)$</td>
</tr>
<tr>
<td>Gamma dist. with known shape $k$</td>
<td>$f(x^N; \mu)$</td>
<td>$x$</td>
<td>$\eta = -\frac{k}{\mu} \in (-\infty, 0)$</td>
<td>$\frac{1}{(-\eta)^k}$</td>
<td>$(kN)^{\frac{1}{2}} N \exp(-kN) \Gamma(kN) \times \int_{0}^{+\infty} d\mu \frac{w(\mu)}{\mu}$</td>
</tr>
<tr>
<td>Weibull dist. with known shape $k$</td>
<td>$f(x^N; L)$</td>
<td>$x^k$</td>
<td>$\eta = -\frac{1}{L} \in (-\infty, 0)$</td>
<td>$\frac{1}{-\eta}$</td>
<td>$\frac{N^N \exp(-N)}{\Gamma(N)} \times \int_{0}^{+\infty} dL \frac{w(L)}{L}$</td>
</tr>
<tr>
<td>Gamma dist. with known scale $\theta$</td>
<td>$f(x^N; \eta)$</td>
<td>$\log x$</td>
<td>$\eta = \psi^{-1}(\lambda - \log \theta) - 1 \in (-1, +\infty)$</td>
<td>$\Gamma(\eta + 1) \theta^{\eta+1}$</td>
<td>See (22)</td>
</tr>
</tbody>
</table>
3.3. Latent Variable Model Selection
Motivation 1. Topic Model

How many topics lie in a document?

LDA: Latent Dirichlet Allocations

#topics = 3

Document 1

Document 2

Document D

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine: 0.05</td>
<td>Matrix: 0.04</td>
<td>Bio: 0.05</td>
</tr>
<tr>
<td>Computer: 0.03</td>
<td>Math: 0.03</td>
<td>Gene: 0.04</td>
</tr>
<tr>
<td>Algorithm: 0.02</td>
<td>equation: 0.01</td>
<td>Array: 0.01</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Latent Variable Model Selection

Motivation 2: Relational Model

Permutation

#blocks = 6

How many blocks lie in a relational data?
Latent Variable Models

\( X \): observed variable, \( Z \): latent variable

\[
p(X; \theta) = \sum_Z p(X, Z; \theta)
\]

Example 3.4.1 (Gaussian mixture model)

\[
p(X, Z = k; \theta) = P(Z = k) \mathcal{N}(X; \mu_k, \Sigma_k)
\]

\( Z \): cluster assignment, \( K \): \# components
\( \mathcal{N}(\mu_k, \Sigma_k) \): normal dist. with mean \( \mu_k \) and variance-covariance matrix \( \Sigma_k \)

\[
\pi_k = P(Z = k), \quad \theta = (\pi_k, \mu_k, \Sigma_k) \quad (k = 1, \ldots, K)
\]
Nonidentifiability of Latent Variable Model

Marginalized model

\[ p(X; \theta) = \sum_Z p(X, Z; \theta) = \sum_Z p(Z)p(X|Z; \theta_Z) \]

Finite mixture model

\[ p(X; \pi, \theta_1, \theta_2) = \pi p(X; \theta_1) + (1 - \pi) p(X; \theta_2) \]

Non-identifiable

\[ \theta_1 = \theta_2 \implies \text{Prob. dist. is identical for any } \pi. \]

Complete variable model

\[ p(X, Z; \theta) \]

Estimate \( Z \) from \( X \)

Identifiable
3.3.1. Latent Stochastic Complexity

Normalized maximum likelihood (NML) Codelegngh for complete variable model

\[ L_{NML}(X, Z; K) \]

Latent Stochastic Complexity (LSC)

\[ -\log \max_{\theta} p(\mathbf{x}, z; \theta, M) + \log \sum_{\mathbf{y}} \sum_{\mathbf{w}} \max_{\theta} p(\mathbf{y}, \mathbf{w}; \theta, M) \]

\[ \Rightarrow \min \text{ w.r.t. } M \]

\( \mathbf{x} = x_1, \ldots, x_n \): data sequence
\( z = z_1, \ldots, z_n \): latent variable sequence
Finite Mixture Model

\[ p(X, Z; \theta, K) = p(X|Z; \theta_1)p(Z; \theta_2) \]

\[ p(X; \theta, K) = \sum_{Z=1}^{K} p(X|Z; \theta_1)p(Z; \theta_2) \quad K: \text{# components} \]

\[ p(Z = i; \theta_2) = \phi_i \quad (i = 1, \ldots, K), \quad \theta_2 = (\phi_1, \ldots, \phi_K) \]

\[ C_n(K) = \sum_z \left( \int \max_{\theta} p(\mathbf{x}, z; \theta, K) d\mathbf{x} \right) \]

\[ = \sum_{\theta_2} \max_{\theta_2} p(z; \theta_2) \int p(\mathbf{x}|z; \theta_1) d\mathbf{x} \]

\[ = \sum_{n_1 + \cdots + n_K = n \atop n_i \geq 0 \; i = 1, \ldots, K} \frac{n!}{n_1! \cdots n_K!} \left( \frac{n_1}{n} \right)^{n_1} \cdots \left( \frac{n_K}{n} \right)^{n_K} \prod_{k=1}^{K} C_{n_k} \]

\[ \Rightarrow O(K^n) \text{ computation time} \]

where \[ \bar{C}_n = \int_{\theta_1} \max_{\theta_1} p(\mathbf{x}|z; \theta_1) d\mathbf{x} \]
Parametric Complexity of FMM

Theorem 3.3.1 (Recurrence relation for parametric complexity fo FMM) [Hirai and Yamanishi IEEE IT2013]

\[ C_n(K + 1) = \sum \frac{n!}{r_1!r_2!} \left( \frac{r_1}{n} \right)^{r_1} \left( \frac{r_2}{n} \right)^{r_2} C_{r_1}(K)\overline{C}_{r_2} \]

\( r_1 + r_2 = n \)
\( r_1 \geq 0, r_2 \geq 0 \)

\( K: \# \) components, \( n: \) data length

\[ \implies \text{computable in } O(n^2K) \text{ time} \]
Example 3.3.1 (Parametric complexity of multivariate GMM)

\[ C_n(K + 1) = \sum_{r_1 + r_2 = n, r_1 \geq 0, r_2 \geq 0} \frac{n!}{r_1!r_2!} \left( \frac{r_1}{n} \right)^{r_1} \left( \frac{r_2}{n} \right)^{r_2} C_{r_1}(K) \overline{C}_{r_2} \]

where

\[ \overline{C}_n = \int_{\mathbf{x} \in \mathcal{X}} p(\mathbf{x}; \hat{\mu}(\mathbf{x}), \hat{\Sigma}(\mathbf{x})) d\mathbf{x} \]

\[ \leq \frac{2^{d+1} R^d}{d^{d+1} \Gamma \left( \frac{d}{2} \right) \Gamma_d \left( \frac{n-1}{2} \right)} \left( \frac{n}{2e} \right)^{\frac{dn}{2}} \prod_{j=1}^{d} \left( \lambda_{\min}^{(j)} \right)^{-\frac{d}{2}} \]

\( d \): data dimension \( \mathcal{X} \triangleq \{ \mathbf{x} : \|\hat{\mu}(\mathbf{x})\| \leq R, \lambda_{\min}^{(j)} \leq \hat{\lambda}_j(\mathbf{x}) \} \)

\( R > 0, \lambda_{\min}^{(j)} (j = 1, \ldots, d) \): given, \( \lambda_j \): the \( j \)-th largest eigenvalue of \( \hat{\Sigma} \)
Latent Variable Model Selection with Latent Stochastic Complexity (1/2)

• Naïve Bayes Model
  [Kontkanen and Myllymaki 2008]

• Gaussian Mixture Model
  [Kyrgyzov, Kyrgyzov, Maître, Campedel MLDM2007]
  [Hirai and Yamanishi IEEE IT 2013, 2019]

• General Relation Model
  [Sakai, Yamanishi IEEE BigData 2013]

• Principal Component Analysis/Canonical Component Analysis
  [Archambeau, Bach NIPS2009]
  [Virtanen, Klami, Kaski ICML2011]
  [Nakamura, Iwata, Yamanishi DSAA2017]
Latent Variable Model Selection with Latent Stochastic Complexity(2/2)

- Non-negative Matrix Factorization (NMF) Rank Estimation
  [Miettinen, Vreeken KDD2011]
  [Yamauchi, Kawakita, Takeuchi ICONIP2012]
  [Ito, Oeda, Yamanishi SDM2016]
  [Squire, Benett, Niranjyan NeuroComput2017]
  c.f. [Cemgil CIN2009] Variational Bayes
  [Hoffman, Blei,Cook 2010] Non-parametric Bayes

- Convolutional NMF
  [Suzuki, Miyaguchi, Yamanishi ICDM2016]

- Non-negative Tensor Factorization (NTF) Rank Estimation
  [Fu, Matsushima, Yamanishi Entropy2019]
3.3.2. Decomposed Normalized Maximum Likelihood (DNML) Codelength

Computable efficiently and non-asymptotically

\[ L_{DNML}(X, Z; K) \]

\[ L_{NML}(Z; K) + \]

\[ L_{NML}(X|Z; K) \]

[Wu, Sugawara, Yamanishi KDD2017]
[Yamanishi, Wu, Sugawara, Okada DAMI2019]
How to Compute DNML

\[ p(X, Z; \theta, K) = p(X|Z; \theta_1)p(Z; \theta_2) \]

\( K: \# \text{ latent variables} \)
\[ x = x_1, \ldots, x_n: \text{ observed data sequence} \]
\[ z = z_1, \ldots, z_n: \text{ latent variable sequence} \]

DNML criterion

\[ \mathcal{L}_{\text{DNML}}(x, z; K) \overset{\text{def}}{=} \mathcal{L}_{\text{NML}}(x|z; K) + \mathcal{L}_{\text{NML}}(z; K) \]

\[ \implies \min \text{ w.r.t. } K \]

\[ \mathcal{L}_{\text{NML}}(x|z; K) = -\log \max_{\theta_1} p(x|z; \theta_1) + \log \sum_{\theta_1} \max_{x} p(x|z; \theta_1) \]

\[ \mathcal{L}_{\text{NML}}(z; K) = -\log \max_{\theta_2} p(z; \theta_2) + \log \sum_{\theta_2} \max_{z} p(z; \theta_2) \]
Finite Mixture Models

\( z_k \) subsequence of \( z \) s.t. \( Z = k \)
\( x_k \) data sequence corresponding to \( z_k \)

\[
\mathcal{L}_{\text{NML}}(x | z; K) = - \log \prod_{k} \max_{\theta_1} p(x_k | z_k; \theta_1) + \log \prod_{k} \left( \sum_{x_k} \max_{\theta_1} p(x_k | z_k; \theta_1) \right)
\]

\[
= \sum \mathcal{L}_{\text{NML}}(x_k | z_k)
\]

where \( \mathcal{L}_{\text{NML}}(x_k | z_k) = - \log \max_{\theta_1} p(x_k | z_k; \theta_1) + \log \sum_{x_k} \max_{\theta_1} p(x_k | z_k; \theta_1). \)

\[\theta_2 = (\phi_1, \ldots, \phi_K) \quad (\sum_k \phi_k = 1, \phi_k \geq 0) \quad \hat{\theta}_2 = \left( \frac{n_1}{n}, \ldots, \frac{n_K}{n} \right)\]

Multinomial distribution

\[
\mathcal{L}_{\text{NML}}(z; K) = - \log \prod_{l=1}^{K} \left( \frac{n_k}{n} \right)^{n_k} + \log C_{n}^{\text{MN}}(K)
\]

\[
C_{n}^{\text{MN}}(K) \overset{\text{def}}{=} \sum \frac{n!}{n_1! \cdots n_K!} \prod_{k=1}^{K} \left( \frac{n_k}{n} \right)^{n_k}
\]

\( n_1 + \cdots + n_K = n \)
\( n_i \geq 0, \ i = 1, \ldots, K \)
Efficient Computation of Parametric Complexity for Multinomial Distribution

Theorem 3.3.2 (Recurrence relation of parametric complexity of multinomial distribution) [Kontkanen, Myllymaki IPL 2006]

\[
C_n^{MN}(K) \overset{\text{def}}{=} \sum_{n_1 + \cdots + n_K = n} \frac{n!}{n_1! \cdots n_K!} \prod_{k=1}^{K} \left( \frac{n_k}{n} \right)^{n_k}
\]

\[n_i \geq 0, \quad i = 1, \ldots, K\]

\[
C_n^{MN}(K) = C_n^{MN}(K - 1) + \frac{n}{K - 2} C_n^{MN}(K - 2)
\]

\[\Rightarrow O(n + K) \text{ comput. time}\]
3.3.3 Applications to a Variety of Classes

Example 3.3.2 (Latent Dirichlet Allocation: LDA)

1. For topic $k = 1, \ldots, K$:
   - Generate a word distribution $\phi_k \sim \text{Dir}(\beta)$.

2. For document $d = 1, \ldots, D$:
   a. Generate a topic mixture $\theta_d \sim \text{Dir}(\alpha)$.
   b. For word $i = 1, \ldots, n_d$ in document $d$:
      i. Generate a latent variable $z_{di} \sim \text{Multi}(\theta_d)$.
      ii. Generate an observed variable $x_{di} \sim \text{Multi}(\phi_{z_{di}})$.

\[
\mathcal{L}_{\text{DNML}}(x, z; K) = \sum_d \sum_k n_{kv} \left( \log n_k - \log n_{kv} \right) + \sum_k \log C_{\text{MN}}(n_k, V) \\
+ \sum_d \sum_k n_{dk} \left( \log n_d - \log n_{dk} \right) + \sum_d \log C_{\text{MN}}(n_d, K),
\]

where

- $n_{kv}$: \# word $v$ in topic $k$
- $n_k$: \# words in topic $k$
- $n_{dk}$: \# words in topic $k$ from document $d$
- $n_d$: \# words in document $d$

$\Longrightarrow$ computable in time $O(d(n + K + V))$
Empirical Evaluation

-Synthetic Data-

DNML is able to identify the true # topics of LDA most robustly

Estimated #topics

DNML converges to the true model as rapidly as LSC, but slower than AIC.

Error Rate

DNML is more robust against noise than AIC.

Estimated number of latent components vs. Sample size

True # topics

Estimated #topics

Sample

Empirical Evaluation

-Synthetic Data-

True # topics

DNML is able to identify the true # topics of LDA most robustly

Estimated number of latent components vs. Sample size

Estimated #topics vs. Sample size

Error Rate

DNML is more robust against noise than AIC.

[Wu, Sugawara, Yamanishi KDD2017]
Empirical Evaluation
-Benchmark Data-

DNML is able to identify the true # topics most exactly

Newsgroup Dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>2 topics</th>
<th>3 topics</th>
<th>4 topics</th>
<th>5 topics</th>
<th>6 topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNML</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>AIC</td>
<td>7</td>
<td>4</td>
<td>9</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Laplace</td>
<td>8</td>
<td>4</td>
<td>8</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>ELBO</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>CV</td>
<td>9</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>HDP</td>
<td>80</td>
<td>98</td>
<td>94</td>
<td>92</td>
<td>98</td>
</tr>
</tbody>
</table>

[Wu, Sugawara, Yamanishi  KDD2017]
Example 3.3.3 (Stochastic Block Models: SBM)

(1) For vertex \( i = 1, \ldots, n \):
   - Generate a latent variable \( z_i \sim \text{Multi}(\pi) \).

(2) For vertex \( i_1 = 1, \ldots, n \):
   - For vertex \( i_2 = 1, \ldots, n \):
     - Generate a variable \( x_{i_1 i_2} \sim \text{Ber}(\eta_{z_{i_1} z_{i_2}}) \).

\[ L_{\text{DNML}}(x, z; K) \]

\[ = \sum_{k_1} \sum_{k_2} \left( n_{k_1 k_2} \log n_{k_1 k_2} - n_{k_1 k_2}^1 \log n_{k_1 k_2}^1 - n_{k_1 k_2}^0 \log n_{k_1 k_2}^0 \right) \]

\[ + \sum_{k_1} \sum_{k_2} \log C_{\text{MN}}(n_{k_1 k_2}, 2) \]

\[ + \sum_k n_k \left( \log n - \log n_k \right) + \log C_{\text{MN}}(n, K), \]

\[ \implies \text{computable in time } O(n + K) \]

where

- \( n_{k_1 k_2} \): \# occurrences in \((k_1, k_2)\) cluster
- \( n_{k_1 k_2}^1 \): \# links in \((k_1, k_2)\) cluster
- \( n_{k_1 k_2}^0 \): \# no-links in \((k_1, k_2)\) cluster
- \( n_k \): \# occurrences in \(k\) cluster
Empirical Evaluation
-Synthetic Data-

Estimated #topics

AIC increases most rapidly but overfit data for large sample size
DNML increases more slowly but converges to true one.

[Wu, Sugawara, Yamanishi KDD2017]

Figure 3: SBM: Estimated number $K$ vs sample size $n$ with $K_{true} = 5$
Example 3.3.4 (Gaussian Mixture Models: GMM)

For observations $i = 1, \ldots, n$:

1. Generate a latent variable $z_i \sim \text{Multi}(\pi)$ with $\pi = (\pi_1, \ldots, \pi_K)$.
2. Generate $x_i \sim \mathcal{N}(\mu_{z_i}, \Sigma_{z_i})$.

\[
L_{\text{DNML}}(x, z; K) = \sum_{k=1}^{K} L_{\text{NML}}(x_k) + \sum_{k=1}^{K} n_k (\log n - \log n_k) + \log C_{\text{MN}}(n, K),
\]

\[
L_{\text{NML}}(x_k) \leq \frac{m n_k}{2} \log(2\pi) + \frac{n_k}{2} \log |\hat{\Sigma}_k| + \frac{m n_k}{2}
\]

\[
+ \frac{m n_k}{2} \log \frac{n_k}{2e} - \frac{m(m-1)}{4} \log \pi - \sum_{j=1}^{m} \log \Gamma \left( \frac{n_k - j}{2} \right)
\]

\[-m \log \frac{m}{2} - \log \Gamma \left( \frac{m}{2} + 1 \right)
\]

\[+ \frac{m}{2} \log ||\hat{\mu}_k||_2^2 - \frac{m}{2} \sum_{j=1}^{m} \log \lambda^{(j)}(\hat{\Sigma}_k)
\]

\[+(m + 1) \log \frac{m}{2} + \log \log \frac{R_2}{R_1} + m \log \log \frac{\lambda_2}{\lambda_1},\]
Empirical Evaluation
-Synthetic Data-

DNML performs best and can identify small components exactly.

\[
\text{Benefit}(\hat{K}, K_{true}) = \max \left\{ 0, 1 - \frac{|\hat{K} - K_{true}|}{2} \right\}
\]

[Yamanishi et al. DAMI 2019]

\[m=5 \quad K^*=10\]

\[m=5 \quad K^*=5\]

With small components
Minimax Optimality of DNML

Theorem 3.3.3 (Estimation Optimality of DNML) [Yamanishi, Wu, Sugawara, Okada DAMI 2019]

\( \hat{K}(\cdot): \) DNML model estimator
\[ \hat{K}(x, z) = \arg\max_{K} \{ p_{\text{NML}}(x|z; K)p_{\text{NML}}(z; K) \} \]

DNML dist. associated with \( \hat{K} \)
\[ \hat{p}(x, z) = \frac{p_{\text{NML}}(x|z; \hat{K}(x, z))p_{\text{NML}}(z; \hat{K}(x, z))}{\hat{C}_{X,Z}^n} \]

Then DNML dist. \( \hat{p}_{\text{DNML}} \) achieves the minimum of
\[ \min_{\bar{p}} \max_{\theta, K} D(p_{\theta,K}||\bar{p}) \quad \text{KL-Divergence} \]
Wide Applicability of DNML

DNML is universally applicable to model selection for a wide class of hierarchical latent variable models.

DNML is a conclusive solution to latent variable model selection.

Network, relations

Stochastic block model

Censored block model

Mixed membership SBM

Finite mixture models (Naïve Bayes, GMM)

Independent mixtures

Mixed membership model

Grade of membership model

LDA

Author-topic model

Correspondence LDA

Text data

DNML is a conclusive solution to latent variable model selection.
Classes of Hierarchical Latent Variable Models

- **a)** Finite mixture models
- **b)** LDA
- **c)** SBM
- **d)** MMSBM
3.4. High-dimensional Sparse Model Selection
3.4.1. High-dimensional Penalty Selection

**Goal** Identify “essential” parameter subset in high dim. models
- To achieve better generalization
- To realize knowledge discovery in lower dimensional space

**Method** Minimize regularized log loss:

\[
x = x_1, \ldots, x_n: \text{data sequence}, \ \Theta: \text{parameter space}
\]

\[
\hat{\theta}(x, \lambda) = \arg\min_{\theta \in \Theta} \left\{ - \log p(x; \theta) + g(\theta, \lambda) \right\}
\]

where

\[
g(\theta, \lambda) = \frac{1}{2} \sum_{j=1}^{d} \lambda_j \theta_j^2
\]

- Small \(\lambda_j\) for essential parameters
- Large \(\lambda_j\) for unnecessary parameters
Luckiness NML Code length (LNML)

\[ LR_n(\mathcal{P}) \overset{\text{def}}{=} \min_{\mathcal{L}: \text{prefix code}} \max_x \{ \mathcal{L}(x) - \min_\theta (-\log p(x; \theta) + g(\theta, \lambda)) \} \]

The minimum of \( LR_n(\mathcal{P}) \) is achieved by Luckiness NML (LNML):

\[ \mathcal{L}_{\text{LNML}}(x) = \min_\theta (-\log p(x; \theta) + g(\theta, \lambda)) + \log Z_n(\lambda) \]

where \( Z_n(\lambda) \overset{\text{def}}{=} \int \max_\theta p(x; \theta) e^{-g(\theta, \lambda)} \, dx \)

Problem:
1) \( Z_n(\lambda) \) is analytically intractable in general
2) How to choose \( \lambda \)?
Upper Bounding on LNML

Theorem 3.3.5  (uLNML: An upper bound on LNML)  
[Miyaguchi, Yamanishi  *Machine Learning* 2018]

LNML code length is uniformly upper-bounded by

\[
L_{\text{LNML}}(\mathbf{x}) \leq \min_{\theta} \left( - \log p(\mathbf{x}; \theta) + g(\theta, \lambda) \right)
\]

\[
\left\{ \frac{1}{2} \log |H(\lambda)| + \log \int e^{-g(\theta, \lambda)} d\theta \right\} + \text{const}
\]

an upper bound on \( Z_n(\lambda) \)

where \(|H(\lambda)|\) is an upper bound on \(|\left. - \frac{\partial^2 \log p(\mathbf{x}; \theta) g(\theta, \lambda)}{\partial \theta \partial \theta} \right|\).
Optimization Algorithm

[Miyaguchi, Yamanishi *Machine Learning* 2018]

• Upper bound on LNML = concave + convex function

\[
\underline{L}_{LNML}(x; \lambda) \overset{\text{def}}{=} \min_\theta (-\log p(x; \theta) + g(\theta, \lambda)) + \frac{1}{2} \log |H(\lambda)| + \log \int e^{-g(\theta, \lambda)} d\theta
\]

\[\text{Concave} \quad \underline{\rightarrow} \quad \text{min w.r.t. } \lambda \quad \underline{\rightarrow} \quad \text{Convex}\]

• Concave-convex procedure (CCCP) [Yuille, Rangarajan NC 02]
  
  ✅ Monotone non-increasing optimization
  ✅ Convergence to a saddle point
Example 3.3.4 (Linear regression model)

\[ y = X\beta + \sigma \epsilon, \quad \epsilon \sim \mathcal{N}_n[0, I_n] \]

The upper bound on LNML is calculated as

\[
\overline{\mathcal{L}}_{\text{LNML}}(\lambda) = \min_{\beta, \sigma^2} \frac{1}{2\sigma^2} \| y - X\beta \|^2 + \frac{n}{2} \ln \sigma^2 + \frac{1}{2\sigma^2} \sum_{j=1}^{m} \lambda_j \beta_j^2 + \frac{1}{2} \ln \frac{\det(X^\top X + \text{diag } \lambda)}{\det \text{diag } \lambda}
\]

\[ \implies \min \text{ w.r.t. } \lambda \]
Experiments: Linear Regression

— UCI Repository —

• Better than grid-search-based methods (esp. if $d \geq n$)
• More robust than RVM (relevant vector machine)

Example 3.3.5  (Graphical Model Estimation)

- \( m \)-variate Gaussian model
- Precision \( \Theta \) represents conditional dependencies

\[
x_i \sim \mathcal{N}_m(\mathbf{0}, \Theta^{-1}) \quad (i = 1, \ldots, n), \quad \Theta \succeq R^{-1}I_m
\]

- The upper bound on LNML

\[
\overline{\mathcal{L}}_{\text{LNML}}(\lambda) = \min_{\Theta} \left\{ \frac{1}{2} \text{Tr}[X^\top X \Theta] - \frac{n}{2} \ln \det \Theta + \right. \\
\left. \frac{1}{2} \sum_{i<j} \lambda_{ij} \Theta_{ij}^2 + \frac{1}{2} \sum_{i<j} \ln \left( 1 + \frac{n R^2}{\lambda_{ij}} \right) \right\}
\]

where \( \lambda_{ij} \) are the eigenvalues of the precision matrix \( \Theta \).
Experiments: Graphical Model
—Synthetic Data—

Generated w/ double-ring model, Can handle $>10^3$ dimension

[Miyaguchi, Yamanishi *Machine Learning* 2018]
**Luckiness NML for L1-Penalty**

[Miyaguchi, Matsushima and Yamanishi SDM2016]

- **Luckiness Minimax Regret**

\[
\mathcal{L}_{\text{LNML}}(x) = \min_{\theta} \left( - \log p(x; \theta) + g(\theta, \lambda) \right) + \log Z_n(\lambda)
\]

where

\[
Z_n(\lambda) \overset{\text{def}}{=} \int \max_{\theta} p(x; \theta) e^{-g(\theta, \lambda)} \, dx \quad g(\theta, \lambda) = \sum_{j=1}^{d} \lambda_j |\theta_j|
\]

**Stochastic Gradient Descent**

\[
\lambda \leftarrow \lambda - \eta \frac{\partial \mathcal{L}_{\text{LNML}}(x; \lambda)}{\partial \lambda} = \lambda - \eta (|\bar{\theta}(x, \lambda)| - E[|\bar{\theta}(Y, \lambda)|])
\]

\[
\approx \lambda - \eta (|\bar{\theta}(x, \lambda)| - |\bar{\theta}(y, \lambda)|)
\]

sampled by \( p(y; \theta) e^{-g(\theta, \lambda)} \)

\[
\bar{\theta} = \arg\max_{\theta} p(x; \lambda) e^{-g(\theta, \lambda)}
\]
3.4.2. High-dimensional Minimax Prediction

Problem:

\( \mathbf{x} = x_1, x_2, \ldots, x_t, \ldots: \) data sequence

\( \mathcal{P} = \{p(X; \theta) : \theta \in \Theta\} : \) probability model

Construct on-line predictor \( \mathcal{A} : \{p(X|x^{t-1})\} \)

to attain the minimax regret:

\[
\min_{\mathcal{A}} \max_{\mathbf{x}} \left\{ \sum_{t=1}^{n} (- \log p(x_t|x^{t-1})) - \min_{\theta} \sum_{t=1}^{n} (- \log p(x_t; \theta)) \right\}
\]
On-line Bayesian Prediction Strategy

In conventional low dimensionality setting, the minimax regret is attained by e.g. the Bayesian prediction strategy

\[ p_{\text{Bayes}}(X|x^{t-1}) = \int p(X; \theta)p(\theta|x^{t-1})d\theta \]

\[ p(\theta|x^{t-1}) = \frac{\pi_{\text{Jeff}}(\theta)p(x^{t-1}; \theta)}{\int \pi_{\text{Jeff}}(\theta)p(x^{t-1}; \theta)d\theta} \]

\[ \pi_{\text{Jeff}}(\theta) = \frac{|I(\theta)|^{1/2}}{\int |I(\theta)|^{1/2}d\theta} : \text{Jeffreys' prior} \]

Then cumulative log loss amounts

\[ \sum_{t=1}^{n} (-\log p_{\text{Bayes}}(x_t|x^{t-1})) = -\log \int \pi_{\text{1eff}}(\theta)p(x; \theta)d\theta \]

\[ \approx -\log \max_{\theta} p(x; \theta) + \frac{k}{2} \log \frac{n}{2\pi} + \log \int |I(\theta)|^{1/2}d\theta \]

Problem: No counterpart in high-dim setting \((d \geq n)\)
High-dimensional Asymptotics

Assumption

- Twice differential of loss is uniformly upper bound by $L$
- $\ell_1$-radius of $P$ is at most $R < +\infty$

Worst-case regret for on-line prediction algorithm $A$:

$$R_n(A) \overset{\text{def}}{=} \max_x \left\{ \sum_{t=1}^n (- \log p_A(x_t | x_t^{t-1})) - \min_{\theta} \sum_{t=1}^n (- \log p(x_t; \theta)) \right\}$$

Theorem 3.3.6 (Asymptotics on worst regret)

[Miyaguchi, Yamanishi AISTATS 2019]

Under the high-dim limit $\omega(\sqrt{n}) = d = e^{o(n)}$,

For the Bayesian prediction alg. with ST prior,

$$R_n(A_{ST}) \leq R \sqrt{2Ln \log \left( \frac{d}{\sqrt{n}} \right)} (1 + o(1)).$$

For some L-smooth model, for any prediction alg. $A$,

$$R_n(A) \geq \frac{R}{2} \sqrt{2Ln \log \left( \frac{d}{\sqrt{n}} \right)} (1 + o(1)).$$

Optimal within a factor of 2
Minimax Predictor in High-dimensional Setting

Bayesian prediction with Spike-and-Tails (ST) prior

\[
\sum_{t=1}^{n} (- \log p_{ST}(x_t|x^{t-1})) = - \log \int \pi_{ST}(\theta)p(x; \theta) d\theta
\]

• One-dim illustration of ST prior
  Gap is wide when \( d \gg \sqrt{n} \)

\[\sqrt{2 \ln \frac{d}{\sqrt{n}}} \]

\[\text{Spike} \]

\[\text{Exponential tail } \propto e^{-\lambda |\theta|} \]
Summary

• Stochastic complexity (SC), namely the NML codelength, is the well-defined key information quantity of data relative to the model. MDL is to minimize SC.

• Techniques for efficient computation of SC are established: 1) asymptotic formula, 2) g-functions, 3) Fourier transform, 4) combinatorial method.

• When applying MDL to latent variable models, use complete variable models, and apply Latent Stochastic Complexity (LSC) or Decomposed NML (DNML) to realize efficient and optimal estimation.

• When applying MDL to high-dimensional sparse models, apply LNML to realize optimal penalty selection.
References

3.1. Stochastic Complexity and NML Codelength

References

3.2. G-function and Fourier Transformation Method


3.3.1. Latent Stochastic Complexity

References

3.3.1. Latent Stochastic Complexity (Cont.)

References

3.3.1. Latent Stochastic Complexity (Cont.)


3.3.2. Decomposed Normalized Maximum Likelihood Codelength


3.3.3. High-dimensional Sparse Model Selection


C.f. Bayesian prediction related to MDL