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?

■Case1: Data distribution is known

Data sequence : $\boldsymbol{x} = x_1, \ldots, x_n$

Probability mass (density) function : $p(\boldsymbol{x})$

Shannon information $I(\boldsymbol{x}) = -\log p(\boldsymbol{x})$

Characterization of Shannon Information

Theorem 3.1.1 （Shannon's Source Coding Theorem）

$\forall \mathcal{L}$: prefix code-length function,

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

Entropy

Note:

Prefix code-length function $\mathcal{L} \overset{\text{def}}{\Longleftrightarrow} \mathcal{L}(\boldsymbol{x}) \geq 0$ and $\sum_{\boldsymbol{x}} 2^{-\mathcal{L}(\boldsymbol{x})} \leq 1$

Kraft's Inequality

Shannon information gives optimal codelength when the true distribution is known in advance.

■Case2: Data distribution is unknown.

Data seq. : $\boldsymbol{x} = x_1, \ldots, x_n$

Probabilistic model class. : $\mathcal{P}_M = \{p(\boldsymbol{x};\theta) : \theta \in \Theta_M\}$ M: model

Normalized Maximum Likelihood (NML) Distribution

$$p_{\mathrm{NML}}(\boldsymbol{x}; M) = \frac{\max_\theta p(\boldsymbol{x};\theta)}{\sum_{\boldsymbol{y}} \max_\theta p(\boldsymbol{y};\theta)}$$

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

Stochastic Complexity of ***x*** relative to P_M =NML codelength

$$\begin{aligned} I(x^n; M) &= -\log p_{\mathrm{NML}}(\boldsymbol{x}; M) \\ &= -\log \max_\theta p(\boldsymbol{x};\theta) + \log \mathcal{C}_n(M) \end{aligned}$$

Parametric Complexity of P_M

$$\mathcal{C}_n(M) = \sum_{\boldsymbol{x}} \max_\theta p(\boldsymbol{x};\theta)$$

Characterization of NML Codelength

Theorem 3.1.2 （Minimax optimality of NML codelength)
[Shtarkov Probl. Inf. Trans. 87]

NML codelength achieves the minimum of the risk

$$I(\boldsymbol{x};M) = \arg \min_{\substack{\mathcal{L} \\ \text{prefix code}}} \max_{\boldsymbol{x}} \left\{ \mathcal{L}(\boldsymbol{x}) - \left(-\log \max_{\theta} p(\boldsymbol{x};\theta) \right) \right\}$$

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:

$\sqrt{n}(\hat{\theta}(\boldsymbol{x}) - \theta) \sim \mathcal{N}(0, I^{-1}(\theta))$,

($\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, p(\boldsymbol{x};\theta)$:maximum likelihood estimator), it holds:

$$\begin{aligned} \log C_n(M) &= \log \sum_{\boldsymbol{x}} \max_{\theta} p(\boldsymbol{x};\theta) \\ &= \frac{k}{2}\log\frac{n}{2\pi} + \log\int |I(\theta)|^{1/2} d\theta \end{aligned}$$

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

k: $\sharp$parameters, n: data length,

$I(\theta) = \lim_{n\to\infty} E_p[-\frac{1}{n}\frac{\partial^2 \log p(\boldsymbol{x};\theta)}{\partial\theta\partial\theta^T}]$ (Fisher informtion matrix)

Example 3.1.1 (Multinomial Distribution)

$\mathcal{X} = \{0, 1, \ldots, K\}$

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

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

$\boldsymbol{x} = x_1, \ldots, x_n$: data sequence

n_i: $\sharp$ occurrences of $X = i$

$\hat{\theta} = (\frac{n_0}{n}, \ldots, \frac{n_K}{n})$: m.l.e. of θ

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

Stochastic Complexity

$$\begin{aligned} I(x^n; K) &= -\log p(\boldsymbol{x}; \hat{\theta}) + \frac{K}{2} \log \frac{n}{2\pi} + \log \int \sqrt{|I(\theta)|} d\theta \\ &= nH\left(\frac{n_0}{n}, \ldots, \frac{n_K}{n}\right) + \frac{K}{2} \log \frac{n}{2\pi} + \log \frac{\pi^{\frac{K+1}{2}}}{\Gamma\left(\frac{K+1}{2}\right)}, \end{aligned}$$

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

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

m.l.e.

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

$$s,r:\ \text{smallest integers such that}\ \hat{\mu}\le 2^s,\ \ \hat{\tau}\ge 2^{-2r}$$

Restricted Parameter Space

$$\tilde{\Theta}=\{(\mu,\tau):\ \mu\le 2^s,\ \tau\ge 2^{-2r}\}$$

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

Stochastic complexity

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

$\log^* x=\log c+\log x+\log\log x+\ldots(c=2.865)$: 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 Θ_0 such that $vol(\Theta_0) \to 0\ (n \to \infty)$,
$\forall \epsilon > 0$, $\forall \mathcal{L}$: prefix codelength function, it holds:

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

Stochastic complexity gives optimal codelength when the true distribution is unknown.

Sequential NML Codelength

Sequential NML (SNML)

=Cumulative codelength for sequential NML coding

$\boldsymbol{x} = x_1, \ldots, x_n$: data sequence

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

Sequential NML

Theorem 3.1.5（Property of SNML）

For model classes under some regular conditions

$$I(\boldsymbol{x}; M) = \tilde{I}(\boldsymbol{x}; M) + o(\log n)$$

E.g. Regression model [Rissanen IEEE IT2000]

[Roos, Myllymaki, Rissanen MVA 2009]

Example 3.1.3.(Auto-regression model)

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

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

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

where

$$V_t\;\stackrel{\text{def}}{=}\;\left(\sum_{j=1}^{t}\bar{x}_j\bar{x}_j^T\right)^{-1}=V_{t-1}-\frac{V_{t-1}\bar{x}_t\bar{x}_t^TV_{t-1}}{1+c_t},\qquad c_t\stackrel{\text{def}}{=}\bar{x}_t^TV_{t-1}\bar{x}_t.$$

When σ is fixed, : $\hat{x}_t=\hat{a}_t^T\bar{x}_t$ then SNNL dist. becomes

$$\boxed{\begin{aligned}p_{\text{SNML}}(x_t|x^{t-1}) &= \frac{p(x_t|x^{t-1};\hat{a}_t)}{K(x^{t-1})}\\ &= \frac{1}{\sqrt{2\pi(1+c_t)^2\sigma^2}}\exp\left(-\frac{(y_t-\hat{a}_{t-1}^T\bar{x}_t)^2}{2(1+c_t)^2\sigma^2}\right).\end{aligned}}$$

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 $\mathcal{M}$:

$$\begin{aligned} I(\boldsymbol{x}; \mathcal{M}) &= -\log \frac{\max_{M \in \mathcal{M}} \{\hat{p}(\boldsymbol{x}; M)\}}{C} \\ &= \boxed{\min_{M \in \mathcal{M}} \left\{ -\max_\theta \log p(\boldsymbol{x}; \theta, M) + \log C_n(M) \right\}} + \log C \end{aligned}$$

MDL(Minimum Description Length）criterion

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

MDL Criterion（2/2）

$\mathcal{M}=\{M\}$: model class [Rissanen IEEE IT 1996]

Under the condition of central limit theorem:

$\sqrt{n}(\hat{\theta}(\boldsymbol{x})-\theta)\sim\mathcal{N}(0,1/I(\theta))$,

$\boldsymbol{x}=x_1,\ldots,x_n$: given

$$
\begin{aligned}
&I(\boldsymbol{x};M)+\ell(M)\\
&=-\log\max_{\theta} p(\boldsymbol{x};\theta)+\frac{k}{2}\log\frac{n}{2\pi}+\log\int\sqrt{|I(\theta)|}d\theta+\ell(M)\\
&\Longrightarrow \min\ \text{w.r.t. } M.
\end{aligned}
$$

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

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

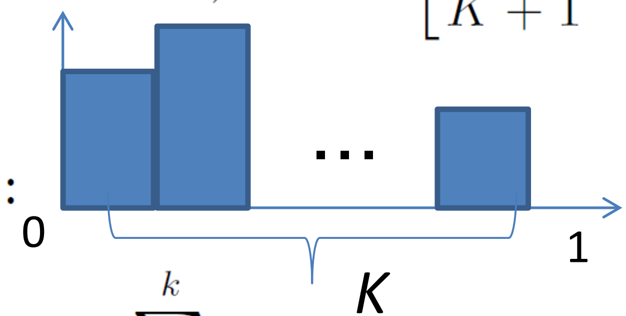
Example 3.1.4. （Histogram Density）

$\mathcal{X} = [0,1]$....K disjoint cells

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

$\theta_i = Prob(X \in C_i)$

Class of histogram densities with equal length cells :



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

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

$$I(x^n; K) = \min_{\theta}\{-\log p(x^n;\theta)\} + \frac{K}{2}\log\frac{n}{2\pi} + \log\int\sqrt{|I(\theta)|}d\theta + \mathcal{L}(K)$$

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

$\Longrightarrow$ min w.r.t. K (optimal K)

Optimality of MDL Estimation

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

$\bar{\theta},\ \bar{M}$: any given estimators

Define a general normlized distribution associated with $\bar{\theta}$ and $\bar{M}$ by

$$\bar{p}(\boldsymbol{x}) = \frac{p(\boldsymbol{x};\bar{\theta},\bar{M})}{\sum_{\boldsymbol{y}} p(\boldsymbol{y};\bar{\theta},\bar{M})},$$

then it holds:

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

$D(f||g) = E_f[\log(f(x)/g(x))]$ (Kullback-Leiber divergence)

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

$$\hat{C}_n(M) \stackrel{\text{def}}{=} \sum_{\boldsymbol{x}} \max_{\theta} p(\boldsymbol{x};\theta,M)$$

$$\hat{C}_n \stackrel{\text{def}}{=} \sum_{\boldsymbol{x}} \max_{M}\{\max_{\theta}(p(\boldsymbol{x};\theta,M)/C_n(M)\}$$

Why is the MDL?

- Optimal solution to Shtarkov minimax risk
- Attaining Rissanen's lower bound
- Consistency [Rissanen Auto. Control 1978]
  [Barron 1989]


Andrew Barron

- 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

$$\mathcal{C}_n(M) \;=\; \sum_{\boldsymbol{x}} \max_{\theta} p(\boldsymbol{x};\theta)$$

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 [Rissanen 2007, 2012]

$$p(\boldsymbol{x};\theta) = p(\boldsymbol{x}|\hat{\theta}(\boldsymbol{x}))g(\hat{\theta}(\boldsymbol{x});\theta)$$

where $\hat{\theta}(\boldsymbol{x}) \stackrel{\text{def}}{=} \operatorname*{argmax}_{\theta} p(\boldsymbol{x};\theta)$: m.l.e.

$$\boxed{g(\bar{\theta};\theta) \stackrel{\text{def}}{=} \sum_{\boldsymbol{x}:\hat{\theta}(\boldsymbol{x})=\bar{\theta}} p(\boldsymbol{x};\theta)}$$

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

$$\int g(\bar{\theta};\theta)d\bar{\theta} = \int d\bar{\theta}\left(\sum_{\boldsymbol{y}:\hat{\theta}(\boldsymbol{y})=\bar{\theta}} p(\boldsymbol{y};\theta)\right) = 1.$$

Calculation of Parametric Complexity via g-function

$$
\begin{aligned}
C_n(M) &= \sum_{\boldsymbol{x}} p(\boldsymbol{x}; \hat{\theta}(\boldsymbol{x})) \\
&= \int d\hat{\theta} \sum_{\boldsymbol{y}:\hat{\theta}(\boldsymbol{y})=\hat{\theta}} p(\boldsymbol{y}; \hat{\theta}) \\
&= \int g(\hat{\theta}; \hat{\theta}) d\hat{\theta}.
\end{aligned}
$$

Variable transformation

Example 3.2.1 （g-function for exponential distributions）

$$\mathcal{P}_{\mathrm{Exp}} \stackrel{\mathrm{def}}{=} \{p(X;\theta) = \theta \exp(-\theta X) : \ \theta \in \mathbf{R}\}$$

$\boldsymbol{x} = x_1, \ldots, x_n$: data sequence $\quad \hat{\theta}(x^n) = \dfrac{n}{\sum_{i=1}^n x_i}$: m.l.e.

$$\begin{aligned} p(\boldsymbol{x};\theta) &= \exp\left\{-\theta\sum_{i=1}^n x_i + n\log\theta\right\} \\ &= \theta^n \exp\left\{-\frac{n\theta}{\hat{\theta}}\right\} \\ &= f(\boldsymbol{x}|\hat{\theta}(\boldsymbol{x}))g(\hat{\theta}(\boldsymbol{x});\theta). \end{aligned}$$

$\Rightarrow$ g.. Gamma dist. of $n/\hat{\theta}(x^n)$ with shape n and scale $1/\theta$.

$$\boxed{g(\hat{\theta}(\boldsymbol{x});\theta) = \frac{\theta^n n^n}{\Gamma(n)\hat{\theta}(\boldsymbol{x})^{n+1}} \exp\left\{-\theta \cdot \frac{n}{\hat{\theta}(\boldsymbol{x})}\right\}}$$

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

$$Y(\theta_{\min},\theta_{\max}) \stackrel{\text{def}}{=} \{\hat{\theta} : \theta_{\min} \leq \hat{\theta} \leq \theta_{\max}\}$$

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

Then

$$\begin{aligned} C_n &= \int_{Y(\theta_{\min},\theta_{\max})} g(\hat{\theta};\hat{\theta})d\hat{\theta} \\ &= \frac{n^n}{e^n(n-1)!} \int_{\theta_{\min}}^{\theta_{\max}} \frac{1}{\hat{\theta}} d\hat{\theta} \\ &= \frac{n^n}{e^n(n-1)!} \log \frac{\theta_{\max}}{\theta_{\min}}. \end{aligned}$$

$\Rightarrow$ 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]

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

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

$$h(\theta) \stackrel{\text{def}}{=} \frac{1}{(2\pi)^k} \int d\xi \sum_{\boldsymbol{x}} p(\boldsymbol{x};\theta) \exp(i\xi^\top(\hat{\theta}(\boldsymbol{x}) - \theta)),$$

where $\hat{\theta}(\boldsymbol{x}) = \underset{\theta}{\operatorname{argmax}}\, p(\boldsymbol{x};\theta)$.

Fourier Transformation Method(2/2)

Proof Sketch.

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

$$\begin{aligned}\tilde{p}(\boldsymbol{x};\xi) &= \frac{1}{(2\pi)^{k/2}}\int d\theta \exp(-i\xi^\top\theta)p(\boldsymbol{x};\theta),\\ p(\boldsymbol{x};\theta) &= \frac{1}{(2\pi)^{k/2}}\int d\xi \exp(i\xi^\top\theta)\tilde{p}(\boldsymbol{x};\xi).\end{aligned}$$

$$\begin{aligned}\sum_{\boldsymbol{x}}\max_{\theta} p(\boldsymbol{x};\theta) &= \sum_{\boldsymbol{x}}\frac{1}{(2\pi)^{k/2}}\int d\xi \exp(i\xi^\top\hat{\theta}(\boldsymbol{x}))\tilde{p}(\boldsymbol{x};\xi)\\ &= \frac{1}{(2\pi)^{k/2}}\int d\xi \sum_{\boldsymbol{x}}\int d\theta \exp(i\xi^\top(\hat{\theta}(\boldsymbol{x})-\theta))p(\boldsymbol{x};\theta)\\ &= \frac{1}{(2\pi)^{k/2}}\int d\theta \int d\xi \sum_{\boldsymbol{x}} p(\boldsymbol{x};\theta)\exp(i\xi^\top(\hat{\theta}(\boldsymbol{x})-\theta))\\ &= \int d\theta h(\theta).\end{aligned}$$

Exponential Family

$$p(x;\eta) = m(x)\exp(\eta^\top t(x))/Z(\eta).$$

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

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

$$\boldsymbol{x} = x_1, \ldots, x_n$$

$$\int d\boldsymbol{x} \max_{\tau} p(\boldsymbol{x};\eta(\tau)) = \frac{1}{(2\pi)^k} \int d\tau \int d\xi \exp(-\xi^\top \tau) \left(\frac{Z(\eta(\tau) + i\xi/n)}{Z(\eta(\tau))} \right)^n$$

Parametric Complexity of Exponential Family Computed with Fourier Method

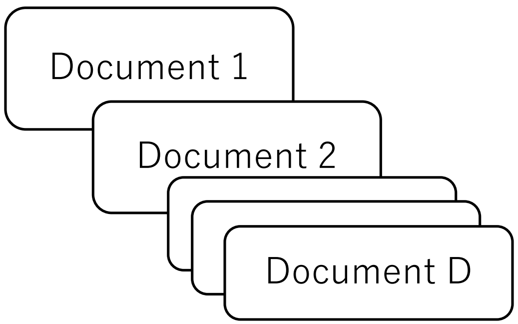
[Suzuki and Yamanishi ISIT 2018]

Distribution	Density function $f\left(\boldsymbol{x}^N;\boldsymbol{\theta}\right)$	Sufficient statistics $u_k\left(\boldsymbol{x}\right)$	Canonical parameter $\boldsymbol{\eta}$ Expectation parameter $\boldsymbol{\tau}=\mathbb{E}\left[u_k\left(\boldsymbol{x}\right)\right]$	Partition function $Z\left(\boldsymbol{\eta}\right)$	Parametric complexity
Normal dist. with known variance v	$f\left(\boldsymbol{x}^N;\mu\right)$ $=\frac{1}{\sqrt{2\pi v}}\exp\left(-\frac{(x-\mu)^2}{2v}\right)$	x	$\eta=\frac{\mu}{v}\in(-\infty,+\infty)$ $\mu=v\eta\in(-\infty,+\infty)$	$v^{\frac{1}{2}}\exp\left(\frac{1}{2}v\eta^2\right)$	$\int_{-\infty}^{+\infty}\mathrm{d}\mu\frac{w(\mu)}{\sqrt{2\pi\frac{v}{N}}}$
Normal dist. with known mean μ	$f\left(\boldsymbol{x}^N;v\right)$ $=\frac{1}{\sqrt{2v}}\exp\left(-\frac{(x-\mu)^2}{2v}\right)$	$(x-\mu)^2$	$\eta=-\frac{1}{2v}\in(-\infty,0)$ $v=-\frac{1}{2\eta}\in(0,\infty)$	$\frac{1}{\sqrt{-\eta}}$	$\frac{\left(\frac{1}{2}N\right)^{\frac{1}{2}N}\exp\left(-\frac{1}{2}N\right)}{\Gamma\left(\frac{1}{2}N\right)}$ $\times\int_0^{+\infty}\mathrm{d}v\frac{w(v)}{v}$
Laplace dist. with known mean μ	$f\left(\boldsymbol{x}^N;b\right)$ $=\frac{1}{2b}\exp\left(-\frac{\lvert x-\mu\rvert}{b}\right)$	$\lvert x-\mu\rvert$	$\eta=-\frac{1}{b}\in(-\infty,0)$ $b=-\frac{1}{\eta}\in(0,\infty)$	$\frac{2}{-\eta}$	$\frac{N^N\exp(-N)}{\Gamma(N)}$ $\times\int_0^{+\infty}\mathrm{d}b\frac{w(b)}{b}$
Gamma dist. with known shape k [1] [2]	$f\left(\boldsymbol{x}^N;\mu\right)$ $=\frac{k^kx^{k-1}}{\Gamma(k)\mu^k}\exp\left(-\frac{kx}{\mu}\right)$	x	$\eta=-\frac{k}{\mu}\in(-\infty,0)$ $\mu=-\frac{k}{\eta}\in(0,\infty)$	$\frac{1}{(-\eta)^k}$	$\frac{(kN)^{kN}\exp(-kN)}{\Gamma(kN)}$ $\times\int_0^{+\infty}\mathrm{d}\mu\frac{w(\mu)}{\mu}$
Weibull dist. with known shape k	$f\left(\boldsymbol{x}^N;L\right)$ $=kL^{-\frac{k+1}{k}}x^k\exp\left(-\frac{x^k}{L}\right)$	x^k	$\eta=-\frac{1}{L}\in(-\infty,0)$ $L=-\frac{1}{\eta}\in(0,\infty)$	$\frac{1}{-\eta}$	$\frac{N^N\exp(-N)}{\Gamma(N)}$ $\times\int_0^{+\infty}\mathrm{d}L\frac{w(L)}{L}$
Gamma dist. with known scale θ [3]	$f\left(\boldsymbol{x}^N;\eta\right)$ $=\frac{x^\eta}{\Gamma(\eta+1)\theta^{\eta+1}}\exp\left(-\frac{x}{\theta}\right)$	$\log x$	$\eta=\psi^{-1}\left(\lambda-\log\theta\right)-1\in(-1,+\infty)$ $\lambda=-\psi\left(\eta+1\right)+\log\theta\in(-\infty,+\infty)$	$\Gamma\left(\eta+1\right)\theta^{\eta+}$	See (22)

3.3. Latent Variable Model Selection

Latent Variable Model Selection

Motivation 1. Topic Model

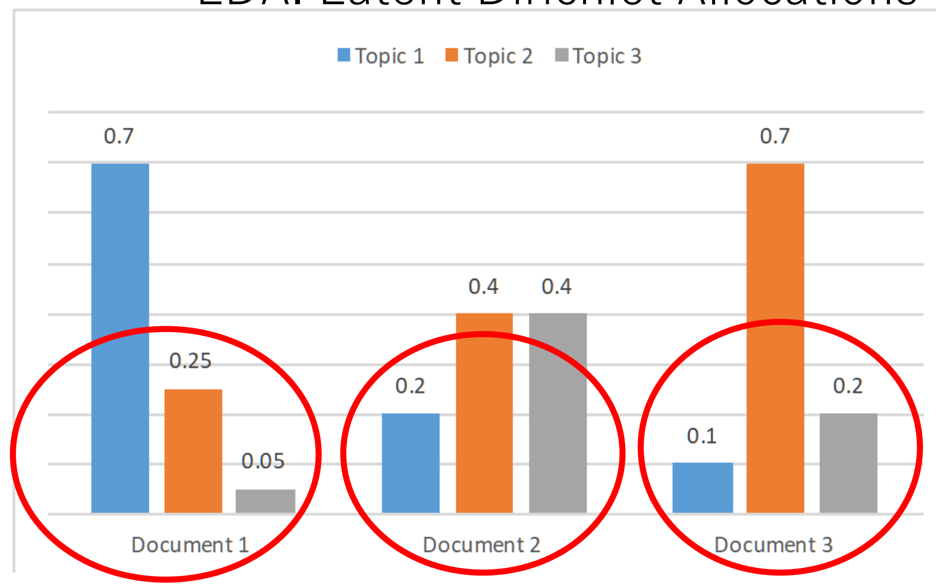


Topic 1
Machine: 0.05
Computer: 0.03
Algorithm: 0.02
…

Topic 2
Matrix: 0.04
Math: 0.03
equation: 0.01
…

Topic K
Bio: 0.05
Gene: 0.04
Array: 0.01
…

LDA: Latent Dirichlet Allocations

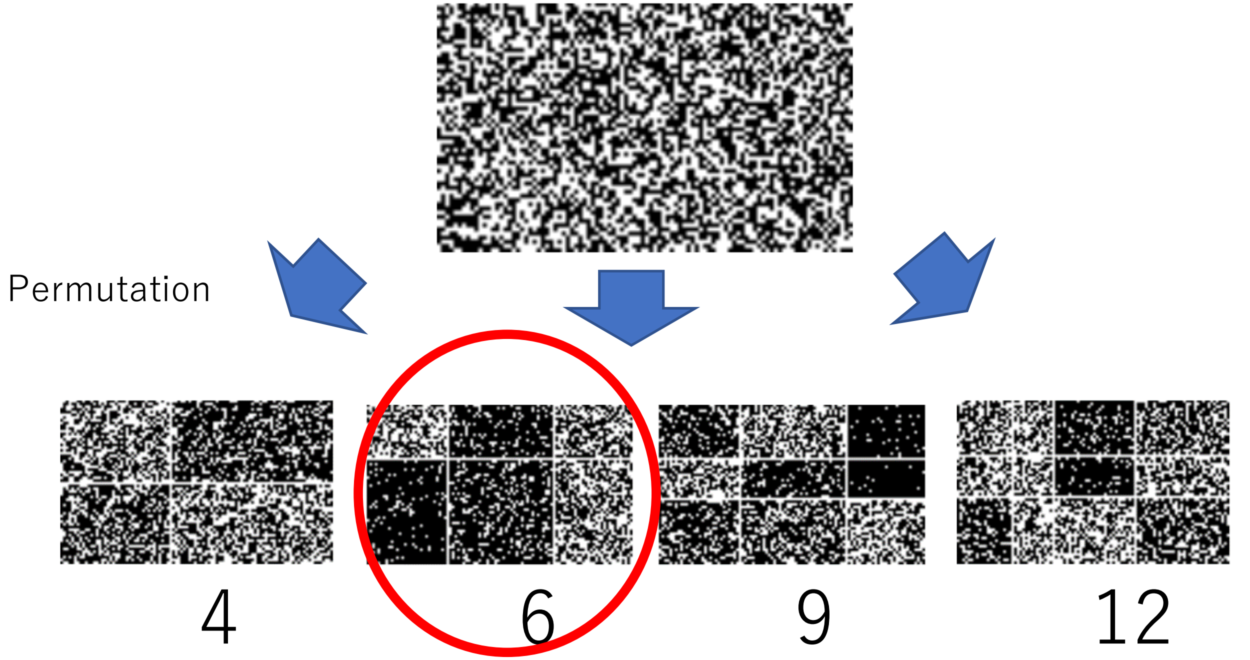


#topics =3

How many topics lie in a document?

Latent Variable Model Selection

Motivation 2： Relational Model



#blocks = 6

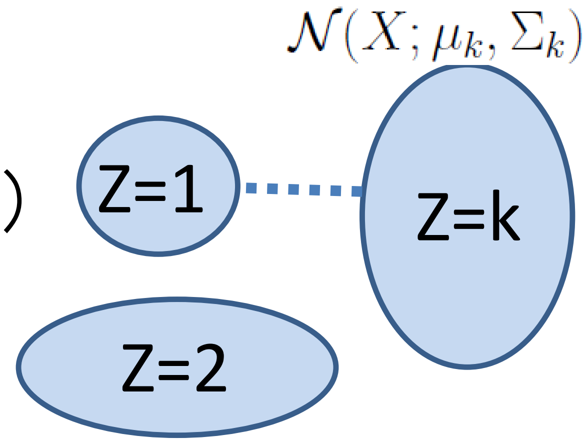
How many blocks lie in a relational data?

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: $\sharp$ components

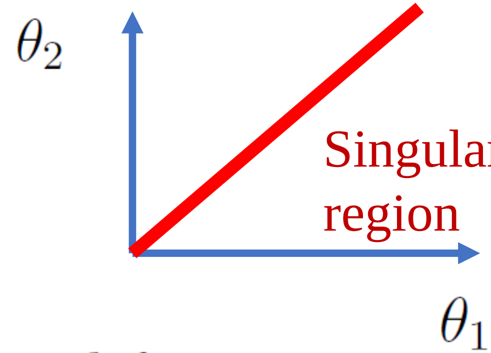
$\mathcal{N}(\mu_k, \Sigma_k)$: normal dist. with mean μ_k and variance-covariance matrix Σ_k

$$\pi_k = P(Z = k),\ \theta = (\pi_k, \mu_k, \Sigma_k)\ (k = 1, \ldots, K)$$

Nonidentifiability of Latent Variable Model

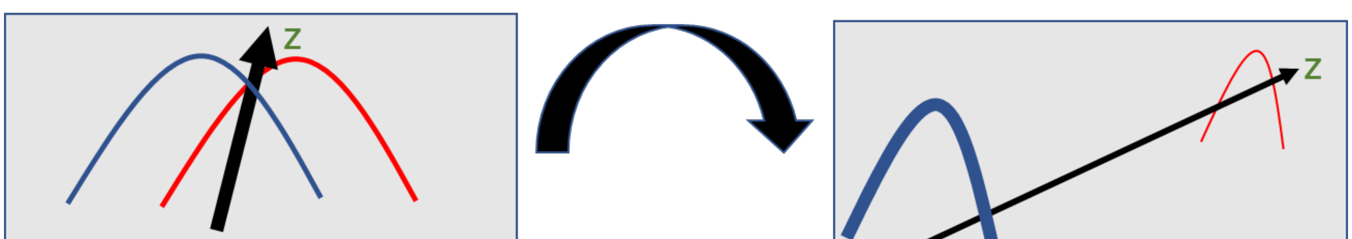
Marginalized model $p(X;\theta) = \sum_Z p(X,Z;\theta) = \sum_Z p(Z)p(X|Z;\theta_Z)$

Finite mixture model $p(X;\pi,\theta_1,\theta_2) = \pi p(X;\theta_1) + (1-\pi)p(X;\theta_2)$



Non-identifiable $\theta_1 = \theta_2 \Longrightarrow \text{Prob. dist. is identical for any } \pi.$

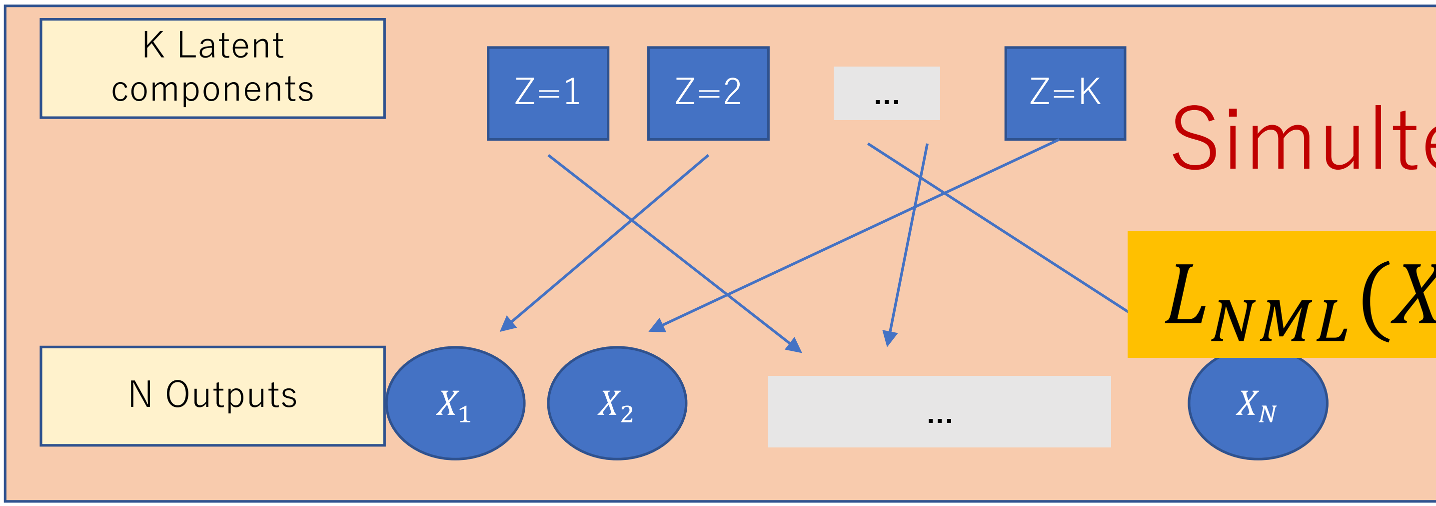
Complete variable model $p(X,Z;\theta)$ Estimate Z from X



Identifiable

3.3.1. Latent Stochastic Complexity

Normalized maximum likelihood (NML) Codelength for complete variable model



Simulteneously

$L_{NML}(X, Z; K)$

Code-length

$\boldsymbol{x} = x_1, \ldots, x_n$: data sequece $\quad \boldsymbol{z} = z_1, \ldots, z_n$: latent variable sequence

Latent Stochastic Complexity(LSC)

$$-\log \max_{\theta} p(\boldsymbol{x}, \boldsymbol{z}; \theta, M) + \log \sum_{\boldsymbol{y}} \sum_{\boldsymbol{w}} \max_{\theta} p(\boldsymbol{y}, \boldsymbol{w}; \theta, M)$$

$$\Longrightarrow \min \quad \text{w.r.t. } M$$

Latent Parametric Complexity

Finite Mixture Model

Finite Mixture Model

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

$$p(X;\theta,K)=\sum_{Z=1}^{K}p(X|Z;\theta_1)p(Z;\theta_2) \qquad K\text{: }\sharp\text{ components}$$

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

Latent Parametric Complexity

$$\begin{aligned}C_n(K) &= \sum_{\boldsymbol{z}}\left(\int \max_{\theta} p(\boldsymbol{x},\boldsymbol{z};\theta,K)d\boldsymbol{x}\right)\\ &= \sum_{\boldsymbol{z}}\max_{\theta_2} p(\boldsymbol{z};\theta_2)\int p(\boldsymbol{x}|\boldsymbol{z};\theta_1)d\boldsymbol{x}\\ &= \sum_{\substack{n_1+\cdots+n_K=n\\ n_i\geq 0\ i=1,\ldots,K}} \frac{n!}{n_1!\cdots n_K!}\left(\frac{n_1}{n}\right)^{n_1}\cdots\left(\frac{n_K}{n}\right)^{n_K}\prod_{k=1}^{K}\overline{C}_{n_k}\end{aligned}$$

$$\text{where}\quad \overline{C}_n = \int \max_{\theta_1} p(\boldsymbol{x}|\boldsymbol{z};\theta_1)d\boldsymbol{x}$$

$\Longrightarrow O(K^n)$ computation time

Parametric Complexity of FMM

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

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

K: $\sharp$ components, n: data length

$\Longrightarrow$ computable in $O(n^2K)$ time

Gaussian Mixture Models

[Hirai Yamanishi IEEE IT 2013, 2019]

Example 3.3.1（Parametric complexity of multivariate GMM）

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

where

$$\begin{aligned} \overline{\mathcal{C}}_n &= \int_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{x}; \hat{\mu}(\boldsymbol{x}), \hat{\Sigma}(\boldsymbol{x}) d\boldsymbol{x} \\ &\leq \frac{2^{d+1} R^{\frac{d}{2}} \prod_{j=1}^{d} \left(\lambda_{\min}^{(j)}\right)^{-\frac{d}{2}}}{d^{d+1} \Gamma\left(\frac{d}{2}\right) \Gamma_d\left(\frac{n-1}{2}\right)} \left(\frac{n}{2e}\right)^{\frac{dn}{2}} \end{aligned}$$

d: data dimension $\qquad \mathcal{X} \stackrel{\text{def}}{=} \{\boldsymbol{x} : ||\hat{\mu}(\boldsymbol{x})|| \leq R,\ \lambda_{\min}^{(j)} \leq \hat{\lambda}_j(\boldsymbol{x})\}$

$R > 0,\ \lambda_{\min}^{(j)}\ (j = 1, \ldots, d)$: given, λ_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 IEEEBigData 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

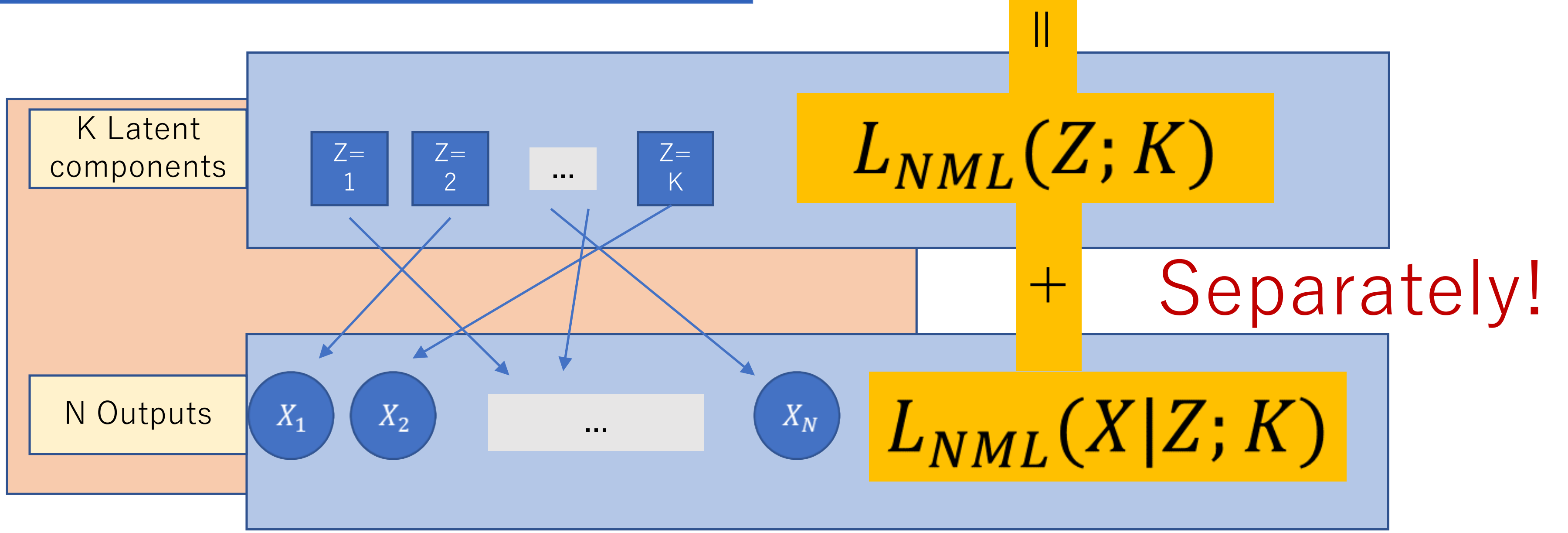
[Fu, Matsushima, Yamanishi Entropy2019]

3.3.2. Decomposed Normalized Maximum Likelihood (DNML) Codelength

[Wu, Sugawara, Yamanishi KDD2017]
[Yamanishi, Wu, Sugawara, Okada DAMI2019]

Computable efficiently and non-asymptotically

$L_{DNML}(X, Z; K)$ Code-length



How to Compute DNML

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

K: $\sharp$ latent variables

$\boldsymbol{x} = x_1, \ldots, x_n$: observed data sequence

$\boldsymbol{z} = z_1, \ldots, z_n$: latent variable sequence

DNML criterion

$$\mathcal{L}_{\mathrm{DNML}}(\boldsymbol{x}, \boldsymbol{z}; K) \overset{\mathrm{def}}{=} \mathcal{L}_{\mathrm{NML}}(\boldsymbol{x}|\boldsymbol{z}; K) + \mathcal{L}_{\mathrm{NML}}(\boldsymbol{z}; K)$$

$$\Longrightarrow \min \text{w.r.t. } K$$

$$\mathcal{L}_{\mathrm{NML}}(\boldsymbol{x}|\boldsymbol{z}; K) = -\log \max_{\theta_1} p(\boldsymbol{x}|\boldsymbol{z}; \theta_1) + \boxed{\log \sum_{\boldsymbol{x}} \max_{\theta_1} p(\boldsymbol{x}|\boldsymbol{z}; \theta_1)}$$

$$\mathcal{L}_{\mathrm{NML}}(\boldsymbol{z}; K) = -\log \max_{\theta_2} p(\boldsymbol{z}; \theta_2) + \boxed{\log \sum_{\boldsymbol{z}} \max_{\theta_2} p(\boldsymbol{z}; \theta_2)}$$

Finite Mixture Models

$\boldsymbol{z}_k$ subsequence of $\boldsymbol{z}$ s.t. $Z = k$

$\boldsymbol{x}_k$ data sequence corresponding to $\boldsymbol{z}_k$

$$\begin{aligned}\mathcal{L}_{\mathrm{NML}}(\boldsymbol{x}|\boldsymbol{z};K) &= -\log \prod_k \max_{\theta_1} p(\boldsymbol{x}_k|\boldsymbol{z}_k;\theta_1) + \log \prod_k \left(\sum_{\boldsymbol{x}_k} \max_{\theta_1} p(\boldsymbol{x}_k|\boldsymbol{z}_k;\theta_1) \right) \\ &= \sum_k \mathcal{L}_{\mathrm{NML}}(\boldsymbol{x}_k|\boldsymbol{z}_k)\end{aligned}$$

where $\mathcal{L}_{\mathrm{NML}}(\boldsymbol{x}_k|\boldsymbol{z}_k) = -\log \max_{\theta_1} p(\boldsymbol{x}_k|\boldsymbol{z}_k;\theta_1) + \log \sum_{\boldsymbol{x}_k} \max_{\theta_1} p(\boldsymbol{x}_k|\boldsymbol{z}_k;\theta_1)$.

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

Multinomial distribution

$$\mathcal{L}_{\mathrm{NML}}(\boldsymbol{z};K) = -\log \prod_{k=1}^{K} \left(\frac{n_k}{n}\right)^{n_k} + \log C_n^{\mathrm{MN}}(K)$$

$$C_n^{\mathrm{MN}}(K) \stackrel{\mathrm{def}}{=} \sum_{\substack{n_1 + \cdots + n_K = n \\ n_i \geq 0,\ i = 1, \ldots, K}} \frac{n!}{n_1! \cdots n_K!} \prod_{k=1}^{K} \left(\frac{n_k}{n}\right)^{n_k}$$

Efficient Computation of Parametric Complexity for Multinomial Distribution

$$C_n^{\mathrm{MN}}(K) \stackrel{\mathrm{def}}{=} \sum_{\substack{n_1+\cdots+n_K=n \\ n_i \geq 0,\ i=1,\ldots,K}} \frac{n!}{n_1!\cdots n_K!} \prod_{k=1}^{K} \left(\frac{n_k}{n}\right)^{n_k}$$

Petri Myllimaki

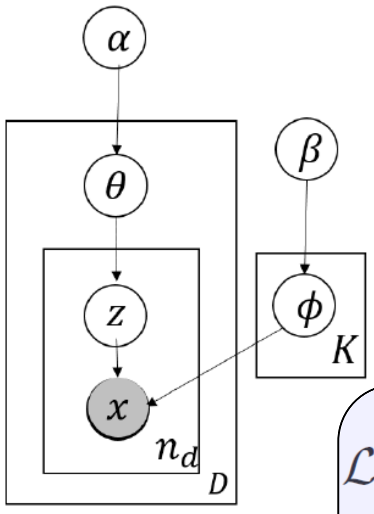
Theorem 3.3.2 (Recurrence relation of parametric complexity of multinomial distribution) [Kontkanen, Myllymaki IPL 2006]

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

$$\Longrightarrow O(n+K) \text{ comput. time}$$

3.3.3 Applications to a Variety of Classes

Example 3.3.2 (Latent Dirichlet Allocation: LDA)



b) LDA

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

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

$$\mathcal{L}_{\mathrm{DNML}}(\boldsymbol{x}, \boldsymbol{z}; K) = \sum_k \sum_v n_{kv} (\log n_k - \log n_{kv}) + \sum_k \log C_{\mathrm{MN}}(n_k, V) + \sum_d \sum_k n_{dk} (\log n_d - \log n_{dk}) + \sum_d \log C_{\mathrm{MN}}(n_d, K),$$

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

where

$C_{\mathrm{MN}}(n, K) = C_n^{\mathrm{MN}}(K)$

n_{kv}: ♯ word v in topic k

n_k: ♯ words in topic k

n_{dk}: ♯ words in topic k from document d

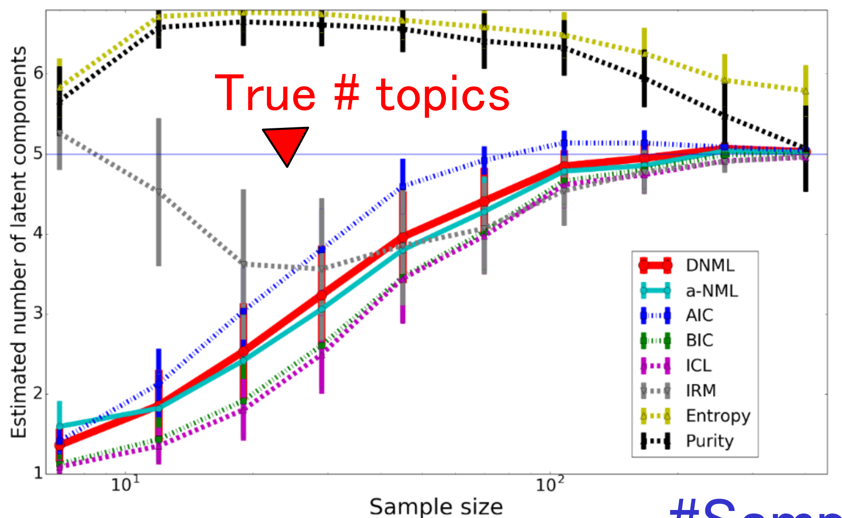
n_d: ♯ words in document d

Empirical Evaluation

-Synthetic Data-

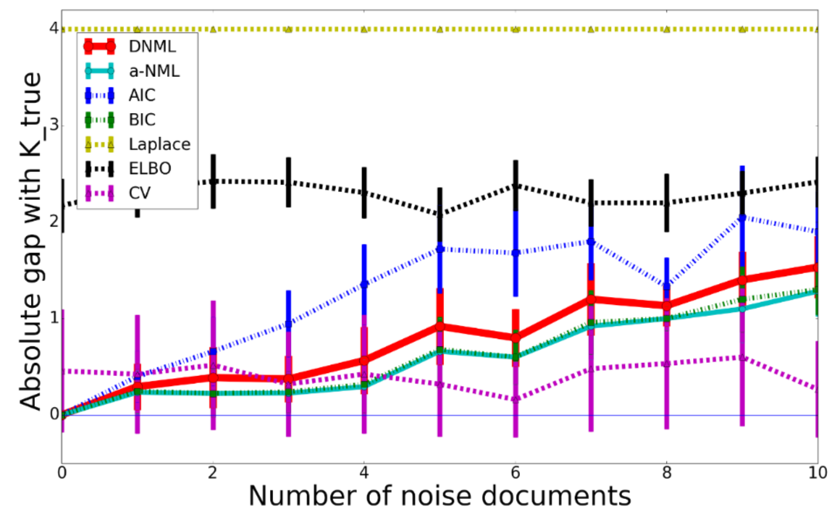
DNML is able to identify the true # topics of LDA most robustly

Estimated #topics



#Sample

Error Rate



Noise Ratio

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

DNML is more robust against noise than AIC.

[Wu, Sugawara, Yamanishi KDD2017

Empirical Evaluation

-Benchmark Data-

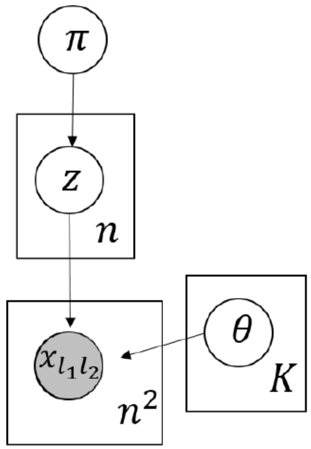
DNML is able to identify the true # topics most exactly

Newsgroup Dataset

Method	2 topics	3 topics	4 topics	5 topics	6 topics
DNML	**2**	**3**	**4**	**5**	**7**
AIC	7	4	9	7	**7**
Laplace	8	4	8	9	5
ELBO	3	4	**4**	4	3
CV	9	1	3	7	1
HDP	80	98	94	92	98

[Wu, Sugawara, Yamanishi KDD2017]

Example 3.3.3 (Stochastic Block Models: SBM)



c) SBM

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

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

$$
\begin{aligned}
&\mathcal{L}_{\mathrm{DNML}}(\boldsymbol{x}, \boldsymbol{z}; K) \\
&= \sum_{k_1} \sum_{k_2} \left(n_{k_1 k_2} \log n_{k_1 k_2} - n^1_{k_1 k_2} \log n^1_{k_1 k_2} - n^0_{k_1 k_2} \log n^0_{k_1 k_2} \right) \\
&\quad + \sum_{k_1} \sum_{k_2} \log C_{\mathrm{MN}}(n_{k_1 k_2}, 2) \\
&\quad + \sum_{k} n_k (\log n - \log n_k) + \log C_{\mathrm{MN}}(n, K),
\end{aligned}
$$

$\Longrightarrow$ computable in time $O(n + K)$

where

$n_{k_1 k_2}$: $\sharp$ occurrences in (k_1, k_2) cluster

$n^1_{k_1 k_2}$: $\sharp$ links in (k_1, k_2) cluster

$n^0_{k_1 k_2}$: $\sharp$ no-links in (k_1, k_2) cluster

n_k: $\sharp$ occurrences in k cluster

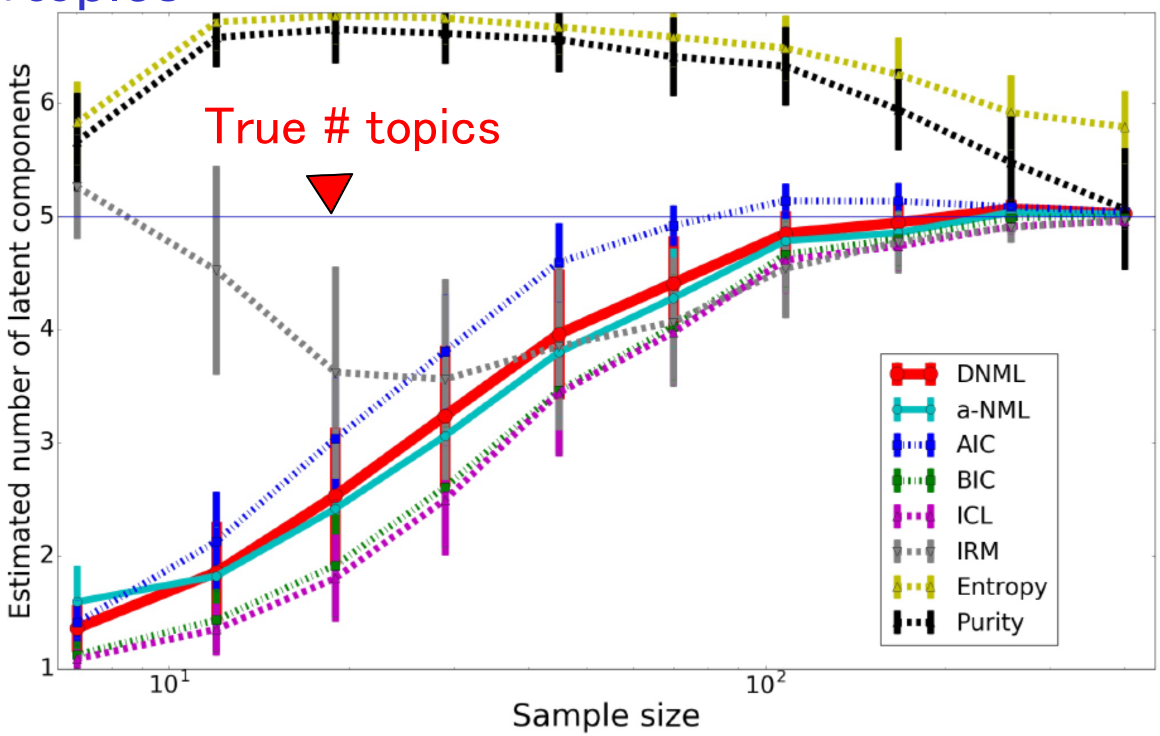
Empirical Evaluation

-Synthetic Data-

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

[Wu, Sugawara, Yamanishi KDD2017]

Estimated #topics



#Sample

Figure 3: SBM: Estimated number K vs sample size n with $K_{true} = 5$

Example 3.3.4 (Gaussian Mixture Models: GMM)

For observations $i = 1, ..., n$:

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

$$L_{\mathrm{DNML}}(\boldsymbol{x}, \boldsymbol{z}; K) = \sum_{k=1}^{K} L_{\mathrm{NML}}(\boldsymbol{x}_k) + \sum_{k=1}^{K} n_k(\log n - \log n_k) + \log C_{\mathrm{MN}}(n, K),$$

$$\begin{aligned} L_{\mathrm{NML}}(\boldsymbol{x}_k) \leq\ & \frac{mn_k}{2}\log(2\pi) + \frac{n_k}{2}\log|\hat{\Sigma}_k| + \frac{mn_k}{2} \\ & + \frac{mn_k}{2}\log\frac{n_k}{2e} - \frac{m(m-1)}{4}\log\pi - \sum_{j=1}^{m}\log\Gamma\left(\frac{n_k - j}{2}\right) \\ & - m\log\frac{m}{2} - \log\Gamma\left(\frac{m}{2} + 1\right) \\ & + \frac{m}{2}\log||\hat{\mu}_k||_2^2 - \frac{m}{2}\sum_{j=1}^{m}\log\lambda^{(j)}(\hat{\Sigma}_k) \\ & + (m+1)\log\frac{m}{2} + \log\log\frac{R_2}{R_1} + m\log\log\frac{\lambda_2}{\lambda_1}, \end{aligned}$$

Empirical Evaluation
-Synthetic Data-

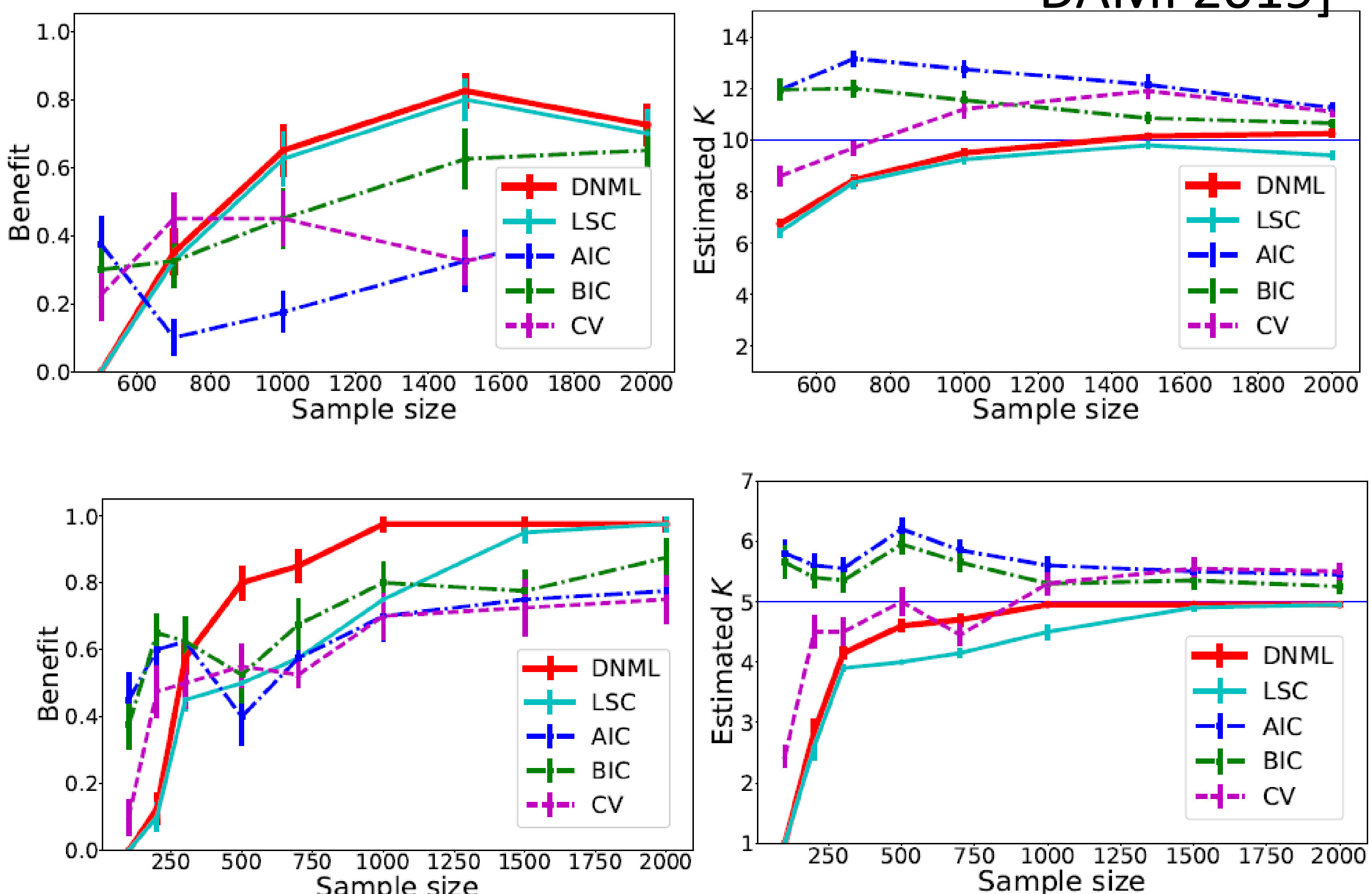
DNML performs best and can identify small components exactly.

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

[Yamanishi et al. DAMI 2019]

m=5
K*=10

m=5
K*=5
With small components



Minimax Optimality of DNML

Theorem 3.3.3 (Estimation Optimality of DNML)
[Yamanishi, Wu,Sugawara,Okada DAMI2019]

$\bar{K}(\cdot)$: Arbitrary model estimator

NML dist. associated with $\bar{K}$

$$\bar{p}(\boldsymbol{x},\boldsymbol{z})=\frac{\bar{p}(\boldsymbol{x}|\boldsymbol{z};\bar{K}(\boldsymbol{x},\boldsymbol{z}))\bar{p}(\boldsymbol{z};\bar{K}(\boldsymbol{x},\boldsymbol{z}))}{\bar{C}^n_{X,Z}},\quad \bar{C}^n_{X,Z}=\sum_{\boldsymbol{x},\boldsymbol{z}}\bar{p}(\boldsymbol{x},\boldsymbol{z};\bar{K}(\boldsymbol{x},\boldsymbol{z}))\bar{p}(\boldsymbol{z};\bar{K}(\boldsymbol{x},\boldsymbol{z})).$$

$\hat{K}(\cdot)$: DNML model estimator $\quad \hat{K}(\boldsymbol{x},\boldsymbol{z})=\underset{K}{\mathrm{argmax}}\,\{p_{\mathrm{NML}}(\boldsymbol{x}|\boldsymbol{z};K)p_{\mathrm{NML}}(\boldsymbol{z};K)\}.$

DNML dist. associated with $\hat{K}$

$$\hat{p}(\boldsymbol{x},\boldsymbol{z})=\frac{p_{\mathrm{NML}}(\boldsymbol{x}|\boldsymbol{z};\hat{K}(\boldsymbol{x},\boldsymbol{z}))p_{\mathrm{NML}}(\boldsymbol{z};\hat{K}(\boldsymbol{x},\boldsymbol{z}))}{\hat{C}^n_{X,Z}}$$

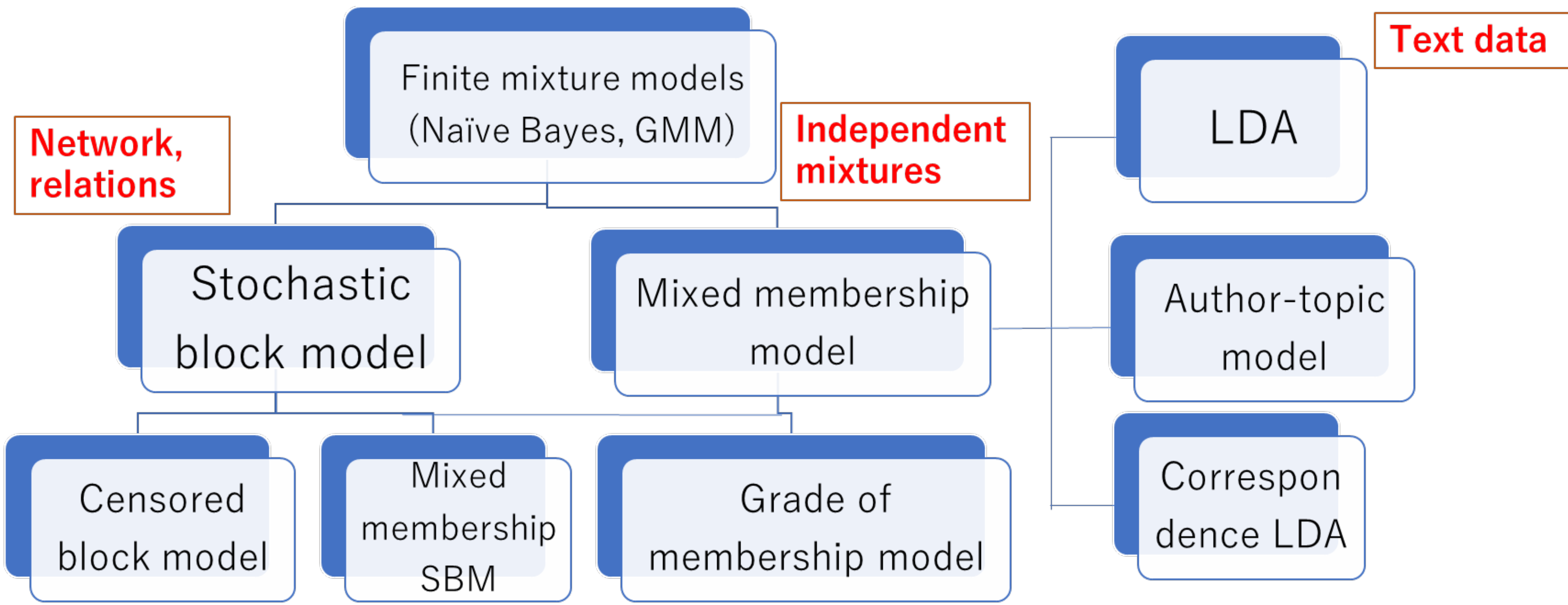
Then DNML dist. $\hat{p}_{\mathrm{DNML}}$ achieves the minimum of

$$\min_{\bar{p}}\max_{\theta,K} D(p_{\theta,K}||\bar{p})$$

KL－Divergence

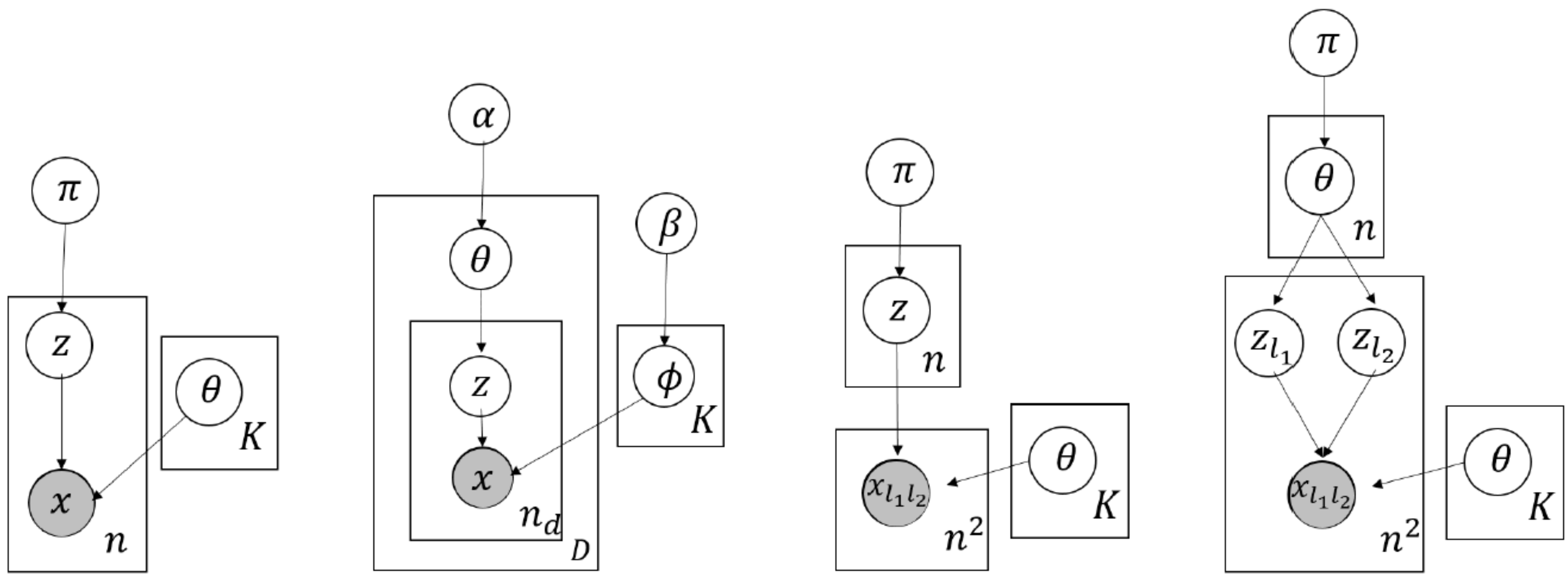
Wide Applicability of DNML

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.

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:

$\boldsymbol{x} = x_1, \ldots, x_n$: data sequence, Θ: parameter space

$$\hat{\theta}(\boldsymbol{x}, \lambda) = \underset{\theta \in \Theta}{\mathrm{argmin}} \{-\log p(\boldsymbol{x}; \theta) + g(\theta, \lambda)\}$$

where $$g(\theta, \lambda) = \frac{1}{2} \sum_{j=1}^{d} \lambda_j \theta_j^2$$ penalty

- Small λ_j for essential parameters
- Large λ_j for unnecessary parameters

Luckiness NML Codelength (LNML)

[Grűnwald 2007]

Peter Grunwald

- Luckiness Minimax Regret

$$\mathcal{P} = \{p(\boldsymbol{x};\theta) : \theta \in \Theta\}$$

$$LR_n(\mathcal{P}) \stackrel{\text{def}}{=} \min_{\mathcal{L}:\text{prefix code}} \max_{\boldsymbol{x}} \{\mathcal{L}(\boldsymbol{x}) - \min_{\theta}(-\log p(\boldsymbol{x};\theta) + g(\theta,\lambda)\}$$

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

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

where $\quad Z_n(\lambda) \stackrel{\text{def}}{=} \int \max_{\theta} p(\boldsymbol{x};\theta) e^{-g(\theta,\lambda)} d\boldsymbol{x}$

Problem:

1) $Z_n(\lambda)$ is anlytically intractable in general

2) How to choose λ?

Upper Bounding on LNML

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

LNML codelength is uniformly upper-bounded by

$$L_{\mathrm{LNML}}(\boldsymbol{x}) \leq \min_{\theta}(-\log p(\boldsymbol{x};\theta) + g(\theta,\lambda)) + \underbrace{\frac{1}{2}\log|H(\lambda)| + \log\int e^{-g(\theta,\lambda)}d\theta}_{\text{an upper bound on } Z_n(\lambda)} + \text{const}$$

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

Optimization Algorithm

[Miyaguchi, Yamanishi *Machine Learning* 2018]

- Upper bound on LNML = concave + convex function

$$\overline{L_{\mathrm{LNML}}}(\boldsymbol{x};\lambda) \stackrel{\mathrm{def}}{=} \underbrace{\min_{\theta}(-\log p(\boldsymbol{x};\theta)+g(\theta,\lambda))}_{\text{Concave}} + \underbrace{\frac{1}{2}\log|H(\lambda)| + \log\int e^{-g(\theta,\lambda)}d\theta}_{\text{Convex}}$$

$$\Longrightarrow \min \text{ w.r.t. } \lambda$$

- Concave-convex procedure (CCCP) [Yuille, Rangarajan NC 02]
  - ☑ Monotone non-increasing optimization
  - ☑ Convergence to a saddle point

Example 3.3.4 (Linear regression model)

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

The upper bound on LNML is calculated as

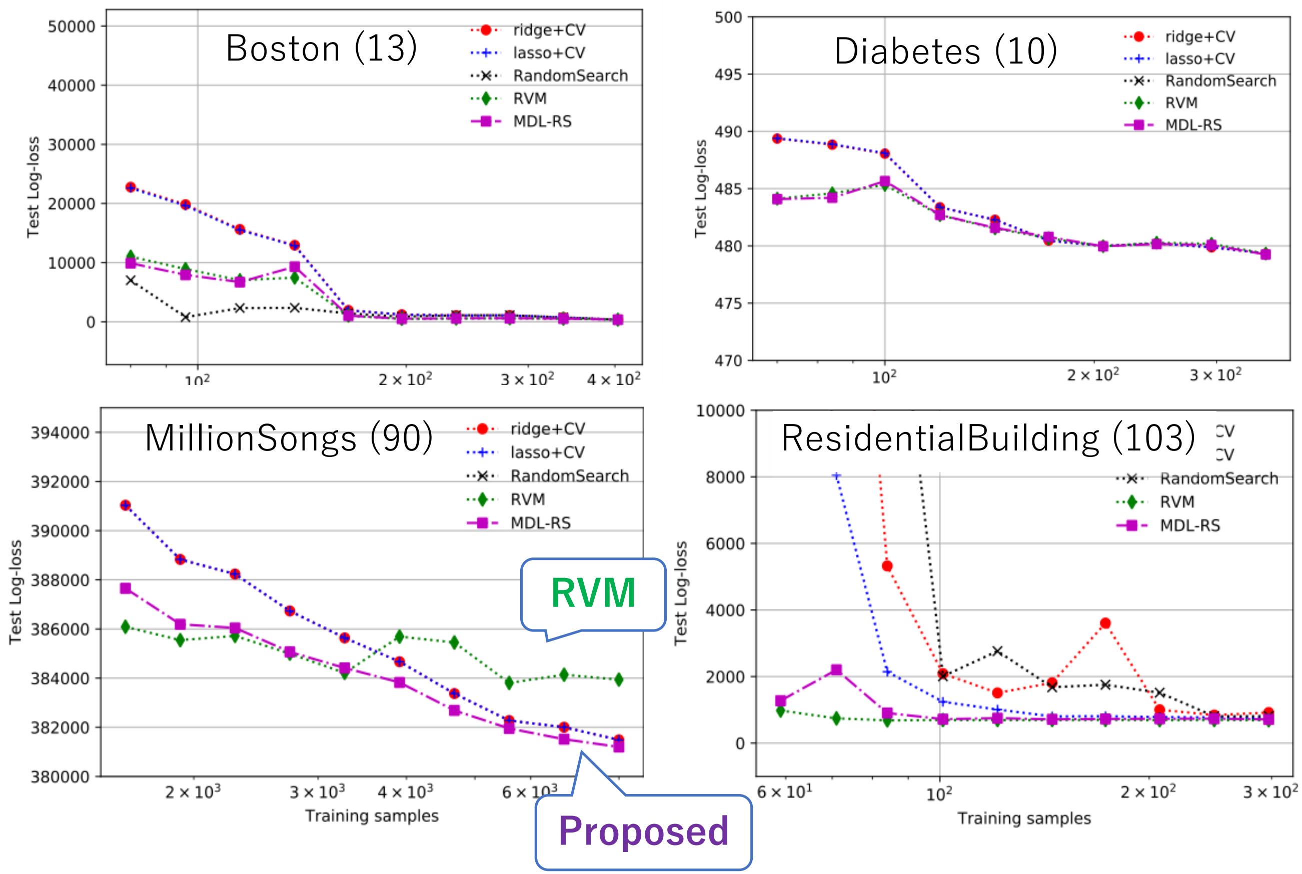
$$\overline{\mathcal{L}_{\mathrm{LNML}}}(\lambda) = \min_{\beta,\sigma^2} \underbrace{\frac{1}{2\sigma^2}\|y - X\beta\|^2 + \frac{n}{2}\ln\sigma^2}_{\text{loss}} + \underbrace{\frac{1}{2\sigma^2}\sum_{j=1}^{m}\lambda_j\beta_j^2}_{\text{penalty}} + \underbrace{\boxed{\frac{1}{2}\ln\frac{\det(X^\top X + \operatorname{diag}\lambda)}{\det \operatorname{diag}\lambda}}}_{\text{normalizing term}}$$

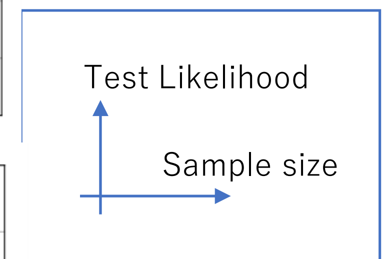
$$\Longrightarrow \min \text{w.r.t. } \lambda$$

Experiments: Linear Regression
—UCI Repository—

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

[Miyaguchi, Yamanishi *Machine Learning* 2018]





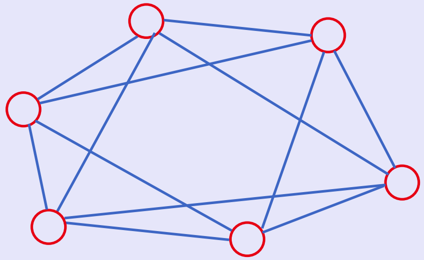
Example 3.3.5 (Graphical Model Estimation)

- m-variate Gaussian model
- Precision Θ represents conditional dependencies

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

- The upper bound on LNML

$$\overline{\mathcal{L}_{\mathrm{LNML}}}(\lambda) = \min_{\Theta} \underbrace{\frac{1}{2}\mathrm{Tr}[X^\top X\Theta] - \frac{n}{2}\ln\det\Theta}_{\text{loss}} + \underbrace{\frac{1}{2}\sum_{i<j}\lambda_{ij}\Theta_{ij}^2}_{\text{penalty}} + \underbrace{\boxed{\frac{1}{2}\sum_{i<j}\ln\left(1 + \frac{nR^2}{\lambda_{ij}}\right)}}_{\text{normalizing term}}$$

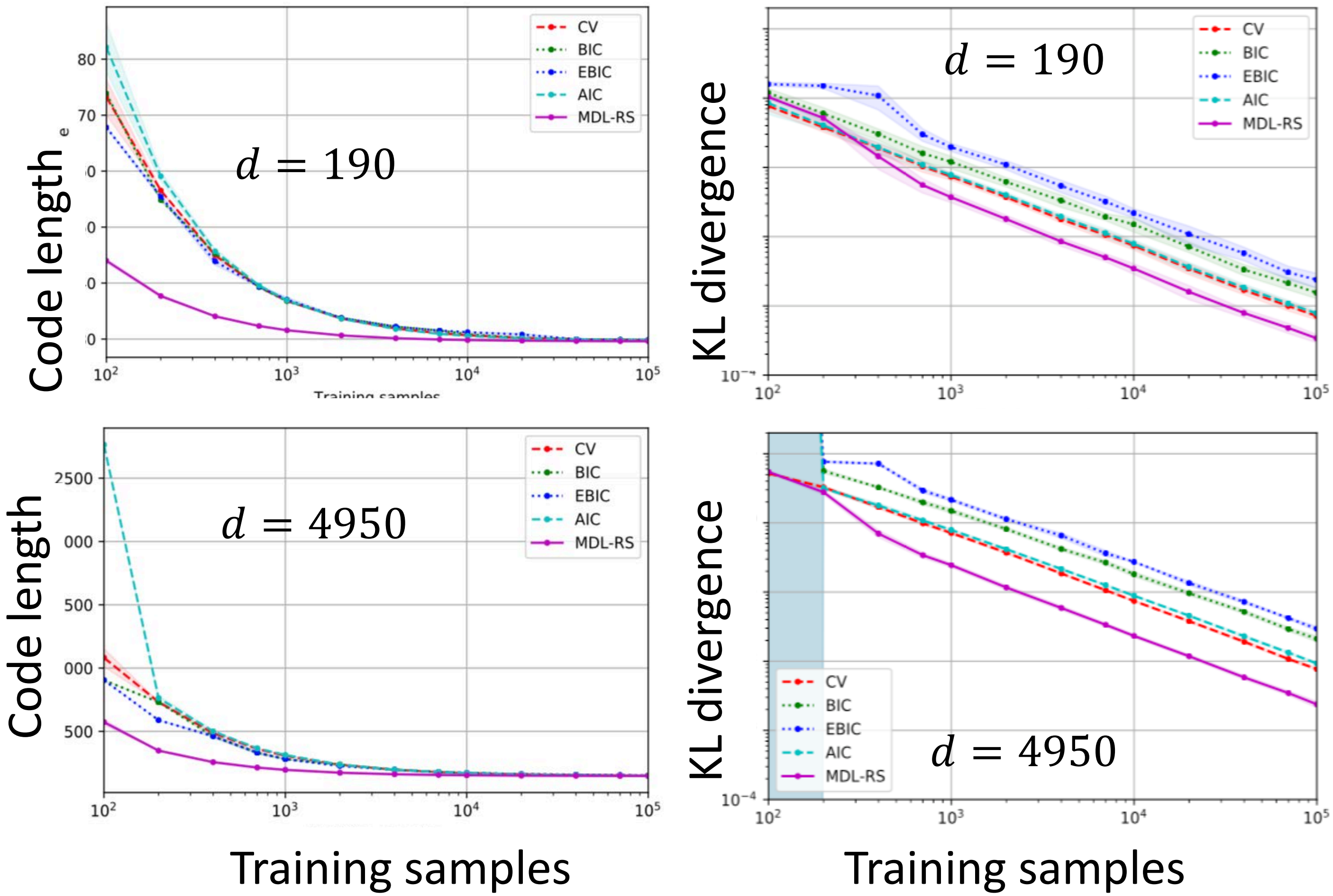


Experiments: Graphical Model

—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

$$\mathcal{L}_{\mathrm{LNML}}(\boldsymbol{x}) = \min_{\theta}(-\log p(\boldsymbol{x};\theta) + g(\theta,\lambda)) + \log Z_n(\lambda)$$

where $Z_n(\lambda) \stackrel{\mathrm{def}}{=} \int \max_{\theta} p(\boldsymbol{x};\theta) e^{-g(\theta,\lambda)} d\boldsymbol{x} \quad g(\theta,\lambda) = \sum_{j=1}^{d} \lambda_j |\theta_j|$

Stochastc Gradient Descent

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

$$\approx \quad \lambda - \eta(|\bar{\theta}(\boldsymbol{x},\lambda)| - |\bar{\theta}(\boldsymbol{y},\lambda)|)$$

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

$$\bar{\theta} = \operatorname*{argmax}_{\theta} p(\boldsymbol{x};\lambda) e^{-g(\theta,\lambda)}$$

3.4.2. High-dimensional Minimax Prediction

Problem:

$\boldsymbol{x} = x_1, x_2, \ldots, x_t, \ldots$: data sequence

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

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

attaining the minimax regret:

$$\min_{\mathcal{A}} \max_{\boldsymbol{x}} \left\{ \sum_{t=1}^{n} (-\log p(x_t|x^{t-1})) - \min_{\theta} \sum_{t=1}^{n} (-\log p(x_t;\theta)) \right\}$$

On-line Bayesian Prediction Strategy

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

[Clarke and Barron JSPI 1994, Takeuchi and Barron ISIT1998]

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

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

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

Then cumulative log loss amounts

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

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

Problem: No counterpart in high-dim setting ($d \gtrsim n$)

High-dimensional Asymptotics

Assumption

- Twice differetial of loss is uniformly upper bound by L
- ℓ_1-radius of P is at most $R < +\infty$

Worst-case regret for on-line prediction algorithm $\mathcal{A}$:

$$R_n(\mathcal{A}) \stackrel{\text{def}}{=} \max_{\boldsymbol{x}} \left\{ \sum_{t=1}^{n} (-\log p_{\mathcal{A}}(x_t|x^{t-1})) - \min_{\theta} \sum_{t=1}^{n} (-\log p(x_t;\theta)) \right\}$$

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,

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

For some L-smooth model, for any prediction alg. $\mathcal{A}$,

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

Optimal within a factor of 2

Minimax Predictor in High-dimensional Setting

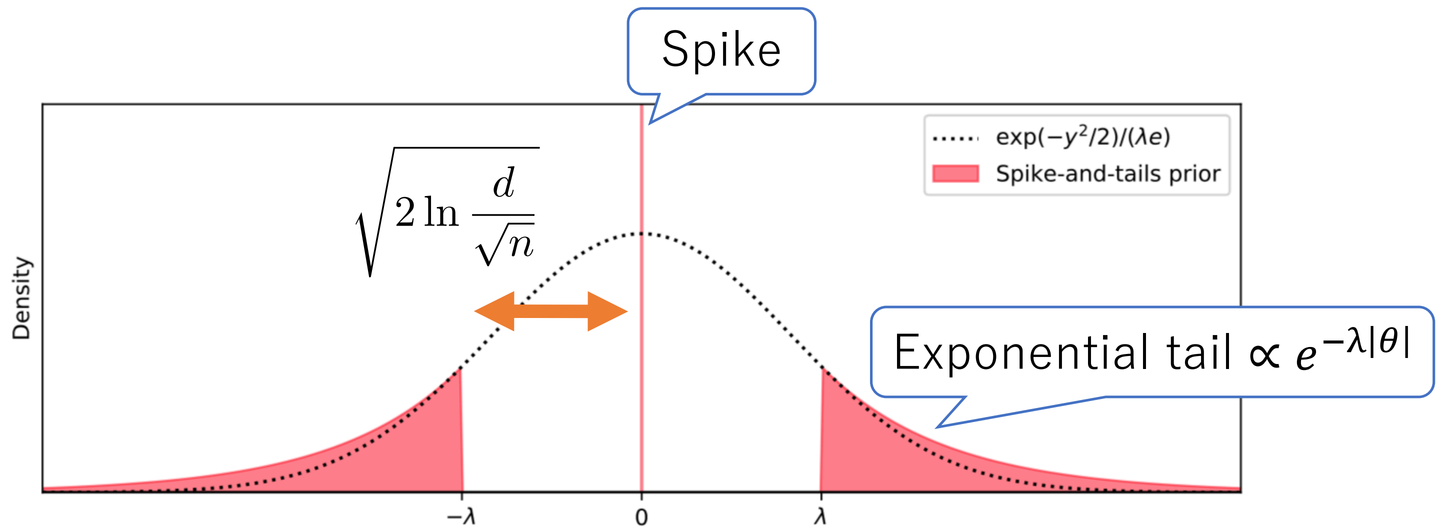
[Miyaguchi and Yamanishi AISTATS2019]

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

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

- One-dim illustration of ST prior

  Gap is wide when $d \gg \sqrt{n}$



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

- J.Rissanen: “Modeing by shortest description length,”*Automatica,*14:465–471,1978.
- Y.M. Shtar’kov: “Universal sequential coding of single messages,” *Problemy Peredachi Informatsii*, 23(3):3–17, 1987.
- J.Rissanen: “*Stochastic Complexity in Statistical Inquiries*,” Wiley, 1989.
- A.Barron and T.Cover: “Minimum complexity density estimator,” *IEEE Transactions on Information Theory*, Vol.37, 4, pp:1034-1054, 1991.
- K.Yamanishi： “A learning criterion for stochastic rules,” *Machine Learning*, Vol.9, pp:165-203, 1992.
- J.Rissanen: “Fisher information and stochastic complexity,” *IEEE Transactions on Information Theory*, 42(1):40–47, 1996.
- P.D. Grünwald: “*Minimum Description Length Principle*,” MIT Press, 2007.
- J.Rissanen： “*Optimal Estimation of Parameters*,” Cambridge, 2012.
- M.Kawakita and J.Takeuchi: “Barron and Cover’s theory in supervised learning and its applications to lasso,” in *Proceedings of International Conference on Machine Learning*, 2016.

References

■ 3.2. G-function and Fourier Transformation Method

- J. Rissanen: "MDL denoising," *IEEE Transactions on Information Theory*, 46(7):2537–2543, 2000.
- J.Rissanen: "*Information and Complexity in Statistical Modeling*," Springer, 2007.
- A. Suzuki and K. Yamanishi : "Exact calculation of normalized maximum likelihood code length using Fourier analysis" *Proceedings of IEEE Symposium on Information Theory (ISIT2018),* pp: 1211-1215, 2018.

■ 3.3.1. Latent Stochastic Complexity

- P. Kontkanen, P. Myllymaki, W. Buntine, J. Rissanen, and H. Tirri: "An mdl framework for data clustering," *In Advances in Minimum Description Length: Theory and Applications. MIT Press*, pp: 323–35, 2005.
- P. Kontkanen and P. Myllymäki : ˵A linear-time algorithm for computing the multinomial stochastic complexity," *Information Processing Letters*, 103(6), pp:227–233, 2007.
- I.O.Kyrgyzov, O.O.Kyrgyzov, H.Maître, M.Campedel : "Kernel MDL to determine the number of clusters," in *Proceedings of Workshop on Machine Learning and Data Mining in Pattern Recognition*. Lecture Notes in Computer Science, vol 4571, pp:2013-217, 2007.

References

■ 3.3.1. Latent Stochastic Complexity(Cont.)

・S. Hirai and K. Yamanishi: ˋEfficient computation of normalized maximum likelihood codes for gaussian mixture models with its applications to clustering,” *IEEE Transactions on Information Theory*, 59(11):7718–7727, 2013.

・Y.Sakai and K.Yamanishi: "An NML-based model selection criterion for general relational data modeling," *Proceedings of IEEE International Conference on Big Data (BigData 2013)*, pp:421--429, 2013.

・P.Miettenin and J.Vreeken: “MDL4BMF: Minimum Description Length for Boolean Matrix Factorization,” *ACM Transactions on Knowledge Discovery from Data*, Vol. 8 Issue 4, 2014.

・A. Suzuki, K.Miyaguchi, and K. Yamanishi: “Structure selection convolutive non-negative matrix factorization using normalized maximum likelihood coding," *Proceedings of IEEE International Conference on Data Mining (ICDM2016),* pp:1221-1226, 2016.

・Y.Ito, S.Oeda, and K.Yamanishi: “Rank selection for non-negative matrix factorization with normalized maximum likelihood coding." *Proceedings of SIAM International Conference on Data Mining (SDM2016),* pp:720-728, Mar. 2016.

・S.Squires, A.Prügel-Bennett, M.Niranjan: “Rank selection in nonnegative matrix factorization using minimum description length,” *Neural Computation,* 29, 2164–2176, 2017.

References

■ 3.3.1. Latent Stochastic Complexity(Cont.)

・T.Nakamura, T.Iwata, and K.Yamanishi: "Latent dimensionality estimation for probabilistic canonical correlation analysis using normalized maximum likelihood code-length," *Proceedings of the 4th IEEE International Conference on Data Science and Advanced Analytics (DSAA2017*), pp:716-725, 2017.

・Y. Fu, S. Matsushima, K. Yamanishi: "Model selection for non-negative tensor factorization with minimum description length," *Entropy*, Jul. 2019.

・S.Hirai and K.Yamanishi: "Correction to Efficient computation of normalized maximum likelihood codes for Gaussian mixture models with its applications to clustering," IEEE Transactions on Information Theory, 2019.

■ 3.3.2. Decomposed Normalized Maximum Likelihood Codelength

・T. Wu, S. Sugawara, K.Yamanishi: "Decomposed normalized maximum likelihood codelength criterion for selecting hierarchical latent variable models," *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2017),* pp:1165--1174, 2017.

・K. Yamanishi, T. Wu, S.Sugawara, M.Okada: "The decomposed normalized maximum likelihood code-length criterion for selecting hierarchical latent variable models", *Data Mining and Knowledge Discovery*. 33(4): 1017-1058, 2019.

References

■ 3.3.3. High-dimensional Sparse Model Selection

- S. Chatterjee and A.Barron: "Information-theoretic validity of penalized likelihood," in *Proceedings of IEEE International Symposium on Information Theory*, pp:3027-3031, 2014.
- K. Miyaguchi, S.Matsushima, and K. Yamanishi: "Sparse graphical modeling via stochastic complexity," *Proceedings of 2017 SIAM International Conference on Data Mining (SDM2017)*, pp:723-731, 2017.
- K. Miyaguchi and K. Yamanishi: "High-dimensional penalty selection via minimum description length principle" Machine Learning, 107(8-10), pp:1283-1302, 2018.
- K. Miyaguchi and K. Yamanishi: "Adaptive minimax regret against smooth logarithmic losses over high-dimensional l1-balls via envelope complexity", *Proceedings of Artificial Intelligence and Statistics (AISTATS 2019)*, pp:3440-3448, 2019.

C.f. Bayesian prediction related to MDL

- B.Clarke and A.Barron: "Jeffreys' prior is asymptotically least favorable under entropy risk," *Journal of Statistical Planning and Inference,* Vol. 41, Issue 1, pp:37-60,1994
- J.Takeuchi and A.Barron: "Asymptotically minimax regret by Bayes mixtures," in Proceedings of IEEE International Symposium on Information Theory, 1998.