+

fNML Criterion for Learning Bayesian Network Structures

Tomi Silander
Teemu Roos
Petri Kontkanen
Petri Myllymaki

PGM-08
Hirtshals

September 17-19 2008

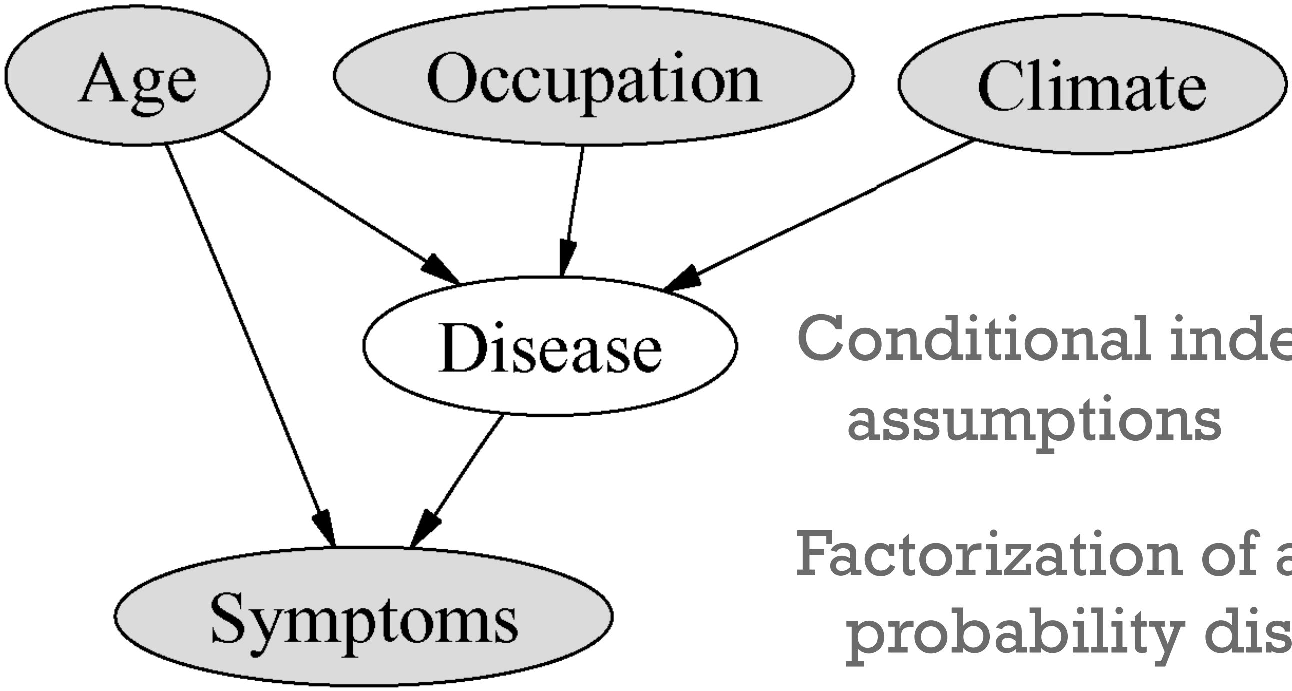
HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

Helsinki Institute for Information Technology HIIT
FINLAND

Outline:

1. Bayesian Networks
2. Model Selection Scores
3. New Stuff: fNML Score

Bayesian Networks



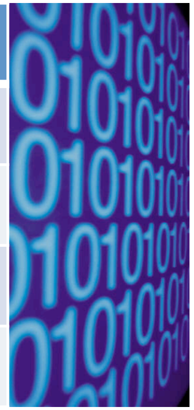
Conditional independence assumptions

Factorization of a joint probability distribution:

$$P(x \mid G) = \prod_{i=1}^{m} P(x_i \mid G_i).$$

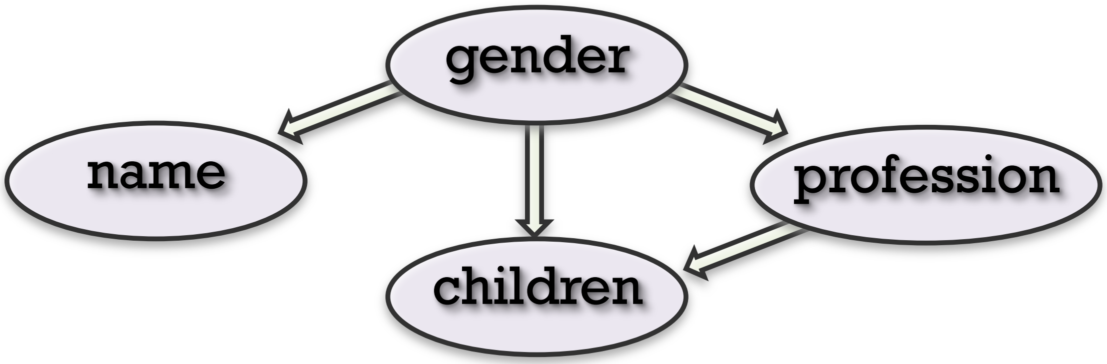
Data

NAME	GENDER	PROFESSION	CHILDREN
Teemu	male	researcher	2
Clark	male	reporter	0
Margrethe	female	queen	2
:	:	:	:



Data

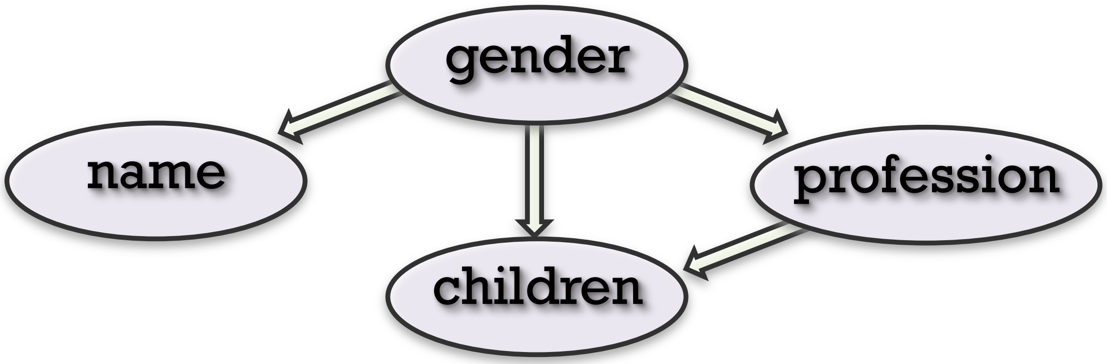
NAME	GENDER	PROFESSION	CHILDREN
Teemu	male	researcher	2
Clark	male	reporter	0
Margrethe	female	queen	2
:	:	:	:



Data

NAME	GENDER	PROFESSION	CHILDREN
Teemu	male	researcher	2
Clark	male	reporter	0
Margrethe	female	queen	2
:	:	:	:

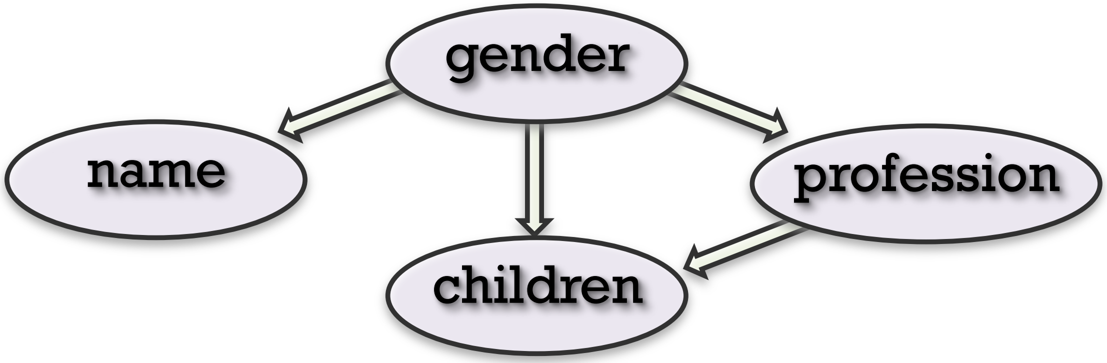
D



Data

NAME	GENDER	PROFESSION	CHILDREN
Teemu	male	researcher	2
Clark	male	reporter	0
Margrethe	female	queen	2
:	:	:	:

D_i



Model Selection: Scores

- Bayes (BDe)
- BIC & AIC
- MDL

Bayesian Score

The state-of-the-art model selection criterion:

Bayesian Dirichlet equivalent (BDe) score

Assumes Dirichlet prior on model parameters θ .

Evaluate marginal likelihood of data given model

$$P(D \mid G, \alpha) = \int P(D \mid G, \theta) P(\theta \mid G, \alpha) \, d\theta.$$

Depends on hyper-parameter α.

BIC & AIC

BIC: Asymptotic approximation of marginal likelihood:

$$BIC(G, D) = \log \hat{P}(D \mid G) - \frac{k}{2} \log n.$$

AIC: Asymptotic approximation of estimated prediction error:

$$AIC(G, D) = \log \hat{P}(D \mid G) - k.$$

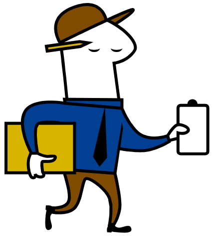
MDL

Minimum Description Length (MDL) Principle:

Choose the model that yields the shortest description of the data together with the model.

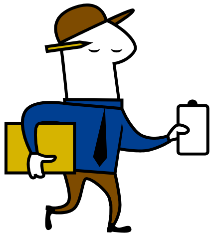
Too simple model	data long, model short
"Just right"	data short, model short
Too complex model	data short, model long

Flavours of MDL



1. "Pedestrian"
   Asymptotic two-part code-length same as BIC.

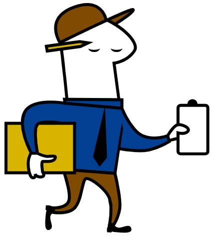
Flavours of MDL



1. "Pedestrian"
   Asymptotic two-part code-length same as BIC.

2. "Sophisticated"
   Bayesian marginal likelihood.

+ Flavours of MDL



1. "Pedestrian"
   Asymptotic two-part code-length same as BIC.

2. "Sophisticated"
   Bayesian marginal likelihood.

3. "Champions League"
   Modern (minimax regret optimal) code

   normalized maximum likelihood (NML)

Problem: NML computationally very hard.

Bayes vs. MDL (minimax regret)

The Bayesian decision principle is **minimization of expected loss**:

$$min_A \; E_X [loss(A,X)]$$

MDL (especially NML) is based on **minimization of worst-case regret**:

$$min_A \; max_X \underbrace{[loss(A,X) - min_{A'} \; loss(A',X)]}_{\text{"regret"}}$$

New stuff:
fNML Score

- fNML = "factorized NML"
- computation
- consistency

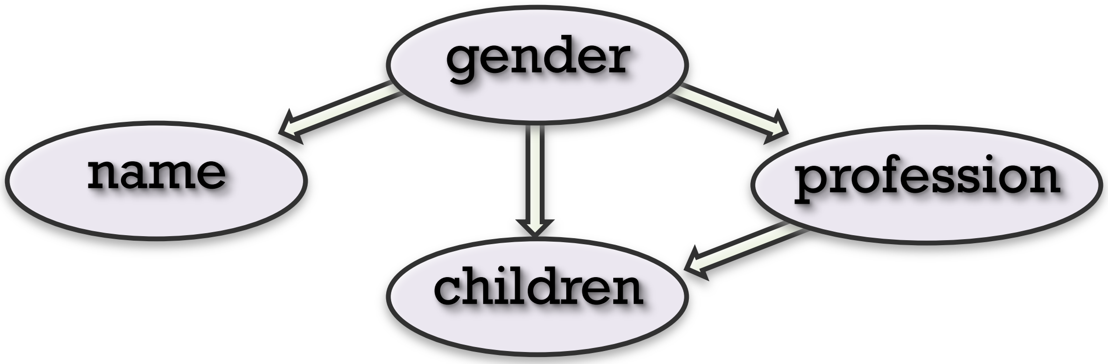
fNML Score

We propose a new MDL score, factorized NML, which is

1. easy to compute,
2. decomposable (allowing fast search),
3. robust (experimentally).

fNML vs. NML: what's new?

NAME	GENDER	PROFESSION	CHILDREN
Teemu	male	researcher	2
Clark	male	reporter	0
Margrethe	female	queen	2
:	:	:	:

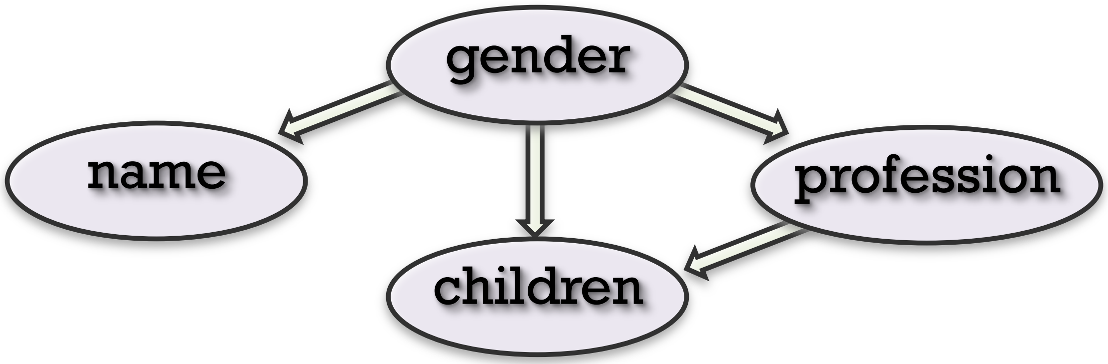


fNML vs. NML: what's new?

NML: Minimax code applied to whole data as one block

NAME	GENDER	PROFESSION	CHILDREN
Teemu	male	researcher	2
Clark	male	reporter	0
Margrethe	female	queen	2
:	:	:	:

D

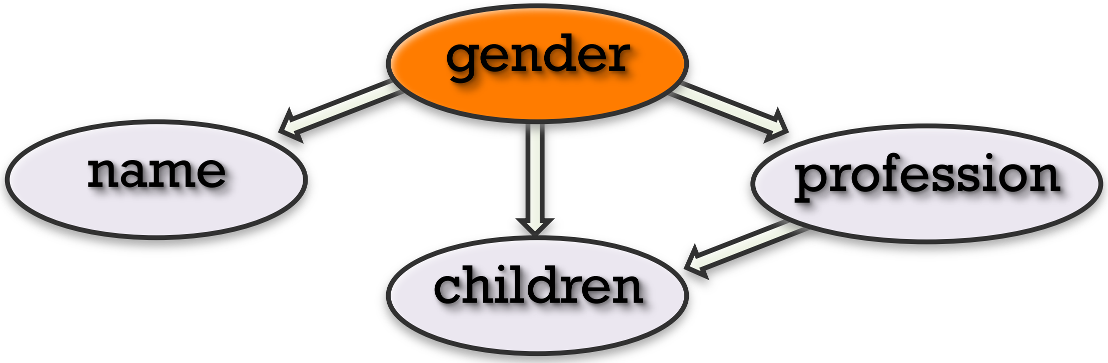


fNML vs. NML: what's new?

fNML: minimax code applied
column by column

D_2

NAME	GENDER	PROFESSION	CHILDREN
Teemu	male	researcher	2
Clark	male	reporter	0
Margrethe	female	queen	2
:	:	:	:

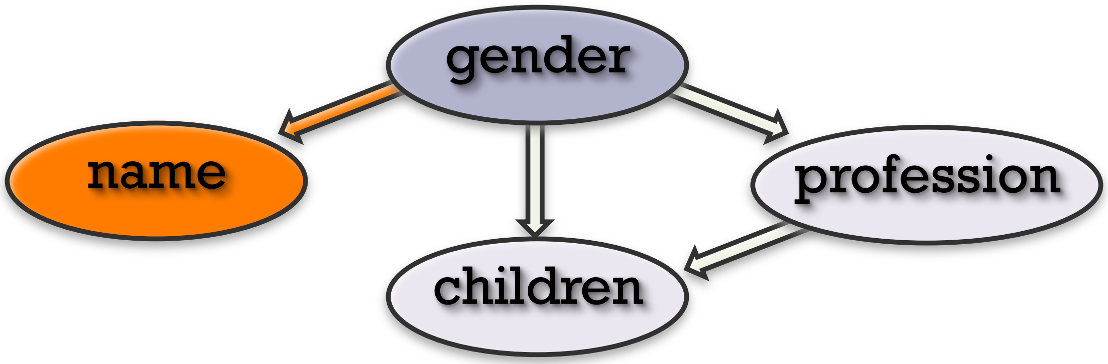


fNML vs. NML: what's new?

fNML: Conditional minimax code when parent(s) exist.

NAME	GENDER	PROFESSION	CHILDREN
Teemu	male	researcher	2
Clark	male	reporter	0
Margrethe	female	queen	2
:	:	:	:

D_1

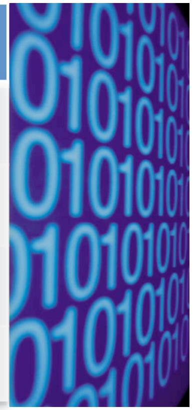


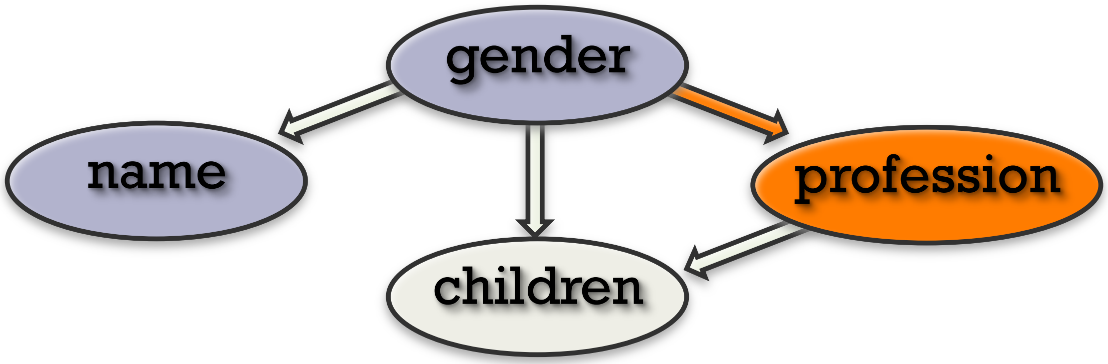
fNML vs. NML: what's new?

fNML: Conditional minimax code when parent(s) exist.

NAME	GENDER	PROFESSION	CHILDREN
Teemu	male	researcher	2
Clark	male	reporter	0
Margrethe	female	queen	2
:	:	:	:

D_3



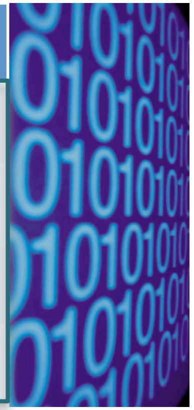


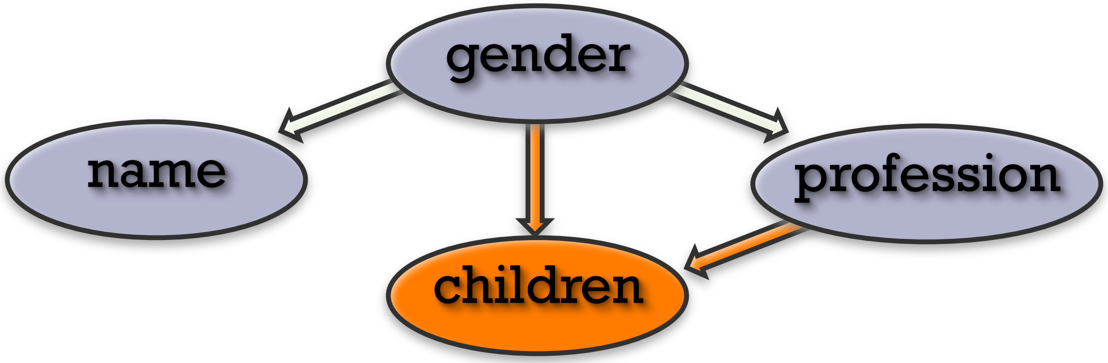
fNML vs. NML: what's new?

fNML: Conditional minimax code when parent(s) exist.

NAME	GENDER	PROFESSION	CHILDREN
Teemu	male	researcher	2
Clark	male	reporter	0
Margrethe	female	queen	2
:	:	:	:

D_4



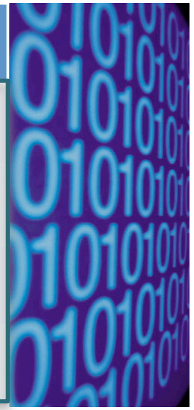


fNML vs. NML: what's new?

fNML: Conditional minimax code when parent(s) exist.

NAME	GENDER	PROFESSION	CHILDREN
Teemu	male	researcher	2
Clark	male	reporter	0
Margrethe	female	queen	2
:	:	:	:

D_4



Each column is encoded using the minimax code for multinomials.

Using fast NML algorithms, this takes O(n log n) per column.

fNML: Consistency

(Haughton, 1988): Any penalized likelihood score of the form

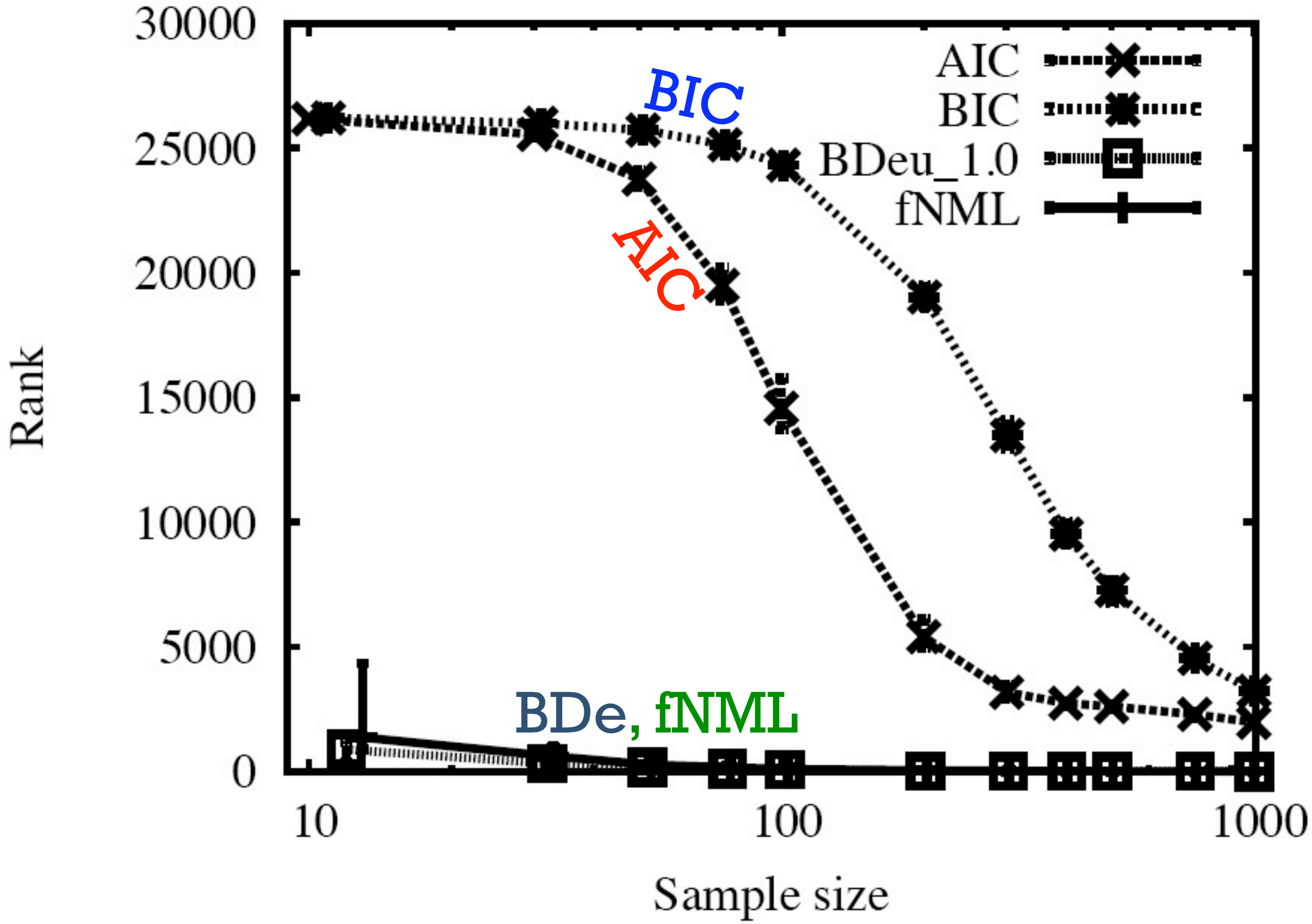
$$SCORE(G, D) = \log \hat{P}(D \mid G) - \frac{k}{2} a_n,$$

where a_n satisfies $a_n / n \to 0$ and $a_n \to \infty$, is consistent.

Theorem: fNML behaves asymptotically like BIC, i.e., $a_n = log\ n$.

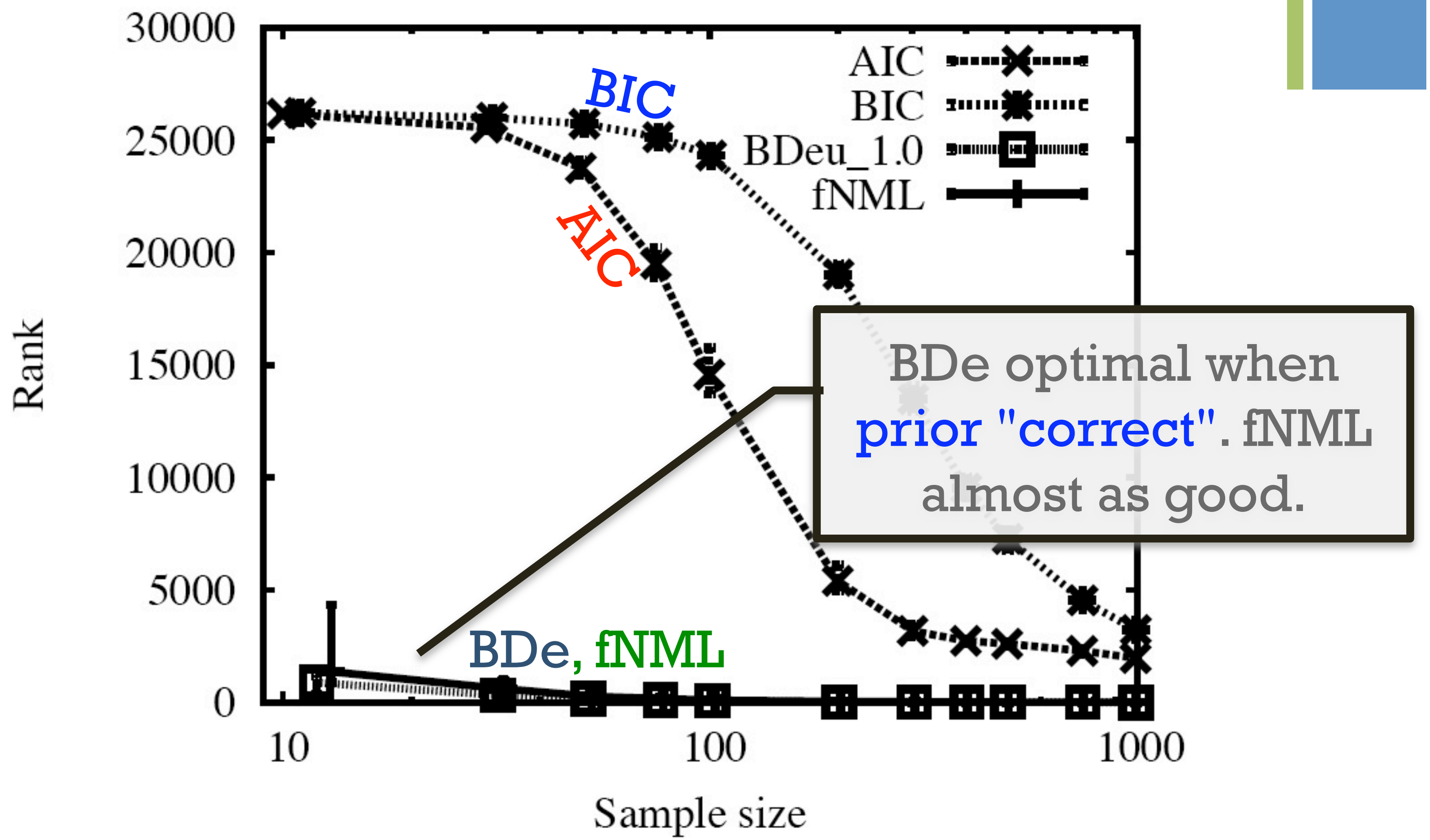
Hence, fNML is consistent.

Robustness



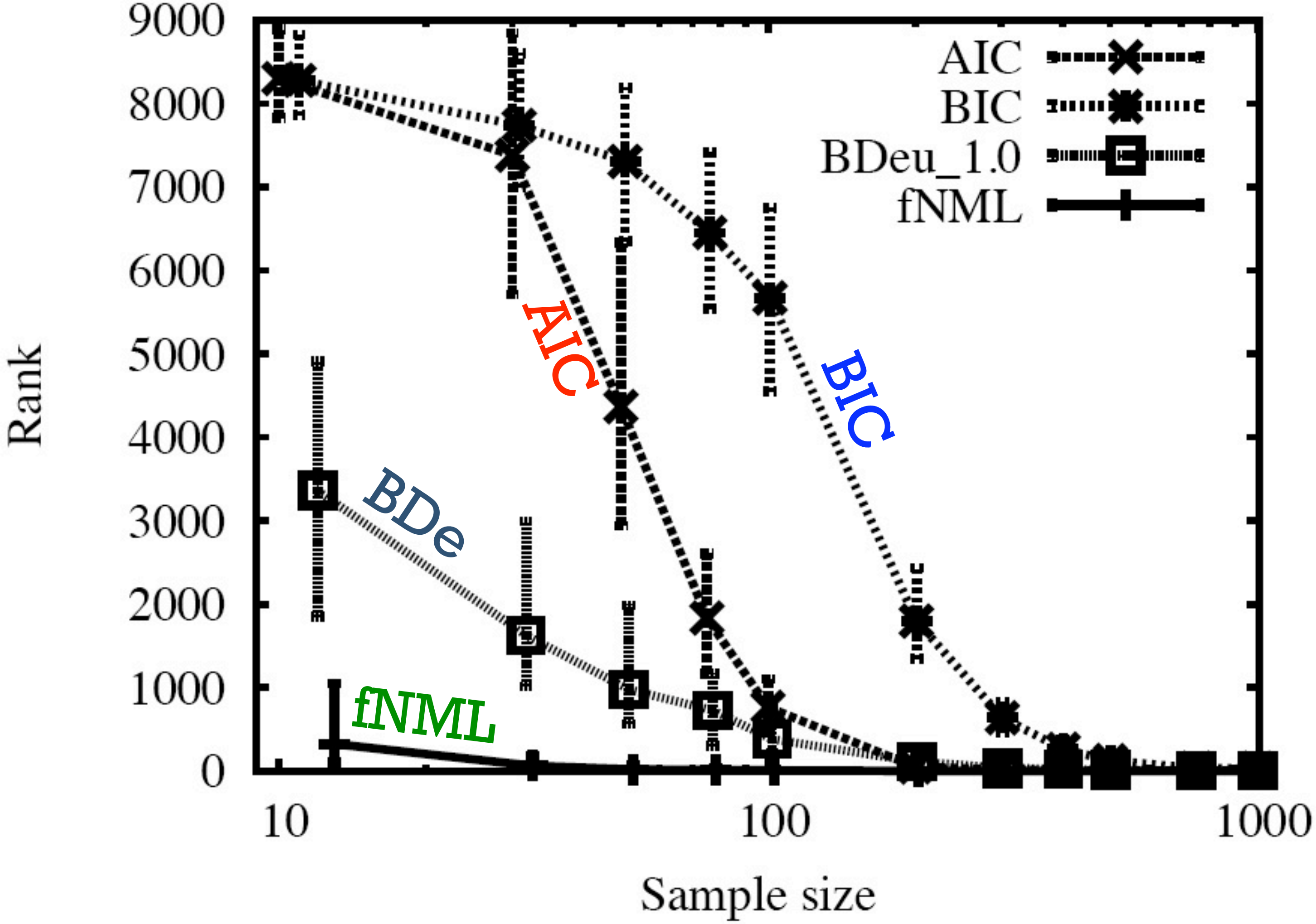
(a) BDeu Scheme

Robustness



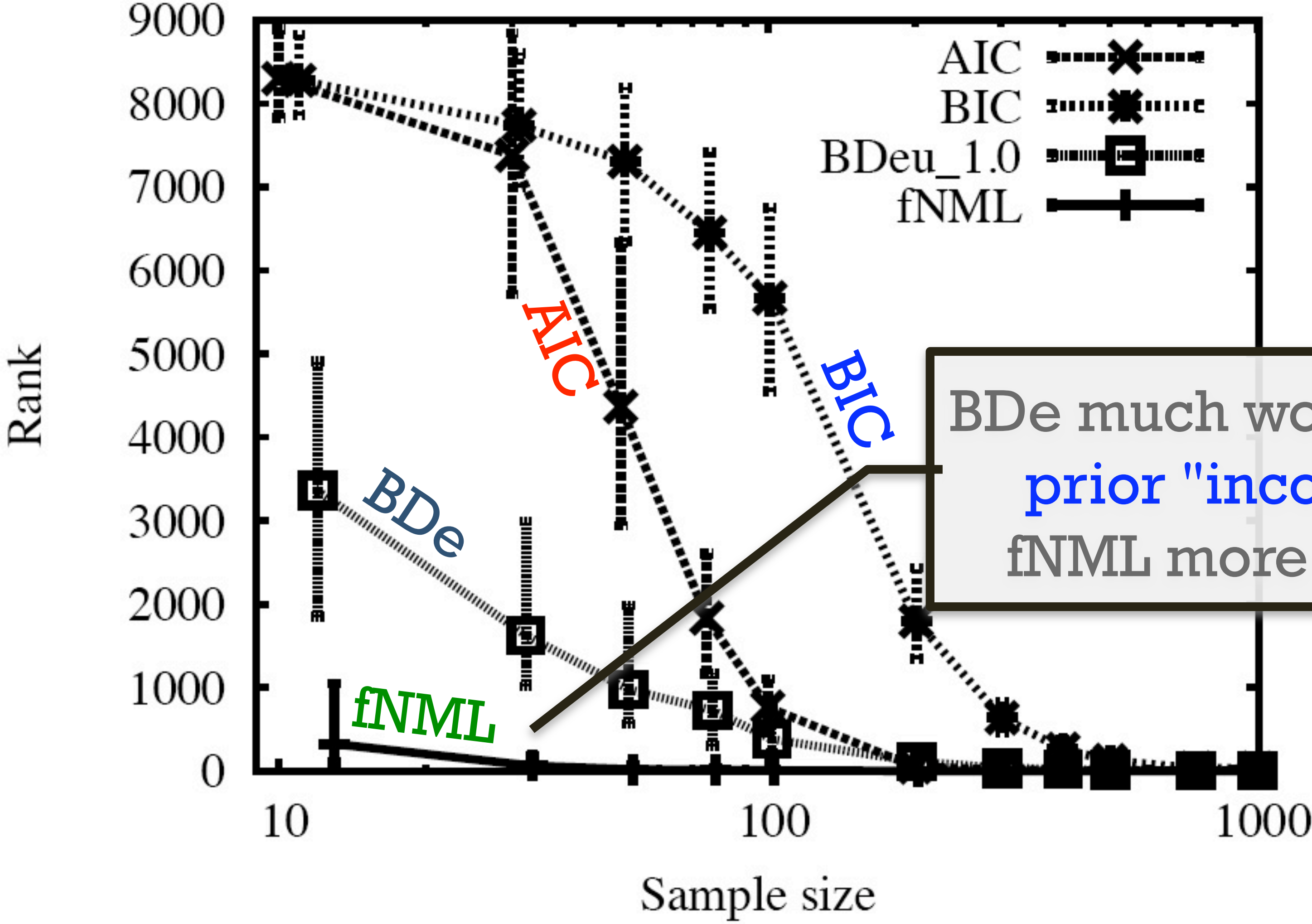
(a) BDeu Scheme

Robustness



(b) Dir(1/2,. . . ,1/2) Scheme

Robustness



(b) Dir(1/2,. . . ,1/2) Scheme

BDe much worse when **prior "incorrect"**.
fNML more **robust**.


Questions?

Decomposable Scores

Problem: Super-exponential search space.

Solution: Decomposable scores

$$SCORE(G,D) = \sum_{i=1}^{m} S(D_i, D_{Gi})$$

For decomposable scores, exact search (global optimum) can be done for about $m \leq 30$ **nodes** (Koivisto & Sood, 2004; Silander and Myllymäki, 2006).