Confident Bayesian Learning of Graphical Models

***

MIKKO KOIVISTO

Machine Learning Coffee Seminar
22 January 2018

AGENDA

I Graphical Models

II Confident Bayesian Structure Learning

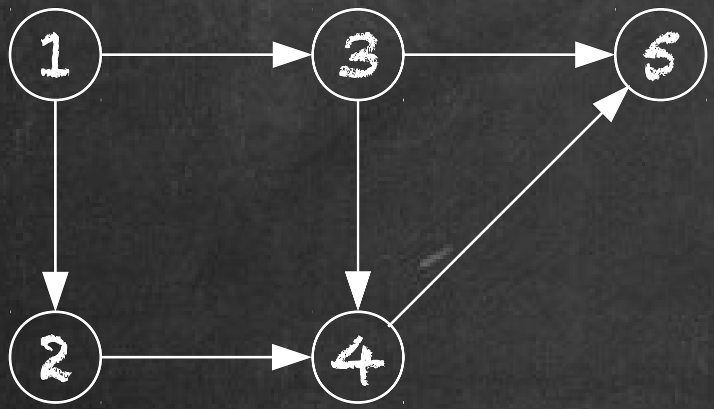
III State of the Art - Summary

IV Computational Techniques & Sample Results

Part I

Graphical Models

BAYESIAN NETWORK (BN)



Structure

G : a DAG

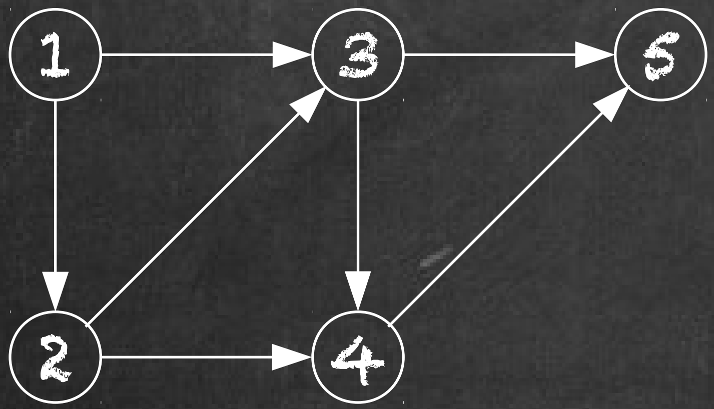
G_v: parents of v

$G_4 = \{2, 3\}$

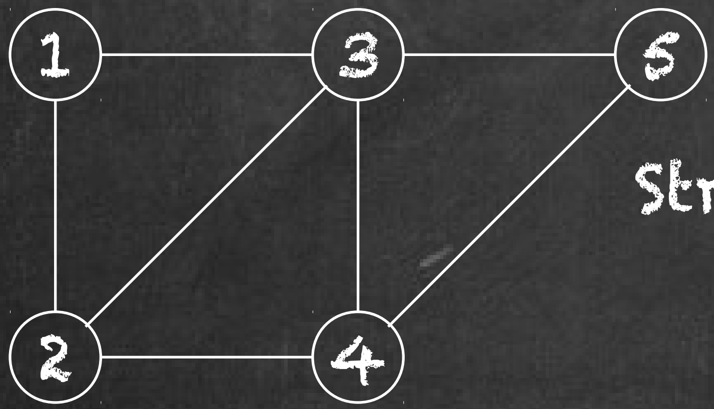
Local conditional distributions

$$p(x) = \prod_{v=1}^{d} p(x_v | x_{G_v})$$

CHORDAL MARKOV NETWORK (CMN)



CHORDAL MARKOV NETWORK (CMN)



Structure

G : a chordal graph

C : a clique in G

Cliques: {1, 2, 3}, {2, 3, 4}, {3, 4, 5}

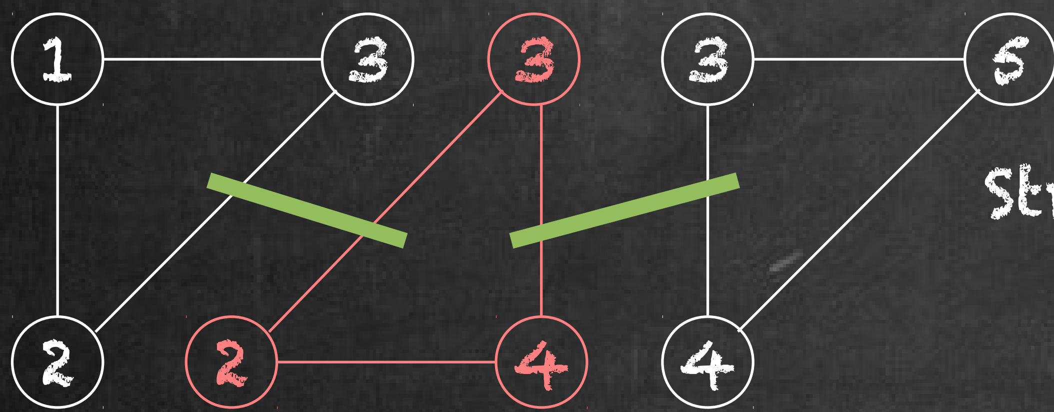
Separators: {2, 3}, {3, 4}

Local marginal distributions

$$p(x) = \frac{\prod_{\text{clique } C} p(x_C)}{\prod_{\text{separator } S} p(x_S)}$$

CHORDAL MARKOV NETWORK (CMN)

Clique tree



Structure

G : a chordal graph

C : a clique in G

Cliques: {1, 2, 3}, {2, 3, 4}, {3, 4, 5}

Separators: {2, 3}, {3, 4}

Local marginal distributions

$$p(x) = \frac{\prod_{\text{clique } C} p(x_C)}{\prod_{\text{separator } S} p(x_S)}$$

Part II

Confident Bayesian Structure Learning

BAYESIAN STRUCTURE LEARNING

Given a data set $X = (x^1, \ldots, x^N) = (X_1, \ldots, X_d)$, summarize the posterior of G.

Mode

$$\operatorname{argmax} P(G)\, P(X \mid G)$$

Expectations

$$E[\, f(G) \mid X \,]$$

Example: $f(G) = 1\{u \text{ is a parent of } v\}$

Normalizing constant

$$\sum_G P(G)\, P(X \mid G) = P(X)$$

CONFIDENT BAYESIAN LEARNING

Attach the estimate of the target quantity Q with a characterization of accuracy.

Example (confidence interval):

Output an interval I and a number p such that Q belongs to I with probability at least p

Example (exact deterministic):

$I = \{Q\}$ and $p = 1$

Characterization required only *a posteriori* !

Part III

State of the Art - Summary

BAYESIAN NETWORKS

NP-hard [Chickering 1996, Chickering et al. 2004]

Exact algorithms, time 2^d poly(d) or 3^d poly(d)
[Ott et al. 2004, K. & Sood 2004, Singh & Moore 2005, Silander & Myllymäki 2006, Tian & He 2009]

Complete solvers based on B&B, A*, [redacted]
programming [de Campos & Ji 2011, [redacted] 2013,
Bartlett & Cussens 2015, van Beek & Hoffm[redacted]

Tractable subclasses
[Gaspers et al. 2014, Korhonen & Parviai[redacted]

Markov chain Monte Carlo (MCMC) [redacted]
1995, Friedman & Koller 2003, Niinimäki [redacted]

Mode only!

Mode only!

Mostly <u>un</u>confident!

CHORDAL MARKOV NETWORKS

NP-hard [Srebro 2003]

Exact algorithms, time 4^d poly(d)
[Kangas et al. 2014, 2015]

Complete solvers based on constrai[redacted]
programming, ILP, B&B [Corande[redacted]
Cussens 2015, Rantanen et al. 2017]

Mode only!

For bounded-treewidth network[redacted]
Parviainen 2013, Berg et al. 2014, Parviai[redacted]

Mode only!

Markov chain Monte Carlo (MCMC)
[Giudici & Green 1999, Green & Thomas [redacted]

Unconfident!

Part IV

Computational Techniques & Sample Results

THE COVERING TECHNIQUE

Cover the structure space by "nice-behaving" sets.

Example (MCMC for BNs):

$$\{\text{DAGs}\} = \bigcup_{\text{constraint } \partial} \{\text{DAGs compatible with } \partial\}$$

Constraint families

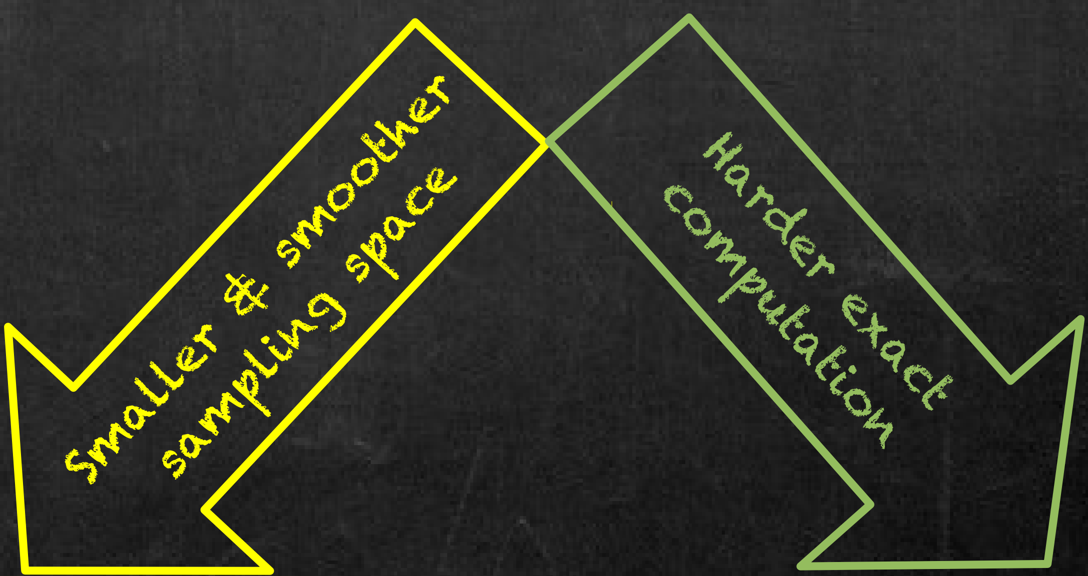
DAGs
[Madigan & York 1995]

Linear orders
[Friedman & Koller 2003]

Partial orders
[Niinimäki et al. 2011-16]

No constraints
[K. & Sood 2004]

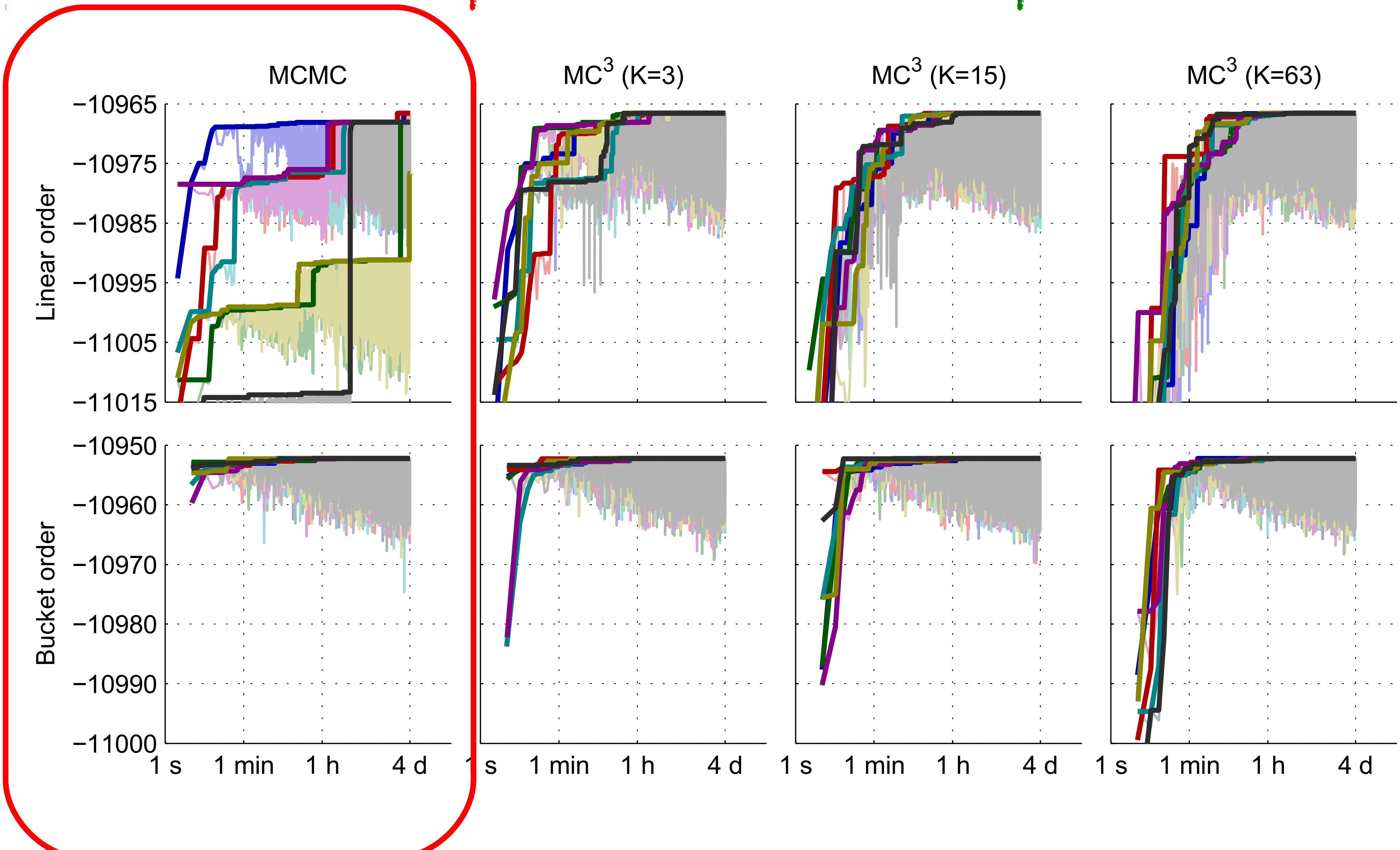
Tradeoffs



MCMC CONVERGENCE [Niinimäki et al. 2016]

Dataset: Mushroom, 1000 points, d = 22 variables

Bucket orders superior, even with tempered MCMC



DYNAMIC PROGRAMMING

Exploit a factorization to memorize repeated subproblems.

Chordal Markov networks:

- Call a function f decomposable if

$$f(G) = \frac{\prod_{\text{clique } C} f'(C)}{\prod_{\text{separator } S} f''(S)}$$

- Require a decomposable structure prior $P(G)$.

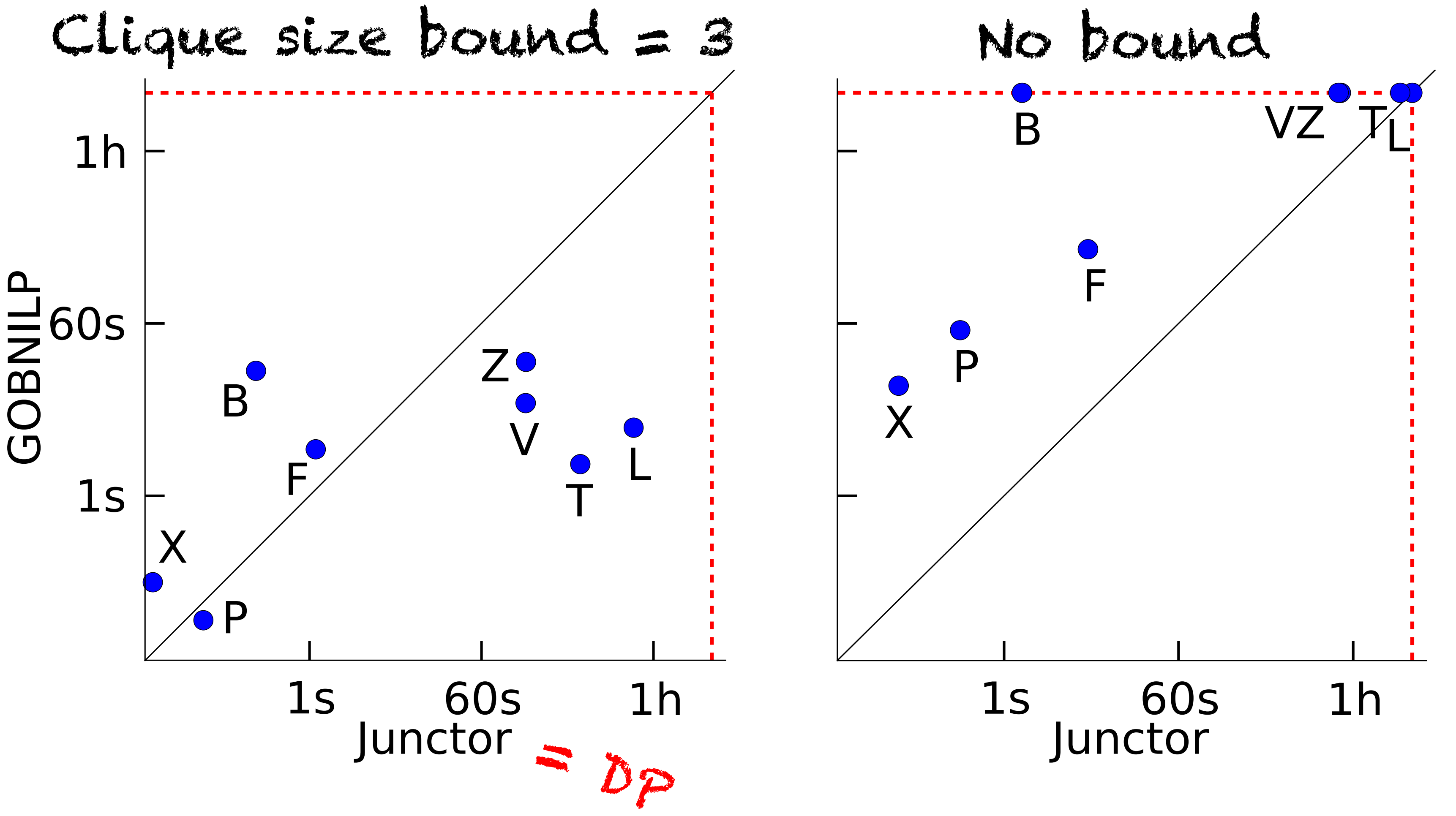
- Require a "nice" parameter prior of $p(x_C)$ for each clique C to get a decomposable $P(X|G)$.

- (Only expectations of decomposable functions.)

FINDING A MODE [Kangas et al. 2014]

Eight benchmark datasets, d = 10 to 19 variables

DP superior, especially if allowing large cliques



IMPORTANCE SAMPLING & BIAS CORRECTION

Sample from a "nice" proxy distribution q.
Correct bias by importance weighting.

Draw $G^1, \ldots, G^T$ independently from

$$q(G) = P(G|X)\, b(G) \cdot \text{Constant}$$

Put $w^t = 1/b(G^t)$ and estimate

$$E[f(G)|X] \approx \sum_t f(G^t)\, w^t \Big/ \sum_t w^t$$

Examples:

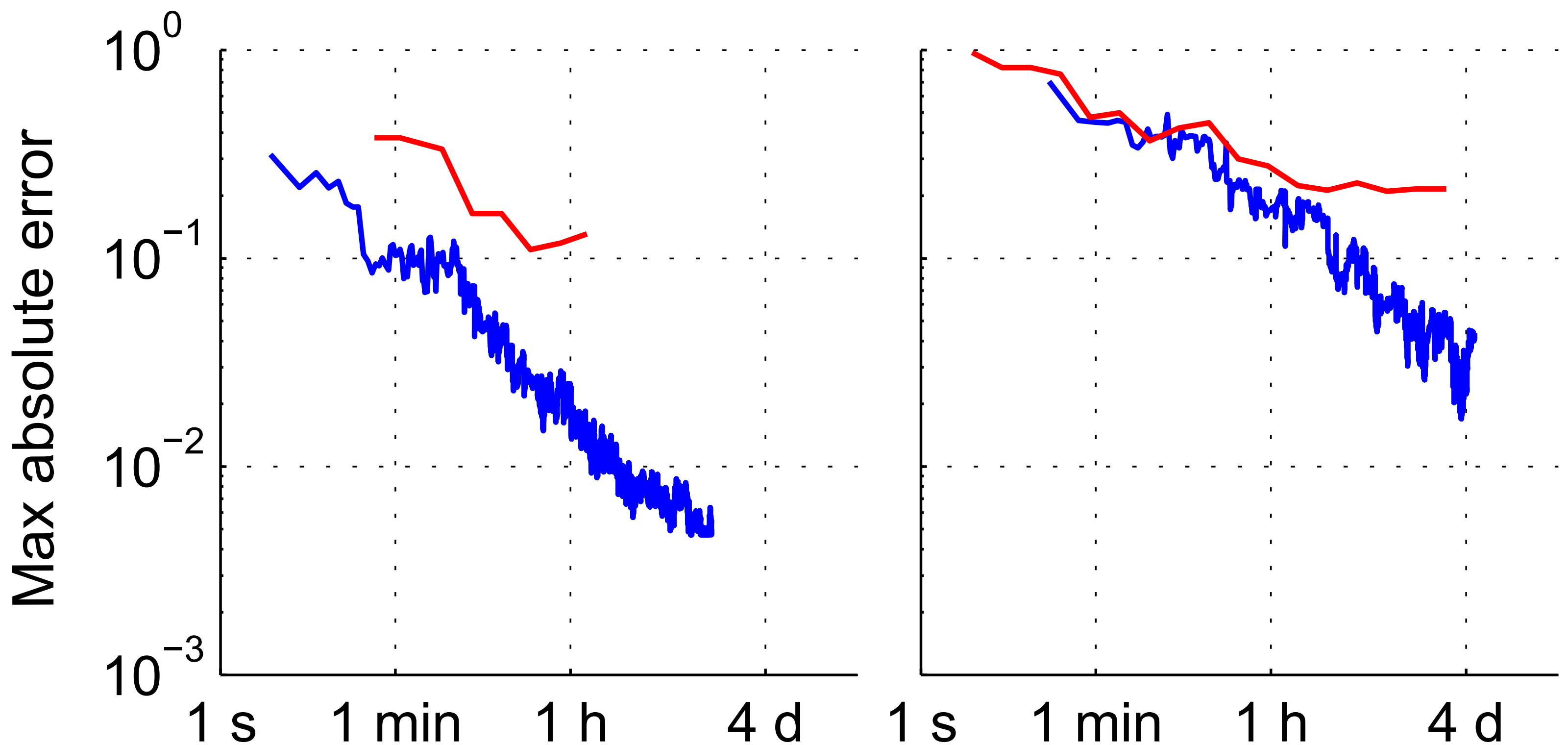
BNs: $b(G)$ = # topological sorts of G

CMNs: $b(G)$ = # rooted clique trees of G

ARC POSTERIOR [Niinimäki et al. 2016]

Importance weighting is superior to a heuristic by Ellis & Wong [JASA 2008]

Datasets: Flare, d = 13   German, d = 20



ANNEALING & MARKOV'S INEQUALITY

Construct a sampling distribution q close to the target. Independent samples yield lower bounds.

Annealed importance sampling [Neal 2001]

Generate ∂_K along a Markov chain:

$$q(\partial_1, \ldots, \partial_K) = q_1(\partial_1)\, q_2(\partial_2 | \partial_1) \ldots q_K(\partial_K | \partial_{K-1})$$

Theorem (Markov lower bound) [Gomes at al. 2007]

Let $Z_1, \ldots, Z_T$ be independent and nonnegative with mean μ. Then with probability at least p:

$$(1-p)^{1/T} \min\{Z_1, \ldots, Z_T\} \leq \mu$$

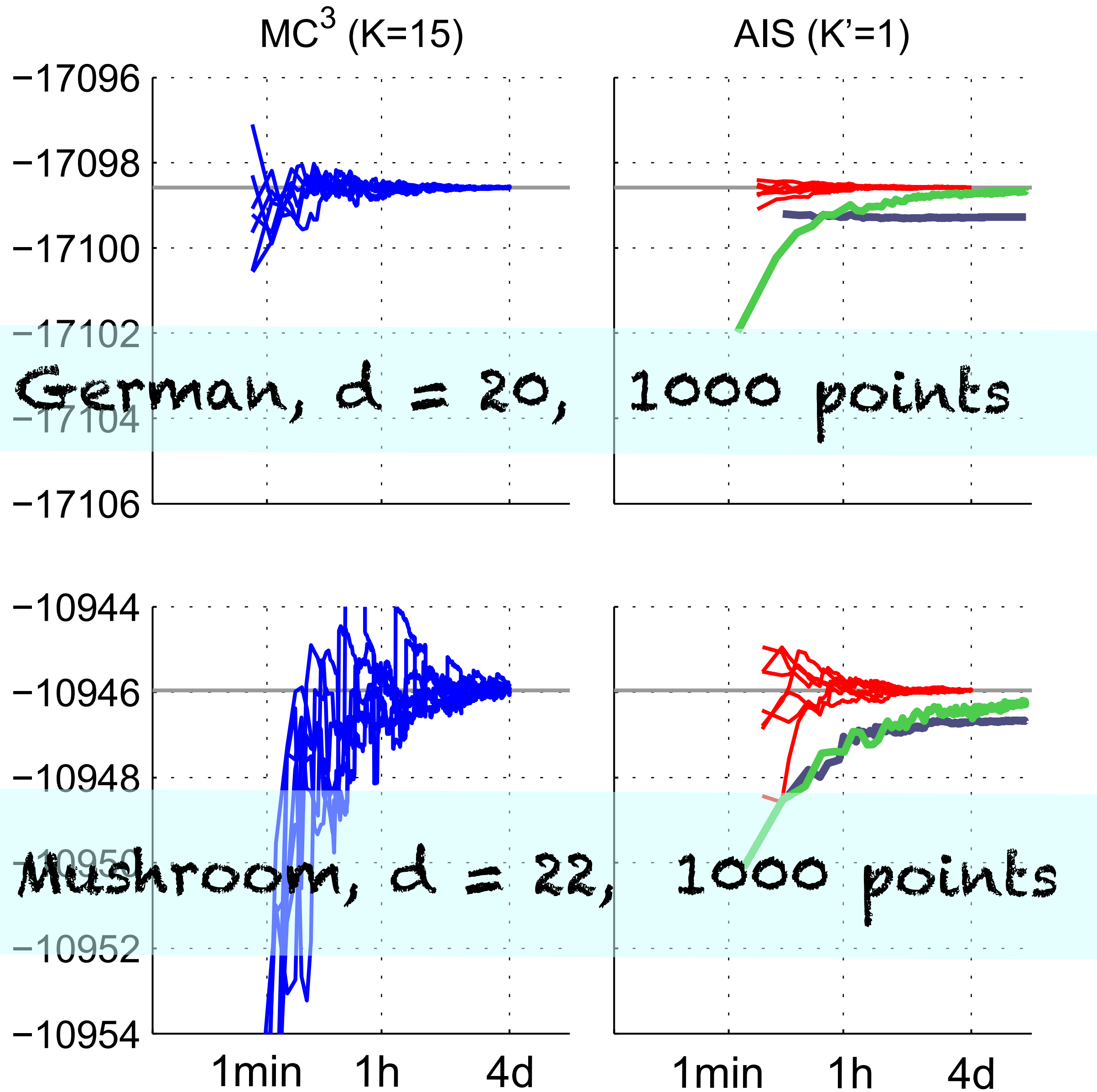
NORMALIZING CONSTANT [Niinimäki et al. 2016]

AIS is accurate and yields good lower bounds

Use B averages, each over T/B samples:

$B = 5$

$B = \sqrt{T}$



SUMMARY

Confident Bayesian learning is hard but desirable

Some recent progress in structure learning

Bidirectional approach: enhance unconfident schemes - expand exact algorithms

For details, see the PhD theses of Pekka P. [2012], Janne K. [2014], Teppo N. [2015], and Kustaa K. [2016]