Variational Scoring of Graphical Model Structures

Matthew J. Beal

Work with
Zoubin Ghahramani
& Carl Rasmussen

ML Meeting, Toronto. 15th September 2003

Overview

Bayesian model selection

Approximations using Variational Bayesian EM

Annealed Importance Sampling

Structure scoring in discrete Directed Acyclic Graphs

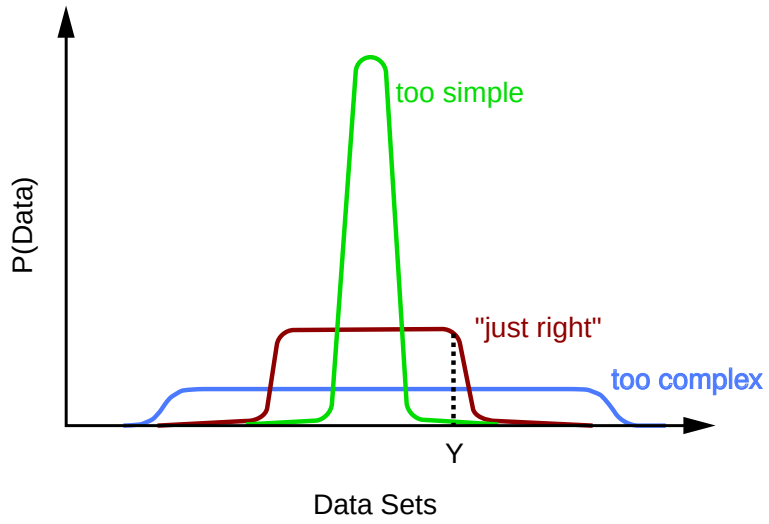
Cheeseman-Stutz vs. Variational Bayes (if time)

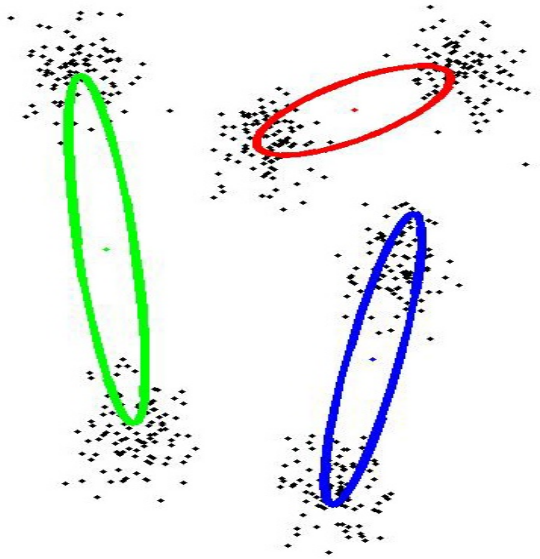
Bayesian Model Selection

Select the model class m_j with the highest probability given the data $\mathbf{y}$:

$$p(m_j|\mathbf{y}) = \frac{p(m_j)p(\mathbf{y}|m_j)}{p(\mathbf{y})}, \qquad p(\mathbf{y}|m_j) = \int d\boldsymbol{\theta}_j \, p(\boldsymbol{\theta}_j|m_j)p(\mathbf{y}|\boldsymbol{\theta}_j, m_j)$$

Interpretation: The probability that *randomly selected* parameter values from the model class would generate data set $\mathbf{y}$.





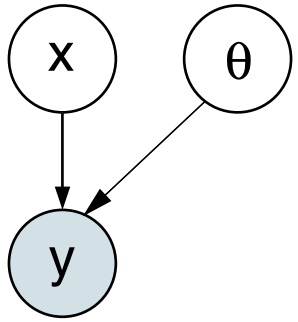
Model classes that are too simple are unlikely to generate the data set.

Model classes that are too complex can generate many possible data sets, so again, they are unlikely to generate that particular data set at random.

Intractable Likelihoods

The marginal likelihood is often a difficult integral

$$p(\mathbf{y}|m) = \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta}|m)p(\mathbf{y}|\boldsymbol{\theta})$$



- because of the high dimensionality of the parameter space
- analytical intractability
- and also due to the presence of hidden variables:

$$\begin{aligned} p(\mathbf{y}|m) &= \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta}|m)p(\mathbf{y}|\boldsymbol{\theta}) \\ &= \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta}|m) \int d\mathbf{x}\; p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}, m) \end{aligned}$$

Practical Bayesian methods

- Laplace approximations:
  - Appeal to Central Limit Theorem, making a Gaussian approximation about the maximum *a posteriori* parameter estimate, $\hat{\boldsymbol{\theta}}$.

$$\ln p(\mathbf{y}|m) \approx \ln p(\hat{\boldsymbol{\theta}} \,|\, m) + \ln p(\mathbf{y} \,|\, \hat{\boldsymbol{\theta}}) + \tfrac{d}{2} \ln 2\pi - \tfrac{1}{2} \ln |H|$$

- Large sample approximations:
  - e.g. BIC: as $n \to \infty$, $\ln p(\mathbf{y}|m) \approx \ln p(\mathbf{y} \,|\, \hat{\boldsymbol{\theta}}) - \frac{d}{2} \ln n$

- Markov chain Monte Carlo (MCMC):
  - Guaranteed to converge in the limit.
  - Many samples required for accurate results.
  - Hard to assess convergence.

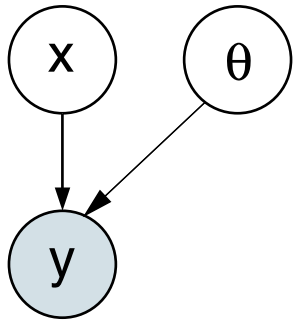
- **Variational approximations ...** *this changes the cost function*

Other deterministic approximations are also available now: e.g. Bethe approximations (Yedidia, Freeman & Weiss, 2000) and Expectation Propagation (Minka, 2001).

Lower Bounding the Marginal Likelihood

Variational Bayesian Learning

Let the hidden states be $\mathbf{x}$, data $\mathbf{y}$ and the parameters $\boldsymbol{\theta}$.



We can lower bound the marginal likelihood (Jensen's inequality):

$$\begin{aligned}
\ln p(\mathbf{y}|m) &= \ln \int d\mathbf{x}\, d\boldsymbol{\theta}\; p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta} \,|\, m) \\
&= \ln \int d\mathbf{x}\, d\boldsymbol{\theta}\; q(\mathbf{x}, \boldsymbol{\theta}) \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta} \,|\, m)}{q(\mathbf{x}, \boldsymbol{\theta})} \\
&\geq \int d\mathbf{x}\, d\boldsymbol{\theta}\; q(\mathbf{x}, \boldsymbol{\theta}) \ln \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta} \,|\, m)}{q(\mathbf{x}, \boldsymbol{\theta})}.
\end{aligned}$$

Use a simpler, factorised approximation to $q(\mathbf{x}, \boldsymbol{\theta})$:

$$\begin{aligned}
\ln p(\mathbf{y} \,|\, m) &\geq \int d\mathbf{x}\, d\boldsymbol{\theta}\; q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta} \,|\, m)}{q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \\
&= \mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{y}).
\end{aligned}$$

Updating the variational approximation . . .

Maximizing this lower bound, $\mathcal{F}_m$, leads to **EM-like** updates:

$$q^*_{\mathbf{x}}(\mathbf{x}) \quad \propto \quad \exp\left[\int d\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln p(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\theta})\right] \qquad E-like\ \ step$$

$$q^*_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \quad \propto \quad p(\boldsymbol{\theta}) \exp\left[\int d\mathbf{x}\; q_{\mathbf{x}}(\mathbf{x}) \ln p(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\theta})\right] \qquad M-like\ \ step$$

Maximizing $\mathcal{F}_m$ is equivalent to minimizing KL-divergence between the *approximate posterior*, $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\; q_{\mathbf{x}}(\mathbf{x})$ and the *true posterior*, $p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y}, m)$:

$$\underbrace{\ln p(\mathbf{y}|m)}_{\substack{\text{desired} \\ \text{quantity}}} - \underbrace{\mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{y})}_{\text{computable}} = \underbrace{\int d\mathbf{x}\, d\boldsymbol{\theta}\; q_{\mathbf{x}}(\mathbf{x})\; q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{q_{\mathbf{x}}(\mathbf{x})\, q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{p(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y}, m)} = \mathbf{KL}(q\|p)}_{\text{measure of inaccuracy of approximation}}$$

In the limit as $n \to \infty$, for identifiable models, the variational lower bound approaches Schwartz's (1978) BIC criterion.

Conjugate-Exponential models

Let's focus on *conjugate-exponential* (**CE**) models, which satisfy **(1)** and **(2)**:

Condition (1). The joint probability over *variables* is in the exponential family:

$$p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{x}, \mathbf{y})\, g(\boldsymbol{\theta}) \exp\left\{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \mathbf{u}(\mathbf{x}, \mathbf{y})\right\}$$

where $\boldsymbol{\phi}(\boldsymbol{\theta})$ is the vector of *natural parameters*, $\mathbf{u}$ are *sufficient statistics*

Condition (2). The prior over *parameters* is conjugate to this joint probability:

$$p(\boldsymbol{\theta}|\eta, \boldsymbol{\nu}) = h(\eta, \boldsymbol{\nu})\, g(\boldsymbol{\theta})^\eta \exp\left\{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \boldsymbol{\nu}\right\}$$

where η and $\boldsymbol{\nu}$ are hyperparameters of the prior.

Conjugate priors are computationally convenient and have an intuitive interpretation:

- η: number of pseudo-observations
- $\boldsymbol{\nu}$: values of pseudo-observations

Conjugate-Exponential examples

In the **CE** family:

- Gaussian mixtures
- factor analysis, probabilistic PCA
- hidden Markov models and factorial HMMs
- linear dynamical systems and switching models
- discrete-variable belief networks

Other as yet undreamt-of models can combine Gaussian, Gamma, Poisson, Dirichlet, Wishart, Multinomial and others.

Not in the **CE** family:

- Boltzmann machines, MRFs (no conjugacy)
- logistic regression (no conjugacy)
- sigmoid belief networks (not exponential)
- independent components analysis (not exponential)

One can often approximate these models with models in the **CE** family e.g. IFA (Attias, 1998).

A very useful result

Theorem Given an iid data set $\mathbf{y} = (\mathbf{y}_1, \ldots \mathbf{y}_n)$, if the model is **CE** then:

(a) $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is also conjugate, *i.e.*

$$q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = h(\tilde{\eta}, \tilde{\boldsymbol{\nu}}) g(\boldsymbol{\theta})^{\tilde{\eta}} \exp\left\{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \tilde{\boldsymbol{\nu}}\right\}$$

(b) $q_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^n q_{\mathbf{x}_i}(\mathbf{x}_i)$ is of the same form as in the E step of regular EM, but using pseudo parameters computed by averaging over $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$

$$q_{\mathbf{x}_i}(\mathbf{x}_i) \propto f(\mathbf{x}_i, \mathbf{y}_i) \exp\left\{\overline{\boldsymbol{\phi}}^\top \mathbf{u}(\mathbf{x}_i, \mathbf{y}_i)\right\} = p(\mathbf{x}_i | \mathbf{y}_i, \tilde{\boldsymbol{\theta}})$$

where $\overline{\boldsymbol{\phi}} = \langle \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle_{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \stackrel{?}{=} \boldsymbol{\phi}(\tilde{\boldsymbol{\theta}})$

KEY points:
(a) the approximate parameter posterior is of the same form as the prior;
(b) the approximate hidden variable posterior, *averaging over all parameters*, is of the same form as the *exact* hidden variable posterior under $\tilde{\boldsymbol{\theta}}$.

The Variational Bayesian EM algorithm

EM for MAP estimation

Goal: maximize $p(\boldsymbol{\theta}|\mathbf{y}, m)$ w.r.t. $\boldsymbol{\theta}$

E Step: compute

$$q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}^{(t)})$$

M Step:

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \int d\mathbf{x}\; q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \ln p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})$$

Variational Bayesian EM

Goal: lower bound $p(\mathbf{y}|m)$

VB-E Step: compute

$$q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}, \bar{\boldsymbol{\phi}}^{(t)})$$

VB-M Step:

$$q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}) \propto \exp\left[\int d\mathbf{x}\; q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \ln p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})\right]$$

Properties:

- Reduces to the EM algorithm if $q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$.
- $\mathcal{F}_m$ increases monotonically, and incorporates the model complexity penalty.
- Analytical parameter distributions (but not constrained to be Gaussian).
- VB-E step has same complexity as corresponding E step.
- We can use the junction tree, belief propagation, Kalman filter, etc, algorithms in the VB-E step of VB-EM, but **using expected natural parameters**, $\bar{\boldsymbol{\phi}}$.

Variational Bayesian EM

The Variational Bayesian EM algorithm has been used to approximate Bayesian learning in a wide range of models, such as:

- probabilistic PCA and factor analysis (Bishop, 1999)
- mixtures of Gaussians (Attias, 1999)
- mixtures of factor analysers (Ghahramani & Beal, 1999)
- state-space models (Ghahramani & Beal, 2000; Beal, 2003)
- ICA, IFA (Attias, 1999; Miskin & MacKay, 2000; Valpola 2000)
- mixtures of experts (Ueda & Ghahramani, 2000)
- hidden Markov models (MacKay, 1995; Beal, 2003)

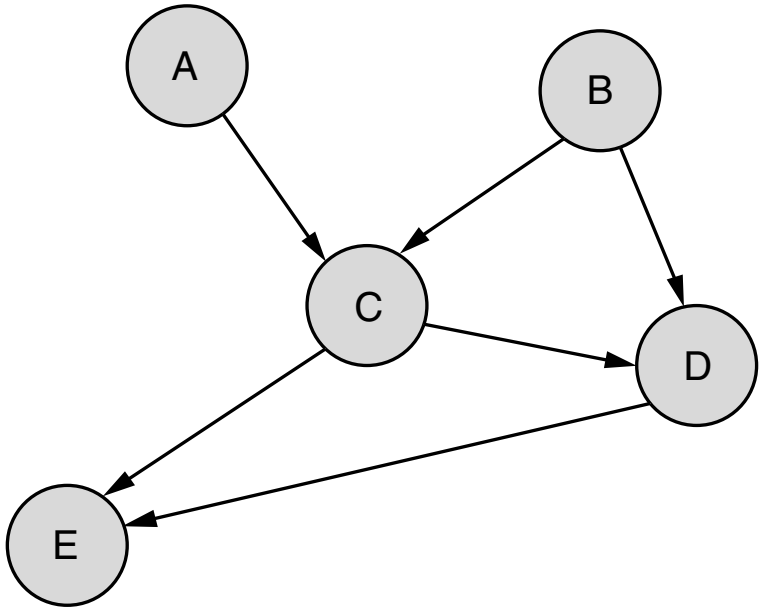
The main advantage is that it can be used to **automatically do model selection** and does not suffer from overfitting to the same extent as ML methods do.

Also it is about as computationally demanding as the usual EM algorithm.

See: www.variational-bayes.org

Graphical Models

(Bayesian Networks / Directed Acyclic Graphical Models)



A Bayesian network corresponds to a factorization of the joint distribution:

$$p(A,B,C,D,E) = p(A)p(B)p(C|A,B) \\ p(D|B,C)p(E|C,D)$$

In general:
$$p(X_1,\ldots,X_n) = \prod_{i=1}^{n} p(X_i|X_{\mathsf{pa}(i)})$$

where $\mathsf{pa}(i)$ are the parents of node i.

Semantics: $X \perp\!\!\!\perp Y|V$ if V d-separates X from Y.

Two advantages: interpretability and efficiency.

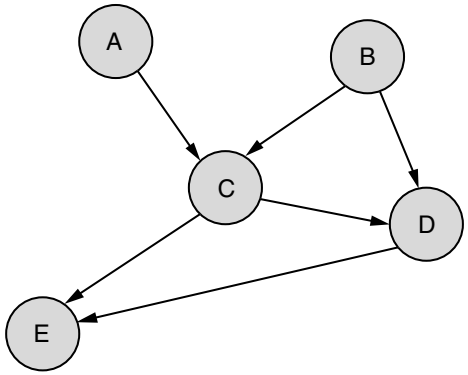
Model Selection Task

Which of the following graphical models is the data generating process?

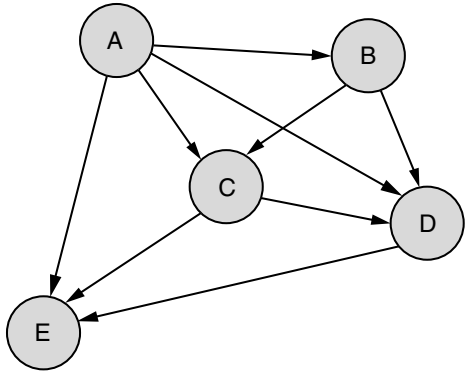
Discrete directed acyclic graphical models: data $\mathbf{y} = (A, B, C, D, E)^n$

ALL OBSERVED

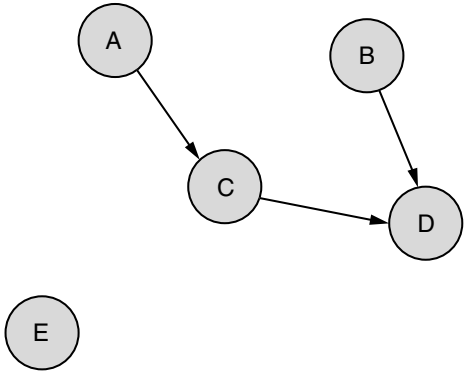
true



complex



simple

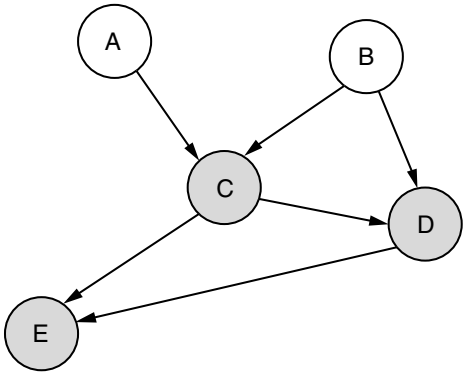


marginal likelihood tractable

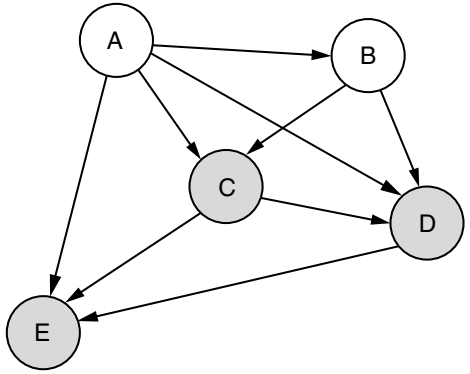
If the data are just $\mathbf{y} = (C, D, E)^n$, and $\mathbf{s} = (A, B)^n$ are **hidden** variables... ?

OBS.+HIDDEN

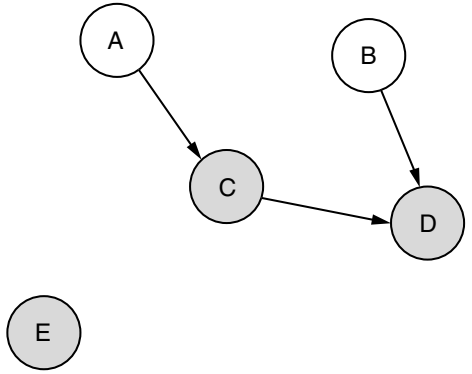
true



complex



simple



marginal likelihood intractable

Case study: Discrete DAGs

Let the hidden and observed variables be denoted with $\mathbf{z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\} = \{\mathbf{s}_1, \mathbf{y}_1, \ldots, \mathbf{s}_n, \mathbf{y}_n\}$, of which some $j \in \mathcal{H}$ are hidden and $j \in \mathcal{V}$ are observed variables, i.e. $\mathbf{s}_i = \{\mathbf{z}_{ij}\}_{j \in \mathcal{H}}$ and $\mathbf{y}_i = \{\mathbf{z}_{ij}\}_{j \in \mathcal{V}}$.

Complete-data likelihood

$$p(\mathbf{z} \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{j=1}^{|\mathbf{z}_i|} p(\mathbf{z}_{ij} \,|\, \mathbf{z}_{i\mathbf{pa}(j)}, \boldsymbol{\theta}) \ ,$$

Complete-data marginal likelihood

$$p(\mathbf{z} \,|\, m) = \int d\boldsymbol{\theta} \; p(\boldsymbol{\theta} \,|\, m) \prod_{i=1}^{n} \prod_{j=1}^{|\mathbf{z}_i|} p(\mathbf{z}_{ij} \,|\, \mathbf{z}_{i\mathbf{pa}(j)}, \boldsymbol{\theta})$$

Incomplete-data likelihood

$$p(\mathbf{y} \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{y}_i \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{\{\mathbf{z}_{ij}\}_{j \in \mathcal{H}}} \prod_{j=1}^{|\mathbf{z}_i|} p(\mathbf{z}_{ij} \,|\, \mathbf{z}_{i\mathbf{pa}(j)}, \boldsymbol{\theta})$$

Incomplete-data marginal likelihood

$$p(\mathbf{y} \,|\, m) = \int d\boldsymbol{\theta} \; p(\boldsymbol{\theta} \,|\, m) \prod_{i=1}^{n} \sum_{\{\mathbf{z}_{ij}\}_{j \in \mathcal{H}}} \prod_{j=1}^{|\mathbf{z}_i|} p(\mathbf{z}_{ij} \,|\, \mathbf{z}_{i\mathbf{pa}(j)}, \boldsymbol{\theta}) \qquad \text{(intractable!)}$$

The BIC and CS approximations

BIC - Bayesian Information Criterion

$$p(\mathbf{y}|m) = \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta}\,|\,m)p(\mathbf{y}\,|\,\boldsymbol{\theta})\;,$$

$$\ln p(\mathbf{y}|m) \approx \ln p(\mathbf{y}|m)_{\mathsf{BIC}} = \underbrace{\ln p(\mathbf{y}\,|\,\hat{\boldsymbol{\theta}})}_{\substack{\text{use EM}\\ \text{to find } \hat{\boldsymbol{\theta}}}} - \frac{d}{2}\ln n\;.$$

CS - Cheeseman-Stutz criterion

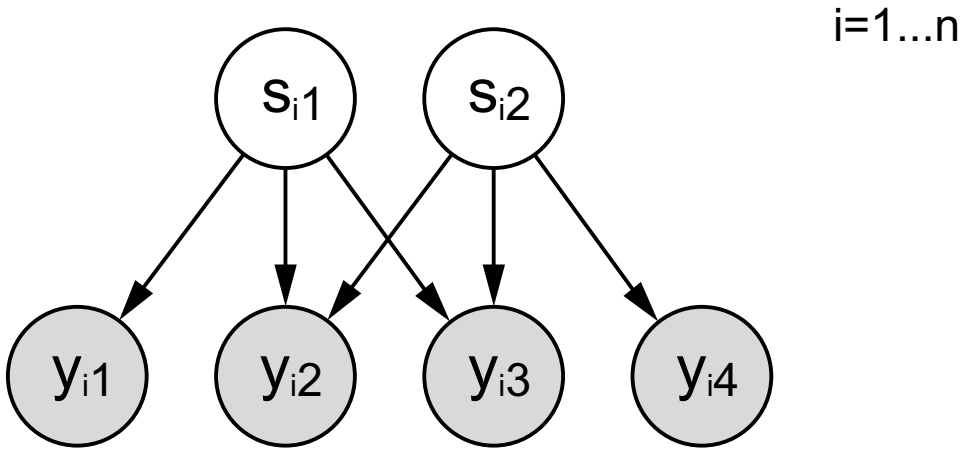
$$p(\mathbf{y}\,|\,m) = p(\mathbf{z}\,|\,m)\frac{p(\mathbf{y}\,|\,m)}{p(\mathbf{z}\,|\,m)} = p(\mathbf{s},\mathbf{y}\,|\,m)\frac{\int d\boldsymbol{\theta}\; p(\boldsymbol{\theta}\,|\,m)p(\mathbf{y}\,|\,\boldsymbol{\theta})}{\int d\boldsymbol{\theta}'\; p(\boldsymbol{\theta}'\,|\,m)p(\mathbf{s},\mathbf{y}\,|\,\boldsymbol{\theta}')}\;, \qquad (*)$$

$$\ln p(\mathbf{y}|m) \approx \ln p(\mathbf{y}\,|\,m)_{\mathsf{CS}} = \ln p(\hat{\mathbf{s}},\mathbf{y}\,|\,m) + \ln p(\mathbf{y}\,|\,\hat{\boldsymbol{\theta}}) - \ln p(\hat{\mathbf{s}},\mathbf{y}\,|\,\hat{\boldsymbol{\theta}})\;.$$

$(*)$ is correct for *any* completion $\mathbf{s}$ of the hidden variables, so what completion $\hat{\mathbf{s}}$ should we use? [CS uses result of E-step from the EM algorithm.]

Experiments

- Bipartite structure: only hidden variables can be parents of observed variables. **Two** binary hidden variables, and **four** five-valued discrete observed variables.



- Conjugate prior is Dirichlet, Conjugate-Exponential model, so the VB-EM algorithm is a straightforward modification of EM.

- **Experiment:** There are 136 distinct structures (out of 256) with 2 latent variables as potential parents of 4 conditionally independent observed vars.

- **Score** each structure for twenty varying size data sets:

$n \in \{10, 20, 40, 80, 110, 160, 230, 320, 400, 430, \underline{480}, 560, 640, 800, 960, 1120, 1280, 2560, 5120, 10240\}$

using 4 methods: BIC, CS, VB, and a **gold standard** AIS

- 2720 graphs to score, times for each: BIC (1.5s), CS (1.5s), VB (4s), AIS (400s).

Annealed Importance Sampling (AIS)

AIS is a state-of-the-art method for estimating marginal likelihoods, by breaking a difficult integral into a series of easier ones.

Combines ideas from importance sampling, Markov chain Monte Carlo, & annealing.

$$\text{Define} \qquad \mathcal{Z}_k = \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta}\,|\,m) p(\mathbf{y}\,|\,\boldsymbol{\theta})^{\tau(k)} = \int d\boldsymbol{\theta}\; f_k(\boldsymbol{\theta})$$

with $\tau(0) = 0 \implies \mathcal{Z}_0 = \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta}\,|\,m) = 1$ ← normalisation of prior,

and $\tau(K) = 1 \implies \mathcal{Z}_K = p(\mathbf{y}\,|\,m)$ ← marginal likelihood.

$$\text{Schedule: } \{\tau(k)\}_{k=1}^K, \qquad \frac{\mathcal{Z}_K}{\mathcal{Z}_0} \equiv \frac{\mathcal{Z}_1}{\mathcal{Z}_0}\,\frac{\mathcal{Z}_2}{\mathcal{Z}_1}\,\cdots\,\frac{\mathcal{Z}_K}{\mathcal{Z}_{K-1}}\,.$$

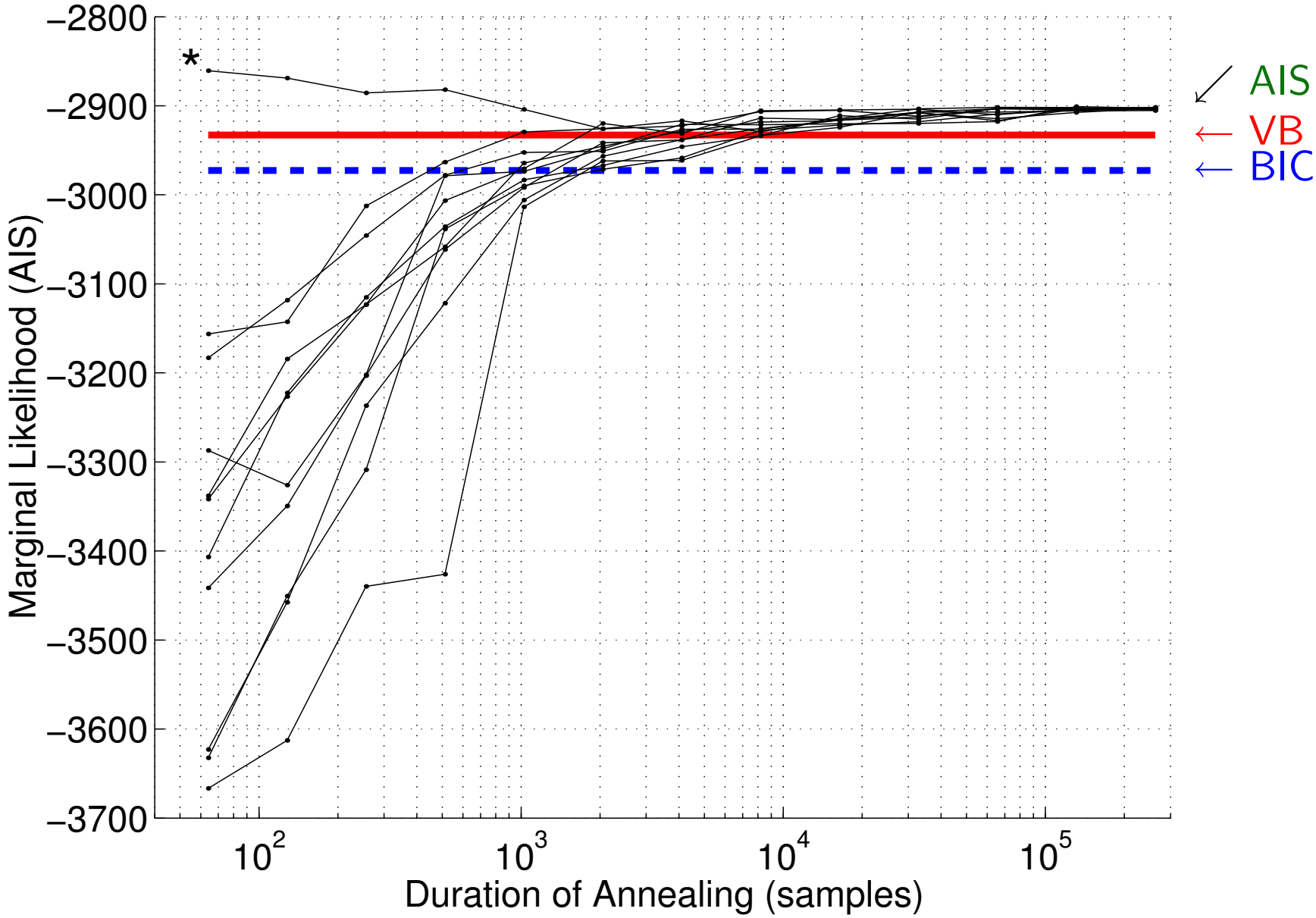
Importance sample from $f_{k-1}(\boldsymbol{\theta})$ as follows: with $\boldsymbol{\theta}^{(r)} \sim f_{k-1}(\boldsymbol{\theta})$,

$$\frac{\mathcal{Z}_k}{\mathcal{Z}_{k-1}} \equiv \int d\boldsymbol{\theta}\; \frac{f_k(\boldsymbol{\theta})}{f_{k-1}(\boldsymbol{\theta})}\frac{f_{k-1}(\boldsymbol{\theta})}{\mathcal{Z}_{k-1}} \approx \frac{1}{R}\sum_{r=1}^{R} \frac{f_k(\boldsymbol{\theta}^{(r)})}{f_{k-1}(\boldsymbol{\theta}^{(r)})} = \frac{1}{R}\sum_{r=1}^{R} p(\mathbf{y}\,|\,\boldsymbol{\theta}^{(r)}, m)^{\tau(k)-\tau(k-1)}$$

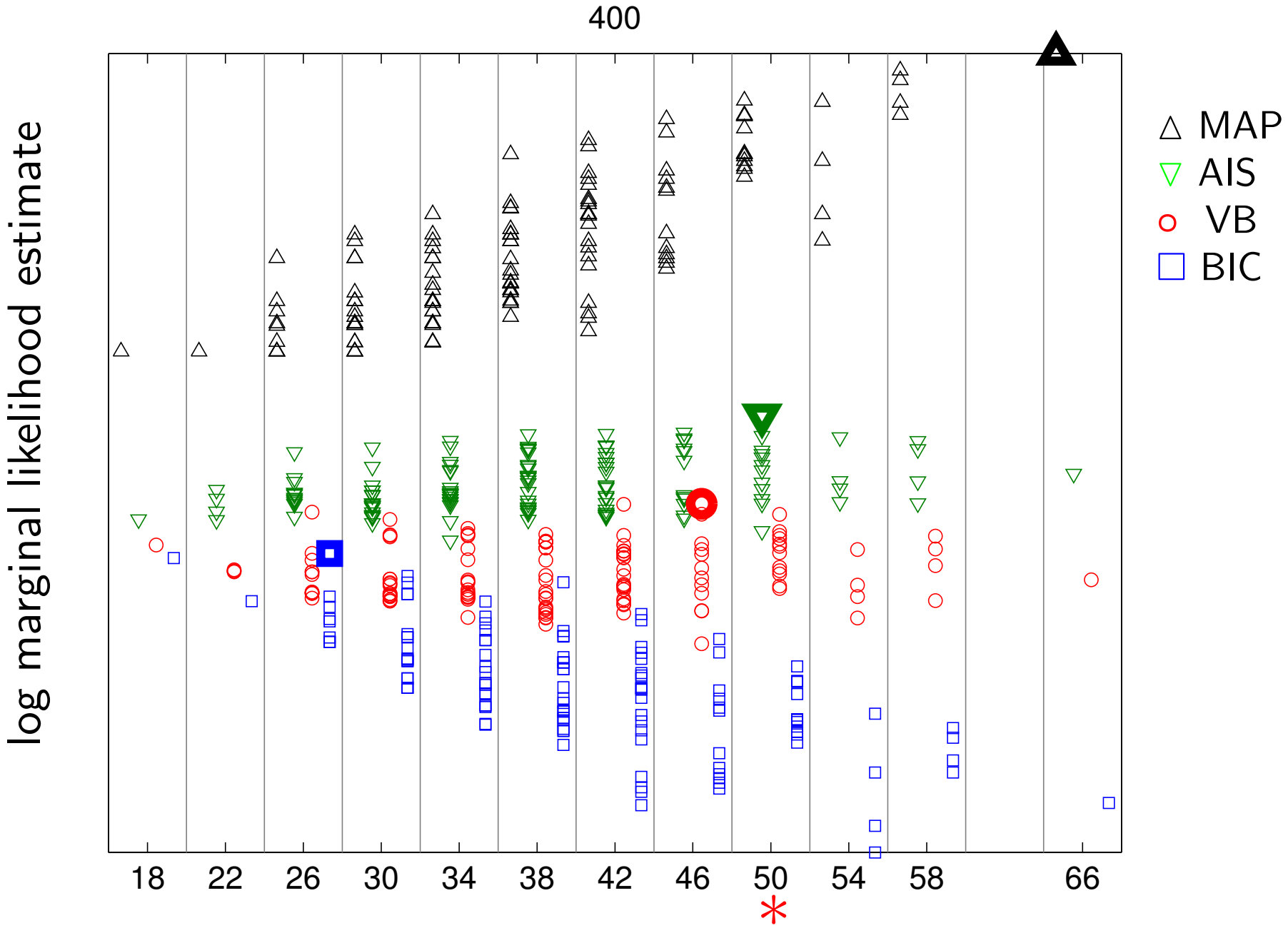
- How reliable is AIS? How tight are the variational bounds?

How reliable is the AIS for this problem?

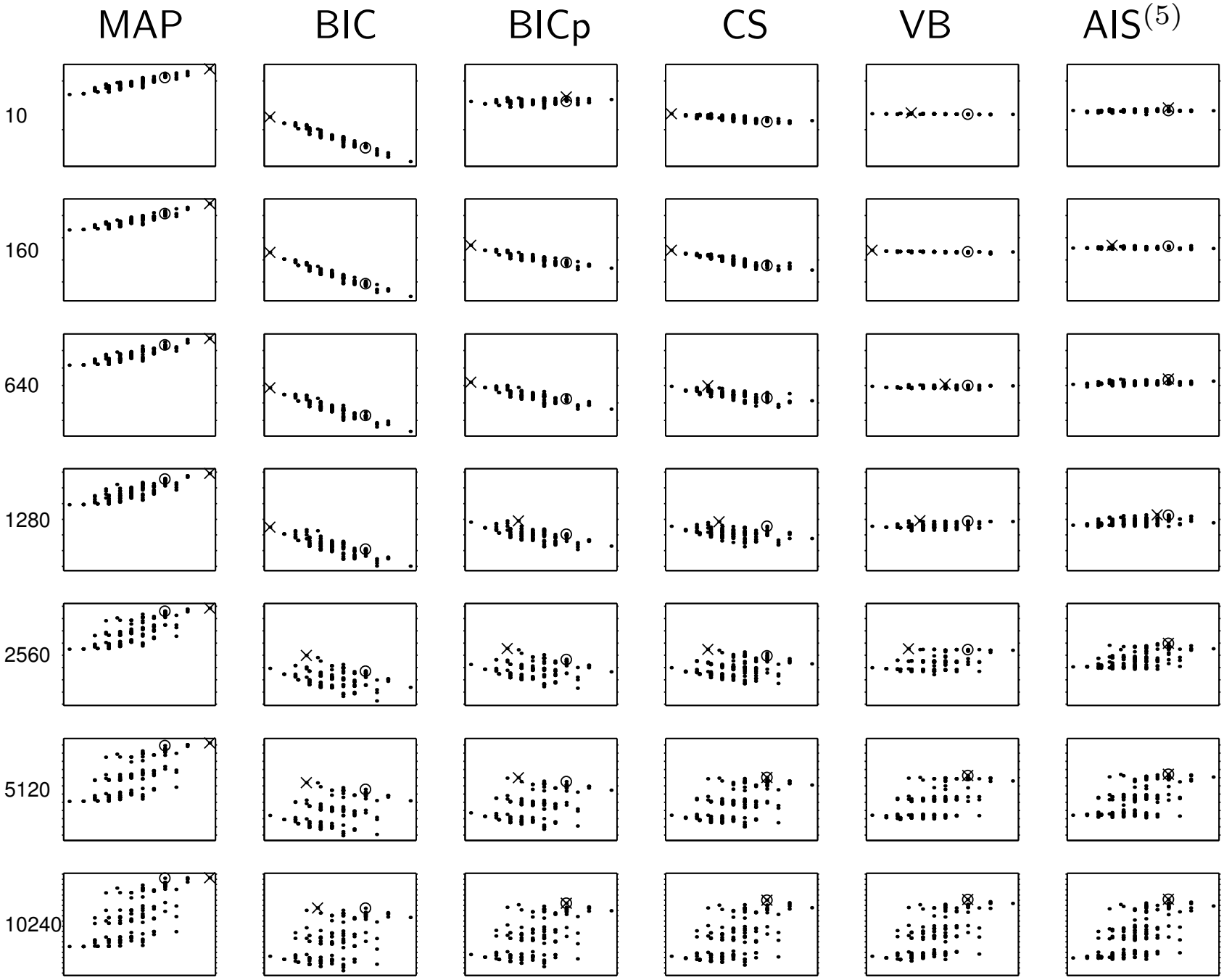
Varying the annealing schedule with random initialisation. $n = 480, K = 2^6 \ldots 2^{18}$



Scoring all structures by every method



Every method scoring every structure

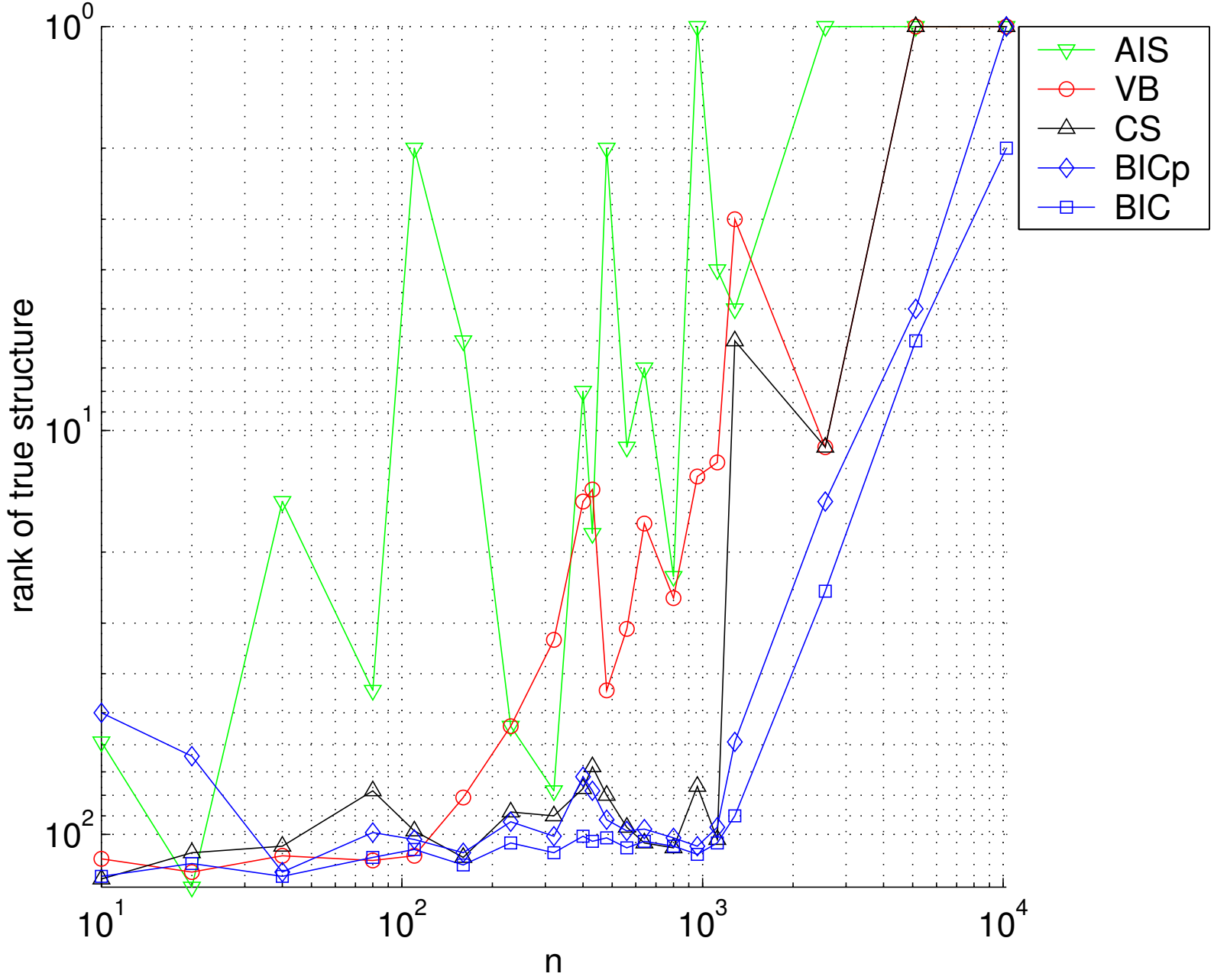


Ranking of the true structure by each method

n	MAP	BIC	BICp	CS	VB	AIS$^{(5)}$
10	21	127	50	129	115	59
20	12	118	64	111	124	135
40	28	127	124	107	113	15
80	8	114	99	78	116	44
110	8	109	103	98	113	2
160	13	119	111	114	81	6
230	8	105	93	88	54	54
320	8	111	101	90	33	78
400	6	101	72	77	15	8
430	7	104	78	68	14	18
480	7	102	92	80	44	2
560	9	108	98	96	31	11
640	7	104	97	105	17	7
800	9	107	102	108	26	23
960	13	112	107	76	13	1
1120	8	105	96	103	12	4
1280	7	90	59	6	3	5
2560	6	25	15	11	11	1
5120	5	6	5	1	1	1
10240	3	2	1	1	1	1

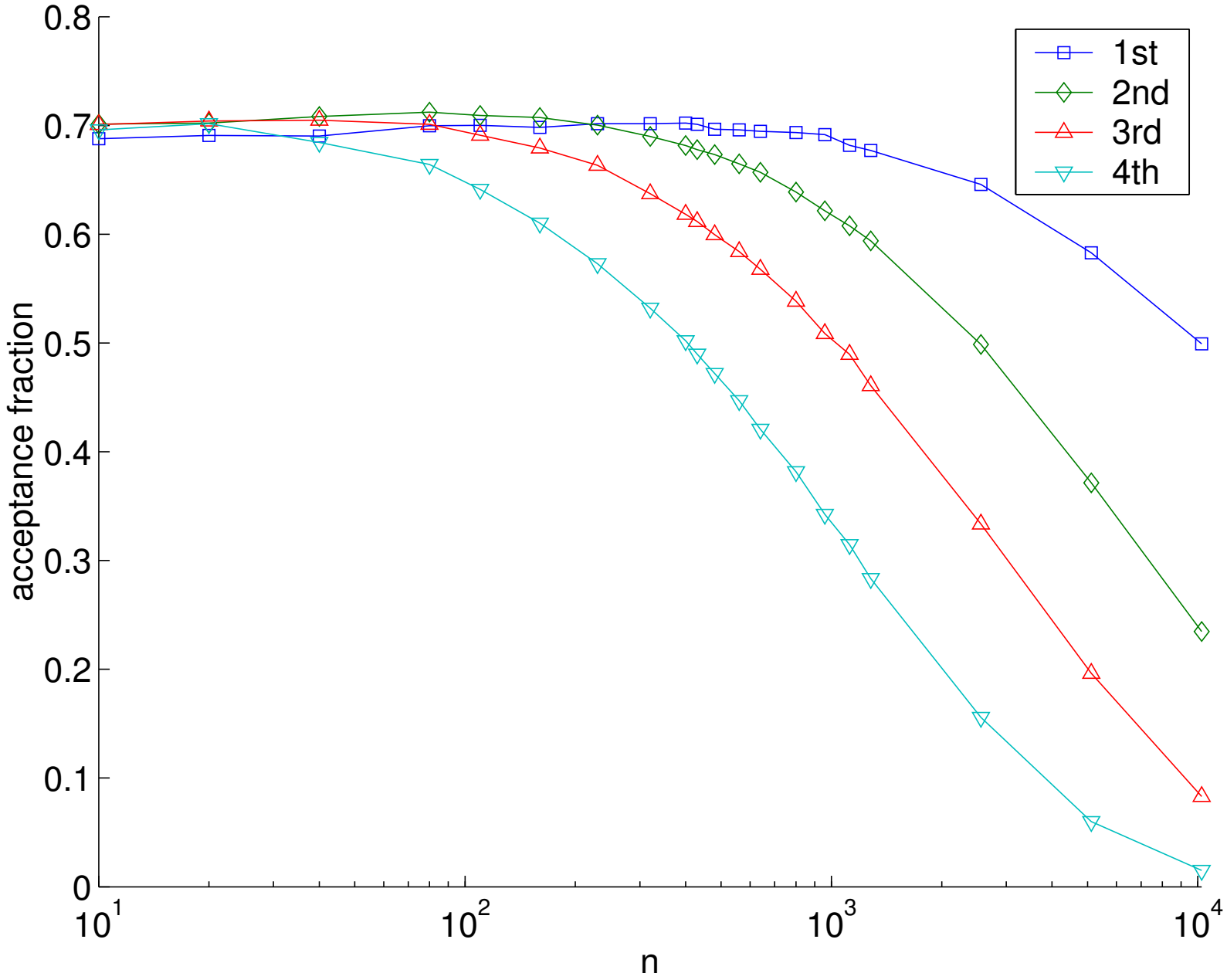
Ranking the true structure

VB score finds correct structure earlier, and more reliably

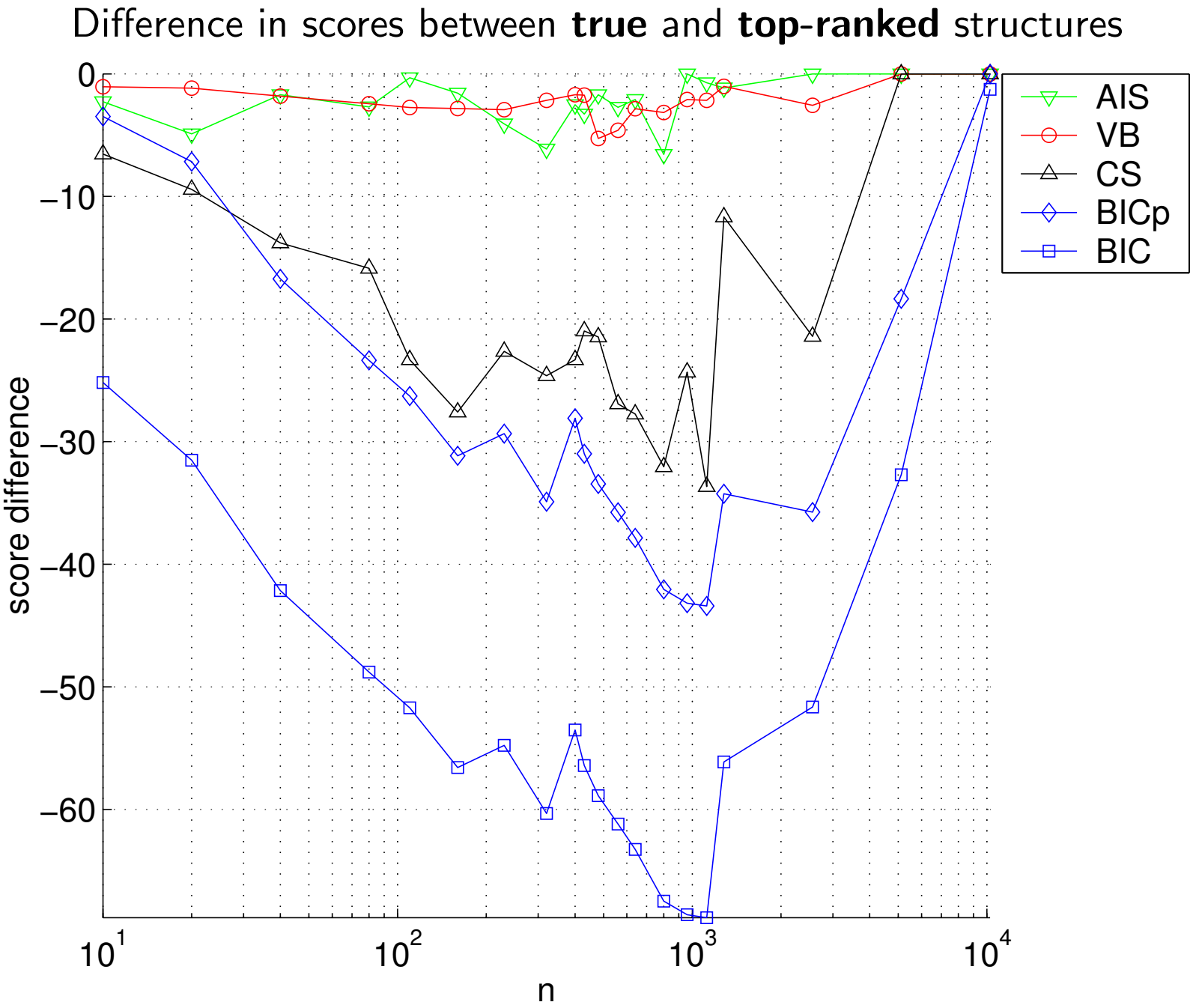


Acceptance rate of the AIS sampler

M-H acceptance fraction, measured over each of four quarters of the annealing schedule

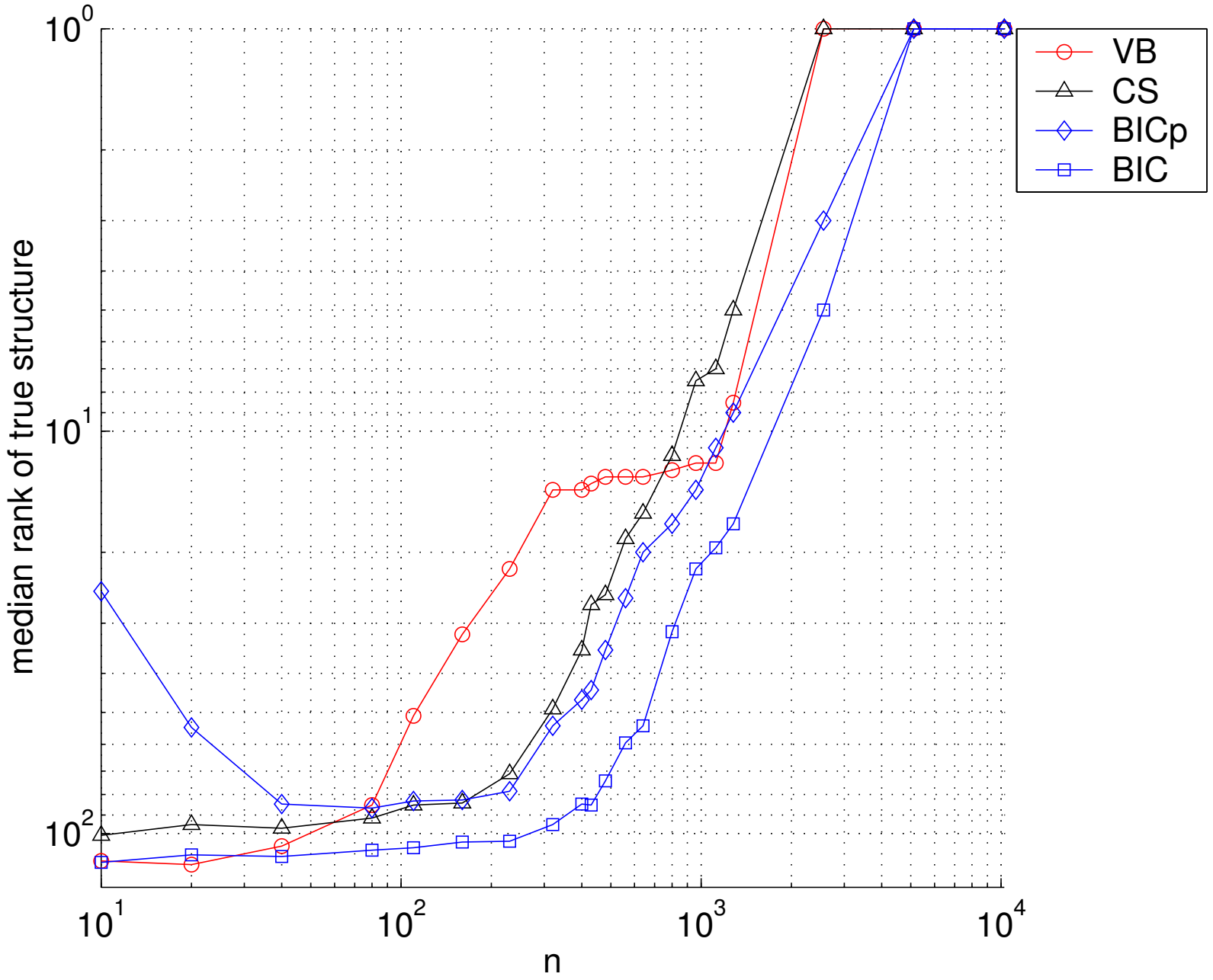


How true are the various scores?



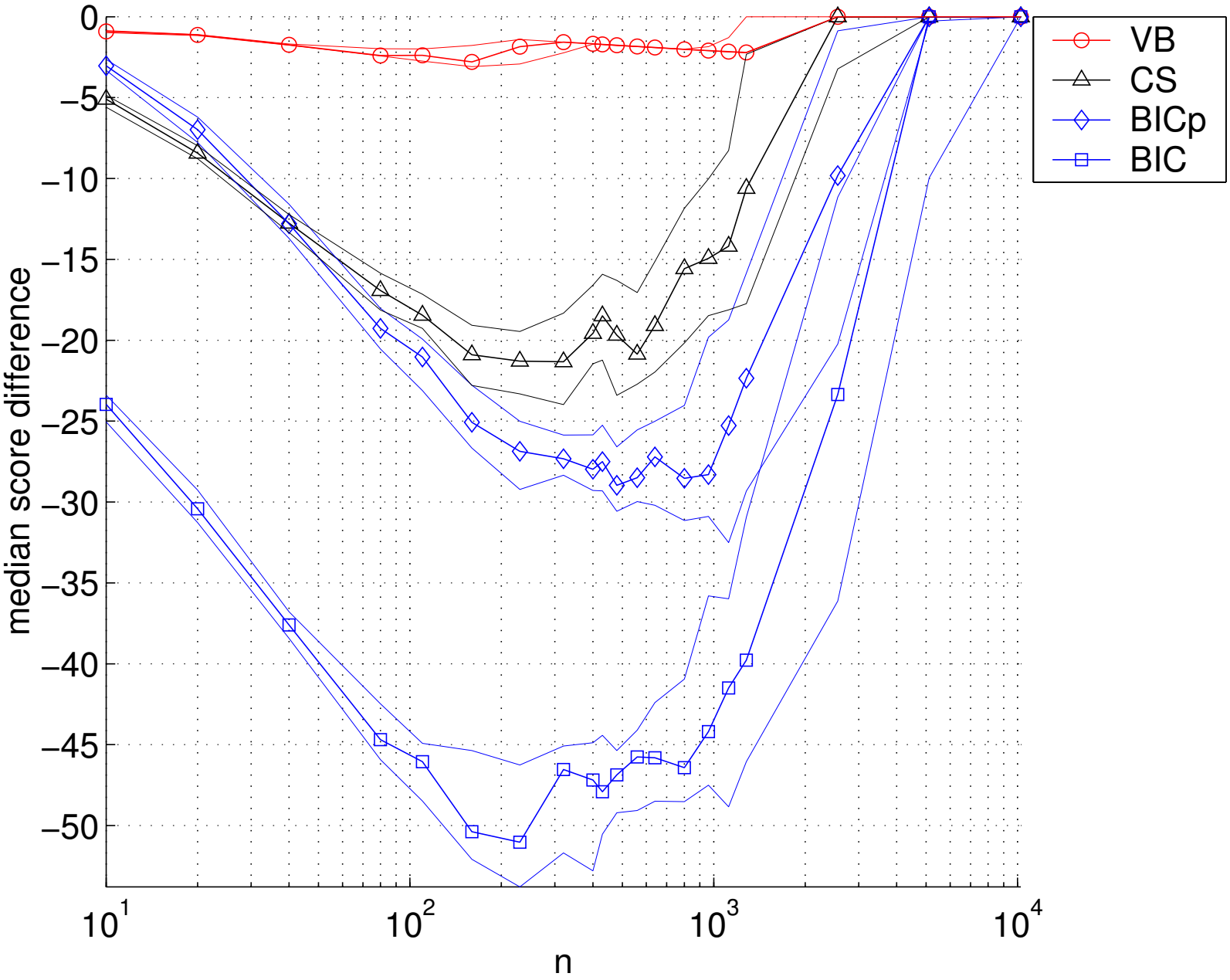
Average Rank of the true structure

Averaged over true structure parameters drawn from the prior (106 instances)



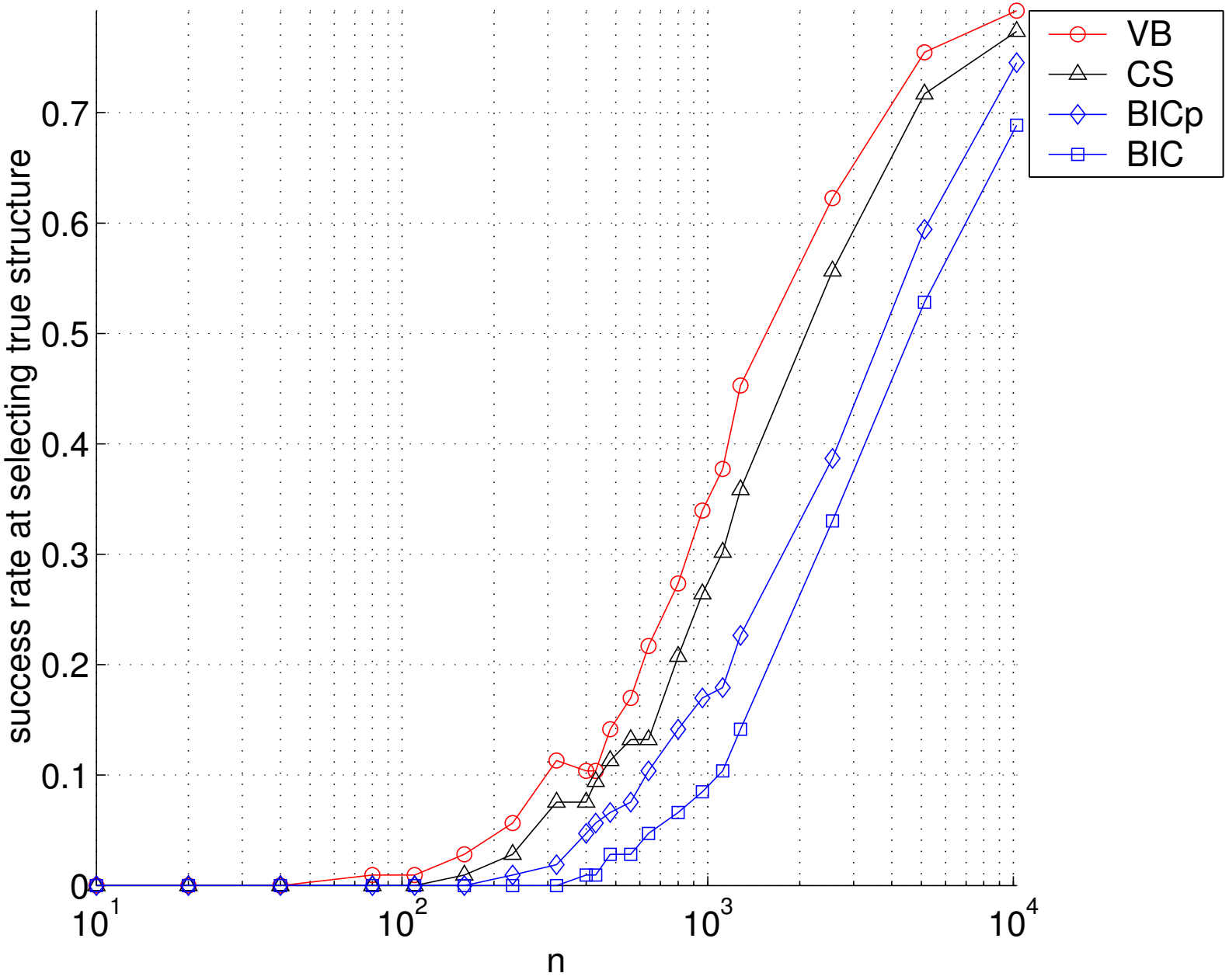
True $\Longleftrightarrow$ Top-ranked score differences

Averaged over true structure parameters drawn from the prior (106 instances)



Overall Success Rate of each method

Averaged over true structure parameters drawn from the prior (106 instances)



Results summary

- VB outperforms BIC on ranking: 20 data sets, and 95 instances.

% times that \ than	BIC*	BICp*	CS*	BIC	BICp	CS
VB ranks worse	16.9	30.2	31.8	15.1	29.6	30.9
same	11.1	15.0	20.2	11.7	15.5	20.9
better	72.0	54.8	48.0	73.2	55.0	48.2

- AIS standard can break down at high n, violating VB strict lower bound when scoring the 136 structures:

n	10 . . . 560	640	800	960	1120	1280	2560	5120	10240
% M-H rej.	$<$40.3	41.5	43.7	45.9	47.7	49.6	59.2	69.7	79.2
single #AIS$^{(1)}<$VB	$\leq$7.5	15.1	9.4	14.2	12.3	20.8	31.1	59.4	74.5
averaged #AIS$^{(5)}<$VB	$\leq$1.9	0.0	0.0	0.0	1.5	2.2	5.1	19.9	52.9

- AIS has many **parameters to tune**: Metropolis-Hastings proposal widths/shapes, annealing schedules (non-linear), # samples, reaching equilibrium...

  VB has **none**!

Cheeseman-Stutz is a lower bound

$$p(\mathbf{y} \mid m)_{\mathsf{CS}} = p(\hat{\mathbf{s}}, \mathbf{y} \mid m) \frac{p(\mathbf{y} \mid \hat{\boldsymbol{\theta}})}{p(\hat{\mathbf{s}}, \mathbf{y} \mid \hat{\boldsymbol{\theta}})} \leq p(\mathbf{y} \mid m) \ .$$

$$p(\mathbf{y} \mid m) = \int d\boldsymbol{\theta} \ p(\boldsymbol{\theta}) \prod_{i=1}^{n} p(\mathbf{y}_i \mid \boldsymbol{\theta}) \geq \int d\boldsymbol{\theta} \ p(\boldsymbol{\theta}) \prod_{i=1}^{n} \exp \left\{ \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln \frac{p(\mathbf{s}_i, \mathbf{y}_i \mid \boldsymbol{\theta})}{q_{\mathbf{s}_i}(\mathbf{s}_i)} \right\} \ .$$

$$p(\mathbf{y} \mid \hat{\boldsymbol{\theta}}) = \prod_{i=1}^{n} p(\mathbf{y}_i \mid \hat{\boldsymbol{\theta}}) = \prod_{i=1}^{n} \exp \left\{ \sum_{\mathbf{s}_i} \hat{q}_{\mathbf{s}_i}(\mathbf{s}_i) \ln \frac{p(\mathbf{s}_i, \mathbf{y}_i \mid \hat{\boldsymbol{\theta}})}{\hat{q}_{\mathbf{s}_i}(\mathbf{s}_i)} \right\} \ .$$

where $\hat{q}_{\mathbf{s}_i}(\mathbf{s}_i) \equiv p(\mathbf{s}_i \mid \mathbf{y}, \hat{\boldsymbol{\theta}}) \ , \qquad \hat{\mathbf{s}}_i : \quad \ln p(\hat{\mathbf{s}}_i, \mathbf{y} \mid \hat{\boldsymbol{\theta}}) = \sum_{\mathbf{s}_i} \hat{q}_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{s}_i, \mathbf{y}_i \mid \boldsymbol{\theta}) \ .$

$$p(\mathbf{y} \mid m) \geq \prod_{i=1}^{n} \exp \left\{ \sum_{\mathbf{s}_i} \hat{q}_{\mathbf{s}_i}(\mathbf{s}_i) \ln \frac{1}{\hat{q}_{\mathbf{s}_i}(\mathbf{s}_i)} \right\} \cdot \int d\boldsymbol{\theta} \ p(\boldsymbol{\theta}) \prod_{i=1}^{n} \exp \left\{ \sum_{\mathbf{s}_i} \hat{q}_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{s}_i, \mathbf{y}_i \mid \boldsymbol{\theta}) \right\}$$

$$= \frac{p(\mathbf{y} \mid \hat{\boldsymbol{\theta}})}{\prod_{i=1}^{n} \exp \left\{ \sum_{\mathbf{s}_i} \hat{q}_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{s}_i, \mathbf{y}_i \mid \hat{\boldsymbol{\theta}}) \right\}} \int d\boldsymbol{\theta} \ p(\boldsymbol{\theta}) \prod_{i=1}^{n} \exp \left\{ \sum_{\mathbf{s}_i} \hat{q}_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{s}_i, \mathbf{y}_i \mid \boldsymbol{\theta}) \right\}$$

$$= \frac{p(\mathbf{y} \mid \hat{\boldsymbol{\theta}})}{\prod_{i=1}^{n} p(\hat{\mathbf{s}}_i, \mathbf{y}_i \mid \hat{\boldsymbol{\theta}})} \int d\boldsymbol{\theta} \ p(\boldsymbol{\theta}) \prod_{i=1}^{n} p(\hat{\mathbf{s}}_i, \mathbf{y}_i \mid \boldsymbol{\theta}) \ .$$

VB can be made universally tighter than CS

$$\ln p(\mathbf{y} \,|\, m)_{\mathsf{CS}} \leq \mathcal{F}_m(q_{\mathbf{s}}(\mathbf{s}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) \leq \ln p(\mathbf{y} \,|\, m) \ .$$

Let's approach this result by considering the following forms for $q_{\mathbf{s}}(\mathbf{s})$ and $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$:

$$q_{\mathbf{s}}(\mathbf{s}) = \prod_{i=1}^{n} q_{\mathbf{s}_i}(\mathbf{s}_i) \ , \quad \text{with} \quad q_{\mathbf{s}_i}(\mathbf{s}_i) = p(\mathbf{s}_i \,|\, \mathbf{y}_i, \hat{\boldsymbol{\theta}}) \ ,$$

$$q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \propto \left\langle \ln p(\boldsymbol{\theta}) p(\mathbf{s}, \mathbf{y} \,|\, \boldsymbol{\theta}) \right\rangle_{q_{\mathbf{s}}(\mathbf{s})} \ .$$

We write the form for $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ explicitly:

$$q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{p(\boldsymbol{\theta}) \prod_{i=1}^{n} \exp\left\{ \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{s}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}) \right\}}{\int d\boldsymbol{\theta}' \ p(\boldsymbol{\theta}') \prod_{i=1}^{n} \exp\left\{ \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{s}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}') \right\}} \ ,$$

and then substitute $q_{\mathbf{s}}(\mathbf{s})$ and $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ into the variational lower bound $\mathcal{F}_m$.

$$\mathcal{F}_m(q_{\mathbf{s}}(\mathbf{s}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) = \int d\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \sum_{i=1}^{n} \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln \frac{p(\mathbf{s}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta})}{q_{\mathbf{s}_i}(\mathbf{s}_i)} + \int d\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{\theta})}{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}$$

$$= \int d\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \sum_{i=1}^{n} \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln \frac{1}{q_{\mathbf{s}_i}(\mathbf{s}_i)}$$

$$+ \int d\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \int d\boldsymbol{\theta}'\; p(\boldsymbol{\theta}') \prod_{i=1}^{n} \exp \left\{ \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{s}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}') \right\}$$

$$= \sum_{i=1}^{n} \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln \frac{1}{q_{\mathbf{s}_i}(\mathbf{s}_i)} + \ln \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta}) \prod_{i=1}^{n} \exp \left\{ \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{s}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}) \right\} \quad .$$

With this choice of $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and $q_{\mathbf{s}}(\mathbf{s})$ we acheive equality between the CS and VB approximations.

We complete the proof by noting that at the very next step of VBEM (VBE-step) is guaranteed to increase or leave unchanged $\mathcal{F}_m$, and hence surpass the CS bound.

Summary & Conclusions

- Bayesian learning avoids overfitting and can be used to do model selection.
- Variational Bayesian EM for CE models and propagation algorithms.
- These methods have advantages over MCMC in that they can provide fast approximate Bayesian inference. Especially important in machine learning applications with large data sets.
- Results: VB outperforms BIC and CS in scoring discrete DAGs.
- VB approaches the capability of AIS sampling, at little computational cost.
  - Finds the true structure consistently, whereas AIS needs tuning (e.g. large n).
  - **Compute time**:

	BIC	CS	VB	AIS
time to compute each graph score	1.5s	1.5s	4s	400s
total time for all 2720 graphs	1hr	1hr	3hrs	**13days**

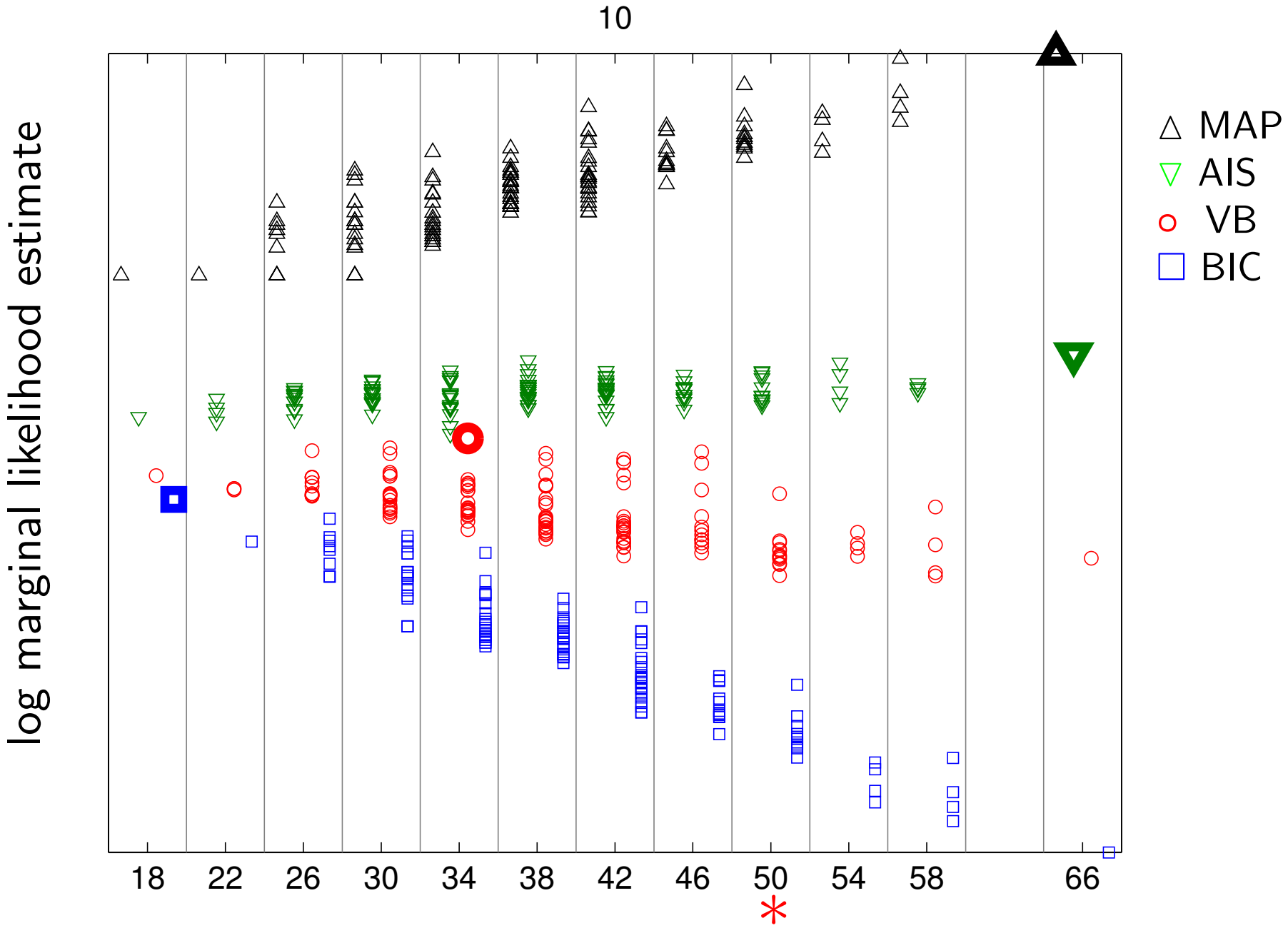
- No need to use CS because VB is provably better!

Future Work

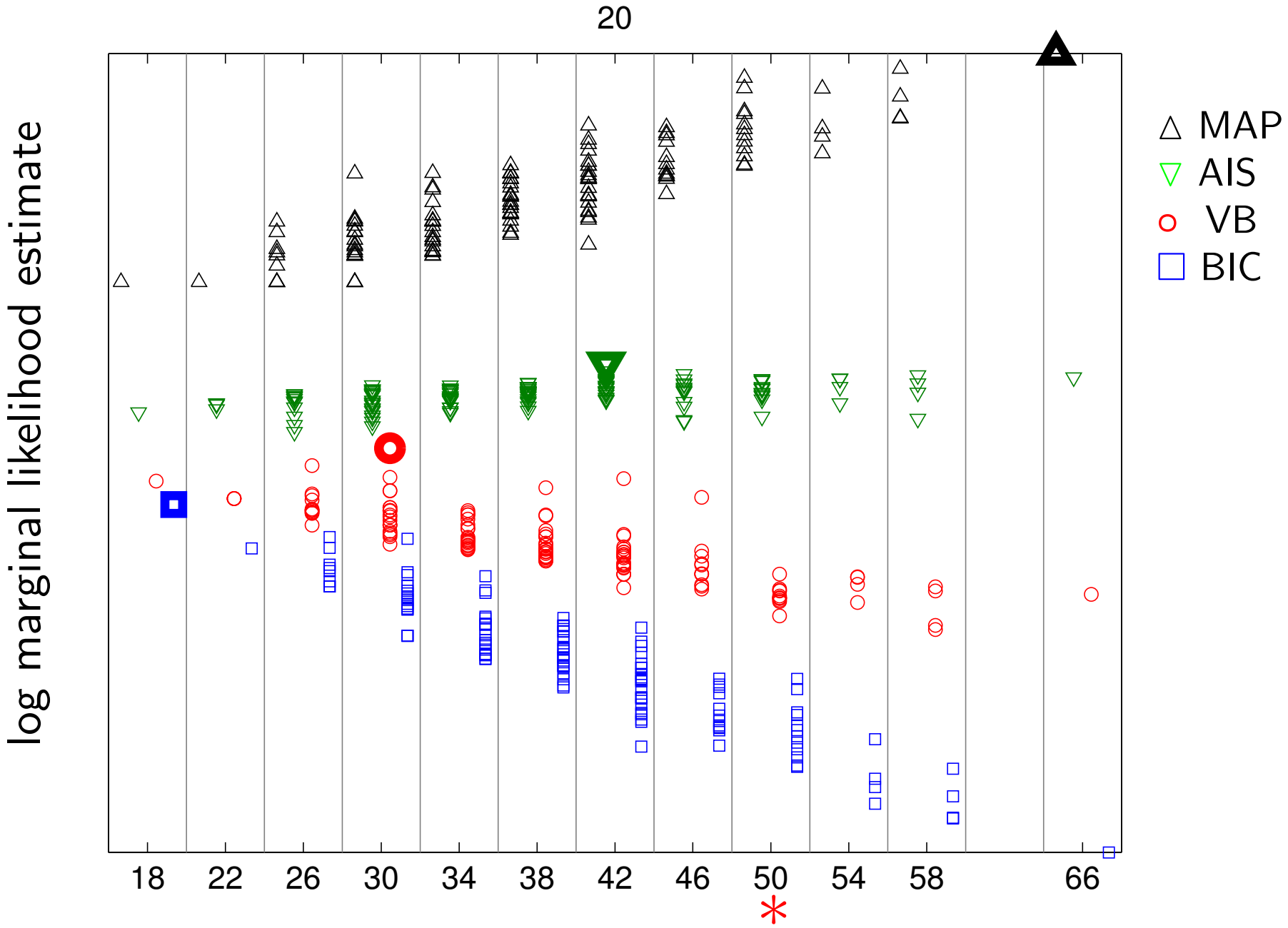
- Comparison to other methods:
  - Laplace Method
  - Other more sophisticated MCMC variants? (e.g. slice sampling)
  - Bethe/Kikuchi approximations to marginal likelihood for discrete models (Heskes)
- Incorporate into a local search algorithm over structures (exhaustive enumeration done here is only of academic interest!).
- Extend to Gaussian and other non-discrete graphical models.
- Apply to real-world data sets.
- VB tools development:
  - Overlay an AIS module into a general variational inference engine, such as *VIBES* (Winn et al.)
  - Automated algorithm derivation, such as *AutoBayes* (Gray et al.)

no more slides... coffee

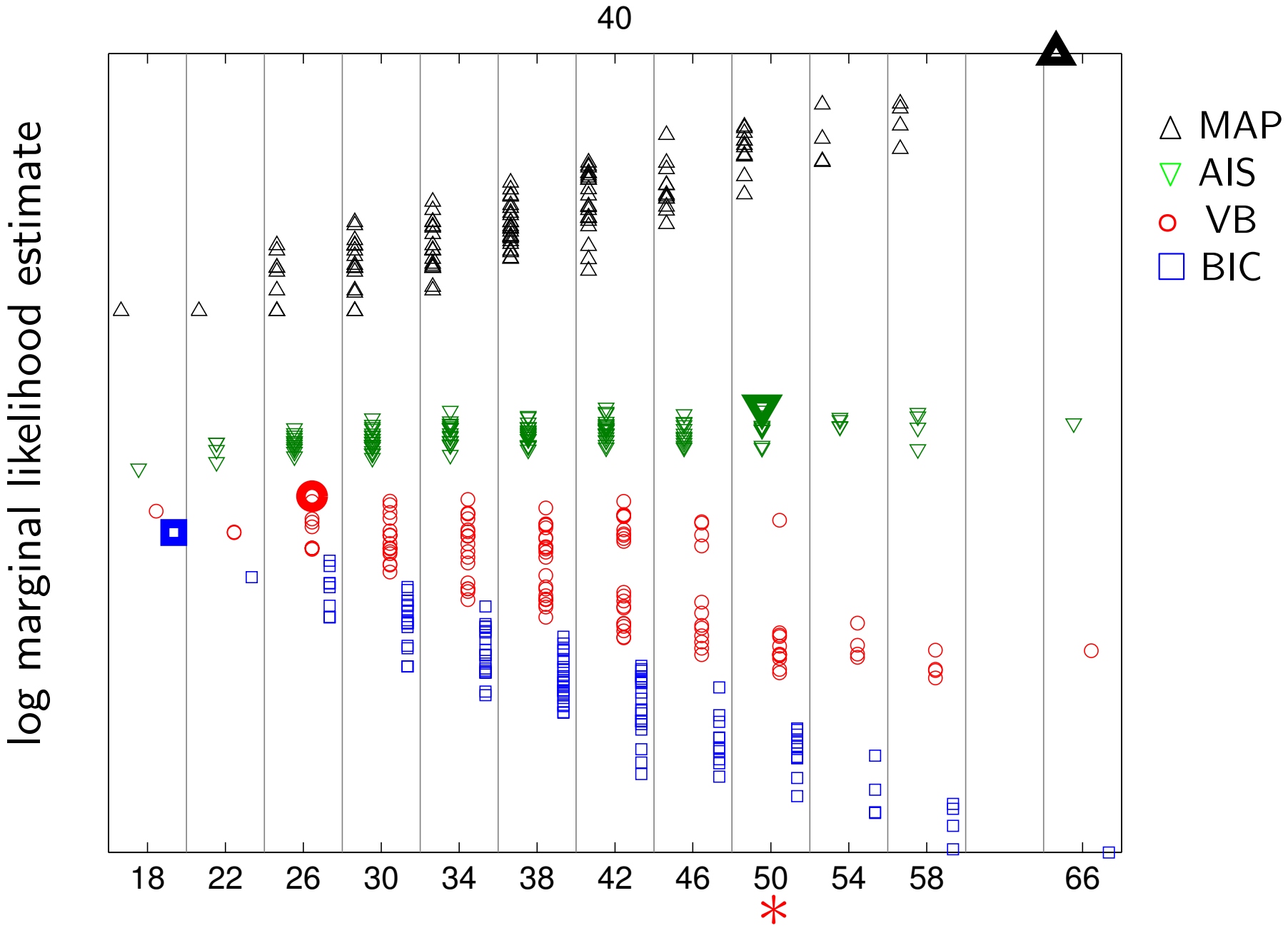
Scoring all structures by every method



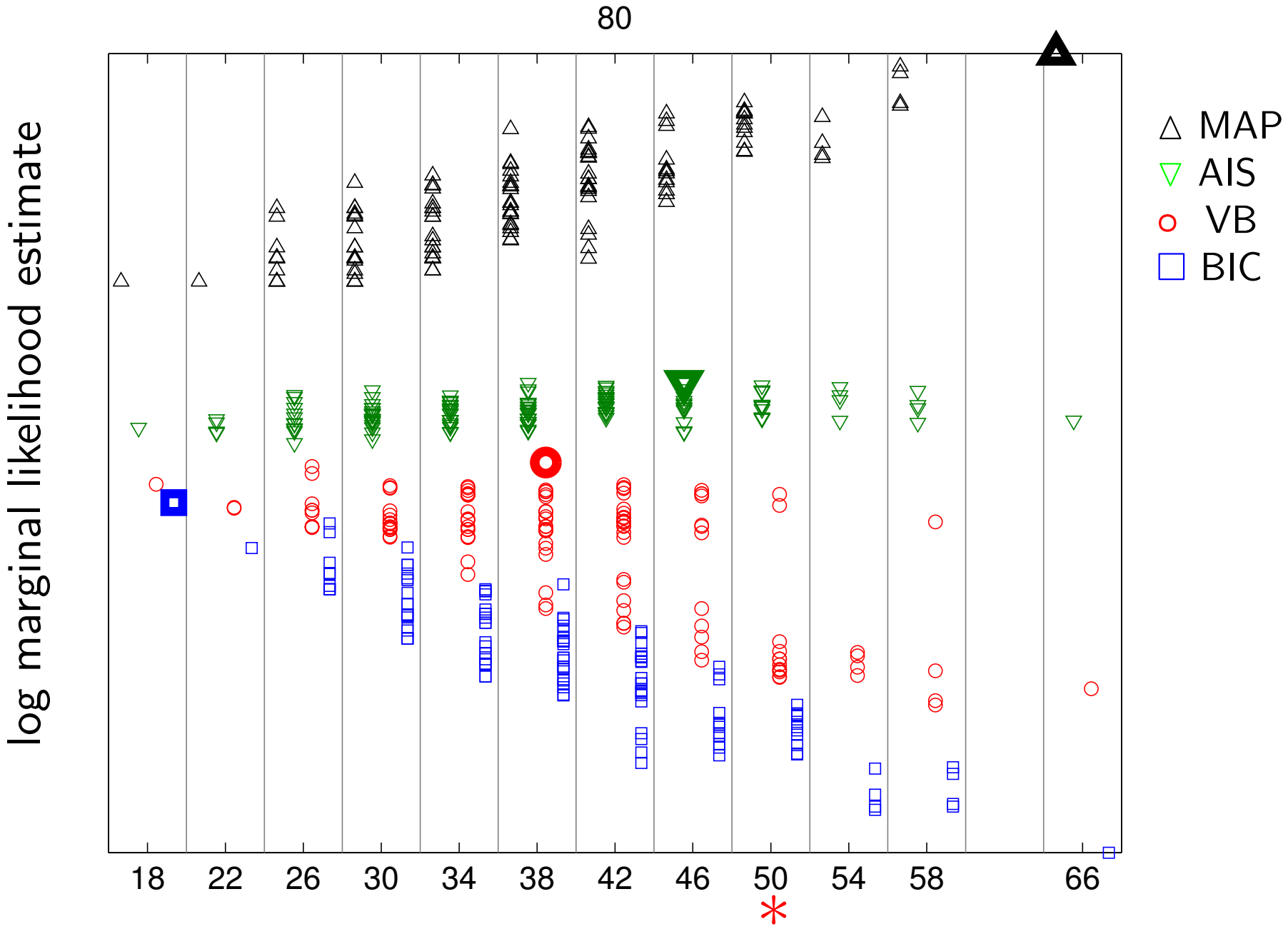
Scoring all structures by every method



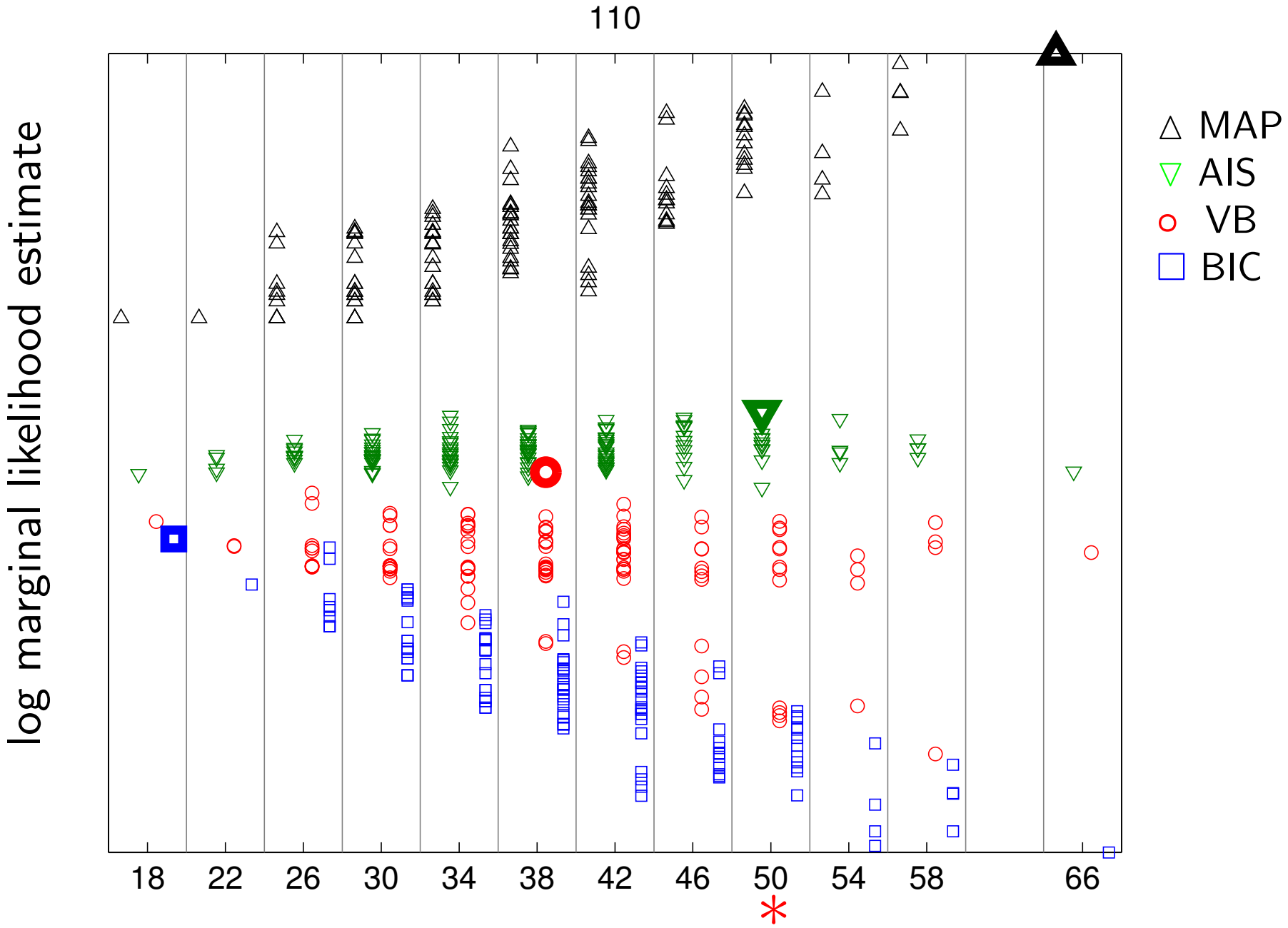
Scoring all structures by every method



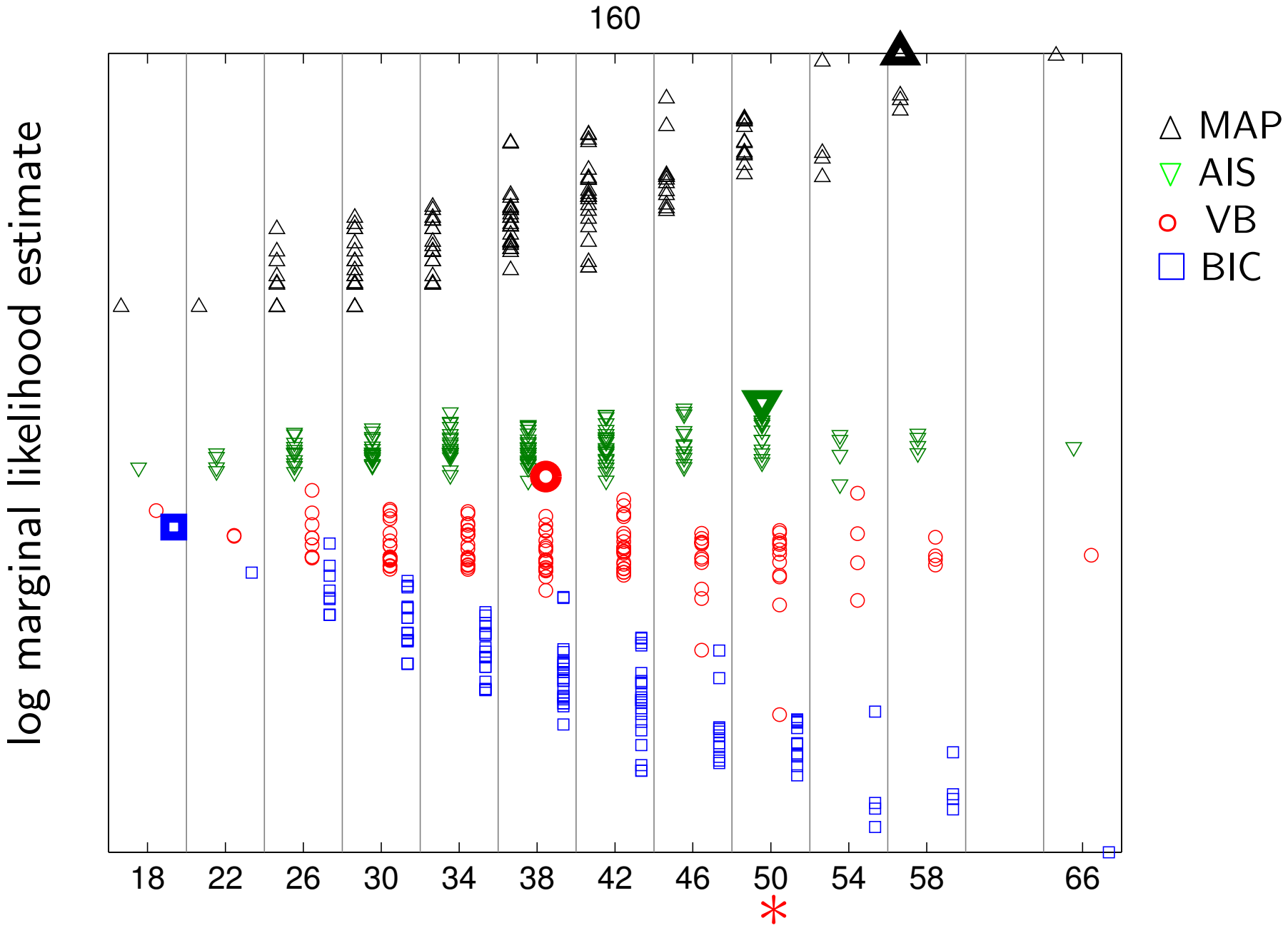
Scoring all structures by every method



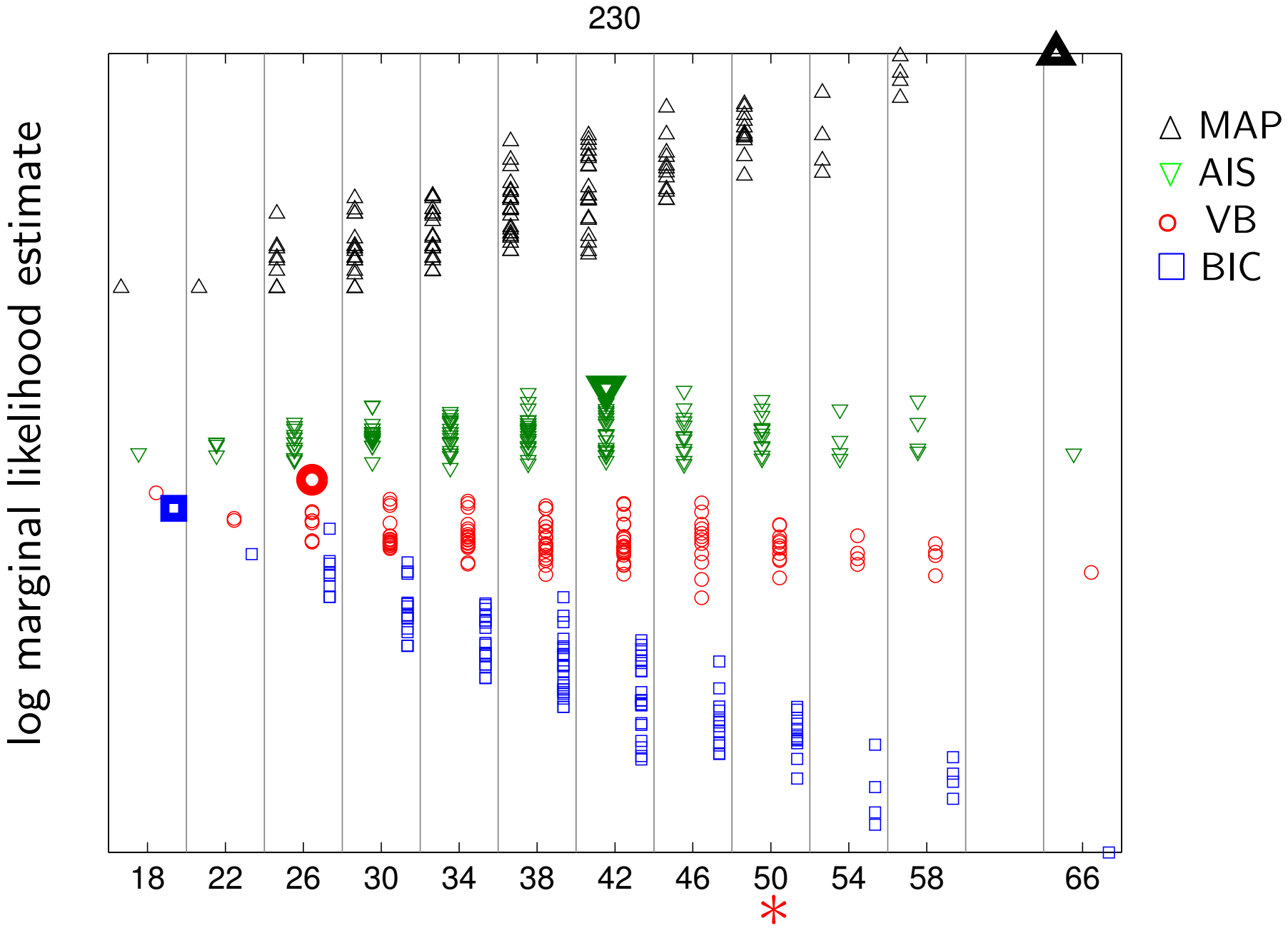
Scoring all structures by every method



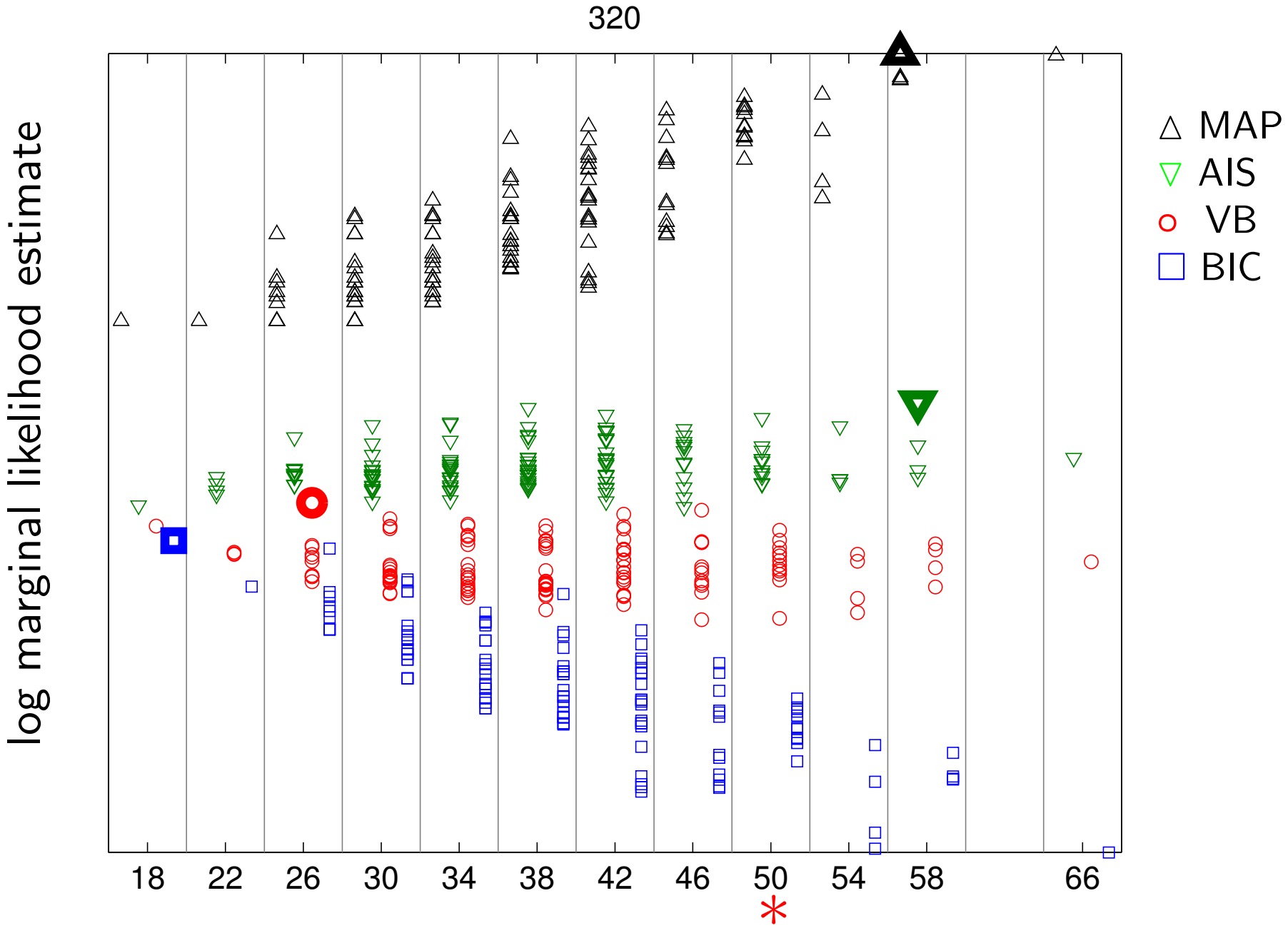
Scoring all structures by every method



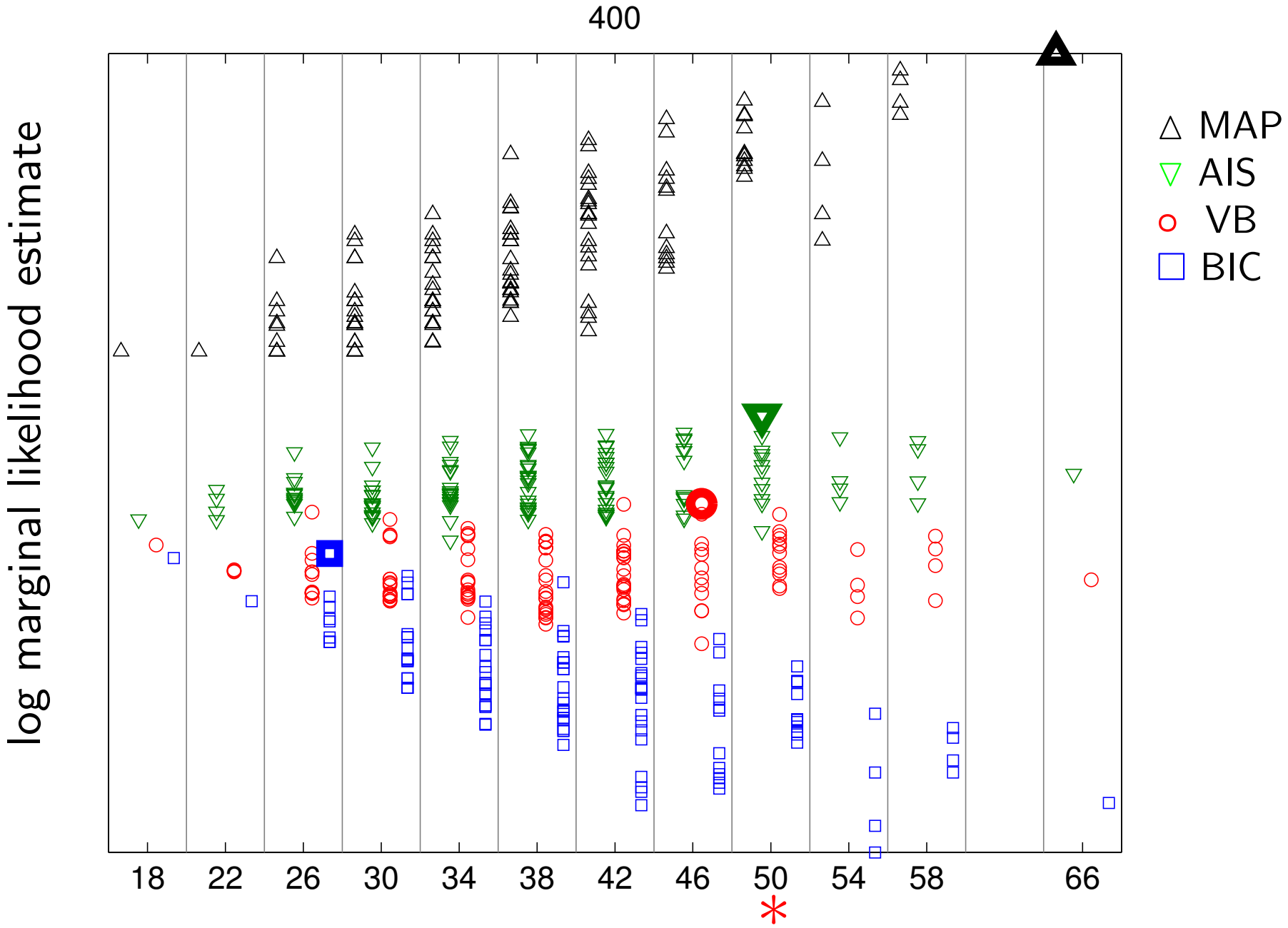
Scoring all structures by every method



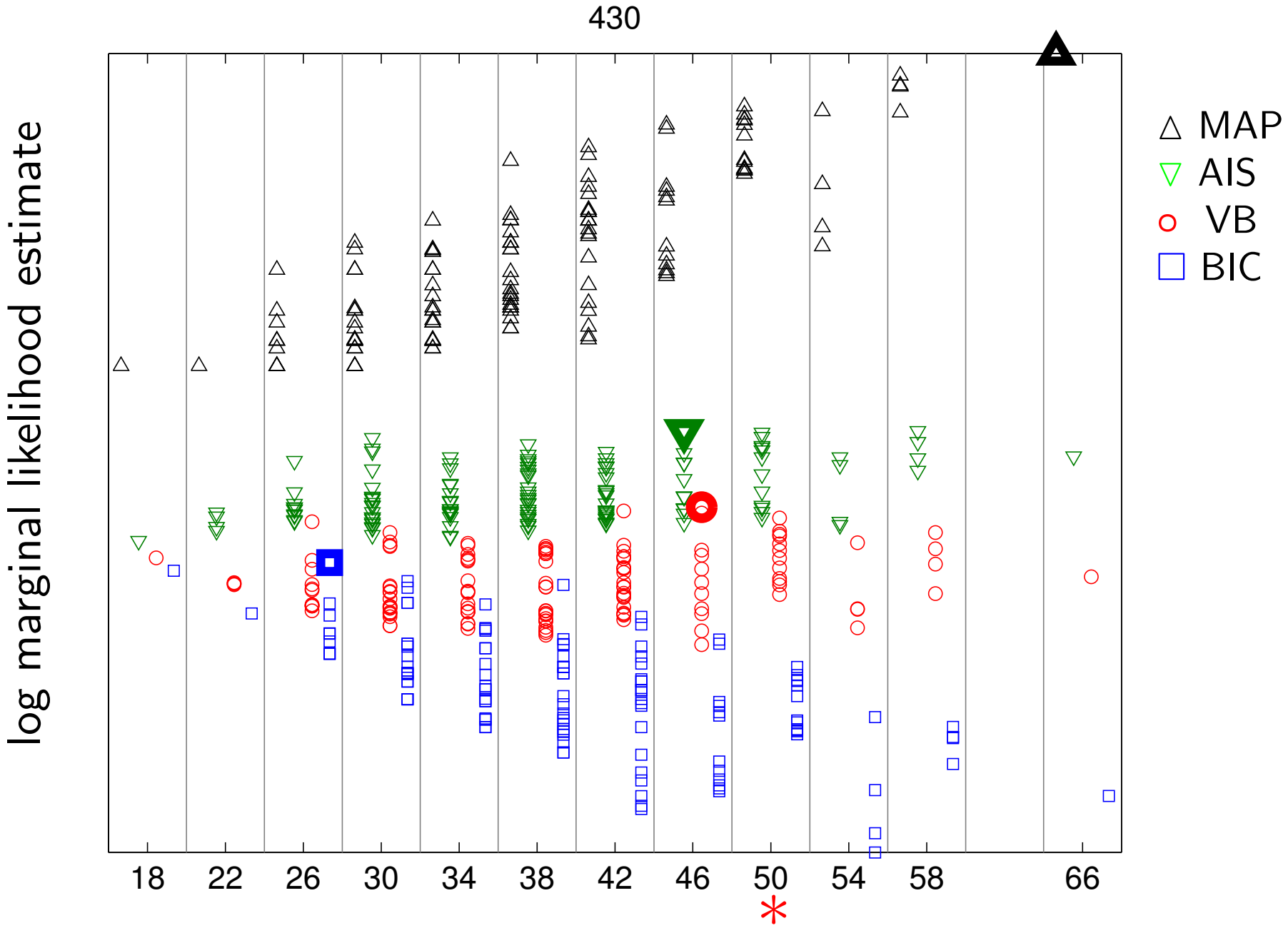
Scoring all structures by every method



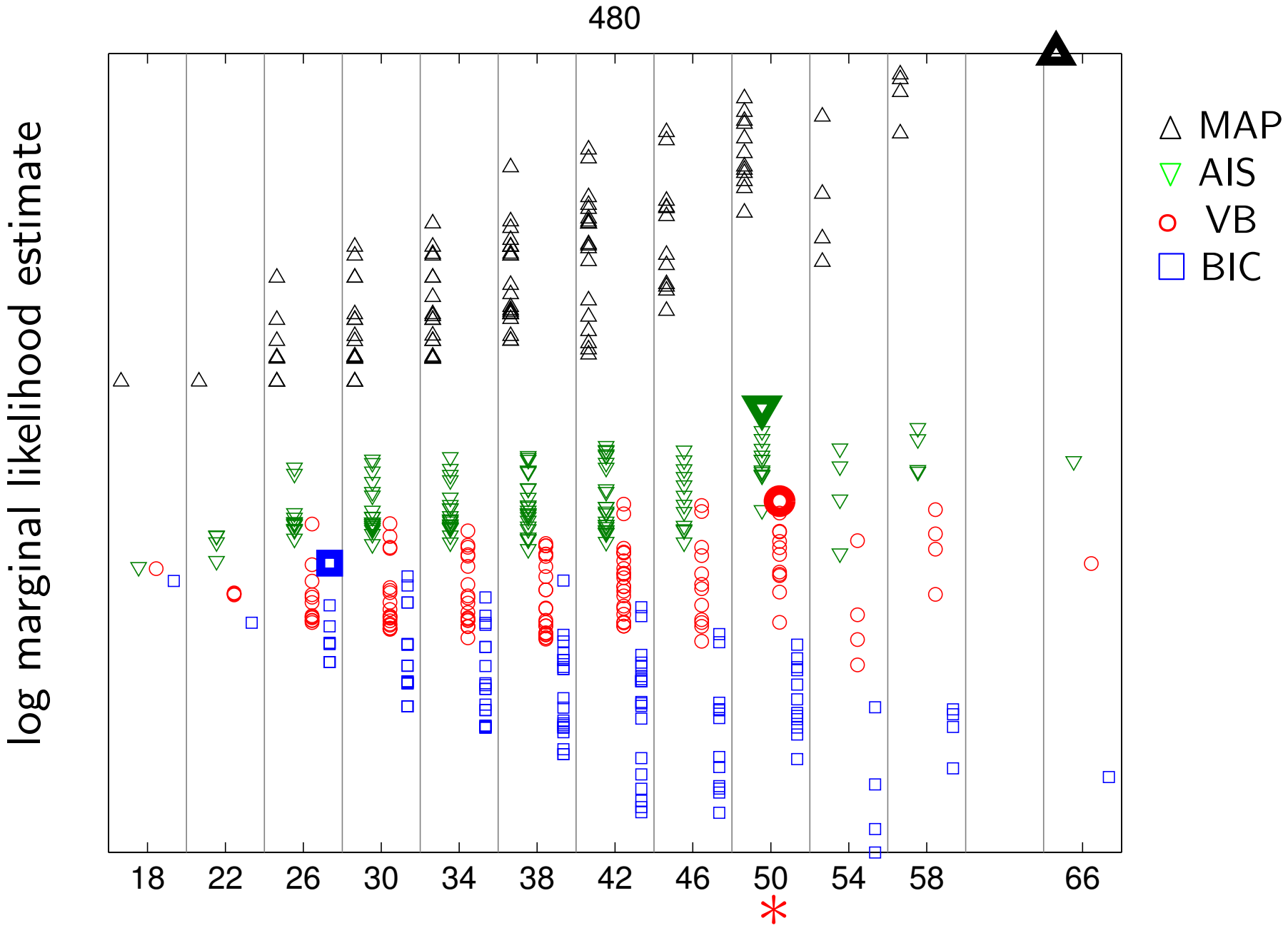
Scoring all structures by every method



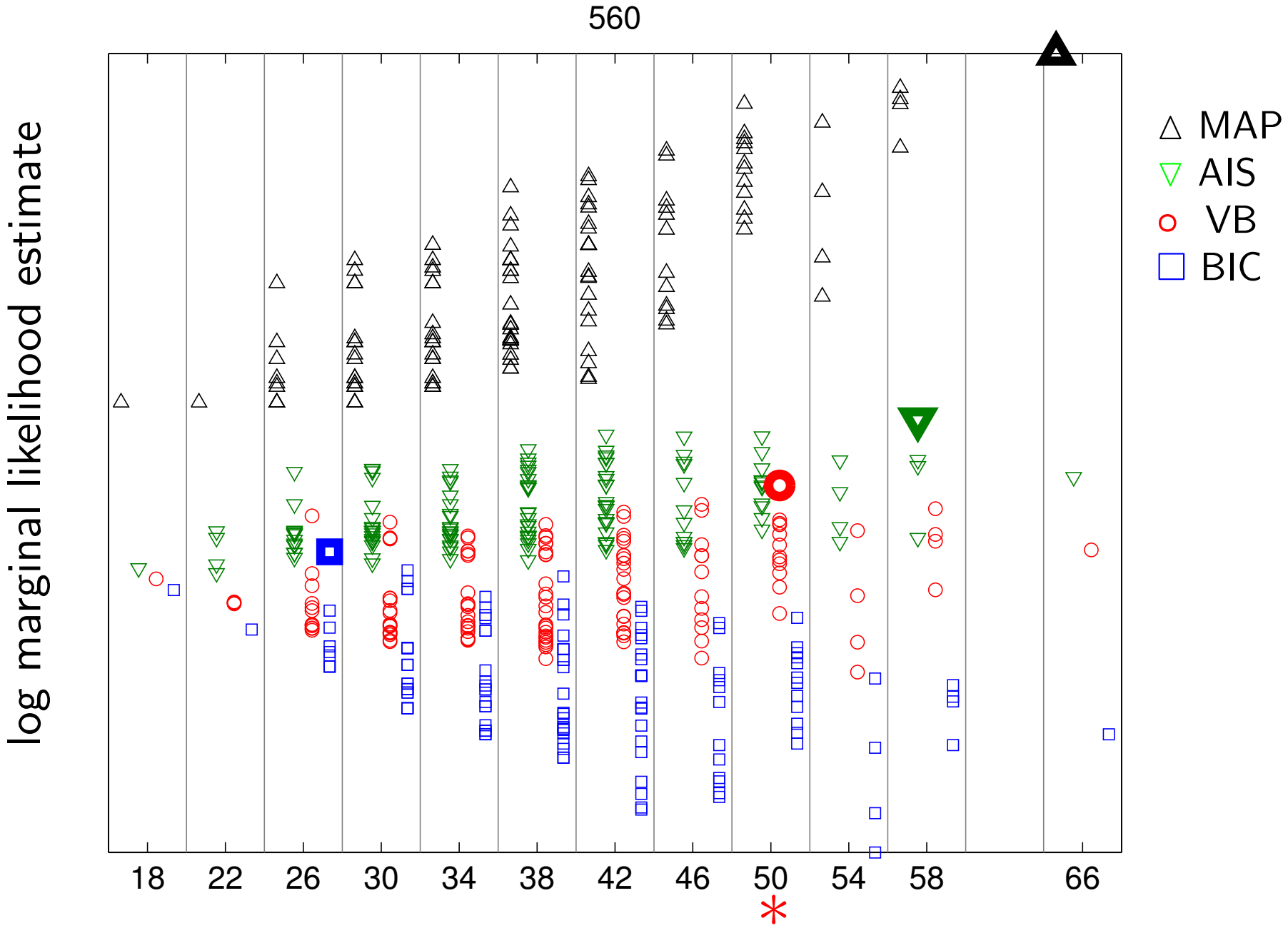
Scoring all structures by every method



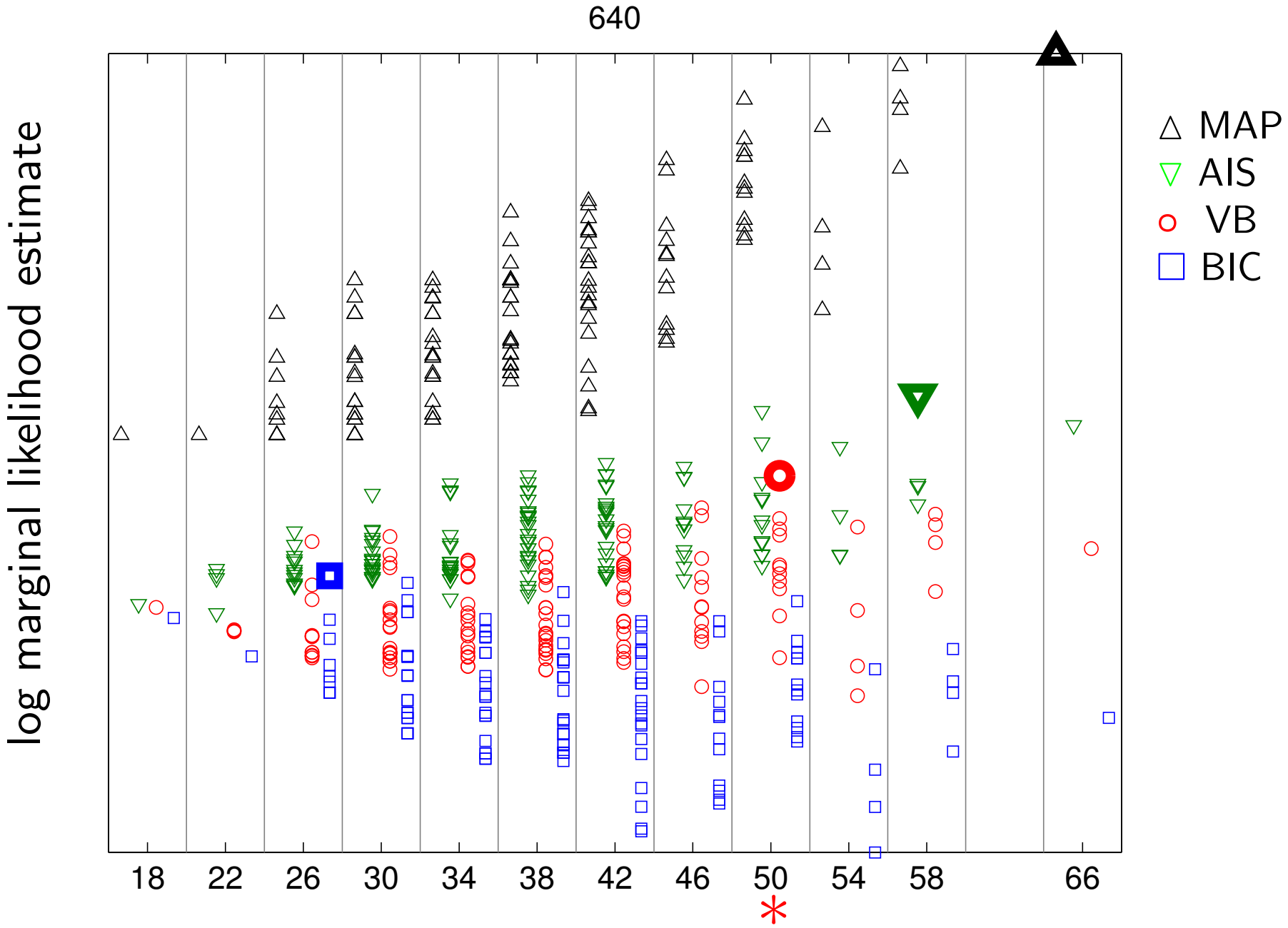
Scoring all structures by every method



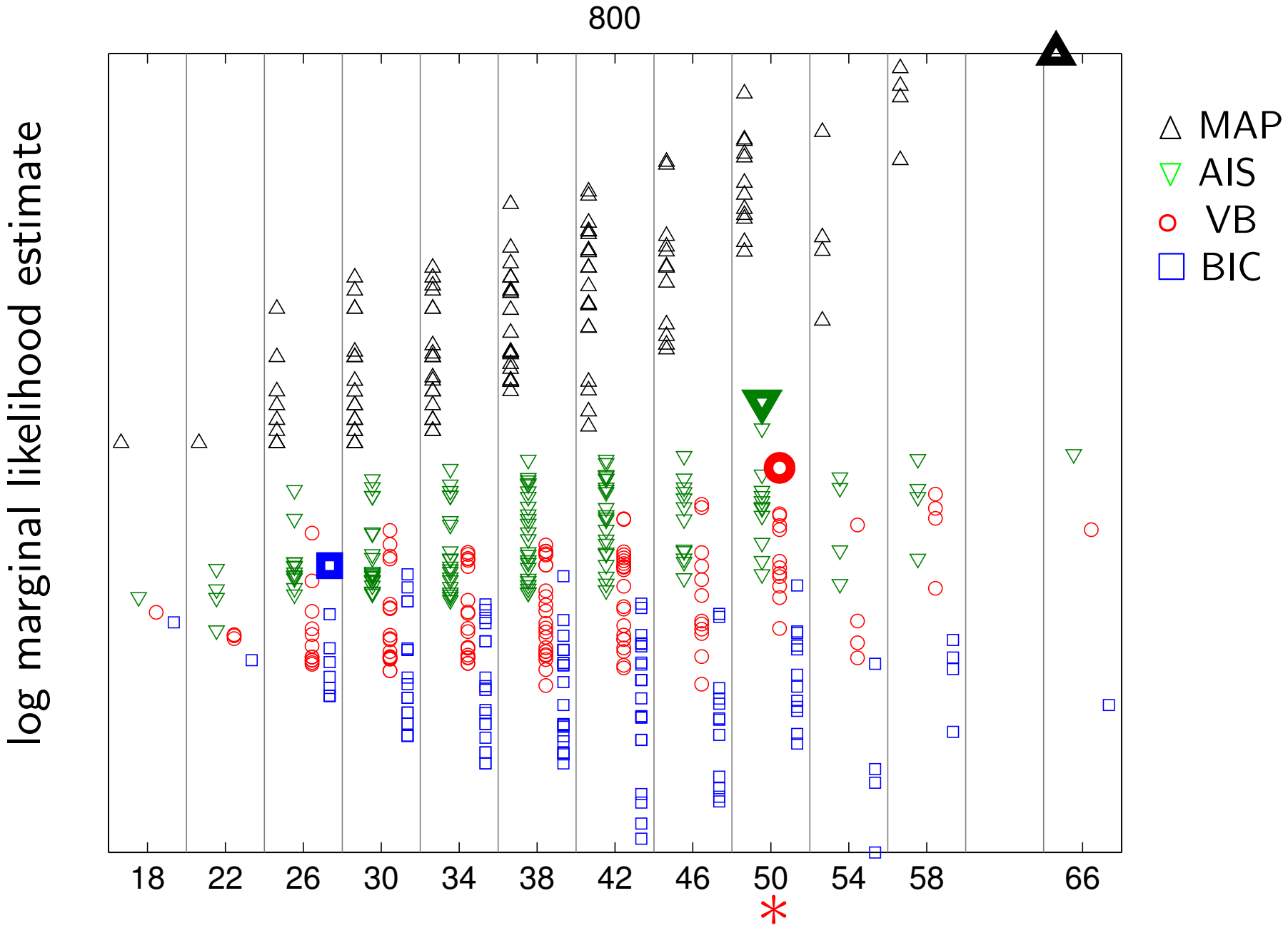
Scoring all structures by every method



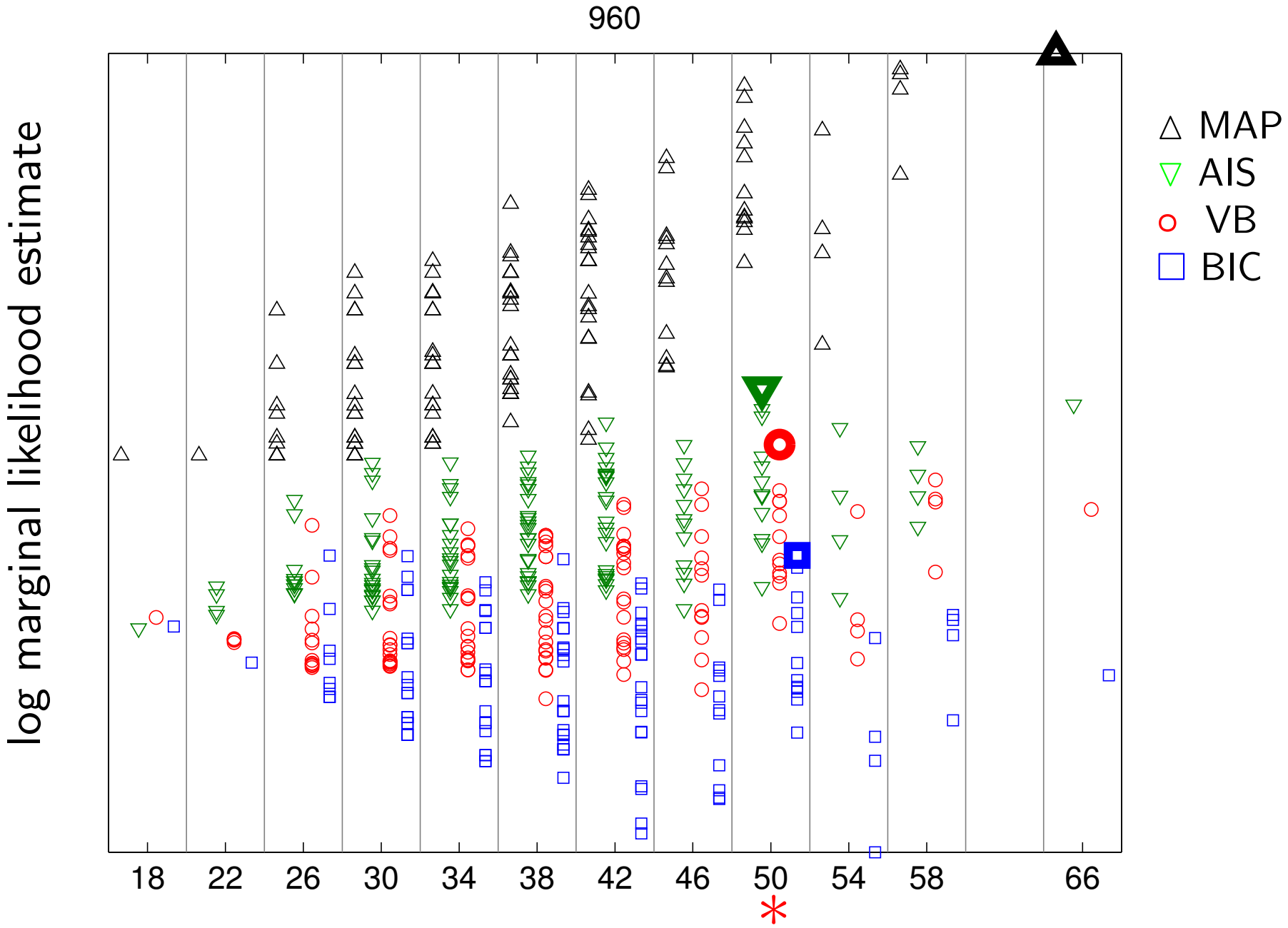
Scoring all structures by every method



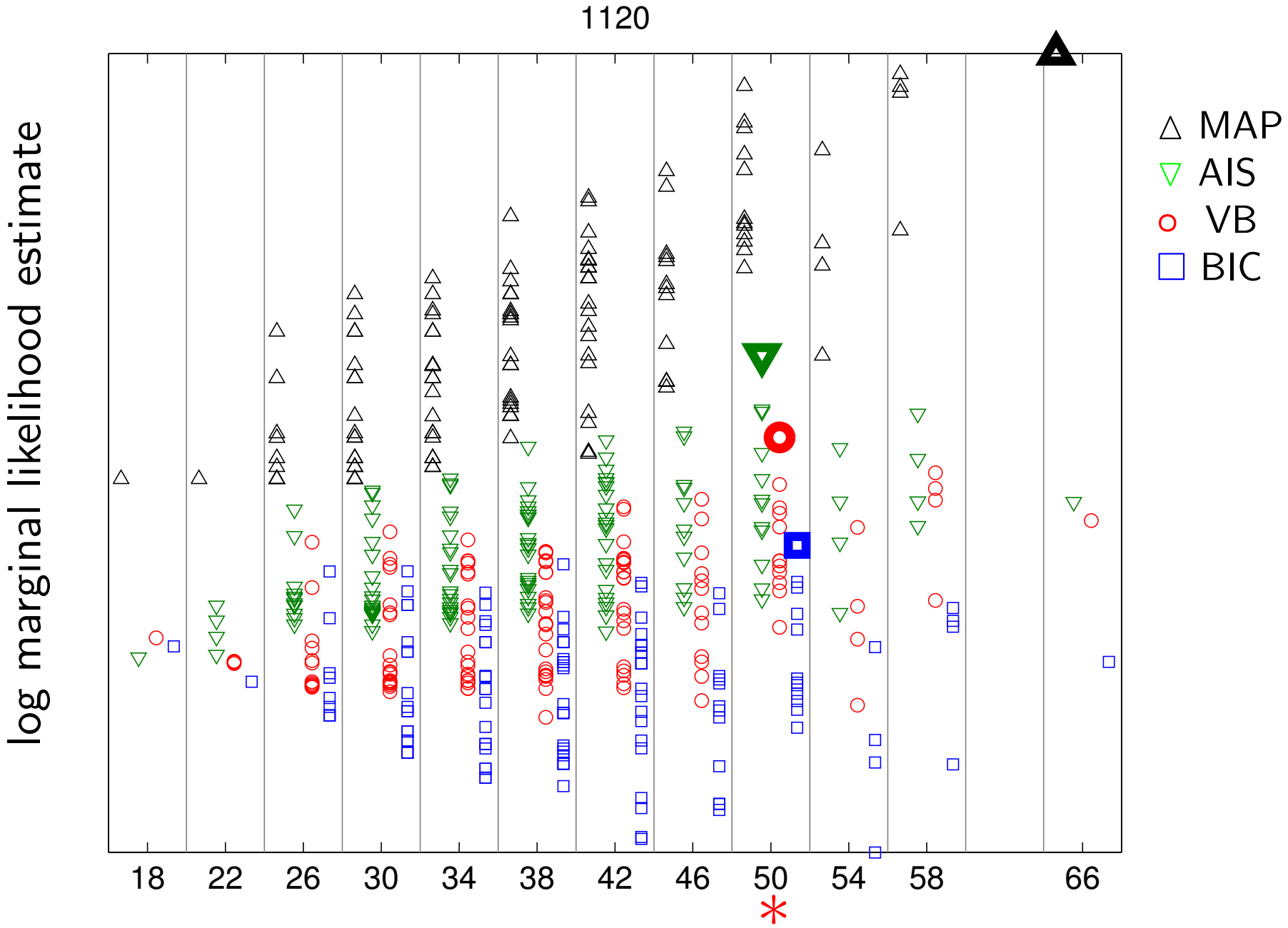
Scoring all structures by every method



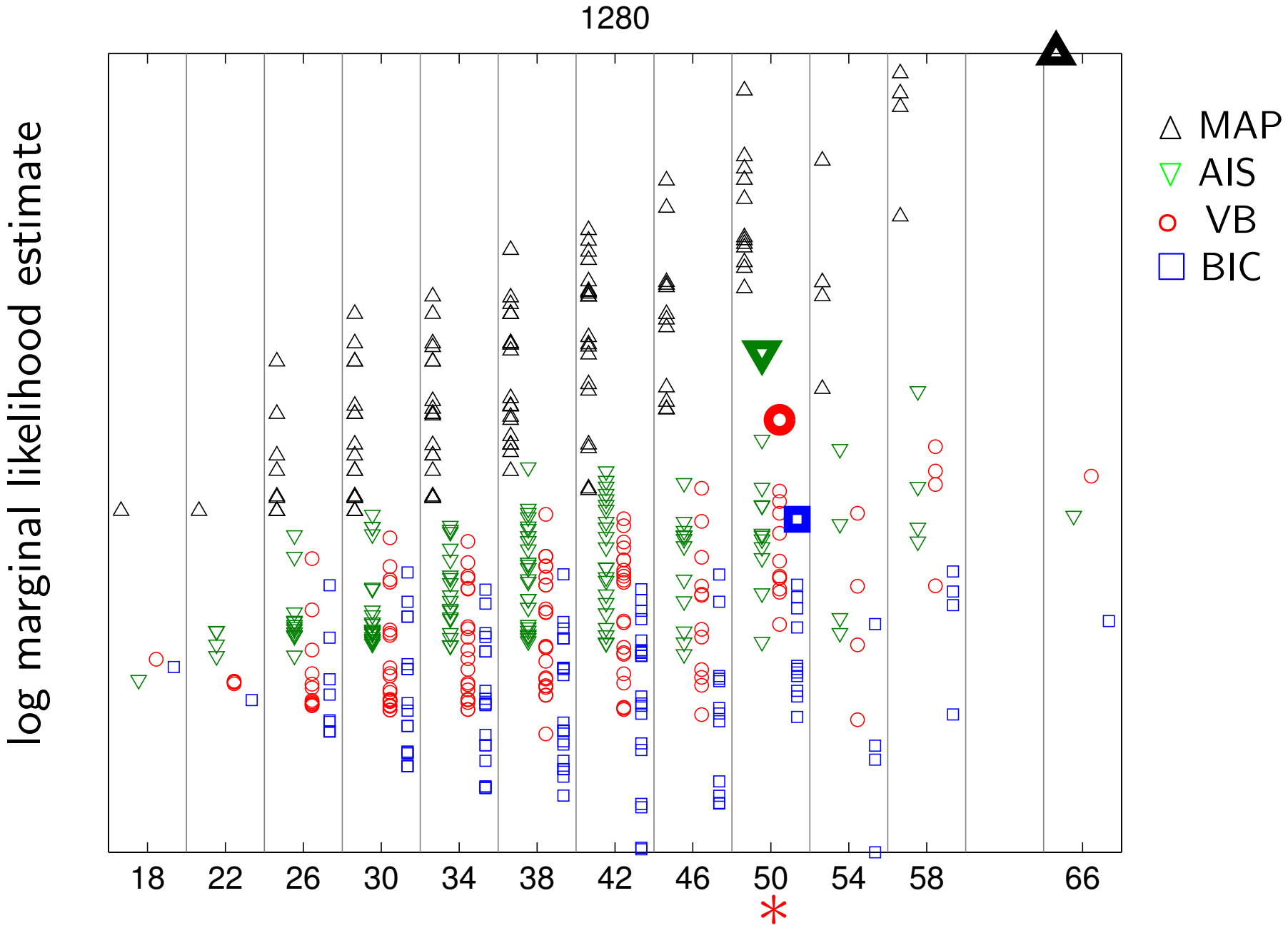
Scoring all structures by every method



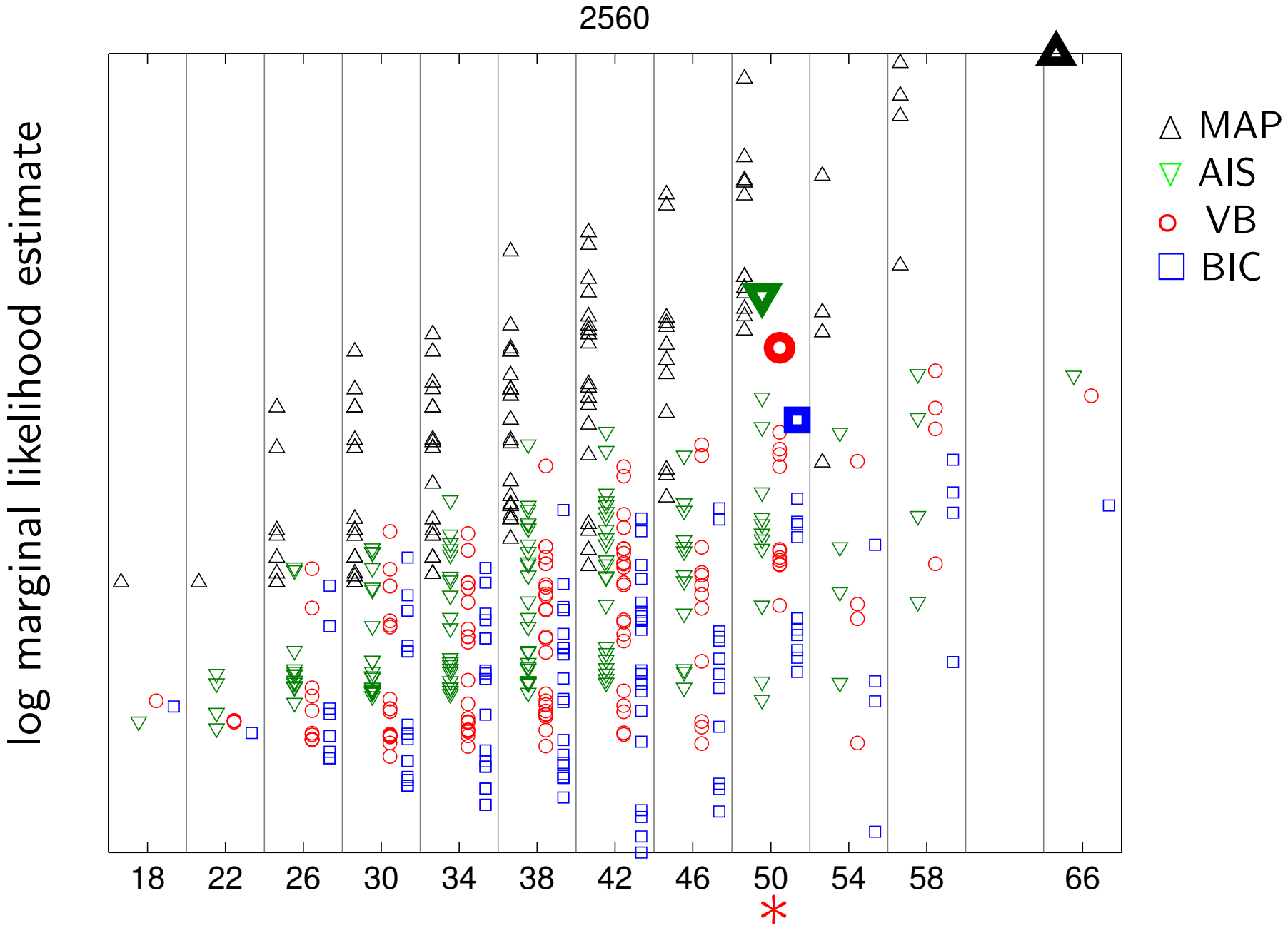
Scoring all structures by every method



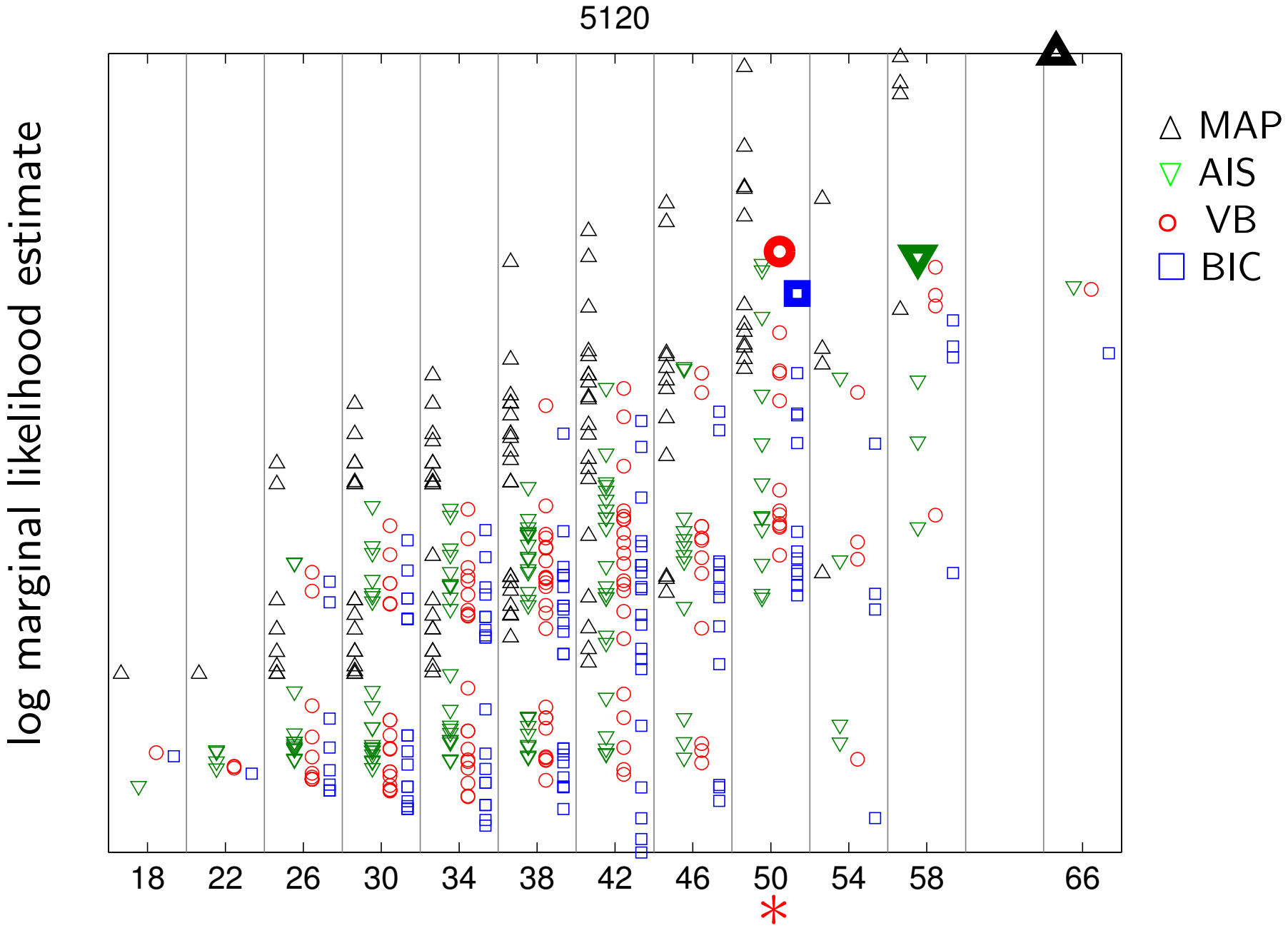
Scoring all structures by every method



Scoring all structures by every method



Scoring all structures by every method



Scoring all structures by every method