Supplementary Material for the article *Quotient Normalized Maximum Likelihood Criterion for Learning Bayesian Network Structures*

A Score equivalency proof

Chickering showed that any equivalent structure can be reached from another by a series of arc reversal operations without leaving the equivalence class Chickering (1995). This means that in order to show that equivalent structures leads to the same qNML score, we only need to prove that this is the case for equivalent structures that differ by a single arc reversal.

The network without any arcs is the sole member of its equivalence class. All the other networks have at least one arc. We will first present some lemmas that characterize the relations of the parent sets before and after the arc reversal. Let G be a network structure and G' an equivalent network structure after the arc from A to B has been reversed to point from B to A. We will denote the parent sets of A and B in G by G_A and G_B, and the parent sets of A and B in G' by G'_A and G'_B. Note that not all the arc reversals lead to the equivalent structures. By saying that G and G' are equivalent we imply that our arc reversal does not destroy existing V-structures or create new ones. The crucial observation is that if reversing the arc that goes from A to B does not create or destroy V-structures, it must be that the parents of B other than A are exactly the same as parents of A (statement 3 below).

Lemma 1.

$G_B = G_A \cup \{A\}$

Proof. We show the direction $G_A \cup \{A\} \subset G_B$ by contradiction. First of all, a parent of A cannot be a child of B or otherwise there would be a loop in G. If A had a parent Z that is not a parent of B, then B and Z were not adjacent and the reversal of the arc would create a new V-structure (B, A, Z). Since this was forbidden by the equivalence of G and G', Z must also be a parent of B.

Also, B cannot have parents other than A that are not also parents of A, i.e., $G_B \subset G_A \cup A$. If there were such a parent Z it should be anyway adjacent to A, otherwise the V-structure (A, B, Z) would be destroyed in the reversal. Since the Z was not a parent of A, the only possibility for adjacency is that Z were a child of A. However, in this case arc reversal would lead to a loop B, A, Z, B. □

We will next list some simple consequences of the equivalence preserving arc reversal:

Lemma 2.

1. $G'_B = G_B \setminus \{A\}$
2. $G'_A = G_A \cup \{B\}$
3. $G'_B = G_A$
4. $\{B\} \cup G_B = \{B\} \cup G_A \cup \{A\} = G'_A \cup \{A\}$
5. $G'_A = G_A \cup \{B\} = G'_B \cup \{B\}$

Proof. The statements 1 and 2 are trivial since they just state the fact that an arc reversal removes A from the parents of B and adds B to the parents of A.

The statement 3 follows directly by using the lemma 1 to the statement 1. Combining these equalities we can generate more of them such as statements 4 and 5. □

Let us take structures G and G'. Since we can assume the data to be fixed we lighten up the notations and write $P^1_{NML}(i, G_i) := P^1_{NML}(D[\cdot, (i, G_i)]; G)$ and $P^1_{NML}(G_i) := P^1_{NML}(D[\cdot, G_i]; G)$.

Theorem 1. $s^{qNML}(D; G) = s^{qNML}(D; G')$.

Proof. Now the scores for structures can be decomposed as the $s^{qNML}(D;G) = \sum_{i=1}^{n} s_i^{qNML}(D;G)$ and $s^{qNML}(D;G') = \sum_{i=1}^{n} s_i^{qNML}(D;G')$.

Since only the terms corresponding to the variables A and B in these sums are different, it is enough to show that

$$s_A^{qNML}(D;G) + s_B^{qNML}(D;G) = s_A^{qNML}(D;G') + s_B^{qNML}(D;G')$$

Now

$$\begin{aligned} s_A^{qNML}(D;G) + s_B^{qNML}(D;G) &= \log \frac{P^1_{NML}(A,G_A)}{P^1_{NML}(G_A)} \frac{P^1_{NML}(B,G_B)}{P^1_{NML}(G_B)} \\ &= \log 1 \cdot \frac{P^1_{NML}(B,G_B)}{P^1_{NML}(G_A)} \\ &= \log \frac{P^1_{NML}(B,G'_B)}{P^1_{NML}(G'_A)} \frac{P^1_{NML}(A,G'_A)}{P^1_{NML}(G'_B)} \\ &= s_A^{qNML}(D;G') + s_B^{qNML}(D;G'). \end{aligned}$$

The second equation follows from the lemma 1, and the third from the statements 5 and 4 of the lemma 2. □

B Regularity proof

B.1 Preliminaries

We start by recalling the definition of regularity (Suzuki, 2017):

Definition 1. *Assume $H_N(X \mid Y') \leq H_N(X \mid Y)$, where $Y' \subset Y$. We say that the scoring function $Q_N(\cdot \mid \cdot)$ is regular if $Q_N(X \mid Y') \geq Q_N(X \mid Y)$.*

In the definition, N denotes the sample size, X is some random variable, Y denotes the proposed parent set for X, and $H_N(X \mid Y)$ refers to the empirical conditional entropy based on N samples of variables X and Y.

Let X be a categorical random variable with r possible values. Let U denote a possible parent set with q different combinations of values for the variables, and V a set with m different configurations. Assume that we have observed N samples of (X, U, V) (denoted by x_N, u_N and v_N) and $H_N(X \mid U) \leq H_N(X \mid U \cup V)$ holds.

Recall the definition of the qNML score:

$$\begin{aligned} Q_N^{qnml}(X \mid U) &= \log P(x_N \mid \hat{\theta}_{X|U}\ ) - (reg(N, rq) - reg(N, q)) \\ &= \log P(x_N \mid \hat{\theta}_{X|U}\ ) - \log \frac{C(N, rq)}{C(N, q)}, \end{aligned}$$

where $C(N, r)$ is the normalizing constant the of the NML distribution for a categorical variable with r possible values and sample size N and $\hat{\theta}_{X|U}$ denotes the maximum likelihood parameters of the conditional distribution of X given U which are computed from the data (x_N, u_N).

In order to prove the regularity, we need the following three lemmas:

Lemma 3. *We can write $C(N, k)$ as a polynomial of k, formally*

$$C(N, k) = \sum_{j=1}^{N} a_j k^j,$$

where $a_j > 0$.

Lemma 4. *Assume $H_N(X \mid Y') \leq H_N(X \mid Y)$, where $Y' \subset Y$. Now $\log P(x_N \mid \hat{\theta}_{X|Y}\ ) = \log P(x_N \mid \hat{\theta}_{X|Y'}\ )$.*

Lemma 5. *Let $r \in \mathbb{N}, r \geq 2$. The function $k \mapsto \frac{C(N,rk)}{C(N,k)}$ is increasing for every $k \geq 2$.*

We present the proofs of these lemmas in Section B.3.

B.2 The main proof

Theorem 2. *qNML score is regular.*

Proof. We want to show that

$$Q_N^{qnml}(X \mid U) \geq Q_N^{qnml}(X \mid U \cup V).$$

assuming $H_N(X \mid U) \leq H_N(X \mid U \cup V)$. Using the entropy assumption and Lemma 4 implies that the maximized likelihood terms are equal. In order to prove the claim, it suffices to study the penalty terms, and we want to show that

$$\begin{aligned} -\left(reg(N, rq) - reg(N, q)\right) &\geq -\left(reg(N, rqm) - reg(N, qm)\right) \\ \log \frac{C(N, rq)}{C(N, q)} &\leq \log \frac{C(N, rqm)}{C(N, qm)}. \end{aligned}$$

This holds, since logarithm is an increasing function, and $q \leq qm$, so we can apply Lemma 5 to conclude the proof.

□

B.3 Proofs of lemmas

Lemma 1. *$C(N, k)$ can be written as a polynomial of k, formally*

$$C(N, k) = \sum_{j=1}^{N} a_j k^j,$$

where $a_j > 0$.

Proof. Mononen and Myllymäki (2008) derive the following representation for the normalizing constant

$$C(N, k) = \sum_{l=0}^{N-1} \frac{(N-1)^{\underline{l}} k^{\overline{l+1}}}{N^{l+1}\ l!},$$

where $x^{\underline{l}}$ and $x^{\overline{l}}$ denote falling and rising factorials, respectively.

We utilize the fact that the rising factorial can be represented as polynomial using unsigned Stirling numbers of the first kind (see Adamchik (1997), for instance)

$$\begin{aligned}
C(N,k) &= \sum_{l=0}^{N-1} \frac{(N-1)^{\underline{l}} k^{\overline{l+1}}}{N^{l+1}\ l!} \\
&= \sum_{l=0}^{N-1} b_l\ k^{\overline{l+1}} \\
&= \sum_{l=0}^{N-1} b_l \left(\sum_{j=1}^{l+1} |s(l+1,j)|\ k^j \right) \\
&= \sum_{l=0}^{N-1} \left(\sum_{j=1}^{N} b_l\ |s(l+1,j)|\ k^j \right) \\
&= \sum_{j=1}^{N} \left(\sum_{l=0}^{N-1} b_l\ |s(l+1,j)|\ k^j \right) \\
&= \sum_{j=1}^{N} \left(\sum_{l=0}^{N-1} b_l\ |s(l+1,j)| \right) k^j \\
&= \sum_{j=1}^{N} a_j k^j,
\end{aligned}$$

where $s(i,j)$ denotes the (signed) Stirling number of the first kind and

$$a_j = \left(\sum_{l=0}^{N-1} \frac{(N-1)^{\underline{l}}}{N^{l+1} l!}\ |s(l+1,j)| \right),$$

$a_j > 0$ for all j. On the second row, we denoted $b_l = (N-1)^{\underline{l}}/(N^{l+1}l!)$. On the row 4, we used the property of Stirling numbers: $s(i,j) = 0$ for all $j > i$. □

Lemma 2. *Assume $H_N(X \mid Y') \leq H_N(X \mid Y)$, where $Y' \subset Y$. Now $\log P(x_N \mid \hat{\theta}_{X|Y}\ ) = \log P(x_N \mid \hat{\theta}_{X|Y'})$.*

Proof. We can write the logarithm of the maximized likelihood,

$\log P(x_N \mid \hat{\theta}_{X|Y}\ )$, as follows (Koller and Friedman, 2009)

$$\begin{aligned}\log P(x_N \mid \hat{\theta}_{X|Y}\ ) &= -N\left(H_N(X) - I_N(X;Y)\right)\\ &= -N\ H_N(X \mid Y),\end{aligned}$$

where $I_N(\cdot;\cdot)$ is the empirical mutual information. This implies that the assumption

$$H_N(X \mid Y') \leq H_N(X \mid Y)$$

is equivalent to

$$\log P(x_N \mid \hat{\theta}_{X|Y}\ ) \leq \log P(x_N \mid \hat{\theta}_{X|Y'}\ ).$$

Actually we must have the equality holding in the above expression, since

$$H_N(X \mid Y') < H_N(X \mid Y)$$

would imply that

$$I_N(X; Z \mid Y) < 0,$$

where $Z = Y \setminus Y'$, which is impossible. □

Lemma 3. *Let $r \in \mathbb{N}, r \geq 2$. The function $k \mapsto \frac{C(N,rk)}{C(N,k)}$ is increasing for every $k \geq 2$.*

Proof. Lemma 3 lets us to write

$$C(N,k) = \sum_{j=1}^{N} a_j k^j \tag{1}$$

and, similarly,

$$C(N,rk) = \sum_{j=1}^{N} a_j r^j k^j. \tag{2}$$

In the following, we assume that k is a real number. From (1) and (2), it is easy see that the derivative of the quotient, $d/dk(C(N,rk)/C(N,k))$, will be a ratio of two polynomials of k. Our goal is to show that the polynomial in the numerator has positive coefficients, which will guarantee the positivity of derivative for every $k > 0$, and thus imply that the original function is increasing (polynomial in the denominator is squared and non-zero for $k > 0$, so it can be ignored).

Derivatives of (1) and (2) are obtained easily:

$$\begin{aligned}\frac{d}{dk}C(N,k) &= \sum_{j=1}^{N} j a_j k^{j-1} \\ &= \sum_{j=0}^{N-1} (j+1) a_{j+1} k^j\end{aligned}$$

and

$$\begin{aligned}\frac{d}{dk}C(N,rk) &= \sum_{j=1}^{N} j a_j r^j k^{j-1} \\ &= \sum_{j=0}^{N-1} (j+1) a_{j+1} r^{j+1} k^j.\end{aligned}$$

Consider next the products found in the derivative of the quotient. We obtain

$$\begin{aligned}\left(\frac{d}{dk}C(N,rk)\right) C(N,k) &= \left(\sum_{j=0}^{N-1} (j+1) a_{j+1} r^{j+1} k^j\right)\left(\sum_{l=1}^{N} a_l k^l\right) \\ &= \sum_{i=1}^{2N-1} \left(\sum_{j+l=i} (j+1) a_{j+1} r^{j+1} a_l\right) k^i\end{aligned}$$

and

$$\begin{aligned}\left(\frac{d}{dk}C(N,k)\right) C(N,rk) &= \left(\sum_{j=0}^{N-1} (j+1) a_{j+1} k^j\right)\left(\sum_{l=1}^{N} a_l r^l k^l\right) \\ &= \sum_{i=1}^{2N-1} \left(\sum_{j+l=i} (j+1) a_{j+1} a_l r^l\right) k^i.\end{aligned}$$

Subtracting these two expression yields

$$\begin{aligned}&\left(\frac{d}{dk}C(N,rk)\right) C(N,k) - \left(\frac{d}{dk}C(N,k)\right) C(N,rk) \\ &= \sum_{i=1}^{2N-1} \left(\sum_{j+l=i} (j+1) a_{j+1} r^{j+1} a_l\right) k^i - \sum_{i=1}^{2N-1} \left(\sum_{j+l=i} (j+1) a_{j+1} a_l r^l\right) k^i \\ &= \sum_{i=1}^{2N-1} \left(\sum_{j+l=i} (j+1) a_{j+1} a_l (r^{j+1} - r^l)\right) k^i\end{aligned}$$

which is the polynomial in the numerator of the derivative of $C(N,rk)/C(N,k)$. Next, we study the coefficient of k^i, if $i \leq N$

$$\begin{aligned}\sum_{j+l=i}(j+1)a_{j+1}a_l(r^{j+1}-r^l) &= \sum_{l=1}^{i}(i-l+1)a_{i-l+1}a_l(r^{i-l+1}-r^l)\\ &= \sum_{l=1}^{i}(i-l+1)c_l\\ &= \sum_{t=1}^{\lfloor i/2\rfloor}(i-t+1)c_t+(i-(i-t+1)+1)c_{i-t+1}\\ &= \sum_{t=1}^{\lfloor i/2\rfloor}(i-t+1)c_t+tc_{i-t+1}\\ &= \sum_{t=1}^{\lfloor i/2\rfloor}(i-t+1)c_t-tc_t\\ &= \sum_{t=1}^{\lfloor i/2\rfloor}(i-2t+1)c_t.\end{aligned}$$

On the first row, we re-wrote sum using only one running index. On the second row we denoted $c_l = a_{i-l+1}a_l(r^{i-l+1}-r^l)$. On the third row, we re-arranged the sum so that we are summing over pairs of terms of the original sum: the first and the last term, the second and the second to last, and so on. This resulting sum has $\lfloor i/2\rfloor$ terms. We have to use the floor-function since if i is odd, there exists an index l' in the original sum such that $r^{i-l'+1}-r^{l'}=0$. On the fifth row, we make use of the identity $c_t=-c_{i-t+1}$ which is straightforward to verify. From the last row, we can observe that every term of the sum is positive since $i-2t+1$ and $r^{i-t+1}-r^t$ are both positive if $t \leq (i+1)/2$ which holds since t ranges from 1 to $\lfloor i/2\rfloor$.

Let us now consider the situation where $n < i \leq 2N-1$. We start with the special case where $i=2N-1$. Then, we have only one term in the sum

$$\begin{aligned}\sum_{j+l=i}(j+1)a_{j+1}a_l(r^{j+1}-r^l) &= \sum_{l=N}^{N}(2N-1-l+1)a_{2N-1-l+1}a_l(r^{2N-1-l+1}-r^l)\\ &= Na_Na_N(r^N-r^N)\\ &= 0.\end{aligned}$$

Now, let $N < i < 2N-1$, we follow a similar procedure as before to manipulate the sum

$$\begin{aligned}
\sum_{j+l=i}(j+1)a_{j+1}a_l(r^{j+1}-r^l) &= \sum_{l=i-N+1}^{N}(i-l+1)a_{i-l+1}a_l(r^{i-l+1}-r^l) \\
&= \sum_{l=i-N+1}^{N}(i-l+1)c_l \\
&= \sum_{t=1}^{2N-i}(N-t+1)c_{i-N+t} \\
&= \sum_{t=1}^{\lfloor N-i/2\rfloor}(N-t+1)c_{i-N+t} \\
&+ (N-(2N-i-t+1)+1)c_{i-N+(2N-i-t+1)} \\
&= \sum_{t=1}^{\lfloor N-i/2\rfloor}(N-t+1)c_{i-N+t}+(i-N+t)c_{N-t+1} \\
&= \sum_{t=1}^{\lfloor N-i/2\rfloor}(N-t+1)c_{i-N+t}-(i-N+t)c_{i-N+t} \\
&= \sum_{t=1}^{\lfloor N-i/2\rfloor}(N-t+1-(i-N+t))c_{i-N+t} \\
&= \sum_{t=1}^{\lfloor N-i/2\rfloor}(2N-i-2t+1)c_{i-N+t}.
\end{aligned}$$

It is now easy to verify that $(2N-i-2t+1)$ and c_{i-N+t} are positive if $t \leq N-(i-1)/2$ which holds since t ranges from 1 to $\lfloor N-i/2\rfloor$. The floor function is again used when we sum over pairs of terms since if i is odd there is a zero-term. Since all the coefficients are non-negative and the $k \geq 2$, the derivative is positive. This implies that the original function is increasing. □

B.4 fNML is Regular

In this section, we will show that fNML score is regular. Recall first the definition of fNML local score

$$Q_N^{fnml}(X|\,U) = \log P(x_N \mid \hat{\theta}_{X|U}\ ) - \sum_{j=1}^{r_U} reg(N_j, r_X), \tag{3}$$

where r_X is the number of categories for X, r_U denotes the number of possible configurations of variables in U, and N_j is the number of times the j^{th} configuration is observed in our samples u_N. Note that $reg(N_j, r_X) = 0$ for any configuration not observed in our sample. fNML criterion differs from qNML only by how the penalty term is defined. We can follow a highly similar strategy in order to prove the regularity for fNML. To this end, we need the following Lemma:

Lemma 4. *Let $N = N_1 + N_2$ and $r \in \mathbb{N}, r \geq 2$. Now,*

$$reg(N, r) \leq reg(N_1, r) + reg(N_2, r).$$

Proof. We start with the definition of $C(N_1, r) = \exp(reg(N_1, r))$,

$$\begin{aligned} C(N_1, r)C(N_2, r) &= \sum_{x_{N_1}} P(x_{N_1}|\hat{\theta}(x_{N_1})) \sum_{x_{N_2}} P(x_{N_2}|\hat{\theta}(x_{N_2})) \\ &\geq \sum_{x_{N_1}} P(x_{N_1}|\hat{\theta}(x_N)) \sum_{x_{N_2}} P(x_{N_2}|\hat{\theta}(x_N)) \\ &= \sum_{x_N} P(x_N|\hat{\theta}(x_N)) \\ &= C(N, r). \end{aligned}$$

Taking the logarithm on both sides yields our claim. On the second row, we used the definition of maximum likelihood parameters: the probability of the data vector x_{N_1} is maximized under parameters $\hat{\theta}(x_{N_1})$, and if we use different parameters, $\hat{\theta}(x_N)$, the probability can only go down or stay the same. On the third row, we used the i.i.d., assumption, which allows us to take the single sum over data vectors of length $N = N_1 + N_2$. □

Now we can proceed to the actual proof.

Theorem 3. *fNML score is regular.*

Proof. Using the entropy assumption and Lemma 2, we can ignore the maximized likelihood terms and study the penalty terms. We want show that

$$\sum_{j=1}^{r_{Y'}} reg(N'_j, r_X) \leq \sum_{l=1}^{r_Y} reg(N_l, r_X), \tag{4}$$

where we used $r_{Y'}$ and r_Y to denote the numbers of possible configurations for variables in Y' and Y, respectively. Also, N_j refers to the number times the j:th configuration of variables Y is observed in our sample, and N'_j denotes the same for Y'.

Now, since $r_{Y'} \leq r_Y$, and we know that $\sum_j N'_j = \sum_l N_l = N$, we can just apply Lemma (4) multiple times to conclude our proof. □

C qNML coincides with NML for many models

C.1 qNML Equals NML for Many Models

The fNML criterion can be seen as a computationally feasible approximation of the more desirable NML criterion. However, the fNML criterion equals the NML criterion only for the Bayesian network structure with no arcs. We will next show that the qNML criterion equals the NML criterion for all the networks G whose connected components tournaments (i.e., complete directed acyclic subgraphs of G).

Theorem 4. *If G consists of C connected components $(G^1, \dots, G^C)$ with variable sets $(V^1, \dots, V^C)$, then $\log P_{NML}(D; G) = s^{qNML}(D; G)$ for all data sets D.*

Proof. We first show that the NML-criterion for a Bayesian network decomposes by the connected components.

Because the maximum likelihood for the data D decomposes, we can write

$$\begin{aligned}
& P_{NML}(D; G) \\
&= \frac{P(D; \hat{\theta}(D), G)}{\sum_{D'_{V_1}} \cdots \sum_{D'_{V_C}} \prod_{c=1}^{C} P(D'_{V^c}; \hat{\theta}(D'_{V^c}), G)} \\
&= \frac{\prod_{c=1}^{C} P(D_{V^c}; \hat{\theta}(D_{V^c}), G)}{\prod_{c=1}^{C} \sum_{D'_{V^c}} P(D'_{V^c}; \hat{\theta}(D'_{V^c}), G)}. \\
&= \prod_{c=1}^{C} P_{NML}(D_{V^c}; G). \qquad (5)
\end{aligned}$$

Clearly, the qNML score also decomposes by the connected components, so it remains to show that if the (sub)network G is a tournament, then for any data D, $s^{qNML}(D; G) = \log P_{NML}(D; G)$. Due to the score equivalence of the NML criterion and the qNML criterion, we may pick a tournament G such that the linear ordering defined by G matches the ordering of the data columns, i.e., $i < j$ implies $G_i \subset G_j$. Now from the definition (8) of the qNML criterion (in the main paper), we see that for the tournament G, the sum telescopes leaving us with $s^{qNML}(D; G) = \log P^1_{NML}(D_G; G)$, thus it is enough to show that $P^1_{NML}(D; G) = P_{NML}(D; G)$. This follows, since for any data vector x in data D, we have $P^1(x;\hat{\theta}(D),G) = P(x;\hat{\theta}(D),G)$, where P^1 denotes the model that takes n-dimensional vectors to be values of the single (collapsed) categorical variable. Denoting prefixes of data vector x by $x^{:i}$, and the number of times such a prefix appears on N rows $[d_1,\ldots,d_N]$ of the $N \times n$ data matrix D by $N_D(x^{:i})$, (so that $N_D(x^{:0}) = N$), we have

$$\begin{aligned}
P(D;\hat{\theta}(D),G) &= \prod_{j=1}^{N} P(d_j;\hat{\theta}(D),G) \\
&= \prod_{j=1}^{N}\prod_{i=1}^{n} \frac{N_D(d_j^{:i})}{N_D(d_j^{:i-1})} \\
&= \prod_{j=1}^{N} \frac{N_D(d_j^{:n})}{N} \\
&= P^1(D;\hat{\theta}^1(D),G). \qquad (6)
\end{aligned}$$

Since both P^1_{NML} and P_{NML} are defined in terms of the maximum likelihood probabilities, the equality above implies the equality of these distributions.

□

Equality established, we would like to still state the number $a(n)$ of different n-node networks whose connected components are tournaments. We start by generating all the $p(n)$ integer partitions i.e. ways to partition n labelled into parts with different size-profiles. For example, for $n = 4$, we have $p(4) = 5$ partition profiles $[[4],[3,1],[2,2],[2,1,1],[1,1,1,1]]$. Each of these partition size profiles corresponds to many different networks, apart from the last one that corresponds just to the empty network. We count number of networks for one such partition size-profile (and then later sum these counts up). For any such partition profile $(p_1,\ldots,p_k)$ we can count the ways we can assign the nodes to different parts and then order each part. This leads to the product $\binom{n}{p_1}p_1!\binom{n-p_1}{p_2}p_2!\binom{n-p_1-p2}{p_3}p_3!\ldots\binom{n-\sum_{j=1}^{k-1}p_j}{p_k}p_k!$. However, the order of different parts of the same size does not matter, so for all groups of parts having the same size, we have to divide the product above by the factorial of the size of such group. Notice also that the product above telescopes, leaving us a formula for OEIS sequence A000262[1] as described by Thomas Wieder:

[1] https://oeis.org/A000262

With $p(n)$ = the number of integer partitions of n, $d(i)$ = the number of different parts of the i^{th} partition of n, $m(i,j)$ = multiplicity of the j^{th} part of the i^{th} partition of n, one has:

$$a(n) = \sum_{i=1}^{p(n)} \frac{n!}{\prod_{j=1}^{d(i)} m(i,j)!}.$$

For example, $a(4) = \frac{4!}{1!} + \frac{4!}{1!1!} + \frac{4!}{2!} + \frac{4!}{1!2!} + \frac{4!}{4!} = 73$. In general this sequence grows rapidly; $1, 1, 3, 13, 73, 501, 4051, 37633, 394353, 4596553, \ldots$.

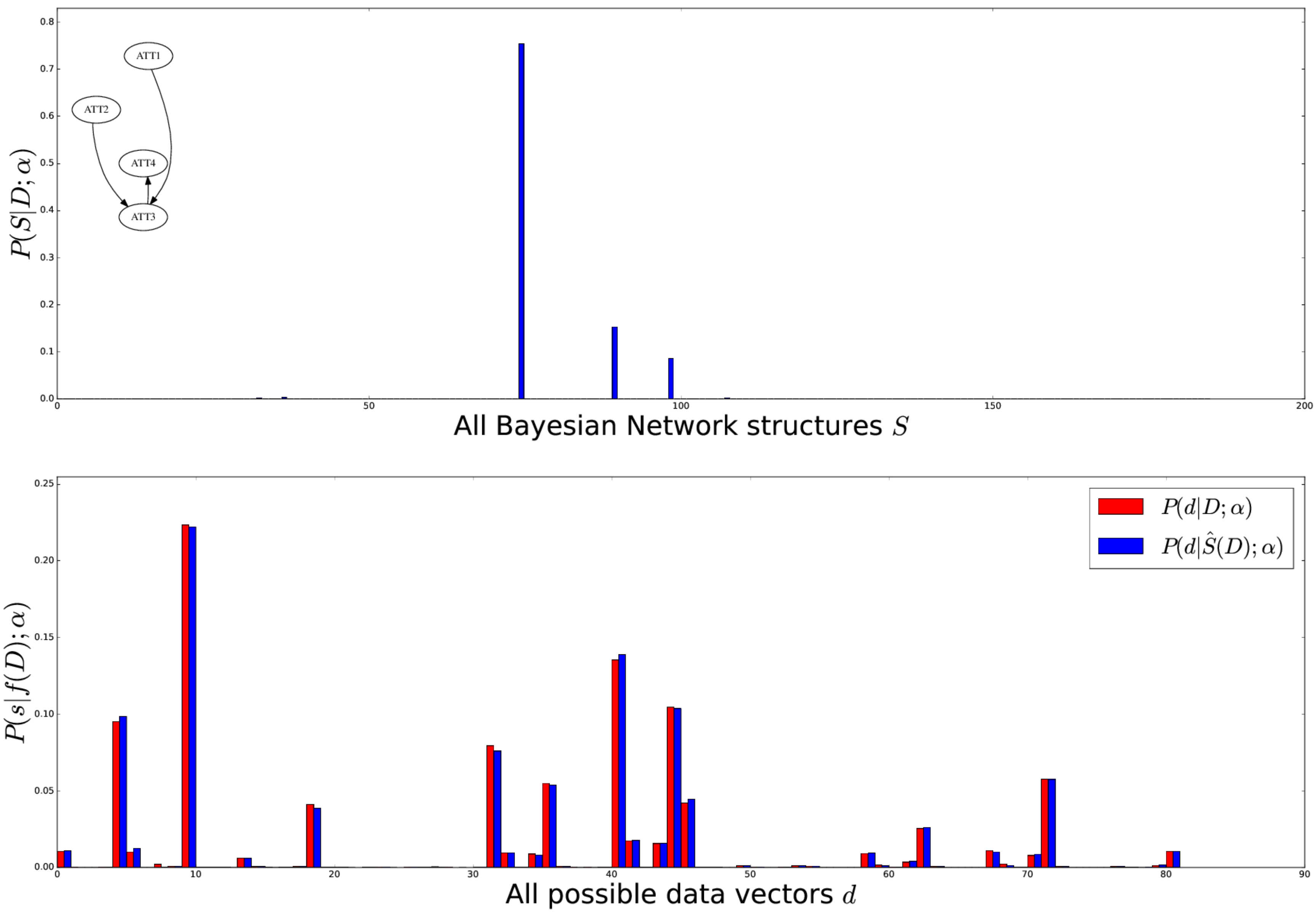
D Bayesian Dirichlet quotient score

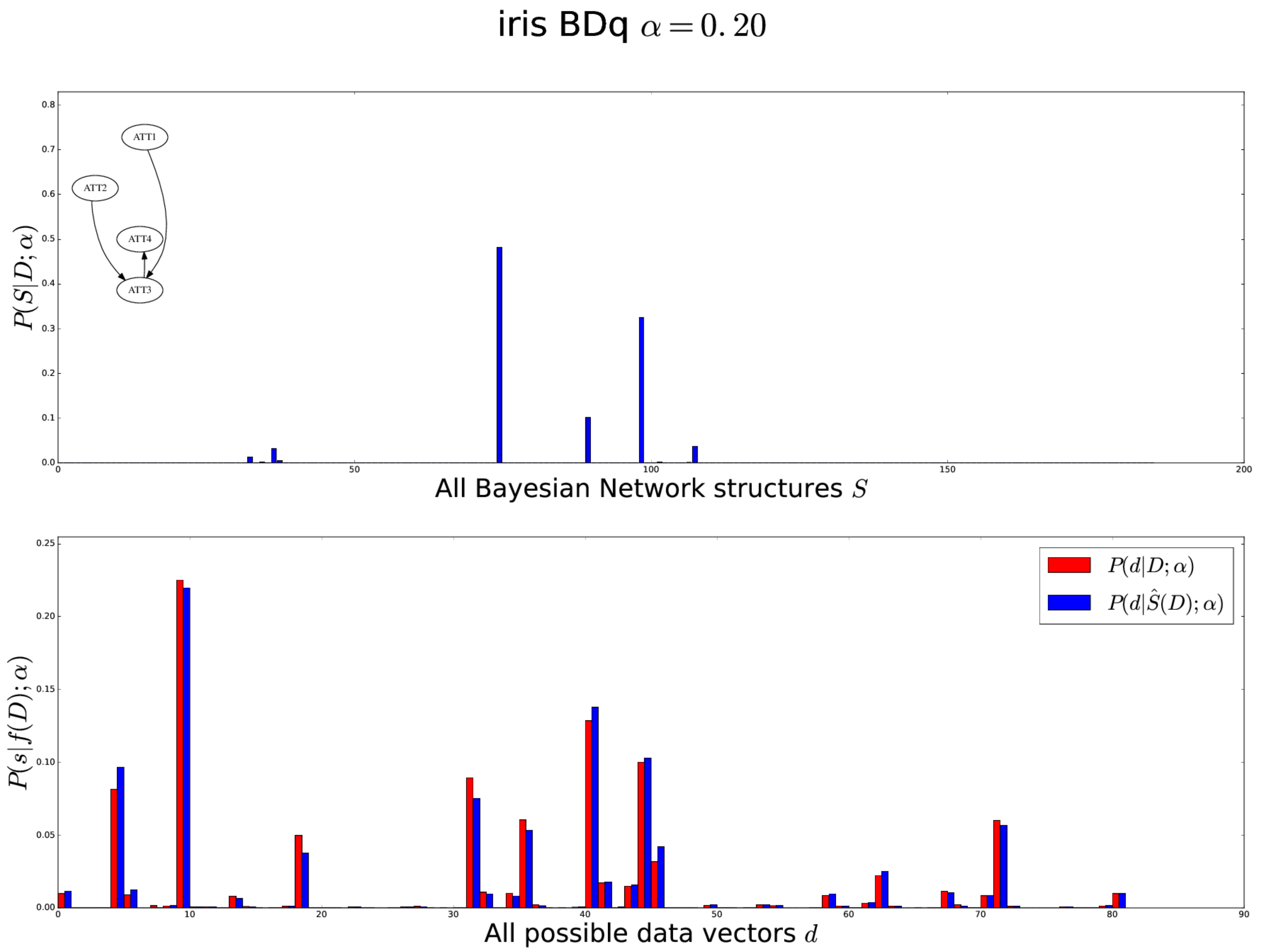
Suzuki (2017) independently suggested a Bayesian Dirichlet quotient score BDq, in which the NML-distributions $P^1_{NML}(D_{i,G_i}; G)$ and $P^1_{NML}(D_{G_i}; G)$ of the qNML are replaced by the Bayesian marginal likelihoods $P^1(D_{i,G_i}; G, \alpha)$ and $P^1_{NML}(D_{G_i}; G, \alpha)$ using parameter prior $\theta_{ij} \sim Dirichlet(\alpha, \ldots, \alpha)$. Suzuki further suggested using the Jeffreys' prior with $\alpha = \frac{1}{2}$. While this is a popular choice in statistics, it is also possible to argue for other values of α like $\alpha = \frac{1}{2} - \frac{\ln 2}{2 \ln N}$ (Watanabe and Roos, 2015) and $\alpha = \frac{1}{3}$ (Jääsaari et al., 2018).

Our initial experiments indicate that model selection by BDq is still highly sensitive to the hyperparameter α. On the next 5 pages we present results for learning the posterior distribution of the network structures for the four predictor variables (discretized to three equal width bins) in the iris data (N=150). The structure prior is assumed to be uniform. Since BDq is score equivalent, we only include one representative from each equivalence class. The images also show the most probable structure and the predictive distribution given by the most probable structure and the predictive distribution when marginalizing over all the structures.

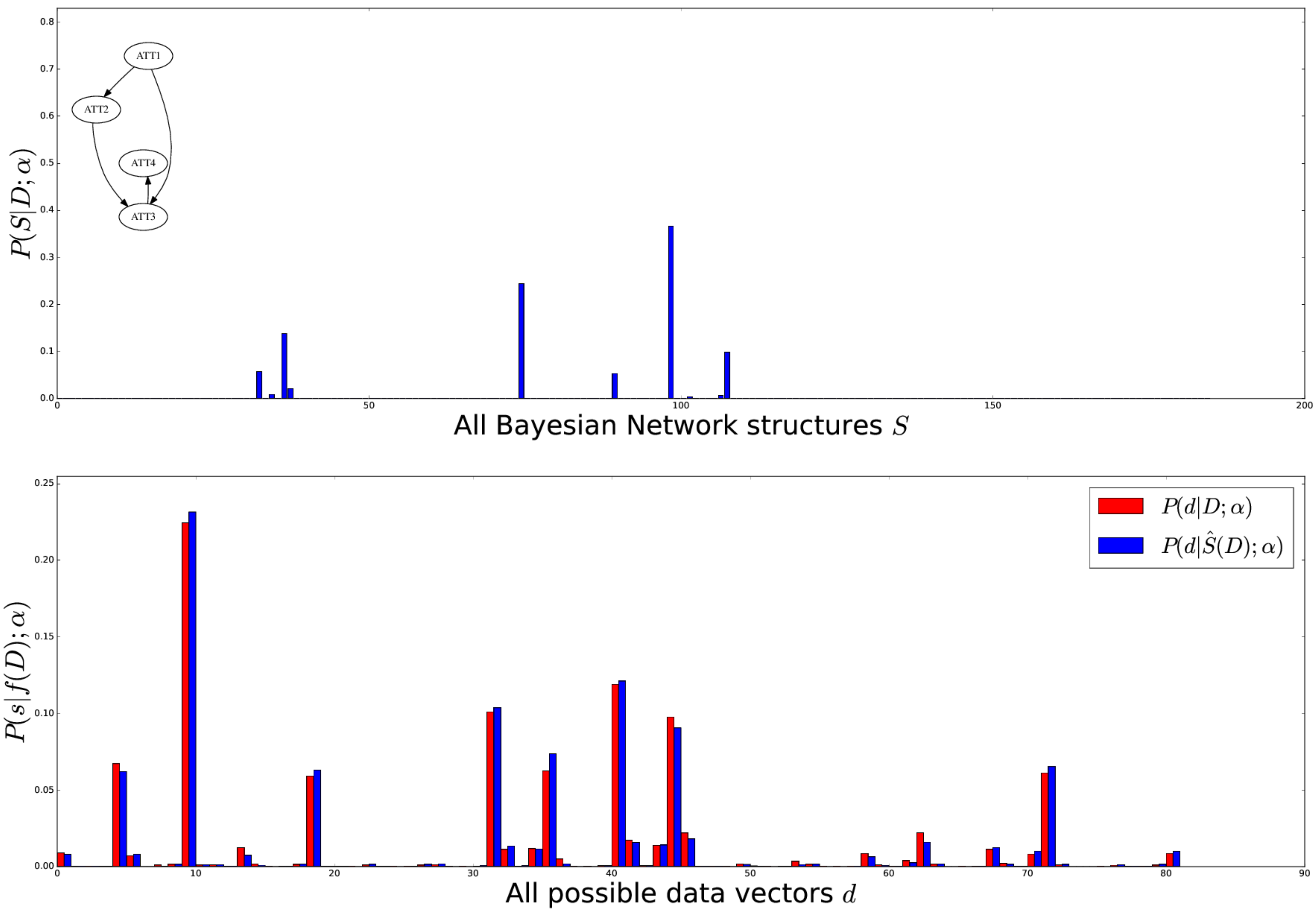
One can see how the structure posterior is quite sensitive to α and the selected network is different for α-values 0.2, 0.3 and 0.5.

iris BDq $\alpha = 0.10$

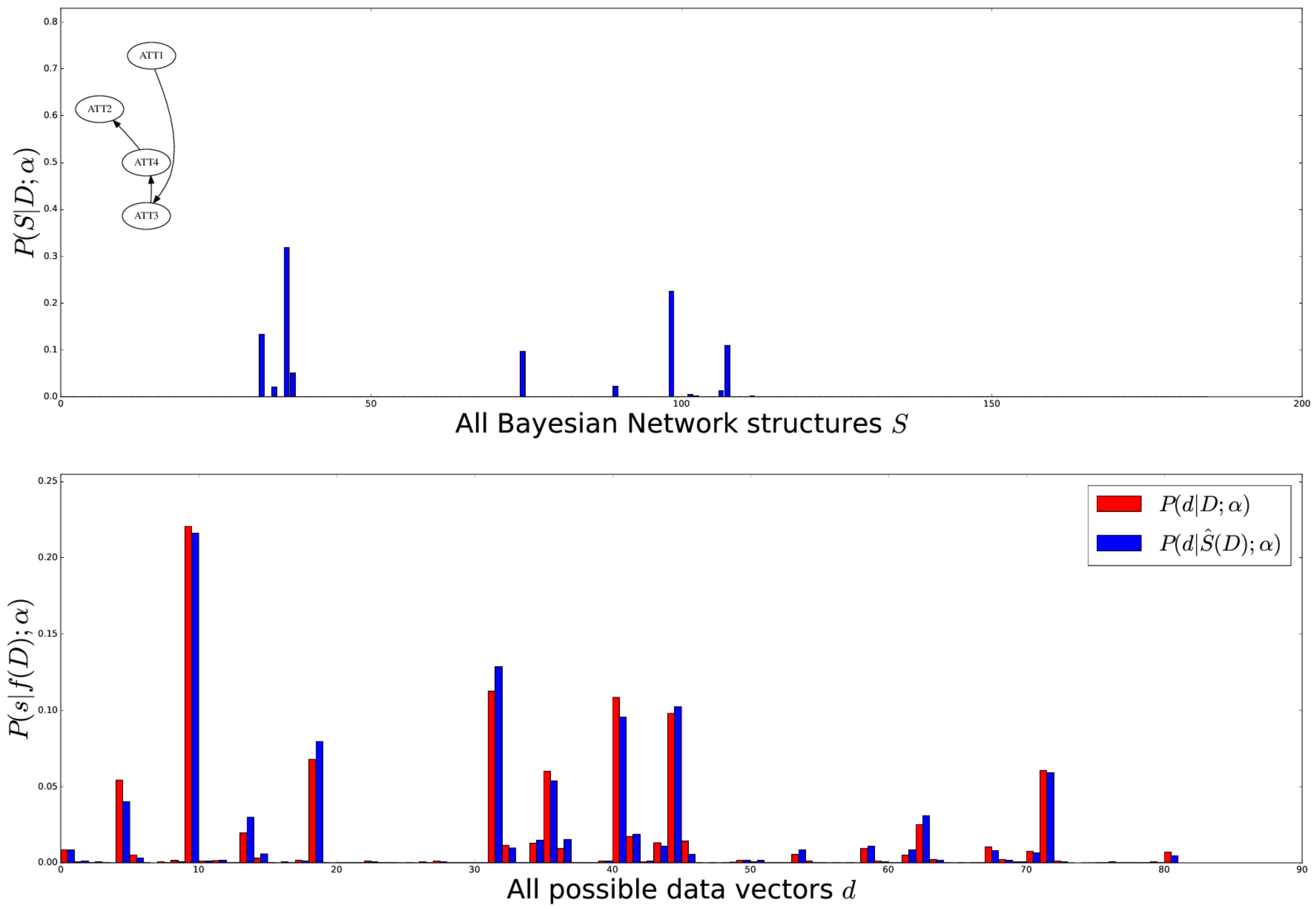



iris BDq $\alpha = 0.20$
$P(S|D;\alpha)$
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
ATT1
ATT2
ATT4
ATT3
0
50
100
150
200
All Bayesian Network structures S
$P(s|f(D);\alpha)$
0.00
0.05
0.10
0.15
0.20
0.25
$P(d|D;\alpha)$
$P(d|\hat{S}(D);\alpha)$
0
10
20
30
40
50
60
70
80
90
All possible data vectors d

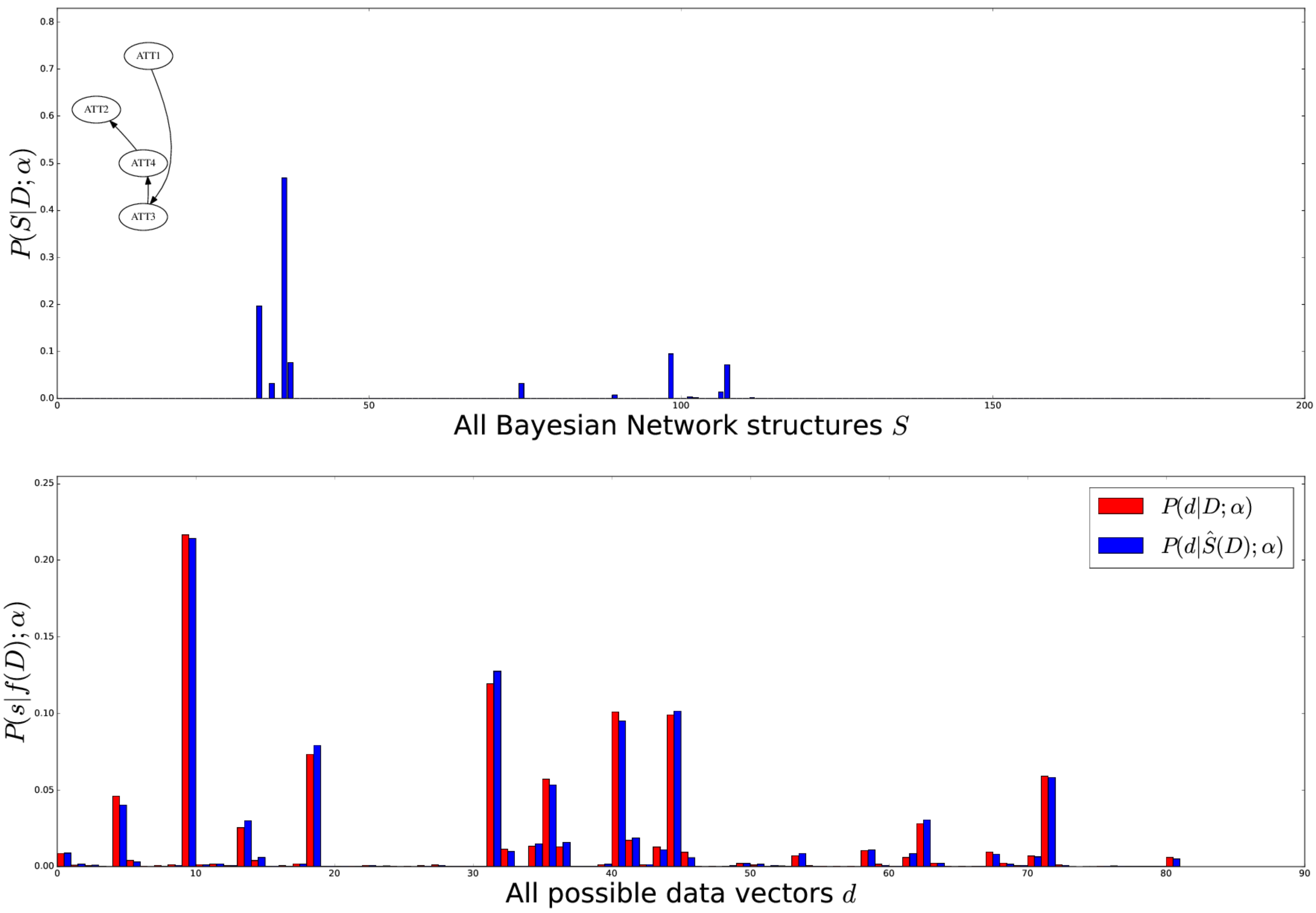
iris BDq $\alpha = 0.30$



iris BDq $\alpha = 0.40$

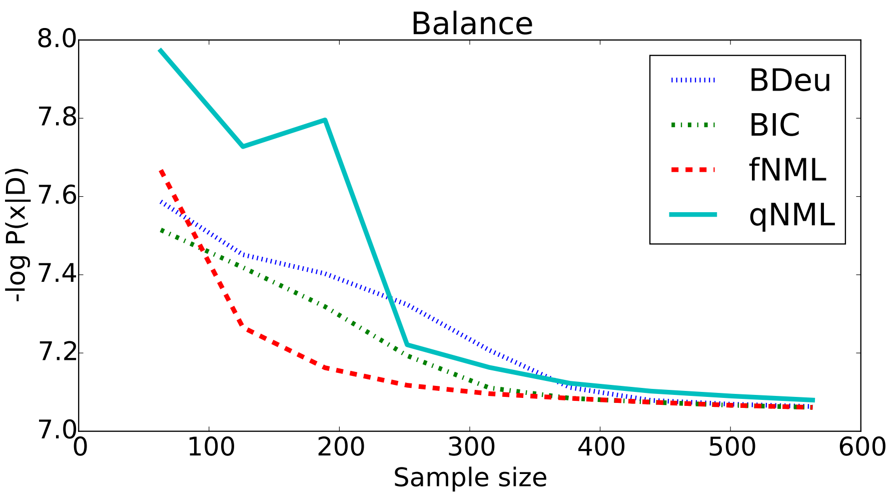


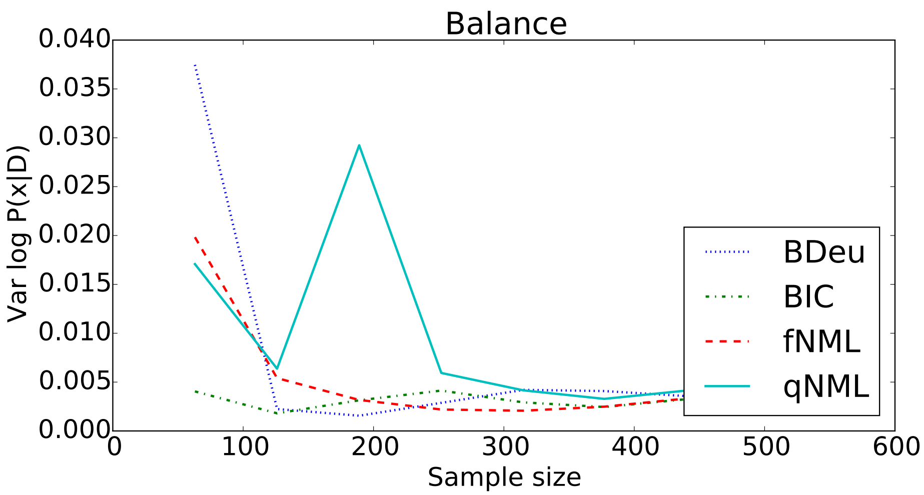
iris BDq $\alpha = 0.50$

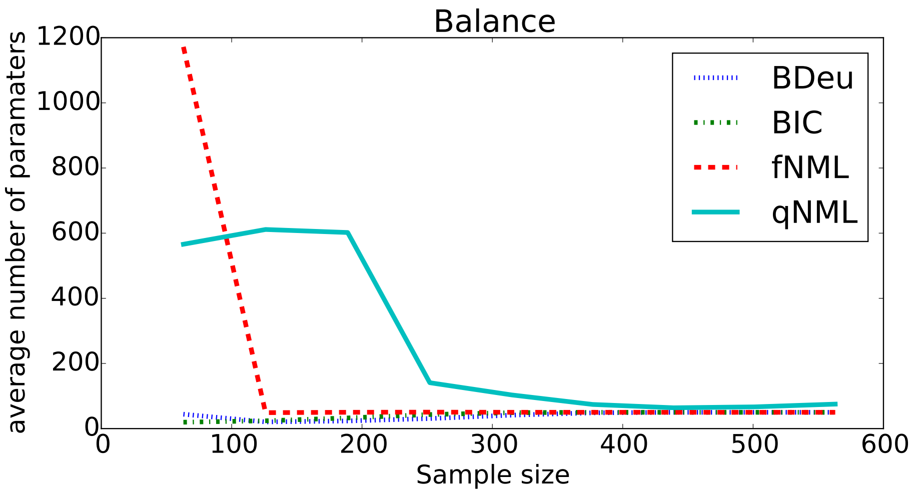


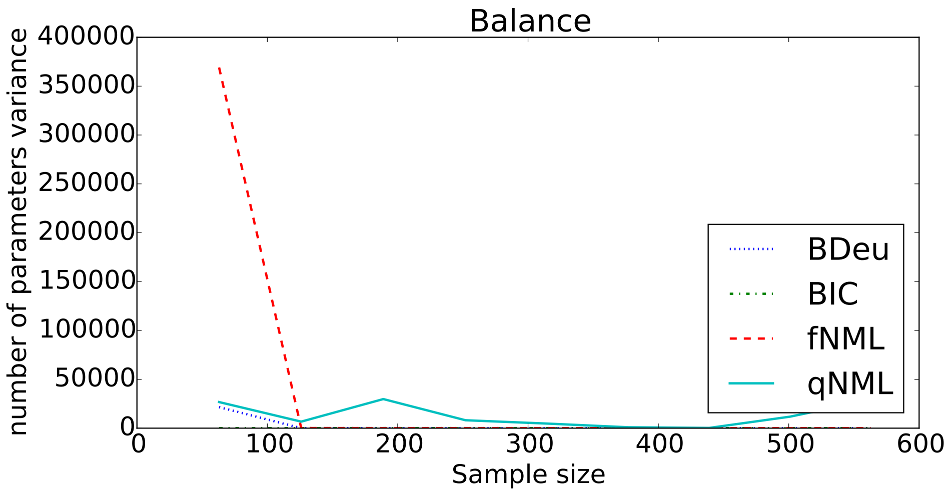
E Prediction and Parsimony

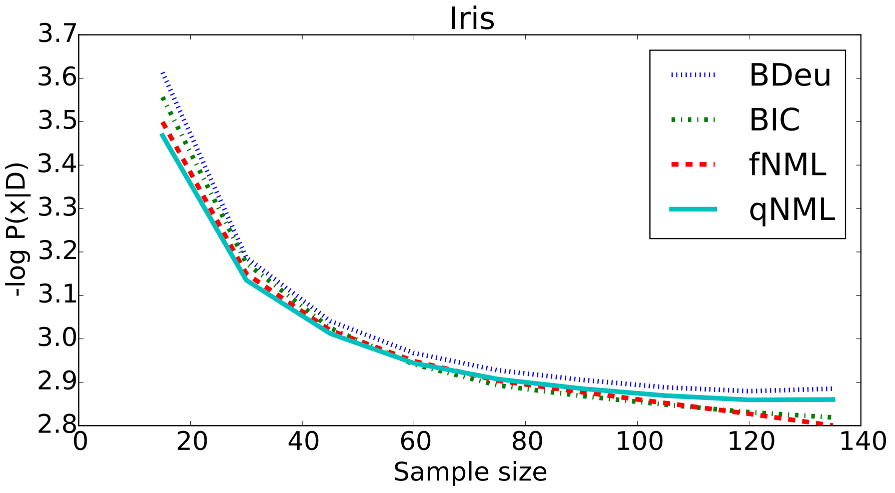
We took 20 UCI data sets and created 1000 permutations of each data set. We then took the first $x\%$ of each permuted data set (x in $10, 20, 30, \ldots, 90$) as training data and used the remaining part as test data. We used exact structure learning to learn the best scoring model using BDeu ($\alpha = 1$), BIC, fNML and qNML, recording the number of parameters in the learnt models and the average predictive log-probability of each test vector. These numbers are collected for 1000 different data permutations, the average and variance of which appear on the following pages, one page per data set.

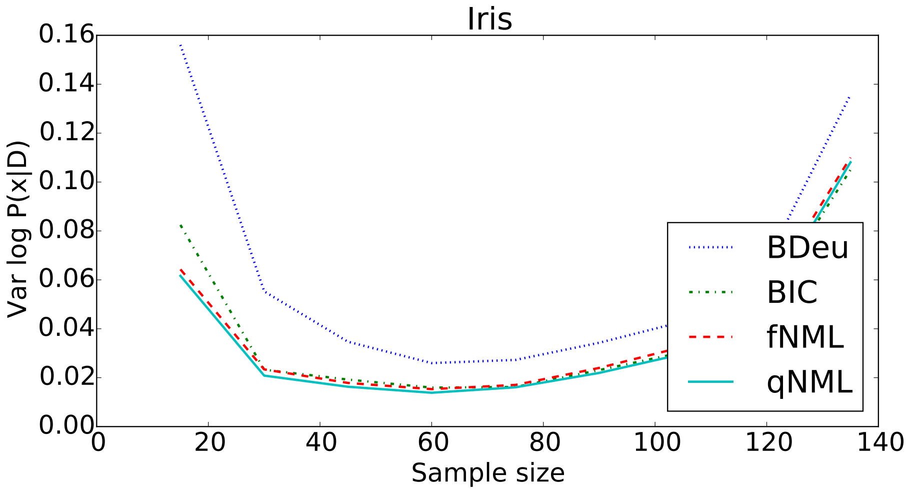

Balance
-log P(x|D)
8.0
7.8
7.6
7.4
7.2
7.0
0
100
200
300
400
500
600
Sample size
BDeu
BIC
fNML
qNML

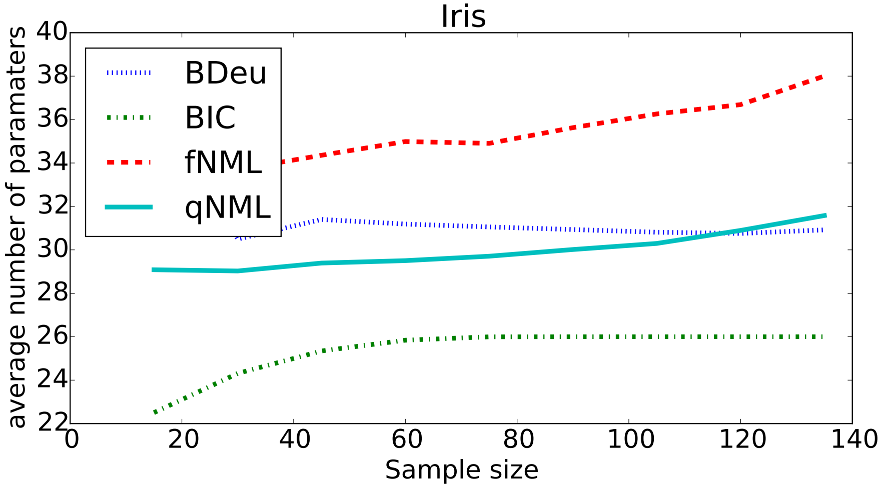

Balance
Var log P(x|D)
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
100
200
300
400
500
600
Sample size
BDeu
BIC
fNML
qNML

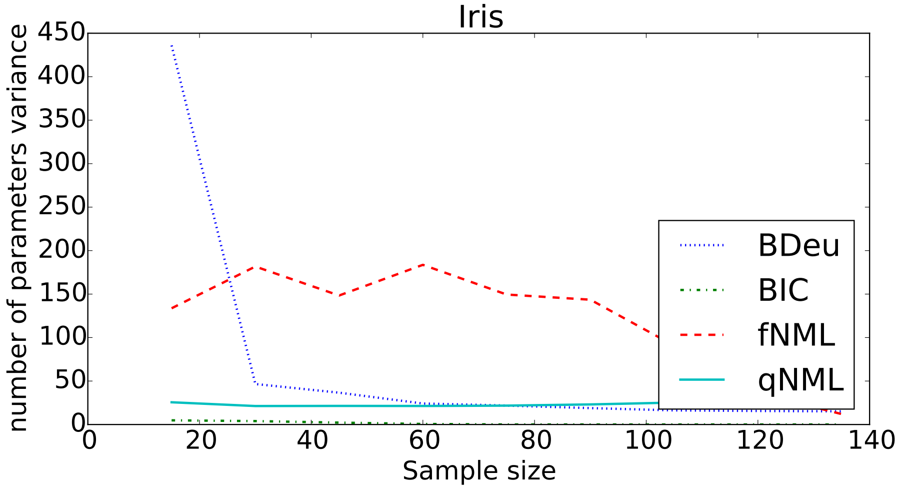

Balance
average number of paramaters
1200
1000
800
600
400
200
0
0
100
200
300
400
500
600
Sample size
BDeu
BIC
fNML
qNML

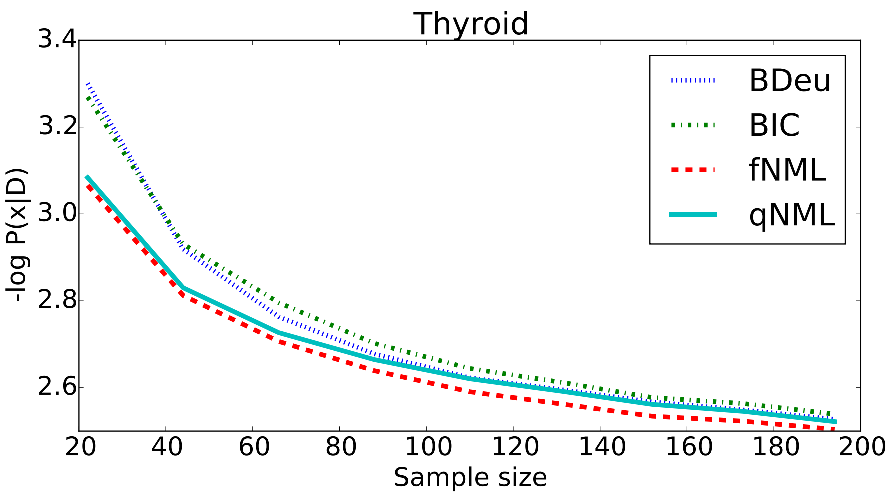

Balance
number of parameters variance
400000
350000
300000
250000
200000
150000
100000
50000
0
0
100
200
300
400
500
600
Sample size
BDeu
BIC
fNML
qNML

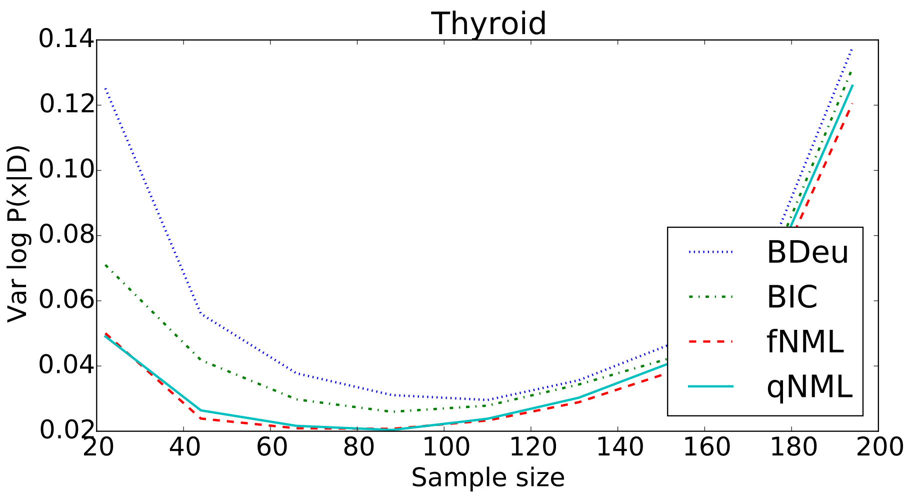

Iris
-log P(x|D)
3.7
3.6
3.5
3.4
3.3
3.2
3.1
3.0
2.9
2.8
0
20
40
60
80
100
120
140
Sample size
BDeu
BIC
fNML
qNML

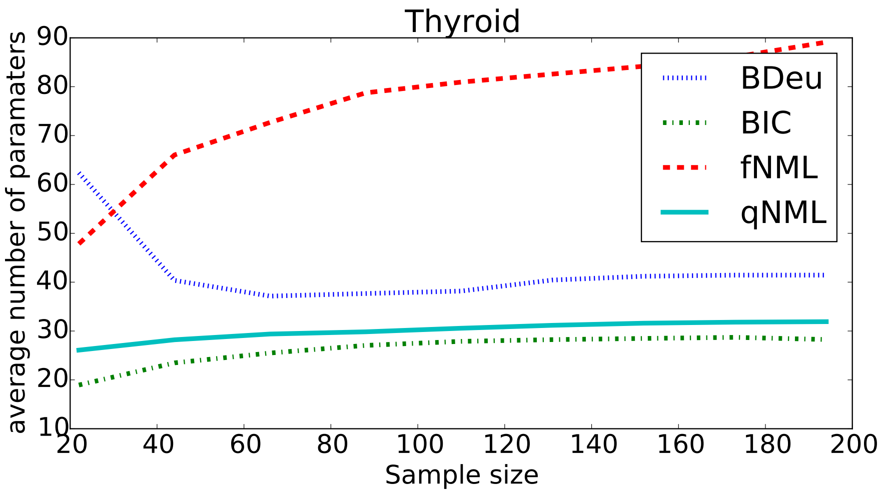

Iris
Var log P(x|D)
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
20
40
60
80
100
120
140
Sample size
BDeu
BIC
fNML
qNML

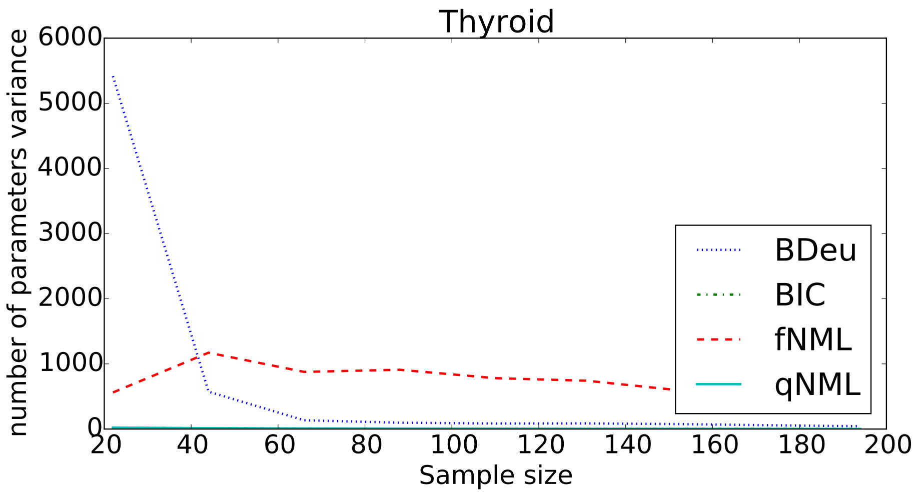

Iris
average number of paramaters
40
38
36
34
32
30
28
26
24
22
0
20
40
60
80
100
120
140
Sample size
BDeu
BIC
fNML
qNML

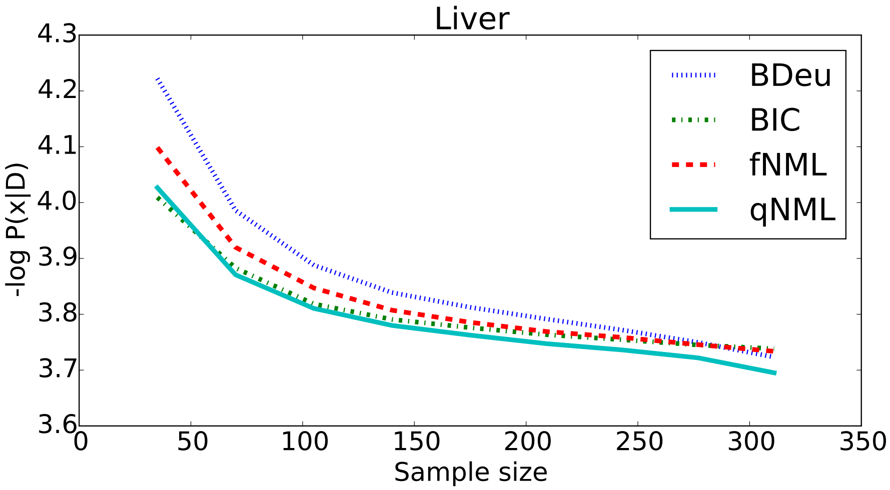

Iris
number of parameters variance
450
400
350
300
250
200
150
100
50
0
0
20
40
60
80
100
120
140
Sample size
BDeu
BIC
fNML
qNML

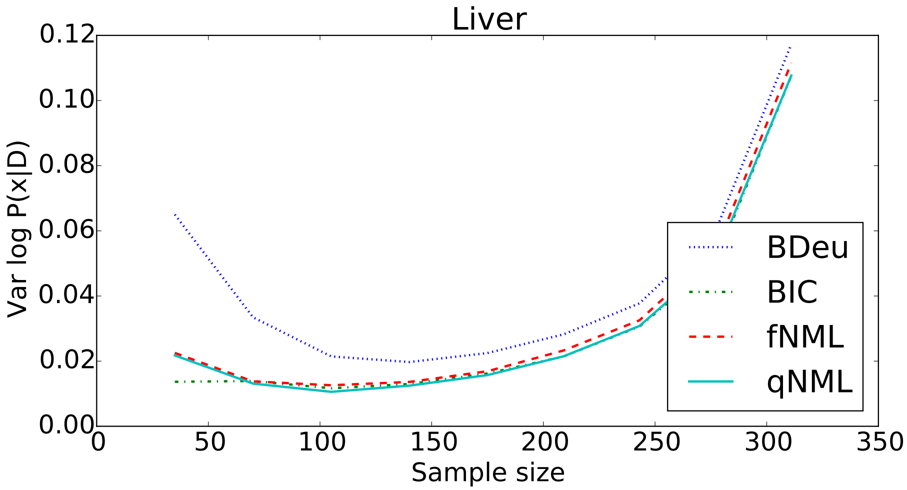

Thyroid
-log P(x|D)
3.4
3.2
3.0
2.8
2.6
20
40
60
80
100
120
140
160
180
200
Sample size
BDeu
BIC
fNML
qNML

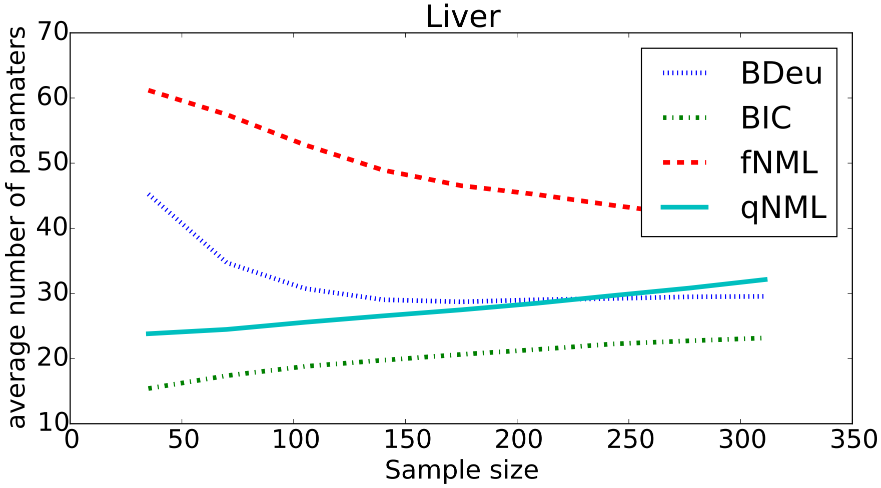

Thyroid
Var log P(x|D)
0.14
0.12
0.10
0.08
0.06
0.04
0.02
20
40
60
80
100
120
140
160
180
200
Sample size
BDeu
BIC
fNML
qNML

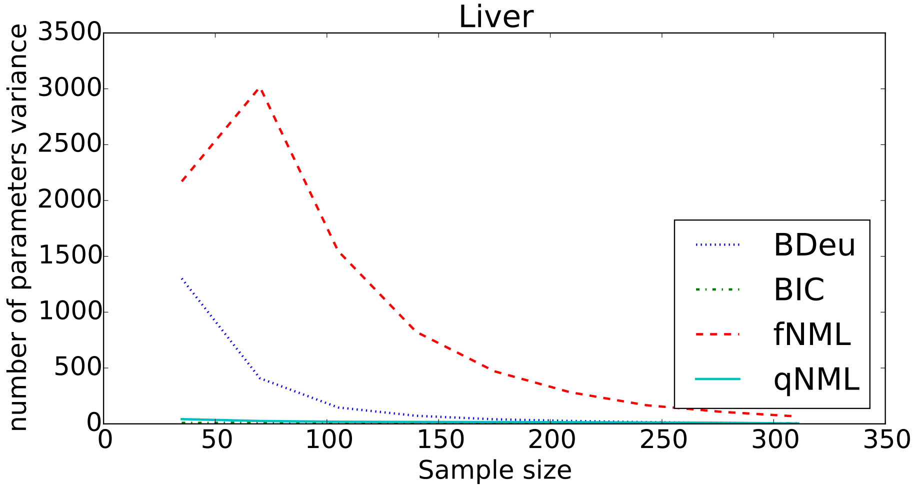

Thyroid
average number of paramaters
90
80
70
60
50
40
30
20
10
20
40
60
80
100
120
140
160
180
200
Sample size
BDeu
BIC
fNML
qNML

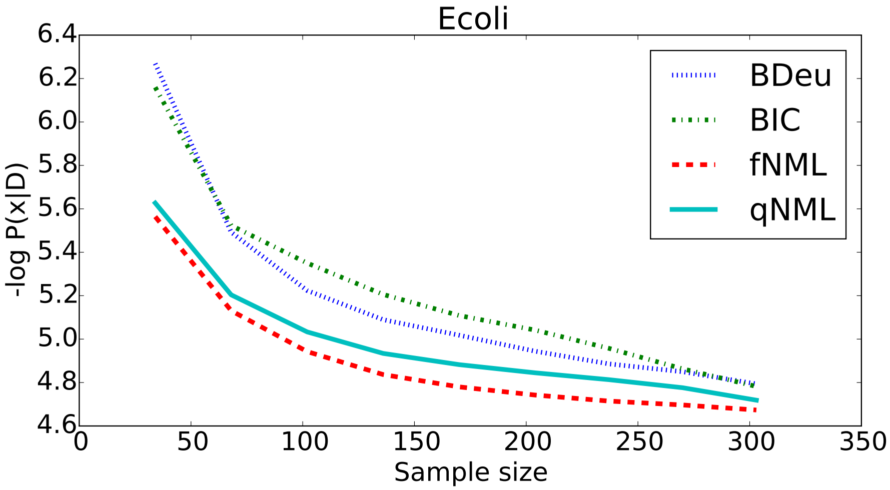

Thyroid
number of parameters variance
6000
5000
4000
3000
2000
1000
0
20
40
60
80
100
120
140
160
180
200
Sample size
BDeu
BIC
fNML
qNML

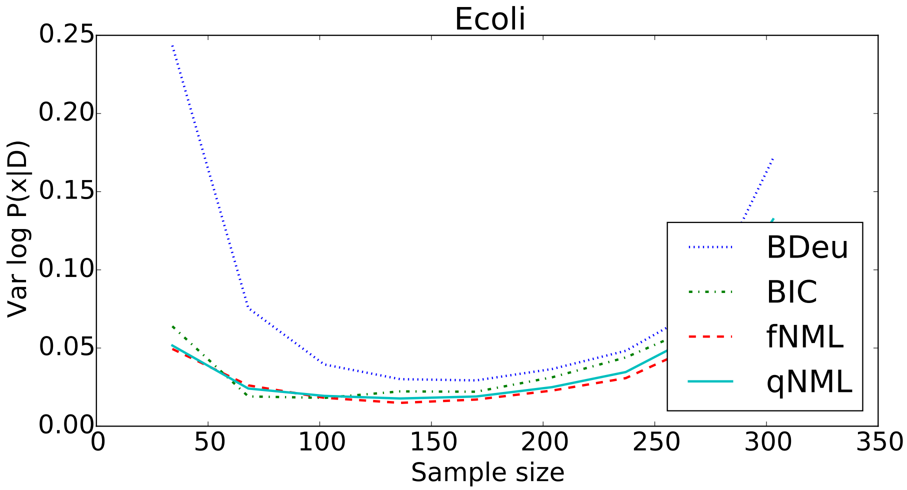

Liver
-log P(x|D)
4.3
4.2
4.1
4.0
3.9
3.8
3.7
3.6
0
50
100
150
200
250
300
350
Sample size
BDeu
BIC
fNML
qNML

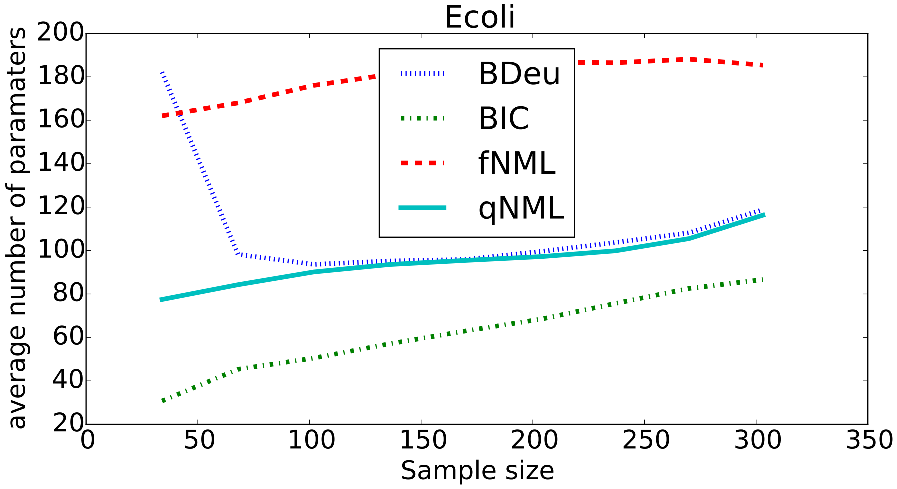

Liver
Var log P(x|D)
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
50
100
150
200
250
300
350
Sample size
BDeu
BIC
fNML
qNML

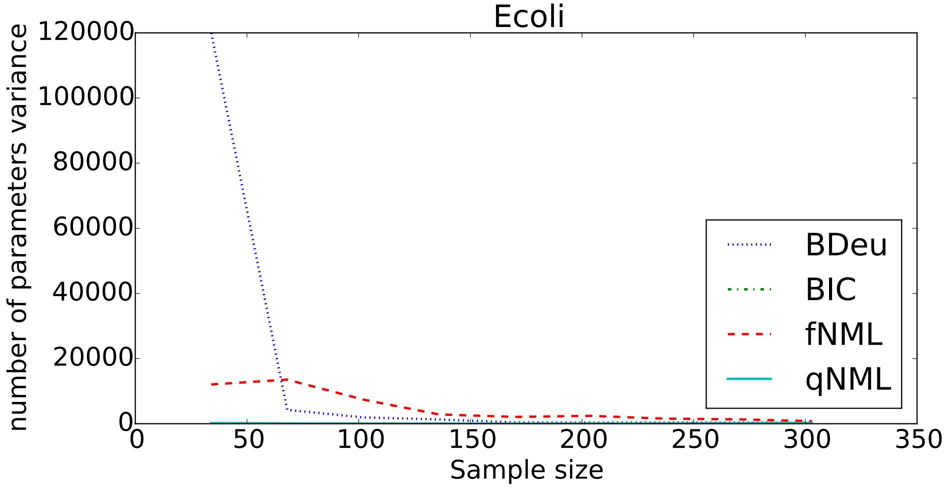

Liver
average number of paramaters
70
60
50
40
30
20
10
0
50
100
150
200
250
300
350
Sample size
BDeu
BIC
fNML
qNML

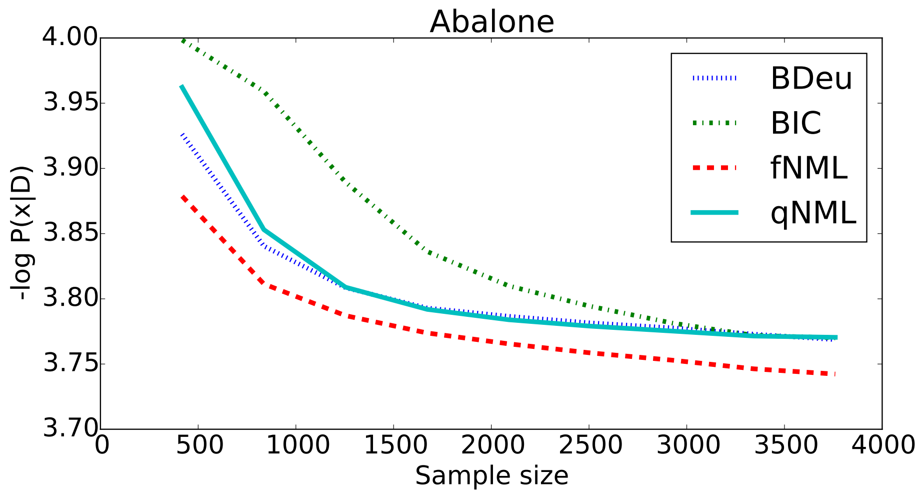

Liver
number of parameters variance
3500
3000
2500
2000
1500
1000
500
0
0
50
100
150
200
250
300
350
Sample size
BDeu
BIC
fNML
qNML

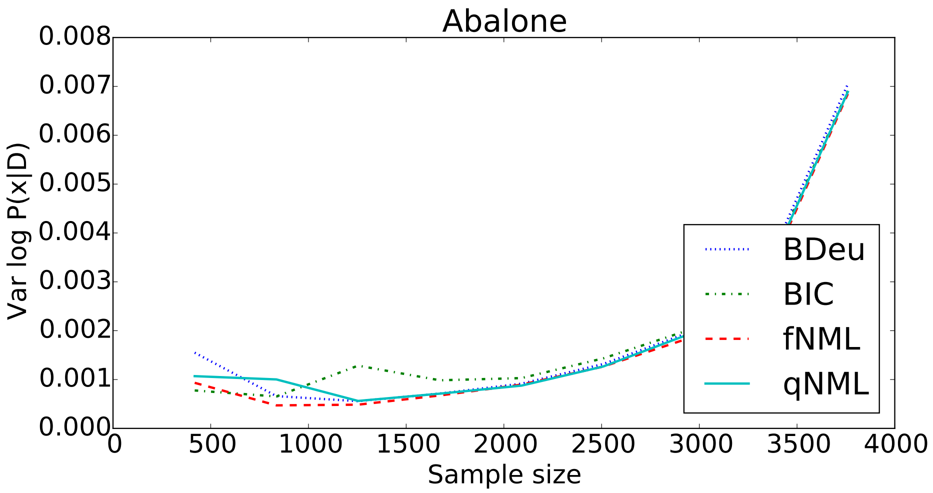

Ecoli
-log P(x|D)
6.4
6.2
6.0
5.8
5.6
5.4
5.2
5.0
4.8
4.6
0
50
100
150
200
250
300
350
Sample size
BDeu
BIC
fNML
qNML

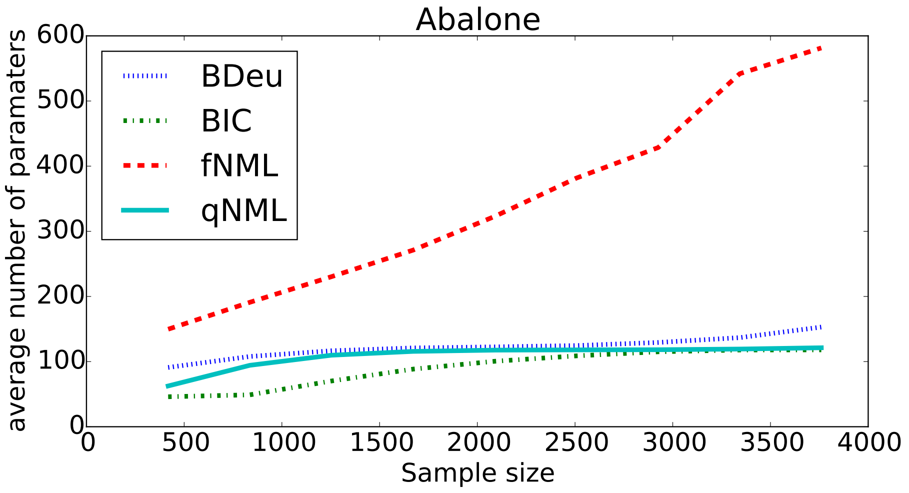

Ecoli
Var log P(x|D)
0.25
0.20
0.15
0.10
0.05
0.00
0
50
100
150
200
250
300
350
Sample size
BDeu
BIC
fNML
qNML

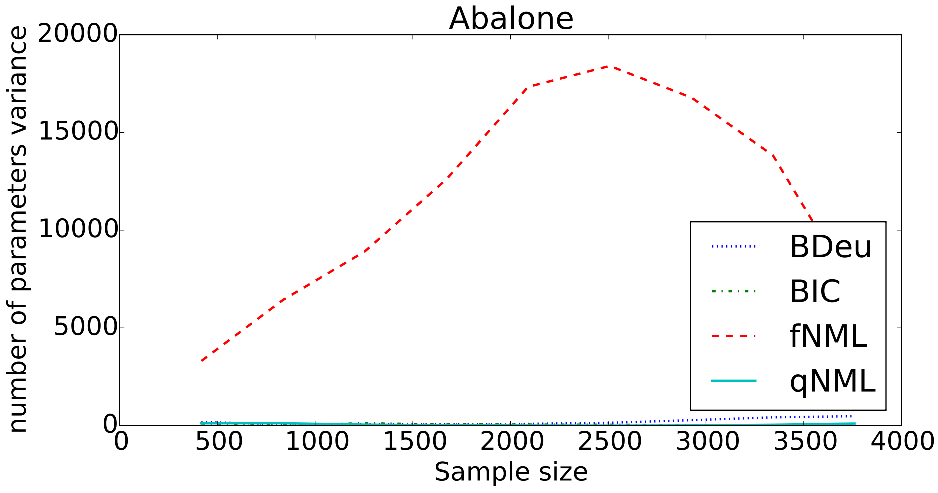

Ecoli
average number of paramaters
200
180
160
140
120
100
80
60
40
20
0
50
100
150
200
250
300
350
Sample size
BDeu
BIC
fNML
qNML

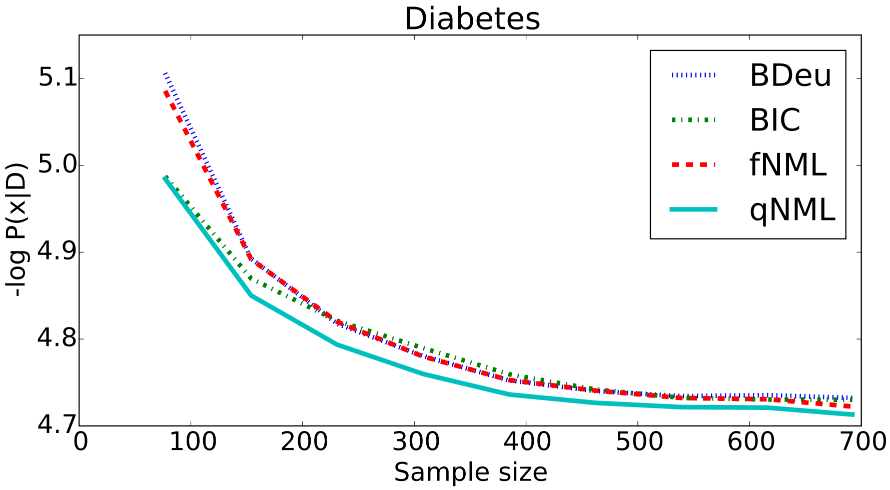

Ecoli
number of parameters variance
120000
100000
80000
60000
40000
20000
0
0
50
100
150
200
250
300
350
Sample size
BDeu
BIC
fNML
qNML

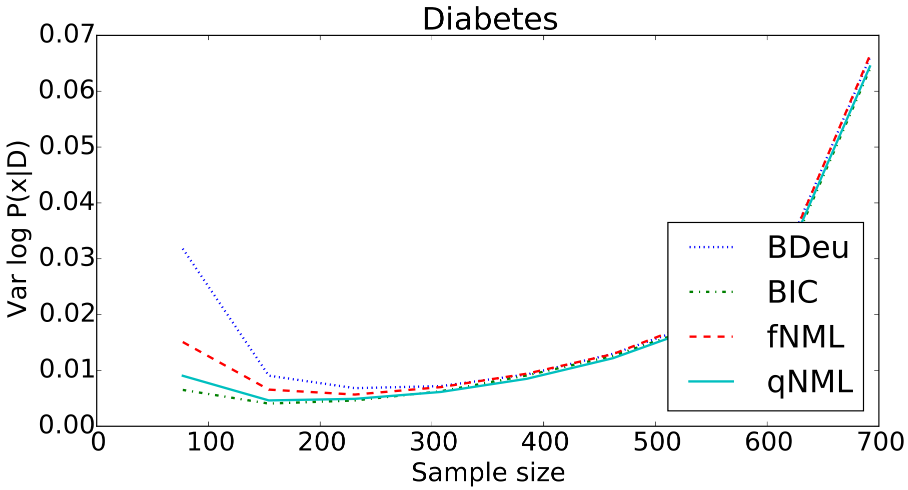

Abalone
-log P(x|D)
4.00
3.95
3.90
3.85
3.80
3.75
3.70
0
500
1000
1500
2000
2500
3000
3500
4000
Sample size
BDeu
BIC
fNML
qNML

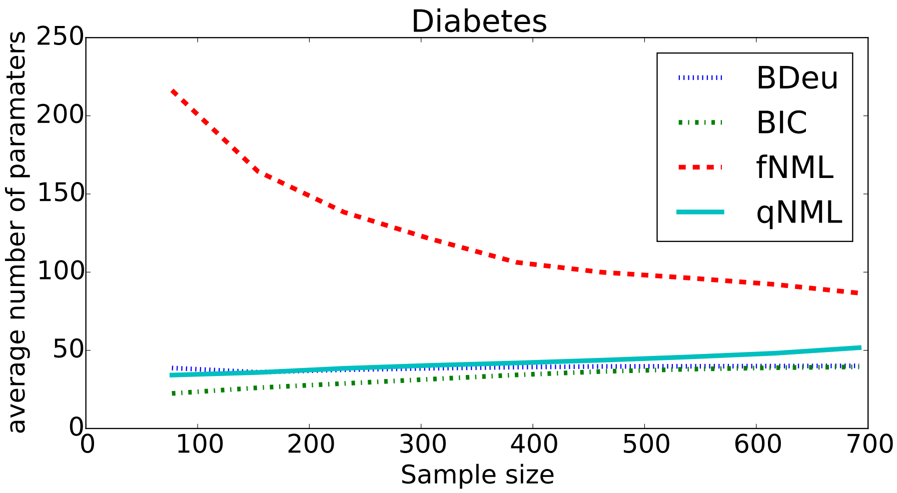

Abalone
Var log P(x|D)
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
500
1000
1500
2000
2500
3000
3500
4000
Sample size
BDeu
BIC
fNML
qNML

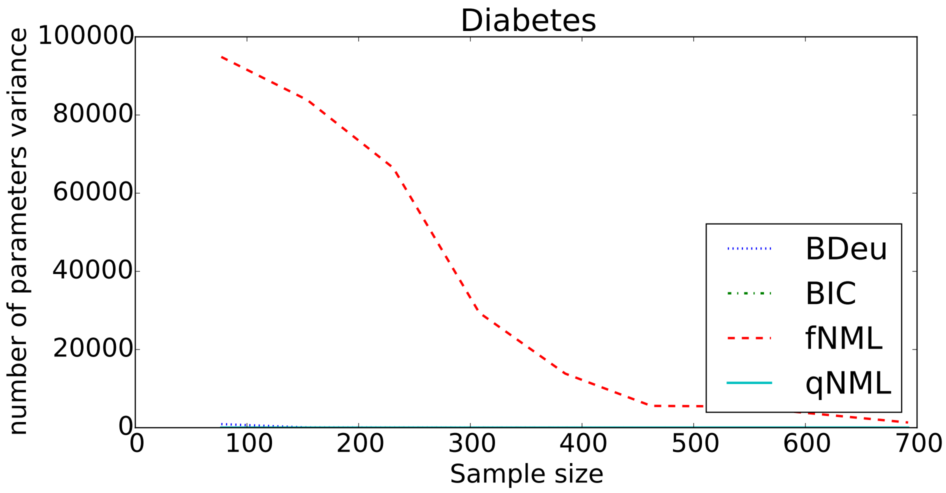

Abalone
average number of paramaters
600
500
400
300
200
100
0
0
500
1000
1500
2000
2500
3000
3500
4000
Sample size
BDeu
BIC
fNML
qNML

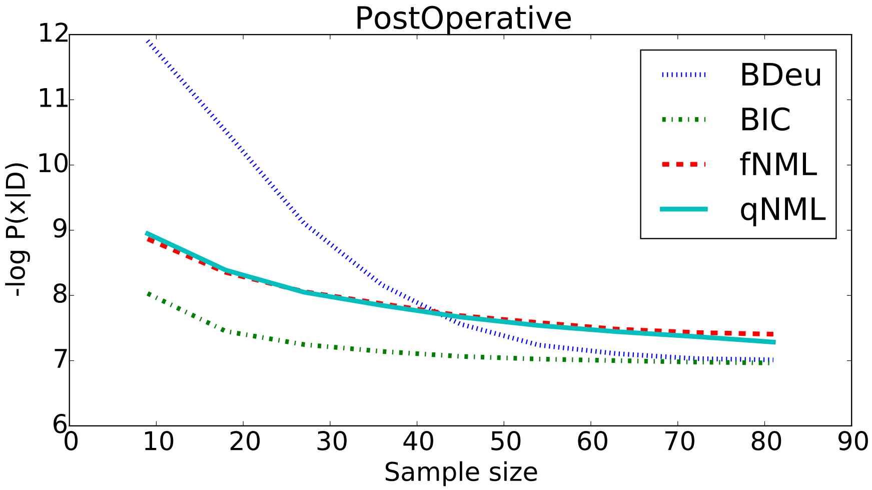

Abalone
number of parameters variance
20000
15000
10000
5000
0
0
500
1000
1500
2000
2500
3000
3500
4000
Sample size
BDeu
BIC
fNML
qNML

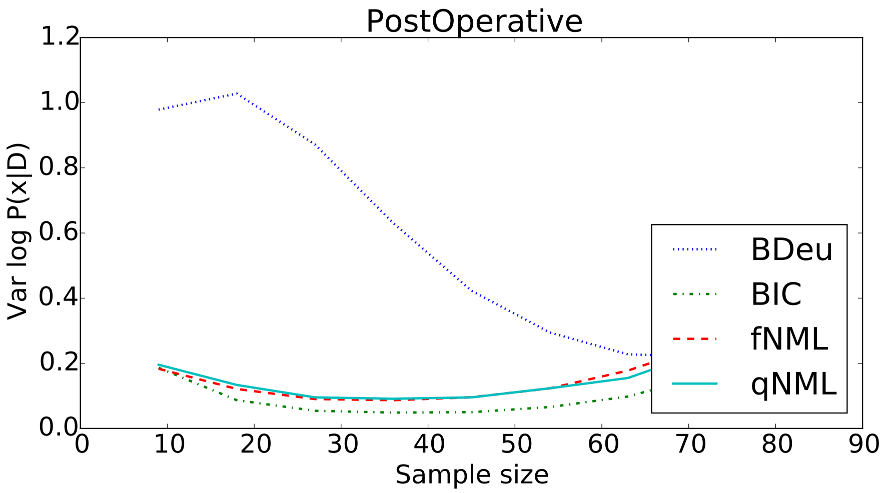

Diabetes
-log P(x|D)
5.1
5.0
4.9
4.8
4.7
0
100
200
300
400
500
600
700
Sample size
BDeu
BIC
fNML
qNML

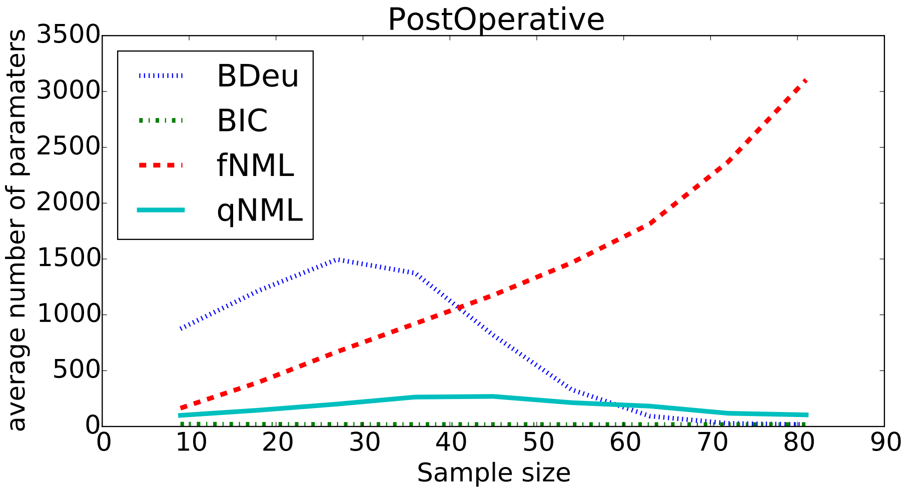

Diabetes
Var log P(x|D)
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
100
200
300
400
500
600
700
Sample size
BDeu
BIC
fNML
qNML

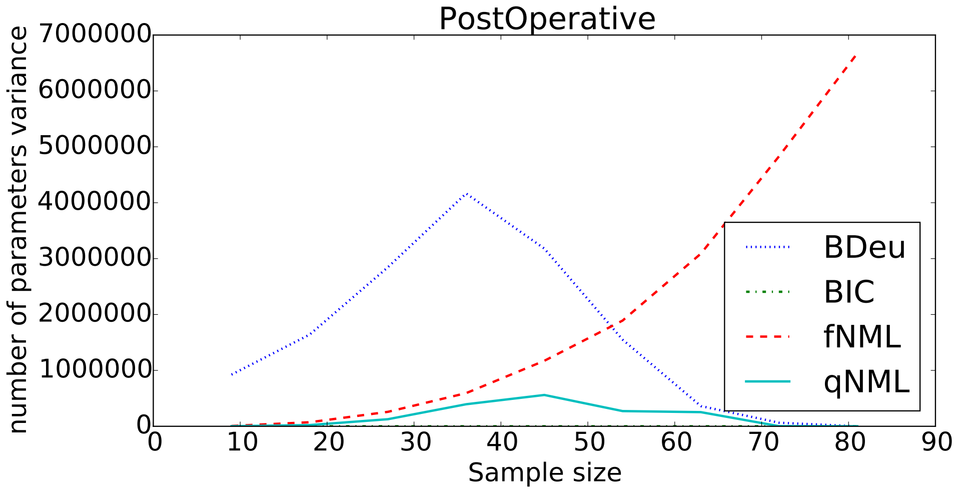

Diabetes
average number of paramaters
250
200
150
100
50
0
0
100
200
300
400
500
600
700
Sample size
BDeu
BIC
fNML
qNML

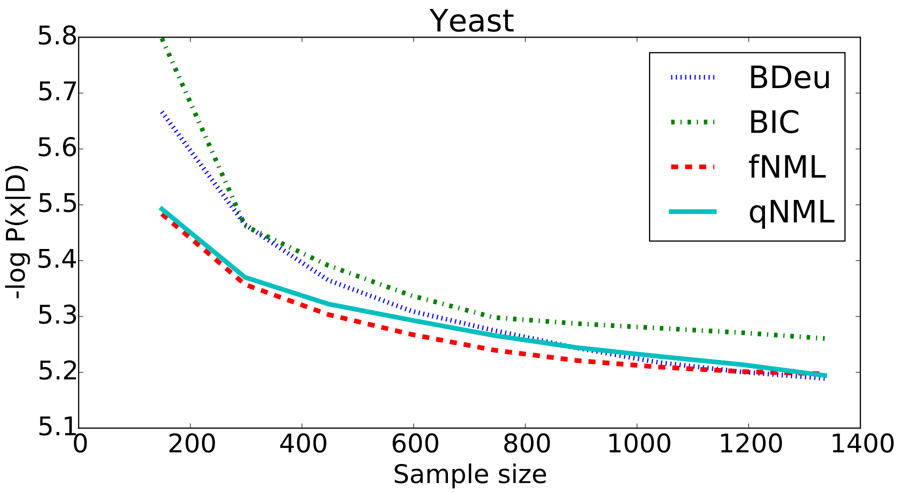

Diabetes
number of parameters variance
100000
80000
60000
40000
20000
0
0
100
200
300
400
500
600
700
Sample size
BDeu
BIC
fNML
qNML

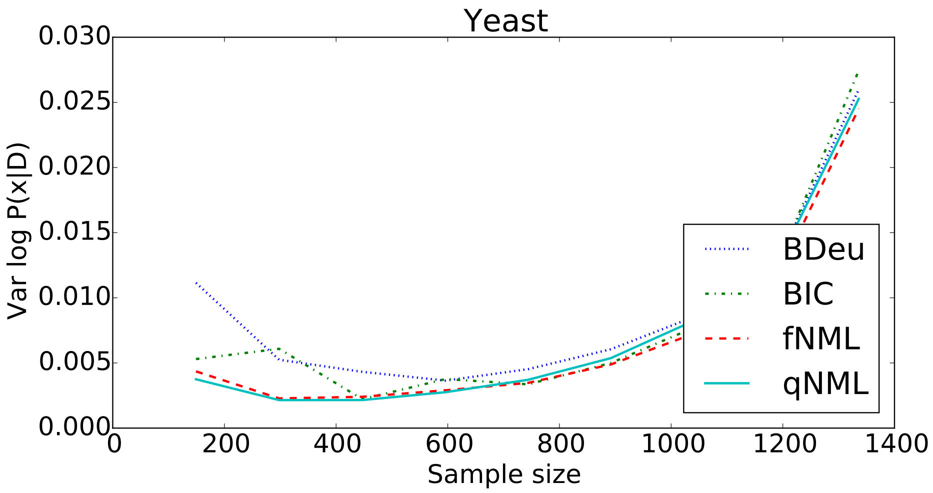

PostOperative
-log P(x|D)
12
11
10
9
8
7
6
0
10
20
30
40
50
60
70
80
90
Sample size
BDeu
BIC
fNML
qNML

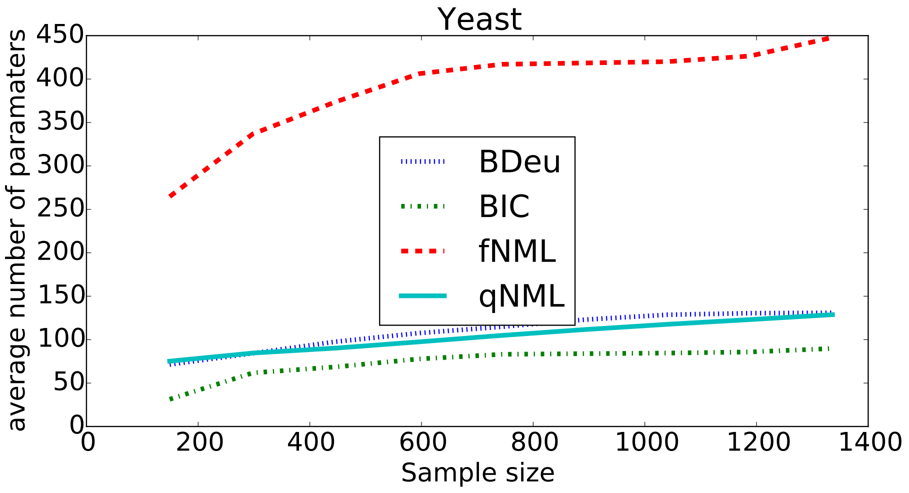

PostOperative
Var log P(x|D)
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
70
80
90
Sample size
BDeu
BIC
fNML
qNML

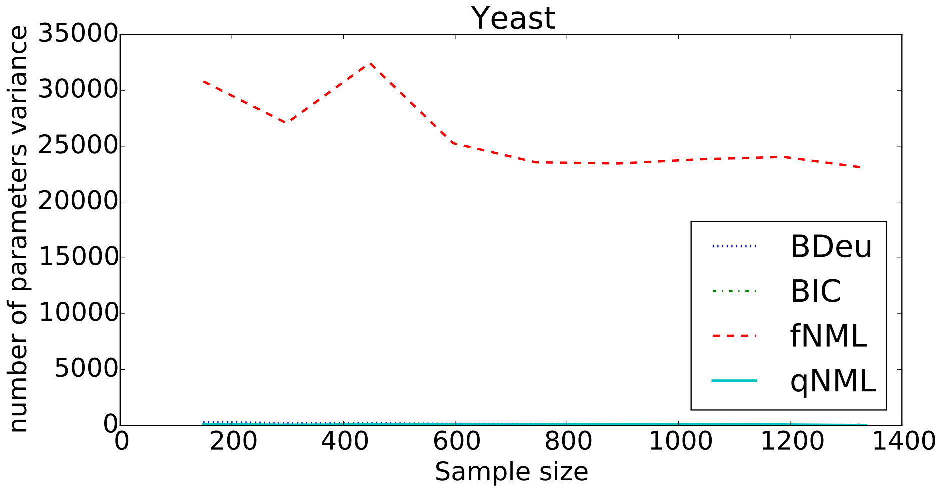

PostOperative
average number of paramaters
3500
3000
2500
2000
1500
1000
500
0
0
10
20
30
40
50
60
70
80
90
Sample size
BDeu
BIC
fNML
qNML

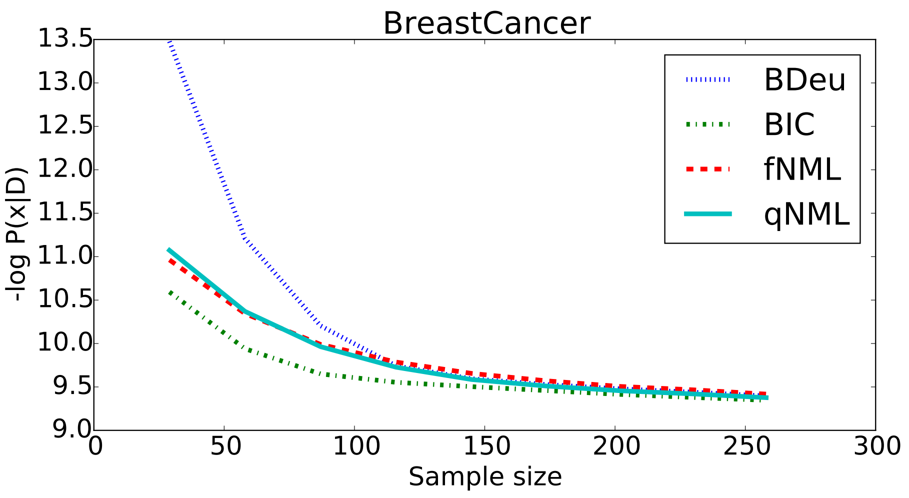

PostOperative
number of parameters variance
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
0
10
20
30
40
50
60
70
80
90
Sample size
BDeu
BIC
fNML
qNML

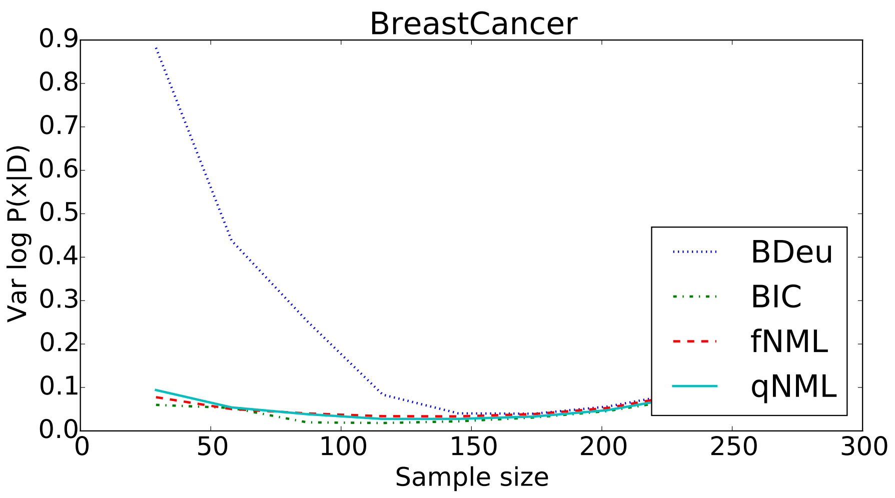

Yeast
-log P(x|D)
5.8
5.7
5.6
5.5
5.4
5.3
5.2
5.1
0
200
400
600
800
1000
1200
1400
Sample size
BDeu
BIC
fNML
qNML

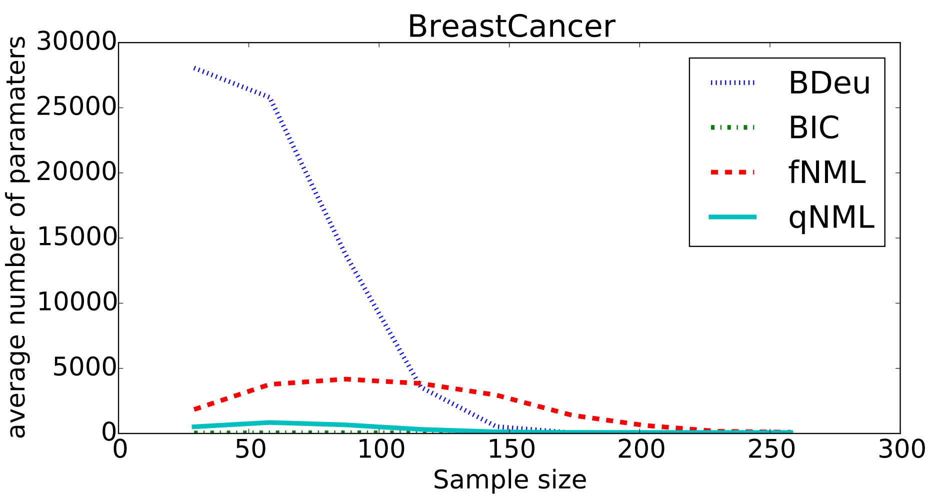

Yeast
Var log P(x|D)
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
200
400
600
800
1000
1200
1400
Sample size
BDeu
BIC
fNML
qNML

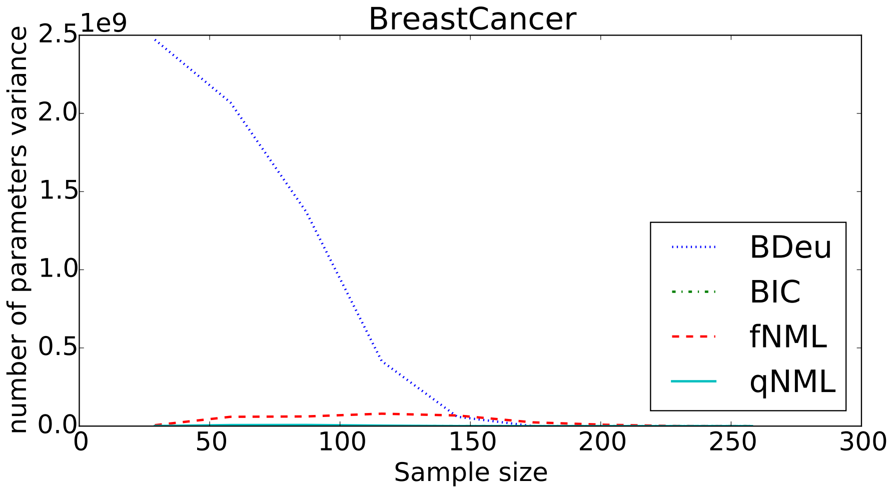

Yeast
average number of paramaters
450
400
350
300
250
200
150
100
50
0
0
200
400
600
800
1000
1200
1400
Sample size
BDeu
BIC
fNML
qNML

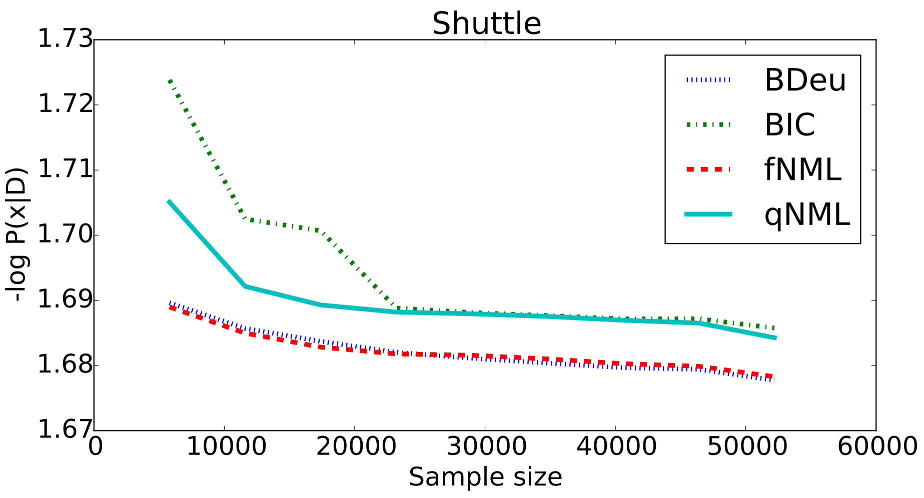

Yeast
number of parameters variance
35000
30000
25000
20000
15000
10000
5000
0
0
200
400
600
800
1000
1200
1400
Sample size
BDeu
BIC
fNML
qNML


BreastCancer
-log P(x|D)
13.5
13.0
12.5
12.0
11.5
11.0
10.5
10.0
9.5
9.0
0
50
100
150
200
250
300
Sample size
BDeu
BIC
fNML
qNML

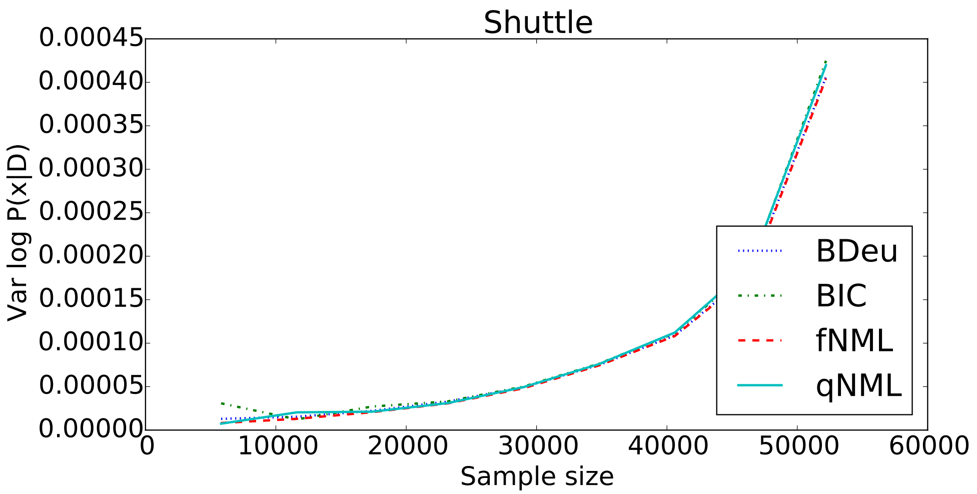

BreastCancer
Var log P(x|D)
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
50
100
150
200
250
300
Sample size
BDeu
BIC
fNML
qNML


BreastCancer
average number of paramaters
30000
25000
20000
15000
10000
5000
0
0
50
100
150
200
250
300
Sample size
BDeu
BIC
fNML
qNML

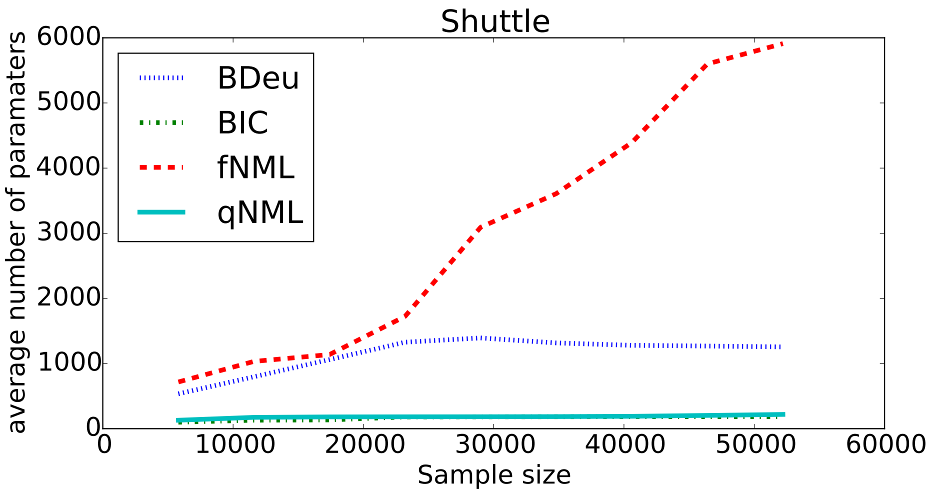

BreastCancer
number of parameters variance
1e9
2.5
2.0
1.5
1.0
0.5
0.0
0
50
100
150
200
250
300
Sample size
BDeu
BIC
fNML
qNML



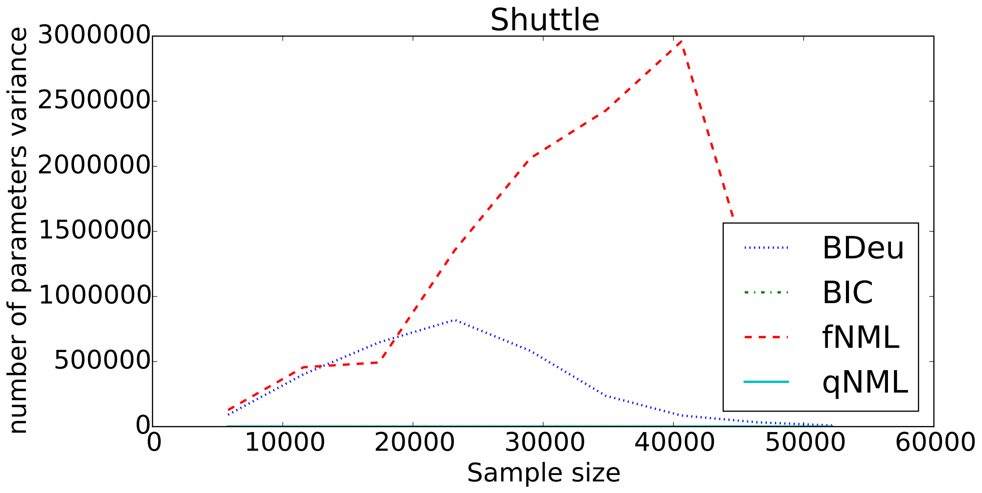
Shuttle



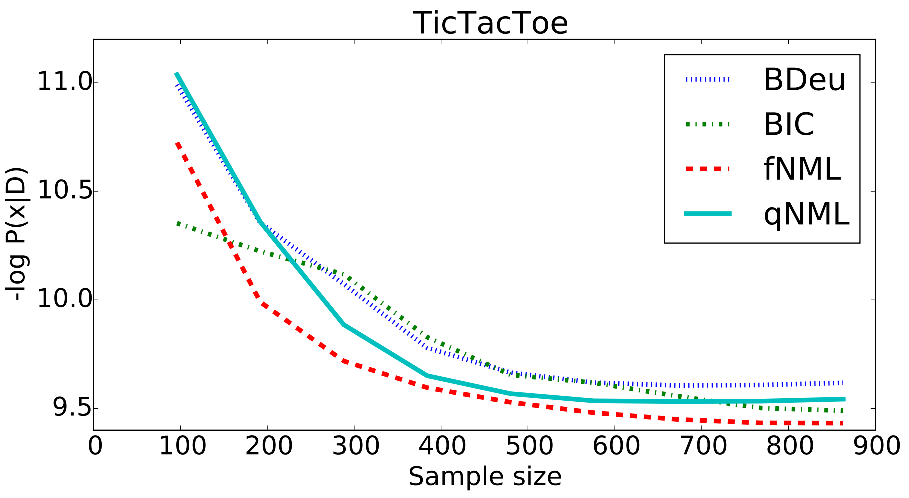
Shuttle

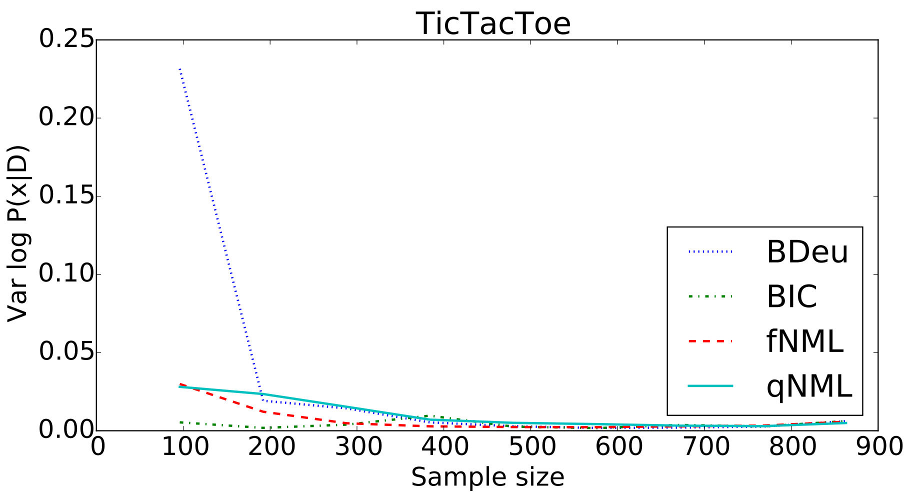


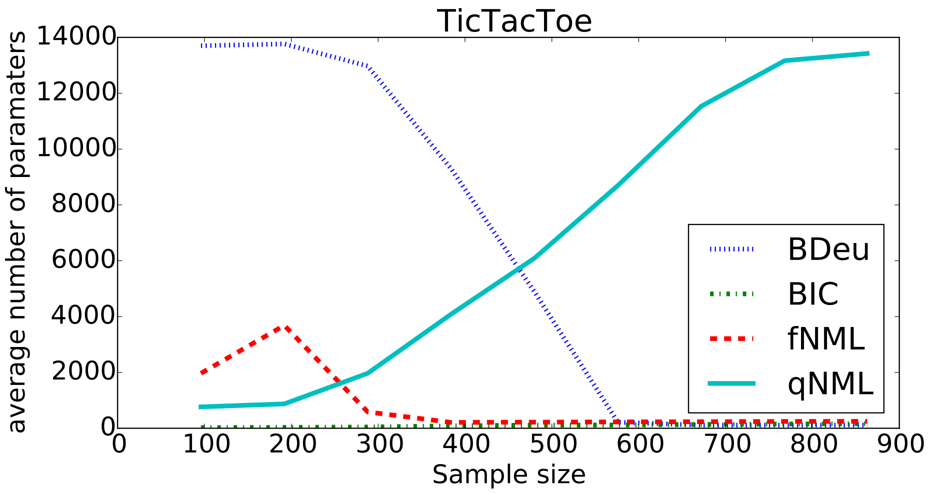
Shuttle

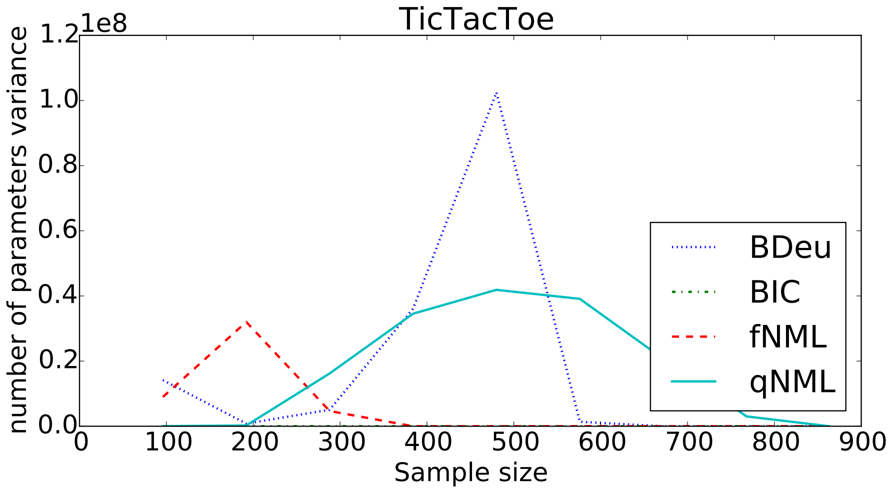


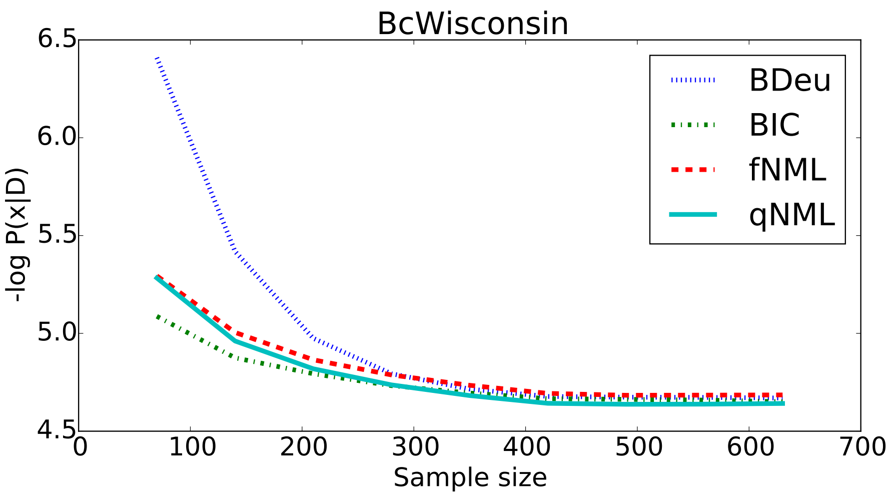
Shuttle

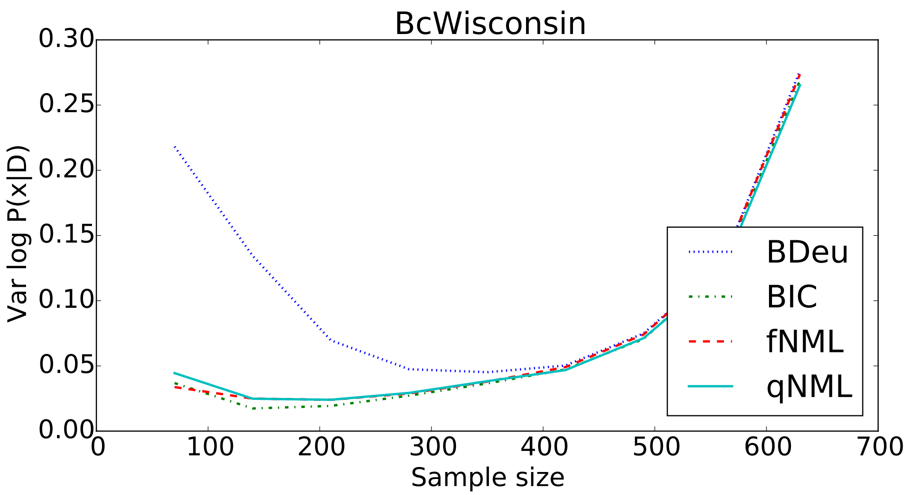

TicTacToe
-log P(x|D)
11.0
10.5
10.0
9.5
0
100
200
300
400
500
600
700
800
900
Sample size
BDeu
BIC
fNML
qNML

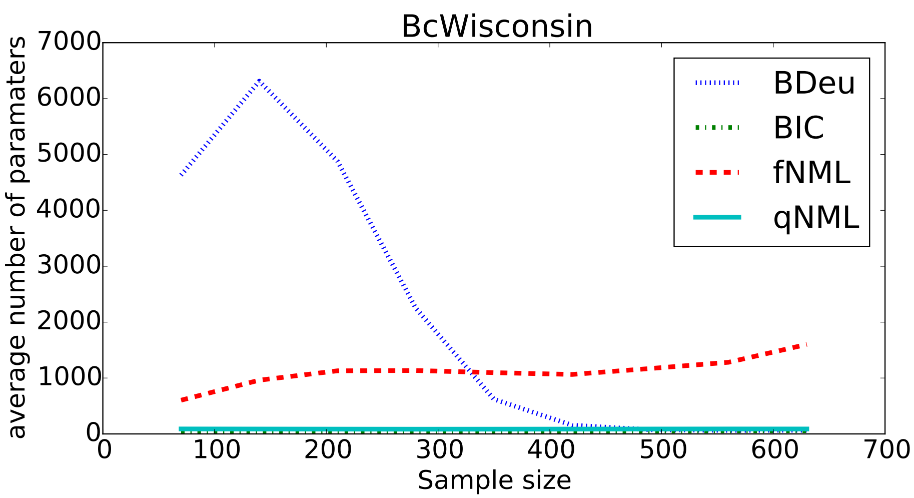

TicTacToe
Var log P(x|D)
0.25
0.20
0.15
0.10
0.05
0.00
0
100
200
300
400
500
600
700
800
900
Sample size
BDeu
BIC
fNML
qNML

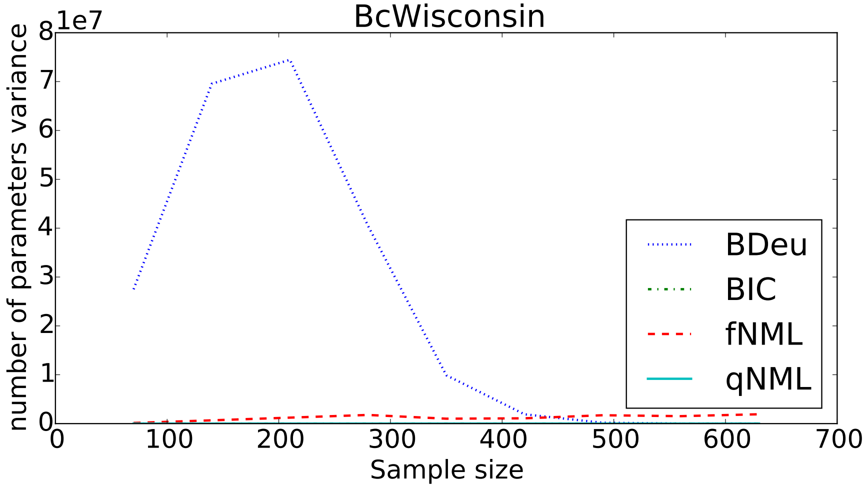

TicTacToe
average number of paramaters
14000
12000
10000
8000
6000
4000
2000
0
0
100
200
300
400
500
600
700
800
900
Sample size
BDeu
BIC
fNML
qNML

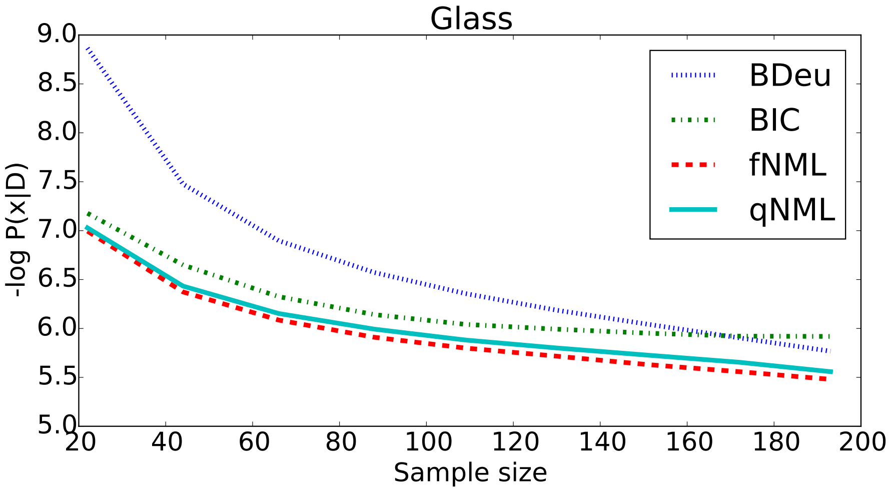

TicTacToe
number of parameters variance
1e8
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
600
700
800
900
Sample size
BDeu
BIC
fNML
qNML

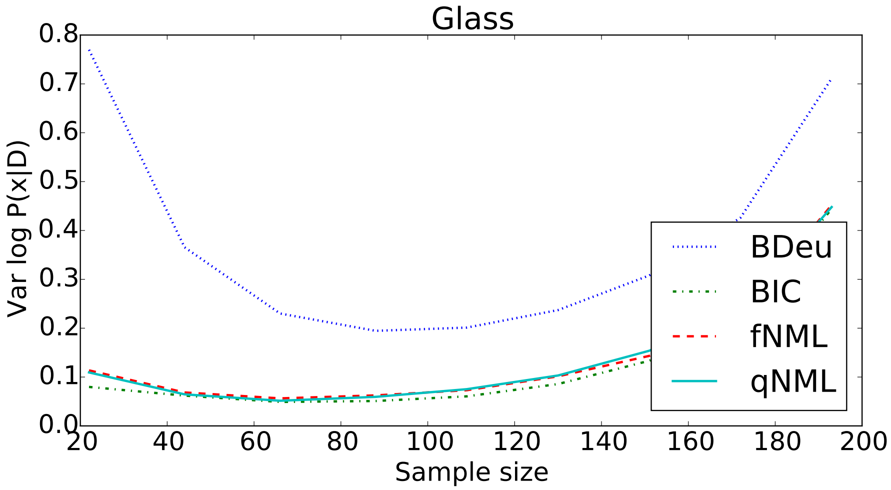

BcWisconsin
-log P(x|D)
6.5
6.0
5.5
5.0
4.5
0
100
200
300
400
500
600
700
Sample size
BDeu
BIC
fNML
qNML

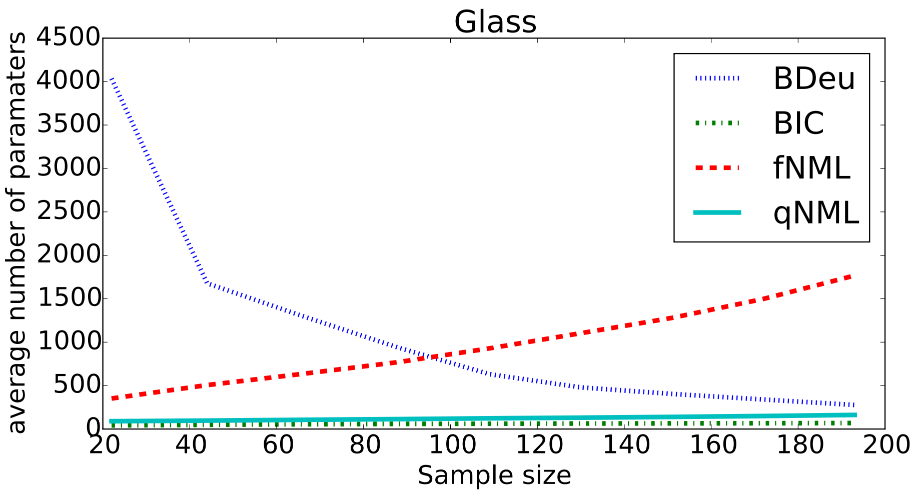

BcWisconsin
Var log P(x|D)
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
100
200
300
400
500
600
700
Sample size
BDeu
BIC
fNML
qNML

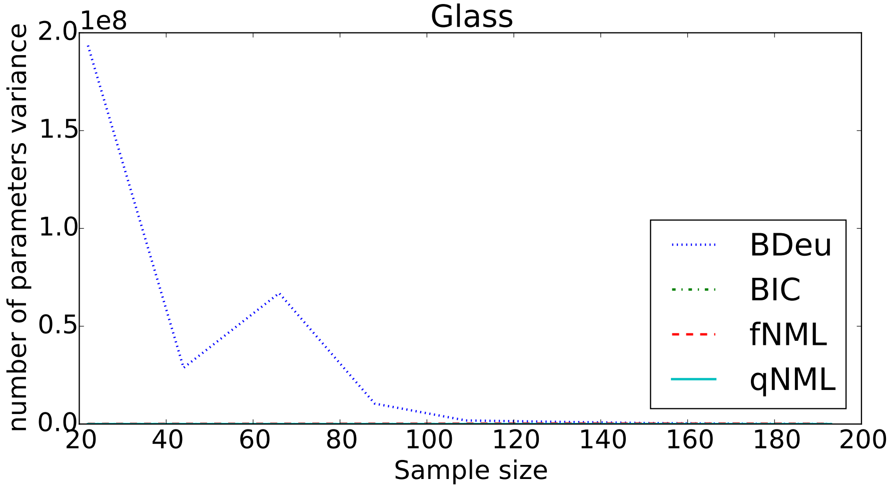

BcWisconsin
average number of paramaters
7000
6000
5000
4000
3000
2000
1000
0
0
100
200
300
400
500
600
700
Sample size
BDeu
BIC
fNML
qNML

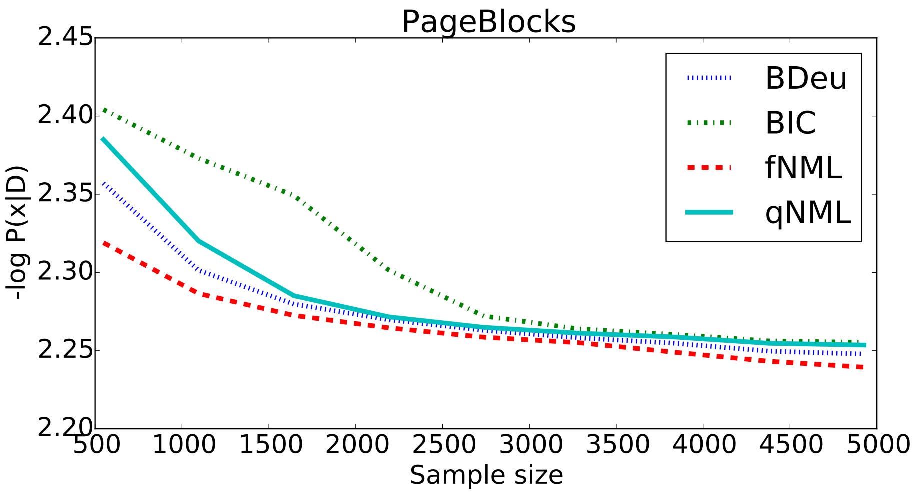

BcWisconsin
1e7
number of parameters variance
8
7
6
5
4
3
2
1
0
0
100
200
300
400
500
600
700
Sample size
BDeu
BIC
fNML
qNML

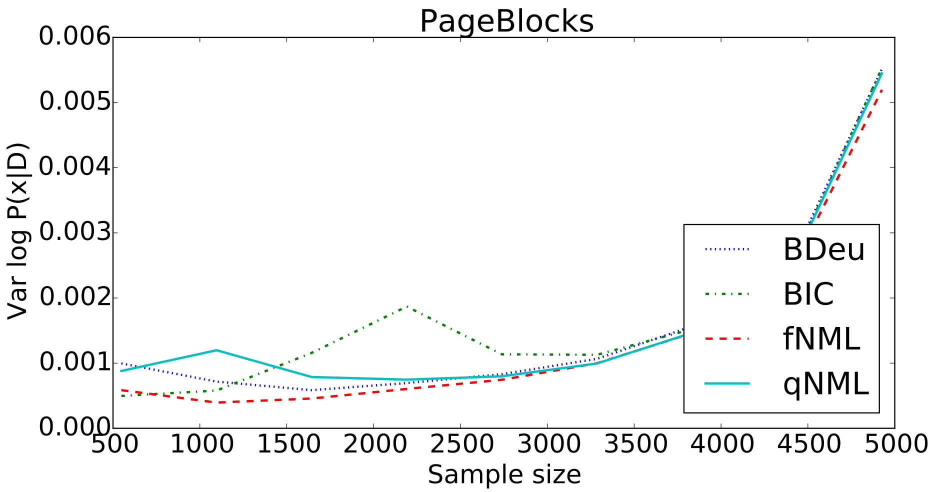

Glass
-log P(x|D)
9.0
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
20
40
60
80
100
120
140
160
180
200
Sample size
BDeu
BIC
fNML
qNML

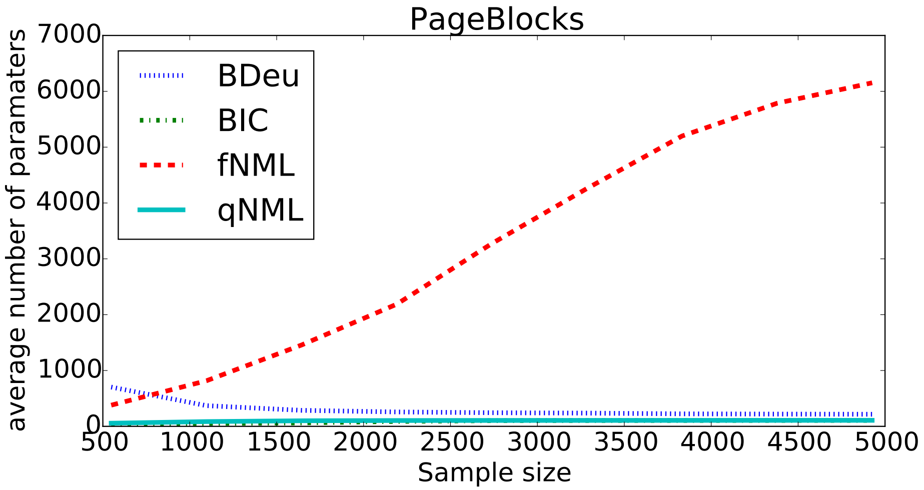

Glass
Var log P(x|D)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
20
40
60
80
100
120
140
160
180
200
Sample size
BDeu
BIC
fNML
qNML

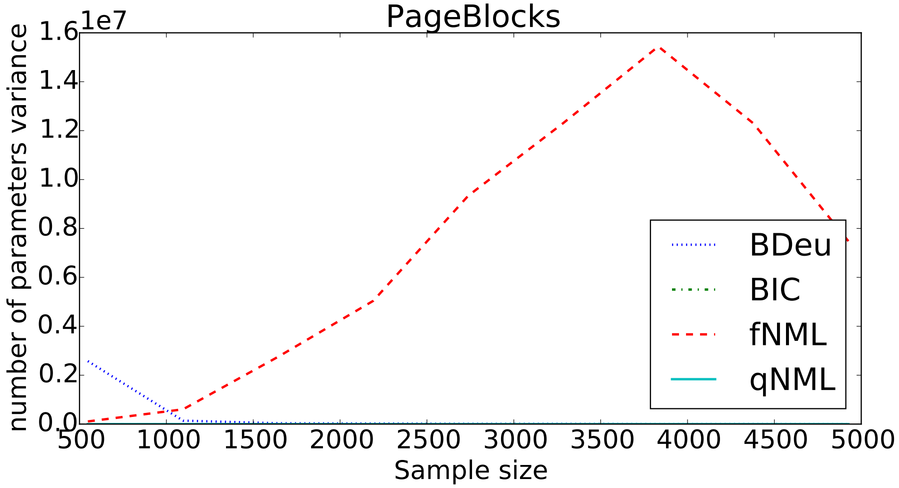

Glass
average number of paramaters
4500
4000
3500
3000
2500
2000
1500
1000
500
0
20
40
60
80
100
120
140
160
180
200
Sample size
BDeu
BIC
fNML
qNML

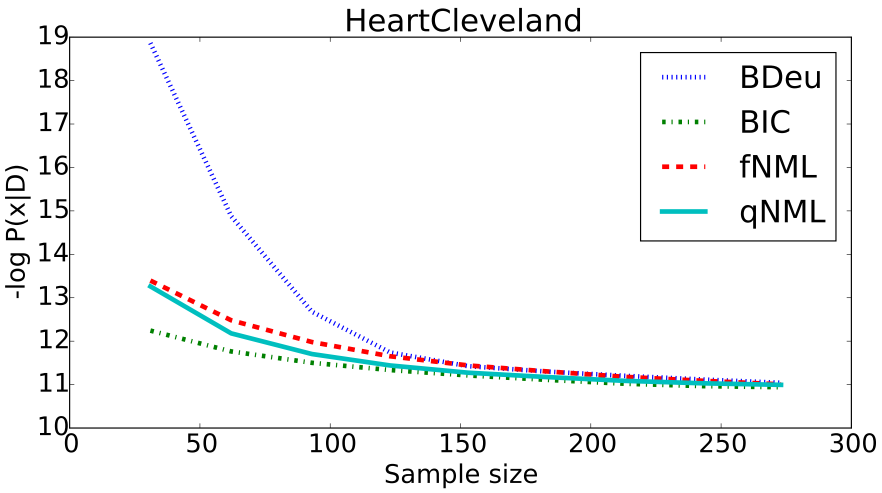

Glass
number of parameters variance
1e8
2.0
1.5
1.0
0.5
0.0
20
40
60
80
100
120
140
160
180
200
Sample size
BDeu
BIC
fNML
qNML

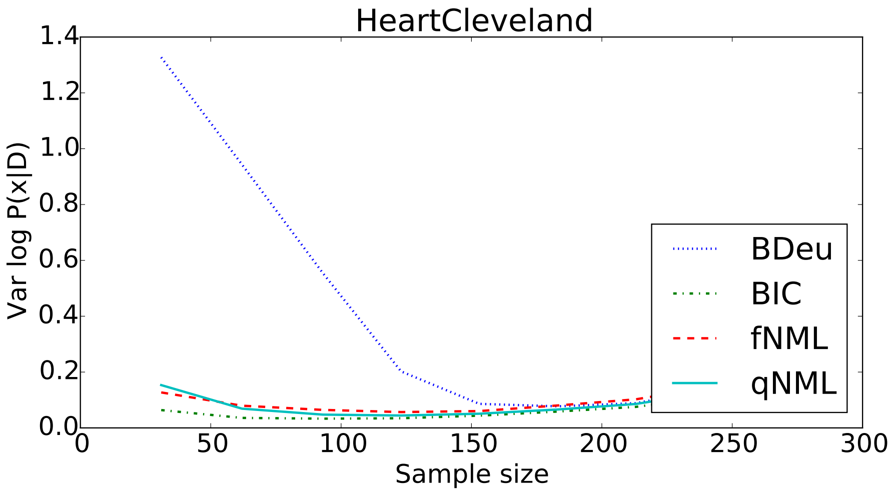

PageBlocks
-log P(x|D)
2.45
2.40
2.35
2.30
2.25
2.20
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
Sample size
BDeu
BIC
fNML
qNML

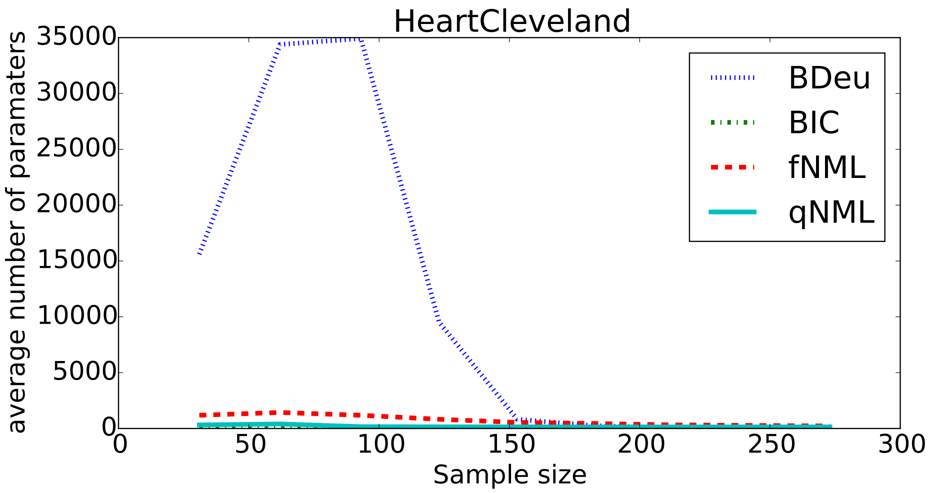

PageBlocks
Var log P(x|D)
0.006
0.005
0.004
0.003
0.002
0.001
0.000
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
Sample size
BDeu
BIC
fNML
qNML

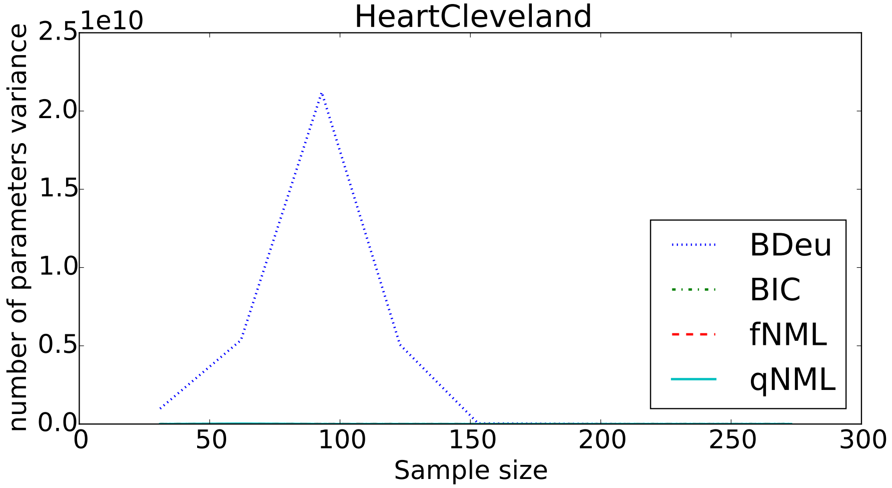

PageBlocks
average number of paramaters
7000
6000
5000
4000
3000
2000
1000
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
Sample size
BDeu
BIC
fNML
qNML

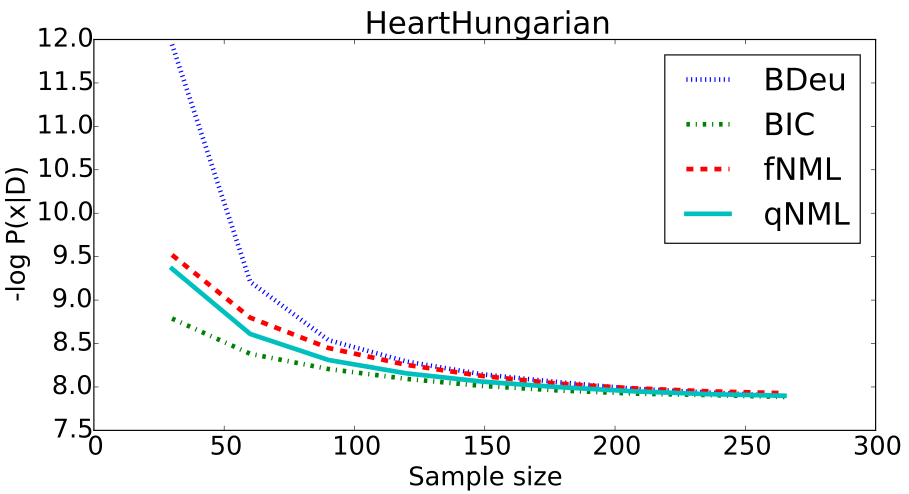

PageBlocks
1e7
number of parameters variance
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
Sample size
BDeu
BIC
fNML
qNML

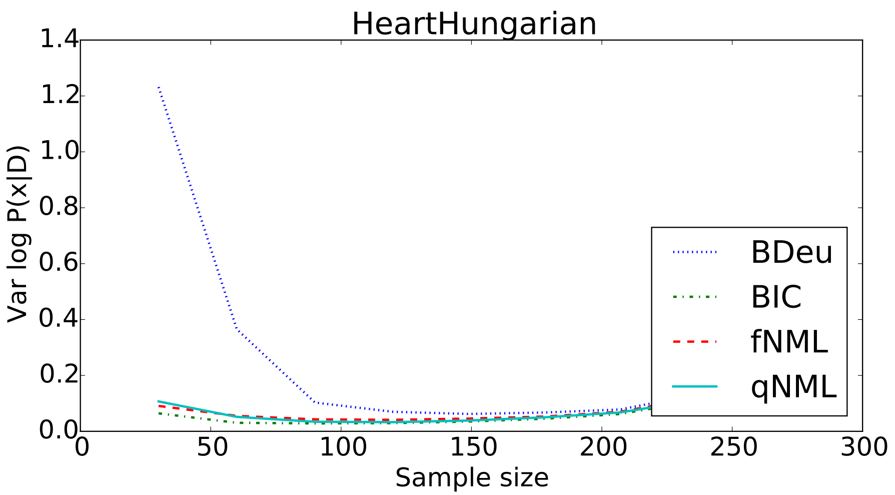

HeartCleveland
-log P(x|D)
19
18
17
16
15
14
13
12
11
10
0
50
100
150
200
250
300
Sample size
BDeu
BIC
fNML
qNML

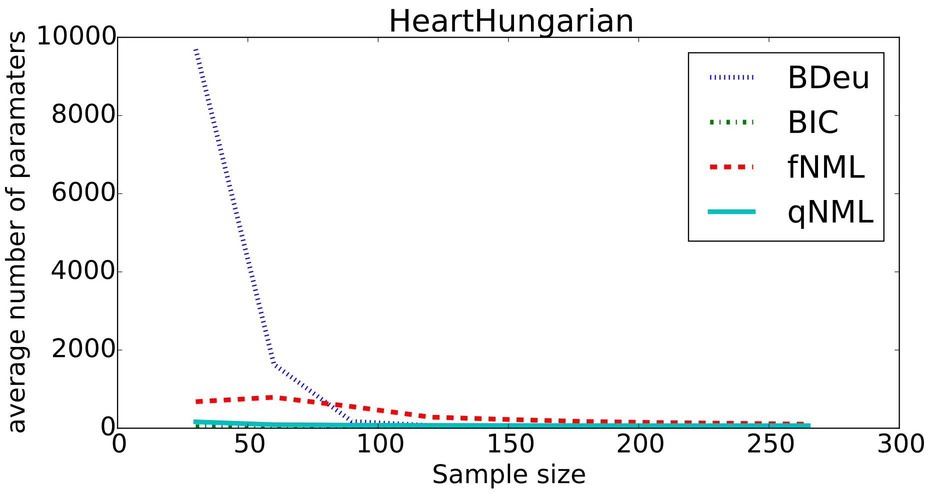

HeartCleveland
Var log P(x|D)
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200
250
300
Sample size
BDeu
BIC
fNML
qNML

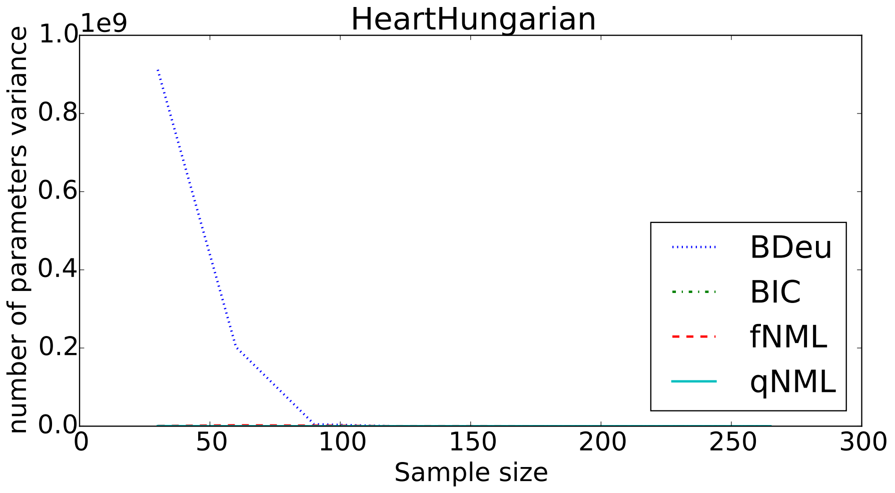

HeartCleveland
average number of paramaters
35000
30000
25000
20000
15000
10000
5000
0
0
50
100
150
200
250
300
Sample size
BDeu
BIC
fNML
qNML

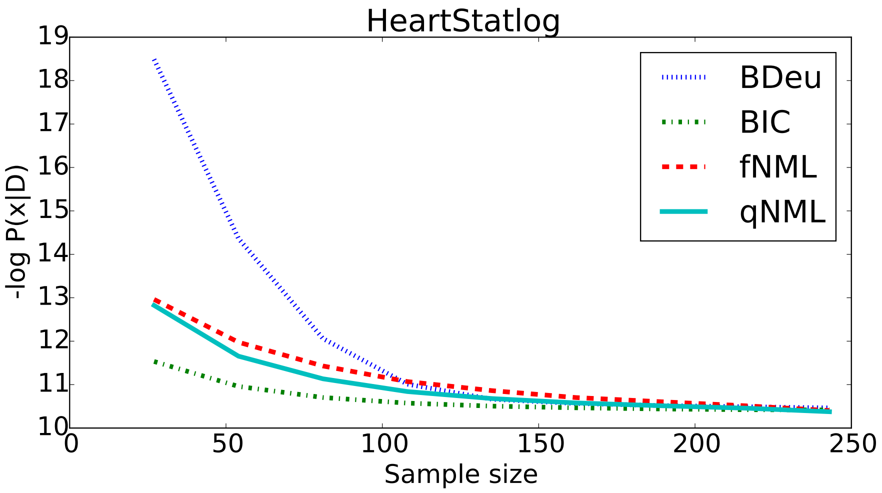

HeartCleveland
number of parameters variance
1e10
2.5
2.0
1.5
1.0
0.5
0.0
0
50
100
150
200
250
300
Sample size
BDeu
BIC
fNML
qNML

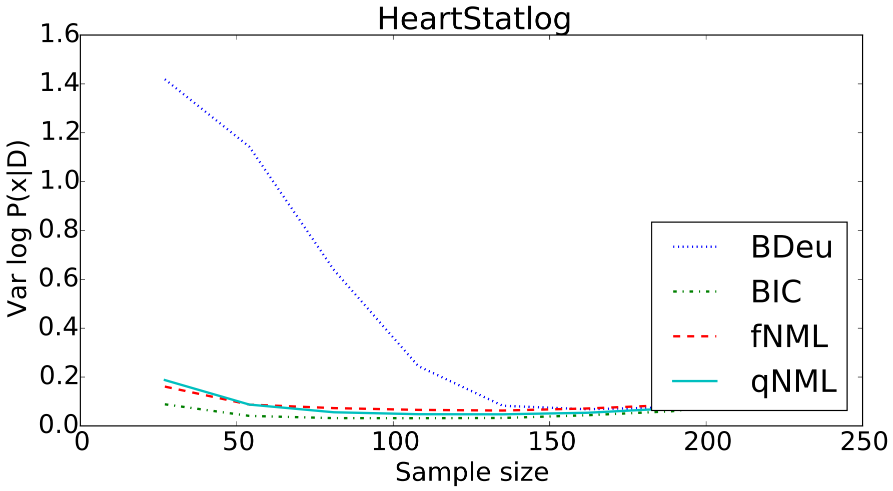

HeartHungarian
-log P(x|D)
12.0
11.5
11.0
10.5
10.0
9.5
9.0
8.5
8.0
7.5
0
50
100
150
200
250
300
Sample size
BDeu
BIC
fNML
qNML

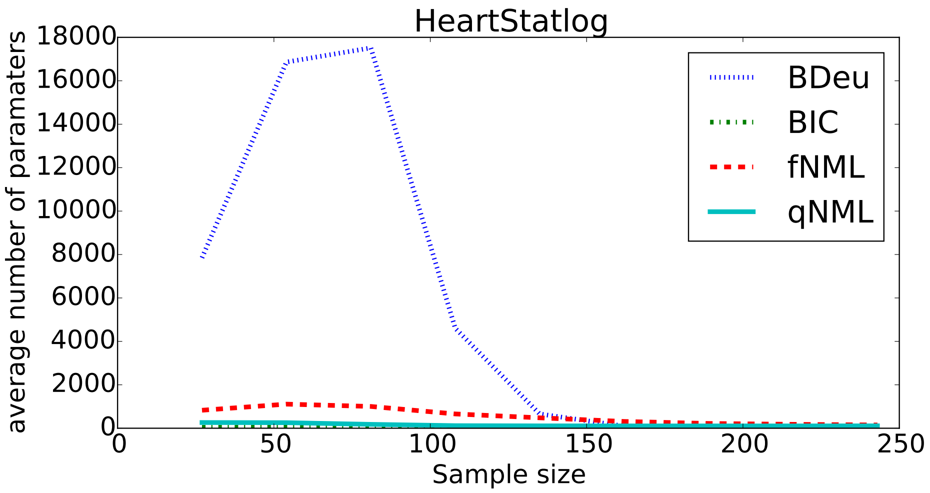

HeartHungarian
Var log P(x|D)
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200
250
300
Sample size
BDeu
BIC
fNML
qNML

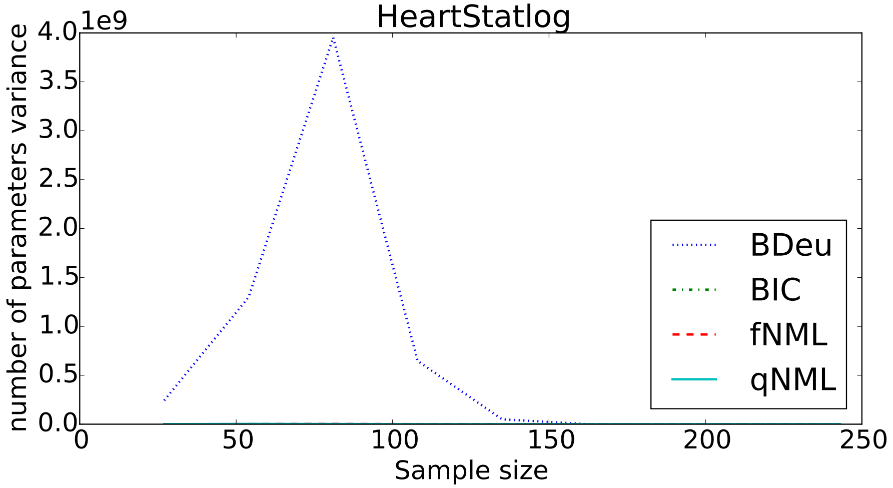

HeartHungarian
average number of paramaters
10000
8000
6000
4000
2000
0
0
50
100
150
200
250
300
Sample size
BDeu
BIC
fNML
qNML

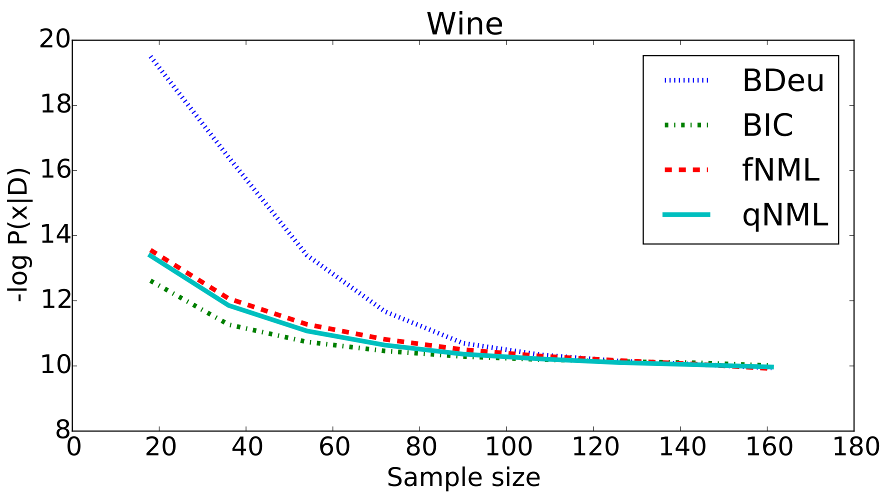

HeartHungarian
number of parameters variance
1e9
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200
250
300
Sample size
BDeu
BIC
fNML
qNML

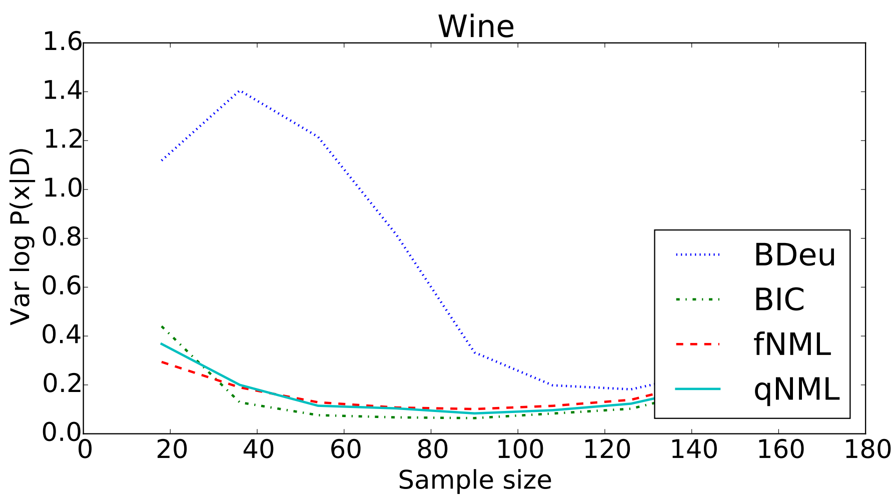

HeartStatlog
-log P(x|D)
19
18
17
16
15
14
13
12
11
10
0
50
100
150
200
250
Sample size
BDeu
BIC
fNML
qNML

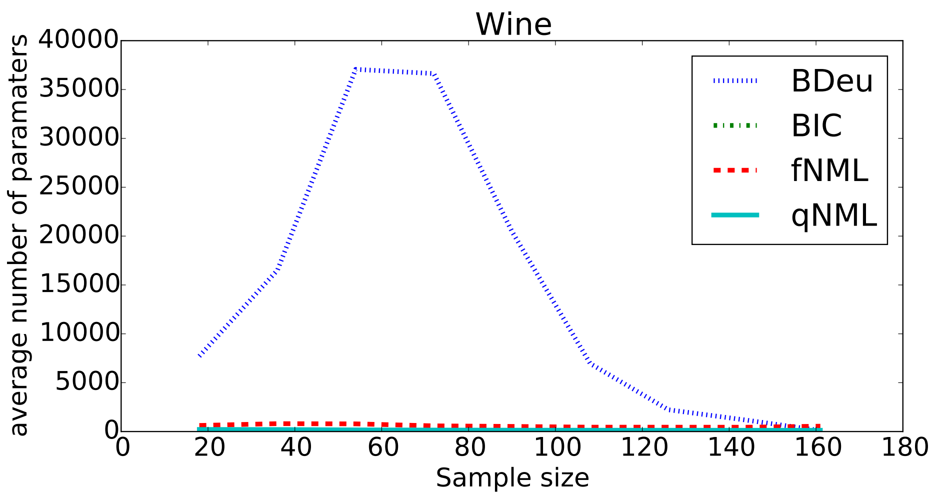

HeartStatlog
Var log P(x|D)
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200
250
Sample size
BDeu
BIC
fNML
qNML

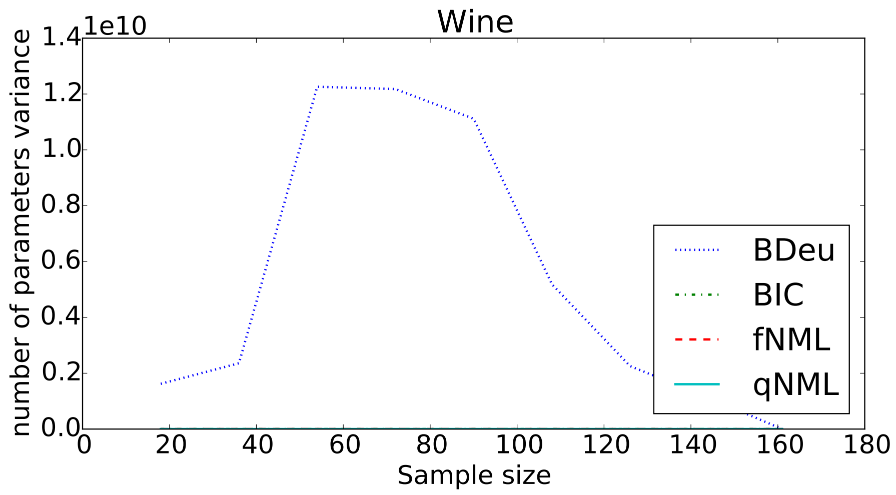

HeartStatlog
average number of paramaters
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
0
50
100
150
200
250
Sample size
BDeu
BIC
fNML
qNML

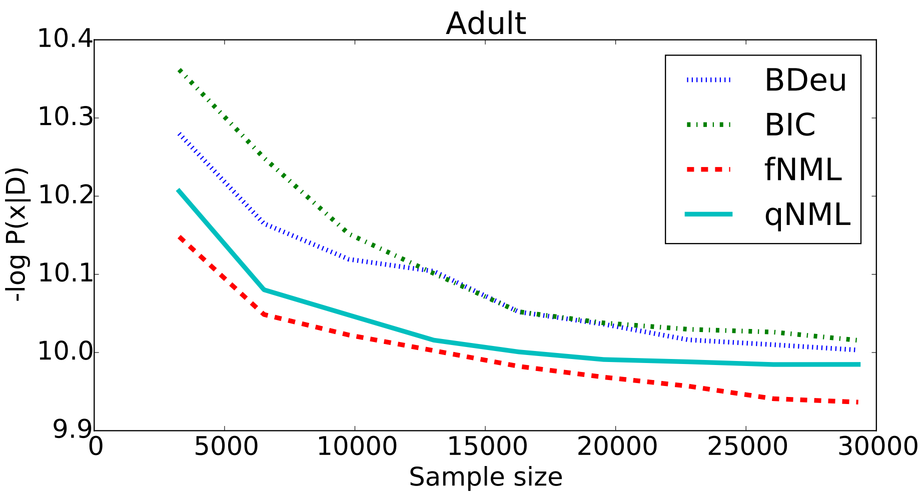

HeartStatlog
number of parameters variance
1e9
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
50
100
150
200
250
Sample size
BDeu
BIC
fNML
qNML

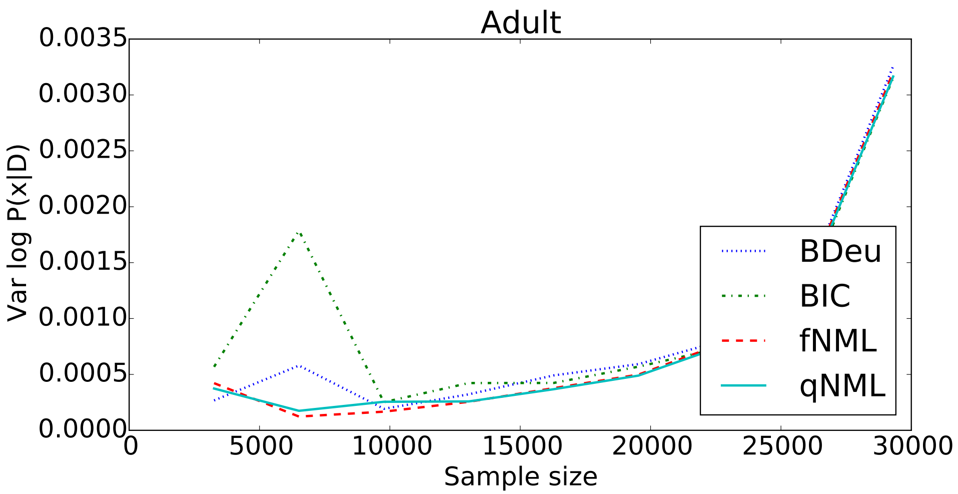

Wine
-log P(x|D)
20
18
16
14
12
10
8
0
20
40
60
80
100
120
140
160
180
Sample size
BDeu
BIC
fNML
qNML

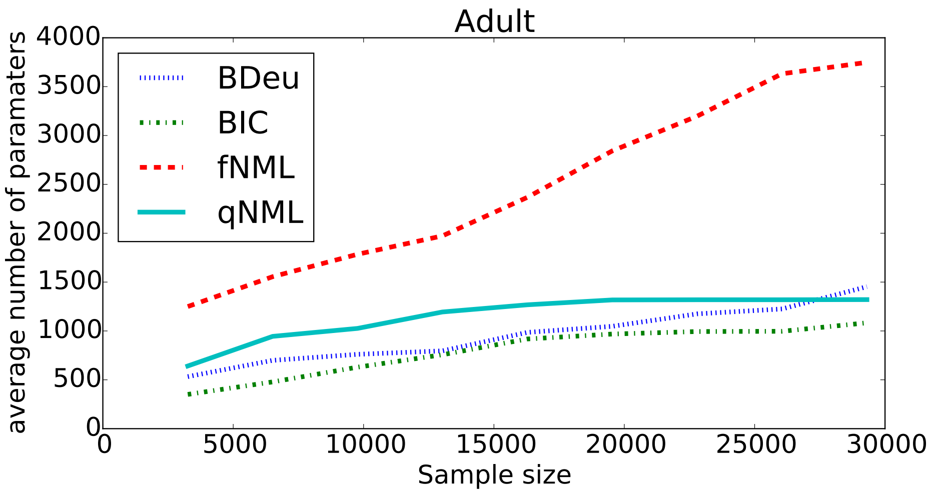

Wine
Var log P(x|D)
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
120
140
160
180
Sample size
BDeu
BIC
fNML
qNML

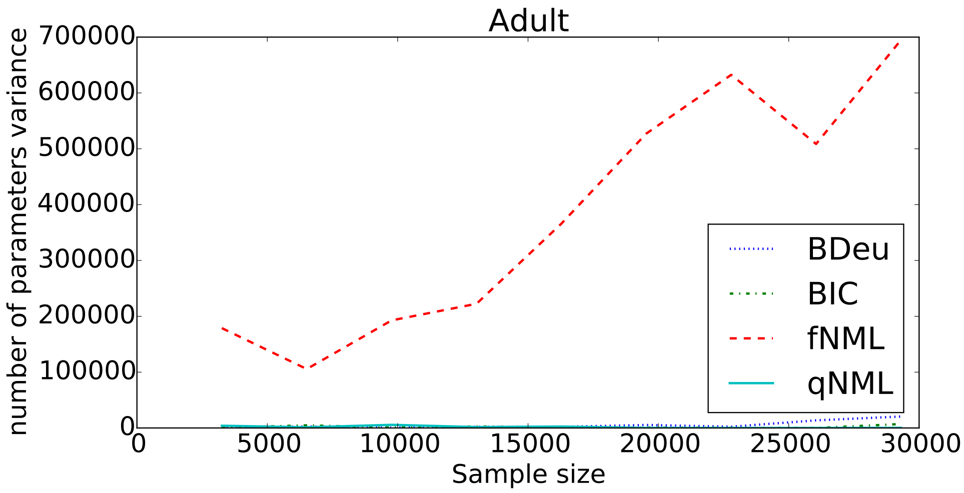

Wine
average number of paramaters
40000
35000
30000
25000
20000
15000
10000
5000
0
0
20
40
60
80
100
120
140
160
180
Sample size
BDeu
BIC
fNML
qNML


Wine
1e10
number of parameters variance
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
120
140
160
180
Sample size
BDeu
BIC
fNML
qNML


Adult
-log P(x|D)
10.4
10.3
10.2
10.1
10.0
9.9
0
5000
10000
15000
20000
25000
30000
Sample size
BDeu
BIC
fNML
qNML


Adult
Var log P(x|D)
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
5000
10000
15000
20000
25000
30000
Sample size
BDeu
BIC
fNML
qNML


Adult
average number of paramaters
4000
3500
3000
2500
2000
1500
1000
500
0
0
5000
10000
15000
20000
25000
30000
Sample size
BDeu
BIC
fNML
qNML


Adult
number of parameters variance
700000
600000
500000
400000
300000
200000
100000
0
0
5000
10000
15000
20000
25000
30000
Sample size
BDeu
BIC
fNML
qNML

References

Adamchik, V. (1997). On Stirling numbers and Euler sums. *Journal of Computational and Applied Mathematics*, 79(1):119 – 130.

Chickering, D. M. (1995). A Transformational Characterization of Equivalent Bayesian Network Structures. In *UAI '95: Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence*, pages 87–98. Morgan Kaufmann.

Jääsaari, E., Leppä-aho, J., Silander, T., and Roos, T. (2018). Minimax optimal Bayes mixtures for memoryless sources over large alphabets. Accepted for the *29th International Conference on Algorithmic Learning Theory (ALT 2018).*

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques.* MIT Press.

Mononen, T. and Myllymäki, P. (2008). On the multinomial stochastic complexity and its connection to the birthday problem. In *Proceedings of the International Conference on Information Theory and Statistical Learning*, Las Vegas, NV.

Suzuki, J. (2017). A theoretical analysis of the BDeu scores in Bayesian network structure learning. *Behaviormetrika*, 44(1):97–116.

Watanabe, K. and Roos, T. (2015). Achievability of asymptotic minimax regret by horizon-dependent and horizon-independent strategies. *Journal of Machine Learning Research*, 16:2357–2375.