Robust learning of inhomogeneous PMMs

Ralf Eggeling
Martin Luther University
Halle-Wittenberg
Germany

Teemu Roos
Helsinki Insitute for
Information Technology
Finland

Petri Myllymäki
Helsinki Insitute for
Information Technology
Finland

Ivo Grosse
Martin Luther University
Halle-Wittenberg
Germany

Abstract

Inhomogeneous parsimonious Markov models have recently been introduced for modeling symbolic sequences, with a main application being DNA sequence analysis. Structure and parameter learning of these models has been proposed using a Bayesian approach, which entails the practically challenging choice of the prior distribution. Cross validation is a possible way of tuning the prior hyperparameters towards a specific task such as prediction or classification, but it is overly time-consuming. On this account, robust learning methods, which do not require explicit prior specification and – in the absence of prior knowledge – no hyperparameter tuning, are of interest. In this work, we empirically investigate the performance of robust alternatives for structure and parameter learning that extend the practical applicability of inhomogeneous parsimonious Markov models to more complex settings than before.

1 INTRODUCTION

Modeling statistical dependencies and independencies among a set of random variables is a common task in data analysis. Parsimonious Markov models (PMMs) have been proposed by Bourguignon and Robelin [2004] as an extension of variable order Markov models [Rissanen, 1983] for effectively capturing dependencies in sequential data. Recently, inhomogeneous PMMs have been proposed for modeling short sequence patterns by taking into account position-specific higher-order dependencies, allowing a favorable tradeoff between modeling dependencies and avoiding overfitting [Eggeling et al., 2013].

Learning inhomogeneous PMMs, which involves structure and parameter learning, has been proposed using a Bayesian approach, which allows taking into account prior knowledge. In practice, however, there is often either only vague prior knowledge available, or existing prior knowledge cannot be translated into a mathematically convenient form. As a consequence, the functional form of the prior is often chosen based on arithmetical convenience, and further restrictions finally reduce the prior choice to choosing the values of one or two hyperparameters.

Since choosing appropriate values of prior hyperparameters is difficult, and since inappropriate values may yield a significantly degraded performance, hyperparameters are commonly tuned based on, e.g., repeated hold-out or cross validation techniques. From a practical perspective, cross validation or similar approaches are overly time-consuming in relation to a single estimation step. While this procedure is certainly doable for a few highly relevant data sets, it nevertheless limits large-scale applicability of inhomogeneous PMMs, especially in situations where structure and parameter learning are only subtasks in a more complex learning procedure.

Alternative to the Bayesian approach, methods for both structure and parameter learning that require no prior specification have been proposed. One prominent example is the Minimum Description Lengh (MDL) principle [Rissanen, 1978], which is motivated by the information theoretic argument that learning from data is equivalent to compressing data [Grünwald, 2007]. Modern MDL is based on the minimax optimal Normalized Maximum Likelihood (NML) distribution [Shtarkov, 1987], which involves no prior distribution. Since the NML can be computed exactly only for a few simple models, approximations of it have been used for learning complex models such as Bayesian networks [Silander et al., 2010].

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

The purpose of this work is (i) to apply different learning approaches to inhomogeneous PMMs and compare their predictive power (Section 4) and (ii) to demonstrate that extensive cross validation procedures for hyperparameter tuning can be avoided when robust learning methods are used (Section 5). To this end, we first recap the definition of inhomogeneous PMMs in the following section and specify the learning approaches that we investigate in Section 3.

2 MODEL DESCRIPTION

Here, we briefly recap inhomogeneous parsimonious Markov models (PMMs) as proposed by Eggeling et al. [2013]. We denote a single observation from the alphabet by $x \in \mathcal{A}$, a sequence of observations of length L by $\vec{x} = (x_1, \ldots, x_L)$, and a data set of N sequences of length L by $\mathbf{x} = (\vec{x}_1, \ldots, \vec{x}_N)$.

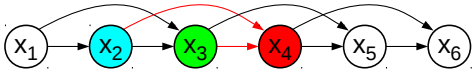
Parsimonious context trees (PCTs) as proposed by Bourguignon and Robelin [2004] are the central data structures in the model. A PCT partitions the set of context words into subsets, called *contexts*. We denote a single context, represented by a leaf in a PCT, by c. An inhomogeneous PMM for sequences of length L is based on L PCTs, which we denote by $\vec{\tau} = (\tau_1, \ldots, \tau_L)$.

For each PCT, we associate a conditional probability distribution over $\mathcal{A}$ to each of its contexts, and we denote the conditional probability of observing symbol $a \in \mathcal{A}$ at position ℓ given that the concatenation of the preceding d symbols is in c by $\theta_{\ell ca}$. We denote the model parameters of the ℓ-th position by $\Theta_\ell = \left(\tau_\ell, (\vec{\theta}_{\ell c})_{c \in \tau_\ell}\right)$, and all model parameters by $\vec{\Theta} = (\Theta_1, \ldots, \Theta_L)$. The likelihood of an inhomogeneous parsimonious Markov model of order d is then given by

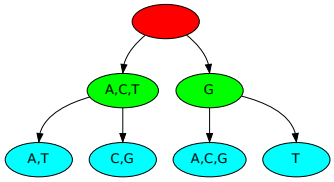
$$P(\mathbf{x}|\vec{\Theta}) = \prod_{\ell=1}^{L} \prod_{c \in \tau_\ell} \prod_{a \in \mathcal{A}} (\theta_{\ell ca})^{N_{\ell ca}}, \qquad (1)$$

where $N_{\ell ca}$ is the number of occurrences of symbol a at position ℓ in all sequences in data set $\mathbf{x}$ where the symbols from position $\ell - d$ to position $\ell - 1$ are in c [Eggeling et al., 2013].

An inhomogeneous PMM can be understood as a Bayesian network (BN) of fixed structure (Fig. 1(a)) that uses a PCT for each random variable for reducing the parameter space of each conditional probability table (Fig. 1(b)). Whereas in an inhomogeneous PMM the parent nodes are fixed, the structural flexibility and thus the model selection problem arises from the choice of an appropriate PCT at each position. Loosely speaking, the incentive of a PCT is to choose the smallest possible set of contexts that capture relevant dependencies in the data, yielding a statistical model with only a few parameters that might be less prone to overfitting than alternative models with a higher number of parameters.



(a) General dependency structure



(b) Parsimonious context tree

Figure 1: The general dependency structure of a second order inhomogeneous PMM for a sequence of length 6 is shown in Fig. 1(a). PCTs are used for reducing the number of conditional probability parameters. Figure 1(b) shows an example PCT for DNA alphabet at position 4. The nodes are colored according to the random variables they correspond to. Each position in the model may have a different PCT.

3 LEARNING APPROACHES

Learning inhomogeneous PMMs consists of structure learning for each PCT in the model and estimation of the corresponding conditional probability parameters.

3.1 Structure scores

Structure learning is algorithmically challenging, but can be solved by the dynamic programming algorithm of Bourguignon and Robelin [2004]. This algorithm can be used for optimizing any score function satisfying the so called decomposability property [Heckerman et al., 1995].

In the Bayesian setting, the structure score of the ℓ-th PCT τ_ℓ for data $\mathbf{x}$ is the local posterior probability $P(\tau_\ell|\mathbf{x})$. Using the prior specification of Eggeling et al. [2013], we obtain

$$S_{\mathrm{BDeu}}(\tau_\ell|\eta, \kappa, \mathbf{x}) = \sum_{c \in \tau_\ell} \log \frac{\kappa \mathcal{B}\left(\vec{N}_{\ell c} + (\alpha, \ldots, \alpha)_{|\mathcal{A}|}\right)}{\mathcal{B}\left((\alpha, \ldots, \alpha)_{|\mathcal{A}|}\right)}, \qquad (2)$$

with $\alpha = \frac{\eta |c|}{|\mathcal{A}|^{d+1}}$, where η is the equivalent sample size (ESS), which is the sole hyperparameter of the symmetric Dirichlet prior over the probability param-

eters; see Buntine [1991] and Heckerman et al. [1995] for an introduction into the ESS concept in graphical models. $\mathcal{B}$ denotes the multinomial beta function, and $(a, \ldots, a)_b$ denotes a b-dimensional vector filled with constant a. This score is equivalent to the *BDeu* score of BNs and we name it accordingly. The only conceptual difference is the existence of a second hyperparameter, κ, which originates from the structure prior and is as such a particularity of PMMs. Setting $\kappa = 1$ corresponds to a uniform distribution over all structures, yielding a BDeu score in its most widely used form [Buntine, 1991].

Since the Bayesian scoring criterion is equivalent to that of BNs, it is worthwile to consider applying other scoring criteria for BNs to inhomogeneous PMMs as well. A possible alternative is based on the Normalized Maximum Likelihood (NML) distribution, motivated by the information theoretical argument of minimizing the worst case regret [Shtarkov, 1987]. However, exact computation of the NML is difficult since it involves a normalization of all possible data sets, which can only be done for a few simple models. Silander et al. [2008] have proposed the *factorized NML* (fNML) criterion for learning BNs in order to approximate the NML distribution of the full model by a product of independently normalized terms. For inhomogeneous PMMs, the fNML score can be written as

$$S_{\mathrm{fNML}}(\tau_\ell|\mathbf{x}) = \sum_{c\in\tau_\ell} \log\left(\left(\frac{N_{\ell ca}}{N_{\ell c\cdot}}\right)^{N_{\ell ca}}\right) - C^{|\mathcal{A}|}_{N_{\ell c\cdot}}, \quad (3)$$

where C^a_b is the stochastic complexity of a multinomial distribution with a being the number of categories and b being the sample size. C^b_a can be computed using the linear-time algorithm of Kontkanen and Myllymäki [2007] or the so called Szpankowski approximation [Kontkanen et al., 2003]. The fNML score as such does not require any explicit prior assumptions. However, it is close to the Bayesian marginal likelihood using a Jeffreys prior, which is a Dirichlet distribution with hyperparameters $\frac{1}{2}$ in this setting, thus violating the equivalent sample size condition. In such a Bayesian interpretation, using fNML would also imply using a uniform prior over all model structures. The fNML score yields a consistent estimator of the model structure [Silander et al., 2008].

A generally applicable score, which has an interpretation both in Bayesian statistics and in information theory, is the *Bayesian Information Criterion* (BIC) of Schwarz [1978], which is also referred to as MDL score, since it corresponds to a coarse minimum description length approximation using a so called two-part encoding of model structure and data given the model [Grünwald, 2007]. The BIC score can be written as

$$\begin{aligned} S_{\mathrm{BIC}}(\tau_\ell|\mathbf{x}) &= \sum_{c\in\tau_\ell} 2\log\left(\left(\frac{N_{\ell ca}}{N_{\ell c\cdot}}\right)^{N_{\ell ca}}\right) \\ &\quad -|\tau_\ell|(|\mathcal{A}|-1)\log(N), \end{aligned} \quad (4)$$

and it is known to penalize additional parameters rather strictly, so it is typically more prone to underfitting than to overfitting. Furthermore, BIC can be seen as an approximation of both the fNML score and the Bayesian marginal likelihood.

For BNs, Silander et al. [2008] have shown that BIC is inferior to BDeu and fNML, when attempting to find the true model structure. However, dismissing BIC for model selection in PMMs might not be justified, especially when the task is not finding a true model structure but rather a structure that is suitable for prediction. Even though it has often been observed that structures that are good for prediction are overly rich and lack interpretability, this may not apply for situations where there is only little training data in relation to maximal model complexity.

3.2 Parameter estimates

Once the model structure is learned, we also need to fix the model parameters in order to be able to use the resulting fully specified model for tasks such as prediction.

In the Bayesian setting, the parameters can be estimated according to the mean posterior principle [Jaynes, 2003], yielding

$$\hat{\theta}^{\mathrm{MP}}_{\ell ca}(\eta, \mathbf{x}) = \frac{N_{\ell ca} + \frac{\eta|c|}{|\mathcal{A}|^{d+1}}}{N_{\ell c\cdot} + \frac{\eta|c|}{|\mathcal{A}|^{d}}} \quad (5)$$

when using the prior specification of Eggeling et al. [2013]. For PMMs, mean posterior estimates directly correspond to a prediction method integrating over the parameter space, so they are in resonance to structure learning using the Bayesian marginal likelihood of Eq. 2.

An alternative is here also offered by the NML distribution. For parameter learning in BNs, *factorized sequential NML* (fsNML) estimates have been proposed by Silander et al. [2009] to estimate probability parameters in accordance with the fNML structure learning score. For inhomogeneous PMMs, the fsNML estimate writes as

$$\hat{\theta}^{\mathrm{fsNML}}_{\ell ca}(\mathbf{x}) = \frac{e(N_{\ell ca})(N_{\ell ca}+1)}{\sum_{b\in\mathcal{A}} e(N_{\ell cb})(N_{\ell cb}+1)}, \quad (6)$$

where $e(N) = (\frac{N+1}{N})^N$ for $N > 0$ and $e(0) = 1$. Due to its minimax optimality properties, using fsNML for sequential prediction, where the t-th prediction is based

on the $t-1$ previous observations, $t \in \{1,\ldots,N\}$, yields a predictive performance almost as good as the optimal parameter estimates, which are obtained using the maximum likelihood estimator with full data [Silander et al., 2009].

4 EMPIRICAL COMPARISON

In this section, we empirically compare the Bayesian and MDL-based methods for structure learning and parameter estimation. We study the behavior of the aforementioned methods w.r.t. the choice of the prior hyperparameter (Section 4.1) and the sample size (Section 4.2) using the splice site data set of Yeo and Burge [2004]. This data set is particularly suitable for comparing scoring criteria for PMMs since (i) it is known that comparatively strong dependencies among adjacent sequence positions exist in relation to other types of functional oligonucleotides and (ii) the large number of data points originating from the same source offers the possibility to study the influence of the sample size.

4.1 Influence of the ESS

In a first study, we compare the performance of Bayesian and NML-approximating scoring criteria, using the data set of Yeo and Burge [2004] consisting of 12,624 experimentally verified human splice donor sites. The sequences in the data set have length $L = 7$ over the four letter alphabet $\mathcal{A} = \{\text{A,C,G,T}\}$. They have been split by Yeo and Burge [2004] into training data ($\mathbf{x}_{\text{train}}$) and test data ($\mathbf{x}_{\text{test}}$) at a ratio of 2:1, and we rely on the same partion for the following experiments.

We sample $N = 500$ sequences from $\mathbf{x}_{\text{train}}$, learn structure $\hat{\vec{\tau}}$ and probability parameters $\hat{\vec{\theta}}_{\hat{\vec{\tau}}}$ of a third-order inhomogeneous PMM, and compute $\log P(\mathbf{x}_{\text{test}}|\hat{\vec{\Theta}})$, where $\hat{\vec{\Theta}} = (\hat{\vec{\tau}}, \hat{\vec{\theta}}_{\hat{\vec{\tau}}})$. We repeat this procedure 10^3 times, and average the resulting log predictive probabilities in order to let the standard error caused by randomly selecting data points for the training data set become negligible.

Using this procedure, we compare the performance of the NML-approximating method (using fNML structure score and fsNML parameter estimate) with the Bayesian method (using BDeu structure score and MP parameter estimate). The latter offers the possibility to incorporate prior knowledge through the equivalent sample size η and the structure prior hyperparameter κ. Recall that setting the structure prior hyperparameter $\kappa = 1$ results in a uniform structure prior, so that we are able to separately investigate the influence of the ESS. In addition, we perform cross-comparisons by combining fNML structure score with MP parameter estimate and BDeu structure score with fsNML parameter estimate. While theoretically difficult to justify, this cross-comparison might be helpful to evaluate the influence of the ESS on structure and parameter learning separately. In Figure 2, we plot the prediction performance of the various combinations of structure scores and parameter estimates against the ESS (ranging from 10^{-1} to 10^3).

We observe that the performance of the Bayesian method (BDeu-MP) depends strongly on the ESS, with too large values leading to more dramatic degradation in performance than too small values. Moreover, the NML-approximation yields a higher prediction than the Bayesian method irrespective of the chosen ESS. This is surprising, since intuitively the Bayesian method should excel for at least some particular choices of the prior - even if those well-performing choices are typically unknown.

We find a possible explanation of this observation by investigating the cross-comparisons of Bayesian and NML-approximating methods. Since the optimum of the BDeu-fsNML curve (dashed, red) is located at a smaller ESS value than the fNML-MP curve (solid, blue), the optimal ESS for structure learning seems to be not necessarily the optimal ESS for parameter learning. With the comparatively small sample size of $N = 500$, structure learning requires a small ESS, whereas parameter learning requires a larger ESS for better parameter-smoothing.

We also observe that non-optimal choices of the ESS affect parameter learning more severely than structure learning. The error arising from a false model structure is bounded by the prediction performance of the most overfitted model structure, which corresponds here to a third-order inhomogeneous Markov model. However, the error that may arise from parameter learning is virtually unlimited, since the parameter estimates may get dominated by the prior, either yielding estimates close to maximum likelihood (very small ESS) or blurred towards a uniform distribution (very large ESS). We may thus conclude that fsNML is here a safe choice for parameter learning that avoids the explicit specification of any parameter prior.

4.2 Influence of the sample size

After having studied the effect of the ESS at one particular sample size, we now focus on the influence of the sample size on structure learning and prediction. We take a closer look at structure learning by investigating the influence of different structure scores on the complexity of the learned models. To this end, we use a similar experimental setting as described in Section 4.1. We sample N sequences from $\mathbf{x}_{\text{train}}$, learn the

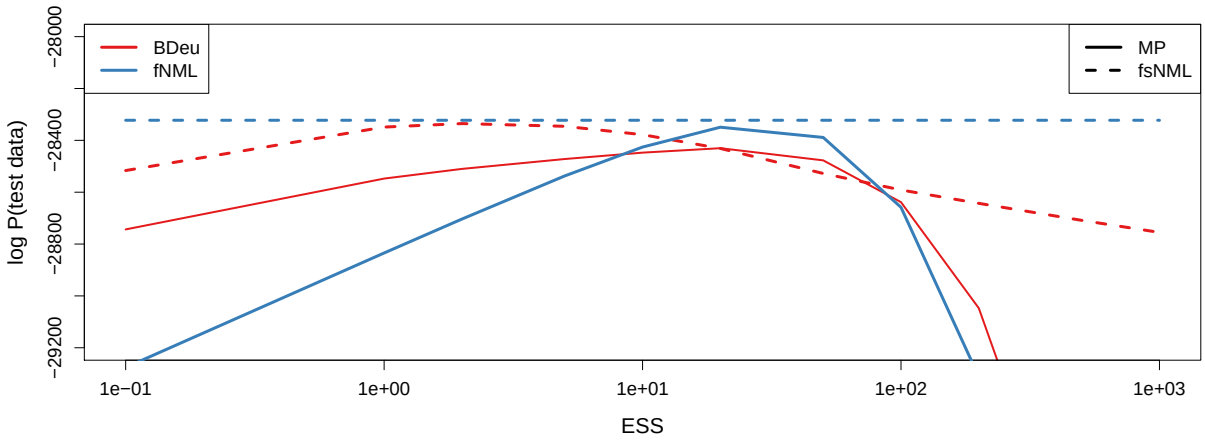


Figure 2: Prediction performance versus ESS of different combinations of structure scores and parameter estimates on the splice site data. The structure score is indicated by the color of each line, with BDeu displayed in red and fNML displayed in blue. The parameter estimate is indicated by the shape of the line, with solid being MP and dashed being fsNML. Hence, the solid red line displays the traditional Bayesian method, whereas the dashed blue line displays the hyperparameter-free NML method. The other two lines are influenced by the ESS either only in structure learning (dashed, red) or in parameter learning (solid, blue).

structures of third-order inhomogeneous PMMs with different scoring criteria, and compute the total number of leaves of all PCTs, which is proportional to the number of model parameters (with a factor of $|\mathcal{A}|$) and can thus be referred to as *model complexity*. We repeat this procedure 10^3 times and average the resulting model complexities. We perform this procedure for different values of N (ranging from 50 to 5000) and plot the model complexity against the sample size (Fig. 3).

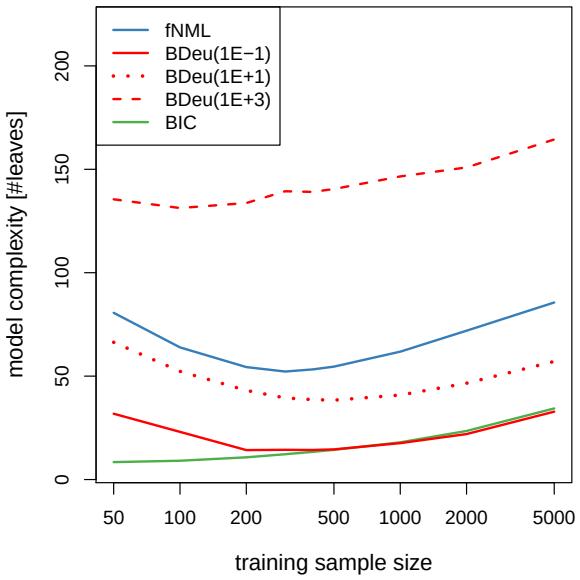


Figure 3: Complexity of the learned model structures w.r.t. the sample size and different structure scores: BIC, fNML, and BDeu (three different ESS values).

For the BDeu score, we observe that the model complexity depends on the ESS: the larger ESS, the larger the model; see the three BDeu curves in Fig. 3. This effect is in agreement to a similar observation made for BNs [Silander et al., 2007].

For all methods, there is a general trend that model complexity increases with increasing sample size from $N = 500$ onwards, but even when using 5000 data points, we do not get close to the maximal model, which has 277 leaves. However, for BDeu and also fNML, the complexity also increases when the sample size becomes very small ($N < 300$). Whereas this seems to be counter-intuitive at first glance, it can be explained with an extreme case: when the sample size decreases to zero, there is no observed data, all terms originating from the parameter prior cancel out, and as a consequence the structure prior dominates model selection. If we assume that for identical optimal scores one of the candidate structures is selected at random, we obtain models of average complexity (w.r.t. the total space of candidate structures). When observing few data points, BDeu collects evidence in favor of either simple or complex models, but the starting point is – in accordance with the uniform structure prior – a model of average complexity. Since fNML implicitly also assumes a uniform structure prior, a similar effect appears for that score as well. Even if the data was actually generated by an independence model, and thus contained no statistical dependencies at all, many data points would be required to to consistently identify the simple model as the correct one, because, unlike BIC, neither of them has a built-in bias towards simple models. BIC, on the other hand, shows a different behav-

ior, as it is known to penalize model complexity heavily in general. But in contrast to BDeu(1E-1), which is similar to BIC for large sample sizes, the model complexity for BIC is monotically increasing, and for very small sample sizes BIC almost selects an independence model.

In a third study, we investigate how different structure scores influence the prediction performance when the sample size varies. We focus compare BIC and fNML structure scores with the extreme cases of independence model and third-order inhomogeneous Markov model, using the fsNML parameter estimate in all four cases. We also include the Bayesian method of BDeu structure score and MP parameter estimate in the comparison, using the same ESS values as in Fig. 3. The experimental setup is here identical to that of the previous experiment, but now we compute predictive probabilities as for different sample sizes N (Fig. 4).

First, we observe that the minimal model (black), which corresponds to an independence model, is optimal when the sample size is smaller than 100 data points. This is intuitively clear since all more complex models are overfitted when the sample size becomes very small. Conversely, the maximal model (purple), which corresponds to a third-order inhomogeneous Markov model, is strongly overfitted when the sample sizes are small, performing significantly worse than the minimal model until the sample size increases to more than approximately 500 data points. For larger sample sizes, the maximal model clearly outperforms the minimal model, which shows that statistical dependencies among adjacent sequence positions exist in the splice site data set.

Using PMMs, we are capable of interpolating between both special cases, if the structure score yields reasonable models. BIC and fsNML scores perform well for large sample sizes, as they are superior to the maximal and to the minimal model.

There are differences for small sample sizes though, which could be expected by merely considering the model complexity for varying sample sizes (Fig. 3), and by the hypothesis that small models might be good for small sample sizes, as the comparison of the fixed structure models illustrates. For small sample sizes, BIC yields model structures that are only slightly more complex than the minimal model, so they also come close to it in terms of predictive probability. However, with increasing sample size, BIC is still capable of capturing dependencies, so it does not suffer from learning very sparse models. In comparison, fNML performs significantly worse for small sample sizes, and the method yields a similar prediction performance compared to BIC only for more than 500 data points.

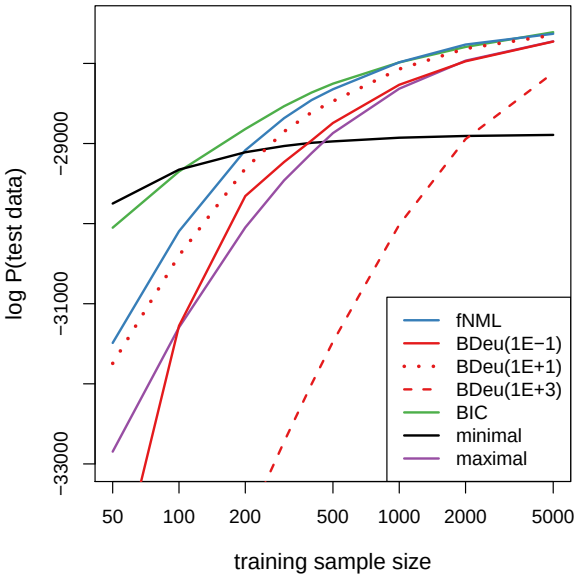


Figure 4: Prediction performance versus sample size for fNML, BIC and BDeu scores with different ESS. In addition the prediction performance of the minimal and the maximal model are shown.

We find that the Bayesian approaches are here inferior to the other three methods that are capable of structure learning, irrespective of the chosen ESS. Whereas $\eta = 10^{-1}$ yields a model complexity that is rather close to that of BIC, it nevertheless suffers from (i) the effect of rising model complexity for small sample sizes and (ii) unfavorable parameter estimates. An ESS value of $\eta = 10^3$ yields for most sample sizes (i) too large model structures and (ii) parameter estimates that are close to a uniform distribution. A reasonable ESS value of $\eta = 10$ performs similar to fNML, yet it suffers from the problem of different ESS optima for structure and parameter learning as discussed in Section 4.1.

In summary, we find that on this data set the minimal model is optimal if there are less than 100 data points available for estimating the distribution from. For sample sizes between 100 and 500, BIC yields the best tradeoff in finding a model structure that captures dependencies while avoiding overfitting, and even for sample sizes below 100 the difference in prediction compared to that of the minimal model is rather small. For $N > 500$ most structure scores perform similar, even though the learned model complexities still differ to a large extent (Fig. 3). For large sample sizes, the negative effect of some dispensable parameter sets may be negligible, which is supported by the comparatively good performance of the maximal model for $N > 2000$, even though it never catches up to the parsimonious models yet.

5 AVOIDING CROSS VALIDATION

We have observed in the previous section that the Bayesian methods for structure learning and parameter estimation of inhomogeneous PMMs are hampered by the influence of the ESS. Especially regarding the cross comparison in Fig. 2, we speculate that it might be impossible to find one ESS value that is optimal for both structure and parameter learning. In practice, this can be dealt with by modifying the structure prior hyperparameter κ, which then overshadows the ESS influence on the model complexity. However, it is intuitively neither clear which value κ has to be chosen to obtain certain model complexities, nor which combination of κ and η may be optimal for prediction. To this end, an internal cross validation on the training data can be used for determining an optimal prior choice [Eggeling et al., 2013]. However, this increases the computational effort dramatically and may limit the large-scale applicability of inhomogeneous PMMs.

5.1 Comparing prediction performances

We speculate that a prediction framework using BIC or fNML structure score in combination with fsNML parameter estimate could represent a reasonable alternative to the Bayesian approach with hyperparameters optimized via internal cross validation. In order to test this hypothesis, we perform a study on several real world data sets. We focus on transcription factor binding sites (TFBS) from the publicly available database Jaspar [Sandelin et al., 2004], since modeling TFBS is the most important application of inhomogeneous PMMs to date. We select all available data sets containing more than 100 sequences, since we have seen in Fig. 4 that below that size the independence model can be expected to yield optimal predictions even if strong dependencies exist in the data. We obtain 20 different data sets, which vary in sequence length from 8 to 21 and in sample size from 101 to 4311, covering the whole range of complexity currently known to appear in TFBS biology. For all data sets, we evaluate the prediction performance of the different methods by cross validation.

For the Bayesian approach, we use an additional internal cross validation (leave-one-out cross validation for all data sets with $N < 1000$, 10-fold cross validation for the rest) for optimizing the hyperparameters. We use three different values (1,10,100) for the ESS and 20 different values for the structure prior hyperparameter κ, interpolating between minimal and maximal model.

The results are shown in Table 1. In most cases, all three methods (BIC-fsNML, fNML-fsNML, and BDeu-MP with double cross validation) show a similar performance, implying that the hyperparameter-free methods are indeed as good as the Bayesian method that uses an exhaustive internal cross validation for hyperparameter tuning.

Table 1: Prediction performance of the three different methods using 20 different TFBS data sets. BDeu uses an internal cross validation on training data for tuning hyperparameters κ and η.

data set	L	N	BIC	fNML	BDeu
EWSR	18	101	-0.32	-0.40	**-0.21**
HIF1A	8	103	-4.51	-4.64	**-4.50**
NFYA	16	116	**-13.39**	-14.27	-13.55
Myc	10	227	**-6.25**	-6.29	-6.27
ESR2	18	357	**-15.46**	-15.62	-15.58
ESR1	20	475	**-18.40**	-18.52	-18.44
Zfx	14	481	**-10.37**	-10.41	-10.38
Stat3	10	613	-4.44	**-4.42**	-4.44
Sox2	15	669	-11.59	-11.67	**-11.58**
Foxa2	12	808	-7.35	-7.36	**-7.32**
PPARG	15	864	-12.58	**-12.52**	-12.54
FOXA1	11	897	**-6.46**	-6.53	-6.54
CTCF	19	908	-13.65	-13.72	**-13.62**
GABPA	11	993	-5.94	-5.96	**-5.93**
Pou5f1	15	1356	-10.44	**-10.42**	-10.44
REST	21	1607	-12.60	-12.63	**-12.59**
STAT1	15	2085	**-11.97**	-11.99	-11.99
Esrrb	12	3661	-7.44	**-7.43**	-7.45
Tcfcp2l1	14	4079	-11.06	-11.02	**-11.01**
Klf4	10	4311	-5.07	**-5.07**	-5.07

One interesting example is NFYA, where BIC leads to a clearly increased prediction performance compared to fNML and where it also predicts better than the Bayesian approach. This can be explained by the fact that a small data set may require a rather simple model despite containing strong statistical dependencies. So NFYA is an example of a situation that we have simulated in Section 4.2 by subsampling (Fig. 4). As we have seen before, fNML and also BDeu may have difficulties learning simple models when there is not a sufficient amount of data available to consistently identify the simple model as the correct one.

Hence, BIC might be the best structure score if it is combined with fsNML parameter estimates and if the sample size is small ($N < 500$) in relation to the maximal model complexity and the total number of possible structures. However, it might be possible that BIC underfits when there are strong and diverse statistical dependencies, requiring large model structures. If the optimal model complexity is above average or if sample size is comparatively large ($N > 500$), fNML might be the more robust choice.

These observations actually yield a vague prior knowledge about model complexity for analyzing further data sets, as we now expect comparatively sparse models to perform well. However, since it is difficult to translate this vague knowledge into precise values for the structure prior hyperparameter κ, using BIC constitutes a reliable method of obtaining rather sparse model structures and thus expressing our recently gained vague prior belief.

5.2 Runtime considerations

We have seen that BIC and fNML structure scores in combination with fsNML parameter estimares might be a alternative learning approach to Bayesian methods if there is no or only vague prior knowledge available, and the prior must be tuned by cross validation. Cross validation multiplies the time complexity of the entire learning algorithm by KC, where K is the number of holdouts, and C is the number of different prior values to be tested. In Section 5.1, we used $K = 10$ for the large data sets with more than 1000 data points, and $K = N - 1$ for the remaining data sets. C is the product of different ESS values η (3 in our studies) and different structure prior hyperparameter values κ (20 in our studies).

In practice, learning one third-order inhomogeneous PMM (implemented in Java using Jstacs [Grau et al., 2012]) from the Pou5f1 data set on a 2.5 GHz processor takes 1.8 seconds using BIC-fsNML and 1.6 seconds using fNML-fsNML. The runtime is here dominated by the dynamic programming algorithm for finding the optimal model structures. Learning a similar model with the Bayesian approach using cross validation takes 1036 seconds, which is indeed close to a factor of $3 \times 20 \times 10 = 600$ times slower. For other large data sets, the runtime ratio is similar close to the expectation. For smaller data sets, this is even more unfavorable for the Bayesian method, since leave-one-out cross validation must be used to obtain robust hyperparameter estimates.

The runtime depends on the numbers K and C, but without a doubt compromises w.r.t. both values, such as reducing the number of tested candidates, might reduce the runtime difference between methods and yet not decrease the performance significantly. However, it entails the danger of obtaining unreliable estimates, which is probably worse than needlessly investing more time. In addition, transforming the problem of choosing good hyperparameter values into choosing an appropriate selection of hyperparameter candidate values for the internal cross validation is essentially not avoiding user interference at all.

6 CONCLUSIONS

In this work, we have studied the empirical performance of different learning approaches for inhomogeneous PMMs. The current state of the art is a Bayesian approach, which may be problematic since the prior choice can influence both structure learning and parameter estimation dramatically. We examined alternative learning methods, motivated by the MDL principle, which were originally proposed for Bayesian networks.

We found that the factorized sequential NML estimate is a safe choice for obtaining probability parameters, as it always provides a certain parameter smoothing without dominating the data, whereas mean posterior estimates perform poorly if the ESS parameter is chosen inappropriately. For structure learning, BIC is a surprisingly good choice, especially when sample sizes are small or the expected optimal model complexity is below average. In those cases, it outperforms all other alternatives, which may be surprising at first glance, since BIC can be thought of as a large sample approximation of more sophisticated scores. However, whereas the Bayesian and the fNML score may be tailored towards finding the true model structure, and may outperform BIC in that respect, this true structure may not be optimal for prediction using limited training data. Prediction on real-world data sets such as DNA binding sites seems to favor comparatively sparse model structures, which in turn favor BIC as structure score for obtaining them.

The Bayesian approach for learning inhomogeneous PMMs offers the possibility of tuning the hyperparameters of structure and parameter prior via cross validation in case that no or only vague prior knowledge is available. Whereas a generally robust behavior of the MDL-techniques compared to the Bayesian approach was promised by theory, it was not clear how these methods perform in comparison to an exhaustive prior optimization. The results from this work suggest that the extensive cross validation, which is used to compensate the shortcomings of Bayesian methods, may be unnessary since a similar – and sometimes even better – performance can be achieved by using robust methods for structure learning and parameter estimation that avoid explicit prior specification.

Acknowlegdments

This work was funded by *Reisestipendium des allg. Stiftungsfonds der MLU Halle-Wittenberg* and the Academy of Finland (Centre of Excellence COIN and Project PRIME). This work was carried out while the first author was visiting Helsinki Institute for Information Technology (HIIT).

References

P.-Y. Bourguignon and D. Robelin. Modèles de Markov parcimonieux. In *Proceedings of JOBIM*, 2004.

W. Buntine. Theory refinement on Bayesian networks. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, pages 52–60. Morgan Kaufmann, 1991.

R. Eggeling, A. Gohr, P.-Y. Bourguignon, E. Wingender, and I. Grosse. Inhomogeneous parsimonious Markov models. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013*, volume 1, pages 321–336. Springer, 2013.

J. Grau, J. Keilwagen, A. Gohr, B. Haldemann, S. Posch, and I. Grosse. Jstacs: A Java framework for statistical analysis and classification of biological sequences. *Journal of Machine Learning Research*, 13:1967–1971, 2012.

P. Grünwald. *The Minimum Description Length Principle*. MIT Press, June 2007.

G. Heckerman, D. Geiger, and D. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

E. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

P. Kontkanen and P. Myllymäki. A linear-time algorithm for computing the multinomial stochastic complexity. *Information Processing Letters*, 103(6): 227–233, September 2007.

P. Kontkanen, W. Buntine, P. Myllymäki, J. Rissanen, and H. Tirri. Efficient computation of stochastic complexity. In *Proceedings of the Ninth International Conference on Artificial Intelligence and Statistics*, 2003.

J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

J. Rissanen. A universal data compression system. *IEEE Trans. Inform. Theory*, 29(5):656–664, 1983.

A. Sandelin, W. Alkema, P. Engström, W. Wasserman, and B. Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32:D91–D94, 2004.

G. E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 2:461–464, 1978.

Y. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23: 3–17, 1987.

T. Silander, P. Kontkanen, and P. Myllymäki. On sensitivity of the MAP Bayesian network structure to the equivalent sample size parameter. In *Proceedings of the The 23rd Conference on Uncertainty in Artificial Intelligence (UAI-2007)*, pages 360–367, 2007.

T. Silander, T. Roos, P. Kontkanen, and P. Myllymäki. Factorized NML criterion for learning Bayesian network structures. In *Proceedings of the 4th European Workshop on Probabilistic Graphical Models (PGM-08)*, 2008.

T. Silander, T. Roos, and P. Myllymäki. Locally minimax optimal predictive modeling with Bayesian networks. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, pages 504–511, 2009.

T. Silander, T. Roos, and P. Myllymäki. Learning locally minimax optimal Bayesian networks. *International Journal of Approximate Reasoning*, 51:544–557, 2010.

G. Yeo and C. Burge. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of Computational Biology*, 11(2/3):377–394, 2004.