

Adaptive simulated annealing EM algorithm for inference on non-homogeneous hidden Markov models

Aliaksandr Hubin

aliaksandr.hubin@nr.no

Norwegian Computing Center

Oslo, Norway

ABSTRACT

Non-homogeneous hidden Markov models (NHHMM) are a subclass of dependent mixture models used for semi-supervised learning, where both transition probabilities between the latent states and mean parameter of the probability distribution of the responses (for a given state) depend on the set of p covariates. A priori we do not know which (and how) covariates influence the transition probabilities and the mean parameters. This induces a complex combinatorial optimization problem for model selection with 4^p potential configurations. To address the problem, in this article we propose an adaptive (A) simulated annealing (SA) expectation maximization (EM) algorithm (ASA-EM) for joint optimization of models and their parameters with respect to a criterion of interest.

CCS CONCEPTS

- Mathematics of computing → Combinatorial optimization; Variable elimination; Expectation maximization; Simulated annealing;
- Computing methodologies → Probabilistic reasoning; Latent variable models.

KEYWORDS

simulated annealing, hidden Markov models, model selection, adaptive parameter tuning, expectation maximization

ACM Reference Format:

Aliaksandr Hubin. 2019. Adaptive simulated annealing EM algorithm for inference on non-homogeneous hidden Markov models. In *AIIPCC2019: 2019 International Conference on Artificial Intelligence, Information Processing and Cloud Computing, December 19–21, 2019, Sanya, China*. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3357750.3357764>

1 INTRODUCTION

Hidden Markov models (HMM) represent a tool used for pattern recognition, inference and predictions in both social and natural sciences, in different areas and for different purposes. For example, in economics and econometrics, they are used for modelling regime-switching processes [10] and various jump processes [9]. In psychology, they are used for learning process modelling [28]. Other applications include speech recognition [11], biology and

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

AIIPCC2019, December 19–21, 2019, Sanya, China

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-7633-4...\$15.00

<https://doi.org/10.1145/3357750.3357764>

genetics [20]. Non-homogeneous hidden Markov models (NHHMMs) are the extension of standard HMMs that allow to model both transition probabilities between the latent states and mean parameters of the probability distribution of the responses with respect to a set of covariates [18]. NHHMMs typically have the same applications as standard HMMs [21], but allow additional flexibility. In this article, we address the model (variable) selection problem in NHHMMs. Model selection problem in terms of the choice of observations influenced by the transitions of a hidden Markov chain is well studied for simple HMMs and hidden semi-Markov models (HSMMs). Existing methods include variational Bayes (VB) [29], Markov chain Monte Carlo (MCMC) [3] and adaptations of expectation maximization (EM) [2], and are summarized in Adams and Beling [1]. At the same time, the inverse problem of selecting exogenous variables that, conditionally on the hidden state, influence the distributions of the observations as well as the problem of selecting exogenous variables influencing transition probabilities are much less studied. In Paroli and Spezia [23], a comparison of different strategies of variable selection in NHHMMs is given, however, there only Gaussian observations are considered, the problem is limited to only inference (with no guidance to predictions), and only variable selection for transition probabilities is allowed. The extended MCMC based approach by the same authors is considered in Paroli and Spezia [24], where both variable selection for transitions and observations as well as the problem of selecting the number of hidden states are studied, however, the approach is still limited in the sense that only Gaussian observations are addressed and no guidance for model-based predictions is given. A more recent prediction driven approach based on reversible jump MCMC [13] is suggested in Meligkotsidou and Dellaportas [22], however, there, also, only Gaussian observations are considered with variable selection only on the level of transition probabilities and no guidance for pattern recognition driven applications. All of the approaches mentioned above are certainly sound mathematically and can be extended to broader problems with a moderate effort, however, they (as of today) remain slightly impractical for several reasons: 1. They are computationally heavy; 2. They are not designed for parallel computing; 3. No implementations or software are publically available for the scientific community; 4. Only Gaussian observations are considered. This results in that a) the approaches are only suited for small problems (6 covariates at most were considered in application of [22–24]), b) even more recent works cannot easily adapt them and still rely on manual variable selection [15], c) it is impossible to benchmark against them on new data sets without reimplementing the approaches from scratch.

The main contribution of this paper is a pragmatic, yet rigorous, solution for variable and model selection problems in NHHMMs.

We consider any distribution of the observations from the exponential family and address both pattern recognition and prediction driven examples. Additionally, we enable practical parallel computing (allowing to perform variable selection for much larger sets of covariates) and provide a GitHub library¹ for common use. In particular, we suggest an adaptive irreducible simulated annealing expectation maximization algorithm (ASA-EM) and its parallel version for joint optimization of models and their parameters with respect to a criterion of interest, which can be chosen among Akaike information criterion (AIC), Bayesian information criterion (BIC), Deviance information criterion (DIC), focused information criterion (FIC), marginal log-likelihood (MLIK), marginal posterior model probability (PMP) and others [5]. Alternatively, a time-aware cross-validation can be used. All of the mentioned model selection strategies induce regularization on the model complexity and hence prevent the model from overfitting. The suggested ASA-EM algorithm has an asymptotic guarantee to find the optimal model and values of its parameters, however, as we will show in the experiments section, it works well even for the limited time computations. To illustrate its performance, we address an example for financial stock prices data from *S&P 500* (30 covariates), and an example dealing with epigenetic observations from *Arabidopsis* plant (17 covariates). Both of the examples address significantly larger sets of covariates than reported in all of the previous studies [22–24].

2 MATHEMATICAL MODEL

We assume a parametric probability distribution \hat{f} of the responses $Y_i, i \in \{1, \dots, n\}$, which come from the exponential family [5]. We define separate mean (μ_{s_i}) and dispersion (ϕ_{s_i}) parameters, conditional on the latent states $s_i \in \{1, \dots, S\}$, which are assumed to have a Markovian dependence structure along indices $\{1, \dots, n\}$ allowing to incorporate a simple *temporal* dependence between the observations:

$$Y_i | \mu_{s_i}, \phi_{s_i}, s_i \sim \hat{f}(y | \mu_{s_i}, \phi_{s_i}), \quad \mu_{s_i} = g^{-1}(\eta_{s_i}), \quad (1)$$

$$\eta_{s_i} = \beta_0^{(s_i)} + \sum_{j=1}^p \gamma_j \beta_j^{(s_i)} x_{ij}, \quad (2)$$

$$p(s_i | s_{i-1}) = \frac{e^{\omega_0^{(s_i, s_{i-1})} + \sum_{j=1}^p \delta_j \omega_j^{(s_i, s_{i-1})}}}{1 + e^{\omega_0^{(s_i, s_{i-1})} + \sum_{j=1}^p \delta_j \omega_j^{(s_i, s_{i-1})}}} x_{ij}. \quad (3)$$

Here $\beta_j^{(s_i)} \in \mathbb{R}, j \in \{0, \dots, p\}$ are regression coefficients of the covariates of the model showing for a given state s_i whether and how the corresponding covariate influences the mean parameter μ_{s_i} of the distribution of the responses, $\gamma_i \in \{0, 1\}, i \in \{1, \dots, M\}$ are latent indicators, defining if (for any state) covariate i is included into the linear predictor of the model ($\gamma_i = 1$) or not ($\gamma_i = 0$). $\omega_j^{(s_i, s_{i-1})} \in \mathbb{R}, j \in \{0, \dots, p\}$ are regression coefficients of the covariates of the model showing for a given state s_{i-1} whether and in which way the corresponding covariate influences the transition probabilities to state s_i . $\delta_i \in \{0, 1\}, i \in \{1, \dots, p\}$ are latent indicators, defining if (for any state) covariate i is included into the model for transition probabilities between the latent states ($\delta_i = 1$) or not ($\delta_i = 0$). Here both the mean parameters of the responses and

¹<https://github.com/aliaksah/depmixS4pp>

the transition probabilities of the Markov chain depend upon up to p covariates and the combinatorics of the choice of the covariates (model configurations) is incorporated via the latent binary indicators, which switch the covariates on and off separately for the latent Markov chain of the states (through δ_i) and the distribution of the responses (through γ_i). At the same time, to avoid super-exponential explosion of the number of candidate models, we assume that the same variables are influencing the observations for every hidden state and the same pattern of variables is influencing transition probabilities between all pairs of hidden states. The following section will present the algorithm for fitting the model (1)-(3).

3 ADAPTIVE SIMULATED ANNEALING EM ALGORITHM

In order to deal with the presence of latent states and latent binary indicators for covariate selection as well as multimodality in the joint space of model configurations $\mathbf{m} = \{\boldsymbol{\delta}, \boldsymbol{\gamma}\} \in \mathcal{M}$ and their parameters $\boldsymbol{\theta}_m = \{\boldsymbol{\beta}, \boldsymbol{\phi} | \mathbf{m}\} \in \Theta_m$ we will combine the adaptive simulated annealing (SA) algorithm (for model space exploration) and the standard expectation maximisation (EM) algorithm (for parameter space exploration) into the novel ASA-EM algorithm, described in Algorithm 2. There we consider the objective criterion $r(\mathbf{m}, \boldsymbol{\theta}_m | Y, \mathbf{x})$ to be minimized across $\mathbf{m} \in \mathcal{M}$ and $\boldsymbol{\theta}_m \in \Theta_m$. The objective $r(\mathbf{m}, \boldsymbol{\theta}_m | Y, \mathbf{x})$ can be chosen as AIC, BIC, DIC, FIC, MLIK or PMP [5], with no loss of generality. Acceptance ratio of ASA-EM Algorithm 2 at temperature τ is defined as

$$a_\tau(r_1, r_2) = \min \left\{ 1, \exp \frac{(r_1 - r_2)}{\tau} \right\}. \quad (4)$$

Here $r_1 = r(\mathbf{m}_1, \boldsymbol{\theta}_{m_1} | Y, \mathbf{x})$ and $r_2 = r(\mathbf{m}_2, \boldsymbol{\theta}_{m_2} | Y, \mathbf{x})$.

3.1 The EM algorithm

In this subsection, we describe the EM part of the ASA-EM algorithm in more detail.

Conditional on having a fixed model m from the model space \mathcal{M} , the standard expectation maximization (EM) algorithm is used to make inference on the parameters $\boldsymbol{\theta}_m$ of the model m [20]. The inference algorithm here works as follows: One first introduces the likelihood of the model parameters $p(Y | \mathbf{x}, \boldsymbol{\theta})$. Here, however, the parameters of the model are not defined without the sequence of latent states \mathbf{s} for all of the observations. We hence add them and then integrate out using the law of total probability as:

$$p(Y | \mathbf{x}, \boldsymbol{\theta}) = \sum_{\mathbf{s} \in \Omega} p(Y, \mathbf{s} | \mathbf{x}, \boldsymbol{\theta}).$$

At this point, one still does not know explicitly $p(Y, \mathbf{s} | \mathbf{x}, \boldsymbol{\theta})$. To resolve this, one can multiply and divide the expression by some auxiliary $q(\mathbf{s} | \boldsymbol{\theta})$:

$$p(Y | \mathbf{x}, \boldsymbol{\theta}) = \sum_{\mathbf{s} \in \Omega} q(\mathbf{s} | \boldsymbol{\theta}) \frac{p(Y, \mathbf{s} | \mathbf{x}, \boldsymbol{\theta})}{q(\mathbf{s} | \boldsymbol{\theta})}.$$

In the E-step of the EM algorithm one obtains the likelihood explicitly by taking expectation with respect to $q(\mathbf{s} | \boldsymbol{\theta}^*)$. This means getting:

$$p(Y | \mathbf{x}, \boldsymbol{\theta}) = \mathbb{E}_{q(\mathbf{s} | \boldsymbol{\theta}^*)} \left(\frac{p(Y, \mathbf{s} | \mathbf{x}, \boldsymbol{\theta})}{q(\mathbf{s} | \boldsymbol{\theta}^*)} \right). \quad (5)$$

Here, θ^* are the values of parameters obtained at a previous M-step. Further the M-step consists of maximizing the expectation (5) with respect to θ to obtain the new θ^* , which will be used in the new EM iteration. Typically, $q(s|\theta)$ utilizes the Markovian properties of the latent sequence of states to make the algorithm computationally more efficient. Also note that in the E-step, analytical expectation can be exchanged with importance sampling, if the former is difficult to obtain.

3.2 SA model exploration algorithm

Having the methodology for estimating θ_m within the model, we can move on to the model selection algorithm, which is an adaptive simulated annealing procedure. We will here first describe the standard SA with fixed hyperparameters.

The standard simulated annealing algorithm generates a sequence of *solutions* $\{z_1, \dots, z_k\}$ minimizing some objective $G(z)$ ² as described in Algorithm 1.

Algorithm 1 Standard simulated annealing optimization

```

 $z \leftarrow z_0;$ 
 $z_b \leftarrow z_0;$ 
 $\text{for } \tau \text{ in } (\tau_{max}, \dots, \tau_{min}) \text{ do}$ 
     $\text{for } k \text{ in } \{1, \dots, K\} \text{ do}$ 
         $z_c \leftarrow N(z);$ 
         $\text{if } a_\tau(G(z), G(z_c)) \geq u \sim Unif[0; 1] \text{ then}$ 
             $z \leftarrow z_c;$ 
             $\text{if } G(x_c) < G(x_b) \text{ then}$ 
                 $z_b \leftarrow z_c;$ 
             $\text{end if}$ 
         $\text{end if}$ 
     $\text{end for}$ 
 $\text{end for}$ 
 $\text{return } z, z_b$ 

```

Here the cooling schedule is assumed to be exponential, i.e.

$$\tau_{i+1} = \tau_i \exp(-\kappa_t).$$

Acceptance probabilities for the simulated annealing kernel are of form (4) but with $r_1 = G(z_1)$ and $r_2 = G(z_2)$. The initial point z_0 is drawn randomly. The limiting distributions of SA for each temperature τ are found (due to the Markovian property of the procedure) as follows:

$$p(z_k = z | \tau) \propto \exp\left(-\frac{G(z)}{\tau}\right). \quad (6)$$

Note that if $G(z)$ is the likelihood multiplied by the prior scaled with τ , i.e. for $G(z) = -\tau \log p(Y|x, z)p(z)$ the SA corresponds to the Metropolis-Hastings MCMC algorithm [12] and hence gives samples from the posterior distribution of z . This can be easily utilized in our approach if one is interested in the whole *posterior distribution of models and parameters* rather than one *best* solution.

3.3 Adaptive SA hyperparameter tuning

Finally, the adaptive part of the ASA-EM acronym means that the algorithm is gradually learning its own hyperparameters over some time before switching to a stationary regime after R_a epochs. The tuning process is described in this subsection.

²In our case the solutions are different models and their parameters, i.e. $z_i = \{\mathbf{m}_i, \theta_{m_i}\}$, but we will in this section address z_i , because the focus for the reader here is understanding the standard simulated annealing procedure. Additionally for simplicity we consider a general objective $G(z)$ to be minimized, whereas the focus of ASA-EM will be a specific $G(z) := r(\mathbf{m}, \theta_m | Y, \mathbf{x})$

♦ For a random variable of the number of iterations of EM algorithm l we set $l \sim \text{Poisson}(\lambda)$ with the parameter prior of the form $\lambda \sim \text{Gamma}(a_\lambda, b_\lambda)$, which is conjugate to the distribution of l . Then throughout training, we draw proposals l from the following posterior distribution:

$$p(l|D_l) \stackrel{d}{=} \text{NB}\left(a_\lambda + \sum_{x \in D_l} x, \frac{1}{1 + b_\lambda + ||D_l||}\right), \quad (7)$$

which we call *adaptive posterior distribution*. Here $||\vec{v}||$ is a L_0 norm representing the length of the vector \vec{v} and D_l is the synthetic data on the (what we call) *successful* iterations, that is if a new global optima in terms of $r(\mathbf{m}, \theta_m | Y, \mathbf{x})$ or the move is accepted by SA algorithm at temperature $\tau < 1$, we append the current l to D_l .

♦ The number of iterations K of SA per temperature τ is also assumed Poisson distributed $K \sim \text{Poisson}(\mu)$ with a conjugate prior $\mu \sim \text{Gamma}(a_\mu, b_\mu)$. During training we draw proposals K from the adaptive posterior distribution:

$$p(K|D_K) \stackrel{d}{=} \text{NB}\left(a_\mu + \sum_{x \in D_K} x, \frac{1}{1 + b_\mu + ||D_K||}\right), \quad (8)$$

where again D_K is the synthetic data, to which we append the current K on *successful* iterations.

♦ Similarly, for the size c of the neighbourhood operator $N_c(\mathbf{m})$ of SA we assume $c \sim \text{Binomial}(C, p_c)$ with $p_c \sim \text{Beta}(a_c, b_c)$. Then the posterior adaptive distribution for c becomes:

$$p(c|D_c) \stackrel{d}{=} \text{BetaBin}\left(a_c + \sum_{x \in D_c} x, b_c + C \cdot ||D_c|| - \sum_{x \in D_c} x\right). \quad (9)$$

Here, for *successful* iterations we append c to the synthetic data D_c .

♦ Finally, for the probabilities of having the covariates included into the model $\Psi_j, j \in \{1, \dots, p\}$ we consider the Multinomial prior $\Psi_j \sim \text{Multinomial}(p_{j1}, p_{j2}, p_{j3}, p_{j4})$, where the four states correspond to $\{\gamma_j = 0, \delta_j = 0\}$, $\{\gamma_j = 1, \delta_j = 0\}$, $\{\gamma_j = 0, \delta_j = 1\}$ and $\{\gamma_j = 1, \delta_j = 1\}$ accordingly. Here, we set a conjugate Dirichlet prior for the hyperparameters of the Multinomial distribution $\{p_{j1}, p_{j2}, p_{j3}, p_{j4}\} \sim \text{Dirichlet}(\zeta_1, \zeta_2, \zeta_3, \zeta_4)$ leading to a Dirichlet-Multinomial posterior adaptive distribution:

$$\begin{aligned}
p(\Psi_j|D_{\Psi_j}) &\stackrel{d}{=} \text{DirMult}(\zeta_1 + \sum_{x \in D_{\Psi_j}} \mathbb{I}(x = \{0, 0\}), \\
&\quad \zeta_2 + \sum_{x \in D_{\Psi_j}} \mathbb{I}(x = \{0, 1\}), \zeta_3 + \sum_{x \in D_{\Psi_j}} \mathbb{I}(x = \{1, 0\}), \\
&\quad \zeta_4 + \sum_{x \in D_{\Psi_j}} \mathbb{I}(x = \{1, 1\})). \quad (10)
\end{aligned}$$

Here, for *successful* iterations we update the synthetic data D_{Ψ_j} with the currently drawn Ψ_j . Other tuning parameters of the algorithm including the cooling schedule and the stopping time for adaptive learning are assumed fixed.

Algorithm 2 Adaptive Simulated Annealing Expectation Maximization

```

initialize { $m^g$ ,  $\theta_{mg}$ };
for  $r$  in  $\{1, \dots, R\}$  do
    initialize { $m^c$ ,  $\theta_{mc}$ };
    for  $\tau$  in  $\{\tau_{max}, \dots, \tau_{min}\}$  do
        sample  $K \sim p(K|D_K)$ ;
        for  $k$  in  $\{1, \dots, K\}$  do
            sample  $c \sim p(c|D_c)$ ;
            choose  $J$  from  $N_C(m)$ ;
            sample  $m_j^*$  from  $p(\Psi_j|D_\Psi)$  for  $j \in J$ ;
        initialize  $\theta_{m^*}$  for  $m^*$ ;
        sample  $l \sim p(l|D_l)$ ;
        run EM for  $I$  iterations and obtain  $\hat{\theta}_{m^*}$ ;
        if  $a_\tau \left( r(m^c, \theta_{mc}|Y, x), r(m^*, \theta_{m^*}|Y, x) \right)$  then
            set  $\{m^c, \theta_{mc}\} \leftarrow \{m^*, \theta_{m^*}\}$ ;
            if  $r < R_a$  and  $\tau < 1$  then
                 $D_l \leftarrow D_l \cup l, D_c \leftarrow D_c \cup c, D_K \leftarrow D_K \cup K, D_\Psi \leftarrow D_\Psi \cup \Psi$ ;
            end if
            if  $r(m^c, \theta_{mc}|Y, x) < r(m^g, \theta_{mg}|Y, x)$  then
                set  $\{m^g, \theta_{mg}\} \leftarrow \{m^c, \theta_{mc}\}$ ;
                if  $r < R_a$  then
                     $D_l \leftarrow D_l \cup l, D_c \leftarrow D_c \cup c, D_K \leftarrow D_K \cup K, D_\Psi \leftarrow D_\Psi \cup \Psi$ ;
                end if
            end if
        end if
    end for
end for

```

The ASA-EM Algorithm 2 converges to a homogeneous Markov chain by design, since the adaptations of the hyperparameters are stopped after R_a epochs. Hence after R_a epochs we end up with a standard simulated annealing algorithm for the model exploration combined with a standard expectation maximization algorithm for parameter estimation. The resulting algorithm is represented by a Markov chain and converges to a stationary distribution at any given temperature τ independently of the starting solution. At the same time, the EM part of the algorithm is greedy and thus it finds only the local optimum, which depends on the starting points. Multiple revisits of the same model m with different initial points for parameters in Θ_m resolves this issue, if there is only a finite countable number of local extrema of the objective in Θ_m . ASA-EM has a guarantee to find the global optimum asymptotically since it is a positively recurrent Markov chain in $\{\mathcal{M} \cup_{m \in \mathcal{M}} \Theta_m\}$. However, in the limiting case, it needs exponential time to do this. Otherwise (within a small finite amount of time) it can be seen as a metaheuristic optimization strategy for model selection in non-homogeneous hidden Markov models.

The algorithm can be embarrassingly parallelized in a straight forward fashion either from the beginning (with the same or different tuning parameters for each thread) or after R_a epochs ³. Note that additionally different number of hidden states can be considered in different threads by e.g. sampling from a uniform integer distribution within a reasonable range, allowing to combine variable selection problem with the problem of selecting the optimal number of states.

4 EXPERIMENTS

To study the performance of the proposed model and algorithm in the case of limited computational time we will address the data from the first chromosome of Arabidopsis plant belonging to several predefined groups of genes and the logreturns data from the *S&P 500* listing. The first example will serve as a pattern recognition

³Whilst we do not give any guidance on how many R_a epochs should be considered, we numerically show that even 2-3 R_a epochs are typically enough, provided that at least one additional epoch is run to guarantee ergodicity of the Markov chain

Table 1: hyperparameters of ASA-EM. Here $U[a, b]$ is a uniform distribution on the interval from a to b .

ζ_1	ζ_2	ζ_3	ζ_4	a_λ	b_λ	a_μ
0	0	0	1	100 * $U[1, 2]$	$U[1, 10]$	$5 * U[3, 7]$
b_μ	a_s	b_s	R_a	τ_{min}	τ_{max}	k_t
2	5	15	3	$5 * U[10^{-7}, 10^{-2}]$	$2 * U[10^4, 10^9]$	$U[2, 6]$

Table 2: The selected model for epigenetics study.

	BIC	AIC	Init. State 1	Init. State 2
	1481.928	1412.839	1.000	0.000
→ State 1	Intercept	X_{M_d}	X_{EXPR}	
State 1	0.000	0.000	0.000	
State 2	-129.264	119.104	0.000	
→ State 2	Intercept	X_{M_d}	X_{EXPR}	
State 1	0.000	0.000	0.000	
State 2	5.478	0.288	0.000	
Observations	Intercept	X_{CHG}	X_{CG}	
State 1	-5.518	-79.583	-1.770	
State 2	-2.303	1.714	2.152	

study alongside the genome, whilst the second example is used as a prediction driven experiment.

4.1 Epigenetic data example

The addressed data set consists of 1502 observations from the first chromosome of Arabidopsis plant belonging to several predefined groups of genes. The observations are represented by the methylated Y_i versus total N_i number of reads inducing the binomial distribution of the responses. The hidden Markov chain here has just two states (1 - methylated and 0 - non-methylated), which results in the following model specification:

$$Y_i | p_{s_i}, s_i, N_i \sim \text{Binomial}(p_{s_i}, N_i), \quad p_{s_i} = \text{logit}^{-1}(\eta_{s_i}), s_i \in \{0, 1\}.$$

We also have data on various exogenous variables (covariates). Among these covariates, we address (following [17]) the factor with 3 levels corresponding to whether the location belongs to CGH, CHH or CHG genetic region, where H is either A, C or T and thus generating two covariates X_{CG} and X_{CHG} . The second group of factors indicates whether the distance to the previous cytosine nucleobase (C) in DNA is 1, 2, 3, 4, 5, from 6 to 20 or greater than 20 inducing six binary covariates $X_{DT1}, X_{DT2}, X_{DT3}, X_{DT4}, X_{DT5}$, and $X_{DT6:20}$. We also include such 1D distance as a continuous covariate X_{DIST} . The third addressed group of factors corresponds to whether the location belongs to a gene from a particular group of genes of biological interest. These groups are indicated as M_a , M_g and M_d , yielding two additional covariates X_{M_a}, X_{M_d} . Additionally, we have a covariate X_{CODE} indicating if the corresponding nucleobase is in the coding region of a gene and a covariate X_{STRD} indicating if the nucleobase is on a "+" or a "-" strand. Finally, we have a continuous covariate $X_{EXPR} \in \mathbb{R}^+$ representing expression level for the corresponding gene and interactions between expression levels and gene groups $X_{EXPR,a}, X_{EXPR,g}, X_{EXPR,d} \in \mathbb{R}^+$. Thus multiple predictors with respect to a strict choice of the reference model in our example induced $M = 17$ potentially important covariates, yielding $4^{17} > 17 \times 10^9$ potential models to consider.

We run 30 parallel threads of ASA-EM algorithm with the choice of hyperparameters shown in Table 1, whilst $p(s|\theta)$ and sampling in

the E-step are the default choices from Visser et al. [27]. The distribution of the values of model selection criteria (AIC and BIC) across the runs is shown in Figure 1 with BIC used as a primary criterion for optimization. Then the best model found in all of the threads in terms of BIC was selected for the given pattern recognition task. According to Figure 1. Exactly this model was found in 10% of the threads, whilst in 87.7% of the threads, the deviation of the BIC value of the found solution is almost neglectable in comparison to the best one found. The structure of the observations together with the discovered pattern of the most probable sequence of states, representing the methylation status of the observations, found by the suggested ASA-EM procedure, are represented in Figure 2, whilst the selected model is summarized in Table 2.

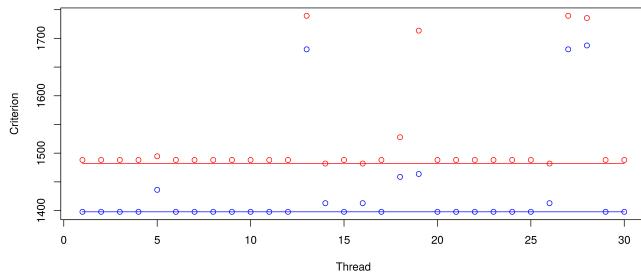


Figure 1: Values of BIC (red) and AIC (blue) across 30 parallel threads of ASA-EM algorithm for epigenetics data.

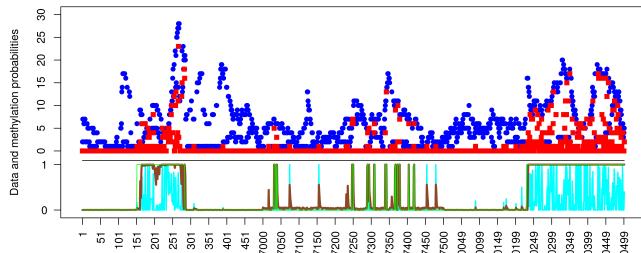


Figure 2: Epigenetic observations, where blue dots are the total number of reads, red dots - number of methylated reads, the light blue line gives naïve probabilities as rates, the green line - most probable latent states of the non-homogeneous hidden Markov model, the brown line - probability of a latent state being 1.

According to the BIC criterion, only factors X_{CHG} and X_{CG} are significant on the level of observations (conditional on the state) of the methylation for the addressed epigenetic region. At the same time, the factor X_{Ma} and the level of expression X_{EXPR} are significant for the transition probabilities between the latent states. The detailed influence on the linear predictors and transition is given in Table 2. Based on the results for the best model we carried out computations of methylation probabilities of the locations along the genome. Furthermore, we compared the results with the naïve approach based on computing the proportion of methylated reads, which is currently addressed in the biological literature as a standard way to evaluate methylation probability of a given nucleobase.

These results are summarized in Figure 2. The results show that the naïve approach should not be trusted in the presence of spatially correlated data and the corresponding to it probabilities are strongly biased. Previously, a similar analysis of exactly the same data was performed in Hubin et al. [17], where the inference was based on the GLMM model fitted by the MJMCMC algorithm developed in Hubin and Storvik [16] also with the aim to capture spatial structure of the methylation probabilities. In Hubin et al. [17] the significant variables of GLMM's linear predictor were X_{CHG} , X_{CG} and X_{CODE} with less significant X_{Ma} and X_{Mg} . Hence variable selection in both of the approaches shows the importance of almost identical sets of covariates, though the models are rather different. This creates insights to be studied further by the biologists. The patterns of the recovered probabilities and most probable states (found by the Viterbi forward algorithm [6]) in our study and in Hubin et al. [17] are also quite similar and strongly agree for locations 1-7000 and 10250 - 10500. Also, both of the approaches are quite sceptical to the methylation status of locations 7100 - 7200, despite some observations with a high proportion of methylated reads in that region. Finally, the approaches disagree in locations 7250 - 7400, where the GLMM suggests an almost continuously methylated region, whilst the non-homogeneous HMM models multiple jumps between the methylated and non-methylated states. This region hence might need careful studies by the biologists in future. It would be also of interest to obtain additional covariates such as whether the corresponding nucleobase belongs to a particular part of the non-coding gene region like promoter, intron or transposon, and whether the nucleobase is within a CpG island.

4.2 S&P500 data example

Our second example is focused upon Amazon stock (AMZN) prediction with respect to logreturns of other $p = 30$ stocks from the S&P500 listing, which have the highest correlations to AMZN. The addressed data⁴ consists of 1258 observations of logreturns for 31 stocks based on the daily close price. The 30 predictors are: X_{ADBE} , X_{AOS} , X_{APH} , X_{ATVI} , X_{BLK} , X_{CA} , X_{CRM} , X_{FB} , X_{FISV} , X_{GOOGL} , X_{GOOG} , X_{ITW} , X_{MA} , X_{MHK} , X_{MMC} , X_{MSFT} , X_{MTD} , X_{NFLX} , X_{PCLN} , X_{PKI} , X_{ROP} , X_{SBUX} , X_{SNPS} , X_{SPGI} , X_{SYK} , X_{TEL} , X_V , X_{TMO} , X_{VRSN} , where the underscripts are representing official S&P500 acronyms of the corresponding stocks. In terms of physical time, the observations are ranged from 11.02.2013 to 07.02.2018. The focus of this example is in both inference and predictions, hence we divided the data into a training data set (before 01.01.2017) and a testing data set (after 01.01.2017). Here, we address the Gaussian NHMM with three classes, which we hypothesize to represent the "buy", "sell", "wait" states of the market:

$$Y_i | \mu_{s_i}, \sigma_{s_i}^2, s_i \sim N(\mu_{s_i}, \sigma_{s_i}^2), \quad \mu_{s_i} = \eta_{s_i}, s_i \in \{0, 1, 2\}.$$

The primary model selection criterion addressed in this example is AIC due to that we are mainly interested in predictions, whilst BIC is the secondary reported criterion. Other hyperparameters of the algorithm are the same as the ones used in the previous example. Here we have in total $4^{30} > 1.15 \times 10^{18}$ candidate solutions, hence 50 parallel threads of ASA-EM algorithm were run. The distribution of the best-found solutions across the threads is depicted in Figure 3.

⁴From www.kaggle.com/camnugent/sandp500.

We clearly see three major local extrema to which the threads are converging. This indicates that more epochs of each thread would be needed to resolve the issue of convergence to multiple local extrema. On the other hand, it would also require significantly more computational effort. In this sense (and since we are interested in one "best" solution) running more parallel threads (50 instead of 30) is a far more practical and pragmatic solution. The best found model is summarized in Table 3. We use the standard deviations

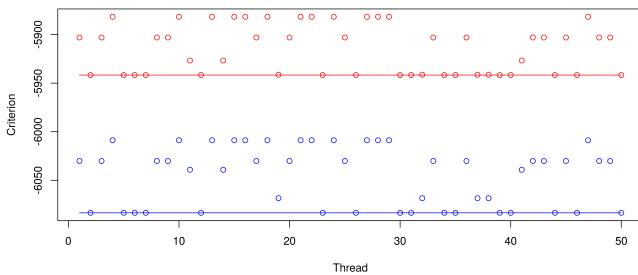


Figure 3: Values of BIC (red) and AIC (blue) across 50 parallel threads of ASA-EM algorithm for S&P500 data.

$\sigma_{s_i}, s_i \in \{0, 1, 2\}$ and means $\mu_{s_i}^0, s_i \in \{0, 1, 2\}$ at zero values of the predictors for distributions of the observations from the 3 different classes in order to resolve the label switching problem and see if the discovered classes are "buy", "wait", and "sell". These standard deviations and means at zero values of the covariates are $\{\sigma_0 = 0.008, \mu_0^0 = 0.000\}$, $\{\sigma_1 = 0.007, \mu_1^0 = -0.001\}$ and $\sigma_2 = 0.019, \mu_2^0 = 0.009$. This indicates that $s_i = 0$ might represent the "wait" or neutral state of the market, $s_i = 1$ - the "buy" state of the market and $s_i = 2$ - the "sell" state of the market. Let us illustrate the market structure, returns and hidden states graphically. In Figure 4 we clearly see that $s_i = 2$ states of the selected by ASA-EM NHHMM (SNHHMM) correspond to the pikes of the price and highly volatile logreturns and hence create the best moment for the investor to sell the stock. The $s_i = 1$ states correspond to the moments when the price starts to grow and volatility of the logreturns is low, which can be clearly interpreted as the best moment to buy the stock. The states $s_i = 0$ correspond to the "wait" situation on the market when the investor should hold either the bought asset or the funds. At the same time, this does not seem to be the case for the full NHHMM (FNHHMM), which has all of the covariates included. FNHHMM has less transparent states in terms of interpretations (hence we will not use states of FNHHMM as investment triggers). Note that the most probable states for both inference and predictions are recovered by the Viterbi algorithm [6, 27]. Consider that the investor during the training period keeps to the following strategy: She invests all her money into AMZN stock when the latent variable of SNHHMM takes the value of 1 and sells all her stocks when the latent variable takes the value of 2. Assume the investor had 1000\$ on 11.02.2013. Then on 01.01.2017 she would have 1915.137\$ in case she used the strategy purely based on SNHHMM, which shows that the latent states indeed correspond to the "buy", "sell", "wait" states.

The goal of this example was not only inference but also predictions, which means we have to check both how the predicted values of AMZN stock price behave and what would have happened

Table 3: The selected model for s&p500 study of AMZN prediction

	BIC	AIC	Init. State 1	Init. State 2	Init. State 3						
→ State 1 Inter.	-5534.046	-6159.914	0.000	1.000	0.000						
State 1	0.00	0.00	0.00	0.00	0.00						
State 2	-5.80	14.66	-13.09	-197.69	54.81						
State 3	-4.81	6.88	-745.00	-67.42	-32.22						
				-90.12	139.29						
				-5.61	63.93						
				-49.40	49.85						
→ State 2 Inter.	X _{BLK}	X _{FISV}	X _{GOOG}	X _{MA}	X _{MHK}						
State 1	0.00	0.00	0.00	0.00	0.00						
State 2	-8.75	-410.08	838.22	154.20	-675.17						
State 3	-14.84	951.36	866.36	-40.28	445.56						
				210.34	-589.45						
				-181.41	-128.87						
				-929.12	-754.37						
→ State 3 Inter.	X _{BLK}	X _{FISV}	X _{GOOG}	X _{MA}	X _{MHK}						
State 1	0.00	0.00	0.00	0.00	0.00						
State 2	-0.29	494.63	-800.14	1449.02	758.39						
State 3	-12.66	192.90	223.25	1463.84	323.61						
				930.92	-1671.98						
				-132.73	626.89						
				-277.53	-383.94						
Observ.	Inter.	X _{ADBE}	X _{APAH}	X _{ATVI}	X _{BLK}	X _{CA}	X _{FB}	X _{GOOGL}	X _{ITW}	X _{MA}	-
State1	0.00	0.10	0.03	0.05	0.00	-0.02	0.11	0.42	-0.07	0.09	-
State 2	0.00	-0.11	-0.72	0.08	0.17	-0.64	-0.04	-0.13	-0.52	0.89	-
State3	0.01	-1.32	-0.07	0.30	-1.56	0.38	-1.03	3.01	1.64	-0.85	-
Observ.	SD	X _{MMC}	X _{NFLX}	X _{PCLN}	X _{SBUX}	X _{SNPS}	X _{SYK}	X _{TEL}	X _{VRSN}	X _V	-
State 1	0.01	0.03	0.05	0.08	0.15	0.02	0.03	-0.02	0.11	-0.03	-
State 2	0.01	0.12	-0.23	0.37	-0.06	0.52	0.32	0.25	0.36	1.16	-
State 3	0.02	3.58	0.18	-0.02	-0.15	0.93	0.50	-1.77	-0.65	-0.91	-

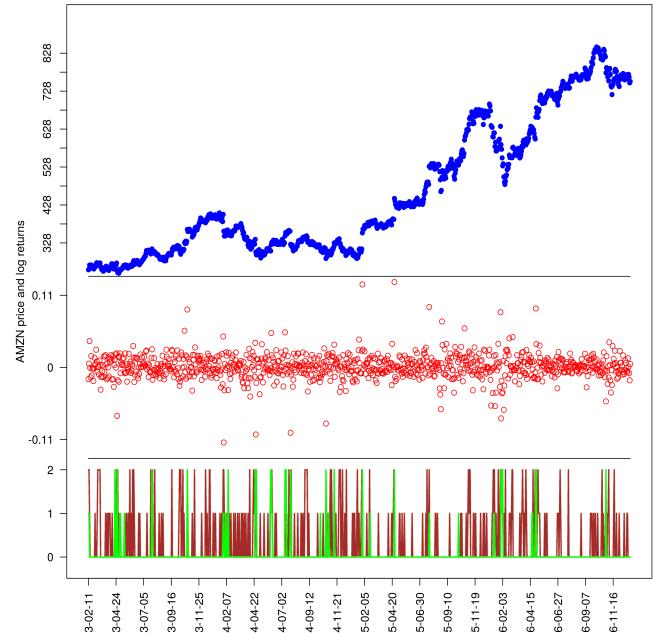


Figure 4: Amazon prices and logreturns for training data, where blue dots are stock prices, red dots - logreturns, green line - most probable latent states of the selected by ASA-EM NHHMM, brown line - most probable latent states of full NHHMM.

should the investor had used the suggested investment strategy on the test data. In terms of the second goal, if the investor kept to the strategy described in the previous paragraph during the test period (from 01.01.2017 to 07.02.2018), she would end up with 1198.636\$ at the end of the test period, provided that she started with 1000\$ on 01.01.2017. This clearly shows that the strategy works in the very long test period, even though both the model and its parameters were fixed on 01.01.2017 and were never updated during test time.

In terms of the other goal, we predict daily stock prices for the whole test period based on the selected by ASA-EM optimal model (SNHHMM) and full NHHMM (FNHHMM). Here, we also

Table 4: Test statistics results. For DM-test 0.1 level of significance corresponds to *, 0.05 - to **, 0.01 - to *.**

Criterion Data	NHHMM	FNHHMM	LASSO	RIDGE	SARIMAX	CAUSAL	PROPHET
RMSE AMZN	78.614	153.391	92.266	89.015*	87.439	159.687	181.475
PVDM AMZN	-	0.0000***	0.0109**	0.0597*	0.0496**	0.0000***	0.0000***
COINT AMZN	0.9999	1.0043	1.0035	1.0036	1.0048	1.0026	1.0039
MONEY AMZN	1198.63	-	-	-	-	-	-
RMSE AAPL	22.795	36.625	19.725	18.738	18.304	17.422	23.601
PVDM AMZN	-	0.0000***	1.0000	0.9998	1.0000	1.0000	0.1738
COINT AAPL	1.0019	1.0055	1.0054	1.0053	1.0050	1.0050	-
MONEY AAPL	1456.46	-	-	-	-	-	-
RMSE AAPL (HALF)	5.4772	33.468	14.333	13.474	13.195	13.124	20.211
PVDM AAPL (HALF)	-	0.0000***	0.0000***	0.0000***	0.0000***	0.0000***	0.0000***
COINT AAPL (HALF)	0.9766	1.0063	1.0105	1.0105	1.0104	1.0102	1.0094
MONEY AAPL (HALF)	1185.59	-	-	-	-	-	-

address several benchmark approaches, namely LASSO [26] and RIDGE regressions [14] combined with AIC used for the hyperparameter selection and SARIMAX model with automatic selection of hyperparameters based on AICc criterion [19]. Finally, as two industrial state-of-the-art models we addressed Bayesian structural time series model (CAUSAL) developed by Google [4] and PROPHET model developed by Facebook [25]⁵. The benchmark approaches were trained and tested using the same splits of the data. For all of the models, the predicted prices of AMZN shares were obtained using the following transformation of logreturns $\hat{p}_i = p_0 \exp(\sum_{j=1}^i \hat{Y}_j)$, $i \in \{1, \dots, n_{\text{test}}\}$, where p_0 is the actual stock price on the last training day. Then test RMSE was evaluated as:

$$\text{RMSE} = \sqrt{\frac{1}{n_{\text{test}}} \sum_{j=1}^n (p_j - \hat{p}_j)^2}.$$

Historically stock price prediction algorithms are compared in terms of RMSE. However, relying only on RMSE might be not enough to decide which of them yields the best predictive performance. This is due to the fact that the variance of squared errors of predictions is completely ignored in such a case. Hence rigorous statistical testing should be applied. Simple t-test or z-test in this context can not be trusted, since forecast errors may be serially correlated, and a robust Diebold-Mariano (DM) test [7], designed to compare forecasts, should be applied. We hence are additionally reporting p-values (PVDM) of one-sided DM-test [8] of that our method is superior to the corresponding competitors. We also tested the absence of auto-correlations of the residuals of the predicted prices using the following cointegration test statistics:

$$\text{COINT} = \max\{|\hat{\rho} + 1.96\hat{\sigma}_{\rho}|, |\hat{\rho} - 1.96\hat{\sigma}_{\rho}|\}.$$

Here $\hat{\rho}$ is estimated from the following equation $\varepsilon_t = \rho\varepsilon_{t-1} + \eta_t$, and ε_t is the noise equal to $\hat{p}_i - p_i$. We want COINT as small as possible and accept the series to be cointegrated if it is below 1. The results for the set of compared approaches are summarized in Table 4. They clearly show that our method (SNHHMM) yields the smallest test RMSE error and its advantage of predictive performance is statistically significant according to DM-test on $\alpha = 0.01$ level for FNHHMM, CAUSAL and PROPHET, on $\alpha = 0.05$ level for LASSO and SARIMAX, and on $\alpha = 0.1$ level for RIDGE. Moreover, SNHHMM is the only method with stationary and cointegrated residuals for AMZN data. We repeated the experiment for Apple data (AAPL) as observations with 30 explanatory stocks (with the

⁵No explanatory variables are addressed by PROPHET due to its model assumptions.

highest correlations to AAPL) and also report the results in Table 4. For AAPL, RMSE of SNHHMM was slightly worse than for most of the competitors except for FNHHMM and PROPHET (also, no significant evidence of that our method is better in accordance with PVDM, except for FNHHMM on 0.01 level of confidence), but cointegration results were still the best. Graphically the results for AAPL are summarized in Figure 6. There one can see that the predictions for SNHHMM worsen in the second part of the test period, where some sort of a causal impact could have happened. This means that the model should be calibrated with some frequency. In particular, the results for all of the methods could be improved should we retrain the models after each test day by adding the corresponding logreturns to improve the following day's predictions, however, we leave this option outside of the scope of the paper. Note, however, that if one only used the first half of the test period for AAPL, both the RMSE (with a significant difference of predictive performance on $\alpha = 0.001$ level according to DM-test), and COINT of SNHHMM would be by far the best on the given set of baselines (see Table 4). Finally, using 1000\$ as the starting capital, the investor would gain 1456.457\$ for AAPL using our strategy based on the hidden states during the whole test period and 1185.59\$ during the first half of the period.

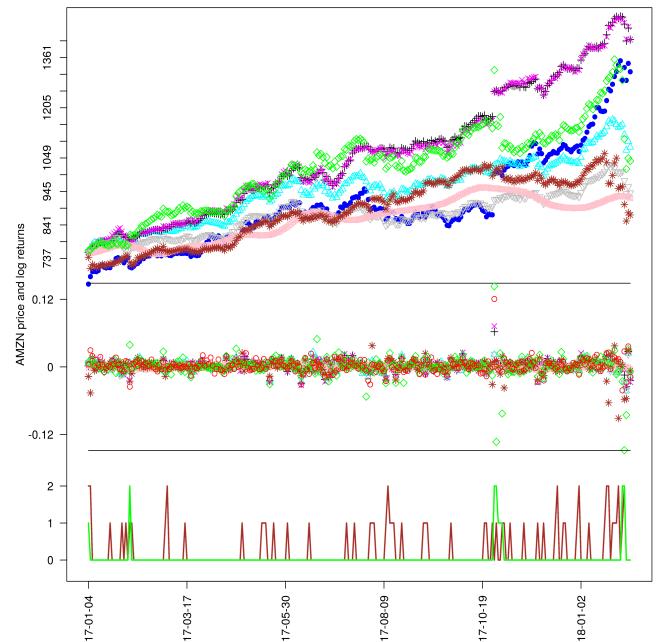


Figure 5: Amazon prices, log-returns and most probable latent states as in caption to Figure 4. Green signs correspond to the selected by ASA-EM NHHHM predictions, brown - to full NHHMM, black - to LASSO, purple - to RIDGE, light blue - to SARIMAX, grey - to CAUSAL, pink - to PROPHET.

5 DISCUSSION

In this article, we have suggested a novel adaptive simulated annealing expectation maximization algorithm for simultaneous exploration of model and parameter space of non-homogeneous hidden

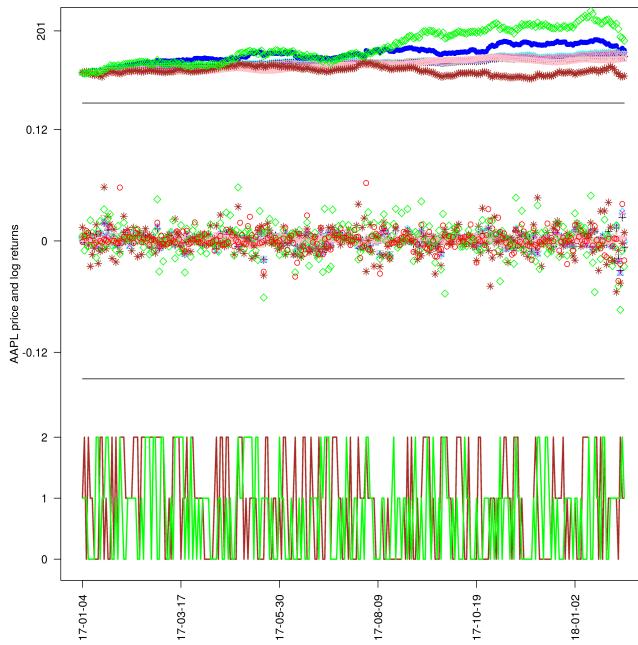


Figure 6: Apple results. See details in the caption to Figure 5

Markov models. Also, parallelization strategies were suggested for ASA-EM. The algorithm has shown to efficiently converge to good solutions within finite time. The solutions both yield high sparsification rates and interpretable results in terms of the meaning of the latent states and inference on the observations. In two applications we found the latent states to be associated with important biological and financial phenomena, namely methylation status of the nucleobases along the genome and extremal/non-extremal points of the stock price, which can be interpreted as "buy"- "wait"- "sell" states. Finally, the resulting model yields good predictive properties and often outperforms the baseline approaches. In future, it would be of interest to allow for sub-sampling from the data in the EM part of the algorithm to deal with large high dimensional data samples.

ACKNOWLEDGMENTS

I would like to acknowledge Professor G. Storvik (UiO) for the discussion of the paper in the first year of my PhD studies. The work resulted in a talk at Klækken workshop (bit.ly/2MaN3l7) back in 2015, but then was put aside for a while. The project got a second birth in 2019 inspired by the work on Exabel (exabel.com) project at NR in collaboration with Dr K. Aas (NR) and Dr P.C. Moan (Exabel), which also led to the experimental design of the S&P500 example. The epigenetic example was inspired by the discussions with Professors P. Grini, O.C. Lingærde, G. Storvik, K. S. Jakobsen and Dr M. Butenko (all from UiO). Finally, I would like to thank Dr. P. Lison (NR) for thoroughly proofreading the paper.

REFERENCES

- [1] Stephen Adams and Peter A Beling. 2019. A survey of feature selection methods for Gaussian mixture models and hidden Markov models. *Artificial Intelligence Review* 52, 3 (2019), 1739–1779.
- [2] Stephen Adams, Peter A Beling, and Randy Cogill. 2016. Feature selection for hidden Markov models and hidden semi-Markov models. *IEEE Access* 4 (2016), 1642–1657.
- [3] Richard J Boys and Daniel A Henderson. 2001. A comparison of reversible jump MCMC algorithms for DNA sequence segmentation using hidden Markov models. *Comp. Sci. and Statist* 33 (2001), 35–49.
- [4] Kay H Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, Steven L Scott, et al. 2015. Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics* 9, 1 (2015), 247–274.
- [5] Gerda Claeskens and Nils Lid Hjort. 2008. *Model selection and model averaging*. Technical Report. Cambridge University Press.
- [6] Michael Collins. 2002. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 1–8.
- [7] Francis X Diebold. 2015. Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests. *Journal of Business & Economic Statistics* 33, 1 (2015), 1–1.
- [8] Francis X Diebold and Robert S Mariano. 2002. Comparing predictive accuracy. *Journal of Business & economic statistics* 20, 1 (2002), 134–144.
- [9] Robert J Elliott and Tak Kuen Siu. 2012. An HMM approach for optimal investment of an insurer. *International Journal of Robust and Nonlinear Control* 22, 7 (2012), 778–807.
- [10] Christina Erlwein, Fred Espen Benth, and Rogemar Mamon. 2010. HMM filtering and parameter estimation of an electricity spot price model. *Energy Economics* 32, 5 (2010), 1034–1043.
- [11] Mark JF Gales. 1998. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer speech & language* 12, 2 (1998), 75–98.
- [12] Charles J Geyer and Elizabeth A Thompson. 1995. Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Amer. Statist. Assoc.* 90, 431 (1995), 909–920.
- [13] Peter J Green and David I Hastie. 2009. Reversible jump MCMC. *Genetics* 155, 3 (2009), 1391–1403.
- [14] Arthur E Hoerl and Robert W Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 1 (1970), 55–67.
- [15] Tracy Holsclaw, Arthur M Greene, Andrew W Robertson, Padhraic Smyth, et al. 2017. Bayesian nonhomogeneous Markov models via Pólya-Gamma data augmentation with applications to rainfall modeling. *The Annals of Applied Statistics* 11, 1 (2017), 393–426.
- [16] Aliaksandr Hubin and Geir Storvik. 2018. Mode jumping MCMC for Bayesian variable selection in GLMM. *Computational Statistics & Data Analysis* 127 (2018), 281–297.
- [17] A. Hubin, G. Storvik, P. Grini, and M. Butenko. 2019. Bayesian binomial regression model with a latent Gaussian field for analysis of epigenetic data. *COMPUTER DATA ANALYSIS AND MODELING. Stochastics & Data Science. Proceedings of the XII International Conference.* (2019).
- [18] James P Hughes, Peter Guttorp, and Stephen P Charles. 1999. A non-homogeneous hidden Markov model for precipitation occurrence. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 48, 1 (1999), 15–30.
- [19] Rob J Hyndman, Yeasmin Khandakar, et al. 2007. *Automatic time series for forecasting: the forecast package for R*. Number 6/07. Monash University, Department of Econometrics and Business Statistics.
- [20] Anders Krogh, Michael Brown, I Saira Mian, Kimmen Sjölander, and David Haussler. 1994. Hidden Markov models in computational biology: Applications to protein modeling. *Journal of molecular biology* 235, 5 (1994), 1501–1531.
- [21] Antonello Maruotti and Roberto Rocci. 2012. A mixed non-homogeneous hidden Markov model for categorical data, with application to alcohol consumption. *Statistics in Medicine* 31, 9 (2012), 871–886.
- [22] Loukia Meligkotsidou and Petros Dellaportas. 2011. Forecasting with non-homogeneous hidden Markov models. *Statistics and Computing* 21, 3 (2011), 439–449.
- [23] Roberta Paroli and Luigi Spezia. 2007. Bayesian variable selection in Markov mixture models. *Communications in Statistics-Simulation and Computation* 37, 1 (2007), 25–47.
- [24] Roberta Paroli and Luigi Spezia. 2008. Bayesian inference in non-homogeneous Markov mixtures of periodic autoregressions with state-dependent exogenous variables. *Computational Statistics & Data Analysis* 52, 5 (2008), 2311–2330.
- [25] Sean J Taylor and Benjamin Letham. 2018. Forecasting at scale. *The American Statistician* 72, 1 (2018), 37–45.
- [26] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 1 (1996), 267–288.
- [27] Ingmar Visser, Maarten Speekenbrink, et al. 2010. depmixS4: an R package for hidden Markov models. *Journal of Statistical Software* 36, 7 (2010), 1–21.
- [28] Thomas D Wickens. 1982. *Models for behavior: Stochastic processes in psychology*. WH Freeman San Francisco.
- [29] Hao Zhu, Zhongshi He, and Henry Leung. 2012. Simultaneous feature and model selection for continuous hidden Markov models. *IEEE Signal Processing Letters* 19, 5 (2012), 279–282.