Variational inference for coupled hidden markov models applied to the joint detection of copy number variations

Xiaoqiang Wang, Emilie Lebarbier, Julie Aubert, Stephane Robin

Variational inference for coupled Hidden Markov Models applied to the joint detection of copy number variations

Xiaoqiang Wang[1,2*] xiaoqiang.wang@sdu.edu.cn, Emilie Lebarbier[2], Julie Aubert[2], and Stéphane Robin[2]
[1]School of Mathematics and Statistics, Shandong University, Weihai, China
[2]UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France

October 8, 2018

Abstract

Hidden Markov models provide a natural statistical framework for the detection of the copy number variations (CNV) in genomics. In this paper, we consider a Hidden Markov Model involving several correlated hidden processes at the same time. When dealing with a large number of series, maximum likelihood inference (performed classically using the EM algorithm) becomes intractable. We thus propose an approximate inference algorithm based on a variational approach (VEM). A simulation study is performed to assess the performance of the proposed method and an application to the detection of structural variations in plant genomes is presented.

Keywords. Coupled Hidden Markov Models; Variational approximation; Copy Number Variation.

1 Introduction

CNV detection Copy Number Variation (CNV) refers to the DNA sequence variation which increasing or decreasing the genomic segment of at least 50bp (Zarrei *et al.* (2015)). As a typical form of structural variation, CNV generally consists of duplication, insertion or deletion events. Since first studies in 2003-2004 (Lucito *et al.* (2003),**?**,**?**), CNV have been prevalently discovered in the human genome (MacDonald *et al.* (2014)). While most of the CNV analyses arise in human health, some have been proved to be associated with, or directly cause diseases or clinical phenotype variations (Weischenfeldt *et al.* (2013),**?**,**?**). Due to their potential functional effect, thus possibly altering phenotypes/traits of interest, CNVs have also been intensively identified in many animal and plant species in the past few years. For example, some studies have also investigated the effect of CNVs on agronomical traits, such as the milk production traits (Xu *et al.* (2014)), the growth traits in cattle (Zhou *et al.* (2016)), the flowering time trait in maize (Lu *et al.* (2015)). All of these CNV analysis in animal and plant species glean some preliminary insights into factors linked to the genomic selection.

Method for CNV detection Alkan *et al.* (2011) review extensively the experimental platforms applied to CNV discovery and genotyping, which include two hybridization-based microarray technologies: array comparative genomic hybridization (CGH), Single Nucleotide Polymorphism (SNP) microarray, and sequencing-based next-generation sequencing (NGS) technologies. Depending on the data architecture resulted from above different platforms, a number of statistical methodologies and software tools have been developed to detect CNV. Moreover, their

performance to detect CNV are usually compared in the literature, for instance, Lai *et al.* (2005) for CGH, Dellinger *et al.* (2010) and Winchester *et al.* (2009) for SNP array, Pinto *et al.* (2011) for cross-platform between CGH and SNP, Zhao *et al.* (2013), Magi *et al.* (2012) and Ji and Chen (2016) for NGS. Among these methods, hidden Markov models-based (HMM) and segmentation-based algorithm are two main types of approaches. Particularly, several studies were investigated to analyse simultaneously multiple individuals in CNV discovery. For instance, Wang *et al.* (2008) and Liu *et al.* (2016) propose some HMM-like methods which making use of family informations from parent-offspring trios. Picard *et al.* (2011), Tai *et al.* (2010), Zhang *et al.* (2010) and Hu *et al.* (2016) propose some segmentation-like methods which focusing on detecting common or rare CNV regions across individuals. These analyses applying on multiple individuals attract our great attention, because they could obviously improve the accuracy of CNV estimation in comparison with analyzing typically single individual. We believe that the positive performance might rely on the fact that CNV is inheritable (Sun *et al.* (2009)). Consequently, altered in the same loci across the individuals with common phylogenetic past, such as between offspring and either of parents in trios cas. These facts imply that the relatedness between individuals is a useful factor in CNV detection.

Measure of relatedness Relatedness is a fundamental concept in genetic association studies, however there does not exist a common way to define them. Astle and Balding (2009) present that kinship is a central concept to measure pairwise genetic relatedness among individuals. The kinship coefficient s_{ij} between two individuals i and j is the probability that an allele selected randomly from i and an allele selected randomly from the same autosomal locus of j are identical by descent (IBD). Nowadays, SNP marker-based relative kinship estimates have proven useful and accurate for quantitative inheritance studies in different populations. This genetic relatedness matrix is called also genetic similarity matrix by Speed and Balding (2015), in particular, the authors summarize the SNP-based measure accounting for minor allele fraction (MAF) of the SNP by a series of formulates as:

$$s_{ij}(\alpha) = \frac{1}{L}\sum_{t=1}^{L}\frac{(Z_{i,t}-2p_t)(Z_{j,t}-2p_t)}{[2p_t(1-p_t)]^{\alpha}},$$

where $Z_{i,t}$ is the minor allele count (0, 1 or 2) of individual i, p_t is the population MAF at the t^{th} SNP and α takes some integer values. The performance for the case of $\alpha = -1, 0, 1, 2$ is compared in Speed and Balding (2015).

Coupled HMM and intractable likelihood As an extension of HMM, coupled HMM (CHMM) model a system of multiple interacting processes, they take into consideration the interactions between variables in the latent space rather than observation space (Rezek *et al.* (2002)). Intuitively, CHMM has the ability to capture the relatedness between individuals in CNV discovery. In fact, CHMM have been applied in several fields such as speech recognition (Nock and Ostendorf (2003)), disease studies (Sherlock *et al.* (2013)), health informatics (Ghahjaverestan *et al.* (2016)), electroencephalogram analysis (Zhong and Ghosh (2002)) and bioinformatics (Choi *et al.* (2013)).

CHMM is an incomplete data model for which the EM algorithm (Dempster *et al.* (1977)) is the most popular algorithm to maximize the likelihood. However, the exact inference in CHMM raises some computational issues. Indeed, when considering Q status, I individuals and T observations for each individuals, CHMM is a HMM, the state space of which consists in all possible combinations of individual status. So the number of hidden states is $K := Q^I$ and the complexity of each E step of a regular EM algorithm is $K^2T = Q^{2I}T$, which becomes intractable when K (that is, either Q or I) becomes large. Many efforts have been made to manage this

complexity, mostly by modeling the $K \times K$ transition matrix in a parsimonious way. (Saul and Jordan (1999)) use a mixture form

$$\mathbb{P}(S_{j,t}|S_{1:I,t-1}) = \sum_{i=1}^{I} \omega_{ij} \mathbb{P}(S_{j,t}|S_{i,t-1}),$$

where $S_{j,t}$ represents the status of individual j at position t and ω_{ij} can be viewed as mixing weights, or strength of effect of chain i on chain j. This strategy reduces the number of transition parameters from one Q^{2I} to $I(I+1)Q/2$.

Variational approximation for CHMM In this paper, we try to keep the original form of CHMM even when K is large and we use variational approximation to make the E step of the EM algorithm computationally tractable. The resulting variational EM (VEM) aims at maximizing a lower bound of the log-likelihood. VEM was first explicitly introduced in machine learning such as Saul *et al.* (1996), now this approach has been routinely applied and generalized in many different ways. Jaakkola (2000) gives a brief introduction and Wainwright and Jordan (2008) provide a very complete overview.

The variational approximation consists in seeking for some tractable distributions $\widetilde{\mathbb{P}}(S)$ to replace $\mathbb{P}(S|X)$, the conditional distribution of hidden status S given the observed data X, in E step. Such an approximation relies on two ingredients. We first need to choose a divergence measure between the true conditional distribution $\mathbb{P}(S|X)$ and the approximated one $\widetilde{\mathbb{P}}(S)$ and the most popular is the Küllback-Leibler divergence. Then we need to define the class of distributions $\mathcal{Q}$ within which $\mathbb{P}(S|X)$ is searched for. Obviously, this second choice is critical as $\mathcal{Q}$ has to be as large as possible to make the approximation better but $\mathcal{Q}$ must also contain only distributions that are computationally manageable. In the context of factorial HMMs, Ghahramani and Jordan (1997) introduced the distribution family of independent heterogeneous Markov chains to approximate the original distribution. Using this distribution class yields, during the E step, at preserving the within individual dependencies while neglecting the between individual dependencies. In this paper, we will adopt the same approximation type.

Joint CNV detection Although several methods considering multiple individuals have been derived in the literature, their relatedness/dependency relationship across individuals is fuzzy because of lack of informations on the status transition structure. Under the framework of CNV discovery, the genetic kinship matrix can be nowadays computed from the SNP genotyping data, therefore, we propose a novel CHMM accounting for this genetic relatedness between individuals. Our model is inspired by the fact that the closer the genetic relationship between any two individuals, the more likely their hidden status.

Paper outline In the following section we define the probabilistic model for CHMM which accounting for the genetic kinship matrix and in Section 3 we present algorithms for exact and variational inference and learning in CHMM. In Section 4 we describe simulation results comparing exact and approximate algorithms for learning on the basis of time complexity and model quality, next, confirming the necessity of taking kinship relationship into account in model. We also apply CHMM to a time series dataset consisting of 336 maize lines. We discuss several generalizations of the probabilistic model in Section 5.

2 Model

We consider a set of I individuals ($i = 1, \ldots, I$). For each individual, we observe a series of (microarray) measurements $X_i = (X_{i,t})$, that is supposed to vary according to the status (copy

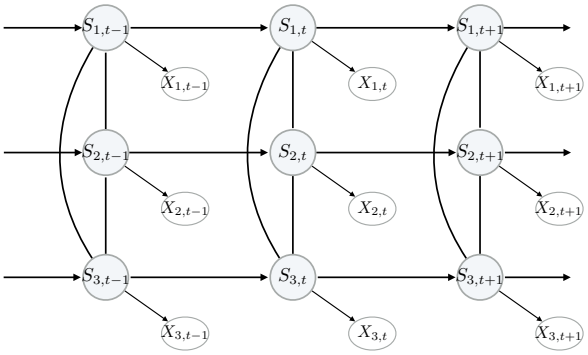

Figure 1: A mixture of directed and undirected graphical model. The directed edges represent the dependency relationship within the individual. The undirected edges represent the correlations among individuals.

number) of the individual at 'time' $t = 1, \dots, T$ (position along the genome). We denote $(S_{i,t})_t$ the hidden process for individual i, where $S_{i,t}$ can take Q different values (e.g. $Q = 3$, $-1 =$ 'deletion', $0 =$ 'normal', $1 =$ 'amplification'). In this setting, the state space of the joint hidden process $(S_t)_t$, with $S_t = (S_{1,t}, \dots S_{I,t})$, consists in $K := Q^I$ possible values.

2.1 Emission distribution

We assume that the observed data are all conditionally independent given the hidden process S, with respective conditional distribution:

$$(X_{i,t}|S_{i,t} = q) \sim \mathcal{N}(\mu_q, \sigma_q^2),$$

where μ_q represents the mean value of state q. In the following, the means will be gathered in the vector $\mu = (\mu_q)_q$. We further denote $S_{i,t}^q = \mathbf{1}_{\{S_{i,t}=q\}}$ and $\phi_q(X_{i,t})$ the conditional probability density function of $X_{i,t}$ given the value q of state $S_{i,t}$.

It is worth noting that dependent process was already considered in Picard *et al.* (2011) in the same segmentation context. In this paper, the dependency was encoded in the joint distribution of the observed signals at each position $(X_{1,t}, \dots X_{i,t}, \dots, X_{I,t})$, making the normality assumption critical to achieve the inference. In our model, the dependency is encoded in the hidden layer so the emission distributions can be chosen arbitrarily. The Gaussian distribution is only chosen here to fit with microarray data. The same model could be easily adapted to sequencing data using Poisson or negative binomial distribution, or to any other type of signal.

2.2 Hidden Markov chain

We also assume that joint hidden process S is distributed according to a Markov chain. One purpose of our work is to introduce a hidden dependency structure as in Figure 1. More specifically, the set of status of all individuals $(S_{i,t})_i$ is a Markov chain and the edges between the status of all individuals at a given time t allows to account for their (phylogenetic) proximity. Note that these edges introduce a coupling between the individual's hidden processes.

The variation of the hidden status along time as well as their correlation from one individual to another is encoded in the $K \times K$ transition matrix P of the joint hidden process (S_t). We consider that the transition probabilities result from the product of two terms: (a) one accounting for the transitions within each individual and (b) one accounting for the similarities

between individuals (both supposed to be constant along time):

$$\mathbb{P}(S_t = \ell | S_{t-1} = k) =: P_{k\ell} \propto \left(\prod_i \pi_{k_i,\ell_i} \right) W_\ell, \tag{1}$$

where

(*a*) π is a $Q \times Q$ transition matrix (each row sums to one) and k_i (resp. ℓ_i) stands for the hidden state of individual i when the joint hidden state is k (resp. ℓ);

(*b*) the dependency relationship among the individuals is encoded in the coefficients

$$W_\ell = \prod_{i,j \neq i} \omega^{s_{ij} \mathbf{1}_{\{\ell_j \neq \ell_i\}}}, \tag{2}$$

where $\omega \leq 1$ and s_{ij} denotes the similarity (e.g. phylogenetic proximity) between individuals i and j. Note that considering $\omega = 1$ means considering the independent case.

In this model, π rules the within individual transitions, while W introduces dependency between the individuals.

We further assume that the initial state $S_1 = (S_{i,1})_i$ has distribution

$$\mathbb{P}(S_1 = \ell) \propto \left(\prod_i m_{\ell_i} \right) W_\ell \tag{3}$$

where (m_q) is a distribution of the states $1 \leq q \leq Q$.

Because the initial distribution (3) and the transitions (1) are not normalized, the distribution of the hidden process S writes

$$\mathbb{P}(S) = \frac{1}{Z} \prod_\ell \left[\left(\prod_i m_{\ell_i} \right) W_\ell \right]^{S_1^\ell} \prod_{\substack{t \geq 2 \\ k,\ell}} \left[\left(\prod_i \pi_{k_i,\ell_i} \right) W_\ell \right]^{S_{t-1}^k S_t^\ell}$$

where Z stands for the normalizing constant. [1]

3 Inference

This section introduces the variational inference algorithm we propose.

[1] It follows that

$$\begin{aligned} \log \mathbb{P}(X, S) \quad = \quad & \sum_{i,q} S_{i,1}^q \log m_q + \sum_{i,t \geq 2,q,r} S_{i,t-1}^q S_{i,t}^r \log \pi_{q,r} \\ & + \sum_{i,t,r} S_{i,t}^r \sum_{j \neq i} (1 - S_{j,t}^r) s_{ij} \log \omega \\ & + \sum_{i,t,r} S_{i,t}^r \log \phi_r(X_{i,t}) - \log Z \end{aligned}$$

3.1 EM algorithm

The EM algorithm aims at estimating all the parameters, denoted θ, for a fixed number of states Q and a fixed parameter ω. Indeed, ω could be estimated along with all other parameters, but this introduces some instability in the behavior of the algorithm. A heuristic to estimate ω outside the EM algorithm is given in Section 4.2.

The EM algorithm alternates two steps:

- **E-step:** evaluate the moments of the conditional distribution of the hidden variables $\mathbb{P}(S|X)$ for a current value of the parameter θ, say θ^h;
- **M-step:** update the parameter θ by maximizing the conditional expectation of the complete log-likelihood with respect to θ

$$\theta^{h+1} = \arg\max_{\theta} \mathbb{E}_{\theta^h} \left[\log \mathbb{P}(X, S; \theta)|X\right].$$

In the case of HMM, the E-step can be achieved via a forward-backward recursion. This step is the critical point for the inference of the model described in Section 2. Indeed, we can face three situations:

1. If we do not take into account the phylogenetic dependency (i.e. if we set $\omega = 1$ in (2)), then the individuals' hidden processes are independent, so the E-step can be achieved using the standard forward-backward recursion for each individual.
2. If we take into account the phylogenetic proximity but if both the number of individuals and the number of states are small, namely if $K = Q^I$ remains bellow few tens, the global model can be considered as one single HMM and the E-step can be achieved using the forward-backward recursion with complexity $\mathcal{O}(TK^2)$.
3. If we take into account the phylogenetic proximity and if K is too large, the complexity of the E-step becomes prohibitive, so some alternative has to be proposed.

In the first two cases a regular EM can be used. Our work focuses on the third case.

3.2 Variational EM algorithm

We follow the line of Jaakkola (2000) and Wainwright and Jordan (2008) to derive our variational approximation. We first observe that, for any distribution $\widetilde{\mathbb{P}}$, we have

$$\begin{aligned}\log \mathbb{P}(X) &\geq \log \mathbb{P}(X) - KL\left[\widetilde{\mathbb{P}}(S)||\mathbb{P}(S|X)\right] \\ &= \widetilde{\mathbb{E}} \log \mathbb{P}(X, S) - \widetilde{\mathbb{E}} \log \widetilde{\mathbb{P}}(S) =: \mathcal{J}(X, \theta, \widetilde{\mathbb{P}}),\end{aligned} \tag{4}$$

where $\widetilde{\mathbb{E}} = \mathbb{E}_{\widetilde{\mathbb{P}}}$ and KL stands for the Küllback-Leibler divergence. The inference strategy then consists in maximizing the lower bound $\mathcal{J}(X, \theta, \widetilde{\mathbb{P}})$ with respect to the parameter θ. As EM algorithm, VEM alternates two steps:

- **VE-step:** update the approximate conditional distribution $\widetilde{\mathbb{P}}$, given the current value of the parameter θ^h, as

$$\widetilde{\mathbb{P}}^{h+1} = \arg\max_{\widetilde{\mathbb{P}}} \mathcal{J}(X, \theta^h, \widetilde{\mathbb{P}}) = \arg\min_{\widetilde{\mathbb{P}}} KL\left[\widetilde{\mathbb{P}}(S)||\mathbb{P}(S|X; \theta^h)\right].$$

- **M-step:** update the parameter estimates as

$$\theta^{h+1} = \arg\max_{\theta} \mathcal{J}(X, \theta, \widetilde{\mathbb{P}}^{h+1}).$$

The quality of this approximation mostly relies on the class of approximating distributions within which $\widetilde{\mathbb{P}}$ is searched for. We adopt here the general approach proposed by Saul and Jordan (1995) and adapted to the coupled HMM by Ghahramani and Jordan (1997), forcing $\widetilde{\mathbb{P}}$ to be a product of independent Markov chains, that is

$$\widetilde{\mathbb{P}}(S) = \prod_i \widetilde{\mathbb{P}}(S_i) \text{ where } \widetilde{\mathbb{P}}(S_i) = \prod_i \widetilde{\mathbb{P}}(S_{i,1}) \prod_{t\geq 2} \widetilde{\mathbb{P}}(S_{i,t}|S_{i,t-1}).$$

We use the same parametrization setting

$$\widetilde{\mathbb{P}}(S_i) = \frac{1}{\widetilde{Z}_i} \left(\prod_q (m_q h_{i1}^q)^{S_{i,1}^q} \right) \prod_{t\geq 2} \left(\prod_{q,r} (\pi_{q,r} h_{it}^r)^{S_{i,t-1}^q S_{i,t}^r} \right)$$

where $\widetilde{Z}_i$ stands for the normalizing constant ensuring that $\widetilde{\mathbb{P}}(S_i)$ sums to one. The variational parameters h_{it}^r can be viewed as correction terms with respect to a Markov chain with parameters (m, π).

We denote $\tau_{it}^r = \widetilde{\mathbb{E}}(S_{i,t}^r)$, $\Delta_{it}^{qr} = \widetilde{\mathbb{E}}(S_{i,t-1}^q S_{i,t}^r)$ and

$$\log \Omega_{it}^r = \left[\sum_{j\neq i} s_{ij}(1 - \tau_{jt}^r) \right] \log \omega.$$

Using the factorization properties of the approximating distribution $\widetilde{\mathbb{P}}$, the lower bound $\mathcal{J}(X, \theta, \widetilde{\mathbb{P}}^h)$ given in (4) becomes:

$$\begin{aligned}
\mathcal{J}(X, \theta, \widetilde{\mathbb{P}}^h) &= \sum_{i,r} \tau_{i1}^r [\log m_r + \log \phi_r(X_{i,1}) - \log(m_r h_{i1}^r)] \\
&\quad + \sum_{i,t\geq 2,q,r} \Delta_{it}^{qr} [\log \pi_{q,r} - \log(\pi_{q,r} h_{it}^r)] \\
&\quad + \sum_{i,t\geq 2,r} \tau_{it}^r [\log \Omega_{it}^r + \log \phi_r(X_{i,t})] \\
&\quad - \log Z + \sum_i \log \widetilde{Z}_i \\
&= \sum_{i,t,r} \tau_{it}^r [\log \phi_r(X_{it}) + \log \Omega_{it}^r - \log h_{it}^r] \\
&\quad - \log Z + \sum_i \log \widetilde{Z}_i,
\end{aligned}$$

since $\widetilde{\mathbb{E}}(S_{it}^r S_{jt}^r) = \tau_{it}^r \tau_{jt}^r$ for all $i \neq j$ and since $\sum_q \Delta_{it}^{qr} = \tau_{it}^r$.

The VE-step consists in both finding the optimal value for the variational parameters (h_{it}^r) and computing the approximate conditional moments τ_{it}^r and Δ_{it}^{qr}. Following Ghahramani and Jordan (1997), Appendix D, we get

$$\begin{aligned}
\frac{\partial \mathcal{J}(X, \theta, \widetilde{\mathbb{P}}^h)}{\partial \log h_{it}^r} &= \Big[\log \phi_r(X_{i,t}) + \log \Omega_{it}^r - \log h_{it}^r \Big] \frac{\partial \tau_{it}^r}{\partial \log h_{it}^r} \\
&\quad - \tau_{it}^r + \tau_{it}^r,
\end{aligned}$$

because Z does not depend on h_{it}^r and $\partial \log \widetilde{Z}_i / \partial \log h_{it}^r = \tau_{it}^r$. This derivative is zero for

$$h_{it}^r = \Omega_{it}^r \ \phi_r(X_{i,t}). \tag{5}$$

The conditional moments, which depend on the normalizing constants $\widetilde{Z}_i$, are then computed using an independent forward-backward recursion for each individual i:

- Forward recursion: set $F_{i,1}^q \propto m_q h_{i1}^q$ and, for $t \geq 2$, compute

$$F_{i,t}^r \propto \sum_q F_{i,t-1}^q \pi_{q,r} h_{it}^r;$$

- Backward recursion: $\tau_{iT}^r = F_{i,T}^r$ holds and, for $1 \leq t \leq T-1$, compute

$$G_{i,t+1}^r = \sum_q F_{i,t}^q \pi_{q,r}, \quad \Delta_{it}^{qr} = \pi_{q,r} \frac{\tau_{i,t+1}^r}{G_{i,t+1}^r} F_{i,t}^q, \quad \tau_{it}^q = \sum_r \Delta_{it}^{qr}.$$

Model selection The number of states Q can be fixed according to the considered problem as in our application study (Section 5). However, it can be difficult to choose in advance. We thus propose a criterion relying on the popular BIC criterion (Schwarz (1978)), which consists at subtracting from the maximized likelihood $\log \widehat{\mathbb{P}}(X)$ the penalty term $0.5D\log(N)$ where N is the number of observations and D the number of free parameters. In our case, we have $N = IT$ and $D = 1 + Q(Q-1)$, so the penalty writes

$$\text{pen}_{BIC} = [1 + Q(Q-1)]\log(IT)/2.$$

Still, the likelihood of the observed data can not be computed in practice so we simply replace it by its variational lower bound and choose Q as

$$\widehat{Q} = \arg\max_Q \mathcal{J}_Q(X, \widehat{\theta}, \widetilde{\mathbb{P}}) - [1 + Q(Q-1)]\log(IT)/2,$$

where $\mathcal{J}_Q(X, \widehat{\theta}, \widetilde{\mathbb{P}})$ is the maximized lower bound of the Q-state model (see e.g. Daudin *et al.* (2008)).

3.3 Classification

The aim of CNV analysis is to associate each genetic locus with a status, e.g. 'deleted', 'normal' or 'amplified'. So, the inference procedure requires a classification step that returns a predicted value for each $S_{i,t}$ to be completed. For a given locus t in a given individual i, the VEM algorithm provides us with $\tau_{it}^r = \widetilde{\mathbb{P}}\{S_{it} = r\}$ that is the variational approximate of $\mathbb{P}\{S_{it} = r|X\}$. A local classification rule then consists in simply taking the most probable status according to the τ_{it}^r, that is to take

$$\widetilde{S}_{it} = \arg\max_r \tau_{it}^r.$$

Still, in many HMM applications, one is often interested in classifying all loci at once, which means retrieving the most probable hidden path $\widehat{S} = \arg\max_S \mathbb{P}(S|X)$. Because $\mathbb{P}(S|X)$ is intractable, we consider its variational approximation $\widetilde{S} = \arg\max_S \widetilde{\mathbb{P}}(S)$, which can be obtained via the following Viterbi recursion. Let us denote

$$\alpha_{i,t}^r = \max_{r_1,\ldots,r_{t-1}} \widetilde{\mathbb{P}}(S_{i,1} = r_1, \ldots, S_{i,t-1} = r_{t-1}, S_{i,t} = r),$$
$$p_{itqr} = \widetilde{\mathbb{P}}(S_{i,t} = r | S_{i,t-1} = q) \propto \pi_{q,r} h_{it}^r.$$

At the first position of each profile, we have that $\alpha_{i,1}^r = \tau_{i1}^r$. Then, we apply the classical recursion

$$\alpha_{i,t}^r = \max_q \alpha_{i,t-1}^q p_{itqr}$$

(for t from 2 to T) and compute $\psi_t(r) = \arg\max_q \alpha_{i,t-1}^q p_{itqr}$. When reaching the last locus, we obtain $\widehat{S}_{i,T} = \arg\max_r \alpha_{i,T}^r$ and the rest of the optimal path is obtained recursively as $\widetilde{S}_{i,t} = \psi_{t+1}(\widetilde{S}_{i,t+1})$ (for t from $T-1$ to 1).

4 Simulation studies

To assess the performance of our approximated inference procedure, so-called here *CHMM-VEM* for Variational EM for Coupled HMM, we perform two simulation studies which aim is to show the advantage of our method in terms of both computational time and classification. In Study 1, we compare the computation time of *CHMM-VEM* to the exact version (the EM algorithm called here *CHMM-EM* for EM for Coupled HMM). In Study 2, we illustrate the importance of accounting for the dependency. To this aim, we consider an independent HMM, but in order to allow a fair comparison we assume moreover that the emission parameters are common among the series. In this case, the parameters can be estimated using an EM algorithm. We denote it *iHMM-EM* , which is equivalent to *CHMM-VEM* with $\omega = 1$.

4.1 Simulation design and quality criteria

Simulation design In Study 1, we considered an increasing number of individuals $I \in \{2, 3, 4, 5\}$, whereas we kept it fixed to $I = 10$ in Study 2. For both studies, the length of the series was set to $T = 1000$ and the number of hidden states was fixed to $Q = 3$.

We considered respectively the homoscedastic case and the heteroscedastic one for residual errors. In homoscedastic case, we used Gaussian emission distributions with respective means -1, 0 and 1, and we considered an increasing sequence of noise standard deviation: $\sigma \in \{0.3, 1, 1.2\}$. The difficulty of the detection problem increases with σ. In heteroscedastic case, we consider two configurations based on (a) a Maize dataset (Bouchet *et al.* (2013)) (b) Illumina HumanHap550 array data (Wang *et al.* (2007)). Chosen means and standard deviation values correspond to estimated HMM parameters. (a) We used Gaussian emission distributions with respective means -2, 0 and 2 and associated $\sigma = 2, 0.25, 2$; (b) we used Gaussian emission distributions with respective means -3.5, 0 and 0.68 and associated $\sigma = 1.3, 0.2, 0.2$.

The correlation term W_ℓ (2) depends on both the similarities and the parameter ω. Here, in order to mimic real data, we extracted the similarity matrix $(s_{ij})_{i,j}$ ($[I \times I]$) from the genetic kinship matrix of 336 maize lines given in Bouchet *et al.* (2013). For ω, we consider two values corresponding to two levels of correlation between individuals: one case with moderate dependency (such that $\log \omega = -0.35$) and one case with weak dependency (such that $\log \omega = -0.2$). We simulated the hidden states in the following way:

1. we fixed central altered positions every 50 positions, i.e. at positions $25, 75, \ldots, 975$;
2. around each central altered position, we set a window with Poisson distributed length (with mean 15) so that alterations have various lengths;
3. for each window, we sampled the combination ℓ ($1 \leq \ell \leq K$) of individual status with probability proportional to W_ℓ, so each alteration is not carried by every individual.

Each configuration (σ, ω) was simulated 100 times. For each simulated dataset, both *iHMM-EM* and *CHMM-VEM* were run and the loci were classified using the Viterbi algorithm.

Comparison criteria To study the computational time, we measure it as the mean of runtime in second. The forward-backward is written in C, the rest is implemented in R. We consider the classification between normal (0) and altered (-1 or +1) loci. To evaluate the performances, we use the following different criteria:

- False positive rate (FPR): the proportion of erroneously detected alterations among the normal status,

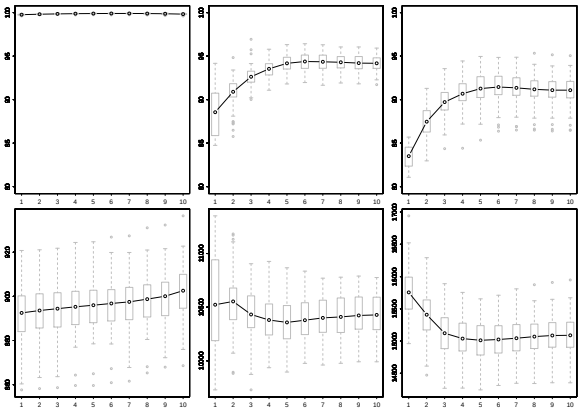

Figure 2: Boxplot of accuracy (%, top) and RSS_ω (bottom) for different values of $\omega \in \{e^{-k/20} | k = 1, 2, \ldots, 10\}$. Left: $\sigma = 0.3$, center: $\sigma = 1$, right: $\sigma = 1.2$.

- False negative rate (FNR): the proportion of erroneously estimated normal status among the alteration status,
- Accuracy: the proportion of correctly estimated status.

'Positive' corresponds to the two alteration status and 'negative' to the normal one.

For each configuration, we consider the average of these criteria over the 100 simulations.

4.2 Choice of the parameter ω

The proposed procedure does not allow to estimate the parameter ω. To select it, we propose the following strategy: we vary ω in a grid of values and select the one that minimizes a weighted Residuals Sum of Squares (RSS_ω) criterion defined by

$$\text{RSS}_\omega = \sum_{i,t,r} \tau_{it}^r (x_{i,t} - \mu_r)^2.$$

Figure 2 gives both the classification accuracy and the RSS_ω criterion for different values of ω ($\log \omega = -k/20, k \in \{1, 2, \ldots, 10\}$) in the simulation case where $I = 10$ and a weak dependency. Recall that a small value of $\omega \leq 1$ indicates a high dependency, so the x-axis of Figure 2 designs a decreasing level of dependency. We observe that when σ is small, the accuracy is not affected by the choice of ω because the segmentation problem is obvious. For larger values of σ, we observe that the RSS_ω curve displays a minimum which is close to the maximum classification accuracy. These phenomenons appear also in the case of the moderate dependency.

4.3 Study 1

Only the results with weak dependency and $\sigma = 1$ are presented, the other configurations lead to the same conclusions. Table 1 gives the median of runtime in second on a PC with 3.2GHz with increasing number of individuals and Figure S1 in Supplementary presents the classification accuracy only with $I = 3$ individuals. As expected, *CHMM-EM* out-beats (slightly) *CHMM-VEM* followed by *iHMM-EM* in terms of accuracy. However, the runtime of *CHMM-EM* is exponential growth as I increases and can not be used for larger (even small) number of individuals. Note that the runtime of *CHMM-VEM* is slightly better compared to *iHMM-EM*.

Supplementary Figure S1 also shows that accounting for dependency between individuals improves the accuracy.

I	*iHMM-EM*	*CHMM-VEM*	*CHMM-EM*
2	0.8	0.4	2.0
3	1.1	0.5	11.2
4	1.2	0.6	79.4
5	1.6	0.8	920.2

Table 1: Runtime depending on the number of individuals I (in second)

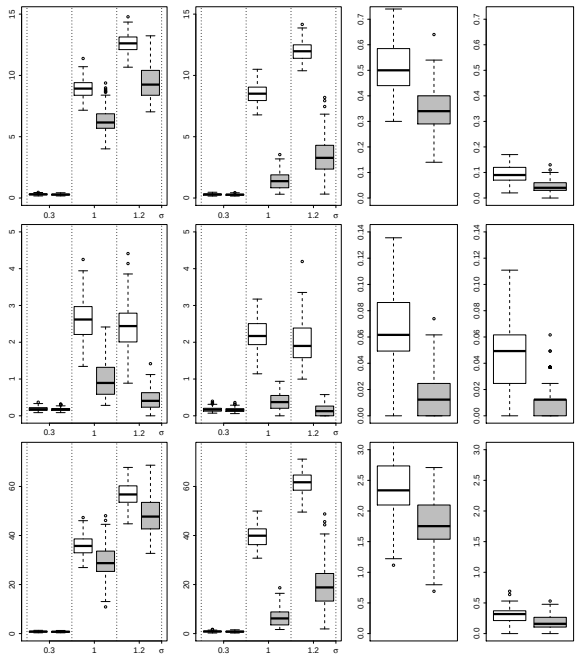

Figure 3: First column: homoscedastic and weak dependency case. Second column: homoscedastic and moderate dependency case. Third column: first heteroscedastic case and weak dependency case. Forth column: second heteroscedastic case and weak dependency case. Boxplots of classification error rate (%, top), FPR (%, center) and FNR (%, bottom) for different values of σ (x-axis). For each σ, we distinguish *iHMM-EM* (white box) and *CHMM-VEM* (gray box).

4.4 Study 2

For each configuration, ω has been chosen following the strategy described in Section 4.2. In Figure 3, we observe for homoscedastic model that when σ small, i.e. when the detection problem is easy, both *iHMM-EM* and *CHMM-VEM* perform well. However, when σ increases, *CHMM-VEM* outperforms *iHMM-EM* whatever the dependency, meaning the importance of taking into account for the existing dependency. This is more marked when this dependency increases.

For heteroscedastic model, we observe from Figure 3 that *CHMM-VEM* outperforms *iHMM-EM* whatever the configurations.

5 Application

Maize is one of the three most cultivated crop in the world and a very interesting model for studying CNV and their impact on phenotype. CNV are very numerous in maize with thousand of CNV harboring hundred of functional gene between two inbred lines (Lai *et al.* (2010),**?**,**?**). Lu *et al.* (2015) evaluated that one third of maize genome could be absent from B73 reference genome but present in another inbred lines.

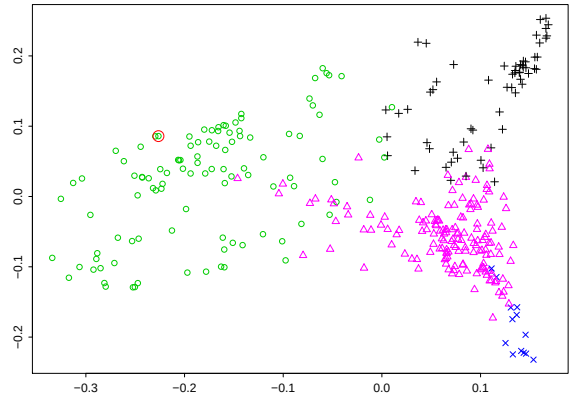

Figure 4: Multidimensional scaling of the kinship matrix (s_{ij}): First two axes. Red circle: Fv2.

Since the maize genotype B73 was sequenced in 2009 (Schnable *et al.* (2009)), B73 is usually considered as a reference genome to identify or understand the CNV in different types of maize. These experimental techniques have been investigated in different platforms such as in CGH platform (Swanson-Wagner *et al.* (2010),**?**,**?**); in NGS platform (Wang *et al.* (2014),**?**).

5.1 Data description

We consider a dataset which consists of Illumina SNP genotyping arrays on $I = 336$ maize lines (Bouchet *et al.* (2013)). The Illumina GenomeStudio software (see `http://support.illumina.com/array/arr/genomestudio/downloads.html`) was used to compute the log R ratio (LRR) defined as

$$X_{it} = \log_2 \left(R_{it}^{\text{observed}} / R_t^{\text{expected}} \right)$$

where R_{it}^{observed} is the normalized signal intensity at locus t in line i and R_t^{expected} is a reference intensity at locus t (Wang *et al.* (2007)). In this panel, two situations are expected: either the tested line i shares locus t with the reference genome and X_{it} is close to zero (normal case) or locus t does not exist in the genome of line i and X_{it} is below zero (altered case).

Among the 336 individuals, the French Fv2 inbred line has been especially studied and 58 deleted loci have been detected in contrast with B73 by sequencing method (Darracq *et al.* (2017)).

In addition to the Illumina array data, we have access to the kinship matrix (s_{ij}) between the lines (Astle and Balding (2009)), which reveals the genetic similarity between them. Figure 4 displays the multidimensional scaling (MDS) based on the similarity matrix. The clustering feature as shown in Figure 4 implies that we should analyze jointly the closed individuals rather than overall individuals.

5.2 Data analysis

Fixing the number of status As explained before, we expect only $Q = 2$ status. To validate this, we fitted an HMM on each of the lines with $Q = 2$, 3 and 4 status. Supplementary Figure S2 displays the histogram of the $Q \times I$ estimated means. Two main modes can be distinguished, which justify our choice of $Q = 2$.

Applying HMM to the analysis of 10 chromosomes in 336 individuals, we estimated the means as $-1.94(\pm 0.28)$ and $-0.05(\pm 0.02)$ for two states, respectively. The corresponding variances of errors are estimated as $3.95(\pm 0.68)$ and $0.05(\pm 0.02)$(standard deviation $1.98(\pm 0.20)$ and $0.22(\pm 0.04)$).

	iHMM-EM	*CHMM-VEM*				
I	1	6	49	80	153	336
$\overline{s}_I$	1.00	0.75	0.71	0.67	0.65	0.64
FPR(%)	12.68	10.43	10.02	9.32	8.89	8.95
FNR(%)	24.14	24.14	24.14	25.86	25.86	25.86

Table 2: Classification accuracy of *iHMM-EM* and *CHMM-VEM*. I : size of the panel. $\overline{s}_I$: mean kinship within the panel. FPR and FNR for the validated 58 Fv2 alterations.

Detecting CNV for 336 individuals As shown in Section 4, analyses jointly for correlated individuals are more effective than analyses independently from each other. Moreover, this effectiveness is obvious when the correlations among the individuals are strong. In order to get some better correlated groups, we divide 336 individuals into 4 groups inspired from hierarchical clustering, then analyse one by one. The distance between these four groups is represented in Figure 4. As shown in Supplementary Table S1, the analysis of the four groups compared to a single analysis gain nearly 1e4 deletion locus.

Supplementary Figure S3 displays the correlation between original similarity matrix and correlation matrix estimated separately by *iHMM-EM* and *CHMM-VEM*. We notice that the analysis accounting for the dependency between individuals by *CHMM-VEM* are much more revealing than that of *iHMM-EM* in terms of similarity structure among individuals.

Supplementary Figure S4 lists the overlapped number of deleted loci for *iHMM-EM* and *CHMM-VEM*.

The simulation results from Figure 3 show that accuracy of *CHMM-VEM* is greater than that of *iHMM-EM* under some different parameter scenarios. Hence, we believe that *CHMM-VEM* gives more exact result than *iHMM-EM*, although *iHMM-EM* can find more deleted loci than *CHMM-VEM*.

Classification accuracy We use the 58 deletions detected in Fv2 by sequencing as references to compare the classification performances of *iHMM-EM* and *CHMM-VEM*. In particular, we study how the selection of the panel of lines does influence the results. To this aim, we ordered the lines by decreasing kinship with Fv2 and defined a sequence of panel with increasing sizes.

The results are given in Table 2. We observe that the joint analysis with correlated lines reduces the proportion of falsely detected alterations.

6 Discussion

In practice, hundreds or thousands of individuals are often simultaneously analyzed to detect the CNV. Especially for animal or plant species, these individuals share usually a common phylogenetic past. Therefore, their similarity relationship motivate us to focus firstly on constructing the probabilistic model on transition structure accounting for the kinship matrix. Next, we use the variational inference for CHMM in order to enable to handle jointly a large size of individuals. Simulation studies and real data analysis demonstrate that the account for the kinship between individuals improves the detection of CNV.

In addition, our transition models are compatible with the heterogeneous transition models and more sophisticated emission models such as in Wang *et al.* (2007),**?**,**?** in the context of CNV detection using SNP genotyping data. Furthermore, the read count data collected by NGS techniques is usually used to detect CNV in recent years. Taking some emission distributions based on Poisson or negative binomial distribution such as Wang *et al.* (2014),**?**, our model can be also easily extended to detect CNV for NGS platforms.

Our method can be widened in CGH platform to detect CNV.That platform is based on the principle of comparative hybridization of two labelled individuals, say test and reference to a set of hybridization targets. The logarithm of signal ratio is used as the data to observe the copy number. For instance, when comparing two individuals test and reference, a deletion in the reference individual is indistinguishable from an amplification in the test individual. In this section, as shown in Supplementary Figure S5, we consider a design which have multiple comparisons such as $(i,j),(i,k),(j,k),\ldots$. We assume $m^+(i)$ the set of test individuals while taking individual i as reference; conversely, we assume $m^-(i)$ the set of reference individuals while taking individual i as test. Similar to the strategy in above sections, we search some independent heterogenous HMMs to approximate the original distribution as shown in Supplementary Figure S5 in terms of Küllback-Leibler divergence $KL(\mathbb{P}||\widetilde{\mathbb{P}})$. Taking the same notations as above, the parameter in approximated HMM can be computed as

$$h^r_{it} = \Omega^r_{it} \prod_{j\in m^-(i),v} \phi_{\gamma_{rv}}(X_{(i,j),t})^{\frac{\tau^v_{jt}}{2}} \prod_{j\in m^+(i),u} \phi_{\gamma_{ur}}(X_{(j,i),t})^{\frac{\tau^u_{jt}}{2}}$$

(the product over all individuals compared with i appears because the observations are paired, so we always have to deal with the joint distribution of (X_{it}, X_{jt}), as opposed as in (5)).

The algorithms in current paper have been implemented in the R (R Core Team (2015)) package `CHMM`. The R package is available from the Comprehensive R Archive Network.

Acknowledgements

This work was supported by the CNV-Maize program funded by the french National Research Agency (ANR-10-GENM-104) and France Agrimer (11000415). Xiaoqiang Wang was financed by CNV-Maize project and National Natural Science Foundation of China (11601286). We are grateful to Stéphane Nicolas for providing the maize dataset.

References

[Alkan *et al.* (2011)] Alkan, C., Coe, B. P. and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics.* **12 (5)** 363–376.

[Astle and Balding (2009)] Astle, W. and Balding, D. J. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science.* **24 (4)** 451–471.

[Bouchet *et al.* (2013)] Bouchet, S., Servin, B., Bertin, P., Madur, D., Combes, V., Dumas, F., Brunel, D., Laborde, J., Charcosset, A. and Nicolas, S. (2013). Adaptation of maize to temperate climates: Mid-density genome-wide association genetics and diversity patterns reveal key genomic regions, with a major contribution of the Vgt2 (ZCN8) locus. *PLoS ONE.* **8 (8)** e71377.

[Choi *et al.* (2013)] Choi, H., Fermin, D., Nesvizhskii, A. I., Ghosh, D. and Qin, Z. S. (2013). Sparsely correlated hidden Markov models with application to genome-wide location studies. *Bioinformatics.* **29 (5)** 533–541.

[Darracq *et al.* (2017)] Darracq, A., Vitte, C., Nicolas, S., Duarte, J., Pichon, J., Aubert, J., Wang, X., Mary-Huard, T., Chevalier, C., Charcosset, A., Lepaslier, M., Rogowsky, P. and Joets, J. (2017). Sequence analysis of European maize inbred line FV2 provides new insights into molecular and chromosomal characteristics of presence/absence variants.

[Daudin *et al.* (2008)] Daudin, J.-J., Picard, F. and Robin, S. (Jun, 2008). A mixture model for random graphs. *Stat. Comput.* **18 (2)** 173–83.

[Dellinger *et al.* (2010)] Dellinger, A. E., Saw, S.-M., Goh, L. K., Seielstad, M., Young, T. L. and Li, Y.-J. (2010). Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nucleic Acids Research.* **38 (9)** e105–e105.

[Dempster *et al.* (1977)] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. **39** 1–38.

[Ghahjaverestan *et al.* (2016)] Ghahjaverestan, N. M., Masoudi, S., Shamsollahi, M. B., Beuchée, A., Pladys, P., Ge, D. and Hernández, A. I. (2016). Coupled hidden Markov model-based method for apnea bradycardia detection. *IEEE Journal of Biomedical and Health Informatics.* **20 (2)** 527–538.

[Ghahramani and Jordan (1997)] Ghahramani, Z. and Jordan, M. I. (1997). Factorial hidden Markov models. *Machine learning.* **29 (2-3)** 245–273.

[Hu *et al.* (2016)] Hu, J., Zhang, L. and Wang, H. J. (2016). Sequential model selection-based segmentation to detect DNA copy number variation. *Biometrics.* **72 (3)** 815–826.

[Jaakkola (2000)] Jaakkola, T. S. (2000). *Advanced mean field methods: theory and practice.* chapter Tutorial on variational approximation methods. MIT Press.

[Ji and Chen (2016)] Ji, T. and Chen, J. (2016). Statistical models for dna copy number variation detection using read-depth data from next generation sequencing experiments. *Aust. N. Z. J. Stat.* **58 (4)** 473–491.

[Lai *et al.* (2010)] Lai, J., Li, R., Xu, X., Jin, W., Xu, M., Zhao, H., Xiang, Z., Song, W., Ying, K., Zhang, M., Jiao, Y., Ni, P., Zhang, J., Li, D., Guo, X., Ye, K., Jian, M., Wang, B., Zheng, H., Liang, H., Zhang, X., Wang, S., Chen, S., Li, J., Fu, Y., Springer, N. M., Yang, H., Wang, J., Dai, J., Schnable, P. S. and Wang, J. (2010). Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat Genet.* **42 (11)** 1027–1030.

[Lai *et al.* (2005)] Lai, W. R., Johnson, M. D., Kucherlapati, R. and Park, P. J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics.* **21 (19)** 3763.

[Liu *et al.* (2016)] Liu, Y., Liu, J., Lu, J., Peng, J., Juan, L., Zhu, X., Li, B. and Wang, Y. (2016). Joint detection of copy number variations in parent-offspring trios. *Bioinformatics.* **32 (8)** 1130–1137.

[Lu *et al.* (2015)] Lu, F., Romay, M. C., Glaubitz, J. C., Bradbury, P. J., Elshire, R. J., Wang, T., Li, Y., Li, Y., Semagn, K., Zhang, X., Hernandez, A. G., Mikel, M. A., Soifer, I., Barad, O. and Buckler, E. S. (2015). High-resolution genetic mapping of maize pan-genome sequence anchors. *Nature Communications.* **6** 6914 EP –.

[Lucito *et al.* (2003)] Lucito, R., Healy, J., Alexander, J., Reiner, A., Esposito, D., Chi, M., Rodgers, L., Brady, A., Sebat, J., Troge, J., West, J. A., Rostan, S., Nguyen, K. C., Powers, S., Ye, K. Q., Olshen, A., Venkatraman, E., Norton, L. and Wigler, M. (2003). Representational oligonucleotide microarray

analysis: A high-resolution method to detect genome copy number variation. *Genome Research.* **13 (10)** 2291–2305.

[MacDonald *et al.* (2014)] MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L. and Scherer, S. W. (2014). The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research.* **42 (D1)** D986–D992.

[Magi *et al.* (2012)] Magi, A., Tattini, L., Pippucci, T., Torricelli, F. and Matteo Benelli2, ., 4. (2012). Read count approach for DNA copy number variants detection. *Bioinformatics.* **28 (4)** 470–478.

[Nock and Ostendorf (2003)] Nock, H. and Ostendorf, M. (2003). Parameter reduction schemes for loosely coupled HMMs. *Computer Speech & Language.* **17 (2–3)** 233 – 262.

[Picard *et al.* (2011)] Picard, F., Lebarbier, E., Budinska, E. and Robin, S. (2011). Joint segmentation of multivariate Gaussian processes using mixed linear models. *Computational Statistics & Data Analysis.* **55 (2)** 1160–1170.

[Pinto *et al.* (2011)] Pinto, D., Darvishi, K., Shi, X., Rajan, D., Rigler, D., Fitzgerald, T., Lionel, A. C., Thiruvahindrapuram, B., MacDonald, J. R., Mills, R., Prasad, A., Noonan, K., Gribble, S., Prigmore, E., Donahoe, P. K., Smith, R. S., Park, J. H., Hurles, M. E., Carter, N. P., Lee, C., Scherer, S. W. and Feuk, L. (2011). Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature Biotechnology.* **29 (6)** 512–520.

[R Core Team (2015)] (2015). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

[Rezek *et al.* (2002)] Rezek, I., Gibbs, M. and Roberts, S. J. (2002). Maximum a posteriori estimation of coupled hidden Markov models. *Journal of VLSI signal processing systems for signal, image and video technology.* **32 (1)** 55–66.

[Saul and Jordan (1995)] Saul, L. and Jordan, M. I. (1995). Exploiting tractable substructures in intractable networks. In *Advances in Neural Information Processing Systems 8*, 486–492. MIT Press.

[Saul *et al.* (1996)] Saul, L. K., Jaakkola, T. and Jordan, M. I. (1996). Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research.* **4 (1)** 61–76.

[Saul and Jordan (1999)] Saul, L. K. and Jordan, M. I. (1999). Mixed memory Markov models: Decomposing complex stochastic processes as mixtures of simpler ones. *Machine Learning.* **37 (1)** 75–87.

[Schnable *et al.* (2009)] Schnable, P., Ware, D., R.S., F. and et al. (2009). The b73 maize genome: Complexity, diversity, and dynamics. *Science.* **326** 1112–1115.

[Schwarz (1978)] Schwarz, G. (1978). Estimating the dimension of a model. **6** 461–4.

[Sherlock *et al.* (2013)] Sherlock, C., Xifara, T., Telfer, S. and Begon, M. (2013). A coupled hidden Markov model for disease interactions. *Journal of the Royal Statistical Society. Series C, Applied Statistics.* **62 (4)** 609–627.

[Speed and Balding (2015)] Speed, D. and Balding, D. J. (2015). Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics*. **16 (1)** 33–44.

[Sun *et al.* (2009)] Sun, W., Wright, F. A., Tang, Z., Nordgard, S. H., Loo, P. V., Yu, T., Kristensen, V. N. and Perou, C. M. (2009). Integrated study of copy number states and genotype calls using high-density SNP arrays. *Nucleic Acids Research*. **37 (16)** 5365–5377.

[Swanson-Wagner *et al.* (2010)] Swanson-Wagner, R. A., Eichten, S. R., Kumari, S., Tiffin, P., Stein, J. C., Ware, D. and Springer, N. M. (2010). Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Research*. **20 (12)** 1689–1699.

[Tai *et al.* (2010)] Tai, Y. C., Kvale, M. N. and Witte, J. S. (2010). Segmentation and estimation for SNP microarrays: A Bayesian multiple change-point approach. *Biometrics*. **66 (3)** 675–683.

[Wainwright and Jordan (2008)] Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1 (1-2)** 1–305.

[Wang *et al.* (2014)] Wang, H., Nettleton, D. and Ying, K. (2014). Copy number variation detection using next generation sequencing read counts. *BMC Bioinformatics*. **15** 109–109.

[Wang *et al.* (2007)] Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F., Hakonarson, H. and Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome snp genotyping data. *Genome research*. **17 (11)** 1665–1674.

[Wang *et al.* (2008)] Wang, K., Chen, Z., Tadesse, M. G., Glessner, J., Grant, S. F. A., Hakonarson, H., Bucan, M. and Li, M. (2008). Modeling genetic inheritance of copy number variations. *Nucleic Acids Research*. **36 (21)** e138–e138.

[Weischenfeldt *et al.* (2013)] Weischenfeldt, J., Symmons, O., Spitz, F. and Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature Reviews Genetics*. **14 (2)** 125–138.

[Winchester *et al.* (2009)] Winchester, L., Yau, C. and Ragoussis, J. (2009). Comparing CNV detection methods for SNP arrays. *Briefings in Functional Genomics and Proteomics*. **8 (5)** 353–366.

[Xu *et al.* (2014)] Xu, L., Cole, J. B., Bickhart, D. M., Hou, Y., Song, J., VanRaden, P. M., Sonstegard, T. S., Van Tassell, C. P. and Liu, G. E. (2014). Genome wide CNV analysis reveals additional variants associated with milk production traits in holsteins. *BMC Genomics*. **15 (1)** 683.

[Zarrei *et al.* (2015)] Zarrei, M., MacDonald, J. R., Merico, D. and Scherer, S. W. (2015). A copy number variation map of the human genome. *Nature Reviews Genetics*. **(3)** 172–183.

[Zhang *et al.* (2010)] Zhang, N. R., Siegmund, D. O., Ji, H. and Li, J. Z. (2010). Detecting simultaneous changepoints in multiple sequences. *Biometrika*. **97 (3)** 631–645.

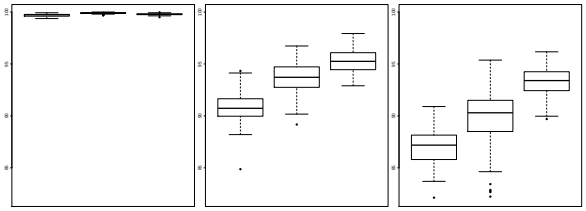

Figure 5: Classification accuracy (%) *iHMM-EM* (left), *CHMM-VEM* (middle) and *CHMM-EM* (right) for $I = 3$. Left: $\sigma = 0.3$. Middle: $\sigma = 1$. Right: $\sigma = 1.2$

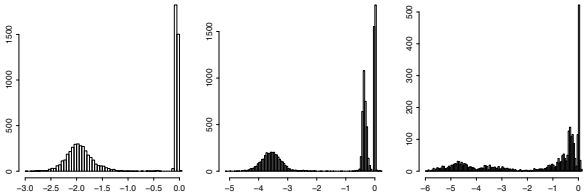

Figure 6: Histogram of the mean estimated by independent HMM for 336 individuals. Left: $Q = 2$, center: $Q = 3$, right: $Q = 4$,

[Zhao *et al.* (2013)] Zhao, M., Wang, Q., Wang, Q., Jia, P. and Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*. **14 (Suppl 11)** S1.

[Zhong and Ghosh (2002)] Zhong, S. and Ghosh, J. (2002). HMMs and coupled HMMs for multi-channel EEG classification. In *Neural Networks, 2002. IJCNN '02. Proceedings of the 2002 International Joint Conference on*, volume 2, 1154–1159.

[Zhou *et al.* (2016)] Zhou, Y., Utsunomiya, Y. T., Xu, L., Hay, E. H. a., Bickhart, D. M., Alexandre, P. A., Rosen, B. D., Schroeder, S. G., Carvalheiro, R., de Rezende Neves, H. H., Sonstegard, T. S., Van Tassell, C. P., Ferraz, J. B. S., Fukumasu, H., Garcia, J. F. and Liu, G. E. (2016). Genome-wide CNV analysis reveals variants associated with growth traits in bos indicus. *BMC Genomics*. **17 (1)** 419.

Supplementary Materials

Table 3: Classification comparison between 4 groups and one 1 group

		4 groups	
		Deletion	Normal
1 group	Deletion	1469821	49679
	Normal	59456	17082820

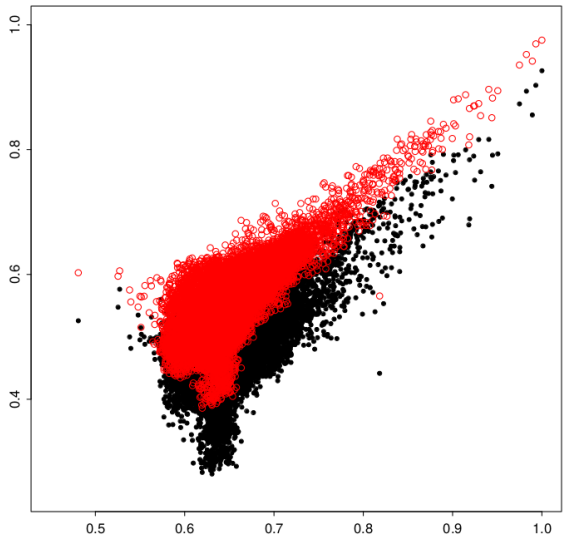

Figure 7: Correlation between original similarity matrix and correlation matrix estimated by *iHMM-EM* (Black), *CHMM-VEM* (Red)

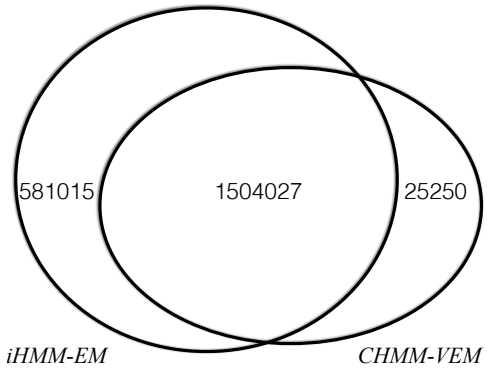

Figure 8: Venn diagram of deleted loci detected by *iHMM-EM* and *CHMM-VEM*

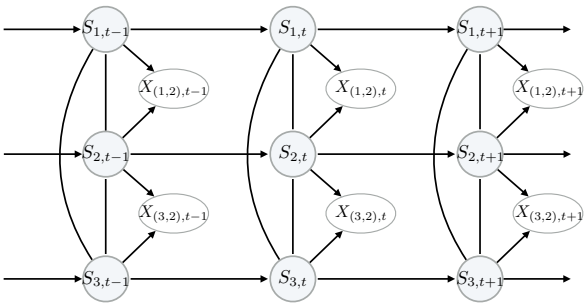

Figure 9: A mixture of directed and undirected graphical model for CGH. The directed edges represent the dependency relationship within the individual. The undirected edges represent the genetical correlation among individuals.