Supplementary Material
A Global Structural EM Algorithm for a Model of Cancer Progression

1 The derivation

A Hidden-variable Oncogenetic Tree (HOT) is a rooted binary tree in which we associate two binary stochastic variables, $Z(u)$ and $X(u)$, with each vertex u. The vector Z represents the hidden variables and X the visible ones. Let $\mathcal{T}$ be a HOT and let r be the root of $\mathcal{T}$. For the root, we set $\Pr[Z(r) = 1] = 1$ and $\Pr[X(r) = 1] = 1$. For a non-root vertex u, the probabability that $Z(u) = 1$ will depend on $Z(p(u))$, i.e., we have two conditional probability distributions associated with the edge $(p(u), u)$, namely

$$\begin{aligned}&\Pr[Z(u)|Z(p(u)) = 0], \quad \text{and}\\&\Pr[Z(u)|Z(p(u)) = 1].\end{aligned}$$

As for the visible variables, two conditional probability distributions are associated with each:

$$\begin{aligned}&\Pr[X(u)|Z(u) = 0], \quad \text{and}\\&\Pr[X(u)|Z(u) = 1].\end{aligned}$$

The standard derivation of the EM-algorithm works for our HOTs as long as $\Pr[Z, X|\mathcal{T}]$ is non-zero for all Zs and Xs. Practically, when implementing the EM-algorithm, we ensure that all of the parameters of a HOT $\mathcal{T}$, i.e., the conditional probabilities associated with the vertices of $\mathcal{T}$, are non-zero by applying a lower bound on the parameters. In other words, after each iteration of the EM-algorithm, which results in a new HOT with higher likelihood, the paramters of the HOT are adjusted to ensure that every probability is at least as large as the lower bound.

To derive the EM-algorithm, we need to be able to maximize the so-called Q-term, i.e., given a set D of observations of the visible variables and a HOT, we need to be able to find the HOT $\mathcal{T}'$ that maximizes

$$
\begin{aligned}
Q(\mathcal{T}';\mathcal{T}) &= \sum_{X\in D}\sum_{Z}\Pr[Z|X,\mathcal{T}]\log(\Pr[Z,X|\mathcal{T}'])\\
&= \sum_{X\in D}\sum_{Z}\Pr[Z|X,\mathcal{T}]\log(\Pr[Z|\mathcal{T}']\Pr[X|Z,\mathcal{T}'])\\
&= \sum_{X\in D}\sum_{Z}\Pr[Z|X,\mathcal{T}]\log(\prod_{(u,v)\in\mathcal{T}'}\Pr[Z(v)|Z(u),\mathcal{T}']\Pr[X(v)|Z(v),\mathcal{T}'])\\
&= \sum_{X\in D}\sum_{Z}\Pr[Z|X,\mathcal{T}]\sum_{(u,v)\in\mathcal{T}'}\Big(\log(\Pr[Z(v)|Z(u),\mathcal{T}'])+\log(\Pr[X(v)|Z(v),\mathcal{T}'])\Big)\\
&= \sum_{(u,v)\in\mathcal{T}'}\left(\sum_{X\in D}\sum_{Z}\Pr[Z|X,\mathcal{T}]\log(\Pr[Z(v)|Z(u),\mathcal{T}'])\quad+\sum_{X\in D}\sum_{Z}\Pr[Z|X,\mathcal{T}]\log(\Pr[X(v)|Z(v),\mathcal{T}'])\right)\\
&= \sum_{(u,v)\in\mathcal{T}'}\Bigg(\sum_{X\in D}\sum_{a,b\in\{0,1\}}\log(\Pr[Z(v)=b|Z(u)=a,\mathcal{T}'])\sum_{\substack{Z:Z(u)=a,\\ Z(v)=b}}\Pr[Z|X,\mathcal{T}]\\
&\qquad\qquad+\sum_{\sigma,a\in\{0,1\}}\log(\Pr[X(v)=\sigma|Z(v)=a,\mathcal{T}'])\sum_{\substack{X\in D:\\ X(v)=\sigma}}\sum_{Z:Z(v)=a}\Pr[Z|X,\mathcal{T}]\Bigg)\\
&= \sum_{(u,v)\in\mathcal{T}'}\Bigg(\sum_{a,b\in\{0,1\}}\sum_{X\in D}\log(\Pr[Z(v)=b|Z(u)=a,\mathcal{T}'])\Pr[Z(u)=a,Z(v)=b|X,\mathcal{T}]\\
&\qquad\qquad+\sum_{\sigma,a\in\{0,1\}}\sum_{\substack{X\in D:\\ X(v)=\sigma}}\log(\Pr[X(v)=\sigma|Z(v)=a,\mathcal{T}'])\Pr[Z(v)=a|X,\mathcal{T}]\Bigg)\\
&= \sum_{(u,v)\in\mathcal{T}'}\Bigg(\sum_{a,b\in\{0,1\}}\log(\Pr[Z(v)=b|Z(u)=a,\mathcal{T}'])\sum_{X\in D}\Pr[Z(u)=a,Z(v)=b|X,\mathcal{T}]\\
&\qquad\qquad+\sum_{\sigma,a\in\{0,1\}}\log(\Pr[X(v)=\sigma|Z(v)=a,\mathcal{T}'])\sum_{\substack{X\in D:\\ X(v)=\sigma}}\Pr[Z(v)=a|X,\mathcal{T}]\Bigg)\\
&= \sum_{(u,v)\in\mathcal{T}'}\Bigg(\sum_{a,b\in\{0,1\}}\log(\Pr[Z(v)=b|Z(u)=a,\mathcal{T}'])A_{(u,v)}(a,b)\\
&\qquad\qquad+\sum_{\sigma,a\in\{0,1\}}\log(\Pr[X(v)=\sigma|Z(v)=a,\mathcal{T}'])B_v(\sigma,a)\Bigg),
\end{aligned}
$$

where

$$
\begin{aligned}
A_{(u,v)}(a,b) &= \sum_{X\in D}\Pr[Z(u)=a,Z(v)=b|X,\mathcal{T}],\quad\text{and}\\
B_v(\sigma,a) &= \sum_{\substack{X\in D:\\ X(v)=\sigma}}\Pr[Z(v)=a|X,\mathcal{T}].
\end{aligned}
$$

Maximizing the Q-term is now seen to be standard procedure. Given $\mathcal{T}$, we are able to compute $A_{(u,v)}(a,b)$ for $a, b \in \{0,1\}$, and for any edge $(u,v) \in \mathcal{T}'$ the parameters are optimized by setting

$$\Pr[Z(v) = a | Z(u) = b] = \frac{A_{(u,v)}(a,b)}{A_{(u,v)}(a,b) + A_{(u,v)}(1-a,b)}.$$

Furthermore, it is clear that the optimal parameters associated with an edge is independant of any other edges present in $\mathcal{T}'$. Therefore, we can find the optimum paramters for each possible edge and find the tree that maximizes the Q-term.

Also, note that associating a weight $f(X)$ with each datapoint $X \in D$ causes us no difficulties in optimizing the Q-term. The only difference is we would have to add the factor $f(x)$ to each term of the definition of $A_{(u,v)}(a,b)$ and $B_v(\sigma, a)$.

2 Experiments

The following figures show some of the detailed results of our experimental analysis.

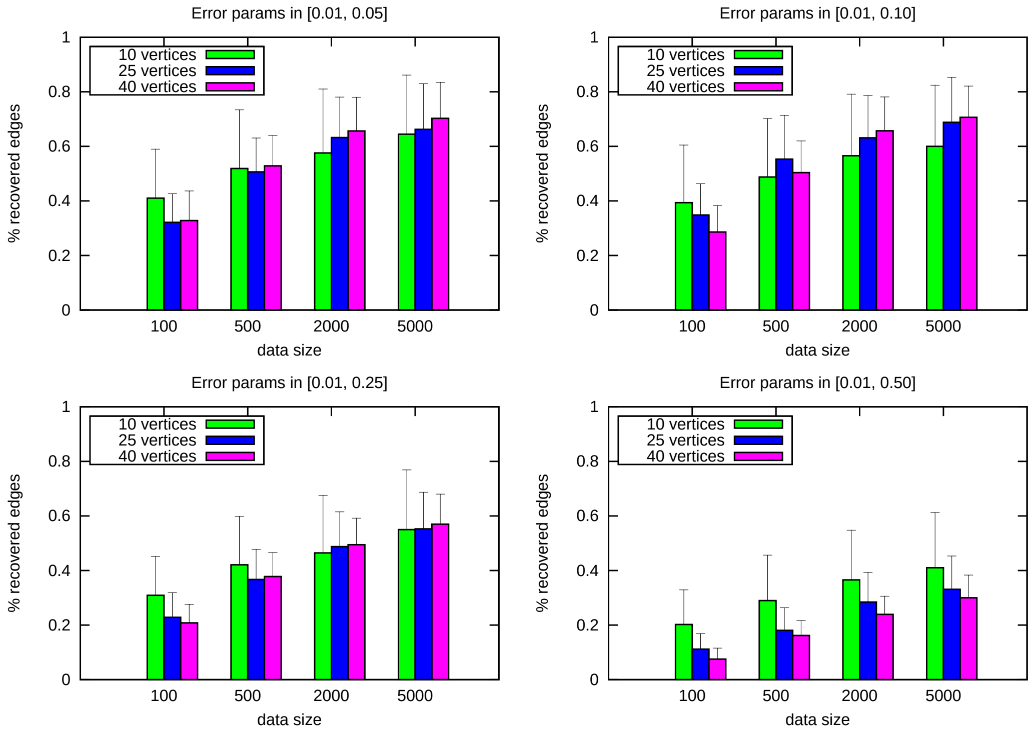

Figure 1: Histograms showing proportion of edges correctly recovered by the EM algorithm with free parameters on synthetic data. Error bars show one standard deviation.

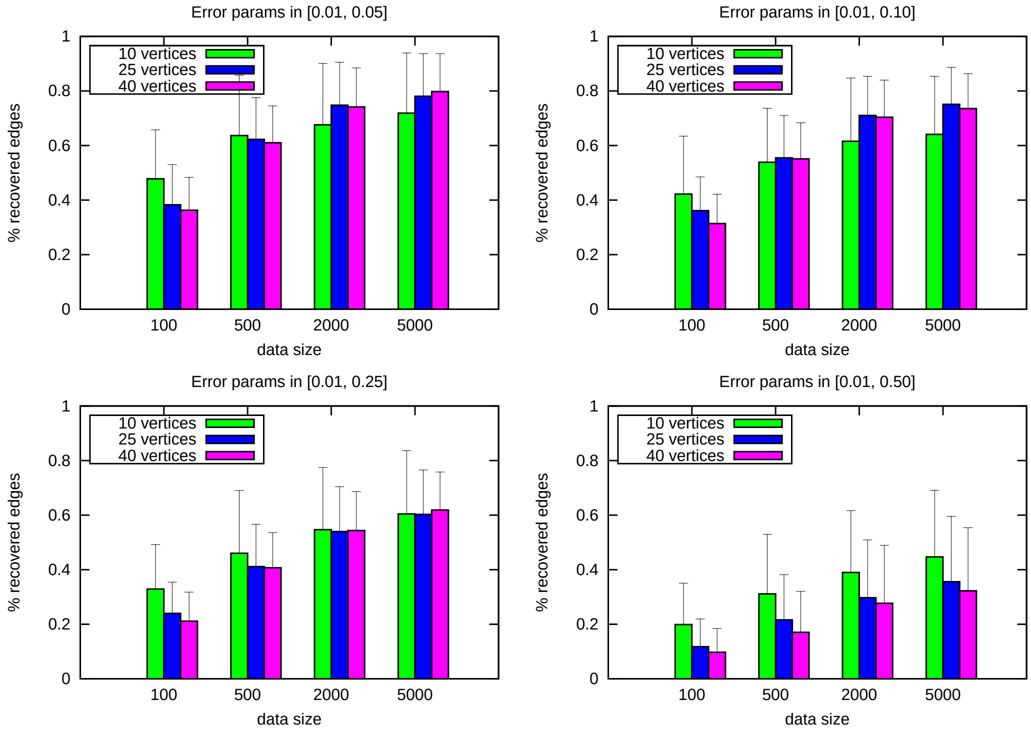

Figure 2: Histograms showing proportion of edges correctly recovered by the EM algorithm with global parameters on synthetic data. Error bars show one standard deviation.

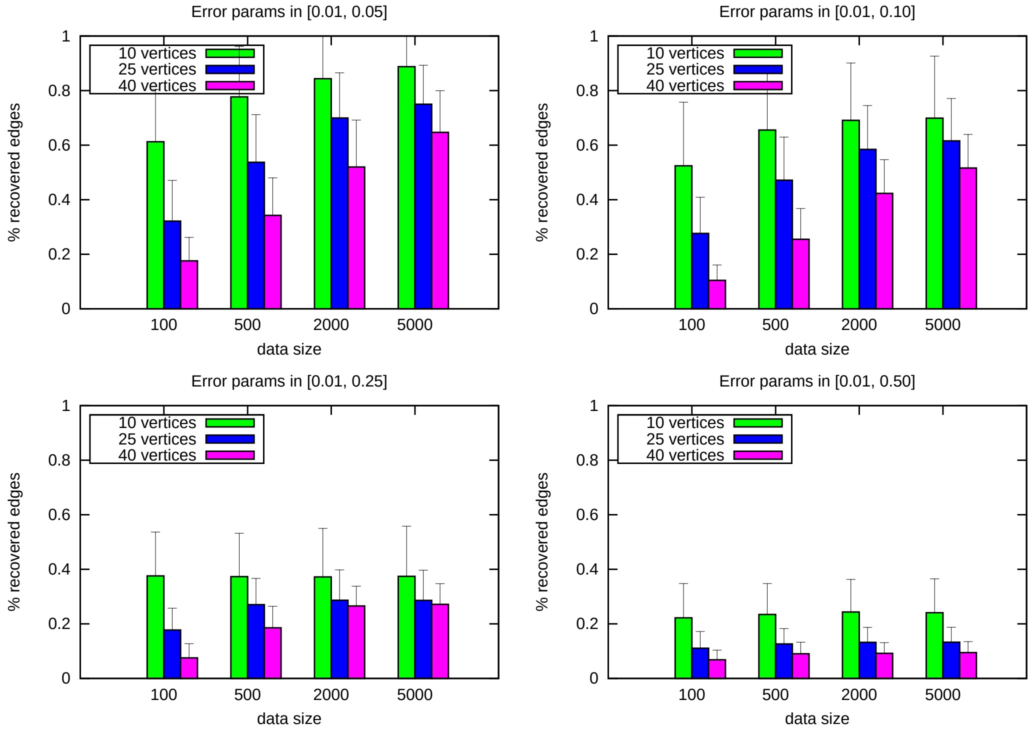

Figure 3: Histograms showing proportion of edges correctly recovered by Mtreemix on the same synthetic data as in Figure 1. Error bars show one standard deviation.

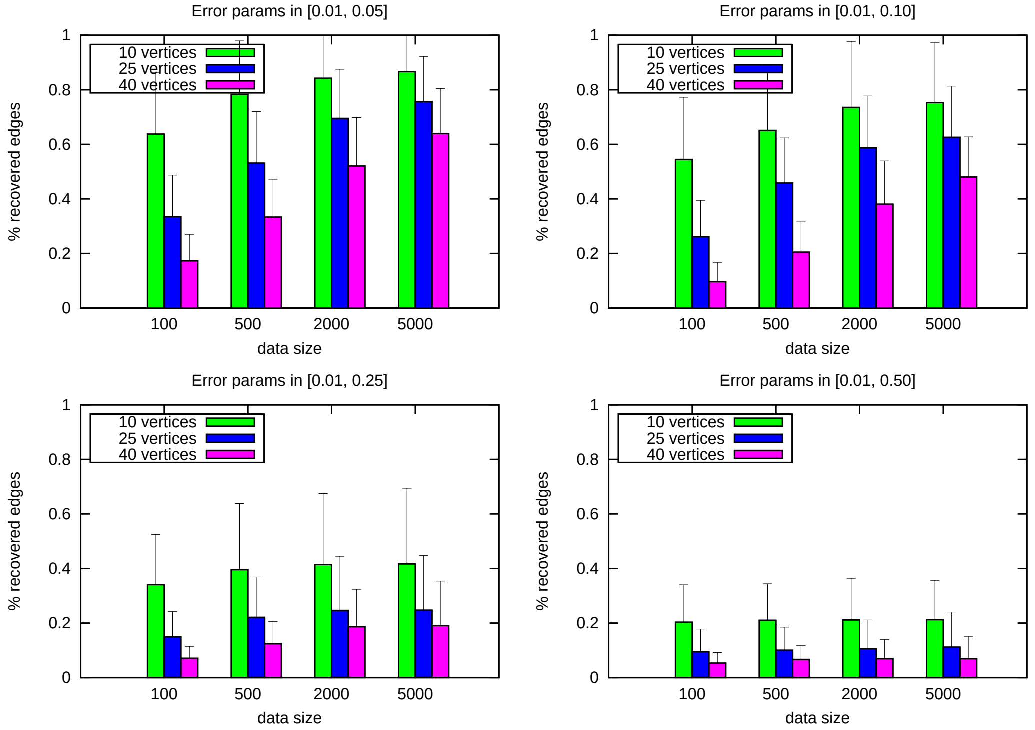

Figure 4: Histograms showing proportion of edges correctly recovered by Mtreemix on the same synthetic data as in Figure 2. Error bars show one standard deviation.

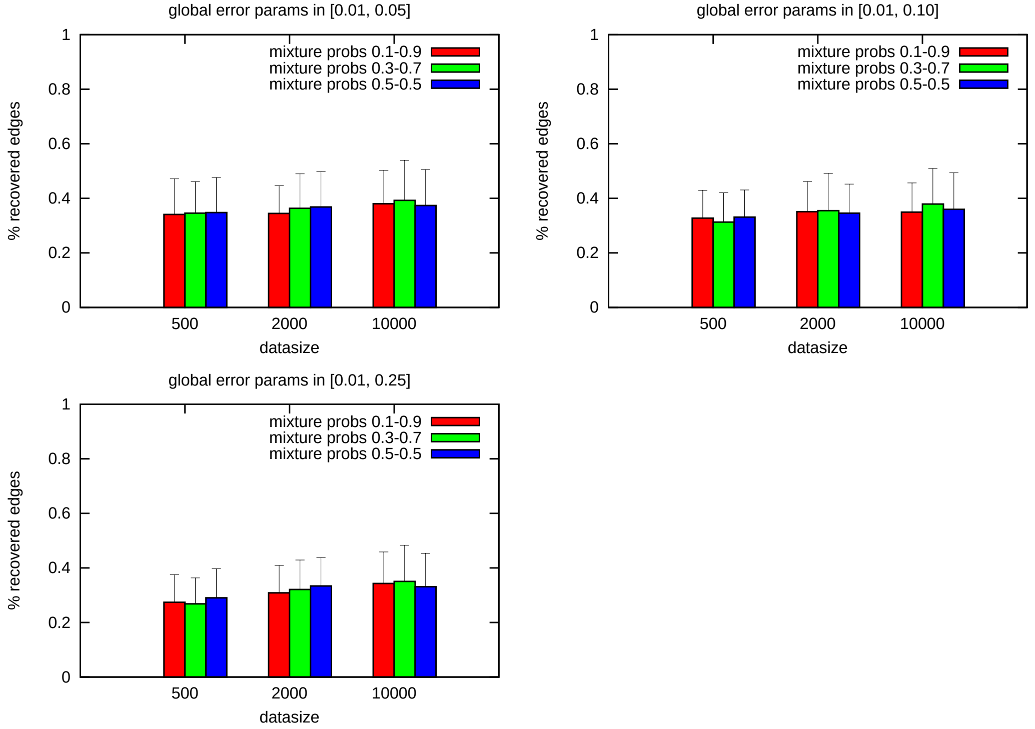

Figure 5: Histograms showing proportion of edges correctly recovered by the EM algorithm for HOT-mixtures with global parameters on two HOTs with 10 vertices each. Each bar represents 100 mixtures. Error bars show one standard deviation.

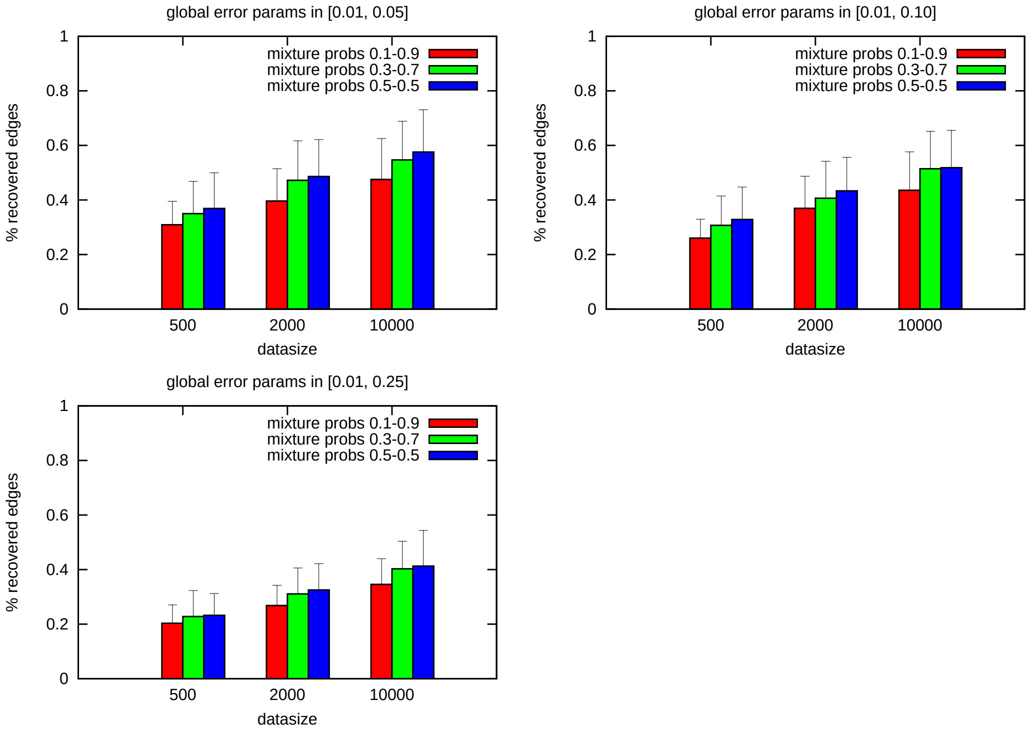

Figure 6: Histograms showing proportion of edges correctly recovered by the EM algorithm for HOT-mixtures with global parameters on two HOTs with 25 vertices each. Each bar represents 100 mixtures. Error bars show one standard deviation.