Motif Finding in DNA Sequences

Bartosz Makaruś

Our goal is to build a simple model which lets us identify "motifs" in DNA sequences. As input we are given a size k sample of sequences $x_i = (x_{i1}, x_{i2}, \dots, x_{iw})$ (of length w, $i = 1, 2, \dots, k$), where each x_{in} is one of four bases $\{A, C, G, T\}$, which we will identify respectively with $\{1, 2, 3, 4\}$.

The model assumes that the n-th term in a sequence is a random variable that follows the distribution $\theta_n = (\theta_{1n}, \theta_{2n}, \theta_{3n}, \theta_{4n})^T$, where

$$P(X_{in} = j) = \theta_{jn}$$

and obviously $\sum_{j=1}^{4} \theta_{jn} = 1$. The matrix $\theta = (\theta_1, \theta_2, \dots, \theta_w)$ is called the position weight matrix of the motif distribution.
We also define a single separate background distribution $\theta^b = (\theta^b_1, \theta^b_2, \theta^b_3, \theta^b_4)^T$, where

$$P(X_{in} = j) = \theta^b_j,$$

meaning that the distribution is not dependent on n.

The model

Model assumption is that whether vector (sequence) X_i is generated from the motif or background distribution depends on the realisation of a latent variable $Z_i \in \{0, 1\}$, such that

$$P(Z_i = 0) = 1 - \alpha, \quad P(Z_i = 1) = \alpha,$$

where if $z_i = 1$, the sequence comes from the motif distribution, and if $z_i = 0$ - the background distribution is used. The variables Z_i are, of course, not observed.

We aim to estimate $\Theta = (\theta, \theta^b)$ (assume α is known), using the EM algorithm:

Algorithm 1: Expectation Maximisation

```
1. Initialise $\Theta^{(0)}$, set $t \leftarrow 0$
2. E-step: For each observation $i = 1, \dots, k$ and class $j = 1, \dots, M$ calculate
       $Q_i^{(t)}(j) = P(Z_i = j | X; \Theta^{(t)}).$
3. M-step: Define
       $Q(\Theta, \Theta^{(t)}) := \sum_{i=1}^{k} \sum_{j=1}^{M} Q_i^{(t)}(j) \log P(x_i, Z_i = j; \Theta),$
   maximise and update:
       $\Theta^{(t+1)} \leftarrow \arg\max_{\Theta} Q(\Theta, \Theta^{(t)}).$
4. If difference between $\Theta^{(t)}$, $\Theta^{(t+1)}$ small enough, $\hat{\Theta} \leftarrow \Theta^{(t+1)}$, and stop. Otherwise, $t \leftarrow t + 1$ and go to step 2.
Result: Estimate $\hat{\Theta}$
```

E-step details

In our model, the values Q calculated in t-th E-step are

$$Q_i^{(t)}(0) = P(Z_i = 0 | x_i; \Theta^{(t)}) = \frac{P(Z_i = 0, x_i; \Theta^{(t)})}{P(x_i; \Theta^{(t)})} = \frac{P(x_i | Z_i = 0; \Theta^{(t)}) P(Z_i = 0; \Theta^{(t)})}{P(x_i; \Theta^{(t)})} = \frac{(1-\alpha) \prod_{j=1}^{w} \theta_{x_{ij}}^{b,t}}{P(x_i; \Theta^{(t)})}$$

$$Q_i^{(t)}(1) = \frac{\alpha \prod_{j=1}^{w} \theta_{x_{ij}j}^{t}}{P(x_i; \Theta^{(t)})},$$

where of course $P(x_i, \Theta^{(t)}) = P(x_i | Z_i = 0; \Theta^{(t)}) + P(x_i | Z_i = 1; \Theta^{(t)}) = \prod_{j=1}^{w} \theta_{x_{ij}}^{b,t} + \prod_{j=1}^{w} \theta_{x_{ij}j}^{t}$.

M-step details

In the M-step we maximise the function Q defined as

$$Q(\Theta, \Theta^{(t)}) = \sum_{i=1}^{k} \sum_{j=0}^{1} Q_i^{(t)}(j) \log P(x_i, Z_i = j; \Theta),$$

with respect to Θ. It can be rewritten as follows

$$Q(\Theta, \Theta^{(t)}) = \sum_{i=1}^{k} Q_i^{(t)}(0) \log P(x_I, Z_i = 0; \Theta) + \sum_{i=1}^{k} Q_i^{(t)}(1) \log P(x_I, Z_i = 1; \Theta) =$$

$$= \sum_{i=1}^{k} Q_i^{(t)}(0) \log((1-\alpha) \prod_{j=1}^{w} \theta_{x_{ij}}^{b}) + \sum_{i=1}^{k} Q_i^{(t)}(1) \log(\alpha \prod_{j=1}^{w} \theta_{x_{ij}j}).$$

The first sum is a function of θ^b only, and the second one - of θ, which means that to maximise $Q(\Theta, \Theta^{(t)})$ is to maximise each sum individually.

Consider the first sum $Q_1(\theta^b)$ dependent only on the background distribution. The task can be formulated as maximisation of $Q_1(\theta^b)$ with a single restriction

$$Q_1(\theta^b) \to \max, \sum_{r=1}^{4} \theta_r^b = 1.$$

We can find extrema of the function using Lagrange's multipliers:

$$\begin{cases} L(\theta^b) = Q_1(\theta^b) - \lambda g(\theta^b) \\ g(\theta^b) = \theta_1^b + \theta_2^b + \theta_3^b + \theta_4^b - 1 \end{cases}$$

The derivative of L with respect to a set θ_r^b is

$$\frac{\partial L}{\partial \theta_r^b} = \frac{\partial}{\partial \theta_r^b}(\sum_{i=1}^{k} Q_i^{(t)}(0)(\log(1-\alpha) + \sum_{j=1}^{w} \log \theta_{x_{ij}}^b) - \lambda(\theta_1^b + \theta_2^b + \theta_3^b + \theta_4^b - 1)),$$

and since $\frac{\partial \log \theta_{x_{ij}}^b}{\partial \theta_r^b} = \begin{cases} 0, & \text{if } x_{ij} \neq r \\ \frac{1}{\theta_r^b}, & \text{if } x_{ij} = r \end{cases}$, we have

$$\frac{\partial L}{\partial \theta_r^b} = \sum_{i=1}^{k} (Q_i^{(t)}(0) \frac{1}{\theta_r^b} |\{j : x_{ij} = r\}|) - \lambda.$$

We equate the derivative to zero to get θ_r^b in closed form:

$$\theta_r^b = \frac{1}{\lambda} \sum_{i=1}^{k} Q_i^{(t)}(0) |\{j : x_{ij} = r\}|,$$

and utilising the restriction $\sum_{r=1}^{4} \theta_r^b = 1$, we also obtain $\lambda = w \sum_{i=1}^{k} Q_i^{(t)}(0)$, which concludes our search.

In similar fashion we aim to maximise $Q_2(\theta)$, dependent only on the parameters of the motif distribution. In this case, the potential solutions are subject to w restrictions

$$\begin{cases} L(\theta) = \sum_{i=1}^{k} Q_1^{(t)}(1)(\log \alpha + \sum_{j=1}^{w} \log \theta_{x_{ij}j}) - \sum_{i=1}^{w} \lambda_i g_i(\theta_i) \\ g_1(\theta_1) = \theta_{11} + \theta_{21} + \theta_{31} + \theta_{41} - 1 \\ \vdots \\ g_w(\theta_w) = \theta_{1w} + \theta_{2w} + \theta_{3w} + \theta_{4w} - 1 \end{cases}$$

The derivative if L with respect to single θ_{rn} is

$$\frac{\partial L}{\partial \theta_{rn}} = \frac{\partial}{\partial \theta_{rn}}(\sum_{i=1}^{k} Q_i^{(t)}(1)(\log \alpha + \sum_{j=1}^{w} \log \theta_{x_{ij}j}) - \sum_{j=1}^{w} \lambda_j g_j(\theta_j)),$$

which, given $\frac{\partial \log \theta_{x_{ij}j}}{\partial \theta_{rn}} = \begin{cases} \frac{1}{\theta_{x_{in}n}}, & \text{if } j = n, x_{in} = r \\ 0, & \text{otherwise} \end{cases}$, simplifies to

$$\frac{\partial L}{\partial \theta_{rn}} = \sum_{i=1}^{k} Q_i^{(t)}(1) \frac{1}{\theta_{x_{in}n}} 1(x_{in} = r)) - \lambda_n.$$

Equating the derivative to zero yields:

$$\frac{\partial L}{\partial \theta_{rn}} = 0 \iff \sum_{i=1}^{k} Q_i^{(t)}(1) \frac{1}{\theta_{x_{in}n}} 1(x_{in} = r) = \sum_{i:\ x_{in}=r} Q_i^{(t)}(1) \frac{1}{\theta_{rn}} = \lambda_n \implies$$

$$\implies \theta_{rn} = \frac{1}{\lambda_n} \sum_{i:\ x_{in}=r} Q_i^{(t)}(1).$$

We still need λ_n to calculate the estimate. To obtain it, we use the restriction $g_n(\theta_n) = \sum_{r=1}^{4} \theta_{rn} = 1$:

$$\sum_{r=1}^{4} \theta_{rn} = 1 \iff \sum_{r=1}^{4} \frac{1}{\lambda_n} \sum_{i:\ x_{in}=r} Q_i^{(t)}(1) = 1 \implies$$

$$\implies \lambda_n = \sum_{r=1}^{4} \sum_{i:\ x_{in}=r} Q_i^{(t)}(1) = \sum_{i=1}^{k} Q_i^{(t)}(1).$$

Notice that it does not depend on n ($\lambda_1 = \lambda_2 = \ldots = \lambda_w$). Having calculated λ_n we are able to evaluate θ_{rn}, which solves the problem in this step.

Algorithm performance

For evaluation of the final estimates, as well as a measure of difference between consecutive epochs, we use d_{tv} - total variance distance, defined as

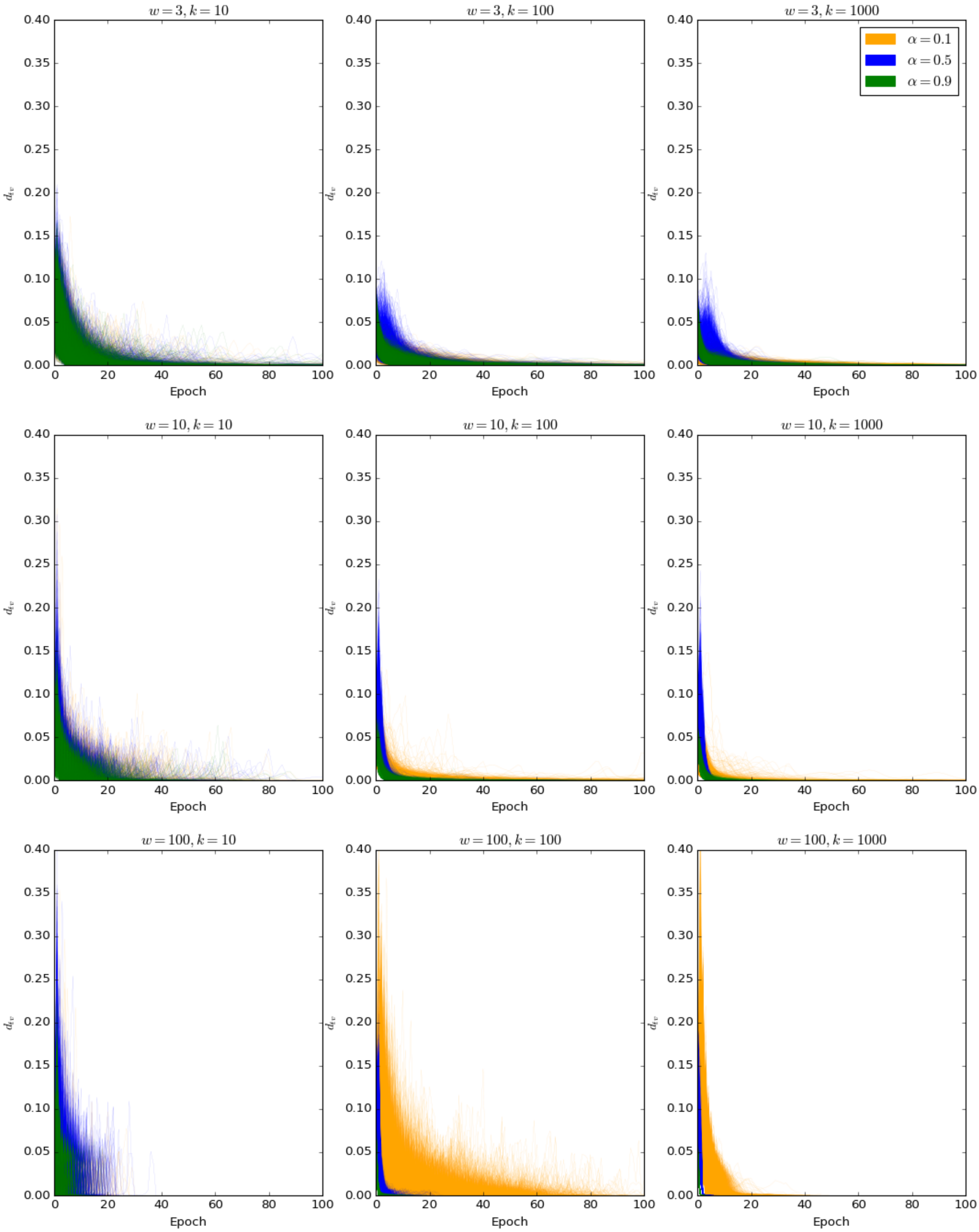
$$d_{tv}(\hat{\Theta}, \Theta) = \frac{1}{w+1} (\frac{1}{2} \sum_{i=1}^{4} |\hat{\theta}_i^b - \theta_i^b| + \sum_{k=1}^{w} (\frac{1}{2} \sum_{j=1}^{4} |\hat{\theta}_{jk} - \theta_{jk}|)).$$

To gauge the performance, we conduct a simulation for each parameter combination spanning sequence length $w = 3, 10, 100$, sample size $k = 10, 100, 1000$ and probability of "drawing" from the motif distribution $\alpha = 0.1, 0.5, 0.9$. In each case, we repeat following procedure 10000 times:

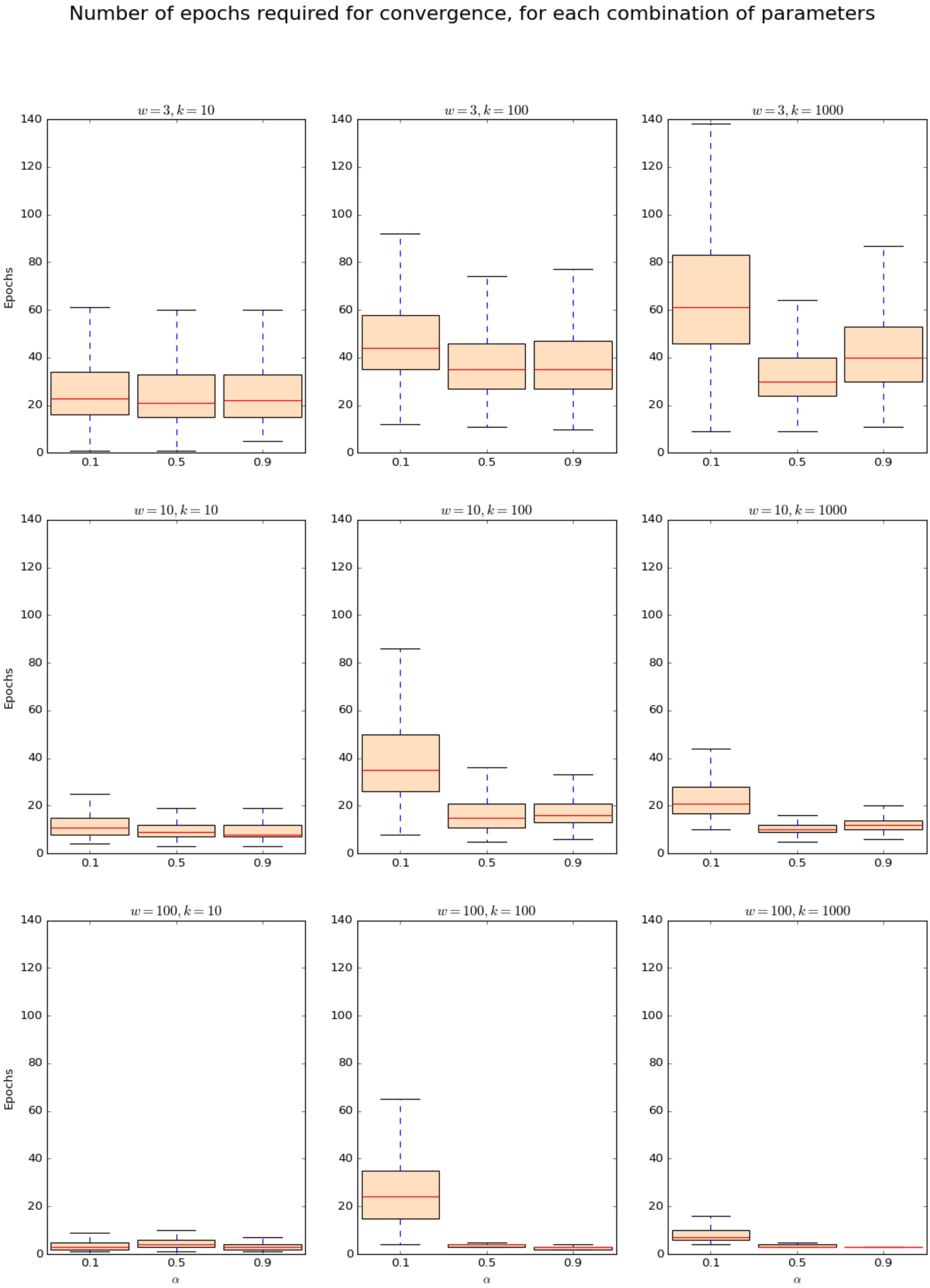
1. Elements of the parameter set Θ are generated from the uniform distribution $U(0, 1)$ and standardised so that they form probability distributions in line with the assumed model (basically so that they sum up to 1).
2. A data set (DNA sequences) are generated in accordance with the distributions and chosen α value.
3. Estimation algorithm is run - as the initial estimate, Maximum Likelihood Estimation is used, under assumption of $\alpha = 0$ or $\alpha = 1$, for θ^b and θ respectively. The algorithm stops when d_{tv} between consecutive estimates falls under 0.0001.
4. Each run yields us: number of iterations needed for the algorithm to converge, running d_{tv} history, evaluation d_{tv} calculated between final estimate and true parameters.

The algorithm convergence, as in running d_{tv} paths are plotted below

Paths of total variance distance by step, for each combination of parameters



We could also take a look at the distribution of the number of steps needed for the algorithm to converge (d_{tv} to fall under 0.0001):



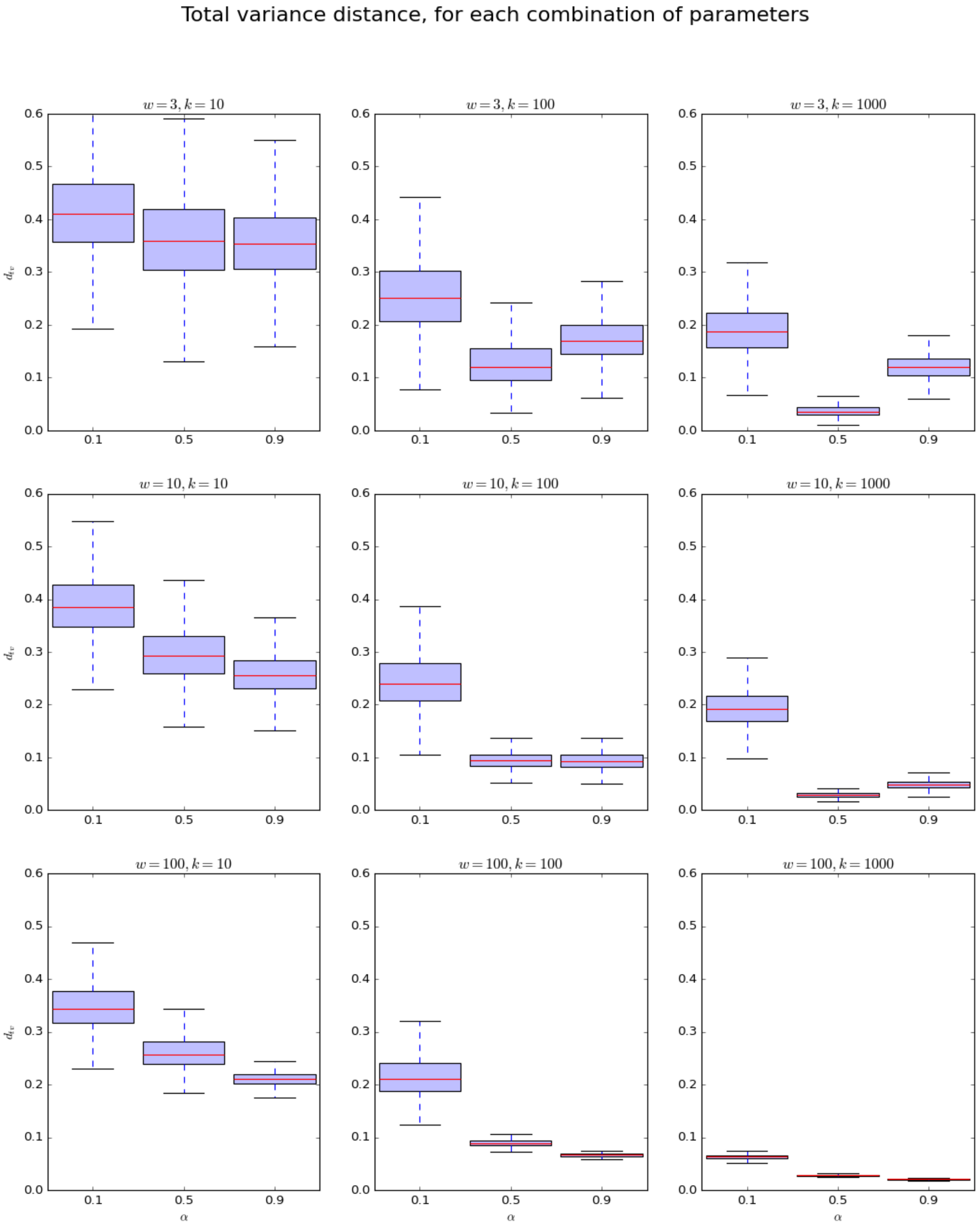
The sequence length w lowers the number of steps needed for the algorithm to converge, even if the intial d_{tv} tends to be larger.

We can also gather that for small sample size ($k = 10$), algorithm stops fast, however, as we will see shortly, the estimates acquired in this setup are much less accurate. We can see that significant fluctuations occur on the

d_{tv} paths in the case of small sample size.

Also, low probability of getting a sequence with a motif ($\alpha = 0.1$) results in much slower convergence, understandably so, since the motif model parameters constitute the vast majority of total parameters to estimate. The exception is, again the $k = 10$ case.

Distribution of the "evaluation" d_{tv}, which measures the actual efficiency of estimation, is visualised on the following boxplots.


Total variance distance, for each combination of parameters

The most obvious and unsurprising pattern is that large sample size k, and to a slightly lesser degree sequence length w lets us obtain more accurate estimates, both in terms of average of the results, as well as their dispersion.

An arguably more interesting observation is how the value of α affects estimators' efficiency - when w is relatively small, and so the proportion of the number of background distribution parameters to the background distribution is relatively large (4 to $4w$), coupled with high probability of "drawing" motif distribution, d_{tv} values are larger than in the case of moderate α. The algorithm gets relatively few observations to estimate a relatively large portion of parameters, hence the estimates for background distribution are significantly less accurate.

With other combinations, the higher the α value, the more accurate the estimates - the number of motif distribution parameters dwarfs the number of background distribution parameters, and the same mechanism applies.