Motif fiding in DNA sequences

Contents

1 Introduction

Genetics is a field that is related to many statistical problems. One of them is motif finding in DNA sequences. Motifs are sequences of nucleotides that are usually assumed to be related to biological function. We are interested in finding distributions of those nucleotides for each position in sequance, in form of a matrix, which is often called in biological sciences a "position weight matrix".

In our later considarations we assume two types of distributions: first for each position in sequance and second called "background" distribution which is the same for every position. Realistically we don't know which distribution is choosen for each position, but in our project we assumed the knowledge of probability of choosing distribution.

1.1 Data

We do not possess any real data, instead we generate it, so that we can test algorithms working in different environments. We are able to generate matrix $X_{k\times w}$, which k rows contain nucleotide sequence of length w. Each row is generated using one distribution, which is choosen with probability α. Data is generated according to parameters:

- $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_w)$ – matrix $4 \times w$ of distributions for each position, where $\boldsymbol{\theta}_i$ is a vector of length 4 and represents distribution of nucleotides on i-th position in sequence.
- $\boldsymbol{\theta}^b = \left(\theta_1^b, \ldots, \theta_4^b\right)^T$ – a vector of length 4 that represents "background" distribution.
- $\alpha = P(Z_i = 1) = 1 - P(Z_i = 0)$ – probability of choosing $\boldsymbol{\theta}$ $(Z_i = 1)$ or $\boldsymbol{\theta}^b$ $(Z_i = 0)$ distribution.

When given α space of all parameters (position weight matrix) has form:

$$\boldsymbol{\Theta} = \left(\boldsymbol{\theta}, \boldsymbol{\theta}^b\right)$$

Generating matrix $X_{k\times w}$ is done by "fliping independently k coins" $(Z_i, i = 1, \ldots, k)$ to choose distribution and then generate nucleotides using corresponding probabilities. Matrix X has form:

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1w} \\ \vdots & \ddots & \vdots \\ x_{k1} & \cdots & x_{kw} \end{pmatrix}$$

Where $x_{ij} \in \{A, C, G, T\}$ (or alternatively 1,2,3,4).

2 Expectation–maximization algorithm

EM algorithm is exactly suited for our purpose – to estimate parameters of position weight matrix given only α (probability of choosing $\boldsymbol{\theta}$ or $\boldsymbol{\theta}^b$ distribution). It's an iterative algorithm for fiding local maximum likelihood over the space of all parameters.

2.1 Intuitions

The basic setup is just like in our case – we have 2 different distributions that we can choose with probability α and let's say they are corresponding with two different coins (example from lecture) which have different probabilities of getting heads. We choose one coin and then count how many are there heads and tails (total of w measurements) and we repeat this k times.

Before we start algorithm, we need to initialize parameters. There are 2 steps in EM algorithm, first one being *expectation*. In this step we calculate probabilities of observations belonging to given coin (distribution) under our parameters $\Theta^{(t)}$ (in this iteration). The second step is *maximization*. In this step we find local maximum of likelihood function (in practice numericaly) and obtain new set of parameters $\Theta^{(t+1)}$ (in next iteration). We are repeating this process unitil some type of convergance, which will be discussed later (sec. 3.3).

2.2 Theretical calculations

We start with input matrix $X = (\mathbf{x}_1, \ldots, \mathbf{x}_w)$, where $\mathbf{x}_i$ are vetors of length k.

1: $t = 0$, initialize $\Theta^{(t)}$ (more about initializing methods later, sec. 3.2).

2: E(xpectation) step: computing $Q_i^{(t)}(j) = P(Z_i = j|\mathbf{x}_i, \Theta^{(t)})$:

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

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

3: M(aximization) step: computing $\mathcal{Q}(\Theta, \Theta^{(t)}) = \sum_{i=1}^{k}\sum_{j=1}^{2} Q_i^{(t)}(j)\log P(\mathbf{x}_i, Z_i = j, \Theta)$ and then finding $\Theta^{(t+1)} = \arg\max_{\Theta}\mathcal{Q}(\Theta, \Theta^{(t)})$

$$\theta_s^{b,(t+1)} = \frac{\sum_{i=1}^{k} Q_i^{(t)}(0)\#\{j|x_{ij} = s\}}{\lambda^b}, \quad \theta_{r,s}^{(t+1)} = \frac{\sum_{i=1}^{k} Q_i^{(t)}(1)\mathbb{1}\{x_{ir} = s\}}{\lambda_r}$$

Where

$$\lambda^b = w\sum_{i=1}^{k} Q_i^{(t)}(0), \quad \lambda_r = \sum_{i=1}^{k} Q_i^{(t)}(1)$$

4: if converged then stop, else go to step 2 (more on convergance later sec. 3.3).

3 Implementation

There are a few aspects of implementation that need to be mentioned.

3.1 Long sequences

When $w \to \infty$ products in function $Q_i^{(t)}(j)$ are aproaching 0. In theory it's never zero, but computationaly it's a problem for computer. In case of our studies, $w > 350$ produced dividing by 0. That could be further improved by implementing calculations on logarithms but it's not implemented in this work.

3.2 Initialization

Important part of EM algorithm is initialization of parameters. The EM algorithm finds *local* maxima so starting point is important and sometimes has tremendous influence on the results. It's because algorithm can end up in a maximum that is far away from global maximum. We will be testing two approaches: uniform initialization and initialization that takes information from input matrix.

3.2.1 Random initialization

Vectors $\boldsymbol{\theta}^{b,(0)}$ and $\boldsymbol{\theta}_i^{(0)}$ are generated using uniform distribution and then normalized, so that sum of its components is equal 1.

3.2.2 Empirical initialization

Vectors $\boldsymbol{\theta}^{b,(0)}$ and $\boldsymbol{\theta}_i^{(0)}$ are calculated based on input matrix X. Vector $\boldsymbol{\theta}^{b,(0)}$ is a vector of frequencies of all appearances of a letter divided by the number of all letters in matrix. Vector $\boldsymbol{\theta}_i^{(0)}$ is a vector of all appearances of a letter in given column divided by the number of all letters in that column.

$$\theta_i^{b,(0)} = \frac{\#\{x_{mj} = i\}}{wk}, \quad \theta_{ij}^{(0)} = \frac{\#\{x_{mj} = i\}}{k}$$

3.3 Convergance criterion

In the algorithm there is briefly mentioned "repeating" until convergance. There could be many ways of stoping iterating, but it was decided to use the sum of absolute values of differences of parameters between iterations. It means, that the algorithm will stop when parameters are not changing drastically. The criterion has form:

$$\sum_{i,j} |\Theta_{ij}^{(t)} - \Theta_{ij}^{(t+1)}| < \varepsilon$$

Where $\varepsilon = 0.01$. Decreasing ε has not proved any better – it caused only more iterations, to get similar result.

4 Experiments

Now we are able to conduct experiments and we will be interested in determining how "good" is EM algorithm under different circumstances. We are interested how is the algorithm behaving when we change the distributions, how the length and number of sequences influence results and if lower parameter α is causing any problems.

4.1 Evaluation

The "goodness" of the algorithm will be evaluated using total variation distance. For vectors $\mathbf{p}$ and $\mathbf{q}$ of length n is has form:

$$d_{tv}(\mathbf{p}, \mathbf{q}) = \frac{1}{2} \sum_{i=1}^{n} |p_i - q_i|$$

We will be applying it for original and estiimated vectors of parameters $\boldsymbol{\theta}^{b,orig}$, $\boldsymbol{\theta}^{orig}$ and $\boldsymbol{\theta}^{b,est}$, $\boldsymbol{\theta}^{est}$ and calculating their averages:

$$d_{final} = \frac{1}{w+1} \left[d_{tv}(\boldsymbol{\theta}^{b,orig}, \boldsymbol{\theta}^{b,est}) + \sum_{i=1}^{w} d_{tv}(\boldsymbol{\theta}_i^{orig}, \boldsymbol{\theta}_i^{est}) \right]$$

4.2 Distributions

It's important to test if changing the distribution will cause any difference. We will use 2 distributions: random (but normalized) and concentrated on randomly chosen letter. Vectors from Θ have form:

$$\boldsymbol{\theta}_i = (U_1, \ldots, U_4)^T / (U_1 + \ldots + U_4)$$

Where U_i are from uniform distribution. In first case they are from $U(0, 1)$ and in other case there is one of them from $U(0.8, 1)$ and others from $U(0, 0.2)$.

4.3 Results

Results are presented in tables 1.9, 1.12 and in plots 4.1. In this section are shown only a few examples and more can be found in appendix 1 and 2. Results contain the final d_{tv} for every combination of parameters w, k, α and in different starting conditions. Every result is computed only once, so there is a possibility (especially in low w and k setup) for EM algorithm to stay in wrong minimum (more on that in sec. 5).

4.3.1 Tables

Tables 1.9, 1.12 present d_{tv} of every combination w (rows) and k (columns).

w/k	10	50	70	100	150	200	500	1000
5	0.424	0.113	0.122	0.135	0.075	0.062	0.055	0.034
10	0.338	0.230	0.136	0.092	0.073	0.070	0.042	0.028
15	0.257	0.155	0.099	0.094	0.084	0.075	0.039	0.027
20	0.241	0.120	0.122	0.094	0.082	0.055	0.041	0.029
25	0.309	0.143	0.109	0.085	0.067	0.070	0.045	0.025
30	0.299	0.121	0.129	0.088	0.067	0.069	0.042	0.030
35	0.224	0.154	0.099	0.099	0.084	0.059	0.040	0.029

Table 4.1: d_{tv} for $\alpha = 0.5$, random distribution random initialization.

w/k	10	50	70	100	150	200	500	1000
5	0.187	0.082	0.065	0.051	0.053	0.042	0.023	0.017
10	0.194	0.067	0.076	0.052	0.043	0.041	0.035	0.016
15	0.415	0.098	0.076	0.057	0.055	0.048	0.027	0.022
20	0.492	0.097	0.093	0.078	0.046	0.045	0.031	0.019
25	0.196	0.095	0.080	0.059	0.053	0.035	0.027	0.020
30	0.256	0.093	0.070	0.066	0.046	0.043	0.031	0.020
35	0.278	0.082	0.076	0.073	0.057	0.043	0.031	0.018

Table 4.2: d_{tv} for $\alpha = 0.5$, concentrated distribution, empirical initialization.

4.3.2 Plots

Plots 4.1 present d_{tv} vs k for fixed w and $\alpha = 0.1, 0.3, 0.5, 0.7, 0.9$ and every combination of starting conditions.

5 Conclusions

The first and the most obvious conclusion – EM algorithm works increasingly better with more observations (k). This holds almost everywhere. Also, longer sequences (w) work out better, but it has any impact only in case of lower k. It is because we do not have a lot of data to recognize what distribution was used, it is especially problematic with lower α. When $k = 1000$ all results are almost equaly good.

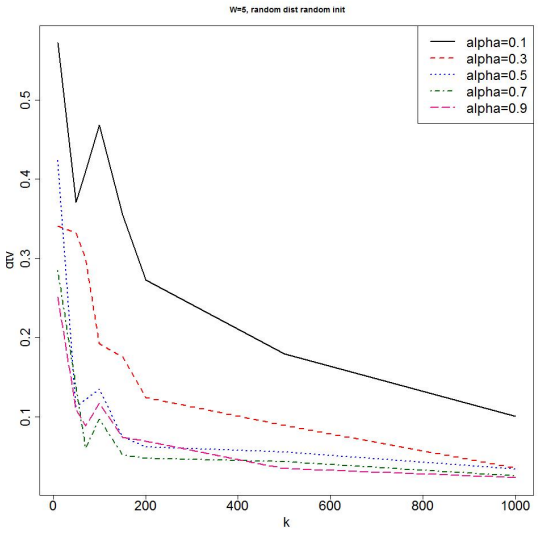
The second conclusion is, that EM alogrithm has a lot of problems dealing with low α. It means that the background distribution is chosen more frequently – it is estimated more precisly, but at cost of the other distribution, that has always more parameters to estimate. With growing α estimations are getting better, because the rest of the parameters are estimated better.

The third conclusion is that EM algorithm is sensitive to starting conditions. The distributions and initial parameters have significant impact on results. In some cases empirical initialization improves results, but it decreases d_tv with lower α. Empirical initialization estimates background distribution precisly, because there are more positions drawn from this distribution. It hinders the EM algorithm – it's often further from the true parameters. Similar thing happens when distributions are concentrated on some letter and we initialize with random parameters.

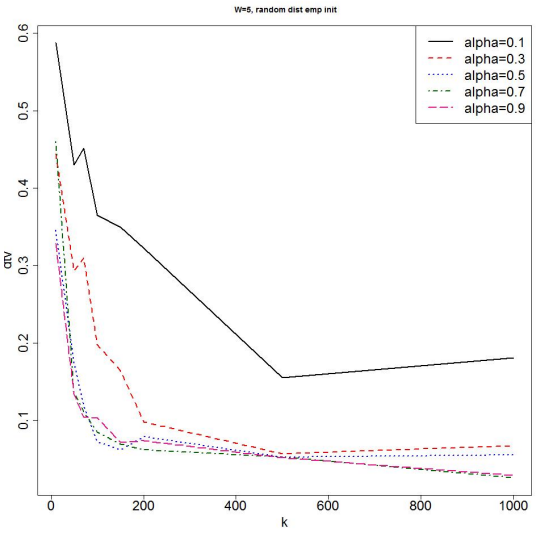
The highest d_{tv} is always when the sequence is short and we have only a few observations, in case of $\alpha = 0.01$ it reaches around 0.6. The lowest d_{tv} is 0.13 when $\alpha = 0.9$ (many parameters of other distribution are estimated better). To reduce error in low dimensional cases, we could repeat n times EM algorithm and average the parameters (it makes sense only in case of random initialization), eliminating the influence of rare wrong minimum.

The most important thing to be concluded is this – to use EM algorithm at its best, it is advised to use appropriate initialization. With knowledge of low α it not worth using empirical initialization. Without any knowledge of α it is safer to use random initialization.

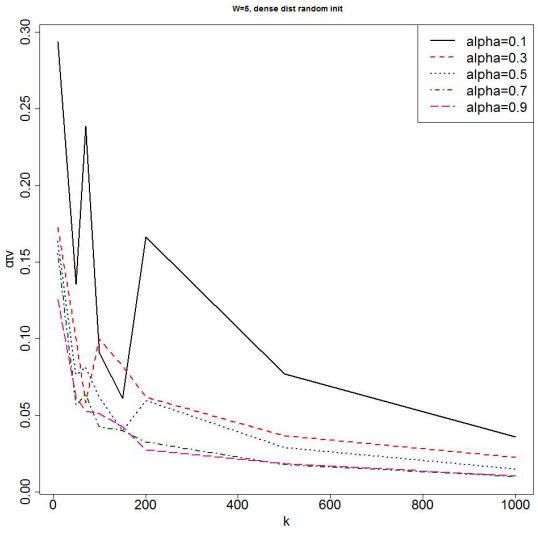
Overall, the EM algorithm proves to be a really useful tool in this kind of problem. It is working fast and when there is a lot of data, it's also precise.



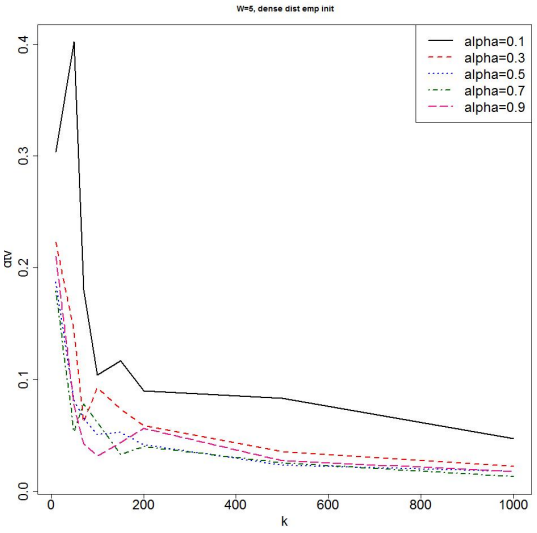
(a) Random distribution, random initialization.



(b) Random distribution, empirical initialization.



(c) Concentrated distribution, random initialization.



(d) Concentrated distribution, empirical initialization.

Figure 4.1: Plots of d_{tv} vs k for $w = 5$ in different types of distribution and initialization.

List of Tables

List of Figures

Appendices

Appendix 1 Tables

Random and concentrated (dense) distributions, random and empirical (emp) initializations.

1.1 $\alpha = 0.1$

	10	50	70	100	150	200	500	1000
5	0.572	0.371	0.410	0.468	0.355	0.272	0.179	0.101
10	0.461	0.426	0.400	0.431	0.396	0.403	0.172	0.085
15	0.409	0.401	0.362	0.352	0.303	0.354	0.160	0.087
20	0.418	0.418	0.329	0.245	0.209	0.199	0.134	0.077
25	0.283	0.216	0.326	0.284	0.230	0.117	0.100	0.080
30	0.372	0.357	0.302	0.246	0.161	0.176	0.105	0.063
35	0.287	0.310	0.260	0.295	0.343	0.116	0.104	0.069

Table 1.1: d_{tv} for $\alpha = 0.1$, random dist random init.

	10	50	70	100	150	200	500	1000
5	0.588	0.430	0.451	0.365	0.349	0.322	0.155	0.181
10	0.585	0.362	0.414	0.391	0.294	0.325	0.178	0.102
15	0.454	0.380	0.400	0.356	0.328	0.337	0.134	0.076
20	0.350	0.261	0.366	0.283	0.203	0.247	0.093	0.080
25	0.319	0.305	0.317	0.212	0.197	0.353	0.126	0.069
30	0.415	0.285	0.275	0.286	0.189	0.156	0.114	0.079
35	0.316	0.285	0.249	0.248	0.281	0.213	0.095	0.075

Table 1.2: d_{tv} for $\alpha = 0.1$, random dist emp init.

	10	50	70	100	150	200	500	1000
5	0.294	0.136	0.239	0.091	0.061	0.166	0.077	0.036
10	0.302	0.209	0.209	0.111	0.083	0.087	0.052	0.031
15	0.490	0.260	0.137	0.147	0.106	0.097	0.072	0.051
20	0.252	0.211	0.166	0.126	0.122	0.117	0.078	0.041
25	0.370	0.219	0.174	0.162	0.107	0.101	0.067	0.045
30	0.422	0.200	0.671	0.141	0.102	0.108	0.072	0.046
35	0.375	0.158	0.173	0.154	0.092	0.113	0.060	0.046

Table 1.3: d_{tv} for $\alpha = 0.1$, dense dist random init.

	10	50	70	100	150	200	500	1000
5	0.304	0.402	0.180	0.104	0.116	0.090	0.083	0.047
10	0.411	0.173	0.148	0.218	0.114	0.129	0.068	0.036
15	0.447	0.154	0.257	0.134	0.114	0.093	0.085	0.044
20	0.261	0.241	0.122	0.195	0.097	0.112	0.060	0.039
25	0.343	0.173	0.170	0.164	0.121	0.119	0.064	0.054
30	0.399	0.245	0.149	0.188	0.104	0.095	0.067	0.041
35	0.621	0.150	0.121	0.112	0.156	0.123	0.064	0.046

Table 1.4: d_{tv} for $\alpha = 0.1$, dense dist emp init.

1.2 $\alpha = 0.3$

	10	50	70	100	150	200	500	1000
5	0.341	0.332	0.299	0.192	0.176	0.124	0.089	0.035
10	0.394	0.204	0.196	0.189	0.128	0.103	0.052	0.039
15	0.439	0.218	0.131	0.106	0.091	0.086	0.056	0.033
20	0.266	0.129	0.135	0.126	0.092	0.079	0.049	0.047
25	0.291	0.268	0.139	0.110	0.099	0.107	0.050	0.042
30	0.292	0.198	0.152	0.108	0.088	0.076	0.047	0.032
35	0.260	0.170	0.145	0.262	0.080	0.087	0.049	0.038

Table 1.5: d_{tv} for $\alpha = 0.3$, random dist random init.

	10	50	70	100	150	200	500	1000
5	0.444	0.293	0.310	0.198	0.164	0.098	0.057	0.067
10	0.302	0.214	0.261	0.150	0.121	0.109	0.057	0.038
15	0.348	0.385	0.162	0.171	0.118	0.078	0.055	0.043
20	0.267	0.203	0.173	0.163	0.126	0.088	0.066	0.043
25	0.270	0.130	0.194	0.123	0.098	0.089	0.053	0.040
30	0.280	0.201	0.137	0.149	0.130	0.082	0.058	0.038
35	0.271	0.158	0.175	0.142	0.116	0.102	0.055	0.035

Table 1.6: d_{tv} for $\alpha = 0.3$, random dist emp init.

	10	50	70	100	150	200	500	1000
5	0.172	0.100	0.058	0.099	0.082	0.062	0.037	0.023
10	0.188	0.095	0.110	0.066	0.073	0.049	0.032	0.019
15	0.202	0.094	0.069	0.085	0.056	0.065	0.038	0.025
20	0.680	0.124	0.092	0.081	0.078	0.058	0.039	0.027
25	0.356	0.122	0.075	0.083	0.065	0.055	0.038	0.028
30	0.191	0.146	0.093	0.078	0.067	0.053	0.039	0.027
35	0.213	0.124	0.108	0.107	0.061	0.052	0.032	0.027

Table 1.7: d_{tv} for $\alpha = 0.3$, dense dist random init.

	10	50	70	100	150	200	500	1000
5	0.223	0.144	0.062	0.093	0.073	0.059	0.035	0.022
10	0.428	0.112	0.090	0.088	0.061	0.037	0.030	0.025
15	0.210	0.128	0.089	0.086	0.064	0.046	0.034	0.021
20	0.251	0.112	0.096	0.087	0.082	0.065	0.034	0.031
25	0.276	0.131	0.118	0.088	0.076	0.055	0.040	0.025
30	0.266	0.105	0.121	0.083	0.068	0.054	0.040	0.023
35	0.302	0.119	0.100	0.083	0.060	0.055	0.038	0.022

Table 1.8: d_{tv} for $\alpha = 0.3$, dense dist emp init.

1.3 $\alpha = 0.5$

w/k	10	50	70	100	150	200	500	1000
5	0.424	0.113	0.122	0.135	0.075	0.062	0.055	0.034
10	0.338	0.230	0.136	0.092	0.073	0.070	0.042	0.028
15	0.257	0.155	0.099	0.094	0.084	0.075	0.039	0.027
20	0.241	0.120	0.122	0.094	0.082	0.055	0.041	0.029
25	0.309	0.143	0.109	0.085	0.067	0.070	0.045	0.025
30	0.299	0.121	0.129	0.088	0.067	0.069	0.042	0.030
35	0.224	0.154	0.099	0.099	0.084	0.059	0.040	0.029

Table 1.9: d_{tv} for $\alpha = 0.5$, random distribution random initialization.

	10	50	70	100	150	200	500	1000
5	0.345	0.176	0.120	0.072	0.062	0.079	0.053	0.055
10	0.339	0.121	0.151	0.081	0.075	0.060	0.045	0.032
15	0.311	0.150	0.131	0.097	0.085	0.076	0.048	0.032
20	0.212	0.141	0.165	0.106	0.067	0.062	0.048	0.027
25	0.237	0.143	0.120	0.101	0.074	0.067	0.038	0.028
30	0.235	0.162	0.098	0.103	0.071	0.057	0.042	0.030
35	0.223	0.128	0.111	0.088	0.077	0.082	0.050	0.029

Table 1.10: d_{tv} for $\alpha = 0.5$, random dist emp init.

	10	50	70	100	150	200	500	1000
5	0.164	0.076	0.082	0.062	0.040	0.060	0.029	0.015
10	0.135	0.084	0.090	0.048	0.039	0.044	0.021	0.020
15	0.194	0.082	0.092	0.068	0.052	0.052	0.028	0.022
20	0.218	0.112	0.090	0.064	0.060	0.040	0.032	0.018
25	0.229	0.089	0.068	0.062	0.051	0.048	0.032	0.025
30	0.210	0.095	0.080	0.063	0.064	0.050	0.026	0.020
35	0.197	0.079	0.078	0.069	0.048	0.053	0.030	0.026

Table 1.11: d_{tv} for $\alpha = 0.5$, dense dist random init.

w/k	10	50	70	100	150	200	500	1000
5	0.187	0.082	0.065	0.051	0.053	0.042	0.023	0.017
10	0.194	0.067	0.076	0.052	0.043	0.041	0.035	0.016
15	0.415	0.098	0.076	0.057	0.055	0.048	0.027	0.022
20	0.492	0.097	0.093	0.078	0.046	0.045	0.031	0.019
25	0.196	0.095	0.080	0.059	0.053	0.035	0.027	0.020
30	0.256	0.093	0.070	0.066	0.046	0.043	0.031	0.020
35	0.278	0.082	0.076	0.073	0.057	0.043	0.031	0.018

Table 1.12: d_{tv} for $\alpha = 0.5$, concentrated distribution, empirical initialization.

1.4 $\alpha = 0.7$

	10	50	70	100	150	200	500	1000
5	0.285	0.138	0.061	0.097	0.052	0.048	0.043	0.025
10	0.267	0.108	0.100	0.107	0.058	0.046	0.025	0.027
15	0.235	0.111	0.080	0.086	0.057	0.050	0.028	0.026
20	0.252	0.107	0.085	0.065	0.068	0.058	0.029	0.027
25	0.181	0.113	0.101	0.086	0.051	0.059	0.036	0.021
30	0.187	0.089	0.088	0.074	0.062	0.061	0.037	0.022
35	0.345	0.115	0.092	0.074	0.062	0.054	0.030	0.023

Table 1.13: d_{tv} for $\alpha = 0.7$, random dist random init.

	10	50	70	100	150	200	500	1000
5	0.460	0.137	0.111	0.085	0.069	0.062	0.052	0.026
10	0.309	0.115	0.090	0.099	0.059	0.064	0.036	0.024
15	0.243	0.120	0.099	0.067	0.066	0.051	0.038	0.026
20	0.265	0.138	0.092	0.076	0.075	0.067	0.034	0.026
25	0.214	0.112	0.092	0.082	0.053	0.049	0.032	0.022
30	0.228	0.107	0.094	0.069	0.063	0.063	0.032	0.023
35	0.209	0.115	0.084	0.078	0.067	0.062	0.035	0.026

Table 1.14: d_{tv} for $\alpha = 0.7$, random dist emp init.

	10	50	70	100	150	200	500	1000
5	0.155	0.057	0.063	0.042	0.040	0.033	0.018	0.010
10	0.117	0.059	0.068	0.042	0.036	0.037	0.024	0.020
15	0.167	0.070	0.057	0.041	0.044	0.041	0.019	0.021
20	0.153	0.071	0.061	0.060	0.045	0.032	0.026	0.016
25	0.158	0.067	0.066	0.054	0.045	0.034	0.024	0.019
30	0.258	0.080	0.073	0.049	0.037	0.040	0.024	0.018
35	0.158	0.087	0.069	0.052	0.043	0.037	0.024	0.019

Table 1.15: d_{tv} for $\alpha = 0.7$, dense dist random init.

	10	50	70	100	150	200	500	1000
5	0.179	0.053	0.078	0.062	0.033	0.040	0.025	0.013
10	0.166	0.059	0.066	0.040	0.040	0.049	0.031	0.020
15	0.219	0.063	0.057	0.051	0.036	0.042	0.017	0.013
20	0.200	0.070	0.074	0.061	0.047	0.043	0.026	0.015
25	0.193	0.063	0.069	0.058	0.045	0.035	0.025	0.016
30	0.160	0.076	0.069	0.055	0.043	0.042	0.023	0.018
35	0.170	0.087	0.068	0.054	0.040	0.034	0.025	0.016

Table 1.16: d_{tv} for $\alpha = 0.7$, dense dist emp init.

1.5 $\alpha = 0.9$

	10	50	70	100	150	200	500	1000
5	0.251	0.110	0.088	0.117	0.074	0.070	0.035	0.023
10	0.199	0.100	0.073	0.075	0.055	0.058	0.030	0.024
15	0.212	0.122	0.085	0.066	0.053	0.048	0.027	0.020
20	0.238	0.097	0.068	0.074	0.046	0.041	0.028	0.020
25	0.167	0.091	0.085	0.069	0.063	0.044	0.031	0.017
30	0.201	0.092	0.086	0.069	0.055	0.045	0.031	0.020
35	0.216	0.080	0.076	0.068	0.058	0.047	0.032	0.021

Table 1.17: d_{tv} for $\alpha = 0.9$, random dist random init.

	10	50	70	100	150	200	500	1000
5	0.328	0.134	0.105	0.103	0.072	0.074	0.052	0.029
10	0.207	0.104	0.083	0.079	0.059	0.041	0.038	0.022
15	0.215	0.102	0.068	0.080	0.039	0.048	0.021	0.022
20	0.238	0.102	0.074	0.080	0.059	0.048	0.031	0.017
25	0.171	0.099	0.097	0.063	0.059	0.053	0.030	0.025
30	0.187	0.104	0.084	0.065	0.051	0.046	0.032	0.021
35	0.231	0.083	0.081	0.071	0.064	0.050	0.030	0.019

Table 1.18: d_{tv} for $\alpha = 0.9$, random dist emp init.

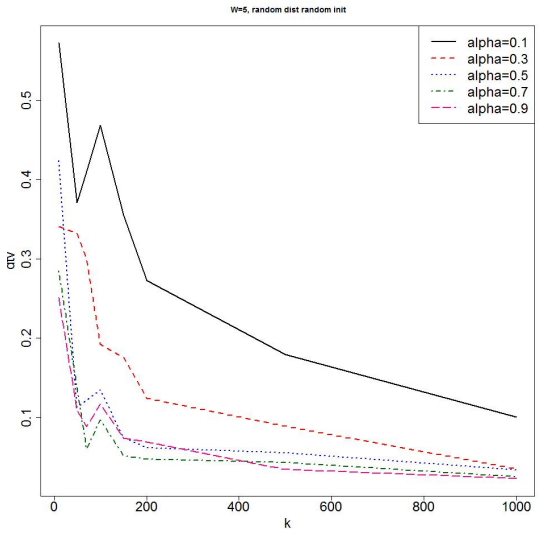
	10	50	70	100	150	200	500	1000
5	0.126	0.062	0.053	0.051	0.043	0.027	0.018	0.010
10	0.138	0.054	0.041	0.054	0.034	0.037	0.020	0.015
15	0.205	0.070	0.058	0.044	0.039	0.028	0.021	0.016
20	0.167	0.077	0.054	0.047	0.040	0.034	0.020	0.013
25	0.160	0.067	0.057	0.047	0.040	0.036	0.020	0.015
30	0.165	0.062	0.056	0.048	0.036	0.035	0.024	0.015
35	0.164	0.077	0.062	0.045	0.043	0.039	0.021	0.013

Table 1.19: d_{tv} for $\alpha = 0.9$, dense dist random init.

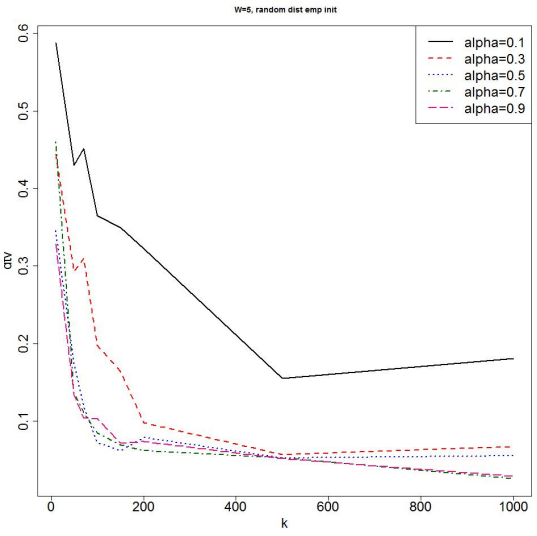
	10	50	70	100	150	200	500	1000
5	0.210	0.077	0.042	0.031	0.043	0.056	0.027	0.018
10	0.159	0.083	0.057	0.041	0.039	0.035	0.018	0.017
15	0.173	0.077	0.061	0.050	0.032	0.032	0.024	0.017
20	0.153	0.071	0.051	0.051	0.035	0.035	0.023	0.016
25	0.192	0.080	0.066	0.052	0.034	0.034	0.022	0.014
30	0.165	0.083	0.054	0.052	0.033	0.036	0.023	0.013
35	0.156	0.071	0.062	0.049	0.041	0.036	0.021	0.015

Table 1.20: d_{tv} for $\alpha = 0.9$, dense dist emp init.

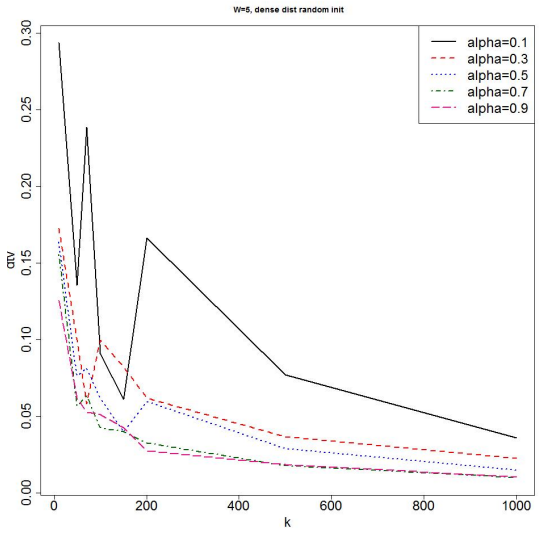
Appendix 2 Plots



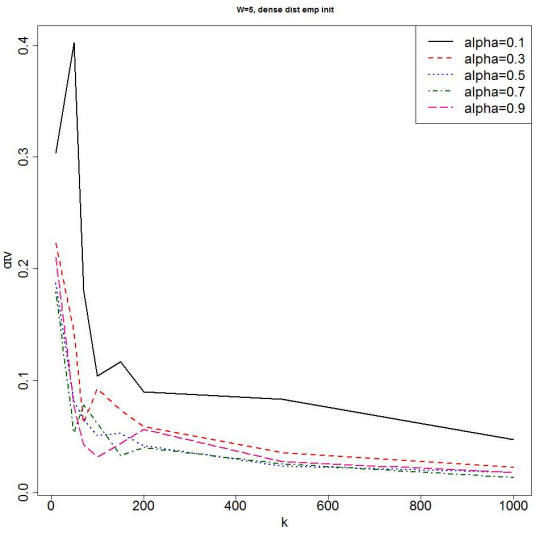
(a) Random distribution, random initialization.



(b) Random distribution, empirical initialization.

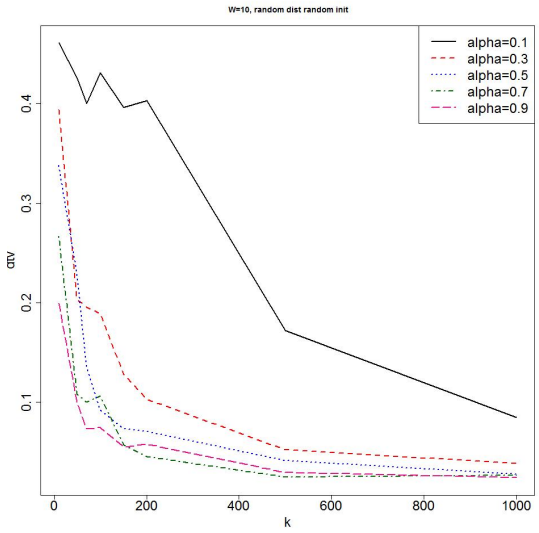


(c) Concentrated distribution, random initialization.

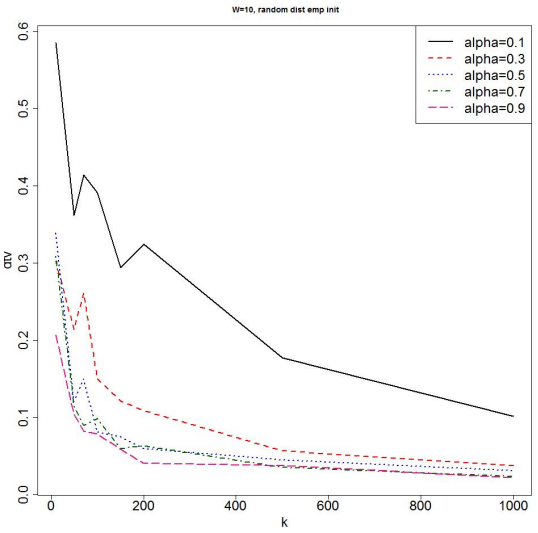


(d) Concentrated distribution, empirical initialization.

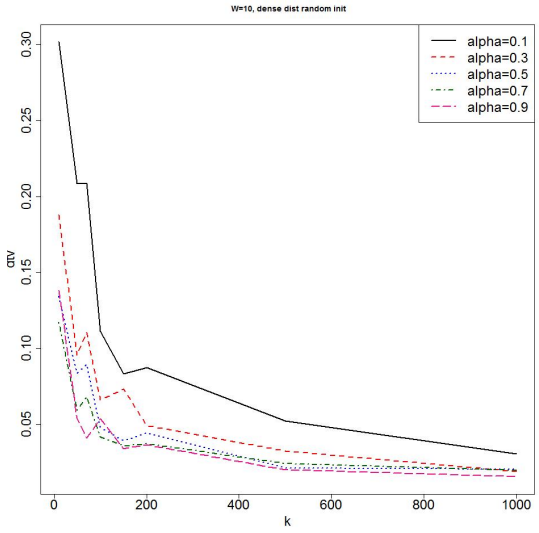
Figure 2.1: Plots of d_{tv} vs k for $w = 5$ in different types of distribution and initialization.



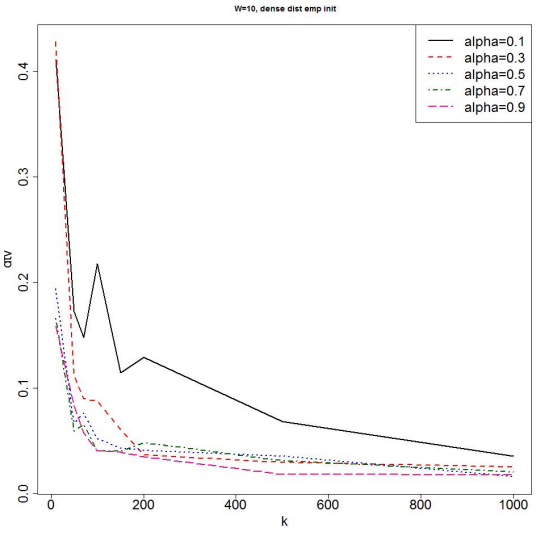
(a) Random distribution, random initialization.



(b) Random distribution, empirical initialization.

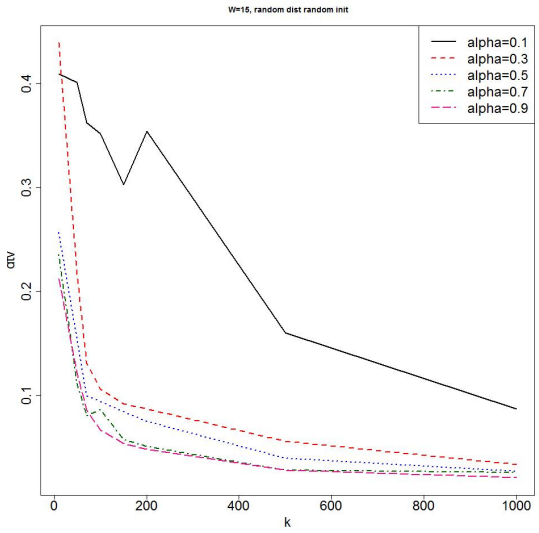


(c) Concentrated distribution, random initialization.

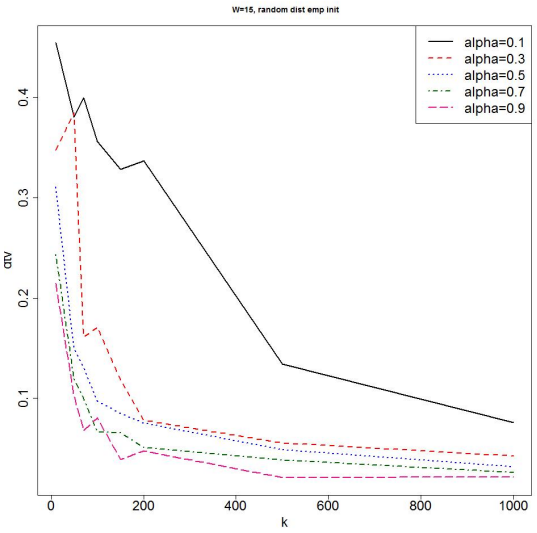


(d) Concentrated distribution, empirical initialization.

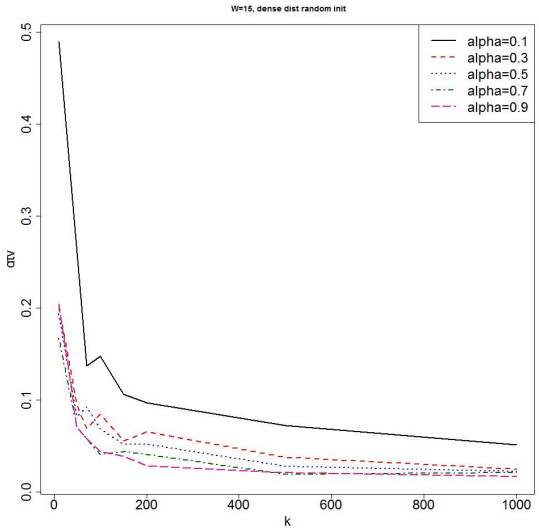
Figure 2.2: Plots of d_{tv} vs k for $w = 10$ in different types of distribution and initialization.



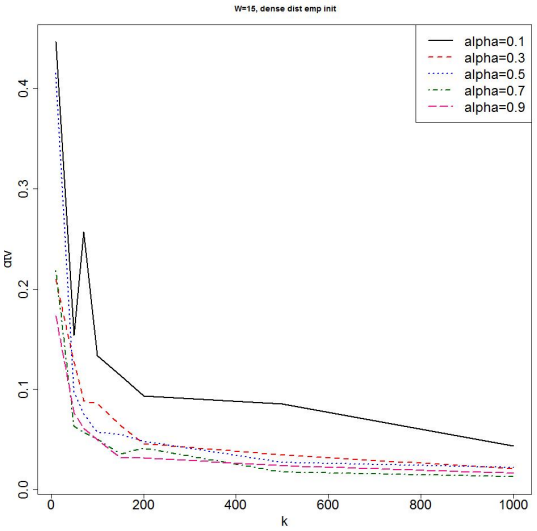
(a) Random distribution, random initialization.



(b) Random distribution, empirical initialization.

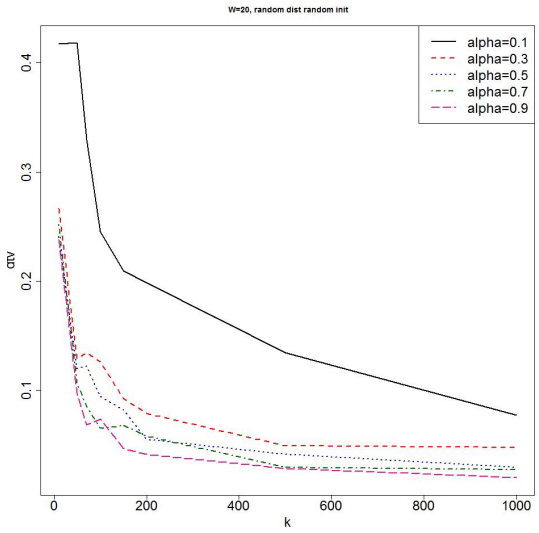


(c) Concentrated distribution, random initialization.

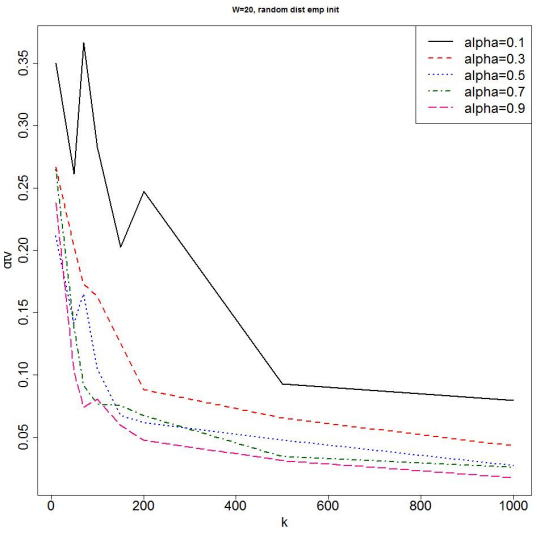


(d) Concentrated distribution, empirical initialization.

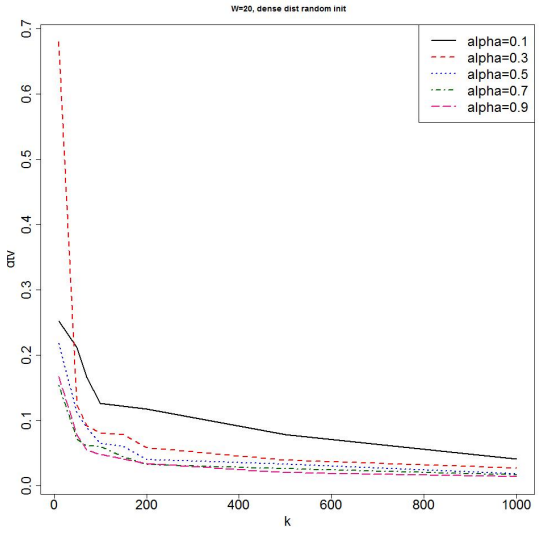
Figure 2.3: Plots of d_{tv} vs k for $w = 15$ in different types of distribution and initialization.



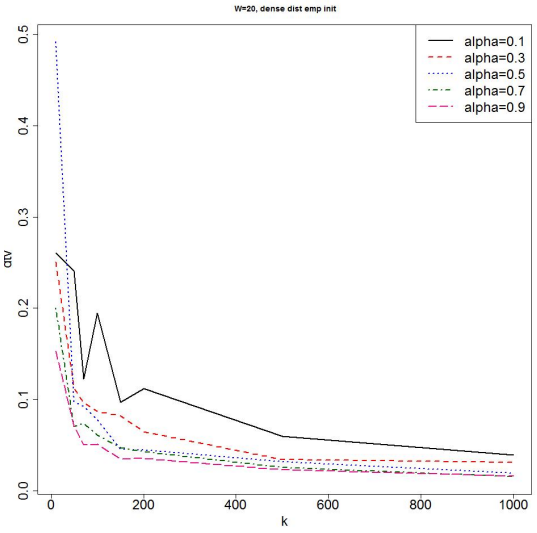
(a) Random distribution, random initialization.



(b) Random distribution, empirical initialization.

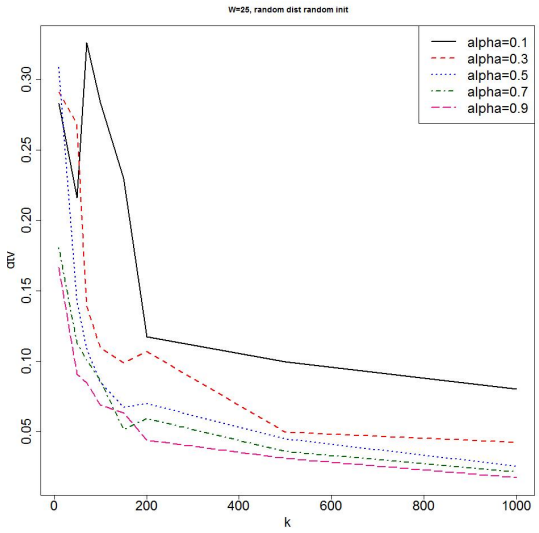


(c) Concentrated distribution, random initialization.

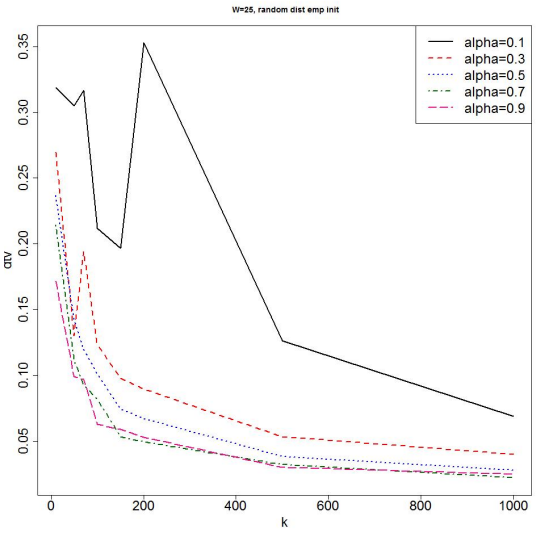


(d) Concentrated distribution, empirical initialization.

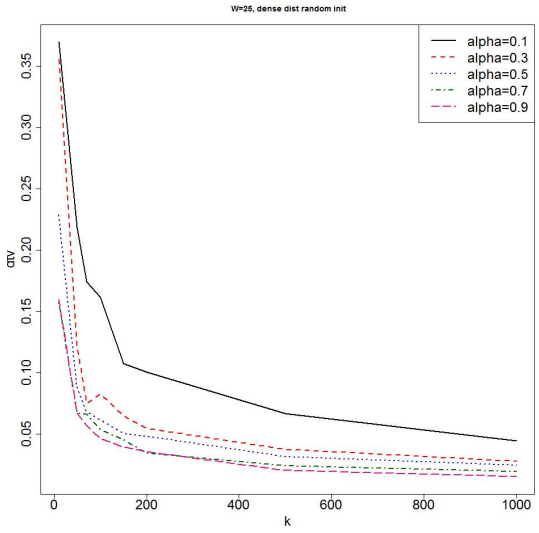
Figure 2.4: Plots of d_{tv} vs k for $w = 20$ in different types of distribution and initialization.



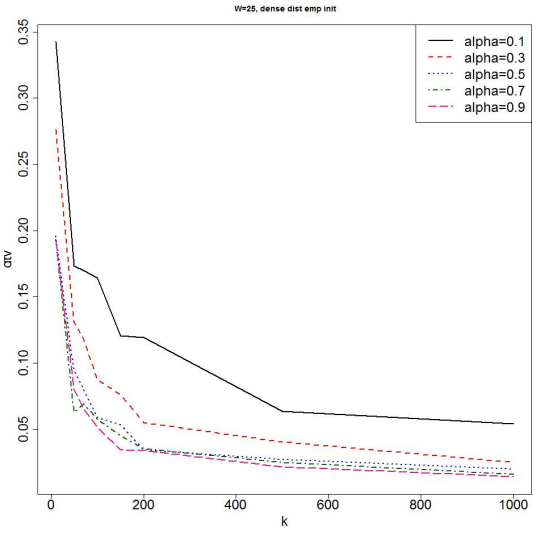
(a) Random distribution, random initialization.



(b) Random distribution, empirical initialization.

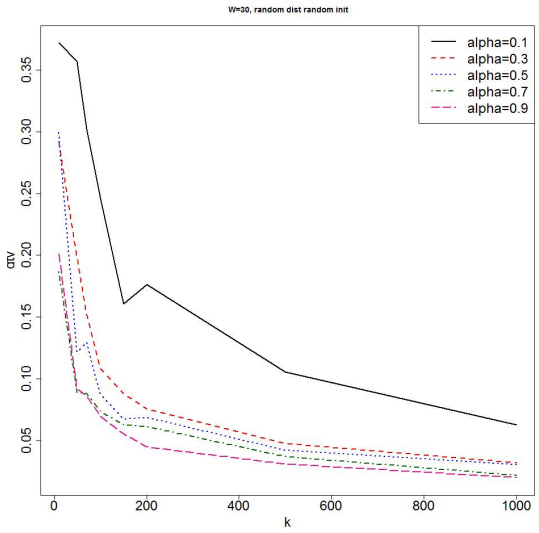


(c) Concentrated distribution, random initialization.

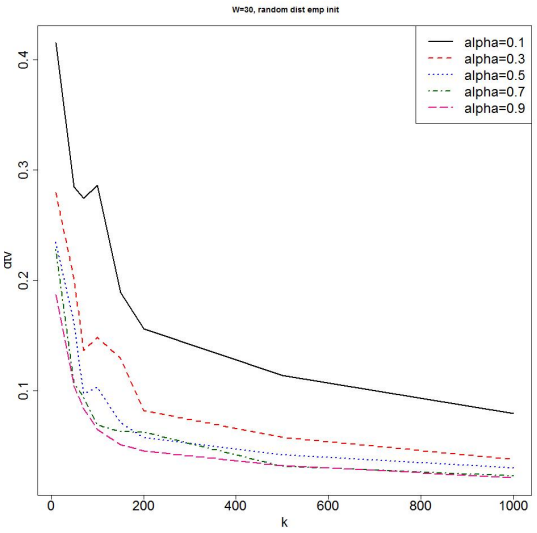


(d) Concentrated distribution, empirical initialization.

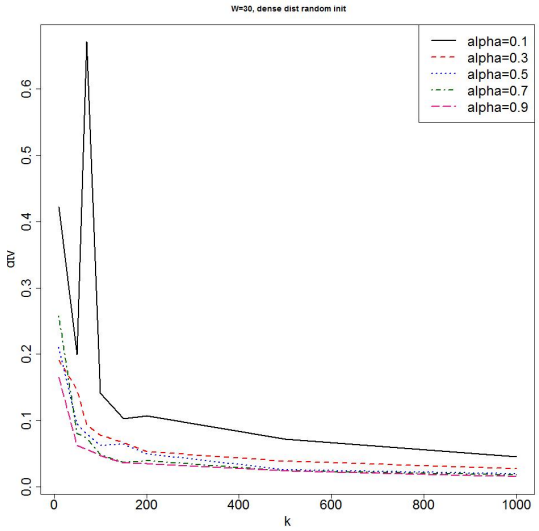
Figure 2.5: Plots of d_{tv} vs k for $w = 25$ in different types of distribution and initialization.



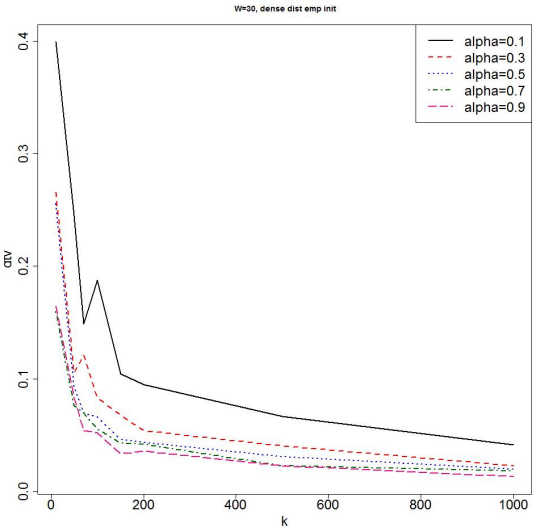
(a) Random distribution, random initialization.



(b) Random distribution, empirical initialization.

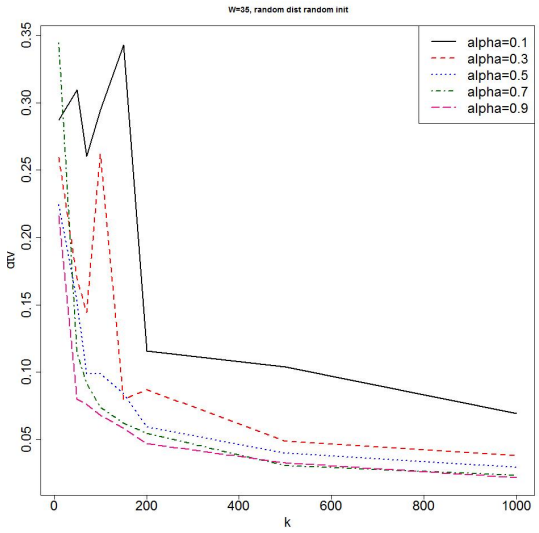


(c) Concentrated distribution, random initialization.

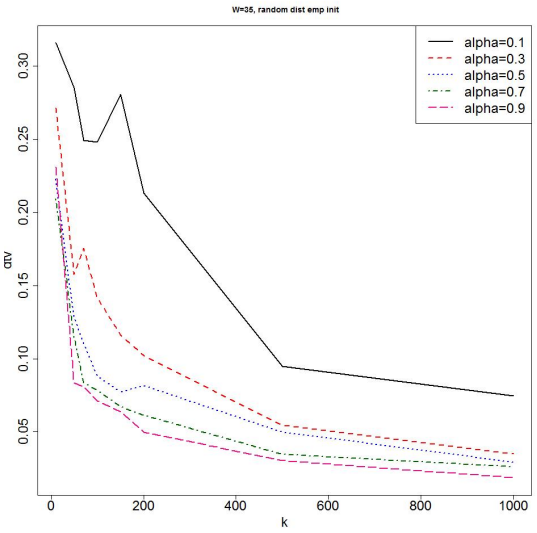


(d) Concentrated distribution, empirical initialization.

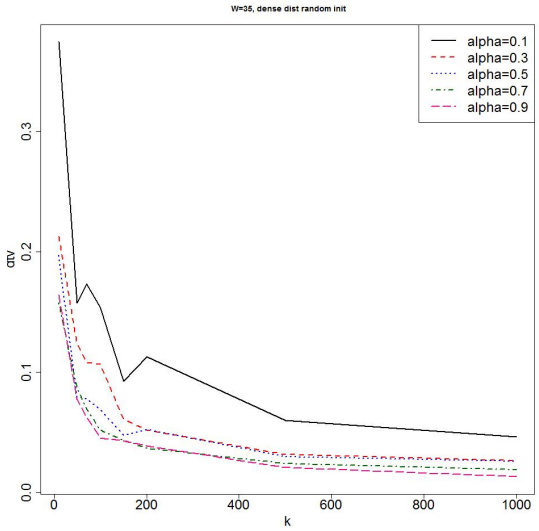
Figure 2.6: Plots of d_{tv} vs k for $w = 30$ in different types of distribution and initialization.



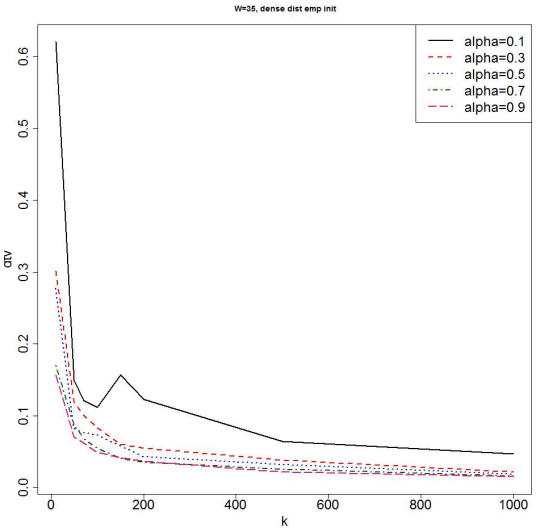
(a) Random distribution, random initialization.



(b) Random distribution, empirical initialization.



(c) Concentrated distribution, random initialization.



(d) Concentrated distribution, empirical initialization.

Figure 2.7: Plots of d_{tv} vs k for $w = 35$ in different types of distribution and initialization.