Supplement to "Automated Clustering of High-dimensional Data with a Feature Weighted Mean Shift Algorithm"

Saptarshi Chakraborty*[1], Debolina Paul*[2], and Swagatam Das†[3]

[1]Department of Statistics, University of California, Berkeley
[2]Indian Statistical Institute, Kolkata
[3]Electronics and Communication Sciences Unit, Indian Statistical Institute, Kolkata

1 Proof of Theorem from Section 2

Theorem 1. *Let $\boldsymbol{w}^*$ be the minimizer of* (5)*, subject to $\boldsymbol{w}^\top \mathbf{1} = 1$. Then,*

$$w_l^* = \frac{\exp\{-\frac{1}{n\lambda}\sum_{i=1}^n (x_{il} - y_{il}^{(t)})^2\}}{\sum_{l'=1}^p \exp\{-\frac{1}{n\lambda}\sum_{i=1}^n (x_{il'} - y_{il'}^{(t)})^2\}}.$$

Proof. Problem (5) can be rewritten as follows:

$$\min_{\boldsymbol{w}} \sum_{l=1}^p D_l w_l + \lambda \sum_{l=1}^p w_l \log w_l$$

where $D_l = \frac{1}{n}\sum_{i=1}^n (x_{il} - y_{il}^{(t)})^2$.

This problem can now be solved using the constraints $\boldsymbol{w}^\top \mathbf{1} = 1$, $w_l \geq 0\ \forall l = 1, \ldots, p$. The Lagrangian is thus given by,

$$L = \sum_{l=1}^p D_l w_l + \lambda \sum_{l=1}^p w_l \log w_l - \alpha\left(\sum_{l=1}^p w_l - 1\right)$$

Putting $\frac{\partial L}{\partial w_l} = 0$, we have,

$$D_l + \lambda(1 + \log w_l) - \alpha = 0$$

Thus, we have $w_l \propto \exp(-\frac{D_l}{\lambda})$.

Using the condition $\boldsymbol{w}^\top \mathbf{1} = 1$, we have,

$$\begin{aligned} w_l^* &= \frac{\exp(-\frac{D_l}{\lambda})}{\sum_{l'=1}^p \exp(-\frac{D_l'}{\lambda})} \\ &= \frac{\exp\{-\frac{1}{n\lambda}\sum_{i=1}^n (x_{il} - y_{il}^{(t)})^2\}}{\sum_{l'=1}^p \exp\{-\frac{1}{n\lambda}\sum_{i=1}^n (x_{il'} - y_{il'}^{(t)})^2\}}. \end{aligned}$$

This solution to the relaxed problem also satisfies the non-negativity condition. Hence the result. □

2 Proofs of Results from Section 3

Theorem 2. *Let $C_t = \mathcal{C}(\{\boldsymbol{y}_1^{(t)}, \ldots, \boldsymbol{y}_n^{(t)}\})$. Then $\{C_t\}_{t=0}^\infty$ constitutes a decreasing sequence of sets, i.e.*

$$C_0 \supseteq C_1 \supseteq \ldots C_t \supseteq C_{t+1} \supseteq \ldots$$

Proof. From equation (3), we note that $\boldsymbol{y}_i^{(t+1)}$ can be represented as a convex combination of $\boldsymbol{y}_1^{(t)}, \ldots, \boldsymbol{y}_n^{(t)}$. This implies that $\boldsymbol{y}_i^{(t+1)} \in C_t$, for all $i = 1, \ldots, n$. Thus C_t is a convex set, which contains $\{\boldsymbol{y}_1^{(t+1)}, \ldots, \boldsymbol{y}_n^{(t+1)}\}$. Since C_{t+1} is the smallest convex set containing $\{\boldsymbol{y}_1^{(t+1)}, \ldots, \boldsymbol{y}_n^{(t+1)}\}$, $C_t \supseteq C_{t+1}$. □

Corollary 1. $\lim_{t\to\infty} C_t$ *exists and is given by,* $\lim_{t\to\infty} C_t = \cap_{t=1}^\infty C_t$.

Proof. $\{C_t\}_{t=1}^\infty$ constitutes a decreasing sequence of sets. The result directly follows from applying monotone convergence theorem for sets (Rudin, 1964). □

*The first authors contributed equally.
†Correspondence to swagatam.das@isical.ac.in

From the above analysis, we observe that the convex hull C_t of $\{\boldsymbol{y}_1^{(t)}, \ldots, \boldsymbol{y}_n^{(t)}\}$ converges as $t \to \infty$.

Theorem 3. *For any pre-fixed tolerance level δ, there exists $T \in \mathbb{N}$ such that*

$$\big|\max_{i,j} \|y_i^{(t+1)} - y_j^{(t+1)}\|_2 - \max_{i,j} \|y_i^{(t)} - y_j^{(t)}\|_2\big| < \delta,$$

for all $t \geq T$.

Proof. From Corollary 1, $\{C_t\}_{t=1}^{\infty}$. This implies that the vertices of C_t converges. Let $a_t = \max_{i,j} \|y_i^{(t)} - y_j^{(t)}\|_2$. Since a_t is a function of the vertices of C_t, $\{a_t\}_{t=1}^{\infty}$ also converges. Since $\{a_t\}_{t=1}^{\infty}$ is a convergent sequence of reals, $\{a_t\}_{t=1}^{\infty}$ is Cauchy (Rudin, 1964). Thus for any $\delta > 0$, there exists $T \in \mathbb{N}$ such that $\big|\max_{i,j} \|y_i^{(t+1)} - y_j^{(t+1)}\|_2 - \max_{i,j} \|y_i^{(t)} - y_j^{(t)}\|_2\big| < \delta$, for all $t \geq T$. □

Theorem 4. *Let $\boldsymbol{x}_0 \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$ and $\boldsymbol{x}_{t+1} = \int \boldsymbol{y} p_t(\boldsymbol{y}|\boldsymbol{x}_t) d\boldsymbol{y}$. Here $p_t(\cdot)$ denotes the distribution of $\boldsymbol{x}_t$ and $p_t(\boldsymbol{y}|\boldsymbol{z}) = \phi(\boldsymbol{z} - \boldsymbol{y}; 0, \frac{h}{2} diag(1/\boldsymbol{w}^{(t)})) q_t(\boldsymbol{y}) / p_t(\boldsymbol{z})$. Then, $\boldsymbol{x}_t \sim \mathcal{N}_p(\mathbf{0}, diag((s_1^{(t)})^2, \ldots, (s_p^{(t)})^2))$, with*

$$s_l^{(t+1)} = (1 + h(s_l^{(t)})^2 / 2w_l^{(t)})^{-1} s_l^{(t)}. \tag{1}$$

Proof. We first find the distribution of $\boldsymbol{x}_t$.

$$p_1(\boldsymbol{y}|\boldsymbol{x}_0) = \phi(\boldsymbol{x}_0 - \boldsymbol{y}; 0, \frac{h}{2} diag(1/\boldsymbol{w}^{(0)})) q_0(\boldsymbol{y}) / p_t(\boldsymbol{x}_0)$$

$$\propto \phi(\boldsymbol{x}_0 - \boldsymbol{y}; 0, \frac{h}{2} diag(1/\boldsymbol{w}^{(0)})) q_0(\boldsymbol{y})$$

$$\propto \exp\{-\frac{1}{h} \sum_{l=1}^{p} w_l^{(0)} (y_l - x_l^{(0)})^2\} \exp\{-\sum_{l=1}^{p} y_l^2 / (2\sigma_l^2)\}$$

$$\propto \exp\Big\{ -\frac{1}{2} \sum_{l=1}^{p} \frac{\Big(y_l - \frac{2w_l^{(0)} x_l^{(0)}/h}{2w_l^{(0)}/h + 1/\sigma_l^2}\Big)^2}{\frac{1}{(2w_l^{(0)}/h + 1/\sigma_l^2)}} \Big\}.$$

Thus, $\boldsymbol{x}_1 = E(\boldsymbol{y}|\boldsymbol{x}_0) = \big(\frac{2w_1^{(0)} x_1^{(0)}/h}{2w_1^{(0)}/h + 1/\sigma_1^2}, \ldots, \frac{2w_p^{(0)} x_p^{(0)}/h}{2w_p^{(0)}/h + 1/\sigma_p^2}\big)$. Thus $\boldsymbol{x}_1$ is also a Gaussian distribution, with mean $\mathbf{0}$ and dispersion matrix as $\text{Var}(\boldsymbol{x}_1) = diag((s_1^{(1)})^2, \ldots, (s_p^{(1)})^2)$, with

$$s_l^{(1)} = \frac{2w_l^{(0)} s_0/h}{2w_l^{(0)}/h + 1/\sigma_l^2} = \frac{1}{1 + \frac{h}{2w_l^{(0)}} (s_l^{(0)})^2} s_l^{(0)}.$$

Here $s_l^{(0)} = \sigma_l$. By an inductive argument, it is easy to see that, $\boldsymbol{x}_{t+1}$ is also a Gaussian distribution, with mean $\mathbf{0}$ and dispersion matrix as $\text{Var}(\boldsymbol{x}_{t+1}) = diag((s_1^{(t+1)})^2, \ldots, (s_p^{(t+1)})^2)$, with $s_l^{(t+1)} = \big(1 + \frac{h}{2w_l^{(t)}} (s_l^{(t)})^2\big)^{-1} s_l^{(t)}$. □

Theorem 5. *Let $\boldsymbol{x}_t$, $s_l^{(t)}$ be as in Theorem 4. Then, $s_l = \lim_{t\to\infty} s_l^{(t)} = 0$ and the order of convergence of $\{s_l^{(t)}\}_{t=1}^{\infty}$ is at least cubic. Moreover, the asymptotic rate of convergence of $\{s_l^{(t)}\}$ is $\frac{2w_l}{h}$.*

Proof. We first consider the case $\lim_{t\to\infty} w_l^{(t)} = 0$. From equation (1), we observe that $\lim_{t\to\infty} s_l^{(t+1)} = 0$. Now if $\lim_{t\to\infty} w_l^{(t)} > 0$, we take limit as $t \to \infty$ on both sides of equation (1) and get,

$$s_l = \frac{1}{1 + \frac{h}{2w_l} (s_l)^2} s_l \implies s_l = 0, \, \forall \, l = 1, \ldots, p.$$

We will now find the order of convergence of s_l to zero. We say that the order of convergence is m if m is the largest integer such that

$$r_s = \lim_{t\to\infty} \frac{|s_l^{(t+1)} - s_l|}{|s_l^{(t)} - s_l|^m} = \lim_{t\to\infty} \frac{(s_l^{(t)})^{3-m}}{(s_l^{(t)})^2 + \frac{h}{2w_l}} < \infty.$$

This occurs at least for the case $m = 3$, i.e. the convergence rate of $\{s^{(t)}\}_{t=1}^{\infty}$ is at least cubic. If $m = 3$, the asymptotic rate of convergence is given by, $r_s = \frac{2w_l}{h}$. □

3 Experimental Studies

3.1 Data Dimensions

3.2 Rankings for the Experiment on Real-life datasets

This is an extension of the experiment done in Section 4.3 of the main paper. The experimental procedure, datasets, and the results are the same as those provided earlier. Table 2 additionally indicates the individual rankings as an explanation of the average ranks provided in the original draft. It can be easily observed from Table 2 that among all the peers, WBMS surely proves itself to be the best performing algorithm in terms of both the NMI and ARI values.

Table 1: Performance Analysis on Real Life Datasets in terms of NMI & ARI values

Datasets	Method	k-Means	G-Means	WG-Means	DP-Means	EWDP	RCC	MS	BMS	WBMS
GLIOMA	NMI	0.499(8)	0.522(6)	0.517(7)	0.576(4)	0.675(2)	0.113(9)	0.580(3)	0.546(5)	**0.706**(1)
	ARI	0.328(8)	0.367(7)	0.373(6)	0.416(4)	0.598(2)	0.004(9)	0.429(3)	0.398(5)	**0.618**(1)
Appendicitis	NMI	0.157(7)	0.165(6)	0.185(5)	0.158(8)	0.189(4)	0.193(3)	0.008(9)	0.195(2)	**0.249**(1)
	ARI	0.230(2)	0.169(7)	0.188(5)	0.231(3)	0.188(6)	0.000(9)	0.014(8)	0.223(4)	**0.434**(1)
Zoo	NMI	0.741(6)	0.801(4)	0.749(5)	0.611(8)	0.859(2)	0.557(9)	0.706(7)	0.841(3)	**0.925**(1)
	ARI	0.459(6)	0.646(4)	0.559(5)	0.452(7)	0.872(2)	0.039(9)	0.436(8)	0.867(3)	**0.953**(1)
Mammo	NMI	0.231(4)	0.181(7)	0.240(3)	0.191(5)	0.273(2)	0.181(6)	0.181(8)	0.008(9)	**0.348**(1)
-graphic	ARI	0.292(3)	0.209(6)	0.293(2)	0.257(4)	0.247(5)	0.001(8)	0.150(7)	0.001(9)	**0.351**(1)
Yale	NMI	0.539(2)	0.417(5)	0.521(4)	0.324(7)	0.334(6)	0.268(8)	0.548(3)	0.222(8)	**0.693**(1)
	ARI	0.271(2)	0.173(4)	0.227(3)	0.040(7)	0.061(6)	0.031(8)	0.069(5)	0.027(9)	**0.485**(1)
nci9	NMI	0.458(2)	0.394(4.5)	0.394(4.5)	0.181(8)	0.211(7)	0.143(9)	0.346(6)	0.394(3)	**0.686**(1)
	ARI	0.187(2)	0.100(3.5)	0.100(3.5)	0.005(9)	0.018(6)	0.005(8)	0.011(7)	0.086(5)	**0.419**(1)
Lymphoma	NMI	0.441(7)	0.592(5)	0.690(2)	0.518(6)	0.648(3)	0.243(8)	0.241(9)	0.595(4)	**0.778**(1)
	ARI	0.269(6)	0.340(5)	0.463(2)	0.088(7)	0.431(4)	0.001(8)	0.000(9)	0.458(3)	**0.604**(1)
Movement	NMI	0.591(3)	0.231(9)	0.328(7)	0.333(6)	0.465(5)	0.639(2)	0.245(8)	0.503(4)	**0.663**(1)
Libras	ARI	0.309(2)	0.068(8)	0.123(5)	0.113(6)	0.217(3)	0.014(9)	0.090(7)	0.185(4)	**0.532**(1)
GCM	NMI	0.532(4)	0.484(7)	0.497(5.5)	0.025(9)	0.497(5.5)	0.637(3)	0.456(8)	0.649(2)	**0.833**(1)
	ARI	0.288(5)	0.248(8)	0.266(6.5)	0.002(9)	0.266(6.5)	0.561(2)	0.413(4)	0.536(3)	**0.714**(1)
Average	NMI	4.94	5.81	4.69	6.43	3.89	6.75	6.63	4.75	**1**
Rank	ARI	3.94	5.56	3.94	5.81	4.3	8.5	6.75	5.25	**1**

3.3 Optimal Parameter values for Real-Life Datasets

In this section, we provide the optimal parameter values for real-life datasets. The experimental results for WBMS that are provided in Table 2 are based on the optimal parameter values provided in Table 4.

4 Ablation Studies

4.1 Abalation Study on h

In this section, we conduct an ablation study to assess the performance of the proposed WBMS algorithm for different values of h. We take $h = 0.1, 0.5, 0.8$ and 1 for our experimental study. The algorithm is run according to the procedure described in Section 4. As a cluster validity index, we use the ARI values between the ground truth and the obtained partition. We choose λ corresponding to the highest ARI value. Table 5 gives the ARI values for all the real-life datasets used in our experiments for different values of h.

4.2 Abalation Study on λ

For each of the optimal h, described in the previous section, we assess the performance of WBMS for different values of λ. Since, $\lambda \in [1, 20]$, we take $\lambda = 1, 5, 10$ and 20. For each different value of λ, we run the algorithm similar to the procedure described in Section 4. We choose h corresponding to the highest ARI value. For ties, any value of h between the ties can be taken as optimal. Table 6 summarizes the ARI values for all real-life datasets used in our experiments for different values of λ.

Sensitivity Analysis We also conducted a sensitivity analysis on a toy example with two clusters, each containing 100 points. The first two features of the dataset fully contains the cluster structure of the data, while the rest 30 are simulated independently from a standard normal distribution. For different values of h and λ, we run the WBMS and compute the NMI values between the ground truth and the obtained partition. In Fig. 1, we show the heatmap of NMI values, plotted against different values of h and λ. The hetmap clearly shows that the WBMS is able to recover the perfect clustering (colored in yellow) over a considerable range of the hyper-parameter values.

Table 2: Runtime Analysis on Real Life Datasets (Time in Seconds)

Datasets	k-Means	G-Means	WG-Means	DP-Means	EWDP	RCC	MS	BMS	WBMS
GLIOMA	2.19	21.57	30.45	17.48	17.73	23.60	26.96	20.97	20.18
Appendicitis	0.56	3.50	3.42	5.64	4.96	4.65	2.13	1.36	1.17
Zoo	0.89	2.81	6.06	2.70	3.36	4.29	2.02	1.36	1.29
Mammographic	2.78	5.27	5.78	4.29	5.61	52.69	103.98	81.04	86.21
Yale	8.45	16.06	28.63	25.67	27.18	37.50	41.92	189.98	209.18
nci9	9.08	17.25	26.07	29.19	28.96	63.79	68.65	53.48	50.95
Lymphoma	1.31	25.01	34.81	33.42	35.70	64.34	56.47	47.80	42.16
Movement Libras	1.98	17.41	21.97	18.09	20.69	43.28	38.27	22.80	24.17
GCM	12.81	80.67	113.10	153.74	147.32	90.56	308.09	340.61	305.76

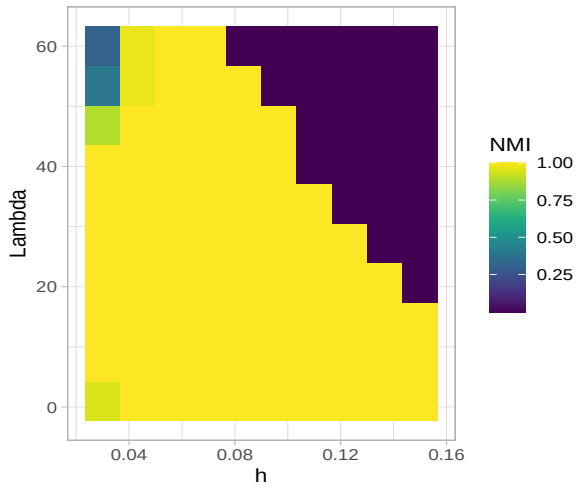

Figure 1: Heatmap of NMI values, plotted against different values of h and λ, showing that the WBMS is able to recover the perfect clustering (colored in yellow) over a considerable range of the hyper-parameter values.

Datasets	# Datapoints	# Features	# Clusters
GLIOMA	50	4434	4
Appendicitis	107	7	2
Zoo	101	16	7
Mammographic	830	5	2
Yale	165	1024	15
nci9	60	9712	9
Lymphoma	62	4026	3
Movement Libras	360	90	15
GCM	191	16063	15

Table 3: The dimensions of various real data, used in our experiments along with the true number of clusters.

5 Additional Experiments

References

Rudin, W. (1964). *Principles of mathematical analysis*, volume 3. McGraw-hill New York.

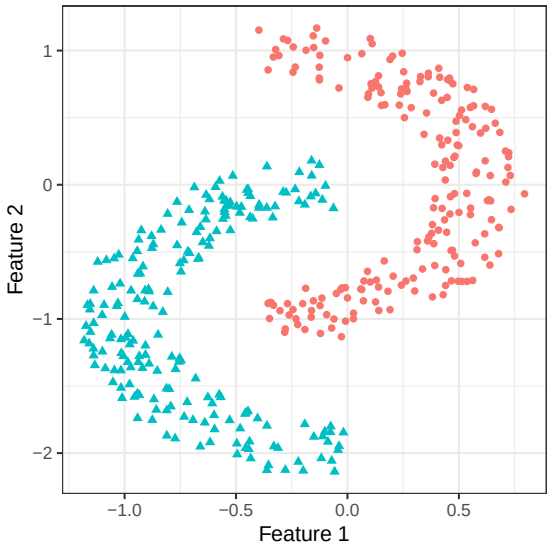

Figure 2: Experimental Results on Twomoons Data

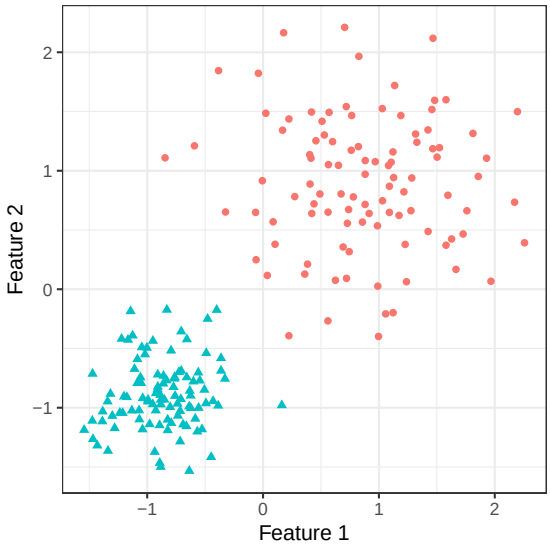

Figure 3: Experimental Results on synthetic data with clusters having different densities and volumes.

Datasets	h	λ
GLIOMA	0.5	1
Appendicitis	1	10
Zoo	0.8	20
Mammographic	0.8	10
Yale	0.1	10
nci9	0.1	5
Lymphoma	0.5	5
Movement Libras	0.1	10

Table 4: Optimal Parameter Values for Real-Life Datasets

Datasets	$h = 0.1$	$h = 0.5$	$h = 0.8$	$h = 1$
GLIOMA	0.587	0.618	0.618	0.524
Appendicitis	0.218	0.324	0.401	0.434
Zoo	0.953	0.912	0.754	0.754
Mammographic	0.212	0.295	0.351	0.320
Yale	0.485	0.414	0.234	0.001
nci9	0.419	0.419	0.340	0.108
Lymphoma	0.589	0.604	0.544	0.330
Movement Libras	0.532	0.414	0.021	0.001
GCM	0.714	0.531	0.531	0.562

Table 5: Average ARI values for WBMS on real-datasets for different values of h

Datasets	$\lambda = 1$	$\lambda = 5$	$\lambda = 10$	$\lambda = 20$
GLIOMA	0.618	0.618	0.618	0.510
Appendicitis	0.257	0.352	0.434	0.398
Zoo	0.717	0.717	0.906	0.953
Mammographic	0.198	0.202	0.351	0.351
Yale	0.055	0.414	0.485	0.423
nci9	0.210	0.419	0.419	0.378
Lymphoma	0.167	0.604	0.503	0.503
Movement Libras	0.251	0.376	0.532	0.532
GCM	0.510	0.714	0.714	0.714

Table 6: Average ARI values for WBMS on real-datasets for different values of λ