Estimation of Cluster Centroids in Presence of Noisy Observations

Khaled Kamal Saab
Department of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia, USA
khaled.saab.95@gmail.com

***Abstract*—This paper considers the problem of clustering vector-valued datasets whose replicate observations are contaminated by weighted additive zero-mean white measurement noise. A corresponding error model of the cluster centroid is developed. Subsequently, an optimal iterative algorithm is proposed for updating cluster centroids obtained by using the k-means algorithm implemented on each set of noisy observations. The gain of the proposed algorithm aims for per-iteration minimization of the mean square estimate error. Three other methods are considered for performance evaluation. A numerical toy example is presented in order to illustrate the performance capabilities of the proposed method.**

Keywords—machine learning; clustering; noise, k-means algorithm; stochastic optimization

I. INTRODUCTION

Problems in clustering algorithms can arise when on top of some cluster structure or groups of similar objects the data also contains an unstructured subset of points or outliers. This unstructured subset is referred to as noise or background noise in literature of machine learning (e.g., see, [1]-[6]). Such noise issues tend to disrupt the recovery of the cluster structure and robustness examination of clustering algorithm in presence of unstructured subset is anticipated (e.g., see, [2]-[5]).

Another class of problems associated with clustering algorithms arises when the datasets are contaminated by noise due to experimental random measurement errors such as in biological processes (e.g., see, [7]-[13]). As illustrated in [13], substantial noisy measurements could lead to misinterpretations of the relationships between members of different clusters. In order to reduce the effect of noise several replicate measurements of datasets are needed [7] and [11]. For example, the study in [11] shows that 10 to15 replicates yield stable results when considering gene expression microarray experiments. A study on the relationship between experimental replication and clustering precision is presented in [10]. One approach for handling noisy data is based on weighted averaging of the noisy data using analytical distribution of the data [8]-[9]. Another approach applies clustering algorithm on the collected data obtained from all the replicates expanded [12]. The most common approach for clustering experimental data is achieved by clustering the averages of replicate measurements [13]. However, the latter may not be applicable when the standard deviation of the noise is large with respect to the norm between neighboring objects.

This paper considers the case where covariance norm of the measurement errors may vary from replicate to replicate. In order to motivate this situation, we consider a measuring device installed on a vehicle where the vehicle is in motion and where the measurement error increases as the distance between the measuring device and targeted objects increases. To the knowledge of the author, no such problem has been studied in the relevant literature. We consider vector-valued datasets whose replicate observations are contaminated by weighted additive zero-mean white measurement noise. The latter is inspired by the error model considered in [14]. The problem under consideration is the estimation of clusters centroids using all the noisy replicate measurements of datasets. We develop a corresponding error model of cluster centroids and propose an optimal iterative algorithm for updating the cluster centroids. We consider the centroids obtained by using the k-means algorithm implemented on each set of noisy observations as the input to our proposed algorithm. The gain of the proposed algorithm is obtained by minimizing the trace of estimate error covariance matrix. Three other methods are considered. One method extract the cluster centroids from observations of datasets associated with smallest covariance norm of the measurement errors, another method resembles the one in [12], and one method averages all cluster centroids obtained using each replicate. We compare the performance of the proposed method with the other three methods under consideration on a two-dimensional toy example.

The rest of the paper is organized as follows. Section II formulates the problem under consideration and presents the proposed algorithm and describes three other methods followed by an example in Section III. The paper ends with conclusions.

II. MAIN RESULTS

This section formulates the problem under consideration and presents the proposed iterative algorithm and three other methods for estimating the cluster centroids in presence of replicate noisy observations.

A. Problem formulation

Let $Z = \{z_1, z_2, \dots, z_N\}$ be a dataset of precise observations where $z_n \in \mathbb{R}^q, 1 \le n \le N$. Let $C = \{C_1, C_2, \dots, C_k\}$ be a set of clusters with $C_j \subset Z$. The cluster centroids, $c_j \in \mathbb{R}^q$, for $1 < j \le k$ are obtained such that

$$\text{argmin}_C \sum_{j=1}^{k} \sum_{z \in C_j} \|z - c_j\|^2 \qquad (1)$$

where $\|.\|$ is the Euclidean distance and c_j is the cluster centroid or center of cluster C_j, which is the average of points in C_j; that is, $c_j = \frac{1}{N_j}\sum_{i=1}^{N_j} z_{j_i}$ where N_j is the number of associated points in C_j and $z_{j_i} \in C_j$. The k-means is one of the simplest learning algorithms that solve this well-known clustering problem (1).

In this work we consider a class of noisy M replicates of the dataset. In particular, we assume that the erroneous observation at replicate, m, is given by

$$\hat{z}_n(m) = z_n + g(m)v_n, \quad 1 \le n \le N, 1 \le m \le M \qquad (2)$$

where $v_n \in \mathbb{R}^q$ is a zero-mean white Gaussian random process, and $g(m) \in \mathbb{R}^{q\times q}$ is a known deterministic function. Therefore, the covariance of the observation error, $\epsilon_z(m) \triangleq z_n - \hat{z}_n(m)$, is $\mathrm{E}[\epsilon_z(m)\epsilon_z^T(m)] = g(m)Rg^T(m)$, where $\mathrm{E}[.]$ is the expectation operator and $R \triangleq \mathrm{E}[v_n v_n^T]$.

It is important to note that when applying a clustering algorithm (e.g., k-means) on the m^{th} noisy replicate (2), then we obtain cluster $\hat{C}_j(m)$ with number of corresponding points $\hat{N}_j(m)$ and centroid $\hat{c}_j(m)$, which are likely different than C_j, N_j and c_j, respectively.

B. Proposed Algorithm

The problem addresses estimation of cluster centroids using noisy M replicates of observation (2). For each noisy set of replicates $\{\hat{z}_1(m), \hat{z}_2(m), \dots, \hat{z}_N(m)\}$ application of the k-means algorithm results in clusters $\hat{C}_j(m)$ with number of points $\hat{N}_j(m)$, and cluster centroid $\hat{c}_j(m)$, where $\hat{c}_j(m) = \frac{1}{\hat{N}_j(m)}\sum_{i=1}^{\hat{N}_j(m)} \hat{z}_{j_i}(m)$. By making use of (2), we obtain

$$\hat{c}_j(m) = \frac{1}{\hat{N}_j(m)}\sum_{i=1}^{\hat{N}_j(m)} (z_{j_i} + g(m)v_{j_i}) \qquad (3)$$

where $z_{j_i} \in \hat{C}_j(m)$. The optimal j^{th} cluster centroid, in the sense of (1), corresponds to $c_j = \frac{1}{N_j}\sum_{l=1}^{N_j} z_{j_l}$ at steady-state of sequential k-means algorithm. Consequently, we consider $\hat{c}_j(m)$ as a measurement to c_j with measurement error given by

$$\epsilon_j(m) \triangleq c_j - \hat{c}_j(m) = \epsilon_\xi(m) - \frac{1}{\hat{N}_j(m)}\sum_{i=1}^{\hat{N}_j(m)} g(m)v_{j_i} \qquad (4)$$

where $\epsilon_\xi(m) \triangleq \frac{1}{N_j}\sum_{l=1}^{N_j} z_{j_l} - \frac{1}{\hat{N}_j(m)}\sum_{i=1}^{\hat{N}_j(m)} z_{j_i}$.

We propose the following iterative algorithm for updating the j^{th} cluster centroid

$$\bar{c}_j(m+1) = \bar{c}_j(m) - K(m)(\bar{c}_j(m) - \hat{c}_j(m)) \qquad (5)$$

where $\bar{c}_j(1) = \hat{c}_j(1)$ and $K(m) \in \mathbb{R}^{q\times q}$ is a learning gain.

Remark 1: It is important to note that $\bar{c}_j(m+1)$ in (5) is not meant to predict the estimate of c_j during the next $(m+1)^{th}$ observation but instead it is meant to estimate c_j while using $\hat{c}_j(m)$. The proposed update (5) resembles the a posteriori state estimate update of a Kalman filter. However, the problem under consideration does not deal with state prediction and the model under consideration does not involve time or a difference/differential equation. In what follows we derive the optimal gain $K(m)$ by minimizing the mean square estimate error, which also resembles the Kalman filter approach. ■

Define $\delta c_j(m) \triangleq c_j - \bar{c}_j(m)$ to be the update error of the algorithm (5). Inserting (4) in (5) and subtracting both sides from c_j yields

$$\delta c_j(m+1) = \delta c_j(m) - K_j(m)(\delta c_j(m) - \epsilon_j(m)) \qquad (6)$$

Since v_n is assumed to be a zero-mean white random process and $\epsilon_j(m) \approx \frac{1}{\hat{N}_j(m)}\sum_{i=1}^{\hat{N}_j(m)} g(m)v_{j_i}$, then $\mathrm{E}[\epsilon_j(m)] \cong 0$, using an induction argument $\mathrm{E}[\delta c_j(m)] = 0$ and for $m > 2$, $\mathrm{E}[\delta c_j(m)\epsilon_j^T(m)] = 0$. Arranging two terms in (6) leads to

$$\delta c_j(m+1) = \big(I - K_j(m)\big)\,\delta c_j(m) + K_j(m)\epsilon_j(m) \qquad (7)$$

Consequently, (7) leads to the following update in the estimate error covariance matrix, $P_j(m) \triangleq \mathrm{E}[\delta c_j(m)\delta c_j^T(m)]$,

$$P_j(m+1) = \big(I - K_j(m)\big)P_j(m)\big(I - K_j(m)\big)^T + K_j(m)R_j(m)K_j^T(m) \qquad (8)$$

where $R_j(m) \triangleq \mathrm{E}[\epsilon_j(m)\epsilon_j^T(m)]$.

Theorem 1. Consider the iterative algorithm proposed in (5). The gain $K_j(m)$ that minimizes the mean-square of $\delta c_j(m+1)$ at each m^{th} observation is given in the following recursive algorithm for all $m > 1$,

$$K_j(m) = P_j(m)(P_j(m) + R_j(m))^{-1} \qquad (9)$$

$$P_j(m+1) = (I - K_j(m))P_j(m) \qquad (10)$$

where $P_j(1)$ is a symmetric positive-definite matrix.

Proof of Theorem 1. Expanding (8) yields

$$P_j(m+1) = P_j(m) + K_j(m)(P_j(m)+R_j(m))K_j^T(m) - K_j(m)P_j(m) - P_j(m)K_j^T(m) \qquad (11)$$

To minimize $\mathrm{tr}(P_j(m+1))$, where $\mathrm{tr}(.)$ is the trace operator, with respect to $K_j(m)$, we set $\frac{\partial \mathrm{tr}(P_j(m+1))}{\partial K_j(m)} \equiv 0$ at each m, $\frac{\partial \mathrm{tr}(P_j(m+1))}{\partial K_j(m)} = 2K_j(m)(P_j(m) + R_j(m)) - 2P_j(m) \equiv 0$.
Therefore, $K_j(m) = P_j(m)(P_j(m) + R_j(m))^{-1}$. Inserting this optimal value of $K_j(m)$ in (11) and collecting terms lead to

$$P_j(m+1) = P_j(m) + P_j(m)(P_j(m) + R_j(m))^{-1}(P_j(m)+R_j(m))K_j^T(m) - K_j(m)P_j(m) - P_j(m)K_j^T(m)$$

Cancelling then collecting terms leads to (10). ■

Remark 2: The recursive algorithm in (9) and (10) require the knowledge of $P_j(1)$ and $R_j(m)$. If cluster $\hat{C}_j(m)$ is the same as C_j, then $\epsilon_\xi(m)$ (4) becomes zero. If we neglect $\epsilon_\xi(m)$, then $\epsilon_j(m) \cong -\frac{1}{\hat{N}_j(m)}\sum_{i=1}^{\hat{N}_j(m)} g(m)v_i$. Consequently, $R_j(m) = \mathrm{E}[\epsilon_j(m)\epsilon_j^T(m)] \cong \frac{1}{\hat{N}_j{}^2(m)}\hat{N}_j(m)g(m)Rg^T(m)$ or $R_j(m) \cong$

$\frac{1}{\hat{N}_j(m)} g(m)Rg^T(m)$. Similarly, we can estimate $P_j(1) \cong \frac{1}{\hat{N}_j(1)} g(1)Rg^T(1)$. However, since in general $\epsilon_\xi(m) \neq 0$, then in order to accommodate for $\epsilon_\xi(m)$, we can add some positive-definite matrix $Q > 0$ to $P_j(1)$ and to $R_j(m)$. Since it is expected that the estimation of c_j improves with additional observations, that is as m increases, then we propose to set

$$P_j(1) \cong \frac{1}{\hat{N}_j(1)} g(1)Rg^T(1) + Q_P \quad (12)$$

$$R_j(m) \cong \frac{1}{\hat{N}_j(m)} g(m)Rg^T(m) + \frac{1}{m} Q_R \quad (13)$$

where $Q_P > Q_R \geq 0$ are considered as tuning parameters. ■

The implementation of proposed strategy is shown below.

for $m = 1{:}\, M$

1. Consider observations $\hat{z}_n(m)$, $1 \leq n \leq N$
2. Implement k-means to do the classification
3. Obtain cluster $\hat{C}_j(m)$ with corresponding number of points $\hat{N}_j(m)$ and cluster centroid $\hat{c}_j(m)$, $1 \leq j \leq k$
4. **if** $m == 1$, $\bar{c}_j(1) = \hat{c}_j(1)$ and obtain $P_j(1)$ from (12), $1 \leq j \leq k$, **end**
5. Compute $R_j(m)$ from (13), $1 \leq j \leq k$
6. Obtain $K_j(m)$ from the recursive algorithm in (9) and (10), $1 \leq j \leq k$
7. Update $\bar{c}_j(m)$ using (5), $1 \leq j \leq k$

end

C. Other Strategies

We also consider three additional methods of handling noise in clustering. For consistency, the clustering engine is based on k-means algorithm, and we denote by $\hat{c}_j(.)$ the center of k^{th}-cluster obtained by the k-means algorithm.

Method A: This method is based on noisy observations associated with the replicate with least amount of noise. Since the observation error covariance matrix $\mathrm{E}[\epsilon_z(m)\epsilon_z^T(m)] = g(m)Rg^T(m)$, then $\overline{m} \triangleq \mathrm{argmin}_m \|g(m)Rg^T(m)\|$. Consequently, the estimate of cluster centroids is given by $\bar{c}_{j,\mathrm{A}} \equiv \hat{c}_j(\overline{m}), 1 \leq j \leq k$.

Method B: In this method we average the clusters' centroids obtained from all replicates. That is, $\bar{c}_{j,\mathrm{B}} = \frac{1}{M}\sum_{m=1}^{M} \hat{c}_j(m),\ 1 \leq j \leq k$.

Method C: The observations from all the replicates are expanded and then clustered using the k-means algorithm, see related work in [12]. That is, we extract $\bar{c}_{j,\mathrm{C}}, 1 \leq j \leq k$ by applying the k-means algorithm on $\cup_{m=1}^{M} \{\hat{z}_1(m), \hat{z}_2(m), \dots, \hat{z}_N(m)\}$.

III. Numerical Example

In this example we consider $z_n \in \mathbb{R}^2$, and generate $2N$ landmark points normally distributed at random in two clusters as follows. $C_1{:}\, z_n = (1,1) + \delta z_n,\ 1 \leq n \leq N$, and $C_2{:}\, z_n = (-1,-1) + \delta z_n,\ N+1 \leq n \leq 2N$, where $\delta z_n \in \aleph(0,1) \times \aleph(0,1)$. The observation errors, $g_i(m)v_n$, $i \in \{1,2\}$, are also generated at random with $v_n \in \aleph(0,\sigma) \times \aleph(0,\sigma)$, where

$g_1(m) = \left(1 + \sqrt[4]{m}\right)I$ or $g_2(m) = \frac{3}{2}\left(1 + \frac{\sin((m-1)\pi)}{M-1}\right)I$, $1 \leq m \leq M$. It is important to note that while considering $g_1(m)$, the measurement errors slowly but monotonically increases with additional replicate observations while $g_2(m)$ increases at a faster rate then decreases at the same rate.

Performance metric: Since the k-means algorithm minimizes the objective function $S \triangleq \sum_{j=1}^{k}\sum_{z\in C_j}\|z - c_j\|^2$ (1), then we implement the k-means algorithm on the modeled precise landmark points $z_n, 1 \leq n \leq 2N$ and compute the corresponding S. Subsequently, we compute $\hat{S} \triangleq \sum_{j=1}^{k}\sum_{z\in C_j}\|z - \hat{c}_j\|^2$ for each method under consideration while using the noisy observations and then normalize $\hat{S}$ with respect to S, that is, $S_{\mathrm{normal}} \triangleq \frac{\hat{S}}{S}$. S_{normal} is the performance metric adopted in this example.

The following three scenarios are considered and performance of Methods A, B, C and proposed method are examined:

Scenario 1: We fix $\sigma = 1$ and $N = 25$, and examine the performance for different values of M; in particular, $M \in \{2,3,\dots,20\}$.

Scenario 2: We fix $\sigma = 1$ and $M = 10$, and examine the performance for different values of N; in particular, $N \in \{10,20,\dots,100\}$.

Scenario 3: We fix $M = 10$ and $N = 25$, and examine the performance for different values of σ; in particular, $\sigma \in \{0.5,0.6,\dots,1.5\}$.

One thousand independent runs are conducted for Scenario 2 and 10,000 independent runs for Scenarios 1 and 3 where in each run we generate $2N$ new landmark points normally distributed at random as described above. We apply the latter while using $g_1(m)$ and another separate experiment while using $g_2(m)$. For the proposed method, we set $R_j(m)$ as in (13) with $Q_R = 0$, and $P_j(1)$ as in (12) with $Q_P = I$. MATLAB is employed throughout all numerical simulations.

Performance comparison: Fig. 1 to Fig. 3 show the performance of all methods corresponding to Scenarios 1 to 3, respectively, where the top plots are based on the employment of $g_1(m)$ and the bottom plots are based on $g_2(m)$. By examining Fig. 1 to Fig. 3, the following is concluded:

- The proposed method outperforms Methods A, B, and C at all levels.
- Unlike the other methods, Fig. 1 shows that S_{normal} corresponding to the proposed method monotonically decreases as M increases while using $g_1(m)$ or $g_2(m)$.
- Although Method B shows overall superiority over Methods A and C, Method A outperforms (Fig. 2) Methods B and C for $N > 70$ while using $g_2(m)$.
- Method C shows overall superiority over Method A when using $g_1(m)$ and Method A shows overall superiority over Method C when using $g_2(m)$.

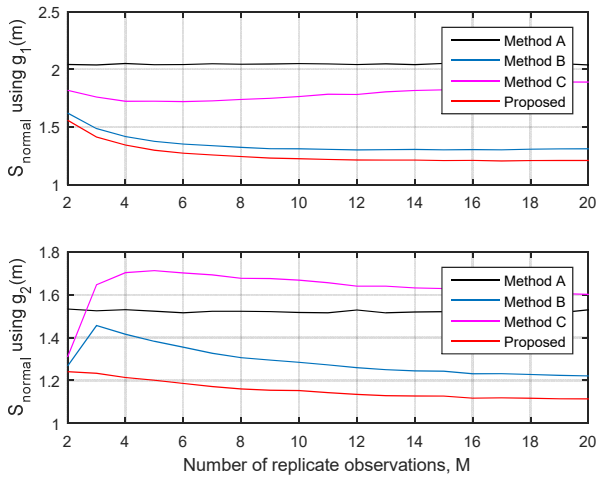

Fig. 1. Scenario 1: S_{normal} in function of M.

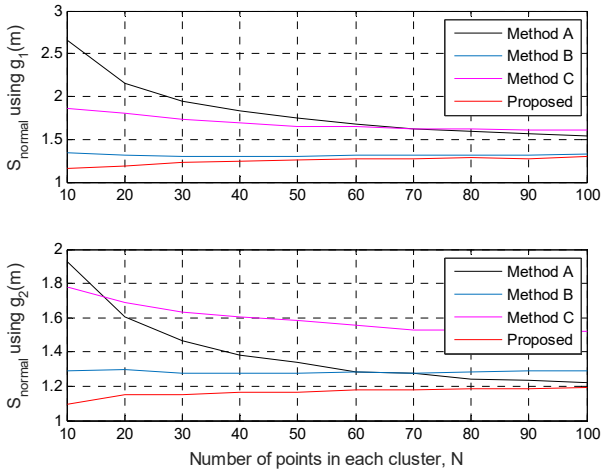

Fig. 2. Scenario 2. S_{normal} in function of N.

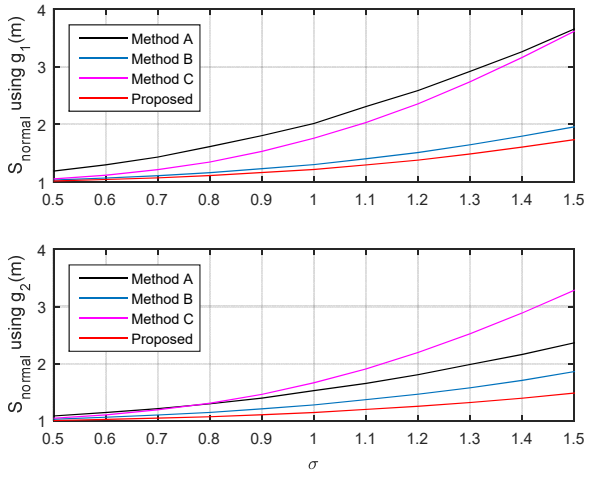

Fig. 3. Scenario 3. S_{normal} in function of σ.

IV. Conclusion

This paper proposed an optimal stochastic iterative algorithm for estimating cluster centroids in presence of noisy replicate measurements of datasets where the covariance of measurement noise may vary from replicate to replicate. Based on numerical simulations the superiority of the proposed approach over three other methods has been demonstrated. In addition, the performance of the proposed algorithm has been shown to monotonically improve with additional replicates where the norm of measurement noise covariance matrix increases at different rates as the number of replicates increases. That is, the proposed algorithm can robustly and effectively make use of the information in any replicate even with a large degree of measurement error variance.

References

[1] R. N. Dave, "Characterization and detection of noise in clustering," *Pattern Rec. Letters*, vol. 12(11), pp 657-664, 1991.

[2] R. N. Dave, "Robust fuzzy clustering algorithms," *in Proc. of the 2nd IEEE International Conference on Fuzzy Systems*, pp. 1281–1286, 1993.

[3] J. A. Cuesta-Albertos, A. Gordaliza, and C. Matran, "Trimmed k-means: An attempt to robustify quantizers," *Annals of Statistics*, vol. 25, no. 2, 1997, pp. 553–576.

[4] L. A. Garcıa-Escudero, and A. Gordaliza, "Robustness properties of k means and trimmed k means," *Journal of the American Statistical Association*, vol. 94, no. 447, 1999, pp. 956–969.

[5] S. Ben-David, and N. Haghtalab, "Clustering in the Presence of Background Noise," *in Proc. of the 31st International Conference on Machine Learning*, vol. 32 Beijing, China, 2014.

[6] S. Sharma, M. Goel, and P. Kaur, "Performance comparison of various robust data clustering algorithms." *International Journal of Intelligent Systems and Applications*, vol. 7, 2013, pp. 63-71.

[7] M. L. Lee, F. C. Kuo, G. A. Whitmore, and J. Sklar, "Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations," *Proceedings of the National Academy of Sciences*, vol. 97, no. 18, 2000, pp. 9834–9839.

[8] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, et al., "Molecular classification of cutaneous malignant melanoma by gene expression profiling," *Nature*, vol. 406, 2000, pp. 536–540.

[9] M. K. Kerr, and G. A. Churchill, "Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments," *Proceedings of the National Academy of Sciences*, vol. 98, no. 16, 2001, pp. 8961–8965.

[10] E. R. Dougherty, J. Barrera, M. Brun, S. Kim, R. M. Cesar, Y. Chen, et al., "Inference from clustering with application to gene-expression microarrays," *Journal of Computational Biology*, vol. 9, no. 1, 2002, pp. 105–126.

[11] P. Pavlidis, Q. Li, and W. S. Noble, "The effect of replication on gene expression microarray experiments," *Bioinformatics*, vol. 19, no. 13, 2003, pp. 1620–1627.

[12] K. Y. Yeung, M. Medvedovic, and R. E. Bumgarner, "Clustering gene-expression data with repeated measurements," *Genome Biology*, vol. 4, no. 5, 2003, pp.1–17.

[13] R. Sloutsky, N. Jimenez, S. J. Swamidass, and K. M. Naegle, "Accounting for noise when clustering biological data," Briefings in bioinformatics, vol. 14, no. 4, 2013, pp. 423–436.

[14] K. K. Saab, and S. S. Saab, Jr, "A stochastic Newton's method with noisy function measurements", *IEEE Signal Processing Letters*, vol. 23, no. 3, pp. 361–365, 2016.