The Bayesian procedure of Olkin, Guttman and Philips (1995) maximizes the a posteriori probability of the data over all possible cluster assignments, over all meaningful numbers of clusters. Suppose instead we minimize the within-cluster sum of squares over all possible assignments of the data to say, k, clusters. Given lab means $x_1, \ldots, x_n$, the objective is

$$(1) \qquad \underset{J_1,\ldots,J_k}{min} \sum_{m=1}^{k} \frac{1}{2|J_m|} \sum_{i,j \in J_m} (x_i - x_j)^2 = \underset{J_1,\ldots,J_k}{min} \sum_{m=1}^{k} \sum_{i \in J_m} (x_i - \bar{x}_{J_m})^2,$$

where the minimization is over all partitions $J_1, \ldots, J_k$ of the indices $1, \ldots, n$ into k sets. $\bar{x}_{J_m}$ denotes the mean of the data corresponding to partition J_m, i.e., the mean of $\{x_i | i \in J_m\}$. Since the lab means generally come with variance estimates $v_1, \ldots, v_n$, a more appropriate objective would weight the sum of squares. (A model is given further below.) Let $w_i = (1/v_i)/(\sum_{i=1}^{n}(1/v_i))$ be the inverse variance weight of the ith mean x_i. The objective becomes:

$$(2) \qquad \underset{J_1,\ldots,J_k}{min} \sum_{m=1}^{k} \frac{1}{2} \sum_{i,j \in J_m} w_i w_j (x_i - x_j)^2 = \underset{J_1,\ldots,J_k}{min} \sum_{m=1}^{k} \sum_{i \in J_m} w_i (x_i - \bar{x}_{J_m})^2,$$

where now $\bar{x}_{J_m} = \sum_{i \in J_m} w_i x_i / (\sum_{i \in J_m} w_i)$ is the weighted mean of cluster J_m. The statistic in (2) being summed is similar to Cochran's Q statistic, but it is now computed on each cluster.

Problem (1) is k-means clustering (Hartigan (1979)): finding the partition of the data that minimizes the sum of the distances (typically Euclidean distane) from each data point to the center of that data point's partition. Problem (2) is analogous to k-means clustering but a distinct problem since (2) does not use a well-defined distance function.

Computing the optimal clusters for k-means clustering is NP-hard when the data has dimension greater than 1 (Mahajan, Nimbhorkar and Varadarajan (2009)), with the only known exact algorithms requiring one to go through all possible cluster assignments. The same must hold for problem 2, which reduces to k-means clustering when all variances/weights are equal. However, as long as the data is scalar, like most interlaboratory data, the problem appears to be completely tractable (computation discussed further below).

A downside of this procedure as an outlier detection method is that the within-cluster sum of squares can only decrease when the number of clusters is increased. With the procedure of Olkin et al. (1995) (data on last page), one finds the maximum a posteriori likelihood over all k clusters, then maximizes this maximum over all values k:

Table 1. Determination of the number k of aberrant results

k	$\max C(j_1, \ldots, j_K)$
1	$.9815 = C(8)$
2	$.9983 = C(1,8)$
3	$.5846 = C(1,6,8)$
4	$.9298 = C(1,6,7,8)$
5	$.2210 = C(1,3,6,7,8)$

When the criterion is the sum of squares, other methods must be used. A common heuristic is a scree-type plot. One plots the minimizing within sum of squares over all k cluster assignments against the number k of clusters. One then looks for a "bend" or "elbow" in the plot to indicate the correct cutoff. For example, in fig. 1 shows scree plots for synthetic data generated with 0,1, and 2 well-separated clusters.

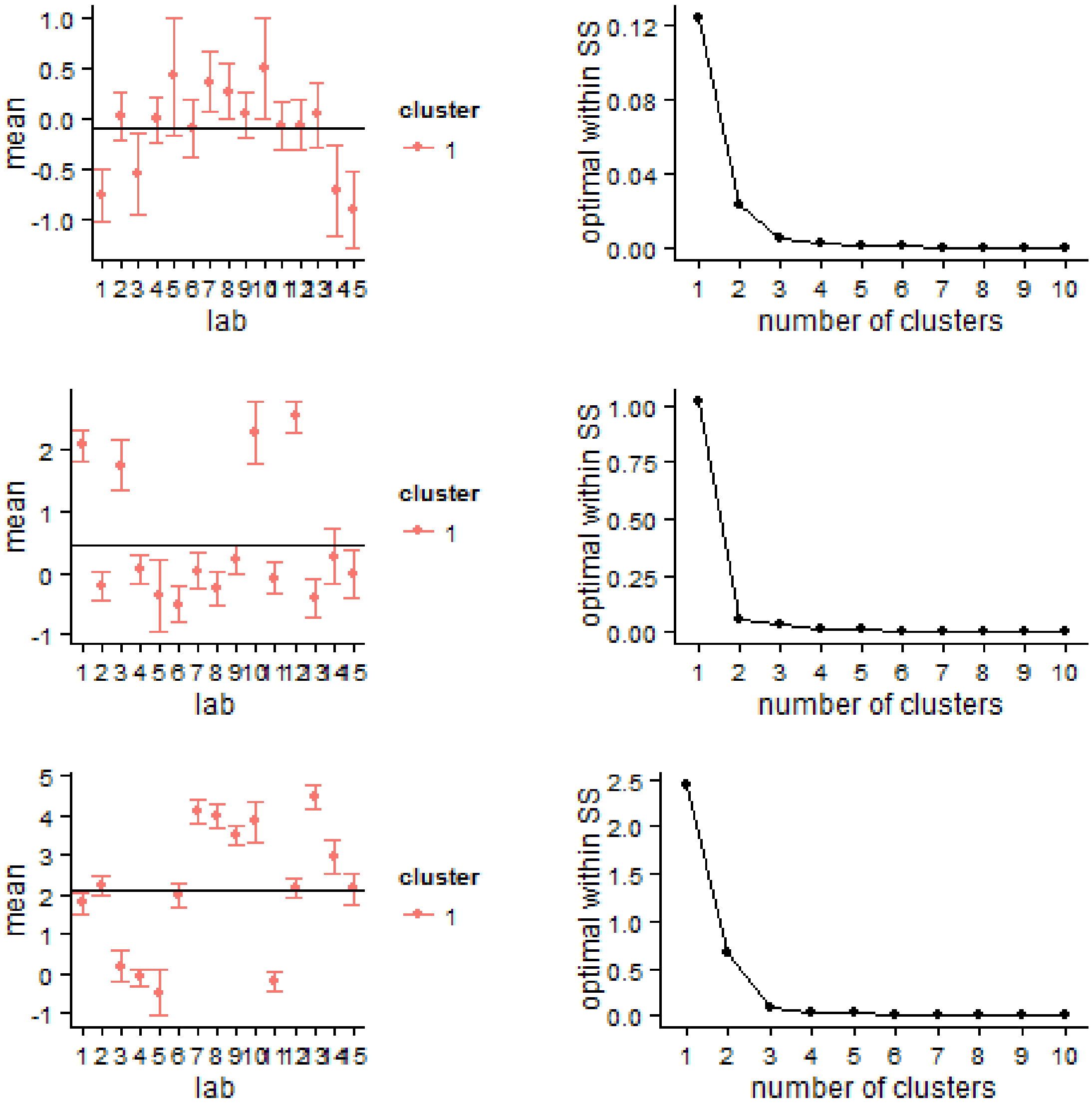


Figure 1. Synthetic data and scree-type plots for 1, 2, and 3 clusters. The "elbow" in the scree plot is easy to locate at the correct number of clusters.

Applied to the data set of Olkin et al. (1995), one obtains the scree plot (fig. 2). The scree method suggests 3 clusters. As in the Bayesian procedure of Olkin et al. (1995), labs #1 and #8 are identified as a separate cluster from the majority. However, under the present analysis another group consisting of labs #6 and #7 are also gouped in a separate cluster.

Comparison with Olkin et al. (1995) procedure. As noted, the procedure of Olkin et al. (1995) has the advantage of giving a natural way to select the number of aberrant laboratories, whereas heuristic methods like the scree plot must be used for the present procedure. On the other hand, an advantage of the present method is that it does not assume common standard errors across the laboratory means. We have noted that the Bayesian procedure gives different results from other outlier detection methods when this assumption is not warranted by the data.

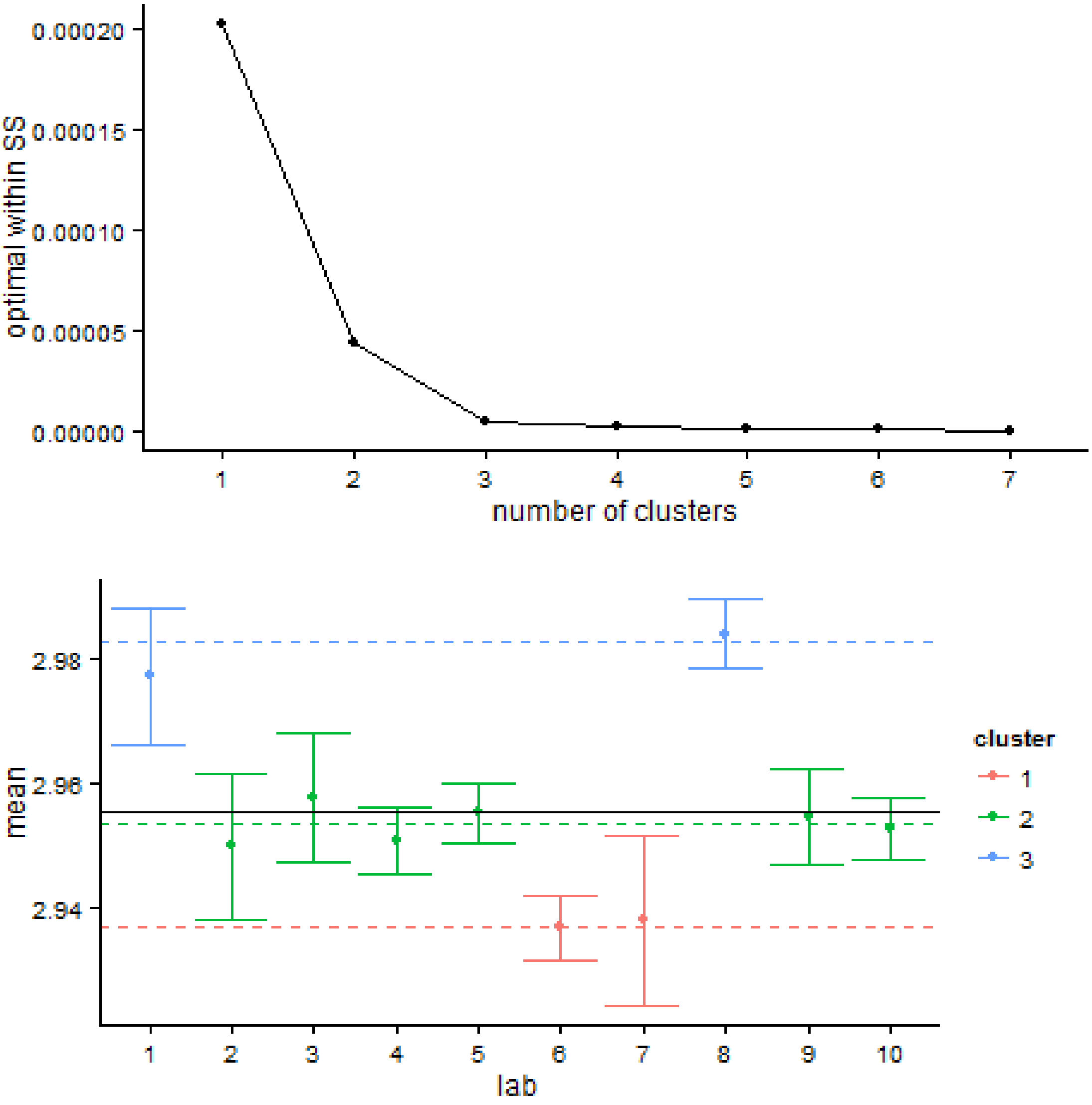


Figure 2. Minimized within sum of squares for a range of values of k the number of clusters, using Olkin et al. (1995) data.

The present procedure is differs conceptually from the Bayesian procedure as well. The Bayesian procedure seeks to identify individual laboratories as aberrant. In the model, each aberrant lab has a mean shifted by some amount from the true mean. The means of the data are divided into as many means as there are aberrant labs plus the mean of the non-aberrant labs. In the present procedure, the labs are grouped into clusters. Lab results from a given cluster share a common mean, as in a random effects model.

Random effects interpretation. Given a partition $J_1, \ldots, J_k$ of $1, \ldots, n$ and k scalars $\theta_1, \ldots, \theta_k$, suppose

$$x_i \sim \mathcal{N}(\theta_m, v_i) \qquad\qquad \text{when } i \in J_m.$$

This is a random effects model with k subject effects $\theta_1, \ldots, \theta_k$, where the variance of each observation is allowed to vary, even within a cluster:

$$x_{ij} = \theta_i + \epsilon_{ij} \qquad\qquad \epsilon_{ij} \sim \mathcal{N}(0, \sigma_{ij}^2)$$

Consider the MLE over Θ, the set of multivariate normal distributions for all such partitions $J_1, \ldots, J_k$, and partition means $\theta_1, \ldots, \theta_k$. The negative log-likelihood is minimized by

$$\begin{aligned}
&\underset{\Theta}{min} \sum_{m=1}^{k} \sum_{i \in J_m} (x_i - \theta_m)^2/(2v_i) + \text{constant} \\
&= \frac{1}{2} \underset{\Theta}{min} \sum_{m=1}^{k} \sum_{i \in J_m} w_i (x_i - \theta_m)^2 + \text{constant} \\
&= \frac{1}{2} \underset{J_1, \ldots, J_k}{min} \sum_{m=1}^{k} \sum_{i \in J_m} w_i (x_i - \bar{x}_{w,m})^2 + \text{constant}
\end{aligned}$$

,

which is the same as problem (2) above. The last equality follows since, for any cluster, the likelihood is maximized for $\theta_1, \ldots, \theta_k$ equal to the cluster (weighted) means $\bar{x}_{w,1}, \ldots, \bar{x}_{w,k}$. Therefore (2) can be thought of as an normal MLE problem, although over a non-convex parameter space.

Polynomial time solution in 1-dimensoinal case. The problem is tractable because, in the 1-dimensional case only, after ordering the data the optimal assignment to clusters is a contiguous partition. That is, there are no interleaved cluster assignments as for example in 3. This necessity is clear in the unweighted/homoscedastic case 1: in 3, re-assigning x_2 to cluster A and x_1 to cluster B lowers the within sum of squares. These arguments extend to the case of weighted means and weighted within sum of squares.

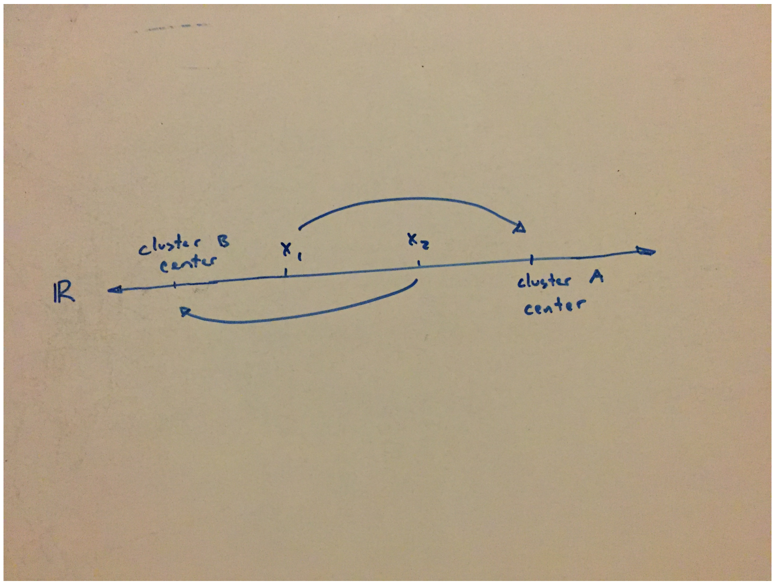
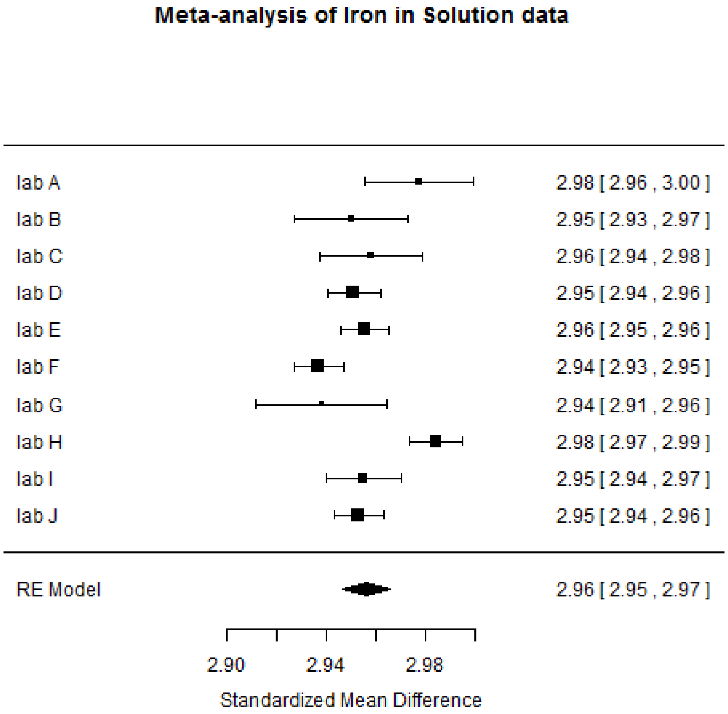


FIGURE 3. Interleaved clusters

Once it is established that the optimal cluster assignments is a contiguous partition, a inductive step, well-known from ordinary k-means clustering, leads to a dynamic programming solution.

Namely, if $J_1, \dots, J_{k_0}$ represents an assignment of $x_1, \dots, x_{n_0}$, $n_0 \geq k_0$, that minimizes the within sum of squares, and $j_0 = minJ_{k_0}$ is the minimum index in the last cluster J_{k_0}, then $J_1, \dots, J_{k_0-1}$ represents a minimizing cluster assignment for $x_1, \dots, x_{j_0-1}$. (Again, the $x_1, \dots, x_n$ are assumed sorted.) This induction gives an algorithm that runs in $O(nk)$ time. It has the nice side effect of obtaining along the way the minimizing cluster assignment for $1, \dots, k$ clusters.



References

Hartigan, J. (1979). Algorithm 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C* 28: 100–108.

Mahajan, M., Nimbhorkar, P. and Varadarajan, K. (2009). The planar k-means problem is np-hard. In *WALCOM '09: Proceedings of the 3rd International Workshop on Algorithms and Computation*. Berlin: Springer-Verlag, 274–285.

Olkin, I., Guttman, I. and Philips, R. (1995). Estimating the number of aberrant laboratories. *Probab. Engrg. Inform. Sci.* 9: 133–150, MR1336806.