Update 25/03/2020

Stephen Coleman

23/03/2020

General model

For data $X = (x_1, \ldots, x_N)$, where each item $x_i = (x_{i1}, \ldots, x_{iP})$, we use a K-component mixture-model paramaterised by θ to describe the data:

$$p(x_i|\theta, \pi) = \sum_{k=1}^{K} \pi_k f(x|\theta_k).$$

Here $\pi = (\pi_1, \ldots, \pi_K)$ is the proportion of items assigned to each component and θ_k is the component specific parameters.

We assume that there is a common probability density function, $f(\cdot)$, associated with each component (e.g. Gaussian). Independence is assumed between the P features, thus:

$$p(x_i|\theta, \pi) = \sum_{k=1}^{K} \pi_k \prod_{p=1}^{P} f(x|\theta_{kp}),$$

where θ_{kp} is the parameters for the p^{th} feature within the k^{th} component (e.g. if we are using a *Gaussian mixture model*, then $\theta_{kp} = (\mu_{kp}, \sigma^2_{kp})$, the mean and standard deviation of the items in the k^{th} component within the p^{th} feature).

In the language of Law, Jain, and Figueiredo (2003), we assume that a subset of the features are *irrelevant*. By this we mean that for a given item x_i,

$$f(x_i|\theta_{kp}) = f(x_i|\theta_{lp}) = g(x_i|\lambda_p) \ \forall \ k, l \in \{1, \ldots, K\}.$$

Thus an irrelevant feature does not contribute any component specific information and is irrelevant to uncovering structure within the data. Let $\Phi = (\phi_1, \ldots, \phi_P)$ be a binary variable indicating the relevance of a feature (i.e. $\phi_p = 1$ if the p^{th} feature is relevant and 0 otherwise). Then our model can be written:

$$p(x_i|\theta, \pi, \Phi) = \sum_{k=1}^{K} \pi_k \prod_{p=1}^{P} f(x_i|\theta_{kp})^{\phi_p} g(x_i|\lambda_p)^{(1-\phi_p)}. \tag{1}$$

If we use an allocation variable $z = (z_1, \ldots, z_N)$ to indicate which component the items belong to, we may write:

$$p(x_i|z_i = k, \theta, \pi, \Phi) = \prod_{p=1}^{P} f(x_i|z_i = k, \theta_{kp})^{\phi_p} g(x_i|\lambda_p)^{(1-\phi_p)}$$

$$p(x|z, \theta, \pi, \Phi) = \prod_{i=1}^{N} \prod_{p=1}^{P} f(x_i|z_i = k, \theta_{kp})^{\phi_p} g(x_i|\lambda_p)^{(1-\phi_p)}$$

Simulations

In our simulations we are interested in testing how *consensus inference* compares to Bayesian inference of mixture models in various circumstances. In each simulation we will assume a generative model that can be described by a finite mixture of Gaussian models.

$$p(x_i|z_i = k, \theta, \pi, \Phi) = \prod_{p=1}^{P} f(x_i|z_i = k, \theta_{kp})^{\phi_p} f(x_i|\theta_p)^{(1-\phi_p)}. \tag{2}$$

where $f(\cdot)$ describes the Gaussian pdf and thus $\theta = (\mu, \sigma^2)$. We assume that the irrelevant variables will be Gaussian in form and that the observed values will be drawn from the same Gaussian distribution for the entire population. Let $P_n = \sum_{p=1}^{P} \phi_p$ be the number of irrelevant features present, and $P_s = P - P_n$ be the number of relevant features present. Then in each simulation we will change various variables associated with this model:

- N: the number of items being clustered;
- P_s: the number of *relevant* features present;
- P_n: the number of *irrelevant* features present;
- K: the true number of subpopulations present;
- π: the proportion of points sampled from each component;
- Δ_μ: the difference between the means associated with each component in each feature;
- σ^2: the standard deviation within each feature for each component; and
- α: the concentration of the Dirichlet distribution π might be generated from (only relevant when π is sampled as explained below).

π will be chosen in one of two ways:

- "Even": a K-vector with all entries equal to $\frac{1}{K}$; or
- "Varying": sampled from a Dirichlet distribution with concetration of *alpha*.

In the second case we will explain our choice of α each time.

I would expect that there is some function of the number of samples, the number of informative features, the number of clusters, the distance between component means and the value of σ^2 used that explains how easy it is to resolve the clutering structure. If C' is the true clustering and C^* is that predicted by the model, and $ARI(X, Y)$ is the adjusted rand index between partitions X and Y, then I expect there to be some relationship of the nature:

$$\begin{aligned}
ARI(C^*, C') &\propto \log(N) \\
ARI(C^*, C') &\propto \sqrt{P_S} \\
ARI(C^*, C') &\propto -K \log(K) \\
ARI(C^*, C') &\propto \Delta_\mu \\
ARI(C^*, C') &\propto \frac{1}{\sigma} \\
ARI(C^*, C') &\propto \frac{\Delta_\mu K \log\left(\frac{N}{K}\right) \sqrt{P_S}}{\sigma}
\end{aligned}$$

I do not expect that the stated nature of these relationships is true, but the directionality of these relationships is expected to hold and something of the relative speed of improvement in resolving the true structure is intended to be indicated by the use of $\log(\cdot)$ and $\sqrt{\cdot}$. Note that the positive linear dependence on K is due to how I am defining Δ_μ. These statements also convey that for many of the variables it is relative values that matter; for instance if Δ_μ increases we expect improved resolution of clusters, but if σ^2 also grows proportionally, then we would not expect the improvement to be at all as significant.

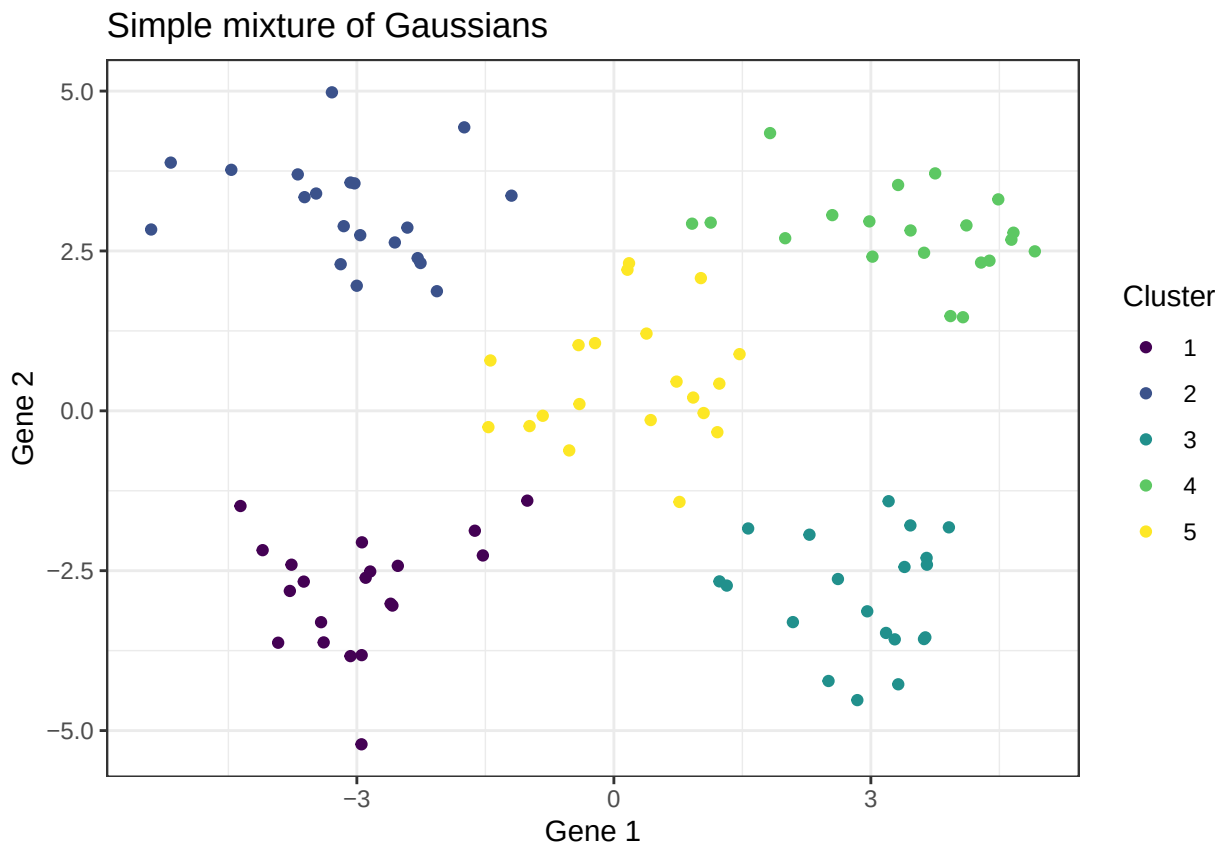
We will test:

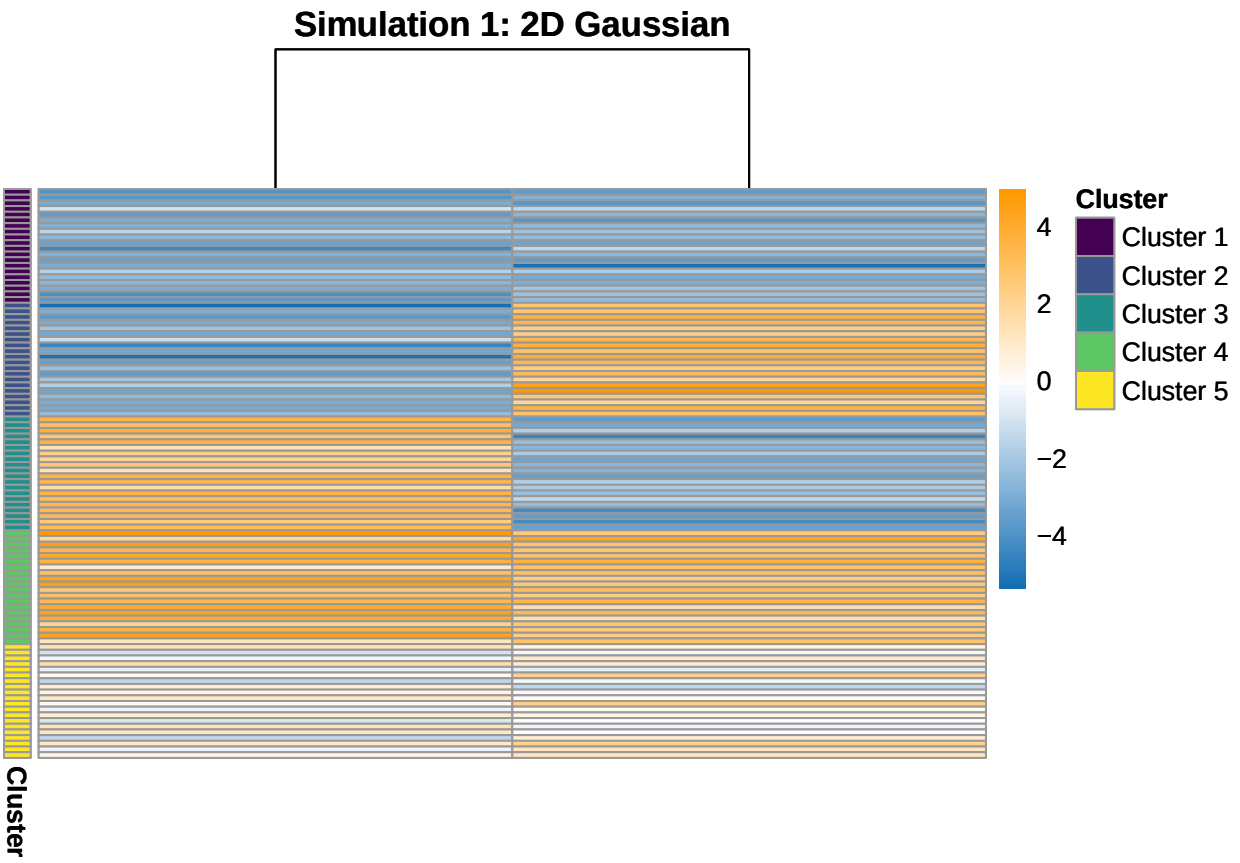
1. The 2D Gaussian case (this is a sense-check);
2. The lack-of-structure case in 2 dimensions;
3. The large N, small P paradigm;
4. Increasing σ^2;
5. Increasing the number of irrelevant features;
6. The small N, large P case; and
7. Varying the proportion of the total population assigned to each sub-population.

Simulation 1: 2D Gaussian

This is a sense-test case. It is the easiest to judge how well sensible the final clustering is as we can visualise the data fully in a 2D setting.

- $N = 100$;
- $P_s = 2$;
- $P_n = 0$;
- $K = 5$;
- $\pi = (0.2, 0.2, 0.2, 0.2, 0.2)^T$;
- $\Delta_\mu = 2$; and
- $\sigma^2_{kp} = 1$.

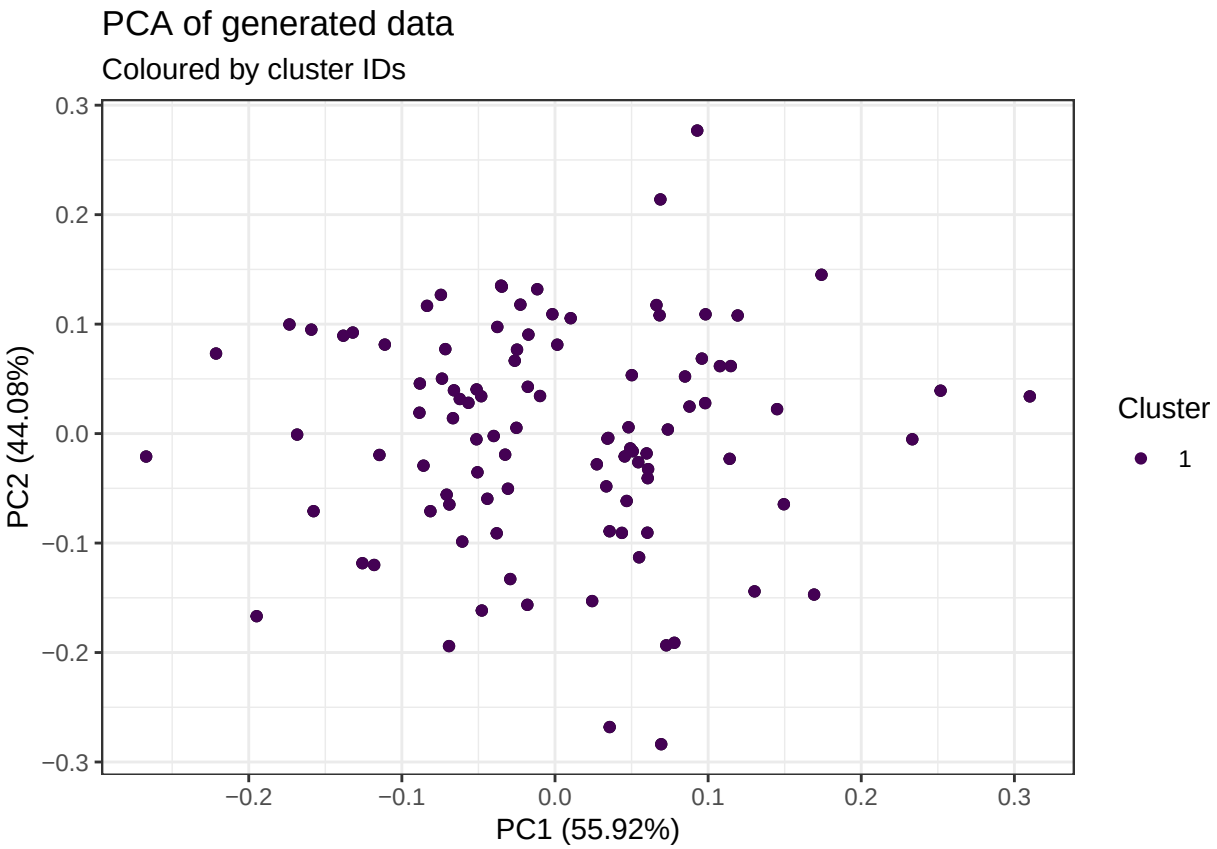

Simple mixture of Gaussians
5.0
2.5
0.0
−2.5
−5.0
Gene 2
−3
0
3
Gene 1
Cluster
1
2
3
4
5

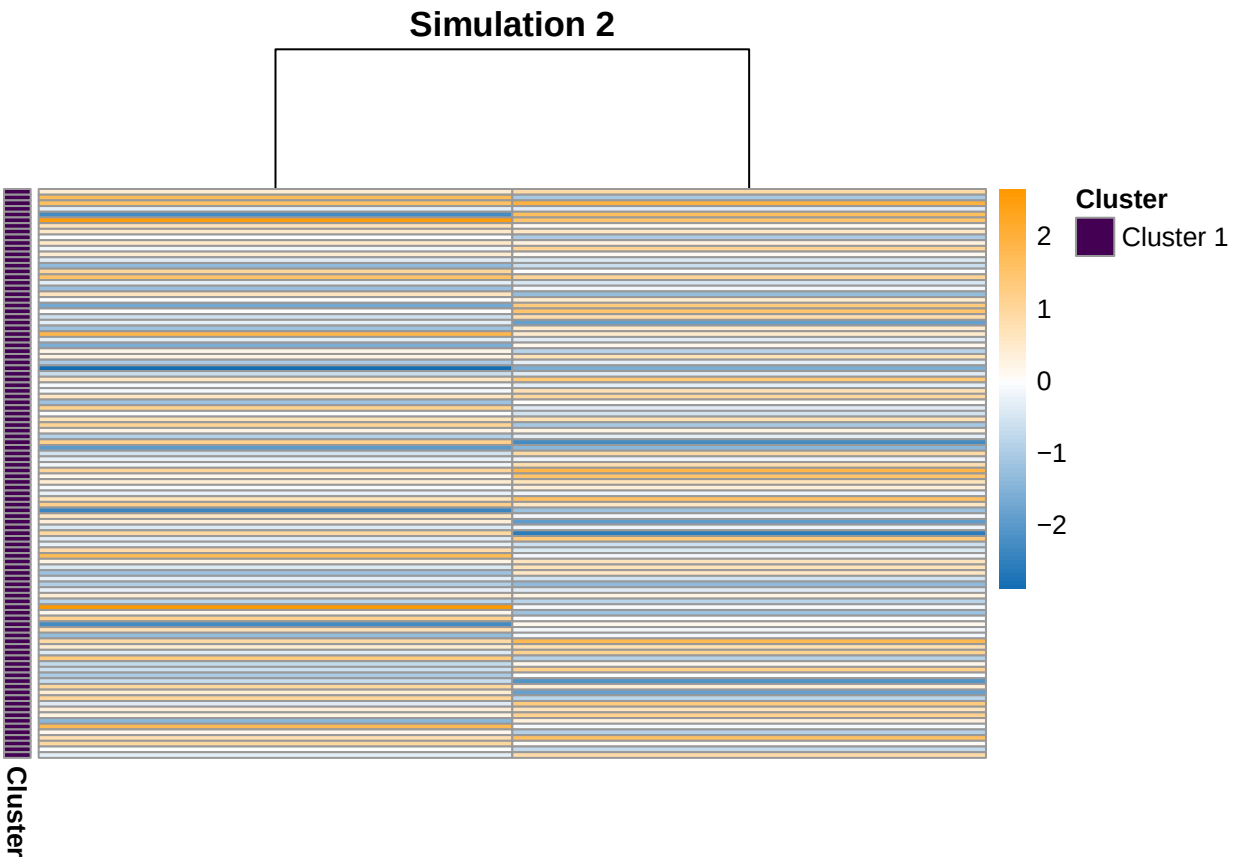


Simulation 2: No structure

We wish to test the case when there is no structure present (i.e. all items are generated from the same Gaussian distirbution). In this scenario there are no subpopulations present so all items should be allocated to the same component.

- $N = 100$;
- $P_s = 0$;
- $P_n = 2$;
- $K = 1$;
- $\pi = 1$;
- $\Delta_\mu = 0$; and
- $\sigma^2_{kp} = 1$.


PCA of generated data
Coloured by cluster IDs
PC2 (44.08%)
PC1 (55.92%)
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
Cluster
1



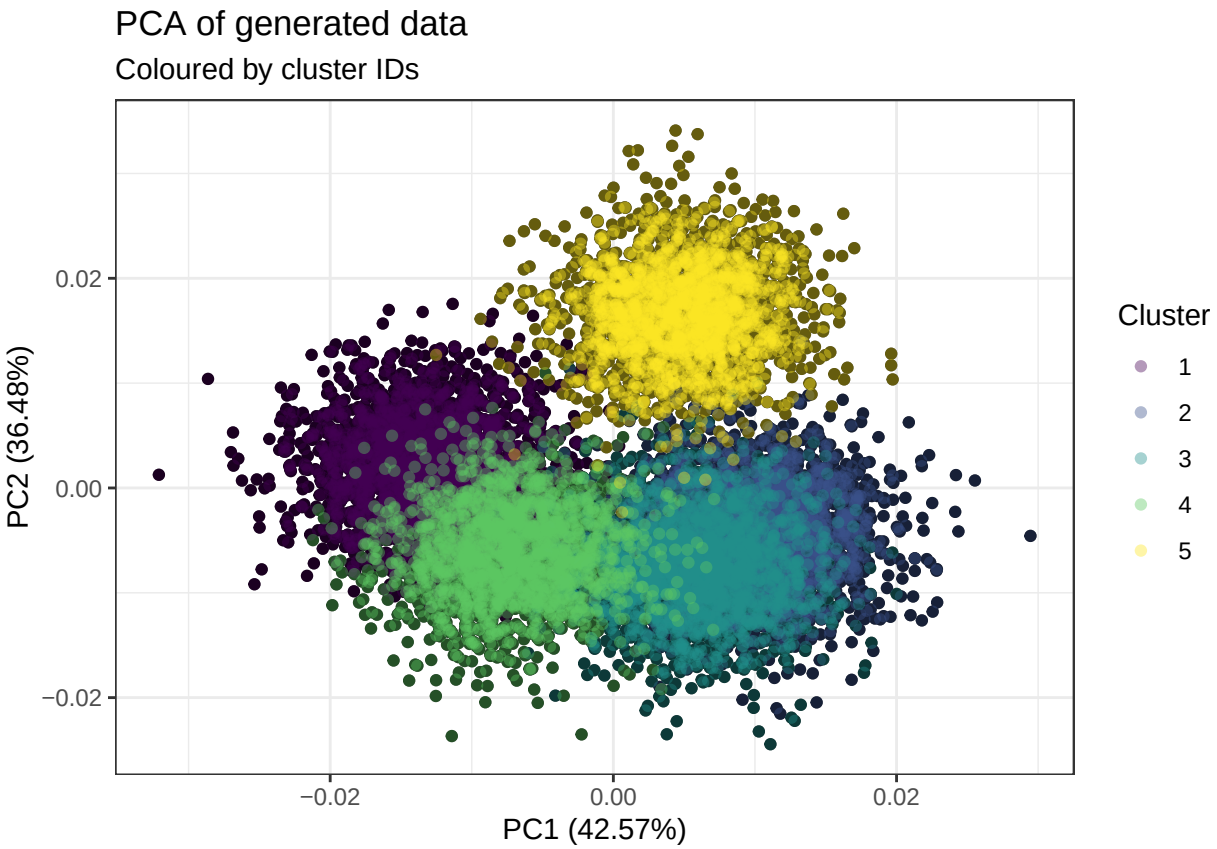
Simulation 3: Large N, Small P

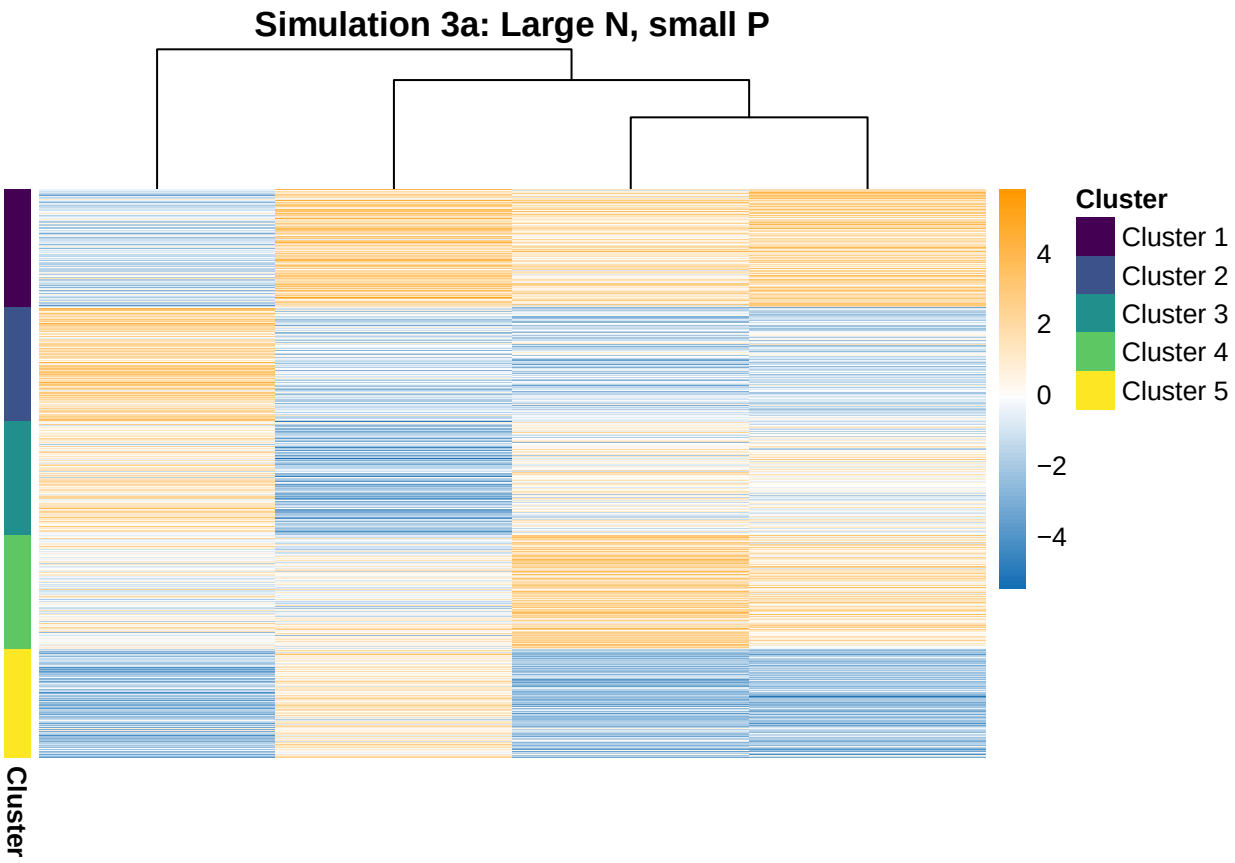
Test performance within the large N, small P paradigm such as is encountered in flow cytometry. Many points will be confidently allocated, but as the number of features is small some items are liable to be on the border of several clusters and therefore harder to allocate (although even $P_S = 4$ might be sufficient for high performance).

Simulation 3a: Large N, Small P dataset

We use a large sample size and a small number of features to investigate the *large N, small P* case.

- $N = 10{,}000$;
- $P_s = 4$;
- $P_n = 0$;
- $K = 5$;
- $\pi = vec(\frac{1}{5})$;
- $\Delta_\mu = 1$; and
- $\sigma^2_{kp} = 1$.

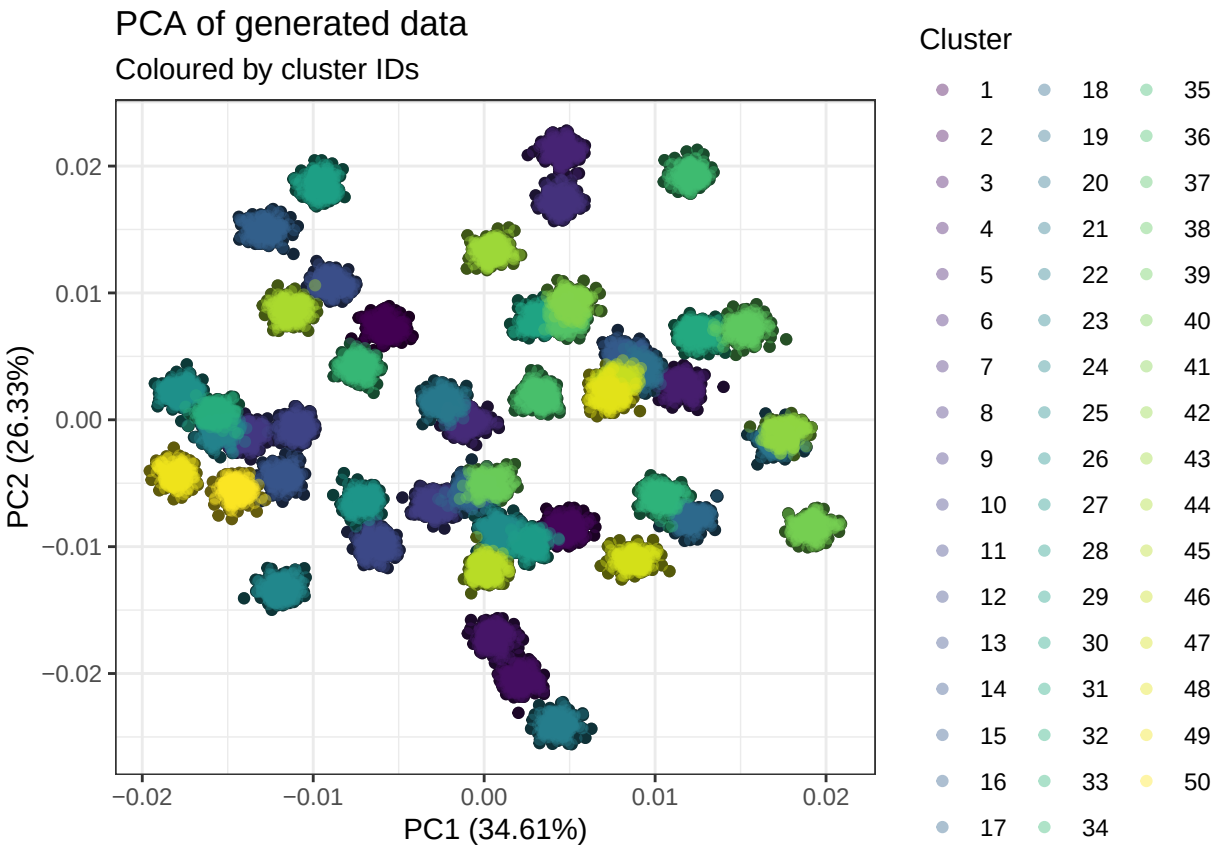

PCA of generated data
Coloured by cluster IDs
PC2 (36.48%)
0.02
0.00
−0.02
−0.02
0.00
0.02
PC1 (42.57%)
Cluster
1
2
3
4
5

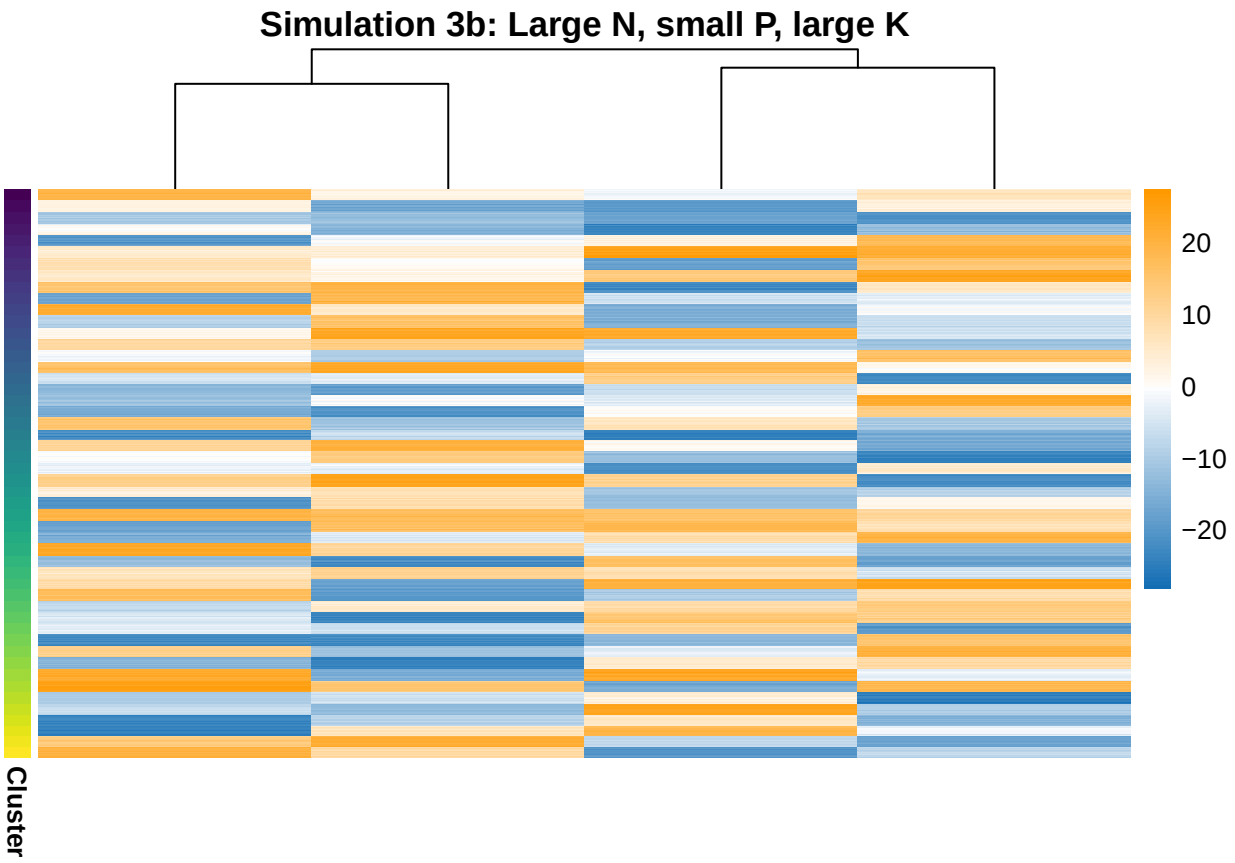


Simulation 3b: Large N, large K, small P dataset

Similarly to 3a, we use a large N and small P but also grow K:

- $N = 10{,}000$;
- $P_s = 4$;
- $P_n = 0$;
- $K = 50$;
- $\pi = vec(\frac{1}{50})$;
- $\Delta_\mu = 1$; and
- $\sigma^2_{kp} = 1$.


PCA of generated data
Coloured by cluster IDs
PC2 (26.33%)
PC1 (34.61%)
0.02
0.01
0.00
−0.01
−0.02
−0.02
−0.01
0.00
0.01
0.02
Cluster
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50

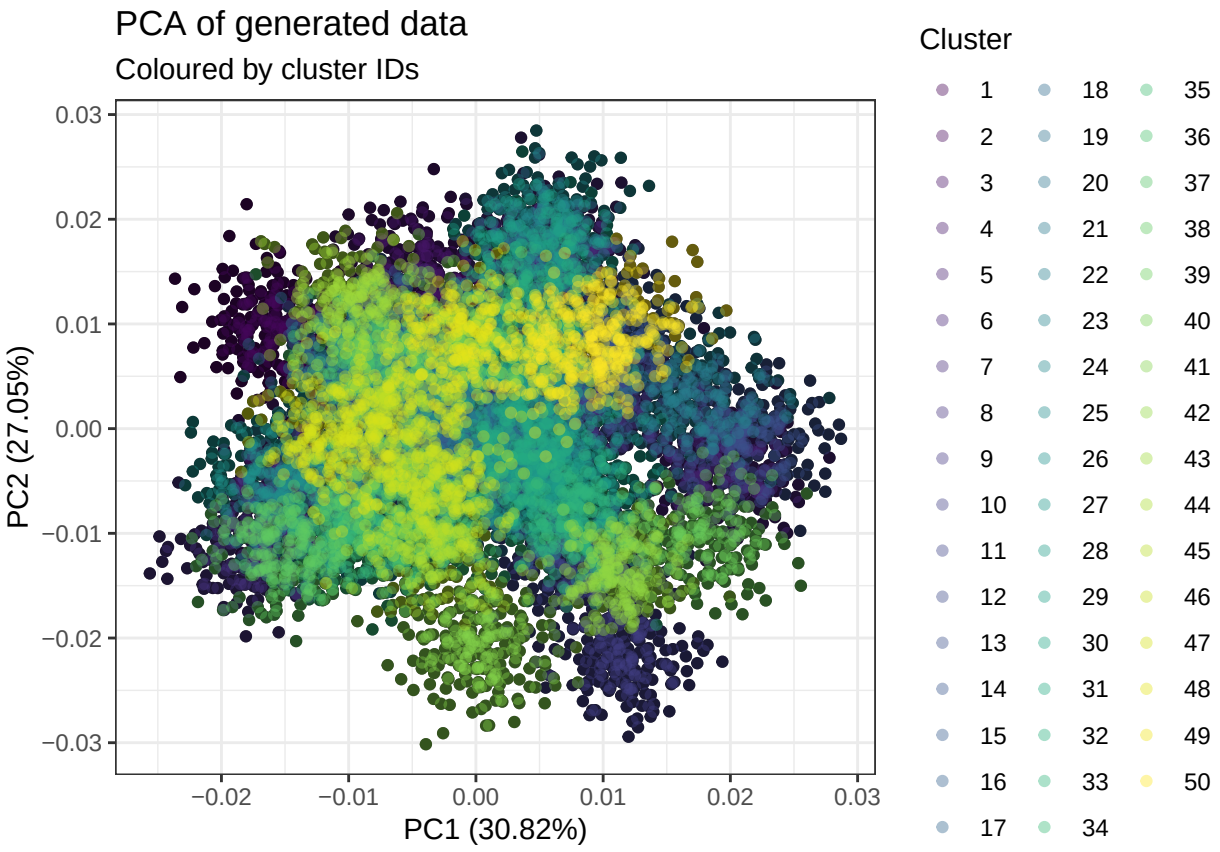


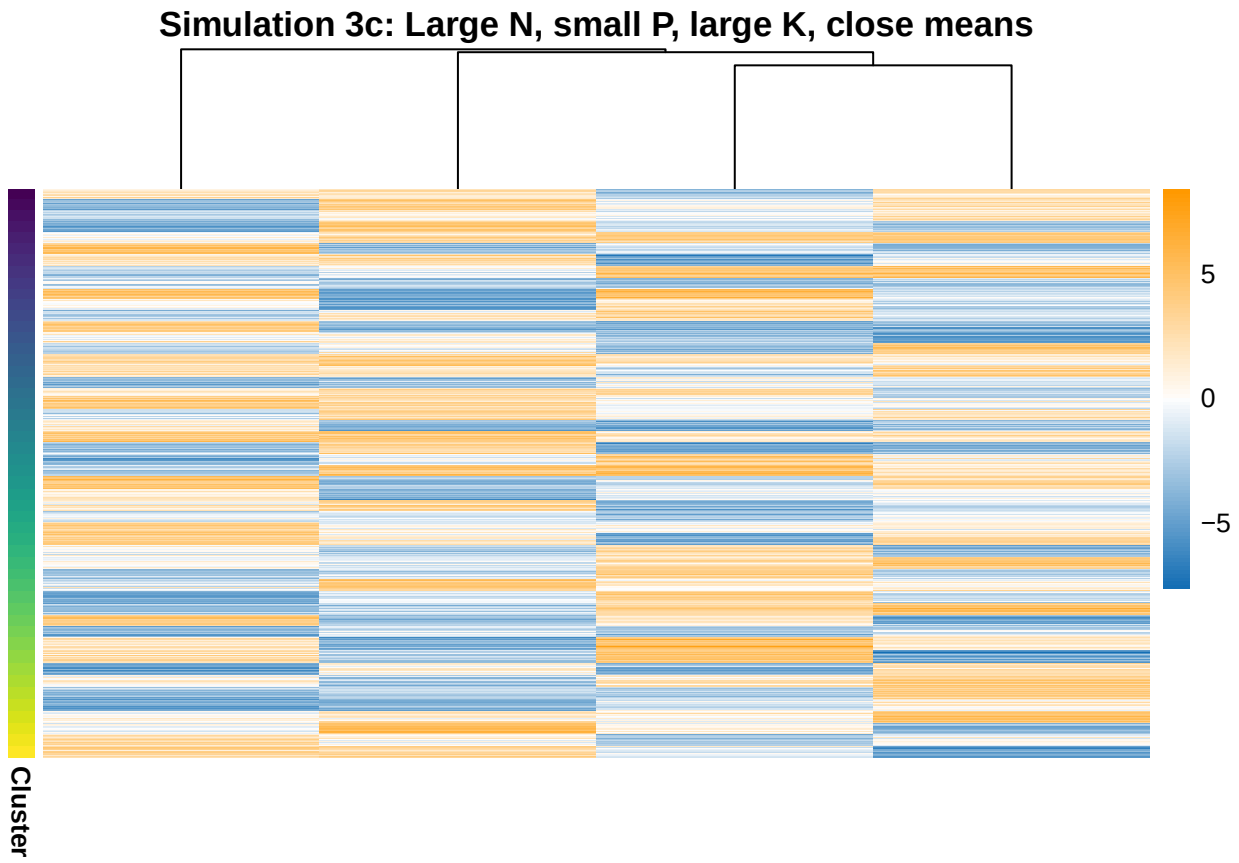
Simulation 3c: Large N, large K, small P, small Δ_μ dataset

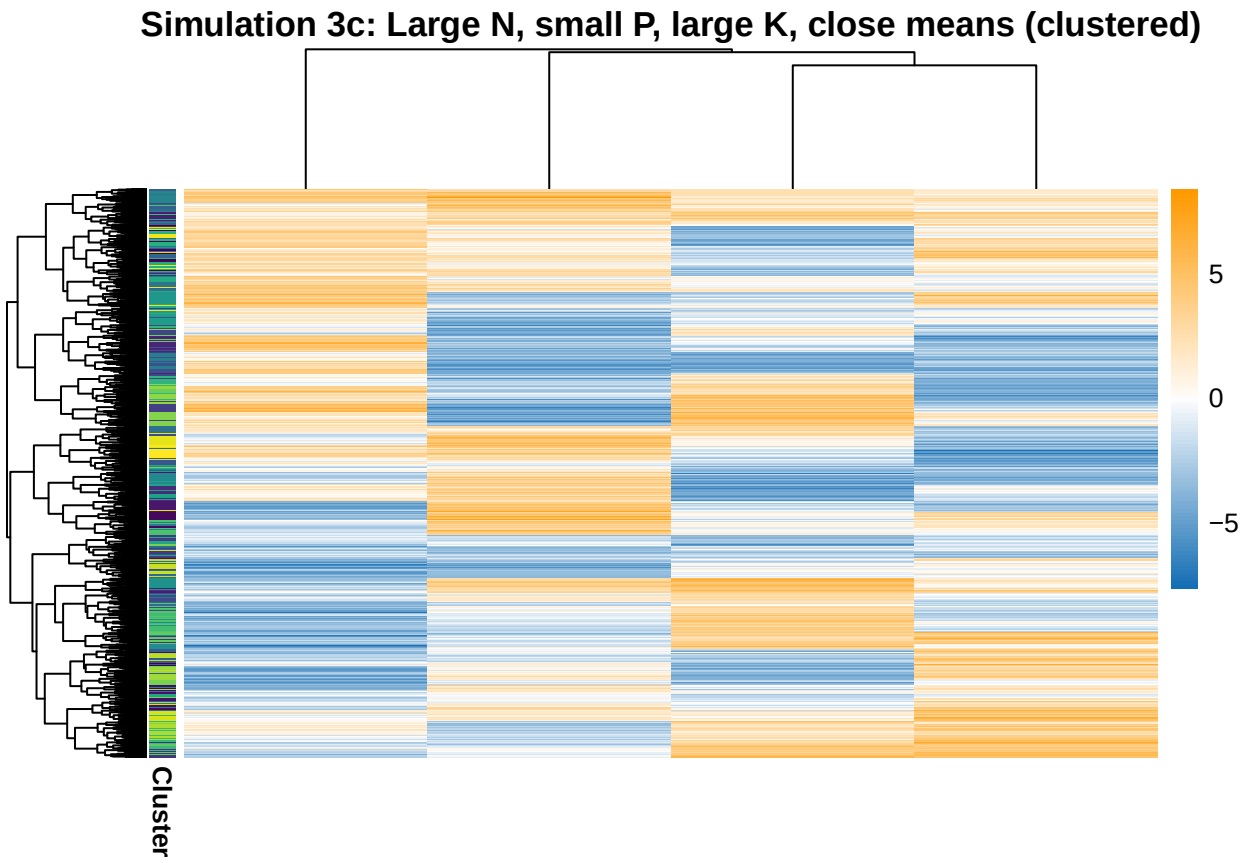
Similar to 3b but now the distance between means is decreased. This should be very hard to cluster well as many of the components overlap on the small number of features. It might be an interesting case to compare many short chains to a single long chain. I suspect that with the small number of features that the Bayesian inference should be less liable to becoming trapped (however, if a cluster is emptied it may struggle to include a previously emptied component). If consensus inference behaves comparatively well here I think that's a good indication of robustness as my a priori expectations are that a long chain would perform better as it explores the different probable clusterings (basically, small number of features but high uncertainty in the clustering structure should favour a long chain I think).

To be clear, *no method will find the large number of "true" subpopulations present*, but there is structure present as well as many valid possible partitions.

- $N = 10{,}000$;
- $P_s = 4$;
- $P_n = 0$;
- $K = 50$;
- $\pi = vec(\frac{1}{50})$;
- $\Delta_\mu = 0.5$; and
- $\sigma^2_{kp} = 1$.


PCA of generated data
Coloured by cluster IDs
PC2 (27.05%)
PC1 (30.82%)
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.02
−0.01
0.00
0.01
0.02
0.03
Cluster
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50


Simulation 3c: Large N, small P, large K, close means
Cluster
5
0
−5

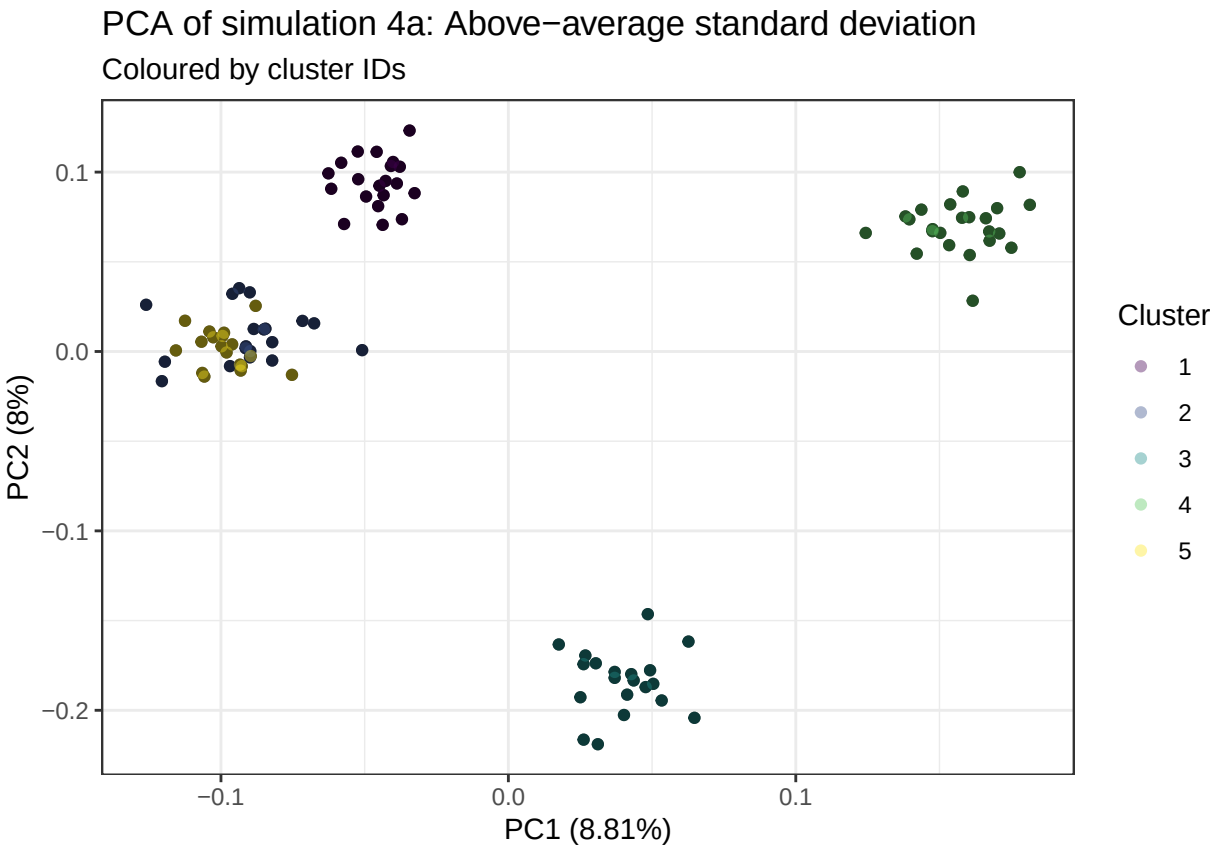


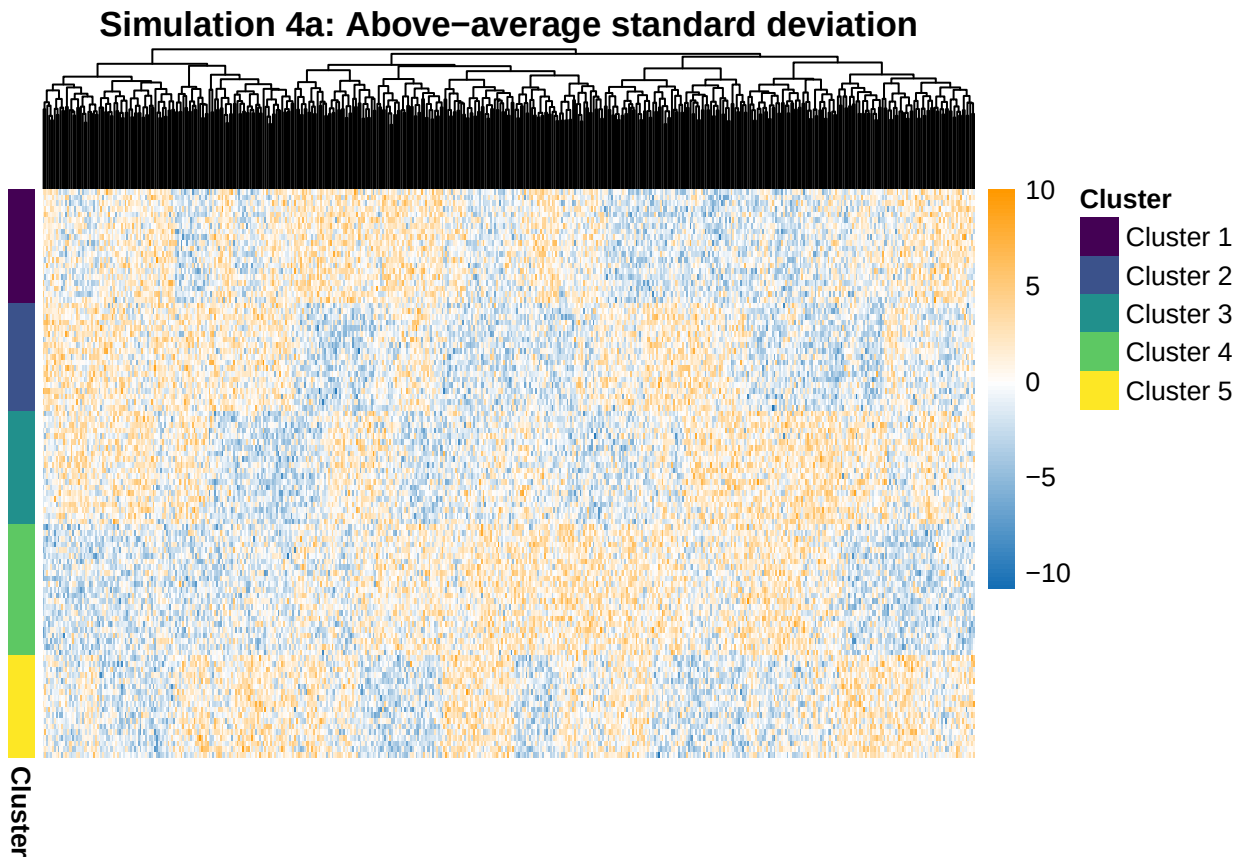
Simulation 4: Increasing σ^2

We test the ratio of μ to σ^2 required for structure to be successfully uncovered in a feature-rich dataset.

Simulation 4a: Large, informative dataset, medium σ^2

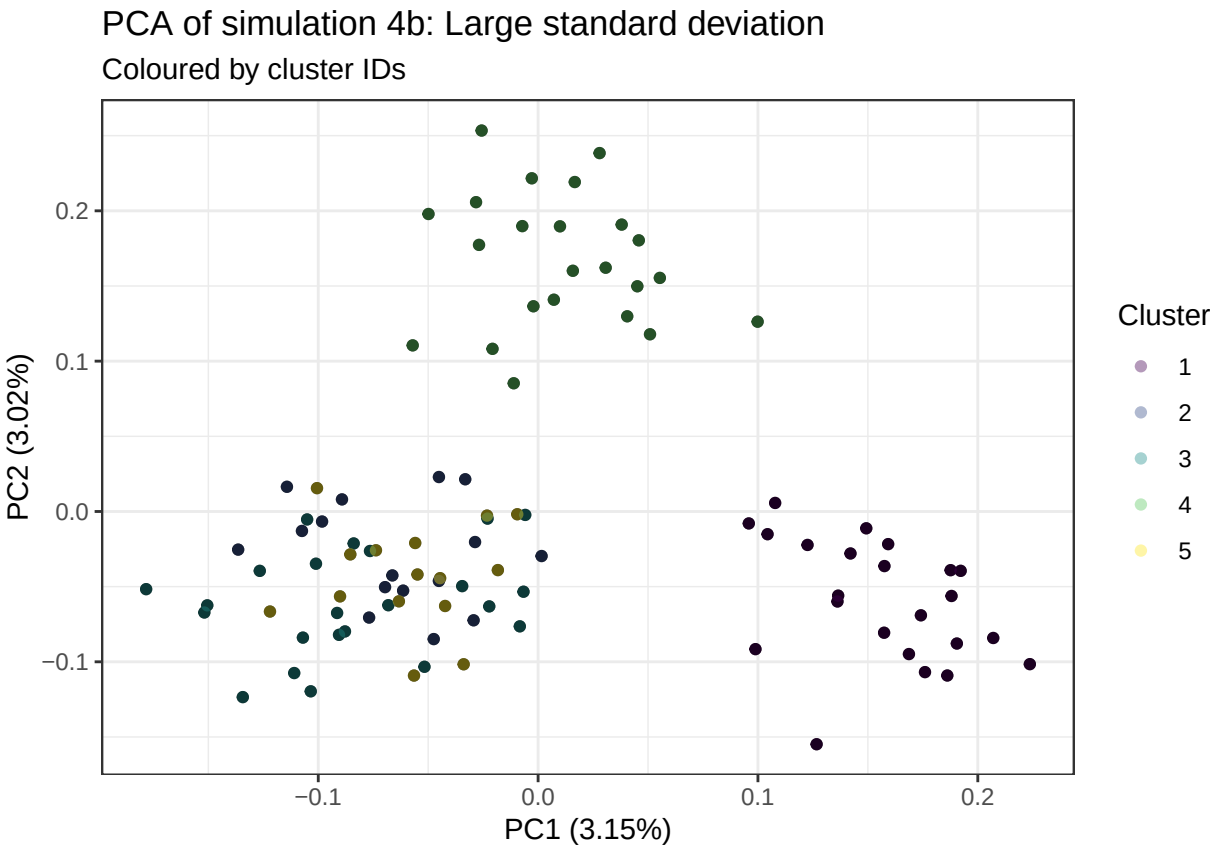
- $N = 100$;
- $P_s = 500$;
- $P_n = 0$;
- $K = 5$;
- $\pi = vec(\frac{1}{5})$;
- $\Delta_\mu = 1$; and
- $\sigma^2_{kp} = 5$.

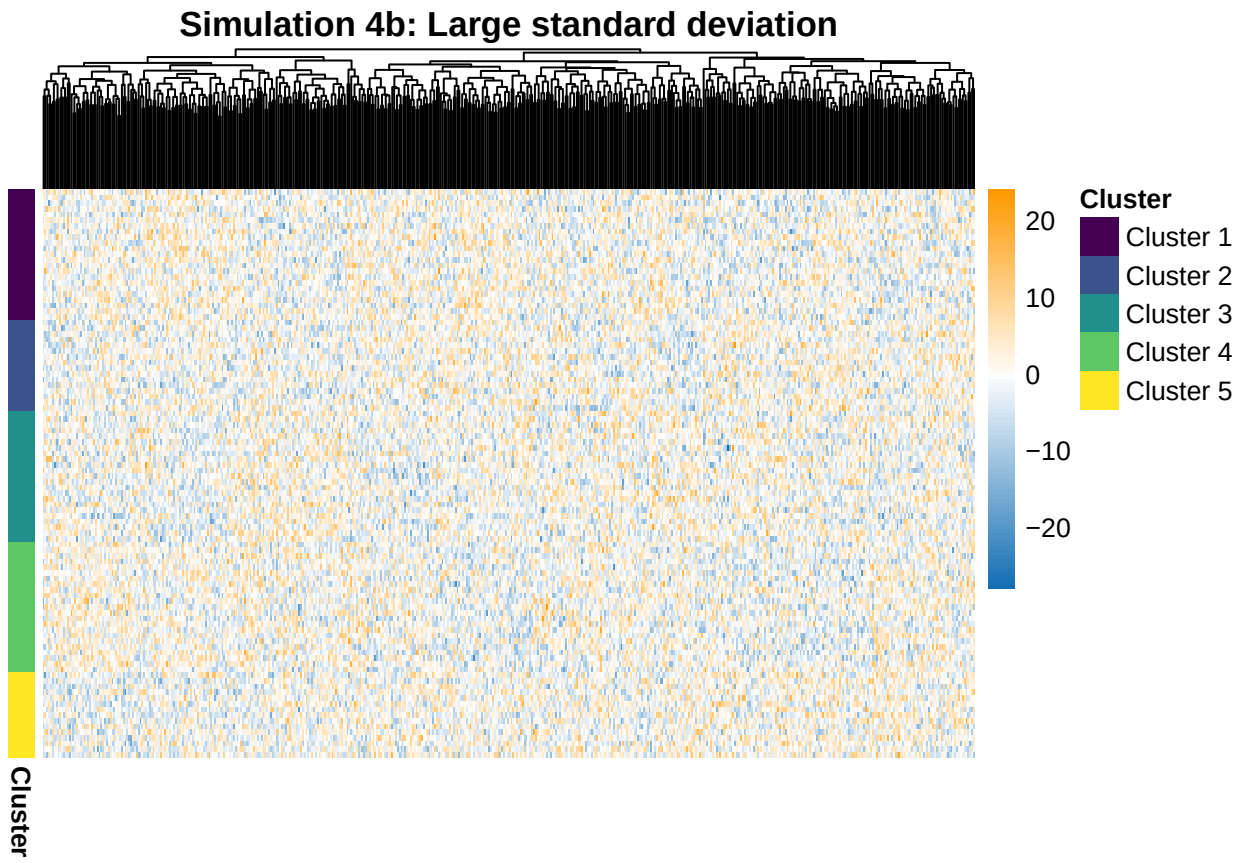

PCA of simulation 4a: Above−average standard deviation
Coloured by cluster IDs
PC2 (8%)
PC1 (8.81%)
0.1
0.0
−0.1
−0.2
−0.1
0.0
0.1
Cluster
1
2
3
4
5



Simulation 4b: Large, informative dataset, larger σ^2

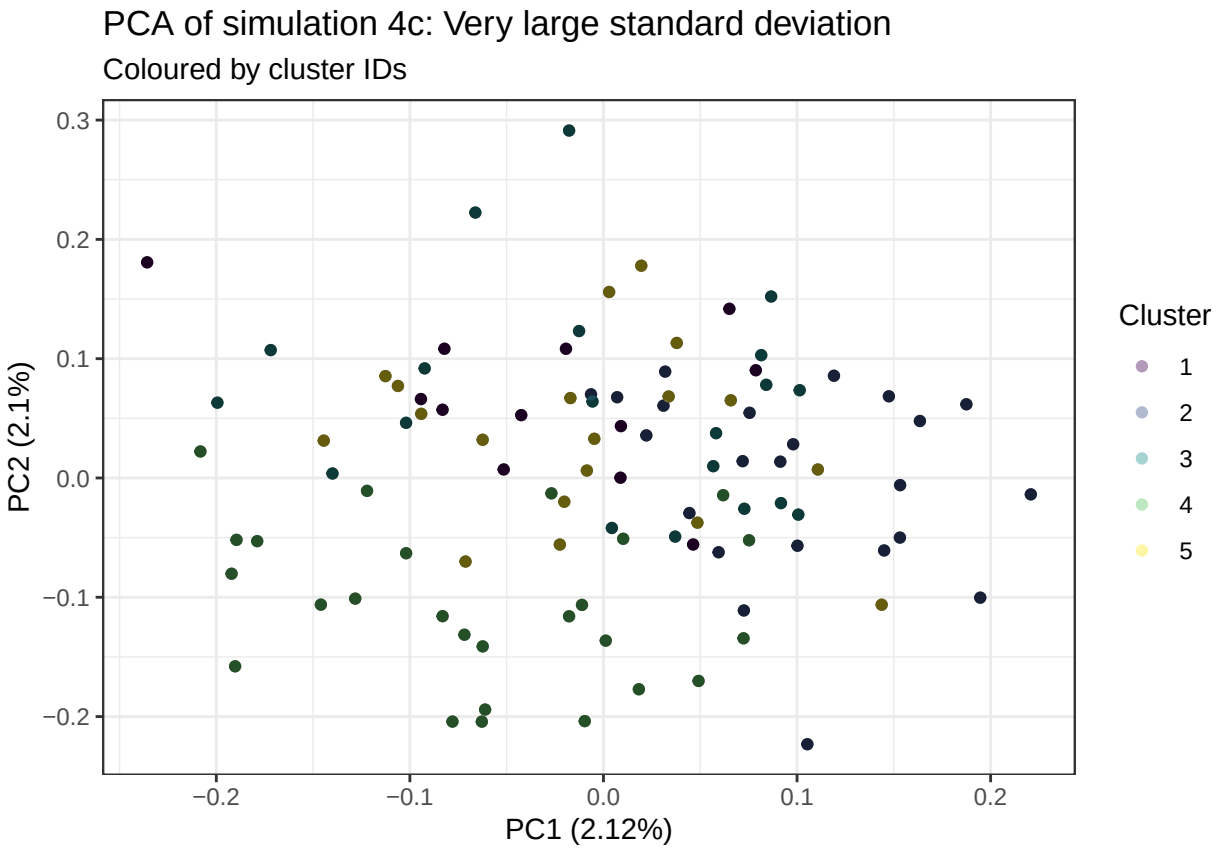
- $N = 100$;
- $P_s = 500$;
- $P_n = 0$;
- $K = 5$;
- $\pi = vec(\frac{1}{5})$;
- $\Delta_\mu = 1$; and
- $\sigma^2_{kp} = 30$.

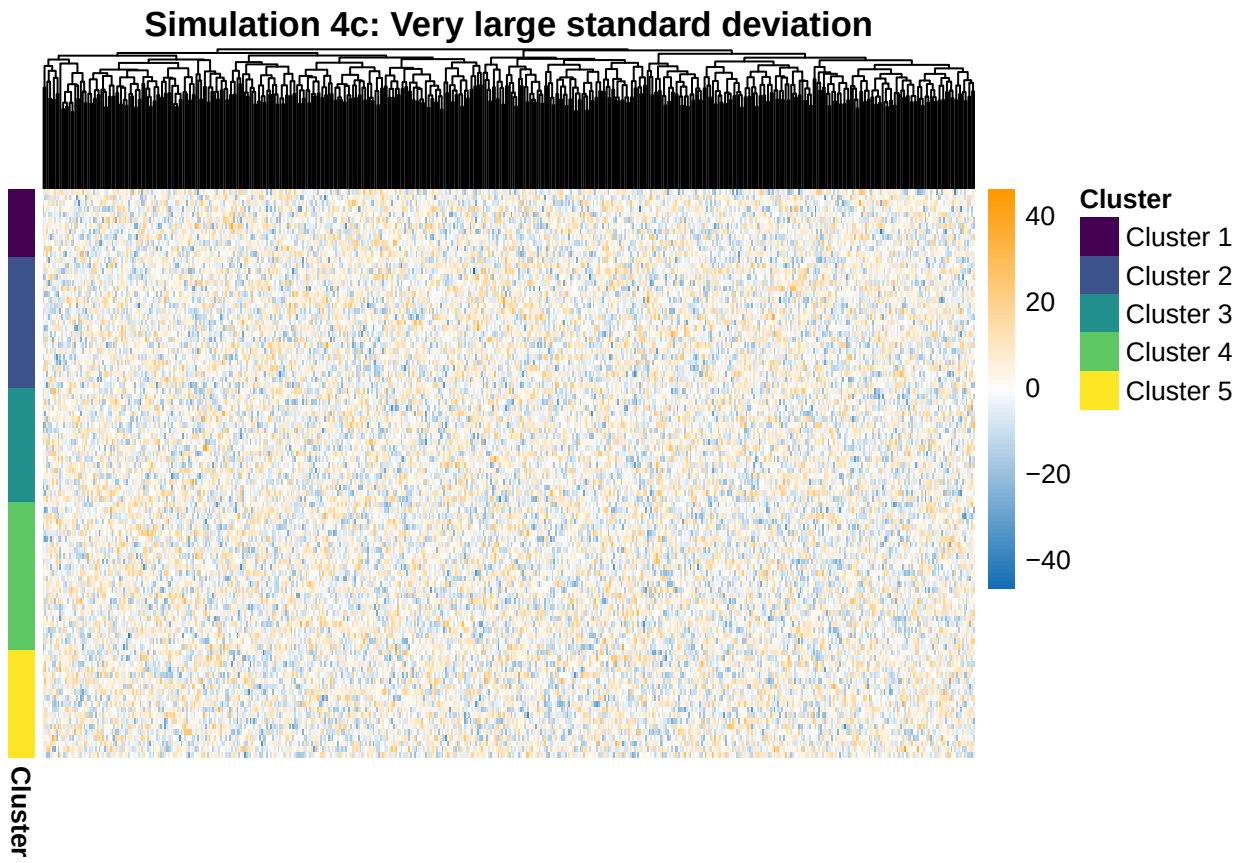

PCA of simulation 4b: Large standard deviation
Coloured by cluster IDs
PC2 (3.02%)
PC1 (3.15%)
0.2
0.1
0.0
−0.1
−0.1
0.0
0.1
0.2
Cluster
1
2
3
4
5

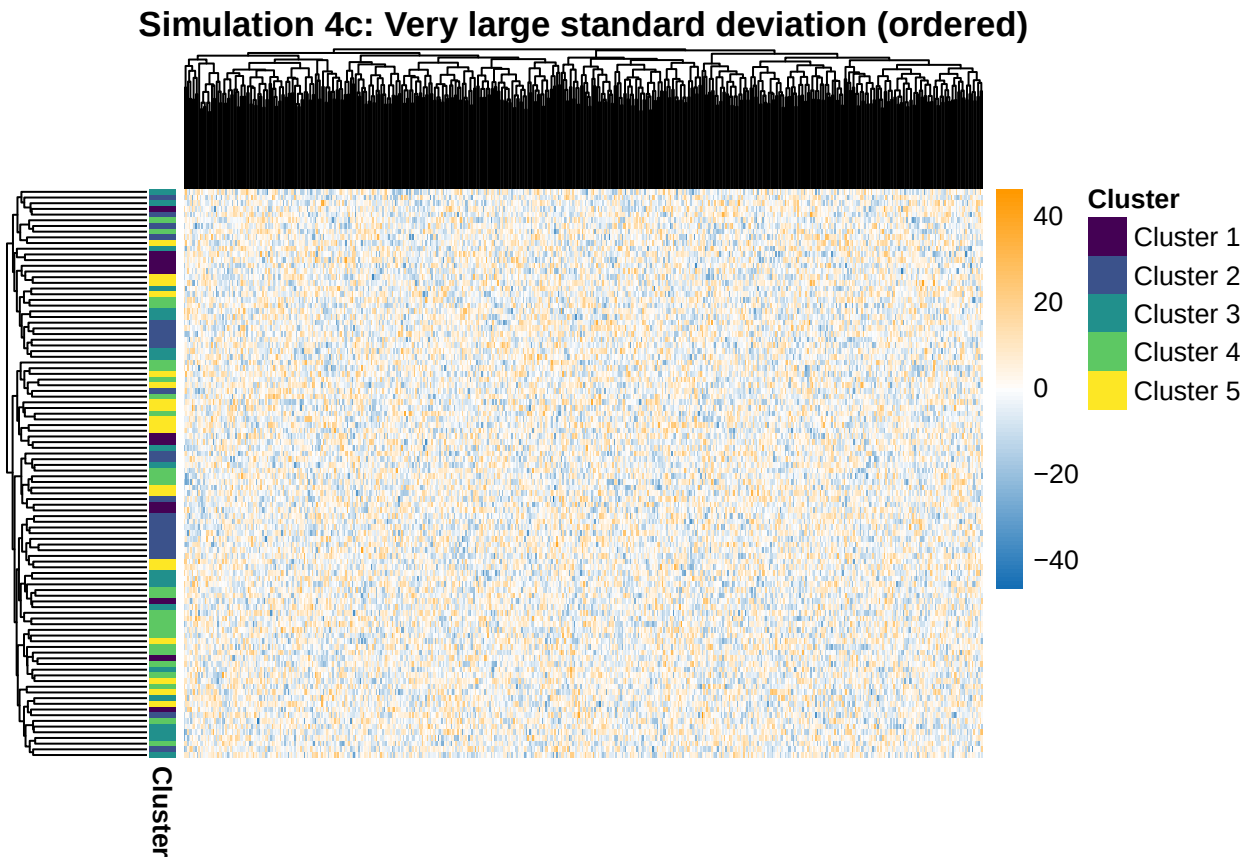


Simulation 4c: Large, informative dataset, largest σ^2

- $N = 100$;
- $P_s = 500$;
- $P_n = 0$;
- $K = 5$;
- $\pi = vec(\frac{1}{5})$;
- $\Delta_\mu = 1$; and
- $\sigma^2_{kp} = 100$.


PCA of simulation 4c: Very large standard deviation
Coloured by cluster IDs
PC2 (2.1%)
PC1 (2.12%)
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.2
−0.1
0.0
0.1
0.2
Cluster
1
2
3
4
5


Simulation 4c: Very large standard deviation
Cluster
40
20
0
−20
−40
Cluster
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5



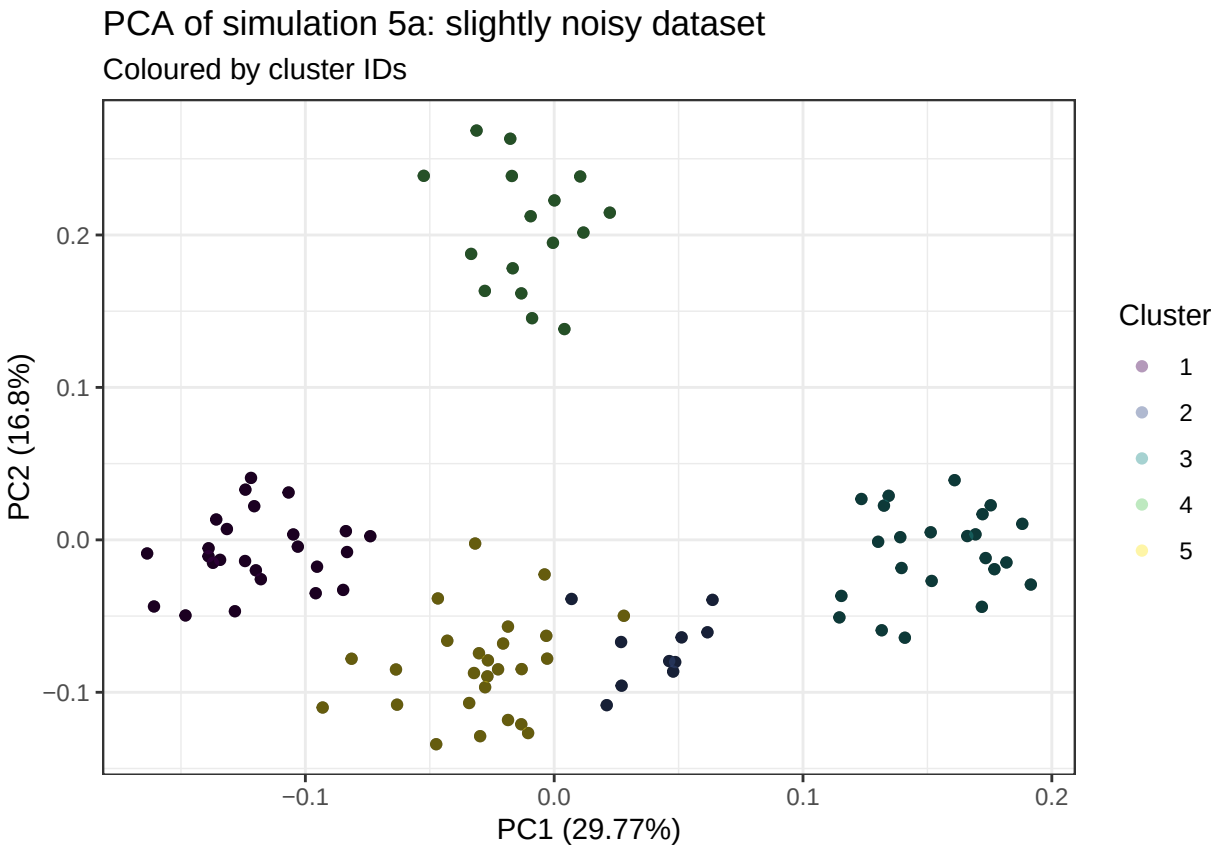
Simulation 5: Large, noisy dataset

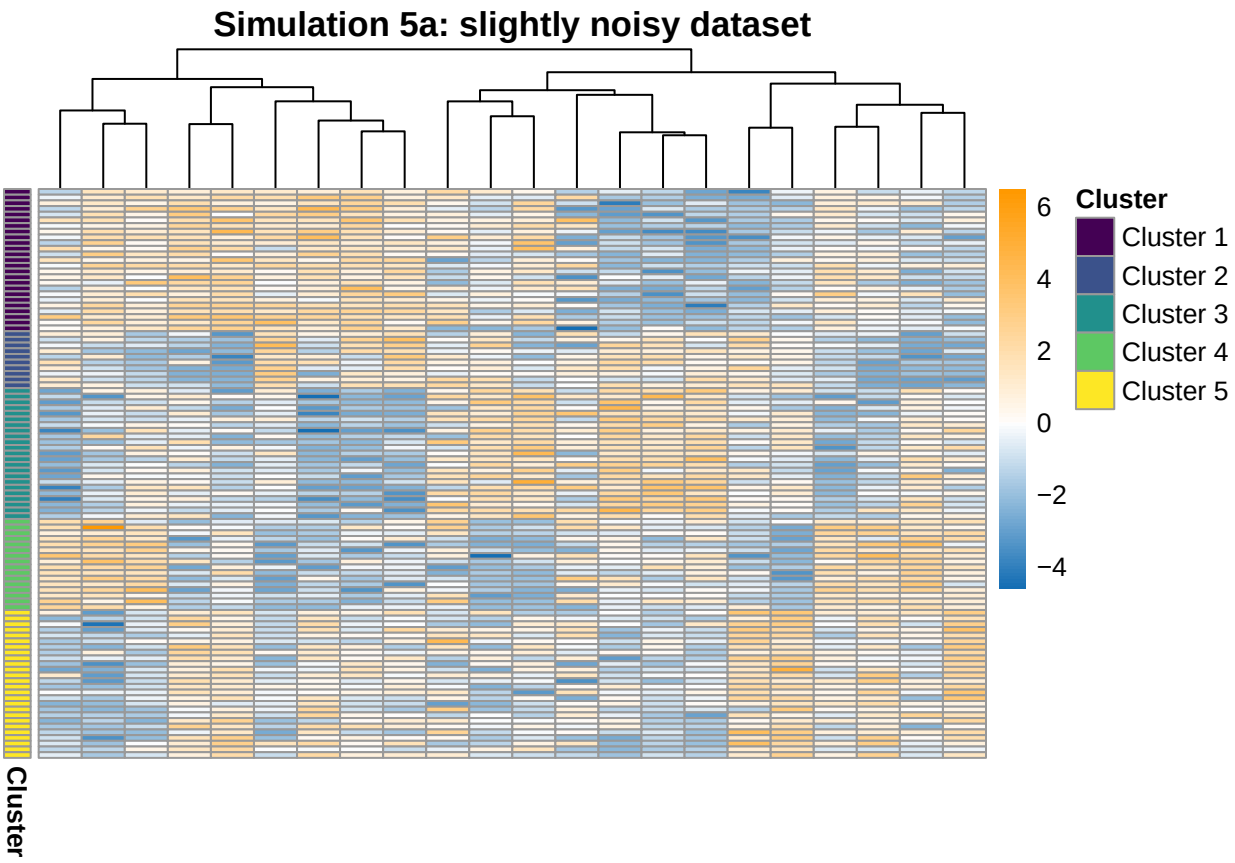
This case is intended to test how well structure can be uncovered as P_n increases. We test for:

$$P_n = 0.1 \times P_s$$
$$P_n = 0.5 \times P_s$$
$$P_n = P_s$$
$$P_n = 5 \times P_s$$
$$P_n = 10 \times P_s$$

Simulation 5a: Large, slightly noisy dataset

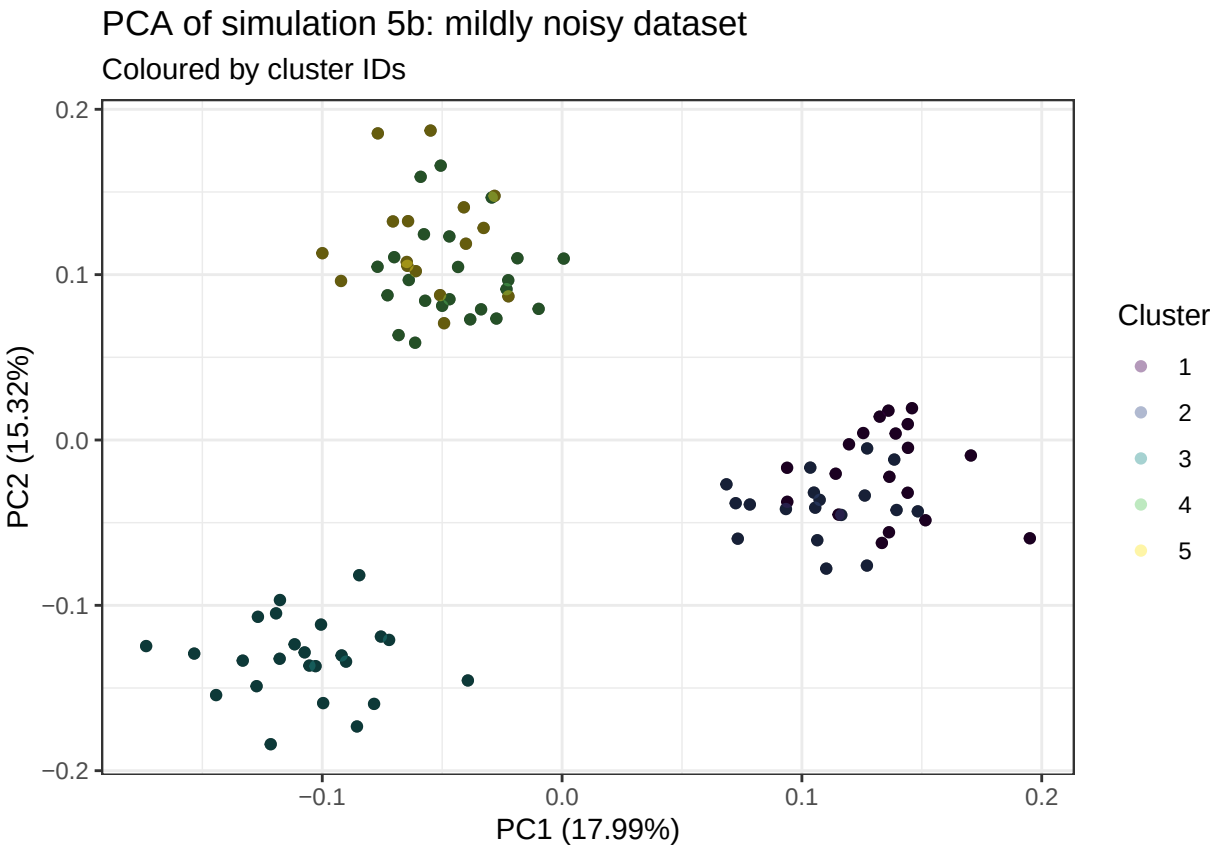
- $N = 100$;
- $P_s = 20$;
- $P_n = 2$;
- $K = 5$;
- $\pi = vec(\frac{1}{5})$;
- $\Delta_\mu = 1$; and
- $\sigma^2_{kp} = 1$.

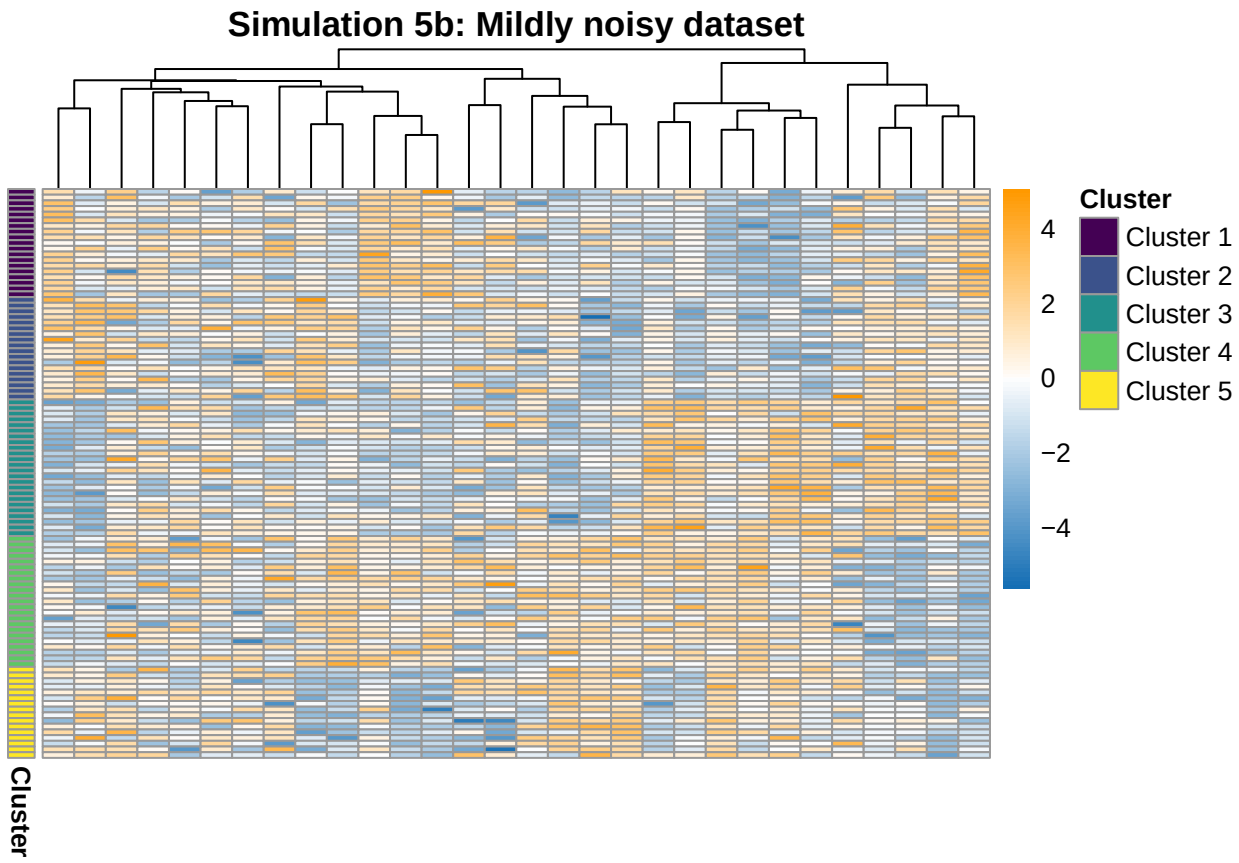

PCA of simulation 5a: slightly noisy dataset
Coloured by cluster IDs
PC2 (16.8%)
PC1 (29.77%)
0.2
0.1
0.0
−0.1
−0.1
0.0
0.1
0.2
Cluster
1
2
3
4
5



Simulation 5b: Large, mildly noisy dataset

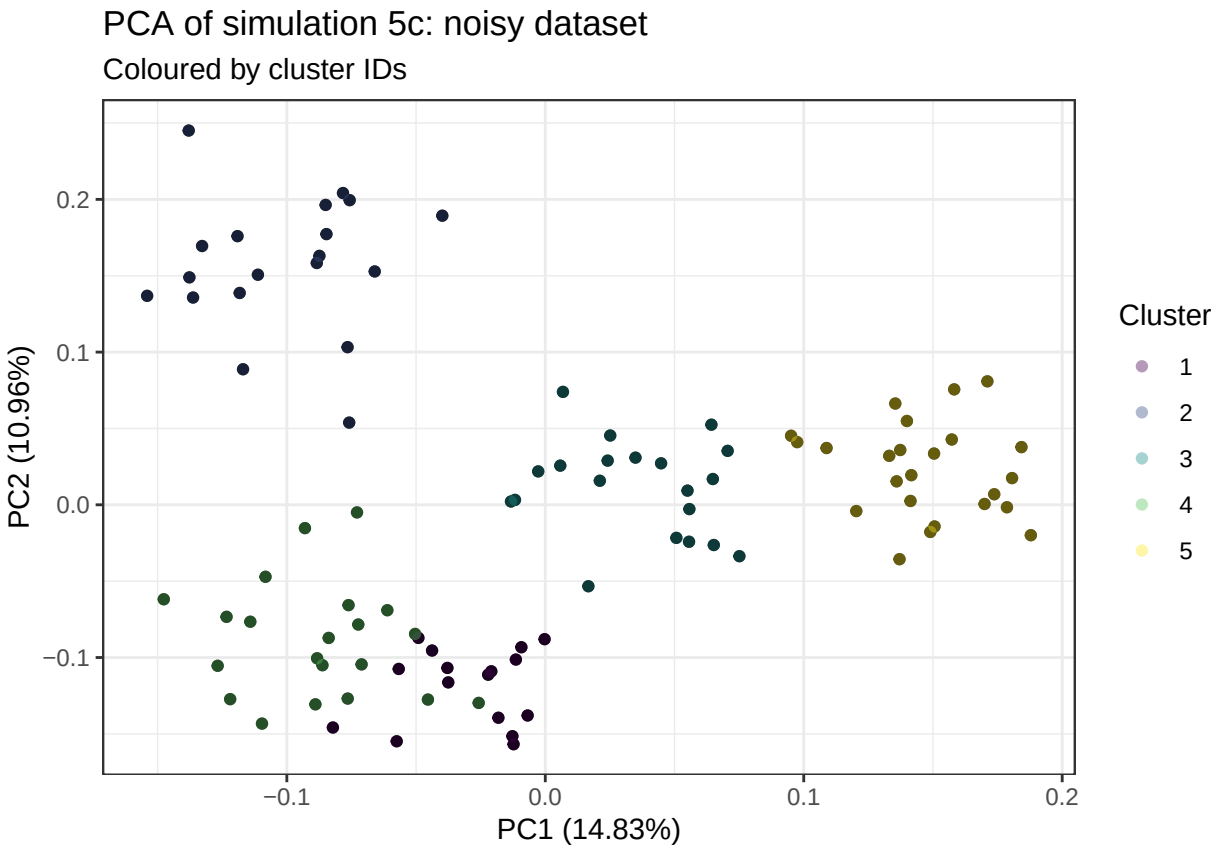
- $N = 100$;
- $P_s = 20$;
- $P_n = 10$;
- $K = 5$;
- $\pi = vec(\frac{1}{5})$;
- $\Delta_\mu = 1$; and
- $\sigma^2_{kp} = 1$.

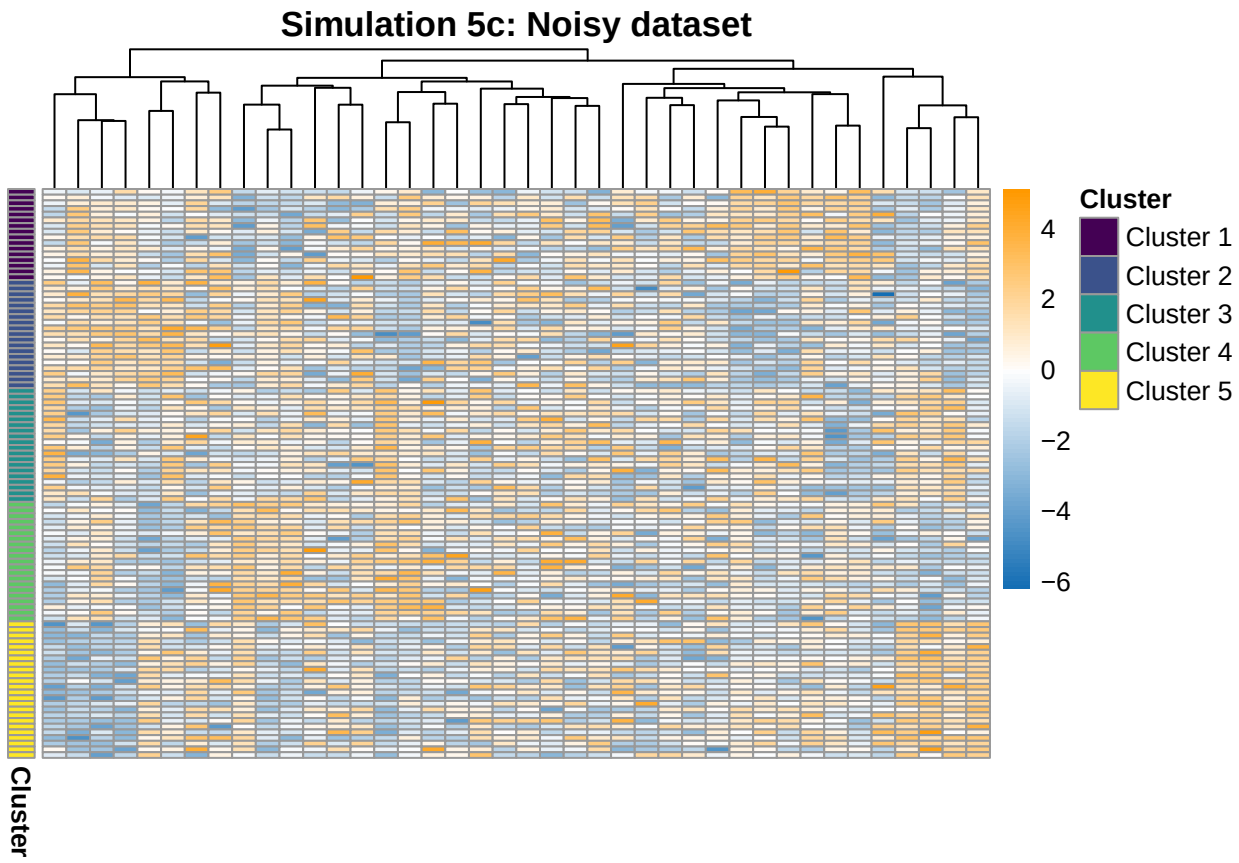

PCA of simulation 5b: mildly noisy dataset
Coloured by cluster IDs
PC2 (15.32%)
PC1 (17.99%)
0.2
0.1
0.0
−0.1
−0.2
−0.1
0.0
0.1
0.2
Cluster
1
2
3
4
5



Simulation 5c: Large, noisy dataset

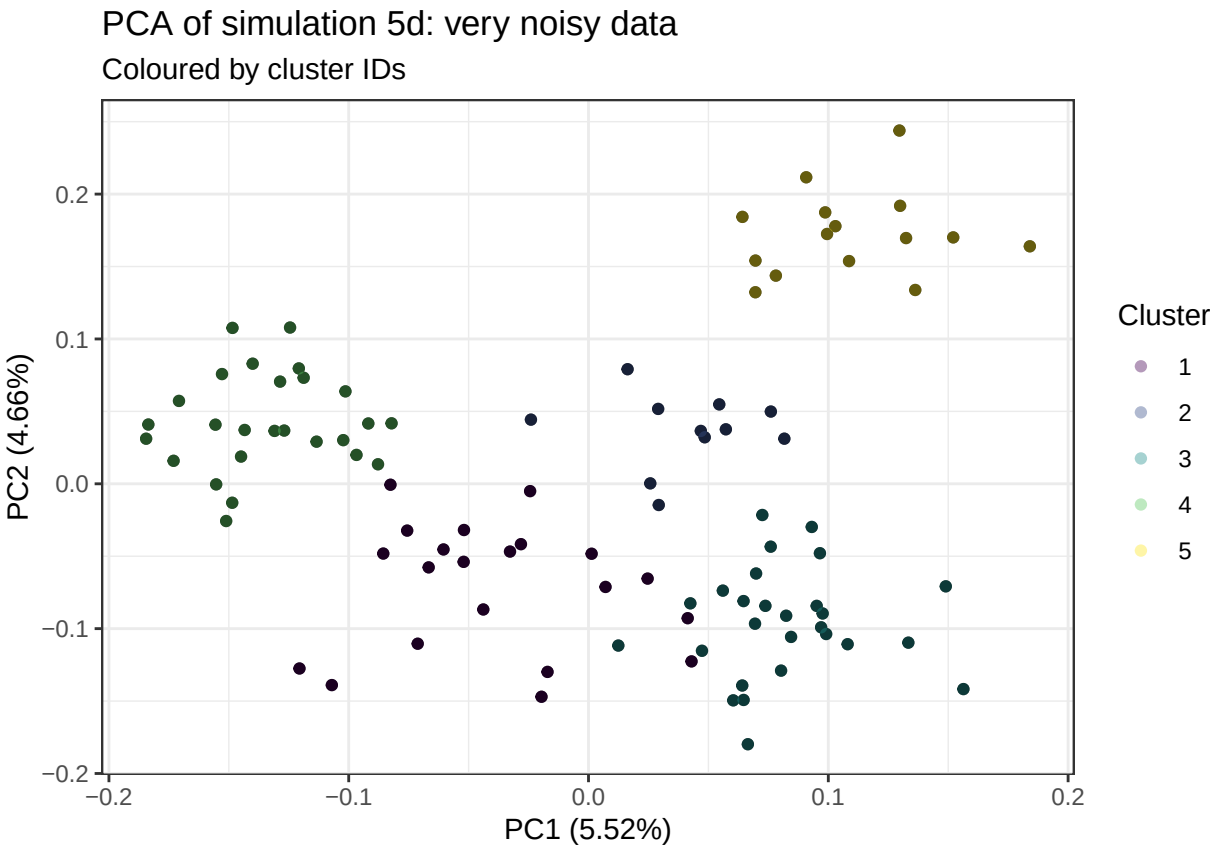
- $N = 100$;
- $P_s = 20$;
- $P_n = 20$;
- $K = 5$;
- $\pi = vec(\frac{1}{5})$;
- $\Delta_\mu = 1$; and
- $\sigma^2_{kp} = 1$.

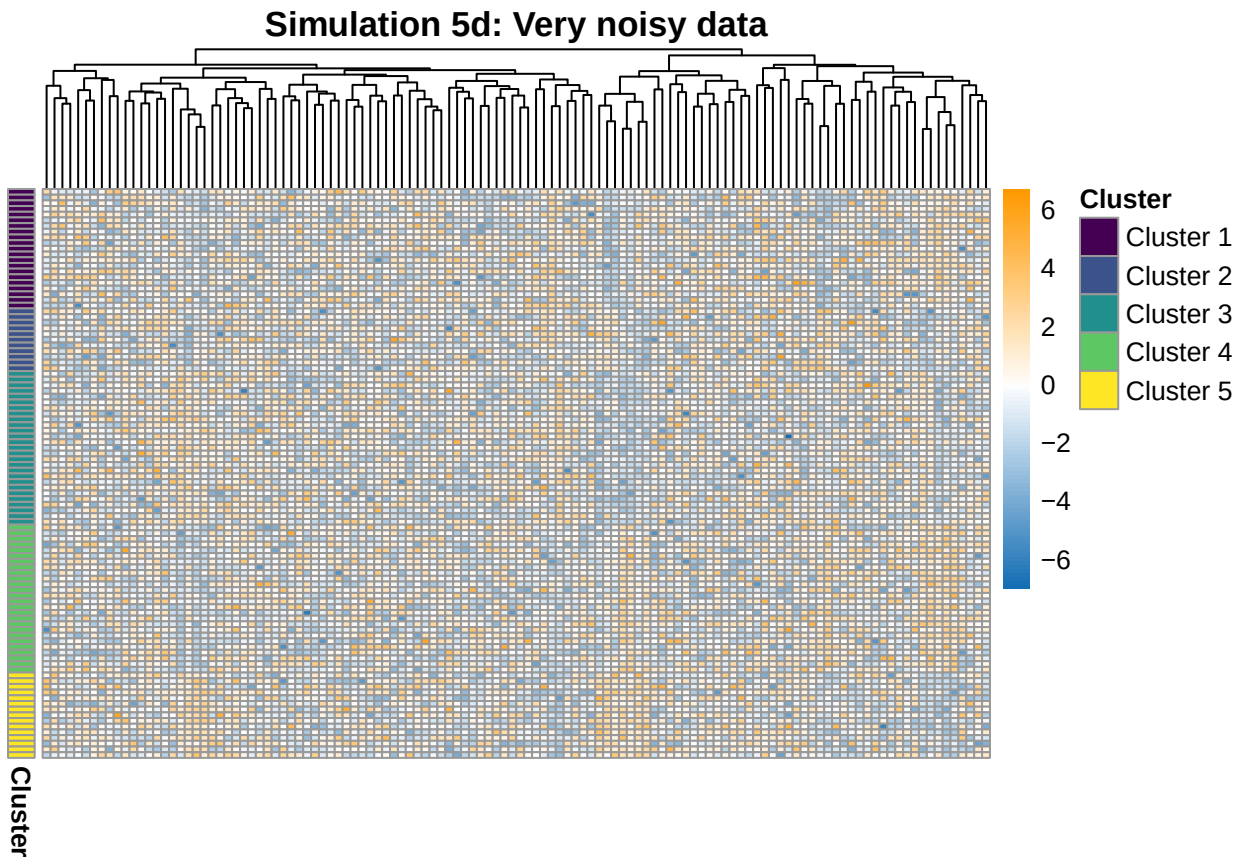

PCA of simulation 5c: noisy dataset
Coloured by cluster IDs
PC2 (10.96%)
PC1 (14.83%)
0.2
0.1
0.0
−0.1
−0.1
0.0
0.1
0.2
Cluster
1
2
3
4
5



Simulation 5d: Large, very noisy dataset

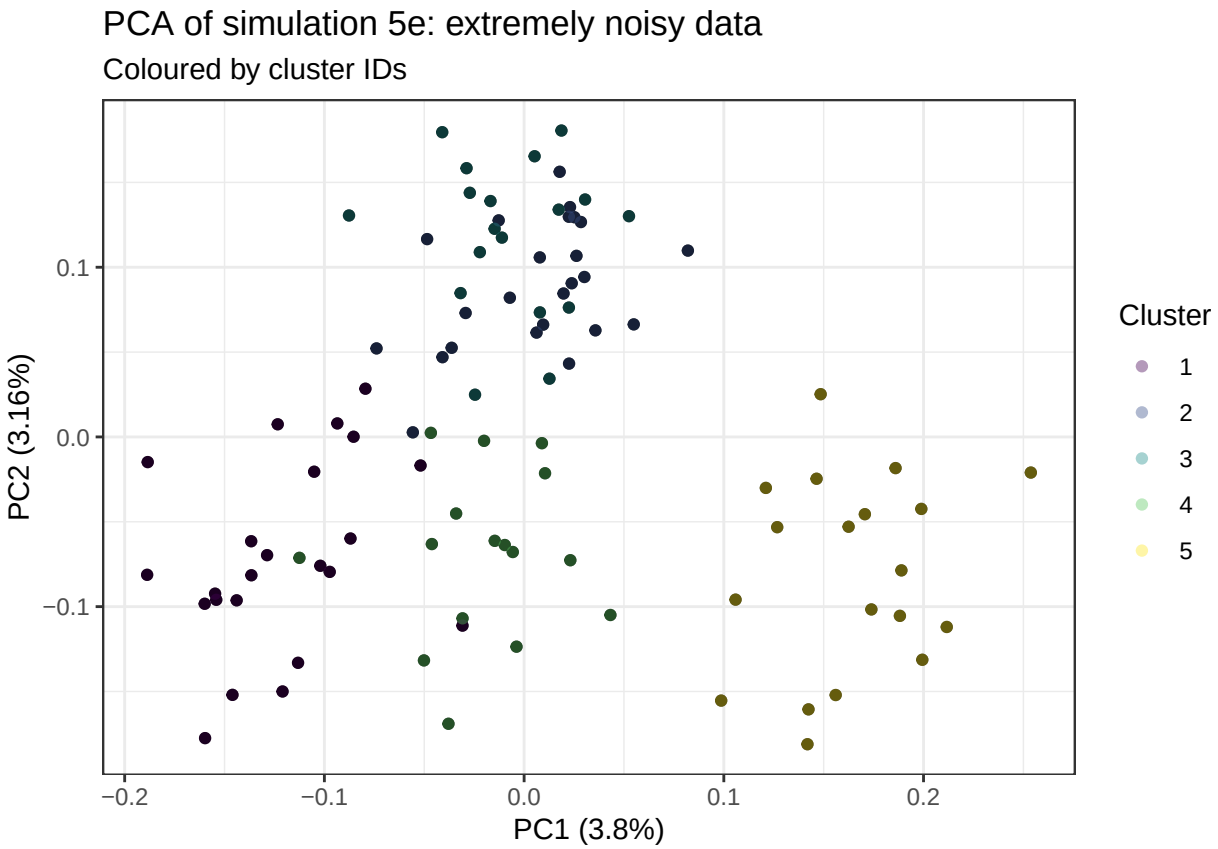
- $N = 100$;
- $P_s = 20$;
- $P_n = 100$;
- $K = 5$;
- $\pi = vec(\frac{1}{5})$;
- $\Delta_\mu = 1$; and
- $\sigma^2_{kp} = 1$.

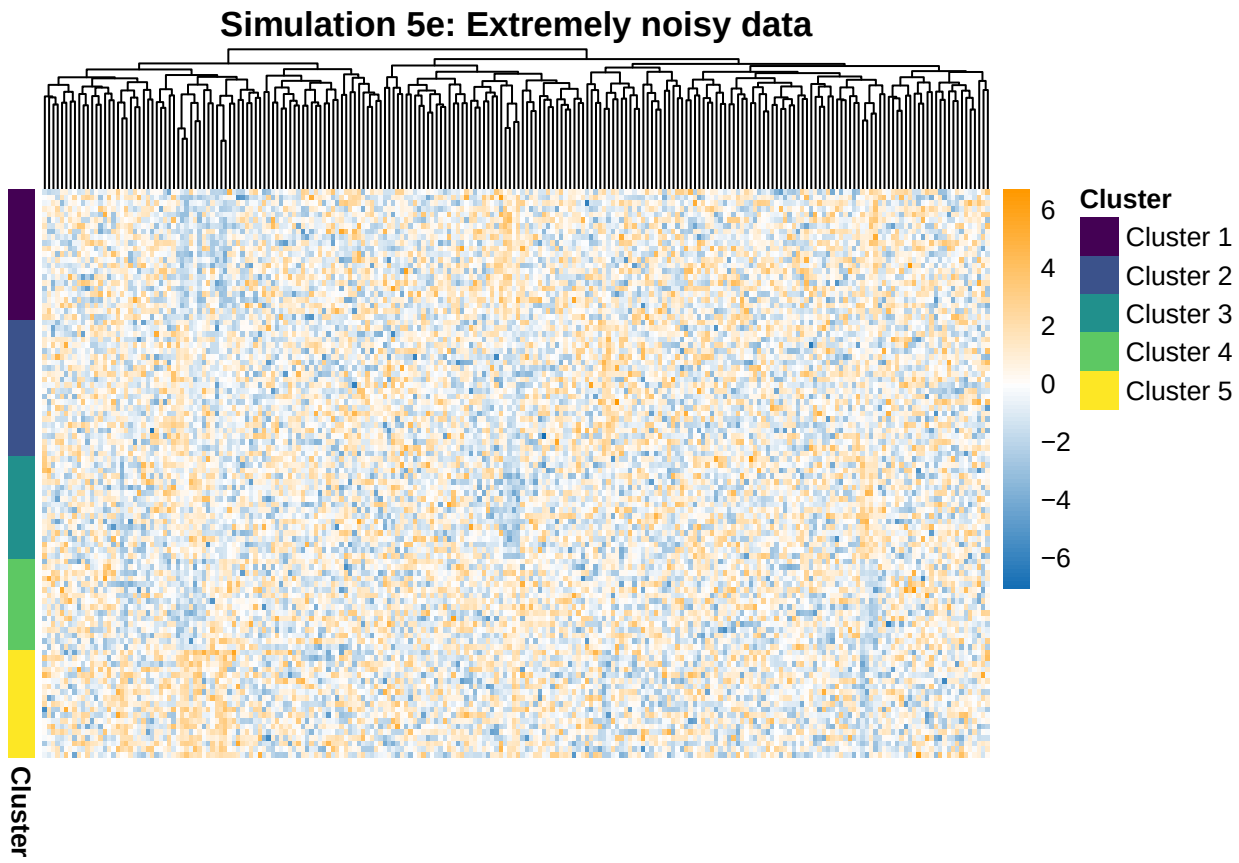

PCA of simulation 5d: very noisy data
Coloured by cluster IDs
PC2 (4.66%)
PC1 (5.52%)
0.2
0.1
0.0
−0.1
−0.2
−0.2
−0.1
0.0
0.1
0.2
Cluster
1
2
3
4
5

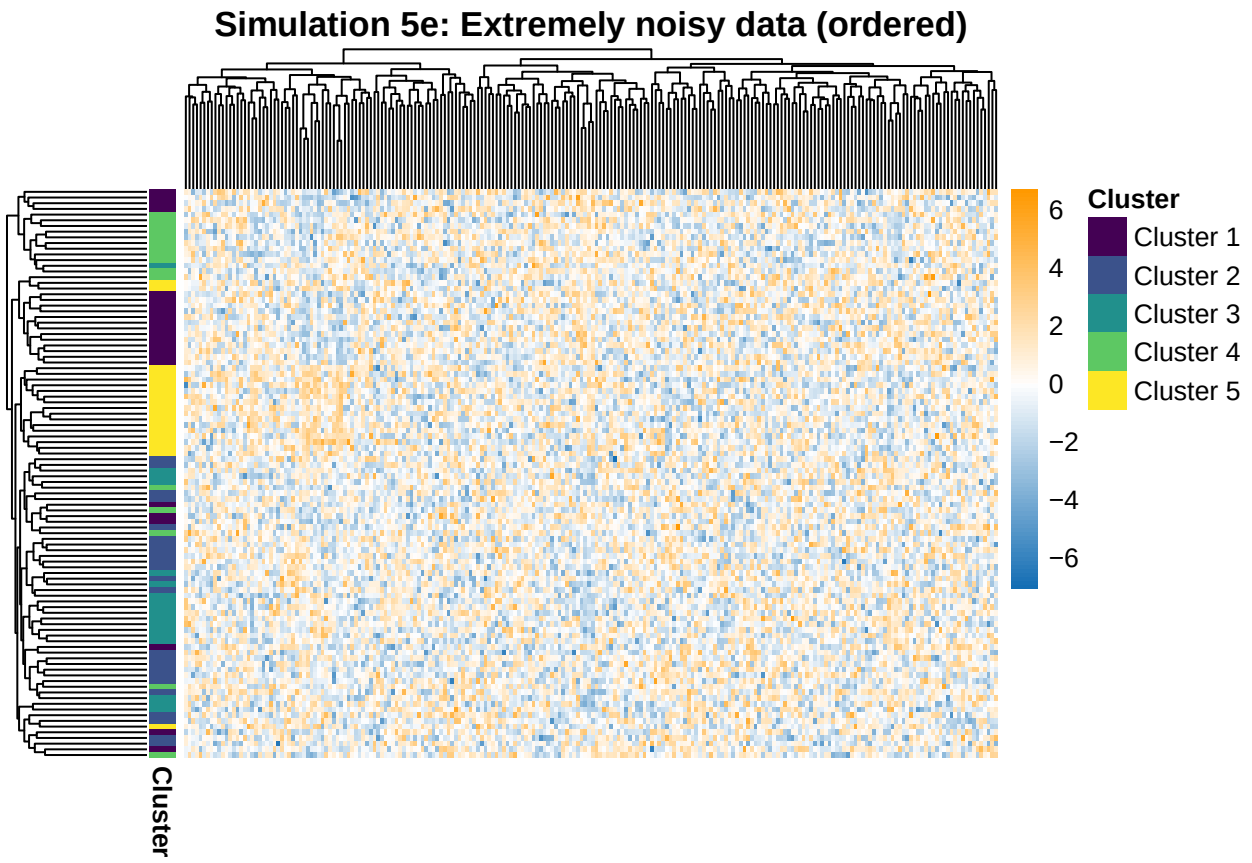


Simulation 5e: Large, extremely noisy dataset

- $N = 100$;
- $P_s = 20$;
- $P_n = 200$;
- $K = 5$;
- $\pi = vec(\frac{1}{5})$;
- $\Delta_\mu = 1$; and
- $\sigma^2_{kp} = 1$.


PCA of simulation 5e: extremely noisy data
Coloured by cluster IDs
PC2 (3.16%)
0.1
0.0
−0.1
−0.2
−0.1
0.0
0.1
0.2
PC1 (3.8%)
Cluster
1
2
3
4
5


Simulation 5e: Extremely noisy data
Cluster
6
4
2
0
−2
−4
−6
Cluster
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5



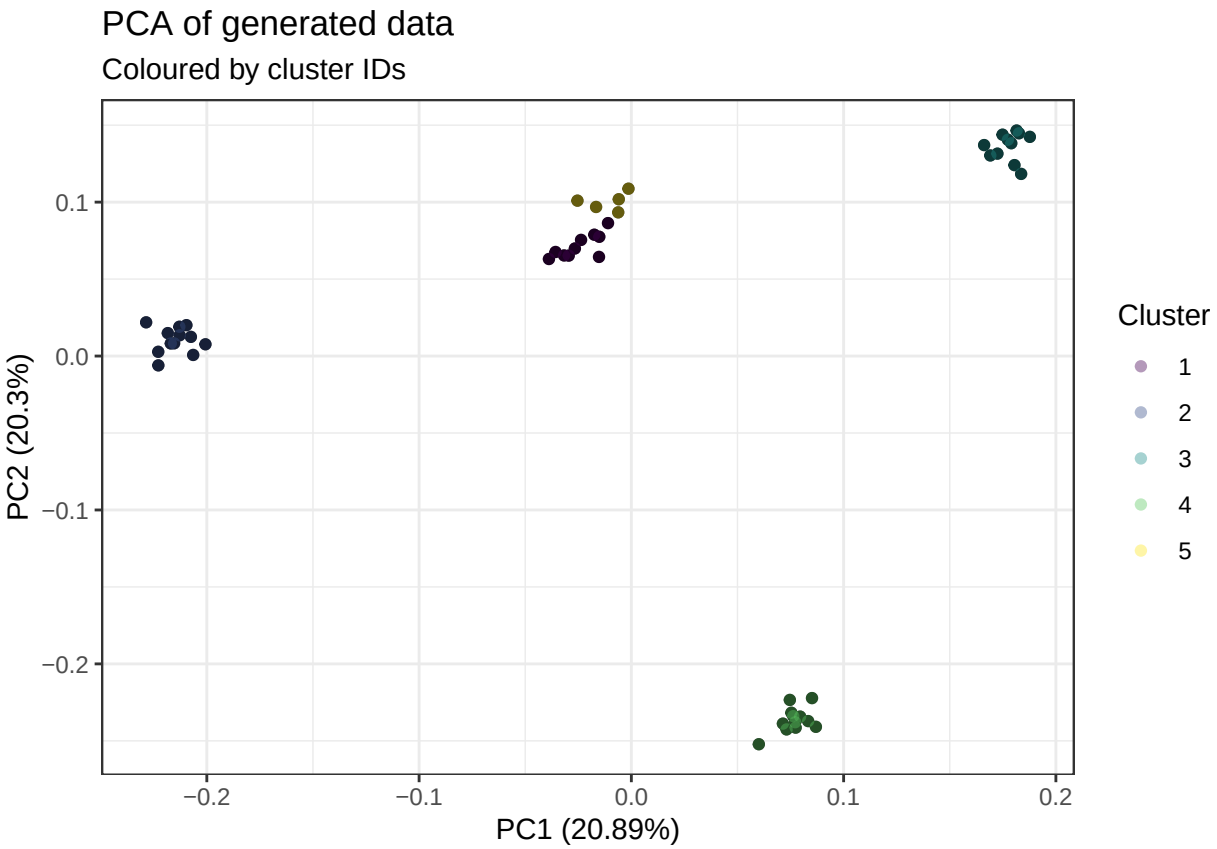
Simulation 6: Small N, large P

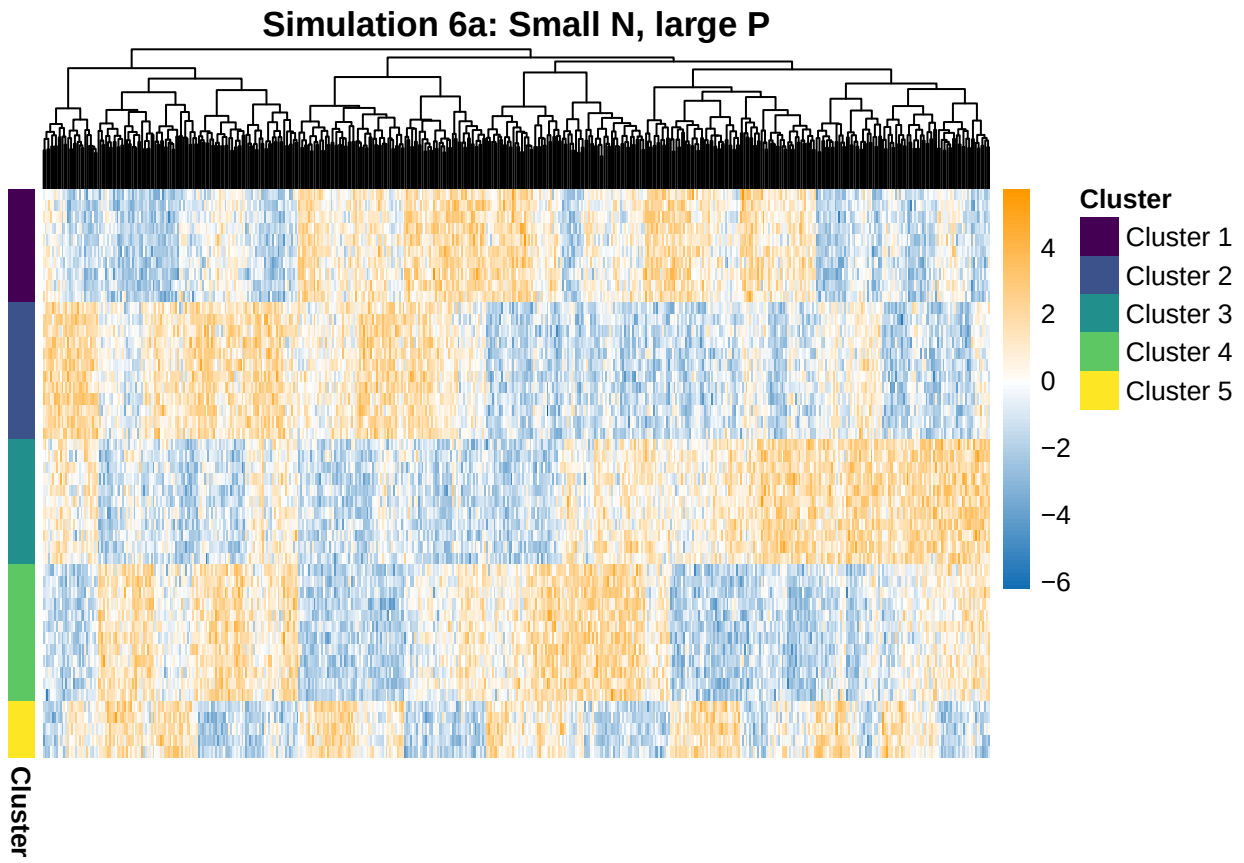
To test the paradigm that gene expression data often fits into. If the features are relevant than as P grows so too does the resolution of the clustering and thus structure can be discovered even as Δ_μ decreases.

Simulation 6a: Small N, large P dataset

We test the small N, large P case.

- $N = 50$;
- $P_s = 500$;
- $P_n = 0$;
- $K = 5$;
- $\pi = vec(\frac{1}{5})$;
- $\Delta_\mu = 1$; and
- $\sigma^2_{kp} = 1$.

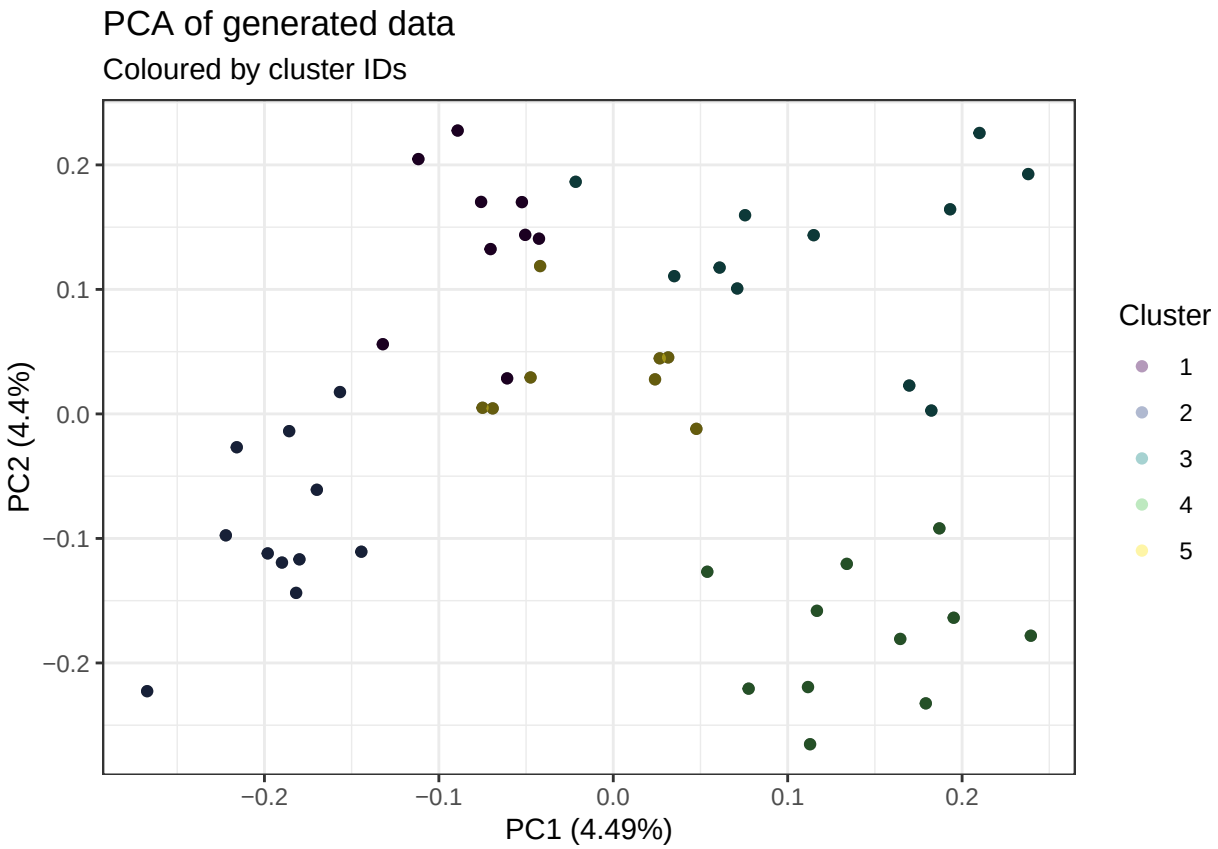

PCA of generated data
Coloured by cluster IDs
PC2 (20.3%)
PC1 (20.89%)
0.1
0.0
−0.1
−0.2
−0.2
−0.1
0.0
0.1
0.2
Cluster
1
2
3
4
5

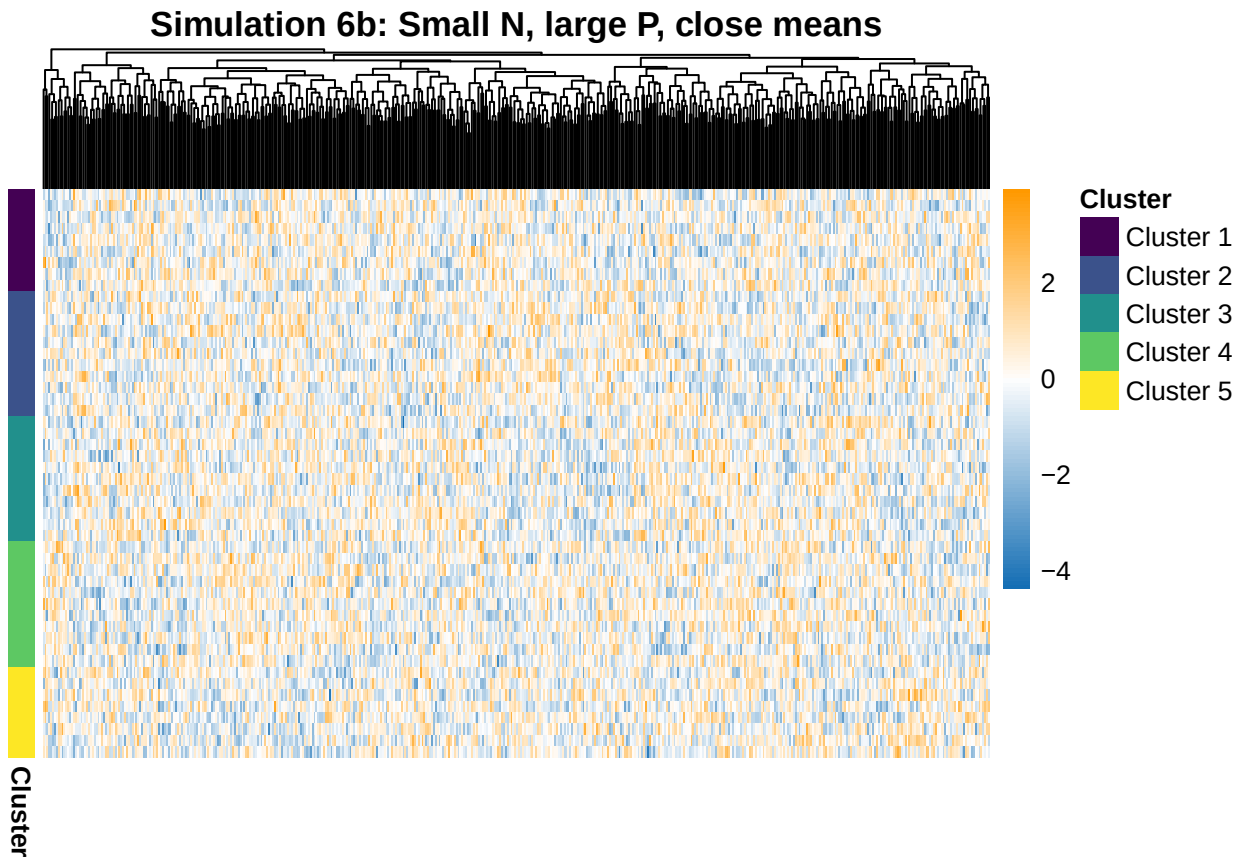


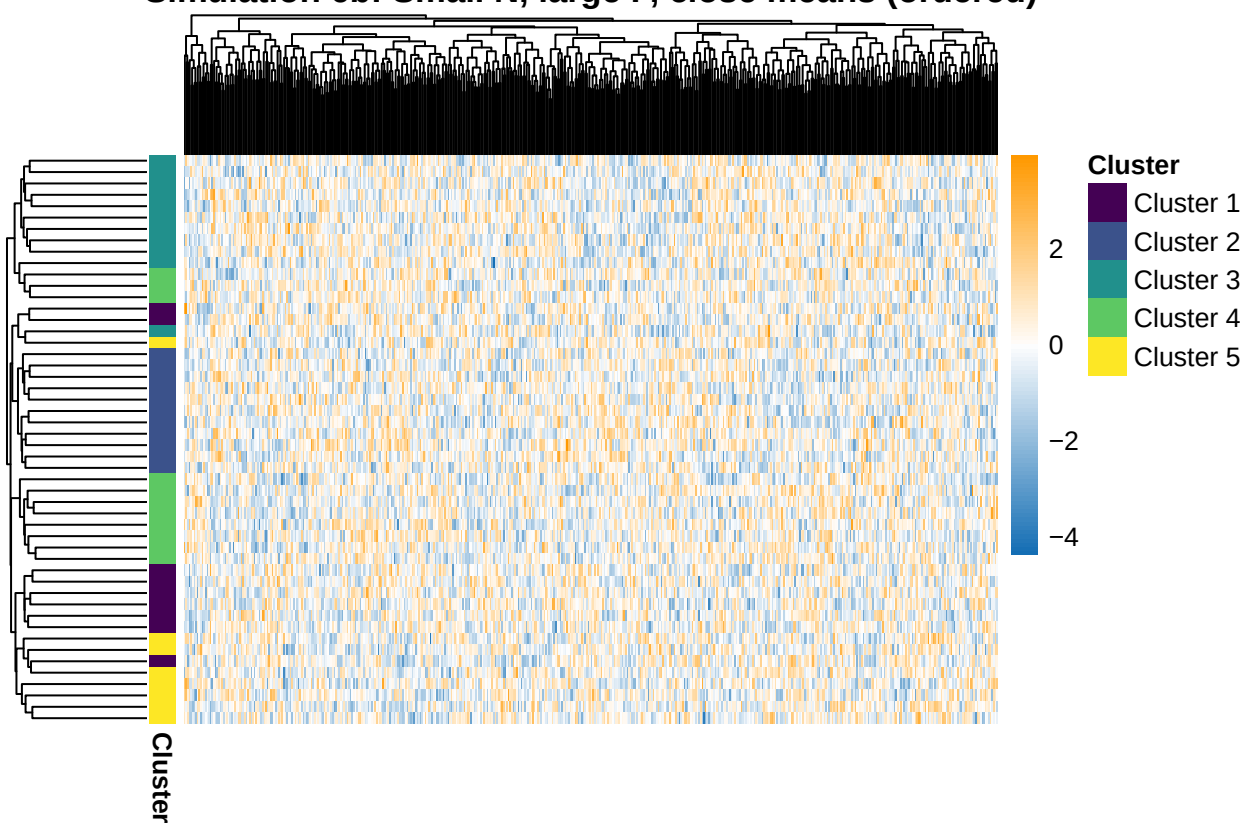
Simulation 6b: Small N, large P, close means dataset

We test how small the Δ_μ can be for the clustering structure to still be possible to resolve. To a degree, the level to which Δ_μ can shrink without structure disappearing is a function of both P_s.

- $N = 50$;
- $P_s = 500$;
- $P_n = 0$;
- $K = 5$;
- $\pi = vec(\frac{1}{5})$;
- $\Delta_\mu = 0.2$; and
- $\sigma^2_{kp} = 1$.

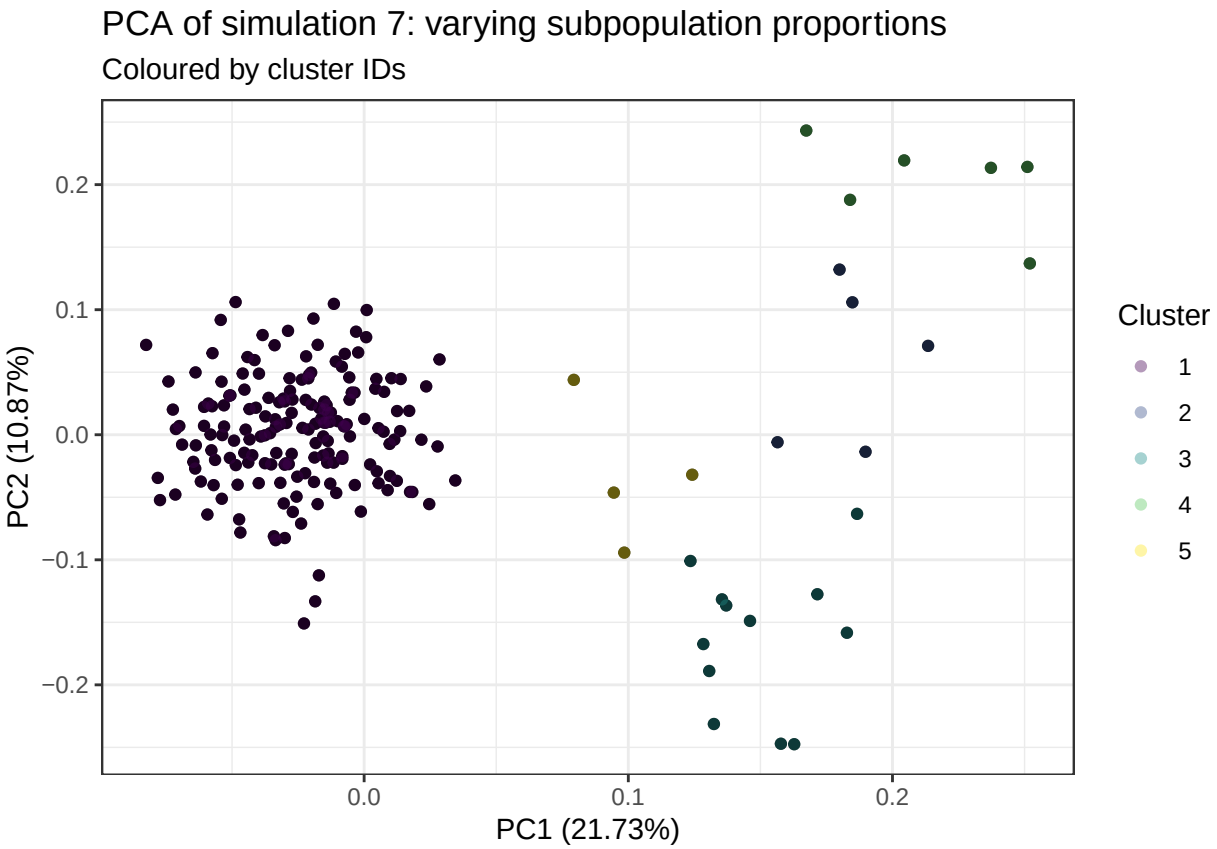

PCA of generated data
Coloured by cluster IDs
PC2 (4.4%)
PC1 (4.49%)
0.2
0.1
0.0
−0.1
−0.2
−0.2
−0.1
0.0
0.1
0.2
Cluster
1
2
3
4
5

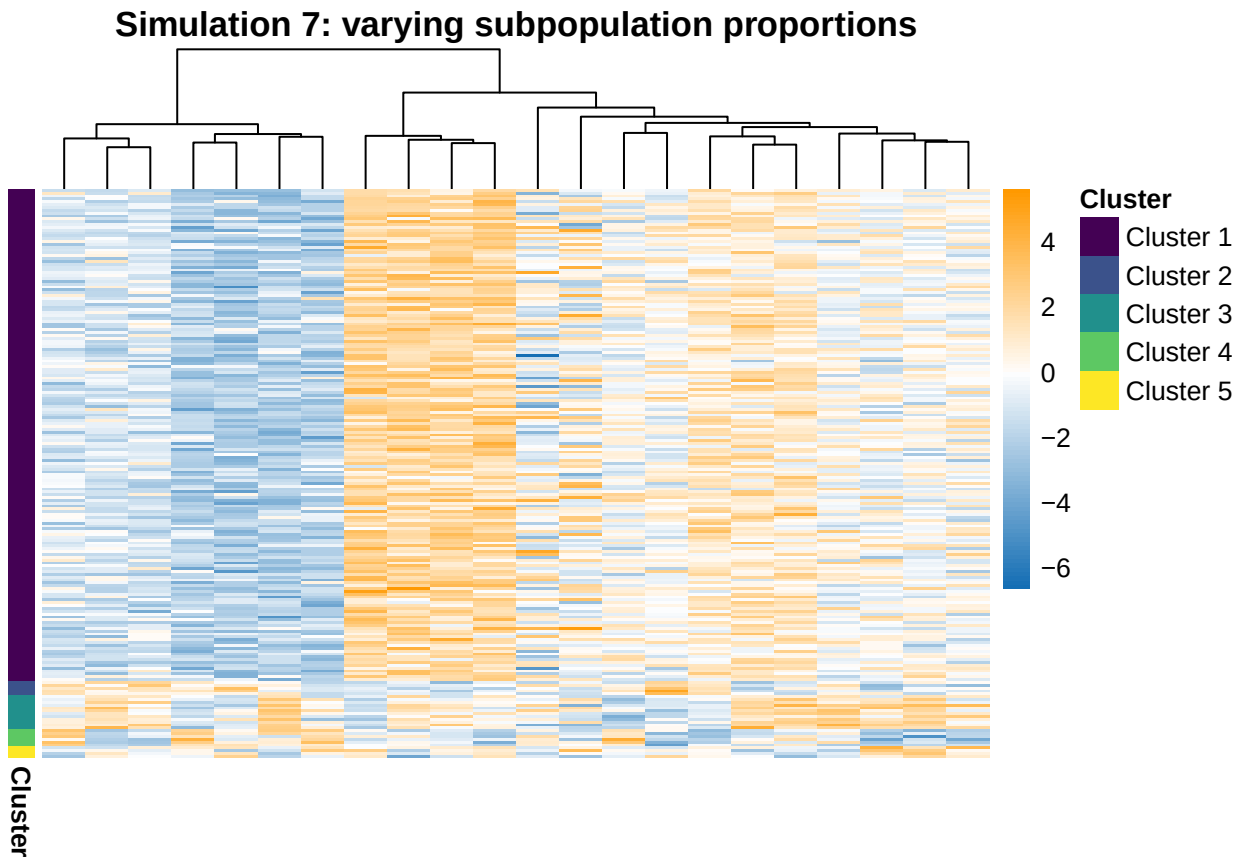

Simulation 6b: Small N, large P, close means
Cluster
2
0
−2
−4
Cluster
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5



Simulation 7: varying subpoputlation proportions

- $N = 200$;
- $P_s = 20$;
- $P_n = 0$;
- $K = 5$;
- $\pi \sim Dirichlet(0.5)$;
- $\Delta_\mu = 1$; and
- $\sigma^2_{kp} = 1$.


PCA of simulation 7: varying subpopulation proportions
Coloured by cluster IDs
PC2 (10.87%)
PC1 (21.73%)
0.2
0.1
0.0
−0.1
−0.2
0.0
0.1
0.2
Cluster
1
2
3
4
5



References

Law, Martin H, Anil K Jain, and Mário Figueiredo. 2003. "Feature Selection in Mixture-Based Clustering." In *Advances in Neural Information Processing Systems*, 641–48.