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:

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

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

$$p(x_i|\theta,\pi,\Phi) = \sum_{k=1}^{K} \pi_k \prod_{p=1}^{P} f(x_i|\theta_{kp})^{\phi_p} g(x_i|\lambda_p)^{(1-\phi_p)}, \tag{2}$$

where $f(\cdot)$ describes the Gaussian pdf and thus $\theta = (\mu, \sigma^2)$. 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 number of components being modeled;
- π: the proportion of points associated with 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
- *alpha*: 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, then I expect there to be some relationship of the nature

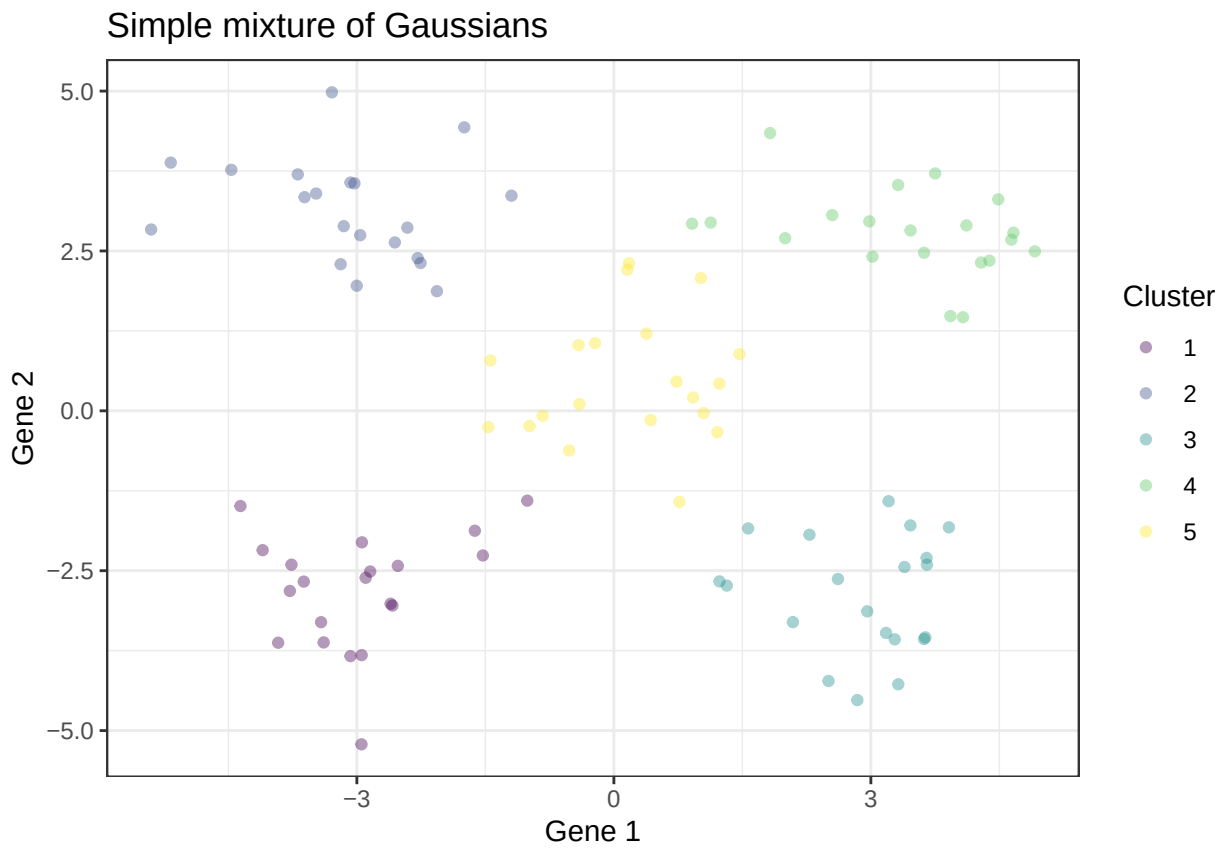
$$\begin{aligned} ARI(C^*, C') &\propto N \\ ARI(C^*, C') &\propto P_S \\ ARI(C^*, C') &\propto \frac{1}{K} \\ ARI(C^*, C') &\propto \Delta_\mu \\ ARI(C^*, C') &\propto \frac{1}{\sigma^2} \end{aligned}$$

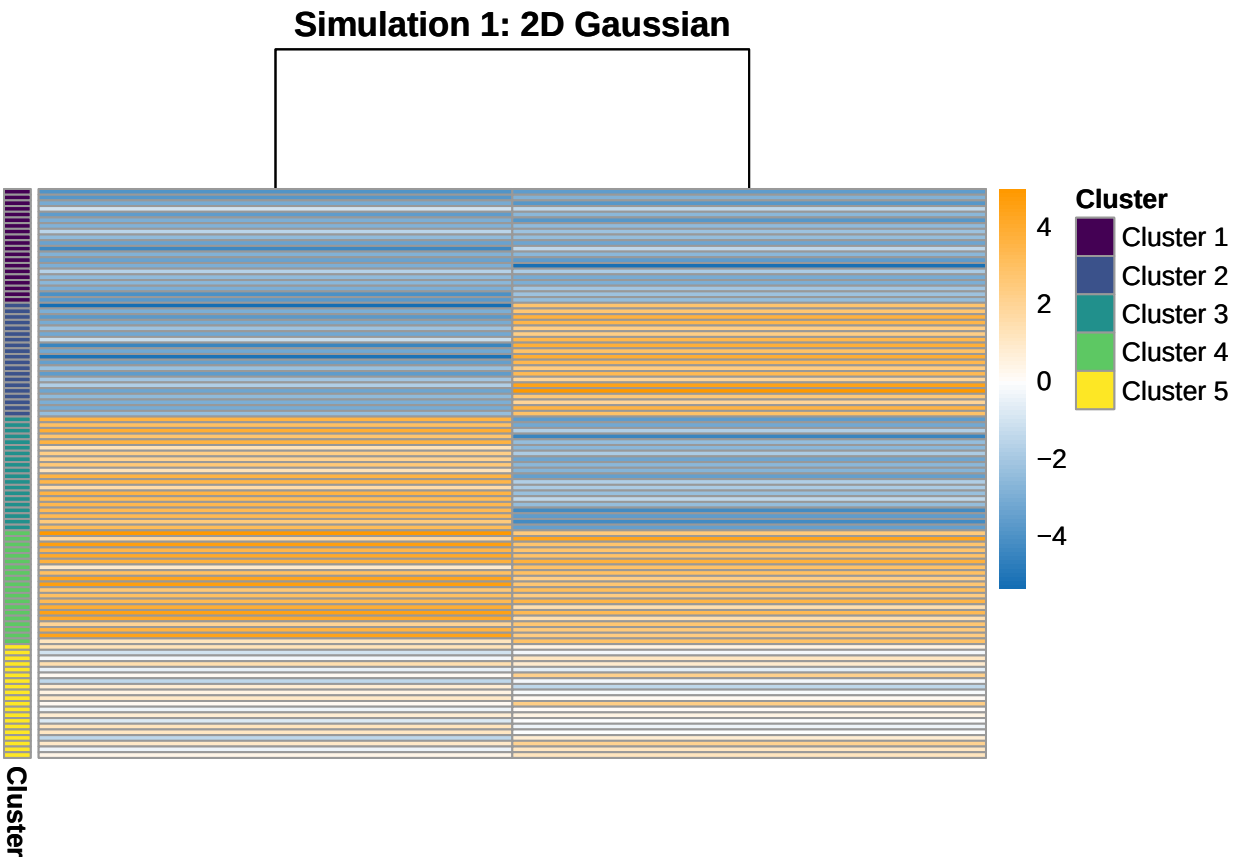
I do not expect that the linear nature of these relationships is true, but the directionality of these relationships is expected to hold. These expectations also convey that for many of the variables it is relative valus that matter; for instance if Δ_μ increases but σ^2 also grows proportionally, then we would not expect the clustering structure to resolve particularly well; similarly we expect that the number of items present contributes as a function of the number of clusters present.

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
- $\sigma^2_{kp} = 1$.

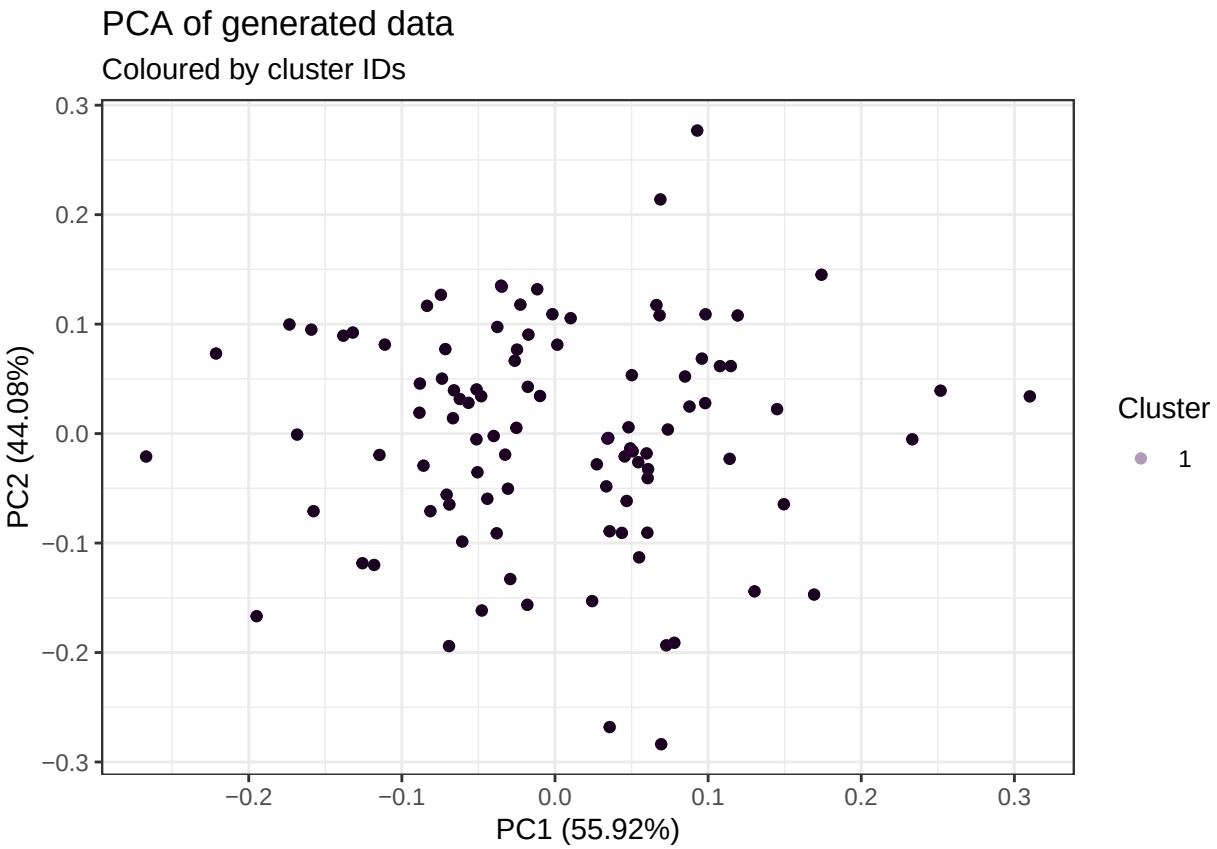

Simple mixture of Gaussians
5.0
2.5
0.0
−2.5
−5.0
Gene 2
−3
0
3
Gene 1
Cluster
1
2
3
4
5

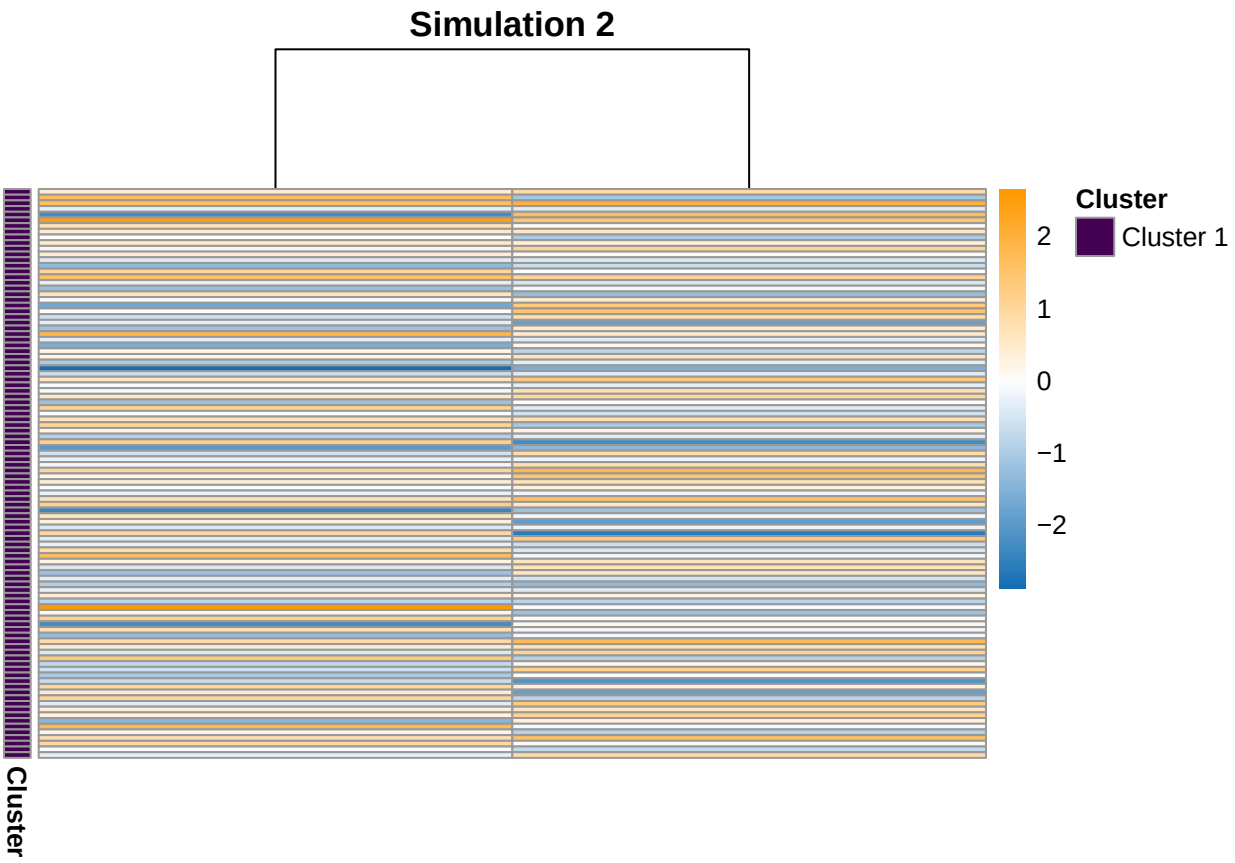


Simulation 2: No structure

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
- $\sigma^2_{kp} = 1$.

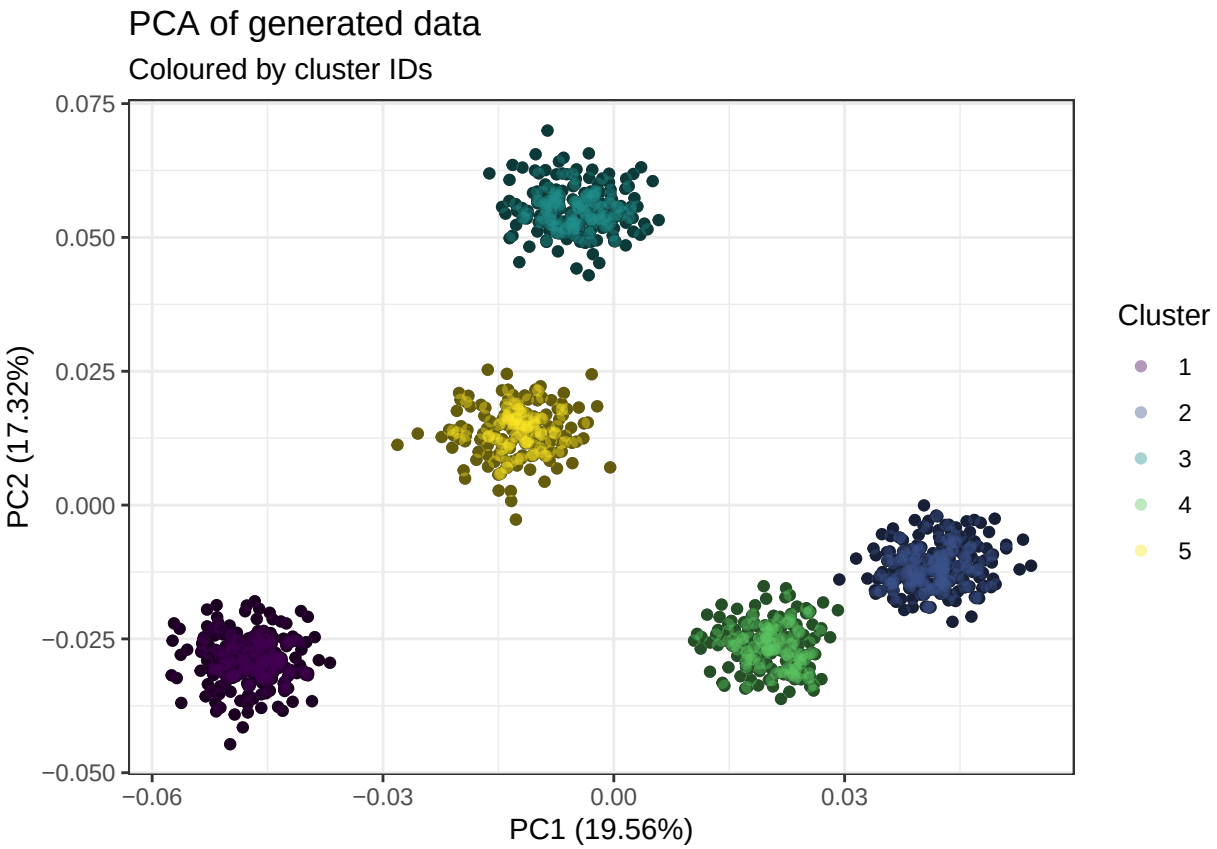

PCA of generated data
Coloured by cluster IDs
PC2 (44.08%)
PC1 (55.92%)
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
Cluster
1

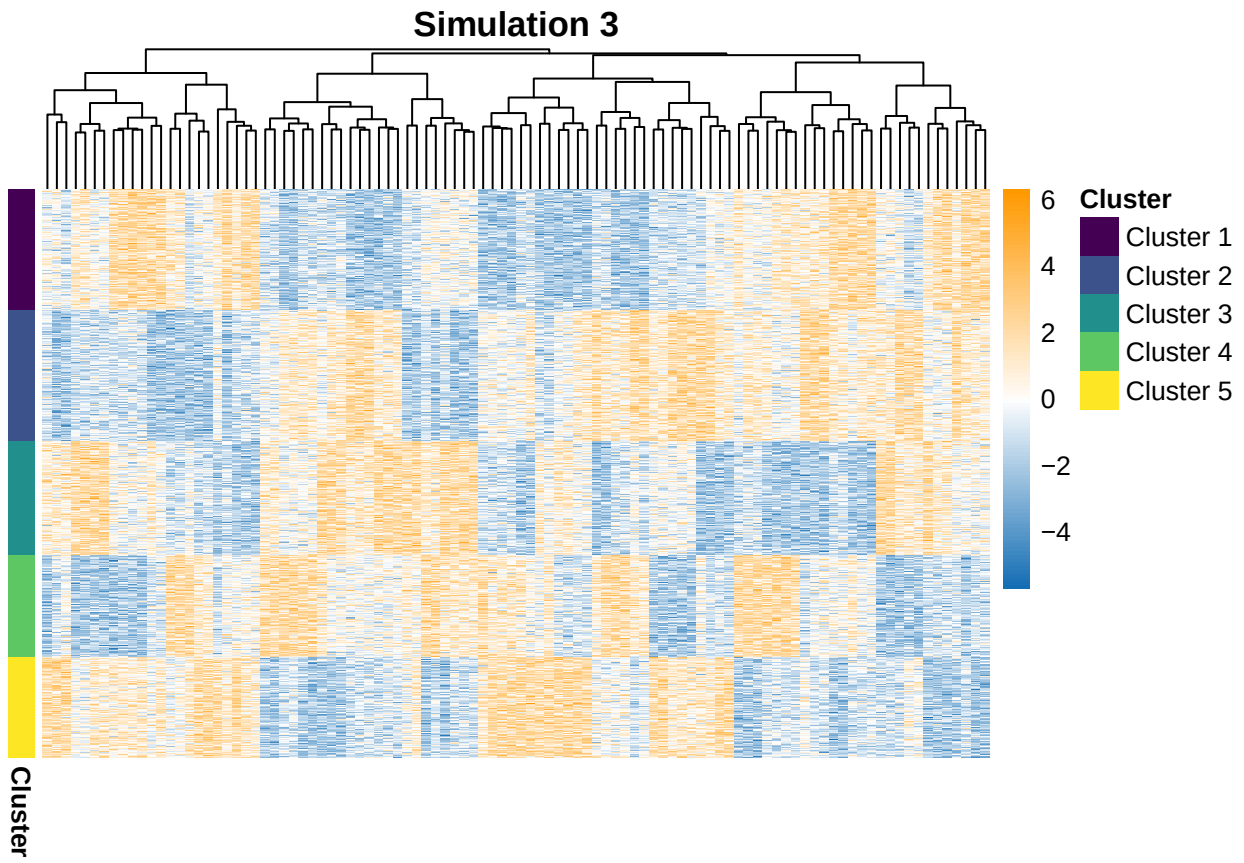


Simulation 3: Large, informative dataset

This case is intended to be more representative of real data. We increase the sample size and the number of features.

- $N = 1000$;
- $P_s = 100$;
- $P_n = 0$;
- $K = 5$;
- $\pi = vec(\frac{1}{5})$;
- $\Delta_\mu = 1$; and
- $\sigma^2_{kp} = 1$.

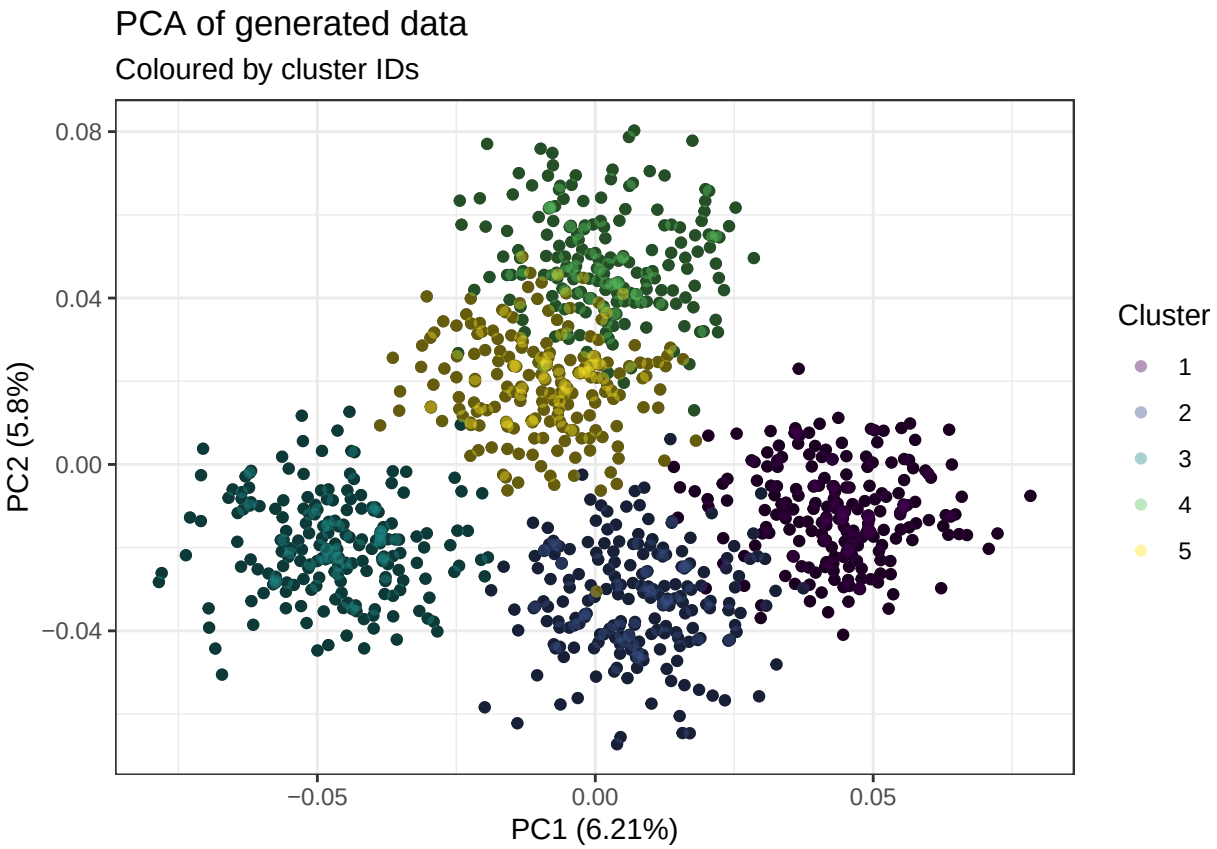

PCA of generated data
Coloured by cluster IDs
0.075
0.050
0.025
0.000
−0.025
−0.050
PC2 (17.32%)
−0.06
−0.03
0.00
0.03
PC1 (19.56%)
Cluster
1
2
3
4
5

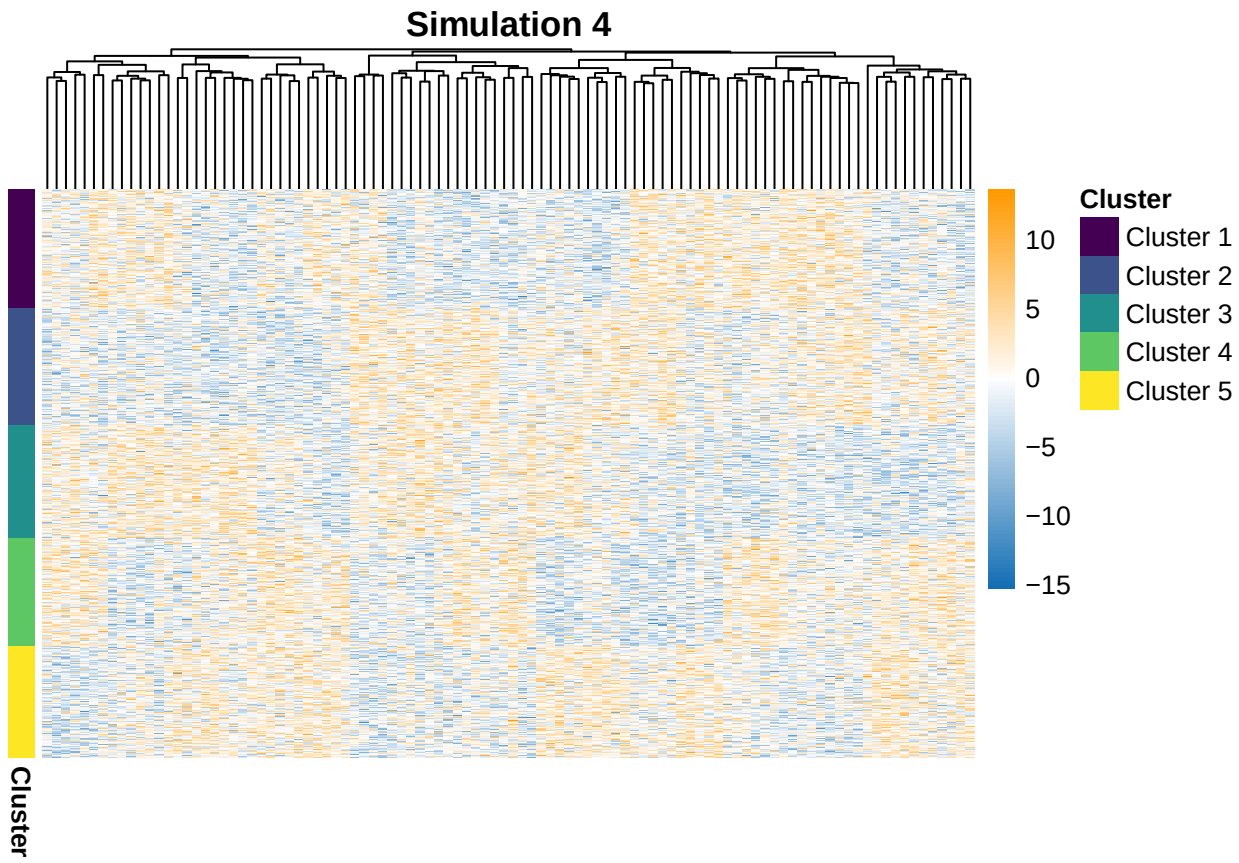


Simulation 4: Large, informative dataset, large σ^2

We test the ratio of μ to σ^2 required for structure to be successfully uncovered.

- $N = 1000$;
- $P_s = 100$;
- $P_n = 0$;
- $K = 5$;
- $\pi = vec(\frac{1}{5})$;
- $\Delta_\mu = 1$; and
- $\sigma^2_{kp} = 3$.


PCA of generated data
Coloured by cluster IDs
0.08
0.04
0.00
−0.04
PC2 (5.8%)
−0.05
0.00
0.05
PC1 (6.21%)
Cluster
1
2
3
4
5



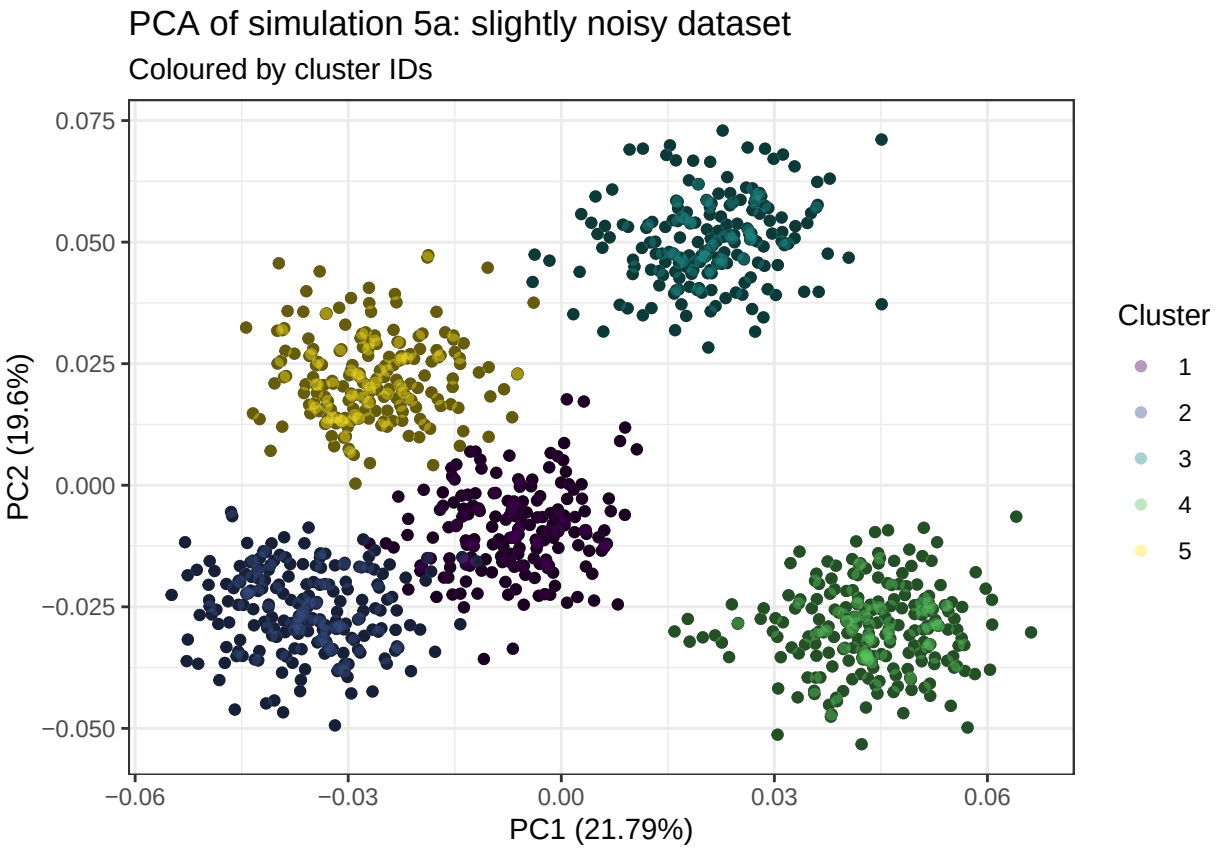
Simulation 5: Large, noisy dataset

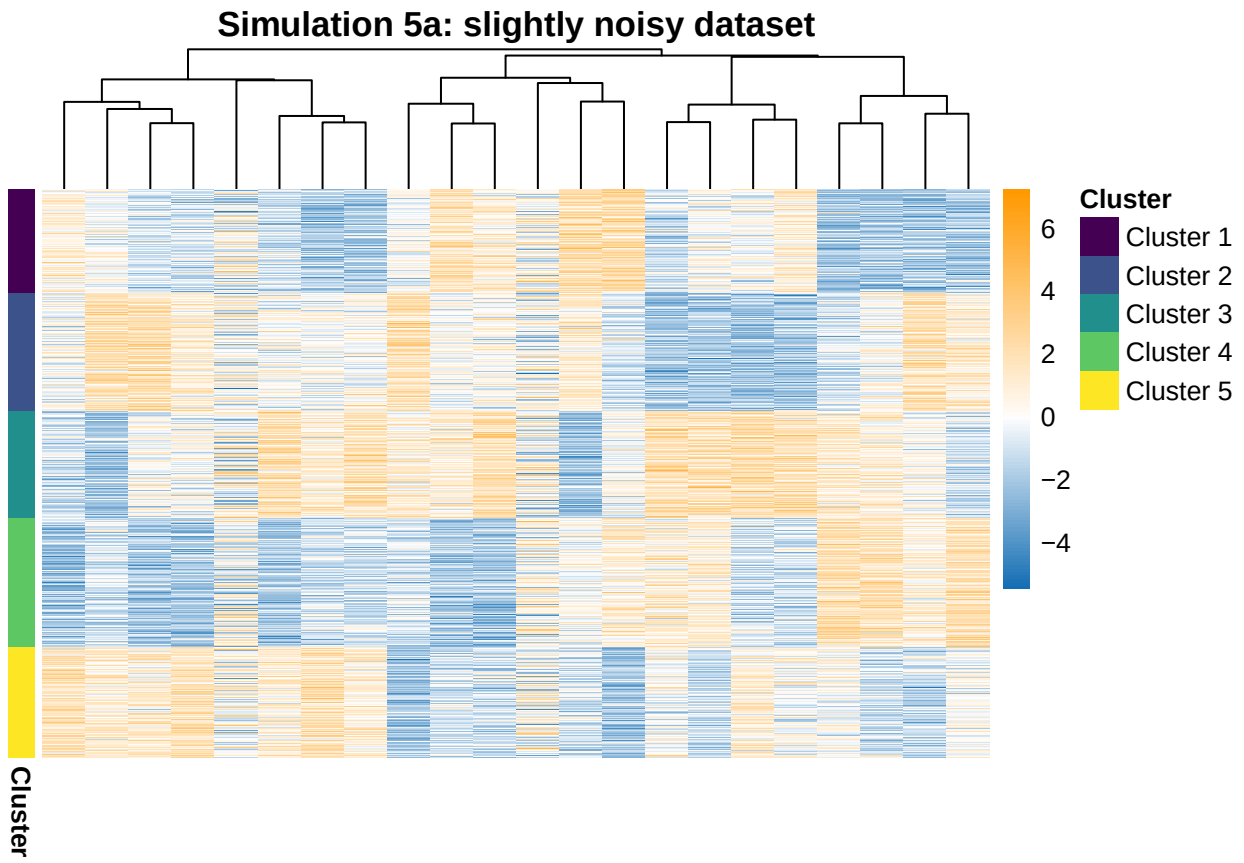
This case is intended to test how well structure can be uncovered as P_n increases. We test for:

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

Simulation 5a: Large, slightly noisy dataset

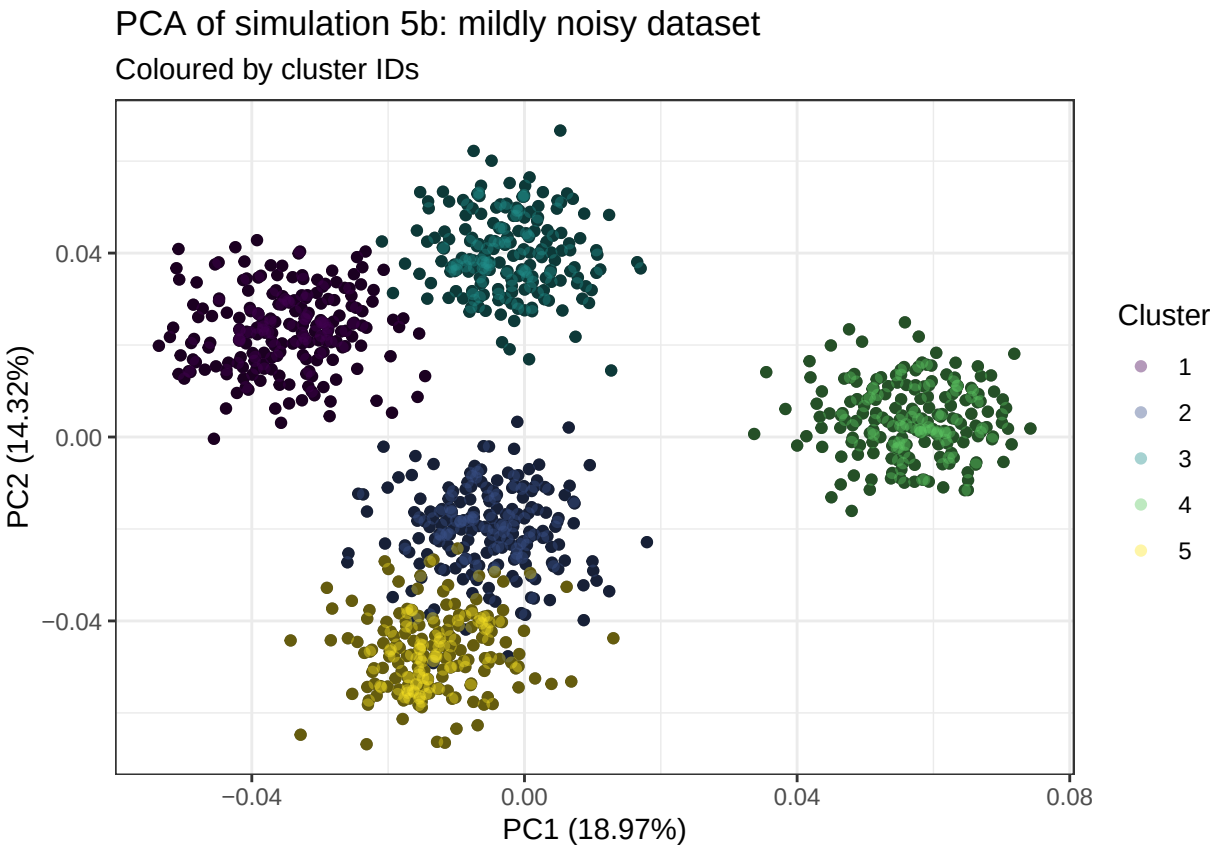
- $N = 1000$;
- $P_s = 20$;
- $P_n = 2$;
- $K = 5$;
- $\pi = vec(\frac{1}{5})$;
- $\Delta_\mu = 1$; and
- $\sigma^2_{kp} = 1$.

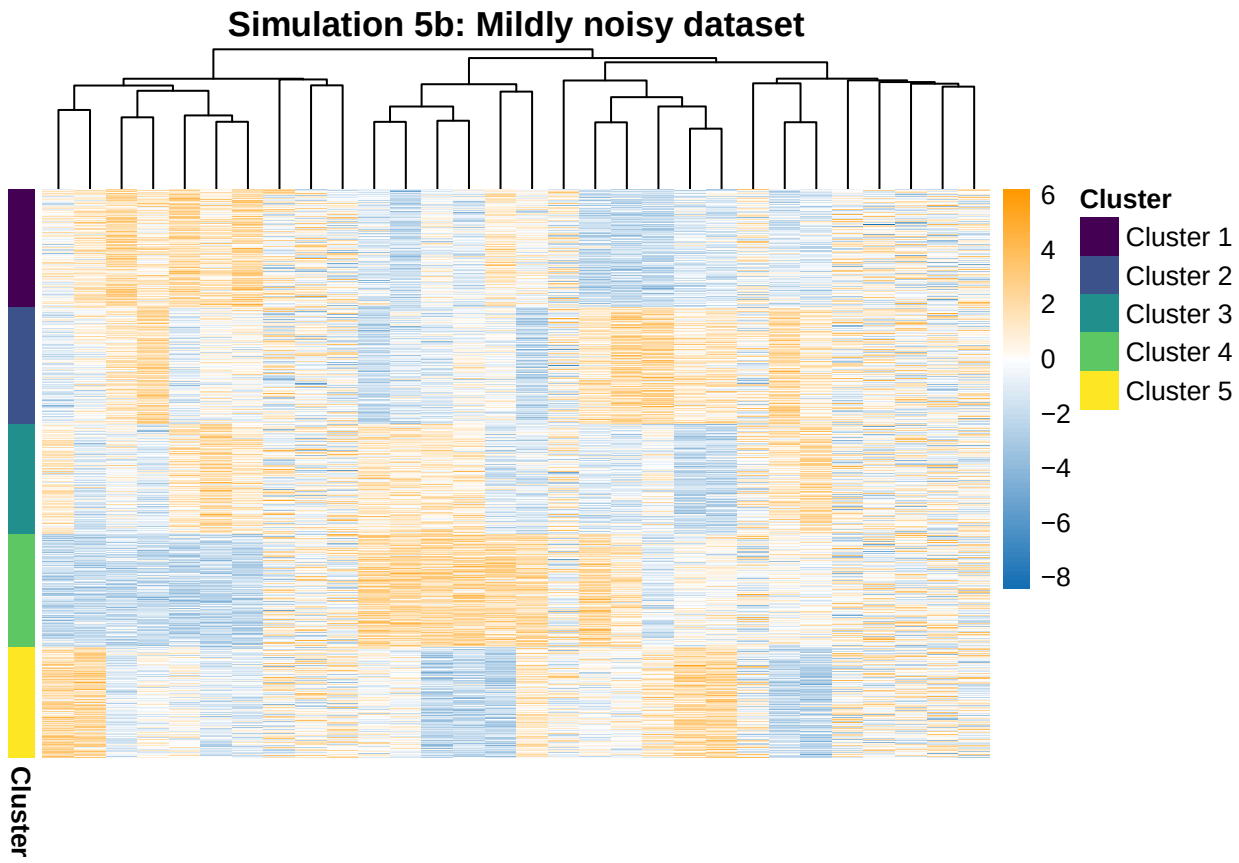

PCA of simulation 5a: slightly noisy dataset
Coloured by cluster IDs
PC2 (19.6%)
PC1 (21.79%)
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.06
−0.03
0.00
0.03
0.06
Cluster
1
2
3
4
5



Simulation 5b: Large, mildly noisy dataset

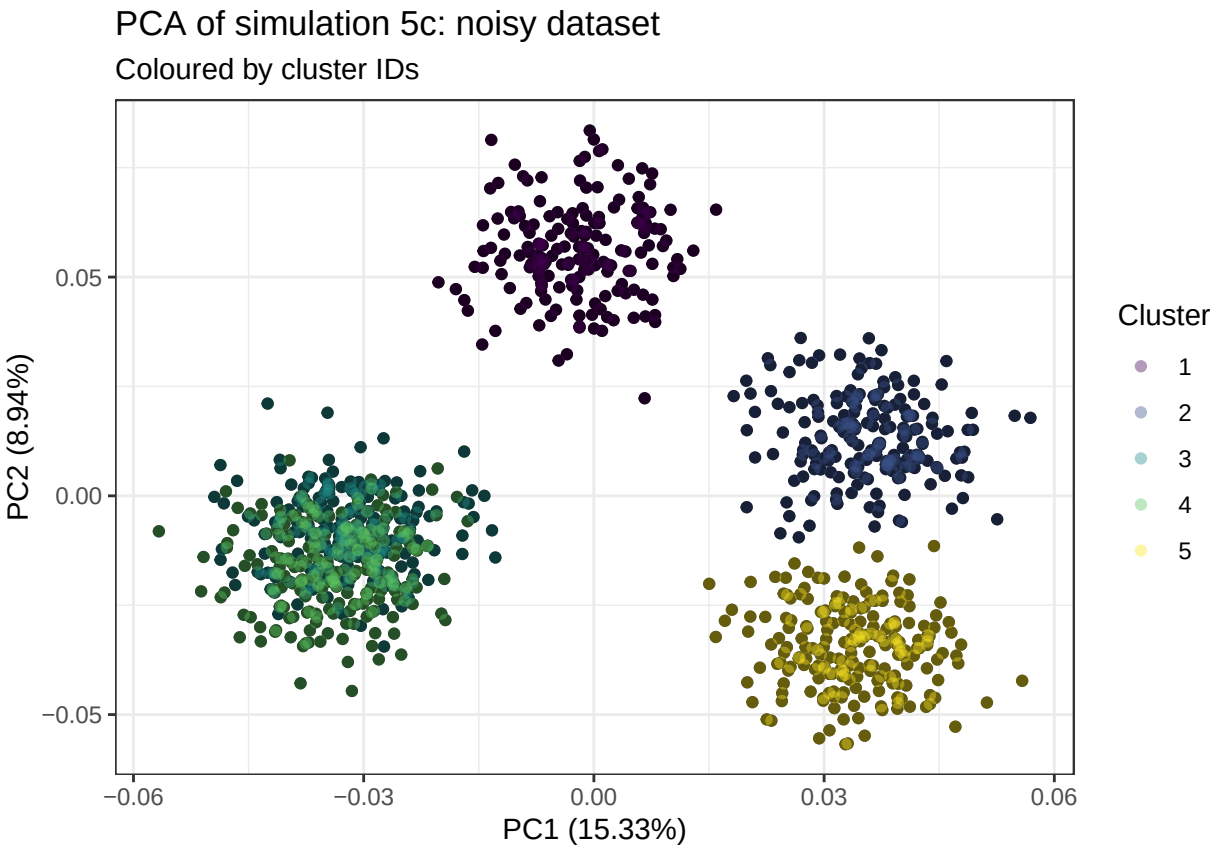
- $N = 1000$;
- $P_s = 20$;
- $P_n = 10$;
- $K = 5$;
- $\pi = vec(\frac{1}{5})$;
- $\Delta_\mu = 1$; and
- $\sigma^2_{kp} = 1$.

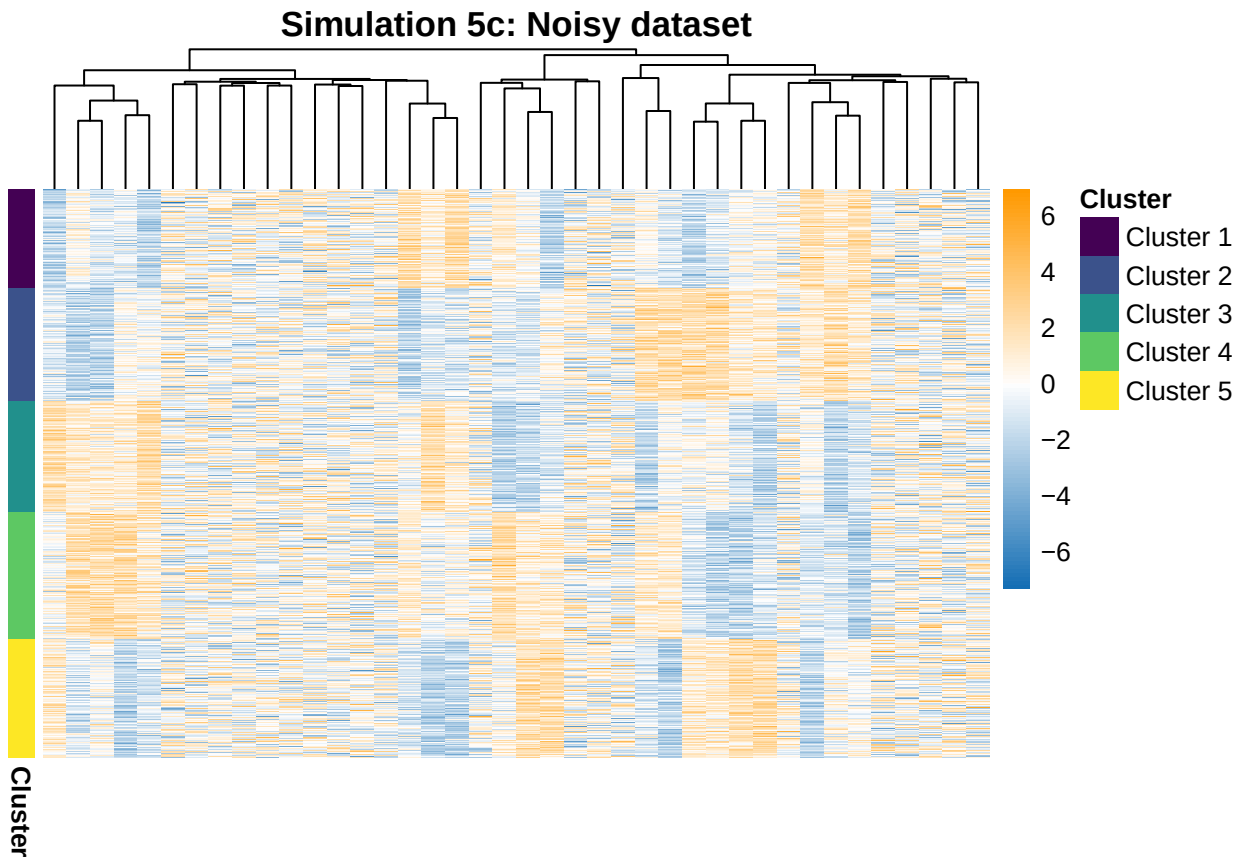

PCA of simulation 5b: mildly noisy dataset
Coloured by cluster IDs
PC2 (14.32%)
0.04
0.00
−0.04
−0.04
0.00
0.04
0.08
PC1 (18.97%)
Cluster
1
2
3
4
5



Simulation 5c: Large, noisy dataset

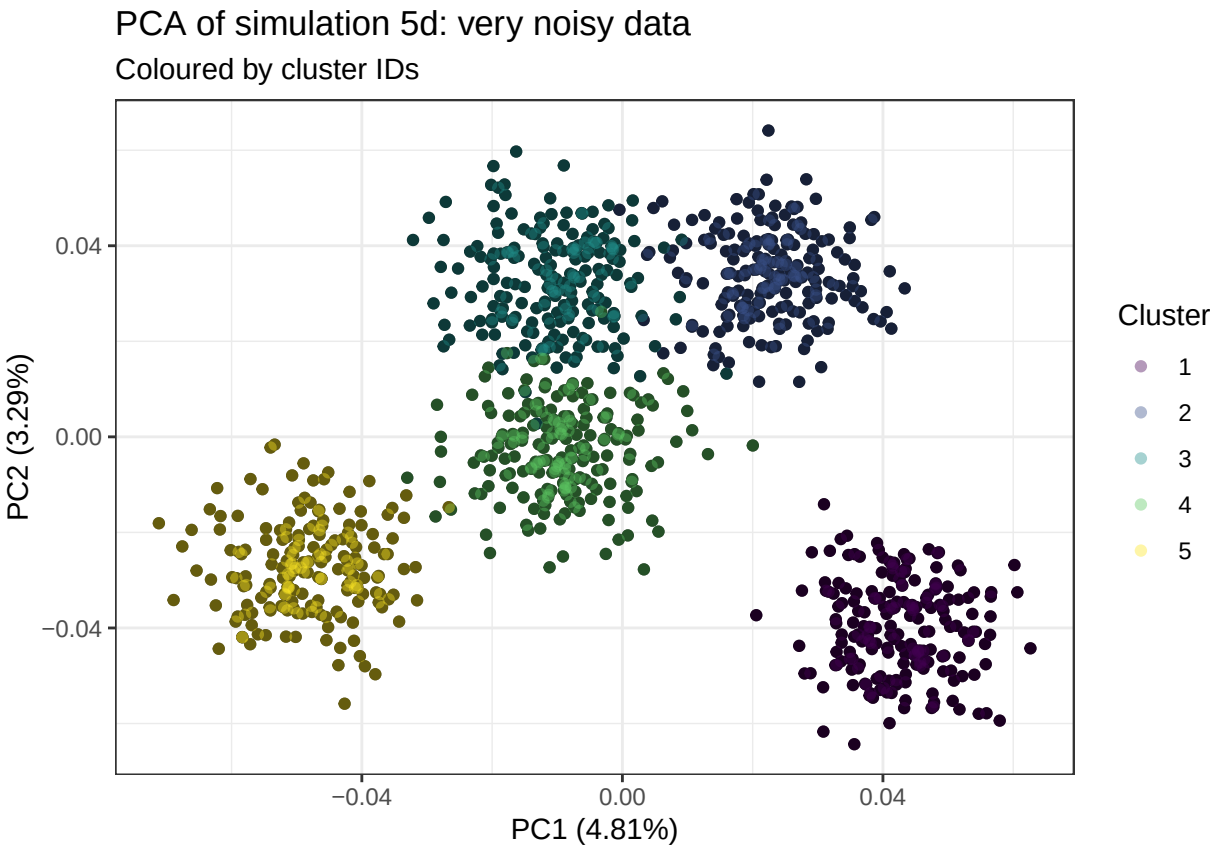
- $N = 1000$;
- $P_s = 20$;
- $P_n = 20$;
- $K = 5$;
- $\pi = vec(\frac{1}{5})$;
- $\Delta_\mu = 1$; and
- $\sigma^2_{kp} = 1$.

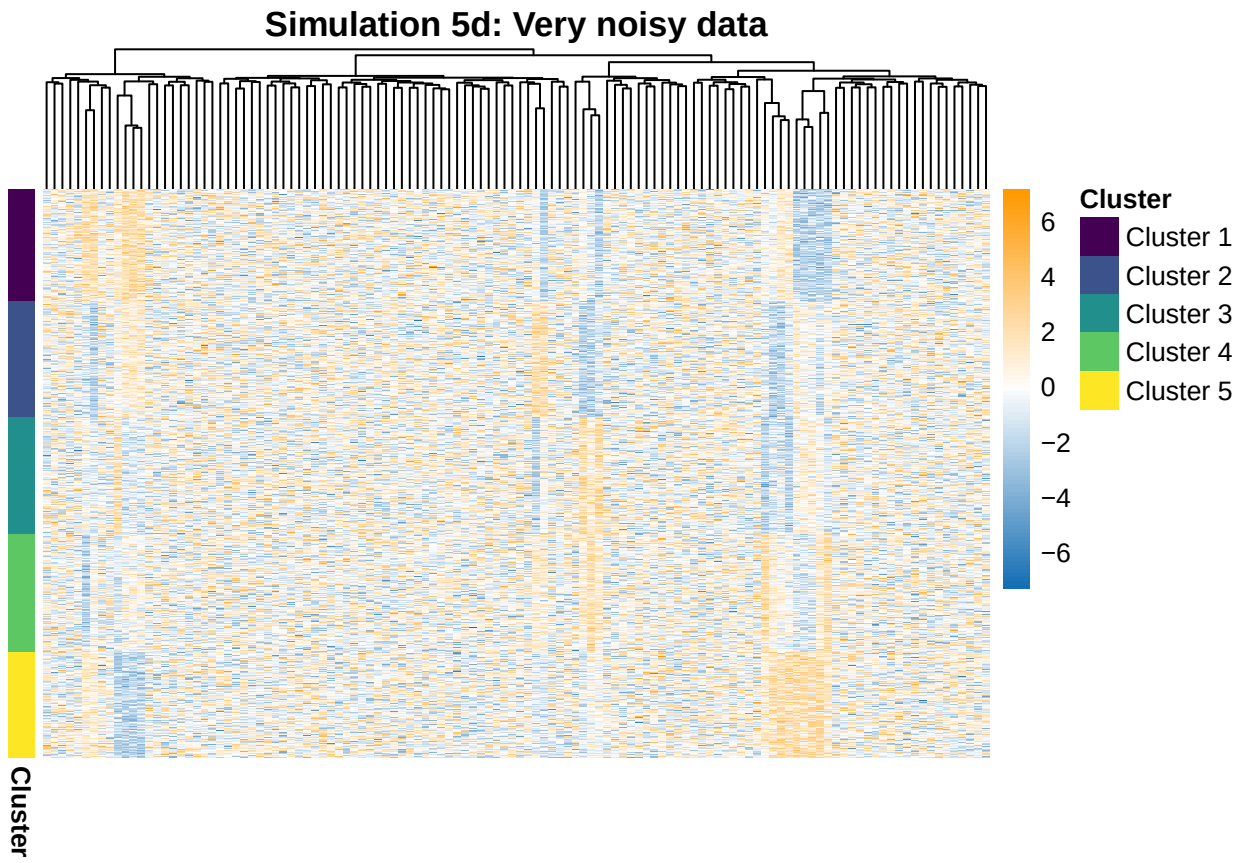

PCA of simulation 5c: noisy dataset
Coloured by cluster IDs
PC2 (8.94%)
0.05
0.00
−0.05
−0.06
−0.03
0.00
0.03
0.06
PC1 (15.33%)
Cluster
1
2
3
4
5



Simulation 5d: Large, very noisy dataset

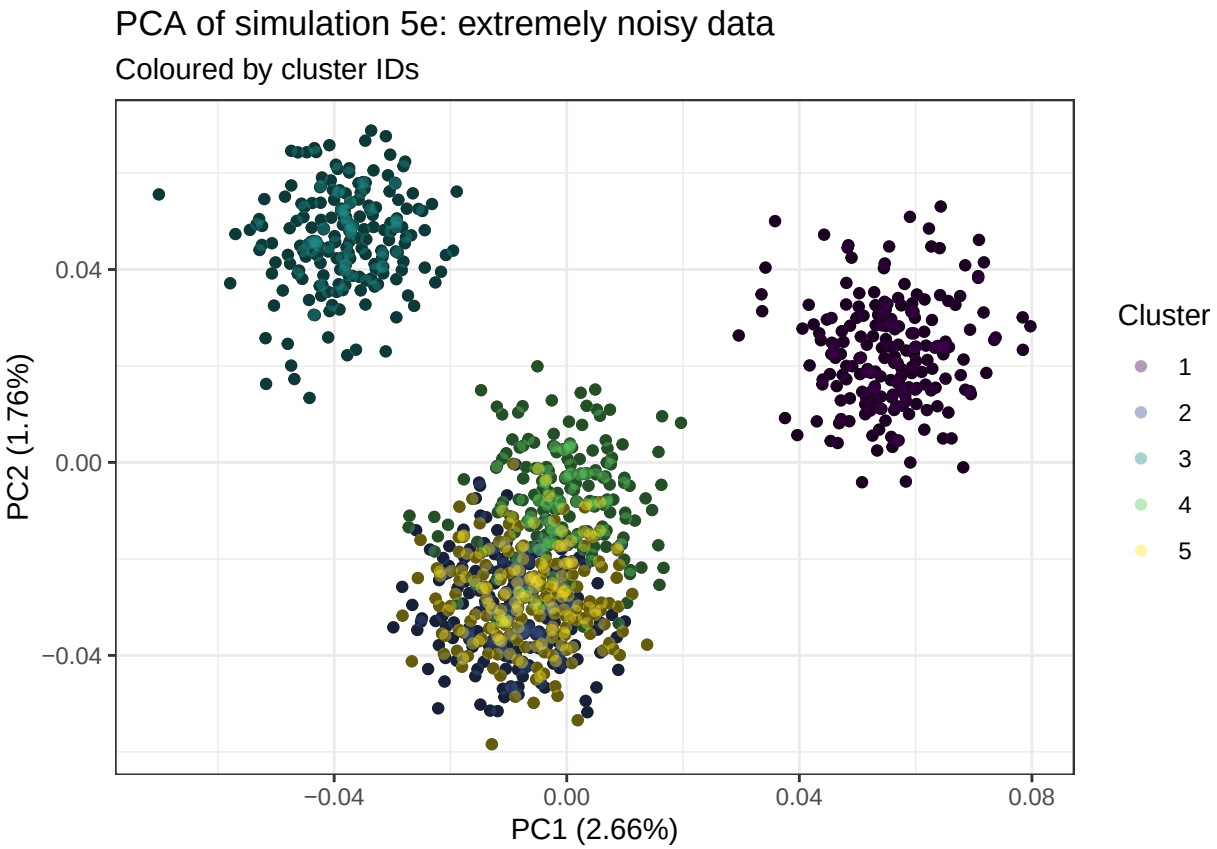
- $N = 1000$;
- $P_s = 20$;
- $P_n = 100$;
- $K = 5$;
- $\pi = vec(\frac{1}{5})$;
- $\Delta_\mu = 1$; and
- $\sigma^2_{kp} = 1$.

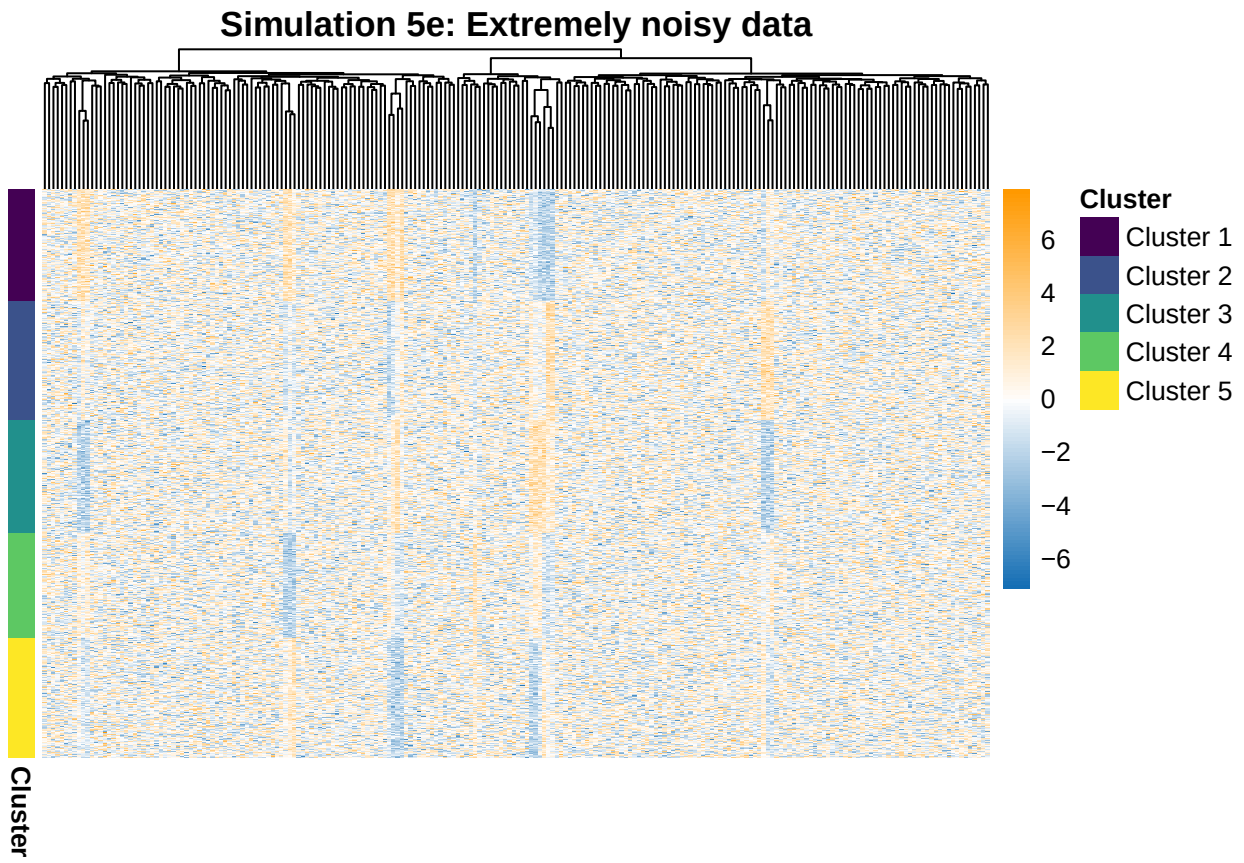

PCA of simulation 5d: very noisy data
Coloured by cluster IDs
PC2 (3.29%)
0.04
0.00
−0.04
−0.04
0.00
0.04
PC1 (4.81%)
Cluster
1
2
3
4
5



Simulation 5e: Large, extremely noisy dataset

- $N = 1000$;
- $P_s = 20$;
- $P_n = 200$;
- $K = 5$;
- $\pi = vec(\frac{1}{5})$;
- $\Delta_\mu = 1$; and
- $\sigma^2_{kp} = 1$.


PCA of simulation 5e: extremely noisy data
Coloured by cluster IDs
PC2 (1.76%)
0.04
0.00
−0.04
−0.04
0.00
0.04
0.08
PC1 (2.66%)
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.