Supplementary Material

A. DP-means objective derivation

First consider the generative model in Section 2. The joint distribution of the observed data x, cluster indicators z, and cluster means μ can be written as follows.

$$\begin{aligned}\mathbb{P}(x,z,\mu) &= \mathbb{P}(x|z,\mu)\mathbb{P}(z)\mathbb{P}(\mu)\\ &= \prod_{k=1}^{K^+}\prod_{n:z_{n,k}=1}\mathcal{N}(x_n|\mu_k,\sigma^2 I_D)\\ &\cdot \theta^{K^+-1}\frac{\Gamma(\theta+1)}{\Gamma(\theta+N)}\prod_{k=1}^{K^+}(S_{N,k}-1)!\\ &\cdot \prod_{k=1}^{K^+}\mathcal{N}(\mu_k|0,\rho^2 I_D).\end{aligned}$$

Then set $\theta := \exp(-\lambda^2/(2\sigma^2))$ and consider the limit $\sigma^2 \to 0$. In the following, $f(\sigma^2) = O(g(\sigma^2))$ denotes that there exist some constants $c, s^2 > 0$ such that $|f(\sigma^2)| \leq c|g(\sigma^2)|$ for all $\sigma^2 < s^2$.

$$\begin{aligned}&-\log \mathbb{P}(x,z,\mu)\\ &= \sum_{k=1}^{K^+}\sum_{n:z_{n,k}=1}\left[O(\log\sigma^2)+\frac{1}{2\sigma^2}\|x_n-\mu_k\|^2\right]\\ &+ (K^+-1)\frac{\lambda^2}{2\sigma^2}+O(1)\\ &+ O(1).\end{aligned}$$

It follows that

$$-2\sigma^2\log\mathbb{P}(x,z,\mu) = \sum_{k=1}^{K^+}\sum_{n:z_{n,k}=1}\|x_n-\mu_k\|^2 + (K^+-1)\lambda^2 + O(\sigma^2\log(\sigma^2)).$$

But since $\sigma^2\log(\sigma^2) \to 0$ as $\sigma^2 \to 0$, we have that the remainder of the righthand side is asymptotically equivalent (as $\sigma^2 \to 0$) to the lefthand side (Eq. (2)).

B. BP-means objective derivation

The recipe is the same as in Sup. Mat. A. This time we start with the generative model in Section 3. The joint distribution of the observed data X, feature indicators Z, and feature means A can be written as follows.

$$\begin{aligned}&\mathbb{P}(X,Z,A) = \mathbb{P}(X|Z,A)\mathbb{P}(Z)\mathbb{P}(A)\\ &= \frac{1}{(2\pi\sigma^2)^{ND/2}}\exp\left\{-\frac{1}{2\sigma^2}\mathbf{tr}((X-ZA)'(X-ZA))\right\}\\ &\cdot \frac{\gamma^{K^+}\exp\left\{-\sum_{n=1}^N\frac{\gamma}{n}\right\}}{\prod_{h=1}^H \tilde{K}_h!}\prod_{k=1}^{K^+}\frac{(S_{N,k}-1)!(N-S_{N,k})!}{N!}\\ &\cdot \frac{1}{(2\pi\rho^2)^{K^+D/2}}\exp\left\{-\frac{1}{2\rho^2}\mathbf{tr}(A'A)\right\}.\end{aligned}$$

Now set $\gamma := \exp(-\lambda^2/(2\sigma^2))$ and consider the limit $\sigma^2 \to 0$. Then

$$\begin{aligned}&-\log\mathbb{P}(X,Z,A)\\ &= O(\log\sigma^2)+\frac{1}{2\sigma^2}\mathbf{tr}((X-ZA)'(X-ZA))\\ &+ K^+\frac{\lambda^2}{2\sigma^2}+\exp(-\lambda^2/(2\sigma^2))\sum_{n=1}^N n^{-1}+O(1)\\ &+ O(1).\end{aligned}$$

It follows that

$$-2\sigma^2\log\mathbb{P}(X,Z,A) = \mathbf{tr}[(X-ZA)'(X-ZA)] + K^+\lambda^2 + O\left(\sigma^2\exp(-\lambda^2/(2\sigma^2))\right) + O(\sigma^2\log(\sigma^2)).$$

But since $\exp(-\lambda^2/(2\sigma^2)) \to 0$ and $\sigma^2\log(\sigma^2) \to 0$ as $\sigma^2 \to 0$, we have that $-2\sigma^2\log\mathbb{P}(X,Z,A) \sim \mathbf{tr}[(X-ZA)'(X-ZA)] + K^+\lambda^2$.

C. Collapsed DP-means objective derivation

We apply the usual recipe as in Sup. Mat. A. The generative model for collapsed DP-means is described in Section 4. The joint distribution of the observed data X and cluster indicators Z can be written as follows:

$$\begin{aligned}&\mathbb{P}(X,Z) = \mathbb{P}(X|Z)\mathbb{P}(Z)\\ &= \left((2\pi)^{ND/2}(\sigma^2)^{(N-K^+)D/2}(\rho^2)^{K^+D/2}|Z'Z+\frac{\sigma^2}{\rho^2}I_D|^{D/2}\right)^{-1}\\ &\cdot \exp\left\{-\frac{1}{2\sigma^2}\mathbf{tr}\left(X'(I_N-Z(Z'Z+\frac{\sigma^2}{\rho^2}I_D)^{-1}Z')X\right)\right\}\\ &\cdot \theta^{K^+-1}\frac{\Gamma(\theta+1)}{\Gamma(\theta+N)}\prod_{k=1}^{K^+}(S_{N,k}-1)!.\end{aligned}$$

Now set $\theta := \exp(-\lambda^2/(2\sigma^2))$ and consider the limit $\sigma^2 \to 0$. Then

$$\begin{aligned}&-\log\mathbb{P}(X,Z) = O(\log(\sigma^2))\\ &+ \frac{1}{2\sigma^2}\mathbf{tr}\left(X'(I_N-Z(Z'Z+\frac{\sigma^2}{\rho^2}I_D)^{-1}Z')X\right)\\ &+ (K^+-1)\frac{\lambda^2}{2\sigma^2}+O(1).\end{aligned}$$

It follows that

$$\begin{aligned}&-2\sigma^2 \log \mathbb{P}(X,Z)\\&= \mathbf{tr}\left(X'(I_N - Z(Z'Z + \frac{\sigma^2}{\rho^2} I_D)^{-1} Z')X\right)\\&+ (K^+ - 1)\lambda^2 + O(\sigma^2 \log(\sigma^2)).\end{aligned}$$

We note that $\sigma^2 \log(\sigma^2) \to 0$ as $\sigma^2 \to 0$. Further note that $Z'Z$ is a diagonal $K \times K$ matrix with (k,k) entry (call it $S_{N,k}$) equal to the number of indices in cluster k. $Z'Z$ is invertible since we assume no empty clusters are represented in Z. Then

$$\begin{aligned}&-2\sigma^2 \log \mathbb{P}(X,Z)\\&\sim \mathbf{tr}\left(X'(I_N - Z(Z'Z)^{-1}Z')X\right) + (K^+ - 1)\lambda^2\end{aligned}$$

as $\sigma^2 \to 0$.

C.1. More interpretable objective

The objective for the collapsed Dirichlet process is more interpretable after some algebraic manipulation. We describe here how the opaque $\mathbf{tr}\left(X'(I_N - Z(Z'Z)^{-1}Z')X\right)$ term can be written in a form more reminiscent of the $\sum_{k=1}^{K^+}\sum_{n:z_{n,k}=1}\|x_n - \mu_k\|^2$ term in the uncollapsed objective. First, recall that $C := Z'Z$ is a $K \times K$ matrix with $C_{k,k} = S_{N,k}$ and $C_{j,k} = 0$ for $j \neq k$. Then $C' := Z(Z'Z)^{-1}Z'$ is an $N \times N$ matrix with $C'_{n,m} = S^{-1}_{N,k}$ if and only if $z_{n,k} = z_{m,k} = 1$ and $C'_{n,m} = 0$ if $z_{n,k} \neq z_{m,k}$.

$$\begin{aligned}&\mathbf{tr}(X'(I_N - Z(Z'Z)^{-1}Z')X)\\&= \mathbf{tr}(X'X) - \mathbf{tr}(X'Z(Z'Z)^{-1}Z'X)\\&= \mathbf{tr}(XX') - \sum_{d=1}^{D}\sum_{k=1}^{K^+}\sum_{n:z_{n,k}=1}\sum_{m:z_{m,k}=1} S^{-1}_{N,k} X_{n,d} X_{m,d}\\&= \sum_{k=1}^{K^+}\left[\sum_{n:z_{n,k}=1} x_n x'_n - 2S^{-1}_{N,k}\sum_{n:z_{n,k}=1} x_n \sum_{m:z_{m,k}=1} x'_m\right.\\&\left. + S^{-1}_{N,k}\sum_{n:z_{n,k}=1} x_n \sum_{m:z_{m,k}=1} x'_m\right]\\&= \sum_{k=1}^{K^+}\sum_{n:z_{n,k}=1}\|x_n - S^{-1}_{N,k}\sum_{m:z_{m,k}=1} x_{m,k}\|^2\\&= \sum_{k=1}^{K^+}\sum_{n:z_{n,k}=1}\|x_n - \bar{x}^{(k)}\|^2,\end{aligned}$$

for the cluster-specific empirical mean defined as $\bar{x}^{(k)} := S^{-1}_{N,k}\sum_{m:z_{m,k}=1} x_{m,k}$, as in the main text.

D. Collapsed BP-means objective derivation

We continue to apply the usual recipe as in Sup. Mat. A. The generative model for collapsed BP-means is described in Section 4. The joint distribution of the observed data X and feature indicators Z can be written as follows:

$$\begin{aligned}&\mathbb{P}(X,Z) = \mathbb{P}(X|Z)\mathbb{P}(Z)\\&= \left((2\pi)^{ND/2}(\sigma^2)^{(N-K^+)D/2}(\rho^2)^{K^+D/2}|Z'Z + \frac{\sigma^2}{\rho^2}I_D|^{D/2}\right)^{-1}\\&\cdot \exp\left\{-\frac{1}{2\sigma^2}\mathbf{tr}\left(X'(I_N - Z(Z'Z + \frac{\sigma^2}{\rho^2}I_D)^{-1}Z')X\right)\right\}\\&\cdot \frac{\gamma^{K^+}\exp\left\{-\sum_{n=1}^{N}\frac{\gamma}{n}\right\}}{\prod_{h=1}^{H}\tilde{K}_h!}\prod_{k=1}^{K^+}\frac{(S_{N,k}-1)!(N-S_{N,k})!}{N!}.\end{aligned}$$

Now set $\gamma := \exp(-\lambda^2/(2\sigma^2))$ and consider the limit $\sigma^2 \to 0$. Then

$$\begin{aligned}&-\log \mathbb{P}(X,Z) = O(\log(\sigma^2))\\&+ \frac{1}{2\sigma^2}\mathbf{tr}\left(X'(I_N - Z(Z'Z + \frac{\sigma^2}{\rho^2}I_D)^{-1}Z')X\right)\\&+ K^+\frac{\lambda^2}{2\sigma^2} + \exp(-\lambda^2/(2\sigma^2))\sum_{n=1}^{N} n^{-1} + O(1).\end{aligned}$$

It follows that

$$\begin{aligned}-2\sigma^2 \log \mathbb{P}(X,Z) = \mathbf{tr}&\left(X'(I_N - Z(Z'Z + \frac{\sigma^2}{\rho^2}I_D)^{-1}Z')X\right)\\&+ K^+\lambda^2 + O\left(\sigma^2\exp(-\lambda^2/(2\sigma^2))\right) + O(\sigma^2\log(\sigma^2)).\end{aligned}$$

But $\exp(-\lambda^2/(2\sigma^2)) \to 0$ and $\sigma^2\log(\sigma^2) \to 0$ as $\sigma^2 \to 0$. And $Z'Z$ is invertible so long as two features do not have identical membership (in which case we collect them into a single feature). So we have that $-2\sigma^2 \log \mathbb{P}(X,Z) \sim \mathbf{tr}\left(X'(I_N - Z(Z'Z)^{-1}Z')X\right) + K^+\lambda^2$.

E. Parametric objectives

First, consider a clustering prior with some fixed maximum number of clusters K. Let $q_{1:K}$ represent a distribution over clusters. Suppose $q_{1:K}$ is drawn from a finite Dirichlet distribution with size $K > 1$ and parameter $\theta > 0$. Further, suppose the cluster for each data point is drawn iid according to $q_{1:K}$. Then, integrating out q, the marginal distribution of the clustering is Dirichlet-multinomial:

$$\mathbb{P}(z) = \frac{\Gamma(K\theta)}{\Gamma(N + K\theta)}\prod_{k=1}^{K}\frac{\Gamma(S_{N,k} + \theta)}{\Gamma(\theta)}. \qquad (11)$$

We again assume a Gaussian mixture likelihood, only now the number of cluster means μ_k has an upper bound of K.

We can find the MAP estimate of z and μ under this model in the limit $\sigma^2 \to 0$. With θ fixed, the clustering prior has no effect, and the resulting optimization problem is $\operatorname{argmin}_{z,\mu} \sum_{k=1}^{K} \sum_{n:z_{nk}=1} \|x_n - \mu_k\|^2$, which is just the usual K-means optimization problem.

We can also try scaling $\theta = \exp(-\lambda^2/(2\sigma^2))$ for some constant $\lambda^2 > 0$ as in the unbounded cardinality case. Then taking the $\sigma^2 \to 0$ limit of the log joint likelihood yields a term of λ^2 for each cluster containing at least one data index in the product in Eq. (11)—except for one such cluster. Call the number of such activated clusters K^+. The resulting optimization problem is

$$\operatorname*{argmin}_{K^+,z,\mu} \sum_{k=1}^{K} \sum_{n:z_{nk}=1} \|x_n - \mu_k\|^2 + (K \wedge K^+ - 1)\lambda^2. \quad (12)$$

This objective caps the number of clusters at K but contains a penalty for each new cluster up to K.

A similar story holds in the feature case. Imagine that we have a fixed maximum of K features. In this finite case, we now let $q_{1:K}$ represent frequencies of each feature and let $q_k \overset{iid}{\sim} \text{Beta}(\gamma, 1)$. We draw $z_{nk} \sim \text{Bern}(q_k)$ iid across n and independently across k. The linear Gaussian likelihood model is as in Eq. (5) except that now the number of features is bounded. If we integrate out the $q_{1:K}$, the resulting marginal prior on Z is

$$\prod_{k=1}^{K} \left(\frac{\Gamma(S_{N,k} + \gamma)\Gamma(N - S_{N,k} + 1)}{\Gamma(N + \gamma + 1)} \frac{\Gamma(\gamma + 1)}{\Gamma(\gamma)\Gamma(1)} \right). \quad (13)$$

Then the limiting MAP problem as $\sigma^2 \to 0$ is

$$\operatorname*{argmin}_{Z,A} \mathbf{tr}[(X - ZA)'(X - ZA)]. \quad (14)$$

This objective is analogous to the K-means objective but holds for the more general problem of feature allocations. Eq. (14) can be solved according to the *K features algorithm* (Alg. 4). Notably, all of the optimizations for n in the first step of the algorithm may be performed in parallel.

We can further set $\gamma = \exp(-\lambda^2/(2\sigma^2))$ as for the unbounded cardinality case before taking the limit $\sigma^2 \to 0$. Then a λ^2 term contributes to the limiting objective for each non-empty feature from the product in Eq. (13):

$$\operatorname*{argmin}_{K^+,Z,A} \mathbf{tr}[(X - ZA)'(X - ZA)] + (K \wedge K^+)\lambda^2, \quad (15)$$

reminiscent of the BP-means objective but with a cap of K possible features.

F. General multivariate Gaussian likelihood

Above, we assumed a multivariate spherical Gaussian likelihood for each cluster. This assumption can be generalized in a number of ways. For instance, assume a general covariance matrix $\sigma^2\Sigma_k$ for positive scalar σ^2 and positive definite $D \times D$ matrix Σ_k. Then we assume the following likelihood model for data points assigned to the kth cluster ($z_{n,k} = 1$): $x_n \sim \mathcal{N}(\mu_k, \sigma^2\Sigma_k)$. Moreover, assume an inverse Wishart prior on the positive definite matrix Σ_k: $\Sigma_k \sim W^{-1}(\Phi, \nu)$ for Φ a positive definite matrix and $\nu > D-1$. Assume a prior $\mathbb{P}(\mu)$ on μ that puts strictly positive density on all real-valued D-length vectors μ. For now we assume K is fixed and that $\mathbb{P}(z)$ puts a prior that has strictly positive density on all valid clusterings of the data points. Then

$$\begin{aligned}
&\mathbb{P}(x, z, \mu, \sigma^2\Sigma) \\
&= \mathbb{P}(x|z, \mu, \sigma^2\Sigma)\mathbb{P}(z)\mathbb{P}(\mu)\mathbb{P}(\Sigma) \\
&= \prod_{k=1}^{K^+} \prod_{n:z_{n,k}=1} \mathcal{N}(x_n|\mu_k, \sigma^2\Sigma_k) \\
&\cdot \mathbb{P}(z)\mathbb{P}(\mu) \cdot \prod_{k=1}^{K} \left[\frac{|\Phi|^{\nu/2}}{2^{\nu D/2}\Gamma_D(\nu/2)} |\Sigma_k|^{-\frac{\nu+D+1}{2}} \right. \\
&\left. \cdot \exp\left\{ -\frac{1}{2}\mathbf{tr}(\Phi\Sigma_k^{-1}) \right\} \right],
\end{aligned}$$

where Γ_D is a multivariate gamma function. Consider the limit $\sigma^2 \to 0$. Set $\nu = \lambda^2/\sigma^2$ for some constant $\lambda^2 : \lambda^2 > 0$. Then

$$\begin{aligned}
&-\log \mathbb{P}(x, z, \mu, \sigma^2\Sigma) \\
&= \sum_{k=1}^{K} \sum_{n:z_{n,k}=1} \left[O(\log \sigma^2) + \frac{1}{2\sigma^2}(x_n - \mu_k)'\Sigma_k^{-1}(x_n - \mu_k) \right] \\
&+ O(1) + \sum_{k=1}^{K} \left[-\frac{1}{2\sigma^2}\lambda^2 \log|\Phi| + \frac{D}{2\sigma^2}\lambda^2 \log 2 \right. \\
&\left. + \log\Gamma_D(\lambda^2/(2\sigma^2)) + \left(\frac{\lambda^2}{2\sigma^2} + \frac{D+1}{2} \right) \log|\Sigma_k| + O(1) \right].
\end{aligned}$$

So we find

$$\begin{aligned}
&-2\sigma^2 \left[\log \mathbb{P}(x, z, \mu, \sigma^2\Sigma) + \log\Gamma_D(\lambda^2/(2\sigma^2)) \right] \\
&\sim \sum_{k=1}^{K} \sum_{n:z_{n,k}=1} (x_n - \mu_k)'\Sigma_k^{-1}(x_n - \mu_k) \\
&+ \sum_{k=1}^{K} \lambda^2 \log|\Sigma_k| + c + O(\sigma^2),
\end{aligned}$$

where c is a constant in z, μ, σ^2, Σ. Letting $\sigma^2 \to 0$, the righthand side becomes

$$\sum_{k=1}^{K} \sum_{n:z_{n,k}=1} (x_n - \mu_k)' \Sigma_k^{-1} (x_n - \mu_k) + \sum_{k=1}^{K} \lambda^2 \log |\Sigma_k| + c.$$

It is equivalent to optimize the same quantity without c.

If the Σ_k are known, they may be inputted and the objective may be optimized over the cluster means and cluster assignments. For unknown Σ_k, though, the resulting optimization problem is

$$\min_{z,\mu,\Sigma} \sum_{k=1}^{K} \left[\sum_{n:z_{n,k}=1} (x_n - \mu_k)' \Sigma_k^{-1} (x_n - \mu_k) + \lambda^2 \log |\Sigma_k| \right].$$

That is, the squared Euclidean distance in the classic K-means objective function has been replaced with a Mahalanobis distance, and we have added a penalty term on the size of the Σ_k matrices (with λ^2 modulating the penalty as in previous examples). This objective is reminiscent of that proposed by Sung & Poggio (1998).

G. Proof of BP-means local convergence

The proof of Proposition 1 is as follows.

Proof. By construction, the first step in any iteration does not increase the objective. The second step starts by deleting any features that have the same index collection as an existing feature. Suppose there are m such features with indices J and we keep feature k. By setting $A_{k,\cdot} \leftarrow \sum_{j \in J} A_{j,\cdot}$, the objective is unchanged.

Next, let $\|Y\|_F = \sqrt{\mathbf{tr}(Y'Y)}$ denote the Frobenius norm of a matrix Y. Then $\|Y\|_F^2$ is a convex function. We check that $f(A) = \mathbf{tr}[(X - ZA)'(X - ZA)]$ is convex. Take $\lambda \in [0, 1]$, and let A and B be $K \times D$ matrices; then,

$$\begin{aligned} & f(\lambda A + (1-\lambda)B) \\ & = \|Z[\lambda A + (1-\lambda)B] - X\|_F^2 \\ & = \|\lambda(ZA - X) + (1-\lambda)(ZB - X)\|_F^2 \\ & \leq \lambda \|ZA - X\|_F^2 + (1-\lambda)\|ZB - X\|_F^2 \\ & = \lambda f(A) + \lambda f(B) \end{aligned}$$

We conclude that $f(A)$ is convex.

With this result in hand, note

$$\nabla_A \mathbf{tr}[(X - ZA)'(X - ZA)] = -2Z'(X - ZA). \quad (16)$$

Setting the gradient to zero, we find that $A = (Z'Z)^{-1}Z'X$ solves the equation for A and therefore minimizes the objective with respect to A when $Z'Z$ is invertible, as we have already guaranteed.

Finally, since there are only finitely many feature allocations in which each data point has at most one feature unique to only that data point and no features containing identical indices (any extra such features would only increase the objective due to the penalty), the algorithm cannot visit more than this many configurations and must finish in a finite number of iterations. □