A. Proof of Proposition 1

Proof. We only prove the proposition for $\boldsymbol{\mu} = 0$. If $\boldsymbol{\mu} \neq 0$, we could simply take the mapping $\boldsymbol{x} \to \boldsymbol{x} - \boldsymbol{\mu}$, $\boldsymbol{y} \to \boldsymbol{y} - \boldsymbol{\mu}$ and complete the proof in a similar manner.

When $\mathbf{W}$ is full rank, it is well known that the projected vector $\boldsymbol{y} \in S \subseteq \mathbb{R}^D$ has the following form:

$$\boldsymbol{y} = \mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top\boldsymbol{x} = (\mathbf{U}_d\mathbf{U}_d^\top)\boldsymbol{x}.$$

Next, note that

$$\mathbf{U}_d\mathbf{U}_d^\top = \mathbf{U}\mathbf{diag}(\mathbf{I}_d, \mathbf{O})\mathbf{U}^\top. \tag{32}$$

Therefore,

$$\begin{aligned}\|\boldsymbol{x} - \boldsymbol{y}\|^2 &= \|\boldsymbol{x} - (\mathbf{U}_d\mathbf{U}_d^\top)\boldsymbol{x}\|^2 \\ &= \|\boldsymbol{x} - \mathbf{U}\mathbf{diag}(1, \cdots, 1, 0, \cdots, 0)\mathbf{U}^\top\boldsymbol{x}\|^2 \\ &= \|\mathbf{U}^\top\boldsymbol{x} - \mathbf{diag}(1, \cdots, 1, 0, \cdots, 0)\mathbf{U}^\top\boldsymbol{x}\|^2 \\ &= \sum_{j=d+1}^{D} [\mathbf{U}^\top\boldsymbol{x}]_j^2.\end{aligned}$$

□

B. Technical details of SVA analysis

B.1. Derivation of the update rule for $\boldsymbol{z}$

For creating a new cluster, we use Laplace approximation to approximate the integration. We first write the conditional distribution as

$$p(z_i = k_{new}|\mathbf{X}, \rho, a, b, r, \alpha) = \frac{\alpha}{Z}\sum_{d=1}^{D} p_0(d|r)\int p(\boldsymbol{x}_i|\mathbf{W}, \boldsymbol{\mu}, \sigma^2)\mathrm{d}p_0(\mathbf{W}, \boldsymbol{\mu}|d, \rho, a, b) =: \frac{1}{Z}\sum_{d=1}^{D} p_0(d|r)J_d. \tag{33}$$

Subsequently, the scaled conditional distribution can be written as

$$p(z_i = k_{new}|\mathbf{X}, \rho, a, b, r, \alpha, \beta) = \frac{\alpha}{Z}\sum_{d=1}^{D} p_0(d|r, \beta')J_d(\beta), \tag{34}$$

where

$$J_d(\beta) = \int p(\boldsymbol{x}_i|\mathbf{W}, \boldsymbol{\mu}, \sigma^2, \beta)\mathrm{d}p_0(\mathbf{W}, \boldsymbol{\mu}|d, \rho, a, b, \beta). \tag{35}$$

Define $\boldsymbol{\theta}_d := (\mathbf{W}, \boldsymbol{\mu})$ with $\mathbf{W} \in \mathbb{R}^{D\times d}$ and

$$f_{d,\beta}(\boldsymbol{\theta}_d) := \beta^{-1} \cdot p(\boldsymbol{x}_i|\mathbf{W}, \boldsymbol{\mu}, \sigma^2, \beta)p_0(\mathbf{W}, \boldsymbol{\mu}|d, a, b, \rho, \beta). \tag{36}$$

Using Laplace's approximation, we have (as $\beta \to \infty$)

$$J_d(\beta) = \int \exp(-\beta f_{d,\beta}(\boldsymbol{\theta}_d))\mathrm{d}\boldsymbol{\theta}_d = \frac{\exp(-\beta f_{d,\beta}(\hat{\boldsymbol{\theta}}_d))}{(2\pi/\beta)^{-D(d+1)/2}}\left(\left|\frac{\partial^2 f_{d,\beta}(\hat{\boldsymbol{\theta}}_d)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right|^{-1/2} + o(1)\right), \tag{37}$$

where $\hat{\boldsymbol{\theta}}_d = \mathrm{argmin}_{\boldsymbol{\theta}_d} f_{d,\beta}(\boldsymbol{\theta}_d)$. Note that [7]

$$\lim_{\beta\to\infty} f_{d,\beta}(\boldsymbol{\theta}_d) = \exp\left(-\sigma^{-2} \cdot \sum_{j=d+1}^{D} [\mathbf{U}^\top(\boldsymbol{x}_i - \boldsymbol{\mu})]_j^2\right) = \exp\left(-\frac{d(\boldsymbol{x}_i, S)^2}{\sigma^2}\right). \tag{38}$$

[7] Recall that $\lim_{x\to\infty} f(x) = g(x)$ means $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1$.

As a result, $f_{d,\beta}(\hat{\boldsymbol{\theta}}_d) = 0$ (taking $\boldsymbol{\mu} = \boldsymbol{x}_i$) and

$$\lim_{\beta\to\infty} J_d(\beta) = (2\pi/\beta)^{D(d+1)/2} \cdot g_d(\boldsymbol{x}_i), \tag{39}$$

where $g_d(\boldsymbol{x}_i)$ only depends on d and $\boldsymbol{x}_i$. Therefore,

$$\lim_{\beta\to\infty} p(z_i = k_{new}) = \lim_{\beta\to\infty} \frac{\alpha}{Z} \sum_{d=0}^{D} \exp(-\beta' d + o(\beta)) = \frac{\alpha}{Z} \exp(o(\beta)). \tag{40}$$

B.2. Derivation of the update rule for W_k

Define

$$K_{d_k}(\beta) := \int p(\mathbf{X}|\mathbf{W}, \boldsymbol{\mu}, \boldsymbol{z}, \sigma^2, \beta) p_0(\mathbf{W}|d_k, a, b, \beta) \mathrm{d}\mathbf{W}. \tag{41}$$

Then the (scaled) posterior distribution of d_k can be written as

$$p(d_k|\mathbf{X}, \boldsymbol{z}, \boldsymbol{\mu}, a, b, r, \beta) = \frac{1}{Z(\mathbf{X})} p_0(d_k|r, \beta') K_{d_k}(\beta). \tag{42}$$

Next, define

$$F_{d_k,\beta}(\mathbf{W}) := \beta^{-1} p(\mathbf{X}|\mathbf{W}, \boldsymbol{\mu}, \boldsymbol{z}, \sigma^2, \beta) p_0(\mathbf{W}|d_k, a, b, \beta). \tag{43}$$

Using Laplace approximation, we have

$$K_{d_k}(\beta) = \int \exp(-\beta F_{d_k,\beta}(\mathbf{W})) \mathrm{d}\mathbf{W} = \frac{\exp(-\beta F_{d_k,\beta}(\hat{\mathbf{W}}))}{(2\pi/\beta)^{-Dd_k/2}} \left(\left| \frac{\partial^2 F_{d_k,\beta}(\hat{\mathbf{W}})}{\partial \mathbf{W} \partial \mathbf{W}^\top} \right|^{-1/2} + o(1) \right), \tag{44}$$

where $\hat{\mathbf{W}}$ is the minimizer of $F_{d_k,\beta}(\cdot)$. Note that for any full-rank $\mathbf{W} \in \mathbb{R}^{D\times d_k}$,

$$\lim_{\beta\to\infty} F_{d_k,\beta}(\mathbf{W}) = \exp\left(-\sum_{i=1}^{n} 1_{[z_i=k]} \sum_{j=d_k+1}^{D} \frac{[\mathbf{U}^\top(\boldsymbol{x}_i - \boldsymbol{\mu}_k)]_j^2}{\sigma^2} - \sum_{j=1}^{d_k} \frac{l_j^{-1}}{b} \right). \tag{45}$$

Taking $l_j \to \infty$, it is then clear that

$$\lim_{\beta\to\infty} F_{d_k,\beta}(\hat{\mathbf{W}}) = \exp\left(-\inf_{\mathbf{U}} \sum_{i=1}^{n} 1_{[z_i=k]} \sum_{j=d_k+1}^{D} \frac{[\mathbf{U}^\top(\boldsymbol{x}_i - \boldsymbol{\mu}_k)]_j^2}{\sigma^2} \right) \tag{46}$$

$$= \exp\left(- \inf_{\mathbf{W}\in\mathbb{R}^{D\times d_k}} \sum_{i=1}^{n} 1_{[z_i=k]} \cdot \frac{d(\boldsymbol{x}_i, S(\mathbf{W}, \boldsymbol{\mu}_k))^2}{\sigma^2} \right). \tag{47}$$

Here the second equation is due to the fact that $d(\cdot, S(\mathbf{W}, \boldsymbol{\mu}))$ does not depend on eigenvalues of $\mathbf{W}$, and hence optimization over $\mathbf{U}$ is equivalent to optimization over $\mathbf{W}$.

B.3. Proof of Theorem 1

Proof. We first prove that for each cluster $k \in [K]$, after updating the subspace projection matrix $\mathbf{W}_k$ (along with its dimension d_k) and the offset $\boldsymbol{\mu}_k$, the loss function $\mathcal{L}$ does not increase. When subspace dimension $d_k = d$ is fixed, the update rule

$$\boldsymbol{\mu}_k = \frac{1}{n_k} \sum_{z_i=k} \boldsymbol{x}_i, \quad \mathbf{U}_{d_k}^{(k)} = \mathbf{A}_d \tag{48}$$

is exactly the same with principle component analysis (PCA) for the top d_k principle directions.. As a result, the subspace S_k given by $S(\mathbf{W}_k, \boldsymbol{\mu}_k)$ minimizes the total squared distance of data points and S_k within the k-th cluster (i.e.,

$\sum_{z_i=k} d(\boldsymbol{x}_i, S_k)^2)$. Note again that the distance $d(\boldsymbol{x}_i, S(\boldsymbol{W}_k, \boldsymbol{\mu}_k))$ only depends on $\boldsymbol{\mu}_k$ and the orthogonal matrix $\mathbf{U}_d^{(k)}$ associated with $\mathbf{W}_k$. The eigenvalues of $\mathbf{W}_k$ do not affect the distance.

We have proved that given $d_k = d$, the update rule given in Eq. (48) chooses $\mathbf{W}_k$ and $\boldsymbol{\mu}_k$ that minimizes the total squared distance for each instance. The update rule for d_k given in Eq. (27) shows that we want to select the dimension d that minimizes the sum of total squared distance and a linear penalty term $s \cdot d$. This is consistent with the deterministic loss function $\mathcal{L}$ shown in Eq. (31). So after updates of $\mathbf{W}$, $\boldsymbol{d}$ and $\boldsymbol{\mu}$ the loss function does not increase.

Next, we turn to the update of cluster assignments $\boldsymbol{z}$. We want to prove that after each update of z_i for some data point $\boldsymbol{x}_i$ the loss function does not increase. This part of analysis resembles the analysis of K-means and DP-means algorithm (Kulis & Jordan, 2012). When we assign z_i to an existing cluster it is clear the distance $d(\boldsymbol{x}_i, S_k)$ does not increase and neither does the total loss. When z_i is assigned to a new cluster, we lose a $d(\boldsymbol{x}_i, S_k)$ cost and gains a λ cost because of creating a new cluster. This does not increase the total loss function $\mathcal{L}$, however, by the definition of $Q_i(k)$ and the update rule of z_i shown in Eq. (23). Note that the new cluster will have a dimension of zero, so no extra penalty term is incurred.

□

C. Details of Hopkins-155 experiments

C.1. Some statistics of the Hopkins-155 dataset

Table 4 gives some statistics of the Hopkins-155 dataset, including the number of sequences (n), the number of points (P) and the number of frames (F) per sequence. In Table 4 the notation *Check-2* refers to all checker board video sequences that contain 2 motions.

Table 4. Some statistics of the Hopkins 155 dataset (Tron & Vidal, 2007)

Dataset	n	P	F
Check-2	78	291	28
Check-3	26	437	28
Traffic-2	31	241	30
Traffic-3	7	332	31
Articul.-1	11	155	40
Articul.-2	2	122	31
All	155	vary	vary

C.2. Detailed performance comparison

In Table 5 we provide detailed performance comparison for the DP-space algorithm and its competitors, including both the mean and median classification error on each of the video sequence groups. Note that the results for EM-MPPCA-m are only included for reference because they are not directly comparable with other performance results.

Table 5. Classification error (%) of several algorithms on the Hopkins 155 dataset

	Check-2		Check-3		Traffic-2		Traffic-3		Articul.-2		Articul.-3		All	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
GPCA (5)	6.09	1.03	31.95	32.93	1.41	**0.00**	19.83	19.55	2.88	**0.00**	16.85	16.85	10.34	2.54
GPCA (4N)	4.78	0.51	36.99	36.26	1.63	**0.00**	39.68	40.92	6.18	3.20	29.62	29.62	11.55	1.36
RANSAC (5)	6.52	1.75	25.78	26.00	2.55	0.21	12.83	11.45	7.25	2.64	21.38	21.38	9.76	3.21
ALC (5)	2.56	**0.00**	6.78	0.92	2.83	0.30	**4.01**	1.35	6.90	0.89	7.25	7.25	3.76	**0.26**
ALC (SP)	**1.49**	0.27	**5.00**	**0.66**	1.75	1.51	8.86	**0.51**	10.70	0.95	21.08	21.08	3.37	0.49
EM-MPPCA (5,a)	18.13	17.48	29.07	30.10	12.84	13.32	18.98	20.32	13.54	15.21	23.49	23.49	18.56	17.56
EM-MPPCA (4N,a)	24.85	24.75	37.01	38.07	18.46	18.15	29.03	26.04	12.90	14.11	32.11	32.11	24.88	23.44
DP-space (5)	2.13	0.48	9.86	7.26	**0.53**	0.20	4.31	2.57	3.79	1.90	**1.75**	**1.75**	**3.32**	0.53
DP-space (4N)	2.08	0.38	8.77	3.94	**1.33**	0.78	7.01	7.27	**2.07**	0.43	16.95	16.95	**3.29**	0.57
EM-MPPCA (5,m)	2.95	0.00	10.76	10.37	0.52	0.00	1.96	0.99	0.46	0.00	9.33	9.33	3.49	0.00
EM-MPPCA (4N,m)	6.56	3.55	19.35	19.96	0.81	0.00	12.03	9.39	0.18	0.00	16.14	16.14	7.28	1.09