Supplementary Information: "BasisVAE: Translation-invariant feature-level clustering with Variational Autoencoders"

A Derivation of the collapsed ELBO

Standard VAE methodology results in the bound

$$\log p(\mathbf{Y}) \geq \sum_{i=1}^{N} \mathbb{E}_{q_\phi(\mathbf{z}_i|\mathbf{y}_i)}[\log p_\theta(\mathbf{y}_i|\mathbf{z}_i)] - \mathrm{KL}(q_\phi(\mathbf{z}_i|\mathbf{y}_i)||p(\mathbf{z}_i))$$

However, unlike for standard VAE, for BasisVAE the $\log p_\theta(\mathbf{y}|\mathbf{z})$ term is intractable due to integration over $\mathbf{w}$ and $\boldsymbol{\pi}$. Now we apply the collapsing strategy of Hensman et al. (2012) to $\log p_\theta(\mathbf{y}_i|\mathbf{z}_i)$ for all data points $i = 1, \ldots, N$.

Knowing that

$$\begin{aligned}
\log p_\theta(\mathbf{y}_i|\mathbf{z}_i, \boldsymbol{\pi}) &= \log \int p_\theta(\mathbf{y}_i|\mathbf{z}_i, \boldsymbol{\pi}, \mathbf{w}) p(\mathbf{w}) d\mathbf{w} \\
&\geq \int q(\mathbf{w}) \log \frac{p_\theta(\mathbf{y}_i|\mathbf{z}_i, \mathbf{w}) p(\mathbf{w}|\boldsymbol{\pi})}{q(\mathbf{w})} d\mathbf{w} \\
&= \mathbb{E}_{q(\mathbf{w})} \log \frac{p_\theta(\mathbf{y}_i|\mathbf{z}_i, \mathbf{w}) p(\mathbf{w}|\boldsymbol{\pi})}{q(\mathbf{w})}
\end{aligned}$$

we can now lower bound

$$\begin{aligned}
\log p_\theta(\mathbf{y}_i|\mathbf{z}_i) &= \log \int p_\theta(\mathbf{y}_i|\mathbf{z}_i, \boldsymbol{\pi}) p(\boldsymbol{\pi}) d\pi \\
&\geq \log \int \exp\left(\mathbb{E}_{q(\mathbf{w})} \log \frac{p_\theta(\mathbf{y}_i|\mathbf{z}_i, \mathbf{w}) p(\mathbf{w}|\boldsymbol{\pi})}{q(\mathbf{w})} \right) p(\boldsymbol{\pi}) d\pi \\
&= \mathbb{E}_{q(\mathbf{w})} \log p_\theta(\mathbf{y}_i|\mathbf{z}_i, \mathbf{w}) + \log \int \exp\left(\mathbb{E}_{q(\mathbf{w})} p(\mathbf{w}|\boldsymbol{\pi}) \right) p(\boldsymbol{\pi}) d\pi - \mathbb{E}_{q(\mathbf{w})} \log q(\mathbf{w})
\end{aligned}$$

where now all integrals can be calculated in closed form. Combining the two lower bounds, we have

$$\begin{aligned}
\log p(\mathbf{Y}) \geq \sum_{i=1}^{N} \mathbb{E}_{q_\phi(\mathbf{z}_i|\mathbf{y}_i)} \mathbb{E}_{q(\mathbf{w})} \log p_\theta(\mathbf{y}_i|\mathbf{z}_i, \mathbf{w}) - \sum_{i=1}^{N} \mathrm{KL}(q_\phi(\mathbf{z}_i|\mathbf{y}_i)||p(\mathbf{z}_i)) \\
+ \log \int \exp\left(\mathbb{E}_{q(\mathbf{w})} p(\mathbf{w}|\boldsymbol{\pi}) \right) p(\boldsymbol{\pi}) d\pi \\
- \mathbb{E}_{q(\mathbf{w})} \log q(\mathbf{w})
\end{aligned}$$

Here, the first term can be calculated as

$$\sum_{i=1}^{N} \mathbb{E}_{q_\phi(\mathbf{z}_i|\mathbf{y}_i)} \mathbb{E}_{q(\mathbf{w})} \log p_\theta(\mathbf{y}_i|\mathbf{z}_i, \mathbf{w}) = \sum_{i=1}^{N} \mathbb{E}_{q_\phi(\mathbf{z}_i|\mathbf{y}_i)} \sum_{j=1}^{P} \sum_{k=1}^{K} \phi_{j,k} \log \mathcal{N}(y_i^{(j)}|\lambda_{j,k} f_{\text{basis}}^{(k)}(\mathbf{z}_i + \delta_{jk}),\ \sigma_j^2)$$

The remaining challenging term is the third one, but it has a closed form as follows: knowing that

$$\mathbb{E}_{q(\mathbf{w})} p(\mathbf{w}|\boldsymbol{\pi}) = \sum_{j=1}^{P} \sum_{k=1}^{K} \phi_{j,k} \log \pi_k = \sum_{k=1}^{K} n_k \log \pi_k$$

where we have denoted $n_k := \sum_{j=1}^{P} \phi_{j,k}$, the term can now be expressed

$$\begin{aligned}
\log \int \exp\left(\mathbb{E}_{q(\mathbf{w})} p(\mathbf{w}|\boldsymbol{\pi})\right) p(\boldsymbol{\pi}) d\pi &= \log \int \exp\left(\sum_{k=1}^{K} n_k \log \pi_k\right) p(\boldsymbol{\pi}) d\pi \\
&= \log \int \prod_{k=1}^{K} \pi_k^{n_k} \frac{1}{B(\boldsymbol{\alpha})} \pi_k^{\alpha_k - 1} d\boldsymbol{\pi} \\
&= \log \frac{1}{B(\boldsymbol{\alpha})} \int \prod_{k=1}^{K} \pi_k^{n_k + \alpha_k - 1} d\boldsymbol{\pi} \\
&= \log B(\mathbf{n} + \boldsymbol{\alpha}) - \log B(\boldsymbol{\alpha})
\end{aligned}$$

where the normalising constant $B(\boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)}$ is the multivariate Beta function.

Note that this hybrid inference scheme combines amortised reparameterisation-based inference for $\mathbf{z}$ and more classical approaches combined with collapsing for inference over $\mathbf{w}, \boldsymbol{\pi}$.

B Adaptation of ELBO for large data sets

For large high-dimensional data sets, the lower bound derived above will be dominated by the data log-likelihood. For large N, also the KL-term $\mathrm{KL}(q_\phi(\mathbf{z}_i|\mathbf{y}_i)||p(\mathbf{z}_i))$ might be large. As a result, the clustering prior that we have introduced will implicitly become relatively less important when N and P increase. While the property that for large data sets the likelihood will dominate the prior is inherent to Bayesian models, it may not always be desirable, especially for mis-specified models.

In practice, one way to alleviate this problem where the likelihood starts to dominate is via introducing weights that either downweigh the likelihood or upweigh the prior. For example, β-VAE modifies the usual VAE lower bound by scaling the KL term by a constant $\beta > 0$ (Higgins et al., 2017). Even though the resulting expression is not a lower bound on the original log marginal likelihood any more, it is closely connected to an ELBO on an alternative formulation with an annealed prior $p(\mathbf{z})^\beta / \int p(\mathbf{z})^\beta d\mathbf{z}$ (Hoffman et al., 2017; Mathieu et al., 2019). Analogously, we propose to achieve a similar effect, by introducing β as part of the following objective

$$\begin{aligned}
\mathcal{L}_\beta := &\sum_{i=1}^{N} \mathbb{E}_{q_\phi(\mathbf{z}_i|\mathbf{y}_i)} \mathbb{E}_{q(\mathbf{w})} \log p_\theta(\mathbf{y}_i|\mathbf{z}_i, \mathbf{w}) + \\
&+ \beta \left(\log \int \exp\left(\mathbb{E}_{q(\mathbf{w})} p(\mathbf{w}|\boldsymbol{\pi})\right) p(\boldsymbol{\pi}) d\pi - \mathbb{E}_{q(\mathbf{w})} \log q(\mathbf{w}) - \sum_{i=1}^{N} \mathrm{KL}(q_\phi(\mathbf{z}_i|\mathbf{y}_i)||p(\mathbf{z}_i)) \right)
\end{aligned}$$

where $\beta = 1$ corresponds to the original ELBO, but we may choose to select $\beta > 1$ in order to increase the relative importance of the sparse clustering prior for large-scale applications. In our experiments, for moderately sized synthetic data we used $\beta = 1$, whereas for single-cell gene expression data we used $\beta = 20$.

References

Hensman, J., Rattray, M., and Lawrence, N. D. (2012). Fast variational inference in the conjugate exponential family. In *Advances in neural information processing systems*, pages 2888–2896.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *International Conference on Learning Representations*, 2(5):6.

Hoffman, M. D., Riquelme, C., and Johnson, M. J. (2017). The beta-VAE's Implicit Prior. In *Workshop on Bayesian Deep Learning, NIPS*, pages 1–5.

Mathieu, E., Rainforth, T., Siddharth, N., and Teh, Y. W. (2019). Disentangling Disentanglement in Variational Autoencoders. In *International Conference on Machine Learning*, pages 4402–4412.

C Additional Figures

C.1 Additional figures for synthetic data

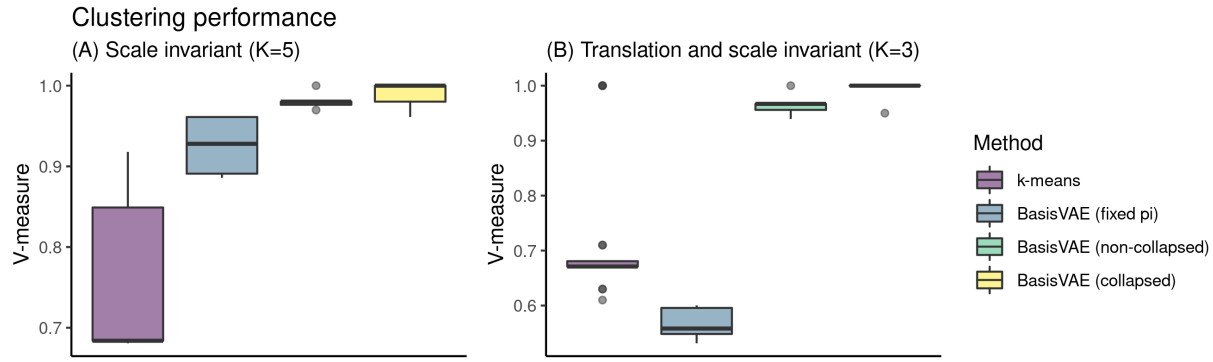

Figure 1: Using the V-measure to quantify the clustering performance of k-means and different versions of BasisVAE (with different inference techniques for $q(\mathbf{w}, \boldsymbol{\pi})$) on the synthetic data from Figure **??** . Results are shown for both (A) scale invariant and (B) translation and scale invariant setting. For k-means, we used the true number of clusters, i.e. (A) $K = 5$ and (B) $K = 3$. For BasisVAE, we used an overspecified $K = 20$.

C.2 Additional figures for single-cell Spermatogenesis data

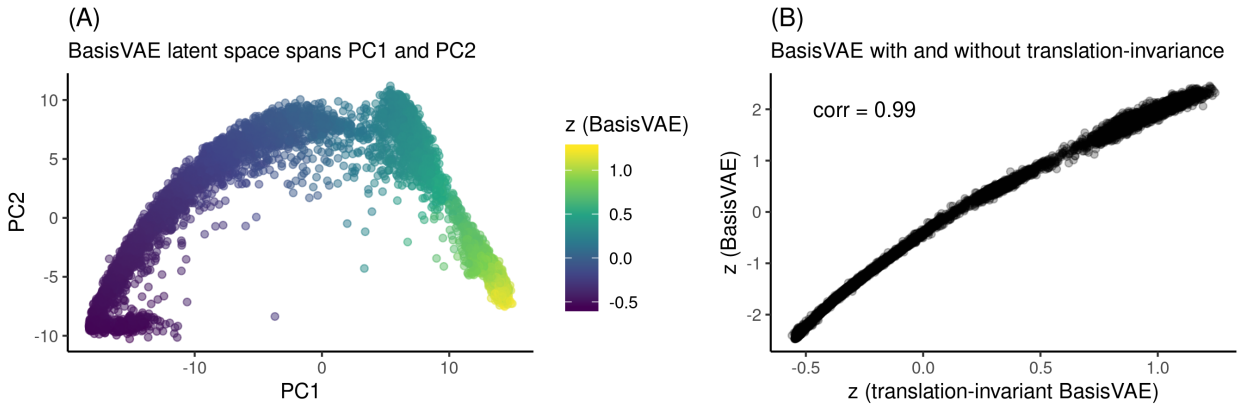

Figure 2: (A) BasisVAE has inferred a latent $\mathbf{z} \in \mathbb{R}$ that captures a trajectory in the (PC1, PC2) space. (B) BasisVAE and its translation-invariant version infer a highly similar latent space (the respective inferred latent coordinates are highly correlated).

D On well-definedness of translation-invariant BasisVAE

As a non-linear latent variable model, even the standard (C)VAE exhibits multiple modes in the sense that different latent configurations $\{\mathbf{z}_i\}_{i=1}^N$ can give rise to the same likelihood. Even in the probabilistic PCA which is a linear model, different rotations of the latent space are indistinguishable in terms of the likelihood. Having introduced additional parameters δ_{jk}, it is natural to ask whether the BasisVAE model is well-defined and whether it is identifiable with respect to δ_{jk} parameters.

Here we assume $\mathbf{z} \in \mathbb{R}$, denoted below as z. First, we note that without any constraints or priors on delta, our model is unidentifiable in two aspects:

- The inputs $z + \delta_{jk}$ can both be shifted by an arbitrary constant, i.e. $z + \delta_{jk} = (z + a_k) + (\delta_{jk} - a_k)$ for any $a_k \in \mathbb{R}$
- Suppose latent variables z are fixed, then for any $f^{(j)}$ and δ_{jk}, there exist $f^{*(j)}$ and δ^*_{jk} such that $f^{(j)}(z + \delta_{jk}) = f^{*(j)}(z + \delta^*_{jk})$

However, having placed a prior over delta, arbitrary reparameterisations as outlined above are not equally likely any more. For example, let us consider a thought experiment where every feature is assigned to its own basis function: even though for every feature j there exist multiple pairs $(f^{(j)}, \delta_{jk})$ of functional representations that are indistinguishable $f^{(j)}(z+\delta_{jk})$ in terms of their outputs, in this particular scenario the solution $f^{*(j)}$ where $\delta^*_{jk} = 0$ is preferred over any other combination. This is because indistinguishable functional representations result all have the same log-likelihood, and thus the configuration with $\delta_{jk} = 0$ achieves the highest ELBO in this scenario. Thus, the prior $p(\delta_{jk})$ alleviates the problem with unidentifiability.

We also note that the existence of these multiple modes would not be a problem from the interpretability perspective, as for this purpose it is the relative difference between δ_{jk} values that is of interest, not their absolute values.

Potential pathological scenario: Here we discuss a corner case where the translation-invariant BasisVAE formulation could potentially be ill-defined due to the added flexibility by δ_{jk}. Even though we have not experienced this in practice, we first describe this scenario, and then propose a solution how this pathological case can be avoided by a simple modification.

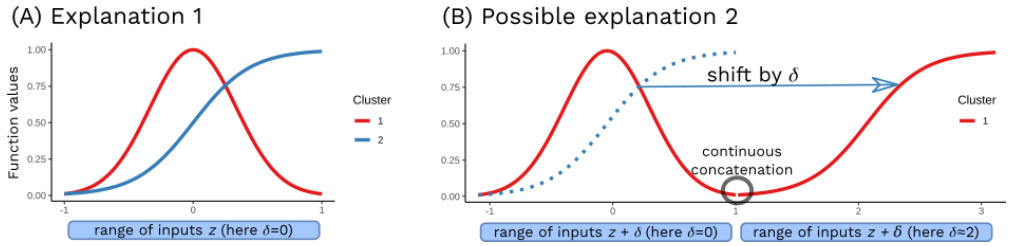

Figure 3: Illustration of an imaginary pathological scenario, where we would expect the two observed features (as shown in (A)) to be assigned to two separate clusters. Panel (B) describes a pathological case, where a large δ value for one feature allows the two curves to be joined and thus the two features to be assigned to the same cluster.

The idea behind the pathological scenario has been illustrated in above: when inputs z are in some interval $[a, b]$, then one can e.g. pick $\delta_{jk} = 0$ for one feature, $\delta_{jk} > (b - a)$ for another feature, and concatenate the two into one basis function (we note that the concatenation of the two functions must be continuous).

While we do not see this scenario as a probable risk in reality, we are able to avoid this potential behaviour altogether by introducing a constraint $|\delta_{jk} - \delta_{j'k'}| < (b-a)$, which can e.g. be easily implemented by explicitly restricting the range of all δ_{jk} values.