oi-VAE: Output Interpretable VAEs for Nonlinear Group Factor Analysis, Supplemental Material

Samuel K. Ainsworth, Nicholas J. Foti, Adrian K. C. Lee, Emily B. Fox

1 Source code

The source code is made available and maintained at https://github.com/samuela/oi-vae.

2 Common experimental details

We found that it was crucial to throttle the variance of the posterior approximation in order stabilize training in the initial stages of optimization for both the VAE and oi-VAE. We did so by multiplying the outputted standard deviations by 0.1 for the first 25 epochs and then resumed training normally after that point. A small $1e-3$ factor was added to all of the outputted standard deviations in order to promote numerical stability when calculating gradients.

In all of our experiments we estimated $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p(x|\mathbf{z}, \mathcal{W}, \theta)\right]$ with one sample. We experimented with using more samples but did not observe any significant benefit from doing so.

When plotting latent component/group interactions, as in Fig. 1, the L2 norm of the column vectors are taken and then these values are normalized to sum to one for each latent component in order to consistently show relative weights across latent components.

3 Synthetic bars data

In addition to the evaluations shown in the paper, we evaluated oi-VAE when the number of latent dimensions K is greater than necessary to fully explain the data. In particular we sample the same 8×8 images but use $K = 16$. See Figure 1. Train and test log likelihoods for the model are given in Table 1.

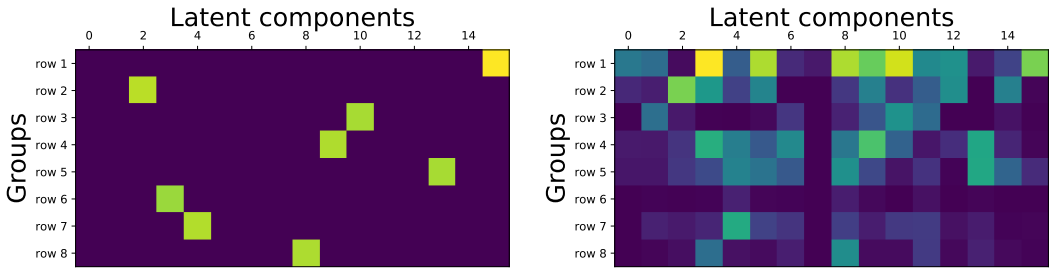
Experimental details For all our synthetic data experiments we sampled 2,048 8×8 images with exactly one bar present uniformly at random. The activated bar was given a value of 0.5, inactive pixels were given values of zero. White noise was added to the entire image with standard deviation 0.05. We set $p = 1$ and $\lambda = 1$.

- Inference model:
  - $\mu(\mathbf{x}) = W_1\mathbf{x} + b_1$.
  - $\sigma(\mathbf{x}) = \exp(W_2\mathbf{x} + b_2)$.
- Generative model:
  - $\mu(\mathbf{z}) = W_3\mathbf{z} + b_3$.
  - $\sigma = \exp(b_4)$.

We ran Adam on the inference and generative net parameters with learning rate $1e-2$. Proximal gradient descent was run on $\mathcal{W}$ with learning rate $1e-4$. We used a batch size of 64 sampled uniformly at random at each iteration and ran for 20,000 iterations.

Table 1: Train and test log likelihoods on the synthetic bars data when K is larger than necessary.

Model	Train log likelihood	Test log likelihood
$\lambda = 1$	**99.9325**	**100.1394**
$\lambda = 0$ and no θ prior	95.0687	95.4285



(a) oi-VAE with $\lambda = 1$

(b) No prior on the $\mathcal{W}$ or θ.

Figure 1: Results with latent dimension $K = 16$ when the effective dimensionality of the data is only 8. Clearly the oi-VAE has learned to use only the sparse set of z_i's that are necessary to explain the data.

4 Motion capture results

Multiple samples from both the VAE and oi-VAE are shown in Figure 2.

Experimental details We used data from subject 7 in the CMU Motion Capture Database. Trials 1-10 were used for training. Trials 11 and 12 were left out to form a test set. Trial 11 is a standard `walk` trial. Trial 12 is a `brisk walk` trial. We set $p = 8$ and $\lambda = 1$.

- Inference model:
  - $\mu(\mathbf{x}) = W_1\mathbf{x} + b_1$.
  - $\sigma(\mathbf{x}) = \exp(W_2\mathbf{x} + b_2)$.
- Generative model:
  - $\mu(\mathbf{z}) = W_3 \tanh(\mathbf{z}) + b_3$.
  - $\sigma = \exp(b_4)$.

We ran Adam on the inference and generative net parameters with learning rate $1e-3$. Proximal gradient descent was run on $\mathcal{W}$ with learning rate $1e-4$. We used a batch size of 64 with batches shuffled before every epoch. Optimization was run for 1,000 epochs.

5 MEG Analysis

We present the three most prominent components determined by summing $||\mathbf{W}^{(g)}_{\cdot,j}||_2$ over all groups g. These components turn out to be harder to interpret than some of the others presented indicating

that the norm of the group-weights may not be the best notion to determine interpretable components. However, this perhaps is not surprising with neuroimaging data. In fact, the strongest components inferred when applying PCA or ICA to neuroimaging data usually correspond to physiological artifacts such as eye movement or cardiac activity [UI97].

We depict the three most prominent latent components according to the group weights. We also depict component 7 which corresponds to the spatial attentional network that consists of a mix of auditory and visual regions. This arises because the auditory attentional network taps into the visual network.

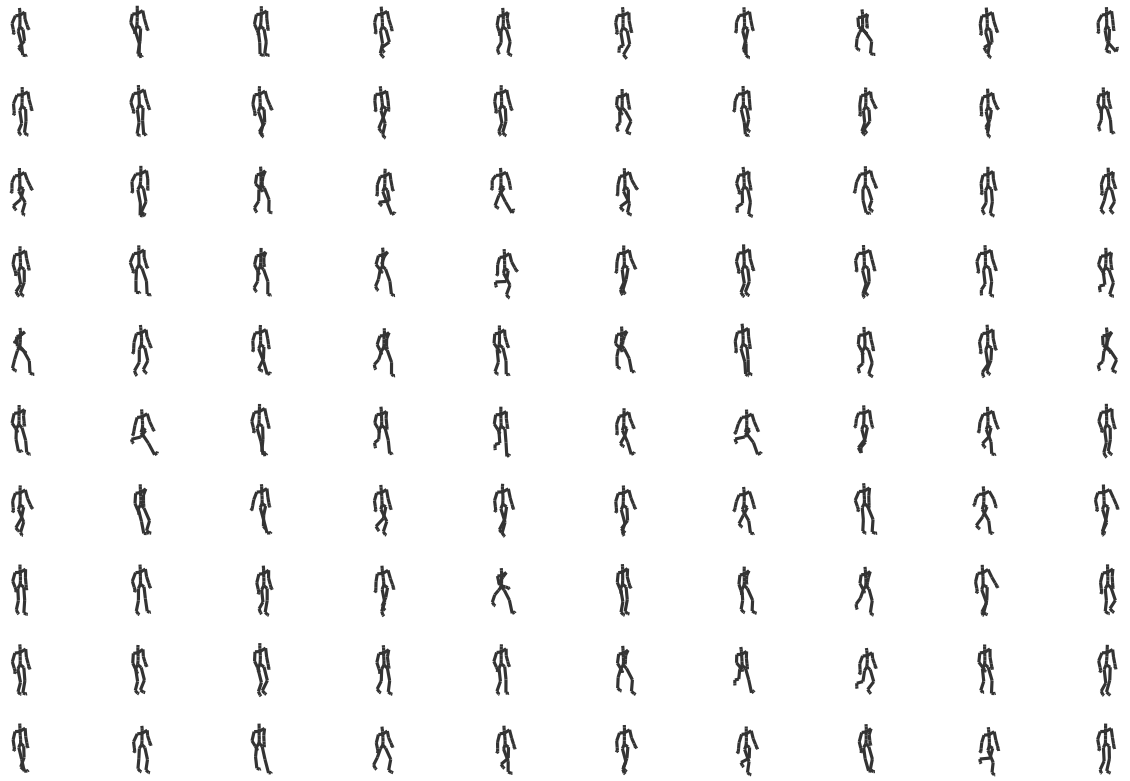
Experimental details We set $p = 10$ and $\lambda = 10$. The inference net was augmented with a hidden layer of 256 units.

- Inference model:
  - $\mu(\mathbf{x}) = W_2 \text{relu}(W_1 \mathbf{x} + b_1) + b_2$.
  - $\sigma(\mathbf{x}) = \exp(W_3 \text{relu}(W_1 \mathbf{x} + b_1) + b_3)$.
- Generative model:
  - $\mu(\mathbf{z}) = W_3 \tanh(\mathbf{z}) + b_3$.
  - $\sigma = \exp(b_4)$.

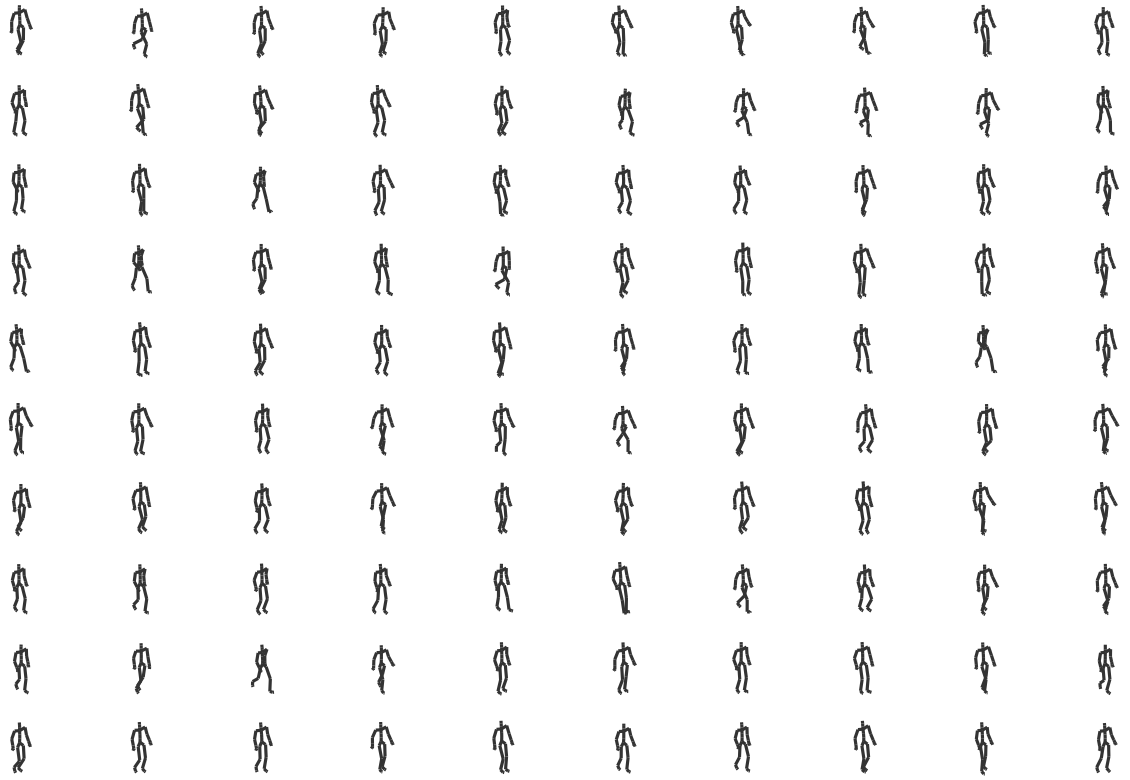
We ran Adam on the inference and generative net parameters with learning rate $1e-3$. Proximal gradient descent was run on $\mathcal{W}$ with learning rate $1e-6$. We used a batch size of 256 with batches shuffled before every epoch. Optimization was run for 40 epochs.

References

[UI97] M. A. Uusitalo and R. J. Ilmoniemi. Signal-space projection method for separating MEG or EEG into components. *Med Biol Eng Comput*, 35(2):135–140, 1997.

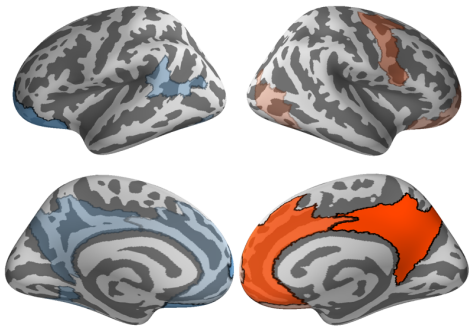


(a)



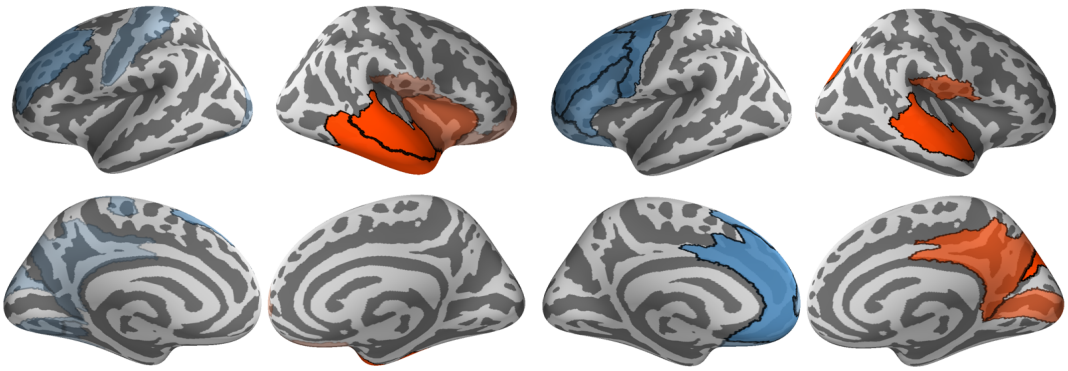
(b)

Figure 2: Samples from the (a) VAE and (b) oi-VAE models. The VAE produces a number of poses apparently inspired by the Ministry of Silly Walks. Some others are even physically impossible. In contrast, results from the oi-VAE are all physically plausible and appear to be representative of walking. Full scale images will be made available on the author's website.



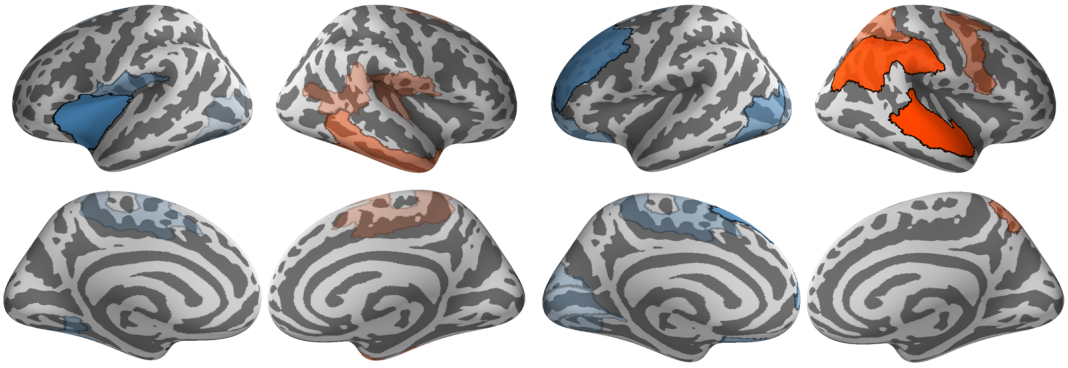
(a) Component 15: Default Mode Network

Figure 3: The projections of two components of $\mathbf{z}$ onto the regions of the HCP-MMP1 parcellation. The regions with the ten largest weights are shaded (blue in the left hemisphere, red in the right hemisphere) with opacity indicating the strength of the weight. Component 15 corresponds to the default mode network.



(a) Component 2. (b) Component 5.

Figure 4: Component 2 has the largest aggregate group weight and component 5 has the second largest.



(a) Component 11 (b) Component 7

Figure 5: Component 11 resembles the ventral stream and has the third largest aggregate group weight. Component 7 has a smaller aggregate group weight but corresponds to the spatial attentional network.