

# Interpretable VAEs for nonlinear group factor analysis

Samuel K. Ainsworth<sup>1</sup>, Nicholas J. Foti<sup>1</sup>, Adrian KC Lee<sup>2</sup>, and Emily B. Fox<sup>1</sup>

<sup>1</sup>School of Computer Science and Engineering, University of Washington

<sup>2</sup>Institute for Learning & Brain Sciences, University of Washington

February 15, 2018

## Abstract

Deep generative models have recently yielded encouraging results in producing subjectively realistic samples of complex data. Far less attention has been paid to making these generative models interpretable. In many scenarios, ranging from scientific applications to finance, the observed variables have a natural grouping. It is often of interest to understand systems of interaction amongst these groups, and latent factor models (LFMs) are an attractive approach. However, traditional LFMs are limited by assuming a linear correlation structure. We present an output interpretable VAE (oi-VAE) for grouped data that models complex, nonlinear latent-to-observed relationships. We combine a structured VAE comprised of group-specific generators with a sparsity-inducing prior. We demonstrate that oi-VAE yields meaningful notions of interpretability in the analysis of motion capture and MEG data. We further show that in these situations, the regularization inherent to oi-VAE can actually lead to improved generalization and learned generative processes.

## 1 Introduction

In many applications there is an inherent notion of groups associated with the observed variables. For example, in the analysis of neuroimaging data, studies are typically done at the level of regions of interest that aggregate over cortically-localized signals. In genomics, there are different treatment regimes. In finance, the data might be described in terms of asset classes (stocks, bonds, ...) or as collections of regional indices. Obtaining interpretable and expressive models of the data is critical to the underlying goals of descriptive analyses and decision making. The challenge arises from the push and pull between interpretability and expressivity in our modeling choices. Methods for extracting interpretability have focused primarily on linear models, resulting in lower expressivity. A popular choice in these settings is to consider sparse linear factor models (Zhao et al., 2016; Carvalho et al., 2008). However, it is well known that neural (Han et al., 2017), genomic (Prill et al., 2010), and financial data (Harvey et al., 1994), for example, exhibit complex nonlinearities.

On the other hand, there has been a significant amount of work on expressive models for complex, high dimensional data. Building on the framework of latent factor models, the Gaussian process latent variable model (GPLVM) (Lawrence, 2003) introduces nonlinear mappings from latent to observed variables. A group-structured GPLVM has also been proposed (Damianou et al., 2012). However, by relying on GPs, these methods do not scale straightforwardly to large datasets. In contrast, *deep generative models* (Kingma & Welling, 2013; Rezende et al., 2014) have proven wildly successful in efficiently modeling complex observations—such as images—as nonlinear mappings of simple latent representations. These nonlinear maps are based on deep neural networks and parameterize an observation distribution. As such, they can be viewed as nonlinear extensions of latent factor models. However, the focus has primarily been on their power as a generative mechanism rather than in the context of traditional latent factor modeling and associated notions of interpretability.

One efficient way of training deep generative models is via the *variational autoencoder* (VAE). The VAE posits an approximate posterior distribution over latent representations that is parameterized by a deep neural network, called the *inference network*, that maps observations to a distribution over latent variables. This direct mapping of observations to latent variables is called *amortized inference* and alleviates the need to determine individual latent variables for all observations. The parameters of both the generator and inference neural networks can then be determined using Monte Carlo variational inference (Kingma & Welling, 2013; Rezende et al., 2014). The VAE can be interpreted as a nonlinear factor model that provides a scalable means of learning the latent representations.

In this work we propose an *output interpretable VAE* (oi-VAE) for grouped data, where the focus is on interpretable interactions amongst the grouped outputs. Here, as in standard latent factor models, interactions are induced through shared latent variables. Interpretability is achieved via sparsity in the latent-to-observed mappings. To this end, we reformulate the VAE as a nonlinear factor model with a generator neural network for each group and incorporate a sparsity inducing penalty encouraging each latent dimension to influence a small number of correlated groups. We develop an amortized variational inference algorithm for a collapsed variational objective and use a proximal update to learn latent-dimension-to-group interactions. As such, our method scales to massive datasets allowing flexible analysis of data arising in many applications.

We evaluate the oi-VAE on motion capture and magnetoencephalography datasets. In these scenarios where there is a natural notion of groupings of observations, we demonstrate the interpretability of the learned features and how these structures of interaction correspond to physically meaningful systems. Furthermore, in such cases, we show that the regularization employed by oi-VAE leads to better generalization and synthesis capabilities, especially in limited training data scenarios or when the training data might not fully capture the observed space of interest.

## 2 Background

The study of deep generative models is an active area of research in the machine learning community and encompasses probabilistic models of data that can be used to generate observations from the underlying distribution. The variational autoencoder (VAE) (Kingma & Welling, 2013) and the related deep Gaussian model (Rezende et al., 2014) both propose the idea of amortized inference to perform variational inference in probabilistic models that are parameterized by deep neural networks. Further details on the VAE specification are provided in Sec. 3. The variational objective is optimized with (stochastic) gradient descent and the intractable expectation arising in the objective is evaluated with Monte Carlo samples from the variational distribution. The method is referred to as *Monte Carlo variational inference*, and has become popular for performing variational inference in generative models. See Sec. 5.

The VAE approach has recently been extended to more complex data such as that arising from dynamical systems (Archer et al., 2015) and also to construct a generative model of cell structure and morphology (Johnson et al., 2017). Though deep generative models and variational autoencoders have demonstrated the ability to produce convincing samples of complex data from complicated distributions, the learned latent representations are not easily interpretable due to the complex interactions from latent dimensions to the observations, as depicted in Fig. 2.

A common approach to encourage simple and interpretable models is through use of *sparsity inducing penalties* such as the *lasso* (Tibshirani, 1994) and *group lasso* (Yuan & Lin, 2006). These methods work by shrinking many model parameters toward zero and have seen great success in regression models, covariance selection (Danaher et al., 2014), and linear factor analysis (Hirose & Konishi, 2012). The group lasso penalty is of particular interest in our group analysis as it simultaneously shrinks entire groups of model parameters toward zero. Commonly, sparsity inducing penalties are considered in the convex optimization literature due to their computational tractability using proximal gradient descent (Parikh & Boyd, 2013).

Though these convex penalties have proven very useful, we cannot apply them directly and obtain a



Figure 1: VAE (*left*) and oi-VAE (*right*) generative models. The oi-VAE considers group-specific generators and a linear latent-to-generator mapping with weights from a single latent dimension to a specific group sharing the same color. The group-sparse prior is applied over these grouped weights.

valid generative model. Instead, we need to consider prior specifications over the parameters of the generator network that likewise yield sparsity. Originally, the Bayesian approach to sparsity was based on the spike-and-slab prior, a two-component mixture that puts some probability on a model parameter being exactly zero (the spike) and some probability of the parameter taking on non-zero values (the slab) (Mitchell & Beauchamp, 1988). Unfortunately, inference in models with the spike-and-slab prior is difficult because of the combinatorial nature of the resulting posterior.

Recently, Bayesian formulations of sparsity inducing penalties take the form of hierarchical prior distributions that shrink many model parameters to small values (though not exactly zero). Such *global-local shrinkage* priors encapsulate a wide variety of hierarchical Bayesian priors that attempt to infer interpretable models, such as the horseshoe prior (Bhadra et al., 2016). These priors also result in efficient inference algorithms.

A sophisticated hierarchical Bayesian prior for sparse group linear factor analysis has recently been developed by (Zhao et al., 2016). This prior encourages both a sparse set of factors to be used as well as having the factors themselves be sparse. Additionally, the prior admits an efficient inference scheme via expectation-maximization. Sparsity inducing hierarchical Bayesian priors have been applied to Bayesian deep learning models to learn the complexity of the deep neural network (Louizos et al., 2017; Ghosh & Doshi-Velez, 2017). Our focus, however, is on using (structured) sparsity-inducing hierarchical Bayesian priors in the context of deep learning for the sake of interpretability, as in linear factor analysis, rather than model selection.

### 3 The OI-VAE model

We frame our proposed output interpretable VAE (oi-VAE) method using the same terminology as the VAE. Let  $\mathbf{x} \in \mathbb{R}^D$  denote a  $D$ -dimensional observation and  $\mathbf{z} \in \mathbb{R}^K$  denote the associated  $K$ -dimensional latent representation. We then write the generative process of the model as:

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (1)$$

$$\mathbf{x} \sim \mathcal{N}(f_{\theta}(\mathbf{z}), \mathbf{D}), \quad (2)$$

where  $\mathbf{D}$  is a diagonal matrix containing the marginal variances of each component of  $\mathbf{x}$ . The generator is encoded with the function  $f_{\theta}(\cdot)$  specified as a deep neural network with parameters  $\theta$ . Note that the formulation in Eq. (1) is simpler than that described in Kingma & Welling (2013) as we assume the observation variances are global parameters and not observation specific. This simplifying assumption follows from that of traditional factor models, but could easily be relaxed.

When our observations  $\mathbf{x}$  admit a natural grouping over the components, we write  $\mathbf{x}$  as  $[\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(G)}]$  for our  $G$  groups. We model the components within each group  $g$  with separate generative networks  $f_{\theta_g}^{(g)}$

parameterized by  $\theta_g$ . It is possible to share generator parameters  $\theta_g$  across groups, however we choose to model each separately. Critically, the latent representation  $\mathbf{z}$  is shared over all the group-specific generators. In particular:

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (3)$$

$$\mathbf{x}^{(g)} \sim \mathcal{N}(f_{\theta_g}^{(g)}(\mathbf{z}), \mathbf{D}_g). \quad (4)$$

To this point, our specified group-structured VAE can leverage within-group correlation structure and between-group independencies. However, one of the main goals of this framework is to capture interpretable relationships between group-specific activations through the latent representation. Note that it is straightforward to apply different likelihoods on different groups, although we did not have reason to do so in our experiments.

Inspired by the sparse factor analysis literature, we extract notions of interpretable interactions through inducing sparse latent-to-group mappings. Specifically, we insert a group-specific linear transformation  $\mathbf{W}^{(g)} \in \mathbb{R}^{p \times K}$  between the latent representation  $\mathbf{z}$  and the group generator  $f^{(g)}$ :

$$\mathbf{x}^{(g)} \sim \mathcal{N}(f_{\theta}^{(g)}(\mathbf{W}^{(g)}\mathbf{z}), \mathbf{D}_g). \quad (5)$$

We refer to  $\mathbf{W}^{(g)}$  as the *latent-to-group matrix*. We assume that the input dimension  $p$  per generator is the same, but this could be generalized. When the  $j$ th column of the group- $g$  latent-to-group matrix,  $\mathbf{W}_{:,j}^{(g)}$ , is all zeros then the  $j$ th latent dimension,  $\mathbf{z}_j$ , will have no influence on group  $g$ . To induce this column-wise sparsity, we place a hierarchical Bayesian prior on the columns  $\mathbf{W}_{:,j}^{(g)}$  as follows (Kyung et al., 2010):

$$\gamma_{gj}^2 \sim \text{Gamma}\left(\frac{p+1}{2}, \frac{\lambda^2}{2}\right) \quad (6)$$

$$\mathbf{W}_{:,j}^{(g)} \sim \mathcal{N}(\mathbf{0}, \gamma_{gj}^2 \mathbf{I}) \quad (7)$$

where  $\text{Gamma}(\cdot, \cdot)$  is defined by shape and rate, and  $p$  denotes the number of rows in each  $\mathbf{W}^{(g)}$ . The rate parameter  $\lambda$  defines the amount of sparsity, with larger  $\lambda$  implying more column-wise sparsity in  $\mathbf{W}^{(g)}$ . Marginalizing over  $\gamma_{gj}^2$  induces group sparsity over the columns of  $\mathbf{W}^{(g)}$ ; the MAP of the resulting posterior is equivalent to a group lasso penalized objective (Kyung et al., 2010).

While we are close to a workable model, one wrinkle remains. Unlike linear factor models, the deep structure of our model permits it to push rescaling across layer boundaries without affecting the end behavior of the network. In particular, it is possible—and in fact encouraged behavior—to learn a set of  $\mathbf{W}^{(g)}$  matrices with very small weights only to have the values revived to “appropriate” magnitudes in the following layers of  $f_{\theta_g}^{(g)}$ . In order to mitigate such behavior we additionally place a standard normal prior on the parameters of each generative network,  $\theta_g \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ , completing the model specification.

**Special cases of the oi-VAE** There are a few notable special cases of our oi-VAE framework. When we treat the observations as forming a single group, the model resembles a traditional VAE since there is a single generator. However, the sparsity inducing prior still has an effect that differs from the standard VAE specification. In particular, by shrinking columns of  $\mathbf{W}$  the prior will essentially encourage a sparse subset of the components of  $\mathbf{z}$  to be used to explain the data, similar to a traditional sparse factor model. Note that the  $\mathbf{z}$ ’s themselves will not necessarily be sparse, but the columns of  $\mathbf{W}$  will indicate which components are used. (Here, we drop the  $g$  superscript on  $\mathbf{W}$ .) This regularization can be advantageous to apply even when the data only has one group as it can provide improved generalization performance in the case of limited training data. Another special case arises when the generator networks are given by the identity mapping. In this case, the only transformation of the latent representation is given by  $\mathbf{W}^{(g)}$  and the oi-VAE reduces to a group sparse linear factor model.

## 4 Interpretability of the oi-VAE

In the oi-VAE, each latent factor influences a sparse set of the observational groups. The interpretability garnered from this sparse structure is two-fold:

**Disentanglement of latent embeddings** By associating each component of  $\mathbf{z}$  with only a sparse subset of the observational groups, we are able to quickly identify *disentangled* representations in the latent space. That is, by penalizing interactions between the components of  $\mathbf{z}$  and each of the groups, we effectively force the model to arrive at a representation that minimizes correlation across the components of  $\mathbf{z}$ , encouraging each dimension to capture distinct modes of variation. For example, in Table 1 we see that each of the dimensions of the latent space learned on motion capture recordings of human motion corresponds to a direction of variation relevant to only a subset of the joints (groups) that are used in specific submotions related to walking. Additionally, it is observed that although the VAE and oi-VAE have similar reconstruction performance the meaningfully disentangled latent representation allows oi-VAE to produce superior unconditional random samples.

**Discovery of group interactions** From the perspective of the latent representation  $\mathbf{z}$ , each latent dimension influences only a sparse subset of the observational groups. As such, we can view the observational groups associated with a specific latent dimension as a related system of sorts. For example, in neuroscience the groups could correspond to different brain regions from a standard parcellation. If a particular dimension of  $\mathbf{z}$  influences the generators of a small set of groups, then those groups can be interpreted as a system of regions that can be treated as a unit of analysis. Such an approach is attractive in the context of analyzing functional connectivity from MEG data where we seek modules of highly correlated regions. See the experiments of Sec. 6.3. Likewise, in our motion capture experiments of Sec. 6.2, we see (again from Table 1) how we can treat collections of joints as a system that covary in meaningful ways within a given human motion category.

Broadly speaking, the relationship between dimensions of  $\mathbf{z}$  and observational groups can be thought of as a bipartite graph in which we can quickly identify correlation and independence relationships among the groups themselves. The ability to expose or refute correlations among observational groups is attractive as an exploratory scientific tool independent of building a generative model. This is especially useful since standard measures of correlation are linear, leaving much to be desired in the face of high-dimensional data with many potential nonlinear relationships. Our hope is that oi-VAE serves as one initial tool to spark a new wave of interest in nonlinear factor models and their application to complicated and rich data across a variety of fields.

It is worth emphasizing that the goal is *not* to learn sparse representations in the  $\mathbf{z}$ ’s. Sparsity in  $\mathbf{z}$  may be desirable in certain contexts, but it does not actually provide any interpretability in the data generating process. In fact, excessive compression of the latent representation  $\mathbf{z}$  through sparsity could be detrimental to interpretability.

## 5 Collapsed variational inference

Traditionally, VAEs are learned by applying stochastic gradient methods directly to the evidence lower bound (ELBO):

$$\log p(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})],$$

where  $q_\phi(\mathbf{z}|\mathbf{x})$  denotes the amortized posterior distribution of  $\mathbf{z}$  given observation  $\mathbf{x}$ , parameterized with a deep neural network with weights  $\phi$ . Using a neural network to parameterize the observation distribution  $p(\mathbf{x}|\mathbf{z})$  as in Eq. (1) makes the expectation in the ELBO intractable. To address this, the VAE employs Monte Carlo variational inference (MCVI) (Kingma & Welling, 2013): The troublesome expectation is approximated with samples of the latent variables from the variational distribution,

$\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$ , where  $q_\phi(\mathbf{z}|\mathbf{x})$  is *reparameterized* to allow differentiating through the expectation operator and reduce gradient variance.

We extend the basic VAE amortized inference procedure to incorporate our sparsity inducing prior over the columns of the latent-to-group matrices. The naive approach of optimizing variational distributions for the  $\gamma_{gj}^2$  and  $\mathbf{W}_{:,j}^{(g)}$  will not result in true sparsity of the columns  $\mathbf{W}_{:,j}^{(g)}$ . Instead, we consider a collapsed variational objective function. Since our sparsity inducing prior over  $\mathbf{W}_{:,j}^{(g)}$  is marginally equivalent to the convex group lasso penalty we can use proximal gradient descent on the collapsed objective and obtain true group sparsity (Parikh & Boyd, 2013). Following the standard VAE approach of Kingma & Welling (2013), we use simple point estimates for the variational distributions on the neural network parameters  $\mathcal{W} = (\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(G)})$  and  $\theta = (\theta_1, \dots, \theta_G)$ . We take  $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mu(\mathbf{x}), \sigma^2(\mathbf{x}))$  where the mean and variances are parameterized by an inference network with parameters  $\phi$ .

## 5.1 The collapsed objective

We construct a collapsed variational objective by marginalizing the  $\gamma_{gj}^2$  to compute  $\log p(\mathbf{x})$  as:

$$\begin{aligned} \log p(\mathbf{x}) &= \log \int p(\mathbf{x}|\mathbf{z}, \mathcal{W}, \theta) p(\mathbf{z}) p(\mathcal{W}|\gamma^2) p(\gamma^2) p(\theta) d\gamma^2 dz \\ &= \log \int \left( \int p(\mathcal{W}, \gamma^2) d\gamma^2 \right) \frac{p(\mathbf{x}|\mathbf{z}, \mathcal{W}, \theta) p(\mathbf{z}) p(\theta)}{q_\phi(\mathbf{z}|\mathbf{x}) / q_\phi(\mathbf{z}|\mathbf{x})} dz \\ &\geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log p(\mathbf{x}|\mathbf{z}, \mathcal{W}, \theta)] - \mathbb{KL}(q_\phi(\mathbf{z}|\mathbf{x}) || p(\mathbf{z})) \\ &\quad + \log p(\theta) - \lambda \sum_{g,j} \|\mathbf{W}_{:,j}^{(g)}\|_2 \\ &\triangleq \mathcal{L}(\phi, \theta, \mathcal{W}). \end{aligned}$$

Importantly, the columns of the latent-to-group matrices  $\mathbf{W}_{:,j}^{(g)}$  appear in a 2-norm penalty in the new collapsed ELBO. This is exactly a group lasso penalty on the columns of  $\mathbf{W}_{:,j}^{(g)}$  and encourages the entire vector to be set to zero.

Now our goal becomes maximizing this collapsed ELBO over  $\phi, \theta, \mathcal{W}$ . Since this objective contains a standard group lasso penalty, we can leverage efficient proximal gradient descent updates on the latent-to-group matrices  $\mathcal{W}$  as detailed in Sec. 5.2. Proximal algorithms achieve better rates of convergence than sub-gradient methods and have shown great success in solving convex objectives with group lasso penalties. We can use any off-the-shelf optimization method for the remaining neural net parameters,  $\theta_g$  and  $\phi$ .

## 5.2 Proximal gradient descent

Proximal gradient descent algorithms are a broad class of optimization techniques for separable objectives with both differentiable and potentially non-differentiable components,

$$\min_x g(x) + h(x), \tag{8}$$

where  $g(x)$  is differentiable and  $h(x)$  is potentially non-smooth or non-differentiable (Parikh & Boyd, 2013). Stochastic proximal algorithms are well-studied for convex optimization problems. Recent work has shown that they are guaranteed to converge to a local stationary point even if the objective is comprised of a non-convex  $g(x)$  as long as the non-smooth  $h(x)$  is convex (Reddi et al., 2016). The usual tactic is to take gradient steps on  $g(x)$  followed by “corrective” *proximal* steps to respect  $h(x)$ :

$$x_{t+1} = \text{prox}_{\eta h}(x_t - \eta \nabla g(x_t)) \tag{9}$$

---

**Algorithm 1** Collapsed VI for oi-VAE

---

**Input:** data  $\mathbf{x}^{(i)}$ , sparsity parameter  $\lambda$   
Let  $\tilde{\mathcal{L}}$  be  $\mathcal{L}(\phi, \theta, \mathcal{W})$  but without  $-\lambda \sum_{g,j} \|\mathbf{W}_{:,j}^{(g)}\|_2$ .  
**repeat**  
    Calculate  $\nabla_{\phi} \tilde{\mathcal{L}}$ ,  $\nabla_{\theta} \tilde{\mathcal{L}}$ , and  $\nabla_{\mathcal{W}} \tilde{\mathcal{L}}$ .  
    Update  $\phi$  and  $\theta$  with an optimizer of your choice.  
    Let  $\mathcal{W}_{t+1} = \mathcal{W}_t - \eta \nabla_{\mathcal{W}} \tilde{\mathcal{L}}$ .  
    **for all** groups  $g, j = 1$  **to**  $K$  **do**  
        Set  $\mathbf{W}_{:,j}^{(g)} \leftarrow \frac{\mathbf{W}_{:,j}^{(g)}}{\|\mathbf{W}_{:,j}^{(g)}\|_2} \left( \|\mathbf{W}_{:,j}^{(g)}\|_2 - \eta \lambda \right)_+$   
    **end for**  
**until** convergence in both  $\hat{\mathcal{L}}$  and  $-\lambda \sum_{g,j} \|\mathbf{W}_{:,j}^{(g)}\|_2$

---

where  $\text{prox}_f(x)$  is the proximal operator for the function  $f$ . For example, if  $h(x)$  is the indicator function for a convex set then the proximal operator is simply the projection operator onto the set and the update in Eq. (9) is projected gradient. Expanding the definition of  $\text{prox}_{\eta h}$  in Eq. (9), one can show that the proximal step corresponds to minimizing  $h(x)$  plus a quadratic approximation to  $g(x)$  centered on  $x_t$ . For  $h(x) = \lambda \|x\|_2$ , the proximal operator is given by

$$\text{prox}_{\eta h}(x) = \frac{x}{\|x\|_2} (\|x\|_2 - \eta \lambda)_+ \quad (10)$$

where  $(v)_+ \triangleq \max(0, v)$  (Parikh & Boyd, 2013). This operator is especially convenient since it is both cheap to compute and results in machine-precision zeros, unlike many hierarchical Bayesian approaches to sparsity that result in small but non-zero values. These methods require an extra thresholding step that our oi-VAE method does not due to attain exact zeros. Geometrically, this operator reduces the norm of  $x$  by  $\eta \lambda$ , and shrinks  $x$ 's with  $\|x\|_2 \leq \eta \lambda$  to zero.

We experimented with standard (non-collapsed) variational inference as well as other schemes, but found that collapsed variational inference with proximal updates provided faster convergence and succeeded in identifying sparser models than other techniques. In practice we apply proximal stochastic gradient updates per Eq. (9) on the  $\mathcal{W}$  matrices and Adam (Kingma & Ba, 2014) on the remaining parameters. See Alg. 1 for oi-VAE pseudocode.

## 6 Experiments

### 6.1 Synthetic data

In order to evaluate oi-VAE's ability to identify sparse models on well-understood data, we generated  $8 \times 8$  images with one randomly selected row of pixels shaded and additive noise corrupting the entire image. We then built and trained an oi-VAE on the images with each group defined as an entire row of pixels in the image. We used an 8-dimensional latent space in order to encourage the model to associate each dimension of  $\mathbf{z}$  with a unique row in the image. Results are shown in Fig. 2. Our oi-VAE successfully disentangles each of the dimensions of  $\mathbf{z}$  to correspond to exactly one row (group) of the image. We also trained an oi-VAE with a 16-dimensional latent space (see the Supplement) and see that when additional latent components are not needed to describe any group they are pruned from the model.

### 6.2 Motion Capture

Using data collected from CMU's motion capture database we evaluated oi-VAE's ability to handle complex physical constraints and interactions across groups of joint angles while simultaneously



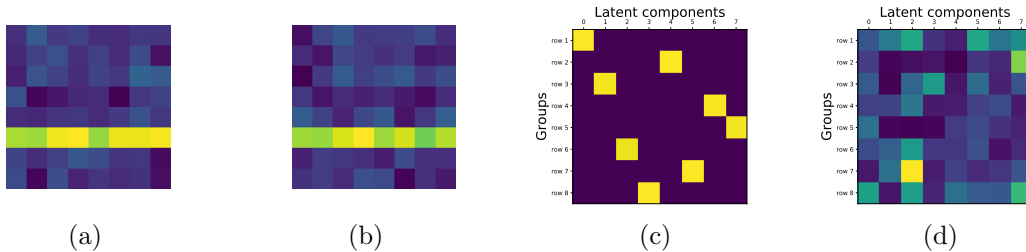


Figure 2: oi-VAE results on synthetic bars data. (a) Example image and (b) oi-VAE reconstruction. Learned oi-VAE  $\mathbf{W}_{:,j}^{(g)}$  for (c)  $\lambda = 1$  and (d)  $\lambda = 0$  (group structure, but no sparsity). In this case, training and test error numbers are nearly identical.

identifying a sparse decomposition of human motion. The dataset consists of 11 examples of **walking** and one example of **brisk walking** from the same subject. The recordings measure 59 joint angles split across 29 distinct joints. The joint angles were normalized from their full ranges to lie between zero and one. We treat the set of measurements from each distinct joint as a group; since each joint has anywhere from 1 to 3 observed degrees of freedom, this setting demonstrates how oi-VAE can handle variable-sized groups. For training, we randomly sample 1 to 10 examples of **walking**, resulting in up to 3791 frames. Our experiments evaluate the following performance metrics: interpretability of the learned interaction structure amongst groups and of the latent representation; test log-likelihood, assessing the model’s generalization ability; and both conditional and unconditional samples to evaluate the quality of the learned generative process. In all experiments, we use  $\lambda = 1$  with the reconstruction loss normalized by the dataset size. For further details on the specification of all considered models (VAE and oi-VAE), see the Supplement.

To begin, we train our oi-VAE on the full set of 10 training trials with the goal of examining the learned latent-to-group mappings. To explore how the learned disentangled latent representation varies with latent dimension  $K$ , we use  $K = 4, 8$ , and  $16$ . The results are summarized in Fig. 3. We see that as  $K$  increases, individual “features” (i.e., components of  $\mathbf{z}$ ) are refined to capture more localized anatomical structures. For example, feature 2 in the  $K = 4$  case turns into feature 7 in the  $K = 16$  case, but in that case we also add feature 3 to capture just variations of **lfingers**, **lthumb** separate from **head**, **upperneck**, **lowerneck**. Likewise, feature 2 when  $K = 16$  represents **head**, **upperneck**, **lowerneck** separately from **lfingers**, **lthumb**. To help interpret the learned disentangled latent representation, for the  $K = 16$  embedding we provide lists of the 3 joints per dimension that are most strongly influenced by that component. From these lists, we see how the learned decomposition of the latent representation has an intuitive anatomical interpretation. For example, in addition to the features described above, one of the very prominent features is feature 14, which jointly influences the **thorax**, **upperback**, and **lowerback**. Collectively, these results clearly demonstrate how the oi-VAE provides meaningful interpretability. We emphasize that it is not even possible to make these types of images or lists for the VAE.

One might be concerned that by gaining interpretability, we lose out on expressivity. However, as we demonstrate in Table 2 and Figs. 4-5, the regularization provided by our sparsity-inducing penalty actually leads to as good or better performance across various metrics of model fit. We first examine oi-VAE and VAE’s ability to generalize to held out data. To examine robustness to different amounts of training data, we consider training on increasing numbers of **walk** trials and testing on a single heldout example of either **walk** or **brisk walk**. The latter represents an example of data that is a slight variation of what was trained on, whereas the former is a heldout example that is very similar to the training examples. In Table 2, we see the benefit of the regularization in oi-VAE in both test scenarios in the limited data regime. Not surprisingly, for the full 10 trials, there are little to no differences between the generalization abilities of oi-VAE and VAE (though of course the oi-VAE still provides interpretability). We highlight that when we have both a limited amount of training



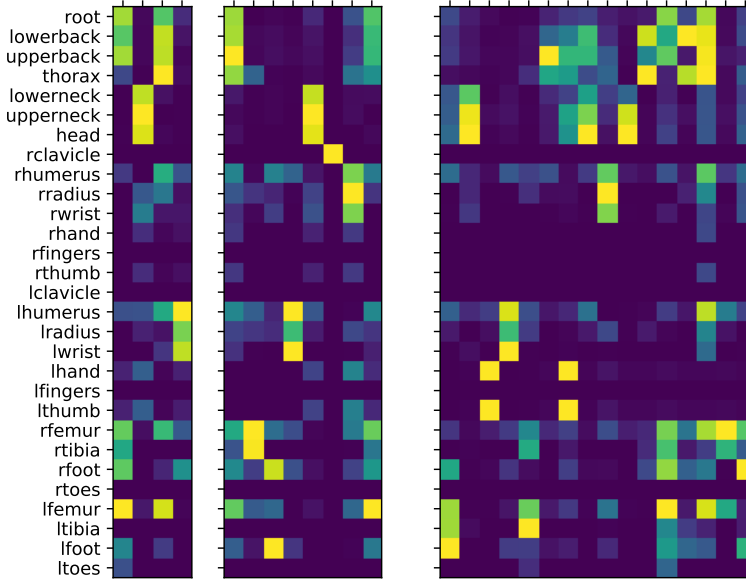


Figure 3: oi-VAE results on motion capture data with  $K = 4, 8$ , and  $16$ . Rows correspond to group generators for each of the joints in the skeleton, columns correspond to individual dimensions of the latent code, and values in the heatmap show the strength of the latent-to-group mappings  $\mathbf{W}_{\cdot,j}^{(g)}$ . Note, joints that experience little motion when walking—clavicles, fingers, and toes—have been effectively pruned from the latent code in all 3 models.

data that might not be fully representative of the full possible dataset of interest (e.g., all types of walking), the regularization provided by oi-VAE provides dramatic improvements for generalization. Finally, in almost all scenarios, the more decomposed oi-VAE  $K = 16$  setting has better or comparable performance to smaller  $K$  settings.

Next, we turn to assessing the learned oi-VAE’s generative process relative to that of the VAE. In Fig. 4 we take our test trial of `walk`, run each frame through the learned inference network to get a set of latent embeddings  $\mathbf{z}$ . For each such  $\mathbf{z}$ , we sample 32 times from  $q_\phi(\mathbf{z}|\mathbf{x})$  and run each through the generator networks to synthesize a new frame mini-“sequence”, where really the elements of this sequence are the perturbed samples about the embedded test frame. To fully explore the space of human motion the learned generators can capture, in Fig. 5 we sample the latent space at random 100 times from the prior. For each *unconditional* sample of  $\mathbf{z}$ , we pass it through the trained generator to create new frames. A representative subset of these frames is shown in Fig. 5. We also show similarly sampled frames from the trained VAE. A full set of 100 random samples from both VAE and oi-VAE are provided in the Supplement. Note that, even when trained on the full set of 10 `walk` trials where we see little to no difference in test log-likelihood between the oi-VAE and VAE, we do see that the learned generator for the oi-VAE is more representative of physically plausible human motion poses. We attribute this to the fact that the generators of the oi-VAE are able to focus on local correlation structure.

### 6.3 Magnetoencephalography

Magnetoencephalography (MEG) records the weak magnetic field produced by the brain during cognitive activity with great temporal resolution and good spatial resolution. Analyzing this data holds great promise for understanding the neural underpinnings of cognitive behaviors and for characterizing neurological disorders such as autism. A common step when analyzing MEG data is to project the

Table 1: Top 3 joints associated with each latent dimension. Grayscale values determined by  $\mathbf{W}_{\cdot,j}^{(g)}$ . We see kinematically associated joints associated with each latent dimension.

DIM.	TOP 3 JOINTS
1	left foot, left lower leg, left upper leg
2	head, upper neck, lower neck
3	left thumb, left hand, left upper arm
4	left wrist, left upper arm, left lower arm
5	left lower leg, left upper leg, right lower leg
6	upper back, thorax, lower back
7	left hand, left thumb, upper back
8	head, upper neck, lower back
9	right lower arm, right wrist, right upper arm
10	head, upper neck, lower neck
11	thorax, lower back, upper back
12	left upper leg, right foot, root
13	lower back, thorax, right upper leg
14	thorax, upper back, lower back
15	right upper leg, right lower leg, left upper leg
16	right foot, right upper leg, left foot

Table 2: Test log-likelihood for VAE and oi-VAE trained on 1,2,5, or 10 trials of **walk** data. Table includes results for a test **walk** (same as training) or **brisk walk** trial (unseen in training). Bold numbers indicate the best performance.

STANDARD WALK				
# TRIALS	1	2	5	10
VAE ( $K = 16$ )	-3,518	-251	18	<b>114</b>
oi-VAE ( $K = 4$ )	<b>-2,722</b>	-214	27	70
oi-VAE ( $K = 8$ )	-3,196	-195	29	75
oi-VAE ( $K = 16$ )	-3,550	<b>-188</b>	<b>31</b>	108

BRISK WALK				
# TRIALS	1	2	5	10
VAE ( $K = 16$ )	-723,795	-15,413,445	-19,302,644	-19,303,072
oi-VAE ( $K = 4$ )	-664,608	-13,438,602	<b>-19,289,548</b>	-19,302,680
oi-VAE ( $K = 8$ )	-283,352	-10,305,693	-19,356,218	-19,302,764
oi-VAE ( $K = 16$ )	<b>-198,663</b>	<b>-6,781,047</b>	-19,299,964	-19,302,924

MEG sensor data into *source-space* where we obtain observations over time on a high-resolution mesh ( $\approx 5$ -10K vertices) of the cortical surface (Gramfort et al., 2013). The resulting source-space signals likely live on a low-dimensional manifold making methods such as the VAE attractive. However, neuroscientists have meticulously studied particular brain regions of interest and what behaviors they are involved in, so that a key problem is inferring groups of interrelated regions.

We apply our oi-VAE method to infer low-rank representations of source-space MEG data where the groups are specified as the  $\approx 40$  regions defined in the HCP-MMP1 brain parcellation (Glasser et al., 2016). See Fig. 6(left). The recordings were collected from a single subject performing an auditory attention task where they were asked to maintain their attention to one of two auditory streams. We

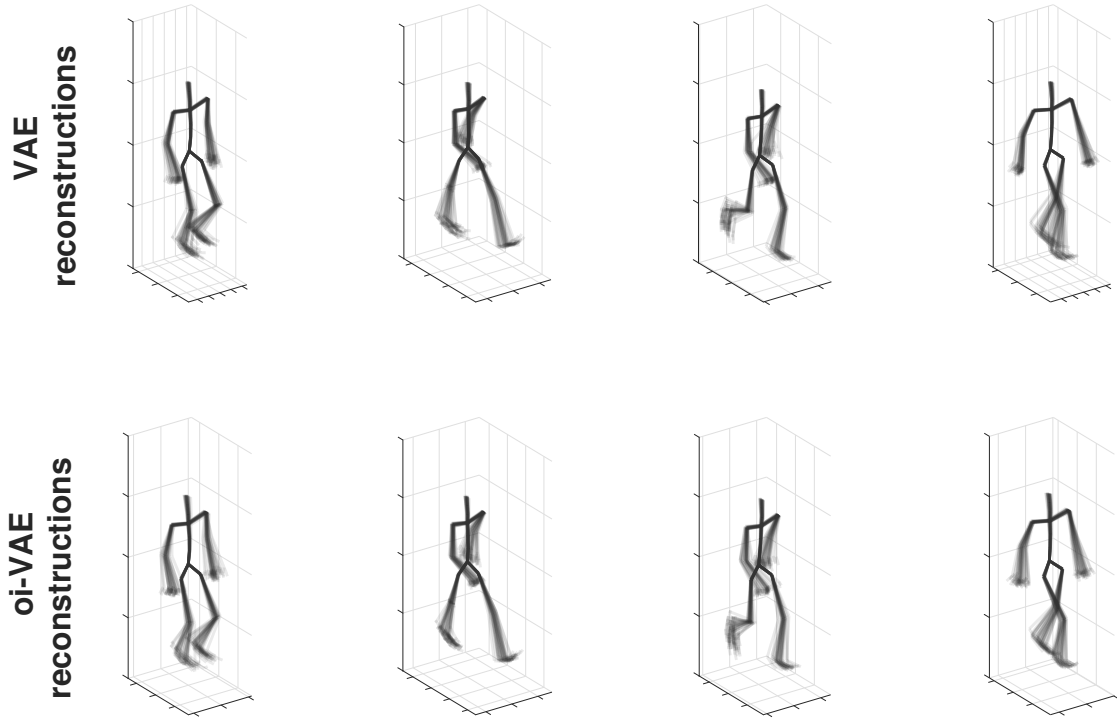


Figure 4: Samples an oi-VAE model trained on walking data and conditioned on an out-of-sample video frame. We can see that oi-VAE has learned noise patterns that reflect the gait as opposed to arbitrary perturbative noise.

use 106 trials each of length 385. We treat each time point of each trial as an i.i.d. observation resulting in  $\approx 41\text{K}$  observations. For details on the specification of all considered models, see the Supplement.

For each region we compute the average sensor-space activity over all vertices in the region resulting in 44-dimensional observations. We applied oi-VAE with  $K = 20$ ,  $\lambda = 1$ , and Alg. 1 for 10,000 iterations. In Fig. 6 we depict the learned group-weights  $\|\mathbf{W}_{:,j}^{(g)}\|_2$  for all groups  $g$  and components  $j$ . We observe that each component manifests itself in a sparse subset of the regions. Next, we dig into specific latent components and evaluate whether each influences a subset of regions in a neuroscientifically interpretable manner.

For a given latent component  $\mathbf{z}_j$ , the value  $\|\mathbf{W}_{:,j}^{(g)}\|_2$  allows us to interpret how much component  $j$  influences region  $g$ . We visualize some of these weights for two prominent learned components in Fig. 7. Specifically, we find that component 6 captures the regions that make up the *dorsal attention network* pertaining to an auditory spatial task, viz., early visual, auditory sensory areas as well as inferior parietal sulcus and the region covering the right temporoparietal junction (Lee et al., 2014). We also find that component 15 corresponds to regions associated with the *default mode network*, viz., medial prefrontal as well as posterior cingulate cortex (Buckner et al., 2008). Again the oi-VAE leads to interpretable results that align with meaningful and previously studied physiological systems. These systems can be further probed through functional connectivity analysis. See the Supplement for the analysis of more components.

## 7 Conclusion

We proposed an output interpretable VAE (oi-VAE) that can be viewed as either a nonlinear group latent factor model or as a structured VAE with disentangled latent embeddings. The approach

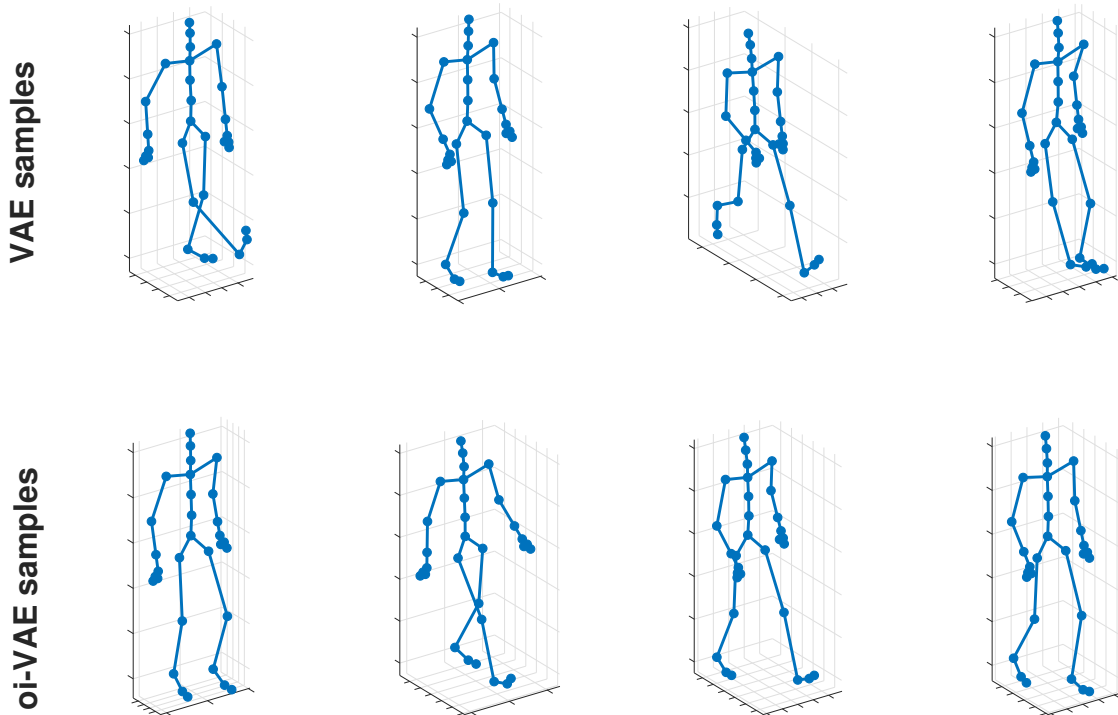


Figure 5: Representative unconditional samples from oi-VAE and VAE trained on `walk` trials. oi-VAE generates physically realistic walking poses while VAE sometimes produces implausible ones.

combines deep generative models with a hierarchical sparsity-inducing prior that leads to our ability to extract meaningful notions of latent-to-observed interactions when the observations are structured into groups. From this interaction structure, **we can infer correlated systems of interaction amongst the observational groups**. In our motion capture and MEG experiments we demonstrated that the resulting systems are physically meaningful. Importantly, this interpretability does not appear to come at the cost of expressivity, and in our group-structured case can actually lead to improved generalization and generative processes.

In contrast to alternative approaches one might consider for nonlinear group sparse factor analysis, leveraging the amortized inference associated with VAEs leads to computational efficiencies. We see even more significant gains through our proposed collapsed objective. The proximal updates we can apply lead to real learned sparsity.

We note that nothing fundamentally prevents applying this architecture to other generative models *du jour*. Extending this work to GANs, for example, should be straightforward. Furthermore, one could consider combining this framework with sparsity inducing priors on  $\mathbf{z}$  to discourage redundant latent dimensions. Oy-vey!

## Acknowledgements

This work was supported by ONR Grant N00014-15-1-2380, NSF CAREER Award IIS-1350133, NSF CRCNS Grant NSF-IIS-1607468, and AFOSR Grant FA9550-1-1-0038. The authors also gratefully acknowledge the support of NVIDIA Corporation for the donated GPU used for this research.



Figure 6: (Left) The regions making up the HCP-MMP1 parcellation defining the groups. (Right) Latent-to-group mappings indicate that each latent component influences a sparse set of regions.

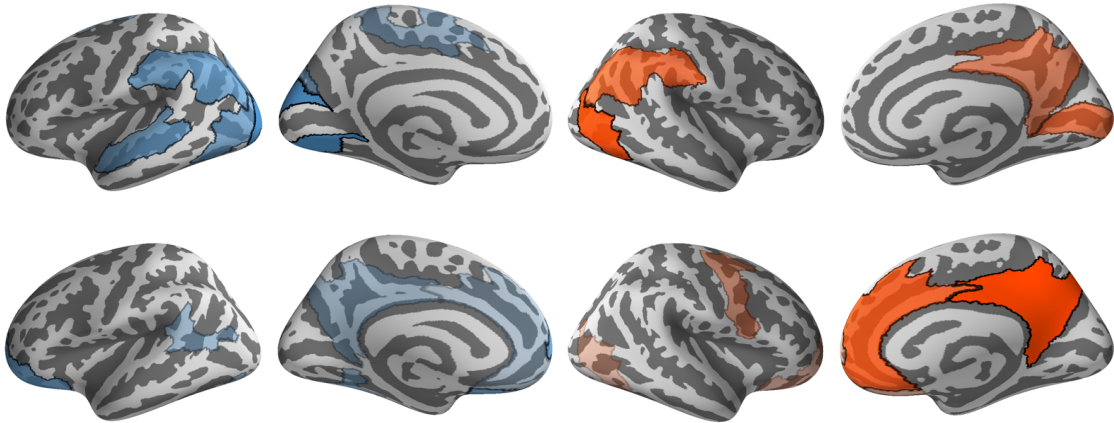


Figure 7: Influence of  $\mathbf{z}_6$  (top) and  $\mathbf{z}_{15}$  (bottom) on the HCP-MMP1 regions. Active regions (shaded) correspond to the *dorsal attention network* and *default mode network*, respectively.

## References

- Archer, E., Park, I. M., Buesing, L., Cunningham, J., and Paninski, L. Black box variational inference for state space models. *CoRR*, abs/1511.07367, 2015.
- Bhadra, A., Datta, J., Polson, N. G., and Willard, B. Default Bayesian analysis with global-local shrinkage priors. *Biometrika*, 103(4):955–969, 2016.
- Buckner, R. L., Andrews-Hanna, J. R., and Schacter, D. L. The brain’s default network: Anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, 1124(1):1–38, 2008.
- Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. High-dimensional sparse factor modeling: Applications in gene expression genomics. *J. Amer. Statist. Assoc.*, 103(434):73–80, 2008.
- Damianou, A. C., Ek, C. H., Titsias, M. K., and Lawrence, N. D. Manifold relevance determination. In *International Conference on Machine Learning*, 2012.
- Danaher, P., Wang, P., and Witten, D. M. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 76(2):373–397, 3 2014.
- Ghosh, S. and Doshi-Velez, F. Model selection in Bayesian neural networks via horseshoe priors. *CoRR*, abs/1705.10388, 2017.

- Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., and Van Essen, D. C. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.
- Gramfort, A., Luessi, M., Larson, E., Engemann, D., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., and Hmlinen, M. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7:267, 2013.
- Han, K., Wen, H., Shi, J., Lu, K.-H., Zhang, Y., and Liu, Z. Variational autoencoder: An unsupervised model for modeling and decoding fmri activity in visual cortex. *bioRxiv*, 2017.
- Harvey, A., Ruiz, E., and Shephard, N. Multivariate stochastic variance models. *Review of Economic Studies*, 61(2):247–264, 1994.
- Hirose, K. and Konishi, S. Variable selection via the weighted group lasso for factor analysis models. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 40(2):345–361, 2012.
- Johnson, G. R., Donovan-Maiye, R. M., and Maleckar, M. M. Building a 3d integrated cell. *bioRxiv*, content/early/2017/12/21/238378, 2017.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.
- Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. *CoRR*, abs/1312.6114, 2013.
- Kyung, M., Gill, J., Ghosh, M., and Casella, G. Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5(2):369–411, 06 2010.
- Lawrence, N. Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in Neural Information Processing Systems*, 2003.
- Lee, A. K. C., Larson, E., Maddox, R. K., and Shinn-Cunningham, B. G. Using neuroimaging to understand the cortical mechanisms of auditory selective attention. *Hearing Research*, 307:111–120, 2014.
- Louizos, C., Ullrich, K., and Welling, M. Bayesian compression for deep learning. *CoRR*, abs/1705.08665, 2017.
- Mitchell, T. J. and Beauchamp, J. J. Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.*, 83:1023–1036, 1988.
- Parikh, N. and Boyd, S. Proximal algorithms, 2013.
- Prill, R. J., Marbach, D., Saez-Rodriguez, J., Sorger, P. K., Alexopoulos, L. G., Xue, X., Clarke, N. D., Altan-Bonnet, G., and Stolovitzky, G. Towards a rigorous assessment of systems biology models: the dream3 challenges. *PLoS one*, 5(2):e9202, 2010.
- Reddi, S. J., Sra, S., Póczos, B., and Smola, A. J. Proximal Stochastic Methods for Nonsmooth Nonconvex Finite-Sum Optimization. In *Advances in Neural Information Processing Systems*, pp. 1145–1153. Curran Associates, Inc., 2016.
- Rezende, Danilo Jimenez, Mohamed, Shakir, and Wierstra, Daan. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32, ICML’14*, pp. II–1278–II–1286. JMLR.org, 2014. URL <http://dl.acm.org/citation.cfm?id=3044805.3045035>.
- Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

- Uusitalo, M. A. and Ilmoniemi, R. J. Signal-space projection method for separating meg or eeg into components. *Med Biol Eng Comput*, 35(2):135–140, 1997.
- Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.
- Zhao, S., Gao, C., Mukherjee, S., and Engelhardt, B. E. Bayesian group factor analysis with structured sparsity. *Journal of Machine Learning Research*, 17, 2016.



## A 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  $8 \times 8$  images but use  $K = 16$ . See Figure 8. Train and test log likelihoods for the model are given in Table 3.

**Experimental details** For all our synthetic data experiments we sampled 2,048  $8 \times 8$  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 3: Train and test log likelihoods on the synthetic bars data when  $K$  is larger than necessary.

MODEL	TRAIN LOG LIKELIHOOD	TEST LOG LIKELIHOOD
$\lambda = 1$	<b>99.9325</b>	<b>100.1394</b>
$\lambda = 0$ and no $\theta$ prior	95.0687	95.4285

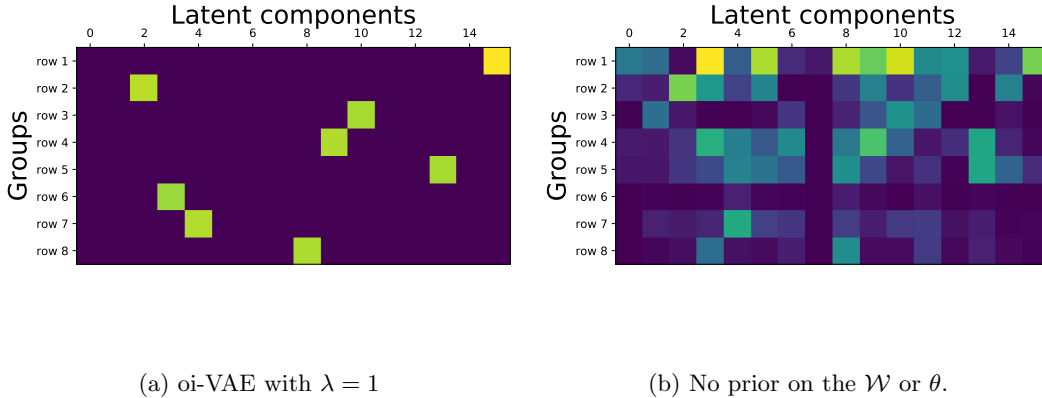


Figure 8: Results with latent dimension  $K = 16$  when the effective dimensionality of the data is only 8. Clearly the oi-VAE has learned to use only the sparse set of  $z_i$ 's that are necessary to explain the data.

## B Motion capture results

Multiple samples from both the VAE and oi-VAE are shown in Figure 9.

**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.

## C MEG Analysis

We present the three most prominent components determined by summing  $\|\mathbf{W}_{:,j}^{(g)}\|_2$  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 (Uusitalo & Ilmoniemi, 1997).

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.

## D 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  $\mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} [\log p(x|\mathbf{z}, \mathcal{W}, \theta)]$  with one sample. We experimented with using more samples but did not observe any significant benefit from doing so.

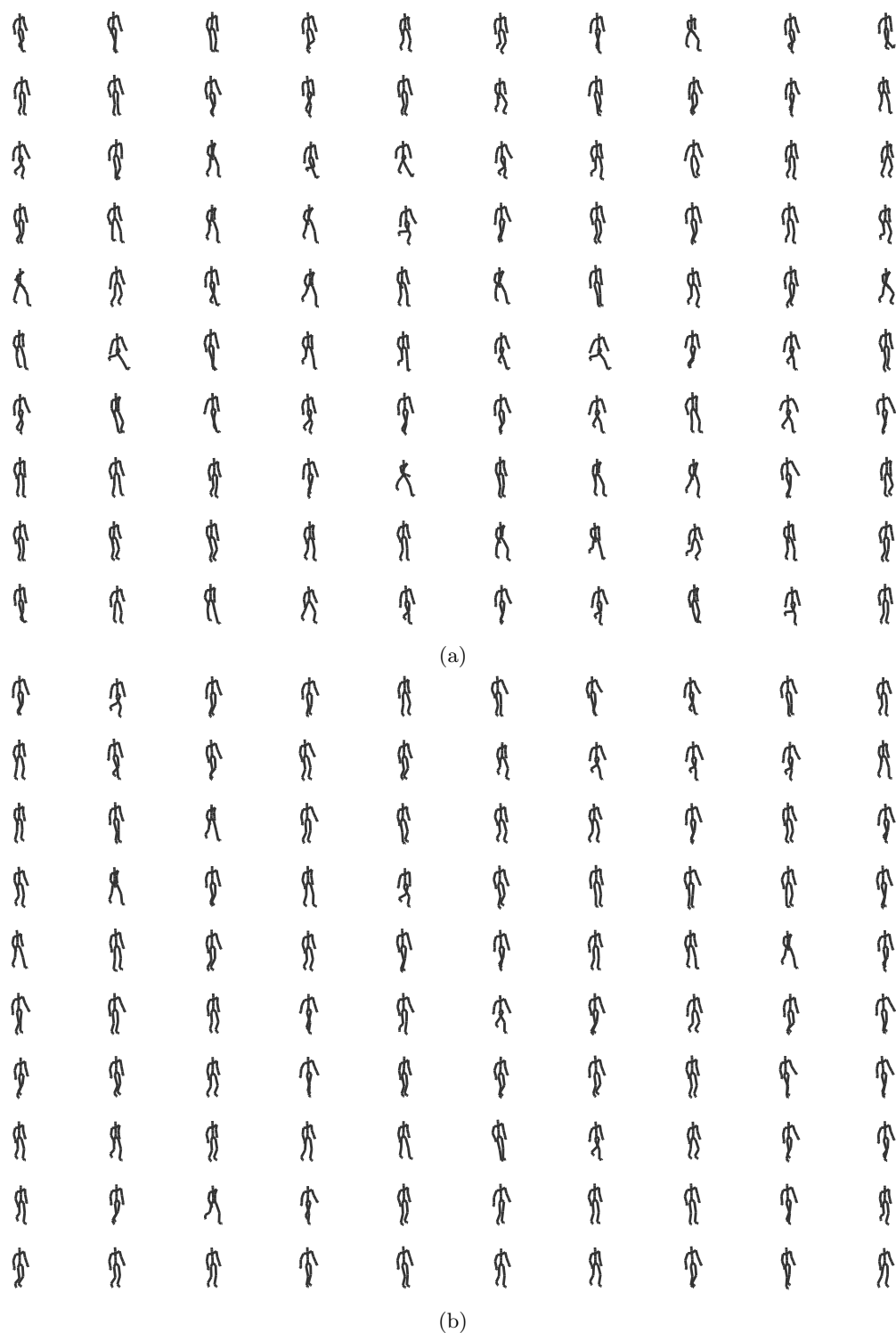
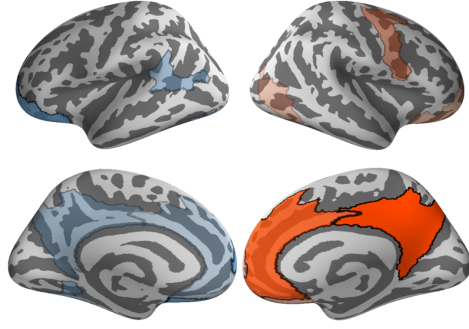
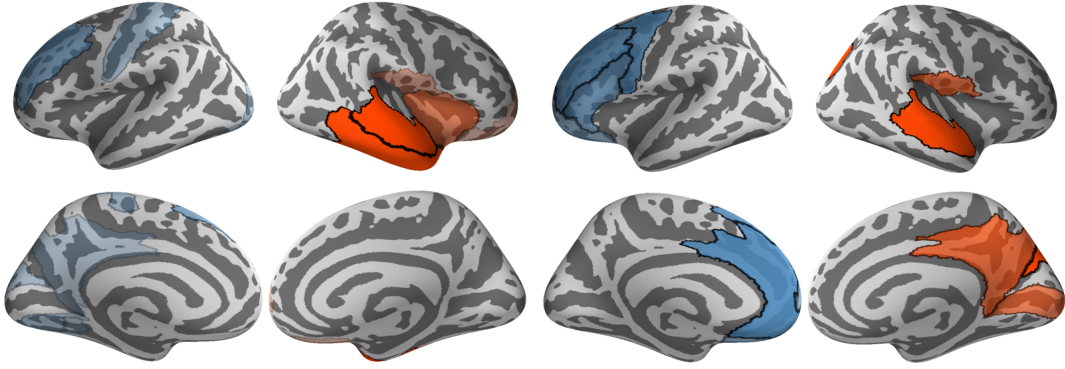


Figure 9: Samples from the (a) VAE and (b) oi-VAE models. The VAE produces a number of poses apparently inspired by the Ministry of Silly Walks. Some others are even physically impossible. In contrast, results from the oi-VAE are all physically plausible and appear to be representative of walking. Full scale images will be made available on the author’s website.



(a) Component 15: Default Mode Network

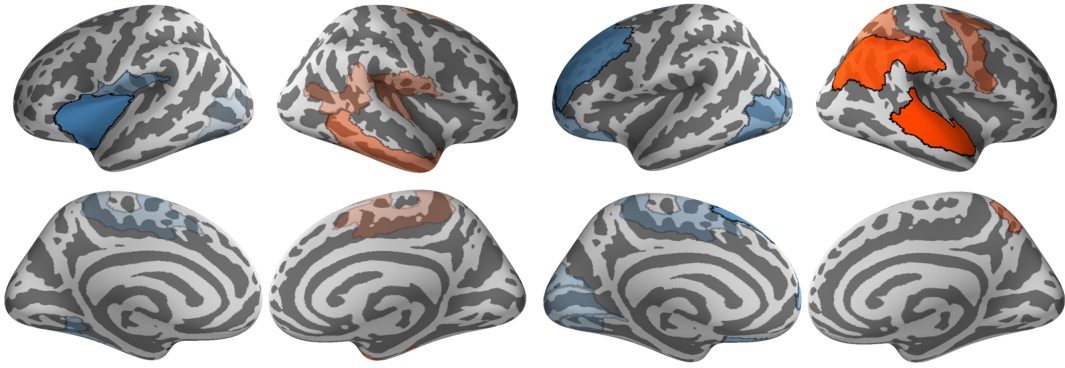
Figure 10: The projections of two components of  $\mathbf{z}$  onto the regions of the HCP-MMP1 parcellation. The regions with the ten largest weights are shaded (blue in the left hemisphere, red in the right hemisphere) with opacity indicating the strength of the weight. Component 15 corresponds to the default mode network.



(a) Component 2.

(b) Component 5.

Figure 11: Component 2 has the largest aggregate group weight and component 5 has the second largest.



(a) Component 11

(b) Component 7

Figure 12: 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.