# oi-VAE: Output Interpretable VAEs for Nonlinear Group Factor Analysis

Samuel K. Ainsworth [1] Nicholas J. Foti [1] Adrian K. C. Lee [2] Emily B. Fox [1]

## 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

Many datasets of interest in machine learning are comprised of high-dimensional, complex objects. Often, one is interested in describing these observations using a low-dimensional latent subspace that captures the statistical variations. Such approaches fall under the umbrella of factor analysis (Bishop, 2016), where we wish to learn a mapping between the latent and observed spaces. The motivation is two-fold: (i) factor models provide a compact representation of the data, and (ii) the mapping can be used to describe the correlation structure of the high-dimensional data. In many applications, we are particularly interested in mappings that elucidate interpretable interactions.

The challenge arises from the push and pull between interpretability and expressivity in factor modeling approaches. Methods emphasizing 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 (Vejmelka et al., 2010), genomic (Prill et al., 2010), and financial data (Harvey et al., 1994), for example, exhibit complex nonlinearities.

Recently, there has been a significant amount of work on expressive models for complex, high dimensional data. In particular, *deep generative models* (Kingma & Welling, 2013; Rezende et al., 2014; Goodfellow et al., 2014; Damianou & Lawrence, 2013) 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 that parameterize an observation distribution, often referred to as the *generator*. We focus on the class of *variational autoencoders* (VAEs) (Kingma & Welling, 2013). Unlike linear models which posit a latent variable per observation, VAEs introduce a mapping from observations to a distribution on the latent space; when parameterized by a deep neural network, this mapping is called the *inference network*. The generator and inference network are jointly trained to minimize a variational objective.

The VAE can be viewed as a nonlinear factor model that provides a scalable means of learning latent representations. The focus, however, has primarily been on their use as a generative mechanism. One shortcoming of the VAE is that, due to the tangled web of connections between neural network layers, it is not possible to interpret how changes in the latent code influence changes in the observations—as in linear latent factor models. For example, imagine you are trying to synthesize human body poses. One might hope to have a disentangled representation where a given latent dimension controls a subset of highly correlated body parts; unfortunately, the standard VAE cannot yield these types of interpretations. Another shortcoming of the VAE

[1] Paul G. Allen School of Computer Science and Engineering, University of Washington [2] Institute for Learning & Brain Sciences and Department of Speech and Hearing Sciences, University of Washington. Correspondence to: Samuel K. Ainsworth <skainsworth@gmail.com>, Nicholas J. Foti <nfoti@uw.edu>, Adrian K. C. Lee <akclee@uw.edu>, Emily B. Fox <ebfox@uw.edu>.

*Proceedings of the $35^{th}$ International Conference on Machine Learning*, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

is that training—as in most neural network-based models—typically requires a massive amount of data. In many applications, we have limited access to training data.

One natural way to encourage disentangled latent representations is by introducing structure and sparsity into the generator. Specifically, we propose an *output interpretable VAE* (oi-VAE) that factorizes the generator across observation dimensions, with a separate generator per group of variables. The generators are coupled both through a shared latent space, and by jointly training with a single inference network. We also introduce a sparsity-inducing penalty that leads each latent dimension to influence a limited subset of groups, resulting in a disentangled latent representation. We develop an amortized variational inference algorithm for a collapsed objective, allowing us to use efficient proximal updates to learn latent-dimension-to-group interactions.

The factorization of generators across dimensions is readily apparent when the data are inherently group structured. There are many applications where this is the case. 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, . . . ). And for motion capture data, multiple angle measurements are grouped by their associated joints. In these group-structured scenarios we may additionally garner *interpretability* from the oi-VAE mappings. For example, we may learn that a given latent dimension controls a collection of highly correlated joints—e.g., joints in a limb—that comprise a system of interest. A side benefit of this structured oi-VAE framework is its ability to handle scenarios with limited amounts of data.

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. In addition, we found that oi-VAE produces unconditional samples that are qualitatively superior to standard VAEs due to oi-VAE's bias towards disentangled representations in the latent space.

## 2. Background

Nonlinear factor analysis aims to relax the strict linearity assumption of classical factor analysis and has a long history in the statistics community. The work of (Gibson, 1959) initially circumvented the issues of linear factor analysis by discretizing continuous nonlinearities. However, (McDonald, 1962) was the first to develop a parametric nonlinear factor analysis model. Significant progress has been made since then as described in Yalcin & Amemiya (2001), including developments in the Bayesian context (Arminger & Muthén, 1998). Recent work in machine learning has also considered similar approaches leveraging Gaussian processes (Lawrence, 2003; Damianou et al., 2012). Despite the resemblance to autoencoding models (Ballard, 1987)—especially in the age of "disentanglement"—little work exists exploring connections between the two.

The study of deep generative models is an active area of research in the machine learning community. The variational autoencoder (VAE) (Kingma & Welling, 2013) is one such example that efficiently trains a generative model via amortized inference (see also Rezende et al., 2014). Though deep generative models like the VAE have demonstrated an ability to produce convincing samples of complex data (cf., Archer et al., 2015; Johnson et al., 2017), the learned latent representations are not readily interpretable due to the entangled interactions between latent dimensions and the observations, as depicted in Fig. 2. We further review the VAE specification in Sec. 3 and its implementation in Sec. 5.

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.), and linear factor analysis (Hirose & Konishi, 2012). The group lasso penalty is of particular interest to us as it simultaneously shrinks entire groups of model parameters toward zero. The usage of group lasso penalties for learning structured inputs to neural networks was explored in Tank et al. (2018) previously, and was inspirational to this work.

To specify a valid generative model, we focus on sparsity-inducing *priors* for the parameters of the generator network. Historically, the spike-and-slab prior (Mitchell & Beauchamp, 1988) was used to encourage sparsity in Bayesian models. The prior consists of a two-component mixture with mass on a model parameter being exactly zero. Unfortunately, inference in spike-and-slab models is difficult because of the combinatorial nature of the resulting posterior. A more computationally tractable family arises from the class of *global-local shrinkage priors* (Polson & Scott, 2010). One popular example is the horseshoe prior (Bhadra et al., 2016). However, these priors do not result in exact zeros, making interpretability difficult.

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

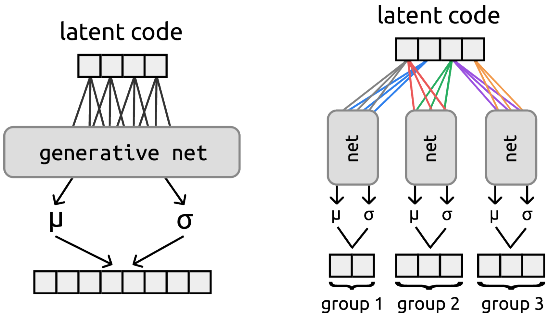


*Figure 1.* VAE (*left*) and oi-VAE (*right*) generative models. The oi-VAE considers group-specific generators and a linear latent-to-generator mapping with weights from a single latent dimension to a specific group sharing the same color. The group-sparse prior is applied over these grouped weights in order to promote a disentangled latent representation in which a particular latent component only interacts with a sparse subset of groups.

be sparse. The resulting model admits an efficient EM algorithm. This builds on previous work on group factor analysis (Virtanen et al., 2012; Klami et al., 2015). Sparsity inducing hierarchical Bayesian priors have also been applied to learn the complexity of the Bayesian deep neural networks (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 latent representation of fixed dimension $K$. We then write the generative process of the model as:

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

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

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. (2) is simpler than that described in Kingma & Welling (2013) where the noise variances were observation specific. This simplifying assumption is common with traditional factor models, but could easily be relaxed.

When our observations $\mathbf{x}$ admit a natural grouping over the components, we write $\mathbf{x} = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(G)}]$ for each of the $G$ groups. We model the components within each group $g \in [G]$ 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 chose to model each separately. Critically, the latent representation $\mathbf{z}$ is shared across all of the group-specific generators. In particular:

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

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

To this point, our specified group-structured VAE can describe within-group and cross-group correlation structure. However, one of the primary goals of this framework is to capture interpretable relationships between groups through the latent representation.

Inspired by the sparse factor analysis literature, we extract notions of interpretable interactions by encouraging 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 $f^{(g)}$:

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

We refer to $\mathbf{W}^{(g)}$ as the latent-to-group matrix. For simplicity, we assume that each generator has input dimension $p$. When the $j$th column of the latent-to-group matrix for group $g$, $\mathbf{W}_{\cdot,j}^{(g)}$, is all zeros then the $j$th latent dimension, $\mathbf{z}_j$, will have no influence on group $g$ in the generative process. To induce this column-wise sparsity, we place a hierarchical Bayesian prior on the columns $\mathbf{W}_{\cdot,j}^{(g)}$ as follows (Kyung et al., 2010):

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

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

where Gamma$(\cdot, \cdot)$ is defined by shape and rate. 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).

Unlike linear factor models, the deep structure of our model allows rescaling of the parameters across layer boundaries without affecting the end behavior of the network (Neyshabur et al., 2015). In particular, it is possible—and in fact encouraged behavior—to learn a set of $\mathbf{W}^{(g)}$ matrices with very small weights and a subsequent layer with very large weights that nullify the shrinkage imposed by the sparsity-inducing prior. In order to mitigate this we additionally place a standard normal prior with fixed scale on the parameters of each generative network, $\theta_g \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

**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. In particular, by shrinking columns of $\mathbf{W}$ (dropping the $g$ superscript) 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 still be standard normal, but the columns of $\mathbf{W}$ will dictate which components are used. This regularization may be advantageous even in the classical, single-group setting 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 classical group sparse linear factor model.

## 4. Interpretability of the oi-VAE

In 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** Disregarding any interest in the learned 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 often our goal is to uncover functionally-connected brain networks. In this setting we may split the signal into groups based on a standard parcellation. Then networks can be identified by inspecting the subset of groups influenced by a component in the latent code, $z_i$. 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. Still, we find that oi-VAE does prune dimensions that are not necessary in synthetic examples.

## 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 by a 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 in order to 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 $\gamma_{gj}^2$ and $\mathbf{W}_{\cdot,j}^{(g)}$ will not result in true sparsity of the columns $\mathbf{W}_{\cdot,j}^{(g)}$. Instead, we consider a collapsed variational objective function. Since our sparsity inducing prior over $\mathbf{W}_{\cdot,j}^{(g)}$ 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 $\mathcal{W} = \left(\mathbf{W}^{(1)}, \cdots, \mathbf{W}^{(G)}\right)$ and $\theta = (\theta_1, \ldots, \theta_G)$. 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 $\phi$.

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

```
Input: data $\mathbf{x}^{(i)}$, sparsity parameter $\lambda$
Let $\tilde{\mathcal{L}} = \mathcal{L}(\phi, \theta, \mathcal{W}) + \lambda \sum_{g,j} ||\mathbf{W}^{(g)}_{\cdot,j}||_2$.
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 $g$ and $j = 1$ to $K$ do
    Set $\mathbf{W}^{(g)}_{\cdot,j} \leftarrow \frac{\mathbf{W}^{(g)}_{\cdot,j}}{||\mathbf{W}^{(g)}_{\cdot,j}||_2} \left(||\mathbf{W}^{(g)}_{\cdot,j}||_2 - \eta\lambda\right)_+$
  end for
until convergence in both $\hat{\mathcal{L}}$ and $-\lambda \sum_{g,j} ||\mathbf{W}^{(g)}_{\cdot,j}||_2$
```

### 5.1. The collapsed objective

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

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

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

Now our goal becomes maximizing this collapsed ELBO over $\phi, \theta$, and $\mathcal{W}$. 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 some variants are guaranteed to converge to a first-order 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 $h(x)$:

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

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 see 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

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

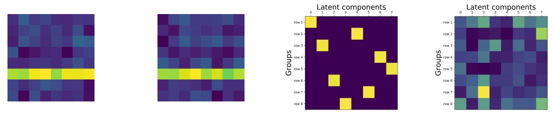
where $(v)_+ \triangleq \max(0, v)$ (Parikh & Boyd, 2013). Geometrically, this operator reduces the norm of $x$ by $\eta\lambda$, and shrinks $x$'s with $||x||_2 \leq \eta\lambda$ to zero. This operator is especially convenient since it is both cheap to compute and results in machine-precision zeros, unlike many Bayesian approaches to sparsity that result in small but non-zero values and thus require an extra thresholding step to attain exact zeros.

We experimented with standard (non-collapsed) variational inference as well 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 complete pseudocode.

## 6. Experiments

### 6.1. Synthetic data

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 model 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



(a) (b) (c) (d)

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

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 model 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 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 `walking` trials, 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 $\lambda = 1$. 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, 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.

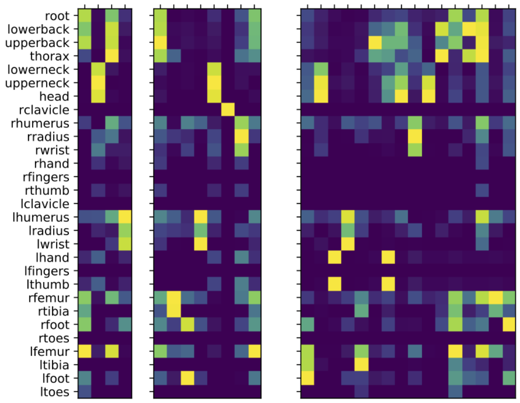


*Figure 3.* oi-VAE results on motion capture data with $K = 4$, 8, and 16 latent dimensions. Rows correspond to group generators for each of the joints in the skeleton, columns correspond to individual dimensions of the latent code, and values in the heatmap show the strength of the latent-to-group mappings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

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. 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 variation of what was trained on, whereas the former is a heldout example, very similar to the training data. In Table 2, we see the benefit of the regularization in oi-VAE in both test scenarios in the limited data regime. Unsurprisingly, 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

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

when we have both a limited amount of training 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. We leave choosing $K$ and investigating the effects of pruning to future work.

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 approximate posteriors. For every such $q_\phi(\mathbf{z}|\mathbf{x})$, we sample 32 times from the distribution and run each sample through the generator networks to synthesize a batch of reconstructions. To fully explore the space of human motion the learned generators can capture, we take 100 *unconditional* samples from both the oi-VAE and VAE models and show a representative subset in Fig. 5. The full set of 100 random samples from both oi-VAE and 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 test log-likelihood does not encourage quality unconditional samples, but a disentangled latent representation should yield qualitatively better results on samples from the prior.

### 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 MEG sensor data into *source-space* where we obtain observations over time on a mesh (≈ 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. Still, neuroscientists have meticulously studied particular brain regions of interest and what behaviors they are involved in by hand.

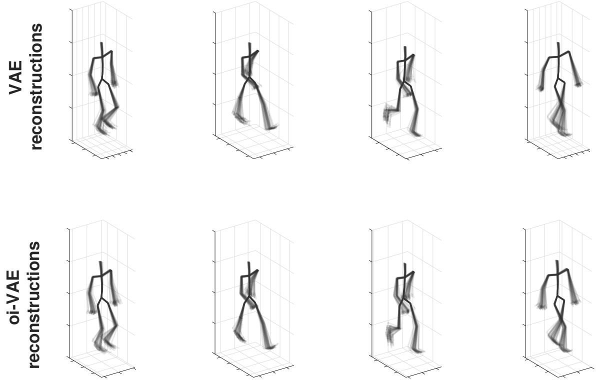


*Figure 4.* Samples an oi-VAE model trained on walking data and conditioned on an out-of-sample video frame. We can see that oi-VAE has learned noise patterns that reflect the natural gait.

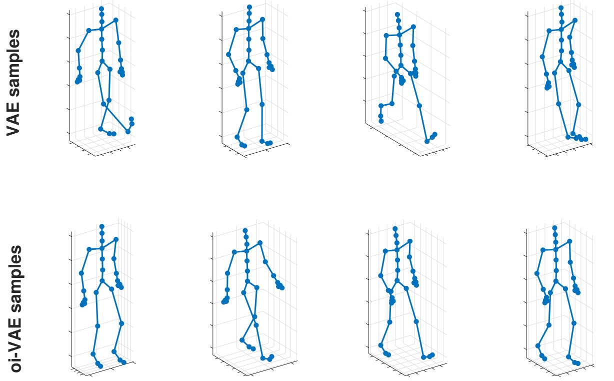


*Figure 5.* Representative unconditional samples from oi-VAE and VAE trained on `walk` trials. oi-VAE generates physically realistic walking poses while VAE sometimes produces implausible ones.

We apply our oi-VAE method to infer low-rank representations of source-space MEG data where the groups are specified as the ≈ 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 use 106 trials each of length 385. We treat each time point

*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. The standard VAE uses the same structure as oi-VAE for a consistent comparison (equivalent to $\lambda = 0$).

| | STANDARD WALK | | | | BRISK WALK | | | |
|---|---|---|---|---|---|---|---|---|
| # TRIALS | 1 | 2 | 5 | 10 | 1 | 2 | 5 | 10 |
| VAE ($K = 16$) | $-3,518$ | $-251$ | 18 | **114** | $-723,795$ | $-15,413,445$ | $-19,302,644$ | $-19,303,072$ |
| OI-VAE ($K = 4$) | $\mathbf{-2,722}$ | $-214$ | 27 | 70 | $-664,608$ | $-13,438,602$ | $\mathbf{-19,289,548}$ | $-19,302,680$ |
| OI-VAE ($K = 8$) | $-3,196$ | $-195$ | 29 | 75 | $-283,352$ | $-10,305,693$ | $-19,356,218$ | $-19,302,764$ |
| OI-VAE ($K = 16$) | $-3,550$ | $\mathbf{-188}$ | **31** | 108 | $\mathbf{-198,663}$ | $\mathbf{-6,781,047}$ | $-19,299,964$ | $-19,302,924$ |

of each trial as an i.i.d. observation resulting in $\approx 41$K observations. For details on the specification of all considered models, see the Supplement.

For each region we compute the average source-space activity over all vertices in each region resulting in 44-dimensional observations. We applied oi-VAE with $K = 20$, $\lambda = 10$, and Alg. 1 for $10,000$ iterations. In Fig. 6 we depict the learned group-weights $||\mathbf{W}_{\cdot,j}^{(g)}||_2$ 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.

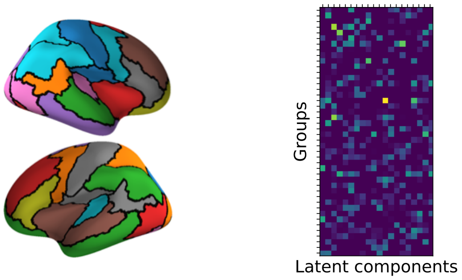


*Figure 6.* (Left) The regions making up the HCP-MMP1 parcellation defining the groups. (Right) Latent-to-group mappings indicate that each latent component influences a sparse set of regions.

For a given latent component $\mathbf{z}_j$, the value $||\mathbf{W}_{\cdot,j}^{(g)}||_2$ 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.

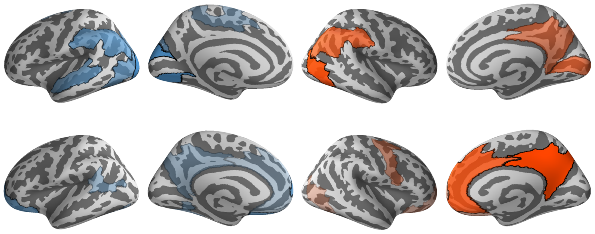


*Figure 7.* Influence of $\mathbf{z}_6$ (top) and $\mathbf{z}_{15}$ (bottom) on the HCP-MMP1 regions. Active regions (shaded) correspond to the *dorsal attention network* and *default mode network*, respectively.

## 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 combines deep generative models with a 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 expressiveness, and in our group-structured case can actually lead to improved generalization and generative processes.

In contrast to alternative approaches 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 quickly to true sparsity.

Note that nothing fundamentally prevents applying this architecture to other generative models *du jour*. Extending this work to generative adversarial models, for example, should be straightforward (Goodfellow et al., 2014). Oy-vey!

## Acknowledgements

Special thanks to Ian Covert and Alex Tank for helpful discussions and insight. 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 FA9550-16-1-0038. The authors also gratefully acknowledge the support of NVIDIA Corporation for the donated GPU used for this research.

## 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.

Arminger, G. and Muthén, B. O. A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. *Psychometrika*, 63(3):271–300, 1998.

Ballard, D. H. Modular Learning in Neural Networks. 1987.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. Default Bayesian analysis with global-local shrinkage priors. *Biometrika*, 103(4):955–969, 2016.

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 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. and Lawrence, N. Deep Gaussian Processes. In *Artificial Intelligence and Statistics*, pp. 207–215, 2013.

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.

Ghosh, S. and Doshi-Velez, F. Model selection in Bayesian neural networks via horseshoe priors. *CoRR*, abs/1705.10388, 2017.

Gibson, W. A. Three multivariate models: Factor analysis, latent structure analysis, and latent profile analysis. *Psychometrika*, 24(3):229–252, 1959.

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.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative Adversarial Nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

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.

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.

Klami, A., Virtanen, S., Leppäaho, E., and Kaski, S. Group factor analysis. *IEEE transactions on neural networks and learning systems*, 26(9):2136–2147, 2015.

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.

McDonald, R. P. A general approach to nonlinear factor analysis. *Psychometrika*, 27(4):397–415, 1962.

Mitchell, T. J. and Beauchamp, J. J. Bayesian Variable Selection in Linear Regression. *J. Amer. Statist. Assoc.*, 83:1023–1036, 1988.

Neyshabur, B., Salakhutdinov, R. R., and Srebro, N. Path-SGD: Path-Normalized Optimization in Deep Neural Networks. In *Advances in Neural Information Processing Systems*, pp. 2422–2430, 2015.

Parikh, N. and Boyd, S. *Proximal Algorithms*, volume 1 of *Foundations and Trends in Optimization*. 2013.

Polson, N. G. and Scott, J. G. Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction. *Bayesian statistics*, 9:501–538, 2010.

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., Poczos, B., and Smola, A. J. Proximal Stochastic Methods for Nonsmooth Nonconvex Finite-Sum Optimization. In *Advances in Neural Information Processing Systems*, pp. 1145–1153. 2016.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pp. 1278–1286, 2014.

Tank, A., Covert, I., Foti, N., Shojaie, A., and Fox, E. Neural Granger Causality for Nonlinear Time Series. *arXiv preprint arXiv:1802.05842*, 2018.

Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

Vejmelka, M., Paluš, M., and Šušmáková, K. Identification of nonlinear oscillatory activity embedded in broadband neural signals. *International journal of neural systems*, 20(02):117–128, 2010.

Virtanen, S., Klami, A., Khan, S., and Kaski, S. Bayesian Group Factor Analysis. In *Artificial Intelligence and Statistics*, pp. 1269–1277, 2012.

Yalcin, I. and Amemiya, Y. Nonlinear Factor Analysis as a Statistical Method. *Statistical science*, pp. 275–294, 2001.

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.