Unsupervised Learning with Contrastive Latent Variable Models

Kristen A. Severson, Soumya Ghosh, Kenney Ng
Center for Computational Health and MIT-IBM Watson AI Lab,
IBM Research, 75 Binney St. Cambridge, Massachusetts, 02142

Abstract

In unsupervised learning, dimensionality reduction is an important tool for data exploration and visualization. Because these aims are typically open-ended, it can be useful to frame the problem as looking for patterns that are enriched in one dataset relative to another. These pairs of datasets occur commonly, for instance a population of interest vs. control or signal vs. signal free recordings. However, there are few methods that work on *sets* of data as opposed to data points or sequences. Here, we present a probabilistic model for dimensionality reduction to discover signal that is enriched in the *target* dataset relative to the *background* dataset. The data in these sets do not need to be paired or grouped beyond set membership. By using a probabilistic model where some structure is shared amongst the two datasets and some is unique to the target dataset, we are able to recover interesting structure in the latent space of the target dataset. The method also has the advantages of a probabilistic model, namely that it allows for the incorporation of prior information, handles missing data, and can be generalized to different distributional assumptions. We describe several possible variations of the model and demonstrate the application of the technique to de-noising, feature selection, and subgroup discovery settings.

Introduction

In unsupervised learning, the goal is often to learn what is unique or interesting about a dataset. Given the subjective nature of this question, it can be useful to frame the problem in the context of what signal is enriched in one dataset, referred to as the *target*, relative to a second dataset, referred to as the *background*. An example of this is an exploration of a heterogeneous disease population, such as patients with Parkinson's disease. The interesting sources of variation are those that are unique to the disease population. However, it is likely that some sources of variation are unrelated to the disease state, for instance variation due to aging. This is difficult to assess without a baseline population, therefore, it is useful to contrast the disease population with a population of healthy controls. Such *contrastive analysis* can discover nuisance variation that is common amongst the two populations and is uninteresting for the problem while highlighting variation unique to the disease population enabling downstream applications such as subgroup discovery.

Despite this natural setting for unsupervised learning, most techniques address individual data points, sequences, or paired data points. Few techniques generalize to the contrastive scenario where we have sets of data but no obvious correspondence between their members. Yet, there are many cases where datasets that can be used in a comparative setting arise naturally: control vs. study populations, pre- and post-intervention groups, and signal vs. signal free groups (Abid et al. 2018). Each of these settings has possible nuisance variation, for example, population level variation, effects unrelated to intervention, and sensor noise variation.

The recently published contrastive principal component approach (cPCA) (Abid et al. 2018) is one example of a technique that can be used for sets of data. cPCA builds on principal component analysis (PCA) (Hotelling 1933), a dimensionality reduction technique which projects data into a lower dimensional space while minimizing the squared loss. PCA and other dimensionality reduction techniques are popular because they allow high-dimensional data to be visualized while removing noise. cPCA seeks to find a projection to a lower dimensional space that discovers variation that is enriched in one dataset as compared to another by applying PCA to the empirical covariance matrix

$$C = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} - \alpha\frac{1}{m}\sum_{j=1}^{m} \mathbf{y}_j\mathbf{y}_j^{\mathrm{T}} \tag{1}$$

where $\{\mathbf{x}_i\}$ are the observations of interest, $\{\mathbf{y}_j\}$ are the comparison data, and α is a tuning parameter. The choice of α is a trade-off between maximizing the retained variance of the target set and minimizing the retained variance of the background set.

In this work, we develop probabilistic latent variable models applicable to the setting where contrastive analysis is desired. These models are based on the insight that it is possible to emphasize latent structures of interest while suppressing spurious, uninteresting variance in the data through carefully designed statistical models. Such models have several key advantages over deterministic approaches: it is straight forward to incorporate prior domain knowledge, missing and noisy data can naturally be modeled through appropriate

Supplemental information is available at https://arxiv.org/

noise distributions, model and feature selection can be performed through sparsity promoting prior distributions, and the model can more easily be incorporated into larger probabilistic systems in a principled manner. Through this paper, we advance the state-of-the-art in several ways. First, we develop latent variable models capable of contrastive analysis. We then demonstrate the generality of our framework by demonstrating how robust and sparse contrastive variants can be developed, learned and how automatic model selection can be performed. We also develop contrastive variants of the variational autoencoder, a deep generative model, and demonstrate its utility in modeling the density of noisy data. Finally, we vet our proposed models through extensive experiments on real world scientific data to demonstrate the utility of the proposed framework.

Contrastive Latent Variable Models

To achieve the aim of discovering patterns that are enriched in one dataset relative to another, we propose a latent variable model where some structure is shared across the two datasets and some structure is unique to the target dataset. Given a target dataset $\{\mathbf{x}_i\}_{i=1}^n$ and a background dataset $\{\mathbf{y}_j\}_{j=1}^m$, the model is specified

$$\begin{aligned}\mathbf{x}_i &= \mathbf{S}\mathbf{z}_i + \mathbf{W}\mathbf{t}_i + \boldsymbol{\mu}_x + \boldsymbol{\epsilon}_i, \;\; i = 1 \dots n \\ \mathbf{y}_j &= \mathbf{S}\mathbf{z}_j + \boldsymbol{\mu}_y + \boldsymbol{\epsilon}_j, \;\; j = 1 \dots m\end{aligned} \tag{2}$$

where $\mathbf{x}_i, \mathbf{y}_j \in \mathbb{R}^d$ are the observed data, $\mathbf{z}_i, \mathbf{z}_j \in \mathbb{R}^k$ and $\mathbf{t}_i \in \mathbb{R}^t$ are the latent variables, $\mathbf{S} \in \mathbb{R}^{d \times k}$ and $\mathbf{W} \in \mathbb{R}^{d \times t}$ are the corresponding factor loadings, $\boldsymbol{\mu}_x, \boldsymbol{\mu}_y \in \mathbb{R}^d$ are the dataset-specific means and $\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_j \in \mathbb{R}^d$ are the noise. In general, we do not expect the number of samples in the two datasets to be the same, i.e. $n \neq m$. Furthermore, there is no special relationship between the samples i and j in equation 2. The primary variables of interest are $\{\mathbf{t}_i\}_{i=1}^n$, which are the lower dimensional representation that is unique to the target dataset.

Gaussian likelihood and priors

To provide intuition into why eqn. 2 meets our goal of capturing patterns enriched in the target with respect to the background, we consider the case where the noise follows isotropic Gaussian distributions, $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \sigma^2\mathbf{I}_d)$ and $\boldsymbol{\epsilon}_j \sim \mathcal{N}(0, \sigma^2\mathbf{I}_d)$ and the latent variables are modeled using standard Gaussian distributions

$$\begin{aligned}\mathbf{x}_i|\mathbf{z}_i, \mathbf{t}_i &\sim \mathcal{N}(\mathbf{S}\mathbf{z}_i + \mathbf{W}\mathbf{t}_i + \boldsymbol{\mu}_x, \sigma^2\mathbf{I}_d) \\ \mathbf{y}_j|\mathbf{z}_j &\sim \mathcal{N}(\mathbf{S}\mathbf{z}_j + \boldsymbol{\mu}_y, \sigma^2\mathbf{I}_d) \\ \mathbf{z}_i \sim \mathcal{N}(0, \mathbf{I}_k), \quad \mathbf{z}_j &\sim \mathcal{N}(0, \mathbf{I}_k), \quad \mathbf{t}_i \sim \mathcal{N}(0, \mathbf{I}_t),\end{aligned} \tag{3}$$

where $\mathcal{N}(\mu, \Sigma)$ is a multivariate normal distribution parameterized by mean μ and covariance Σ and $\mathbf{I}_d$ denotes a $d \times d$ identity matrix. The resulting marginal distributions for the observed data are

$$\begin{aligned}\mathbf{x}_i &\sim \mathcal{N}(\boldsymbol{\mu}_x, \mathbf{W}\mathbf{W}^{\mathrm{T}} + \mathbf{S}\mathbf{S}^{\mathrm{T}} + \sigma^2\mathbf{I}_d) \\ \mathbf{y}_j &\sim \mathcal{N}(\boldsymbol{\mu}_y, \mathbf{S}\mathbf{S}^{\mathrm{T}} + \sigma^2\mathbf{I}_d).\end{aligned} \tag{4}$$

The covariance structure for the target data is additive and contains a term ($\mathbf{S}\mathbf{S}^{\mathrm{T}}$) that is shared with the background data and a term that is unique to the target data ($\mathbf{W}\mathbf{W}^{\mathrm{T}}$). This constructions allows the factor loading $\mathbf{W}$ to model the structure unique to the target. The model closely mirrors probabilistic PCA (PPCA) (Tipping and Bishop 1999; Roweis 1998) and is exactly PPCA applied to the combined datasets when the target factor loading dimensionality t is zero. Similarly, this model is exactly PPCA applied to only the target dataset when the shared factor loading dimensionality k is zero. Expectation-maximization (EM) (Dempster, Laird, and Rubin 1977) can be used to solve for the model parameters. Because EM requires conjugacy, most model formulations will not be solved this way. However, we present a summary of the EM steps to provide an intuition about the model. To provide interpretable equations in the below description, we consider the case where the factor loading matrices $\mathbf{W}$ and $\mathbf{S}$ are orthogonal.

The model parameters are $\mathbf{S}$, $\mathbf{W}$, μ_x, μ_y, σ^2 and the latent variables are $\mathbf{z}_i$, $\mathbf{z}_j$, $\mathbf{t}_i$. The lower bound of the likelihood is

$$\begin{aligned}\mathcal{L} = \sum_{i=1}^{n} \mathbb{E}_{p(\mathbf{z}_i, \mathbf{t}_i|\mathbf{x}_i)}[\ln p(\mathbf{z}_i, \mathbf{t}_i, \mathbf{x}_i)] + \\ \sum_{j=1}^{m} \mathbb{E}_{p(\mathbf{z}_j|\mathbf{y}_j)}[\ln p(\mathbf{z}_j, \mathbf{y}_j)]\end{aligned} \tag{5}$$

The M-step maximizes the lower bound of the likelihood with respect to the parameters. The update step for the shared factor loading is

$$\begin{aligned}\tilde{\mathbf{S}} = \big[(\mathbf{B} + (\mathbf{I} - \mathbf{W}\mathbf{R}^{-1}\mathbf{W}^{\mathrm{T}})\mathbf{T})\mathbf{S}\big] \\ (\sigma^2\mathbf{I} + \mathbf{M}^{-1}\mathbf{S}^{\mathrm{T}}(\mathbf{B} + \mathbf{T})\mathbf{S})^{-1}\end{aligned} \tag{6}$$

where $\mathbf{B}$ is the sample covariance of the background data, $\mathbf{T}$ is the sample covariance of the target data, $\mathbf{M} = \sigma^2\mathbf{I}_k + \mathbf{S}^{\mathrm{T}}\mathbf{S}$, and $\mathbf{R} = \sigma^2\mathbf{I}_t + \mathbf{W}^{\mathrm{T}}\mathbf{W}$. The update step for the target factor loading is

$$\tilde{\mathbf{W}} = ((\mathbf{I} - \mathbf{S}\mathbf{M}^{-1}\mathbf{S})\mathbf{T}\mathbf{W})(\sigma^2\mathbf{I} + \mathbf{R}^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{T}\mathbf{W})^{-1} \tag{7}$$

Details on the derivation can be found in the supplemental information. It is useful to recall that the orthogonal projection onto the range space of a matrix $\mathbf{A}$ is given by $\mathbf{P} = \mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}$ and the orthogonal projection onto the nullspace of $\mathbf{A}$ is given by $\mathbf{I} - \mathbf{P}$. In eqn. 7, $\mathbf{I} - \mathbf{S}\mathbf{M}^{-1}\mathbf{S}$ can be expanded using the definition of $\mathbf{M}$ to $\mathbf{I} - \mathbf{S}(\sigma^2\mathbf{I}_k + \mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}$. Similarly, in eqn. 6, $\mathbf{I} - \mathbf{W}\mathbf{Q}^{-1}\mathbf{W}^{\mathrm{T}}$ can be expanded using the definition of $\mathbf{Q}$ to $\mathbf{I} - \mathbf{W}(\sigma^2\mathbf{I}_t + \mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}$. When σ^2 is small, these equations are similar to the projection onto the nullspace of $\mathbf{S}$ and $\mathbf{W}$, respectively. This matches our intuition as to how these factor loading matrices are updated: in a sense, the part of the target data captured by the target factor loading space is *projected away* before updating the shared factor loading space, and vice versa. This behavior is similar to cPCA. In eqn. 1, as α goes to infinity, directions not in the null space of the background data covariance are given an infinite penalty. When this is the case, cPCA projects the target data onto the null space of the background data and then performs PCA (Abid et al. 2018).

The update steps can also be compared to the PPCA updates. For the factor loading matrix $\mathbf{W}$, the update step is:

$$\tilde{\mathbf{W}} = \mathbf{T}\mathbf{W}(\sigma^2\mathbf{I} + \mathbf{R}^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{T}\mathbf{W})^{-1} \tag{8}$$

which is the same as eqn. 7, except for the projection term.

Beyond Gaussian Models

The assumptions of Gaussianity are not necessary for recovering latent structure that enriches desired patterns in the target dataset. We can more generally express the proposed model as:

$$\begin{aligned} p(\mathcal{D},&\{\mathbf{z}_i, \mathbf{t}_i\}_{i=1}^n, \{\mathbf{z}_j\}_{j=1}^m; \Theta) = \\ &p(\Theta) \prod_{i=1}^n p(\mathbf{x}_i | \mathbf{z}_i, \mathbf{t}_i; \mathbf{W}, \mathbf{S}, \boldsymbol{\mu}_x, \sigma^2) p(\mathbf{z}_i) p(\mathbf{t}_i) \\ &\prod_{j=1}^m p(\mathbf{y}_j | \mathbf{z}_j; \mathbf{S}, \boldsymbol{\mu}_y, \sigma^2) p(\mathbf{z}_j), \end{aligned} \tag{9}$$

where $\mathcal{D} = \{\{\mathbf{x}_i\}_{i=1}^n, \{\mathbf{y}_j\}_{j=1}^m\}$ and $\Theta = \{\mathbf{W}, \mathbf{S}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \sigma^2\}$. The primary modeling decisions are to choose the appropriate likelihoods and priors on the loading matrices. The particular choices are governed by the application and domain specific knowledge.

However, this flexibility comes at a price: the posterior distributions $p(\mathbf{t}_i, \mathbf{z}_i, \mathbf{z}_j | \mathcal{D})$ are no longer guaranteed to be tractable. Consequently, the EM algorithm sketched in the previous section is no longer available and instead, we use variational inference (Wainwright and Jordan 2008). In summary, the intractable posteriors are approximated with tractable surrogates $q(\mathbf{t}_i | \lambda_{\mathbf{t}_i}) q(\mathbf{z}_i | \lambda_{\mathbf{z}_i}) q(\mathbf{z}_j | \lambda_{\mathbf{z}_j})$ and divergence $\mathrm{KL}(q \,||\, p)$ is minimized with respect to the variational parameters $\lambda = \{\{\lambda_{\mathbf{z}_i}, \lambda_{\mathbf{t}_i}\}_{i=1}^n, \{\lambda_{\mathbf{z}_j}\}_{j=1}^m\}$. This is equivalent to maximizing the lower bound of the marginal likelihood,

$$\begin{aligned} p(\mathcal{D}; \Theta) &\geq \mathcal{L}(\lambda, \Theta) \\ &= \sum_i \mathbb{E}_{q(\mathbf{z}_i; \lambda_{\mathbf{z}_i}) q(\mathbf{t}_i; \lambda_{\mathbf{t}_i})} [\ln p(\mathbf{x}_i | \mathbf{z}_i, \mathbf{t}_i; \Theta_{\setminus \{\boldsymbol{\mu}_y\}})] \\ &\quad - \mathrm{KL}(q(z_i; \lambda_{\mathbf{z}_i}) \,||\, p(\mathbf{z}_i)) - \mathrm{KL}(q(\mathbf{t}_i; \lambda_{\mathbf{t}_i}) \,||\, p(\mathbf{t}_i)) \\ &\quad + \sum_j \mathbb{E}_{q(\mathbf{z}_j; \lambda_{\mathbf{z}_j})} [\ln p(\mathbf{y}_j | \mathbf{z}_j; \Theta_{\setminus \{\boldsymbol{\mu}_x, \theta_x\}})] \\ &\quad - \mathrm{KL}(q(\mathbf{z}_j; \lambda_{\mathbf{z}_j}) \,||\, p(\mathbf{z}_j)) + \ln p(\Theta) \end{aligned} \tag{10}$$

where $\Theta_{\setminus \{\cdot\}}$ implies the parameters in Θ except the parameters denoted in the set. Depending on the choice of q and p the expectations required for computing $\mathcal{L}(\lambda, \Theta)$ may themselves be intractable. We use recently proposed black box techniques (Ranganath, Gerrish, and Blei 2014; Kingma and Welling 2014; Rezende, Mohamed, and Wierstra 2014; Titsias and Lázaro-Gredilla 2014) to sidestep this additional complication. In particular, we approximate the intractable expectations in $\mathcal{L}(\Theta, \lambda)$ with unbiased Monte-Carlo estimates, $\tilde{\mathcal{L}}(\Theta, \lambda)$. Because the latent variables of interest are continuous, we are able to use reparameterization gradients (Kingma and Welling 2014; Rezende, Mohamed, and Wierstra 2014) to differentiate through the sampling process and obtain low variance estimates of $\nabla_{\lambda,\Theta} \mathcal{L}(\Theta, \lambda)$, $\nabla_{\lambda,\Theta} \tilde{\mathcal{L}}(\Theta, \lambda)$. Using the noisy but unbiased gradients, optimization can proceed using a stochastic gradient ascent variant, e.g. ADAM (Kingma and Ba 2014). In our experiments we use Edward (Tran et al. 2016), a library for probabilistic modeling, to implement these inference strategies for the proposed models. We sketch the pseudocode for variational learning in Algorithm 1.

Algorithm 1 Pseudocode

```
**Input** Model $p(\mathcal{D}; \Theta)$, variational approximations $q(\{z_i, t_i\}_{i=1}^n, \{z_j\}_{j=1}^m \mid \lambda)$
**Output**: Optimized $\Theta$ and variational parameters $\lambda$
Initialize $\lambda$ and $\Theta$.
**repeat**
    Use reparameterization trick to compute unbiased estimates of the gradients of the objective in Eqn. 10, $\nabla_{\lambda,\Theta} \tilde{\mathcal{L}}(\lambda, \Theta)$
    Update $\lambda^{(l+1)} \leftarrow \mathrm{ADAM}(\lambda^{(l)}, \nabla_{\lambda} \tilde{\mathcal{L}}(\lambda, \Theta))$, $\Theta^{(l+1)} \leftarrow \mathrm{ADAM}(\Theta^{(l)}, \nabla_{\Theta} \tilde{\mathcal{L}}(\lambda, \Theta))$
**until** convergence
```

Finally, we note that the black box inference framework does not restrict us to point estimates of Θ. As we will illustrate in the next section, it is possible to infer variational distributions over Θ by specifying an appropriate approximation $q(\Theta \mid \lambda_{\Theta})$.

cLVM Variants

We refer to the base structure of the model as provided in eqn. 9 as a contrastive latent variable model, cLVM. As previously noted, different choices for the distributions in eqn. 9 can be made to address the specific challenges of the application. Several models are introduced here and are summarized in Table 1.

Sparse cLVM One application-specific problem is feature selection. In unsupervised learning, there is often a secondary goal of learning a subset of measurements that are of interest which is motivated by improved interpretability. This is especially important when the observed data is very high-dimensional. For instance, many biological assays result in datasets that have tens of thousands of measurements such as SNP and RNA-Seq data. During data exploration, discovering a subset of these measurements that is important to the target population can help guide further analysis. To learn a latent representation that is only a function of a subset of the observed dimensions, certain rows of the target factor loading, $\mathbf{W}$, must be zero. The observed data corresponding to the zero rows in $\mathbf{W}$ then have no contribution to the latent representation $\mathbf{t}$. Because there is no restriction on $\mathbf{S}$, a sparsity requirement for $\mathbf{W}$ does not imply that the corresponding observation is zero.

One way to achieve this behavior is by using a regularization penalty on the model parameters. The penalty is added to the objective function to incite certain behavior. Regularization penalties can be related to priors by noting that

Model name	Prior	Likelihood	Variational Approximation
cLVM	–	Gaussian	–
Sparse cLVM	$p(\mathbf{W}) = \prod_{i=1}^{d} \mathcal{N}(\mathbf{W}_{i:} \vert \rho_i, \tau)$ $C^+(\rho_i \vert 0, 1) C^+(\tau \vert 0, b_g)$	Gaussian	$q(\mathbf{W}) = \mathcal{N}(\cdot, \cdot)$ $q(\ln \boldsymbol{\rho}) = \mathcal{N}(\cdot, \cdot)$ $q(\ln \tau) = \mathcal{N}(\cdot, \cdot)$
cLVM with model selection	$p(\mathbf{S}) = \prod_{i=1}^{d-1} \mathcal{N}(\mathbf{S}_{:j} \vert 0, \alpha_j) \mathrm{IG}(\alpha_j \vert a, b)$	Gaussian	$q(\mathbf{S}) = \mathcal{N}(\cdot, \cdot)$ $q(\ln \boldsymbol{\alpha}) = \mathcal{N}(\cdot, \cdot)$
Robust cLVM	$p(\sigma^2) = \mathrm{IG}(a, b)$	Student's t	$q(\ln \sigma^2) = \mathcal{N}(\cdot, \cdot)$
cVAE	–	Gaussian parameterized by neural network	$q(\mathbf{z}_i, \mathbf{t}_i) = \mathcal{N}(g_\mu(\cdot), g_\sigma(\cdot))$

Table 1: Summary of the model variants. For all of the models in the table, the latent variables $\{\mathbf{z}_i, \mathbf{t}_i\}_{i=1}^{n}, \{\mathbf{z}_j\}_{j=1}^{m}$ are modeled as standard Gaussians and the variational distributions are also Gaussian, unless otherwise noted. The model choice depends on the application. The various models are not mutually exclusive and may also be combined.

$\log p(\mathbf{W}) \propto r(\mathbf{W})$, where $r(\cdot)$ is the penalty function. For feature selection, a group sparsity penalty (Yuan and Lin 2007) could be used. The rows of $\mathbf{W} \in \mathbb{R}^{d \times t}$ are penalized:

$$r(\mathbf{W}) = \rho \sum_{i=1}^{d} \sqrt{p_i} \|\mathbf{W}_{i:}\|_2 \tag{11}$$

where $\mathbf{W}_{i:}$ is the i^{th} row of $\mathbf{W}$. This functional form is known to lead to sparsity at the group level, i.e. all members of a group are zero or non-zero. For increasing values of ρ, the target factor loading matrix has a larger number of zero-valued rows.

Sparsity inducing priors such as the automatic relevance determination (ARD) (Bishop 1999a; Virtanen et al. 2011; Klami, Virtanen, and Kaski 2013) or global-local shrinkage priors such as the horseshoe (Carvalho, Polson, and Scott 2009; 2010) can also be easily incorporated into the frameworkÜsing the horseshoe prior as an example, the i^{th} row of $\mathbf{W}$ is modeled,

$$\begin{gathered} \mathbf{W}_{i:} \vert \rho_i, \tau \sim \mathcal{N}(0, \rho_i^2 \tau^2 \mathbf{I}_t) \\ \rho_i \sim C^+(0, 1), \quad \tau \sim C^+(0, b_g) \end{gathered} \tag{12}$$

where $a \sim C^+(0, b)$ is the half-Cauchy distribution with density $p(a \vert b) = \frac{2}{\pi} b(1 + \frac{a^2}{b^2})$ for $a > 0$. The horseshoe prior is useful for subset selection because it has heavy tails and an infinite spike at zero. Further discussion can be found in the supplemental information. For both the prior and regularization formulations, *groups* of rows in $\mathbf{W}$ could also be used instead of single rows if such a grouping exists.

cLVM with Automatic Model Selection The ARD prior is more typically applied to the columns of a factor loading matrix. This use allows for automatic selection of the dimension of the matrix. This could also be done in the cLVM model. Although both latent spaces can have any dimension less than d, which must be selected, we generally recommend setting the target dimension to two for visualization purposes. To select the dimension of the shared space, the percent variance explained can be analyzed or a prior, such as the ARD prior can be used. The columns of $\mathbf{S}$ are modeled

$$\mathbf{S}_{:j} \vert \alpha_j \sim \mathcal{N}(0, \alpha_j \mathbf{I}_d), \quad \alpha_j \sim \mathrm{IG}(a_0, b_0). \tag{13}$$

The ARD prior has been shown to be effective at model selection for PPCA models (Bishop 1999b).

Robust cLVM Another application-specific goal may be to systematically handle outliers in the dataset. Similar to PPCA, the cLVM model is sensitive to outliers and can produce poor results if outliers are not addressed. It may be possible to remove outliers from the dataset, however this is typically a manual process that requires domain expertise and an understanding of the process that generated the data. A more general approach to handling outliers uses a heavy-tailed distribution to describe the data. One approach for constructing heavy tailed distributions is through scale mixtures of Gaussians (West 1987). Consider,

$$\sigma^2 \sim \mathrm{IG}(a, b). \tag{14}$$

The resulting marginal distribution of the observed data is

$$\begin{aligned} p(\mathbf{x}_i \vert \mu, a, b) &= \prod_{k=1}^{d} \int_0^\infty \mathcal{N}(\mathbf{x}_{ik} \vert \boldsymbol{\mu}, \sigma^2) \mathrm{IG}(\sigma^2 \vert a, b) d\sigma^2 \\ &= \prod_{k=1}^{d} \mathrm{St}(\mathbf{x}_{ik} \vert \boldsymbol{\mu}, \nu = 2a, \lambda = \frac{a}{b}) \end{aligned} \tag{15}$$

where St indicates a Student's t-distribution (Archambeau, Delannay, and Verleysen 2006). The larger probability mass in the tails of the Student's t-distribution, as compared to the normal distribution, allows the model to be more robust to outliers.

Beyond Linear Models

Contrastive Variational Autoencoders Thus far we have only considered models that linearly map latent variables $\mathbf{z}$ and $\mathbf{t}$ to the observed space. The linearity constraint can be relaxed, and doing so leads to powerful generative models capable of accounting for nuisance variance.

$$\begin{aligned} \mathbf{x}_i &= f_{\theta_s}(\mathbf{z}_i) + f_{\theta_t}(\mathbf{t}_i) + \boldsymbol{\epsilon}_i, \quad i = 1 \ldots n \\ \mathbf{y}_j &= f_{\theta_s}(\mathbf{z}_j) + \boldsymbol{\epsilon}_j, \quad j = 1 \ldots m, \end{aligned} \tag{16}$$

where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \sigma^2)$, $\boldsymbol{\epsilon}_j \sim \mathcal{N}(0, \sigma^2)$, and f_{θ_s}, f_{θ_t} represent non-linear transformations parameterized by neural networks. The latent variables are modeled using standard Gaussian distributions, as before. Observe that similar to the linear case (eqn. 2) the target and background data share the projection f_{θ_s} while the target retains a private projection f_{θ_t}. This construction forces f_{θ_s} to model commonalities between the target and background data while allowing f_{θ_t} to capture structure unique to the target.

This model can be learned by maximizing the lower bound to the marginal likelihood $p(\mathcal{D}|\Theta)$, $\Theta = \{\theta_s, \theta_t, \mu_x, \mu_y, \sigma^2\}$, analogously to eqn. 10. However, a large amount of data is typically required to learn such a non-linear model well. Moreover, since the number of latent variables proliferate with increasing data, it is computationally more efficient to amortize the cost of inferring the latent variables through inference networks shared between the data instances. In particular, we parametrize the variational posteriors $q_{\lambda_t}(\mathbf{z}_i, \mathbf{t}_i|\mathbf{x}_i) = \mathcal{N}(\mathbf{z}_i|g^{\mu}_{\lambda_t}(\mathbf{x}_i), g^{\sigma}_{\lambda_t}(\mathbf{x}_i))\mathcal{N}(\mathbf{t}_i|g^{\mu}_{\lambda_t}(\mathbf{x}_i), g^{\sigma}_{\lambda_t}(\mathbf{x}_i)$ and $q_{;\lambda_s}(\mathbf{z}_j|\mathbf{y}_j) = \mathcal{N}(\mathbf{z}_j|g^{\mu}_{\lambda_s}(\mathbf{y}_j), g^{\sigma}_{\lambda_s}(\mathbf{y}_j))$, where λ_t and λ_s are inference network parameters. Unlike eqn. 10 where the variational parameters grow with the number of data instances, the variational parameters λ_t and λ_s do not. λ_t is shared amongst the target instances while λ_s is shared between the background examples. This is an example of *amortized variational inference* (Dayan et al. 1995; Gershman and Goodman 2014). Finally, learning proceeds by maximizing the evidence lower bound,

$$
\begin{aligned}
&p(\mathcal{D};\Theta) \geq \mathcal{L}(\Theta, \lambda_s, \lambda_t)\\
&= \sum_i \mathbb{E}_{q_{\lambda_t}(\mathbf{z}_i,\mathbf{t}_i|\mathbf{x}_i)}[\ln p(\mathbf{x}_i|\mathbf{z}_i, \mathbf{t}_i; \Theta_{\backslash\{\boldsymbol{\mu}_y\}})]\\
&- \mathrm{KL}(q_{\lambda_t}(\mathbf{z}_i, \mathbf{t}_i|\mathbf{x}_i) \mid\mid p(\mathbf{z}_i)p(\mathbf{t}_i))\\
&+ \sum_j \mathbb{E}_{q_{\lambda_s}(\mathbf{z}_j|\mathbf{y}_j)}[\ln p(\mathbf{y}_j|\mathbf{z}_j; \Theta_{\backslash\{\boldsymbol{\mu}_x, \theta_x\}})]\\
&- \mathrm{KL}(q_{\lambda_s}(\mathbf{z}_j|\mathbf{y}_j) \mid\mid p(\mathbf{z}_j)) + \ln p(\Theta), \qquad (17)
\end{aligned}
$$

with respect to Θ and λ_s, λ_t. The KL terms are available to us in closed form, however the expectation terms are intractable and we again resort to Monte Carlo approximations and re-parameterized gradients to enable stochastic gradient ascent. We refer to this combination of the non-linear model and the amortized variational inference scheme as the contrastive variational auto encoder (cVAE).

Related Work

There are many techniques for dimensionality reduction, e.g. (Hotelling 1933; van der Maaten and Hinton 2008; Cox and Cox 2008). This review focuses on dimensionality techniques that use sets of data and/or address issues related to nuisance variation. Canonical correlation analysis (CCA) (Hotelling 1936) and its probabilistic variant (PCCA) (Bach and Jordan 2005) use two (or more) sets of data, however requires that samples are paired views (or higher dimensional sets of views) of the same sample. For instance perhaps several tests are run on a single patient and therefore the tests are linked via the patient identity. In CCA, the number of samples in the sets must be equal, $n = m$, however the dimensionality of each sample does not need to be the same. Damianou, Lawrence, and Ek proposed a nonlinear extension of PCCA where the mappings are sampled from a Gaussian process. The resulting model is a multi-view extension of GP-LVM (Lawrence 2005), but still requires linking the samples across datasets.

In this work, we propose addressing nuisance variation in the dataset by introducing a structure to the latent representation. Schulam and Saria investigate a similar idea with respect to sharing representations across different parts of the full data. In their work, a hierarchical model for disease trajectory is proposed where some of the model coefficients are shared across subsets of the data, e.g. total population and individual. This idea has also been proposed for the unsupervised analysis of time series data (Hsu, Zhang, and Glass 2017; Li and Mandt 2018). Data samples are assumed to have a latent representation that can be partitioned into static and dynamic contributions. None of these works have considered a contrastive setting. There has also been work in addressing explicit sources of nuisance variation. Louizos et al. explores a setting where certain variables within the dataset are a priori identified as nuisance and the remaining variables contribute to the latent representation. The observed data is modeled $\mathbf{x} \sim p_\theta(\mathbf{z}, \mathbf{s})$ where $\mathbf{s}$ are the observed nuisance variables.

Experiments

Contrastive latent variable models have applications in subgroup discovery, feature selection, and de-noising, each of which is demonstrated here leveraging different modeling choices. We use examples from Abid et al. to highlight the similarities and differences between the two approaches. The results of cLVM as applied to synthetic datasets can be found in the supplemental information.

Subgroup Discovery for Incomplete Data

To demonstrate the use of cLVM for subgroup discovery, we use a dataset of mice protein expression levels (Higuera, Gardiner, and Cios 2015). The target dataset has 270 samples of two unknown classes of mice: trisomic (Down Syndrome model) and control. The background dataset has 135 known control samples. Each sample has 77 measurements. The dataset contains missing values at a level of approximately 1.6% due to technical artifacts and sampling that cannot be repeated. One of the advantages of the probabilistic approach is that it naturally handles missing data. Depending why the data is missing, missing data can either be ignored, marginalized, or explicitly modeled. For the mice protein dataset, we marginalize over the missing values by treating the missing data as latent variables and adding a corresponding normal variational approximation. Increasing levels of missing data were tested by artificially removing data from the target dataset. The robust variation of the model is applied to account for other possible data issues. The target and shared dimensionalities are both set to two. Fig. 1 shows the latent representation using cPCA and robust cLVM for the naturally occurring missing level, 25%, 50%, and 75%

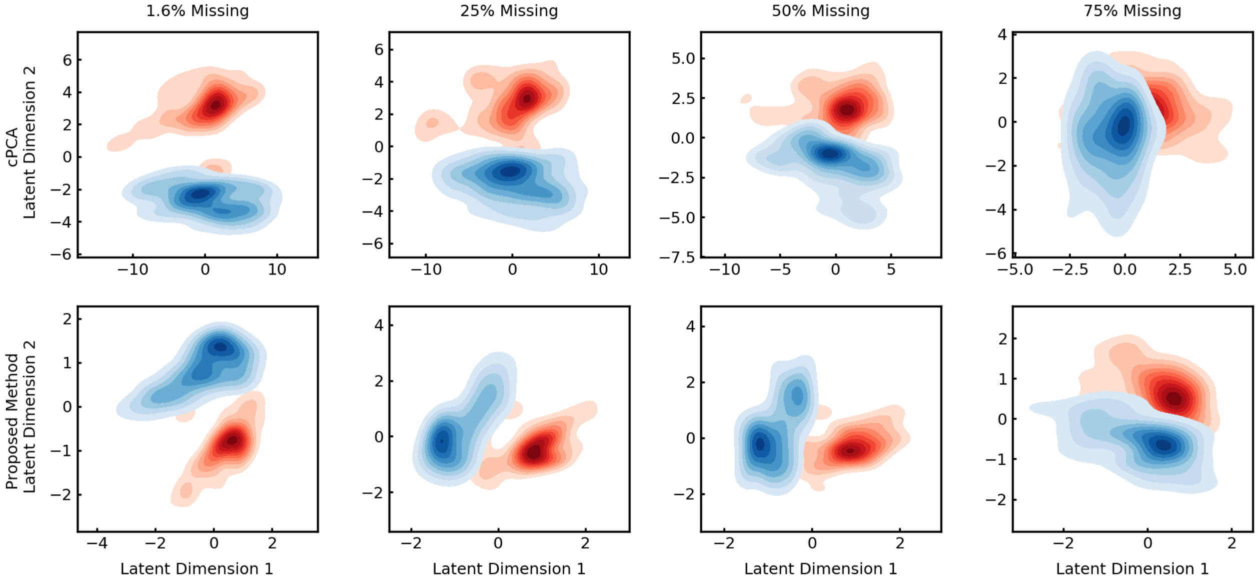

Figure 1: cLVM is robust to missing data. Density plots of the subgroups revealed in the target latent representation of the mice protein expression data. Red and blue points are the control and trisomic mice samples, respectively. The rows use cPCA and robust cLVM to learn the latent representation, respectively. Each column uses a different level of missing data, starting with the leftmost column containing the natural level of missing data. PCA is unable to perform subgroup discovery (see supplemental information) and robust cLVM is better able to perform subgroup discovery in the presence of missing data.

missing data. cPCA does not have natural handling for missing data therefore mean imputation was used to first fill-in. PCA is unable to recover the structure in the dataset (see supplemental information for results). Both cPCA and robust cLVM find the subsets, however, the proposed method is better able to discover the subgroups as the amount of missing data increases.

Subgroup Discovery for High Dimensional Data

To highlight the use of cLVM for subgroup discovery in high-dimensional data, we use a dataset of single cell RNA-Seq measurements (Zheng et al. 2017). The target dataset consists of expression levels from 7,898 samples of bone marrow mononuclear cells before and after stem cell transplant from a leukemia patient. The background contains expression levels from 1,985 samples from a healthy individual. Pre-processing of the data reduces the dimensionality from 32,738 to 500 (Zheng et al. 2017; Abid et al. 2018). Given the size of the data to explore, it is useful in this setting to use an ARD prior to automatically select the dimensionality of the shared latent space. The target latent space is set to two and an $\text{IG}(10^{-3}, 10^{-3})$ prior is used for the columns of the shared factor loading. Fig. 2a shows the resulting latent representation, which is able to discover the subgroups, whereas PCA is not (see supplemental information). Fig. 2b compares the percent of variance explained in the ranked columns as compared to the cLVM model without model selection. The model with ARD uses over 100 fewer columns in the shared factor loading matrix and avoids an analysis to manually select the dimension.

Automatic Feature Selection using Sparse cLVM

The third example uses a dataset, referred to as mHealth, that contains 23 measurements of body motion and vital signs from four types of signals (Banos et al. 2014; 2015). The participants in the study complete a variety of activities. The target data is composed of the unknown classes of cycling and squatting and the background data is composed of the subjects lying still. In this application, we demonstrate feature selection by learning a latent representation that both separates the two activities and uses only a subset of the signals. A group sparsity penalty is used, as described in the methodology, on the target factor loading. The target dimension is two, the shared dimension is twenty, and ρ is 400. ρ is selected by varying its value and inspecting the latent representation. The latent representation using regularization is shown in Fig. 2c. The two classes are clearly separated. Fig. 2d shows the row-wise norms of the target factor loading. The last six dimensions, corresponding to the magnetometer readings, are all zero which indicates that the magnetometer measurements are not important for differentiating the two classes and can be excluded from further analysis.

De-noised Generative Modeling using cVAE

Finally, to demonstrate the utility of cVAE, we consider a dataset of corrupted images (see Fig. 3a). This dataset was created by overlaying a randomly selected set of $30,000$ MNIST (LeCun et al. 1998) digits on randomly selected images of the grass category from Imagenet (Russakovsky et al. 2015). The background is $30,000$ grass images. We train

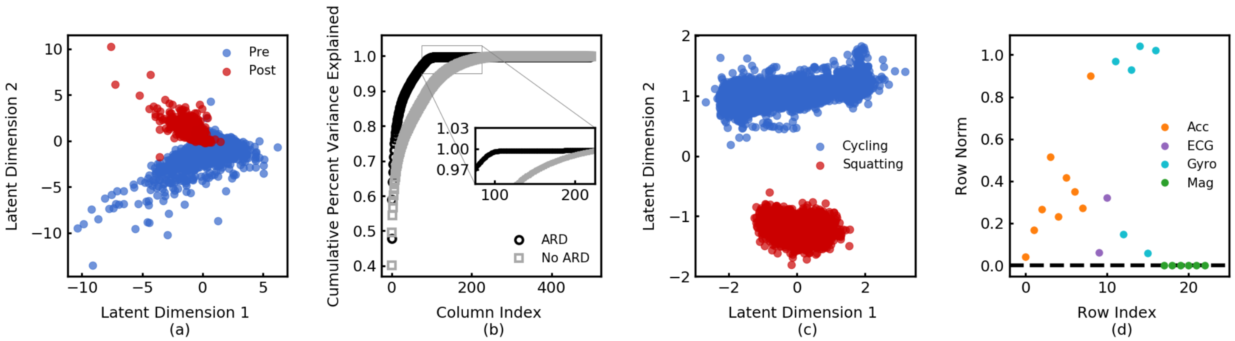

Figure 2: cLVM variants allow for model and feature selection. (a) Subgroups revealed in the target latent representation for the RNA-Seq dataset using the model selection cLVM variant. (b) The percent variance explained by the ordered columns of the shared factor loading for LVM with and without ARD (model selection). The ARD model has over 100 fewer non-zero columns in the shared factor loading. (c) Subgroups revealed in the target latent representation for the mHealth dataset using sparse cLVM. (d) The norms of the rows of the target factor loading for sparse LVM where the different colors correspond to different sensor types. The six dimensions with zero-valued norms correspond to magnetometer readings.

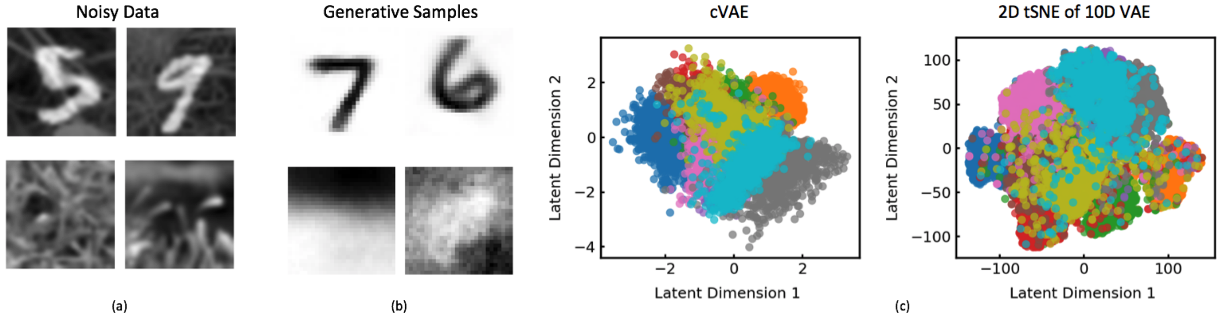

Figure 3: cVAE recovers meaningful structure from noisy data. (a) Samples of the target noisy images of digits on grass and background grass images. (b) Generative samples of the de-noised target (top row) and background (bottom row) which are enabled by the cVAE structure. Note there is no correspondence between the samples in (a) and (b). (c) The 2D cVAE projection and a 2D tSNE projection of a VAE with 10 dimensional space. The colors represent different digits.

a cVAE with a two-dimensional target latent space and an eight-dimensional shared space. We use fully connected encoder and decoder networks with two hidden layers with 128 and 256 hidden units employing rectified-linear nonlinearities. For the cVAE, both the target and shared decoders θ_s and θ_t use identical architectures. We compare against a standard variational autoencoder with an identical architecture and employ a latent dimensionality of ten, to match the combined dimensionality of the shared and target spaces of the contrastive variant. Fig. 3c presents the results of this experiment. The latent projections for the cVAE cluster according to the digit labels. VAE on the other hand confounds the digits with the background and fails to recover meaningful latent projections. Moreover, cVAE allows us to selectively generate samples from the target or the background space, Fig. 3b. The samples from the target space capture the digits, while the background samples capture the coarse texture seen in the grass images. Additional comparisons with a VAE using a two dimensional latent space is available in the supplemental.

Conclusions

Dimensionality reduction methods are important tools for unsupervised data exploration and visualization. We propose a probabilistic model for improved visualization when the goal is to learn structure in one dataset that is enriched as compared to another. The latent variable model's core characteristic is that it shares some structure across the two datasets and maintains unique structure for the dataset of interest. The resulting cLVM model is demonstrated using robust, sparse, and nonlinear variations. The method is well-suited to scenarios where there is a control dataset, which is common in scientific and industrial applications.

References

Abid, A.; Zhang, M. J.; Bagaria, V. K.; and Zou, J. 2018. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature Communications* 9:2134.

Archambeau, C.; Delannay, N.; and Verleysen, M. 2006. Robust probabilistic projections. In *ICML*, 33–40. ACM.

Bach, F. R., and Jordan, M. I. 2005. A probabilistic interpretation of canonical correlation analysis. Technical report, University of California, Berkeley.

Banos, O.; Garcia, R.; Holgado, J. A.; Damas, M.; Pomares, H.; Rojas, I.; Saez, A.; and Villalonga, C. 2014. mhealthdroid: A novel framework for agile development of mobile health applications. In *6th International Work-conference on Ambient Assisted Living and Daily Activities*, 91–98. Springer.

Banos, O.; Villaonga, C.; Rafael, G.; Saez, A.; Damas, M.; Holgado-Terriza, J. A.; Lee, S.; Pomares, H.; and Rojas, I. 2015. Design, implementation, and validation of a novel open framework for agile development of mobile health applications. *BioMedical Engineering Online* 14:1–20.

Bishop, C. M. 1999a. Variational principal components. In *ICANN*, 509–514. IEE.

Bishop, C. M. 1999b. Bayesian PCA. In *NIPS*, 382–388.

Carvalho, C. M.; Polson, N. G.; and Scott, J. G. 2009. Handing sparsity via the horseshoe. In *AISTATS*, 73–80. JMLR.

Carvalho, C. M.; Polson, N. G.; and Scott, J. G. 2010. The horseshoe estimator for sparse signals. *Biometrika* 97:465–480.

Cox, M. A. A., and Cox, T. F. 2008. Multidimensional scaling. In *Handbook of Data Visualizations*. Springer. 315–347.

Damianou, A.; Lawrence, N. D.; and Ek, C. H. 2016. Multi-view learning as a nonparametric nonlinear inter-battery factor analysis. *arXiv preprint arXiv:1604.04939*.

Dayan, P.; Hinton, G. E.; Neal, R. M.; and Zemel, R. S. 1995. The helmholtz machine. *Neural computation* 7(5):889–904.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistics Society. Series B (Methodological)* 39:1–38.

Gershman, S., and Goodman, N. 2014. Amortized inference in probabilistic reasoning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

Higuera, C.; Gardiner, K. J.; and Cios, K. J. 2015. Self-organizing feature maps identify proteins critical to learning in a mouse model of down syndrome. *PLOS ONE* 10:e0129126.

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24:417.

Hotelling, H. 1936. Relations between two sets of variables. *Biometrika* 28:321–377.

Hsu, W.-N.; Zhang, Y.; and Glass, J. 2017. Unsupervised learning of disentangled and interpretable representations from sequential data. In *NIPS*, 1878–1889.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., and Welling, M. 2014. Stochastic gradient vb and the variational auto-encoder. In *ICLR*.

Klami, A.; Virtanen, S.; and Kaski, S. 2013. Bayesian canonical correlation analysis. *Journal of Machine Learning Research* 14:965–1003.

Lawrence, N. 2005. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research* 6:1783–1816.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, 2278–2324.

Li, Y., and Mandt, S. 2018. Disentangled sequential autoencoder. In *ICML*, 5656–5665. PMLR.

Louizos, C.; Swersky, K.; Li, Y.; Welling, M.; and Zemel, R. 2016. The variational fair autoencoder. In *ICLR*.

Ranganath, R.; Gerrish, S.; and Blei, D. M. 2014. Black box variational inference. In *AISTATS*, 814–822.

Rezende, D. J.; Mohamed, S.; and Wierstra, D. 2014. Stochastic backpropogration and approximate inference in deep generative models. In *ICML*, 1278–1286. PMLR.

Roweis, S. T. 1998. EM algorithms for PCA and SPCA. In *NIPS*, 626–632.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Li, F.-F. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115:211–252.

Schulam, P., and Saria, S. 2015. A framework for individualizing of disease trajectories by exploiting multi-resolution structure. In *NIPS*, 748–756.

Tipping, M. E., and Bishop, C. M. 1999. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B* 61:611–622.

Titsias, M., and Lázaro-Gredilla, M. 2014. Doubly stochastic variational Bayes for non-conjugate inference. In *ICML*, 1971–1979. PMLR.

Tran, D.; Kucukelbir, A.; Dieng, A. B.; Rudolph, M.; Liang, D.; and Blei, D. M. 2016. Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*.

van der Maaten, L., and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9:2579–2605.

Virtanen, S.; Jia, J.; Klami, A.; and Darrell, T. 2011. Factorized multi-modal topic model. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 843–851. ACM.

Wainwright, M. J., and Jordan, M. I. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning* 1(1–2):1–305.

West, M. 1987. On scale mixtures of normal distributions. *Biometrika* 74:646–648.

Yuan, M., and Lin, Y. 2007. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* 68:49–67.

Zheng, G. X. Y.; Terry, J. M.; Belgrader, P.; and Ryvkin, P. E. 2017. Massively parallel digital transcriptional profiling of single cells. *Nature Communications* 8:14049.