Emergent Sparsity in Variational Autoencoder Models

Bin Dai  
Tsinghua University  
daib13@mails.tsinghua.edu.cn

David Wipf  
Microsoft Research  
davidwipf@gmail.com

I. INTRODUCTION

We begin with a data set \( X = \{x^{(i)}\}_{i=1}^{n} \) composed of \( n \) i.i.d. samples of some random variable \( x \in \mathbb{R}^{d} \) of interest. The variational autoencoder (VAE) has recently been proposed to provide a tractable approximation for \( p_{0}(x) \), knowledge of which would allow us to generate new samples of \( x \) [1], [2]. Moreover we assume that each sample is governed by unobserved latent variables \( z \in \mathbb{R}^{n} \), such that

\[
p_{0}(x) = \int p_{0}(x|z)p(z)dz,
\]

where \( \theta \) are parameters defining the unknown distribution.

Given that this integral is intractable in all but the simplest cases, the VAE optimizes with respect to \( \theta \) a tractable upper bound on \( -\log p_{0}(x) \). This variational bound is given by

\[
\mathcal{L}(\theta, \phi) \triangleq \sum_{i} \left[ \text{KL} \left[ q_{\phi} \left( z|x^{(i)} \right) \| p(z) \right] - E_{q_{\phi} \left( z|x^{(i)} \right)} \left[ \log p_{0} \left( x^{(i)}|z \right) \right] \right],
\]

where \( q_{\phi} \left( z|x^{(i)} \right) \), parameterized by \( \phi \), defines an arbitrary approximating distribution controlling the tightness of the bound, with equality iff \( q_{\phi} \left( z|x^{(i)} \right) = p_{0} \left( z|x^{(i)} \right) \). Additionally, \( q_{\phi} \left( z|x \right) \) can be viewed as an encoder model that defines a conditional distribution over the latent ‘code’ \( z \), while \( p_{0} \left( x|z \right) \) can be interpreted as a decoder model since, given a code \( z \) it quantifies the distribution over \( x \).

By far the most common distributional assumptions are that \( p(z) = \mathcal{N}(z; 0, I) \) and the encoder model satisfies \( q_{\phi} \left( z|x \right) = \mathcal{N}(z; \mu_{x}, \Sigma_{x}) \) for continuous data, with means and covariances defined analogously. In practice, these encoder/decoder moments can all be parameterized by differentiable deep neural networks to introduce arbitrary representational power. Furthermore, (1) can be minimized via stochastic gradient descent and a simple reparameterization trick [1], [2].

II. EMERGENT SPARSITY IN VAE MODELS

Although originally motivated as a form of generative model for approximating intractable distributions, we now demonstrate subtle attributes of the VAE cost that are particularly useful for finding low-dimensional structure in high-dimensional data corrupted by sparse outliers. For this purpose, we consider the simplified situation where the decoder mean function is affine, i.e., \( \mu_{x} = Wz + b \), while all other VAE models are arbitrary, high-capacity neural networks. While this administers considerable capacity to the model at the potential risk of overfitting, we will soon see that the VAE is nonetheless able to self-regularize in a very precise sense: Global minimizers of the VAE objective will ultimately correspond with optimal solutions to

\[
\min_{L, S} \ n \cdot \text{rank} \left[ L \right] + \| S \|_{0}, \quad \text{s.t.} \ X = L + S,
\]

where \( \| \cdot \|_{0} \) denotes the \( \ell_{0} \) norm. This problem represents the canonical form of robust principal component analysis (RPCA) [3], [4], decomposing a data matrix \( X \) into low-rank principal factors \( L = UV \), with \( U \) and \( V \) low-rank matrices of appropriate dimension, and a sparse outlier component \( S \). It is seemingly quite remarkable that the probabilistic VAE model shares any kinship with (2), even more so given that some of the distracting local minimizers can be smoothed away, a key VAE advantage as we argue below.

Before elucidating this relationship, we require one additional technical caveat. Specifically, since \( \log 0 \) and \( \frac{1}{0} \) are both undefined, and yet we will soon require an alliance with degenerate (or nearly so) covariance matrices that mimic the behavior of sparse and low-rank factors through log-det and inverse terms concealed within (1), we must place the mildest of restrictions on the minimal allowable singular values of \( \Sigma_{x} \) and \( \Sigma_{z} \). For this purpose we define \( S_{\alpha}^{\infty} \) as the set of \( m \times m \) covariance matrices with singular values all greater than or equal to \( \alpha \), and likewise \( S_{\alpha}^{m} \) as the subset of \( S_{\alpha}^{\infty} \) containing only diagonal matrices. We also define \( \text{supp}_{\alpha} \left( x \right) = \{ i : |x_{i}| > \alpha \} \), noting that per this definition, \( \text{supp}_{\alpha} \left( x \right) = \text{supp}(x) \), meaning we recover the standard definition of support: the set of indices associated with nonzero elements.

Given the affine assumption from above, and the mild, oft-used restriction \( \Sigma_{x} \in S_{\alpha}^{\infty} \) and \( \Sigma_{z} \in S_{\alpha}^{m} \) for some small \( \alpha > 0 \), the resulting constrained minimization of (1) can be expressed as

\[
\min_{\theta, \phi} \mathcal{L} \left( W, b = 0, \Sigma_{x} \in S_{\alpha}^{\infty}, \mu_{x}, \Sigma_{z} \in S_{\alpha}^{m} \right),
\]

where now \( \theta \) includes \( W \) as well as all the parameters embedded in \( \Sigma_{x} \), while \( \mu_{x} \) and \( \Sigma_{z} \) are parameterized via \( \phi \). We have also set \( b = 0 \) merely for ease of presentation as its role is minor. And finally, for convenience we define \( \Sigma_{x}^{(i)} \) as the value of the decoder covariance function \( \Sigma_{x} \) evaluated at \( \mu_{x}^{(i)} \), where \( \mu_{x}^{(i)} \) denotes the encoder mean function \( \mu_{x} \) evaluated at \( x^{(i)} \). We then have the following:

**Theorem 1:** Suppose that \( X = \{x^{(i)}\}_{i=1}^{n} \) admits a feasible decomposition \( X = UV + S \) that uniquely optimizes (2). Then for some \( \alpha \) sufficiently small, and all \( \alpha \in (0, \bar{\alpha}) \), any global minimum \( \{ W, \Sigma_{x}, \mu_{x}, \Sigma_{z} \} \) of (3) will be such that

\[
\text{span}[W] = \text{span}[U] \quad \text{and} \quad \text{supp}_{\alpha} \left( \Sigma_{x}^{(i)} \right) = \text{supp}[s^{(i)}]
\]

for all \( i \) provided that the latent representation satisfies \( \kappa \geq \text{rank} \left[ U \right] \).

From this result, it is trivial to extract the optimal solution to (2) using any global minimizer of the VAE objective. Moreover, as we have recently shown in detail elsewhere [5], [6], the VAE maintains intrinsic regulatory mechanisms for pruning unnecessary latent dimensions (i.e., when \( \kappa \) is larger than needed) and for smoothing away suboptimal, distracting local minimizers.

Additionally, when we relax the affine decoder mean assumption, the VAE is also capable of generalizing RPCA to handle arbitrary nonlinear manifolds that serve as a complex inlier replacement for \( L \). We empirically demonstrate this capability in Figures 1 and 2 below.
Fig. 1. Results recovering ground-truth low-dimensional non-linear manifolds across different outlier ratios (x-axis) and manifold dimensions (y-axis) for (a) the VAE, and (b) Convex relaxation of RPCA, namely, the standard nuclear norm plus $\ell_1$-norm decomposition [3], [4]. In all cases, white color indicates normalized MSE near 0.0, while dark blue represents 1.0 or failure. The VAE is dramatically superior. Not surprisingly, convex RPCA fails because it cannot accurately capture the underlying nonlinear manifold using a linear subspace inlier model. Note also that RPCA performance is mostly independent of the manifold dimension because the generated data is nearly full rank across all cases. Of course with zero noise/outliers even a linear subspace model is trivially capable of good reconstructions.

Fig. 2. Analogous results to Figure 1, but using deterministic autoencoder models (AE) involving the exact same inlier capacity as the VAE. (a) An AE with an $\ell_2$-norm regularized latent space and typical outlier robust data fitting term; this is labeled as AE-$\ell_2$. (b) Same as AE-$\ell_2$ but with an $\ell_1$-norm latent space penalty; this is labeled as AE-$\ell_1$. In both cases, the AE performs poorly relative to the stochastic VAE results from Figure 1. Unlike the RPCA model, both AE-$\ell_1$ and AE-$\ell_2$ have the exact same inlier model capacity as the VAE and can in principle represent uncorrupted points perfectly; however, they have inferior agency for pruning superfluous latent dimensions, discarding outliers, or generally avoiding bad locally-optimal solutions.

REFERENCES