Hidden Talents of the Variational Autoencoder

Bin Dai   DAIB13@MAILS.TSINGHUA.EDU.CN
Institute for Advanced Study
Tsinghua University
Beijing, China

Yu Wang   YW323@CAM.AC.UK
Department of Pure Mathematics and Mathematical Statistics
University of Cambridge
Cambridge, UK

John Aston   J.ASTON@STATSLAB.CAM.AC.UK
Department of Pure Mathematics and Mathematical Statistics
University of Cambridge
Cambridge, UK

Gang Hua   GANGHUA@MICROSOFT.COM
Microsoft Research
Redmond, USA

David Wipf   DAVIDWIP@MICROSOFT.COM
Microsoft Research
Beijing, China

Abstract

Variational autoencoders (VAE) represent a popular, flexible form of deep generative model that can be stochastically fit to samples from a given random process using an information-theoretic variational bound on the true underlying distribution. Once so-obtained, the model can be putatively used to generate new samples from this distribution, or to provide a low-dimensional latent representation of existing samples. While quite effective in numerous application domains, certain important mechanisms which govern the behavior of the VAE are obfuscated by the intractable integrals and resulting stochastic approximations involved. Moreover, as a highly non-convex model, it remains unclear exactly how minima of the underlying energy relate to original design purposes. We attempt to better quantify these issues by analyzing a series of tractable special cases of increasing complexity. In doing so, we unveil interesting connections with more traditional dimensionality reduction models, as well as an intrinsic yet underappreciated propensity for robustly dismissing sparse outliers when estimating latent manifolds. With respect to the latter, we demonstrate that the VAE can be viewed as the natural evolution of recent robust PCA models, capable of learning nonlinear manifolds of unknown dimension obscured by gross corruptions.

Keywords: Variational Autoencoder, Deep Generative Model, Robust PCA

1. Introduction

We begin with a dataset $\boldsymbol{X} = \{\boldsymbol{x}^{(i)}\}_{i=1}^n$ composed of n i.i.d. samples of some random variable $\boldsymbol{x} \in \mathbb{R}^d$ of interest, with the goal of estimating a tractable approximation for $p_{\boldsymbol{\theta}}(\boldsymbol{x})$, knowledge of which would allow us to generate new samples of $\boldsymbol{x}$. Moreover we

assume that each sample is governed by unobserved latent variables $\boldsymbol{z} \in \mathbb{R}^\kappa$, such that $p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$, where $\boldsymbol{\theta}$ are the parameters defining the distribution we would like to estimate.

Given that this integral is intractable in all but the simplest cases, variational autoencoders (VAE) represent a powerful means of optimizing with respect to $\boldsymbol{\theta}$ a tractable upper bound on $-\log p_{\boldsymbol{\theta}}(\boldsymbol{x})$ [23, 36]. Once these parameters are obtained, we can then generate new samples from $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ by first drawing some $\boldsymbol{z}^{(i)}$ from $p(\boldsymbol{z})$, and then a new $\boldsymbol{x}^{(i)}$ from $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}^{(i)})$. The VAE upper bound itself is constructed as

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) \;=\; \sum_i \left\{ \mathbb{KL}\left[q_{\boldsymbol{\phi}}\left(\boldsymbol{z}|\boldsymbol{x}^{(i)}\right) ||p_{\boldsymbol{\theta}}\left(\boldsymbol{z}|\boldsymbol{x}^{(i)}\right)\right] - \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) \right\} \;\geq\; -\sum_i \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}), \quad (1)$$

where $q_{\boldsymbol{\phi}}\left(\boldsymbol{z}|\boldsymbol{x}^{(i)}\right)$ defines an arbitrary approximating distribution, parameterized by $\boldsymbol{\phi}$, and $\mathbb{KL}\left[\cdot||\cdot\right]$ denotes the KL divergence between two distributions, which is always a non-negative quantity. For optimization purposes, it is often convenient to re-express this bound as

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) \;\equiv\; \sum_i \left(\mathbb{KL}\left[q_{\boldsymbol{\phi}}\left(\boldsymbol{z}|\boldsymbol{x}^{(i)}\right) ||p(\boldsymbol{z})\right] - \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}^{(i)})}\left[\log p_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}|\boldsymbol{z}\right)\right] \right). \quad (2)$$

In these expressions, $q_{\boldsymbol{\phi}}\left(\boldsymbol{z}|\boldsymbol{x}\right)$ can be viewed as an *encoder* model that defines a conditional distribution over the latent 'code' $\boldsymbol{z}$, while $p_{\boldsymbol{\theta}}\left(\boldsymbol{x}|\boldsymbol{z}\right)$ can be interpreted as a *decoder* model since, given a code $\boldsymbol{z}$ it quantifies the distribution over $\boldsymbol{x}$.

By far the most common distributional assumptions are that $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I})$ and the encoder model satisfies $q_{\boldsymbol{\phi}}\left(\boldsymbol{z}|\boldsymbol{x}\right) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$, where the mean $\boldsymbol{\mu}_z$ and covariance $\boldsymbol{\Sigma}_z$ are some function of model parameters $\boldsymbol{\phi}$ and the random variable $\boldsymbol{x}$. Likewise, for the decoder model we assume $p_{\boldsymbol{\theta}}\left(\boldsymbol{x}|\boldsymbol{z}\right) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ for continuous data, with means and covariances defined analogously.[1]

For arbitrarily parameterized moments $\boldsymbol{\mu}_z$, $\boldsymbol{\Sigma}_z$, $\boldsymbol{\mu}_x$, and $\boldsymbol{\Sigma}_x$, the KL divergence in (2) computes to

$$2\mathbb{KL}\left[q_{\boldsymbol{\phi}}\left(\boldsymbol{z}|\boldsymbol{x}\right) ||p(\boldsymbol{z})\right] \;\equiv\; \mathrm{tr}\left[\boldsymbol{\Sigma}_z\right] + \|\boldsymbol{\mu}_z\|_2^2 - \log|\boldsymbol{\Sigma}_z|, \quad (3)$$

excluding irrelevant constants. However, the remaining integral from the expectation term admits no closed-form solution, making direct optimization over $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ intractable. Likewise, any detailed analysis of the underlying objective function becomes problematic as well.

At least for practical purposes, one way around this is to replace the troublesome expectation with a Monte Carlo stochastic approximation [23, 36]. More specifically we utilize

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}^{(i)})}\left[\log p_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}|\boldsymbol{z}\right)\right] \;\approx\; \tfrac{1}{\tau}\sum_{t=1}^{\tau} \log p_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}|\boldsymbol{z}^{(i,t)}\right), \quad (4)$$

where $\boldsymbol{z}^{(i,t)}$ are samples drawn from $q_{\boldsymbol{\phi}}\left(\boldsymbol{z}|\boldsymbol{x}^{(i)}\right)$. Using a simple reparameterization trick, these samples can be constructed such that gradients with respect to $\boldsymbol{\mu}_z$ and $\boldsymbol{\Sigma}_z$ can be propagated through the righthand side of (4). Therefore, assuming all the required moments $\boldsymbol{\mu}_z$, $\boldsymbol{\Sigma}_z$, $\boldsymbol{\mu}_x$, and $\boldsymbol{\Sigma}_x$ are differentiable with respect to $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, the entire model can be updated using SGD [3].

---

1. For discrete data, a Bernoulli distribution is sometimes adopted instead.

While quite effective in numerous application domains that can apply generative models, e.g., semi-supervised learning [22, 30, 31], certain important mechanisms which dictate the behavior of the VAE are obfuscated by the required stochastic approximation and the opaque underlying objective with high-dimensional integrals. Moreover, it remains unclear to what extent minima remain anchored at desirable locations in the non-convex energy landscape.

We take a step towards better quantifying such issues by probing the basic VAE model under a few simplifying assumptions of increasing complexity whereby closed-form integrations are (partially) possible. This process unveils a number of interesting connections with more transparent, established generative models, each of which shed light on how the VAE may perform under more challenging conditions. This mirrors the rich tradition of analyzing deep networks under various simplifications such as linear layers or i.i.d. random activation patterns [9, 10, 16, 20, 37], and results in the following key contributions:

1. We demonstrate that the canonical form of the VAE, including the Gaussian distributional assumptions described above, harbors an innate agency for robust outlier removal in the context of learning inlier points constrained to a manifold of unknown dimension. In fact, when the decoder mean $\boldsymbol{\mu}_x$ is restricted to an affine function of $\boldsymbol{z}$, we prove that the VAE model collapses to a form of robust PCA (RPCA) [6, 7], a recently celebrated technique for separating data into low-rank (low-dimensional) inlier and sparse outlier components.[2]

2. We elucidate two central, albeit underappreciated roles of the VAE encoder covariance $\boldsymbol{\Sigma}_z$. First, through subtle multi-tasking efforts in both terms of (2), it facilitates learning the correct inlier manifold dimension. Secondly, $\boldsymbol{\Sigma}_z$ can help to smooth out undesirable minima in the energy landscape of what would otherwise resemble a more traditional deterministic autoencoder (AE) [2]. This is true even in certain situations where it provably does not actually alter the globally optimal solution itself. Note that prior to this work the AE could ostensibly be viewed as the most natural candidate for instantiating extensions of RPCA to handle outlier-robust nonlinear manifold learning. However, our results suggest that the VAE maintains pivotal advantages in mitigating the effects of bad local solutions and over-parameterized latent representations, even in completely deterministic settings that require no generative model per se.

As we will soon see, these points can have profound practical repercussions in terms of how VAE models are interpreted and deployed. For example, one immediate consequence is that even if the decoder capacity is not sufficient to capture the generative distribution *within* some fixed, unknown manifold, the VAE can nonetheless still often find the correct manifold itself, which is sufficient for deterministic recovery of uncorrupted inlier points. This is exactly analogous to RPCA recovery results, whereby it is possible to correctly estimate an unknown low-dimensional linear subspace heavily corrupted with outliers even if in doing so we do not obtain an actual generative model for the inliers within this subspace. We emphasize that this is *not* a job description for which the VAE was originally motivated, but a useful hidden talent nonetheless.

---

2. RPCA represents a rather dramatic departure from vanilla PCA and is characterized by a challenging, combinatorial optimization problem. A formal definition will be provided in Section 3.

The remainder of this paper is organized as follows. In Section 2 we consider two affine decoder models and connections with past probabilistic PCA-like approaches. Note that the seminal work from [36] mentions in passing that a special case of their VAE decoder model reduces to factor analysis [1], a cousin of probabilistic PCA; however, no rigorous, complementary analysis is provided, such as how latent-space sparsity can emerge as we will introduce shortly. Next we examine various partially affine decoder models in Section 3, whereby only the mean $\boldsymbol{\mu}_x$ is affine while $\boldsymbol{\Sigma}_x$ has potentially unlimited complexity; all encoder quantities are likewise unconstrained. We precisely characterize how minimizers of the VAE cost, although not available in closed form, nonetheless are capable of optimally decomposing data into low-rank and sparse factors akin to RPCA while avoiding bad local optima. This section also discusses extensions as well as interesting behavioral properties of the VAE.

Section 4 then considers degeneracies in the full VAE model that can arise even with a trivially simple encoder and corresponding latent representation. Section 5 concludes with experiments that directly corroborate a number of interesting, practically-relevant hypotheses generated by our theoretical analyses, suggesting novel usages (unrelated to generating samples) as a tool for deterministic manifold learning in the presence of outliers. We provide final conclusions in Section 6. Note that our prior conference paper has presented the basic demonstration that VAE models can be applied to tackling generalized robust PCA problems [40]. However this work primarily considers empirical demonstrations and high-level motivations, with minimal analytical support.

Notation: We use a superscript $^{(i)}$ to denote quantities associated with the i-th sample, which at times may correspond with the columns of a matrix, such as the data $\boldsymbol{X}$ or related. For a general matrix $\boldsymbol{M}$, we refer to the i-th row as $\boldsymbol{m}_{i\cdot}$ and the j-th column as $\boldsymbol{m}_{\cdot j}$. Although technically speaking posterior moments are functions of the parameters $\{\boldsymbol{\theta}, \boldsymbol{\phi}\}$, the random variables $\boldsymbol{x}$, and the latent $\boldsymbol{z}$, i.e., $\boldsymbol{\mu}_x \equiv \boldsymbol{\mu}_x(\boldsymbol{z}; \boldsymbol{\theta})$, $\boldsymbol{\Sigma}_x \equiv \boldsymbol{\Sigma}_x(\boldsymbol{z}; \boldsymbol{\theta})$, $\boldsymbol{\mu}_z \equiv \boldsymbol{\mu}_z(\boldsymbol{x}; \boldsymbol{\phi})$, and $\boldsymbol{\Sigma}_z \equiv \boldsymbol{\Sigma}_z(\boldsymbol{x}; \boldsymbol{\phi})$, except in cases where some ambiguity exists regarding the arguments, these dependencies are omitted to avoid undue clutter; likewise for $\boldsymbol{\mu}_z^{(i)} \triangleq \boldsymbol{\mu}_z(\boldsymbol{x}^{(i)}; \boldsymbol{\phi})$ and $\boldsymbol{\Sigma}_z^{(i)} \triangleq \boldsymbol{\Sigma}_z(\boldsymbol{x}^{(i)}; \boldsymbol{\phi})$. Also, with some abuse of notation, we will use $\mathcal{L}$ to denote a number of different VAE-related objective functions and bounds, with varying arguments and context serving as differentiating factors. Finally, the diag$[\cdot]$ operator converts vectors to a diagonal matrix, and vice versa as in the Matlab computing environment.

2. Affine Decoder and Probabilistic PCA

If we assume that $\boldsymbol{\Sigma}_x$ is fixed at some $\lambda \boldsymbol{I}$, and force $\boldsymbol{\Sigma}_z = \boldsymbol{0}$ (while removing the now undefined $\log|\boldsymbol{\Sigma}_z|$ term), then it is readily apparent that the resultant VAE model reduces to a traditional AE with squared-error loss function [2], a common practical assumption. To see this, note that if $\boldsymbol{\Sigma}_z = \boldsymbol{0}$, then $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}^{(i)})$ collapses to $\delta(\boldsymbol{\mu}_z)$, i.e., a delta function at the posterior mean, and $\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}^{(i)})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z})\right] = \log p_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}|\boldsymbol{\mu}_z^{(i)}\right)$, which is just a standard AE with quadratic loss and representation $\boldsymbol{\mu}_x(\boldsymbol{\mu}_z[\boldsymbol{x}])$. Moreover, the only remaining (non-constant) regularization from the KL term is $\sum_i \|\boldsymbol{\mu}_z^{(i)}\|_2^2$. However, given scaling ambiguities that may arise in the decoder when $\boldsymbol{\Sigma}_z = \boldsymbol{0}$, $\boldsymbol{\mu}_z^{(i)}$ can often be made arbitrarily small, and therefore the effect of this quadratic penalty is infinitesimal. With

4

affine encoder and decoder models, the resulting deterministic network will simply learn principal components like vanilla PCA, a well-known special case of the AE [4].

Therefore to understand the VAE, it is crucial to explore the role of non-trivial selections for the encoder and decoder covariances, that serve as both enlightening and differentiating factors. As a step in this direction, we will explore several VAE reductions that lead to more manageable (yet still representative) objective functions and strong connections to existing probabilistic models. In this section we begin with the following simplification:

Lemma 1 *Suppose that the decoder moments satisfy $\boldsymbol{\mu}_x = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{b}$ and $\boldsymbol{\Sigma}_x = \lambda\boldsymbol{I}$ for some parameters $\boldsymbol{\theta} = \{\boldsymbol{W}, \boldsymbol{b}, \lambda\}$ of appropriate dimensions. Furthermore, we assume for the encoder we have $\boldsymbol{\mu}_z = f(\boldsymbol{x}; \boldsymbol{\phi})$, $\boldsymbol{\Sigma}_z = \boldsymbol{S}_z\boldsymbol{S}_z^\top$, and $\boldsymbol{S}_z = g(\boldsymbol{x}; \boldsymbol{\phi})$, where f and g are any parameterized functional forms that include arbitrary affine transformations for some arrangement of parameters. Under these assumptions, the objective from (2) admits optimal, closed-form solutions for $\boldsymbol{\mu}_z$ and $\boldsymbol{\Sigma}_z$ in terms of $\boldsymbol{W}$, $\boldsymbol{b}$, and λ such that the resulting VAE cost collapses to*

$$\mathcal{L}(\boldsymbol{W}, \boldsymbol{b}, \lambda) \;=\; \sum_i \boldsymbol{\Omega}^{(i)}(\boldsymbol{W}, \boldsymbol{b}, \lambda\boldsymbol{I}) + n \log\left|\lambda\boldsymbol{I} + \boldsymbol{W}\boldsymbol{W}^\top\right|, \tag{5}$$

where

$$\boldsymbol{\Omega}^{(i)}(\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{\Psi}) \;\triangleq\; \left(\boldsymbol{x}^{(i)} - \boldsymbol{b}\right)^\top \left(\boldsymbol{\Psi} + \boldsymbol{W}\boldsymbol{W}^\top\right)^{-1} \left(\boldsymbol{x}^{(i)} - \boldsymbol{b}\right). \tag{6}$$

Additionally, if we enforce that off-diagonal elements of $\boldsymbol{\Sigma}_z$ must be equal to zero (i.e., $[\boldsymbol{\Sigma}_z]_{ij} = 0$ for $i \neq j$), then (5) further decouples/separates to

$$\mathcal{L}_{sep}(\boldsymbol{W}, \boldsymbol{b}, \lambda) \;=\; \sum_i \boldsymbol{\Omega}^{(i)}(\boldsymbol{W}, \boldsymbol{b}, \lambda\boldsymbol{I}) + n\left[\sum_j \log\left(\lambda + \|\boldsymbol{w}_{\cdot j}\|_2^2\right) + (d - \kappa)\log\lambda\right]. \tag{7}$$

All proofs are deferred to the appendices. The objective (5) is the same as that used by certain probabilistic PCA models [39], even though the latter is originally derived in a completely different manner. Moreover, it can be shown that any minimum of this objective represents a globally optimal solution (i.e, no minima with suboptimal objective function value exist). And with $\boldsymbol{b}$ and λ fixed, the optimal $\boldsymbol{W}$ will be such that span$[\boldsymbol{W}]$ equals the span of the singular vectors of $\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^\top$ associated with singular values greater than $\sqrt{\lambda}$. So the global optimum produces a principal subspace formed by soft-thresholding the singular values of $\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^\top$, with the rank one offset often used to normalize samples to have zero mean.[3]

In contrast, the alternative cost (7), which arises from the oft-used practical assumption that $\boldsymbol{\Sigma}_z$ is diagonal, represents a rigorous upper bound to (5), since

$$\sum_j \log\left(\lambda + \|\boldsymbol{w}_{\cdot j}\|_2^2\right) + (d - \kappa)\log\lambda \;\geq\; \log\left|\lambda\boldsymbol{I} + \boldsymbol{W}\boldsymbol{W}^\top\right| \tag{8}$$

---

3. While the details are omitted here, optimal solutions for both $\boldsymbol{b}$ and λ can be analyzed as well.

by virtue of Hadamard's inequality (see proof of Theorem 2 below), with equality iff $\boldsymbol{W}^\top \boldsymbol{W}$ is diagonal. Interestingly, all minima of the modified cost nonetheless retain global optimality of the original; however, it can be shown that there will be a combinatorial increase in the actual number of distinct (disconnected) minima:[4]

Theorem 2 *Let $\boldsymbol{R} \in \mathbb{R}^{\kappa \times \kappa}$ denote an arbitrary rotation matrix and $\boldsymbol{P} \in \mathbb{R}^{\kappa \times \kappa}$ an arbitrary permutation matrix. Furthermore let $\boldsymbol{W}^*$ be a minimum of (5) and $\boldsymbol{W}^{**}$ any minimum of (7) with $\boldsymbol{b}$ and λ fixed. Then the following three properties hold:*

1. *$\mathcal{L}(\boldsymbol{W}^*, \boldsymbol{b}, \lambda) = \mathcal{L}(\boldsymbol{W}^* \boldsymbol{R}, \boldsymbol{b}, \lambda) = \mathcal{L}_{sep}(\boldsymbol{W}^{**}, \boldsymbol{b}, \lambda)$*

$$= \mathcal{L}(\boldsymbol{W}^{**} \boldsymbol{P}, \boldsymbol{b}, \lambda) = \mathcal{L}_{sep}(\boldsymbol{W}^{**} \boldsymbol{P}, \boldsymbol{b}, \lambda). \tag{9}$$

2. *For any $\boldsymbol{W}^{**}(\boldsymbol{W}^{**})^\top$ with distinct nonzero eigenvalues, there will exist at least $\frac{\kappa!}{(\kappa-r)!}$ distinct (disconnected) minima of (7) located at some $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{P}$, where $\boldsymbol{U}\boldsymbol{\Lambda}^2\boldsymbol{U}^\top$ represents the SVD of $\boldsymbol{W}^{**}(\boldsymbol{W}^{**})^\top$ and $r = rank[\boldsymbol{W}^{**}]$.*

3. *$\boldsymbol{W}^{**}$ will have at most r nonzero columns, while $\boldsymbol{W}^*$ can have any number in $\{r, \ldots, \kappa\}$.*

Although this result applies to relatively simplistic affine decoders (the encoder need not be so constrained however), it nonetheless highlights a couple interesting principles. First, the diagonalization of $\boldsymbol{\Sigma}_z$ collapses the space of globally minimizing solutions to a subset of the original. While the consequences of this may be minor in the fully affine decoder model where all the solutions are still equally good, we surmise that with more sophisticated parameterizations this partitioning of the energy landscape into distinct basins-of-attraction could potentially introduce suboptimal local extrema. And from a broader perspective, Theorem 2 provides tangible validation of prior conjectures that variational Bayesian factorizations of this sort can fragment the space of local minima [19].

But there is a second, potentially-advantageous counter-affect elucidated by Theorem 2 as well. Specifically, even if $\boldsymbol{W}$ is overparameterized, meaning that κ is unnecessarily large, there exists an inherent mechanism to prune superfluous columns to exactly zero, i.e., column-wise sparsity. And once columns of $\boldsymbol{W}$ become sparse, the corresponding elements of $\boldsymbol{\mu}_z$ can no longer influence the data fit. Consequently, the $\|\boldsymbol{\mu}_z\|_2^2$ factor from (3) serves as the only relevant influence, pushing these values to be exactly zero even though ℓ_2 norms in most regularization contexts tend to favor diverse, *non-sparse* representations [34].

So ultimately, sparsity of $\boldsymbol{\mu}_z$ is an artifact of the diagonal $\boldsymbol{\Sigma}_z$ assumption and the interaction of multiple VAE terms, a subtle influence we empirically demonstrate translates to more complex regimes in Section 5. In any event, we have shown that both variants of the affine decoder model lead to reasonable probabilistic PCA-like objectives regardless of how overparameterized $\boldsymbol{\mu}_z$ and $\boldsymbol{\Sigma}_z$ happen to be.

3. Partially Affine Decoder and Robust PCA

Thus far we have considered tight limitations on the complexity allowable in the functional forms of both $\boldsymbol{\mu}_x$ and $\boldsymbol{\Sigma}_x$, while $\boldsymbol{\mu}_z$ and $\boldsymbol{\Sigma}_z$ were free-range variables granted arbitrary flexibility. We now turn our gaze to the case where $\boldsymbol{\Sigma}_x$ can also be any parameterized, diagonal

---

4. By disconnected we mean that, to traverse from one minimum to another, we must ascend the objective function at some point along the way.

matrix[5] while $\boldsymbol{\mu}_x$ remains restricted. Although 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

$$\min_{\boldsymbol{L},\boldsymbol{S}} \quad n \cdot \text{rank}\,[\boldsymbol{L}] + \|\boldsymbol{S}\|_0, \quad \text{s.t. } \boldsymbol{X} = \boldsymbol{L} + \boldsymbol{S}, \tag{10}$$

where $\|\cdot\|_0$ denotes the ℓ_0 norm, or a count of the number of nonzero elements in a vector or matrix. This problem represents the canonical form of robust principal component analysis (RPCA) [6, 7], decomposing a data matrix $\boldsymbol{X}$ into low-rank principal factors $\boldsymbol{L} = \boldsymbol{UV}$, with $\boldsymbol{U}$ and $\boldsymbol{V}$ low-rank matrices of appropriate dimension, and a sparse outlier component $\boldsymbol{S}$. However, we must emphasize that (10), unlike traditional PCA, represents an NP-hard, discontinuous optimization problem with a combinatorial number of potentially bad local minima. Still, it is seemingly quite remarkable that the probabilistic VAE model shares any kinship with (10), even more so given that some of the distracting *local* minimizers can be smoothed away, a key VAE advantage as we will later argue.

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, we must place the mildest of restrictions on the minimal allowable singular values of $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_z$. For this purpose we define $\mathcal{S}_\alpha^m$ as the set of $m \times m$ covariance matrices with singular values all greater than or equal to α, and likewise $\bar{\mathcal{S}}_\alpha^m$ as the subset of $\mathcal{S}_\alpha^m$ containing only diagonal matrices. We also define $\text{supp}_\alpha(\boldsymbol{x}) = \{i : |x_i| > \alpha\}$, noting that per this definition, $\text{supp}_0(\boldsymbol{x}) = \text{supp}(\boldsymbol{x})$, meaning we recover the standard definition of support: the set of indices associated with nonzero elements.

3.1 Main Result and Interpretation

Given the affine assumption from above, and the mild restriction $\boldsymbol{\Sigma}_x \in \bar{\mathcal{S}}_\alpha^d$ and $\boldsymbol{\Sigma}_z \in \mathcal{S}_\alpha^\kappa$ for some small $\alpha > 0$, the resulting constrained VAE minimization problem can be expressed as

$$\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \quad \mathcal{L}\left(\boldsymbol{W}, \boldsymbol{b} = \boldsymbol{0}, \boldsymbol{\Sigma}_x \in \bar{\mathcal{S}}_\alpha^d, \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z \in \mathcal{S}_\alpha^\kappa\right), \tag{11}$$

where now $\boldsymbol{\theta}$ includes $\boldsymbol{W}$ as well as all the parameters embedded in $\boldsymbol{\Sigma}_x$, while $\boldsymbol{\mu}_z$ and $\boldsymbol{\Sigma}_z$ are parameterized as in Lemma 1. We have also set $\boldsymbol{b} = \boldsymbol{0}$ merely for ease of presentation as its role is minor. We then have the following:

Theorem 3 *Suppose that $\boldsymbol{X} = \{\boldsymbol{x}^{(i)}\}_{i=1}^n$ admits a feasible decomposition $\boldsymbol{X} = \boldsymbol{UV} + \boldsymbol{S}$ that uniquely[6] optimizes (10). Then for some $\bar{\alpha}$ sufficiently small, and all $\alpha \in (0, \bar{\alpha}]$, any*

---

5. A full covariance over $\boldsymbol{x}$ is infeasible given the high dimension, and can lead to undesirable degeneracies anyway. Therefore a diagonal covariance is typically, if not always, used in practice.
6. Obviously only $\boldsymbol{L}$ and $\boldsymbol{S}$ will be unique; the actual decomposition of $\boldsymbol{L}$ into $\boldsymbol{U}$ and $\boldsymbol{V}$ is indeterminate up to an inconsequential invertible transform.

global minimum $\{\hat{W}, \hat{\Sigma}_x, \hat{\mu}_z, \hat{\Sigma}_z\}$ of (11) will be such that[7]

$$span[\hat{W}] = span[U] \qquad and \qquad supp_\alpha \left(diag \left[\hat{\Sigma}_x \left(\hat{\mu}_z \left[x^{(i)} \right] \right) \right] \right) = supp[s^{(i)}] \qquad (12)$$

for all i provided that the latent representation satisfies $\kappa \geq rank[U]$.

Several important remarks are warranted here regarding the consequences and interpretation of this result:

- The $\hat{W}$ satisfying (12) forms a linear basis for each inlier component $l^{(i)}$, and likewise, a sample-dependent basis denoted $E^{(i)}$ can be trivially constructed for each outlier component $s^{(i)}$ using $\hat{\Sigma}_x$, and $\hat{\mu}_z$. Specifically, each unique column of $E^{(i)}$ is a vector of zeros with a one in the j-th position, with $j \in \text{supp}_\alpha \left(\text{diag} \left[\hat{\Sigma}_x \left(\hat{\mu}_z \left[x^{(i)} \right] \right) \right] \right)$. It follows that

$$x^{(i)} = l^{(i)} + s^{(i)} = \left[\hat{W} \ \ E^{(i)} \right] \left[\hat{W} \ \ E^{(i)} \right]^\dagger x^{(i)}, \quad \forall i = 1, \ldots, n. \qquad (13)$$

  Therefore if we can globally optimize the VAE objective, we can recover the correct latent representation, or equivalently, the optimal solution to (10).

- The requirements $\Sigma_x \in \bar{\mathcal{S}}_\alpha^d$ and $\Sigma_z \in \mathcal{S}_\alpha^\kappa$ do not portend the need for specialized tuning or brittleness of the result; these are merely technical conditions for dealing with degenerate covariances that occur near optimal solutions. While it might seem natural that Σ_x has diagonal elements pushed to zero in regions where near perfect data fit is possible, less intuitively, global optima of (11) can be achieved with an arbitrarily small Σ_z, e.g., $\Sigma_z = \alpha I$, at least along latent dimensions needed to represent L (see proof construction). And interestingly, this implies that in areas surrounding a global optimum, the VAE objective can *resemble that of a regular AE*. As we will discuss more below, desirable smoothing effects of integration over Σ_z occur *elsewhere* in the energy landscape while preserving extrema anchored at the correct latent representation.

- Even if κ is large, meaning W is possibly overcomplete, the VAE will not overfit in the sense that there exists an inherent regulatory effect pushing span$[W]$ towards span$[U]$.

- If the globally optimal solution to (10) is not unique (this is different from uniqueness regarding the VAE objective), then a low-rank-plus-sparse model may not be the most reasonable, parsimonious representation of the data to begin with, and exact recovery of L and S will not be possible by *any* algorithm without further assumptions. More concretely, an arbitrary data point $x^{(i)} \in \mathbb{R}^d$ requires d degrees of freedom to represent; however, if the data succinctly adheres to the RPCA model, then for properly chosen U, V, and S, we can have $x^{(i)} = Uv^{(i)} + s^{(i)}$, where $\|v^{(i)}\|_0 + \|s^{(i)}\|_0 < d$. Arbitrary data in general position will never admit such a unique decomposition, and we should only expect such structure in data well-represented by our VAE model, or the original RPCA predecessor from (10).

---

7. Although somewhat cumbersome in print, the expression $\hat{\Sigma}_x \left(\hat{\mu}_z \left[x^{(i)} \right] \right)$ refers to $\hat{\Sigma}_x$ evaluated at $\hat{\mu}_z$, where the latter is evaluated at $x^{(i)}$, the i-th sample.

- A number of celebrated results have stipulated conditions [6, 7] whereby global solutions of the convex relaxation into nuclear and ℓ_1 norm components given by

$$\min_{\boldsymbol{L}, \boldsymbol{S}} \quad \sqrt{n} \cdot \mathrm{rank} \, \|\boldsymbol{L}\|_* + \|\boldsymbol{S}\|_1, \quad \text{s.t. } \boldsymbol{X} = \boldsymbol{L} + \boldsymbol{S}, \tag{14}$$

  will equal global solutions of (10). While elegant in theory, and practically relevant given that (10) is discontinuous, non-convex, and difficult to optimize, the required conditions for this equivalence to hold place strong restrictions on the allowable structure in $\boldsymbol{L}$ and support pattern in $\boldsymbol{S}$. In practice these conditions can never be verified and are unlikely to hold, so an alternative modeling approach such as the VAE, which can be viewed as a smoothed version of (10) when an affine decoder mean is used (more on this later), remains attractive. Additionally, there is no clear way to modify (14) to handle nonlinear manifolds, which is obviously the bread and butter of the VAE.

We emphasize that these conclusions are not the product of an overly contrived situation, given that a significant restriction is only placed on $\boldsymbol{\mu}_x$; *all* other posterior quantities are essentially unconstrained provided a sufficient lower complexity bound is exceeded, implying that the result will hold whenever a sufficiently complex deep network is used. Moreover, although we will defer to a formal treatment to future work for purposes of brevity here, with some mild additional conditions, Theorem 3 can naturally be extended to the case where the decoder mean function is generalized to subsume non-linear, union-of-subspace models as commonly assumed in subspace clustering problems [14, 35]. This then deviates substantially from any direct PCA-kinship, and buttresses the argument that the analysis presented here transitions to broader scenarios. The experiments from Section 5 will also provide complementary empirical confirmation.

Moving forward, as a point of further comparison it is also interesting to examine how a traditional AE, which emerges when $\boldsymbol{\Sigma}_z$ is forced to zero, behaves under analogous conditions to Theorem 3.

Corollary 4 *Under the same conditions as Theorem 3, if we remove the $\log |\boldsymbol{\Sigma}_z|$ term and assume $\boldsymbol{\Sigma}_z = \boldsymbol{0}$ elsewhere, then (11) admits a closed-form solution for $\boldsymbol{\Sigma}_x$ in terms of $\boldsymbol{W}$ and $\boldsymbol{\mu}_z$ such that minimizers of the VAE cost are minimizers of*

$$\mathcal{L} \left(\boldsymbol{W}, \boldsymbol{\mu}_z \right) = \sum_i \left\| \boldsymbol{x}^{(i)} - \boldsymbol{W} \boldsymbol{\mu}_z \left(\boldsymbol{x}^{(i)} \right) \right\|_0 \quad \text{in the limit } \alpha \to 0. \tag{15}$$

From this result we immediately observe that, provided $\boldsymbol{\mu}_z$ enjoys a sufficiently rich parameterization, minimization of (15) is just a constrained version of (10), exactly equivalent to solving

$$\min_{\boldsymbol{L}, \boldsymbol{S}} \quad \|\boldsymbol{S}\|_0, \quad \text{s.t. } \boldsymbol{X} = \boldsymbol{L} + \boldsymbol{S}, \quad \mathrm{rank} \, [\boldsymbol{L}] \leq \kappa. \tag{16}$$

This expression immediately exposes one weakness of the AE; namely, if κ is too large, there is no longer any operation in place to prune away unnecessary dimensions, and the trivial solution $\boldsymbol{L} = \boldsymbol{X}$ will be produced. In the large-κ regime then, global VAE and global AE solutions do in fact deviate, ultimately because of the removal of the $-\log |\boldsymbol{\Sigma}_z|$ term in the latter. So $\boldsymbol{\Sigma}_z$ plays a critical role in determining the correct, low-dimensional inlier structure, and ultimately it is this covariance that chaperons $\boldsymbol{W}$ during the learning process.

3.2 Additional Local Minima Smoothing Effects

There is also a more important, yet subtle, advantage of the VAE over both (16) and the original unconstrained RPCA model from (10). For both RPCA constructions, any feasible support pattern, even the trivial ones associated with non-interesting decompositions satisfying $\|\boldsymbol{v}^{(i)}\|_0 + \|\boldsymbol{s}^{(i)}\|_0 \geq d$ for some i, will necessarily represent a local minimum, since there is an infinite gradient to overcome to move from a zero-valued element of $\boldsymbol{S}$ to a nonzero one.

Unlike these deterministic approaches, the behavior of the VAE reflects a form of differential smoothing that rids the model of many of these pitfalls while retaining desirable minima that satisfy (12).[8] Based on details of the proof of Theorem 3, it can be shown that, excluding small-order terms dependent on other variables and a constant scale factor of $-\log \alpha$, then a representative bound on the VAE objective associated with each sample index i behaves like

$$\text{rank}[\boldsymbol{W}] + \text{supp}_\alpha \left(\text{diag} \left[\boldsymbol{\Sigma}_x \left(\boldsymbol{\mu}_z \left[\boldsymbol{x}^{(i)} \right] \right) \right] \right). \tag{17}$$

But crucially, *this behavior lasts only as long as (17) is strictly less than d* and $\boldsymbol{\Sigma}_z$ is forced to be small or degenerate. In contrast, when the value is at or above d, (17) no longer reflects the energy function, which becomes relatively flat because of smoothing via $\boldsymbol{\Sigma}_z$, avoiding the pitfalls described above. This phenomena then has the potential to smooth out a large constellation of bad locally optimal solutions.

To situate things in the narrative of (10), which is useful for illustration purposes, the VAE can be viewed (at least to first order approximation) as minimizing the alternative lower-bounding objective function

$$\sum_i \text{rank} \left[\boldsymbol{L}\boldsymbol{L}^\top + \text{diag} \left(\boldsymbol{s}^{(i)} \right)^2 \right] \leq \sum_i \text{rank} \left[\boldsymbol{L}\boldsymbol{L}^\top \right] + \sum_i \text{rank} \left[\text{diag} \left(\boldsymbol{s}^{(i)} \right)^2 \right]$$

$$= n \cdot \text{rank}[\boldsymbol{L}] + \|\boldsymbol{S}\|_0, \tag{18}$$

or a smooth surrogate thereof, over the constraint set $\boldsymbol{X} = \boldsymbol{L} + \boldsymbol{S}$. The advantages of this lower bound are substantial: As long as a unique solution exists to the RPCA problem, the globally optimal solution with $\|\boldsymbol{v}^{(i)}\|_0 + \|\boldsymbol{s}^{(i)}\|_0 < d$ for all i will be unchanged; however, *any* feasible solution with $\|\boldsymbol{v}^{(i)}\|_0 + \|\boldsymbol{s}^{(i)}\|_0 \geq d$ will have a constant cost via the expression on the left of the inequality, truncating the many erratic peaks that will necessarily occur with the energy on the righthand side.

In fact, away from the strongly attractive basins of optimal VAE solutions, the KL term from (2) is likely to push $\boldsymbol{\Sigma}_z$ more towards

$$\arg \min_{\boldsymbol{\Sigma}_z \succ \boldsymbol{0}} \mathbb{KL}\left[q_\phi \left(\boldsymbol{z} | \boldsymbol{x} \right) \| p(\boldsymbol{z}) \right] \equiv \arg \min_{\boldsymbol{\Sigma}_z \succ \boldsymbol{0}} \text{tr}\left[\boldsymbol{\Sigma}_z \right] - \log |\boldsymbol{\Sigma}_z| = \boldsymbol{I}. \tag{19}$$

Experiments presented in Section 5 confirm that this is indeed the case. And once $\boldsymbol{\Sigma}_z$ moves away from zero, it will generally contribute a strong smoothing effect via the expectation in (2). However, there exists an important previously unobserved caveat here: If the decoder

---

8. A more rudimentary form of this smoothing has been observed in much simpler empirical Bayesian models derived using Fenchel duality [41].

mean function is excessively complex, it can potentially outwit all regulatory persuasions from $\boldsymbol{\Sigma}_z$, leading to undesirable degenerate solutions with no representational value as described next.

4. Degeneracies Arising from a Flexible Decoder Mean

In this section we consider the case where $\boldsymbol{\mu}_x$ is finally released from its affine captivity to join with posterior colleagues in the wild. That simultaneously granting $\boldsymbol{\mu}_x$, $\boldsymbol{\Sigma}_x$, $\boldsymbol{\mu}_z$, and $\boldsymbol{\Sigma}_z$ unlimited freedom leads to overfitting may not come as a surprise; however, it turns out that even if the latter three are severely constrained, overfitting will not be avoided when $\boldsymbol{\mu}_x$ is over-parameterized in a certain sense extending beyond a single affine layer. This is because, at least at a high level, the once-proud regulatory effects of $\boldsymbol{\Sigma}_z$ can be completely squashed in these situations leading to the following:

Theorem 5 *Suppose $\kappa = 1$ (i.e., a latent dimension of only one), $\boldsymbol{\Sigma}_z \equiv \sigma_z^2 = \lambda_z$ (a scalar), $\mu_z = \boldsymbol{a}^\top \boldsymbol{x}$ for some fixed vector $\boldsymbol{a}$, $\boldsymbol{\Sigma}_x = \lambda_x \boldsymbol{I}$, and $\boldsymbol{\mu}_x$ is an arbitrary piecewise linear function with n segments. Then the VAE objective is unbounded from below at a trivial solution $\{\hat{\lambda}_z, \hat{\boldsymbol{a}}, \hat{\lambda}_x, \hat{\boldsymbol{\mu}}_x\}$ such that the resulting posterior mean $\hat{\boldsymbol{\mu}}_x(z; \boldsymbol{\theta})$ will satisfy $\hat{\boldsymbol{\mu}}_x(z; \boldsymbol{\theta}) \in \{\boldsymbol{x}^{(i)}\}_{i=1}^n$ with probability one for any z.*

In this special case, $\boldsymbol{\Sigma}_x$, σ_z^2, and μ_z are all simple affine functions and the latent dimension is minimal, and yet an essentially useless, degenerate solution can arbitrarily optimize the VAE objective. This occurs because the VAE has limited power to corral certain types of heavily over-parameterized decoder mean functions, even when all other degrees of freedom are constrained, and in this regime the VAE essentially has no advantage over a traditional autoencoder (its natural self-regulatory agency may sometimes break down). In contrast, as we saw in a previous section, there is no problem taming the influences of an unlimited latent representation (meaning κ is large, e.g., even $\kappa > n$) and its huge, attendant parameterized mean function, provided the latter is affine, as in $\boldsymbol{\mu}_x = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{b}$.

Indeed then, the issue is clearly not the *degree* of over-parameterization in $\boldsymbol{\mu}_x$ per se, but the actual structures in place. And the key problem is that, at least in some situations, the model can circumvent the entire regulatory mechanism of the KL term, pushing the latent variances towards zero even around *undesirable* solutions. For example, in the context of Theorem 5, the piecewise linear structure of $\boldsymbol{\mu}_x$ allows the decoder to act much like a vector quantization process, encouraging z towards a scalar code that selects for piecewise linear segments matched to training samples $\boldsymbol{x}^{(i)}$. And because this will lead to perfect reconstruction error if an optimal segment is found for a particular $z^{(i)}$, $\boldsymbol{\Sigma}_x = \lambda_x \boldsymbol{I} \approx \boldsymbol{0}$ serves as a reasonable characterization of posterior uncertainty, pushing $p(\boldsymbol{x}^{(i)}|z^{(i)}) \rightarrow \delta\left(\boldsymbol{x}^{(i)}\right)$ provided that $z^{(i)} \approx \mu_z\left(\boldsymbol{x}^{(i)}; \boldsymbol{a}\right) = \boldsymbol{a}^\top \boldsymbol{x}^{(i)}$, meaning that $\sigma_z^2 = \lambda_z$ is not too large.

In this situation, loosely speaking the data term from (2) will behave like $nd \log \lambda_x$, bullying the over-matched KL term that will scale only as $-n \log \lambda_z$. This in turn leads to a useless, degenerate solution as $\lambda_x = \lambda_z \rightarrow 0$, either for the purposes of generating representative samples, or for outlier removal as we have described herein.

One helpful caveat though, is that actually implementing such a complex piecewise linear function $\hat{\boldsymbol{\mu}}_x(z; \boldsymbol{\theta})$ using typical neural network components would require extremely wide and/or deep structure beyond the first decoder mean layer. And the degrees of freedom in

such higher-layer structures would need to scale proportionally with the size of the training data, which is not a practical VAE operational regime to begin with. In contrast, the first layer of the decoder mean network more or less self-regularizes, at least in the affine and related cases as described above. And we conjecture that this self-regularization preserves in more complex networks of reasonable practical size as will be empirically demonstrated in Section 5. So really it is *excessive* complexity in higher decoder mean layers, unrelated to the dimensionality of the latent z bottleneck, where overfitting problems are more likely to arise.

Of course an analogous issue exists with generative adversarial networks (GAN) as well, a popular competing deep generative model composed of a generator network analogous to the VAE decoder, and a discriminator network that replaces the VAE encoder in a loose sense [17]. If the generator network merely learns a segmentation of z-space such that all points in the i-th partition map to $x^{(i)}$, the discriminator will be helpless to avert this degenerate situation even in principle. But there is an asymmetry when it comes to the GAN discriminator network and the VAE encoder: Over-parameterization of the former can be problematic (e.g., it can easily out-wit an affine or other proportionally simple generator), but the latter not so, at least in the sense that a highly flexible VAE encoder need not bully a simple decoder into trivial solutions as we have shown in previous sections.

5. Experiments and Analysis

Theoretical examination of simplified cases can be viewed as a powerful vehicle for generating accessible hypotheses that describe likely behavior in more realistic, practical situations. In this section we empirically evaluate and analyze three concrete hypotheses that directly emanate from our previous technical results and the tight connections between RPCA and VAE models. In aggregate, these hypotheses have wide-ranging consequences in terms of how VAEs should be applied and interpreted.

Before stating these hypotheses, we summarize what can be viewed as two, theoretically-accessible boundary cases considered thus far. First, building on Section 2, Section 3 demonstrated that the VAE can self-regularize and produce useful, robust models provided that restrictions are placed on only the decoder mean network. Conversely, Section 4 demonstrated that, regardless of other model components, if the decoder mean network is unreasonably complex beyond the first layer, then overfitting emerges as a potential concern. But between these two extremes, there exists a large operational regime whereby practical VAE behavior is both worth exploring and likely still informed by the original analysis of these boundary cases.

Within this context then, we conjecture that the desirable VAE properties exposed in Sections 2 and 3 are inherited by models involving deeper decoder mean networks, but at least constrained to practically-sized hidden-layer $\boldsymbol{\mu}_x$ complexity such that the concerns from Section 4 are not a significant factor (e.g., no networks where the degrees of freedom in higher decoder mean layers scales as $d \times n$, an absurd VAE structure by any measure). More specifically, in this section will empirically examine the following three hypotheses:

(i) When the decoder mean function is allowed to have multiple hidden layers of sensible size/depth, the VAE should behave like a nonlinear extension of RPCA, but with natural regularization effects in place that help to avoid local minima and/or overfitting

to outliers. It is therefore likely to outperform either RPCA algorithms or, more importantly, an AE on diverse manifold recovery/outlier discovery problems unrelated to the probabilistic generative modeling tasks the VAE was originally designed for.

(ii) If the VAE latent representation $\boldsymbol{z}$ is larger than needed (meaning its dimension κ is higher than the true data manifold dimension), we have proven that unnecessary columns of $\boldsymbol{W}$ in a certain affine decoder mean model $\boldsymbol{\mu}_x = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{b}$ will automatically be pruned as desired. Analogously, in the extended nonlinear case we would then expect that columns of the weight matrix from the first layer of the decoder mean network should be pushed to zero, again effectively pruning away the impact of any superfluous elements of $\boldsymbol{z}$.

(iii) When granted sufficient capacity in both $\boldsymbol{\mu}_x\left(\boldsymbol{\mu}_z\left[\boldsymbol{x}\right]\right)$ and $\boldsymbol{\Sigma}_x$ to model inliers and outliers respectively, the VAE should have a tendency to push elements of the encoder covariance $\boldsymbol{\Sigma}_z$ to arbitrarily near zero along latent dimensions needed for representing inlier points, *selectively* overriding the KL regularizer that would otherwise push these values towards one. This counterintuitive behavior directly facilitates the VAE's utility as a nonlinear outlier removal tool (per Hypothesis (i)) by preserving exact adherence to the manifold in the neighborhood of optimal solutions.

5.1 Hypothesis (i) Evaluation Using Specially-Designed Ground-Truth Manifolds

If our theory is generally applicable, then a VAE with suitable parameterization should be able to significantly outperform an analogous deterministic AE (i.e., an equivalent VAE but with $\boldsymbol{\Sigma}_z = \boldsymbol{0}$) on the task of recovering data points drawn from a low-dimensional nonlinear manifold, but corrupted with gross outliers. In other words, even if both models have equivalent capacity to capture the intrinsic underlying manifold in principle, the VAE is more likely to avoid bad minima and correctly estimate it. We demonstrate this VAE capability here for the first time across an array of manifold dimensions and corruption percentages, recreating a nonlinear version of what are commonly termed *phase transition plots* in the vast RPCA literature [6, 11, 21, 41]. These plots evaluate the reconstruction quality of competing algorithms for every pairing of subspace dimension and outlier ratio, creating a heat map that differentiates success and failure regions.

Of course explicit knowledge of ground-truth low-dimensional manifolds is required to accomplish this. With linear subspaces it is trivial to generate appropriate synthetic data by simply creating two low-rank random matrices $\boldsymbol{U} \in \mathbb{R}^{d \times \kappa}$ and $\boldsymbol{V} \in \mathbb{R}^{\kappa \times n}$, a sparse outlier matrix $\boldsymbol{S}$, and then computing $\boldsymbol{X} = \boldsymbol{L} + \boldsymbol{S}$ with $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{V}$. Algorithms are presented with only $\boldsymbol{X}$ and attempt to reconstruct $\boldsymbol{L}$. Here we generalize this process to the nonlinear regime using deep networks and the following non-trivial steps. In this revised context, the generated $\boldsymbol{L}$ will now represent a data matrix with columns confined to a ground-truth nonlinear manifold.

Data Generation: First we draw n low-dimensional samples $\boldsymbol{z}^{(i)} \in \mathbb{R}^{\kappa}$ from $\mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I})$ and pass them through a 3-layer network with ReLU activations [32]. We express this structure as $\boldsymbol{z}(\kappa)\text{-}\boldsymbol{D}_1(r_1)\text{-}\boldsymbol{D}_2(r_2)\text{-}\boldsymbol{l}(d)$, where $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are hidden layers, $\boldsymbol{l}$ here serves as

the output layer, and the values inside parentheses denote the respective dimensionalities (these experiment-dependent values will be discussed later). Network weights are set using the initialization procedure from [18]. The d-dimensional output produced by $z^{(i)}$ is denoted as $l^{(i)}$, the collection of which form a matrix L, with columns effectively lying on a κ-dimensional nonlinear manifold. This network can be viewed as a *ground-truth decoder*, projecting $z^{(i)}$ to clean samples $l^{(i)}$.

But we must also verify that there exists a known *ground-truth encoder* that can correctly invert the decoder, otherwise we cannot be sure that any given VAE structure provably maintains an optimal encoder within its capacity (this is very unlike the linear RPCA case where an analogous condition is trivially satisfied). To check this, we learn the requisite inverse mapping by training something like an inverted autoencoder. Basically, the decoder described above now acts as an encoder, to which we append a new 3-layer ReLU network structured as $l(d)$-$E_1(r_2)$-$E_2(r_1)$-$\hat{z}(\kappa)$, where now E_1 and E_2 denote candidate hidden layers for a potentially optimal encoder. The entire intverted structure then becomes $z(\kappa)$-$D_1(r_1)$-$D_2(r_2)$-$l(d)$-$E_1(r_2)$-$E_2(r_1)$-$\hat{z}(\kappa)$. If any $z^{(i)}$ passes through this network with zero reconstruction error, it implies that the corresponding $l^{(i)}$ can pass through the flipped network with zero reconstruction error, and we have verified our complete ground truth network.

We could train the entire system end-to-end to accomplish this, which should be easy since $\kappa \ll d$; however, we found that although $z^{(i)} = \hat{z}^{(i)}$ is obviously not difficult to achieve, the corresponding learned samples $l^{(i)}$ are pushed to very near a low-rank matrix when assembled into L. This would imply that non-linear manifold learning is not actually even required and RPCA would likely be sufficient.

To circumvent this issue, we instead hold the initial $z(\kappa)$-$D_1(r_1)$-$D_2(r_2)$-$l(d)$ structure fixed, which ensures that the rank of L cannot be altered, and only train the second half using a standard ℓ_2 loss. In doing so we are able to obtain an L matrix, extracted from the middle layer, that is both (a) *not* well-represented by a low-rank approximation, and (b) *does* lie on a *known* low-dimensional non-linear manifold. And given that essentially zero reconstruction error is in fact achievable (up to the expected small ripples introduced by stochastic gradient descent or a similar surrogate), the learned decoder from this process implicitly serves as the ground-truth encoder underlying the data structure. Hence any VAE that includes $l(d)$-$E_1(r_2)$-$E_2(r_1)$-$\hat{z}(\kappa)$ within its encoder mean network capacity, as well as $z(\kappa)$-$D_1(r_1)$-$D_2(r_2)$-$\hat{l}(d)$ within its decoder mean network capacity, will at least in principle have the capability of zero reconstruction error as well.

Finally, once L has been created in this manner, we then generate the noisy data matrix X by randomly corrupting $100 \cdot \nu\%$ of the entries, replacing the original value with samples from a standardized Gaussian distribution. In doing so, the original 'signal' component from L is completely swamped out at these locations.

Experimental Design: Given a data matrix X as generated above, we test the relative performance of four competing models:

1. _VAE_: We form a VAE architecture with the cascaded encoder/decoder mean networks $\mu_x(\mu_z[x])$ assembled as $x(100)$-$E_1(2000)$-$E_2(1000)$-$\mu_z(50)$-$D_1(1000)$-$D_2(2000)$-$\mu_x(100)$. This mirrors the high-level structure used to generate the outlier-free data, and ultimately will ensure that the ground-truth manifold is included within the network pa-

rameterization. Consistent with the design in [23], a diagonal encoder covariance $\boldsymbol{\Sigma}_z$ is produced by sharing just the first two mean network layers. An exponential layer is also appended at the output to produce non-negative values. For consistency with AE models, the decoder covariance $\boldsymbol{\Sigma}_x$ is addressed separately via a special process described below.

2. _AE-ℓ_2_: We begin with the VAE model from above and fix $\boldsymbol{\Sigma}_z = \mathbf{0}$. This reduces the KL regularization term from (3) to simply $\|\boldsymbol{\mu}_z\|_2^2$. If no other changes are included, then the scaling ambiguity between $\boldsymbol{\mu}_z$ and decoder layer $\boldsymbol{D}_1$ is such that $\boldsymbol{\mu}_z$ can be made arbitrarily small without any loss of generality, rendering any beneficial regularization effect from $\|\boldsymbol{\mu}_z\|_2^2$ completely moot as discussed at the beginning of Section 2. Therefore we add a standard weight decay term to the AE-ℓ_2 network parameters $\{\boldsymbol{\theta}, \boldsymbol{\phi}\}$ to ameliorate this scaling ambiguity, which is tantamount to including an additional penalty factor $C_1\|\{\boldsymbol{\theta}, \boldsymbol{\phi}\}\|_2^2$. We also balance $\|\boldsymbol{\mu}_z\|_2^2$ with a second tuning parameter C_2, i.e., $C_2\|\boldsymbol{\mu}_z\|_2^2$. For the experiments in this section, we choose $C_1 = 0.0005$, a typical default value for weight decay, and then tune C_2 for optimal performance.[9]

   Note also that once $\boldsymbol{\Sigma}_z = \mathbf{0}$, at every sample $\boldsymbol{\Sigma}_x$ can be solved for in closed form as $\left[\boldsymbol{\Sigma}_x^{(i)}\right]_{jj} = \left(x_j^{(i)} - \mu_{x_j}^{(i)}\right)^2$ for $j = 1, \ldots, d$ assuming sufficient capacity per Corollary 4. We then plug this value into the AE-ℓ_2 cost, effectively optimizing $\boldsymbol{\Sigma}_x^{(i)}$ out of the model altogether making it entirely deterministic. For direct comparison, we apply the same procedure to the VAE from above, which can be interpreted as efficiently modeling the infinite capacity limit for $\boldsymbol{\Sigma}_x$ (i.e., even with infinite capacity in $\boldsymbol{\Sigma}_x$, the VAE model could do no better than this).

3. _AE-ℓ_1_: To explicitly encourage sparse latent representations, which could potentially be helpful in learning the correct manifold dimension, we begin with the AE-ℓ_2 model from above and replace $\|\boldsymbol{\mu}_z\|_2^2$ with the ℓ_1 norm $\|\boldsymbol{\mu}_z\|_1$, a well-known sparsity-promoting penalty function [12]. The corresponding parameter C_2 is likewise independently tuned for optimal performance.

4. RPCA: As an additional baseline, we also apply the convex RPCA formulation from (14) to the same corrupted data. This model is implemented via an augmented Lagrangian method using code from [28].

For the VAE, AE-ℓ_2, and AE-ℓ_1 networks, all model weights were randomly initialized so as not to copy any information from the ground-truth template. Training was conducted over 200 epochs using the Adam optimization technique [24] with a learning rate of 0.0001 and a batch size of 100. We chose $n = 10^6$ training samples for each separate experiment, across which we varied the manifold dimension from $\kappa = 2, 4, \ldots, 20$ while the outlier ratio

---

9. For direct comparison, we include the same weight decay factor $C_1\|\{\boldsymbol{\theta}, \boldsymbol{\phi}\}\|_2^2$ with the VAE model even though there is no equivalent issue with scaling ambiguity. In fact, this can be viewed as an advantage of the VAE regularization mechanism, in that it directly prevents large decoder weights from compensating for arbitrarily small values of $\boldsymbol{\mu}_z$, killing regularization effects. This is because there exists a key dependence between the weights from $\boldsymbol{D}_1$ and the covariance $\boldsymbol{\Sigma}_z$ such that any large weights that would accommodate pushing $\boldsymbol{\mu}_z$ towards zero would equally _amplify_ the random additive noise coming from the stochastic encoder model, nullifying any benefit to the overall cost.
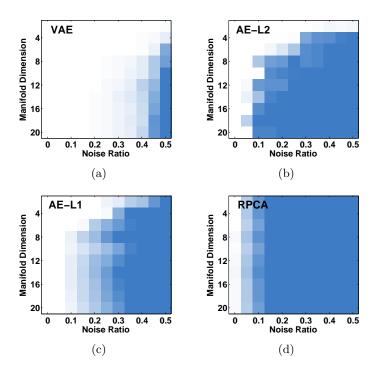
Figure 1: Results recovering synthesized low-dimensional manifolds across different outlier ratios (x-axis) and manifold dimensions (y-axis) for (a) the VAE, (b) the AE-ℓ_2, (c) the AE-ℓ_1, and (d) RPCA. In all cases, white color indicates normalized MSE near 0.0, while dark blue represents 1.0 or failure. The VAE is dramatically superior to each alternative, supporting Hypothesis (i). Additionally, it is crucial to note here that the AE and RPCA solutions perform poorly for quite different reasons. Not surprisingly, convex RPCA fails because it cannot accurately capture the underlying nonlinear manifold using a linear subspace inlier model. In contrast, both AE-ℓ_1 and AE-ℓ_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.

ranged as $\nu = 0.05, 0.10, \ldots, 0.50$. For each pair of experimental conditions, we train/run all four models and measure performance recovering the true $\boldsymbol{L}$ as quantified by the normalized MSE metric

$$\text{NMSE} \triangleq \|\boldsymbol{L} - \hat{\boldsymbol{L}}\|_{\mathcal{F}}^2 / \|\boldsymbol{L}\|_{\mathcal{F}}^2. \tag{20}$$

Note that although in practice we will not generally know the true manifold dimension κ in advance, because we choose $\dim[\boldsymbol{\mu}_z] = 50 > \kappa$ when constructing encoder networks for all experiments, perfect reconstruction is still theoretically possible by any of the VAE or AE models provided that outlier contributions can be successfully mitigated.

16

Results: Figure 1 displays the results estimating $\boldsymbol{L}$, where the VAE outperforms RPCA and the AE models by a wide margin. Perhaps most notably, the VAE performance dominates both AE-ℓ_1 and AE-ℓ_2, supporting our theory that the smoothing effect of integrating over $\boldsymbol{\Sigma}_z$ has immense practical value in avoiding bad minimizing solutions through its unique form of differential regularization. In fact, the AE-ℓ_2 objective is identical to the VAE once $\boldsymbol{\Sigma}_z = \boldsymbol{0}$, at least up to the constant C_2 applied to $\|\boldsymbol{\mu}_z\|_2^2$ which is only tuned to benefit the former while remaining fixed for the latter.[10] So this smoothing effect is essentially the *only* difference between the VAE and AE-ℓ_2 models, and therefore, Figures 1(a) and 1(b) truly isolate the benefits of the VAE in this regard.

To summarize then, by design all VAE and AE network structures are equivalent in terms of their predictive capacity, but only the VAE is able to capitalize on the regularizing effect of $\boldsymbol{\Sigma}_z$ to actually reach a good solution in challenging conditions. Moreover, this is even possible without the hassle of tuning tedious hyperparameters to balance regularization effects as required by AE-ℓ_1 and AE-ℓ_2 models.[11] This confirms Hypothesis (i) and suggests that VAEs are a viable candidate for replacing existing RPCA algorithms [6, 11, 21, 41] in regimes where a single linear subspace is an inadequate signal representation. And we stress that, prior to the analysis herein, it was not at all apparent that a VAE could so dramatically outperform comparable AE models on this type of deterministic outlier removal task.

Note also that perfect reconstruction, as consistently exhibited by the VAE in Figure 1(a), does *not* actually require learning the correct generative model *within* the estimated manifold. Rather it only requires that as $\boldsymbol{\Sigma}_z \to \boldsymbol{0}$ selectively along appropriate dimensions (consistent with Hypothesis (iii) as will be discussed in Section 5.3), the encoder and decoder mean networks project onto the correct manifold while ignoring outliers. Hence although random samples $\boldsymbol{z}$ will likely lie on the true manifold when passed through the decoder network, they need not be perfectly distributed according to the full generative process unless sufficient additional capacity exists beyond that needed to represent the manifold itself. We are not aware of this distinction being discussed in previous works, where VAE and related models are typically evaluated by either the overall quality of their generated samples [13, 26, 33], or by the value of the likelihood bound [23, 25, 5].[12]

Before proceeding to the next set of experiments, we address a tangential issue related to the RPCA performance as exhibited in Figure 1(d). When the outlier ratio is zero, RPCA can recover the ground-truth by simply defaulting to a full-rank inlier model without

---

10. If $C_2 = 1$, the default value as produced by the VAE KL term, the AE-ℓ_2 performance is much worse (not shown) than when using the tuned value of $C_2 = 10^3$ as was adopted in producing Figure 1. In contrast, the VAE requires no such tuning at all, with the default $C_2 = 1$ producing the results shown.

11. Of course we admittedly have not exhaustively ruled out the potential existence of some alternative regularizer capable of outperforming the VAE when carefully tuned to appropriate conditions; however, it is still nonetheless impressive that the VAE can naturally perform so well without such tuning on a task that it was not even originally motivated for.

12. Learning the correct distribution within the manifold, as required for full recovery of the entire generative process and the production of realistic samples, is a topic largely orthogonal to the analysis presented herein. Still, to at least partially address these important issues, we have recently derived relatively broad conditions whereby provable recovery within a manifold itself is possible even in situations where $\boldsymbol{\Sigma}_z$ tends towards zero. However, we defer presentation of this topic, which involves many additional subtleties, to a future publication.

actually learning anything about the true manifold itself (the VAE and AE models do not have this luxury since they are forced to represent $\boldsymbol{L}$ using at most $\dim[\boldsymbol{\mu}_z] = 50 <$ rank$[\boldsymbol{L}] = 100$ dimensions by design). In contrast, as the outlier ratio increases, it becomes increasingly difficult for RPCA to find any linear subspace representation that is both sufficiently high dimensional to include the majority of the inlier variance along the manifold while simultaneously excluding the outlier contributions. This explains the steep drop-off in performance moving from left to right within Figure 1(d). But there is noticeably no change in RPCA performance as we move from top to bottom in the same plot. This is because the clean data $\boldsymbol{L}$ is full-rank regardless of the manifold dimension κ, and so any linear subspace approximation is more or less equally bad across all κ.

5.2 Hypothesis (ii) Evaluation Using Ground-Truth Manifolds and MNIST Data

Synthetic Data Example: To evaluate Hypothesis (ii), we train analogous AE and VAE models as the number of decoder and encoder hidden layers vary, in each case with ground-truth available per the procedure described above. To generate each observed data point $\boldsymbol{x}^{(i)}$, we sample $\boldsymbol{z}^{(i)}$ from a 20-dimensional standard Gaussian distribution and pass it through a neural network structured as $\boldsymbol{z}(20)$-$\boldsymbol{D}_1(200)$-$\boldsymbol{D}_2(200)$-$\boldsymbol{x}(400)$, again with ReLU activations. We then train VAE models of variable depth, with concatenated mean networks $\boldsymbol{\mu}_x\left(\boldsymbol{\mu}_z[\boldsymbol{x}]\right)$ designed as $\boldsymbol{x}(400)$-$\boldsymbol{E}_1(200)$-...-$\boldsymbol{E}_{N_e}(200)$-$\boldsymbol{\mu}_z(30)$-$\boldsymbol{D}_1(200)$-...-$\boldsymbol{D}_{N_d}(200)$-$\boldsymbol{\mu}_x(400)$, where N_e and N_d represent the number of hidden layers in the encoder and decoder respectively. The corresponding covariances are modeled as in Section 5.1, and likewise, the training protocol is unchanged. Note also that $\dim[\boldsymbol{\mu}_z] = 30$ is considerably larger than the ground-truth dimension of 20.

The first layer of the decoder mean network (before the nonlinearity) can be expressed as

$$\boldsymbol{h}_1 = \boldsymbol{W}_1 \boldsymbol{z} + \boldsymbol{b}_1, \tag{21}$$

which in isolation is equivalent to the affine decoder mean model. If the VAE has the ability to find the true underlying manifold dimension, then the number of nonzero columns in $\boldsymbol{W}_1$ should be 20, indicating that $30 - 20 = 10$ dimensions of $\boldsymbol{z}$ are actually useless for any subsequent representation, i.e., we can estimate the intrinsic dimension of the latent code by counting the number of nonzero columns in $\boldsymbol{W}_1$, exactly analogous to the affine case. Of course in practice it is unlikely that a column of $\boldsymbol{W}_1$ converges all the way to exactly $\boldsymbol{0}$ via any stochastic optimization method. Therefore we define a simple threshold as thr $= 0.05 \times \max_{j=1}^{\kappa} \|\boldsymbol{w}_{\cdot j}\|_2$. If $\|\boldsymbol{w}_{\cdot j}\|_2 <$ thr, we regard it as a zero column. But this heuristic notwithstanding, the partition between zero and non-zero columns is generally quite obvious as will be illustrated later.

Table 1 reports the estimated number of non-zero columns in $\boldsymbol{W}_1$ as N_e and N_d are varied, where we have run 10 trials for every pairing and averaged the results. When there is no hidden layer in the decoder (i.e., $N_d = 0$), which implies that the decoder mean is affine, all the columns are nonzero since the network is overly-simplistic and all degrees of freedom are being utilized to compensate. However, once we increase the depth, especially of the decoder within which $\boldsymbol{W}_1$ actually resides, the number of nonzero columns of $\boldsymbol{W}_1$ tends to exactly 20, which is the correct ground-truth manifold dimension by design, directly

	$N_d = 0$	$N_d = 1$	$N_d = 2$	$N_d = 3$
$N_e = 0$	30.0	21.1	21.0	20.0
$N_e = 1$	30.0	21.0	20.0	20.0
$N_e = 2$	30.0	21.0	20.0	20.0
$N_e = 3$	30.0	20.4	20.0	20.0

Table 1: Number of nonzero columns in the VAE decoder mean first-layer weights $\boldsymbol{W}_1$ learned using different encoder and decoder depths applied to data with a ground-truth latent dimension of 20. Provided that the VAE model is sufficiently complex, the correct estimate is automatically obtained. We have not found an analogous AE model with similar capability.

supporting Hypothesis (ii). Similar conclusions can be drawn from models of different sizes and configurations as well (not shown). In contrast, we did not find a corresponding AE model with this capability.

Note that prior work has loosely suggested that the KL regularizer indigenous to VAEs could potentially mute the impact of superfluous latent dimensions as part of the model optimization process [5, 38]. However, there has been no theoretical or empirical demonstration of why this should happen, nor any rigorous explanation of a precise pruning mechanism built into the aggregate VAE cost function itself. And as mentioned previously, the KL term is characterized by an ℓ_2 norm penalty on $\boldsymbol{\mu}_z$ (see (3)), which we would normally expect to promote low-energy latent representations with mostly small, but nonzero values [8], the exact *opposite* of any sparsity-promotion or pruning agency. But of course if columns of $\boldsymbol{W}_1$ are set to zero, then no information about $\boldsymbol{z}$ can pass through these dimensions to the hidden layers of the decoder. Therefore the KL term can now be minimized in isolation along these dimensions with the corresponding elements of $\boldsymbol{\mu}_z$ set to exactly zero. Hence it is only the counterintuitive co-mingling of *all* energy terms that leads to this desirable VAE pruning effect as we have meticulously characterized.

Finally, Figure 2(a) provides validation for our heuristic criterion for classifying columns of $\boldsymbol{W}_1$ as zero or not. Under the same experimental conditions as were used for creating Table 1, we plot the sorted column norms of $\boldsymbol{W}_1$ for the cases where $N_e = 3$ and $N_d \in \{0, 1, 2, 3\}$. Especially when $N_d \in \{2, 3\}$, meaning the model is of (or nearly of) sufficient capacity, zero and nonzero values are easily distinguishable and any reasonable thresholding heuristic would be adequate. Likewise for $N_d = 0$ it is clear that all values are significantly distant from zero. In contrast, when $N_d = 1$ (green curve) it is admittedly more subjective whether or not the smallest 9 or 10 elements should be classified as zero. Regardless, the overall trend is unequivocal, with any heuristic threshold only influencing the $N_d = 1$ boundary case.

MNIST Example: To further verify Hypothesis (ii), we train VAE models on the MNIST dataset of handwritten digit images [27] as κ is varied. We use all $n = 70000$ samples, each of size 28×28. We structure 4-layer cascaded VAE mean networks $\boldsymbol{\mu}_x(\boldsymbol{\mu}_z[\boldsymbol{x}])$ as $\boldsymbol{x}(d)$-$\boldsymbol{E}_1(1000)$-$\boldsymbol{E}_2(500)$-$\boldsymbol{E}_3(250)$-$\boldsymbol{\mu}_z(\kappa)$-$\boldsymbol{D}_1(250)$-$\boldsymbol{D}_2(500)$-$\boldsymbol{D}_3(1000)$-$\boldsymbol{\mu}_x(d)$, where $d =$
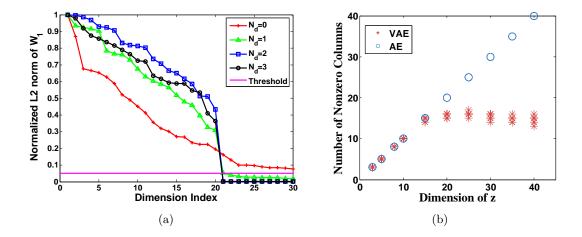
(a)   (b)

Figure 2: (*a*) Validation of thresholding heuristic for determining nonzero columns in $\boldsymbol{W}_1$. With $N_e = 3$ and the settings from Table 1, the sorted column norms of $\boldsymbol{W}_1$ are plotted. Clearly for $N_d \in \{2, 3\}$ the gap between zero and nonzero values is extremely clear and any reasonable thresholding heuristic will suffice. (*b*) Number of nonzero columns in the decoder mean first-layer weights $\boldsymbol{W}_1$ as the latent dimension κ is varied for both AE and VAE models trained on MNIST data. Only the VAE automatically self-regularizes when κ becomes sufficiently large (here at $\kappa \approx 15$), consistent with Hypothesis (ii).

$28 \times 28 = 784$ and ReLU activations are used. Covariances and training protocols are handled as before. We draw values of κ from $\{3, 5, 8, 10, 15, 20, 25, 30, 35, 40\}$.

Figure 2(b) displays the number of nonzero columns in $\boldsymbol{W}_1$ produced by each κ-dependent model, again across 10 trials. We observe that when $\kappa > 15$, the number of nonzero columns plateaus for the VAE consistent with Hypothesis (ii). Of course unlike the synthetic case, we no longer have access to ground truth for determining what the optimal manifold dimension should be.

We also applied an analogous AE model trained with $C_2 = 0$, i.e., a standard AE with no additional regularization penalty added. Not surprisingly, the number of nonzero columns in $\boldsymbol{W}_1$ is always equal to κ since there is no equivalent agency for column-wise pruning as implicitly instilled by the VAE. Note that tuning C_2 with either ℓ_1- or ℓ_2-norm penalties is of course always possible; however, the optimal value can be κ-dependent making subsequent results less interpretable. Moreover, in general we have not found a setting whereby the penalties lead to correct latent dimensionality estimation in situations where the ground-truth is known.
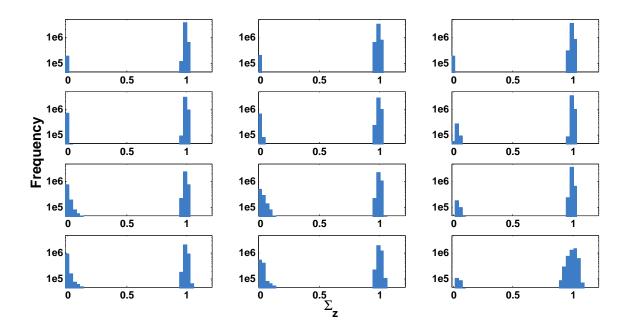
Figure 3: Log-scale histograms of $\left\{\boldsymbol{\Sigma}_z^{(i)}\right\}_{i=1}^n$ diagonal elements as outlier ratios and manifold dimensions are varying for the corrupted manifold recovery experiment corresponding with Figure 1. The three columns represent outlier ratios of $\nu \in \{0.0, 0.25, 0.50\}$ from left to right. The four rows represent manifold dimensions of $\kappa \in \{2, 8, 14, 20\}$ from top to bottom. All plots demonstrate the predicted clustering of variance values around either zero or one. Likewise, the relative sizes of these clusters, including observed changes across experimental conditions, conforms with our theoretical predictions (see detailed description in Section 5.3).

5.3 Hypothesis (iii) Evaluation Using Covariance Statistics from Corrupted Manifold Recovery Task

If some columns of $\boldsymbol{W}_1$ tend to zero as we have argued both empirically and theoretically, then the corresponding diagonal elements of $\boldsymbol{\Sigma}_z$, like $\boldsymbol{\mu}_z$, can no longer influence the decoder. And with only the lingering KL term to offer guidance, along these coordinates the optimal variance will then equal one by virtue of (19). But for nonzero columns of $\boldsymbol{W}_1$, the behavior of $\boldsymbol{\Sigma}_z$ is much more counter-intuitive. Despite the $-\log|\boldsymbol{\Sigma}_z|$ factor from the KL divergence that contributes an unbounded cost as any $[\boldsymbol{\Sigma}_z]_{jj} \to 0$, we nonetheless have proven for the affine decoder mean case a natural tendency of the VAE to push these variance values arbitrarily close to zero when approaching globally optimal solutions, at least along latent dimensions required for representing inlier points lying on ground-truth manifolds (i.e, dimensions where $\boldsymbol{W}_1$ is nonzero).

We now empirically verify that this same effect is inherited by general VAE models with more sophisticated, nonlinear decoder mean networks. For this purpose, we created histograms of all diagonal elements of $\left\{\boldsymbol{\Sigma}_z^{(i)}\right\}_{i=1}^n$ obtained from the experiments described in Section 5.1 where the outlier ratios and manifold dimensions vary. The results are plotted in Figure 3 for all pairs of outlier ratios $\nu \in \{0.0, 0.25, 0.50\}$ (columns) and ground-truth manifold dimensions $\kappa \in \{2, 8, 14, 20\}$ (rows). These results directly conform with our theoretical predictions per the following explanations.

First, consider the upper-left panel displaying the simplest case from an estimation standpoint, since $\nu = 0.0$ (no outliers) and $\kappa = 2$ (very low-dimensional manifold). Here we observe a clear partitioning between elements of $\left\{\boldsymbol{\Sigma}_z^{(i)}\right\}_{i=1}^n$ going to either zero or one. Moreover, given that $\dim[\boldsymbol{\mu}_z] = 50$ while the ground-truth involves $\kappa = 2$, 48 out of 50 dimensions are actually unnecessary. Hence we should expect that only about 4% of variance values should concentrate around zero, with the remainder forced towards one. In fact, this is precisely the general partitioning we observe (note the log scaling of the y-axis). Additionally, if we examine the other panels in the left-most column of Figure 3, we notice that as the ground-truth κ increases, the percentage of variance values shifts from one to zero roughly proportional to $\kappa/50$. In other words, as more dimensions are required to represent the more challenging, higher-dimensional manifolds, more diagonal elements of each $\boldsymbol{\Sigma}_z^{(i)}$ are pushed towards zero to enforce accurate reconstructions.

Next, we observe that in the top row of Figure 3, each of the three panels are more or less the same, indicating that the inclusion of outliers has not disrupted the VAE's ability to model the ground-truth manifold. In contrast, the bottom row presents a somewhat different story. Given the more challenging conditions with a much higher dimensional ground-truth manifold ($\kappa = 20$), the inclusion of additional outliers (as we move from left to right) shifts more variance elements from zero to one. This implies that the VAE, when confronted with both a higher-dimensional manifold and severe outliers (bottom-right panel), is settling on a relatively lower-dimensional approximation. This behavior is reasonable in the sense that accurately estimating a complex manifold via any method becomes problematic when 50% of the data is corrupted, and a low-dimensional approximation is all that is feasible to avoid simply fitting all the outliers. In this situation some manifold dimensions of lesser
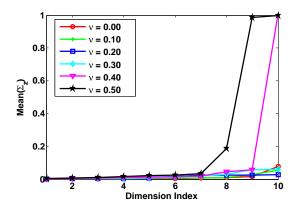
Figure 4: Diagonal values of $\frac{1}{n}\sum_{i=1}^{n}\mathbf{\Sigma}_z^{(i)}$ sorted in ascending order for a VAE model trained with ground-truth $\kappa = \dim[\boldsymbol{\mu}_z] = 10$ on the recovery task from Section 5.1. When the outlier proportion is $\nu \leq 0.30$, the average variance is near zero across all latent dimensions. However, for $\nu > 0.30$, some variance values are pushed towards one, indicating that the VAE is defaulting to a lower-dimensional approximation.

importance can be viewed as expendable, and consequently we will likely have additional elements of $\left\{\mathbf{\Sigma}_z^{(i)}\right\}_{i=1}^{n}$ tending to one.

To further elucidate this phenomena, we include one additional supporting visualization involving the special case where the ground-truth manifold dimension equals 10 while the outlier ratio ν varies. However, we slightly modify the testing conditions from Section 5.1. Instead of choosing $\dim[\boldsymbol{\mu}_z] = 50$, we set this value to the ground-truth value $\kappa = 10$. In this constrained setting, we expect that perfect recovery should require *all* diagonal elements of $\mathbf{\Sigma}_z$ to be pushed towards zero, since there are no longer any superfluous degrees of freedom. Therefore, if any covariance elements tend to one, we have isolated the emergence of a low-dimensional approximation as presumably necessitated by increasing outlier levels.

Figure 4 displays the diagonal values of $\frac{1}{n}\sum_{i=1}^{n}\mathbf{\Sigma}_z^{(i)}$ sorted in ascending order along the x-axis. When $\nu \leq 0.30$, the average variance is near zero across all latent dimensions. However, for $\nu > 0.30$, some variance values are pushed towards one, indicating that the VAE is defaulting to a lower-dimensional approximation.

To summarize then, the results of this section help to confirm a rather curious behavior of the VAE: If $\boldsymbol{\mu}_x\left(\boldsymbol{\mu}_z\left[\boldsymbol{x}\right]\right)$ is suitably parameterized to model inlier samples, and $\mathbf{\Sigma}_x$ is sufficiently complex to model outlier locations, then elements of $\mathbf{\Sigma}_z$ can be selectively pushed towards zero in the neighborhood of global minima. This involves overpowering the $-\log|\mathbf{\Sigma}_z\|$ factor from the KL divergence that would otherwise seemingly prevent this from happening. Moreover, in this degenerate regime, the VAE will exhibit deterministic behavior and can perfectly represent original clean training data samples via a low-dimensional manifold provided that the outlier level is not too high.
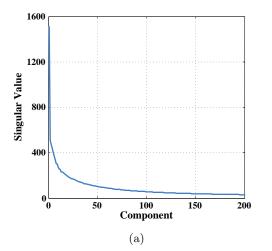
6. Discussion

Although originally developed as a viable deep generative model or tractable bound on the data likelihood, in this work we have revealed certain properties and abilities of the VAE that are not obvious from first inspection. For example, in addition to its putative role in driving diversity into the learned generative process, the latent covariance $\boldsymbol{\Sigma}_z$ also serves as an important smoothing mechanism that aids in the robust recovery of corrupted samples, even if sometimes this requires exhibiting behavior (i.e., selective convergence towards zero) that may seem counterintuitive. And although the VAE only adopts an ℓ_2 norm penalty on $\boldsymbol{\mu}_z$ that in isolation should favor low energy solutions with all or mostly nonzero values, the latent mean estimator nonetheless tends to be highly sparse because of subtle, non-obvious interactions with other factors in the energy function such as the first-layer decoder mean network weights $\boldsymbol{W}_1$. Likewise, outliers can be estimated and completely removed via the action of $\boldsymbol{\Sigma}_x$ despite no traditional, additive sparsity penalty applied across each data point.

In general, our results speak to many under-appreciated aspects of VAE behavior, have wide ranging practical consequences, and suggest novel usages beyond the original VAE design principles. These include:

- The VAE can be applied to estimating deterministic nonlinear manifolds heavily corrupted with outliers.

- The self-regularization effects of the VAE can largely handle excessive degrees of freedom when it comes to the latent representation $\boldsymbol{z}$ as produced by the full encoder and processed by the first layer of the decoder mean network, as well as an arbitrarily-parameterized decoder covariance $\boldsymbol{\Sigma}_x$. Conversely, only excessive complexity specifically localized in higher decoder mean network layers can, at least in principle, lead to potential problems with overfitting.

- The latent covariance $\boldsymbol{\Sigma}_z$ can serve as an approximate bellwether for determining the true dimensionality of a manifold, provided that excessive outliers/corruptions do not lead to an under-estimate. This is because typically near global solutions, we observe $[\boldsymbol{\Sigma}_z]_{jj} \to 0$ for *useful* dimensions, while for *useless* dimensions we have shown that $[\boldsymbol{\Sigma}_z]_{jj} \to 1$, a clear bifurcation.

Although the primary purpose of this paper is not to build a better generative model per se, we nevertheless hope that ideas introduced here will help to ensure that VAEs are not under or improperly utilized. Additionally, in closing we should also mention that the focus herein has been almost entirely on the analysis of the VAE energy function itself, independent of the specifics of how this energy function might ultimately be optimized in practice. But we believe the latter to be an equally-important, complementary topic, and further study is undeniably warranted. For example, if an optimization trajectory is somehow lured astray by the Siren's song of a bad local minimum in the VAE energy landscape, then obviously many of our conclusions predicated on global optima will not necessarily still hold.
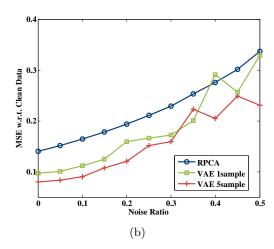
Acknowledgments

(a) (b)

Figure 5: (a) Singular value spectrum of MNIST data revealing (approximately) low-rank structure. (b) Normalized MSE recovering MNIST digits from corrupted samples. The VAE is able to reduce the reconstruction error by better modeling more fine-grain details occupying the low end of the singular value spectrum.

Appendix A. Additional MNIST Dataset Experiment

Here we examine practical denoising of MNIST data corrupted with outliers using a VAE model. Outliers are added to MNIST handwritten digit data [27] by randomly replacing from 5% to 50% of the pixels with a value uniformly sampled from $[0, 255]$ to create $\boldsymbol{X}$. We choose $\kappa = 30$ for the dimension of $\boldsymbol{z}$ and apply the same VAE structure as applied to MNIST data in Section 5.2. The model is trained using both $\tau = 1$ and $\tau = 5$ latent samples $\{\boldsymbol{z}^{(i,t)}\}_{t=1}^{\tau}$ for each $\boldsymbol{x}^{(i)}$, observing that the latter, which more closely approximates the posterior, should perform significantly better.

We compare the VAE against convex RPCA on the task of recovering the original, uncorrupted digits. Note that RPCA is commonly used for unsupervised cleaning of this type of data [14], and MNIST is known to have significant low-rank structure [29] as shown in Figure 5(a). Regardless, we observe in Figure 5(b) that the VAE performs significantly better in terms of normalized MSE by capturing additional manifold details that deviate from a purely low-rank representation. Furthermore, we hypothesize that using extra latent samples (the $\tau = 5$ case) may work better on outlier removal tasks given the strong need for accurate smoothing of the VAE objective as described previously.

Appendix B. Proof of Lemma 1

Under the stated assumptions, the VAE cost can be simplified as

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) \;=\; & \sum_i \left\{ \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}^{(i)})} \left[\tfrac{1}{\lambda} \left\| \boldsymbol{x}^{(i)} - \boldsymbol{W}\boldsymbol{z} - \boldsymbol{b} \right\|_2^2 \right] + d \log \lambda \right. \\
& + \;\; \mathrm{tr}\left[\boldsymbol{\Sigma}_z^{(i)} \right] - \log \left| \boldsymbol{\Sigma}_z^{(i)} \right| + \| \boldsymbol{\mu}_z^{(i)} \|_2^2 \Big\} . \\
\;=\; & \sum_i \left\{ \tfrac{1}{\lambda} \left\| \boldsymbol{x}^{(i)} - \boldsymbol{W}\boldsymbol{\mu}_z^{(i)} - \boldsymbol{b} \right\|_2^2 + \tfrac{1}{\lambda} \mathrm{tr}\left[\boldsymbol{\Sigma}_z^{(i)} \boldsymbol{W}^\top \boldsymbol{W} \right] + d \log \lambda \right. \\
& + \;\; \mathrm{tr}\left[\boldsymbol{\Sigma}_z^{(i)} \right] - \log \left| \boldsymbol{\Sigma}_z^{(i)} \right| + \| \boldsymbol{\mu}_z^{(i)} \|_2^2 \Big\} ,
\end{aligned}
\tag{22}
$$

where $\boldsymbol{\mu}_z^{(i)} \triangleq \boldsymbol{\mu}_z \left(\boldsymbol{x}^{(i)}; \boldsymbol{\phi} \right)$ and $\boldsymbol{\Sigma}_z^{(i)} \triangleq \boldsymbol{\Sigma}_z \left(\boldsymbol{x}^{(i)}; \boldsymbol{\phi} \right)$. Given that

$$
\log \left| \boldsymbol{A}\boldsymbol{A}^\top \right| = \arg \inf_{\boldsymbol{\Gamma} \succ \mathbf{0}} \; \mathrm{tr}\left[\boldsymbol{A}\boldsymbol{A}^\top \boldsymbol{\Gamma}^{-1} \right] + \log |\boldsymbol{\Gamma}| ,
\tag{23}
$$

when optimization is carried out over positive definite matrices $\boldsymbol{\Gamma}$, minimization of (22) with respect to $\boldsymbol{\Sigma}_z^{(i)}$ leads to the revised objective

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) \;\equiv\; & \sum_i \left\{ \tfrac{1}{\lambda} \left\| \boldsymbol{x}^{(i)} - \boldsymbol{W}\boldsymbol{\mu}_z^{(i)} - \boldsymbol{b} \right\|_2^2 + \log \left| \tfrac{1}{\lambda} \boldsymbol{W}^\top \boldsymbol{W} + \boldsymbol{I} \right| + d \log \lambda + \| \boldsymbol{\mu}_z^{(i)} \|_2^2 \right\} , \\
\;=\; & \sum_i \left\{ \tfrac{1}{\lambda} \left\| \boldsymbol{x}^{(i)} - \boldsymbol{W}\boldsymbol{\mu}_z^{(i)} - \boldsymbol{b} \right\|_2^2 + \log \left| \boldsymbol{W}\boldsymbol{W}^\top + \lambda \boldsymbol{I} \right| + \| \boldsymbol{\mu}_z^{(i)} \|_2^2 \right\} ,
\end{aligned}
\tag{24}
$$

ignoring constant terms. This expression only requires that $\boldsymbol{\Sigma}_z^{(i)} = \left[\tfrac{1}{\lambda} \boldsymbol{W}^\top \boldsymbol{W} + \boldsymbol{I} \right]^{-1}$, or a constant parameterization, independent of $\boldsymbol{x}^{(i)}$. Similarly we can optimize over $\boldsymbol{\mu}_z^{(i)}$ in terms of the other variables. This is just a ridge regression problem, with optimal solution

$$
\boldsymbol{\mu}_z^{(i)} = \boldsymbol{W}^\top \left(\lambda \boldsymbol{I} + \boldsymbol{W}\boldsymbol{W}^\top \right)^{-1} \left(\boldsymbol{x}^{(i)} - \boldsymbol{b} \right) ,
\tag{25}
$$

or a simple linear function of $\boldsymbol{x}^{(i)}$. Hence as long as the parameterization of both $\boldsymbol{\mu}_z^{(i)}$ and $\boldsymbol{\Sigma}_z^{(i)}$ allows for arbitrary affine functions as stipulated in the lemma statement, these optimal solutions are feasible. Plugging (25) into (24) and applying some basic linear algebra, we arrive at

$$
\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) \;\equiv\; \sum_i \left(\boldsymbol{x}^{(i)} - \boldsymbol{b} \right)^\top \left(\boldsymbol{W}\boldsymbol{W}^\top + \lambda \boldsymbol{I} \right)^{-1} \left(\boldsymbol{x}^{(i)} - \boldsymbol{b} \right) + n \log \left| \boldsymbol{W}\boldsymbol{W}^\top + \lambda \boldsymbol{I} \right| .
\tag{26}
$$

Finally, in the event that we enforce that $\boldsymbol{\Sigma}_z^{(i)}$ be diagonal, (24) must be modified via

$$
\boldsymbol{\Sigma}_z^{(i)} = \left[\tfrac{1}{\lambda} \mathrm{diag}\left(\mathrm{diag}\left[\boldsymbol{W}^\top \boldsymbol{W} \right] \right) + \boldsymbol{I} \right]^{-1} = \sum_{j=1}^{\kappa} \log \left(\lambda + \| \boldsymbol{w}_{\cdot j} \|_2^2 \right) - \kappa \log \lambda ,
\tag{27}
$$

where the diag[·] operator converts vectors to diagonal matrices, and a matrix to a vector formed from its diagonal (just as in the Matlab computing environment), leading to the stated result.

Appendix C. Proof of Theorem 2

First, for part 1 on the theorem, given that $\boldsymbol{W}\boldsymbol{R}\boldsymbol{R}^{\top}\boldsymbol{W}^{\top} = \boldsymbol{W}\boldsymbol{P}\boldsymbol{P}^{\top}\boldsymbol{W}^{\top} = \boldsymbol{W}\boldsymbol{W}^{\top}$ for any rotation $\boldsymbol{R}$ and permutation $\boldsymbol{P}$, then obviously if $\boldsymbol{W}^*$ is a minimum of (5), $\boldsymbol{W}^*\boldsymbol{R}$ and $\boldsymbol{W}^*\boldsymbol{P}$ must also be. Likewise, since $\sum_{j=1}^{\kappa} \log\left(\lambda + \|\boldsymbol{w}_{\cdot j}\|_2^2\right)$ is invariant to the order of the summation, then if $\boldsymbol{W}^{**}$ is a minimum of (7), $\boldsymbol{W}^{**}\boldsymbol{P}$ must be as well.

We also have that

$$
\begin{aligned}
\mathcal{L}_{sep}(\boldsymbol{W}^{**}, \boldsymbol{b}, \lambda) &= \sum_i \boldsymbol{\Omega}^{(i)}(\boldsymbol{W}^{**}, \boldsymbol{b}, \lambda\boldsymbol{I}) + n\left[\sum_j \log\left(\lambda + \|\boldsymbol{w}_{\cdot j}^{**}\|_2^2\right) + (d - \kappa)\log\lambda\right] \\
&= \sum_i \boldsymbol{\Omega}^{(i)}(\boldsymbol{W}^{**}, \boldsymbol{b}, \lambda\boldsymbol{I}) + n\left[\sum_j \log\left(1 + \tfrac{1}{\lambda}\|\boldsymbol{w}_{\cdot j}^{**}\|_2^2\right) + d\log\lambda\right] \\
&\geq \sum_i \boldsymbol{\Omega}^{(i)}(\boldsymbol{W}^{**}, \boldsymbol{b}, \lambda\boldsymbol{I}) + n\left[\log\left|\tfrac{1}{\lambda}(\boldsymbol{W}^{**})^{\top}\boldsymbol{W}^{**} + \boldsymbol{I}\right| + d\log\lambda\right] \\
&= \sum_i \boldsymbol{\Omega}^{(i)}(\boldsymbol{W}^{**}, \boldsymbol{b}, \lambda\boldsymbol{I}) + n\log\left|\lambda\boldsymbol{I} + \boldsymbol{W}^{**}\boldsymbol{R}(\boldsymbol{W}^{**}\boldsymbol{R})^{\top}\right| \\
&\geq \sum_i \boldsymbol{\Omega}^{(i)}(\boldsymbol{W}^*, \boldsymbol{b}, \lambda\boldsymbol{I}) + n\log\left|\lambda\boldsymbol{I} + \boldsymbol{W}^*\boldsymbol{R}(\boldsymbol{W}^*\boldsymbol{R})^{\top}\right| \quad (28)
\end{aligned}
$$

where the the second inequality follows from the fact that $\boldsymbol{W}^*$ is an optimal solution to (5). The first inequality stems from Hadamard's inequality [15] applied to

$$
\tfrac{1}{\lambda}(\boldsymbol{W}^{**})^{\top}\boldsymbol{W}^{**} + \boldsymbol{I} = \boldsymbol{M}^{\top}\boldsymbol{M} \quad (29)
$$

for some square matrix $\boldsymbol{M}$ of appropriate dimension. This results in

$$
\log\left|\tfrac{1}{\lambda}(\boldsymbol{W}^{**})^{\top}\boldsymbol{W}^{**} + \boldsymbol{I}\right| = 2\log|\boldsymbol{M}| \leq 2\log\left(\prod_j \|\boldsymbol{m}_{\cdot j}\|_2\right) = \sum_j \log\left(1 + \tfrac{1}{\lambda}\|\boldsymbol{w}_{\cdot j}^{**}\|_2^2\right), \quad (30)
$$

with equality iff $\boldsymbol{M}^{\top}\boldsymbol{M}$ is diagonal. We can further manipulate the log-det term in (28) via

$$
\begin{aligned}
n\log\left|\lambda\boldsymbol{I} + \boldsymbol{W}^*\boldsymbol{R}(\boldsymbol{W}^*\boldsymbol{R})^{\top}\right| &= \log\left|\tfrac{1}{\lambda}(\boldsymbol{W}^*\boldsymbol{R})^{\top}\boldsymbol{W}^*\boldsymbol{R} + \boldsymbol{I}\right| + d\log\lambda \\
&= \log\left|\tfrac{1}{\lambda}\left(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^{\top}\boldsymbol{R}\right)^{\top}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^{\top}\boldsymbol{R} + \boldsymbol{I}\right| + d\log\lambda \\
&= \log\left|\tfrac{1}{\lambda}\boldsymbol{R}^{\top}\boldsymbol{V}\boldsymbol{\Lambda}^2\boldsymbol{V}^{\top}\boldsymbol{R} + \boldsymbol{I}\right| + d\log\lambda,
\end{aligned}
$$

where $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^{\top}$ is the SVD of $\boldsymbol{W}^*$. Now if we choose $\boldsymbol{R} = \boldsymbol{V}$ and define $\bar{\boldsymbol{W}} \triangleq \boldsymbol{W}\boldsymbol{V}$, then $\Lambda_{jj} = \|\bar{\boldsymbol{w}}_{\cdot j}\|_2^2$ and this expression further reduces via

$$
\begin{aligned}
\log\left|\tfrac{1}{\lambda}\boldsymbol{R}^{\top}\boldsymbol{V}\boldsymbol{\Lambda}^2\boldsymbol{V}^{\top}\boldsymbol{R} + \boldsymbol{I}\right| + d\log\lambda &= \sum_j \log\left(1 + \tfrac{1}{\lambda}\Lambda_{jj}\right) + d\log\lambda \\
&= \sum_j \log\left(\lambda + \|\bar{\boldsymbol{w}}_{\cdot j}\|_2^2\right) + (d - \kappa)\log\lambda. \quad (31)
\end{aligned}
$$

Of course we cannot have

$$\sum_i \mathbf{\Omega}^{(i)}(\boldsymbol{W}^{**}, \boldsymbol{b}, \lambda \boldsymbol{I}) + n \left[\sum_j \log\left(\lambda + \|\boldsymbol{w}^{**}_{\cdot j}\|_2^2\right) + (d - \kappa)\log\lambda \right] \tag{32}$$

$$> \sum_i \mathbf{\Omega}^{(i)}(\bar{\boldsymbol{W}}, \boldsymbol{b}, \lambda \boldsymbol{I}) + n \left[\sum_j \log\left(\lambda + \|\bar{\boldsymbol{w}}_{\cdot j}\|_2^2\right) + (d - \kappa)\log\lambda \right],$$

otherwise $\boldsymbol{W}^{**}$ would not be a minimum of (7). Therefore, $\bar{\boldsymbol{W}}$ must also be a minimum of (7), from which the remaining parts of (9) immediately follows.

We next confront the arrangement of disconnected minima for part 2 of the theorem. It is not difficult to show that (5), and by virtue of the analysis above (7), will be uniquely minimized by $\boldsymbol{U}$ and $\boldsymbol{\Lambda}$ arising from the SVD of either $\boldsymbol{W}^*$ or equivalently $\boldsymbol{W}^{**}$. Let $\boldsymbol{W}^{**} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top$ via such a decomposition. So any partitioning into disconnected minimizers must come at the hands of $\boldsymbol{V}$, which only influences the $\sum_j \log\left(\lambda + \|\boldsymbol{w}_{\cdot j}\|_2^2\right)$ term in (7).

At stated above, if $\boldsymbol{W}^{**}$ is a minimum, then $\boldsymbol{W}^{**}\boldsymbol{P}$ must also be a minimum. Assume for the moment that $\boldsymbol{W}^{**}$ is full column rank. There will obviously be $r!$ unique permutations of its columns, with $r = \mathrm{rank}[\boldsymbol{W}^{**}]$. Moreover, any transition from some permutation $\boldsymbol{P}'$ to another $\boldsymbol{P}''$ will necessarily involve some non-permutation-matrix rotation $\boldsymbol{V}$. Given our assumption of distinct eigenvalues, this will ensure that

$$\tfrac{1}{\lambda}\left(\boldsymbol{W}^{**}\right)^\top \boldsymbol{W}^{**} + \boldsymbol{I} = \tfrac{1}{\lambda}\boldsymbol{V}\boldsymbol{\Lambda}^2\boldsymbol{V}^\top + \boldsymbol{I} \tag{33}$$

is non-diagonal. While this will not increase (5), it *must* increase (7) when diagonalized by Hadamard's inequality. Therefore every permutation will reflect a distinct, disconnected minimizer. If $\boldsymbol{W}^{**}$ also has $\kappa - r$ zero-valued columns, then the resulting number of unique permutations increases to $\frac{\kappa!}{\kappa-r}$ by standard rules of combinatorics.

Finally, part 3 of the theorem follows directly from part 2: Given that any minimizer of (7) must be of the form $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{P}$, then there cannot be more than r nonzero columns. In contrast, for (5) we may apply any arbitrary rotation to $\boldsymbol{W}^*$, and hence all columns can be nonzero even if the rank is smaller than κ.

Appendix D. Proof of Theorem 3

For convenience we will adopt the notation $f(x) = \mathrm{O}(h(\epsilon))$ to indicate that $|f(x)| < C|h(\epsilon)|$ for some constant C independent of x or ϵ, and $f(x) = \Omega(h(\epsilon))$ to indicate that $|f(x)| > C|h(\epsilon)|$ for some constant C independent of x or ϵ. Likewise, we say $f(x) = \Theta(h(\epsilon))$ iff $f(x) = \mathrm{O}(h(\epsilon))$ and $f(x) = \Omega(h(\epsilon))$. Additionally, if the input argument to one of these expressions is a vector, the result is understood to apply element-wise.

The basic high-level strategy here is as follows: We first present a candidate solution that satisfies (12) and carefully quantify the achievable objective function value for $\alpha \in (0, \bar{\alpha}]$, and $\bar{\alpha}$ small. We then analyze a lower bound on the VAE cost and demonstrate that no solution can do significantly better, namely, any solution that can match the performance of our original proposal must necessarily also satisfy (12). Given that this is a lower bound, this implies that no other solution can both minimize the VAE objective and not satisfy (12). We now proceed to the details.

Define $\boldsymbol{\mu}_z^{(i)} \triangleq \boldsymbol{\mu}_z \left(\boldsymbol{x}^{(i)}; \boldsymbol{\phi} \right)$ and $\boldsymbol{\Sigma}_z^{(i)} \triangleq \boldsymbol{\Sigma}_z \left(\boldsymbol{x}^{(i)}; \boldsymbol{\phi} \right)$. We first note that if $\boldsymbol{z} = \boldsymbol{\mu}_z^{(i)} + \boldsymbol{S}_z^{(i)} \boldsymbol{\epsilon}$, with $\boldsymbol{S}_z^{(i)}$ satisfying $\boldsymbol{\Sigma}_z^{(i)} = \boldsymbol{S}_z^{(i)} \left(\boldsymbol{S}_z^{(i)} \right)^{\top}$, and $\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}; \boldsymbol{0}, \boldsymbol{I})$, then $\boldsymbol{z} \sim q_{\boldsymbol{\phi}} \left(\boldsymbol{z} | \boldsymbol{x}^{(i)} \right)$. With this reparameterization and

$$
\begin{aligned}
\boldsymbol{\mu}_x^{(i)} &\triangleq \boldsymbol{W} \boldsymbol{\mu}_z^{(i)} + \boldsymbol{W} \boldsymbol{S}_z^{(i)} \boldsymbol{\epsilon}, \\
\mathrm{diag}[\boldsymbol{\Sigma}_x^{(i)}] &\triangleq \nu \left(\boldsymbol{\mu}_z^{(i)} + \boldsymbol{S}_z^{(i)} \boldsymbol{\epsilon}; \boldsymbol{\theta} \right) \text{ for some function } \nu \\
\boldsymbol{\mu}_z^{(i)} &\triangleq f(\boldsymbol{x}^{(i)}; \boldsymbol{\phi}) \text{ for some function } f \\
\boldsymbol{S}_z^{(i)} &\triangleq g(\boldsymbol{x}^{(i)}; \boldsymbol{\phi}) \text{ for some function } g,
\end{aligned}
\tag{34}
$$

the equivalent VAE objective becomes

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_i \Bigg\{ &\mathbb{E}_{p(\boldsymbol{\epsilon})} \left[\left(\boldsymbol{x}^{(i)} - \boldsymbol{W} \boldsymbol{\mu}_z^{(i)} - \boldsymbol{W} \boldsymbol{S}_z^{(i)} \boldsymbol{\epsilon} \right)^{\top} \left(\boldsymbol{\Sigma}_x^{(i)} \right)^{-1} \left(\boldsymbol{x}^{(i)} - \boldsymbol{W} \boldsymbol{\mu}_z^{(i)} - \boldsymbol{W} \boldsymbol{S}_z^{(i)} \boldsymbol{\epsilon} \right) \right] \\
&+ \mathbb{E}_{p(\boldsymbol{\epsilon})} \left[\log \left| \boldsymbol{\Sigma}_x^{(i)} \right| \right] + \mathrm{tr} \left[\boldsymbol{\Sigma}_z^{(i)} \right] - \log \left| \boldsymbol{\Sigma}_z^{(i)} \right| + \| \boldsymbol{\mu}_z^{(i)} \|_2^2 \Bigg\}
\end{aligned}
\tag{35}
$$

when $\boldsymbol{b} = \boldsymbol{0}$ as stipulated.[13] For now assume that κ, the dimension of the latent $\boldsymbol{z}$, satisfies $\kappa = \mathrm{rank}[\boldsymbol{U}]$ (later we will relax this assumption).

D.1 A Candidate Solution

Here we consider a candidate solution that, by design, satisfies (12). For the encoder parameters we choose

$$
\hat{\boldsymbol{\mu}}_z^{(i)} = \boldsymbol{\pi}^{(i)}, \quad \hat{\boldsymbol{\Sigma}}_z^{(i)} = \alpha \boldsymbol{I}.
\tag{36}
$$

where α is a non-negative scalar and $\boldsymbol{\pi}^{(i)}$ is defined in conjunction with a matrix $\boldsymbol{\Psi}$ such that

$$
\begin{aligned}
\mathrm{supp}_\alpha \left[\boldsymbol{x}^{(i)} - \boldsymbol{\Psi} \boldsymbol{\pi}^{(i)} \right] &= \mathrm{supp} \left[\boldsymbol{s}^{(i)} \right] \\
\mathrm{span} \left[\boldsymbol{U} \right] &= \mathrm{span} \left[\boldsymbol{\Psi} \right].
\end{aligned}
\tag{37}
$$

All quantities in (36) can be readily computed via $\boldsymbol{X}$ applied to an encoder module provided that $\kappa = \dim[\boldsymbol{z}] = \mathrm{rank}[\boldsymbol{U}]$ as stipulated, and sufficient representational complexity for $\boldsymbol{\mu}_z$ and $\boldsymbol{\Sigma}_z$. Additionally, for the encoder we only need to define the posterior moments at specific points $\boldsymbol{x}^{(i)}$, hence the indexing via i in (36).

In contrast, for the decoder we consider the solution defined over any $\boldsymbol{z}$ given by

$$
\begin{aligned}
\hat{\boldsymbol{W}} &= \boldsymbol{\Psi} \\
\hat{\boldsymbol{\mu}}_x &= \hat{\boldsymbol{W}} \boldsymbol{z} \\
\mathrm{diag} \left[\hat{\boldsymbol{\Sigma}}_x \right] &= \boldsymbol{\Lambda}^{(h_\pi(\boldsymbol{z}))},
\end{aligned}
\tag{38}
$$

where $\boldsymbol{\Lambda}^{(i)} \in \mathbb{R}^{d \times d}$ is a diagonal matrix with

$$
\left[\boldsymbol{\Lambda}^{(i)} \right]_{jj} = \begin{cases} \alpha, & \text{if } s_j^{(i)} = 0, \\ 1, & \text{otherwise}, \quad \forall j. \end{cases}
\tag{39}
$$

---

13. The extension to arbitrary $\boldsymbol{b}$ is trivial but clutters the presentation.

and $h_\pi : \mathbb{R}^\kappa \to \{1, \ldots, n\}$ is a function satisfying

$$h_\pi(\boldsymbol{z}) \triangleq \arg \min_{i \in \{1,\ldots,n\}} \|\boldsymbol{z} - \boldsymbol{\pi}^{(i)}\|_2. \tag{40}$$

Again, given sufficient capacity, this function can always be learned by the decoder such that (38) is computable for any $\boldsymbol{z}$. Given these definitions, then the index-specific moments $\hat{\boldsymbol{\mu}}_x^{(i)}$ and $\hat{\boldsymbol{\Sigma}}_x^{(i)}$ are of course reduced to functions of $\boldsymbol{\epsilon}$ given by

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_x^{(i)} &= \hat{\boldsymbol{\mu}}_x \left(\hat{\boldsymbol{\mu}}_z^{(i)} + \hat{\boldsymbol{S}}_z^{(i)} \boldsymbol{\epsilon}; \boldsymbol{\theta} \right) \\
\hat{\boldsymbol{\Sigma}}_x^{(i)} &= \hat{\boldsymbol{\Sigma}}_x \left(\hat{\boldsymbol{\mu}}_z^{(i)} + \hat{\boldsymbol{S}}_z^{(i)} \boldsymbol{\epsilon}; \boldsymbol{\theta} \right).
\end{aligned} \tag{41}$$

.

We next analyze the behavior of (35) at this specially parameterized solution as $\bar{\alpha}$ becomes small, in which case by design all covariances will be feasible by design. For this purpose, we first consider the integration across all cases where $\hat{\boldsymbol{\Sigma}}_x^{(i)}$ does not reflect the correct support, meaning $\boldsymbol{\epsilon} \notin \mathcal{S}^{(i)}$, where

$$\mathcal{S}^{(i)} \triangleq \left\{ \boldsymbol{\epsilon} : \quad \left[\boldsymbol{\Sigma}_x^{(i)} \right]_{jj} = \alpha \;\; \text{iff} \;\; s_j^{(i)} = 0, \;\; \forall j \right\}. \tag{42}$$

With this segmentation in mind, the VAE objection naturally partitions as

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_i \left\{ \mathcal{L}^{(i)}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\epsilon} \notin \mathcal{S}^{(i)}) + \mathcal{L}^{(i)}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\epsilon} \in \mathcal{S}^{(i)}) \right\} \tag{43}$$

where $\mathcal{L}^{(i)}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\epsilon} \notin \mathcal{S}^{(i)})$ denotes the cost for the i-th sample when integrated across those samples not in $\mathcal{S}^{(i)}$, and $\mathcal{L}^{(i)}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\epsilon} \in \mathcal{S}^{(i)})$ is the associated complement.

D.2 Evaluation of $\mathcal{L}^{(i)}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\epsilon} \notin \mathcal{S}^{(i)})$

First we define

$$\rho = \min_{i,j \in \{1,\ldots,n\}, i \neq j} \tfrac{1}{2} \|\boldsymbol{\pi}^{(i)} - \boldsymbol{\pi}^{(j)}\|_2, \tag{44}$$

which is just half the minimum distance between any two distinct coefficient expansions. If any $\boldsymbol{z}$ is within this distance of $\boldsymbol{\pi}^{(i)}$, it will necessarily be quantized to this value per our previous definitions. Therefore if $\|\hat{\boldsymbol{S}}_z^{(i)} \boldsymbol{\epsilon}\|_2 < \rho$, we are guaranteed that the correct generating support pattern will be mapped to $\hat{\boldsymbol{\Sigma}}_x^{(i)}$, and so it follows that

$$P\left(\boldsymbol{\epsilon} \notin \mathcal{S}^{(i)} \right) \leq P\left(\left\| \hat{\boldsymbol{S}}_z^{(i)} \boldsymbol{\epsilon} \right\|_2 > \rho \right) = P(\|\sqrt{\alpha} \boldsymbol{\epsilon}\|_2 > \rho) \tag{45}$$

at our candidate solution. We also make use of the quantity

$$\eta \triangleq \max_{i \in \{1,\ldots,n\}} \|\boldsymbol{x}^{(i)} - \boldsymbol{\Psi} \boldsymbol{\pi}^{(i)}\|_2^2, \tag{46}$$

which represents the maximum data-fitting error. Then for the i-th sample we have

$$
\begin{aligned}
\mathcal{L}^{(i)}(\boldsymbol{\theta}, \boldsymbol{\phi}; & \boldsymbol{\epsilon} \notin \mathcal{S}^{(i)}) \\
= & \int_{\boldsymbol{\epsilon} \notin \mathcal{S}^{(i)}} \left[\left(\boldsymbol{x}^{(i)} - \hat{\boldsymbol{W}} \hat{\boldsymbol{\mu}}_z^{(i)} - \hat{\boldsymbol{W}} \hat{\boldsymbol{S}}_z^{(i)} \boldsymbol{\epsilon} \right)^\top \left(\hat{\boldsymbol{\Sigma}}_x^{(i)} \right)^{-1} \left(\boldsymbol{x}^{(i)} - \hat{\boldsymbol{W}} \hat{\boldsymbol{\mu}}_z^{(i)} - \hat{\boldsymbol{W}} \hat{\boldsymbol{S}}_z^{(i)} \boldsymbol{\epsilon} \right) \right. \\
& + \left. \log \left| \hat{\boldsymbol{\Sigma}}_x^{(i)} \right| + \operatorname{tr} \left[\hat{\boldsymbol{\Sigma}}_z^{(i)} \right] - \log \left| \hat{\boldsymbol{\Sigma}}_z^{(i)} \right| + \| \hat{\boldsymbol{\mu}}_z^{(i)} \|_2^2 \right] p(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon} \\
\leq & \int_{\| \sqrt{\alpha} \boldsymbol{\epsilon} \|_2 > \rho} \left[\left(\boldsymbol{x}^{(i)} - \hat{\boldsymbol{W}} \hat{\boldsymbol{\mu}}_z^{(i)} - \hat{\boldsymbol{W}} \hat{\boldsymbol{S}}_z^{(i)} \boldsymbol{\epsilon} \right)^\top \left(\hat{\boldsymbol{\Sigma}}_x^{(i)} \right)^{-1} \left(\boldsymbol{x}^{(i)} - \hat{\boldsymbol{W}} \hat{\boldsymbol{\mu}}_z^{(i)} - \hat{\boldsymbol{W}} \hat{\boldsymbol{S}}_z^{(i)} \boldsymbol{\epsilon} \right) \right. \\
& + \left. \log \left| \hat{\boldsymbol{\Sigma}}_x^{(i)} \right| + \operatorname{tr} \left[\hat{\boldsymbol{\Sigma}}_z^{(i)} \right] - \log \left| \hat{\boldsymbol{\Sigma}}_z^{(i)} \right| + \| \hat{\boldsymbol{\mu}}_z^{(i)} \|_2^2 \right] p(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon} \\
\leq & \int_{\| \sqrt{\alpha} \boldsymbol{\epsilon} \|_2 > \rho} \left[\frac{1}{\alpha} \left(\boldsymbol{x}^{(i)} - \boldsymbol{\Psi} \boldsymbol{\pi}^{(i)} - \sqrt{\alpha} \boldsymbol{\Psi} \boldsymbol{\epsilon} \right)^\top \left(\boldsymbol{x}^{(i)} - \boldsymbol{\Psi} \boldsymbol{\pi}^{(i)} - \sqrt{\alpha} \boldsymbol{\Psi} \boldsymbol{\epsilon} \right) \right. \\
& + \left. \kappa \alpha - \kappa \log \alpha + \| \boldsymbol{\pi}^{(i)} \|_2^2 \right] p(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon}, \tag{47}
\end{aligned}
$$

where the second inequality comes from setting $\hat{\boldsymbol{\Sigma}}_x^{(i)} = \alpha \boldsymbol{I}$ (its smallest possible value) in the inverse term and $\hat{\boldsymbol{\Sigma}}_x^{(i)} = \boldsymbol{I}$ (its largest value) in the log-det term. Next, given that

$$
\begin{aligned}
\| \boldsymbol{x}^{(i)} - \boldsymbol{\Psi} \boldsymbol{\pi}^{(i)} \|_2^2 & \leq \eta \\
\int_{\| \sqrt{\alpha} \boldsymbol{\epsilon} \|_2 > \rho} \left(\boldsymbol{\pi}^{(i)} \right)^\top \boldsymbol{\Psi}^\top \boldsymbol{\Psi} \boldsymbol{\epsilon} \cdot p(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon} & = 0 \\
\int_{\| \sqrt{\alpha} \boldsymbol{\epsilon} \|_2 > \rho} \| \boldsymbol{\Psi} \boldsymbol{\epsilon} \|_2^2 p(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon} & \leq \operatorname{tr} \left[\boldsymbol{\Psi}^\top \boldsymbol{\Psi} \right],
\end{aligned} \tag{48}
$$

it follows that the bound from (47) can be further reduced via

$$
\begin{aligned}
\mathcal{L}^{(i)}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\epsilon} \notin \mathcal{S}^{(i)}) & \leq \operatorname{tr} \left[\boldsymbol{\Psi}^\top \boldsymbol{\Psi} \right] + \int_{\| \sqrt{\alpha} \boldsymbol{\epsilon} \|_2 > \rho} \left[\frac{1}{\alpha} \eta + \kappa \alpha - \kappa \log \alpha + \| \boldsymbol{\pi}^{(i)} \|_2^2 \right] p(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon} \\
& = \Theta(1) + \left[\frac{1}{\alpha} \eta + \kappa \alpha - \kappa \log \alpha + \| \boldsymbol{\pi}^{(i)} \|_2^2 \right] \int_{\| \sqrt{\alpha} \boldsymbol{\epsilon} \|_2 > \rho} p(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon} \\
& \leq \Theta(1) + \left[\frac{1}{\alpha} \eta + \kappa \alpha - \kappa \log \alpha + \| \boldsymbol{\pi}^{(i)} \|_2^2 \right] \frac{\alpha}{\rho^2} \\
& = \Theta(1) + \Theta(\alpha^2) - \Theta(\alpha \log \alpha) \\
& = \Theta(1) \quad \text{as } \alpha \to 0, \tag{49}
\end{aligned}
$$

where the second inequality holds based on the vector version of Chebyshev's inequality, which ensures that

$$
\int_{\| \sqrt{\alpha} \boldsymbol{\epsilon} \|_2 > \rho} p(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon} = P(\| \sqrt{\alpha} \boldsymbol{\epsilon} \|_2 > \rho) \leq \frac{\alpha}{\rho^2}. \tag{50}
$$

Clearly then, as α becomes small, we have established that

$$
\mathcal{L}^{(i)}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\epsilon} \notin \mathcal{S}^{(i)}) \to O(1). \tag{51}
$$

D.3 Evaluation of $\mathcal{L}^{(i)}(\boldsymbol{\theta}, \boldsymbol{\phi}; \epsilon \in \mathcal{S}^{(i)})$

In analyzing $\mathcal{L}^{(i)}(\boldsymbol{\theta}, \boldsymbol{\phi}; \epsilon \in \mathcal{S}^{(i)})$, we note that

$$
\begin{aligned}
\int_{\boldsymbol{\epsilon} \in \mathcal{S}^{(i)}} & \left(\boldsymbol{x}^{(i)} - \hat{\boldsymbol{W}}\hat{\boldsymbol{\mu}}_z^{(i)} - \hat{\boldsymbol{W}}\hat{\boldsymbol{S}}_z^{(i)}\boldsymbol{\epsilon}\right)^{\top} \left(\hat{\boldsymbol{\Sigma}}_x^{(i)}\right)^{-1} \left(\boldsymbol{x}^{(i)} - \hat{\boldsymbol{W}}\hat{\boldsymbol{\mu}}_z^{(i)} - \hat{\boldsymbol{W}}\hat{\boldsymbol{S}}_z^{(i)}\boldsymbol{\epsilon}\right) p(\boldsymbol{\epsilon})d\boldsymbol{\epsilon} \\
& \leq \quad \int \left(\boldsymbol{x}^{(i)} - \boldsymbol{\Psi}\boldsymbol{\pi}^{(i)} - \sqrt{\alpha}\boldsymbol{\Psi}\boldsymbol{\epsilon}\right)^{\top} \left(\boldsymbol{\Lambda}^{(i)}\right)^{-1} \left(\boldsymbol{x}^{(i)} - \boldsymbol{\Psi}\boldsymbol{\pi}^{(i)} - \sqrt{\alpha}\boldsymbol{\Psi}\boldsymbol{\epsilon}\right) p(\boldsymbol{\epsilon})d\boldsymbol{\epsilon} \\
& \leq \quad \int \left(\boldsymbol{x}^{(i)} - \boldsymbol{\Psi}\boldsymbol{\pi}^{(i)}\right)^{\top} \left(\boldsymbol{\Lambda}^{(i)}\right)^{-1} \left(\boldsymbol{x}^{(i)} - \boldsymbol{\Psi}\boldsymbol{\pi}^{(i)}\right) p(\boldsymbol{\epsilon})d\boldsymbol{\epsilon} + \mathrm{tr}\left[\boldsymbol{\Psi}^{\top}\boldsymbol{\Psi}\right] \\
& \leq \quad \eta + \mathrm{tr}\left[\boldsymbol{\Psi}^{\top}\boldsymbol{\Psi}\right] \\
& = \quad \Theta(1)
\end{aligned}
\tag{52}
$$

given the alignment of $\boldsymbol{\Lambda}^{(i)}$ with zero-valued elements in $\boldsymbol{x}^{(i)} - \boldsymbol{\Psi}\boldsymbol{\pi}^{(i)}$. Furthermore, the remaining terms in $\mathcal{L}^{(i)}(\boldsymbol{\theta}, \boldsymbol{\phi}; \epsilon \in \mathcal{S}^{(i)})$ are independent of $\boldsymbol{\epsilon}$ giving

$$
\begin{aligned}
\int_{\boldsymbol{\epsilon} \in \mathcal{S}^{(i)}} & \left[\log\left|\hat{\boldsymbol{\Sigma}}_x^{(i)}\right| + \mathrm{tr}\left[\hat{\boldsymbol{\Sigma}}_z^{(i)}\right] - \log\left|\hat{\boldsymbol{\Sigma}}_z^{(i)}\right| + \|\hat{\boldsymbol{\mu}}_z^{(i)}\|_2^2\right] p(\boldsymbol{\epsilon})d\boldsymbol{\epsilon} \\
& = \quad \left[\log\left|\boldsymbol{\Lambda}^{(i)}\right| + \kappa\alpha - \kappa\log\alpha + \|\boldsymbol{\pi}^{(i)}\|_2^2\right] \int_{\|\sqrt{\alpha}\boldsymbol{\epsilon}\|_2 < \rho} p(\boldsymbol{\epsilon})d\boldsymbol{\epsilon} \\
& = \quad \left[(r^{(i)} - \kappa)\log\alpha + \kappa\alpha + \|\boldsymbol{\pi}^{(i)}\|_2^2\right] \int_{\|\sqrt{\alpha}\boldsymbol{\epsilon}\|_2 < \rho} p(\boldsymbol{\epsilon})d\boldsymbol{\epsilon} \\
& = \quad \left[(r^{(i)} - \kappa)\log\alpha\right] \int_{\|\sqrt{\alpha}\boldsymbol{\epsilon}\|_2 < \rho} p(\boldsymbol{\epsilon})d\boldsymbol{\epsilon} + O(\alpha) + O(1),
\end{aligned}
\tag{53}
$$

where

$$
r^{(i)} \triangleq \left|\left\{j : \Lambda_{jj}^{(i)} = \alpha\right\}\right| = d - \|\boldsymbol{s}^{(i)}\|_0.
\tag{54}
$$

Therefore, since $\int_{\|\sqrt{\alpha}\boldsymbol{\epsilon}\|_2 < \rho} p(\boldsymbol{\epsilon})d\boldsymbol{\epsilon} \to 1$ as α becomes small, we may conclude that

$$
\mathcal{L}^{(i)}(\boldsymbol{\theta}, \boldsymbol{\phi}; \epsilon \in \mathcal{S}^{(i)}) \to \left(d - \kappa - \|\boldsymbol{s}^{(i)}\|_0\right)\log\alpha + O(1),
\tag{55}
$$

D.4 Compilation of Candidate Solution Cost

After combining (51) and (55) across all i we find that

$$
\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) \to \sum_i \left(d - \kappa - \|\boldsymbol{s}^{(i)}\|_0\right)\log\alpha + O(1)
\tag{56}
$$

for any $\alpha \in (0, \bar{\alpha}]$ as $\bar{\alpha}$ becomes small. If $d > \kappa + \|\boldsymbol{s}^{(i)}\|_0$, then this expression will tend towards minus infinity, indicative of an objective value that is unbounded from below, certainly a fertile region for candidate minimizers. Note that per the theorem statement, $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{V}$ and $\boldsymbol{S}$ must represent a unique feasible solution to

$$
\min_{\boldsymbol{L}, \boldsymbol{S}} \quad d \cdot \mathrm{rank}[\boldsymbol{L}] + \|\boldsymbol{S}\|_0 \quad \text{s.t.} \quad \boldsymbol{X} = \boldsymbol{L} + \boldsymbol{S}.
\tag{57}
$$

Given that each column $\boldsymbol{x}^{(i)}$ has d degrees of freedom, then with $\boldsymbol{U}$ fixed there will be an infinite number of feasible solutions $\boldsymbol{x}^{(i)} = \boldsymbol{U}\boldsymbol{v}^{(i)} + \boldsymbol{s}^{(i)}$ such that $\dim[\boldsymbol{v}^{(i)}] + \|\boldsymbol{s}^{(i)}\|_0 = \kappa + \|\boldsymbol{s}^{(i)}\|_0 > d$ and a combinatorial number such that $k + \|\boldsymbol{s}^{(i)}\|_0 = d$. Therefore for uniqueness we require that $k + \|\boldsymbol{s}^{(i)}\|_0 < d$, so it follows that indeed $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi})$ will be unbounded from below as $\bar{\alpha}$ and therefore α becomes small, with cost given by (56) as a candidate solution satisfying the conditions of the theorem.

Of course it still remains possible that some other candidate solution could exist that violates one of these conditions and yet still achieves (56) or an even lower cost. We tackle this issue next. For this purpose our basic strategy will be to examine a *lower* bound on $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi})$ and show that essentially any candidate solution violating the theorem conditions will be worse than (56).

D.5 Evaluation of Other Candidate Solutions

To begin, we first observe that if granted the flexibility to optimize $\boldsymbol{\Sigma}_x^{(i)}$ independently over all values of $\boldsymbol{\epsilon}$ inside the integral for computing $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi})$, we immediately obtain a rigorous lower bound.[14] For this purpose we must effectively solve decoupled problems of the form

$$\inf_{\gamma > \alpha} \frac{c}{\gamma} + \log \gamma, \tag{58}$$

to which the optimal solution is just

$$\gamma^* = \xi_\alpha(x) \triangleq [c - \alpha]_+ + \alpha, \tag{59}$$

where the operator $[\cdot]_+$ retains only the positive part of its argument, setting negative values to zero. Plugging this solution back into (58), we find that

$$\inf_{\gamma > \alpha} \frac{c}{\gamma} + \log \gamma = \log \xi_\alpha(c) + O(1). \tag{60}$$

In the context of our bound, this leads to

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) \geq \sum_i \left\{ \mathbb{E}_{p(\boldsymbol{\epsilon})} \left[\sum_j \log \xi_\alpha \left(\left[x_j^{(i)} - \boldsymbol{w}_{j.}\boldsymbol{\mu}_z^{(i)} - \boldsymbol{w}_{j.}\boldsymbol{S}_z^{(i)}\boldsymbol{\epsilon} \right]^2 \right) \right] \right.$$
$$\left. + \operatorname{tr}\left[\boldsymbol{\Sigma}_z^{(i)} \right] - \log \left| \boldsymbol{\Sigma}_z^{(i)} \right| + \|\boldsymbol{\mu}_z^{(i)}\|_2^2 \right\} + O(1). \tag{61}$$

From this expression, it is clear that the lowest objective value we could ever hope to obtain cannot involve arbitrarily large values of $\boldsymbol{\Sigma}_z^{(i)}$ and $\boldsymbol{\mu}_z^{(i)}$ since the respective trace and quadratic terms grow faster than log-det terms. Likewise $\boldsymbol{\mu}_z^{(i)}$ cannot be unbounded for analogous reasons. Therefore, optimal solutions to (61) that will be unbounded from below must involve the first term becoming small, at least over a range of $\boldsymbol{\epsilon}$ values with significant probability measure. Although the required integral admits no closed-form solution, we can simplify things further using refinements of the above bound.

---

14. Note that this is never exactly achievable in practice, even with an infinite capacity network for computing $\boldsymbol{\Sigma}_x^{(i)}$, since it would require a unique network for each data sample; however, it nonetheless serves as a useful analysis tool.

For this purpose consider any possible candidate solution $\hat{\boldsymbol{W}} = \boldsymbol{\Psi}$ and $\hat{\boldsymbol{\mu}}_z^{(i)} = \boldsymbol{\pi}^{(i)}$ (not necessarily one that coincides with $\boldsymbol{U}$ and the optimal subspace), and define

$$\Delta_\alpha^{(i)}(\boldsymbol{\Psi}, \boldsymbol{\pi}) \triangleq \operatorname{supp}_\alpha \left[\boldsymbol{x}^{(i)} - \boldsymbol{\Psi}\boldsymbol{\pi} \right]. \tag{62}$$

Without loss of generality we also specify that

$$\boldsymbol{S}_z^{(i)} \triangleq \boldsymbol{\Xi}^{(i)} \boldsymbol{D}^{(i)}, \tag{63}$$

where $\boldsymbol{\Xi}^{(i)} \in \mathbb{R}^{d \times \kappa}$ has orthonormal columns and $\boldsymbol{D}^{(i)}$ is a diagonal matrix with

$$\left[\boldsymbol{D}^{(i)} \right]_{kk} = \xi_{\sqrt{\alpha}} \left(\sigma_k^{(i)} \right), \tag{64}$$

and $\boldsymbol{\sigma}^{(i)} = [\sigma_1^{(i)}, \ldots, \sigma_\kappa^{(i)}]^\top \in \mathbb{R}_+^\kappa$ is an arbitrary non-negative vector. Any general $\boldsymbol{\Sigma}_z^{(i)} = \boldsymbol{S}_z^{(i)} \left(\boldsymbol{S}_z^{(i)} \right)^\top$, with singular values bounded by α, is expressible via this format. We then reexpress (61) as

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) \geq{}& \sum_i \Bigg\{ \mathbb{E}_{p(\boldsymbol{\epsilon})} \Bigg[\sum_{j \in \Delta_\alpha^{(i)}(\boldsymbol{\Psi}, \boldsymbol{\pi}^{(i)})} \log \xi_\alpha \left(\left[x_j^{(i)} - \boldsymbol{\psi}_{j \cdot} \boldsymbol{\pi}^{(i)} - \boldsymbol{\psi}_{j \cdot} \boldsymbol{\Xi}^{(i)} \boldsymbol{D}^{(i)} \boldsymbol{\epsilon} \right]^2 \right) \Bigg] \\
&+ \mathbb{E}_{p(\boldsymbol{\epsilon})} \Bigg[\sum_{j \notin \Delta_\alpha^{(i)}(\boldsymbol{\Psi}, \boldsymbol{\pi}^{(i)})} \log \xi_\alpha \left(\left[\boldsymbol{\psi}_{j \cdot} \boldsymbol{\Xi}^{(i)} \boldsymbol{D}^{(i)} \boldsymbol{\epsilon} \right]^2 \right) \Bigg] \\
&+ \operatorname{tr} \left[\boldsymbol{\Xi}^{(i)} \left(\boldsymbol{D}^{(i)} \right)^2 \left(\boldsymbol{\Xi}^{(i)} \right)^\top \right] - \log \left| \boldsymbol{\Xi}^{(i)} \left(\boldsymbol{D}^{(i)} \right)^2 \left(\boldsymbol{\Xi}^{(i)} \right)^\top \right| + \|\boldsymbol{\pi}^{(i)}\|_2^2 \Bigg\} + O(1), \\
={}& \sum_i \Bigg\{ \mathbb{E}_{p(\boldsymbol{\epsilon})} \Bigg[\sum_{j \in \Delta_\alpha^{(i)}(\boldsymbol{\Psi}, \boldsymbol{\pi}^{(i)})} \log \xi_\alpha \left(\left[x_j^{(i)} - \boldsymbol{\psi}_{j \cdot} \boldsymbol{\pi}^{(i)} - \sum_k \bar{\psi}_{jk} \cdot \xi_{\sqrt{\alpha}} \left(\sigma_k^{(i)} \right) \cdot \epsilon_k \right]^2 \right) \Bigg] \\
&+ \mathbb{E}_{p(\boldsymbol{\epsilon})} \Bigg[\sum_{j \notin \Delta_\alpha^{(i)}(\boldsymbol{\Psi}, \boldsymbol{\pi}^{(i)})} \log \xi_\alpha \left(\left[\sum_k \bar{\psi}_{jk}^{(i)} \cdot \xi_{\sqrt{\alpha}} \left(\sigma_k^{(i)} \right) \cdot \epsilon_k \right]^2 \right) \Bigg] \\
&+ \sum_k \xi_\alpha \left[\left(\sigma_k^{(i)} \right)^2 \right] - \sum_k \log \xi_\alpha \left[\left(\sigma_k^{(i)} \right)^2 \right] + \|\boldsymbol{\pi}^{(i)}\|_2^2 \Bigg\} + O(1),
\end{aligned}
$$

(65)

(66)

where $\bar{\psi}_{jk}^{(i)}$ is the k-th element of the vector $\boldsymbol{\psi}_{j \cdot} \boldsymbol{\Xi}^{(i)}$. We can now analyze any given point $\{\boldsymbol{\Psi}, \boldsymbol{\pi}^{(i)}, \boldsymbol{\Xi}^{(i)}, \boldsymbol{\sigma}^{(i)}\}_{i=1}^n$ as α becomes small. The first term can be shown to be $\Theta(1)$ with all

other variables fixed,[15] leading to the revised bound

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) \;\geq\; \sum_i \left\{ \mathbb{E}_{p(\boldsymbol{\epsilon})} \left[\sum_{j \notin \Delta_\alpha^{(i)}(\boldsymbol{\Psi}, \boldsymbol{\pi}^{(i)})} \log \xi_\alpha \left(\left[\sum_k \bar{\psi}_{jk}^{(i)} \cdot \xi_{\sqrt{\alpha}} \left(\sigma_k^{(i)} \right) \cdot \epsilon_k \right]^2 \right) \right] \right. \tag{67}$$

$$\left. - \sum_k \log \xi_\alpha \left[\left(\sigma_k^{(i)} \right)^2 \right] \right\} + \Theta(1).$$

where the terms $\sum_k \xi_\alpha \left[\left(\sigma_k^{(i)} \right)^2 \right]$ and $\| \boldsymbol{\pi}^{(i)} \|_2^2$ have also been absorbed into $\Theta(1)$.

Given that

$$\mathbb{E}_{p(\boldsymbol{\epsilon})} \left[\log \xi_\alpha \left(\left[\boldsymbol{a}^\top \boldsymbol{\epsilon} \right]^2 \right) \right] = \log \xi_\alpha \left[\max_k \; (a_k)^2 \right] + O(1) \geq \log \alpha + O(1) \tag{68}$$

for any vector $\boldsymbol{a}$, we have the new bound

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) \tag{69}$$

$$\geq \sum_i \left\{ \sum_{j \notin \Delta_\alpha^{(i)}(\boldsymbol{\Psi}, \boldsymbol{\pi}^{(i)})} \log \xi_\alpha \left(\max_k \left[\bar{\psi}_{jk}^{(i)} \cdot \xi_{\sqrt{\alpha}} \left(\sigma_k^{(i)} \right) \right]^2 \right) - \sum_k \log \xi_\alpha \left[\left(\sigma_k^{(i)} \right)^2 \right] \right\} + \Theta(1).$$

If then we choose $\sigma_k^{(i)} = 0$ for all $i = 1, \ldots n$ and $k = 1, \ldots, \kappa$, then

$$\max_k \left[\bar{\psi}_{jk}^{(i)} \cdot \xi_{\sqrt{\alpha}} \left(\sigma_k^{(i)} \right) \right]^2 = \log \alpha + \Theta(1) \tag{70}$$

and we obtain the lower bound

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) \;\geq\; \sum_i \left(d - \kappa - \left| \Delta_\alpha^{(i)} \left(\boldsymbol{\Psi}, \boldsymbol{\pi}^{(i)} \right) \right| \right) \log \alpha + \Theta(1). \tag{71}$$

Additionally, if any set $\Delta_\alpha^{(i)} \left(\boldsymbol{\Psi}, \boldsymbol{\pi}^{(i)} \right)$ exists such that

$$\sum_i \left(d - \kappa - \left| \Delta_\alpha^{(i)} \left(\boldsymbol{\Psi}, \boldsymbol{\pi}^{(i)} \right) \right| \right) \leq \sum_i \left(d - \kappa - \| \boldsymbol{s}^{(i)} \|_0 \right), \tag{72}$$

then $\boldsymbol{s}^{(i)}$ cannot be part of the unique, feasible solution to (10), i.e., we could use the support pattern from each $\Delta_\alpha^{(i)} \left(\boldsymbol{\Psi}, \boldsymbol{\pi}^{(i)} \right)$ to find a different feasible solution with equal or lower value of $n \cdot \mathrm{rank} \left[\boldsymbol{L} \right] + \| \boldsymbol{S} \|_0$, which would violate either the uniqueness or optimality of the original solution. Therefore, we have established that with $\sigma_k^{(i)} = O(\alpha)$ for all i and k, the resulting bound on $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi})$ is essentially no better than (56), or the same bound we had before from our feasible trial solution. Moreover, the resulting $\hat{\boldsymbol{W}} = \boldsymbol{\Psi}$ that maximizes this bound, as well as the implicit

$$\hat{\boldsymbol{\Sigma}}_x^{(i)} \left(\hat{\boldsymbol{\mu}}_z \left[\boldsymbol{x}^{(i)} \right] \right) = \mathrm{diag} \left[\left(\boldsymbol{x}^{(i)} - \boldsymbol{\Psi} \boldsymbol{\pi}^{(i)} \right)^2 \right], \tag{73}$$

---

15. This is ultimately because $\int_{|c_1 + x| > \alpha} \log |c_1 + x| \mathcal{N}(x; 0, 1) dx < c_2 \int_{|c_1 + x| > \alpha} \log |c_1 + x| dx = \Theta(1)$ for any c_1 and a c_2 sufficiently large.

will necessarily satisfy (12). We then only need consider whether other choices for $\sigma_k^{(i)}$ can do better.

Let $\widetilde{\boldsymbol{\Psi}}^{(i)}$ denote the the rows of $\boldsymbol{\Psi}^{(i)}$ associated with row indeces $j \notin \Delta_\alpha^{(i)}\left(\boldsymbol{\Psi}, \boldsymbol{\pi}^{(i)}\right)$, meaning the indices at which we assume no sparse corruption term exists. Additionally, define $\boldsymbol{B}^{(i)} \triangleq \widetilde{\boldsymbol{\Psi}}^{(i)} \boldsymbol{\Xi}^{(i)}$. This implies that

$$
\sum_{j \notin \Delta_\alpha^{(i)}(\boldsymbol{\Psi}, \boldsymbol{\pi}^{(i)})} \log \xi_\alpha \left(\max_k \left[\bar{\psi}_{jk}^{(i)} \cdot \xi_{\sqrt{\alpha}} \left(\sigma_k^{(i)} \right) \right]^2 \right) - \sum_k \log \xi_\alpha \left[\left(\sigma_k^{(i)} \right)^2 \right] \tag{74}
$$
$$
= \sum_j \log \xi_\alpha \left(\max_k \left[B_{jk}^{(i)} \cdot \xi_{\sqrt{\alpha}} \left(\sigma_k^{(i)} \right) \right]^2 \right) - \sum_k \log \xi_\alpha \left[\left(\sigma_k^{(i)} \right)^2 \right].
$$

Contrary to our prior assumption $\boldsymbol{\sigma}^{(i)} = \boldsymbol{0}$, now consider any solution with $\|\boldsymbol{\sigma}^{(i)}\|_0 = \beta > 0$. For the time being, we also assume that $\boldsymbol{B}^{(i)}$ is full column rank. These conditions imply that

$$
\sum_j \log \xi_\alpha \left(\max_k \left[B_{jk}^{(i)} \cdot \xi_{\sqrt{\alpha}} \left(\sigma_k^{(i)} \right) \right]^2 \right) \geq \left(d - \beta - \left| \Delta_\alpha^{(i)} \left(\boldsymbol{\Psi}, \boldsymbol{\pi}^{(i)} \right) \right| \right) \log \alpha + \Theta(1) \tag{75}
$$

since at least β elements of the summation over j must now be order $\Theta(1)$. By assumption we also have $\sum_k \log \xi_\alpha \left[\left(\sigma_k^{(i)} \right)^2 \right] = (\kappa - \beta) \log \alpha + \Theta(1)$. Combining with (75), we see that such a solution is equivalent or worse than (71). So the former is the best we can do at any value of $\{\boldsymbol{\Psi}, \boldsymbol{\pi}^{(i)}, \boldsymbol{\Xi}^{(i)}, \boldsymbol{\sigma}^{(i)}\}_{i=1}^n$, provided that $\boldsymbol{B}^{(i)}$ is full rank, and obtaining the optimal value of $\Delta_\alpha^{(i)}\left(\boldsymbol{\Psi}, \boldsymbol{\pi}^{(i)}\right)$ implies that (12) holds.

However, if $\boldsymbol{B}^{(i)}$ is not full rank it would indeed entail that (74) could be reduced further, since a nonzero element of $\boldsymbol{\sigma}^{(i)}$ would not increase the first summation, while it would reduce the second. But if such a solution were to exist, it would violate the uniqueness assumption of the theorem statement. To see this, note that $\text{rank}[\boldsymbol{B}^{(i)}] = \text{rank}[\widetilde{\boldsymbol{\Psi}}^{(i)}]$ since $\boldsymbol{\Xi}^{(i)}$ is orthogonal, so if the former is not full column rank, neither is the latter. And if $\widetilde{\boldsymbol{\Psi}}^{(i)}$ is not full column rank, there will exist multiple solutions such that $\|\boldsymbol{x}^{(i)} - \boldsymbol{\Psi}\boldsymbol{\pi}^{(i)}\|_0 = \|\boldsymbol{s}^{(i)}\|_0$ or equivalently $\|\boldsymbol{x}^{(i)} - \boldsymbol{U}\boldsymbol{v}^{(i)}\|_0 = \|\boldsymbol{s}^{(i)}\|_0$ in direct violation of the uniqueness clause.

Therefore to conclude, a lower bound on the VAE cost is in fact the same order as that obtainable by our original trial solution. If this lower bound is not achieved, we cannot be at a minimizing solution, and any solution achieving this bound must satisfy (12).

D.6 Generalization to Case where $\kappa > \text{rank}[\boldsymbol{U}]$

Finally, we briefly consider the case where $\kappa > \text{rank}[\boldsymbol{U}] \triangleq \tau$, meaning that $\boldsymbol{W}$ contains redundant columns that are unnecessary in producing an optimal solution to (10). The candidate solution described in Section D.1 can be expanded via $\hat{\boldsymbol{W}} = \left[\boldsymbol{\Psi}, \ \boldsymbol{0}_{[d \times (\kappa - \tau)]} \right]$, $\boldsymbol{\mu}_z^{(i)} = \left[(\boldsymbol{\pi}^{(i)})^\top, \ \boldsymbol{0}_{[1 \times (\kappa - \tau)]} \right]^\top$, and $\hat{\boldsymbol{\Sigma}}_z^{(i)} = \text{diag}\left[\alpha \boldsymbol{1}_{[\tau \times 1]}^\top, \ \boldsymbol{1}_{[(\kappa - \tau) \times 1]}^\top \right]$ such that the same objective function value is obtained.

Now consider the general case where $\kappa \geq \text{rank}[\hat{\boldsymbol{W}}] > \tau$. If we review the lower bound described in Section D.5, with this general $\hat{\boldsymbol{W}}$ replacing $\boldsymbol{\Psi}$, it can be shown that $\hat{\boldsymbol{\Sigma}}_z^{(i)}$ will be

forced to have additional diagonal elements lowered to α, increasing the achievable objective by at least $-\log \alpha$ per sample. The details are not especially enlightening and we omit them here for brevity. Consequently, at any minimizer we must have $\text{rank}[\hat{\boldsymbol{W}}] = \tau$.

Appendix E. Proof of Corollary 4

Under the stated conditions, the partially-affine VAE cost simplifies to the function

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{W}, \boldsymbol{\Sigma}_x, \boldsymbol{\mu}_z) &= \sum_i \left\{ \left(\boldsymbol{x}^{(i)} - \boldsymbol{W}\boldsymbol{\mu}_z^{(i)}\right)^\top \left(\boldsymbol{\Sigma}_x^{(i)}\right)^{-1} \left(\boldsymbol{x}^{(i)} - \boldsymbol{W}\boldsymbol{\mu}_z^{(i)}\right) + \log \left|\boldsymbol{\Sigma}_x^{(i)}\right| + \|\boldsymbol{\mu}_z^{(i)}\|_2^2 \right\} \\
&= \sum_i \left\{ \left(\boldsymbol{x}^{(i)} - \boldsymbol{W}\beta^{-1}\beta\boldsymbol{\mu}_z^{(i)}\right)^\top \left(\boldsymbol{\Sigma}_x^{(i)}\right)^{-1} \left(\boldsymbol{x}^{(i)} - \boldsymbol{W}\beta^{-1}\beta\boldsymbol{\mu}_z^{(i)}\right) \right. \\
&\quad + \left. \log \left|\boldsymbol{\Sigma}_x^{(i)}\right| + \beta^2 \|\boldsymbol{\mu}_z^{(i)}\|_2^2 \right\},
\end{aligned}
\tag{76}
$$

where $\beta > 0$ is an arbitrary scaler, $\boldsymbol{\Sigma}_x^{(i)} \triangleq \boldsymbol{\Sigma}_x\left(\boldsymbol{\mu}_z^{(i)}; \boldsymbol{\theta}\right)$, and $\boldsymbol{\mu}_z^{(i)} \triangleq \boldsymbol{\mu}_z(\boldsymbol{x}^{(i)}; \boldsymbol{\phi})$. Taking the limit as $\beta \to 0^+$, we can minimize (76) while ignoring the $\beta^2 \|\boldsymbol{\mu}_z^{(i)}\|_2^2$ regularization factor. Consequently, we can without loss of generality consider minimization of

$$
\mathcal{L}(\boldsymbol{W}, \boldsymbol{\Sigma}_x, \boldsymbol{\mu}_z) \equiv \sum_i \left\{ \left(\boldsymbol{x}^{(i)} - \boldsymbol{W}\boldsymbol{\mu}_z^{(i)}\right)^\top \left(\boldsymbol{\Sigma}_x^{(i)}\right)^{-1} \left(\boldsymbol{x}^{(i)} - \boldsymbol{W}\boldsymbol{\mu}_z^{(i)}\right) + \log \left|\boldsymbol{\Sigma}_x^{(i)}\right| \right\},
\tag{77}
$$

ignoring any explicit reparameterization by β for convenience. If we optimize over $\boldsymbol{\Sigma}_x^{(i)}$ in the feasible region $\bar{\mathcal{S}}_\alpha^d$ and plug in the resulting value, then (77) reduces to the new cost

$$
\mathcal{L}(\boldsymbol{W}, \boldsymbol{\mu}_z) \equiv \sum_{i,j} \log \xi_\alpha \left(\left[x_j^{(i)} - \boldsymbol{w}_j.\boldsymbol{\mu}_z^{(i)}\right]^2\right),
\tag{78}
$$

an immaterial constant notwithstanding. Given that $\lim_{t\to 0} \frac{1}{t}\left(|x|^t - 1\right) = \log|x|$, and $\lim_{t\to 0} \sum_j |x_j|^p = \|\boldsymbol{x}\|_0$, then up to an irrelevant scaling factor and additive constant, the stated result follows.

Appendix F. Proof of Theorem 5

Based on the stated conditions, the VAE objective simplifies to

$$
\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_i \left\{ \mathbb{E}_{q_\phi(z|\boldsymbol{x}^{(i)})}\left[\frac{1}{\lambda_x}\left\|\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_x(z; \boldsymbol{\theta})\right\|^2\right] + d\log\lambda_x + \lambda_z - \log\lambda_z + \left(\boldsymbol{a}^\top \boldsymbol{x}^{(i)}\right)^2 \right\}.
\tag{79}
$$

Now choose some $\hat{\boldsymbol{a}}$ such that $\hat{\mu}_z^{(i)} = \hat{\boldsymbol{a}}^\top \boldsymbol{x}^{(i)}$ has a unique value for every sample $\boldsymbol{x}^{(i)}$ (here we assume that each sample is unique, although this assumption can be relaxed). We then define the function $h : \mathbb{R} \to \{1, \dots, n\}$ as

$$
h(z) \triangleq \arg \min_{i \in \{1,\dots,n\}} \|z - \mu_z^{(i)}\|_2.
\tag{80}
$$

and the piecewise linear decoder mean function

$$\boldsymbol{\mu}_x(z; \boldsymbol{\theta}) = \boldsymbol{x}^{(h(z))}. \tag{81}$$

Given these definitions, (79) becomes

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_i \left\{ \mathbb{E}_{q_\phi(z|\boldsymbol{x}^{(i)})} \left[\frac{1}{\lambda_x} \left\| \boldsymbol{x}^{(i)} - \boldsymbol{x}^{(h(z))} \right\|^2 \right] + d \log \lambda_x + \lambda_z - \log \lambda_z + \left(\hat{\mu}_z^{(i)} \right)^2 \right\}$$

$$= \sum_i \left\{ \mathbb{E}_{p(\epsilon)} \left[\frac{1}{\lambda_x} \left\| \boldsymbol{x}^{(i)} - \boldsymbol{x}^{(h[\hat{\mu}_z^{(i)} + \sqrt{\lambda_z}\epsilon])} \right\|^2 \right] + d \log \lambda_x + \lambda_z - \log \lambda_z + \left(\hat{\mu}_z^{(i)} \right)^2 \right\}. \tag{82}$$

Now define the set

$$\mathcal{S}^{(i)} \triangleq \left\{ \epsilon : h\left(\hat{\mu}_z^{(i)} + \sqrt{\lambda_z}\epsilon \right) = i \right\}, \tag{83}$$

which represents the set of ϵ that quantize to the correct index. We then have

$$\mathbb{E}_{p(\epsilon)} \left[\frac{1}{\lambda_x} \left\| \boldsymbol{x}^{(i)} - \boldsymbol{x}^{(h[\hat{\mu}_z^{(i)} + \sqrt{\lambda_z}\epsilon])} \right\|^2 \right]$$

$$= \int_{\epsilon \in \mathcal{S}^{(i)}} \left[\frac{1}{\lambda_x} \left\| \boldsymbol{x}^{(i)} - \boldsymbol{x}^{(h[\hat{\mu}_z^{(i)} + \sqrt{\lambda_z}\epsilon])} \right\|^2 \right] p(\epsilon) d\epsilon + \int_{\epsilon \notin \mathcal{S}^{(i)}} \left[\frac{1}{\lambda_x} \left\| \boldsymbol{x}^{(i)} - \boldsymbol{x}^{(h[\hat{\mu}_z^{(i)} + \sqrt{\lambda_z}\epsilon])} \right\|^2 \right] p(\epsilon) d\epsilon$$

$$= \int_{\epsilon \notin \mathcal{S}^{(i)}} \left[\frac{1}{\lambda_x} \left\| \boldsymbol{x}^{(i)} - \boldsymbol{x}^{(h[\hat{\mu}_z^{(i)} + \sqrt{\lambda_z}\epsilon])} \right\|^2 \right] p(\epsilon) d\epsilon$$

$$\leq \int_{\epsilon \notin \mathcal{S}^{(i)}} \frac{\eta}{\lambda_x} p(\epsilon) d\epsilon \tag{84}$$

$$= \frac{\eta}{\lambda_x} P\left(\epsilon \notin \mathcal{S}^{(i)} \right),$$

where

$$\eta \triangleq \max_{i,j \in \{1,\dots,n\}, i \neq j} \| \boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)} \|_2^2, \tag{85}$$

the maximal possible quantization error. Now we also define

$$\rho \triangleq \max_{i,j \in \{1,\dots,n\}, i \neq j} \frac{1}{2} \| \hat{\mu}_z^{(i)} - \hat{\mu}_z^{(j)} \|_2^2, \tag{86}$$

which is half the minimum distance between any two $\hat{\mu}_z^{(i)}$ and $\hat{\mu}_z^{(j)}$, with $i \neq j$. Then

$$P\left(\epsilon \notin \mathcal{S}^{(i)} \right) \leq P\left(\sqrt{\lambda_z}\epsilon > \rho \right)$$

$$\leq \frac{\lambda_z}{\rho^2} \tag{87}$$

by Chebyshev's inequality as was used in proving Theorem 3. This implies that (82) can be bounded via

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) \leq \sum_i \left\{ \frac{\eta}{\lambda_x} P\left(\epsilon \notin \mathcal{S}^{(i)} \right) + d \log \lambda_x + \lambda_z - \log \lambda_z + \left(\hat{\mu}_z^{(i)} \right)^2 \right\}$$

$$= \sum_i \left\{ \frac{\eta}{\rho^2} + (d-1) \log \alpha + \alpha + \left(\hat{\mu}_z^{(i)} \right)^2 \right\} \tag{88}$$

assuming we are at the trial solution $\hat{\lambda}_x = \hat{\lambda}_z = \alpha$. As we allow $\alpha \to 0$, this expression is unbounded from below, and as an upper bound on the VAE objective, the theorem follows. Incidentally, it should also be possible to prove that for α sufficiently small, no other solution can do appreciably better in terms of the dominate $(d-1)\log\alpha$ factor, but we will reserve this for future work.

References

[1] D.J. Bartholomew and M. Knott. Latent variable models and factor analysis. In *Kendalls library of statistics 7, 2nd Edition*, 1999.

[2] Y. Bengio. Learning deep architectures for ai. In *Foundations and Trends in Machine Learning*, 2009.

[3] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *International Conference on Computational Statistics*, page 177187, 2010.

[4] H. Bourlard and Y. Kamp. Auto-association by multilayer perceptrons and singular value decomposition. In *Biological Cybernetics*, 1988.

[5] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. 2015.

[6] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(2), May 2011.

[7] V. Chandrasekaran, S. Sanghavi, P.A. Parrilo, and A.S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optimization*, 21(3), June 2011.

[8] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1), 1999.

[9] A. Choromanska, M. Henaff, M. Mathieu, G. Ben Arous, and Y. LeCun. The loss surfaces of multilayer networks. In *International Conference on Artificial Intelligence and Statistics*, page 192204, 2015.

[10] A. Choromanska, Y. LeCun, and G. Ben Arous. Open problem: The landscape of the loss surfaces of multilayer networks. In *Conference on Learning Theory*, page 17561760, 2015.

[11] X. Ding, L. He, and L. Carin. Bayesian robust principal component analysis. *IEEE Trans. Image Processing*, 20(12), Dec. 2011.

[12] D.L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proc. National Academy of Sciences*, 100(5), 2003.

[13] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *Advances in Neural Information Processing Systems*, pages 658–666, 2016.

[14] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 35(11), 2013.

[15] D.J.H. Garling. *A Journey into Linear Analysis*. Cambridge University Press, 2007.

[16] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[17] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. In *arXiv:1406.2661*, 2014.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision*, 2015.

[19] M. Hoffman. Why variational inference gives bad parameter estimates. In *Advances in Variational Inference, NIPS Worshop*, 2014.

[20] K. Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems 29*, pages 586–594, 2016.

[21] T.H. Ohand H. Kim, Y.W. Tai, J.C. Bazin, and I.S. Kweon. Partial sum minimization of singular values in RPCA for low-level vision. In *International Conference on Computer Vision*, 2013.

[22] D. Kingma, D. Rezende, S. Mohamed, and M. Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, 2014.

[23] D. Kingma and M. Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[25] Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, pages 4743–4751, 2016.

[26] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015.

[27] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998.

[28] Zhouchen Lin, Minming Chen, and Yi Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[29] C. Lu, J. Feng, Z. Lin, and S. Yan. Correlation adaptive subspace segmentation by trace lasso. In *IEEE International Conference on Computer Vision*, 2013.

[30] L. Maaløe, C.K. Sønderby, S.K. Sønderby, and O. Winther. Auxiliary deep generative models. In *International Conference on Machine Learning*, 2016.

[31] E. Mansimov, E. Parisotto, J.L. Ba, and R. Salakhutdinov. Generating images from captions with attention. In *International Conference on Learning Representations*, 2016.

[32] V. Nair and G.E. Hinton. Rectified linear units improve restricted Boltzman machines. In *International conference on Machine Learning*, 2010.

[33] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.

[34] B.D. Rao, K. Engan, S. F. Cotter, J. Palmer, and K. Kreutz-Delgado. Subset selection in noise based on diversity measure minimization. *IEEE Trans. Signal Processing*, 51(3):760–770, March 2003.

[35] S. Rao, R. Tron, R. Vidal, and Y. Ma. Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(10), 2010.

[36] D.J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

[37] A.M. Saxe, J.L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

[38] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. How to train deep variational autoencoders and probabilistic ladder networks. In *arXiv:1602.02282*, 2016.

[39] M. Tipping and C. Bishop. Probabilistic principal component analysis. *J. Royal Statistical Society, Series B*, 61(3):611–622, 1999.

[40] Y. Wang, B. Dai, G. Hua, J. Aston, and D. Wipf. Green generative modeling: Recycling dirty data using recurrent variational autoencoders. In *Uncertainty in Artificial Intelligence*, 2017.

[41] D. Wipf. Non-convex rank minimization via an empirical Bayesian approach. In *Uncertainty in Artificial Intelligence*, 2012.