A Unified Latent Space disentangled VAE Framework with Robust Disentanglement Effectiveness Evaluation

Anonymous authors
Paper under double-blind review

Abstract

Evaluating and interpreting latent representations, such as variational autoencoders (VAEs), remains a significant challenge for diverse data types, especially when ground-truth generative factors are unknown. To address this, we unify several state-of-the-art disentangled VAE approaches for latent space disentanglement into one framework – bfVAE. To assess the effectiveness of a disentangled VAE model and enhance latent space interpretability, we propose Feature Variance Heterogeneity via Latent Traversal (FVH-LT) and Dirty Block Sparse Regression in Latent Space (DBSR-LS). To ensure robust interpretability of learned latent space, we develop a greedy alignment strategy (GAS) that mitigates label switching and aligns latent dimensions across runs to set the foundation of result aggregation. We also introduce a convenient scalar latent space separation index (LSSI) based on the GAS-aligned outputs of FVH-LT and DBSR-LS to summarize the overall latent structural separation without knowledge of the ground-truth generative factors. We compare bfVAE to five VAE models and validate the effectiveness FVH-LT, DBSR-LS, and LSSI in on seven tabular and image datasets. Under our examined experimental settings, bfVAE provides a more flexible disentanglement framework achieves more favorable overall trade-off between disentanglement and reconstruction than the benchmark VAE models; FVH-LT and DBSR-LS reliably uncover semantically meaningful and domain-relevant latent structures and generally yield consistent results; and LSSI makes an effective quantitative summary of latent structural separation.

1 Introduction

1.1 Background and Motivation

Understanding and interpreting latent representations in deep generative models (DGMs) remains a fundamental challenge in the development of efficient and trustworthy AI systems, primarily due to their black-box nature. Disentangled representation learning offers a promising solution by aiming to uncover independent and interpretable generative factors underlying complex data, providing a structured and interpretable latent space (Wang et al., 2024; Shen et al., 2022; Higgins et al., 2018; Paige et al., 2017; Bengio et al., 2013). Studies in different domains further underscore the practical benefits of disentangled representation learning, including medical imaging, single-cell transcriptomics, autonomous driving, and large language models (Liu et al., 2022; Baek et al., 2025; Qin et al., 2023; Pourkeshavarz et al., 2024), highlighting their positive impacts not only for interpretability but also for enhancing robustness in real-world AI systems.

We focus on variational autoencoders (VAEs) (Kingma & Welling, 2013), a widely popular DGM that utilizes latent spaces, in this work. In particular, disentangled VAEs aim to learn structured latent spaces in which distinct latent dimensions (LDs) capture independent and semantically meaningful factors of variation in the data, while retaining reconstruction quality. Disentangled VAEs have been employed in different domains and continue to drive methodological advances, such as in high-resolution image reconstruction (Uppal et al., 2025) and data compression and controlled data augmentation in Global Navigation Satellite System interference classification (Heublein et al., 2025).

Most existing work on disentangled VAE focuses on image data, with relatively limited attention to other data modality, tabular data included. Our empirical studies suggest current disentangled VAE formulations,

although effective for image data, may be ineffective for tabular data. Additionally, current disentanglement evaluation metrics are applicable only to synthetic data where ground-truth generative factors are known to enable quantitative evaluation (Higgins et al., 2017; Kim & Mnih, 2018). Other approaches rely largely on qualitative assessment by visualizing reconstructed images to assign semantic meaning to each LD – an approach not applicable to other data types that lack natural visual representations. In summary, both lines of work have key limitations in their disentanglement evaluation in practice (Locatello et al., 2018).

In summary, there remains a lack of a general framework for disentangled VAEs and disentanglement evaluation techniques and metrics for different data modalities especially when ground-truth generative factors can not be precisely defined – the typical case in real-world applications. Furthermore, the inherent instability of deep neural networks training (e.g., random initialization, stochastic optimization, and local minima, etc) exacerbates the evaluation challenges. Some previous work may ignore the instability and report results from a single run, making results hard to reproduce. Finally, latent spaces are subject to label-switching, where indexing of latent factors changes across independent runs, further complicating the interpretability and reproducibility of the learned representations (Rodríguez & Walker, 2014; Stephens, 2000).

1.2 Our Contributions and Findings

We propose *bfVAE* that unifies several state-of-the-art disentangled VAE models in a coherent framework. To address the challenge of interpreting latent representations when ground-truth generative factors are unknown, we introduce *Feature Variance Heterogeneity via Latent Traversal (FVH-LT)* and *Dirty Block Sparse Regression in Latent Space (DBSR-LS)*. Building on the outputs of FVH-LT and DSBR, we further define the *Latent Space Separation Index (LSSI)*, a scalar non-causal surrogate summary of structural separation among latent dimensions (LDs) when the underlying generative factors and their correspondence to latent dimensions are unavailable. To address the instability and uncertainty during VAE training and ensure robustness of FVH-LT and DBSR-LS, we recommend training VAEs multiple times and aggregating the resulting interpretations. Since LDs are subject to label-switching across independent runs, we develop a *greedy alignment strategy (GAS)* to align LDs across models before result aggregation. For each proposed method, we provide the underlying intuition and, where appropriate, the corresponding mathematical rationale.

We run extensive experiments on synthetic and real data of various types and sizes to evaluate the united bfVAE framework and the proposed interpretability tools FVH-LT, DBSR-LS, GAS and LSSI. Across all datasets, bfVAE provides overall more favorable trade-off between disentanglement and reconstruction compared to the benchmark VAE frameworks in the studied experimental settings due to the flexibility in its formulation. LSSI agrees well with existing supervised disentanglement benchmarks, supporting its effectiveness as a metric for latent structural separation. Both FVH-LT and DBSR-LS consistently and reliably uncover disentangled and semantically meaningful LDs. To our knowledge, FVH-LT and DBSR-LS are the first techniques, coupled with the GAS procedure, that evaluate the latent space disentanglement effectiveness in VAEs for multiple data types without precise knowledge of ground-truth generative factors.

1.3 Related Work

Building on the classical VAE formulation, various extensions have been proposed to promote disentanglement of latent space for enhanced interpretability, while balancing reconstruction fidelity. β-VAE (Higgins et al., 2017) introduces a hyperparameter β that scales the Kullback–Leibler (KL) divergence in the classical VAE objective with $\beta = 1$ – which we refer to as vanilla VAE in this work. Larger β may lead to a higher degree of disentanglement. Disentangled β-VAE (Burgess et al., 2018) introduces an additional capacity parameter C to β-VAE that controls the information transmitted from the input space to the latent space. Factor-VAE (Kim & Mnih, 2018) adds a total correlation penalty to the vanilla VAE, reducing dependencies among the LDs. β-TCVAE (Chen et al., 2018) arrives at a closely related objective by decomposing the KL term in the ELBO into index-code mutual information, total correlation, and dimension-wise KL, and promotes disentanglement by imposing a stronger penalty on the total correlation term. DIP-VAE (Kumar et al., 2017) penalizes the off-diagonal elements in the covariance matrix of the expected posterior means over the marginal distribution of data (DIP-VAE-I) or the posterior covariance matrix over its marginal variational distribution (DIP-VAE-II). Li et al. (2025) generalizes the total correlation penalty to a partial correlation penalty. Disentanglement is also explored in conditional VAE (cVAE) models (Sohn et al.,

2015; Harvey et al., 2021; Zhang et al., 2020) and integrated structural causal models with a bidirectional generative process to achieve causal disentanglement (Shen et al., 2022).

Prior work has proposed disentanglement metrics that require ground-truth generative factors. Higgins et al. (2017) measure disentanglement by fixing one factor, varying the others, computing average latent differences, and training a classifier to identify the fixed factor. The factor-VAE metric (Kim & Mnih, 2018) similarly uses latent differences, selecting the LD with the smallest empirical variance and applying majority voting. SAP (Kumar et al., 2017) measures the gap between the top two LDs predicting each factor, while Chen et al. (2018) use mutual information between top LDs and factors. DCI (Eastwood & Williams, 2018) trains regressors to estimate each LD's importance for predicting generative factors.

2 A Unified Disentangled VAE framework

In this section, we first introduce bfVAE as a unified framework for disentangled VAE, and then provide an information bottleneck perspective on the formulation.

2.1 bfVAE

bfVAE extends the existing β-VAE and factor-VAE models (this the name the name bfVAE). Denote the input by $\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^n$, where $\mathbf{x}_i$ denotes the i-th observed sample. Consider a standard VAE setting with prior $p(\mathbf{z}_i) = \mathcal{N}_K(0, I)\,\forall i$ for the latent variables $\mathbf{Z}$ of dimension K. Denote the variational posterior of $\mathbf{Z}$ by $q(\mathbf{z}_i|\mathbf{x})$. The bfVAE loss is defined as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\omega}) =& \frac{1}{n}\sum_{i=1}^{n}\Big\{\underbrace{-\mathbb{E}_{q(\mathbf{z}_i|\mathbf{x}_i,\boldsymbol{\phi})}\log p(\mathbf{x}_i \mid \mathbf{z}_i, \boldsymbol{\omega})}_{\text{T1: reconstruction error}} + \underbrace{\beta \cdot |\mathrm{KL}\left(q(\mathbf{z}_i \mid \mathbf{x}_i, \boldsymbol{\phi})\|p(\mathbf{z}_i)\right) - C|}_{\text{T2: capacity constraint}}\Big\} \\ &+ \quad \gamma \cdot \underbrace{\mathrm{KL}\left(q(\mathbf{z})\|\bar{q}(\mathbf{z})\right)}_{\text{T3: disentanglement regularization}},\end{aligned} \tag{1}$$

where $\boldsymbol{\omega}$ and $\boldsymbol{\phi}$ are unknown model parameters; and $\beta \geq 0, \gamma \geq 0$, and $C \geq 0$ are hyperparameters.

Different terms in Eq. 1 serve different purposes. T1 ensures reconstruction fidelity by maximizing the expected log-likelihood. T2 penalizes the distance between the prior-posterior KL divergence and target C: if KL $= C$, T2 disappears. C is referred to as the target "capacity in that it constrains the latent space's capacity – larger C allows the posterior to extract more information from the input data and to drift further from the prior. During training, C starts at 0 and gradually increases to reach the pre-specified value for a controlled capacity expansion. T3 encourages statistical independence among the LDs, where $q(\mathbf{z}) := n^{-1}\sum_i q(\mathbf{z}|\mathbf{x}_i)$ and $\bar{q}(\mathbf{z}) := \prod_{j=1}^{K} q(z_j)$. Some existing VAE models are special cases of Eq. 1 by setting β, γ, and C at specific values: when $\gamma = 0, C = 0$, bfVAE reduces to β-VAE and when $\beta = 1, C = 0$, it reduces to factor-VAE; when $\gamma = 0, C = 0, \beta = 1$, it becomes the vanilla VAE.

2.2 An Information Bottleneck Perspective on bfVAE

We can interpret the bfVAE formulation in Eq. 1 via the variational information bottleneck framework (Alemi et al., 2016; Burgess et al., 2018). First, the bfVAE formulation in Eq. 1 is not a classical information bottleneck objective in the form of max $\left\{I(\mathbf{Z};\mathbf{Y}) - \beta I(\mathbf{X};\mathbf{Z})\right\}$[1] (Tishby et al., 2000), but inspired by it. Each of the three terms in Eq. 1 plays a different but complementary role in shaping the latent space in bfVAE.

Specifically, T1 in Eq. 1 encourages $\mathbf{z}$ to retain information needed to reconstruct the input $\mathbf{x}$. If optimized alone, this term would favor highly informative latent representations, resembling a standard autoencoder, with little pressure to compress or organize the latent space. T2 constraints the mutual information between $\mathbf{x}$ and $\mathbf{z}$ and encourages the aggregated posterior $q(\mathbf{z})$ to match the prior, which can be easily seen by decomposing the KL divergence $\mathbb{E}_{p(\mathbf{x})}\left[\mathrm{KL}(q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})\|p(\mathbf{z}))\right] = I_q(\mathbf{x};\mathbf{z}) + \mathrm{KL}(q(\mathbf{z})\|p(\mathbf{z}))$. C and β in T2 have distinct roles. C specifies the *target bottleneck capacity* or the desired amount of information that $\mathbf{z}$ is allowed to carry: smaller C imposes a narrow bottleneck and forces aggressive compression from $\mathbf{x}$ to $\mathbf{z}$, whereas

[1] $\mathbf{X}$ is the input, $\mathbf{Y}$ is the target, and $\mathbf{Z}$ is a representation of $\mathbf{X}$ that is maximally informative about $\mathbf{Y}$ while being as compressed as possible with respect to $\mathbf{X}$. β acts as the Lagrange multiplier that regulates information of $\mathbf{X}$ preserved in $\mathbf{Z}$.

larger C permits more information to pass through. β, on the other hand, controls *how strongly this target is enforced*. In β-VAE, the KL term is simply controlled by β. By introducing C, the model can control the bottleneck more directly in the sense that it is not merely encouraged to reduce KL, but rather to allocate approximately C units of latent capacity, leading to more direct control over the reconstruction-compression tradeoff. T3 penalizes redundancy and dependence among $\mathbf{z}$, encouraging the retained information to be organized into approximately independent and factorized components. Different from T2, which primarily controls how much information is transmitted, T3 determines how the transmitted information is structured after passing through the bottleneck. In summary, the bfVAE formulation extends the standard variational bottleneck from controlling only the quantity of latent information to also shaping its structure across dimensions. The resulting latent space is expected to be not only informative and compressed, but also more interpretable and disentangled.

The interpretation above considers the latent representation as a whole. At a finer level, the effect of Eq. 1 can be expressed unevenly across individual LDs. Specifically, under the joint action of T1, T2, and T3, some LDs actively encodes nontrivial information in $\mathbf{x}$, whereas others contribute little. This motivates a dimension-wise characterization of the latent space, formalized in Def. 1. This distinction provides a more refined view of how the bottleneck allocates information across LDs.

Definition 1 (informative and noninformative latent dimensions) *A latent dimension j in a VAE is informative if the expected KL divergence between its posterior and prior distributions exceeds an information threshold ϵ:*

$$\mathbb{E}_{p(\mathbf{x})}\left[\mathrm{KL}(q_{\boldsymbol{\phi}}(z_j|\mathbf{x},\boldsymbol{\phi})\|p(z_j))\right] > \epsilon, \text{ where } \epsilon \gg 0. \tag{2}$$

A latent dimension j is non-informative (or collapsed) if its posterior conveys negligible information about the input $\mathbf{x}$:

$$\mathbb{E}_{p(\mathbf{x})}\left[\mathrm{KL}(q_{\boldsymbol{\phi}}(z_j|\mathbf{x},\boldsymbol{\phi})\|p(z_j))\right] < \delta, \text{ for a vanishingly small } \delta. \tag{3}$$

In the case of $p(\mathbf{z}) = \mathcal{N}(0,1)$ and $q(z_j|\mathbf{x}) = \mathcal{N}(\mu_j(\mathbf{x}), \sigma_j^2(\mathbf{x}))$ – common choices in VAEs, an informative LD j means $q(z_j|\mathbf{x})$ has successfully "escaped the information bottleneck, achieved through either a non-zero $\mu_j(\mathbf{x})$ or $\sigma_j^2(\mathbf{x})$ away from 1. When $q(z_j|\mathbf{x},\boldsymbol{\phi}) \approx p(z_j)$, the j-th LD effectively behaves as noise. We note that while an informative LD encodes meaningful information in $\mathbf{x}$, it does not automatically offer semantic interpretation on what information it encodes (thus the need for techniques like FVH-LT and DBSR-LS that are introduced in Sections 3.1 and 3.2).

We hypothesize that the relative prevalence of informative and non-informative LDs is governed jointly by C and β of T2, but also relates to the input type and characteristics of $\mathbf{x}$. Larger C generally leads more informative LDs, particularly when accompanied by large β, which strongly enforces the target. For fixed large C, a small β may yield an intermediate regime in which the separation between informative and non-informative LDs is less distinct. In contrast, smaller C imposes a tighter bottleneck and leads to more non-informative LDs; similarly, this effect becomes stronger as β increases. Previous work on β-VAE for image data often adopts relatively large β; since $C = 0$ in standard β-VAE, this can be viewed as placing strong pressure toward a low-capacity information bottleneck. This forces the model to encode only the aspects of $\mathbf{x}$ that most improve reconstruction, which may work well for images as high-dimensional images often contain substantial redundancy across input features. In contrast, for relatively low-dimensional $\mathbf{x}$ or when semantics are more dispersed across $\mathbf{x}$, as is often the case for tabular data, there is typically less redundancy and less compressible information. Applying the same bottleneck setting may discard substantial signal, reduce the number of "true informative LDs, and degrade reconstructed or synthetic data. Instead, a larger C with an appropriately tuned β may allow the latent space to retain more nontrivial information from $\mathbf{x}$. Our experiments in Sec. 4 provide empirical evidence in support of this claim.

3 Latent Space Interpretability Techniques

While bfVAE promotes informative and orthogonal LDs through capacity and dependence regularization, it itself does not explicitly attach semantic meaning to each LD or identify feature–latent associations, motivating us to propose FVH-LT and DBSR-LS as dedicated latent space interpretability techniques to quantify, visualize, and interpret the disentangled LDs To enable stable aggregation of FVH-LT and DBSR-LS results across multiple evaluation runs, we introduce the GAS procedure to resolve LD misalignment and

set the foundation for result aggregation. Finally, we proposed the LSSI to quantify the degree of structural separation captured by the latent–feature association matrix output from FVH-LT and DBSR-LS.

3.1 FVH-LT

FVH-LT quantifies the associations between input features and LDs through LT. As illustrated in Fig. 1, the trained VAE encoder first maps each input instance to its latent representation. FVH-LT then varies one LD at a time while holding the remaining LDs fixed, and measures the resulting changes in the outputs reconstructed by the decoder. The resulting LD-feature association matrix $\mathbf{S}$ summarizes how strongly each feature responds to variations in each LD and can be conveniently visualized as a heatmap, providing an interpretable understanding of individual LD. FVH-LT does not require access to ground-truth generative factors or external supervision, making it applicable to real-world data where such information is unavailable.

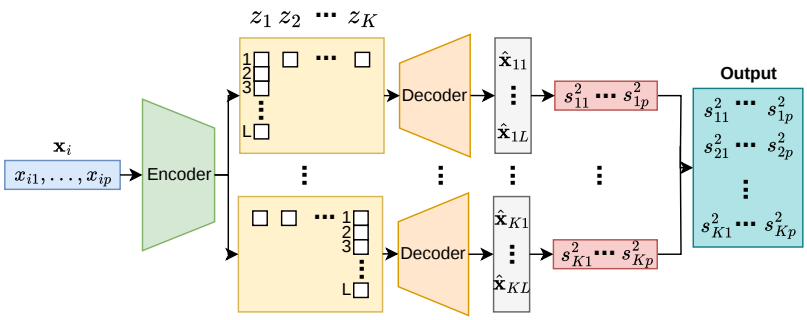


Figure 1: The FVH-LT Workflow. $\mathbf{x}_i$ contains input features for the i-th observation; z_k is the k-th LD and $\hat{\mathbf{x}}_{ikl}$ for $k=1,\ldots,K$ and $l=1,\ldots,L$ is reconstructed output for $\mathbf{x}_i$ at l-th fixed value from the LT of z_k; s_{kj} is the sample variance of the restructured data for the j-th feature from the LT on LD k.

Alg. 1 presents the pseudo-code for FVH-LT. The LT ranges can be defined in several ways: 1) a fixed range across all n input samples and K LDs; 2) a data- and dimension-dependent range, such as $[\mu_{ik} - c\sigma_{ik}, \mu_{ik} + c\sigma_{ik}]$, where $\hat{\mu}_{ik}$ and $\hat{\sigma}_{ik}^2$ are the learned posterior mean and variance of the k-th LD for observation i; 3) a combination of the previous two: range centered at μ_{ik} with a global interval half-width. In our experiments, we find the latter two strategies are more effective in identifying informative LDs as the fixed LT ranges in the first may fail to probe high-density regions of the posterior when μ_{ik} deviates from 0, reducing sensitivity of the LT. As for the number of LT steps L, a large L creates a fine grid for LT to better detect pattern changes across the LD; a small L would lead to a coarse grid, potentially obscuring meaningful signals, but at a lower computational cost. Because the true number of informative LDs K_0 is unknown, one needs to supply an estimate K. The good news is that FVH-LT is robust to the choice of K as long as $K > K_0$.

FVH-LT can be interpreted as a decoder-mediated latent-to-feature response analysis. For notation simplicity, we drop the index r in the following analysis. Denote the learned decoder by $\hat{g} : \mathbb{R}^K \to \mathbb{R}^p$. Applying a first-order Taylor expansion and the variance operator to g at a fixed point z_{ik}^* along LD k for sample $\mathbf{x}_i$, we have $\hat{g}_j(z_{ik}^* + \delta_{ikl}) \approx \hat{g}_j(z_{ik}^*) + \delta_{ikl}\left[J_{\hat{g}}(\mathbf{z}_i^*)\right]_{jk}$, where $\mathbf{z}_i^* = (\mathbf{z}_{i,-k}, z_{ik}^*)$, $J_{\hat{g}}(\mathbf{z}_i^*) \in \mathbb{R}^{p\times K}$ is the decoder Jacobian, and δ_{ikl} can be regarded as the adversarial perturbation in LD k from LT, and

$$s_{ijk}^2 = \hat{\mathrm{V}}\left[\hat{g}_j(z_{ik}^* + \delta_{ikl})\right] \approx \left[J_{\hat{g}}(\mathbf{z}_i^*)\right]_{jk}^2 \hat{\mathrm{V}}(\delta_{ikl}).$$

If the LT points are equally spaced over the LT range $(-a_{ik}, a_{ik})$, which is what we use in the experiments, then $\widehat{\mathrm{Var}}(\delta_{ikl}) = (2a_{ik})^2 L(L+1)/(12(L-1)^2) = a_{ik}^2 L(L+1)/(3(L-1)^2)$. In other words, FVH-LT measures the feature-level variation induced by controlled finite-grid perturbations of each LD. When the LT-grid variance is fixed across LDs, the FVH-LT entry is proportional to the squared local sensitivity of reconstructed feature j to perturbations in LD k.

3.2 DBSR-LS

DBSR promotes shared-sparsity pattern and task-specific sparsity pattern in regression coefficients in a multi-task supervised learning setting (Jalali et al., 2010). We make a novel use of DBSR towards latent

Algorithm 1 The FVH-LT procedure

Input: training data $\mathbf{x}_{n\times p}$, prior $\mathbf{z} = (z_1 \ldots, z_K) \sim f(\mathbf{z})$, LT range $[-a_{ik}, a_{ik}]$ for $k = 1, \ldots, K$ and $i=1, \ldots, n$, LT steps L, run number R, hyper-parameter initialization
Output: Variance matrix S of reconstructed $\hat{\mathbf{x}}$ given LT in $\mathbf{z}$; posterior-prior KL divergence
Pre-processing: standardize $\mathbf{x}$ as needed
for $r = 1$ to R **do**
    Train VAE$_r$ on $\mathbf{x}$ with the loss function in Eq. 1
    for $i = 1$ to n **do**
        for $k = 1$ to K **do**
            Calculate $d_{i,rk} = D_{\mathrm{KL}}(\hat{q}_k(z_{ik}|\mathbf{x}_i)||p(z_{ik}))$
            Generate $\mathbf{z}_{ik} = \{z_{ikl}\}_{k=1,\ldots,L}$ over its LT range $[-a_{ik}, a_{ik}]$, fixing $\mathbf{z}_{i,-k}$ at randomly sampled values from $\hat{q}_{-k}(\mathbf{z}_{i,-k}|\mathbf{x}_i)$
            Feed $\{z_{ikl}, \mathbf{z}_{i,-k}\}$ to the decoder of VAE$_r$ to reconstruct $\{\hat{x}_{irjl,k}\}_{j=1,\ldots,p}$ for $l = 1$ to L
            Calculate sample variance $s^2_{irj,k}$ of $(\hat{x}_{irj1}, \ldots, \hat{x}_{irjL})$ for $j = 1$ to p
        end for
    end for
    Calculate $\bar{s}^2_{rjk} = n^{-1} \sum_{i=1}^{n} s^2_{irj,k}$
end for
Apply GAS (Alg. 3) to align $\bar{s}^2_{rj,k}$ across $r = 1, \ldots, R$ runs
Compute $s^2_{jk} = \frac{1}{R}\sum_{r=1}^{R} \bar{s}^2_{rj,k}$ for $j=1, \ldots, p$ and $d_{ik} = \frac{1}{R}\sum_{r=1}^{R} d_{i,rk}$ for $k=1, \ldots, K$; $i=1, \ldots, n$
Return variance matrix $S = [s^2_{jk}]_{j=1,\ldots,p;k=1,\ldots,K}$ and KL divergence $\mathbf{d} = \{d_{ik}\}_{i=1,\ldots,n;k=1,\ldots,K}$

space separation and interpretability by treating each LD as the outcome of a different regression task and leveraging the shared and task-specific sparsity to identify feature–latent associations. This formulation decomposes the associations into components unique to individual LDs and those shared across multiple LDs, providing a structured view of how information is distributed across LDs. In particular, DBSR-LS estimates two regression coefficient matrices D and B via minimizing the following loss

$$\mathcal{L}(D, B) = (2n)^{-1} \sum_{k=1}^{K} \left\| \boldsymbol{\mu}_Z^{(k)} - \left(D_{[k,]} + B_{[k,]}\mathbf{x}^{(k)}\right) \right\|_2^2 + \lambda_D \|D\|_{1,1} + \lambda_B \|B\|_{1,\infty}. \tag{4}$$

$\boldsymbol{\mu}^{(k)}_{Z,n\times 1}$ contains the posterior means of the k-th latent variables of the n samples and $\mathbf{x}^{(k)}_{n\times p}$ is the corresponding design matrix comprising the observed features. D and B are both of dimension $K \times p$ with $D_{[k,]}$ and $B_{[k,]}$ denoting their k-th columns. While both $\|D\|_{1,1} = \sum_{j=1}^{p}\sum_{k=1}^{K} |D_{kj}|$ and $\|B\|_{1,\infty} = \sum_{j=1}^{p} \max_k |B_{kj}|$ promotes sparsity, the former captures the latent-dimension-specific sparsity whereas the latter represents shared sparse structures across the LDs. As a result, nonzero entries in D indicate features associated distinct LDs. Eq. 4 is formulated using the posterior means $\boldsymbol{\mu}_Z$ as the response variable. An alternative is to use random samples from the posterior distribution though we expect $\boldsymbol{\mu}_Z$ yields more stable estimates.

Alg. 2 lists the pseudo-code for DBSR-LS. For hyperparameters shared with Alg. 1, their specifications are similar. For those unique to Alg. 2 (i.e., $\lambda_B \geq 0, \lambda_D \geq 0$), they can be tuned to balance sparsity in the regression coefficients with prediction accuracy – large values may lead to overly sparse $\hat{D}$ while small values may limit the disentanglement capacity of DBSR-LS.

DBSR-LS is proposed as an interpretable linear surrogate for summarizing first-order LD-feature associations and is not intended to replace the encoder or provide a complete predictive model of $q_\phi(z|x)$ (the encoder itself would be sufficient for that reason and there would be no need to train a linear surrogate). The dirty-model structure is useful because it separates shared associations from LD-specific associations. If different LDs are genuinely separated, their sparse feature-association profiles should differ; if multiple LDs rely on the same features, the DBSR-LS matrix will reveal this overlap. DBSR-LS estimates LD-feature associations directly from learned latent space and is not decoder-mediated like FVH-LT; it may thus serve as a useful complementary approach when decoder-based LT is unstable.

Algorithm 2 The DBSR-LS procedure

1: **Input**: training data $\mathbf{x}_{n\times p}$, prior $\mathbf{z} = (z_1 \ldots, z_r) \sim f(\mathbf{z})$, initialized hyper-parameters (λ_B, λ_D, those in VAE), run number R
2: **Output**: latent-dimension-specific aestimated coefficient matrix $\hat{D}$ and its absolute value $|\hat{D}|$
3: Pre-processing: standardize $\mathbf{x}$ as needed
4: **for** $r = 1$ to R **do**
5: Train VAE$_r$ on $\mathbf{x}$ with the loss in Eq. 1 and extract posterior means $\boldsymbol{\mu}_{n\times K,r}$ of $q(\mathbf{z}|\mathbf{x})$
6: Compute $d_{r,ik} = D_{\mathrm{KL}}(q_k(z_{ik}|\mathbf{x})||p(z_{ik}))$ for $k = 1, \ldots, K$ and $i = 1, \ldots, n$
7: Run DBSR-LS with $\mathbf{x}$ as the design matrix and $\boldsymbol{\mu}_r[, k]$ as $Y^{(k)}$ to estimate D_r and B_r
8: **end for**
9: Apply GAS (Alg. 3) to align D_r and $|D_r|$, respectively, across $r = 1, \ldots, R$
10: Calculate the averages $|\hat{D}| = R^{-1} \sum_{r=1}^{R} |D_r|$.

3.3 GAS for Latent Space Alignment

Aggregating FVH-LT output S and the DBSR-LS output $|\hat{D}|$ in Algs. 1 and 2 across multiple runs can stabilize learned LD-feature associations results. However, such aggregation is complicated by label switching in the latent space.[2] Consequently, latent representations are not directly comparable on a dimension-by-dimension basis cross runs, and naïve aggregation would yield misleading results.

To address this issue, we introduce GAS to align LDs in S and $|\hat{D}|$ before their aggregation. In a nutshell, GAS clusters LDs in each run based on their informativeness defined in Def. 1, selects the run with the maximum number of informative LDs as the reference, and uses it as an anchor for alignment. The matched LDs are then re-indexed to establish consistent LD correspondence across runs. Alg. 3 lists the pseudo-code, which has been integrated in Alg. 1 (Line 16) and Alg. 2 (Line 9), and Fig. 2 illustrates the workflow.

Algorithm 3 The GAS

1: **Input**: KL divergence $\mathbf{d} = \{d_{rk}\}_{r=1,\ldots,R;\ k=1,\ldots,K}$, matrix $\{A_r\}_{r=1,\ldots,R}$ to be aligned (variance matrix S from FVH-LT, coefficient matrix $|D|$ from DBSR-LS), correlation threshold ρ
2: **Output**: aligned A
3: In each run $r = 1$ to R, cluster the LDs into two groups based on $\{d_{rk}\}_{k=1,\ldots,K}$; define the cluster with the higher average KL as the informative set $\mathcal{I}_r$.
4: Define $r^* \triangleq \arg\max_{r=1,\ldots,R} |\mathcal{I}_r|$
5: **for** $r \neq r^*$ **do**
6: Define $\mathbf{C}_{|\mathcal{I}_{r^*}|\times|\mathcal{I}_r|}$, where $C_{ij} = \mathrm{corr}(v_{r^*i}, v_{rj})$ for $i \in \mathcal{I}_{r^*}, j \in \mathcal{I}_r$
7: Initialize position mapping $\mathcal{F}_r = \emptyset$
8: **while** unmatched indices remain in both $\mathcal{I}_{r^*}$ and $\mathcal{I}_r$ & $\max(\mathbf{C}) > \rho$ **do**
9: $(k, j) \triangleq \arg\max_{i',j'} C_{i'j'}$
10: $\mathcal{F}_r \leftarrow \{\mathcal{F}_r, j \to i\}$; $\mathbf{C} \leftarrow \mathbf{C}_{-i,-j}$
11: **end while**
12: Randomly match remaining positions in (A_{r^*}, A_r)
13: Re-align by LDs: $A_r \leftarrow \mathcal{F}_r(A_r)$
14: **end for**

The GAS procedure is designed as a greedy and computationally efficient (see Sec. 3.4) alignment strategy for repeated VAE runs. Its purpose is not to solve a generic assignment problem over all LDs, but to align informative LDs of the same LD-feature association structure to set the foundations to aggregate disentanglement results. Standard post-hoc matching methods, such as Hungarian matching (Kuhn, 1955; Munkres, 1957), could be considered as alternative solvers for the alignment subproblem within GAS, but they can not replace GAS and several critical steps from Alg. 3 remain essential for properly solving the

[2]Specifically, LDs obtained across different runs are not necessarily aligned even when K is fixed. For example, the first LD in one run may correspond to the third LD in another run.

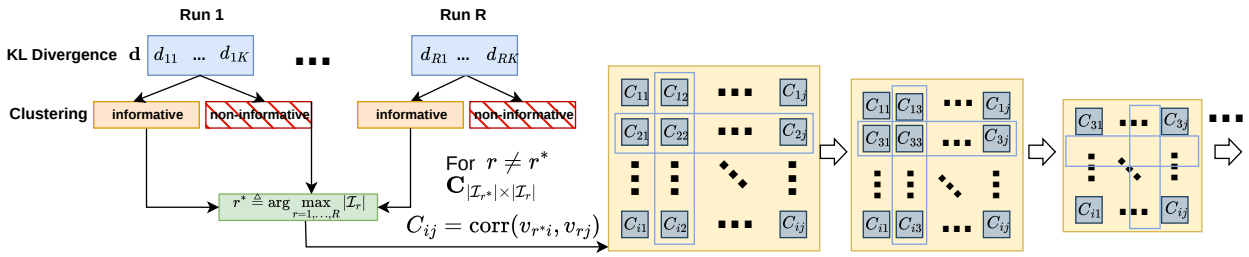


Figure 2: The GAS procedure. d_{rk} denotes the prior-posterior KL divergence of k-th LD; C_{ij} represents the correlation between the i-th LD in reference run r^* and the j-th LD in run r in the assocation matrix output from FVH-LT (variance matrix S) or DBSR-LS (regression coefficient matrix $|D|$).

alignment problem in this setting [3]. More broadly, comparing GAS with alternative alignment strategies, including Hungarian matching, optimal transport, and unbalanced matching methods (Villani, 2009; Peyré & Cuturi, 2019; Chizat et al., 2018; Liero et al., 2018), is an interesting direction for future work.[4]

3.4 Computational Cost and Scalability of FVH-LT, DBSR-LS, and GAS

Let n denote the number of samples, p the number of input features, K the number of latent dimensions (LDs), L the number of LT steps, R the number of repeated training runs, and T the number of solver iterations used by DBSR-LS. FVH-LT evaluates each trained model by traversing each LD over L grid points for each sample and then computing feature-wise variances of the decoded outputs. Ignoring the cost of VAE training itself, the FVH-LT evaluation cost is $\mathcal{O}(RnKL(c_{\text{dec}} + p)\}$, where c_{dec} denotes the cost of one decoder forward pass. If K and the decoder size are treated as fixed, this reduces to $\mathcal{O}(RLn)$.

For DBSR-LS, the main cost comes from solving the sparse multi-task regression problem (the original DBSR work (Jalali et al., 2010) does not provide a computational complexity analysis). For matrix-free proximal-gradient implementation of DBSR-LS, it requires $\mathcal{O}(Knp)$ operations per iteration to compute the least-squares gradients. The proximal update costs $\mathcal{O}(Kp)$ for the element-wise $\ell_{1,1}$ penalty and $\mathcal{O}(pK \log K)$ or $\mathcal{O}(pK)$ for the $\ell_{1,\infty}$ penalty, depending on whether a sorting-based or linear-time projection is used. The total cost is thus $\mathcal{O}\left(T(Knp + pK \log K)\right)$. For a Gram-matrix implementation, it first forms $X_k^\top X_k$ and $X_k^\top y_k$ for each task with cost $\mathcal{O}(Knp^2)$. Each iteration costs $\mathcal{O}(Kp^2 + Knp + pK \log K)$, where $\mathcal{O}(Kp^2)$ comes from multiplying by the Gram matrices and $\mathcal{O}(Knp)$ term may arise from objective evaluation or backtracking line search. The total complexity is $\mathcal{O}\left(Knp^2 + T(Kp^2 + Knp + +pK \log K)\right)$, which is consistent with follow-up analyses of dirty-model implementations (Rose, 2014). This explains why DBSR-LS is more expensive than FVH-LT when p, K or T is large.

In contrast, the additional cost of GAS is comparatively small. GAS is applied after FVH-LT or DBSR-LS has already produced a $K \times p$ LD-feature association matrix per run. Let $k_r = |\mathcal{I}_r| \leq K$ denote the number of informative LDs in run r, then computing the correlation matrix between the reference run r^* and run r costs $\mathcal{O}(k_{r^*} k_r p)$; and the greedy matching step costs $\mathcal{O}(\max {k_r}^2 \log(\max k_r))$ if the pairwise correlations are sorted once. So the total GAS cost across R runs is approximately $\mathcal{O}(\sum_{r \neq r^*} p k_{r^*} k_r) + \mathcal{O}(R(\max {k_r}^2) \log(\max k_r))$, Since $K_r \leq K$ is typically very small compared with n and p, this GAS cost is usually much smaller than the cost of VAE training, FVH-LT, and DBSR-LS.

For practical applications, there are multiple way to reduce computational time if needed. Since repeated runs are independent, they can be parallelized across CPUs/GPUs; within FVH-LT, LT can be also batched and parallelized over samples, LDs, and LD grid points; GAS is easily vectorized because it consists primarily of matrix correlations and greedy matching over a small matrix. For high-dimensional data of large p, we recommend FVH-LT as the default procedure and using DBSR-LS selectively, for example on preselected features or domain-defined feature groups.

[3]Hungarian matching provides a globally optimal one-to-one assignment once the cost matrix and the two sets of objects are fixed. A naïve application to all LDs would force matches even among noninformative LDs. A more meaningful application would still require the followings steps used in GAS: LD clustering, reference-run selection, handling unequal numbers of informative LDs across runs, and thresholding of weak matches.

[4]Such a comparison would require a dedicated benchmark of post-hoc alignment algorithms for stochastic VAE latent spaces, which is beyond the scope of the present work that focuses on disentangled VAEs and ground-truth-free interpretability.

3.5 Latent Space Separation Index (LSSI)

We introduce LSSI for two reasons. First, as discussed in Sec. 1, existing disentanglement metrics require access to ground-truth generative factors, which is often unavailable in real-world applications. In contrast, LSSI is computed without ground-truth factor information, which also implies that LSSI should be interpreted as a non-causal proxy for disentanglement-oriented separation metrics [5]. Second, while FVH-LT and DBSR-LS produce a quantitative LD-feature association matrix $\in \mathbb{R}^{K\times p}$ that is most effectively visualized as heatmaps. LSSI complements this visual interpretation by proving a convenient single-number summary of the latent structural separation, which is especially useful when p is large.

We motivate the mathematical formulation of LSSI from an axiomatic perspective. A well-separated latent representation should have the following properties: 1) different informative LDs should have minimally overlapping feature-association profiles; 2) LDs with similar feature-association profiles should be discounted; 3) The level of informativeness of LDs should be considered; 4) non-informative and weakly informative LDs should not be treated as meaningful factors and be discounted; and 5) the score should be invariant to the overall scale of the association matrix A.

Definition 2 (LSSI) *Let $A \in \mathbb{R}^{K\times p}$ denote the output matrix from FVH-LT or DBSR-LS, $A_{[k,]} \in \mathbb{R}^p$ denote the k-th row of A, and ℓ_1-norm $\|A_{[k,]}\|_1 = \sum_{i=1}^{p} |A_{k,i}|$. Let τ be the p-th largest value among all entries in $|A|$ and $K_{\max}$ be the maximum number informative LD from R runs.*

- *When true generative factors are unknown (the most likely practical setting): Define $\tilde{A}_{ki}^{(\tau)} := |A_{ki}| \cdot \mathbf{1}\{|A_{ki}| \geq \tau\}$, $\rho_k := \|\widetilde{A}_{[k,]}^{(\tau)}\|_1$, $\pi_k := \rho_k \big(\sum_{\ell=1}^{K} \rho_\ell\big)^{-1}$, and $K_{\text{eff}} := \big(\sum_{k=1}^{K} \pi_k^2\big)^{-1}$ (the effective number of informative LD), let $\mathcal{P} = \{(k,j) : 1 \leq k < j \leq K\}$ denote the set of all unordered pairs of LDs, then*

$$\text{LSSI}(A) = c_1 c_2 \left(\frac{\sum_{(k,j)\in\mathcal{P}} \big\| |A_{[k,]}| - |A_{[j,]}| \big\|_1}{\sum_{(k,j)\in\mathcal{P}} \big(\|A_{[k,]}\|_1 + \|A_{[j,]}\|_1 \big)} \right) \in [0,1], \tag{5}$$

$$\text{where } c_1 = \frac{|\{i : \exists k \text{ such that } |A_{ki}| \geq \tau\}|}{p} \text{ and } c_2 := \min\left\{ \frac{K_{\text{eff}} - 1}{K_{\max} - 1}, 1 \right\}, \tag{6}$$

- *When ground-truth generative factors are unknown (the benchmark setting for comparison with existing disentanglement metrics that rely on ground-truth factors): let $\mathcal{P}_0 = \{(k,j) : 1 \leq k < j \leq K_0\}$ denote the set of all unordered pairs of LDs that correspond to the K_0 ground-truth factors, then*

$$\text{LSSI}_0(A) = c_1 c_2 \left(\frac{\sum_{(k,j)\in\mathcal{P}_0} \big\| |A_{[k,]}| - |A_{[j,]}| \big\|_1}{\sum_{(k,j)\in\mathcal{P}_0} \big(\|A_{[k,]}\|_1 + \|A_{[j,]}\|_1 \big)} \right) \in [0,1],$$

$$\text{where } c_1 \text{ is the same as in Eq. 6 and } c_2 = \frac{|\{k : \exists i \text{ such that } |A_{ki}| \geq \tau\} - 1|}{K_0 - 1}. \tag{7}$$

K_{eff} in Def. 2 is known as the participation-ratio or inverse-Simpson effective number (Bell & Dean, 1970; Wegner, 1980; Hill, 1973; Jost, 2006). It equals $K_{\max}$ when the signal is evenly distributed over $K_{\max}$ LDs and approaches 1 when the signal is concentrated in a single LD. Thus, it provides a soft measure of how many LDs meaningfully contribute, while penalizing highly concentrated row-mass distributions.

The pairwise component of LSSI is $d(k,j) = (\| |A_{[k,]}| - |A_{[j,]}| \|_1)(\|A_{[k,]}\|_1 + \|A_{[j,]}\|_1)^{-1}$, which measures the normalized non-overlap between two the feature association profiles in rows k and j. Let $w_{kj} := \|A_{[k,]}\|_1 + \|A_{[j,]}\|_1$, then third term in LSSI is $\sum_{k<j} w_{jk} d(k,j) (\sum_{k<j} w_{kj})^{-1}$, which is a mass-weighted average and measures how distinct the LD-feature association profiles are across LDs. The factors c_1 and c_2 serve as correction terms for this raw separation score. In particular, c_1 measures feature coverage, ensuring that a high LSSI value is assigned only when the learned LDs explain a substantial portion of the input features.

[5] In the strongest sense, disentanglement is often understood as recovering independent or causally meaningful factors of variation (Bengio et al., 2013). Establishing such a causal interpretation generally requires known ground-truth factors, interventions, or additional domain assumptions. Existing disentangled VAE objectives, β-VAE and factor-VAE included, encourage separation of LDs through KL-based or total-correlation regularizers which target statistical properties (e.g., independence) of the learned representation, rather than causal recovery of the true generative factors. The corresponding β-VAE and factor-VAE disentanglement metrics are post-hoc evaluations applied after model training.

c_2 measures LD utilization and penalizes solutions in which most signal is concentrated in only one or a few LDs. Together, c_1 and c_2 ensure that LSSI reflects not only separation among LD-feature profiles, but also sufficient feature coverage and non-degenerate use of the latent space.

In the case of $\|A_{[k,]}\|_1 = 0$ and $\|A_{[j,]}\|_1 = 0$ for a pair of (k, j) or when there exists a row i such that $A_{k,i} \geq A_{j,i}$ for all i and $j \neq k$, then $\text{LSSI}(A) = 0$ and $\text{LSSI}_0(A) = 0$ (illustrated in Fig. 3). Second,

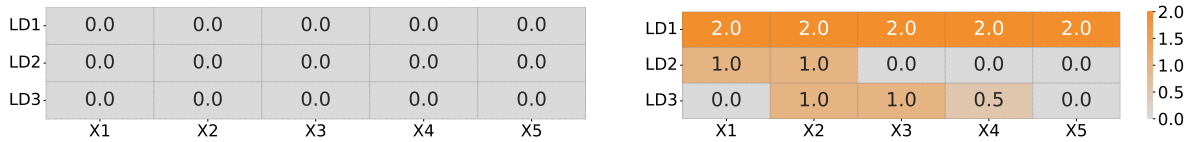


Figure 3: Illustration of two degenerate cases resulting in LSSI = 0. Left: All entries of matrix A are 0, indicating the complete absence of informative LDs. Right: A single LD dominates all other LDs across all input features in A, resulting in no disentanglement.

LSSI is conveniently bounded in $[0, 1]$, which facilitates interpretation, model comparison, and hyperparameter tuning. A value of 1 corresponds to a perfectly separated latent space, where A has a block-separable structure: each informative LD is associated with a distinct, non-overlapping subset of input features while non-informative LDs have negligible feature associations. A value of 0 corresponds to a degenerate latent space with no meaningful separation among LD-feature association profiles, such as when a single dominant LD is associated with most features or multiple LDs encode highly overlapping information. Intermediate values quantify the continuum between these endpoints: larger values indicate stronger structural separation, whereas smaller values indicate more overlapping or entangled feature encoding or one or a very few LDs are highly dominant among all informative LDs.

4 Experiments

We benchmark the proposed bfVAE in its disentanglement efficiency against the vanilla VAE and four existing disentangled VAE models, and assess the robustness of FVH-LT, DBSR-LS, and LSSI in quantifying the disentanglement effect. We use image and tabular data (Tab. 1) with a particular emphasis on tabular data. Image data is included as it is the common benchmark data type for VAE disentanglement assessment. We benchmarks LSSI against two representative existing disentanglement metrics in Higgins et al. (2017) and Kim & Mnih (2018).[6] The two synthetic datasets vary in p and n. For the five real datasets, RNA-seq (Fiorini, 2016) contains high-dimensional cancer gene-expression profiles across multiple tumor types; white wine (Cortez et al., 2009) includes features describing white wine characteristics with a label on the quality score; FIFA 2018 (Mathan, 2018) contains various soccer match statistics; and MNIST (Deng, 2012) and CelebA (Liu et al., 2015) are widely used benchmark image data.

Table 1: Datasets used in the experiments

Dataset	**Training n; test n**	**No. of Features p**	**Known Ground Truth?**
FA15 (tabular)†	800/200	15	Yes ($K_0 = 4$ factors)
FA100 (tabular)†	40,000/10,000	100	Yes ($K_0 = 6$ factors)
White wine (tabular)#	3,918/980	12	No
FIFA 2018 (tabular)#	102/26 (51/13 matches)	16	No
RNA-seq (tabular)#	640/161	20,264	No
MNIST (image)#	50,000/10,000	784 (28×28)	No
CelebA (image)#	35,000/5,000	$12,288$ ($64 \times 64 \times 3$)	No

† We simulated the FA15 and FA100 from factor analysis (FA) models.

Publicly available at Kaggle, PyTorch, or UC Irvine Machine Learning Repository.

In each experiment, the dataset is split into a training set and a test set: the training set is used to train VAE and perform FVH-LT and DBSR-LS and the test set evaluates model fitting and convergence of VAE

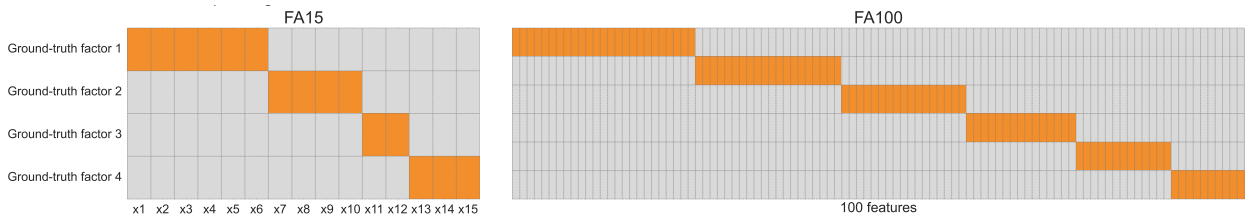
[6]Other disentanglement metrics based on complementary criteria exist (Sec 1.3), and they also require known ground-truth factors. Because these alternatives introduce additional implementation choices and are not universally regarded as superior across settings, we consider the metrics in Higgins et al. (2017) and Kim & Mnih (2018) sufficient as representative benchmarks.

training. R was set at 10 in all experiments except for CelebA, where $R = 5$ due to computational constraints. More details on model configurations and algorithmic hyperparameters are provided in App. B; We provide a subset of the experimental results in the main text and list the comprehensive results in App. G.

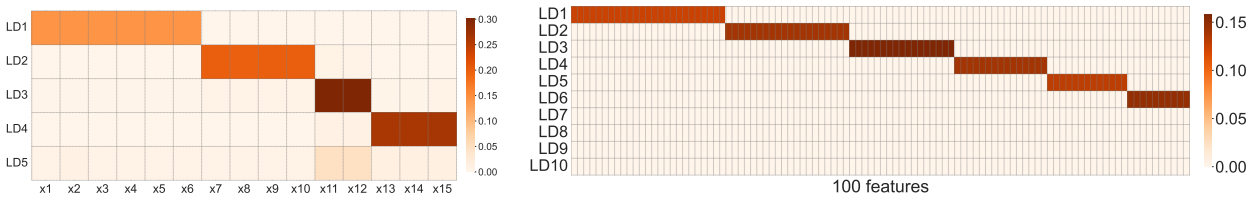
The main findings are as follows. 1) Across synthetic tabular, real tabular, and image datasets, bfVAE provides the most favorable overall trade-off between disentanglement and reconstruction among the benchmark VAE frameworks in the studied experimental settings. 2) In the synthetic experiments with known ground-truth factors, the trends of LSSI across different models agrees well with existing supervised disentanglement benchmarks, supporting its effectiveness as a metric for latent structural separation without ground-truth factor knowledge. 3) Both FVH-LT and DBSR-LS can uncover semantically meaningful or domain-aligned latent structures and generally yield consistent conclusions. FVH-LT is computationally more efficient at large p [7] and often gives clearer informative-LD signals (LDBS-LS can be faster at small p). 4) Overall, bfVAE with FVH-LT and LSSI offers a practical and interpretable framework for learning separated latent representations without sacrificing reconstruction quality.

4.1 Tabular data with knowledge on true generative factors

Fig. 4 the results in the disentanglement effect via bfVAE assessed by FVH-LT in FA15 and FA100. bfVAE effectively disentangles the latent structure, with the learned associations between informative LDs and input features closely aligned with the ground truth. It is worth noting the robustness of FVH-LT to over-specified K; that is, it does not mis-identify non-informative LDs as informative, resulting in a low false positive rate.



(a) Ground-truth factor-feature association in FA15 and FA100



(b) Discovered LD-feature association via bfVAE

Figure 4: Ground-truth factor-feature association and discovered LD-feature association via bfVAE assessed by FVH-LT in FA15 and FA100. Darker cells indicate stronger associations between LDs and input features. Values in the LD-feature FVH-LT matrix are not shown due to readability reasons. DBSR-LS results for bfVAE and FVH-LT and DBSR-LS in the VAE benchmarks are in Apps. D.1- D.4.

The LSSI, reconstruction ℓ_2 error on the test sets, and the estimated number of generative factors are reported in Tab. 2 for bfVAE and the benchmark VAEs. A further break-down of the LSSI by c_1, c_2 and the raw LSSI score in Eq. 5 is provided in Appendix. E. Overall, bfVAE achieves the best trade-off: it ranks first or second in LSSI across datasets and methods, while delivering substantially lower reconstruction error than all competing models; and recovers the true number of generative factors. Among the benchmarks, Factor-VAE achieves competitive LSSI scores – often second-best – but at the cost of poor reconstruction fidelity and under-discover generative factors 50% of the time. Conversely, β-VAE reconstructs well but scores poorly on LSSI and over-identities generative factors 50% of the time. Vanilla VAE and the two DIP-VAE variants perform the worst on both disentanglement and reconstruction.

[7] In the CelebA dataset ($p = 12,288$) and RNA-seq ($p = 20,264$), the high computational costs for DBSR-LS actually exceed our resources. Due to space limitations and for the reasons listed above, we focus on the FVH-LT results in the main text; the DBSR-LS results can be found in Appendix.

Table 2: LSSI, reconstruction error, and uncovered informative LD counts in synthetic FA15 and FA100.

Method	Data	Metric	**bfVAE**	factor-VAE	β-VAE	Vanilla VAE	DIP-VAE-I	DIP-VAE-II
FVH-LT	FA15	↑ LSSI	**0.891**	***0.807***	0.366	0.544	0.633	0.498
		$(\rightarrow 0)\ K_{\max} - K_0$ †	**0**	-1	**0**	**0**	**0**	**0**
	FA100	↑ LSSI	0.913	**0.940**	0.252	0.866	0.874	***0.929***
		$(\rightarrow 0)\ K_{\max} - K_0$ †	**0**	**0**	3	**0**	**0**	**0**
DBSR-LS	FA15	↑ LSSI	***0.979***	**0.987**	0.906	0.597	0.962	0.841
		$(\rightarrow 0)\ K_{\max} - K_0$ †	**0**	-1	**0**	0	**0**	-1
	FA100	↑ LSSI	***0.636***	0.602	0.217	0.591	0.592	**0.660**
		$(\rightarrow 0)\ K_{\max} - K_0$ †	**0**	**0**	3	**0**	**0**	**0**
FA15		↓ recon. ℓ_2 ‡	$\boldsymbol{0.75 \pm 0.29}$	2.13 ± 0.17	$\mathbf{0.64 \pm 0.03}$	2.18 ± 0.18	1.70 ± 0.06	2.36 ± 0.26
FA100		↓ recon. ℓ_2 ‡	$\mathbf{0.39 \pm 0.01}$	1.28 ± 0.03	$\boldsymbol{0.56 \pm 0.02}$	1.35 ± 0.08	1.49 ± 0.06	1.38 ± 0.06

†: The closer to 0, the better; true factor counts $K_0 = 4$ in FA15 and $K_0 = 6$ in FA100

‡ Per-sample reconstruction ℓ_2 error on the test set: mean ± standard deviation across 10 runs.

Bold denotes the best result (highest LSSI; lowest ℓ_2 error) by dataset and method; ***italic bold*** denotes the second best.

Since FA15 and FA100 are synthetic data, we can benchmark our proposed LSSI and LSSI_0 per Eq. 7 against the existing disentanglement metrics in the literatureHiggins et al. (2017); Kim & Mnih (2018), both of which require known generative factors). The results are provided in Tab. 3. First, regardless of which disentanglement metric is used, bfVAE consistently achieves the highest scores, confirming it as the best-performing disentangled VAE framework. Second, we observe a consistent monotonic relationship between LSSI (Tab. 2), LSSI_0 and the two existing disentanglement metrics: higher LSSI and LSSI_0 correspond to higher scores on both benchmarks, validating LSSI as an effective disentanglement metric. Third, the similarity between LSSI and LSSI_0 suggests that LSSI_0 is robust to over-specification of K.1 and show smaller differences across methods.

Table 3: LSSI_0 and baseline disentanglement scores in synthetic FA15 and FA100

Data	Metric	**bfVAE**	factor-VAE	β-VAE	Vanilla VAE	DIP-VAE-I	DIP-VAE-II
FA15	LSSI_0 FVH-LT	**0.963**	0.811	0.405	0.703	0.661	0.780
	LSSI_0 DBSR-LS	**1.000**	0.999	0.972	0.851	0.995	0.578
	Higgins et al. (2017) ‡	$\mathbf{1.00 \pm 0.00}$	0.99 ± 0.01	0.90 ± 0.07	0.96 ± 0.06	0.84 ± 0.11	0.94 ± 0.07
	Kim & Mnih (2018) ‡	$\mathbf{0.99 \pm 0.03}$	0.90 ± 0.06	0.89 ± 0.09	0.84 ± 0.07	0.84 ± 0.13	0.81 ± 0.06
FA100	LSSI_0 FVH-LT	**0.990**	0.986	0.417	0.860	0.932	0.962
	LSSI_0 DBSR-LS	**0.766**	0.743	0.351	0.702	0.681	0.738
	Higgins et al. (2017) ‡	$\mathbf{1.00 \pm 0.00}$	1.00 ± 0.00	0.64 ± 0.08	0.99 ± 0.03	1.00 ± 0.00	1.00 ± 0.00
	Kim & Mnih (2018) ‡	$\mathbf{1.00 \pm 0.00}$	1.00 ± 0.00	0.77 ± 0.13	0.99 ± 0.03	1.00 ± 0.00	1.00 ± 0.00

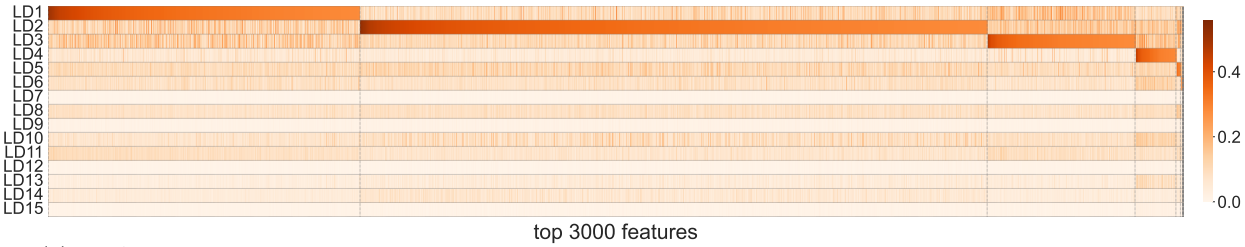
Bold indicates the VAE model with the best score within each dataset and metric.

‡ Scores are reported as mean ± SD over 10 runs. LSSI and LSSI_0 are based on the averaged FVH-LT and DBSR-LS over 10 runs, per definition

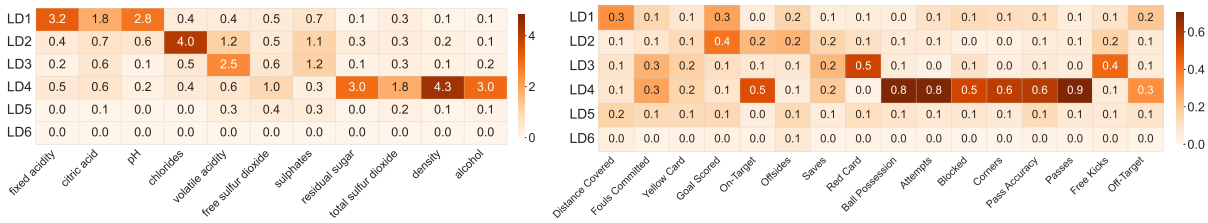
4.2 Tabular data without knowledge on generative factors

Fig. 5 displays the FVH-LT results from bfVAE on the RNA-seq, white wine, the 2018 FIFA datasets. For RNA-seq, the top $3,000$ unique genes are selected fromthe full LD-feature association matrix of $15 \times 20,264$ for better visualization. Tab. 4 reports the LSSI, reconstruction ℓ_2 error on the test sets, and the estimated number of generative factors via bfVAE and the benchmark VAEs. A further break-down of the LSSI by c_1, c_2 and the raw LSSI score in Eq. 5 is provided in Apps. E.3- E.5. The findings are similar to Tab. 2: bfVAE achieves the best trade-off between disentanglement and reconstruction. Specifically, bfVAE ranks mostly the best in LSSI across datasets and methods with the second best reconstruction error on white wine and FIFA 2018 whereas β-VAE achieves the best errors in both; for RNA-seq, the reconstruction errors are about the same across all VAEs models with $\sim$1% differences. The second best LSSI score varies are shared by factor-VAE and DIP-VAE-I where bfVAE is the best; β-VAE achieves the best LSSI for white wine with FVH-LT whereas bfVAE is the second best.

Compared with Tab. 2, the LSSI scores in Tab. 4 are lower than those observed in synthetic settings, which is expected because real-world datasets are typically more complex, noisy, and heterogeneous, and their underlying generative factors may not exhibit a clean block-separable structure. For RNA-seq, the data is very



(a) RNA-seq data: 3,000 unique genes with the largest-valued FVH-LT association matrix entries are shown



(b) White Wine (c) 2018 FIFA

Figure 5: Disentanglement effects of bfVAE assessed by FVH-LT in datasets without knowledge of ground-truth generative factors. Darker cells indicate stronger associations between LD and input features.DBSR-LS results for bfVAE and FVH-LT and DBSR-LS results for the benchmark VAE models are in Apps. D.5- D.9.

Table 4: LSSI, reconstruction error, and uncovered informative LD counts in real tabular data.

Method	Data	Metric	**bfVAE**	factor-VAE	β-VAE	Vanilla VAE	DIP-VAE-I	DIP-VAE-II
FVH-LT	RNA-seq	$\uparrow$ LSSI	**0.081**	0.057	0.056	0.052	***0.072***	0.058
		$K_{\max}$	11	14	12	12	13	14
	White wine	$\uparrow$ LSSI	**0.390**	***0.308***	0.175	0.220	0.090	0.280
		$K_{\max}$	5	4	5	5	4	4
	FIFA 2018	$\uparrow$ LSSI	***0.132***	0.020	**0.197**	0.031	0.104	0
		$K_{\max}$	5	2	5	2	5	3
DBSR-LS	White wine	$\uparrow$ LSSI	**0.736**	0.478	0.466	0.356	***0.719***	0.396
		$K_{\max}$	5	4	5	5	4	4
	FIFA 2018	$\uparrow$ LSSI	**0.455**	0.127	0.404	0.081	***0.450***	0.040
		$K_{\max}$	5	2	5	2	5	3
RNA-seq		$\downarrow$ recon. ℓ_2[‡]	105.51 ± 0.26	104.60 ± 0.15	104.91 ± 0.22	105.15 ± 0.21	104.11 ± 0.17	104.22 ± 0.18
white wine			***2.03 ± 0.04***	2.37 ± 0.06	$\mathbf{1.52 \pm 0.03}$	2.35 ± 0.06	2.08 ± 0.02	2.33 ± 0.04
FIFA 2018			***2.84 ± 0.14***	3.49 ± 0.03	$\mathbf{2.75 \pm 0.08}$	3.53 ± 0.04	3.28 ± 0.09	3.53 ± 0.04

Bold denotes the highest LSSI score by dataset and method; ***bold italic*** denotes the second best.

‡ Per-sample reconstruction ℓ_2 error on the test set: mean $\pm$ standard deviation across 10 runs.

high-dimensional, genes are co-regulated in modules and pathway programs may overlap, and a single gene may participate in multiple biological processes. Therefore, the latent structure may involve overlapping gene sets rather than clean, disjoint feature blocks, providing a biological explanation for the low LSSI values. For FIFA 2018, FactorVAE, vanilla VAE, and DIP-VAE-I exhibit a dominant-LD pattern in FVH-LT and DBSR-LS, where most features associate with one LD while others remain largely non-informative (Apps. D.8–D.9). DIP-VAE-II shows an even more degenerate dominant-row pattern, yielding a zero-valued LSSI. Overall, bf-VAE generally achieves clearer latent-feature association patterns and stronger interpretability on real-world tabular datasets and especially useful for relative model comparison, benchmarking, and hyperparameter tuning is latent structural separation is the main goal. As LSSI is applied to a broader range of real-world datasets, empirical reference ranges can be developed to guide the interpretation of what constitutes strong or moderate latent-space separation in practice.

The outputs from FVH-LT and DBSR-LS can be combined with domain knowledge to yield the highest utility. For RNA-seq, prior pan-cancer studies show that gene-expression variation is strongly organized by tissue/cell-of-origin and by pathway-level programs (Hoadley et al., 2018; Peng et al., 2015; Sanchez-Vega et al., 2018; Thorsson et al., 2018). Thus the uncovered latent structure can be interpreted in light of the known biology of pan-cancer transcriptomes to make the best of the result. For white wine, LD1 is associated with fixed acidity, citric acid, and pH, reflecting the wine acidity profile; LD2 is associated with chlorides; LD3 with volatile acidity; LD4 is strongly associated with density, alcohol, residual sugar, and

total sulfur dioxide, which are characteristics related to wine flavor. LD5 and LD6 is non-informative. In the 2018 FIFA data, LD1 is strongly associated with Distance Covered, Goal Scored, and Off-Target – reflecting match intensity and attacking activity. LD2 encodes Goal Scored, with weaker associations to On-Target, Offsides, Saves, and Free Kicks, representing scoring-related attacking outcomes.LD3 is associated with Fouls Committed, Red Card, Free Kicks, and Ball Possession, reflecting game control. LD4 shows structure with strong associations to On-Target, Ball Possession, Attempts, Blocked, Corners, Pass Accuracy, and Passes, corresponding to attacking buildup and shot creation.LD5 and LD6 are noninformative.

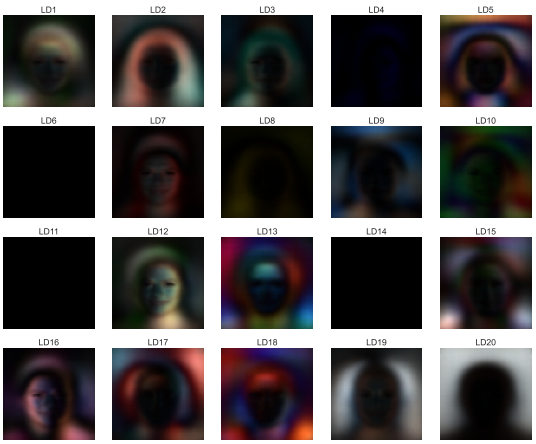
We also applied FVH-LT in the bf-CVAE model on the white wine dataset given that it contains label Y (wine quality). The implementation of bf-CVAE and full FVH-LT results are provided in App. H. Tab. 5 shows the last row of the LD-feature association matrix with LT on the conditioning variable Y. VC, C, D, and A exhibit the highest variance, indicating strong relations with wine quality and consistency with domain knowledge (see App. C). These results suggest a potential use case for conditional generation: a wine manufacture could fix Y at a high-quality score, sample informative LDs that exhibits strong associations between these four features, and feed the samples to the decoder to generate candidate feature profiles for wines targeting that quality level.

Table 5: Variance of reconstructed features with LT on wine quality score via FVH-LT.

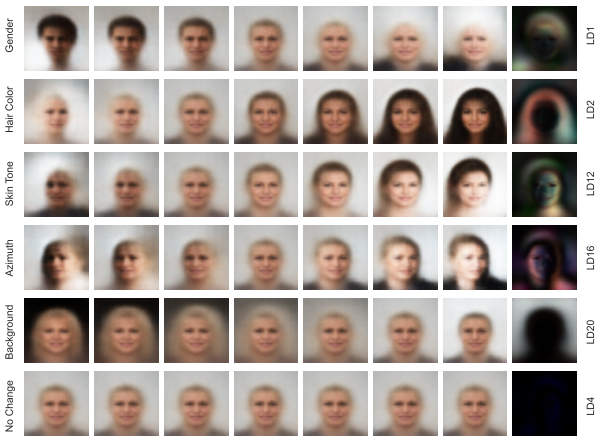
fixed acidity	**volatile acidity** (VC)	citric acid	residual sugar	**chlorides** (C)	free SO_2
0.08	**0.17**	0.08	0.06	**0.14**	0.05
total SO_2	**density** (D)	pH	sulphates	**alcohol** (A)	**quality**
0.09	**0.25**	0.07	0.05	**0.54**	**2.9**

4.3 Image Data CelebA and MNIST

The FVH-LT results in CelebA and MNIST are visualized in Figs. 6 and 7, respectively. CelebA contains three color channels (red, green, and blue). In the FVH-LT analysis, heatmaps were first computed separately by channel, then globally normalized and merged to produce a composite visualization in Fig. 6(a). The LT of a representative image in Figs. 6(b) and 7(b) illustrates how informative LDs correspond to meaningful pixels in the images. The results clearly demonstrate the specific factors each dimension represents, highlighting the disentangled structure captured in the latent space.



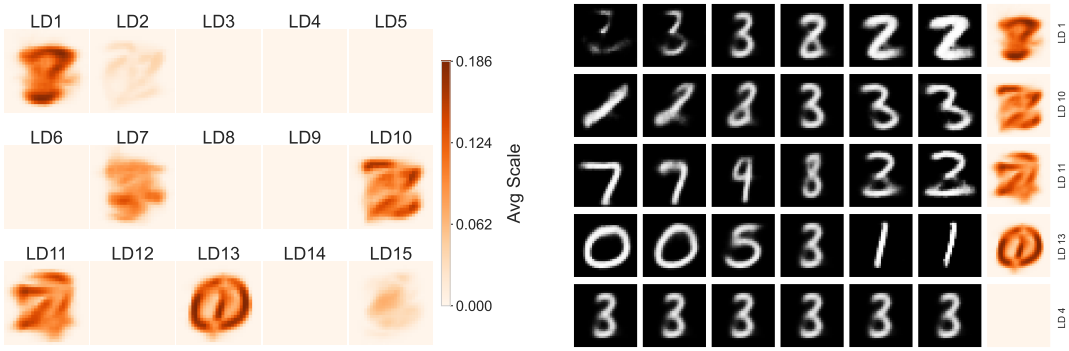
(a) FVH-LT results averaged over 1000 images and 5 runs. Brighter regions indicate higher variance, highlighting image areas that are sensitive to LT.



(b) Example images in Columns 1 to 7 reconstructed from LT on 6 LDs in a single run; the corresponding FVH-LT heatmaps are in Column 8.

Figure 6: Disentanglement effects of bfVAE on CelebA assessed by FVH-LT. FVH-LT shows which facial features correspond to what informative LD: LT of non-informative LD4 produces no visible change in the reconstruction and a completely dark heatmap in (a) and (b); in (b), LD1, 2, 12, 16, and 20 primary encode sex, hair style and color, skin tone and illumination, and azimuth (horizontal head orientation), background information, respectively. FVH-LT results for the benchmark VAE models are in Apps. D.11.

Tab. 6 reports the LSSI, reconstruction l_2 error and $K_{\max}$. bfVAE achieves the highest LSSI score on both datasets and the lowest reconstruction error on MNIST. Although its reconstruction error is not the smallest in CelebA, it remains comparable to the benchmark VAEs with reconstruction errors and is substantially



(a) FVH-LT results (averaged over 1000 images and 10 runs) suggest 7 informative LDs. meaningful pixels are concentrated at the image center, reflecting the fact that the digits are centered in the original images.

(b) Example LT results for an image with digit 3: Columns 1 to 6 are reconstructed from LT on 5 LDs in a single run; Column 7 shows the corresponding FVH-LT heatmaps.

Figure 7: Disentanglement effects of bfVAE in MNIST assessed by FVH-LT. (a) and (b) show that LD1, LD10, LD11, and LD13 capture variation between digits 3 and 2, 1 and 3, 7 and 3, and 0 and 1 respectively. Key structural elements that differentiate the digits are evident such as the loops and center vertical line in LD13 and the curved regions near top-left and bottom-right in LD10. The 7 informative LDs produce obvious changes in reconstructed images under LT (first four rows in (b)) with interpretable semantic patterns. For example, LT of LD1 generates digits 2, 3, 8 and the morphing process is reflected in the last column of (b) that suggests a blend of 2,3,8. LD10 also captures directional variations in handwriting. Along the LT, the generated digit changes from a slanted 1 into a 3 which gradually changes its slant and curvature. FVH-LT results for the benchmark VAE models are in Apps. D.10.

lower than that of β-VAE. Across both datasets, bfVAE shows consistent performance and provides a favorable trade-off between latent-space separation and reconstruction accuracy. In contrast, the performance of the other methods appears more dataset-dependent.

Table 6: LSSI, reconstruction error, and uncovered informative LD counts in CelebA and MNIST.

Data	Metric	**bfVAE**	factor-VAE	β-VAE	Vanilla VAE	DIP-VAE-I	DIP-VAE-II
CelebA	$\uparrow$ LSSI	**0.116**	***0.052***	0	0.029	0.031	0.029
	$\downarrow$ recon. ℓ_2[‡]	11.07 ± 0.23	10.36 ± 0.03	19.58 ± 0.10	10.16 ± 0.03	10.21 ± 0.03	10.21 ± 0.05
	$K_{\max}$	14	8	1	10	10	10
MNIST	$\uparrow$ LSSI	**0.184**	0.089	0.053	0.078	0.088	***0.095***
	$\downarrow$ recon. ℓ_2[‡]	3.10 ± 0.11	3.46 ± 0.02	5.11 ± 0.00	3.28 ± 0.01	3.30 ± 0.02	3.33 ± 0.02
	$K_{\max}$	7	9	14	10	8	9

[‡] Per-sample reconstruction ℓ_2 error on the test set: mean $\pm$ standard deviation across runs.
Bold denotes the highest LSSI score within each dataset; ***bold italic*** denotes the second best.

4.4 Runtime Result for Model Training and Disentanglement Evaluation

Tab. 7 summarizes the approximate computational time per run for VAE training, FVH-LT, and DBSR-LS. Since the experiments were conducted on several platforms, the time should not be compared across datasets that used different platforms.

Table 7: Approximate single-run time for VAE model training, FVH-LT, and DBSR-LS.

	FA15	FIFA 2018	White wine	CVAE White wine	FA100	RNA-seq	MNIST	CelebA
n[‡]	800	102	3,918	3,918	40,000	640	50,000[†]	35,000
Epochs	150	350	150	75	150	100	35	50
model training time (min)	3 (M2)	1 (M2)	10 (M2)	20 (M2)	4 (G4)	1 (G4)	2 (A100)	4 (A100)
FVH-LT[†] time (min)	2 (M2)	<1 (M2)	15 (M2)	16 (M2)	17 (G4)	10 (G4)	14 (A100)	60 (A100)
DBSR-LS time (min)	1 (M2)	<1 (M2)	1 (M2)	–	2 (G4)	–	40 (A100)	–

[‡] n for model training, FVH-LT, and DBSR-LS are the same in all datasets except for MNIST and CelebA, which is n listed is for model training and $n = 1000$ for FVH-LT and DBSR-LS.

[†] The number of LT steps was $L = 500$ RNA-seq and $L = 1,000$ for the rest.

M2: a local machine with an Apple M2 Max CPU and 96 GB RAM; G4: a Google Colab G4 GPU; A100: an A100 GPU node with two NVIDIA A100 80GB PCIe GPUs and 64 Intel Xeon Platinum 8358 CPU cores at 2.60GHz.

Model training time does not include cross-validation or hyperparameter tuning. FVH-LT and DBSR-LS times are reported for one evaluation run. GAS in all datasets took < 1 minute per run.

4.5 Visualizing Posterior Distributions of Informative and Non-informative LDs

Algs. 1 and 2 calculate the posterior-prior KL divergence for each LD as an intermediate product, on which informative vs non-informative LDs in Def. 1 are based. To illustrate how the posterior deviate from the prior $\mathcal{N}(\mathbf{0}, I)$, we plot the posterior distributions from a single run of bfVAE with $\gamma = 0$ and $K = 15$ on a random subset of 1,000 MNIST training images in Fig. 8. The results clearly suggest there are seven informative LDs, the posteriors of which exhibit low variances (< 0.1) and many also have posterior means shifted from 0, leading to a prior-posterior KL divergence of ~ 8 on average. The rest eight LDs are non-informative with an KL divergence of $\sim 3 \times 10^{-3}$ on average. Similar patterns are consistently observed in other runs.

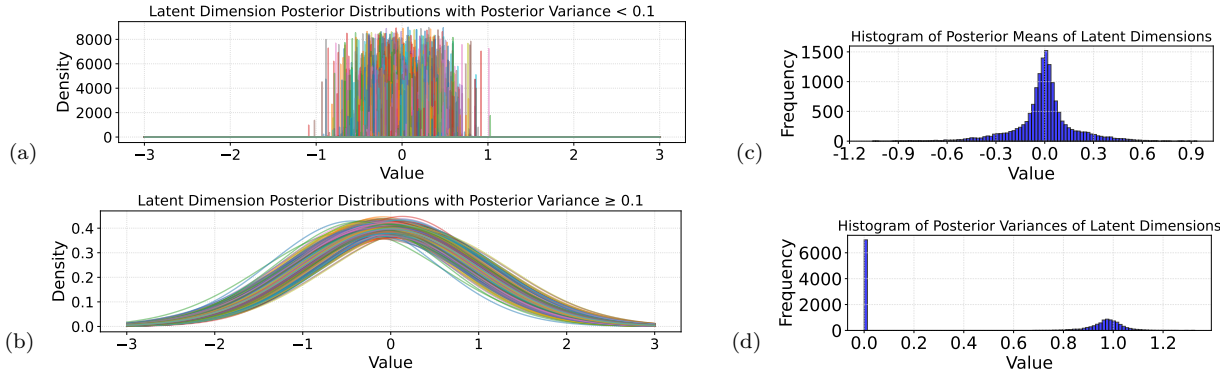


Figure 8: An example of posterior statistics across LDs on MNIST (15 LDs $\times$ 1000 MNIST images): (a) posterior distributions of informative LDs, (b) posterior distributions of non-informative LDs, (c) histogram of posterior means, (d) histogram of posterior variances (among the 15,000 LDs, exactly 7,000 display near-zero variances in their distribution and correspondingly exhibit large KL divergence).

5 Conclusion and Discussion

We proposed bfVAE as a unified framework for disentangling VAE latent space and offered an information-bottleneck perspective on its formulation. We introduced two ground-truth-free evaluation techniques, FVH-LT and DBSR-LS, to support interpreting and quantifying LD-feature associations, along with GAS to address label switching in latent space from multiple VAE runs. We further developed LSSI, a scalar metric derived from the LD-feature association matrices produced by FVH-LT and DBSR-LS to summarize overall latent structural separation. Importantly, FVH-LT, DBSR-LS, GAS or LSSI do not require knowledge of ground-truth generative factor. Across extensive experiments, bfVAE provides an effective framework for promoting latent-space disentanglement and generally achieves more favorable reconstruction-disentanglement trade-off relative to benchmark VAEs, as assessed by FVH-LT, DBSR-LS, LSSI, and reconstruction error.

While both FVH-LT and DBSR-LS effectively uncover can semantically meaningful, domain-aligned latent structures, we recommend FVH-LT as the default technique from both computational and methodological perspectives. DBSR-LS can be computational intensive given its sparse regression framework especially when p is large, and methodological, it relies on a linear surrogate to characterize the LD-feature relation for straightforward interpretability. Because DBSR-LS estimates LD-feature associations directly from learned latent space and is not decoder-mediated like FVH-LT, it can serve as a useful alternative when decoder-based LT is unstable, especially when p is not large.

Future directions include applying our procedures to other data types (e.g., graphs), adapting the procedures to other deep learning models that may benefit from interpretable latent spaces disentanglement, and extending DBSR-LS to non-linear regression settings. A focused assessment of alternative alignment strategies to the greedy strategy proposed in this paper, such as Hungarian matching and optimal transport, would make another interesting direction for future work.

References

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Seungheun Baek, Soyon Park, Yan Ting Chok, Junhyun Lee, Jueon Park, Mogan Gim, and Jaewoo Kang. Cradle-vae: Enhancing single-cell gene perturbation modeling with counterfactual reasoning-based artifact disentanglement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39 (15), pp. 15445–15452, 2025.

R. J. Bell and P. Dean. Atomic vibrations in vitreous silica. *Discussions of the Faraday Society*, 50:55–61, 1970. doi: 10.1039/DF9705000055.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599*, 2018.

Clément Chadebec, Louis Vincent, and Stephanie Allassonniere. Pythae: Unifying generative autoencoders in python-a benchmarking use case. *Advances in Neural Information Processing Systems*, 35:21575–21589, 2022.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

Paulo Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Wine Quality. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C56S3T.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. doi: 10.1109/MSP.2012.2211477.

Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *6th International Conference on Learning Representations*, 2018.

Samuele Fiorini. gene expression cancer RNA-Seq. UCI Machine Learning Repository, 2016. DOI: https://doi.org/10.24432/C5R88H.

William Harvey, Saeid Naderiparizi, and Frank Wood. Conditional image generation by conditioning variational auto-encoders. *arXiv preprint arXiv:2102.12037*, 2021.

Lucas Heublein, Simon Kocher, Tobias Feigl, Alexander Rügamer, Christopher Mutschler, and Felix Ott. VAE-based feature disentanglement for data augmentation and compression in generalized gnss interference classification. *arXiv preprint arXiv:2504.10556*, 2025.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

M. O. Hill. Diversity and evenness: A unifying notation and its consequences. *Ecology*, 54(2):427–432, 1973. doi: 10.2307/1934352.

Katherine A. Hoadley, Christina Yau, Toshinori Hinoue, Denise M. Wolf, Alexander J. Lazar, Esther Drill, Ronglai Shen, Alison M. Taylor, Andrew D. Cherniack, Vésteinn Thorsson, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304.e6, 2018. doi: 10.1016/j.cell.2018.03.022.

Ali Jalali, Sujay Sanghavi, Chao Ruan, and Pradeep Ravikumar. A dirty model for multi-task learning. *Advances in neural information processing systems*, 23, 2010.

Lou Jost. Entropy and diversity. *Oikos*, 113(2):363–375, 2006. doi: 10.1111/j.2006.0030-1299.14714.x.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International conference on machine learning*, pp. 2649–2658. PMLR, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955. doi: 10.1002/nav.3800020109.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017.

Chengrui Li, Yunmiao Wang, Yule Wang, Weihan Li, Dieter Jaeger, and Anqi Wu. A revisit of total correlation in disentangled variational auto-encoder with partial disentanglement. *arXiv preprint arXiv:2502.02279*, 2025.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones Mathematicae*, 211:969–1117, 2018.

Xiao Liu, Pedro Sanchez, Spyridon Thermos, Alison Q O'Neil, and Sotirios A Tsaftaris. Learning disentangled representations in the imaging domain. *Medical Image Analysis*, 80:102516, 2022.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

F Locatello, S Bauer, M Lucic, G Rätsch, S Gelly, B Schölkopf, and O Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. arxiv preprint arxiv: 1811.12359, 2018.

Mathan. Predict FIFA 2018 man of the match. `https://www.kaggle.com/datasets/mathan/fifa-2018-match-statistics`, 2018. Accessed: 2025-05-02.

James Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957. doi: 10.1137/0105003.

Brooks Paige, Jan-Willem Van De Meent, Alban Desmaison, Noah Goodman, Pushmeet Kohli, Frank Wood, Philip Torr, et al. Learning disentangled representations with semi-supervised deep generative models. *Advances in neural information processing systems*, 30, 2017.

Li Peng, Xiu Wu Bian, Di Kang Li, Chuan Xu, Guang Ming Wang, Qing You Xia, and Qing Xiong. Large-scale RNA-Seq transcriptome analysis of 4043 cancers and 548 normal tissue controls across 12 TCGA cancer types. *Scientific Reports*, 5:13413, 2015. doi: 10.1038/srep13413.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

Mozhgan Pourkeshavarz, Junrui Zhang, and Amir Rasouli. CaDeT: a causal disentanglement approach for robust trajectory prediction in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14874–14884, 2024.

Yijian Qin, Xin Wang, Ziwei Zhang, and Wenwu Zhu. Disentangled representation learning with large language models for text-attributed graphs. *arXiv preprint arXiv:2310.18152*, 2023.

Carlos E. Rodríguez and Stephen G. Walker. Label switching in bayesian mixture models: Deterministic relabeling strategies. *Journal of Computational and Graphical Statistics*, 23(1):25–45, 2014.

Rhiannon Rose. gLOP: A cleaner dirty model for multitask learning. Master's thesis, University of Waterloo, Waterloo, Ontario, Canada, 2014. URL `https://uwspace.uwaterloo.ca/bitstreams/10ceb6f1-ce3c-45e0-8a39-e414e797e3e8/download`.

Francisco Sanchez-Vega, Marco Mina, Joshua Armenia, Walid K. Chatila, Augustin Luna, Konnor C. La, Sofia Dimitriadoy, David L. Liu, Havish S. Kantheti, Sadegh Saghafinia, et al. Oncogenic signaling pathways in The Cancer Genome Atlas. *Cell*, 173(2):321–337.e10, 2018. doi: 10.1016/j.cell.2018.03.035.

Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 23(241): 1–55, 2022.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

Matthew Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B*, 62(4):795–809, 2000.

Vésteinn Thorsson, David L. Gibbs, Scott D. Brown, Denise Wolf, Dante S. Bortone, Tai-Hsien Ou Yang, Eduard Porta-Pardo, Galen F. Gao, Christopher L. Plaisier, James A. Eddy, et al. The immune landscape of cancer. *Immunity*, 48(4):812–830.e14, 2018. doi: 10.1016/j.immuni.2018.03.023.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Anshuk Uppal, Yuhta Takida, Chieh-Hsin Lai, and Yuki Mitsufuji. Denoising multi-beta vae: Representation learning for disentanglement and generation. *arXiv preprint arXiv:2507.06613*, 2025.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009. doi: 10.1007/978-3-540-71050-9.

Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

F. Wegner. Inverse participation ratio in $2 + \epsilon$ dimensions. *Zeitschrift für Physik B Condensed Matter*, 36: 209–214, 1980.

Ziye Zhang, Li Sun, Zhilin Zheng, and Qingli Li. Disentangling the spatial structure and style in conditional VAE. In *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 1626–1630. IEEE, 2020.