Beyond Mean-Field: Tree-Copula Variational Autoencoders for Structured Latent Dependencies

Yarden Rachamim[1] **Shai Fine**[2]

[1]Computer Science Dept., Reichman University, Herzliya, Israel
[2]Data Science Institute, Reichman University, Herzliya, Israel

Abstract

Variational Autoencoders (VAEs) typically assume a factorized Gaussian posterior and isotropic normal prior, implicitly imposing mean-field independence in the latent space. Rewriting the ELBO in copula form, we isolate a dependence-mismatch term that contributes to approximation and amortization gaps. We propose a Tree-Copula VAE that replaces the mean-field posterior with a tree-structured copula while retaining flexible learned marginals. The tree is inferred per datapoint via a Chow–Liu maximum-weight spanning tree, exploiting the monotonic relationship between mutual information and copula parameters. We further study an Average-of-Trees (AoT) copula prior for prior–posterior alignment and a rank-1 Gaussian-copula likelihood for low-rank observation correlations. Experiments on correlated dSprites and Fashion-MNIST show tighter importance-weighted bounds than mean-field VAEs when latent dependence is present, and highlight the role of prior alignment and likelihood modeling beyond mean-field inference.

1 INTRODUCTION

Deep generative models represent complex data distributions via latent variables that explain observed variability. *Variational Autoencoders* (VAEs) [Kingma and Welling, 2014, Rezende et al., 2014] are a central framework for such generative modeling. A VAE specifies a prior $p_\lambda(\mathbf{z})$ over latent variables, a likelihood (decoder) $p_\theta(\mathbf{x} \mid \mathbf{z})$ over observations, and an inference network (encoder) with variational posterior $q_\phi(\mathbf{z} \mid \mathbf{x})$. Training maximizes the *Evidence Lower Bound* (ELBO), a tractable lower bound on $\log p_\theta(\mathbf{x})$ [Jordan et al., 1999] that balances reconstruction and KL regularization using amortized variational inference and the reparameterization trick.

In practice, VAEs typically rely on three simplifying assumptions: (i) a standard normal prior, $p_\lambda(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$; (ii) a mean-field (MF) variational posterior, typically a diagonal Gaussian factorized across latent dimensions; and (iii) a factorized likelihood over the observation space, $p_\theta(\mathbf{x} \mid \mathbf{z}) = \prod_i p_\theta(x_i \mid \mathbf{z})$. These choices yield closed-form KL terms in the ELBO, simple sampling, and stable training in many settings [Fortuin, 2022], but they impose strong (conditional-)independence assumptions in the prior, variational, and likelihood.

These standard choices give rise to three related challenges: First, the mean-field restriction prevents $q_\phi(\mathbf{z} \mid \mathbf{x})$ from capturing posterior dependencies, which can degrade inference when the true posterior is strongly coupled. Second, the prior may be mismatched to the encoder-induced marginal over latent codes (the aggregated posterior $q(\mathbf{z}) = \frac{1}{N}\sum_{n=1}^{N} q_\phi(\mathbf{z} \mid \mathbf{x}_n)$), creating regularization pressure that suppresses learned dependencies [Takahashi et al., 2019]. Third, the decoder architecture controls how much information is carried by $\mathbf{z}$: overly expressive decoders can lead to posterior collapse [Bowman et al., 2016, Chen et al., 2017], whereas overly restrictive decoders can force the variational approximation to compensate.

Related work addresses these limitations by increasing flexibility in different components of the VAE, often modeling dependencies implicitly rather than exposing an explicit and interpretable dependency structure. On the inference side, normalizing flows enrich variational posteriors by transforming simple base distributions into more flexible ones [Rezende and Mohamed, 2015, Durkan et al., 2019]. On the prior side, learned priors such as VampPrior and implicit optimal-prior approaches reduce mismatch to the aggregated posterior [Tomczak and Welling, 2018, Takahashi et al., 2019]. Other approaches increase generative flexibility through hierarchical latent-variable structure or expressive autoregressive decoders, allowing dependencies to be captured by the generative model or likelihood rather than by

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:5601–5622.

an explicit structure in the variational posterior [Sønderby et al., 2016, Gulrajani et al., 2017].

A complementary line of work models dependencies explicitly through copulas, which separate univariate marginal behavior from multivariate dependence. Copula variational inference uses this separation to construct dependent variational posteriors from separately specified univariate marginals [Tran et al., 2015]. Within VAEs, copulas have been used on the decoder side, for example to model dependence among mixed-type observed variables [Suh and Choi, 2016], and on the inference side, for example to model correlations among latent variables in language modeling and collaborative filtering VAEs [Wang and Wang, 2019, Zhong et al., 2021]. Our work is closest to the inference-side copula direction, but differs in imposing a sparse, sample-specific tree structure on the variational copula. This makes latent dependence an explicit object of study, while allowing us to examine how structured inference interacts with prior alignment and likelihood-side dependence modeling.

To model latent dependencies explicitly, we work in the copula domain, which separates univariate marginals from the dependence structure [Nelsen, 2006]. In our VAE, we replace the mean-field variational posterior with a *tree-copula* variational posterior [Kirshner, 2007], allowing $q_\phi(\mathbf{z} \mid \mathbf{x})$ to capture sparse pairwise dependencies between latent dimensions; the resulting dependency tree provides a compact and interpretable summary of latent coupling. We further study prior–posterior alignment via an Average-of-Trees (AoT) copula prior in selected configurations, and a low-rank copula-based decoder [Suh and Choi, 2016] as a middle ground between factorized and fully autoregressive likelihoods. We evaluate these design choices using IWAE bounds [Burda et al., 2016] on image benchmarks, with a supplementary text-domain experiment in the appendix.[1]

The main contributions of this work are as follows.

Copula perspective on the ELBO. We decompose $D_{KL}\left(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p_\lambda(\mathbf{z})\right)$ into marginal KL terms and a copula-level mismatch term (Equation (6)), clarifying dependence regularization in both directions: with an independent prior, the dependence term reduces to the total correlation of $q_\phi(\mathbf{z} \mid \mathbf{x})$, while a dependent prior paired with an independent variational copula induces additional regularization through the variational marginals.

Tree-copula variational posterior. We introduce an amortized variational family in which $q_\phi(\mathbf{z} \mid \mathbf{x})$ uses flexible learned marginals and a tree-structured copula to capture sparse latent dependencies. The encoder predicts pair-copula dependence parameters, from which we construct the Chow–Liu tree [Chow and Liu, 1968], while retaining tractable sampling and density evaluation.

[1]Code is available at `https://github.com/YardenRachamim/tree-copula-vae`.

AoT-based copula prior. Since tree-copula variational posteriors induce a Mixture-of-Trees distribution at the aggregated-posterior level [Takahashi et al., 2019], we use an Average-of-Trees (AoT) copula [Kirshner, 2007] as a tractable surrogate prior, with efficient density evaluation via the Matrix-Tree theorem [Tutte, 1984].

Empirical evaluation. We evaluate on correlated dSprites [Matthey et al., 2017] and Fashion-MNIST [Xiao et al., 2017] using IWAE bounds, studying when structured dependence in $q_\phi(\mathbf{z} \mid \mathbf{x})$, prior alignment, and a rank-1 Gaussian-copula decoder are each beneficial.

2 BACKGROUND

2.1 VARIATIONAL AUTOENCODERS

Variational autoencoders (VAEs) posit a latent-variable generative model with joint density $p_\theta(\mathbf{x}, \mathbf{z}) = p_\lambda(\mathbf{z})\, p_\theta(\mathbf{x} \mid \mathbf{z})$, where $p_\lambda(\mathbf{z})$ is a prior over latents and $p_\theta(\mathbf{x} \mid \mathbf{z})$ is a decoder [Kingma and Welling, 2014]. Since the marginal likelihood $p_\theta(\mathbf{x})$ is generally intractable, VAEs introduce an amortized variational posterior $q_\phi(\mathbf{z} \mid \mathbf{x})$ and maximize the evidence lower bound (ELBO)[2]:

$$\begin{aligned} \log p_\theta(\mathbf{x}) - D_{KL}\left(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p_\theta(\mathbf{z} \mid \mathbf{x})\right) \\ = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right] \\ - D_{KL}\left(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p_\lambda(\mathbf{z})\right) . \end{aligned} \tag{1}$$

The first term on the RHS encourages reconstructions that explain $\mathbf{x}$ given $\mathbf{z}$, while the KL term regularizes the encoder toward the prior. A standard baseline uses a mean-field (MF) factorization $q_\phi(\mathbf{z} \mid \mathbf{x}) = \prod_{i=1}^{d} q_\phi(z_i \mid \mathbf{x})$ with an isotropic Gaussian prior $p_\lambda(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$.

The Inference Gap. For fixed generative parameters, the gap between the true log-likelihood and the ELBO equals $\log p_\theta(\mathbf{x}) - \mathcal{L}(\theta, \phi, \lambda; \mathbf{x}) = D_{KL}\left(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p_\theta(\mathbf{z} \mid \mathbf{x})\right)$, termed the *inference gap* [Cremer et al., 2018]. This gap further decomposes into an *approximation gap* (the family $\mathcal{Q}$ cannot represent the true posterior) and an *amortization gap* (a shared encoder cannot attain the per-point optimum in $\mathcal{Q}$) [Shu et al., 2018]. Under a mean-field family $\mathcal{Q}$, dependencies in the true posterior $p_\theta(\mathbf{z} \mid \mathbf{x})$ generally induce a nonzero approximation gap that cannot be eliminated by better optimization alone, motivating richer variational families beyond MF.

The Aggregated Posterior. The *aggregated posterior* (encoder marginal) is the latent distribution induced by the encoder under the data distribution, $q(\mathbf{z}) \triangleq \mathbb{E}_{p_{\text{data}}(\mathbf{x})}\left[q_\phi(\mathbf{z} \mid \mathbf{x})\right]$, i.e., a mixture over per-datapoint variationals; given samples $\{\mathbf{x}_n\}_{n=1}^{N}$, it is approximated by

[2]We denote the ELBO by $\mathcal{L}(\theta, \phi, \lambda; \mathbf{x})$

$\hat{q}(\mathbf{z}) = \frac{1}{N}\sum_{n=1}^{N} q_\phi(\mathbf{z} \mid \mathbf{x}_n)$. For fixed encoder and decoder, the ELBO (Equation (1)) is maximized when the prior matches this aggregate, $p^\star_\lambda(\mathbf{z}) = q(\mathbf{z})$ [Takahashi et al., 2019]. Since evaluating an N-component mixture is prohibitive at scale, this principle motivates tractable learned priors that approximate the (often multimodal) geometry and dependence structure of encoder-induced latents. Decoder expressiveness also shapes how strongly the KL term enforces alignment between $q_\phi(\mathbf{z} \mid \mathbf{x})$ and $p_\lambda(\mathbf{z})$. Highly expressive (e.g., autoregressive) decoders may model $\mathbf{x}$ directly, in which case the ELBO is improved by making $q_\phi(\mathbf{z} \mid \mathbf{x})$ match $p_\lambda(\mathbf{z})$, yielding *posterior collapse* [Bowman et al., 2016]. Conversely, overly restrictive factorized decoders can force $\mathbf{z}$ to carry essentially all structure, inducing a true posterior too complex for simple variational families. This tension motivates a middle-ground decoder design (details deferred to Section 3).

2.2 COPULAS

Copula theory provides a complementary viewpoint on multivariate modeling by separating *marginal behavior* from the *dependence structure* that couples coordinates [Nelsen, 2006]. This separation is particularly attractive for latent-variable models: one can retain simple marginal families while enriching the coupling, enabling structured priors and posteriors without sacrificing tractability [Tran et al., 2015].

Sklar's Theorem. Let $\mathbf{z}$ be a continuous random vector with marginal CDFs $\{F_i\}_{i=1}^d$ and marginal densities $\{p(z)_i\}_{i=1}^d$, and define copula coordinates $u_i \triangleq F_i(z_i) \in [0,1]$. Sklar's theorem states that there exists a copula $C : [0,1]^d \to [0,1]$ such that $F_{\mathbf{z}}(\mathbf{z}) = C(u_1, \ldots, u_d)$, and when the marginals are continuous this copula is unique [Sklar, 1959]. If C is differentiable, its copula density $c(\mathbf{u}) = \frac{\partial^d}{\partial u_1 \cdots \partial u_d} C(\mathbf{u})$ yields the density-level factorization

$$p(\mathbf{z}) \;=\; c\big(F_1(z_1), \ldots, F_d(z_d)\big) \prod_{i=1}^{d} p(z_i)\,. \tag{2}$$

Thus dependence is isolated entirely in c, while univariate behavior is carried by the marginals. In particular, a mean-field factorization corresponds (for continuous marginals) to the independence copula $c_\pi(\mathbf{u}) \equiv 1$, so moving beyond MF amounts to learning a nontrivial copula density.

Copula Entropy and Mutual Information. Substituting the density-level Sklar factorization (Equation (2)) into the definition of differential entropy and applying the change of variables $u_i = F_i(z_i)$ yields $\mathbb{H}_p(\mathbf{z}) = \sum_{i=1}^d \mathbb{H}_p(z_i) + H_c(\mathbf{z})$, where the *copula entropy* is $H_c(\mathbf{z}) \triangleq -\mathbb{E}_{c(\mathbf{u})}[\log c(\mathbf{u})]$. Since the copula has uniform marginals on $[0,1]^d$, H_c depends only on the dependence structure (equivalently, it is invariant to strictly increasing marginal reparameterizations). Letting $c_\pi(\mathbf{u}) \equiv 1$ denote the independence copula density, we have $H_c(\mathbf{z}) = -D_{KL}(c(\mathbf{u}) \,\|\, c_\pi(\mathbf{u})) \le 0$ (since $\log c_\pi \equiv 0$), with equality iff the coordinates are independent.

For $d = 2$, combining the decomposition above with the mutual information (MI) $\mathbb{I}_p(z_i; z_j) = \mathbb{H}_p(z_i) + \mathbb{H}_p(z_j) - \mathbb{H}_p((z_i, z_j))$ gives

$$\mathbb{I}_p(z_i;\, z_j) \;=\; -H_c(z_i,\, z_j)\,. \tag{3}$$

More generally, Ma and Sun [2011] showed that total correlation satisfies $\mathbb{I}_p(z_1; \ldots; z_d) \triangleq \sum_{i=1}^d \mathbb{H}_p(z_i) - \mathbb{H}_p(\mathbf{z}) = -H_c(\mathbf{z})$. Thus, Chow–Liu edge weights $w_{ij} = \mathbb{I}_p(z_i; z_j)$ (Section 2.3) can be interpreted equivalently as negative copula entropies of bivariate margins, directly linking MI-based structure learning to dependence modeling in the copula domain.

KL Divergence Decomposition. Beyond entropy and mutual information, the copula view also yields a KL identity that separates marginal effects from dependence mismatch. Consider two d-dimensional densities $q(\mathbf{z})$ and $p(\mathbf{z})$ with continuous marginals $\{q_i(z_i)_i\}_{i=1}^d$ and $\{p_i(z_i)_i\}_{i=1}^d$, marginal CDFs $\{F_i\}_{i=1}^d$ and $\{G_i\}_{i=1}^d$, and copula densities c_q and c_p as in Equation (2). Let $\mathbf{u} = F_1(z_1), \ldots, F_d(z_d)$ with $u_i \triangleq F_i(z_i)$, and define *aligned* copula coordinates $\mathbf{v} \in [0,1]^d$ by $v_i \triangleq G_i\big(F_i^{-1}(u_i)\big)$. Then, as shown by Tran [2018],

$$\begin{aligned} D_{KL}(q(\mathbf{z}) \,\|\, p(\mathbf{z})) = &\sum_{i=1}^{d} D_{KL}(q_i(z_i) \,\|\, p_i(z_i)) \\ &+ D_{KL}(c_q(\mathbf{u}) \,\|\, c_p(\mathbf{v}))\,. \end{aligned} \tag{4}$$

We use this separation in Section 3 to distinguish marginal regularization from dependence-structure regularization when analyzing structured priors and posteriors.

2.3 TREE-STRUCTURED DISTRIBUTIONS

To introduce dependence beyond mean-field while retaining tractable inference, one can restrict the latent joint distribution to factorize over a tree $T = (V, E)$. A distribution $p_T(\mathbf{z})$ is *tree-structured* if it admits the undirected factorization $p_T(\mathbf{z}) = \prod_{i=1}^d p(z_i) \prod_{(i,j)\in E} \frac{p(z_i, z_j)}{p(z_i)p(z_j)}$, or, equivalently (for any rooting of T), the directed form $p_T(\mathbf{z}) = \prod_{i=1}^d p\big(z_i \mid z_{T(i)}\big)$, where $T(i)$ denotes the parent of node i. With only $|E| = d-1$ edge factors, trees capture sparse pairwise deviations from independence while enabling efficient exact density evaluation, marginalization, and sampling via standard message passing [Koller and Friedman, 2009].

Chow–Liu Algorithm. For a target joint distribution $p(\mathbf{z})$, Chow and Liu [1968] showed that the KL-optimal tree-structured approximation is obtained by choosing $T^\star =$

$\arg\max_T \sum_{(i,j)\in E_T} \mathbb{I}_p(z_i; z_j)$, equivalently the maximum-weight spanning tree (MWST) of the complete graph on $\{1, \ldots, d\}$ with edge weights $w_{ij} = \mathbb{I}_p(z_i; z_j)$. This MWST can be computed efficiently (e.g., Kruskal in $\mathcal{O}(d^2 \log d)$, or $\mathcal{O}(d^2)$ with a dense-graph Prim implementation). Thus, learning an optimal tree structure reduces to selecting a sparse set of pairwise dependencies that is globally optimal under the KL projection onto tree-structured distributions.

Matrix Tree Theorem. To enable differentiable learning and averaging over dependency structures, one can define a distribution over *all* spanning trees with nonnegative edge weights, where a tree's probability is proportional to the product of its edge weights. Kirchhoff's Matrix Tree Theorem shows that the partition function equals $\det(L^*)$, where L^* is the reduced weighted graph Laplacian obtained by deleting any one row and column [Tutte, 1984]. Moreover, edge inclusion probabilities (Kirchhoff marginals) can be computed from entries of $(L^*)^{-1}$, so both normalization and marginals are available in $\mathcal{O}(d^3)$ time via a single linear solve. These marginals provide a differentiable surrogate for hard maximum-weight spanning-tree selection and enable tractable averaging over tree structures.

3 METHODOLOGY

We develop copula-based identities that reinterpret Chow–Liu tree selection and decompose the ELBO into marginal and copula terms (Section 3.1). These results guide the Tree-Copula VAE design (Section 3.2): a tree-copula variational posterior with differentiable structure learning, prior options that reduce dependence mismatch with the variational family, and decoder adjustments that control posterior complexity.

3.1 COPULA AND VAE: ANALYSIS

In this section, we derive copula-based analytical identities that justify the modeling choices in Section 3.2. We analyze the Chow–Liu algorithm (Section 2.3) and the Evidence Lower Bound (Section 2.1 and Equation (1)) through the lens of copula theory (Section 2.2). As established in Section 2.2, for continuous marginals the mean-field assumption corresponds to the independence copula $c_\pi(\mathbf{u}) \equiv 1$; moving beyond MF therefore amounts to learning a nontrivial dependence structure while retaining tractable inference.

3.1.1 Chow–Liu via Copulas

Classically, the Chow–Liu theorem states that, for a given joint distribution $p(\mathbf{x})$ over $\{X_i\}_{i=1}^d$, the optimal tree-dependent approximation in $D_{KL}(\cdot \,\|\, \cdot)$ is obtained by the maximum-weight spanning tree (MWST), with edge weights given by pairwise mutual information. For continuous variables, this result can be expressed through the copula framework: by Equation (3), $\mathbb{I}_p(X_u; X_v) = -H_c(X_u, X_v)$, so the Chow–Liu tree equivalently selects the edges with largest negative bivariate copula entropy.

Proof sketch. Fix a tree T. By the standard Chow–Liu KL-projection argument, the optimal projection onto the family of densities that factorize according to T matches the relevant univariate marginals and bivariate edge marginals of p. Applying the Sklar density factorization (Equation (2)), the marginal terms are constant with respect to T, while the tree-dependent part is, up to constants, $-\sum_{(u,v)\in E_T} \mathbb{I}_p(X_u; X_v)$. Therefore, minimizing $D_{KL}(p \,\|\, p_T)$ over tree-dependent approximations is equivalent to maximizing the sum of pairwise mutual informations over tree edges. Using Equation (3), this is equivalently maximizing the sum of negative bivariate copula entropies, hence the MWST is KL-optimal. The full proof is provided in Section A.

This copula-form view justifies the per-datapoint MWST construction used in Section 3.2: for each conditional variational posterior $q_\phi(\mathbf{z} \mid \mathbf{x})$, the selected tree captures the strongest pairwise latent dependencies while retaining tractable inference.

3.1.2 ELBO - Copula Perspective

By decomposing the ELBO using the copula-based KL divergence (Equation (4)), we distinguish between the regularization of marginals and that of the dependence structure. This formulation reveals that an independence assumption in either the prior or variational distribution implicitly penalizes *unmatched* dependencies in the other:

$$\mathcal{L}(\theta, \phi, \lambda; \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \sum_{i=1}^{d} D_{KL}\left(q_\phi(z_i|\mathbf{x}) \,\|\, p_\lambda(z_i)\right) \tag{5}$$

$$- \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{c\big(\{Q_\phi(z_i|\mathbf{x})\}_{i=1}^d\big)}{c\big(\{P_\lambda(z_i)\}_{i=1}^d\big)}\right] \tag{6}$$

We analyze Equation (6) under two scenarios (with c_π denoting the independence copula): (1) $c_\lambda = c_\pi$ while c_ϕ admits dependence; (2) $c_\phi = c_\pi$ while c_λ admits dependence.

First scenario (independent prior copula). Let $\mathbf{u} := \big(Q_\phi(z_i|\mathbf{x})\big)_{i=1}^d$. If $c_\lambda = c_\pi$, then the copula-mismatch term reduces to the negative copula entropy, i.e., the *total correlation* of the variational posterior:

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log c_\phi(\mathbf{u})\right] = -\mathbb{H}_{c_\phi}(\mathbf{u}) = \mathbb{I}_{q_\phi(\mathbf{z}|\mathbf{x})}(Z_1; \ldots; Z_d)\,.$$

Thus, maximizing the ELBO with an independent prior implicitly penalizes dependence in $q_\phi(\mathbf{z}|\mathbf{x})$, pushing its copula

toward independence.

Second scenario (independent variational). In this scenario, Equation (6) becomes (see Appendix Section C for the derivation)

$$-\mathbb{E}_{c_\pi(\mathbf{u})}\left[\log c_\lambda\Big(\left\{P_\lambda(Q_\phi^{-1}(u_i|\mathbf{x}))\right\}_{i=1}^d\Big)\right].$$

In this case ($c_\phi = c_\pi$), the prior copula is evaluated on a transformed space determined by the marginal mapping $P_\lambda \circ Q_\phi^{-1}$. Two limiting regimes clarify its effect. **(i) Matched marginals:** if $Q_\phi(z_i) = P_\lambda(z_i)\ \forall i$, then this transformation is the identity and the term reduces to a cross-entropy between the (independent) variational copula and c_λ; consequently, maximizing the ELBO encourages c_λ to move toward the independence copula c_π. **(ii) Fixed prior marginals / learned variational marginals:** the ELBO can be improved either by adapting c_λ or by changing Q_ϕ so that the transformed coordinates fall in higher-density regions of c_λ. Therefore, the variational marginals are regularized both (i) explicitly via the marginal KL terms (Equation (5)) and (ii) implicitly through the copula term (Equation (6)) via $P_\lambda \circ Q_\phi^{-1}$, which can create tension between these objectives.

Implications for Model Design. Chow–Liu optimality motivates a tree-copula variational family as a minimal, tractable upgrade to mean-field. Furthermore, the ELBO decomposition above clarifies the role of the prior: aligning p_λ with q_ϕ mitigates mismatch penalties, whereas an independent prior acts as a regularizer that penalizes variational total correlation. Accordingly, in Section 3.2 we introduce a tree-structured variational copula and explore both standard and structure-aligned priors.

3.2 VAE WITH TREE COPULA

Variational Autoencoders (VAEs) comprise a prior $p_\lambda(\mathbf{z})$, a likelihood $p_\theta(\mathbf{x} \mid \mathbf{z})$, and a variational posterior $q_\phi(\mathbf{z} \mid \mathbf{x})$. In this section, we motivate our choices for q_ϕ, p_λ, and p_θ using prior work and the copula/inference-gap analyses in Sections 2.1 and 3.1. Concretely, Chow–Liu optimality (Section 3.1.1) motivates a tree-structured copula encoder, enlarging $\mathcal{Q}$ beyond factorized Gaussians to target the *approximation gap*. The copula-ELBO decomposition guides the corresponding prior choice, while the inference-gap perspective motivates decoder choices that control the complexity of the induced posterior $p_\theta(\mathbf{z} \mid \mathbf{x})$ so that it remains amenable to approximation by our tree-copula variational family. Accordingly, we next specify the variational family by parameterizing $q_\phi(\mathbf{z} \mid \mathbf{x})$ with continuous marginals and a Chow–Liu tree copula whose edges are predicted by an amortized encoder.

3.2.1 Variational Tree Distribution

We model the dependence structure of the variational posterior using a tree-structured copula. Unlike a Gaussian distribution, which captures only linear correlations, copulas can represent richer dependence patterns (e.g., tail dependence) while retaining tractable inference on trees [Nelsen, 2006].

To construct $q_\phi(\mathbf{z} \mid \mathbf{x})$, we employ an amortized inference network (Figure 1). Given $\mathbf{x}$, the encoder outputs (i) parameters of the univariate marginals and (ii) pairwise edge scores (copula parameters) for all pairs (i, j), totaling $\binom{d}{2}$ pairwise copula parameters. We use these scores as MWST edge weights; since they are monotone in the pairwise mutual informations implied by the chosen copula family [Tenzer and Elidan, 2016], the resulting MWST coincides with the Chow–Liu optimal tree for the pairwise marginals implied by the current amortized variational distribution $q_\phi(\mathbf{z} \mid \mathbf{x})$.

We restrict marginals to continuous families with available CDF and inverse CDF (e.g., Gaussian or univariate spline flows [Durkan et al., 2019]), and select copula families satisfying the monotonicity conditions required for MI-based edge scoring and optimization. The resulting tree-structured variational distribution supports efficient sampling and pathwise gradients via reparameterization [Tran et al., 2015].

Analytic Evaluation of Variational Entropy. A key computational advantage of the tree-copula parameterization is the analytic tractability of the variational copula (negative-) entropy, a distinct term within the copula-KL decomposition (Equation (6)). This term corresponds to the total correlation of the variational posterior, which, for a tree-structured distribution, decomposes into a sum of pairwise mutual information terms over the tree edges E_T:

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log c_\phi(\mathbf{u})\right] = \sum_{(u,v)\in E_T} \mathbb{I}_{q_\phi}(z_u; z_v). \tag{7}$$

Consequently, for copula families with closed-form bivariate mutual information, such as Gaussian or Student-t [Calsaverini and Vicente, 2009], this component of the training objective can be computed analytically.

Differentiable structure learning. For each datapoint $\mathbf{x}$, the encoder predicts bivariate copula dependence parameters $\{\psi_{ij}(\mathbf{x})\}_{i<j}$ for all candidate edges (i, j). We treat these copula parameters as *edge scores* and, in the hard variant, obtain a sample-specific Chow–Liu tree by running MWST on $\psi(\mathbf{x})$. However, MWST selection is discrete and therefore yields no useful gradients w.r.t. the scores.

Following the differentiable spanning-tree relaxations based on the stochastic softmax trick [Paulus et al., 2020] and Matrix-Tree computations [Tutte, 1984], we replace hard MWST selection with a temperature-controlled distribution over spanning trees:

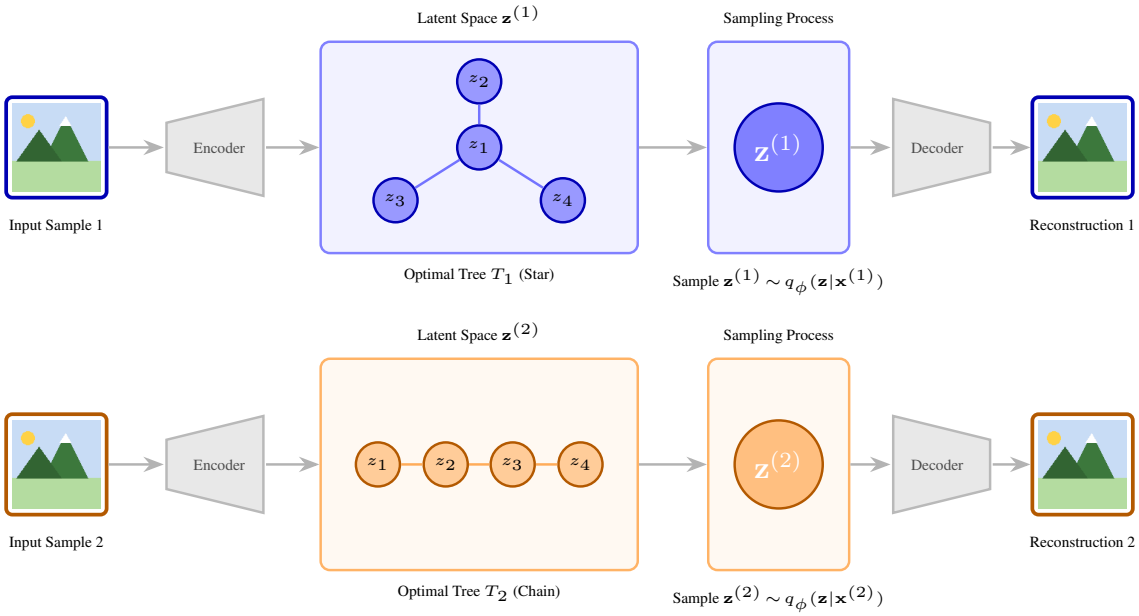


Figure 1: High-level overview of the Tree-Copula VAE framework.

$$p_\tau(T \mid \mathbf{x}) \propto \prod_{(i,j)\in E_T} \exp\big(\psi_{ij}(\mathbf{x})/\tau\big), \qquad \beta_{ij}(\mathbf{x};\tau) \triangleq \mathbb{P}_{T\sim p_\tau(\cdot|\mathbf{x})}\big[(i,j) \in E_T\big]. \tag{8}$$

where $\beta_{ij}(\mathbf{x};\tau)$ denotes the *Kirchhoff marginal* (edge inclusion probability), computed via the Matrix-Tree theorem [Tutte, 1984]. As $\tau \downarrow 0$, $p_\tau(T \mid \mathbf{x})$ concentrates on the MWST, recovering hard-tree behavior.

Applying the surrogate objective. Recall that the copula-ELBO decomposition in Equation (6) separates the KL term into marginal mismatch and a *dependence-structure* contribution involving the copula factors. We use the soft tree marginals $\beta(\mathbf{x};\tau)$ from Equation (8) to obtain a differentiable estimator of this dependence term via the AoT copula density induced by the same edge marginals, $\log c_{\mathrm{AoT}}\big(\mathbf{u};\beta(\mathbf{x};\tau)\big)$ [Kirshner, 2007]. When the chosen bivariate copula family admits closed-form pairwise mutual information (e.g., Gaussian or Student-t), we approximate this using closed-form pairwise MI as $\sum_{i<j}\beta_{ij}(\mathbf{x};\tau)\,\mathbb{I}_{q_\phi}(z_i;z_j)$.

The full training step, including the hard-tree posterior construction, soft-MWST surrogate, and prior-alignment terms, is summarized in Algorithm 1.

3.2.2 Prior Selection

The second component in the VAE that controls the latent space is the prior $p_\lambda(\mathbf{z})$. For any dependent variational family $\mathcal{Q}$, p_λ induces a dependence-mismatch penalty via the copula term in the ELBO (Equation (6)); in our case, restricting $\mathcal{Q}$ to tree copulas renders this term amenable to efficient evaluation and structure learning.

Independent prior. A standard baseline uses an independent prior copula, $c_\lambda \equiv c_\pi$ (e.g., $\mathcal{N}(\mathbf{0},\mathbf{I})$). In this case, the second term in Equation (6) vanishes, and the dependence regularizer reduces exactly to the variational total correlation. For tree copulas this term can be computed analytically, rendering the entire KL divergence closed-form. However, this choice introduces a fundamental tension: Chow–Liu selects the tree by *maximizing* pairwise MI, while the ELBO with $c_\lambda \equiv c_\pi$ *penalizes* the same MI through the KL term. Consequently, the model retains dependencies only when the reconstruction gain outweighs this penalty.

Average-of-Trees (AoT) Prior for Aggregated-Posterior Alignment. To reduce this tension, we seek a dependence-aware prior ($c_\lambda \neq c_\pi$) aligned with the encoder family. In principle, the ELBO-optimal prior is the aggregated posterior $q(\mathbf{z})$ [Takahashi et al., 2019]; since our amortized encoder predicts sample-dependent trees, $q(\mathbf{z})$ effectively forms a mixture of tree-copula posteriors. Direct evaluation of this mixture is intractable, so we approximate it with the *AoT Copula*. AoT defines a distribution over all spanning trees with global (datapoint-independent) edge parameters, aligning with the encoder's structural inductive bias while remaining efficiently computable in $\mathcal{O}(d^3)$ via the Matrix-Tree theorem. Importantly, AoT does not aim to match $q(\mathbf{z})$ pointwise in $\mathbf{x}$; rather, it captures structural variability by averaging over tree structures while learning shared edge parameters, thereby approximating $q(\mathbf{z})$ *in expectation* at

the level of dependence structure.

Alternative Priors. Our framework is modular and supports other advanced priors. **Normalizing Flows** [Rezende and Mohamed, 2015] can model complex non-linear dependencies but often incur higher computational costs than tree-based methods. The **VAMP Prior** [Tomczak and Welling, 2018] approximates the aggregated posterior via a mixture of variational components; while highly effective for multimodal data, it lacks the specific mixture-of-trees inductive bias of the AoT copula. More generally, moving beyond an independent (including AoT copula), typically sacrifices full analytic evaluation of the copula-KL and requires Monte Carlo estimation of the copula cross-entropy term in Equation (6), which can increase estimator variance.

3.2.3 Decoder Adjustments

While the variational family $\mathcal{Q}$ determines the capacity of the encoder, the difficulty of the inference task is dictated by the generative model itself. Specifically, the choice of decoder $p_\theta(\mathbf{x} \mid \mathbf{z})$ fundamentally shapes the topology of the true posterior Using Equation (1) and the inference-gap identity in Section 2.1, the training objective can be viewed as maximizing the marginal likelihood subject to the minimum achievable inference gap allowed by $\mathcal{Q}$:

$$\max_{\theta,\phi} \mathbb{E}_{p_{\text{data}}(\mathbf{x})}\left[\mathcal{L}(\theta,\phi,\lambda;\mathbf{x})\right] = \max_{\theta}\Bigg[\mathbb{E}_{p_{\text{data}}(\mathbf{x})}\left[\log p_\theta(\mathbf{x})\right] - \min_{\phi}\ \mathbb{E}_{p_{\text{data}}(\mathbf{x})}\left[D_{KL}\left(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p_\theta(\mathbf{z} \mid \mathbf{x})\right)\right]\Bigg]$$

This perspective recasts ELBO maximization as approximate maximum-likelihood estimation of $p_\theta(\mathbf{x})$, regularized by the minimum inference gap that $\mathcal{Q}$ can achieve [Cremer et al., 2018]. The decoder controls the difficulty of this regularizer through its effect on the true posterior.

Since $p_\theta(\mathbf{z} \mid \mathbf{x}) \propto p_\theta(\mathbf{x} \mid \mathbf{z})\, p_\lambda(\mathbf{z})$, the posterior dependence is induced by the likelihood–prior product (and under an independent prior, primarily by the decoder). A factorized decoder $p_\theta(\mathbf{x} \mid \mathbf{z}) = \prod_{i=1}^{D} p_\theta(x_i \mid \mathbf{z})$ makes $\mathbf{z}$ the *sole* source of dependence among output dimensions: all correlations in $\mathbf{x}$ must be routed through $\mathbf{z}$, inducing a highly coupled true posterior that may exceed the capacity of even a tree-structured variational family. Conversely, powerful autoregressive decoders can model local dependencies directly, but often absorb so much signal that the optimal posterior collapses to the prior, $p_\theta(\mathbf{z} \mid \mathbf{x}) \approx p_\lambda(\mathbf{z})$, rendering the structured encoder redundant [Bowman et al., 2016].

To balance these extremes, we adopt a *Rank-1 Gaussian Copula decoder* [Suh and Choi, 2016], which enriches a factorized Gaussian likelihood with a rank-1 Gaussian copula in observation space; for Gaussian marginals this corresponds to a covariance of the form $\sigma^2\mathbf{I} + \mathbf{w}_\theta(\mathbf{z})\mathbf{w}_\theta(\mathbf{z})^\top$ for a decoder-predicted rank-1 factor $\mathbf{w}_\theta(\mathbf{z})$. This decoder explicitly models low-rank local correlations between output dimensions, effectively explaining away local noise dependencies. This design aims to better align the complexity of $p_\theta(\mathbf{z} \mid \mathbf{x})$ with the expressive regime of the tree-copula variational family, thereby *potentially* reducing the expected inference gap defined above while mitigating collapse.

4 EXPERIMENTS

We evaluate Tree-Copula VAE on two public image datasets that isolate the effect of structured dependence in the variational posterior, the prior, and the likelihood: a controlled synthetic setting (Correlated dSprites [Matthey et al., 2017]) and a natural image benchmark (Fashion-MNIST [Xiao et al., 2017]). A supplementary text-domain experiment on 20 Newsgroups [Lang, 1995] with ProdLDA [Srivastava and Sutton, 2017] is reported in the appendix as an additional prior-mismatch check (Section F). We report held-out IWAE@512 [Burda et al., 2016] and standard ELBO diagnostics for all experiments.

Setup Each model variant is specified by the triple `Prior-Variational-Likelihood`, where `-` separates components and `_` appears only within a component name. Table 1 lists the component codes. For example, `SN-TC_ST-R1GC_G` denotes a standard-normal prior, a Student-t tree-copula variational family, and a rank-1 Gaussian-copula likelihood. Within each dataset we hold the encoder/decoder backbone, latent dimensionality, and optimizer fixed. All models use a β-annealed objective [Higgins et al., 2017]; tree-copula variants further anneal the soft tree-selection temperature τ. Full experimental setup is in Section E.

4.1 CORRELATED DSPRITES

We construct a controlled subset of dSprites by filtering to a thin, constraint-defined manifold that couples three ground-truth factors (scale, orientation, and horizontal position), inducing strong and known dependence among these factors (construction details in Section E). With latent dimensionality d=3, the resulting dependence cannot be captured by a `MF` posterior, making this a targeted test of structured variational inference.

As shown in Table 2, the Student-t tree copula (`TC_ST`) achieves the strongest held-out density under both decoder families, while `TC_GC` underperforms `MF` with the factorized decoder but surpasses it once the dependent likelihood is introduced. This is consistent with the motivation in Section 3.2.3: modeling limited observation-space dependence can reduce unnecessary posterior coupling from a strictly factorized decoder, placing the true posterior in a regime more amenable to tree-structured approximation. Indepen-

Table 1: Component Codes for the Naming Convention.

Slot	Code	Description
Prior	SN	Standard normal prior
	AoT_SG_GC	Average-of-Trees copula prior with Gaussian pair-copula edges
	AoT_SG_ST	Average-of-Trees copula prior with Student-t pair-copula edges
Variational posterior	MF	Mean-field diagonal Gaussian variational posterior
	TC_GC	Tree-copula variational posterior with Gaussian pair-copula edges
	TC_ST	Tree-copula variational posterior with Student-t pair-copula edges
Likelihood	FG	Factorized Gaussian likelihood
	R1GC_G	Rank-1 Gaussian-copula likelihood with Gaussian marginals

Table 2: Correlated dSprites: Test-Set Results.

Model	IWAE@512↑	$\log p_\theta(\mathbf{x} \mid \mathbf{z})$ ↑	KL↓
SN-MF-FG	4983.26	4961.50	27.68
SN-TC_GC-FG	4958.74	4942.36	**25.95**
SN-TC_ST-FG	**5062.38**	**5046.28**	29.33
SN-MF-R1GC_G	5683.74	5668.10	**30.01**
SN-TC_GC-R1GC_G	5726.95	5711.13	31.16
SN-TC_ST-R1GC_G	**5859.30**	**5842.94**	33.39

dently of the inference family, moving from FG to R1GC_G yields a substantial IWAE improvement, highlighting the importance of modeling local pixel correlations.

To attribute gains specifically to *variational dependence*, we compare SN-TC_ST-R1GC_G against SN-MF-R1GC_G, which share the same prior and likelihood and differ only in the variational. We compute a per-pixel marginal NLL gap map (formal definition in Section E.5) at the variational mean. As shown in Figure 2, positive gains concentrate near object–background boundaries, where errors in correlated global factors most strongly translate into sharp intensity mismatches; additional shapes are in Section E.7.

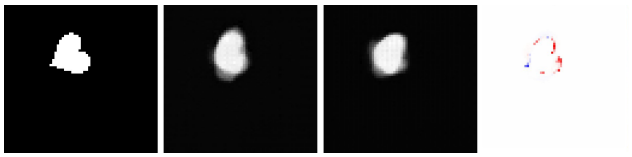


Figure 2: Correlated dSprites (heart): per-pixel marginal NLL gap under the same R1GC_G decoder. Left to right: input, TC_ST reconstruction, MF reconstruction, gap map. Warm colors indicate pixels where tree-copula inference assigns higher marginal likelihood.

4.2 FASHION-MNIST

We use Fashion-MNIST (28×28 grayscale, 10 classes) with uniform dequantization [Theis et al., 2016] and latent dimensionality d=4 to evaluate held-out density and interpretability of the learned tree structures.

Table 3: Fashion-MNIST: Test Results.

Model	IWAE@512↑	$\log p_\theta(\mathbf{x} \mid \mathbf{z})$ ↑	KL↓
SN-MF-R1GC_G	481.03	487.45	15.04
SN-TC_GC-R1GC_G	484.42	490.35	**14.90**
SN-TC_ST-R1GC_G	**485.69**	**491.90**	15.12

As shown in Table 3, tree-copula inference yields consistent density improvements over the MF baseline under the dependent likelihood (R1GC_G), with TC_ST achieving the highest IWAE@512. Absolute gains over MF are more modest than in the synthetic setting, likely because improvements on the informative foreground are diluted by near-constant background pixels, but the consistent direction supports the value of structured variational dependence on natural data.

Beyond density, we analyze whether the inferred latent trees capture semantic structure. For each test example, we extract the hard-MWST over latent dimensions with edges weighted by the pairwise mutual information implied by the predicted bivariate copulas, and encode its topology via the Prüfer sequence [Wang et al., 2009], yielding $d^{d-2} = 16$ labeled trees for d=4 that we treat as discrete assignments (details in Section E.4). As shown in Table 4, the variant with the strongest class–topology association (SN-TC_GC-R1GC_G) achieves NMI $= 0.17$ and weighted purity 0.32, indicating a modest but non-trivial alignment between class labels and inferred topology.

Table 4: Fashion-MNIST: tree–label association for d=4.

Model	NMI↑	Weighted purity↑
SN-TC_GC-FG	0.11	0.27
SN-TC_ST-FG	0.13	0.28
SN-TC_GC-R1GC_G	**0.17**	**0.32**
SN-TC_ST-R1GC_G	0.13	0.28

As visualized in Figure 3, the row-normalized conditional

$\widehat{P}(T \mid Y)$ shows class-dependent concentration on a small subset of topologies; in some cases a single topology accounts for close to half of the mass.

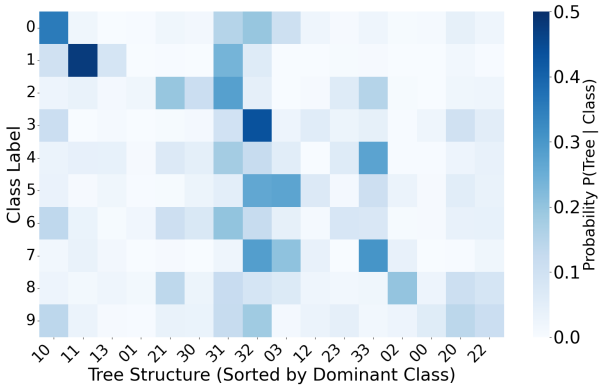


Figure 3: Fashion-MNIST latent-topology conditional frequencies (d=4, `SN-TC_GC-R1GC_G`). Each cell shows $\widehat{P}(T{=}t \mid Y{=}y)$, where T is the MWST topology over latent dimensions (Prüfer-encoded) and Y is the class label.

Prior alignment. The copula-ELBO analysis in Section 3.1.2 predicts that pairing a dependence-aware variational family with an independence prior implicitly penalizes variational total correlation. We test this on Fashion-MNIST by comparing the `SN` prior against the proposed `AoT` prior, keeping the dependent decoder (`R1GC_G`) fixed. These prior-learning runs use the same backbone and optimization as Section 4.2, but add i.i.d. Gumbel noise to the edge scores (the stochastic softmax trick; Paulus et al. 2020) to prevent early topology cmollapse and encourage exploration. all rows in Table 5 share the same training protocol for a fair within-table comparison.

Table 5: Fashion-MNIST (Prior Alignment): Test Results.

Model	IWAE@512 ↑
`SN-TC_GC-R1GC_G`	482.00
`AoT_SG_GC-TC_GC-R1GC_G`	**488.80**
`SN-TC_ST-R1GC_G`	483.58
`AoT_SG_ST-TC_ST-R1GC_G`	484.15

As shown in Table 5, the `AoT` prior yields a notable improvement for the Gaussian tree-copula variant, consistent with a structurally matched prior reducing the dependence-mismatch penalty. The gain is smaller for the Student-t variant, suggesting that heavier tails partially compensate for prior misalignment. An additional evaluation on 20 Newsgroups with a ProdLDA decoder is reported in Section F.

Summary. Across the benchmarks, tree-copula inference improves density when latent dependencies are present and yields interpretable discrete topologies with non-trivial class alignment. On Fashion-MNIST, learning a structurally matched AoT prior improves density for the Gaussian tree-copula variant, consistent with mitigating the dependence-mismatch penalty.

5 LIMITATIONS AND FUTURE WORK

Limitations. The tractability of our framework comes from restricting each model component to a simple dependence structure. The variational posterior uses a single tree-copula per datapoint, the prior is either independent or AoT, and the likelihood captures only rank-1 Gaussian-copula dependence in observation space. These choices make it possible to study the interaction between structured inference, prior alignment, and decoder-side dependence, but they cannot represent higher-order or strongly non-tree-like dependencies. The latent tree is also selected as a deterministic function of $\mathbf{x}$, so $q_\phi(\mathbf{z} \mid \mathbf{x})$ cannot express uncertainty over alternative tree structures. Scalability is limited by the complete candidate graph used here: dependence learning requires $\binom{d}{2}$ pairwise estimates, and Soft-MWST/AoT steps involve $O(d^3)$ operations. Thus, the current implementation is most appropriate for moderate-dimensional latent spaces with sparse pairwise dependence; at larger d, edge estimates may become costly and noisy, and training can be sensitive to KL annealing, Soft-MWST temperature schedules, and structure-learning noise.

Future work. Natural extensions are to relax these restrictions while preserving tractability. Richer copula families, such as vine copulas or mixtures of trees, could model dependencies beyond a single tree [Bedford and Cooke, 2002]. Structural uncertainty could be introduced by treating the tree T as a discrete latent variable and learning a joint $q_\phi(\mathbf{z}, T \mid \mathbf{x})$, using Gumbel-softmax-style relaxations to retain differentiability during training [Jang et al., 2017]. For scaling, trees could be learned over sparse candidate graphs rather than complete graphs, avoiding exhaustive evaluation of all $\binom{d}{2}$ edges in high dimensions. Sparse Gaussian graphical-model proposals offer one possible route, since precision-matrix sparsity encodes graph structure and could guide mutual-information-based edge scoring.

References

Tim Bedford and Roger M Cooke. Vines–a new graphical model for dependent random variables. *The Annals of Statistics*, 30(4):1031–1068, 2002.

Samuel Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, 2016.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *International Conference on Learning Representations*, 2016.

Rafael S Calsaverini and Renato Vicente. An information-theoretic approach to statistical dependence: Copula information. *EPL (Europhysics Letters)*, 88(6):68003, 2009.

Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. In *International Conference on Learning Representations*, 2017.

C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

Chris Cremer, Xuechen Li, and David Duvenaud. Inference suboptimality in variational autoencoders. In *International Conference on Machine Learning*, 2018.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In *Advances in Neural Information Processing Systems*, 2019.

Vincent Fortuin. Priors in Bayesian deep learning: A review. *International Statistical Review*, 90(3):563–591, 2022.

Ishaan Gulrajani, Kundan Kumar, Faruk Ahmed, Adrien Ali Taiga, Francesco Visin, David Vazquez, and Aaron Courville. PixelVAE: A latent variable model for natural images. In *International Conference on Learning Representations*, 2017.

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

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*, 2017.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2): 183–233, 1999.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Sergey Kirshner. Learning with tree-averaged densities and distributions. In *Advances in Neural Information Processing Systems*, 2007.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT Press, 2009.

Ken Lang. NewsWeeder: Learning to filter netnews. In *Proceedings of the 12th International Conference on Machine Learning*, 1995.

Jian Ma and Zengqi Sun. Mutual information is copula entropy. *Tsinghua Science & Technology*, 16(1):51–54, 2011.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dSprites: Disentanglement testing sprites dataset. `https://github.com/deepmind/dsprites-dataset`, 2017.

Marina Meilă and Tommi S. Jaakkola. Tractable Bayesian learning of tree belief networks. *Statistics and Computing*, 16(1):77–92, 2006. doi: 10.1007/s11222-006-5195-2.

Roger B Nelsen. *An introduction to copulas*. Springer, 2006.

Max B. Paulus, Dami Choi, Daniel Tarlow, Andreas Krause, and Chris J. Maddison. Gradient estimation with stochastic softmax tricks. In *Advances in Neural Information Processing Systems*, 2020.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, 2015.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

Rui Shu, Hung H. Bui, Shengjia Zhao, Mykel J. Kochenderfer, and Stefano Ermon. Amortized inference regularization. In *Advances in Neural Information Processing Systems*, 2018.

M Sklar. Fonctions de répartition à n dimensions et leurs marges. In *Annales de l'ISUP*, 1959.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *Advances in Neural Information Processing Systems*, 2016.

Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. In *International Conference on Learning Representations*, 2017.

Suwon Suh and Seungjin Choi. Gaussian copula variational autoencoders for mixed data. arXiv preprint arXiv:1604.04960, 2016.

Hiroshi Takahashi, Tomoharu Iwata, Yuki Yamanaka, Masanori Yamada, and Satoshi Yagi. Variational autoencoder with implicit optimal priors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Yaniv Tenzer and Gal Elidan. On the monotonicity of the copula entropy. arXiv preprint arXiv:1611.06714, 2016.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations*, 2016.

Jakub Tomczak and Max Welling. VAE with a VampPrior. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, 2018.

Dustin Tran, David M. Blei, and Edoardo M. Airoldi. Copula variational inference. In *Advances in Neural Information Processing Systems*, 2015.

Viet Hung Tran. Copula variational Bayes inference via information geometry. arXiv preprint arXiv:1803.10998, 2018.

William Thomas Tutte. *Graph theory*, volume 21. Addison-Wesley Reading, 1984.

Prince Zizhuang Wang and William Yang Wang. Neural Gaussian copula for variational autoencoder. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

Xiaodong Wang, Lei Wang, Yingjie Wu, et al. An optimal algorithm for Prüfer codes. *J. Softw. Eng. Appl.*, 2(2): 111–115, 2009.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747, 2017.

Ting Zhong, Guanyu Wang, Joojo Walker, Kunpeng Zhang, and Fan Zhou. Variational autoencoder with copula for collaborative filtering. In *Workshop on Deep Learning Practice for High-Dimensional Sparse Data with KDD*, 2021.

Appendix

Yarden Rachamim[1] **Shai Fine**[2]

[1]Computer Science Dept., Reichman University, Herzliya, Israel
[2]Data Science Institute, Reichman University, Herzliya, Israel

A CHOW-LIU TREE OPTIMALITY

In Chow and Liu [1968], Chow and Liu proved that the optimal tree-dependent approximation (in KL divergence) can be obtained by maximizing pairwise mutual information for discrete distributions. Here, we prove the analogous result for continuous random variables from a copula perspective.

Lemma A.1 (Chow-Liu Optimal Tree-Dependent Approximation Property)**.** *Let $\mathbf{x} = (x_1, \ldots, x_d)$ be a random vector with joint density $p(\mathbf{x})$, and let $T = (V, E)$ be a fixed tree structure over the index set $\{1, \ldots, d\}$. Let $\mathcal{P}_T$ denote the family of tree-dependent distributions (as in Section 2.3).*

The optimal tree-dependent approximation defined by:

$$p_T^*(\mathbf{x}) = \arg\min_{p_T \in \mathcal{P}_T} D_{KL}\left(p(\mathbf{x}) \,\|\, p_T(\mathbf{x})\right)$$

satisfies $\forall v \in V$,

$$p_T\big(x_v \mid x_{\mathrm{T}(v)}\big) \;=\; p\big(x_v \mid x_{\mathrm{T}(v)}\big),$$

where $\mathrm{T}(v)$ denotes the parent of node v in the tree T.

Proof. Let T be the tree structure of $p_T^*(\mathbf{x})$. By definition of the KL divergence,

$$D_{KL}\left(p(\mathbf{x}) \,\|\, p_T^*(\mathbf{x})\right) = \mathbb{E}_{p(\mathbf{x})}\left[\log \frac{p(\mathbf{x})}{p_T^*(\mathbf{x})}\right]$$

using the tree factorization

$$= \mathbb{E}_{p(\mathbf{x})}\left[\log \frac{p(\mathbf{x})}{\prod_{v \in V}[p_T^*(x_v|x_{T(v)})]}\right]$$

Using properties of logarithms and the linearity of expectation

$$\begin{aligned} &= \mathbb{E}_{p(\mathbf{x})}\left[\log p(\mathbf{x})\right] - \sum_{v \in V} \mathbb{E}_{p(\mathbf{x})}\left[\log p_T^*\big(x_v \mid x_{\mathrm{T}(v)}\big)\right] \\ &= -\mathbb{H}_p\left(X\right) - \sum_{v \in V} \mathbb{E}_{p(\mathbf{x})}\left[\log p_T^*\big(x_v \mid x_{\mathrm{T}(v)}\big)\right] \end{aligned}$$

The first term is the negative differential entropy $\mathbb{H}_p\left(X\right) = -\int p(\mathbf{x}) \, \log p(\mathbf{x}) \, d\mathbf{x}$

Marginalizing over the variables not appearing in $\big(x_v, x_{\mathrm{T}(v)}\big)$, we obtain

$$
\begin{aligned}
&= -\mathbb{H}_p\left(X\right) - \sum_{v \in V} \mathbb{E}_{p(x_v, x_{\mathrm{T}(v)})}\left[\log p_T^*\big(x_v \mid x_{\mathrm{T}(v)}\big)\right] \\
&= -\mathbb{H}_p\left(X\right) - \sum_{v \in V} \mathbb{E}_{p(x_{\mathrm{T}(v)})}\left[\mathbb{E}_{p(x_v \mid x_{T(v)})}\left[\log p_T^*\big(x_v \mid x_{\mathrm{T}(v)}\big)\right]\right]
\end{aligned}
$$

The first term $-\mathbb{H}_p\left(X\right)$, is constant with respect to p_T^*. For each v and fixed $x_{\mathrm{T}(v)}$, the inner integral

$$\int p\big(x_v \mid x_{\mathrm{T}(v)}\big)\, \log p_T^*\big(x_v \mid x_{\mathrm{T}(v)}\big)\, dx_v$$

is (up to a sign) the cross-entropy between $p(\cdot \mid x_{\mathrm{T}(v)})$ and $p_T(\cdot \mid x_{\mathrm{T}(v)})$. By Gibbs' inequality, this term is maximized if and only if $p(\cdot \mid x_{\mathrm{T}(v)}) = p_T(\cdot \mid x_{\mathrm{T}(v)})$, almost everywhere.

Therefore, the whole expression $D_{KL}\left(p(\mathbf{x}) \,\|\, p_T^*(\mathbf{x})\right)$ is minimized if and only if, for all $v \in V$,

$$p\big(x_v \mid x_{\mathrm{T}(v)}\big) = p_T^*\big(x_v \mid x_{\mathrm{T}(v)}\big),$$

which proves the lemma. □

To derive the full Chow–Liu result using copulas, we rewrite the lemma from a copula perspective:

Lemma A.2 (Chow-Liu Optimal Tree-Dependent Approximation Property with Copula)**.** *Given a joint probability distribution $p(\mathbf{x})$ and a fixed tree structure T, the best T-dependent approximation $p_T(\mathbf{x})$ satisfies*

$$\forall v \in V \qquad p_T(x_v) \,\cdot\, c_T\big(P_T(x_v),\, P_T(x_{\mathrm{T}(v)})\big) = p(x_v) \,\cdot\, c\big(P(x_v),\, P(x_{\mathrm{T}(v)})\big),$$

where $p_T(x_v)$ is the marginal distribution of X_v, $P_T(x_v)$ is its corresponding CDF, and $c_T\big(P_T(x_v),\, P_T(x_{\mathrm{T}(v)})\big)$ is the copula density of the edge $(X_v, X_T(v))$, all under $p_T(\mathbf{x})$

Proof. From Lemma A.1, the best T-dependent approximation satisfies for all $v \in V$,

$$p\big(x_v \mid x_{\mathrm{T}(v)}\big) = p_T\big(x_v \mid x_{\mathrm{T}(v)}\big),$$

Applying probability rules and Sklar's Theorem [Sklar, 1959] yields $\forall v \in V$:

$$
\begin{aligned}
p(x_v | x_{T(v)}) &= \frac{p(x_v, x_{T(v)})}{p(x_{T(v)})} \\
&= \frac{p(x_v) \cdot p(x_{T(v)}) \cdot c(P_v(x_v), P_{T(v)}(x_{T(v)}))}{p(x_{T(v)})} \\
&= p(x_v) \cdot c(P_v(x_v), P_{T(v)}(x_{T(v)}))
\end{aligned}
$$

We can now apply the same rules to $p_T(x_v) \,\cdot\, c_T\big(P_T(x_v),\, P_T(x_{\mathrm{T}(v)})\big)$ yielding:

$$p_T(x_v) \,\cdot\, c_T\big(P_T(x_v),\, P_T(x_{\mathrm{T}(v)})\big) = p(x_v) \,\cdot\, c\big(P(x_v),\, P(x_{\mathrm{T}(v)})\big)$$

which proves the lemma.

□

Theorem A.1. *Given joint distribution $p(\mathbf{x})$ over d RVs $\{X_i\}_{i=1}^d$, consider $\mathcal{P}_T$ to be the family of all tree-dependent distributions. Then the optimal tree-dependent approximation defined as:*

$$p_T^*(\mathbf{x}) = \arg\min_{p_T \in \mathcal{P}_T} D_{KL}\left(p(\mathbf{x}) \,\|\, p_T(\mathbf{x})\right)$$

is obtained by the maximum-weight spanning tree, where each edge (v, u) has weight:

$$w_{v,u} = \mathbb{I}(X_v; X_u)$$

where $\mathbb{I}(\cdot;\cdot)$ is the Mutual Information between the RVs X_v and X_u under the true distribution $p(\mathbf{x})$.

Proof.

$$
\begin{aligned}
D_{KL}\left(p(\mathbf{x}) \,\|\, p_T(\mathbf{x})\right) &= \mathbb{E}_{p(\mathbf{x})}\left[\log \frac{p(\mathbf{x})}{p_T(\mathbf{x})}\right] \\
&= \mathbb{E}_{p(\mathbf{x})}\left[\log \frac{p(\mathbf{x})}{\prod_{v\in V}\left[p_T(x_v)\cdot c_T\left(P_T(x_v), P_T(x_{T(u)})\right)\right]}\right] \\
&= \mathbb{E}_{p(\mathbf{x})}\left[\log p(\mathbf{x})\right] \\
&\quad - \mathbb{E}_{p(\mathbf{x})}\left[\log \prod_{v\in V}\left[p_T(x_v)\cdot c_T\left(P_T(x_v), P_T(x_{T(u)})\right)\right]\right]
\end{aligned}
$$

Now since we looking for the optimal tree-dependent approximation, we can apply Lemma A.2 and get

$$
= -\mathbb{E}_{p(\mathbf{x})}\left[\log \prod_{v\in V}\left[p(x_v)\cdot c\left(P(x_v), P(x_{T(u)})\right)\right]\right] + C
$$

Let's focus on the non-constant element. We can use log rules to split it to marginals part and copula part and get:

$$
\begin{aligned}
&= -\mathbb{E}_{p(\mathbf{x})}\left[\log \prod_{v\in V}\left[p(x_v)\right]\right] \\
&\quad - \mathbb{E}_{p(\mathbf{x})}\left[\log \prod_{v\in V}\left[c\left(P(x_v), P(x_{T(v)})\right)\right]\right] \\
&\quad + C
\end{aligned}
$$

The first term can be expand to the sum of the entropy of the marginals $\sum_{v\in V}\mathbb{H}_p\left(X_v\right)$ which is constant in terms of T. Hence we will focus on the second term

$$
\begin{aligned}
&= -\mathbb{E}_{p(\mathbf{x})}\left[\log \prod_{v\in V}\left[c\left(P(x_v), P(x_{T(v)})\right)\right]\right] + C \\
&= -\mathbb{E}_{p(\mathbf{x})}\left[\sum_{v\in V}\log c\left(P(x_v), P(x_{T(v)})\right)\right] + C \\
&= -\sum_{v\in V}\mathbb{E}_{p(\mathbf{x})}\left[\log c\left(P(x_v), P(x_{T(v)})\right)\right] + C
\end{aligned}
$$

using marginalization

$$
= \sum_{v\in V} -\mathbb{E}_{p(x_v, x_{T(v)})}\left[\log c\left(P(x_v), P(x_{T(v)})\right)\right] + C
$$

This is exactly the copula entropy (Section 2.2), hence:

$$
\arg\min_T D_{KL}\left(p(\mathbf{x}) \,\|\, p_T(\mathbf{x})\right) = \arg\min_T \sum_{v\in V}\mathbb{H}_c\left(X_v, X_T(v)\right)
$$

where $\mathbb{H}_c$ is the copula entropy. Given the relation between Copula Entropy and Mutual Information Ma and Sun [2011] we can say that

$$
\arg\min_T D_{KL}\left(p(\mathbf{x}) \,\|\, p_T(\mathbf{x})\right) = \arg\min_T \sum_{v\in V} -\mathbb{I}(X_v; X_T(v))
$$

This recovers the original result of: minimizing the KL divergence is equivalent to maximizing the sum of pairwise mutual information. □

B COPULA-TREE MI

A key advantage of tree copulas is the decomposition of the joint density into pairwise terms. If the mutual information (MI) for the bivariate copula family is analytically tractable, the MI of the entire tree can be computed in closed form. First we will prove a general property for an expectation of copula.

Lemma B.1. *Consider the element-wise change of variables $u_i = F(x_i)$. Since each u_i depends solely on x_i, the resulting Jacobian matrix is diagonal with entries corresponding to the marginal densities $J_{ii} = f(x_i)$. Consequently, the differential volume element transforms as the product of the marginals:*

$$d\mathbf{u} = \left|J\right| d\mathbf{x} = \left(\prod_{i=1}^{d} f(x_i)\right) d\mathbf{x}$$

This relationship holds regardless of the dependence structure within the joint distribution $f(\mathbf{x})$.

Hence, for any expectation with respect to $f(\mathbf{x})$ of a function that depends only on $\mathbf{u}$, we can write:

$$\begin{aligned}
\mathbb{E}_{f(\mathbf{x})}\left[f(\mathbf{u})\right] &= \int f(\mathbf{u}) \cdot \prod_{i=1}^{d} \left[f(x_i)\right] \cdot c(F(x_1), \ldots, F(x_d)) d\mathbf{x} \\
&= \int f(\mathbf{u}) \cdot c(\mathbf{u}) d\mathbf{u} \\
&= \mathbb{E}_{c(\mathbf{u})}\left[f(\mathbf{u})\right]
\end{aligned}$$

$$\begin{aligned}
\mathbb{I}\left(X_1; \ldots; X_d\right) &= \mathbb{E}_{p_T(\mathbf{x})}\left[\log \frac{p_T(\mathbf{x})}{\prod_{i=1}^{d} p(x_i)}\right] \\
&= \mathbb{E}_{p_T(\mathbf{x})}\left[\log \frac{\prod_{i=1}^{d} p(x_i) \cdot c_T(P(x_i), P(x_{T(i)}))}{\prod_{i=1}^{d} p(x_i)}\right] \\
&= \mathbb{E}_{p_T(\mathbf{x})}\left[\log \prod_{i=1}^{d} c_T\left(P(x_i), P(x_{T(i)})\right)\right] \\
&= \mathbb{E}_{p_T(\mathbf{x})}\left[\sum_{i=1}^{d} \log c_T\left(P(x_i), P(x_{T(i)})\right)\right] \\
&= \sum_{i=1}^{d} \mathbb{E}_{p_T(\mathbf{x})}\left[\log c_T\left(P(x_i), P(x_{T(i)})\right)\right]
\end{aligned}$$

Now, using marginalization, we obtain

$$\begin{aligned}
&= \sum_{i=1}^{d} \mathbb{E}_{p_T(x_i, x_{T(i)})}\left[\log c_T\left(P(x_i), P(x_{T(i)})\right)\right] \\
&= \sum_{i=1}^{d} -\mathbb{H}_c\left(X_i, X_{T(i)}\right) \quad \text{(by Lemma B.1)} \\
&= \sum_{i=1}^{d} \mathbb{I}(X_i; X_{T(i)})
\end{aligned}$$

Overall we can conclude that calculating the MI for the entire tree is summing the MI pairs of the relevant edges

$$\mathbb{I}\left(X_1; \ldots; X_d\right) = \sum_{i=1}^{d} \mathbb{I}(X_i; X_{T(i)}) \tag{9}$$

For example in the Gaussian copula case with the correlation parameter $\rho_{i,j}$ we get:

$$\mathbb{H}_c\left(X_i, X_j\right) = 0.5 \cdot \log(1 - \rho_{i,j}^2)$$

And for the entire tree T the MI is just

$$\mathbb{I}\left(X_1; \ldots; X_d\right) = \sum_{i=1}^{d} -0.5 \cdot \log(1 - \rho_{i,T(i)}^2)$$

When this analytical form is available, we can replace the Monte Carlo estimate of the variational mutual information term in the ELBO (the first term of the decomposition in Equation (6)) with this closed-form expression.

C ELBO KL ANALYSIS SECOND SCENARIO DERIVATION

$$\begin{aligned}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{c_\pi\left(\{Q_\phi(z_i|\mathbf{x})\}_{i=1}^d\right)}{c_\lambda\left(\{P_\lambda(z_i)\}_{i=1}^d\right)}\right] &= -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log c_\lambda\left(\{P_\lambda(z_i)\}_{i=1}^d\right)\right] \\ &= -\mathbb{E}_{c_\pi(\mathbf{u})}\left[\log c_\lambda\left(\left\{P_\lambda\left(Q_\phi^{-1}(u_i|\mathbf{x})\right)\right\}_{i=1}^d\right)\right] \quad \text{(by Lemma B.1)}\end{aligned}$$

D SOFT-MWST AS COPULA TREE APPROXIMATOR

While the Maximum Weight Spanning Tree (MWST) provides the variational optimal tree structure, the discrete selection process is non-differentiable and can lead to training instability; small changes in edge weights may cause the optimal tree topology to switch abruptly, hindering the decoder's ability to learn a stable representation. To address this, we employ Kirchhoff Marginals (KM) and Tree-Averaged Copulas (AoT) Kirshner [2007] during training. As discussed in Section 2.3, KM produces a matrix of edge probabilities $\beta = \text{KM}(\eta, \tau)$ based on edge scores η and temperature τ. These marginals represent a "Soft-MWST" distribution over the space of spanning trees Paulus et al. [2020], Meilă and Jaakkola [2006]. As $\tau \to 0^+$, the distribution concentrates mass on the Maximum Weight Spanning Tree, and the marginals β_{uv} converge to 1 if the edge (u, v) is in the MWST and 0 otherwise. During training, we define the tree copula density as an AoT copula weighted by these marginals:

$$c_T(\mathbf{u}) = \lim_{\tau \to 0^+} c_{AoT}(\mathbf{u}; KM(\eta, \tau))$$

where η is a $\binom{d}{2}$ vector represent the edge weights we used to establish T using some MWST algorithm such as Prim. This formulation allows gradients to propagate to all edges in the graph during the initial phases of training (high τ), gradually focusing on the optimal structure as τ is annealed.

A similar strategy applies to estimating the Mutual Information of the tree during training (Section B). If the pairwise MI can be computed in closed form, we can replace the hard summation over tree edges (Section B) with a weighted sum using the Kirchhoff marginals $\beta = \text{KM}(\eta, \tau)$ as edge weights:

$$\mathbb{I}\left(X_1; \ldots; X_d\right) = \lim_{\tau \to 0^+} \sum_{(i,j) \in \mathcal{E}} \beta_{i,j} \cdot \mathbb{I}(X_i; X_j),$$

where $\mathcal{E}$ is the set of all possible edges in the complete graph of d nodes.

E IMPLEMENTATION DETAILS

This appendix provides dataset constructions, preprocessing, training schedules, and architectural details for the experiments in Section 4.

Algorithm 1: Training with hard variational trees and a soft-MWST surrogate.

Model : Tree-copula posterior $q_\phi(\mathbf{z} \mid \mathbf{x})$, likelihood $p_\theta(\mathbf{x} \mid \mathbf{z})$, and prior $p_\lambda(\mathbf{z})$.
Input : Minibatch $\{\mathbf{x}_b\}_{b=1}^B$, temperature τ, KL weight β
for $b = 1, \ldots, B$ **do**
  Hard variational posterior construction.
  Encode $\mathbf{x}_b$ into marginal parameters and pair-copula parameters $\{\rho_{ij}(\mathbf{x}_b)\}_{i<j}$
  Compute MWST scores $s_{ij}(\mathbf{x}_b)$ from the pair-copula parameters: analytic pairwise MI when available, otherwise a monotone copula-parameter score
  Construct the hard sample-specific variational tree $T_b^\star = \arg\max_T \sum_{(i,j)\in T} s_{ij}(\mathbf{x}_b)$ by MWST
  Define $q_\phi(\mathbf{z} \mid \mathbf{x}_b)$ using the learned marginals, pair-copula parameters, and hard tree $T_b^\star$
  Sample $\mathbf{z}_b \sim q_\phi(\mathbf{z} \mid \mathbf{x}_b)$ and evaluate $\log p_\theta(\mathbf{x}_b \mid \mathbf{z}_b)$
  Soft surrogate for the training objective.
  Compute soft-MWST edge weights $\alpha_{ij}(\mathbf{x}_b; \tau)$ from the temperature-scaled scores $s_{ij}(\mathbf{x}_b)$ using Kirchhoff marginals
  if *analytic pairwise MI is available* **then**
    Use the differentiable soft tree-MI surrogate $\sum_{i<j} \alpha_{ij}(\mathbf{x}_b; \tau) I_{q_\phi}(z_i; z_j \mid \mathbf{x}_b)$
  else
    Evaluate the copula-dependence term through a differentiable soft-AoT surrogate induced by $\alpha_{ij}(\mathbf{x}_b; \tau)$
  Compute the marginal KL term and, when using a copula prior, the prior-copula cross-entropy term
Form the minibatch objective and update ϕ, θ, and, when applicable, λ by maximizing the resulting β-annealed ELBO

E.1 DATASETS AND PREPROCESSING

Correlated dSprites. We form a correlated subset of dSprites (64×64 grayscale) by filtering examples to satisfy approximate equality constraints among three ground-truth generative factors (scale, orientation, horizontal position). After min–max normalizing each factor to $[0, 1]$ over the full dataset, we retain only samples with

$$\left|\tilde{s} - \tilde{o}\right| < t \qquad \text{and} \qquad \left|\tilde{o} - (1 - \widetilde{x}_{pos})\right| < t,$$

using $t = 0.15$. We use a deterministic split by shuffling filtered indices with seed 42 and taking $80\%/10\%/10\%$ train/val/test. This yields 49,920 total examples.

Fashion-MNIST. We use the canonical train/test split and construct a validation set by splitting the training set with seed 42. Pixels are discrete intensities $\mathbf{x}^p \in \{0, \ldots, 255\}^{28\times 28}$; we apply uniform dequantization:

$$\mathbf{u} \sim \mathrm{Unif}([0, 1]^{28\times 28}), \qquad \mathbf{x} \triangleq \frac{\mathbf{x}^p + \mathbf{u}}{256},$$

so $\mathbf{x} \in [0, 1]^{28\times 28}$.

20 Newsgroups (ProdLDA). We represent documents as bag-of-words and preprocess by removing standard English stop-words and filtering terms with document frequency > 0.5 or < 20. We use $K = 20$ latent topics.

E.2 OPTIMIZATION AND ANNEALING SCHEDULES

Unless stated otherwise, models are trained with a β-annealed objective:

$$\mathcal{L}_\beta(\mathbf{x}) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\log p(\mathbf{x} \mid \mathbf{z})\right] - \beta\, D_{KL}\left(q(\mathbf{z} \mid \mathbf{x}) \,\|\, p(\mathbf{z})\right),$$

where β is warmed up to 1. Tree-copula variants additionally anneal the Soft-MWST temperature τ toward near-hard MWST behavior (Section 3.2.1).

Fashion-MNIST models additionally use gradient clipping with norm 0.5.

Prior-learning stability (Gumbel exploration). For prior-learning experiments (Section 4), we inject i.i.d. Gumbel noise into edge scores (the stochastic softmax trick, Section 3.2.1) to increase exploration and reduce early collapse to suboptimal tree topologies.

Table 6: Experiment Hyperparameters (See Dataset Subsections for Schedules; ProdLDA Schedules Are Not Specified Explicitly in the Main Text).

Dataset	d	Epochs	Batch	Adam lr	β warmup	τ schedule
Correlated dSprites	3	50	128	5e-3	10 ep (linear 0→1)	2.0 (first 20%), exp.→0.1
Fashion-MNIST	4	50	128	1e-3	25 ep (linear)	cosine 2.0→0.1
20 Newsgroups (ProdLDA)	20	50	32	1e-3	—	—

E.3 BACKBONE ARCHITECTURES

E.3.1 Correlated dSprites Encoder/Decoder

Encoder. The encoder maps $\mathbf{x} \in [0,1]^{1\times64\times64}$ to $h(\mathbf{x}) \in \mathbb{R}^{256}$ with four conv blocks:

$$\mathrm{Conv}(4{\times}4,\ \mathrm{stride}=2,\ \mathrm{pad}=1)\ \rightarrow\ \mathrm{BN}\ \rightarrow\ \mathrm{LeakyReLU}(0.2),$$

with channels $1\rightarrow32\rightarrow64\rightarrow128\rightarrow256$. We flatten the $256\times4\times4$ map (4096 dims) and apply a linear layer to $H=256$. The variational head outputs $(\mu, \log\sigma) \in \mathbb{R}^d$ for `MF`; for `TC_GC`/`TC_ST` it additionally outputs $\binom{d}{2}$ pair-copula parameters per datapoint via a linear projection followed by a scaled $\tanh$.

Decoder. We map $\mathbf{z} \in \mathbb{R}^d$ through a linear layer to 4096, reshape to $256\times4\times4$, then apply four transpose-conv blocks ($4{\times}4$, stride 2, pad 1) with BN and LeakyReLU(0.2): $256\rightarrow128\rightarrow64\rightarrow32\rightarrow32$, producing a $32\times64\times64$ feature map.

Likelihood heads. For `FG`, a $1{\times}1$ conv outputs the mean $\mu_\theta(\mathbf{z})$; the variance is a global learned scalar. For `R1GC_G`, a second $1{\times}1$ conv outputs rank-1 factors $W_\theta(\mathbf{z}) \in \mathbb{R}^{1\times64\times64}$; the diagonal variance component remains a global learned scalar.

E.3.2 Fashion-MNIST Encoder/Decoder Backbone

Encoder backbone. The encoder maps $\mathbf{x} \in [0,1]^{1\times28\times28}$ to a feature vector $h(\mathbf{x}) \in \mathbb{R}^H$ using the following stack:

$$\begin{aligned}&\mathrm{Conv}(4{\times}4,\ \mathrm{s}=2,\ \mathrm{p}=1,\ 32) \rightarrow \mathrm{BN} \rightarrow \mathrm{LeakyReLU}\\ &\quad\rightarrow \mathrm{Conv}(4{\times}4,\ \mathrm{s}=2,\ \mathrm{p}=1,\ 64) \rightarrow \mathrm{BN} \rightarrow \mathrm{LeakyReLU}\\ &\quad\rightarrow \mathrm{Flatten}\ (64\cdot7\cdot7=3136) \rightarrow \mathrm{Linear}(3136\rightarrow H) \rightarrow \mathrm{LeakyReLU}.\end{aligned}$$

A variational-family head then maps $h(\mathbf{x})$ to the parameters of $q_\phi(\mathbf{z}\mid\mathbf{x})$: for `MF` it outputs $\{(\mu_i(\mathbf{x}), \sigma_i(\mathbf{x}))\}_{i=1}^d$, while for `TC_GC`/`TC_ST` it additionally outputs the $\binom{d}{2}$ candidate pair-copula parameters per datapoint, which define the edge scores used for (soft) tree selection.

Decoder backbone. The decoder maps $\mathbf{z} \in \mathbb{R}^d$ to a spatial feature map via:

$$\begin{aligned}&\mathrm{Linear}(d\rightarrow64\cdot7\cdot7) \text{ and reshape to } 64\times7\times7\\ &\quad\rightarrow \mathrm{Deconv}(4{\times}4,\ \mathrm{s}=2,\ \mathrm{p}=1,\ 32) \rightarrow \mathrm{BN} \rightarrow \mathrm{LeakyReLU}\\ &\quad\rightarrow \mathrm{Deconv}(4{\times}4,\ \mathrm{s}=2,\ \mathrm{p}=1,\ 32) \rightarrow \mathrm{BN} \rightarrow \mathrm{LeakyReLU},\end{aligned}$$

yielding a $32\times28\times28$ feature map that is passed to the likelihood head. Under `R1GC_G`, the head additionally outputs the rank-1 dependence factor needed to induce pixel dependencies through a Gaussian copula.

E.3.3 ProdLDA Architecture (20 Newsgroups)

The ProdLDA model [Srivastava and Sutton, 2017] approximates the Dirichlet prior of LDA using a *Logistic Normal* distribution. For each document, latent log-proportions are sampled as $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and mapped to the simplex via $\boldsymbol{\theta} = \mathrm{softmax}(\mathbf{z})$. Document word counts are modeled with a multinomial likelihood parameterized by a linear decoder.

Encoder. We use a feed-forward encoder with two hidden layers of 100 units each, with **Softplus** activations and **Batch Normalization (with `affine=False`)** to prevent component collapse. We apply **Dropout** with rate 0.2 after the activation functions. The encoder maps the input bag-of-words vector to the variational parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (or to the copula dependence parameters for tree-copula variants).

Decoder. The decoder is a single linear layer mapping $\boldsymbol{\theta} = \mathrm{softmax}(\mathbf{z})$ to vocabulary-size logits, followed by Batch Normalization and a Softmax activation to model word probabilities.

Distributional components and optimization. For the `VAMP` prior, we use $K = 10$ learnable pseudo-inputs initialized from a standard normal distribution. We train using Adam with learning rate 10^{-3}, batch size 32, for 50 epochs.

E.4 TREE EXTRACTION AND TOPOLOGY METRICS

Edge weights and MWST extraction. For each datapoint, we build a complete graph on latent dimensions with edge weights given by pairwise mutual information implied by the predicted bivariate copulas (Section 3.2.1). In the Gaussian-copula case, if the edge parameter is a correlation ρ_{ij}, the closed-form mutual information is

$$I_{ij} \;=\; -\tfrac{1}{2}\log\!\big(1-\rho_{ij}^2\big),$$

which we use as the Chow–Liu weight for MWST extraction.

Prüfer encoding. We encode each extracted labeled tree T by its Prüfer sequence (length $d-2$), yielding d^{d-2} distinct labeled trees. For $d = 4$, there are 16 possible topologies, which we treat as discrete assignments.

NMI and weighted purity. Let T denote the discrete topology assignment and Y the class label. We report

$$\mathrm{NMI}(T, Y) \triangleq \frac{2\, I(T;Y)}{H(T) + H(Y)}$$

and weighted purity

$$\mathrm{Purity} \triangleq \sum_t \frac{n_t}{N} \max_y \frac{n_{t,y}}{n_t},$$

where n_t is the number of samples assigned topology t, $n_{t,y}$ counts samples with topology t and label y, and N is the dataset size. For conditional-topology heatmaps, we visualize $\widehat{P}(T = t \mid Y = y) = \frac{n_{t,y}}{\sum_{t'} n_{t',y}}$ and sort topology columns by dominant label for readability.

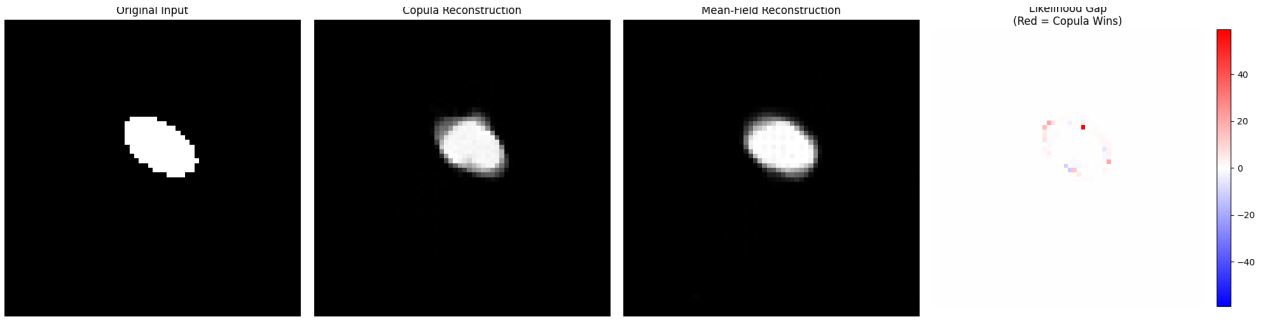
E.5 IWAE ESTIMATION AND PIXELWISE ATTRIBUTION

IWAE@512. We estimate held-out density with the IWAE bound using $K = 512$ importance samples per datapoint. For numerical stability we compute the log-mean-exp form of the importance weights and average over the test set.

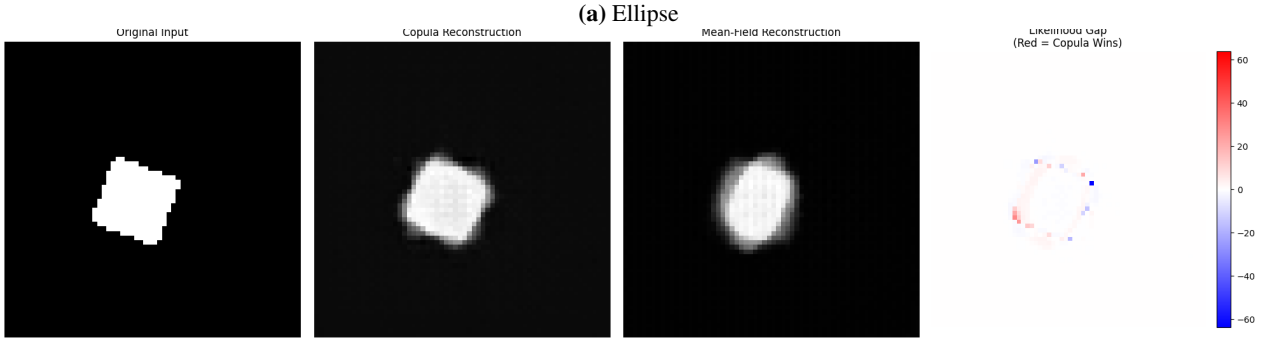
Per-pixel marginal NLL gap under `R1GC_G`. For the attribution maps in Figure 2, we evaluate marginal pixel likelihoods at a deterministic latent code $\hat{\mathbf{z}} = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\mathbf{z}]$. Under the rank-1 Gaussian-copula decoder (`R1GC_G`), the induced observation model is a multivariate Gaussian with covariance $\Sigma(\hat{\mathbf{z}}) = \sigma^2 I + \boldsymbol{w}\boldsymbol{w}^\top$, so pixel j has marginal variance $\sigma^2 + w_j^2$. We compute

$$\Delta_i(j) = \mathrm{NLL}^{\mathrm{marg}}_{\mathrm{MF}}(x_i(j) \mid \hat{\mathbf{z}}_i^{\mathrm{MF}}) - \mathrm{NLL}^{\mathrm{marg}}_{\mathrm{TC}}(x_i(j) \mid \hat{\mathbf{z}}_i^{\mathrm{TC}}),$$

where $\Delta_i(j) > 0$ indicates higher marginal likelihood under tree-copula inference, and emphasize that $\sum_j \Delta_i(j)$ does not equal the IWAE improvement because the joint likelihood includes dependence terms.



(a) Ellipse



(b) Square

Figure 4: Correlated dSprites: additional per-example marginal NLL gap maps under the same `R1GC_G` decoder. Each image shows (left to right): input, `TC_ST` reconstruction, `MF` reconstruction, and the per-pixel marginal NLL gap $\Delta_i(j) = \mathrm{NLL}_{\mathrm{MF}}^{\mathrm{marg}}(x_i(j)) - \mathrm{NLL}_{\mathrm{TC}}^{\mathrm{marg}}(x_i(j))$. Warm colors ($\Delta_i > 0$) indicate pixels where tree-copula inference assigns higher marginal likelihood.

E.6 FASHION-MNIST RESULTS UNDER `FG`

For completeness, we report Fashion-MNIST results under the factorized Gaussian likelihood (Table 7) using the same backbone, latent dimensionality (d=4), and training protocol as in Section 4.2.

Table 7: Fashion-MNIST: Test IWAE@512 and ELBO Diagnostics Under the Factorized Gaussian Likelihood (`FG`).

Model	**IWAE@512**↑	$\log p_\theta(\mathbf{x} \mid \mathbf{z})$↑	**KL**↓
`SN-MF-FG`	453.10	457.44	15.20
`SN-TC_GC-FG`	455.76	459.51	**15.19**
`SN-TC_ST-FG`	**457.48**	**462.46**	15.31

E.7 ADDITIONAL QUALITATIVE RESULTS

Correlated dSprites: additional marginal NLL gap maps. Figure 4 provide additional shapes for the per-pixel marginal NLL gap analysis described in Section E.5.

Fashion-MNIST: additional topology heatmap. Figure 5 shows the conditional topology distribution for `SN-TC_ST-R1GC_G`, complementing Figure 3 in the main text.

F PRIOR MISMATCH SANITY CHECK ON 20 NEWSGROUPS

We report a supplementary prior-mismatch ablation on 20 Newsgroups [Lang, 1995] with a ProdLDA decoder [Srivastava and Sutton, 2017] (K=20 topics), comparing priors under mean-field vs. tree-copula inference (Table 8).

Table 8: 20 Newsgroups (ProdLDA): Test IWAE@512.

Model	**IWAE@512** ↑
SN-MF-ProdLDA	−580.24
SN-TC_GC-ProdLDA	−581.39
VAMP-MF-ProdLDA	−627.23
VAMP-TC_GC-ProdLDA	−599.87
AoT_SG_GC-TC_GC-ProdLDA	**−580.22**

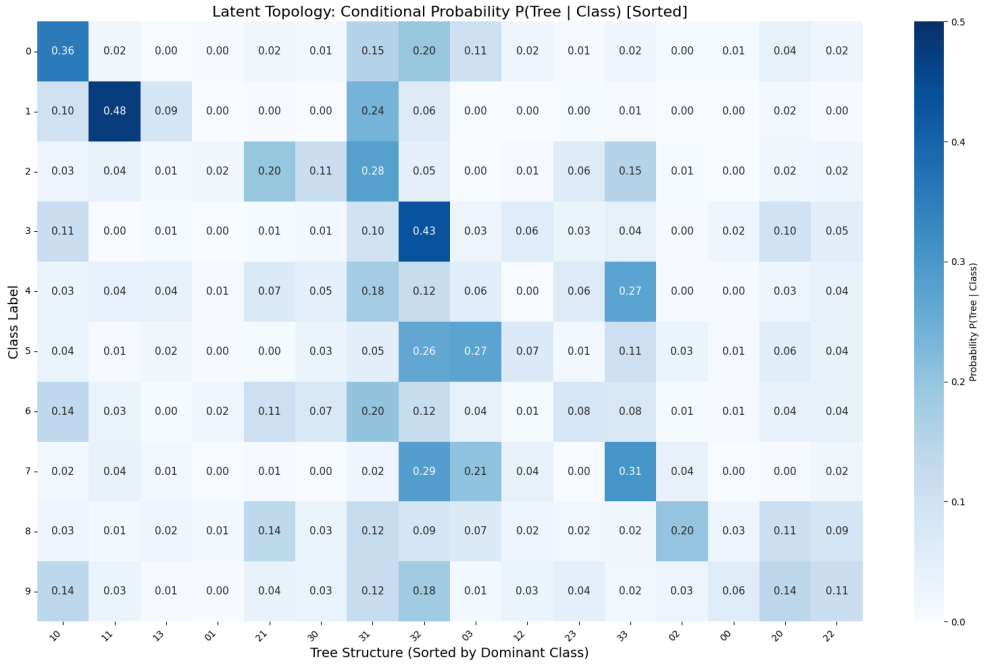


Figure 5: Fashion-MNIST latent-topology conditional frequencies for $d = 4$ under the dependent likelihood `R1GC_G`. Each panel shows the per-class normalized empirical conditional distribution $\widehat{P}(T \mid Y)$, where T is the MWST topology extracted from the variational model (16 Prüfer-encoded columns) and Y is the class label (10 rows). Darker cells indicate that a larger fraction of examples within that class select the corresponding topology (each row sums to 1); columns are sorted by dominant class independently per panel for readability.