

On the Joint Minimization of Regularization Loss Functions in Deep Variational Bayesian Methods for Attribute-Controlled Symbolic Music Generation

Matteo Pettenò
Politecnico di Milano
Milan, Italy
matteo.petteno@mail.polimi.it

Alessandro Ilic Mezza
Politecnico di Milano
Milan, Italy
alessandroilic.mezza@polimi.it

Alberto Bernardini
Politecnico di Milano
Milan, Italy
alberto.bernardini@polimi.it

Abstract—Explicit latent variable models provide a flexible yet powerful framework for data synthesis, enabling controlled manipulation of generative factors. With latent variables drawn from a tractable probability density function that can be further constrained, these models enable continuous and semantically rich exploration of the output space by navigating their latent spaces. Structured latent representations are typically obtained through the joint minimization of regularization loss functions. In variational information bottleneck models, reconstruction loss and Kullback-Leibler Divergence (KLD) are often linearly combined with an auxiliary Attribute-Regularization (AR) loss. However, balancing KLD and AR turns out to be a very delicate matter. When KLD dominates over AR, generative models tend to lack controllability; when AR dominates over KLD, the stochastic encoder is encouraged to violate the standard normal prior. We explore this trade-off in the context of symbolic music generation with explicit control over continuous musical attributes. We show that existing approaches struggle to jointly minimize both regularization objectives, whereas suitable attribute transformations can help achieve both controllability and regularization of the target latent dimensions.

Index Terms—symbolic music generation, attribute-controlled generation, latent space regularization, power transforms

I. INTRODUCTION

Symbolic music generation has witnessed remarkable advancements in recent years with the rise of deep latent variable models [1]–[5], yet the challenge of controlling high-level musical attributes at inference time remains an active area of research. Latent variable models can be broadly categorized into *explicit* density models [6], such as Variational Autoencoders (VAEs), Variational Information Bottleneck (VIB) models, normalizing flows, and diffusion models, and *implicit* density models, notably Generative Adversarial Networks (GANs).

Explicit density models offer the advantage of tractable likelihood estimation and a well-structured latent space. In particular, deep variational Bayesian methods [7] learn smooth and continuous representations of musical data [1], [2], leading to meaningful interpolation and targeted manipulation of specific characteristics of the output [8]. Moreover, VAEs [9] and VIB models [10] are able to learn disentangled representations, i.e., having independent and interpretable latent factors, which, in turn, allow individual musical attributes to be adjusted with-

out unintentionally altering others. This is often accomplished through supervised multi-task learning techniques [11]–[15], which structure the latent space by ensuring that specific attributes are encoded so as to be proportional to changes along designated dimensions. These approaches entail introducing auxiliary attribute-regularization objectives alongside the reconstruction loss, ensuring data fidelity, and Kullback-Leibler Divergence (KLD), which constrains the latent space to be continuous and evenly dense, in addition to enabling efficient and reliable sampling from a predefined prior distribution.

To the best of our knowledge, however, the trade-off between these regularization losses has not been thoroughly explored in the literature, with training recipes being often provided without an accompanying analysis of the ensuing behavior of the regularized dimensions. In this work, by focusing on high-level musical attributes such as Contour, Pitch Range, and Rhythm Complexity, we show that existing attribute-controlled symbolic music generation models are highly sensitive to hyperparameter tuning and tend to fail in satisfying both objectives at the same time. Furthermore, we show that invertible attribute mappings based on power transforms may help mitigate the issue.

II. RELATED WORK

Let X , Y , Z be random variables. The Information Bottleneck (IB) [16] optimization problem is given by [10] $\max_{\theta} I_{\theta}(Z, Y) - \beta I_{\theta}(Z, X)$, where $I(\cdot, \cdot)$ denotes the mutual information. The IB framework entails finding $p_{\theta}(z|x)$ parameterized by θ that defines a compressed representation of X , i.e., the latent variable Z , which retains as much information as possible about Y while discarding irrelevant information about X . This trade-off is controlled by the Lagrange multiplier $\beta \geq 0$.

The VIB approach [10] finds a variational lower bound on the IB objective, resulting in the minimization of the following loss function:

$$\mathcal{L}_{\text{VIB}} = -\mathbb{E}_{p_{\theta}(z|x)} [\log q_{\phi}(y|z)] + \beta D_{\text{KL}} [p_{\theta}(z|x) \parallel r(z)], \quad (1)$$

where $p_{\theta}(z|x)$ is a stochastic encoder, x is the encoder input, $q_{\phi}(y|z)$ is a variational approximation of $p(y|z)$ parameterized

by the weights ϕ of a decoder network, y is the decoder output, $r(z)$ is the variational approximation of the marginal distribution $p(z)$, and $D_{\text{KL}}[\cdot\|\cdot]$ denotes the KLD. If $Y = X$, (1) coincides with the classic β -VAE objective [9]. In most practical cases, $p_\theta(z|x)$ and $r(z)$ are chosen to be multivariate normal distributions, namely, $p_\theta(z|x) \sim \mathcal{N}(\mu_{z|x}, \Sigma_{z|x})$ and $r(z) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

With little loss of generality, we will focus here on the one-dimensional continuous case where an *attribute* is defined as a scalar $a \in \mathcal{A}$ that can be algorithmically computed from an instance x , i.e., there exists $f : x \mapsto f(x)$ such that $f(x) = a$, with $\mathcal{A} \subseteq \mathbb{R}$ the image of f .

To develop *attribute-controlled* generative models, [11], [12] propose to encode a onto the i -th dimension of the latent space by adding a supervised Attribute Regularization (AR) term to (1), i.e.,

$$\mathcal{L}_{\text{AR-VIB}} = \mathcal{L}_{\text{VIB}} + \gamma \mathcal{L}_{\text{AR}}, \quad (2)$$

where $\gamma \geq 0$ is a tunable hyperparameter.

In [12], Mezza et al. define such a regularization term as

$$\mathcal{L}_{\text{AR}} = \text{MAE}(z_i, \tilde{a}), \quad (3)$$

where MAE denotes the mean absolute error, and \tilde{a} is the z -score of the attribute a . This simple pointwise error approach proves effective as long as a is normally distributed [12]. If not, the regularization losses in (2) turn out to have contrasting objectives. If \mathcal{L}_{AR} dominates over the KLD, the learned posterior $p_\theta(z_i|x)$ tends to naively approximate the attribute's sample distribution $p(a)$ and disregard the target prior $r(z_i)$, with the covariance $\Sigma_{z|x}$ collapsing toward a zero matrix. Conversely, if the KLD dominates over \mathcal{L}_{AR} , then $p_\theta(z_i|x)$ follows $r(z_i)$ at the cost of poor attribute regularization.

In [11], Pati and Lerch propose a novel term that relaxes the regularization constraint by enforcing a monotonic relationship between a and z_i , i.e.,

$$\mathcal{L}_{\text{AR}} = \text{MAE}(\tanh(\delta \mathbf{D}_z), \text{sign}(\mathbf{D}_a)), \quad (4)$$

where \mathbf{D}_z and \mathbf{D}_a are the pairwise distance matrices between z_i and a of all samples in a given mini-batch, respectively, and $\delta > 0$ is a hyperparameter controlling the spread of the posterior distribution. Since enforcing an arbitrary monotonic relationship may still result in a distribution diverging from the chosen prior, though, using (4) requires careful balancing of β and γ in (2) in order to learn θ such that $p_\theta(z_i|x)$ approximates $r(z_i)$. In fact, the posterior is encouraged to be Gaussian only as long as the KLD is minimized, which, in turn, may negatively affect attribute control.

III. POWER TRANSFORM-BASED REGULARIZATION

In the previous section, we outlined practical problems arising from the minimization of (2) when having two conflicting terms in \mathcal{L}_{AR} and $D_{\text{KL}}[p_\theta(z|x) \| r(z)]$. In fact, when the prior and the attribute distribution are too dissimilar, we will later show that $p_\theta(z_i|x)$ fails to jointly model $p(a)$ and $r(z_i)$. However, we argue that, the opposite being true, a simple distance measure may prove sufficient as a regularizer. We can achieve

this by means of a parametric isomorphism $T_\lambda : \mathcal{A} \rightarrow \mathbb{R}$ such that $p(T_\lambda(a))$ is (at least approximately) distributed like $r(z_i)$. Hence, we define the following regularization loss:

$$\mathcal{L}_{\text{AR}} = \text{MAE}(z_i, T_\lambda(a)). \quad (5)$$

By requiring T_λ to be invertible, we can seamlessly go back and forth between the latent space and the input space; this enables attribute manipulation within the original, human-interpretable domain, while ensuring the consistency of the component-wise marginals of the posterior.

In case $r(z_i) \sim \mathcal{N}(0, 1)$, T_λ has to transform data so as to make them more normal-like. We refer to this process as *data Gaussianization*. Gaussianization can be implemented using both parametric [17], [18] and non-parametric methods [19], [20]. In the one-dimensional case, a simple approach is to use a family of functions that define monotonic data transformations, known in statistics as Power Transforms (PT).

First introduced in [17], one of the most widely used PT is the Box-Cox transformation, whose two-parameter formulation is

$$g_{\lambda_1, \lambda_2}(u) = \begin{cases} \frac{(u + \lambda_2)^{\lambda_1} - 1}{\lambda_1} & \text{if } \lambda_1 \neq 0, \\ \ln(u + \lambda_2) & \text{if } \lambda_1 = 0, \end{cases} \quad (6)$$

where the shift parameter $\lambda_2 \geq 0$ extends the domain of g from $\{u \in \mathbb{R} \mid u > 0\}$ to $\{u \in \mathbb{R} \mid u > -\lambda_2\}$.¹

The proposed transformation of the attribute distribution applies (6) followed by a Batch Normalization layer with scale and shift parameters set to one and zero, respectively, i.e.,

$$\text{BN}_{1,0}(u) = \frac{u - \mu}{\sqrt{\sigma^2 + \epsilon}}, \quad (7)$$

where ϵ is a small constant, and μ and σ correspond to the mean and standard deviation of the current mini-batch during training, and their moving average estimates across all batches at inference time. Thus, the whole transformation is given by the composition

$$T_\lambda(a) = (\text{BN}_{1,0} \circ g_{\lambda_1, \lambda_2})(a). \quad (8)$$

The transformation parameters $\lambda = \{\lambda_1, \lambda_2\}$ are determined using training data prior to the training phase, rather than being learned alongside the VIB parameters. Specifically, the shift parameter λ_2 is found through a grid search, and for each value in the grid, the power parameter λ_1 is estimated by minimizing the negative log-likelihood using Brent's method [21]. The optimal pair (λ_1, λ_2) is selected by picking the parameters for which the transformed distribution has the lowest negentropy, which can be approximated by $J(\tilde{v}) \approx (\mathbb{E}[\psi(\tilde{v})] - \mathbb{E}[\psi(\nu)])^2$ [22], where \tilde{v} is a standardized instance of a random variable V , $\nu \sim \mathcal{N}(0, 1)$, and ψ is a nonquadratic function that we chose to be $\psi(u) = -\exp(-u^2/2)$ [22]. Once computed, $\{\lambda_1, \lambda_2\}$ is kept fixed during both training and inference, so applying (8) incurs no additional cost.

¹In this study, we only consider attributes that, by construction, are bounded from below. If not, the Box-Cox transformation could be replaced, e.g., by the Yeo-Johnson transformation [18], which is defined on all of \mathbb{R} .

Looking back at (8), we may notice that T_λ is the composition of differentiable, monotone, and invertible functions. As such, we may interpret T_λ^{-1} as a special case of one-dimensional normalizing flows where a nonnormal distribution $p(a)$ is obtained by letting the simple distribution $r(z_i)$ flow through the diffeomorphism. While we do not explicitly require $T_\lambda(a)$ to be differentiable, using diffeomorphic attribute transformations would ultimately allow to learn λ via stochastic gradient descent. We leave this study for future work.

IV. EVALUATION

A. Dataset

We train attribute-controlled models to learn pitch sequence representations of four-bar monophonic melodies. We create a large-scale dataset of monophonic melodies using 176,581 MIDI files from the Lakh MIDI Dataset [23]. First, we check whether the original piece contains changes in its time signature. If so, we split the file where these variations occur and keep only the parts with a 4/4 time signature. Next, we quantize MIDI events to the nearest sixteenth note. We then define a melody as a sequence of notes within the pitch range of a canonical 88-key piano and performed by an instrument associated to a valid MIDI program. Considering a melody ended if one bar of silence is found, we extract only melodies that are at least four bars long and contain at least three different pitches. If more than one note is played simultaneously, we follow [1] and obtain a monophonic sequence by keeping only the note with the highest pitch. Finally, we extract four-bar melodies with a stride of one bar. For each melody thus obtained, we compute 13 musical attributes, including those discussed in Section IV-B. Melodies are encoded as sequences of 64 integers in the range $[0, 129]$, with each step corresponding to the MIDI note number if an onset is present or one of two additional tokens for *note off* (128) and *note hold* (129) events. The corpus is then split into training, validation, and test sets, with the training data augmented through transposition by a random number of semitones in the range of ± 1 octaves. The resulting dataset contains about 10 million unique melodies.²

B. Musical Attributes

In this work, we focus on three of the 13 musical attributes included in our dataset: (i) **Contour**, which captures the extent of melodic movement by averaging the pitch differences between consecutive notes; (ii) **Rhythm Complexity**, evaluated using Toussaint’s metrical complexity measure [24]; (iii) **Pitch Range**, determined by the difference between the highest and lowest MIDI pitch values in the sequence.

C. Evaluation Metrics

We evaluate each attribute-controlled model on two aspects: the degree of regularization of the latent dimension z_i onto which the attribute a is encoded and the ability of controlling the output attributes at inference time.

²M. Pectenò, Aug. 2024, “4 Bars Monophonic Melodies Dataset (Pitch Sequence),” Zenodo, doi: <https://doi.org/10.5281/zenodo.13369389>

On the one hand, a well-regularized latent dimension means that the univariate posterior $p_\theta(z_i|x)$ is as close to $r(z_i)$ as possible. In practice, we apply kernel density estimation with normal kernels to the empirical posterior, and compare the resulting PDF to the target prior $\mathcal{N}(0, 1)$. We measure the (dis)similarity between these two probability distributions with three metrics: (i) **Maximum Mean Discrepancy (MMD)** with a polynomial kernel [4]; (ii) **Overlapping Area (OA)** as previously done in [25], [26] to compare musical attribute distributions; and the (iii) **Jensen–Shannon Divergence (JSD)**, a symmetrized and bounded KLD-based similarity measure.

On the other hand, controllability is assessed through the **Spearman’s rank correlation coefficient** (ρ_s) between the regularized dimension z_i and the attributes a^* of all 22,000 decoded test sequences. A coefficient ρ_s close to one indicates that there exists a monotonic relationship between z_i and a^* , which, in turn, means that we can generate samples with a desired attribute by navigating the latent space along the i -th dimension.

D. Implementation Details

We implement the VIB models so as to reconstruct the input melodies. The encoder and decoder architectures follow the design proposed in [1]. All models are trained for 40,000 iterations with a batch size of 512. The objective in (2), with cross-entropy as the reconstruction loss, is minimized using Adam. The learning rate is decreased exponentially from 10^{-3} to 10^{-5} with a rate of 0.9999. The hyperparameter β is annealed exponentially from 0 to 10^{-3} , which encourages the model to focus on reconstructing the sequence with high accuracy in the first part of the training [1]. Additionally, we apply teacher forcing within the bottom-level decoder with a probability following a logistic schedule [1]. λ_1 and λ_2 are determined for each attribute as described in Section III.

To study the effect of different design choices in balancing β and γ , respectively KLD and \mathcal{L}_{AR} , we train all models with $\gamma = 10^{-3}$, i.e., equal to the maximum value of β , and $\gamma = 1$, as in [11]. These values are kept fixed throughout the training, as our experiments showed that annealing γ in the same way as β yields equivalent results.

V. RESULTS

In this section, we compare three regularization approaches: the method discussed in Section III, denoted here as “PT”, and those from [12] and [11]. Specifically, with “NM” we refer to the method in [12], given by (3), and with “P&L” we refer to the method in [11], given by (4) with $\delta = 10$ [11].

Overall, we train 18 models, considering three regularization losses (NM, P&L, PT), three musical attributes (Contour, Rhythm Complexity, Pitch Range), and two different values for the weight γ (1 and 10^{-3}). Table I reports the evaluation metrics discussed in Section IV-C for all such models.

When $\gamma = 10^{-3}$, i.e., when the weighting of the AR loss is comparable to that of the KLD, OA approaches the theoretical maximum of 1, while JSD and MMD are low. This indicates that the minimization of the KLD successfully

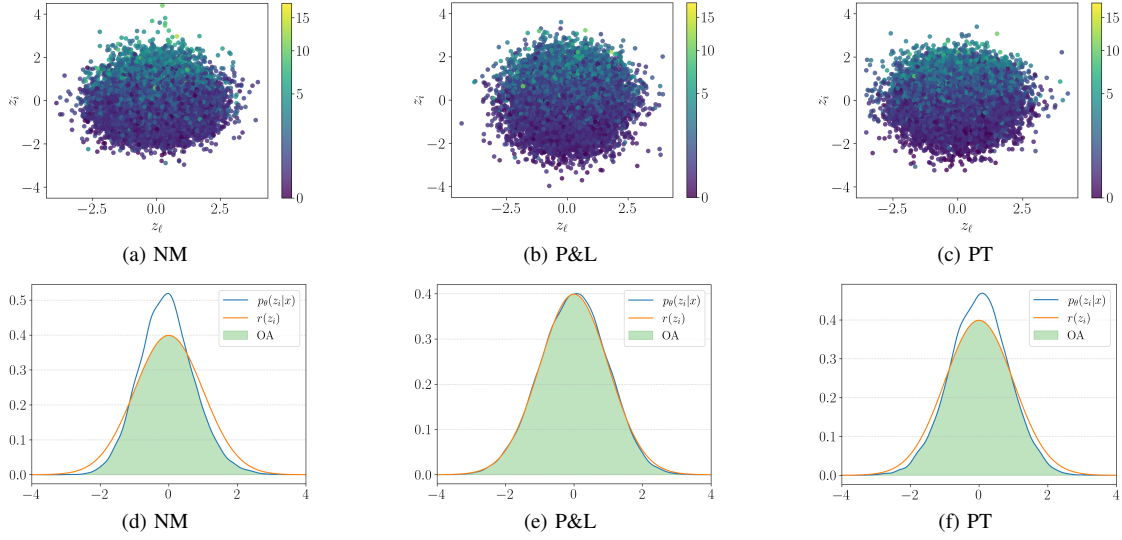


Fig. 1. **Contour** attribute regularization with $\beta \rightarrow 10^{-3}$ and $\gamma = 10^{-3}$

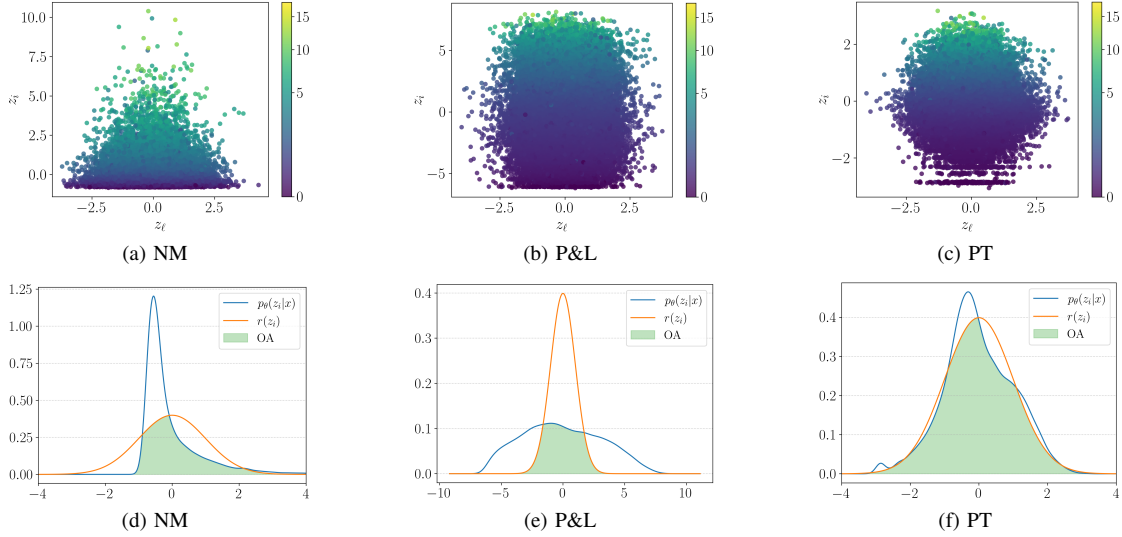


Fig. 2. **Contour** attribute regularization with $\beta \rightarrow 10^{-3}$ and $\gamma = 1$

regularized all latent dimensions, including z_i . Conversely, ρ_s is low, consistently between 0.4 and 0.66, i.e., far from the maximum possible value of 1; this suggests some degree of attribute regularization, yet not enough to ensure fine-grained control over the output attributes in practical applications.

These phenomena are illustrated for Contour³ in Fig. 1. For all 22,000 sequences in the held-out test set, Figs. 1a, 1b, 1c depict z_i against another latent dimension z_l that we choose to be the least correlated with z_i , with brighter colors corresponding to higher a^* values and vice versa. The circular-like distributions of points centered at the origin indicate that, at least approximately, z_i and z_l are jointly Gaussian, as prescribed by the KLD with a normal prior. The lack of a smooth color gradient along z_i , however, indicates poor

AR, corroborating the values of ρ_s ranging from 0.40 to just below 0.54 in Table I. Figs. 1d, 1e, 1f show the comparison between $p_\theta(z_i|x)$ (in blue) and $r(z_i)$ (in orange), as well as the overlap between the two distributions (in green). These plots provide visual evidence of the similarity measured by OA, JSD, and MMD. Fig. 1e also reveals that, with $\gamma = 10^{-3}$, P&L exhibits a better synergy with the KLD compared to NM and PT, possibly because (4) enforces a monotonic rather than pointwise error constraint.

When $\gamma = 1$, instead, AR appears to govern the training dynamics. As such, ρ_s turns out consistently over 0.99, indicating an almost perfect monotonic relationship. On the downside, the regularization metrics for NM and P&L significantly worsen. This can be observed for Contour in Figs. 2a and 2b, where despite a distinct gradient from dark to bright colors, the univariate latent distributions appear to be far from jointly Gaussian. Likewise, Fig. 2d shows that NM encodes

³Here, we focus on the Contour attribute. Additional results, along with audio examples, are available at <https://mpetteno.github.io/box-cox-latent-reg>

TABLE I
SPEARMAN’S RANK CORRELATION COEFFICIENTS AND REGULARIZATION METRICS. (\uparrow) THE HIGHER THE BETTER; (\downarrow) THE LOWER THE BETTER.

	γ	Contour				Rhythm Complexity				Pitch Range			
		ρ_s (\uparrow)	OA (\uparrow)	JSD (\downarrow)	MMD (\downarrow)	ρ_s (\uparrow)	OA (\uparrow)	JSD (\downarrow)	MMD (\downarrow)	ρ_s (\uparrow)	OA (\uparrow)	JSD (\downarrow)	MMD (\downarrow)
NM	10^{-3}	0.4048	0.8803	0.012652	0.3989	0.4547	0.9055	0.007747	0.2752	0.4494	0.8797	0.012855	0.423175
P&L		0.5345	0.9888	0.000366	0.0217	0.6408	0.9537	0.002618	0.1131	0.6646	0.8591	0.017486	0.973104
PT		0.5394	0.9247	0.005627	0.2273	0.4982	0.9133	0.007481	0.3038	0.4028	0.8962	0.009497	0.350608
NM	1.0	0.9994	0.4382	0.237599	256.854	0.9981	0.7946	0.035963	1.9022	0.9982	0.7829	0.040768	1.2870
P&L		0.9999	0.6041	0.121547	6.4854	0.9981	0.6699	0.088264	30.3475	0.9982	0.7871	0.044671	14.4203
PT		0.9998	0.9224	0.004662	0.0085	0.9981	0.8793	0.013633	0.0523	0.9982	0.9474	0.002732	0.0249

an exponential distribution onto z_i , thus drastically reducing the OA, while Fig. 2e reveals that P&L leads to monotonic yet sub-Gaussian attribute encoding.

This means that, when using these regularization losses, it is not sufficient to sample the prior $r(z)$ at inference time, but one should either have prior knowledge of the true attribute distribution or empirically estimate the latent marginal from representative encodings of real-world data before being able to apply meaningful manipulation. In contrast, when it comes to the PT regularization method, OA, JSD, and MMD are comparable if not better with $\gamma = 1$ than with $\gamma = 10^{-3}$, all while maintaining a Spearman’s rank correlation coefficient over 0.99. Therefore, PT-regularized models appear to offer greater flexibility when it comes to hyperparameter tuning, allowing one to prioritize controllability without penalizing regularization, and thus enabling inference by sampling the prior and navigating the latent space as in a regular VIB.

VI. CONCLUSIONS

In this work, we explored the challenges of jointly minimizing regularization loss functions in variational information bottleneck models for symbolic music generation. We examined the trade-off between the Kullback-Leibler divergence and existing attribute-regularization losses, highlighting how an imbalance between these objectives can compromise either controllability or adherence to the prior along the target dimensions. Our findings indicate that existing approaches struggle to balance these constraints, resulting in either poor control over musical attributes or deviations from the desired latent structure. However, experimental results demonstrate that invertible mappings between attributes and latents based on the Box-Cox power transform can alleviate this issue, enabling both improved controllability and robust latent regularization with minimal hyperparameter tuning. Future work includes learning transformation parameters via backpropagation, extending the method to handle multiple attributes, and validating the framework across different signal domains.

REFERENCES

- [1] A. Roberts, J. Engel, C. Raffel, C. Hawthorne, and D. Eck, “A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music,” in *ICML*, 2018, pp. 4364–4373.
- [2] J. Gillick, A. Roberts, J. Engel, D. Eck, and D. Bamman, “Learning to groove with inverse sequence transformations,” in *ICML*, vol. 97, 2019, pp. 2269–2279.
- [3] J. Engel, M. Hoffman, and A. Roberts, “Latent constraints: Learning to generate conditionally from unconditional generative models,” in *ICLR*, 2018.
- [4] G. Mittal, J. Engel, C. Hawthorne, and I. Simon, “Symbolic music generation with diffusion models,” in *ISMIR*, 2021.
- [5] M. Dinculescu, J. Engel, and A. Roberts, “MidiMe: Personalizing a musicvae model with user data,” in *Workshop on Machine Learning for Creativity and Design*, *NeurIPS*, 2019.
- [6] Q. Wu, R. Gao, and H. Zha, “Bridging explicit and implicit deep generative models via neural Stein estimators,” in *NeurIPS*, 2021.
- [7] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” in *ICLR*, 2014.
- [8] A. Roberts, J. Engel, S. Oore, and D. Eck, “Learning latent representations of music to generate interactive musical palettes,” in *IUI Workshops*, 2018.
- [9] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, “beta-vae: Learning basic visual concepts with a constrained variational framework,” in *ICLR*, 2017.
- [10] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, “Deep Variational Information Bottleneck,” in *ICLR*, 2017.
- [11] A. Pati and A. Lerch, “Attribute-based regularization of latent spaces for variational auto-encoders,” *Neural Comput. Appl.*, vol. 33, no. 9, pp. 4429–4444, 2021.
- [12] A. I. Mezza, M. Zanoni, and A. Sarti, “A latent rhythm complexity model for attribute-controlled drum pattern generation,” *J. Audio Speech Music Proc.*, vol. 2023, no. 1, p. 11, 2023.
- [13] G. Hadjeres, F. Nielsen, and F. Pachet, “GLSR-VAE: Geodesic latent space regularization for variational autoencoder architectures,” in *2017 IEEE Symp. Ser. Comput. Intell.*, 2017, pp. 1–7.
- [14] I. Cetin, M. Stephens, O. Camara, and M. A. González Ballester, “Attri-VAE: Attribute-based interpretable representations of medical images with variational autoencoders,” *Comput. Med. Imaging Graph.*, vol. 104, p. 102158, 2023.
- [15] B. Banar, N. Bryan-Kinns, and S. Colton, “A tool for generating controllable variations of musical themes using variational autoencoders with latent space regularisation,” *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 13, p. 16401–16403, Jun. 2023.
- [16] N. Tishby and W. Pereira, Fernando C. and Bialek, “The information bottleneck method,” in *37th Annu. Allerton Conf. Commun. Control Comput.*, 1999.
- [17] G. E. P. Box and D. R. Cox, “An Analysis of Transformations,” *J. R. Stat. Soc. Ser. B Methodol.*, vol. 26, no. 2, pp. 211–243, 1964.
- [18] I.-K. Yeo and R. A. Johnson, “A New Family of Power Transformations to Improve Normality or Symmetry,” *Biometrika*, vol. 87, no. 4, pp. 954–959, 2000.
- [19] S. Chen and R. Gopinath, “Gaussianization,” in *NeurIPS*, vol. 13, 2000.
- [20] V. Laparra, G. Camps-Valls, and J. Malo, “Iterative Gaussianization: From ICA to Random Rotations,” *IEEE Trans. Neural Netw.*, vol. 22, no. 4, pp. 537–549, 2011.
- [21] R. P. Brent, “An algorithm with guaranteed convergence for finding a zero of a function,” *Comput. J.*, vol. 14, no. 4, pp. 422–425, 1971.
- [22] A. Hyvärinen and E. Oja, “Independent component analysis: Algorithms and applications,” *Neural Netw.*, vol. 13, no. 4–5, pp. 411–430, 2000.
- [23] C. Raffel, “Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching,” Ph.D. dissertation, Columbia University, 2016.
- [24] G. Toussaint, “A mathematical analysis of African, Brazilian, and Cuban clave rhythms,” in *Bridges: Mathematical Connections in Art, Music, and Science*, 2002, pp. 157–168.
- [25] L.-C. Yang and A. Lerch, “On the evaluation of generative models in music,” *Neural Comput. Appl.*, vol. 32, no. 9, p. 4773–4784, 2020.
- [26] K. Choi, C. Hawthorne, I. Simon, M. Dinculescu, and J. Engel, “Encoding musical style with transformer autoencoders,” in *ICML*, vol. 119, 2020, pp. 1899–1908.