2nd Symposium on Advances in Approximate Bayesian Inference, 2019 1–6

HM-VAEs: a Deep Generative Model for Real-valued Data with Heterogeneous Marginals

Chao Ma CM905@CAM.AC.UK
University of Cambridge, Cambridge, UK
Sebastian Tschiatschek SEBASTIAN.TSCHIATSCHEK@MICROSOFT.COM
Microsoft Research Cambridge, Cambridge, UK
Yingzhen Li YINGZHEN.LI@MICROSOFT.COM
Microsoft Research Cambridge, Cambridge, UK
Richard Turner RET26@CAM.AC.UK
University of Cambridge, Cambridge, UK
José Miguel Hernández-Lobato JMH233@CAM.AC.UK
University of Cambridge, Cambridge, UK
Cheng Zhang CHENG.ZHANG@MICROSOFT.COM
Microsoft Research Cambridge, Cambridge, UK

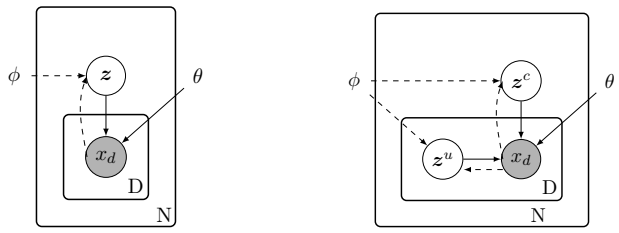
Abstract

In this paper, we propose a very simple but effective VAE model (HM-VAE) that can handle real-valued data with heterogeneous marginals, meaning that they have drastically distinct marginal distributions, statistical properties as well as semantics. Preliminary results show that the HM-VAE can learn distributions with heterogeneous marginal distributions, whereas the vanilla VAEs fails.

1. Introduction

Learning realistic generative models that are well calibrated to uncertainty, is important for understanding the data and applications to downstream tasks. Among many models of choice, Variational Autoencoders (VAEs) (Kingma and Welling, 2013) is a popular method for learning representations and capturing the correlations of high-dimensional data. VAE naturally enables uncertainty estimation in latent space, which is crucial for downstream applications (Ma et al., 2018; Gong et al., 2019).

However, VAE suffers when it is used to model real-world data with heterogenous marginals. Most real world data are statistically more complex than images and speech data where VAEs are typically applied to. Take image data as example, each dimensions of the data shares similar physical meaning (pixels), and share similar marginal distributions. We call this type of data *marginally homogeneous*. Many real world datasets, on the contrary, are marginally heterogeneous. For instance, medical databases often record lab test results of the patients. Even if all lab test variables are continuous and normalized in the pre-processing step, they might still have drastically distinct marginal distributions and statistical properties. We call them *real-valued datasets with heterogeneous marginals*. Naively applying vanilla VAEs to this type of data would fail (Figure 2 (b)).



(a) VAE (b) HM-VAE

Figure 1: Graphical representations of vanilla VAE and our HM-VAE. Note that in this graph, D is the dimension of observations $\boldsymbol{x}$, x_d is the dth dimension of $\boldsymbol{x}$

Contributions In this paper, we focus on improving VAEs for real-valued data with heterogeneous marginals. We introduce a method called heterogeneous-marginal VAE (HM-VAE) that explicitly decomposes intra-variable uncertainties (account for heterogeneous variations) and inter-variable uncertainties (account for correlation). We experimentally observe that the HM-VAEs can improve the data generation quality and generate realistic data with nearly indistinguishable marginals compared with real data.

2. Method

We first review the basic idea of variational auto-encoders, and then introduce our method.

Variational Autoencoders Variational autoencoders (VAEs) (Kingma and Welling, 2013) is a class of probabilistic generative model where the joint distribution is defined as $p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) = \prod_n p_{\boldsymbol{\theta}}(\boldsymbol{x}_n|\boldsymbol{z}_n)p(\boldsymbol{z}_n)$.In other words, each data $\boldsymbol{x}$ is generated from latent variables $\boldsymbol{z}$. Here, $p_{\boldsymbol{\theta}}(\boldsymbol{x}_n|\boldsymbol{z}_n)$ is often induced by the following generative process:

$$\boldsymbol{x} = f_\theta(\boldsymbol{z}) + \mathcal{N}(0, \boldsymbol{\epsilon}), \tag{1}$$

where f_θ is a neural network know as the *decoder*, and $\boldsymbol{\epsilon}$ is known as the *homogeneous noise level* that models the aleatoric uncertainty of the model. To approximate the posterior $p_\theta(\boldsymbol{z}_n|\boldsymbol{x}_n)$, VAEs uses an encoder, which takes the data $\boldsymbol{x}_n$ as input to produce the variational parameters of the posterior $q_\phi(\boldsymbol{z}_n|\boldsymbol{x}_n)$. The posterior statistics $\mu(\boldsymbol{x})$ and $\sigma(\boldsymbol{x})$ are parameterized by a deep neural network. Finally, the VAE model can be trained by optimizing the following variational lower bound (ELBO): $\log p_\theta(\boldsymbol{x}) \geq \mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}$.

Mixture of posteriors as prior In order to handle complicated heterogeneous data we make the VAE prior more flexible, our model choses a mixture of Gaussians (MoGs) as the prior distribution $p(\boldsymbol{z})$ for latent variable. We parameterize this MoG to be an aggregated posterior distribution (Makhzani et al., 2015; Tomczak and Welling, 2017), i.e., $p(\boldsymbol{z}) = \frac{1}{K}\sum_k q_\phi(\boldsymbol{z}|\boldsymbol{u_k})$. When $K \ll N$ and $\boldsymbol{u_k}$ are inducing locations to be optimized, this is known as the VampPrior approach (Tomczak and Welling, 2017).

Decomposing inter-variable and intra-variable uncertainties As a probabilistic tool for modelling multivariate data, fitting of VAEs to data can be decomposed in two tasks: i), *inter-variable uncertainties*: modelling the probabilistic inter-variable correlations; and ii), *intra-variable uncertainties*: modelling the epistemic uncertainties of each variables, which account for homogeneous statistics. We will separately model these *inter-variable* and *intra-variable* uncertainties in VAEs, respectively.

In vanilla VAEs (Eq 1 and Figure 1(a)), the representation for inter-variable and intra-variable uncertainties are entangled: the latent variables $\boldsymbol{z}$ are assumed to be responsible for explaining both shared correlations and private uncertainties at the same time. This assumption is useful when VAEs are applied to data with homogeneous marginals. For example, for image data, each dimension has the same semantics (pixels) and share similar statistical properties. However, for data with heterogeneous marginals, vanilla VAEs often struggle and result in a poor marginal distribution approximations (Figure 2 (b)).

Therefore, we propose to explicitly enforce the separation between shared correlation and private uncertainties. Let $\boldsymbol{z}$ be the set of latent variables of the model. As shown in Figure 1(b), we partite $\boldsymbol{z}$ into $\boldsymbol{z} = \boldsymbol{z}^c \cup \boldsymbol{z}^u$, and the new VAE model is defined as

$$p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) = \prod_n \prod_d p_{\boldsymbol{\theta}}(x_{n,d}|z^u_{n,d}, \boldsymbol{z}^c_n)p(\boldsymbol{z}_n) \tag{2}$$

$$p_{\boldsymbol{\theta}}(x_{n,d}|z^u_{n,d}, \boldsymbol{z}^c_n) = F_{\theta_u}(\tilde{x}_{n,d}, z^u_{n,d}) + \mathcal{N}(0, \boldsymbol{\epsilon}) \tag{3}$$

$$\tilde{x}_{n,d} = f_{\theta_d}(\boldsymbol{z}^c_n) \tag{4}$$

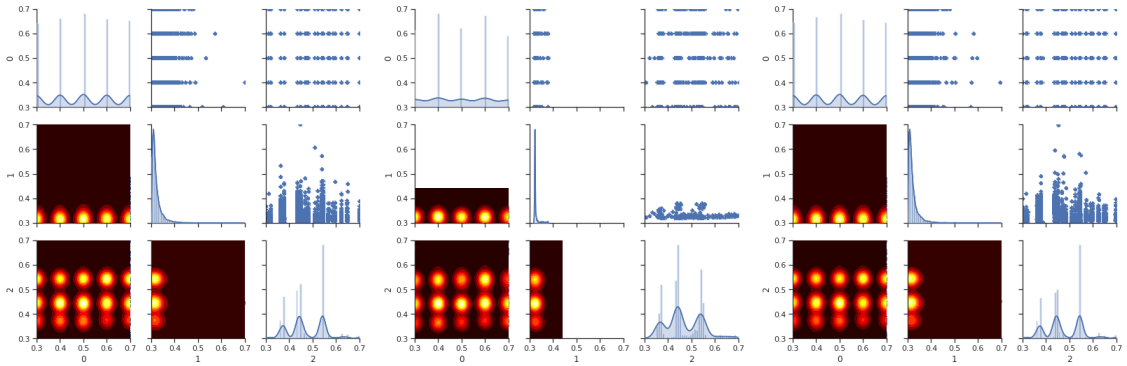
Where $x_{n,d}$ is the d th dimension scalar of the n th data point. Both F_{θ_u} and f_{θ_d} are deterministic neural networks with one dimensional output. $\boldsymbol{z}^c$ is shared across different x_d and is encouraged to be only responsible for modeling shared correlation. $\boldsymbol{z}^u_d$ is private to x_d, hence is only responsible for modeling the private epistemic uncertainty of the dth variable x_d. In order to generate $x_{n,d}$ from $p_{\boldsymbol{\theta}}(x_{n,d}|z^u_{n,d}, \boldsymbol{z}^c_n)$, a "homogeneous" version $\tilde{x}_{n,d}$ of $x_{n,d}$ is first generated from $f_{\theta_d}(\boldsymbol{z}^{\boldsymbol{c}}_{\boldsymbol{n}})$, which is just a vanilla VAE decoder. Then, together with the private uncertainty $z^u_{n,d}$, $\tilde{x}_{n,d}$ is passed to an aggregation network F_{θ_u}(shared across dimensions), which generates the final output $x_{n,d}$. We choose to parameterize $p_{\boldsymbol{\theta}}(x_{n,d}|z^u_{n,d}, \boldsymbol{z}^c_n)$ in this way, so that a vectorized implementation can be easily used to speed up computation. For variational inference, we use Gaussian inference nets for both $\boldsymbol{z}^u$ and $\boldsymbol{z}^c$. We call this improved model the heterogeneous-marginal VAE (HM-VAE).

Related work A closely related work is the Heterogeneous-Incomplete VAE (HI-VAE) (Nazabal et al., 2018). In HI-VAE, the term "heterogeneous" has different meanings. HI-VAE mainly focus on data with variables that might have *different types* (continuous, discrete, categorical, ordinal, etc). If all the variables have continuous marginals (which is the case in our paper) and are pre-normalized, HI-VAE degrades into vanilla VAE with a mixture of Gaussian prior (which forms the baseline in our experiments).

3. Experiment

3.1. Data and Settings

We apply our proposed method (HM-VAE) on a dataset with heterogeneous marginals called Bank Marketing Data Set (Moro et al., 2014). The Bank dataset is a real world dataset related to marketing campaigns of banking institutions. We only focus on the subset of 13 continuous variables since dealing with mixed type is out of the scope of this work. Figure 2 (a) shows the pair-wise plots for 3 of the variables from this dataset. Note that in the Bank dataset, the marginal distributions are drastically different from each other, each have different properties and number of modes.



(*a*) Real data (*b*) Vanilla VAE (*c*) HM-VAE

Figure 2: This figure visualizes the pair plots of three variables selected from the data. **(a): real data**; **(b): vanilla VAE with VampPrior**; **(c): HM-VAE**. Diagonal plots shows the marginal histograms (accompanied by kernel density estimates) of each variable. The upper-triangular part shows sample scatter plots of each variable pair. The lower-triangular part shows heat maps of high-probability regions. Note that vanilla VAEs struggle and result in mismatch in marginal distributions, while HM-VAE does not have this issue.

We train a HM-VAE and vanilla VAE (equipped with VampPrior) on training set, and quantitatively compare their performance on test set using a 90%-10% split. Note that the Heterogeneous-Incomplete VAE (Nazabal et al., 2018) degrades to our baseline under this setting. With the same decoder structure (two layers, 50-100 structure). All models are trained with the Adam optimizer with a learning rate of 0.001. All data are pre-normalized and we set aleatoric noise levels to be $\epsilon = 0.02$.

3.2. Results

As an example, we first visualize the data generation quality of each model. Due to limited space, Figure 2 only visualizes the pair plots of three variables selected from the data. Full plots on the complete dataset can be found in Appendix A. In each subfigure of Figure 2 (a)-(c), diagonal plots shows the marginal histograms (accompanied by kernel density estimates) of each variable. The upper-triangular part shows sample scatter plots of each variable pair. The lower-triangular part shows heat maps of high-probability regions.

Note that the second variable, which corresponds to the "duration" feature of the dataset, is a very important variable that has a heavy tail. The vanilla VAE (Figure 2) is able to identify the high-probability regions, but fails to mimic this heavy tail behaviour of the variable. On the other hand, HM-VAE (Figure 2) is able to exactly reproduce the heavy tail behaviour. It can also generate marginals/second order correlations that are nearly indistinguishable from the real data.

To evaluate the data generation quality quantitatively, we compute the MMD distance (Gretton et al., 2012) between real and generated samples. As shown in Table 2, we both compute the MMD distance on the full data, and on the marginal distribution of each variable, respectively. Based on these results, HM-VAE can consistently generate more realistic samples, both jointly and marginally.

Variable	Full	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13
VAE	1.2e-2	9.9e-4	7.9e-3	7.3e-2	1.9e-1	9.9e-1	4.0e-1	2.4e-1	3.2e-1	1.1e+0	9.4e-1	9.0e-1	6.8e-1	1.4e+0
HM-VAE	**5e-3**	1.1e-3	**1.7e-3**	**2.9e-4**	**3.0e-4**	**2.9e-4**	**9.3e-4**	**1.3e-3**	**1.3e-3**	**9.9e-4**	**7.4e-4**	**2.3e-3**	**1.5e-3**	**1.7e-4**

Table 1: MMD estimation (real vs generated) on Bank data

Table 2: The MMD on the full data (real vs generated), as well as on the marginal distribution of each variables are computed, respectively.

4. Conclusion

In this paper, we focused on improving VAEs for real-valued data that has heterogeneous marginal distributions. We propose the heterogeneous-marginal VAE (HM-VAE), a method that explicitly decomposes intra-variable uncertainties and inter-variable uncertainties. We experimentally observe that the HM-VAEs can generate realistic data with nearly indistinguishable marginals when compared with real data.

References

Wenbo Gong, Sebastian Tschiatschek, Richard Turner, Sebastian Nowozin, and José Miguel Hernández-Lobato. Icebreaker: Element-wise active information acquisition with bayesian deep latent gaussian model. *arXiv preprint arXiv:1908.04537*, 2019.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

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

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, Jose Miguel Hernandez Lobato, Sebastian Nowozin, and Cheng Zhang. Eddi: Efficient dynamic discovery of high-value information with partial vae. *arXiv preprint arXiv:1809.11142*, 2018.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

Alfredo Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using vaes. *arXiv preprint arXiv:1807.03653*, 2018.

Jakub M Tomczak and Max Welling. Vae with a vampprior. *arXiv preprint arXiv:1705.07120*, 2017.

Appendix A. Full Pair plots

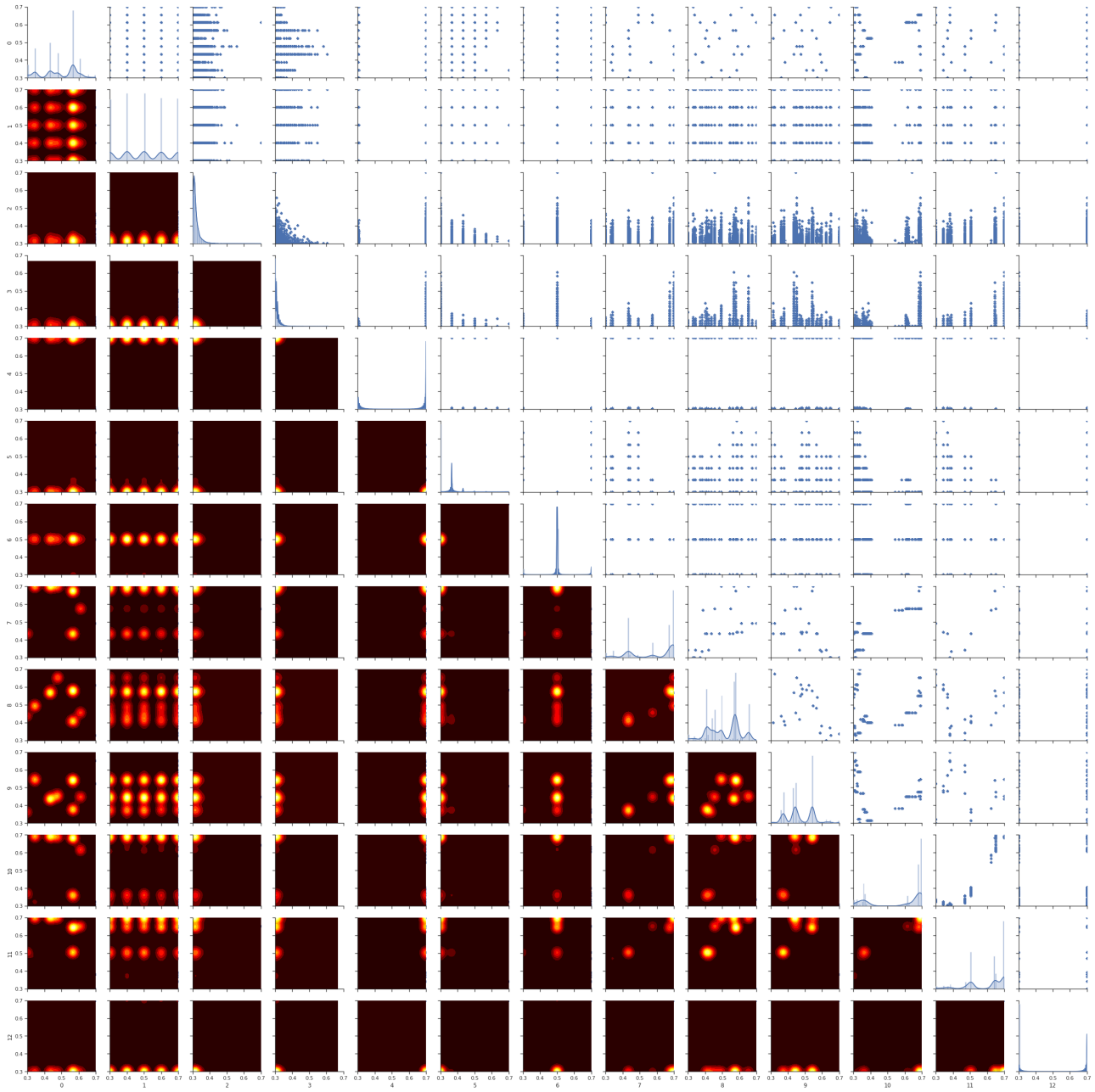


Figure 3: pair plots of all variables from the real Bank dataset. Diagonal plots shows the marginal histograms (accompanied by kernel density estimates) of each variable. The upper-triangular part shows sample scatter plots of each variable pair. The lower-triangular part shows heat maps of high-probability regions.

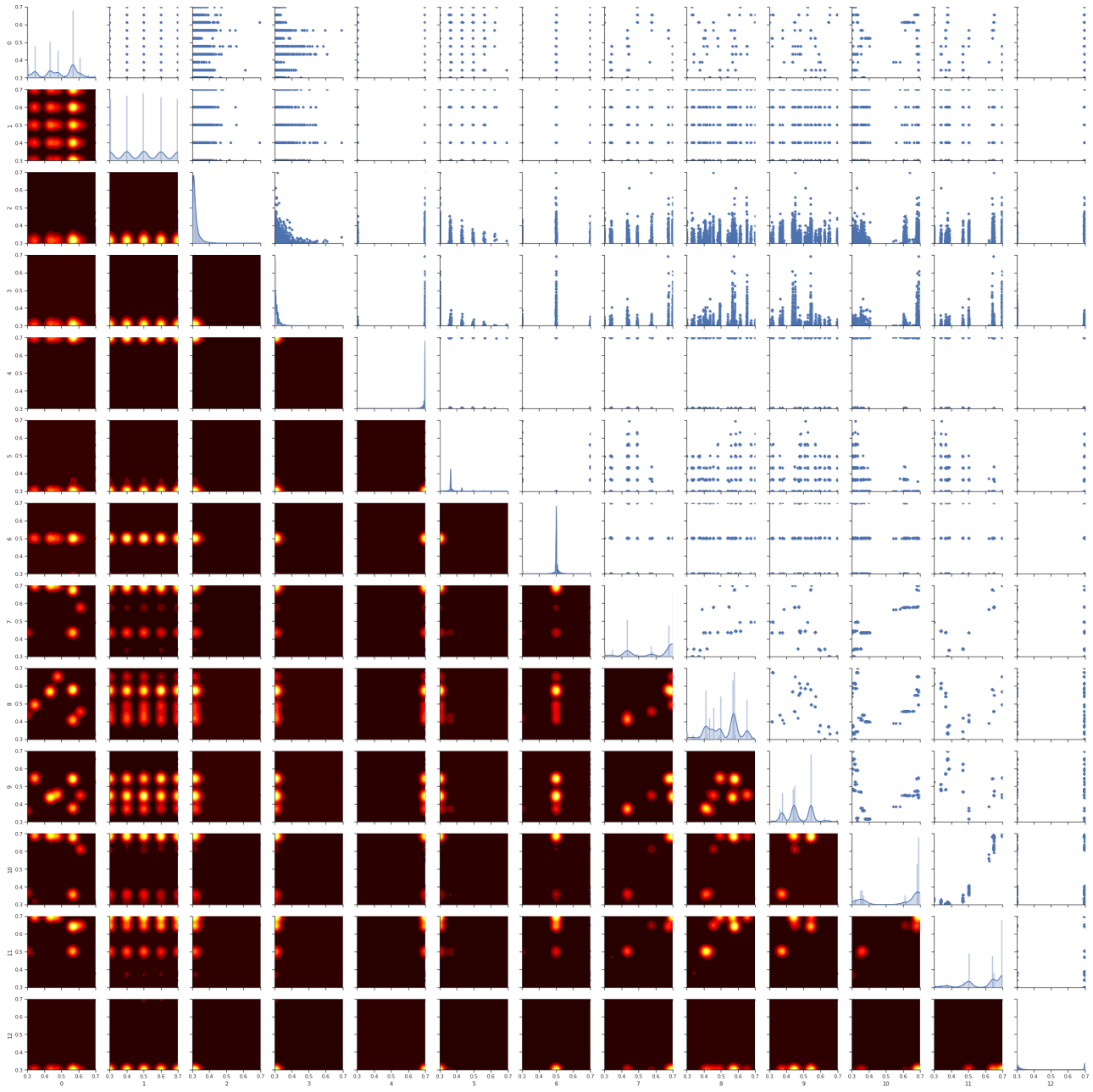


Figure 4: pair plots of all variables generated by HM-VAE. Diagonal plots shows the marginal histograms (accompanied by kernel density estimates) of each variable. The upper-triangular part shows sample scatter plots of each variable pair. The lower-triangular part shows heat maps of high-probability regions.

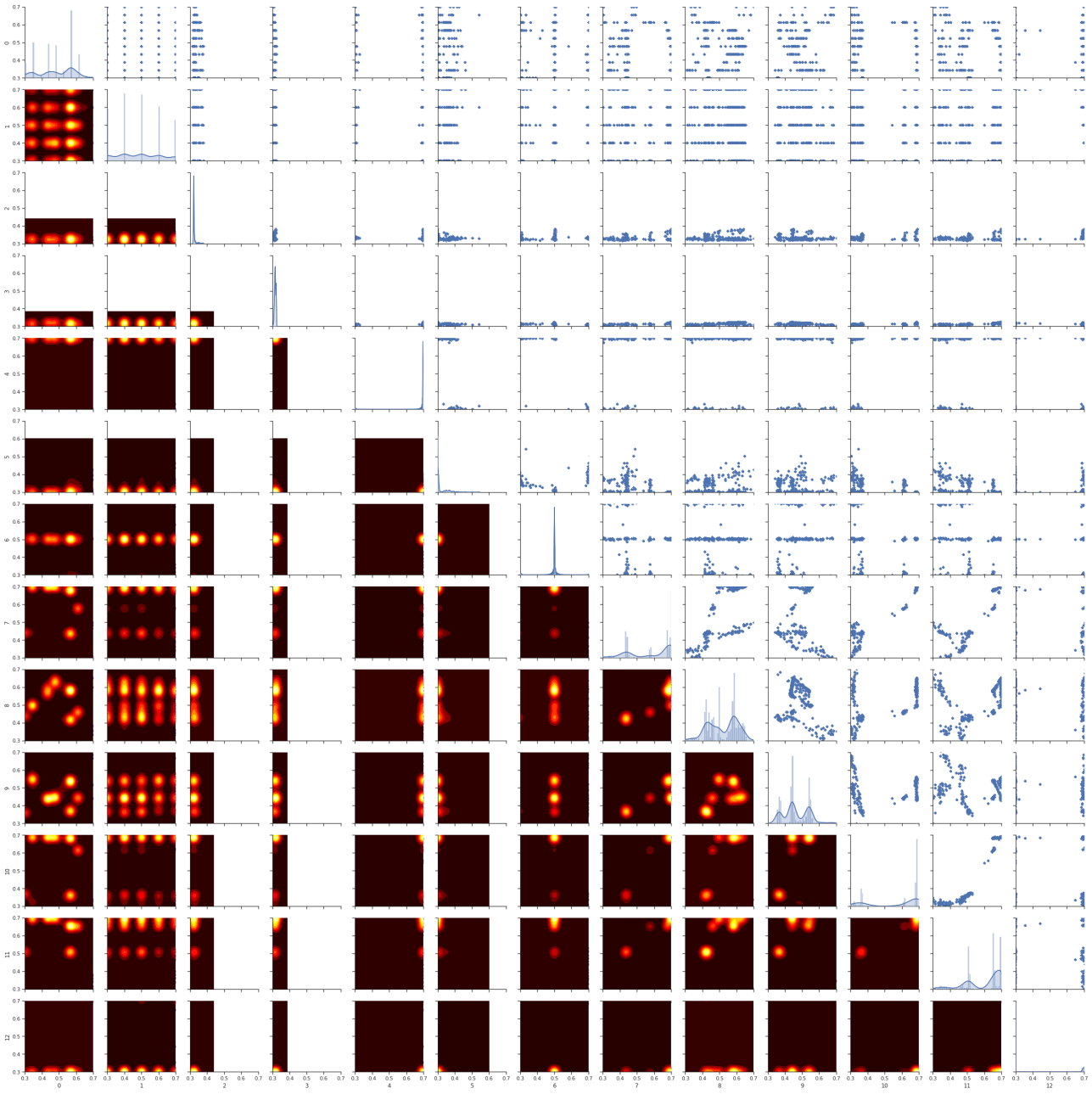


Figure 5: pair plots of all variables generated by vanilla VAE with VampPrior. Diagonal plots shows the marginal histograms (accompanied by kernel density estimates) of each variable. The upper-triangular part shows sample scatter plots of each variable pair. The lower-triangular part shows heat maps of high-probability regions.