Appendix: Discriminative Regularization for Latent Variable Models with Applications to Electrocardiography

Andrew C. Miller [1] **Ziad Obermeyer** [2] **John P. Cunningham** [3] **Sendhil Mullainathan** [4]

A. Related Work (cont'd)

Interpreting and understanding black-box models is an active area of research. Doshi-Velez & Kim (2017) describe guidelines to assess and contextualize work on interpretability of machine learning systems. Within this framework, we view our approach as a way to address the incompleteness of scientific understanding by forming data-driven "cognitive chunks" with a particular focus on "intrinsic task explanations" (Narayanan et al., 2018).

Another approach to interpreting a black-box model is to consider a simpler surrogate model. The local interpretable model agnostic explanation (LIME) approach builds a locally linear (and sometimes sparse) approximation of the original model at a particular input (Ribeiro et al., 2016). One can also estimate Shapley values to explain important predictors (Lundberg & Lee, 2017). In addition to simple local models, program induction has been proposed as a representation of interpretable explanations of complex algorithms (Singh et al., 2016). These approaches aim to provide insight into how *important* input variables contribute to the model prediction. However, when the input is a highly structured high-dimensional observation (e.g an image or a time series) it can be unclear what constitutes a distinct "variable" — these kind of inputs require a learned representation before these (local) variable importance measures become applicable. Visualizing intermediate layer activations of deep neural networks is another way to peer into the structure of a prediction algorithm (Olah et al., 2018).

Wagstaff (2013) details an approach somewhat similar to our discriminative representation inversion, however, does not explicitly represent unlabeled latent variation. Similarly, the "lensing" approach incorporates interpretable latent variable mappings into the model building, fitting, and criticism loop (Dinakar, 2017).

In unsupervised settings, maximum likelihood (or approximate maximum likelihood) is commonly used to train generative models. However, for representation learning, maximum likelihood optimization with flexible generative models may not yield useful representations $\mathbf{z}$ without task-specific constraints (Huszàr, 2017; Alemi et al., 2018).

Similarly, learning interpretable representations with DGMs has received recent attention. These unsupervised approaches use a particular type of regularization on the latent space — a group-sparsity penalty (Ainsworth et al., 2018), an information theoretic penalty on the total correlation (Chen et al., 2018), or additional penalties on the variational objective (Higgins et al., 2016).

To compute model-based morphs, the gradient is one of many directions that increase the model output $m(\cdot)$; we leave the construction of other trajectories for future study. We also note that one can better incorporate the geometry of the surface model represented by $g_{\theta}(\mathbf{z})$ using the Riemannian metric tensor $J_{\mathbf{z}}^{\intercal}J_{\mathbf{z}}$ where $J_{\mathbf{z}}$ is the Jacobian of $g_{\theta}(\mathbf{z})$ with respect to input $\mathbf{z}$, as suggested in Arvanitidis et al. (2017). For computational reasons, however, we form paths using the standard gradient and leave comparisons to geodesics formed with the Riemannian metric tensor for future investigation.

B. Synthetic Experiment Details

The synthetic data have a structured covariance matrix. The dimensions in $\mathbf{x}$ are correlated via a non-stationary squared exponential kernel. We construct the covariance over synthetic data $\mathbf{x}$ as follows

$$\Sigma_{i,j} = \sigma_i \sigma_j \exp\left(-\frac{1}{2}(s_i - s_j)^2/\ell^2\right) \tag{1}$$

where s_i is on an evenly spaced grid between -2 and 2, σ_i is on a fixed grid from .1 to 1, and the length scale $\ell = (1/15)^{1/2} = .258$. This is a squared exponential kernel covariance function with heteroscedastic marginal variance — the later dimensions will have much larger marginal variance than the first dimensions.

[1]Data Science Institute, Columbia University, New York, NY, USA [2]Department of Public Health, UC Berkeley, Berkeley, CA, USA [3]Department of Statistics, Columbia University, New York, NY, USA [4]Booth School of Business, University of Chicago, Chicago, IL, USA. Correspondence to: ACM <am5171@columbia.edu>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

C. Modified MNIST Experiments

To bridge the gap between the totally synthetic setting and a real-world example, we apply a DR-VAE to a modified MNIST data set. We construct a semi-synthetic dataset based on MNIST where the predictive pixels form a cross (5 pixels in the + configuration) in the upper left corner of the image. The intensity of the cross is uniformly distributed between 0 and 1, and this constitutes the only true "predictive variation" — the higher the intensity, the higher the predicted value. Figure 1a depicts this training dataset.

We fit deep generative models with a single 400-unit hidden layer with a ReLU non-linearity and a sigmoid output layer. We fix the latent variable size to $K = 10$ across all models, and train each model for 200 epochs on the MNIST training examples.

In high-dimensional data, specific directions of variation can easily be buried. Without guidance, the VAE reverts those pixels to their population average values, and focuses on predicting the more complex (but more rewarding in the loss function) variation in the digits themselves. The variation in the cross is ignored by the VAE even when these are the five most variable pixels in the entire dataset.

In Figure 1b we visualize test sample reconstructions from a VAE — a generative model without guidance from the discriminative model. In these reconstructions, we see the + is reconstructed, however each reconstruction simply takes on the *average* value in those pixels.

With discriminative regularization, we see in Figure 1c that the variation in each + is represented (from 0 to 1). We quantify this predictive variation captured by the DR-VAE (vs. the VAE) in Figure 1d. We see that setting β to between .001 and .0025, we recover variation in those pixels close to the population variation (the rightmost bar). The VAE generative model, without this additional constraint, is unable to explore the predictive variation in the + pixels by perturbing the latent variable $\mathbf{z}$. The DR-VAE trained model does contain this variation with a sufficiently large β coefficient.

D. EKG Data and Experimental Setup

The details of each EKG data set are presented in Figure 2. We model four outcomes — age, bundle branch block, major adverse cardiac events, and ST elevation.

E. EKG Morphing Statistics

This section contains further summaries and visualizations of model-morphed data. In Figure 3 we visualize the mean morphing and marginal standard deviation in all three leads for the VAE, and DR-VAE with $\beta = 1, 5$, and 10 for ST Elevation. In Figure 4 we visualize the variance of the principal components of $\hat{\Sigma}_{morph}$.

F. Expert Test Experimental Setup

To further validate our generative model of EKGs, we ran an expert labeling experiment. The goal of the experiment is to test how realistic the synthetic data and the model-morphing induced feature look to a physician expert.

To answer these questions, we construct a two-alternative forced choice labeling task. We present an expert — an emergency medicine physician with experience interpretting EKG tracings — with N trials, where each trial compares a pair of EKG beats. In each trial we present a real EKG beat and a fake EKG beat (in a randomized order). One task is straightforward — can the expert distinguish the real from the fake beat? We compare the empirical rate of accurately labeled trials to random guessing.

We generate four different types of synthetic examples

- $\bar{\mathbf{x}}_{lo}$: model reconstruction of test data point — both $m(\mathbf{x})$ and $m(\tilde{\mathbf{x}})$ are "low" (i.e. $Pr(\mathbf{y} = 1|\mathbf{x}_{lo}) = \epsilon$)
- $\bar{\mathbf{x}}_{hi}$: model reconstruction of test data point — both $m(\mathbf{x}_{hi})$ and $m(\tilde{\mathbf{x}}_{hi})$ are "high" (i.e. $Pr(\mathbf{y} = 1|\mathbf{x}_{hi}) = 1 - \epsilon$).
- $\tilde{\mathbf{x}}_{\uparrow}$: morphed data — the original reconstruction started "low" (i.e. $Pr(\mathbf{y} = 1 \,|\, \mathbf{x}_{start}) = \epsilon$)) and the resulting morphed image is "high".
- $\tilde{\mathbf{x}}_{\downarrow}$: same as above, but the other direction.

EKGs were labeled "low ST Elevation" if the predictor $m(\mathbf{x})$ put them in the second decile. EKGs were labeled "high ST Elevation" if the predictor $m(\mathbf{x})$ put them in the ninth decile. The morphed images started from the second decile and were morphed to the ninth (or vice versa).

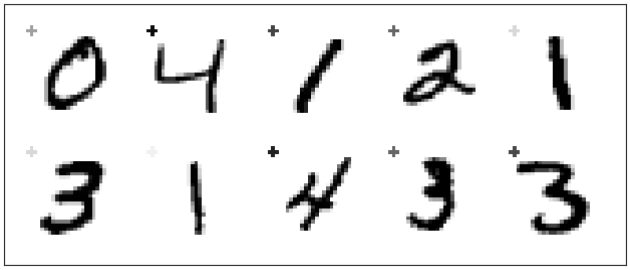
We reproduce the results of the expert labeling task in Figure 5 — the real-vs-fake test in Table 5a and the high-vs-low ST elevation results in Table 5b.

References

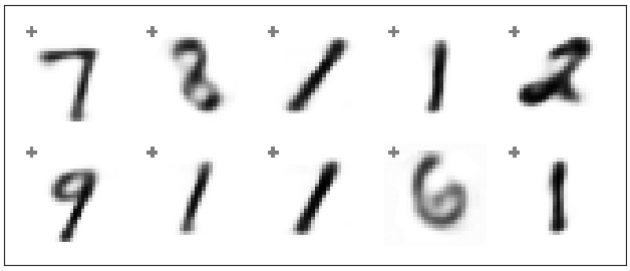
Ainsworth, S., Foti, N., Lee, A. K., and Fox, E. Interpretable vaes for nonlinear group factor analysis. *arXiv preprint arXiv:1802.06765*, 2018.

Alemi, A., Poole, B., Fischer, I., Dillon, J., Saurous, R. A., and Murphy, K. Fixing a broken ELBO. In *International Conference on Machine Learning*, 2018.

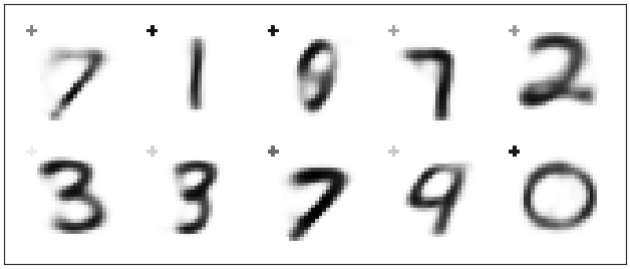
Arvanitidis, G., Hansen, L. K., and Hauberg, S. Latent space oddity: on the curvature of deep generative models. *arXiv preprint arXiv:1710.11379*, 2017.



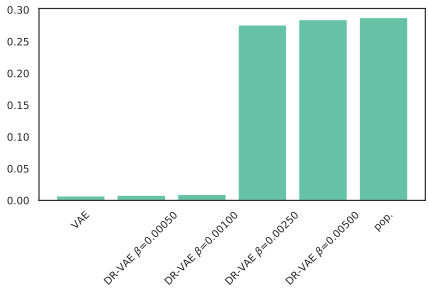
(a) Training examples



(b) VAE reconstructions



(c) DR-VAE reconstructions ($\beta = .005$)



(d) Predictive standard deviation.

Figure 1. Modified MNIST data experiment. (a) The original training examples with variable + pixels. (b) VAE reconstructions. (c) DR-VAE reconstructions. (d) The standard deviation of the predictive pixels — we depict the marginal standard deviation of the 5 predictive + pixels with increasing β values. On the left is the standard VAE, which represents nearly zero variation in those pixels. On the right is the true population marginal standard eviation. With a strong enough regularizer (e.g. $\beta = .0025$), the generative model represents nearly all variation with the generative model structure $g_{\boldsymbol{\theta}}(\mathbf{z})$.

Chen, T. Q., Li, X., Grosse, R., and Duvenaud, D. Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*, 2018.

Dinakar, K. *Lensing Machines: representing perspective in machine learning*. PhD thesis, Massachusetts Institute of Technology, 2017.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint*

Dataset Characteristics	Train	Val	Test
# EKGs	211838	53094	87205
Patient demographics			
# unique patients	14274	3568	5949
mean age (sd)	57.1 (19.1)	57.7 (18.5)	57.7 (18.9)
# female (%)	7852 (55.0 %)	1957 (54.8 %)	3369 (56.6 %)
Beat outcomes: mean (stdev)			
age (standardized)	-0 (1.00 %)	0 (0.97 %)	0 (0.99 %)

(a) age

Dataset Characteristics	Train	Val	Test
# EKGs	183688	44828	75135
Patient demographics			
# unique patients	11916	2979	4966
mean age (sd)	60.1 (19.2)	59.7 (19.0)	59.8 (19.0)
# female (%)	6235 (52.3 %)	1549 (52.0 %)	2630 (53.0 %)
Beat outcomes: total positive (%)			
bbb	61080 (33.25 %)	13857 (30.91 %)	24679 (32.85 %)

(b) Bundle Branch Block

Dataset Characteristics	Train	Val	Test
# EKGs	88908	21308	36383
Patient demographics			
# unique patients	5677	1419	2367
mean age (sd)	58.7 (19.8)	59.3 (19.4)	59.1 (19.5)
# female (%)	3142 (55.3 %)	766 (54.0 %)	1299 (54.9 %)
Beat outcomes: total positive (%)			
mace	45089 (50.71 %)	10915 (51.22 %)	19047 (52.35 %)

(c) MACE

Dataset Characteristics	Train	Val	Test
# EKGs	24767	6075	10301
Patient demographics			
# unique patients	1995	498	833
mean age (sd)	55.9 (19.7)	56.2 (19.1)	55.7 (20.1)
# female (%)	992 (49.7 %)	216 (43.4 %)	390 (46.8 %)
Beat outcomes: total positive (%)			
steleva	8060 (32.54 %)	1965 (32.35 %)	3469 (33.68 %)

(d) ST Elevation

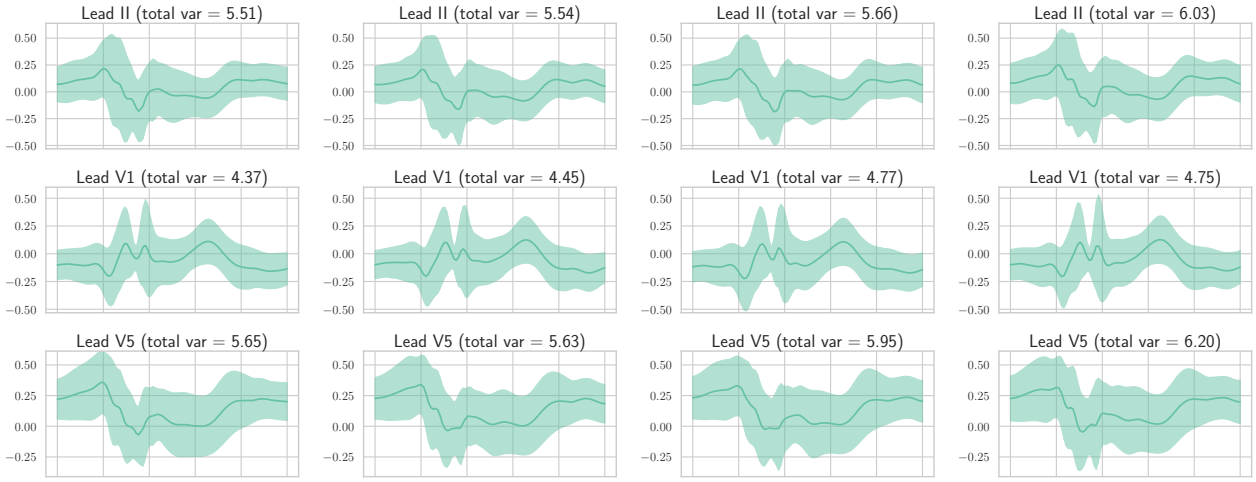
Figure 2. Data set statistics for the four outcomes

arXiv:1702.08608, 2017.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. β-vae: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations*, 2016.

Huszàr, F. Is maximum likelihood useful for representation learning? http://www.inference.vc, 2017.

Lundberg, S. M. and Lee, S.-I. A unified approach to inter-



(a) VAE (b) DR-VAE, $\beta = 1$ (c) DR-VAE, $\beta = 5$ (d) DR-VAE, $\beta = 10$

Figure 3. Model-morph variation, $\hat{\Sigma}_{morph}$, visualization varying β. Above we plot the mean morphing delta (for lead V1) and one standard deviation (computed on 1024 for the ST Elevation predictor). On the left, the standard VAE exhibits small morphing variation. As we increase β, the model morphs exhibit significantly more variation, indicating the DR-VAE represents a richer set of predictive features than the standard VAE. Note that generative reconstruction is similar for all three of these models.

preting model predictions. In *Advances in Neural Information Processing Systems*, pp. 4768–4777, 2017.

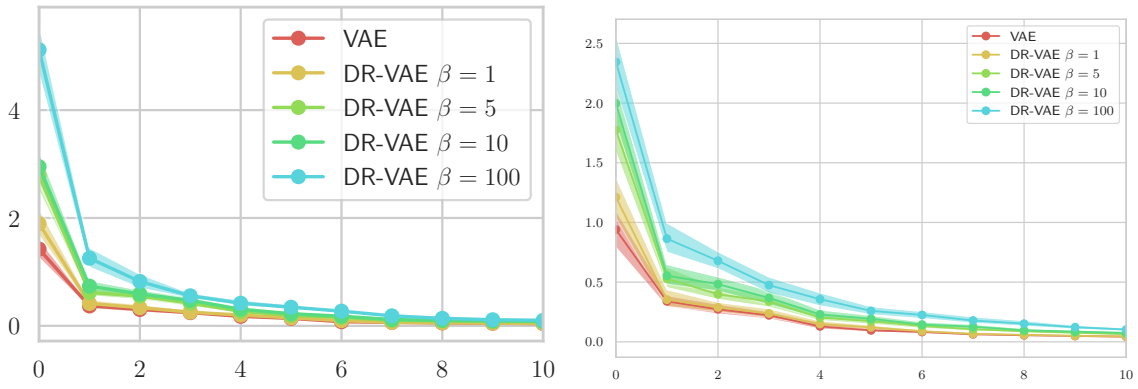
Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., and Doshi-Velez, F. How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*, 2018.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., and Mordvintsev, A. The building blocks of interpretability. *Distill*, 2018. doi: 10.23915/distill.00010. https://distill.pub/2018/building-blocks.

Ribeiro, M. T., Singh, S., and Guestrin, C. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

Singh, S., Ribeiro, M. T., and Guestrin, C. Programs as black-box explanations. *arXiv preprint arXiv:1611.07579*, 2016.

Wagstaff, K. L. Guiding scientific discovery with explanations using demud. 2013.



(a) ST Elevation: low-to-high morph

(b) ST Elevation: high-to-low morph

Figure 4. DR-VAEs have more dimensions of morphing variation than VAEs. Depicted the eigenvalues of the morphing covariance matrix for the two experiments — (a) low-to-high model-morphs and (b) high-to-low. The VAE has the least amount of total variation (summed across all dimension) and fewer dimensions of non-zero variance, indicating that the DR-VAE representation $\mathbf{z}$ is capturing more predictive variation in data $\mathbf{x}$.

type	accuracy
$\bar{\mathbf{x}}_{lo}$	60% [46-74%]
$\bar{\mathbf{x}}_{hi}$	64% [50-78%]
$\tilde{\mathbf{x}}$ low-to-high	76% [64-88%]
$\tilde{\mathbf{x}}$ high-to-low	42% [28-56%]

(a) Expert test real vs. synthetic accuracy results

type	accuracy
$\bar{\mathbf{x}}_{lo}$	100% [100-100%]
$\bar{\mathbf{x}}_{hi}$	92% [84-100%]
$\tilde{\mathbf{x}}$ low-to-high	88% [78-96%]
$\tilde{\mathbf{x}}$ high-to-low	94% [88-100%]

(b) Expert test low vs. high ST accuracy results

Figure 5. Expert labeling results.