End-to-end training of deep probabilistic CCA for joint modeling of paired biomedical observations

Gregory Gundersen
Princeton University
ggundersen@princeton.edu

Bianca Dumitrascu
Princeton University
biancad@princeton.edu

Jordan T. Ash
Princeton University
jordanta@cs.princeton.edu

Barbara E. Engelhardt
Princeton University
bee@princeton.edu

1 Introduction

Many diseases are diagnosed by pathologists using key morphological features in imaging data; but the genes that capture the internal state of cells and are therefore associated with specific morphologies are typically unknown. To address this question, the GTEx Consortium [8; 7] has collected data from over 948 autopsy research subjects, including standardized whole tissue slides and RNA sequencing gene expression levels from approximately 50 different human tissues. These multi-subject, multi-view data provide an opportunity to develop computational tools that quantify how genome-level associations affect morphological features observable in histology slides.

Our work builds upon a line of research rooted in canonical correlation analysis (CCA) [15]. Given two random vectors, CCA aims to find the linear projections for which the projected vectors are maximally correlated. Since its development, CCA has been reframed as a probabilistic model [4; 20] and has been extended to nonlinear settings using kernel methods [1; 14] and neural networks [2]. Probabilistic CCA (PCCA) is particularly attractive for medical applications with small sample sizes by explicitly capturing uncertainty [11]. However, fitting PCCA to high-dimensional data is difficult since the time to fit the model scales superlinearly with the data's dimensionality [17]. In addition, PCCA will ignore possibly important nonlinear structure in data such as medical images; this structure could be extracted with computer vision techniques such as convolutional neural networks [21] that have achieved excellent performance on medical imaging tasks [5; 26; 9; 10; 13].

Given these challenges, two natural extensions would be to either fit a model in a two-stage approach of first embedding the data and then fitting PCCA to the embedded data, or to train a model end-to-end that includes both automatic feature learning and PCCA. For example, a recent two-stage approach fit CCA to gene expression data and image embeddings from a convolutional autoenoder [3]. Another approach computed the principal components of single views and then computed cross-view correlations [6]. But two-stage approaches decouple feature learning from joint modeling, and we hypothesize that end-to-end training will learn features that are good for both reconstruction and joint modeling. Thus, the second extension would be to use neural networks for a nonlinear embedding of the observations before feeding these embedded features to PCCA in a single training epoch. Variants of the end-to-end model exist, including deep canonical correlation analysis (DCCA) [2], and deep variational canonical correlation analysis (DVCCA) [27]. DCCA estimates linear projections on the outputs of neural networks, but the model is not generative. DVCCA is generative, but uses nonlinear observation models for the full network. This makes the model difficult to interpret. While "interpretability" is a broad concept [23], here we mean specifically that we can identify small subsets of genes that are associated with specific image features. Being Bayesian, this means we want a sparse prior. Joint training of probabilistic models and neural networks is challenging [18; 22], and we propose a differentiable, variational take on this problem.

We want a model that is *Bayesian*, meaning it is generative, models uncertainty, and allows for sparsity-inducing priors, and *deep*, meaning it learns features from observations with nonlinear

Third workshop on Bayesian Deep Learning (NeurIPS 2018), Montréal, Canada.

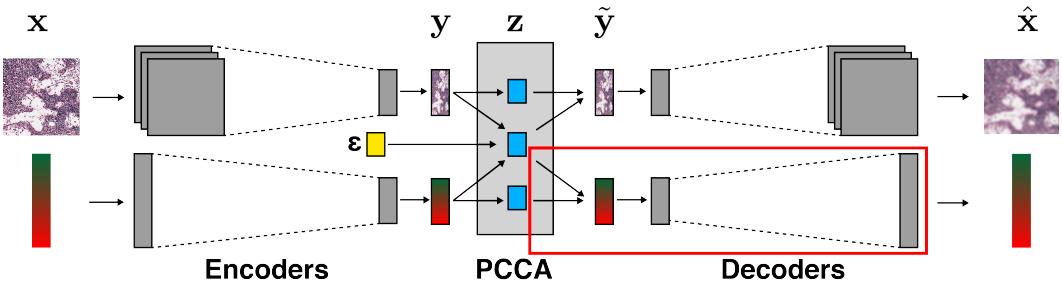

Figure 1: **Diagram of DPCCA.** The input is a paired set of pathology images and gene expression levels. The model is trained by fitting PCCA to embeddings from convolutional and linear autoencoders (gray), sampling from the PCCA model using the reparameterization trick (yellow), and then backpropagating through the model using the reconstruction loss. The model learns common and modality-specific latent variables (blue). Sparsity is induced on the PCCA parameters for the gene expression levels (red box).

structure. To address this challenge, we develop a multimodal modeling and inference framework that estimates shared latent structure of paired gene expression and medical image embeddings while learning these embeddings end-to-end.

2 Deep Probabilistic CCA

Notation. We use subscript i in $\{1, 2, \ldots, n\}$ to index samples and superscript $j \in \{a, b\}$ to index data set views. We denote the set of observations for view j as matrix $\mathbf{X}^j \in \mathbb{R}^{n \times d_j}$, where the ith row of $\mathbf{X}^j$ is a real-valued vector $\mathbf{x}_i^j$. Consider n paired samples $\{(\mathbf{x}_i^a, \mathbf{x}_i^b)\}_{i=1}^n$. In our case, $\mathbf{x}_i^a$ is a histology image and $\mathbf{x}_i^b$ is gene expression levels taken from sample i.

Model and Training. Deep probabilistic CCA (DPCCA) is a deep, generative model that fits PCCA to the embeddings of two autoencoders, training the model end-to-end with backpropogation through the reconstruction loss (Figure 1). The model has a Laplacian prior on the PCCA weights for the gene expression levels to encourage sparsity.

In detail, given samples $\mathbf{x}^a$ and $\mathbf{x}^b$, the encoders $E^a(\cdot; \boldsymbol{\theta}^a_{\text{enc}})$ and $E^b(\cdot; \boldsymbol{\theta}^b_{\text{enc}})$ embed their respective modalities into separate d-dimensional vectors, $\mathbf{y}^a \in \mathbb{R}^d$ and $\mathbf{y}^b \in \mathbb{R}^d$. Each encoder may be modality- and data set-specific. The d-dimensional encoded vectors $\mathbf{y}^a$ and $\mathbf{y}^b$ are the inputs to a PCCA module whose output are three k-dimensional latent embeddings: a *common* latent embedding $\mathbf{z}^c$ shared across modalities, and two *modality-specific* embeddings $\mathbf{z}^a$ and $\mathbf{z}^b$:

$$\mathbf{z}^a, \mathbf{z}^b, \mathbf{z}^c \sim \mathcal{N}(\mathbf{0}_k; \mathbf{I}_k) \qquad \mathbf{y}^j \sim \mathcal{N}(\boldsymbol{\Lambda}^j \mathbf{z}^j + \boldsymbol{\Lambda}^{jc} \mathbf{z}^c; \boldsymbol{\Psi}^j) \tag{1}$$

where $\mathbf{z}^j \in \mathbb{R}^k$, $\boldsymbol{\Psi}^i \in \mathbb{R}^{k \times k}$, and $\boldsymbol{\Lambda}^j \in \mathbb{R}^{d \times k}$. Note that $\boldsymbol{\Lambda}^{jc}$ denotes the parameters for the common latent variable $\mathbf{z}^c$ [24]. The expectation–maximization (EM) parameter updates for PCCA are equivalent to the EM parameter updates for factor analysis given appropriately tiled data: $\mathbf{y} = \langle \mathbf{y}^a, \mathbf{y}^b \rangle^\top$, $\mathbf{z} = \langle \mathbf{z}^a, \mathbf{z}^b, \mathbf{z}^c \rangle$, $\boldsymbol{\Lambda} = \langle \boldsymbol{\Lambda}^a, \boldsymbol{\Lambda}^b \rangle^\top$, and $\boldsymbol{\Psi} = \text{diag}(\langle \boldsymbol{\Psi}^a, \boldsymbol{\Psi}^b \rangle)$. The EM algorithm for parameters $\boldsymbol{\Lambda}^\star$ and $\boldsymbol{\Psi}^\star$ for a single sample $\mathbf{y}$ is [12; 4],

$$\boldsymbol{\Lambda}^\star = \Big(\mathbf{y}\mathbb{E}_{\mathbf{z}|\mathbf{y}}\big[\mathbf{z} \mid \mathbf{y}\big]^\top\Big)\Big(\mathbb{E}_{\mathbf{z}|\mathbf{y}}\big[\mathbf{z}\mathbf{z}^\top \mid \mathbf{y}\big]\Big)^{-1} \qquad \boldsymbol{\Psi}^\star = \frac{1}{n}\text{diag}\Big(\mathbf{y}\mathbf{y}^\top - \boldsymbol{\Lambda}^\star \mathbb{E}_{\mathbf{z}|\mathbf{y}}\big[\mathbf{z} \mid \mathbf{y}\big]\mathbf{y}^\top\Big). \tag{2}$$

The posterior embeddings of $\mathbf{z}$ are sampled from a normal distribution with moments $\mathbb{E}[\mathbf{z} \mid \mathbf{x}] = \boldsymbol{\beta}\mathbf{x}$ and $\mathbb{E}[\mathbf{z}\mathbf{z}^\top \mid \mathbf{x}] = I - \boldsymbol{\beta}\boldsymbol{\Lambda} + \boldsymbol{\beta}\mathbf{x}\mathbf{x}^\top\boldsymbol{\beta}^\top$ where $\boldsymbol{\beta} = \boldsymbol{\Lambda}^\top(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi})^{-1}$. Using these estimated moments, we sample $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\Lambda}^\star \mathbf{z}; \boldsymbol{\Psi}^\star)$ using the reparameterization trick similar to [19], and decode each embedding using a modality-specific decoder $D^j(\cdot; \boldsymbol{\theta}^j_{\text{dec}})$. Finally, let $\mathcal{L}$ be the loss and $\boldsymbol{\Theta}$ be both the neural network and PCCA parameters. We optimize $\nabla_{\boldsymbol{\Theta}} \mathcal{L}$ where

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^n \|\hat{\mathbf{x}}_i^a - \mathbf{x}_i^a\|_2^2 + \frac{1}{n}\sum_{i=1}^n \|\hat{\mathbf{x}}_i^b - \mathbf{x}_i^b\|_2^2 + \gamma\big(\|\boldsymbol{\Lambda}^b\|_1 + \|\boldsymbol{\Lambda}^{bc}\|_1 + \|\boldsymbol{\theta}^b_{\text{dec}}\|_1\big) \tag{3}$$

and $\gamma \geq 0$ is a tuning parameter for the Laplacian prior. When we say that the model is trained *end-to-end*, we mean that this is the single loss that is optimized through backpropagation, while the PCCA parameter estimates act as implicit variational regularization.

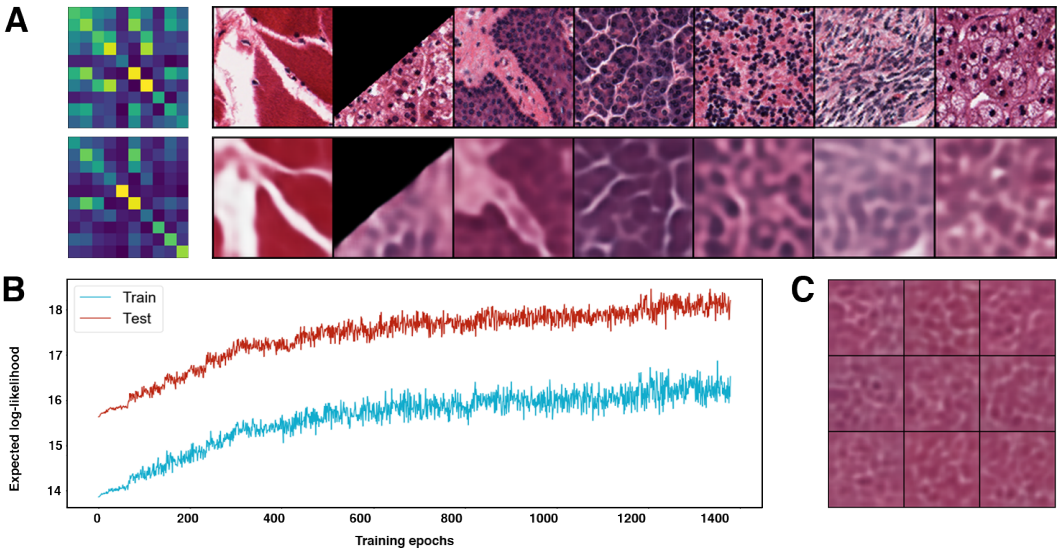

Figure 2: **(A)** Examples of original data (top row) and reconstructions (bottom row) of gene expression covariance matrices (left) and histology slides (right). For clarity, we show the top 10 least sparse columns of the covariance matrix. **(B)** Expected log-likelihood on train (blue) and test (red) data. For clarity since the log-likelihood is a function of sample size, we take the log of both series. **(C)** Generated image samples.

3 Experiments

We trained our model on 2,221 samples from GTEx V6 [8; 7]. Each Aperio whole tissue slide was subsampled once down to a 1000×1000 pixel RGB image. The data set is small and imbalanced; twelve tissues have fewer than 50 samples, and only one tissue has more than 200 samples. To augment the data, the model was trained on 128×128 pixel crops with random rotations and reflections. The image autoencoder is based on [16], and the gene autoencoder is two linear layers.

Before using our model for biomedical analysis, we wanted verify two important properties. First, we wanted to show that our model could reconstruct both modalities from a shared latent variable. To show this, we saved reconstructions of the images and heatmaps of the gene covariance matrices during training. We found that our model was able to reconstruct both modalities (Figure 2A). This suggests that a shared latent variable carries enough information to recapitulate both modalities.

Second, we wanted to verify our variational take on PCCA. With a composite model of neural networks and PCCA, we might ask whether one of the submodules is being ignored due to an imbalance in the numbers of parameters. This was a concern since image samples from the generative model are fairly uniform (Figure 2C). To test this, we computed the expected log-likelihood over training and found that it increases over time (Figure 2B). Taken together, these results suggest that the neural networks and PCCA are jointly learning parameters for embedding and reconstructing data from nonlinear observations while maximizing the log-likelihood of the generative model.

4 Discussion

We describe a method that extends probabilistic CCA with end-to-end training of deep neural networks for joint modeling of high-dimensional paired biomedical data. We plan to address the uniformity of the image samples in a few ways. First, we are currently processing GTEx V8, which has over seven times as many samples. We expect this 7-fold increase in data to substantially improve the generative model. Second, a heavy-tailed prior on $\mathbf{\Psi}$ may prevent the variance from collapsing to small values. Finally, architectural changes such as pretrained convolutional layers or a nonlinear gene encoder

may improve the quality of the embeddings. Our model's latent space can be quantified against given tissue labels by external cluster analysis techniques such as the Rand Index [25], and this would let us benchmark our model against other methods. Our scientific aim is to use this generative model on a number of downstream tasks such as annotating histology slides with predicted gene expression levels and analyzing clusters that capture shared structure in our paired data.

References

[1] S. Akaho. A kernel method for canonical correlation analysis. *arXiv preprint cs/0609071*, 2006.

[2] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255, 2013.

[3] J. Ash, G. Darnell, D. Munro, and B. Engelhardt. Joint analysis of gene expression levels and histological images identifies genes associated with tissue morphology. *bioRxiv*, 2018.

[4] F. R. Bach and M. I. Jordan. A probabilistic interpretation of canonical correlation analysis. 2005.

[5] Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan. Chest pathology detection using deep learning with non-medical training. In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, pages 294–297. IEEE, 2015.

[6] J. D. Barry, M. Fagny, J. N. Paulson, H. J. Aerts, J. Platig, and J. Quackenbush. Histopathological image qtl discovery of immune infiltration variants. *iScience*, 5:80–89, 2018.

[7] L. J. Carithers, K. Ardlie, M. Barcus, P. A. Branton, A. Britton, S. A. Buia, C. C. Compton, D. S. DeLuca, J. Peter-Demchok, E. T. Gelfand, et al. A novel approach to high-quality postmortem tissue procurement: the gtex project. *Biopreservation and biobanking*, 13(5):311–319, 2015.

[8] G. Consortium et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204, 2017.

[9] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

[10] K. J. Geras, S. Wolfson, S. Kim, L. Moy, and K. Cho. High-resolution breast cancer screening with multi-view deep convolutional neural networks. *arXiv preprint arXiv:1703.07047*, 2017.

[11] Z. Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452, 2015.

[12] Z. Ghahramani, G. E. Hinton, et al. The em algorithm for mixtures of factor analyzers. Technical report, Technical Report CRG-TR-96-1, University of Toronto, 1996.

[13] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

[14] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[15] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[16] X. Hou, L. Shen, K. Sun, and G. Qiu. Deep feature consistent variational autoencoder. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 1133–1141. IEEE, 2017.

[17] T. Jendoubi and K. Strimmer. Probabilistic canonical correlation analysis: A whitening approach. *arXiv preprint arXiv:1802.03490*, 2018.

[18] M. Johnson, D. K. Duvenaud, A. Wiltschko, R. P. Adams, and S. R. Datta. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in neural information processing systems*, pages 2946–2954, 2016.

[19] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[20] A. Klami, S. Virtanen, and S. Kaski. Bayesian canonical correlation analysis. *Journal of Machine Learning Research*, 14(Apr):965–1003, 2013.

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[22] W. Lin, N. Hubacher, and M. E. Khan. Variational message passing with structured inference networks. *arXiv preprint arXiv:1803.05589*, 2018.

[23] Z. C. Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

[24] K. P. Murphy. *Machine learning: a probabilistic perspective*. Cambridge, MA, 2012.

[25] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[26] M. Shah, D. Wang, C. Rubadue, D. Suster, and A. Beck. Deep learning assessment of tumor proliferation in breast cancer histological images. In *Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on*, pages 600–603. IEEE, 2017.

[27] W. Wang, X. Yan, H. Lee, and K. Livescu. Deep variational canonical correlation analysis. *arXiv preprint arXiv:1610.03454*, 2016.