Exchangeable Variational Autoencoders with Applications to Genomic Data

Jeffrey Chan*
UC Berkeley

Jeffrey P. Spence*
Stanford University

Yun S. Song
UC Berkeley

Abstract

Exchangeable-structured datapoints (datapoints which contain permutation-invariant symmetries) are ubiquitous in statistical problems ranging from point clouds to graphs to sets. Particularly in biological settings, where multiple experiments derived from a noisy scientific process attempt to measure a latent variable of interest, experimental datapoints are often exchangeable-structured demanding the development of methods which can exploit this structure. Modern machine learning approaches to scalable Bayesian inference typically use autoencoding variational Bayes – marrying ideas from deep learning and probabilistic modeling to achieve practical inference for expressive models. Current VAE-based approaches do not naturally handle exchangeable (but non-i.i.d.) datapoints. Often exchangeable-structured datapoints may contain heterogeneity in datapoint dimensions precluding a staightforward application of the vanilla VAE framework. In this work, we develop the Exchangeable Variational Autoencoder which provides inferential and computational benefits while enabling varying set size data to be robustly handled in the VAE framework. We then demonstrate its efficacy in two settings: (1) on the well-studied Latent Dirichlet Allocation model and (2) on the bootstrapped, isoform-level uncertainty estimates of single-cell RNA-seq data.

1 Introduction

Exchangeable-structured datapoints are a data object which can be found across many real-world applications including sets, graphs, and point clouds. Genomics – and biology in general – showcase many instances of exchangeable-structured (but not i.i.d.) datapoints where each datapoint contains multiple noisy measurements to more accurately infer an underlying biological process. Traditionally, a popular approach to directly inferring the underlying biological process is by explicitly incorporating the structural assumptions of exchangeability between the measurements into a probabilistic model, then performing inference. However, with the advent of high-throughput datasets in genomics, the demand for scalable, yet flexible inference methods has become increasingly necessary rendering many traditional inference techniques for such probabilistic models too computationally inefficient.

Modern machine learning methods over the past decade have had massive success is scaling inference to large datasets. Furthermore, variational autoencoders (VAEs) interweave scalability ideas from machine learning while enabling more flexible probabilistic models to allow for flexible and scalable Bayesian inference – perfect for tackling genomics problems. However, a common underlying assumption to most success stories in machine learning require that the data is i.i.d. In this work, we develop an Exchangeable VAE which directly handles exchangeable-structured data via a permutation-invariant encoder to allow for the direct inference of the de Finetti measure. This method naturally provides many computational and inferential benefits over its nonexchangeable counterpart. More importantly, it allows us to directly perform inference in settings where the dimension of each datapoint varies, which cannot be trivially handled via a nonexchangeable VAE. We demonstrate the efficacy of our method across two settings: (1) the Latent Dirichlet Allocation model which provides

Preprint. Under review.

us with a fixed probabilistic model with no global parameters to optimize allowing us to directly assess the capabilities of our permutation-invariant encoder and (2) single-cell RNA-seq data we improve upon pre-existing VAE-based approaches [1] to single-cell data by accounting for inferential uncertainty in the read-alignment process via bootstrap samples, which has allowed for more precise, isoform-level analyses in the bulk RNA-seq setting.

2 Exchangeable Variational Autoencoder

The exchangeable variational autoencoder directly incorporates the structural symmetries of exchangeable-structured datapoints into its neural network architecture improving statistical and computational efficiency over the vanilla VAE. Furthermore, it enables the sharing of statistical strength across exchangeable-structured datapoints of a variable number of exchangeable elements in the same latent space which is often useful in many setups including missing data regimes.

2.1 Exchangability and de Finetti's Theorem

Exchangeability is defined on a sequence of random variables $x_1, x_2, \ldots, x_n$ that satisfy $p(x_1, \ldots x_n) = p(x_{\sigma(1)}, \ldots x_{\sigma(n)})$ for all permutations σ. De Finetti's theorem states a sequence is exchangeable if and only if, for all n, it satisfies

$$p(x_1 \ldots, x_n) = \int_\Theta \prod_{i=1}^{n} p(x_i|\theta)p(d\theta).$$

In other words, de Finetti's theorem tells us that conditioned on the de Finetti measure, we are guaranteed that our data are drawn i.i.d. Thus, informed by de Finetti, we seek to develop a neural network architecture that maps from the data to a probability distribution over θ.

2.2 Permutation-Invariant Encoder

Let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^T \in \mathbb{R}^{n \times d}$ be a matrix and S_n denote the symmetric group. We wish to construct a neural network layer Φ that is *invariant* with respect to all row-permutations, that is, for every $g \in S_n$,

$$\Phi(\mathbf{X}) = \Phi(g \cdot \mathbf{X}).$$

We use parameter sharing to encode permutation invariance for computational efficiency and to prevent a combinatorial explosion of parameters. Our proposed network is defined as

$$\Phi(\mathbf{X}) = (h \circ g)(f(x_1), \ldots f(x_n))$$

where $f : \mathbb{R}^d \to \mathbb{R}^{d_1}$ is parameterized by a neural network, $g : \mathbb{R}^{n \times d_1} \to \mathbb{R}^{d_2}$ is any symmetric function such as the max, sum, or higher-order moments, and $h : \mathbb{R}^{d_2} \to \mathbb{R}^Z$ is another neural network that maps to the latent space as shown in Figure 2.2. This is equivalent to the permutation invariant layers proposed by [2] and [3] which they show to be able to approximate any permutation-invariant function. The output of the encoder defines a latent probability measure which can be transformed via a neural network to infer the de Finetti measure from which we can draw i.i.d. samples in the generative model to produce exchangeable data.

3 Toy Example: Exchangeable LDA-VAE

First, we illustrate the properties of the exchangeable VAE and its improvements over nonexchangeable architectures on a simple model without any neural networks in its generative model, Latent Dirichlet Allocation(LDA) [4, 5]. By illustrating its performance on a simpler parametric probabilistic model, we can directly evaluate the performance of the permutation-invariant encoder. For a review of the details of the Latent Dirichlet Allocation model, see Appendix A. Under the LDA model, the words within each document are exchangeable and are conditionally iid given the topic distribution and word distribution for each topic. Note that typically when performing VAE-style inference procedures on the LDA model as done in [6, 7], one can simply collapse the words in a document into a lower-dimensional sufficient statistic, a histogram over the dictionary of words reflecting which words are used in each document – thereby, destroying the exchangeable structure. For the purposes

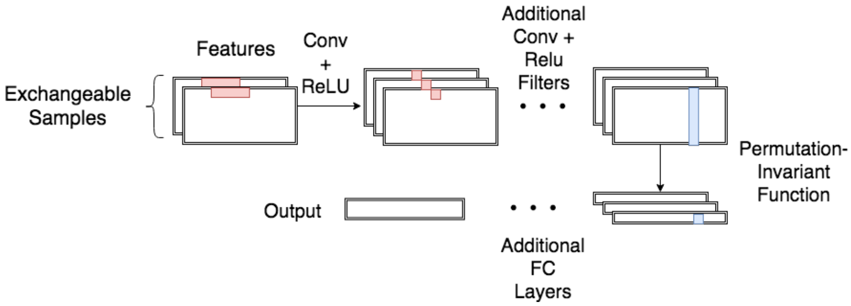

Figure 1: A schematic of the Permutation-Invariant Encoder

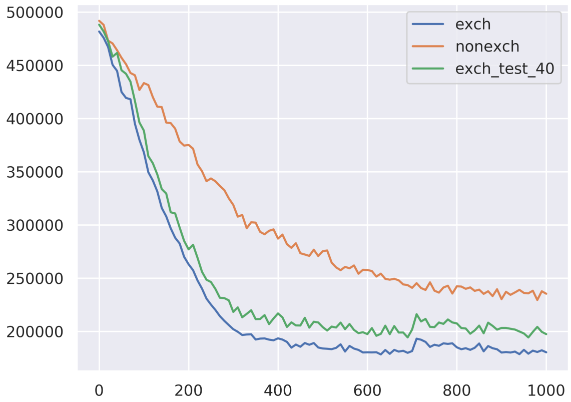

Figure 2: The test log-likelihood of the data under 4 conditions: (1) Nonexchangeable architecture trained and tested on 64-length documents, (2) exchangeable architecture trained and tested on 64-length documents, (3) exchangeable architectured trained on 64-length, tested on 40-length, (4) exchangeable architecture trained on 10-length architecture.

of demonstrating the efficacy of our method, we choose to maintain the words in the document as a one-hot encoding over the dictionary for each word. We can infer the global variables via conjugacy and use stochastic VI on the local variables. The permutation-invariant encoder takes in the one-hot encoded set of words and outputs a reparameterized Dirichlet posterior distribution for the topics corresponding to a given document as done in [8]. Note that this enables amortized inference that can flexibly handle varying number of words per document.

Through our experiments on data simulated from the LDA model, we verify four properties of the Exchangeable VAE over its nonexchangeable counterpart as shown in Figure 3: (1) reduced per-batch compute time, (2) faster convergence to the minima, (3) smaller test negative log-likelihood, and (4) robustness to changes in set sizes. As we can see, the exchangeable VAE demonstrates clear benefits over the nonexchangeable version.

4 Bootstrap Isoform scVI

Single-cell RNA-seq is a sequencing technology that is often used to infer the underlying expression level of each gene within a given cell. While the underlying gene expression levels cannot be directly measured, noisy read data (short chunks of sequence) are typically aligned to a genomic sequence and converted into noisy expression count data. With this data in hand, the goal is often to infer the underlying expression levels and characterize the underlying biological variation due to cell type, cell state, or experimental conditions. Recent success has been achieved by scaling inference of the underlying expression levels to a large number of cells using a VAE [1]. However, many recent approaches do not account for the uncertainty due to aligning the reads to the underlying genomic

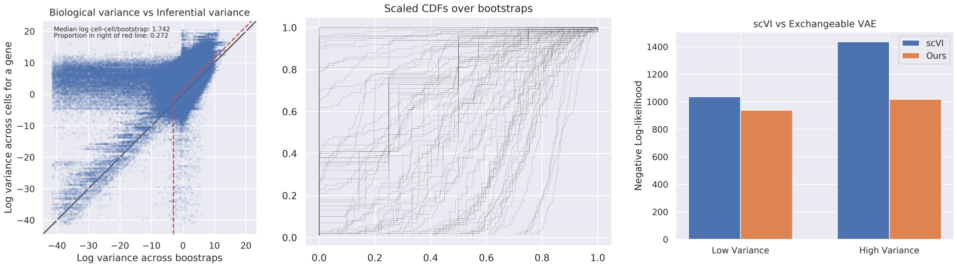

Figure 3: (Left) We have a plot of biological variance (variance over cells) against inferential variance (variance over bootstraps) for each cell-gene pair. Those pairs to the right of the red dotted-line indicate cell-gene pairs from which accounting for inferential variance would be useful. (Middle) The scaled CDFs of each cell-gene pair over the bootstraps. This motivates the use of Gaussin-noise and per-bootstrap zero-inflation. (Right) The negative log-likelihood results on simulated data from SymSim comparing scVI vs our method.

sequence nor the uncertainty of the count data produced via EM or Gibbs-style algorithms as is known to be a problem in bulk RNA-seq [9].

Our contributions with applying the exchangeable VAE to single-cell data are two-fold. First, we account for inferential variation due to the sequencing process to further denoise biological variation from technical variation. Roughly, 27% of cell-gene pairs have inferential variation that is on the same order of magnitude or larger than that of the biological variation across cells as shown in Figure 3 motivating directly accounting for inferential variation in the probabilistic model as it could wash out the biological signal. The uncertainty is measured via bootstrap samples over the original read data. Second, we analyze expression at the isoform (or transcript)-level rather than at the gene-level as is commonly done in the bulk RNA-seq setting allowing for deeper biological insight into the underlying expression mechanisms. There have been a few barriers preventing the adoption of isoform-specific analysis of single-cell data. Droplet-based single-cell technologies are typically targeted to the edges of the gene leading to the lack of reads which recover splicing events. To circumvent this, we focus on transcript-based technologies such as SmartSeq2. Another issue may be that there simply aren't enough reads for a given cell that span a splicing event. To address this, we choose a dataset with a large number of cells to allow for the sharing of statistical strength across cells.

4.1 Probabilistic Model

We augment the scVI probabilistic model (shown in Appendix B) to account for inferential uncertainty related to the sequencing process via bootstraps generated by `kallisto`[9]. We instead observed bootstrapped expression level $x_{n,g,b}$ for cell n, *isoforms* (not genes) g, and bootstrap b with $u_{n,g}$ now as a latent variable.

$$
\begin{aligned}
\mathbf{a}_n &\sim \mathrm{Normal}(0, I_A) \\
\epsilon_{n,g,b} &\sim \mathrm{Normal}(0, f_\sigma^g(\mathbf{a}_n)) \\
h'_{n,g,b} &\sim \mathrm{Bernoulli}(f_{h'}^g(\mathbf{a}_n)) \\
x_{n,g,b} &= \begin{cases} u_{n,g} + \epsilon_{n,g,b} & h'_{n,g,b} = 0 \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

The bootstrapped version of the model includes adding zero-mean Gaussian noise to the unobserved counts given infinite sequencing depth to model the underlying inferential variance resulting from the sequencing process. In addition, we incorporate zero-inflation at the per-bootstrap level in the scenario that the small number of reads mapped to an isoform gets resampled out. The latter modeling choice needs some justification which we show in Figure 3(b). In this figure, we scale the CDF of counts over bootstraps for each cell-gene pair to lie between 0 and 1. We can see from the CDFs that most of the mass lie at the mode which generally occurs in the median motivating a Gaussian noise model. However, due to the resampling of reads, we notice that a proportion of the CDFs have zero-inflation, so we add bootstrap-specific zero-inflation to our model.

4.2 Inference

While the standard training procedure for variational autoencoders is via black-box variational inference, the reconstruction error for the evidence lower bound (ELBO) of our model cannot be analytically expressed since this requires the computing the convolution of a negative binomial and a Gaussian. We choose to approximate the integral by computing the probability of the negative binomial taking on the values in a grid of points up to 10000 and the corresponding $\epsilon_{n,g,b}$ values. This approximation is carried out as a tensor operation which does not have a large effect on performance. Another technical issue with respect to the model is that at test time our generative model can output expression levels that are negative since $\epsilon_{n,g,b}$ can be negative. We found that this is not really an issue in practice and choose to simply set those values to 0.

4.3 Experiments

Our method was evaluated under two regimes: (1) a simulated dataset of 5 cell types simulated via `SymSim` [10] where the true underlying expression counts and cell types is known as is shown on the right panel of Figure 2 and (2) 5000 cells of SmartSeq2 data from marrow tissue derived from the Tabula Muris dataset [11]. We evaluate the usefulness of incorporating bootstraps by comparing the inferred value of $u_{n,g}$ in comparison to that of `scVI` and demonstrating the calibration of inferential uncertainty. The real dataset poses challenges due to the known amplification issues of `SmartSeq2` data in comparison to UMI-based methods.

References

[1] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053, 2018.

[2] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in neural information processing systems*, pages 3391–3401, 2017.

[3] Jeffrey Chan, Valerio Perrone, Jeffrey Spence, Paul Jenkins, Sara Mathieson, and Yun Song. A likelihood-free inference framework for population genetic data using exchangeable neural networks. In *Advances in Neural Information Processing Systems*, pages 8594–8605, 2018.

[4] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[6] Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*, 2017.

[7] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep Universal Probabilistic Programming. *arXiv preprint arXiv:1810.09538*, 2018.

[8] Martin Jankowiak and Fritz Obermeyer. Pathwise derivatives beyond the reparameterization trick. *arXiv preprint arXiv:1806.01851*, 2018.

[9] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic rna-seq quantification. *Nature biotechnology*, 34(5):525, 2016.

[10] Xiuwei Zhang, Chenling Xu, and Nir Yosef. Simulating multiple faceted variability in single cell rna sequencing. *Nature communications*, 10(1):2611, 2019.

[11] Nicholas Schaum, Jim Karkanias, Norma F Neff, Andrew P May, Stephen R Quake, Tony Wyss-Coray, Spyros Darmanis, Joshua Batson, Olga Botvinnik, Michelle B Chen, et al. Single-cell transcriptomics of 20 mouse organs creates a tabula muris: The tabula muris consortium. *Nature*, 562(7727):367, 2018.

A Latent Dirichlet Allocation Model

Briefly, the LDA model can be written as

$$\begin{aligned}\theta_i &\sim \text{Dirichlet}(\alpha) \text{ for } i \in \{1 \dots M\}\\ \phi_k &\sim \text{Dirichlet}(\beta) \text{ for } i \in \{1 \dots K\}\\ z_{i,j} &\sim \text{Multinomial}(\theta_i)\\ w_{i,j} &\sim \text{Multinomial}(\phi_{z_{i,j}})\end{aligned}$$

where $i \in \{1 \dots M\}$ and $j \in \{1, \dots, N_i\}$ denotes the document and the document position. θ_i is the topic distribution, and ϕ_k denotes the word distribution for each topic k.

B scVI Model

The probabilistic model for scVI [1] is a zero-inflated negative binomial(ZINB) model with neural networks parameterizing the zero-inflation dropout rate and the parameters of the negative binomial. In addition, the ZINB is scaled by the library size. It is written as follows for observed expression level $u_{n,g}$ cell n and gene g:

$$\begin{aligned}z_n &\sim \text{Normal}(0, I_Z)\\ l_n &\sim \text{LogNormal}(l_\mu, l_{\sigma^2})\\ w_{n,g} &\sim \text{Gamma}(f_\mu^g(z_n, s_n), f_\theta^g(z_n, s_n))\\ y_{n,g} &\sim \text{Poisson}(l_n w_{n,g})\\ h_{n,g} &\sim \text{Bernoulli}(f_h^g(z_n, s_n))\\ u_{n,g} &= \begin{cases} y_{n,g} & h_{n,g} = 0\\ 0 & \text{otherwise}\end{cases}\end{aligned}$$

where z_n and l_n are typically 10- and 1-dimensional latent variables of the VAE. $y_{n,g}$ is drawn from a negative binomial, and $h_{n,g}$ controls the zero-inflation. The functions f are parameterized by fully-connected neural networks.