Scalable and Checkable Models for Multiomics Data

Matthew Broerman
Department of Statistics
University of Pittsburgh
Pittsburgh, PA 15213
mjb305@pitt.edu

Abstract

Recent probabilistic models take advantage of the performance of approximate inference for single-cell analysis. After some discussion of new forms of data, models, and their algorithms, we test whether single-cell Hierarchical Poisson Factorization is suited to non-RNA-seq data. Using posterior predictive checks, we find that the model is not.

1 Introduction

Advances in the last 15 years in single-cell analysis have enabled the vast collection of complementary datasets that capture several dimensions of cellular state at unprecedented resolution. Many of these datasets consist of count data where each entry in a matrix represents the number of reads of a sequence that uniquely corresponds to a feature of a single cell, and where larger counts correspond to greater prevalence in the cell. The earliest single-cell datasets were single-cell RNA sequence (scRNA-seq), where the feature is directly the mRNA transcription sequences of a gene []. More recently, Stoeckius et al. [2017] has developed Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq) where cells are first stained with a library of engineered antibodies that bind to surface proteins. Since these antibodies have unique oligonucleotide tails whose sequencing is compatible with existing single-cell sequencing methods, they could count the prevalence of both mRNA sequences and some surface proteins within a single pass in the traditional sequencing workflows.

Stoeckius et al. [2017] have shown that this method of protein detection is consistent with data from flow cytometry. But it is more flexible since the throughput of flow cytometry is limited both by the available library of antibody-bound fluorophores which fluoresces laser light, and the capacity of sensors to distinguish between fluorescent spectra. Beside epitopic data, CITE-seq can be used to tag subpopulations of cells, and so enables multiplex experimental data within a single lane of existing single-cell sequencing technology, as well as the detection of multiplets in sequencing droplets that have so far limited the loading of cells per run of microfluidic sequencing systems. This innovation and others such as ATAC-seq generate multimodal datasets that have redundant information, and they are orders of magnitude larger than earlier datasets though many are also very sparse. Matrix factorization is a dimension reduction method to consolidate this information and project it into a smaller and more manageable subspace for further analysis. In this paper, we use a probabilistic matrix factorization model called single-cell Hierarchical Poisson Factorization (scHPF) that is especially well-suited to sparse count matrices.

Over roughly the same period, advances in approximate probabilistic models have enabled the estimation of more complicated models with checks for accuracy and reliability. One subset, called a hierarchical models, are built up out of mixtures of probability distributions where the parameter for one distribution, say the proportion of a binomial distribution, is itself modeled as a draw from another distribution, say a beta distribution. The advantage of these probabilistic models over non-probabilistic models is at least twofold. First, most models already specify distributions for

Preprint. Under review.

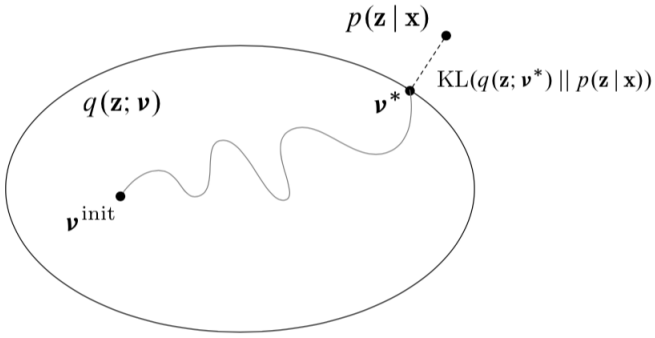


Figure 1: Representation of Variational Inference. Blei et al. [2017]

$$q^*(\mathbf{z}) = \arg\min_{q(\mathbf{z}) \in \mathcal{Q}} \mathrm{KL}\left(q(\mathbf{z}) \| p(\mathbf{z} \mid \mathbf{x})\right).$$

$$\mathrm{KL}(q(\mathbf{z}) \| p(\mathbf{z} \mid \mathbf{x})) = \mathbb{E}\left[\log q(\mathbf{z})\right] - \mathbb{E}\left[\log p(\mathbf{z} \mid \mathbf{x})\right],$$

$$\mathrm{KL}\left(q(\mathbf{z}) \| p(\mathbf{z} \mid \mathbf{x})\right) = \mathbb{E}\left[\log q(\mathbf{z})\right] - \mathbb{E}\left[\log p(\mathbf{z}, \mathbf{x})\right] + \log p(\mathbf{x}).$$

Figure 2: Versions of the objective. Blei et al. [2017]

parameters, only they do this covertly as assumptions that are not available for criticism. By contrast, using probability distributions to model phenomena directly allows us to ask whether each component of the overall model is reasonable and properly specified. Second, since training the model is learning parameters of distributions (sometimes called posterior inference), the trained model is a generative model that can simulate data. Thus we can check whether the simulated data matches the distribution of actual data. These posterior predictive checks (PPCs) again enable a kind of model criticism that is not available to non-probabilistic models. As a demonstration of their utility, we will first use PPCs to first assess whether scHPF is a reasonable probabilistic model of scRNA-seq data of about 12,000 protein coding genes from nearly 6,000 Peripheral Blood Mononuclear Cells (PBMC). Second, we will assess whether the count of cell-surface protein tags generated by a version of CITE-seq can be reasonably modeled by the same probabilistic model that we use to model to the scRNA-seq count data.

2 Methods

2.1 Variational Inference

Although there have been improvements in the performance of Markov Chain Monte Carlo (MCMC), it can still be slow on large and complicated datasets such as single-cell data. An alternative is variational inference (VI). A more tractable family of approximations of the model, the variational family $\mathcal{D}$ consisting of candidate densities $q(\cdot)$ with parameters $\mathbf{z}$, is substituted for the actual distribution, $p(z|x)$, in the training task. The difference of a candidate density from the target true distribution is measured by the Evidence Lower-bound (ELBO) which itself is a tractable approximation, up to an intractable constant $\log p(x)$, of the Kullback–Leibler divergence.

The parameters for the next candidate variational density are varied according to coordinate ascent optimization by optimizing one parameter in turn while holding all other parameters fixed. This requires that we can factor the approximate density into a form such that each parameter belongs to a density that is conditionally independent of the other parameters z as in (15) and (16). This form is called the complete conditional. For each update step of the coordinate ascent algorithm in (17), we

$$q(\mathbf{z}) = \prod_{j=1}^{m} q_j(z_j).$$

$$q_j^*(z_j) \propto \exp\left\{\mathbb{E}_{-j}\left[\log p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})\right]\right\}.$$

Figure 3: Complete Condition and Algorithm Update. Blei et al. [2017]

set the new variable parameter to the exponentiated expectation over the fixed parameters of the log probability of the old variable parameter given the other fix parameters and the data.[1]

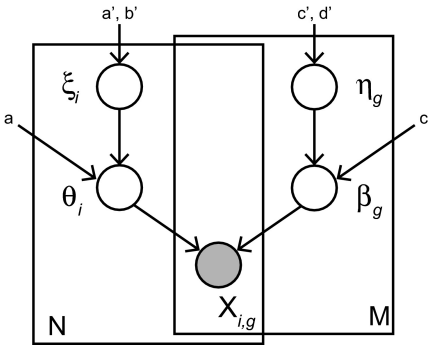
One challenge of VI is that the choice of variational method is not independent of model we wish to approximate, since models differ in their possible variational family. The model we used was first developed by Gopalan et al. [2014] and then refined for application to single-cell data by Levitin et al. [2019]. This factorization is especially suited to sparse count matrices and has an efficient variational family and method called mean-field approximation. But since VI effectively recasts probabilistic inference as an optimization task, the promise of VI for single-cell data rests on the advantage it can take of developments in stochastic optimization for more complex models that do not themselves have a correspondingly efficient variational families and methods Lopez et al. [2018].

2.2 Single-cell Hierarchical Poisson Factorization

In the paradigmatic matrix factorization problem, each entry in a matrix of n users and m items is a count x of a particular item i for a particular user u–this may be views of a movie for Alice, or purchases of a product for Bob. The count can be decomposed into a sum of k weighted scores, or the inner product for k factors. The k weights might be understood as the intensity of k preferences for each user, while the k scores might be the amount of some feature or genre for each item, but we emphasize that the k factors need not correspond to any such features or preferences. As a result, we can see the n,m matrix of counts $x_{i,j}$ as the matrix product of two lower dimension factor matrices, (n, k) and (k, m).

In hierarchical Poisson factorization, each count is treated as a draw from a Poisson distribution whose rate parameter λ_i is itself drawn from a gamma distribution. As in traditional matrix factorization, this parameter is the inner product of k user and item factors, but each of the k user "preference" factors is characterized as an allocation from a gamma distributed budget (likewise for the k item "genre" factors). Intuitively this makes sense in cases where factors are constrained like we might imagine preferences and genres to be. That is, a strong preference in one component might sideline preferences elsewhere, just as a movie that emphasizes one genre might be precluded from other genres. Moreover, treating factors as allocations from budgets explicitly models the source of sparsity in count data, since budgets over many factors will necessarily include several small weights, and these in turn contribute to small rate parameters for Poisson distributions. Finally, hierarchical Poisson distributions with variable rates drawn from gamma distribution are marginalized as negative binomials which is a common choice for overdispersed count data.

Figure 4: Fig. 2. Probabilistic Graphical Model



In scHPF, we apply this hierarchical framework to scRNA-seq data: mRNA counts are drawn from Poisson distributions with variable rate parameters, while cell and gene factors are allocations from their respective gamma distributed budgets. Each set of cell and gene budgets, in turn, are themselves 'budgeted' by an additional scale parameter that is drawn from its own gamma distributed 'capacity.'

[1] My intuition is that we are trying to stretch a sheet over a lumpy bed single-handedly. By moving one corner of the sheet while the other parts are stuck in place, we both find an optimal position for the corner we moving at this step, and we set a new fixed position by which to adjust the next corner. We have to loop through the corners several times to get the overall optimal fit.

An important question is whether the prevalence of tagged cell-surface proteins captured in CITE-seq can be modelled by scHPF. On the one hand, cell-surface proteins share features with mRNA: they serve as an important mechanisms of cell-differentiation, and they are presumably metabolically 'budgeted' at least as much as mRNA. So we might think that their prevalence is a weighted sum of latent budgets analogous to scRNA-seq data. On the other hand, recent studies have shown that the frequency of some mRNA and coordinated cell-surface proteins are out of joint. For example, when cells are treated with phorbol myristate acetate (PMA) and their levels of RNA PD-LI and protein FACS MRI are measured at common intervals post-treatment, the RNA peaks and declines while the proteins levels continue to climb Burr et al. [2017].

2.3 Posterior Predictive Checks

The strength of probabilistic modeling is that, however indispensable domain knowledge is guiding model formulation, simulations from our trained model must confront the distribution of the observed data. PPCs begin by drawing a set of parameters from the posterior density of parameters θ given data x. These θ_i then parameterized the data generating distribution of $x_i \sim p(\hat{x}|\theta_i)$ Gelman et al. [2013]. Taken together, we have the posterior predictive distribution:

In order to compare the actual and simulated data, we construct a relevant test statistic that summarizes the information in both distributions. For example, we can ask what is the probability that the maximum value of the simulated distribution is greater than the simulated value in the observed data: $Pr(T_{max}(x_{sim}) > T_{max}(x_{obs}))$.

2.4 Data

We used a 10X Genomics dataset of 5k Peripheral blood mononuclear cells from a healthy donor with cell surface proteins. We converted these to a loom file with the loompy package. We then segmented the dataset into RNA-seq and cell-surface protein ($n = 32$) datasets. We filtered the RNA-seq dataset by a whitelist for protein coding genes that was provided by the authors of the scHPF package. scHPF does is free from any other correction or normalization steps, which recent studies suggest has the biggest impact on single-cell analysis[]

3 Results

3.1 Selecting and Scoring K

We trained with factors $k = 7, 8, 9$ for the RNA-seq dataset and found that we got the best separation with $k = 7$. We judged separation by counting the greatest pairwise overlaps in top scores for genes and cells for each factor. Following Levitin et al. [2019], each gene and cell's score for a factor k is the expected values of its factor loading bg, k or hi, k scaled by its respective its capacity $\eta_{g,k}$ or $\xi_{i,k}$, respectively:

3.2 Posterior Predictive Checks

For PPCs, we took 10 samples of the learned parameters theta and beta for each factor k, resulting in two arrays with dimensions $(ncells, 7factors, 10samples)$ and $(mgenes, 7factors, 10samples)$. These matrices were multiplied to form an $nxmx10$ matrix of Poisson rate parameters. We drew a Poisson variable for each rate parameter and take the average these across samples. Since we were interested in the dispersion across Poisson variables, we use the Kolmogorov-Smirnov test, a non-parametric test of similarity between cumulative distribution functions, as our test statistics of the samples from Poisson distributions. We report our results in Table 1. Tests on both mRNA and Protein return a very small and presumably significant p-value, though there is an order of magnitude difference in the statistics.

$$\text{cell\_score}_{i,k} = E[\xi_i|\boldsymbol{x}] * E[\theta_{i,k}|\boldsymbol{x}]$$

$$\text{gene\_score}_{i,k} = E[\eta_g|\boldsymbol{x}] * E[\beta_{g,k}|\boldsymbol{x}].$$

Figure 5: Cell and gene scores

In addition, to get an overall sense of the distribution of the matrix, we calculated the coefficient of variation for each row and column of the n x m matrix: $std(x_{i,.})/mean(x_{i,.})$ and $std(x_{.,g})/mean(x_{.,g})$

Table 1: Goodness of Fit

Part		
Name	KS-stat	P-value
RNA-seq	0.00432	0.0
Protein	0.07794	0.0

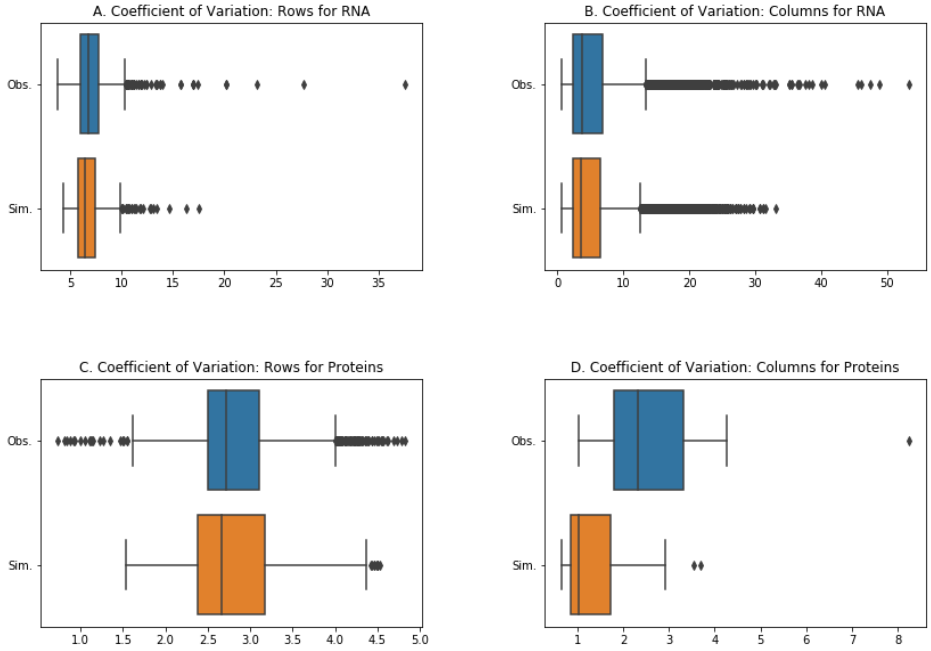


Figure 6: Observed and Simulated Compared

(Fig. 6a-d). In the case of RNA-seq data, we see that the rowwise and columnwise variation closely resembles those of our data. But we do see that our model still underestimates the variation and so also the dispersion. This would be a good reason to change the scale hyperparameter $a\prime$ of the 'capacity' parameter, which indirectly influences dispersion of x.

The box plots of the protein data tell a different story. Although the KS-statistic was strong, we can see clear differences between the observed and simulated protein data. The observed row coefficients are much more dispersed than the simulated ones, and the simulated column coefficients are much more right skewed than the observed. This suggests that scHPF is not a good choice of model for cell-surface protein counts. Moreover, the KS-statistic may report false-significance and is probably not a good test statistic in this application.

4 Conclusions and further work

We have replicated one state-of-the-art machine learning model and an algorithm for probabilistic modelling on a new dataset. We tested scHPF in a novel application by implementing posterior predictive checks with test-statistics and graphical displays. Although there was reason to think scHPF might perform well on counts of cell-surface proteins, it does not seem to, and this casts doubt on whether it handle other modalities of single-cell data. It must also be asked whether such 'bespoke' models are useful in general practice, or whether they should be reserved only for otherwise intractable modeling problems. In future work, we could explore more accurate alternatives to the KS-statistic, and in the case of proteins, we could adjust the hyperparameters in order to force less sparsity on the protein counts since they are in general larger.

References

Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K. Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, September 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4380. URL https://www.nature.com/articles/nmeth.4380.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, April 2017. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2017.1285773. URL http://arxiv.org/abs/1601.00670. arXiv: 1601.00670.

Prem Gopalan, Jake M. Hofman, and David M. Blei. Scalable Recommendation with Poisson Factorization. *arXiv:1311.1704 [cs, stat]*, May 2014. URL http://arxiv.org/abs/1311.1704. arXiv: 1311.1704.

Hanna Mendes Levitin, Jinzhou Yuan, Yim Ling Cheng, Francisco JR Ruiz, Erin C Bush, Jeffrey N Bruce, Peter Canoll, Antonio Iavarone, Anna Lasorella, David M Blei, and Peter A Sims. De novo gene signature identification from single-cell RNA-seq with hierarchical Poisson factorization. *Molecular Systems Biology*, 15(2):e8557, February 2019. ISSN 1744-4292. doi: 10.15252/msb.20188557. URL https://www.embopress.org/doi/full/10.15252/msb.20188557.

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, December 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0229-2. URL https://www.nature.com/articles/s41592-018-0229-2.

Marian L. Burr, Christina E. Sparbier, Yih-Chih Chan, James C. Williamson, Katherine Woods, Paul A. Beavis, Enid Y. N. Lam, Melissa A. Henderson, Charles C. Bell, Sabine Stolzenburg, Omer Gilan, Stuart Bloor, Tahereh Noori, David W. Morgens, Michael C. Bassik, Paul J. Neeson, Andreas Behren, Phillip K. Darcy, Sarah-Jane Dawson, Ilia Voskoboinik, Joseph A. Trapani, Jonathan Cebon, Paul J. Lehner, and Mark A. Dawson. CMTM6 maintains the expression of PD-L1 and regulates anti-tumour immunity. *Nature*, 549(7670):101–105, September 2017. ISSN 1476-4687. doi: 10.1038/nature23643. URL https://www.nature.com/articles/nature23643.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, 3 edition edition, November 2013. ISBN 978-1-4398-4095-5.