Latent Variable Models for Simulated Gene Expression Data with Interventions

Max Medina** [1] Jake Yeung [2] Russell Littman [2]

[1]UC Berkeley [2]Genentech

Motivation

Gene programs, i.e. groups of genes collaborating to perform specific cellular functions, are central to provide insights into how cells respond to different stimuli or perturbations. They are crucial for advancing biomedical research in areas like disease progression and drug development [2].

Traditional methods often rely on manually curated gene sets that can be confounded by overlapping genes, noisy measurements and might only be applicable to unique contexts [2].

Data-driven methods can discover and interpret gene programs directly from single-cell RNA sequencing (scRNA-seq) data. When cells are subjected to specific **interventions** or perturbations, scRNA-seq can reveal how gene expression patterns change. This information is invaluable for identifying active gene programs and understanding cellular processes on a granular level [1].

Latent Variable Models reduce high-dimensional data into lower-dimensional representations called *factors*. If the factors capture the most significant variations in the data and integrate prior knowledge–such as existing gene sets and known cell types–we ensure that the extracted gene programs both explain the actual data and are aligned with established biological understanding [3].

Simulated Data

We simulate gene expression under different interventions by sampling normals with different means. We also add common variability across the groups. Even though gene counts are positive integers, often after scaling and applying the $\log(x+1)$ transformation, data more appropriately resembles normals.

Background: Probabilistic PCA

The basis of our work is the simplest **Bayesian Linear Factor** model called Probabilistic PCA (PPCA) [4].

Let y be an $n-$dimensional vector with the gene expression of a cell, x be the $m-$th dimensional latent factor, W the feature-to-factor map called the loading matrix, and an error term ε.

The **model** is

$$y|x, W, \sigma^2 \sim N(Wx, \sigma^2 I_n), \tag{1}$$

where $x \sim N(0, I_m), w_{j,k} \sim N(0,1)$, and $\sigma^2 \sim$ InverseGamma$(1,1)$. The reconstruction of our data will be $\hat{y} = \hat{W}\hat{x}$.

We use the mean or mode of the posterior, $p(x, W, \sigma^2|y)$, as **point estimates**. We compute them in two approaches:

1. Variational Inference (VI). We find tractable approximate posteriors, $q(x), q(W), q(\sigma^2)$, and compute the mean of each. We minimize:

$$-ELBO = \mathbb{E}_{x\sim q} NLL(y|x) + KL(q(x), p(x)) + KL(q(W), p(W)) + KL(q(\sigma^2), p(\sigma^2))$$

where $NLL(\cdot)$ represents the negative log-likelihood and $KL(\cdot,\cdot)$ the Kullback-Leibler divergence.

2. Maximum A Posteriori (MAP). Directly find the values that maximize the posterior by minimizing

$$NLL(x, W, \sigma^2|y) \propto NLL(y|x) + NLL(x) + NLL(W) + NLL(\sigma^2)$$

The Maximum-Likelihood Estimate (MLE) of W is $\hat{W} = U_m(\Lambda_m - \sigma^2 I_m)^{1/2}$,, where Λ_m and U_m contain the first m eigenvalues and eigenvectors of the covariance matrix. As $\sigma^2 \to 0$, $\hat{W}$ tends to the **projection matrix of PCA**. However, this never happens in practice, so PCA and PPCA differ.

Multi-View Domain-Informed (MuVi)

In [3], the authors propose a **Multi-View Domain-Informed** Latent Variable (MuVi) model. If we denote by z a vector of covariates associated with y, we model:

$$y|x, W, \beta, \Psi \sim N(Wx + \beta' z, \Psi), \tag{2}$$

where $\Psi = diag(\sigma_j^2), \beta \sim N(0, I_v)$ and W has a sparse prior known as the *horseshoe prior*:

$$w_{jk} \sim N(0, (\tau\delta_k\lambda_{jk})^2),$$
$$\tau, \delta_k, \lambda_{jk} \sim Cauchy^+(1).$$

This extends PPCA in two important ways:

1. **Supervision**: Allows for labels (i.e. interventions, cell type, etc.) in the gene expression data, thus adding supervision to an unsupervised problem.
2. **Sparsity**: The prior on W sparsifies the loadings. It *shrinks* the variance of w by multiplying heavy-tailed distributions. Sparsity favours interpretability.

Finding 1: Data Reconstruction

We simulate clusters of 100 points in $n = 2$ dimensions, with $m = 1$ true latent variables and noise $\sigma^2 = 0.5$.

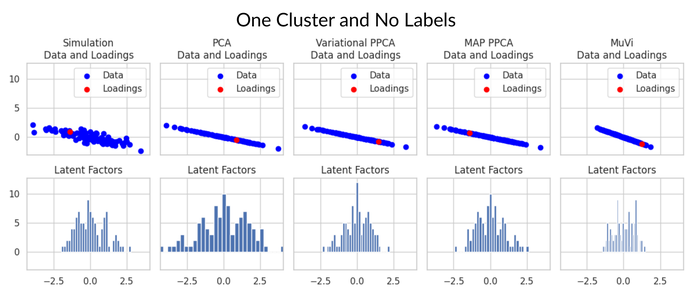

Figure 1. The first column is ground truth data (on blue), loadings (on red) and latent factors. Columns 2 to 4 show reconstructed data and estimated loadings and latent factors. MAP PPCA is very close to the ground truth and MuVi estimates are biased.

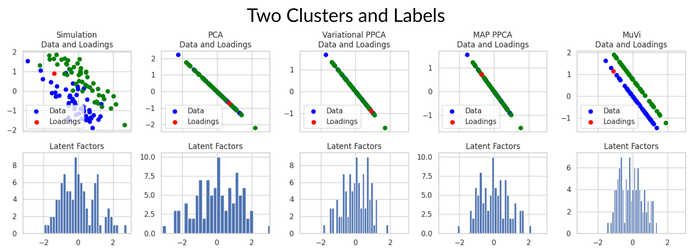

Figure 2. MuVi is effective at separating the data by using labels while PCA and PPCA fail at separating the data. MuVi is able to separate them by using labels

Finding 2: MuVi as Linear Regression + PPCA

If we assume W and β are independent, the model can be interpreted as two steps:

1. OLS Regression $y \sim z$: to capture portion of variance of the gene expression explained by the covariates
2. PPCA on $\hat{r} = y - \hat{\beta}' z$: to capture residual structure with latent factors.

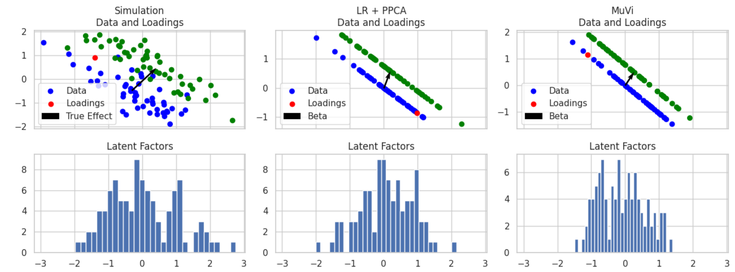

Figure 3. In MuVi, the approximate posterior is a fully parametrized family. The independence of β and W allows for this interpretation. However, it biases the estimation of β.

Conclusion

Algorithm Performance. We demonstrated the strengths and limitations of MuVi, by exploring its ability to reconstruct data and separate clusters using labels. We highlight its comparative advantages and limitations over PCA and PPCA, which are industry standards.

Conceptual Understanding. By interpreting MuVi as a combination of PPCA and linear regression, we clarified its inner workings, boosting application across different scenarios.

Limitations. In MuVi, the authors adopt independent approximate posteriors for W and β to simplify computations. However, this approach introduces bias in the estimation process and underestimates the variance, failing to capture the interactions between these parameters. To address this, we propose exploring a *blocked* factorization, where W and β are estimated jointly. Furthermore, additional simulations using high-dimensional, sparse, and non-normal data, as well as real-world datasets—such as [1], which contains gene expression profiles for *cytokine* interventions—are essential to validate the proposed approach.

References

[1] Ang Cui, Teddy Huang, Shuqiang Li, Aileen Ma, Jorge L Pérez, Chris Sander, Derin B Keskin, Catherine J Wu, Ernest Fraenkel, and Nir Hacohen.
Dictionary of immune responses to cytokines at single-cell resolution.
Nature, 625(7994):377–384, 2024.

[2] Russell Z Kunes, Thomas Walle, Max Land, Tal Nawy, and Dana Pe'er.
Supervised discovery of interpretable gene programs from single-cell data.
Nature Biotechnology, 42(7):1084–1095, 2024.

[3] Arber Qoku and Florian Buettner.
Encoding domain knowledge in multi-view latent variable models: A bayesian approach with structured sparsity.
In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 11545–11562. PMLR, 25–27 Apr 2023.

[4] Michael E Tipping and Christopher M Bishop.
Probabilistic principal component analysis.
Journal of the Royal Statistical Society Series B: Statistical Methodology, 61(3):611–622, 1999.