GPNSF: A model using GP priors and NMF to facilitate multi-omics data integration

Linxun GAO

January 6, 2026

Contents

1 Related Work

Jin et al. [1] proposed joint matrix factorization method to integrate multi omics data. The model aims to address two major challenges simultaneously: (i) the extremely sparse and near-binary nature of single-cell epigenomic data and (ii) the integration of this binary epigenomic data with the scRNA-seq data, which are often continuous after being normalized.

$$\begin{aligned}&\min_{W_1,W_2,H,Z\geq 0}\alpha\left\|X_1-W_1H\right\|_F^2\\&+\left\|X_2(Z\circ R)-W_2H\right\|_F^2+\lambda\left\|Z-H^TH\right\|_F^2\\&+\gamma\sum_j\left\|H_{\cdot j}\right\|_1^2,\end{aligned}$$

Figure 1: scAI Loss Function

A method considers the spatial information of single-cell data by using Gaussian process [3], but it's applied to single omic data (spatial count data).

$$y_{ij}\sim \text{Poi}(v_i\lambda_{ij})$$

$$\lambda_{ij}=\sum_{l=1}^{L}w_{jl}e^{f_{il}}$$

$$f_{il}=f_l(\mathbf{x}_i)\sim \text{GP}\left(\mu_l(\mathbf{x}_i),k_l(\mathbf{x}_i,X)\right).$$

Figure 2: NSF

2 Model

2.1 Model Overview

We have two data matrices $X_1 = (x_{ij}^{(1)}) \in \mathbb{R}_+^{n \times p}$ and $X_2 = (x_{ij}^{(2)}) \in \mathbb{R}_+^{n \times q}$ with the same number of samples n (they are samples from the same cells), but different numbers of features p and q. We assume that both matrices can be approximated by a low-dimensional representation $H \in \mathbb{R}^{n \times K}$ and factor loading matrices $W_1 = (w_{ij}^{(1)}) \in \mathbb{R}_+^{K \times p}$ and $W_2 = (w_{ij}^{(2)}) \in \mathbb{R}_+^{K \times q}$.

Let the spatial 2-D coordinates of the i-th sample in the low-dimensional space be denoted as $\boldsymbol{s_i} = (s_{i1}, s_{i2})$. The approximation we suppose is different from the standard NMF in three aspects:

(1) H is no longer a matrix of numbers, but a matrix of distributions priored by the spatial coordinates of the samples, i.e.,

$$h_{ik} = h_k(\boldsymbol{s_i}) = \mathcal{GP}(\mu_k(\boldsymbol{s_i}), K_k(\boldsymbol{s_i}, \boldsymbol{S})) \tag{1}$$

$$\mu_k(\boldsymbol{s_i}) = \beta_{0k} + \boldsymbol{s_i}^\top \boldsymbol{\beta_{1k}},\ i = 1, \ldots, n,\ k = 1, \ldots, K \tag{2}$$

$$K_k(\boldsymbol{s_i}, \boldsymbol{S}) = \gamma_1 \mathbf{1}_{\{i=j\}} + \gamma_2 \exp\left(-\frac{||\boldsymbol{s_i} - \boldsymbol{s_j}||}{2l}\right) \tag{3}$$

We sample from these K multi-variated Gaussians to get H each time.

(2) While W is restricted to non-negative, H is not because it is sampled from

some Gaussians. To compromise this, we calculate:

$$y_{ij}^{(r)} = \sum_{k=1}^{K} \exp(h_{ik}) w_{kj}^{(r)}, \ r = 1, 2 \tag{4}$$

(3) X is generated by under likelihood assumptions, if X_1 is scRNA-seq data, X_2 is ATAC data, then:

$$x_{ij}^{(1)} | y_{ij}^{(1)} \sim NB(y_{ij}^{(1)}, \theta_j) \tag{5}$$

$$x_{ij}^{(2)} | y_{ij}^{(2)} \sim Bernoulli(y_{ij}^{(2)} \lambda_j) \tag{6}$$

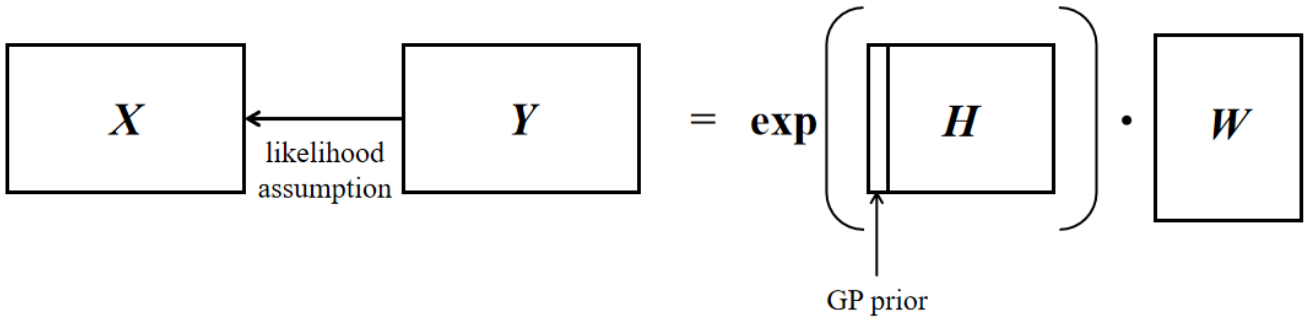

Figure 3: Model Skeleton

2.2 Variants

- Data enhancing step for scATAC data can be added to the model if we define $Z = (\mu_1(S), \ldots, \mu_K(S))^\top (\mu_1(S), \ldots, \mu_K(S))$ as the cell-to-cell similarity matrix.

- Non-spatial latent features can be add to the model if we separate the rows of H to two parts: one has GP prior and another has standard Gaussian prior.

3 Optimization

3.1 Inducing Point Locations

We assume a set of inducing point locations $\boldsymbol{z_m}$ indexed by $m = 1, \ldots, M$. If $M = n$ we set $\boldsymbol{z_m}$ to be the spatial coordinates S. Otherwise, for $M < N$ we set $\boldsymbol{z_m}$ to be the center points of a k-means clustering (with $k = M$) applied to S. Let

$$u_{mk} = \mathcal{GP}\left(\mu_k(\boldsymbol{z_m}), K_k(\boldsymbol{z_m}, Z)\right) \tag{7}$$

be the inducing points, i.e., the Gaussian process evaluation of the inducing locations.

The prior of the inducing points followed by 7 is

$$p(U; Z) = \prod_{k=1}^{K} p(u_k; Z) \tag{8}$$

$$p(u_k; Z) = \mathcal{N}(\mu_k(Z), K_{uuk}) \tag{9}$$

$$[K_{uuk}]_{m,m'} = K_k(\boldsymbol{z_m}, \boldsymbol{z'_m}) \tag{10}$$

Since H and U are jointly normal-distributed, the conditional probability of H

can be calculated given U:

$$p(H|U;S,Z) = \prod_{k=1}^{K} p(h_k|u_k;S,Z) \tag{11}$$

$$p(h_k|u_k;S,Z) = \mathcal{N}(\mu_{h|u,k}, K_{h|u,k}) \tag{12}$$

$$\mu_{h|u,k} = \mu_k(S) + K_{uhk}^{\top} K_{uuk}^{-1} \left(u_k - \mu_k(Z)\right) \tag{13}$$

$$K_{h|u,k} = K_{hhk} - K_{uhk}^{\top} K_{uuk}^{-1} K_{uhk} \tag{14}$$

$$[K_{uhk}]_{m,i} = K_k(\boldsymbol{z_m}, \boldsymbol{x_i}) \tag{15}$$

3.2 Variational Inference

We use the following approximation to facilitate variational inference:

$$q(U,H;S,Z) = p(H|U;S,Z)q(U;Z) \tag{16}$$

where $q(U;Z)$ is a parameterized Gaussian:

$$q(U;Z) = \prod_{k=1}^{K} q(u_k;Z) = \prod_{k=1}^{K} \mathcal{N}(\delta_k, \Omega_k) \tag{17}$$

We need to draw samples from the distribution of H under the variational ap-

proximation, which means we need to marginalize U in $q(U, H; S, Z)$ [2]:

$$q(H; S, Z) = \prod_{k=1}^{K} q(h_k|\delta_k, \Omega_k; S, Z) \tag{18}$$

$$q(h_k|\delta_k, \Omega_k; S, Z) = \int_{u_k} q(u_k, h_k; S, Z) = \mathcal{N}(\tilde{\mu}_k, \tilde{\Sigma}_k) \tag{19}$$

$$\tilde{\mu}_k = \mu_k(S) + K_{uhk}^{\top} K_{uuk}^{-1} \left(\delta_k - \mu_k(Z)\right) \tag{20}$$

$$\tilde{\Sigma}_k = K_{hhk} - K_{uhk}^{\top} K_{uuk}^{-1} \left(K_{uuk} - \Omega_k\right) K_{uuk}^{-1} K_{uhk} \tag{21}$$

Next, we find ELBO as the loss function [2]. Note that we ignore the subscript of X_1 and X_2 and drop the notations of S and Z:

$$\log p(X) = \mathbb{E}_{q(U,H)}[\log p(X)] \tag{22}$$

$$= \mathbb{E}_{q(U,H)}\left[\log \frac{p(X|H)p(H|U)p(U)}{p(U,H)}\right] \tag{23}$$

$$= \mathbb{E}_{q(U,H)}\left[\log \frac{p(X|H)p(H|U)p(U)}{p(U,H)} \cdot \frac{q(H,U)}{q(H,U)}\right] \tag{24}$$

$$= \mathbb{E}_{q(U,H)}\left[\log \frac{p(X|H)p(H|U)p(U)}{p(H|U)q(U)} \cdot \frac{q(H,U)}{p(H,U)}\right] \tag{25}$$

$$= \underbrace{\mathbb{E}_{q(U,H)}\big[\log p(X|H)\big] - \sum_{k=1}^{K} \mathrm{KL}\left(q(u_k)||p(u_k)\right) + \mathrm{KL}\left(q(H,U)||p(H,U)\right)}_{\text{ELBO}} \tag{26}$$

3.3 Loss Function

The loss function for whole model should both consider the reconstruction loss of X_1 and X_2:

$$\mathcal{L} \triangleq \mathbb{E}_{q(U,H)}\big[\log p(X_1|H)\big] + \mathbb{E}_{q(U,H)}\big[\log p(X_2|H)\big] - \eta \sum_{k=1}^{K} \mathrm{KL}\left(q(u_k)||p(u_k)\right) \quad (27)$$

where η is a hyperparameter to determine the proportion of reconstruction term and KL term.

The KL term in ELBO has a closed form since $q(u_k)$ and $p(u_k)$ are both Gaussians:

$$\begin{aligned}&\mathrm{KL}\left(q(u_k)||p(u_k)\right) = \\ &\frac{1}{2}\left[\log\frac{|K_{uuk}|}{|\Omega_k|} - M + \mathrm{tr}(K_{uuk}^{-1}\Omega_k) + (\delta_k - \mu_k(Z))^\top K_{uuk}(\delta_k - \mu_k(Z))\right] \end{aligned} \quad (28)$$

The expected log likelihood in ELBO takes more effect to compute.

$$\mathbb{E}_{q(U,H)}\big[\log p(X|H)\big] = \sum_{i=1}^{n}\sum_{j=1}^{p} \mathbb{E}_{q(U,H)}\Big[\zeta\Big(x_{ij}|\nu_j \sum_{k=1}^{K} \exp(h_{ik})w_{kj}\Big)\Big] \quad (29)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{p} \mathbb{E}_{q(H)}\Big[\zeta\Big(x_{ij}|\nu_j \sum_{k=1}^{K} \exp(h_{ik})w_{kj}\Big)\Big] \quad (30)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{p} \mathbb{E}_{q(h_{i1},\ldots,h_{iK})}\Big[\zeta\Big(x_{ij}|\nu_j \sum_{k=1}^{K} \exp(h_{ik})w_{kj}\Big)\Big] \quad (31)$$

Note that $h_{i1}, \ldots, h_{iK}$ are independent if i is fixed, so we can compute $q(h_{i,:})$ by:

$$q(h_{i,:}) = \prod_{k=1}^{K} q(h_{ik}) = \prod_{k=1}^{K} \mathcal{N}([\tilde{\mu}_k]_i, [\tilde{\Sigma}_k]_{i,i}) \tag{32}$$

$$[\tilde{\mu}_k]_i = \mu_k(\boldsymbol{s_i}) + \alpha_k(\boldsymbol{s_i})^\top (\delta_k - \mu_k(Z)) \tag{33}$$

$$[\tilde{\Sigma}_k]_{i,i} = K_k(\boldsymbol{s_i}, \boldsymbol{s_i}) - \alpha_k(\boldsymbol{s_i})^\top (K_{uuk} - \Omega_k)\alpha_k(\boldsymbol{s_i}) \tag{34}$$

$$\text{where } \alpha_k(\boldsymbol{s_i}) \triangleq K_{uuk}^{-1}[K_{uhk}]_{:,i} \tag{35}$$

So, instead of sampling H following equation 18, we actually sample H row by row following equation 32. Finally, the expectation is estimated by Monte Carlo procedure, i.e., we draw s samples from the distribution of H and calculate the mean of s log likelihoods.

References

[1] Suoqin Jin, Lihua Zhang, and Qing Nie. scai: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. *Genome biology*, 21(1):25, 2020.

[2] Kevin P Murphy. *Probabilistic machine learning: an introduction*, pages 679–732. MIT press, 2022.

[3] F William Townes and Barbara E Engelhardt. Nonnegative spatial factorization applied to spatial genomics. *Nature methods*, 20(2):229–238, 2023.