Linear Relations between Gene Expression in Several Tissues

Derek Modzelewski and Ben Pikus

December 13, 2017

Changes to Model

Added tissue centers

Normalize patient representations

Weight samples

Adaptive PCA

Adding Tissue Centers: Old Notation, Model

Notation:

$$
\begin{aligned}
Y_{iu} &:= \text{Gene Expression for patient } u \text{ in tissue } i \\
S_u &:= \text{Euclidean vector representation of patient } u \\
h &:= \text{Dimensionality of patient representations} \\
d &:= \text{Dimensionality of expression representations} \\
F_i &:= \text{Linear transform } (\mathbb{R}^{d \times h}) \text{ associated with tissue } i \\
c_i &:= \text{Tissue center}
\end{aligned}
$$

Model:

$$Y_{iu} = F_i S_u + c_i + \mathcal{N}(0, \sigma^2)$$

Adding Tissue Centers: Reframed Model

Notation:

$$S_u^* := \begin{pmatrix} 1 \\ S_u \end{pmatrix}$$

$$F_i* := \begin{pmatrix} c_i & F_i \end{pmatrix}$$

Model:

$$Y_{iu} = F_i^* S_u^* + \mathcal{N}(0, \sigma^2)$$

Adding Tissue Centers: New Algorithm

Given either F_i^* or S_u^*, becomes least-squares linear regression.
For solving for F and c, we use:

$$Y_{iu} = F_i^* S_u^* + \mathcal{N}(0, \sigma^2)$$

For solving for S, we use:

$$Y_{iu} = F_i S_u + c_i + \mathcal{N}(0, \sigma^2)$$

Alternate:

$$\text{set } F_i^* \leftarrow Y_i S_{(i)}^{*T} \left(S_{(i)}^* S_{(i)}^{*T} \right)^{-1} \qquad \forall i$$

$$\text{set } S_u \leftarrow \left(F_{(u)}^T F_{(u)} \right)^{-1} F_{(u)}^T (Y_{iu} - c_i) \qquad \forall u$$

Normalizing Patient Representations: Centering

Let $\bar{S} = |S|^{-1} \sum_u S_u$ Note that $S - \bar{S}$ has mean $\mathbf{0}$
Recall our model:

$$\begin{aligned} Y_{iu} &= F_i S_u + c_i + \mathcal{N}(0, \sigma^2) \\ &= F_i(S_u - \bar{S}) + (F_i \bar{S} + c_i) + \mathcal{N}(0, \sigma^2) \\ &= F_i S'_u + c'_i + \mathcal{N}(0, \sigma^2) \end{aligned}$$

where

$$\begin{aligned} S'_u &:= (S_u - \bar{S}) \\ c'_i &:= (F_i \bar{S} + c_i) \end{aligned}$$

Normalizing Patient Representations: Normalizing Variance

Let E be a matrix *s.t.* ES has identity covariance matrix.
Recall our model:

$$\begin{aligned} Y_{iu} &= F_i S_u + c_i + \mathcal{N}(0, \sigma^2) \\ &= F_i E^{-1} E S_u + c_i + \mathcal{N}(0, \sigma^2) \\ &= F_i' S_u' + c_i + \mathcal{N}(0, \sigma^2) \end{aligned}$$

where

$$\begin{aligned} S_u' &:= ES_u \\ F_i' &:= F_i E^{-1} \end{aligned}$$

Weighting Samples

Intuition: Some tissues have very few samples, and so we are uncertain of their tissue transforms (the F matrix). When estimating a patient's representation, we should not value these tissues as much since they are noisier sources of information.
Assumptions: Each tissue has its own persistent noise (bias), and its own variance when estimating a patient's representation. The variance is inversely proportional to the number of samples in that tissue.

$$\Rightarrow w_i \propto \frac{n_i}{\lambda + n_i}$$

where n_i is the number of samples in tissue i. In total, we use:

$$w_{iu} = \left(\frac{n_i}{\lambda_t + n_i}\right)\left(\frac{n_u}{\lambda_p + n_u}\right)$$

Note that this weight tends to 1 as the number of samples in both the tissue and the patient go to ∞

Weighting Samples: Model

This method of weighting the samples is equivalent to solving the MLE of a related model:

$$w_{iu} Y_{iu} = w_{iu}\left(F_i S_u + c_i\right) + \mathcal{N}(0, \sigma^2)$$

Which, without going into details, has nearly the same training algorithm as the previous models.
In the future, I might train for the λ_t and λ_p values, but currently they are priors.

Adaptive PCA

Motivation: We run PCA on each tissue separately. This minimizes the number of parameters estimated when representing tissues, and also aims to maximize how much of the variance of the data is kept. Instead of keeping a set number of PCs for each tissue, I globally choose which set of PCs across **all** tissues capture the most variance. Formally, to maximize:

$$\sum_i \sum_u Y_{iu}^T Y_{iu}$$

Does this overly penalize the tissues with few samples?

Methods and Results

Reconstruction

Subsequent Inference

Reconstruction: LOOR

Issues with reconstructing after throwing out necessary data.
LOOR (mostly) avoids that issue, but takes nearly 3 days to run.
Preliminary results:
$h = 4$, $\lambda_t = \lambda_p = 0$: %explained $= 16.37977 \pm 2.31300705$
$h = 5$, $\lambda_t = \lambda_p = 0$: %explained $= 18.91337 \pm 2.57074079$
$h = 5$, $\lambda_t = \lambda_p = 10$: %explained $= 19.46173 \pm 5.497874442$
$h = 4$, $\lambda_t = \lambda_p = 10$: %explained $= 17.613365704 \pm 4$
$h = 7$, $\lambda_t = \lambda_p = 10$: %explained $= 23.43076 \pm 10.78931$

Subsequent Inference

Ultimately, we wish to use these patient representations to relate patients or to predict features of those patients. When inferring sex, age, and DTHHRDY, we find that logistic classifiers on the patient representations are no better than predicting the most common class in the training set (naive classification). With weighting, we achieve a whole 3% better than random result. NOTE: these results are old, in need of update.