Linear Relations between Gene Expression in Several Tissues

Derek Modzelewski and Ben Pikus

May 3, 2017

Notation, Model

Notation:

$Y_{iu} :=$ Gene Expression for patient u in tissue i

$S_u :=$ Euclidean vector representation of patient u

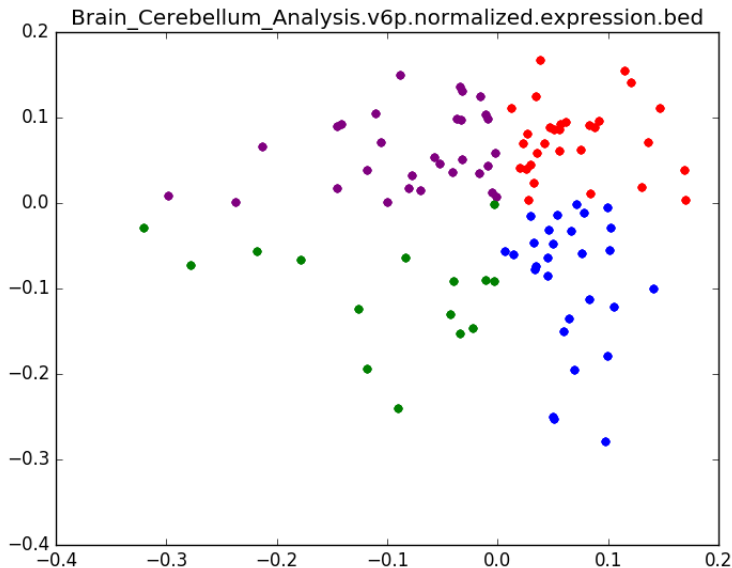$h :=$ Dimensionality of patient representations

$d :=$ Dimensionality of expression representations

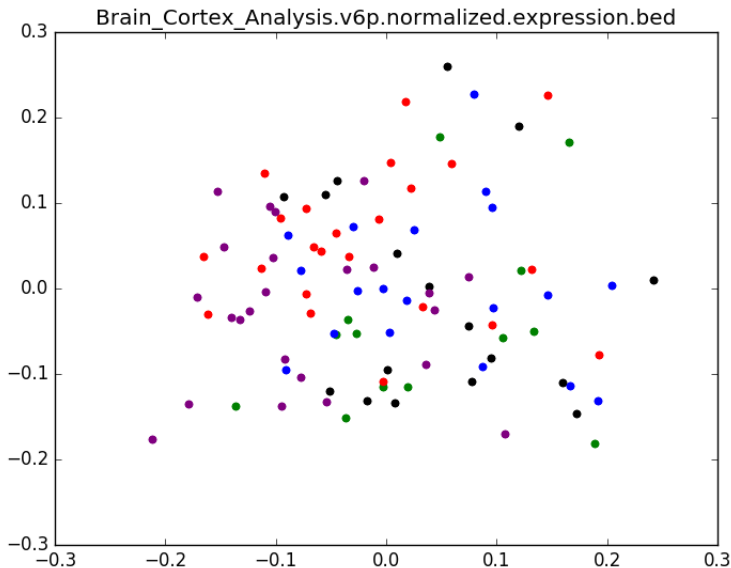$F_i :=$ Linear transform ($\mathbb{R}^{d \times h}$) associated with tissue i

Model:

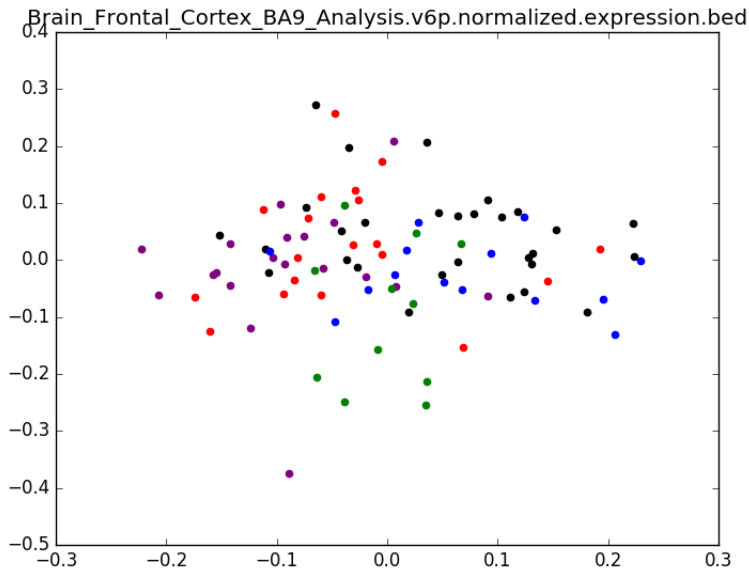$$Y_{iu} \sim F_i S_u + \mathcal{N}(0, \sigma^2)$$
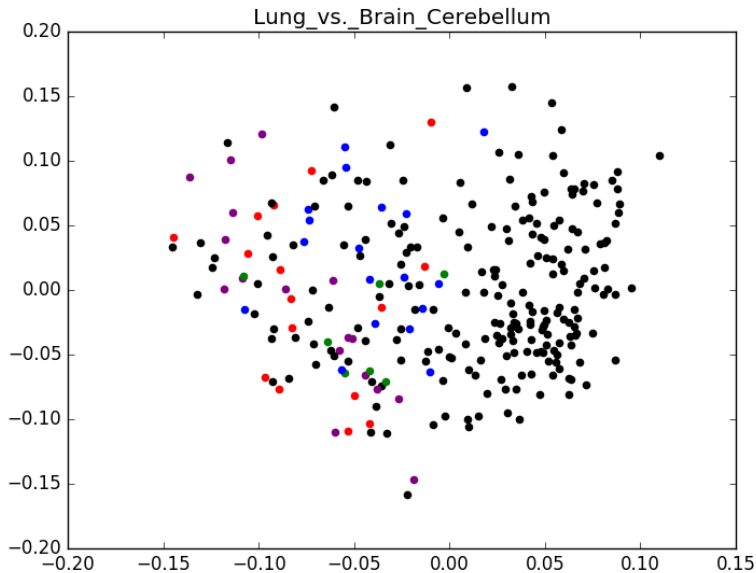
Comparing Tissues by Patients (Cerebellum)

Brain_Cerebellum_Analysis.v6p.normalized.expression.bed

Comparing Tissues by Patients (Cortex)

Brain_Cortex_Analysis.v6p.normalized.expression.bed

Comparing Tissues by Patients (Frontal Cortex)

Brain_Frontal_Cortex_BA9_Analysis.v6p.normalized.expression.bed

Comparing Tissues by Patients (Lung)

Lung_vs._Brain_Cerebellum

Comparing Tissues by Patients (Pancreas)

Pancreas_vs._Brain_Cerebellum
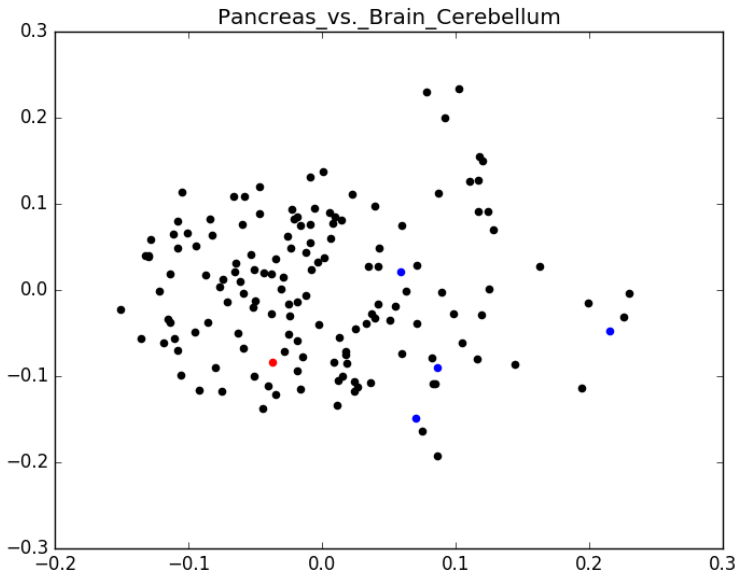
Our algorithm

Recall our model:

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

Given either F_i or S_u, looks like least-squares linear regression!
Alternate:

set $F_i \leftarrow Y_i S^T (SS^T)^{-1}$ $\qquad \forall i$

set $S_u \leftarrow \frac{1}{|I_u|} \sum_{i \in I_u} (F_i^T F_i)^{-1} F_i^T Y_{iu}$ $\qquad \forall u$
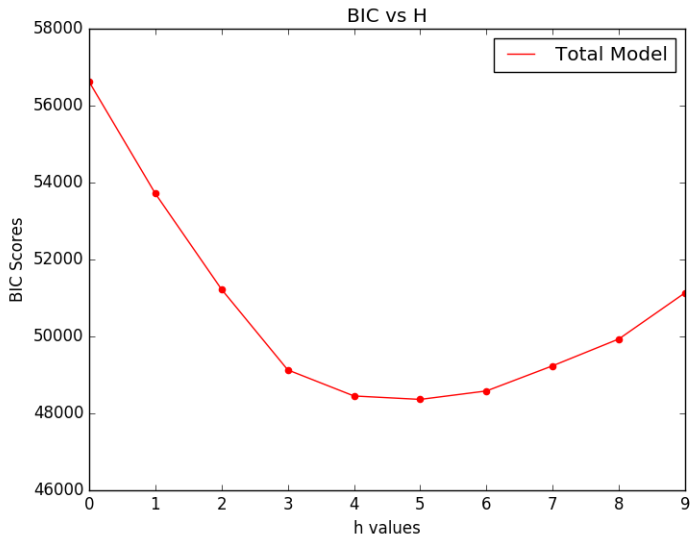
BIC evaluation

Since F is derived from S in closed form, so F is not a free parameter.

S has nh parameters, where n is the number of patients and h is dimensionality of representation
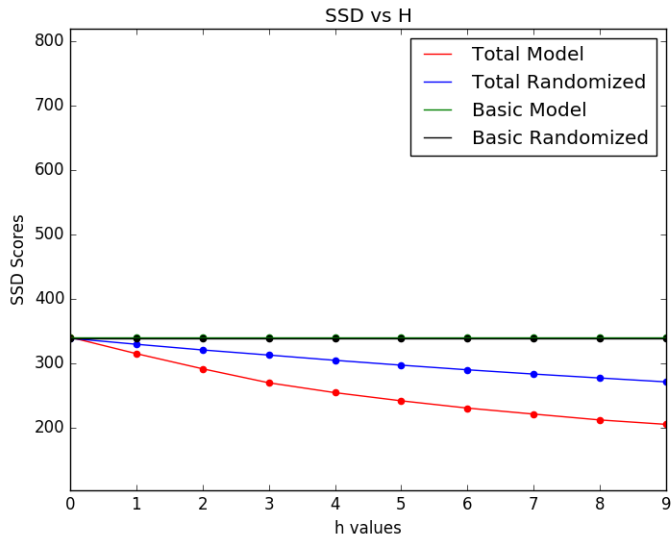
But every solution is rotationally equivalent with solutions along h degrees of freedom, leaving us $(n-1)h$ free parameters.

$$\Rightarrow \mathrm{BIC} = \ln n(n-1)h - 2\ln p(Y:S)$$
$$= \big(\ln n\big)(n-1)h + n\ln(2\pi\sigma^2) + \frac{\mathrm{SSD}}{\sigma^2}$$

BIC score across h settings

SSD score across h settings

Applications

Filling gaps in data: if missing gene expression for some patient for some tissue, can just model it.

Comparing patients: Can calculate cosine similarity between patients' vector representations to compare their similarity.

Deciding on patients: Since all patient representations are in the same vector space, we may solve any decision problem there.