Graphic Lasso: Data analysis

Jiaxin Hu

February 25, 2021

1 Pre-processed data with unbalanced weight

The unbalanced sample size has some effects to the pre-processed tensor data clustering. For comparison, Table 1 is the result with balanced sample size, and Table 2 is the result with unbalanced sample size.

$A_0 : u_{kl} = 0$	Brain - Cerebellum Brain - Cerebellar Hemisphere
$A_{11} : u_{k1} > 0$	Brain - Cortex Brain - Anterior cingulate cortex (BA24) Brain - Frontal Cortex (BA9)
$A_{12} : u_{k1} < 0$	Brain - Hippocampus, Brain - Hypothalamus Brain - Spinal cord (cervical c-1) Brain - Amygdala
$A_2 : u_{k2} \neq 0$	Brain - Caudate (basal ganglia) Brain - Nucleus accumbens (basal ganglia) Brain - Putamen (basal ganglia)
$A_3 : u_{k3} \neq 0$	Brain - Substantia nigra

Table 1: Membership result for 13 brain tissues with $r = 3, \rho = 1000$ with balanced sample size.

$A_{11} : u_{k1} > 0.2$	Brain - Cortex Brain - Anterior cingulate cortex (BA24) Brain - Frontal Cortex (BA9) Brain - Amygdala
$A_{12} : 0.2 > u_{k1} > 0$	Brain - Cerebellum Brain - Cerebellar Hemisphere
$A_{13} : 0 > u_{k1} > -0.2$	Brain - Caudate (basal ganglia) Brain - Nucleus accumbens (basal ganglia) Brain - Putamen (basal ganglia) Brain - Substantia nigra
$A_{14} : -0.2 > u_{k1}$	Brain - Hippocampus, Brain - Hypothalamus Brain - Spinal cord (cervical c-1)

Table 2: Membership result for 13 brain tissues with $r = 3, \rho = 1000$ with unbalanced sample size.

Though the clustering results are similar, the criterion to distinguish each group becomes tricky. In balanced case, we only use signs to distinguish groups share the same Θ, while in balanced case, the rank degenerates to 1 and we need to use the threshold 0.2 to separate the tissues share the common Θ.

Different settings such as $r = 2, \rho = 100$ and $r = 3, \rho = 500$ give similar results.

2 Gtex data with balanced weight

The Gtex data with pre-processed genes has sightly improvement in clustering when we set the sample size is balanced, for example, `sample size = rep(100,13)`. For comparison, under $r = 3, \rho = 1000$, Table 3 is the clustering results of Gtex data with true unbalanced sample size, and Table 4 is the clustering results with balanced sample size `rep(100,13)`.

$A_0 : u_{kl} = 0$	Brain - Putamen (basal ganglia) Brain - Amygdala
$A_{11} : u_{k1} > 0$	Brain - Spinal cord (cervical c-1)
$A_{12} : u_{k1} < 0$	Brain - Anterior cingulate cortex (BA24) Brain - Frontal Cortex (BA9)
$A_{21} : u_{k2} > 0$	Brain - Hippocampus, Brain - Hypothalamus Brain - Substantia nigra
$A_{21} : u_{k2} < 0$	Brain - Cortex Brain - Caudate (basal ganglia)
$A_{31} : u_{k3} > 0$	Brain - Cerebellum Brain - Cerebellar Hemisphere
$A_{31} : u_{k3} < 0$	Brain - Nucleus accumbens (basal ganglia)

Table 3: Membership result for 13 Gtex brain tissues with $r = 3, \rho = 1000$ with unbalanced sample size.

$A_{11} : u_{k1} > 0$	Brain - Hippocampus Brain - Substantia nigra
$A_{12} : u_{k1} < 0$	Brain - Anterior cingulate cortex (BA24) Brain - Frontal Cortex (BA9) Brain - Caudate (basal ganglia) Brain - Cortex Brain - Amygdala
$A_{21} : u_{k2} > 0$	Brain - Cerebellum Brain - Cerebellar Hemisphere
$A_{21} : u_{k2} < 0$	Brain - Nucleus accumbens (basal ganglia) Brain - Putamen (basal ganglia) Brain - Hypothalamus Brain - Spinal cord (cervical c-1)

Table 4: Membership result for 13 Gtex brain tissues with $r = 3, \rho = 1000$ with balanced sample size.

With balanced sample size, the cortex tissues are grouped together. Under $\rho = 2, \rho = 1000$, there are also slight improvement with balanced sample size, for example, cortex are separated from cerebellum under balanced size.

Though there exist slight improvements with balances size, there are still some problems with Gtex data. For example, under $r = 3$, we can not separate cortex with basal ganglia tissue; under $r = 2$, we can not separate cerebellum with basal ganglia. Therefore, we may choose other genes in the Gtex dataset.