A compressed sensing approach to determine MDS to AML progression rate

Pratik Worah
Problem

- Given a set of 26K gene expression data (tpm) for 26 patients:
  - 13 patients have MDS progressing to AML in < 3 years. (rapid progression)
  - 13 patients have MDS progressing to AML in > 5 years. (slow progression)
- **Goal**: Use this data to predict (for a new patient) whether MDS will progress to AML rapidly or slowly.
  - Split data into 8+8 sample for training and 5+5 sample for test.
- **Constraint**: Predictions can not depend on large set of genes, to avoid overfitting, as well as for practical reasons.
Existing methods

- The method outlined in these slides achieves 60% prediction accuracy (with some caveats).
- It also gives a way of visualizing the clusters in 3D.
- **Typical approach:** Regressions, SVMs, Random forests etc.
- **Problems:** Either depend on too many genes or need large number of samples.
- Combining usual clustering ideas with compressed sensing [Candes-Tao] may help restrict to few genes, but it needs a slight modification.
Algorithm

- **Input**: A $n \times m$ matrix $Y$ of gene expressions in tpm; $n$ genes, $m$ examples (positive and negative)
- **Output**: A $n \times p$ matrix $M$ of measurement coefficients, so that the $m$ $p$-dimensional points $YM$ are separated into two clusters (based on progression rate of MDS).
- **Objective**: $\min_M \max_v |vM|_0$, where $v$ is a $(p+1)$-dimensional vector.
- Also impose L2 norm constraints separate the rapid and slow progressing MDS patients.
- **Mathematical Result**: minimax SDP with L0 objective function, we can show that under restricted isometry type property on $Y$, L1 optimization suffices!
Relaxation

\[
\min_{M \in \mathbb{R}^{n \times p}} \max_{v \in \mathbb{R}^{p+1}} \|vM\|_0 \\
\text{s.t. } \| (y_i - y_j)M \|_2 = \text{dist}(i, j), \\
\text{dist}(i, j) \in \{1 - \epsilon, \epsilon\}
\]

- The SDP typically has 200K variables - not solvable in reasonable time!
- It can be converted into a “dual” Quadratic Program, which is more tractable.
Guide to interpreting pictures on next slide:

- Each point corresponds to a patient.
- There are 8 Red + 8 Blue = 16 points on the training set.
- There are 5 Red + 5 Blue = 10 points on test set.
- The Red points correspond to slow progressors i.e., patients who progressed to AML after at least 5 years.
- The Blue points correspond to rapid progressors i.e., patients who progressed to AML within 3 years.
- The idea is to cluster the training set of patients by mapping each patient to a point in 3D space, based on a small number of genes.
- Each patient in test set is mapped to rapid or slow progressor, based on their similarity to the closest training set patient.
Training range: genes expressions ranked 251-350

- Rapidly progressing MDS points closer to origin, so can’t use this to identify singular genes that predict rapid progression (but it does give a formula).
- However, **HSPA8 and HISTH1D over-expression indicates slow progression**.
Test

- Test has about 60% accuracy. 80% accurate for predicting slow, 40% accurate for predicting rapidly progressing MDS.
The formula for the coordinates of the points

For each point

- Coordinate[1] = -1.1SNORD67+1.1RPL13-GNAS-1.8HSPA8-1.7HISTH1D+...
- Coordinate[2] = RPL13 -1.8HSPA8 -1.7HISTH1D+...
- Coordinate[3] = -1.1SNORD67 + 1.8CALR -1.5HSPA8 -2.3HISTH1D+...

How to draw conclusions?

- Observe that higher CALR conc. increases coordinate[3], driving it toward the origin i.e., towards blue points in the training data.
- So high CALR leads to greater chances of a patient having rapid MDS.
Another example

- Observe that Blue points are further from the origin in coordinate[3]
- Therefore, large positive values in coordinate[3] for a gene indicates that the gene leads to rapidly progressing disease.
LAPTM5 and scaRNA22

- Effect of removing scaRNA22 from the coordinate calculation.
- Both LAPTM5 and scaRNA22 are responsible for rapid progression!
Conclusions

- Increased HSPA8 and HISTH1D expression indicates slow progression.
- **Rapid progression can be harder to detect, usually no single gene is responsible - the blue clusters are typically present close to the origin, compared to the red clusters!**
- That said, increased FOS, RPL27, LAPT M5, CALR seem indicative of rapid progression.
References

1. [Candes-Tao] Stable signal recovery from incomplete and inaccurate measurements, Candes et al., 2006.
2. [scaRNA22] The expression pattern of small nucleolar and small Cajal body-specific RNAs characterizes distinct molecular subtypes of multiple myeloma, Ronchetti et al., 2012.
4. [RPL27] https://www.proteinatlas.org/ENSG00000131469-RPL27/pathology
5. ...