Ultrahigh dimensional variable screening with measurement error.

Magne Thoresen

University of Oslo

High-dimensional regression problems (p>>n) has become common in statistics, and penalized regression models may be a good solution.

However, regression situations where the number of potential explanatory variables is far too high to be included in any regression model, even penalized models, start to emerge.

In such situations, one will need to do some initial screening to bring the number of candidate variables down to something we can manage, without losing the important ones.

A number of methods have been suggested, many of them building in the Sure Independence Screening (SIS) idea (Fan & Lv, 2008), and their extension Iterated SIS.

But what if we have measurement error?

Example

Sample of approx. 250 individuals, approx. 4 000 000 methylation sites.

Measurements based on sequencing technology.

Methylation is a part of epigenetics, regulating gene transcription.

Methylation is in principle on / off, but as we typically analyze a mix of molecules for each site, our measurements are proportions (proportion being methylated), and we talk about methylation rate.

Ultimate goal: Predict age from methylation (forensic purposes).

The estimated methylation rate is based on a number of reads, which will vary between sites and across people, leading to varying degree of sample variation.

For a given site (position), say we have m reads of which T is methylated; estimated methylation rate is T/m which we denote W.

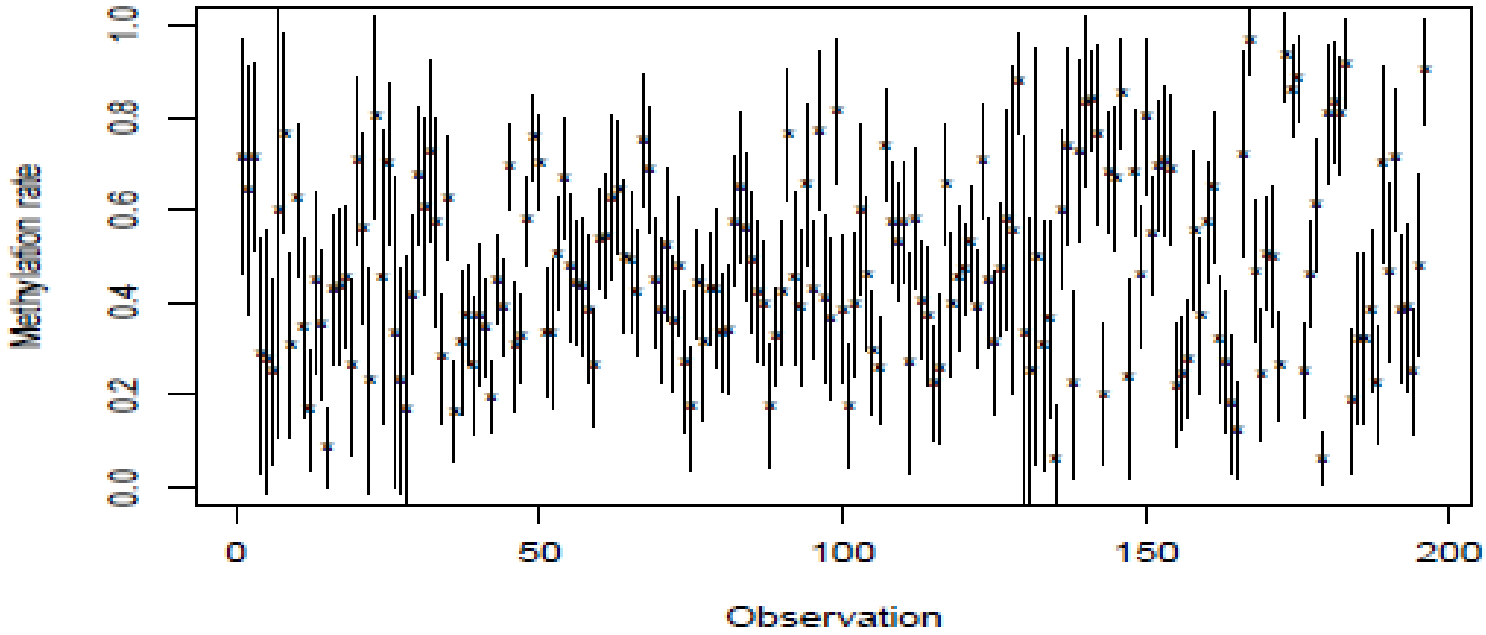
Typically, estimated methylation rates from sites with too few reads (<20, 30) are discarded from the analysis.

These data lead to a number of interesting methodological problems, both of the high-dimensional type and when studying one position / site at the time.

We can look at the sampling variability as measurement error. If we let the true methylation rate be denoted X, we have additive measurement error in the sense that

$$E(W_i \mid x_i) = x_i.$$

The error has binomial variation, varying with methylation rate and number of reads (so heteroscedastic error).


Methylation rate
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
Observation

The ultimate goal is to fit a regression model

$$Y_i \mid \mathbf{x}_i = f(\boldsymbol{\beta}, \mathbf{x}_i) + \varepsilon_i,\ i=1 \text{ to } n,$$

where $\boldsymbol{\beta}, \mathbf{x}_i$ are of dimension $k < n$.

Due to the extremely high dimension (ultrahigh dimensional problem) one will need to do some screening / filtering before running any regression model.

We would probably like to base this screening on some measure related to the squared error, like R^2. However, in our data we will have to deal with

- measurement error (sampling variability)
- missing data (meaning analyzing different samples for each position)
- non-linear associations
- residual variance varying with methylation rate

In essence, we need a model for the residual variance, and we need a way to rank sites according to the “output” of this model.

Focusing on the measurement error, we need to estimate this variance function, possibly under non-linear relationships, corrected for measurement error, and we need a method for variable filtering based on this correction, which is fast enough for practical use.

In general, ultrahigh dimensional variable filtering with measurement error.