Given big data **,**  a data matrix with

data points and covariates and is the response vector.

A modelthat is assumed to describe the big data

(**,** where is some function of ).

Subsampling probabilities for each data point based on the

assumed model are determined and a subsample based on

these probabilities are obtained.

Given big data

Consider a set of Q models which can be constructed to

encapsulate a variety of scenarios that may be observed

within the data with model probabilities , where .

Assumed Model 1 Assumed Model q Assumed Model Q

Determine Determine Determine

subsampling subsampling subsampling

probabilities probabilities probabilities

Obtain model robust subsampling probabilities via a

weighted average such that for

and a subsample based on these probabilities are obtained.

Data Model:

1. Random sampling to estimate the parameters of the data model.
2. Optimal subsampling under the data model.
3. to 5. Optimal subsampling under alternative models (All models in Table 1 except the data model).

6. Model robust subsampling of the data model under the assumption that each of the models are equally likely a priori.

Using the estimated model parameters from the subsamples evaluate the simulated mean squared error, where

is an matrix, is the number of parameters in the data model.

Data Model:

1. Random sampling 2. Optimal subsampling 6. Model robust subsampling

to estimate the under the data

parameters of the model ().

data model ().