Title:
Optimal subset selection without outliers

Authors:
Elena Pesce (Swiss Re Institute, Swiss Re Management Ltd, Zurich, Switzerland) Laura Deldossi (Dipartimento di Scienze Statistiche, Università Cattolica del Sacro Cuore, Milan, Italy) and Chiara Tommasi (Dipartimento di Economia, Management e Metodi Quantitativi,Università degli Studi di Milano, Milan, Italy)

Abstract:
With the advent of 'Big Data', massive data sets are becoming increasingly prevalent. Several subdata selection are proposed in these last few years both to reduce the computational burden and to improve cost effectiveness and learning of the phenomenon. Some of these proposals (Drovandi et al., 2017; Wang et al., 2019; Deldossi and Tommasi (2021) among others) are inspired to Optimal Experimental Design (OED). However, differently from the OED context - where researchers have typically complete control over the predictors - in subsampling methods these, and the responses as well, are passively observed. Thus if outliers are present in the 'Big Data', it is likely that they could be included in the sample selected applying the D-criterion, being the D-optimal design points on the boundary of the design space.

In regression analysis, outliers - and more in general influential points – could have a large impact on the estimates; identify and exclude them in advance, especially in large datasets, is generally not an easy task. In this study, we propose an exchange procedure to select a compromise-optimal subset which is informative for the inferential goal and avoids outliers and 'bad' influential points.