In traditional robust statistics it is generally assumed that the majority of the observations is free of contamination. An observation that deviates from the model is as a whole flagged as an outlier even if only one component of the observation is contaminated. Since flagged outliers will be down-weighted in the estimation, this leads to a huge loss of information on the intact components. For a data set with a large number of features, it is more common to observe that components are randomly contaminated and that the proportion of component-wise contaminated observations can be well beyond a half.

In this setting, most traditional robust methods may fail because they naively assume less than half the observations being contaminated. The objective of my proposal is therefore to revamp the traditional robust methods of multiple regression to adapt to the component-wise contamination.

To my best knowledge, there are very limited existing regression methods that can handle random component-wise outliers. Recently, Ollerer et al. (2013) proposed to first decompose the minimization problem in multiple regression into a sequence of minimization problems in a simple regression setting. Then they replace the residual in the problem by its robust counterpart to achieve robustness against component-wise outliers. On the other hand, Xu et al. (2010) proposed to rewrite the minimization problem to include sparsity terms for every cells in the design matrix in order to remove the effect of component-wise outliers on the parameter estimation. Neither of the methods can deal with dummy predictors. Moreover, neither of them allows statistical inference about the estimates.

To address the drawback of the previous methods, I propose the following method that: (1) first uses a univariate filtering by comparing the tail distributions of the predictors with an adaptively chosen Pareto distribution to detect and eliminate large component-wise outliers; (2) then uses a robust estimator for missing data to down-weight undetected contaminated observations in the estimation of regression parameters. With a mild assumption that the tail distributions of the predictors are lighter than that of the Pareto, the proposed filtering can be shown to be efficient. That is, the filtering will not mistakenly remove any components in an observation when the data is generated from the model distribution, i.e., not contaminated. This leads to a robust estimation for complete data in the second step. Therefore, the traditional asymptotic theories for robust estimators for complete data can be adopted and statistical inference about the estimates is doable.

* Ollerer, V., Alfons, A., & Croux, C. (2013). The shooting S-estimator for robust regression. Technical report. SSRN: http://ssrn.com/abstract=2381960 or <http://dx.doi.org/10.2139/ssrn.2381960>.
* H. Xu, C. Caramanis, and S. Mannor (2010), Robust regression and Lasso, IEEE Trans. Inform. Theory, 56, pp. 3561–3574.