# Abstract

Nonparametric regression is a very popular tool for data analysis because these techniques impose few assumptions about the shape of the mean function. Hence, they are extremely flexible tools for uncovering nonlinear relationships between variables. A disadvantage of these methods is their computational complexity when considering large data sets. In order to reduce the complexity for least squares support vector machines (LS-SVM), we propose a method called Fixed-Size LS-SVM which is capable of handling large data set on standard personal computers.

We study the properties of the LS-SVM regression when relaxing the Gauss-Markov conditions. We propose a robust version of LS-SVM based on iterative reweighting with weights based on the distribution of the error variables. We show that the empirical maxbias of the proposed robust estimator increases slightly with the number of outliers in region and stays bounded right up to the breakdown point. We also establish three conditions to obtain a fully robust nonparametric estimator.

We investigate the consequences when the i.i.d. assumptions is violated. We show that, for nonparametric kernel based regression, classical model selection procedures such as cross-validation, generalized cross-validation and $v$-fold cross-validation break down in the presence of correlated data and not the chosen smoothing method. Therefore, we develop a model selection procedure for LS-SVM in order to effectively handle correlation in the data without requiring any prior knowledge about the correlation structure.

Next, we propose bias-corrected $100(1-\alpha)\%$ approximate confidence and prediction intervals (pointwise and uniform) for linear smoothers, in particularly for LS-SVM. We prove, under certain conditions, the asymptotic normality of LS-SVM. Further, we show the practical use of these interval estimates by means of toy examples for regression and classification.

Finally, we illustrate the capabilities of the proposed methods on a number of applications i.e. system identification, hypothesis testing and density estimation.