### The need for nonparametric estimation and big data in fluctuation period

In the strong effort to free discriminatory analysis from tight assumptions by applying nonparametric estimates, Fix & Hodges (1951) theoretically proved that with the large enough dataset, the nonparametric estimates are consistent (consistency in the sense of decision functions) with likelihood ratio procedure (procedures, as the author said, are believed to have ideal properties concerning control of possibility of misclassification). They considered that in the case of data without normal distribution or equal covariance matrices, the validity of the estimates of the linear discriminant function (most familiar parametric procedure at that time) is doubtful. This insight is supported by both theoretical and empirical works. In the book that presented the topic of nonparametric analysis with density estimation, Silverman (1998) said that the linear rule is not able to perform well with data with non-normal distribution, such as clusters; the nonparametric rule has the availability to deal with this feature in the data. He pointed out by generalizing the ideas of density estimation, nonparametric discrimination can be extended to deal with discrete and mixed data. However, this procedure requires extremely large sample sizes, especially if the data are of high dimensionality. Fortunately, the smoothing dimensional algorithm of density estimation procedures can resolve this problem.

Empirically, using data of US ISM production index and Eurozone manufacturing PMI output index, Vermeulen (2012) found out that estimations based on models constructed by these data have better performance than simple autoregressive in forecasting exercise of actual industrial production growth in US and Euro area. The data of the ISM production index is reported based on monthly interviews from the ISM website with more than 400 industrial companies in the US. While Eurozone Manufacturing PMI output index is collected from Markit, which provides answers from more than 3000 manufacturing firms. The author applied the Carlson-Parkin method to deal with data from the survey. This method tries to make a parametric assumption on the underlying distribution for the answers of the survey, then that distribution is used to provide an estimate of the mean of the distribution based on the aggregate response shares. Three different distributions were applied on cross-sectional data samples: uniform, logistic and Laplace distribution. Root mean squared forecast errors (RMSFE) are adapted to test the performance of forecasting models. The result shows that compared with the fitted values of a benchmark autoregressive process, the output of survey data models is better. However, in a dramatic abrupt, and massive period like the financial crisis, the RMSFE has a large increase, which shows the more volatile and less predictable forecasting output growth, and the additional survey data has less meaning. The results considered that nonlinear transformation of the data applied in the Carlson-Parking method leads to its outliers, and it leads to that result. Moreover, Vermeulen (2012) also considered that during a financial crisis, the Laplace distribution constructs the best estimates.

With the large efforts put into reviewing applying big data to economic research, Giannone et al. (2018) found out that density models are more effective than sparsity models in using high-frequency data. They defined sparse-modeling techniques "focus on the selection a small set of explanatory variables with the highest predictive power, out of a much larger pool of regressors", while dense-modeling techniques "recognize that all possible explanatory variables might be important for the prediction, although their impact might be small." Using six popular datasets with large information in macroeconomics, finance, and microeconomics, Giannone et al. (2018) constructed a model to estimate economic activity. For macroeconomics, they studied the predictability of economic activity in the US, and the determinants of economic growth in a cross-section of countries. In finance application, they investigate the predictability of the US equity premium and the determinants of the cross-sectional variation of US stock returns. Lastly, for microeconomics, they want to find out the factors behind the decline in the crime rate in a cross-section of US states, and the determinants of rulings in the matter of government taking of private property in US judicial circuits. Applying sparse predictive models and Bayesian inferential method, they found out that: more than a handful of predictors, more preferable to achieve good forecasting accuracy; larger-scale models forecast well in their framework; and "while the appropriate degree of shrinkage and model size is quite well identified, the data are much less informative about the specific set of predictors that should be included in the model." Moreover, they considered that because sparsity is essentially assumed, the conclusion about predictions of sparsity is better cannot be accepted with these methods. In the addition, the output from sparse estimators is not clear if the true data-generating process is dense. Sometimes the true model is not sparse, but because of reducing estimation uncertainty, alleviating the curse of dimensionality, and improving prediction accuracy, they select a small set of explanatory or the sparsity approach.