Big data analytics using semi-supervised learning methods

The expanding availability of complex data structures requires development of new analysis methods for process understanding and monitoring. In manufacturing, this is primarily due to high-frequency and high-dimensional data available through automated data collection schemes and sensors. However, particularly for fast production rate situations, data on the quality characteristics of the process output tend to be scarcer than the available process data. There has been a considerable effort in incorporating latent structure–based methods in the context of complex data. The research question addressed in this paper is to make use of latent structure–based methods in the pursuit of better predictions using all available data including the process data for which there are no corresponding output measurements, ie, unlabeled data. Inspiration for the research question comes from an industrial setting where there is a need for prediction with extremely low tolerances. A semi-supervised principal component regression method is compared against benchmark latent structure–based methods, principal components regression, and partial least squares, on simulated and experimental data. In the analysis, we show the circumstances in which it becomes more advantageous to use the semi-supervised principal component regression over these competing methods.