Finding the hidden link: Statistical methods for multi-source high-dimensional data

Katrijn Van Deun
Tilburg University, Tilburg, The Netherlands

Research in many disciplines, including the behavioural and social sciences, has entered the era of big data. Many detailed measurements are taken and multiple sources of information are used to unravel complex multivariate relations. For example, in studying obesity or depression as the outcome of environmental and genetic influences, researchers increasingly collect survey, dietary, biomarker and genetic data from the same individuals. Revealing the variables that are linked throughout these different types of data gives crucial insight in the complex interplay between the multiple factors that determine human behavior, e.g., the concerted action of genes and environment in the emergence of obesity or depression.

Although linked more-variables-than-samples (or, high-dimensional) multi-source data form an extremely rich resource for research, extracting meaningful and integrated information is challenging and not appropriately addressed by current statistical methods. The challenge is to select - in an automated way - those variables that are linked throughout the different blocks and this eludes current available methods for data analysis. A first problem is that relevant information is hidden in a bulk of irrelevant variables with a high risk of finding incidental associations. Second, the sources are often very heterogeneous, which may obscure apparent links between the shared mechanisms.

In this presentation we will discuss the challenges associated to the analysis of large scale multi-source data and present a sparse common and distinctive components approach to address the challenges.