Assessing the uncertainty in $k$nn Data Fusion

Tomàs Aluja-Banet*, Josep Daunis-i-Estadella**, Enric Ripoll***

*Universitat Politècnica de Catalunya, Campus Nord, C5-204, E-08034 Barcelona
tomas.aluja@upc.edu
**Universitat de Girona, Campus de Montilivi, Edifici P4, E-17071 Girona
josep.daunis@udg.edu
***Institut d'Estadística de Catalunya, Via Laietana 58, E-08003 Barcelona
eripoll@idescat.net

Data fusion, also known as statistical matching, is a technological operation whose aim is to integrate the information of two independent data sources. Let $(X_0; Y_0)$ the *donor* file and (X_1) the *recipient* file, where the X are the *common* variables and the Y are the *specific* ones. The goal is to complete the recipient file $(X_1, \hat{Y}_1)$ in such a way that it can a be a realization of the joint density function $f(X, Y)$.

There are three basic approaches for data fusion. The first one consists of embedding the common and specific variables within a *parametric* multivariate distribution $f(X, Y|\theta)$, assuming donors and receptors independently and randomly draw from this distribution. This distribution can be factored into $f(X, Y|\theta) = f(Y|X, \theta_{Y|X})f(X, \theta_X)$; hence, it is possible to estimate its parameters θ_X and $\theta_{Y|X}$ from the available information and use them to impute the missing block of data. The second approach consists of directly *modelling* the relationship between the Y variables and the X variables in the donor file by means a regression function: $E(Y|X) = r(X) + \varepsilon$ and applying this model in the *recipient* file (*explicit modelling*). The last approach consists of finding for each individual of the recipient file one or more donor individuals as similar as possible, and then in some way, transferring the values of Y variables to the *recipient* individual (*implicit modelling*). This method is known as *hot deck*, a term borrowed from data editing in data bases.

Validity of the imputation

We will say that a data fusion is valid if the fused data set $(X_1, \hat{Y}_1)$ is an instance of the distribution function $f(X, Y)$. In general the distribution function f is unknown, thus we are compelled to compare the empirical distribution functions $ef(X_1, \hat{Y}_1)$ with the $ef(X_1, Y_1)$.

We call the discrepancy between both distributions *matching noise*; following Paass (1985) the matching noise depends on the rightness of the imputation function $i(X)$ to approximate instances of the joint true distribution, which in the parametric case it depends on how well the assumed multivariate distribution represents the true data, and in the hot deck methodology depends on, as before, the assumed imputation model $i(X)$ and in addition to the existing discrepancies between the recipients and their corresponding donors.

However, whatever the imputation method chosen, imputed data is not like observed data, since it has inherent uncertainty, this is **the uncertainty problem**. Imputed values $\hat{Y}_1$ are estimates, thus, to be realistic, we need to take into account the variability of the imputed data when analyzing it. This variability comes from the random fluctuation of the distribution