Data Fusion in Several Algorithms

Stan Lipovetsky
GfK Custom Research North America
8401 Golden Valley Road, Minneapolis, MN 55427

Abstract

Data fusion consists of the process of integrating several datasets with some common variables, and other variables available only in partial datasets. The main problem of data fusion can be described as follows. From one source, having X^0 and Y^0 datasets (with N^0 observations by multiple x and y variables, n and m of those, respectively), and from another source, having X^1 data (with N^1 observations by the same n x-variables), we need to estimate the missing portion of the Y^1 data (of size N^1 by m variables) in order to combine all the data into one set. Several algorithms are considered in this work, including estimation of weights proportional to the distances from each i-th observation in the X^1 "recipients" dataset to all observations in the X^0 "donors" dataset. Or we can use a sample balancing technique with the maximum effective base performed by applying ridge-regression for the Gifi system of binaries obtained from the x-variables for the best fit of the "donors" X^0 data to the margins defined by each respondent in the "recipients" X^1 dataset. Then the weighted regressions of each y in the Y^0 dataset by all variables in the X^0 are constructed. For each i-th observation in the dataset X^0, these regressions are used for predicting the y-variables in the Y^1 "recipients" dataset. If X and Y are the same n variables from different sources, the dual partial least squares technique and a special regression model with dummies defining each of the three available sets are used for prediction of the Y^1 data.

Keywords: data fusion, distances, weighting, dual canonical correlations and PLS, ridge-regression, regression with dummies.

I. Introduction

Data fusion consists of integrating multiple data sources, some with variables common in different sources, and other variables available only in partial datasets, to achieve a synergy in the information acquired from all the sources. Such problems appear in various fields of data analysis and information fusion, for instance, in sensor observations (Jilkov and Li, 2009; Tian and Bar-Shalom, 2009; Carvalho and Chang, 2012), and databases combined into an integrated data knowledge source in economics and other applied statistical fields (Hastie et al., 2001; Moriariti and Scheuren, 2001; Rassler, 2002, 2004; Smarandache and Dezert, 2009). Various useful practical approaches have been recently suggested in marketing research (Kamakura and Wedel, 1997; Gilula et al., 2006; Bosch, 2011a,b; Collins, 2012; Maier, 2012).

This current paper considers several algorithms for practical implementation of data fusion needed in marketing research. A typical main problem of data fusion can be described as follows: from one source, having X^0 and Y^0 datasets (with N^0 observations by multiple x and y variables, n and m of those, respectively), and from another source, having

X^1 data (with N^1 observations by the same n x-variables), we need to estimate the missing portion of the Y^1 data (of size N^1 by m variables) for combining all the data into one set. Several algorithms are considered in this work. In one approach, the weights are taken proportionally to the distances from each i-th multivariate observation in the X^1 "recipients" dataset to each observation in the X^0 "donors" dataset. The distances are estimated with the pooled covariance matrix for Mahalanobis distance, or with Euclidean or Manhattan distance. Another approach is based on sample balancing technique with the maximum effective base (Lipovetsky, 2007, 2010a) performed by applying a ridge-regression model for the Gifi system (Gifi, 1990; Lipovetsky, 2012a) of binaries obtained from the x-variables for the best fit of the donors X^0 data to the margins defined by each respondent in the recipients X^1 dataset. With the obtained weights, the regressions of each y in the Y^0 dataset by all xs in the X^0 dataset are constructed, and predictions are made by each i-th observation in the X^1 "recipients" dataset for all y-variable for the Y^1 data segment. This procedure is repeated for each i-th observation in the X^1 dataset, filling all the missing positions in the Y^1 "recipients" dataset.

i	x_1 x_2 … x_n		y_1 y_2 … y_m
1 2 … N^0	donors dataset is given in matrix X^0		donors dataset is given in matrix Y^0
1 2 … N^1	recipients dataset is given in matrix X^1		target dataset to fill - matrix Y^1

In some cases, all data sets X^0, Y^0, and X^1 contain the same n variables from different sources, for instance, X^0 and Y^0 might correspond to panel data elicited via phone and internet interviewing in a previous year, respectively, while X^1 is the data elicited by phone in the current year, and the objective is to predict the Y^1 data from the internet under the new conditions. In such a case, the dual robust canonical correlation analysis (CCA), also called dual partial least squares (PLS), can be applied. The regular CCA considers two sets of x and y variables (n_1 and n_2 of those, respectively) with the same number of observations N from the same respondents. In contrast to this, the dual PLS technique is applicable when there are two groups of data with a different number of observations from different respondents obtained for the same set of variables, which is a special instance for a data fusion type problem. The dual PLS eigenproblem technique and predictions by the robust CCA are described in detail in (Tishler and Lipovetsky, 2000; Lipovetsky, 2012b). The current work considers another approach in which all three

available datasets are stacked in rows, and two dummy variables are added to identify the period of observations (zero for X^0 and Y^0, and one for X^1), and the source of the data (zero for phone X^0 and X^1, and one for the internet Y^0). Then we construct a regression model of each of the stacked variables by all others with the two additional binary variables. With these regressions, using X^1 data and the 1 values for the dummy variables, we can predict the Y^1 online values. This procedure is repeated for each of the *n* variables subsequently, getting all the Y^1 "recipient" variables predicted. It is easy to add the mixed-effects of each predictor with the dummy variables, increasing the precision of data fit.

II. Weighted by Distances Multivariate Regression

Suppose there are three data sets: from one source, we have got X^0 and Y^0 data sets (let us call them "Donors") with the N^0 responses (from the same respondents) by *n* variables $x_1^0, x_2^0,..., x_n^0$, and by *m* variables $y_1^0, y_2^0,..., y_m^0$; and from another source, we have X^1 data set (called "Recipients") with the N^1 responses by the same *n* variables $x_1^1, x_2^1,..., x_n^1$. The problem of data fusion consists in estimation of the data in a set Y^1 (called "Target") with the N^1 observations, like in X^1, by the same *m* variables as in Y^0.

Consider at first the following simple algorithm. Find the distances from one *i*-th row (multivariate point of observation) in X^1 to each of the rows in X^0, to define which of "donors" X_j^0 points lay in the vicinity of the *i*-th "recipient" X_i^1 point. For this task, we stack the sets X^0 and X^1 in rows, and find the total covariance matrix *B*, and its inverted matrix B^{-1}. Then using Mahalanobis distance in a general case of different scales of the variables we find distance d_{ij} from each *i*-th point in X_i^1 data to all points of X_j^0 data:

$$d_{ij}^2 = (X_i^1 - X_j^0)' B^{-1} (X_i^1 - X_j^0), \qquad (1)$$

where X_i^1 and X_j^0 are *n*-dimensional vectors of *i*-th recepient and *j*-th donor data sets. Weights of the distance from the *i*-th point to each of *j*-th points are defined by the proportions:

$$w_{ij} = d_{ij}^{-2} N^0 / \sum_{k=1}^{N^0} d_{ik}^{-2} \quad . \qquad (2)$$

If a distance equals zero the weight equals one. It is also convenient to use weights taken proportional to the $\exp(-d_{ij}^2)$, or to $1/(1+d_{ij}^2)$. With the weights defined for each current value of *i*, using the donors data, we construct the weighted regression models of each of the variables $y_1^0, y_2^0,..., y_m^0$ by all the predictors $x_1^0, x_2^0,..., x_n^0$. The coefficients of all these *m* models can be found by the matrix solution:

$$A^{(i)} = ((X^0)' W^{(i)} X^0)^{-1} (X^0)' W^{(i)} Y^0 \ , \qquad (3)$$

where $W^{(i)}$ is a diagonal matrix of the weights, and an identity column for having the intercepts can be added into design matrix X^0. The matrix $A^{(i)}$ contains in its columns the coefficients of regressions of each y^0 by all x^0 variables.

Taking an *i*-th row of the X_i^1 and multiplying it by matrix (3) we find m predicted Y_i^1 values for the target data:

$$Y_i^1 = X_i^1 A^{(i)} = X_i^1 ((X^0)' W^{(i)} X^0)^{-1} (X^0)' W^{(i)} Y^0 . \qquad (4)$$

Repeating the procedure (1)-(4) for each current row i=1, 2, 3., …, N^1, we produce all the predicted values for the matrix Y^1 of N^1 by m order.

When all the variables are measured in the same scale, instead of Mahalanobis we can use the Euclidean distance (putting $B=I$ identity matrix in (1)). It is also possible to split each variable into the Gifi system of binary variables. For instance, a variable measured on a 1 to 5 Likert scale is represented in five binary variables corresponding to each of the five levels. Then we can use Manhattan distance in the procedure described above.

III. Sample Balancing in Ridge Regression

Instead of weighting by distances (2), some works suggest to define weights by the equalizing

$$(X^0)'w = t\,, \tag{5}$$

where t is a vector of n-th order equal to an i-th row in the "recipients" matrix X^1. Also, the maximum of the effective sample size is required, so the weights w should minimize a measure like the variance

$$(w-\omega)'C(w-\omega) \to \min \tag{6}$$

of their deviation from a design vector ω of N^0 order, where C is a diagonal cost matrix of the same order. Then the conditional objective for the vector minimization is:

$$(w-\omega)'C(w-\omega) - \lambda[(X^0)'w - t] \to \min \tag{7}$$

where λ is the n-th order vector of the Lagrange terms. Minimizing (5) by w yields the solution $w = \omega + C^{-1}X^0\lambda$, so multiplying it by the matrix X^0 and using the restrictions $(X^0)'w = t$ yields the solution (Bosch, 2011a,b):

$$w = C^{-1}X^0[(X^0)'C^{-1}X^0]^{-1}t + \left(I - C^{-1}X^0[(X^0)'C^{-1}X^0]^{-1}(X^0)'\right)\omega. \tag{8}$$

where I is the identity matrix of N^0 order. The cost matrix is generally needed for tweaking some negative values in the weights to make them zero or positive, but mostly it is taken as a uniform matrix, so $C^{-1} = I$. The vector of design weights is usually taken as a uniform vector, $\omega = e$, of N^0 order, so the expression (8) reduces to the following:

$$w = e + X^0[(X^0)'X^0]^{-1}\left(t - (X^0)'e\right), \tag{9}$$

and the weights are distributed around 1.

The main problem with solutions (8)-(9) is that being constructed with the requirement of the exact fulfillment of multiple restrictions (5) it can easily produce negative weights, which requires additional adjustments by changing the elements in C through a heuristic attempts. Instead of requiring exact restrictions' fulfillment (5) in the objective (7), it makes sense to minimize the deviations from those restrictions, subject to only one condition of the effective base. It means that in place of objective (7) we can consider an alternative problem as follows:

$$\left\|(X^0)'w - t\right\|^2 + q[(w-\omega)'C(w-\omega)] \to \min\,, \tag{10}$$

with only one Lagrange multiplier q. Taking the derivative of this least squares objective and putting it to zero yields the solution:

$$w = \left(X^0(X^0)' + qC\right)^{-1}\left(X^0t + qC\omega\right). \tag{11}$$

With the uniform matrix $C=I$ and vector $\omega = e$, the problem (11) reduces to:

$$w = \left(X^0 (X^0)' + qI\right)^{-1}\left(X^0 t + qe\right), \quad (12)$$

which is a generalized ridge-regression solution (Lipovetsky, 2010b). Increasing the positive value of the profiling parameter *q* always guarantees reaching the non-negative weights.

The techniques of ridge-regression for finding positive weights while simultaneously keeping the maximum possible effective sample base are given in detail in (Lipovetsky, 2007, 2010a). Using the Chi-squared criterion in place of the least squares (10) leads to a solution which generalizes (11) to the following expression:

$$w = \left(X^0 D^{-1} (X^0)' + qC\right)^{-1}\left(X^0 D^{-1} t + qC\omega\right), \quad (13)$$

where the diagonal matrix $D = diag(\tilde{x})$, and $\tilde{x}$ is defined as the total $\tilde{x} = (X^0)'e$. With the uniform matrix $C=I$ and vector $\omega = e$, the problem (13) corresponds in other notation to the expression obtained in (Lipovetsky, 2007, the formulae (14)-(19) within). As it is shown in that work, the expression (13) can be further simplified to a form similar to (9) where the weights are distributed around 1:

$$w = e + X^0\left((X^0)' X^0 + q\,diag(\bar{X}^0)\right)^{-1}\left(t - \bar{X}^0\right). \quad (14)$$

Instead of the variables in X^0 and X^1 data sets, it is possible to split each variable into the Gifi system of binary variables. Using the Gifi system, we apply the sample balancing procedure to find weights for the best fit of the "donors" X^0 data to the margins defined by each respondent in the "recipients" X^1 data. Sample balancing with the maximum effective base performed by the ridge-regression estimation produces weights with which the same procedure (3)-(4) described in Section II is applied for predicting Y^1.

IV. Dual PLS analysis

If the "donors" data X^0 and Y^0 are obtained from the same respondents, instead of weighted regressions of each *y*-variable by all *x*-variables, as it is described above, we can consider canonical correlation analysis (CCA) for finding the multivariate regressions of all *y* by all *x* variables simultaneously. However, the data for X^0 and Y^0 can be obtained from different sources and contain a different number of observations, N_X^0 and N_Y^0. If both *X* and *Y* data consist of the same variables (measured from different sources, data bases, or in a different moment in time, etc.) it is possible to consider the problem of data fusion in another approach, namely, via the dual robust canonical correlation analysis (CCA), which can also be called the dual partial least squares (PLS) technique. In a regular CCA, we have the same number of observations elicited from the same respondents by two datasets with different sets of variables. We can consider an alternative situation – when there are two groups of data with a *different* number of observations from *different* respondents obtained for the *same* set of variables. This we will call the dual CCA, or dual PLS, and it can be reduced to a special eigenproblem of the product of correlation matrices of the two data sets. This technique can be further generalized to multiple data sets in an eigenproblem of block-matrices, and it also corresponds to the dual PLS (for more detail on this approach, see in Lipovetsky, 2012b). Let us consider it briefly.

Suppose there are two matrices *X* and *Y* of the orders $N_1 \times n$ and $N_2 \times n$, respectively. Here N_1 and N_2 are the number of observations in the first and second groups, and *n* is the number of the same variables in both datasets. It is also assumed that all the variables are standardized within each data set. The loadings *a* and *b* of the order *n* for each data set aggregated with the scores (weights) are:

$$a = X'\xi, \quad b = Y'\eta. \tag{15}$$

The straightforward CCA approach is usually inapplicable for the dual problem in (15), because in most cases the number of variables *n* is significantly less than the number of observations N_1 or N_2. This makes it impossible to invert the matrices like XX' and YY' of the order N_1 and N_2, respectively, while both are of rank *n*. But the robust CCA (also known as the PLS) approach can be used for the dual problem in (15) because it does not require matrix inversion.

With the covariance $a'b$ of loadings in (15), consider a conditional objective:

$$\xi'XY'\eta - \frac{\lambda}{2}(\xi'\xi - 1) - \frac{\mu}{2}(\eta'\eta - 1) \to \max. \tag{16}$$

Maximizing (16) by the vectors of scores yields a system of equations:

$$XY'\eta - \lambda\xi = 0, \quad YX'\xi - \mu\eta = 0. \tag{17}$$

Multiplying ξ' and η' by the first and second equation in (17), respectively, and using the normalizing conditions from (16), yields equality in the Lagrange terms $\lambda = \mu = \xi'XY'\eta$. Then with (15), the system (17) can be represented as follows:

$$Xb = \lambda\xi, \quad Ya = \lambda\eta. \tag{18}$$

These relations connect the loadings of one group with the scores of the other. Multiplying the transposed matrices *X* and *Y* by the first and second equation (18), respectively, yields the system:

$$X'Xb = \lambda X'\xi, \quad Y'Ya = \lambda Y'\eta. \tag{19}$$

The product of the transposed matrix by the initial matrix of the standardized data equals the matrix *R* of correlations, so using notations (15) let us represents system (19) as:

$$R_{xx}b = \lambda a, \quad R_{yy}a = \lambda b. \tag{20}$$

Substituting vector *a* from the first into the second equation in (20), and similarly with vector *b*, yields the eigenproblems:

$$(R_{xx}R_{yy})a = \lambda^2 a, \quad (R_{yy}R_{xx})b = \lambda^2 b. \tag{21}$$

The eigenvectors *a* and *b* present solutions for the dual PLS problem (15). With the *a* and *b* vectors, the scores ξ and η can be found using the relations from (18) by projecting observations onto the vectors of loadings. Both equations in (20) can be presented as one eigenproblem with the block-matrices:

$$\begin{pmatrix} 0 & R_{xx} \\ R_{yy} & 0 \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \lambda\begin{pmatrix} a \\ b \end{pmatrix}. \tag{22}$$

Multiplying the matrix in (22) by this equation itself yields the same solutions as in (21). Having the solution of eigenproblems (21) or (22), it is possible to predict matrix *Y* data by the *X* data in the technique described in (Tishler and Lipovetsky, 2000).

V. Multiple Regressions with Dummy Indices

For the case where we have the same *n* variables in all data sets used in fusion, let us describe a convenient regression model with dummy variables identifying each of the data sets. For an illustration, the *X* data sets correspond to phone, and the *Y* data sets correspond to online results gathered in a previous and a current periods of time (marked by zero and one upper indices) on the same *n* variables from a panel study. In this approach, all the available data sets can have a different numbers of observations, and we stack X^0, Y^0, and X^1

together in rows, creating $N_X^0 + N_Y^0 + N_X^1$ rows of the total data set. We also add a column for a dummy binary variable u={0,1} for identification of X versus Y (phone versus online) data, respectively, and a column v={0,1} for identification of the time moment in X^0 and Y^0 versus X^1 cases, respectively. The stacked data we can present in the following block-matrix:

$$Z = \begin{pmatrix} (X^0)_{N_X^0,n} & (0)_{N_X^0} & (0)_{N_X^0} \\ (Y^0)_{N_Y^0,n} & (1)_{N_Y^0} & (0)_{N_Y^0} \\ (X^1)_{N_X^1,n} & (0)_{N_X^1} & (1)_{N_X^1} \end{pmatrix}, \quad (23)$$

where the indices near each block show the order of the matrices and vectors. Let us denote the variables in the columns of the stacked matrices as z_1, z_2, …, z_n, and the last two columns of the binary variables we denote u and v, as before.

For standardized variables, a linear regression model of each variable z_j by all the other z-variables and u and v variables is as follows:

$$z_j = a_{j0} + a_{j1}z_1 + ... + a_{j,j-1}z_{j-1} + a_{j,j+1}z_{j+1} + ... + a_{jn}z_n + a_{j,n+1}u + a_{j,n+2}v, \quad (24)$$

where the first index denotes coefficients related to each j-th dependent variable (j=1, 2,…, n). A convenient way to produce simultaneously all the needed models (24) can be based on the inverted correlation matrix R^{-1} constructed from the data in the combined matrix (23). This matrix can be presented as following:

$$R^{-1} = \begin{pmatrix} (1-R_{1.}^2)^{-1} & -(1-R_{1.}^2)^{-1}a_{12} & -(1-R_{1.}^2)^{-1}a_{13} & ... & -(1-R_{1.}^2)^{-1}a_{1,n+2} \\ -(1-R_{2.}^2)^{-1}a_{21} & (1-R_{2.}^2)^{-1} & -(1-R_{2.}^2)^{-1}a_{23} & ... & -(1-R_{2.}^2)^{-1}a_{2,n+2} \\ - & - & - & - & - \\ -(1-R_{n+2.}^2)^{-1}a_{n+2,1} & -(1-R_{n+2.}^2)^{-1}a_{n+2,2} & -(1-R_{n+2.}^2)^{-1}a_{n+2,3} & ... & -(1-R_{n+2.}^2)^{-1} \end{pmatrix}. \quad (25)$$

The diagonal elements of this inverted correlation matrix (called variance inflation factors) equal the reciprocal values of the residual sums of squares in the regressions of each variable by all the rest of them, where $R_{j.}^2$ are coefficients of multiple determination in the models of each variable by all the others. The non-diagonal elements in a j-th row of R^{-1}, taken with the opposite signs and divided by the diagonal element in the same j-th row, coincide with the coefficients (24) of the regression of j-th variable by all the others (Kendall and Stuart, 1973). So with one matrix inversion we obtain all the coefficients of the needed n regression models (24) (and the last two rows of coefficients in (25) are not needed because they correspond to the models of the binary dummy variables by the other predictors). When the models (24) are constructed, we can predict the target values Y^1 taking the inputs from the "recipients" data X^1 as z-variable values in model (24) and u=v=1 for the dummy variables identifying the Y^1 data set. This procedure is repeated for each z_j model subsequently, so we get all n variables in Y^1 predicted.

The models (24) can be easily extended by adding mixed-effects for each predictor with the u and v dummy variables, which improves the quality of the data fit by using

three times more parameters of regression and is acceptable for large base sizes. Also, instead of linear models we can use logit and other regressions to accommodate for some special requirements of the predicted variables, for instance, their positive values. And the weighted regressions can be easily constructed as well.

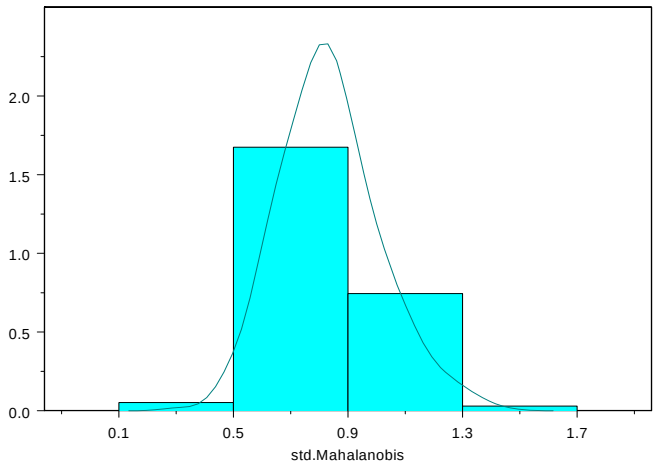
VI. Numerical results

For checking the described methods, the X^0 and Y^0 data sets of 250 observations with 30 and 51 variables, respectively, and the X^1 data set of 766 observations by 30 variables were used to estimate the $Y^1_{predicted}$ data of the size 766 by 51. The variables were measured on 1 to 5 Likert scales. In this test, the actual data set Y^1_{actual} was available, so the results were compared with it. The standard deviations (std) for the difference $(Y^1_{predicted} - Y^1_{actual})_i$ were estimated for each of the 766 observations. The weights for regressions were obtained in (1)-(2) estimation by Mahalanobis, Euclidean, and Manhattan (if for binary Gifi presentation of the variables) distances, and by the sample balance "Sambal" procedure (10)-(14). The descriptive statistics on these standard deviations (std) are given in Table 1. The mean values of the std are smallest for Manhattan and Sambal estimations (by the data presented as Gifi system). Example of such a distribution is shown in Figure 1.

Table 1. Summary statistics for std of prediction compared with actual data.

	Mahalanobis	Euclidean	Manhattan	Sambal
Min	0.298	0.318	0.292	0.264
1st Qu.	0.712	0.720	0.707	0.709
Mean	0.834	0.839	0.826	0.829
Median	0.821	0.826	0.816	0.819
3rd Qu.	0.943	0.943	0.932	0.935
Max	1.452	1.502	1.464	1.400
Std Dev.	0.177	0.176	0.177	0.180

Figure 1. STD distribution for the difference $(Y^1_{predicted} - Y^1_{actual})_i$ by 51 variables, estimated for each of the 766 observations.

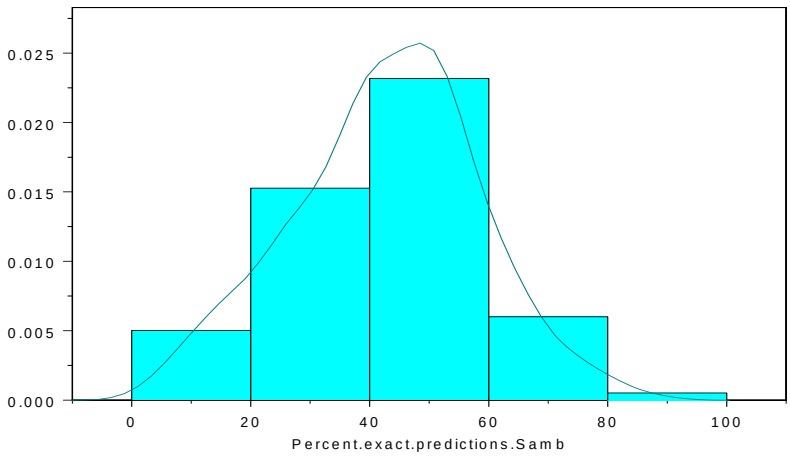


Another measure is the exactness of the prediction. By rounding the predicted values we consider a cross-table of the original and predicted values and estimate the percent of the coinciding values (the total on diagonal of such a table) to the total of 51 predicted values, for each of the 766 observation points. The descriptive statistics on this hit-rate are given in Table 2. We see that predictions are similar by all methods, and the highest hit-rate is achieved with the Sambal estimations. Example of such a distribution is shown in Figure 2.

Table 2. Summary statistics for percent of exact predictions with actual data.

	Mahalanobis	Euclidean	Manhattan	Sambal
Min	3.921	1.960	1.960	1.960
1st Qu.	33.333	33.333	33.333	33.333
Mean	42.627	42.796	42.530	42.814
Median	43.137	43.137	43.137	43.137
3rd Qu.	52.941	52.941	52.941	52.941
Max	88.235	90.196	92.156	92.156
Std Dev.	15.040	15.201	15.631	15.854

Figure 2. Hit rate of the correct predictions $Y^1_{predicted}$ coinciding with Y^1_{actual} by 51 variables, estimated for each of the 766 observations.



The second data set used for checking the methods is the data on pixels, 1060 observations, 366 variables. The "donor" data sets X^0 and Y^0 are of 142 observations with 49 and 317 variables, respectively, and the X^1 data set has 918 observations by 49 variables. They were used to find the $Y^1_{predicted}$ data of size 918 by 317. The variables are measured in continuous numerical scales in the range from 0 to 255 units. In this case, logit models were used giving positive prediction values. The results on std for the deviations of the predicted from control data are presented in Table 3.

The sample balance with maximum effective base technique produces the best results – the smallest means. In contrast to the methods of weight estimation by distances (2), the Sambal technique works 10 times faster (for instance, in this example, it takes 10 minutes versus about 2 hours for each of the distances weighting estimations).

Table 3. Summary statistics for std of prediction compared with actual Pixels data.

	Mahalanobis	Euclidean	Manhattan	Sambal
Min	22.983	22.983	14.909	8.820
1st Qu.	36.442	36.442	30.900	25.308
Mean	44.262	44.262	39.037	35.347
Median	43.366	43.366	37.767	33.750
3rd Qu.	50.102	50.102	44.836	42.037
Max	94.124	94.124	91.832	93.690
Std Dev.	10.870	10.870	11.378	13.384

For a numerical example in the approach (15)-(22) we can refer to (Lipovetsky, 2012). For the case of the same n variables in each of the data sets let us describe the regression models with dummy variables identifying each of the stacked sets (23)-(24). In this example, using data from a real research project on bank customer activity, 17 variables were considered, so each model (24) contains 19 parameters. The X data sets corresponded to phone, and Y data sets to online interviews from a panel study; X^0 and X^1 data sets have 128 and 1114 observations, respectively, and Y^0 has 1028 observations. The problem consists in evaluating the data for Y^1 and their comparison with the data in X^1. Multiple regression models (24) were constructed and prediction for the data set identified by the dummy variables u=v=1 performed. A comparison of the mean values for the variables in X^1 and in Y^1 is presented in Table 4. By the t-statistics and p-values it is easy to identify which variables of the phone and online studies could differ.

Table 4. Comparison of Phone versus Online data.

	Means by Phone data	Means Predicted Online	t-statistics	p-value
z_1	0.21	0.22	0.57	0.57
z_2	0.13	0.15	1.17	0.24
z_3	0.39	0.44	1.95	0.05
z_4	0.61	0.56	2.12	0.03
z_5	0.77	0.72	3.04	0.00
z_6	0.11	0.13	1.54	0.12
z_7	0.07	0.09	1.55	0.12
z_8	0.01	0.02	0.53	0.59
z_9	0.86	0.85	0.73	0.47
z_{10}	0.09	0.10	0.79	0.43
z_{11}	0.03	0.04	0.68	0.50
z_{12}	0.00	0.00	0.45	0.65
z_{13}	0.01	0.01	0.68	0.50
z_{14}	0.35	0.41	2.69	0.01
z_{15}	0.65	0.59	3.00	0.00
z_{16}	0.35	0.27	4.37	0.00
z_{17}	0.65	0.73	3.59	0.00

VII. Summary

Several algorithms for practical solutions to the problem of data fusion are presented (see also Lipovetsky, 2013). Weighting of the "donor" data due to the distances to each multivariate observation in the "recipient" data, and using weighted regression of one data set by another with the following prediction of the needed target variables are considered. The tried distances include Mahalanobis, Euclidean, and also Manhattan metrics for the binary presentation of the variables in the Gifi system. Another approach is based on a sample balancing procedure for finding weights of the "donor" data by their correspondence to the margins defined by the available "recipient" observations. The sample balancing procedure produces better results in prediction and works many times faster than the simple distances' weighting. For the case of the same variables obtained from different sources, the dual robust canonical correlations, or partial least squares analysis is described, together with the suggested technique of stacking all the datasets and using dummy variables for their identification. Such models present a convenient way of accounting for additional requirements in prediction; for instance, it is possible to use not linear, but logistic regressions for obtaining positive values. Numerical applications show that the considered approaches are simple, reliable, and can be applied in any available software. These techniques facilitate practical performance of data fusion and enrich a set of possible tools for various needs of data integration.

References

V. Bosch, Causal Inference with Linear Matching, JSM'11, Joint Statistical Meeting of the American Statistical Association, Session #395 – Advanced Statistical Methods for Marketing Research, Miami, FL, (2011a).

V. Bosch, Linear Fusion, IMSM'11, International Meeting of GfK Statisticians and Methodologists, NYC, NY, (2011b).

R.N. Carvalho, K. Chang, A Fusion Analysis and Evaluation Tool for Multi-Sensor Classification Systems, *J. of Advances in Information Fusion*, 7, 2 (2012), 1-12.

J. Collins, Dynamic Importance Weighting in Data Fusion, IMSM'12, International Meeting of GfK Statisticians and Methodologists, Warsaw, Poland, (2012).

A. Gifi, *Nonlinear multivariate analysis*. Chichester, England: Wiley, 1990.

Z. Gilula, R.E. McCulloch, P.E. Rossi, A Direct Approach to Data Fusion, *J. of Marketing Research*, XLIII, February (2006), 73-83.

T. Hastie, R.Tibshirani, J. Friedman, *The Elements of Statistical Learning*, New York, Springer, 2001.

V.P. Jilkov, X.R. Li, On Fusion of Multiple Objectives for UAV Search & Track Path Optimization, *J. of Advances in Information Fusion*, 4, 1 (2009), 27-39.

W.A. Kamakura, M. Wedel, Statistical Data Fusion for Cross-Tabulation, *J. of Marketing Research*, 34, November (1997), 485-498.

M.G. Kendall, A. Stuart, *The Advanced Theory of Statistics*, vol.2, New York, Hafner Publishing, 1973.

S. Lipovetsky, Ridge Regression Approach to Sample Balancing with Maximum Effective Base, *Model Assisted Statistics and Applications*, 2, (2007), 17 – 26.

S. Lipovetsky, Nonlinear Parameterization in Bi-Criteria Sample Balancing, *J. of Modern Applied Statistical Methods*, 9 (2010a), 198-208.

S. Lipovetsky, Enhanced Ridge Regressions, *Mathematical and Computer Modelling*, 51 (2010b), 338-348.

S. Lipovetsky, Regression Split by Levels of the Dependent Variable, *J. of Modern Applied Statistical Methods*, 11, (2012a), 319-324.

S. Lipovetsky, Dual PLS Analysis, *Int. J. of Information Technology & Decision Making*, 11, (2012b), 879-891.

S. Lipovetsky, Data Fusion in Several Algorithms, *Advances in Adaptive Data Analysis*, 5(3), 2013.

S. Maier, Data-Integrator – a New Toolbox to Enable Faster and More Efficient Data Fusion, IMSM'12, International Meeting of GfK Statisticians and Methodologists, Warsaw, Poland, (2012).

C. Moriariti, S. Scheuren, Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure, *J. of Official Statistics*, 17, 3 (2001), 407-422.

S. Rassler, *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*, New York, Springer, 2002.

S. Rassler, Data Fusion: Identification Problems, Validity, and Multiple Imputation, Austrian J. of Statistics, 33 (2004), 153-171.

F. Smarandache, J. Dezert (Editors), *Applications and Advances of DSmT for Information Fusion*, Vol. 3, American Research Press, Rehoboth, 2009.

X. Tian, Y. Bar-Shalom, Track-to-Track Fusion Configurations and Association in a Sliding Window, *J. of Advances in Information Fusion*, 4, 2 (2009), 146-164.

A. Tishler, S. Lipovetsky, Modelling and Forecasting with Robust Canonical Analysis: Method and Application, *Computers and Operations Research*, 27, (2000), 217-232.