Summary of Latent Variable Modeling for Statistical Data Combination Problem

Graduate School of Economics, Keio University

Masaki Mitsuhiro

Chapter 1
Introduction

To understand human behavior and the causal mechanism behind the phenomena, it is necessary to investigate the relationship between variables obtained simultaneously. Concerning data analysis, when observational data are obtained from the same unit such as consumers and companies, is called a "(quasi) single-source dataset." However, it is often difficult to obtain this single-source dataset, and instead we can obtain "multiple-source datasets." In these datasets, different sets of variables are observed for different datasets. The relationship between interest variables that are not observed simultaneously cannot be investigated with multiple-source datasets, but this problem is solved by integrating these datasets as a single-source dataset. For example, in the marketing field, the purchase history and demographic information are observed for consumers from big data such as ID-POS transactions, while web-ad exposure and demographic information are observed for different number of consumers from marketing research data. Although the web-ad exposure cannot be obtained only from the transaction data, when variables that are not observed at the same time are estimated for each unit, we can understand the relationship between the purchase history and web-ad exposure. Moreover, when missing values of each unit are estimated, the obtained single-source dataset can be used for marketing strategies such as segmentation.

Another example concerns integrating official statistical survey and purchase history data. Japanese government statistics are increasingly failing to capture the reality, as the time series survey framework has become less appropriate over time. In the family income and expenditure survey, studies have been conducted to correct for aggregate values such as the total purchase amount of the entire population using the purchase history held by companies. Integrating multiple-source datasets is also important to bring aggregate values with a large bias closer to the representative values of reality.

We focus on the approaches of statistical data fusion (Kamakura and Wedel, 1997) and statistical data combination in economics (Ridder and Moitt, 2007), which are methods based on a probabilistic model to integrate multiple-source datasets obtained from different units as a single-source dataset. When imputing of missing values for each unit, we can investigate the relationships between variables that are not observed at the same time and be useful for segmentation, recommendation, and customer management in marketing. In economics, a pseudo-panel approach has been proposed to make inferences concern-

ing missing outcome variables from cross-section data for each unit such as consumers, companies and regions (Browning et al., 1985; Moffitt, 1993). Although time series panel surveys can result in panel dropouts and high costs, the pseudo-panel solves this problem and helps in policy panning and decision making. This data combination problem is not limited to marketing and economics, but is common for data analysis in many other fields and several methods have been proposed so far.

Statistical matching, frequently used in practical usage in marketing, matches units from a certain dataset with similar units from another dataset using the distance of covariates between units. However, the prediction accuracy is insufficient owing to the attenuations of correlation, and it is expected to use many covariates to overcome this problem. To improve the accuracy of predicting unobserved variables, statistical modeling such as a regression model based on the Bayesian approach (Gilula et al., 2006) and the latent variable model (Cudeck, 2000; Kamakura and Wedel, 2000) has been proposed. The missing-data mechanism is incorporated into these statistical models and assume "missing at random (MAR)" missingness, but it does not support a "not missing at random (NMAR)" situation. The multiple-source datasets used in data analysis have selection bias owing to the different source, and data combination method applied to NMAR missing data is required. In this study, we describe several data combination problems and we propose four data combination methods to solve them.

Chapter 2
Data combination problem

In Chapter 2, we introduce existing methods used for the data combination problem and explain the relationship between the missing-data mechanism and the data combination methods.

In situations where a single-source dataset is not readily available, the missing problem must be considered when analyzing multiple-source datasets, and various techniques has developed in main related problem domains: data combination problem, split questionnaires, and mixed aggregate-disaggregate data. Among these problems, we focused on data combination, whose main purpose is to predict the missing values of units. Statistical data combination integrates multiple-source datasets obtained from different units into a (quasi) single-source dataset to perform the following: (1) estimate the relationships between the variables, which are not observed simultaneously in the same dataset, and (2)

impute the unobserved missing values for a practical purpose (e.g., for recommendation or customer management in marketing). This data combination problem is very common in applied data analysis in various fields.

The missing-data mechanism is divided into three types (Rubin, 1976). First type is missing completely at random, in which the missing data is random and the probability of being missing is independent of both observed and unobserved data. Second type is MAR wherein the probability of being missing depends only on the observed data and not on the unobserved data. Third type is NMAR wherein that the missing data is not MAR and the probability of being missing depends on both observed and unobserved data.

Although existing methods are known to be useful for MAR missing data, the multiple-source datasets obtained in marketing often have selection bias, and data combination methods for MAR missing data do not work well. It is necessary to understand the characteristics of observed data and consider the missing-data mechanism for the data combination problem. To see what kind of difference it can make on predictions obtained from MAR and NMAR missing data, we compared the results of applying the same methods to two patterns of two-source datasets created from the same single-source dataset. The regression-based model and Mahalanobis matching (Rubin, 1980) have poor prediction accuracy in the case of NMAR missingness and there are limitations to the use of data combination methods for MAR missing data.

Chapter 3
Kernel canonical correlation analysis and statistical matching

In Chapter 3, we propose a data combination method that combines extension of kernel canonical correlation analysis and kernel matching to integrate multiple-source datasets obtained from different units into a single-source dataset.

Statistical matching method matches units from a certain dataset with the same or similar units from another dataset according to the distance of each unit's covariate. By treating the pairs as if they are the same unit, the missing variable for one pair is replaced with the observed variable for the other pair. However, when multiple-source datasets have a large number of covariates, exact matching and Mahalanobis matching (Rubin, 1980) may not perform well. This is because the combination of values of covariates becomes

complicated, the relationships between the variables can be non-linear, and it is difficult to match units with the same or similar units. To solve this problem, various dimension reduction methods that can map the high-dimensional covariates to some composite variables with small dimensions have been developed, but they have not been utilized in the data-combination problem except for the linear factor models.

One multivariate analysis method with dimensional reduction is the canonical correlation analysis (Hotelling, 1936), which estimates canonical variables for two-variable sets. The kernel canonical correlation analysis that can extract the nonlinear relationship between two-variable sets (Akaho, 2001) has been proposed and requires single-source data. For example, these methods can be used when two survey datasets collected from the same respondents or two time-series datasets for the same points in time.

As another problem of statistical matching, it is difficult to predict unobserved variables in the part not overlapping in the distribution of covariates when using one-to-one matching. To solve this problem, kernel matching that assigns observed values weighted by the kernel of the propensity scores (Rosenbaum and Rubin, 1983), a specific function of the covariates, has been proposed (Heckman et al., 1998). Since this method involves many-to-one matching, it is possible to substitute values that have not been observed in the past. However, this method can only be applied to a one-dimensional covariate or propensity score, so this approach is not directly applicable for data combination.

To solve these problems of statistical matching for data combination, we propose canonical variable matching that combines extension of kernel canonical correlation analysis and statistical matching. Although the kernel canonical correlation analysis method is applied to two pairs of variable sets in two-source datasets, the proposed method can estimate canonical variables of a common low-dimensional space that can preserve the relationship between covariates and outcome variables. Our method will be more useful than existing statistical matching when two-source datasets have many covariates or there is a nonlinear relationship between covariates and outcome variables.

In a simulation study, we compared the proposed method with the traditional Mahalanobis matching method using simulated datasets that have many covariates and the nonlinear relationship between covariates and the outcome variables. The study confirmed that the prediction results of the proposed method were better than those of the existing methods, demonstrating the usefulness of the proposed method. Moreover, as a result of applying the proposed method to real-world data, the mean squared error (MSE) ratio and the correlation coefficients between the true values and the predictions were found to

be better than they are in Mahalanobis matching.

Chapter 4
Constrained matching via error minimization approach

In Chapter 4, we propose matrix-based constrained kernel matching in which marginal moments of the important covariates act as constraints.

The statistical matching methods described in the previous chapters are known as unconstrained matching and have no restrictions on matched units. When units from a certain data are associated to the same or similar unit from another data, this unconstrained matching allows for the most suitable assignment with respect to distance between units, and the observed values of the paired unit are imputed to the missing values. However, the distributions of the covariates or outcome variables is not always preserved before and after applying unconstrained matching method to two-source datasets. From a micro perspective, the units are assigned the same or similar units, but when it comes to a macro perspective, the moment information of the covariates or outcome variables may change.

To reduce the gap between the distribution of outcome variables before and after matching, constrained matching has been proposed (Rodgers, 1984) and this method imposes constraints on the marginal moments of outcome variables such as the mean and variance. This constrained matching method is many-to-many matching rather than one-to-one matching, and estimates weighted scores obtained from multiple units. The advantage of this matching method is that the distribution of the observed outcome variables is precisely replicated in the imputed dataset. However, it becomes more complex as the number of covariates increases because the matching algorithm needs to be constructed to satisfy the constraints.

For NMAR missing data, the prediction accuracy of the outcome variables is poorer than for MAR missing data shown in Chapter 2, and the distributions of the covariates or outcome variables before and after matching are likely to be different. For example, when integrating customer relationship management dataset for products that are popular among young users with survey dataset for all generations, the distribution of age variable used as covariates is expected to be different before and after applying the unconstrained matching method. Since one of the benefits of data combination methods is prediction

of individual missing values, we focus on many-to-one matching and consider imposing constraints on the statistical matching method. We extend unconstrained matching by simultaneously optimizing both the many-to-one matching results and the consistency of the mean of covariates before and after matching.

To make the mean of covariates close to their mean before many-to-one matching, we propose matrix-based constrained matching. This proposed method is a two-step approach that estimates the similarity matrix using unconstrained matching and a weights matrix by imposing constraints on the means of the covariates. The weights matrix is estimated using an optimization method with error minimization, which is commonly used in multivariate analysis. Our method can estimate the weights for the means of many covariates with the same updating formula, and it can also restrict the means of outcome variables.

We demonstrated the results of applying MCM with kernel matching to real-world data and confirmed the mean difference between the covariates before and after matching. From the prediction results, the mean difference of the proposed method is better than that of unconstrained matching. The MSE ratio of the proposed method is poorer than unconstrained matching, but for real-world data where the true of missing values is unknown, it is necessary to flexibly determine the positive parameters of the prediction accuracy and the correction of the distribution of covariates before and after matching according to the purpose of the analytics.

Chapter 5
Latent variable model with Gaussian process in MAR missingness

In Chapter 5, we propose a latent variable model with Gaussian process in MAR missingness to improve the prediction accuracy of missing values by avoiding parametric model assumptions.

In latent variable modeling for the data combination problem, the exploratory factor analysis model, which assumes latent factors behind observations, has been proposed (Kamakura and Wedel, 2000). This method incorporates the missing-data mechanism, and assumes a situation in MAR missingness for two-source datasets. Since the latent variable model assumes a common factor behind the observed variables, it can predict missing values for each unit by estimating the latent variables of each unit from the observed variables

even if a certain variable of a unit is missing. This model is a probability generation model based on the distribution of exponential distribution families, which is useful because of the ease of handling discrete and continuous variables. However, this latent variable model assumes MAR situation but does not support NMAR situation, and then the prediction results are affected by selection bias. Moreover, this assumes a linear mapping from latent variables to observed variables and requires parametric estimation, and deviations from the model's assumptions result in poor data fitting. To avoid such parametric model assumptions, we extend the latent variable model to a non-parametric model using the Gaussian process applied in the field of machine learning.

The Gaussian process is a probability distribution that generates a random function, and the function values obtained from the Gaussian process are used to capture the complex nonlinear relationships between variables. To reduce the dimensions of observed data, Gaussian process latent variable model (GPLVM; Lawrence, 2004) has been proposed, which applies this Gaussian process to principal component analysis, and consists of unsupervised learning. In these Gaussian process methods, the similarity between unit's latent variables contributes to the construction of the distribution of observed variables because the Gram matrix with latent variables is used as the covariance function. Using the distance between latent variables is analogous to incorporating the relational data of latent variables into the model and related to techniques such as matching methods between latent variables and graph embedding.

The Gaussian process modeling has been applied to general methods such as regression and classification model, and promising non-parametric Bayesian approaches to them have been provided. Moreover, the log Gaussian Cox process (Møller et al., 1998) and discriminative GPLVM (Urtasun and Darrell, 2007) have been proposed, and are also useful for discrete variables as well as continuous variables. We consider a simple and useful data combination method in a unified framework using widely applied Gaussian process.

We propose Gaussian process data combination model based on GPLVM for multiple-source datasets with mixed observed variables. This proposed method is a probabilistic generative model that estimates latent variables assumed behind observed variables, and generates discrete and continuous observed values from a latent function of the Gaussian process. Our method can predict the missing values of the outcome variables because the latent variables of all units are estimated even if all outcome variables are not observed. The missing-data mechanism is incorporated into this model and assumes a situation of MAR missingness like the exploratory factor analysis model.

Using simulation datasets generated from prior distribution, we compared the prediction results of Gaussian process data combination model for MAR missing data with that of the existing data combination methods and demonstrated the usefulness of the proposed method and the difference in their prediction accuracy. Our method with the Gaussian process is useful for mixed continuous and binary observed data with MAR missing values.

Chapter 6
Latent variable model with Gaussian process in NMAR missingness

In Chapter 6, we extend Gaussian process data combination model for MAR missing data and propose latent variable models with Gaussian process in NMAR missingness.

When integrating multiple-source datasets into a single-source dataset, the observed variables for different data sources often have bias of units obtained from different dataset. For example, large scale transaction data is observed only for users of a particular product or service, while small scale market research data is observed for users of various products and services. The population characteristics of these two-source datasets are naturally different. Although we proposed a latent variable model with Gaussian process for MAR missing data in the previous chapter, it is not suitable to apply this method to these datasets in the case of data combination like in this example. It is necessary to consider NMAR missingness, which is one of the missing-data mechanisms. In the field of labor economics, to adjust selection bias that occurred in the observational data, assuming a probit selection model for determining whether an outcome variable is observed or not, a two stage estimation with a regression model for interest variables has been proposed (Heckman, 1979). This adjustment of selection bias can estimate parameters that are closer to the true values than the values obtained using the regression model alone.

To adjust selection bias in the data combination problem, we focus on latent variable modeling and consider the situation of NMAR missingness where the probability of being missing depends on the latent variables. When dealing with NMAR missing data, the probability of being missing, which depends on both observed and unobserved data, cannot be ignored unlike with MAR missingness. We assume a probabilistic model for the probability of being missing as well as the observed variables. The probability of being missing is expressed by the discrete choice model with Gaussian process latent variables

and this is added to Gaussian process data combination model for MAR missing data as described in the previous chapter. The probabilistic models for both the observed variables and probability of being missing are extended to a non-parametric model that is represented in a unified framework with Gaussian process latent variables.

We propose Gaussian process data combination model for NMAR missing data. This method also consists of probabilistic models for both the observed variables and the probability of being missing. This model is represented in a unified framework by Gaussian process latent functions and generates discrete and continuous observed values from the estimated latent variables. By incorporating the missing-data mechanism into the latent variable model, it can be applied to NMAR missing data as well as MAR missing data.

We prepared simulation datasets assuming a situation of NMAR missingness and compared the prediction results of the proposed methods and the existing Mahalanobis matching. As a result, we can see the effect of the difference in the missing-data mechanism and in the use of Gaussian process latent function. Moreover, the proposed method is applied to real-world data, which are the corporate financial data and corporate brand survey data, and compared it with the data combination methods used in the simulation study. The prediction results show that the proposed method works effectively on units that are not well predicted by Mahalanobis matching due to selection bias. Our method proves to be useful for data combination of multiple-source datasets with selection bias.

Chapter 7
Conclusions

In this study, we focused on the data combination problem, which is treated as a common data analysis problem in various fields, and proposed several new methods to overcome existing data combination method problems. The proposed methods solved the data combination problem, and our methods were confirmed to have better performance than the existing methods through simulation studies and verification with real-world data.

First, we proposed statistical matching combining kernel canonical correlation analysis. Our matching method can estimate low-dimensional canonical variables that can preserve the relationship between covariates and outcome variables. These canonical variables match units with similar units more accurately than do existing matching methods. Second, we proposed matrix-based constrained matching. This method corrects for the difference in the mean of the covariates before and after matching by error minimization.

Third, we proposed a latent variable model with Gaussian process for MAR missingness. Our latent variable model can capture the non-linear relationships between variables using Gaussian process latent function and be applied to observed discrete and continuous variables. Finally, the latent variable model with Gaussian process for MAR missing data was extended, and we proposed a latent variable model with Gaussian process for NMAR missingness. This method is represented by a unified framework of Gaussian process latent function and NMAR missing data can also be handled by assuming the discrete choice model for the probability of being missing.

We consider that data combination methods are useful for integrating the multiple-source datasets as well as two-source datasets. In recent years, there is a growing need to integrate datasets obtained from various sources for data-driven decision making, but to understand and utilize the characteristics of multiple-source datasets, it is necessary to focus on the perspectives described above and select an appropriate data combination methods. Therefore, our proposed methods are expected to contribute to solving the data combination problem.

Reference

[1] Akaho, S. (2001). A kernel method for canonical correlation analysis. *In Proceedings of the International Meeting of the Psychometric Society*, arXiv:cs/0609071.

[2] Browning, M., Deaton, A., and Irish, M. (1985). A profitable approach to labor supply and commodity demands over the life-cycle. *Econometrica: journal of the econometric society*, 503–543.

[3] Cudeck, R. (2000). An estimate of the covariance between variables which are not jointly observed. *Psychometrika*, **65**, 539–546.

[4] Gilula, Z., McCulloch, R. E., and Rossi, P. E. (2006). A direct approach to data fusion. *Journal of Marketing Research*, **43**, 73–83.

[5] Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, 153–161.

[6] Heckman, J. J., Ichimura, H., and Todd, P. (1998). Matching as an econometric evaluation estimator. *The review of economic studies*, **65**, 261–294.

[7] Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, **28**, 321–377.

[8] Kamakura, W. A. and Wedel, M. (1997). Statistical data fusion for cross-tabulation. *Journal of Marketing Research*, **34**, 485–498.

[9] Kamakura, W. A. and Wedel, M. (2000). Factor analysis and missing data. *Journal of Marketing Research*, **37**, 490–498.

[10] Lawrence, N. D. (2004). Gaussian process latent variable models for visualization of high dimensional data. *Advances in Neural Information Processing Systems*, **16**, 329–336.

[11] Moffitt, R. (1993). Identification and estimation of dynamic models with a time series of repeated cross-sections. *Journal of Econometrics*, **59**, 99–123.

[12] Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log Gaussian Cox processes. *Scandinavian journal of statistics*, **25**, 451–482.

[13] Ridder, G. and Moffitt, R. (2007). The econometrics of data combination. In Heckman, J. and E. Leamer (eds.), *Handbook of Econometrics*, 6B (Ch. 75). North-Holland: Elsevier Science.

[14] Rodgers, W. L. (1984). An evaluation of statistical matching. *Journal of Business and Economic Statistics*, **2**, 91–102.

[15] Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.

[16] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581–590.

[17] Rubin, D. B. (1980). Bias reduction using Mahalanobis-metric matching. *Biometrics*, **36**, 293–298.

[18] Urtasun, R. and Darrell, T. (2007). Discriminative Gaussian process latent variable model for classification. *In Proceedings of the 24th international conference on Machine learning*, 927–934.