Which Brand Purchasers Are Lost to Counterfeiters? An Application of New Data Fusion Approaches

Yi Qian* **and Hui Xie†**

Forthcoming in *Marketing Science*

Summary

Firms and organizations often need to collect and analyze sensitive consumer data. A common problem encountered in such evidence-based research is that they cannot collect all essential information from one sample, and may need to link non-overlapping data items across independent samples. We propose an automated nonparametric data fusion solution to this problem. The proposed methods are not restricted to specific types of variables and distributions. They require no prior knowledge about how data at hand may behave differently from standard theoretical distributions, automate the process of generating suitable distributions that match data, and therefore are particularly useful for linking data with complex distributional shapes. In addition, these methods have strong theoretical support, permit highly efficient direct fusion to relate a mixture of continuous, semicontinuous, and discrete variables, and enable nonparametric identification of the entire distributions of fusion variables, including higher moments and tail percentiles. These novel and promising features overcome important limitations of existing methods and have the potential to increase fusion effectiveness. We apply the proposed methods to overcome data constraints in a study of counterfeiting. By combining datasets from multiple sources, data fusion provides a feasible approach to studying the relationship between counterfeit purchases and various marketing elements, such as consumers' purchase motivations, behaviors, and attitudes, brand marketing channels, promotions, and advertisements. Therefore, data fusion sheds light on counterfeit purchase behaviors and suggests ways to counter counterfeits that would not be available if these datasets were analyzed separately.

Key words: Counterfeit; CRM; Database Marketing; Nonparametric Method; Sensitive Data; Underground Economics

* Department of Marketing, Kellogg School of Management, Northwestern University. Email: yiqian@northwestern.edu.

† Division of Epidemiology and Biostatistics, School of Public Health, University of Illinois at Chicago. Email: huixie@uic.edu.

1. Introduction

Firms and organizations frequently need to collect sensitive data from consumers. In product marketing, consumption of counterfeit products, use of pirated digital products, web browsing behavior, fake reviews, and financial assets are examples of sensitive data collected from consumers. In social marketing sensitive data also abound, including those on issues such as smoking, substance use, patient health, criminal acts, abortion, voting, energy conservation, environmental protection, and charitable donations. Data about these sensitive issues are vital for firms and organizations to understand consumer behaviors and to evaluate the effectiveness of their marketing strategies to promote commercial products and social goods.

Studies focusing on these sensitive topics are often challenging to conduct. A common problem that greatly hinders evidence-based business research and practice in these areas is that researchers often find themselves having no joint observations on these sensitive variables and other key variables in a single-source dataset, despite the purpose to study relationship between these variables. This situation can arise for various reasons. In some cases, a single-source comprehensive dataset that includes all essential variables is unavailable, prohibitively expensive to collect, or uninformative because of a small sample. For example, in media planning and targeting, although data on media usage and on sensitive product usage and behaviors (e.g., smoking, voting behavior, purchases of environmentally friendly goods) are readily available from independent survey samples, a single-source dataset containing both sets of variables for the same sample of consumers is typically unavailable. The situation could also arise when there are concerns that collecting sensitive data together with all other relevant data may introduce a bias. For example, firms may prefer to ask sensitive questions about counterfeit consumption in a survey separately from other questions (such as attitudinal questions on authentic products, shopping habits, lifestyles) because of concern that responses to the sensitive questions could affect or be affected by responses to the other questions[1]. In some other cases, the procurement and creation of a comprehensive database are legally constrained or even *prohibited* when sensitive respondent data are involved. In modern legal, business and societal environment, there are increasingly strong

[1] In this respect, Pouta (2004) documents that attitude and belief questions asked in the same survey are a source of context effects that influence responses to questions about willingness to pay for environmental goods.

public concerns, data privacy laws and regulations that limit or even prohibit sharing sensitive respondent data. The data privacy concerns can limit data availability in various ways and contexts, and have a range of important implications for database marketing (Blattberg et al. 2008). For illustration purpose, we quote a concrete example (Winer 2001)

"$\cdots$ *These [privacy] concerns have received more prominence. The defining moment in web privacy occurred in 1999 when the Web ad serving company Doubleclick announced that it was acquiring the direct marketing database company, Abacus Direct, with intentions to cross-reference Web browsing [that many consider sensitive data] and buying behavior with real names and addresses. The public outcry was so strong that Doubleclick had to state that it would not combine information from the two companies.*"

As aforementioned, in many cases an ideal dataset with all essential variables is absent or even prohibited when sensitive data are involved. Instead often available to researchers and policymakers are multiple datasets, each of which contains a different set of essential variables collected from independent samples. In these situations an attractive alternative is to link sensitive data to other relevant variables collected from different samples by using a set of linkage variables common to these datasets (*aka* data fusion). The idea is to match non-overlapping data items from *similar* consumers when matching these data from *same* consumers is impossible. Global marketing research leaders, such as Nielsen in US and Gfk in Europe, have been using data fusion to link data on sensitive patient health information, media and marketing information across different sources[2]. Data fusion has also been used by organizations and agencies to inform public policy through facilitating policy microsimulation and the analysis of economic, social or public health programs (Rassler 2002).

In this article, we propose automated and efficient nonparametric solutions to data fusion problems. These methods are not tied to specific types of variables and distributions, and are useful for linking a mixture of discrete, semicontinuous, and continuous variables that are frequently encountered in marketing databases. In these situations, these nonparametric methods can automatically adapt to complex distributional shapes, require minimal efforts from researchers in specifying proper data distributions, and ensure that fusion results are not artifacts driven by the distributional assumptions imposed on these variables. As a result, these new methods have the potential to improve the accuracy and efficiency of data fusion by overcoming important limitations of existing methods.

[2] E.g., see //www.nielsen.com/us/en/insights/press-room/2007/Nielsen_Introduces_First_Suite_of_NielsenConnect_Services.html.

We apply the methods to overcome data limitation issues when studying consumer counterfeit purchase behaviors. Product counterfeiting has become increasingly pervasive. The magnitude of counterfeiting is larger than the national GDPs of 150 economies (OECD 1998). It has great impacts on nearly all product sectors, ranging from apparel (Qian 2008, 2011) to software (OECD 1998). Despite the increasing prevalence and large impact of counterfeiting, it remains an under-researched area. One of the greatest challenges is the scarcity of detailed and comprehensive consumer-level data, partly due to the underground and sensitive nature of counterfeiting. This paper seeks to fill in the gap by combining datasets from an authentic firm's surveys and internal databases. Although none of the available datasets alone contains all the variables that allow one to investigate how various marketing mix elements relate to counterfeit purchase behaviors, the proposed approach provides a feasible way to relate two sets of variables so that such questions can be examined. Because of their ability to non-parametrically match informative and complex data, the proposed fusion methods improve the identification of consumer behavior patterns in counterfeit consumption, and individual predictions. Consequently, they offer the opportunity for better managerial decisions, such as more effective and accurate anti-counterfeit planning and targeting.

2. Data Fusion Problems and Related Work

Data fusion seeks to relate variables collected from independent samples. Figure 1.1 (a) illustrates the classical data fusion scenario with two independent samples, A and B. Typically the goal is to estimate the marginal relationships between unique variables Y_A and Y_B even if there is no observation on their joint distribution in the concatenated dataset. Data fusion is generally made possible by a set of common variables observed in both datasets, Y_C, which may include demographic and psychological variables. A key working assumption in data fusion is the conditional independence assumption (CIA), which states that the two sets of unique variables, Y_A and Y_B, are independent, given the common variables Y_C. The CIA can be made reasonable if a rich set of relevant common variables are used in fusion.

One conventional method to solve data fusion problems is the hot-deck procedure. It uses a set of rules to select a matching donor, based on common variables, for a recipient whose unique variables are unobserved and need to be filled in. Despite its notable merits, hot-deck suffers several important drawbacks because of its heuristic nature (Kamakura and Wedel

1997). The matching step, a key element of the procedure, is often heuristic and can involve subjective decisions in many aspects. For example, different rules (e.g., different collapsing of levels of categorical matching variables and/or excluding and including of these variables) can be used for different recipients in order for each of them to find an exact donor match. Different metric-matching criteria can be used for continuous matching variables with little knowledge about which one provides optimal matching. These issues are due to the implicit modeling in hot-deck, which makes it hard to identify the underlying modeling assumptions. As a result, the literature on the theoretical properties of the procedure is sparse. Another drawback of implicit modeling is the difficulty in properly quantifying sampling variability associated with fusion results. Model-based methods have been developed to overcome these drawbacks. Kamakura and Wedel (1997, 2000) develop a class of novel fusion procedures based on a finite mixture or a factor model. These methods model all available data (i.e., both the common and unique variables) and can thus be considered a full-information (FI) approach. Gilula et al. (2006) propose a limited-information (LI) direct fusion approach that models only the conditional distribution of those unique variables. Despite considerable progress made in combining multiple-source datasets, more powerful data fusion methods are needed. As noted by Kamakura et al. (2005), "existing [data fusion] methods are strongly dependent on distributional assumptions, and nonparametric approaches are called for." In addition, further research on more efficient and flexible methods is needed. In this work we propose a new class of data fusion methods that tackle these important issues as follows.

(1) *Effective and theoretically sound nonparametric solution to data fusion.*[3] Both hot-deck and our procedure are nonparametric in that neither requires specifying distributional forms for any variable in the datasets. Both can be viewed as assigning probability weights (either implicitly or explicitly) to the set of observed values and using that distribution for prediction. On the other hand, as will be shown in the next section, the weights in our method are determined by principled and coherent rules from statistical models with theoretical justifications and thus our method can overcome major drawbacks of hot-deck,

[3] This benefit is most useful for linking variables that have unknown complex distributional forms, but is minimal for relating binary variables, which have the simplest distributional form. Furthermore, the performance gain due to the nonparametric merit depends on applications and the complexity of underlying phenomena. In our empirical application we only observe moderately strong, but not overwhelming, evidence for the superiority of the proposed methods for a range of fusion analyses and managerial implications.

including its ad hoc nature in critical steps of data fusion, lack of theory support, invalid statistical significance level, poor sampling properties, and difficulty in handling missingness in matching variables. Importantly, our method simultaneously retains some of the main merits of hot-deck contributing to its popularity, such as imposing no distributional assumptions and automatically satisfying the boundary requirements of bounded fusion variables. The end result is a more effective nonparametric fusion method outperforming hot-deck.

(2) *High efficiency.* Our methods improve the efficiency of data fusion in three ways. As nonparametric fusion methods, they obviate the need to identify proper distributions in a variable-by-variable fashion. Second, our approach extends the efficient direct fusion approach (Gilula et al. 2006) to a much broader range of applications to relate a mixture of continuous, semicontinuous, and discrete unique variables. Our direct methods overcome two important difficulties in such extension by (1) eliminating the dependence on strong distributional assumptions and (2) providing closed-form joint predictive distributions even if fusion variables are continuous. Last, our methods require no modeling of completely observed common variables, thereby increasing the efficiency and robustness of data fusion.

A recent work by Qian and Xie (2011) develops a distribution-free Bayesian approach that overcomes the limitations of Chen (2004) and can handle high-dimensional missing covariate problems frequently seen in business applications. As depicted in Fig 1.2, their methods address a different class of problems in which data are from a single source with the regression response variable Y_B completely observed and the covariates Y_A partially observed[4]. The use of their methods requires some observations jointly observed on Y_A and Y_B, which provide essential information for model identification. Thus as it stands, their methods are not applicable for typical data fusion problems with no such joint observation[5]. Our objective here is to develop more effective nonparametric procedures that can solve data fusion problems, and these procedures do not require using missing covariate methods when common variables are fully observed.

[4] This is also true when applying their methods to multiple source data, e.g., to that in Feit et al. (2010).

[5] In this case, because the regression coefficients are inestimable and their posterior distribution equals prior (Rubin 1974), running these missing covariate methods encounters model unidentifiability issue.

3. An Efficient Adaptive Data Fusion Framework

In this section we introduce a robust and flexible odds ratio modeling framework for data fusion. Unlike parametric modeling, the odds ratio model is capable of modeling a joint distribution of highly disparate data elements without making prior restrictions on their distributions in data fusion. Although our fusion approach can be made more general, we restrict our attention here to the leading situation in which there is no joint observation on unique variables Y_A and Y_B (Figure 1.1.a and c). We consider fusion using the commonly used conditional independence assumption (CIA), under which $f(Y_A, Y_B|Y_C) = f(Y_A|Y_C)f(Y_B|Y_C)$. Let $(Y_1, \cdots, Y_k)$ be the set of common variables in Y_C. We first consider how to model the conditional distribution $f(y_A|y_k, ..., y_1)$, where y_A could be either continuous or discrete. As shown in Chen (2004), the conditional distribution can be rewritten as

$$f_{\theta_A}(y_A|y_k, ..., y_1) = \frac{\eta(y_A, y_{A0}; y_k, ..., y_1, y_{k0}, ..., y_{10})f(y_A|y_{k0}, ..., y_{10})}{\int \eta(y_A, y_{A0}; y_k, ..., y_1, y_{k0}, ..., y_{10})f(y_A|y_{k0}, ..., y_{10})dy_A}, \quad (1)$$

where θ_A denotes model parameters and $(y_{10}, \cdots, y_{k0}, y_{A0})$ is an arbitrarily-chosen fixed point in the sample space of $(Y_1, \cdots, Y_k, Y_A)$. The main point of Eqn (1) is to reexpress the conditional distribution as a function of two component functions, which can then be modeled separately[6]. The first component function, $\eta(y_A, y_{A0}; y_k, ..., y_1, y_{k0}, ..., y_{10}) = \frac{f(y_A|y_1,\cdots,y_k)f(y_{A0}|y_{10},\cdots,y_{k0})}{f(y_A|y_{10},\cdots,y_{k0})f(y_{A0}|y_1,\cdots,y_k)}$, is the odds ratio function relative to the reference point $(y_{10}, \cdots, y_{k0}, y_{A0})$ and captures the dependence of Y_A on $(Y_1, \cdots, Y_k)$. It is a constant (one) if Y_A is independent of $(Y_1, \cdots, Y_k)$. The second component function, $f(y_A|y_{k0}, ..., y_{10})$, is the density function of Y_A at the fixed reference point. To motivate the modeling approach for these two component functions, consider the generalized linear models (GLMs) whose density function is

$$f_\theta(y_A|y_k, \cdots, y_1) = \exp\left\{\frac{y_A\Psi(\beta, y_k, \cdots, y_1) - b(\Psi(\beta, y_k, \cdots, y_1))}{a(\tau)} + c(y_A, \tau)\right\}, \quad (2)$$

where Ψ is the canonical parameter as a function of regression parameter β and with canonical link functions $\Psi(\beta, y_k, \cdots, y_1) = \beta_0 + \beta_1 y_1 + \cdots + \beta_k y_k$; functions $b(\cdot)$ and $c(\cdot, \cdot)$ determine a particular distribution in the exponential family; $a(\tau) = \tau/w$, where τ is the dispersion

[6] Chen (2004) shows because the two component functions are variation independent, theoretically the choice of the reference point can be arbitrary. Computationally, choosing a reference point closer to the center of the distribution may lead to faster computation.

parameter and w is a known weight. The odds ratio function for the GLMs is

$$\eta(y_A, y_{A0}; y_k, ..., y_1, y_{k0}, ..., y_{10}) \;=\; \exp\left\{\sum_{j=1}^{k} \frac{\beta_j}{a(\tau)}(y_A - y_{A0})(y_j - y_{j0})\right\}. \qquad (3)$$

The function $f(y_A|y_{k0}, \cdots, y_{10})$ becomes $\exp\left\{\frac{y_A\Psi(\beta, y_{k0}, \cdots, y_{10}) - b(\Psi(\beta, y_{k0}, \cdots, y_{10}))}{a(\tau)} + c(y_A, \tau)\right\}$ which is of a parametric functional form determined by the functions $b(\cdot)$ and $c(\cdot)$. A drawback of using GLMs for data fusion is its dependence on strong distributional assumptions. To overcome this drawback and enhance modeling robustness, notice that $f(y_A|y_{k0}, \cdots, y_{10})$ conditions on the fixed reference point and thus behaves like a marginal distribution. By analogy to using the empirical distribution to estimate a marginal distribution, below $f(y_A|y_{k0}, \cdots, y_{10})$ is modeled nonparametrically like a marginal distribution. Specifically, let $(y_{A1}, ..., y_{AL_A})$ be the unique observed values in the dataset for Y_A. A nonparametric model for $f(y_A|y_{k0}, \cdots, y_{10})$ assigns probability mass $p_A = (p_{A1}, \cdots, p_{AL_A})$ on these unique data points as

$$Prob(Y_A = y_{Al}|y_{k0}, \cdots, y_{10}) = p_{Al},\ l = 1, \cdots, L_A,\ \text{subject to} \sum_{l=1}^{L_A} p_{Al} = 1, \text{ and } 0 < p_{Al} < 1, \forall\, l. \qquad (4)$$

Like the empirical marginal distribution estimates, a reasonably large sample size is needed so that the observed values cover the important range of the sample space. The modeling strategy is therefore well suited for database marketing which typically has a large sample size. To relax the constraint in Eqn (4), we reparameterize p_A as $\lambda_A = (\lambda_{A1}, ..., \lambda_{AL_A})$, such that $\lambda_{Al} = \ln(p_{Al}/p_{AL_A})$ for $l = 1, \cdots, L_A$. Thus, $p_{Al} = \exp(\lambda_{Al})/\sum_{u=1}^{L_A}\exp(\lambda_{Au})$. Motivated by Eqn (3) from the GLMs example, the log-odds-ratio function, $\ln \eta_{\gamma_A}(y_A, y_{A0}; y_k, ..., y_1, y_{k0}, ..., y_{10})$, is modeled in a bi-linear form as

$$\sum_{v=1}^{k}\sum_{m=1}^{M} \gamma_{Avm}(y_A - y_{A0})(y_v - y_{v0})^m + \sum_{v=1}^{k-1}\sum_{u=v+1}^{k} \gamma_{Avu}(y_A - y_{A0})(y_v - y_{v0})(y_u - y_{u0}). \qquad (5)$$

The above odds ratio function includes higher-order and interaction terms to model complex nonlinear relationships. Categorical common variables can be included as a set of dummy variables. As evident from Eqn (3), the GLMs have a log-bilinear form of odds ratio function, and the log-odds parameter γ_A is a reparametrization of the parameters in the GLMs. Like GLMs, the parameters in the odds ratio function can be estimated and tested using likelihood-based inference. On the other hand, the marginal-like distribution,

$f_{\lambda_A}(y_A|y_{k0},\cdots,y_{10})$, is modeled nonparametrically in the odds ratio model but parametrically in the GLMs. Therefore, the odds ratio model nests the commonly used parametric GLMs as special cases by eschewing their distributional assumptions. The above odds ratio model for the conditional density function $f_{\theta_A}(y_A|y_k,...,y_1)$ assigns point mass to the set of the observed values $(y_{A1},\cdots,y_{AL_A})$ according to the following probability mass function

$$f_{\theta_A}(y_A|y_k,...,y_1) = \frac{\sum_{l=1}^{L_A} 1_{\{y_A=y_{Al}\}}\eta_{\gamma_A}(y_{Al},y_{A0};y_k,...,y_1,y_{k0},...,y_{10})\exp(\lambda_{Al})}{\sum_{l=1}^{L_A}\eta_{\gamma_A}(y_{Al},y_{A0};y_k,...,y_1,y_{k0},...,y_{10})\exp(\lambda_{Al})}. \quad (6)$$

Prediction or imputation is simple to perform from this discrete distribution on a finite number of data points. More importantly, as shown above, the odds ratio model nests the GLMs by eschewing their distributional assumptions, and thus provides more robust prediction or imputation by making no prior restrictions on the distributional forms of unique variables in data fusion. The above odds ratio model is also used for $f_{\theta_B}(y_B|y_k,\cdots,y_1)$.

We develop four approaches for data fusion using the above modeling framework. One group uses a direct estimation (DE) approach to data fusion and the other uses multiple imputation (MI). Each group has an FI and LI version, resulting in a total of four approaches. We first describe the two LI approaches (LI-DE and LI-MI) that are designed for the classical data fusion problem as shown in Figure 1.1 (a).

(1) *LI-DE.* Unlike MI, DE does not view data fusion as a missing data problem and needs no imputation of the unobserved unique values. It directly estimates the joint distribution of only the unique variables, which is $f(Y_A,Y_B|D)=\int f(Y_A,Y_B,\theta|D)d\theta$, where θ collects the model parameters, $D=(D_A,D_B)$, $D_A=(y_{iA},y_{iC}^A), i=1,\cdots,N_A$, denotes the N_A observations on (Y_A,Y_C) in dataset A, and $D_B=(y_{iB},y_{iC}^B), i=1,\cdots,N_B$, denotes the N_B observations on (Y_B,Y_C) in dataset B. With CIA, this joint distribution under our LI-DE is

$$f(Y_A,Y_B|D) = \int\int\left[\int f_{\theta_A}(Y_A|Y_C)f_{\theta_B}(Y_B|Y_C)f_{\theta_C}(Y_C)dY_C\right]f(\theta_A,\theta_B|D)d\theta_A d\theta_B,$$
$$\text{where } Y_A\in(y_{A1},\cdots,y_{AL_A}), Y_B\in(y_{B1},\cdots,y_{BL_B}), \quad (7)$$

and $f(\theta_A,\theta_B|D)$ is the posterior distribution of the parameters. Online Appendix A.1 provides an MCMC algorithm to obtain draws from this posterior distribution, which can then be used to evaluate the integration with respect to $f(\theta_A,\theta_B|D)$. With a large sample, a simpler approach is to condition on the estimate of (θ_A,θ_B), rather than integrating them

out, i.e., $f(Y_A, Y_B|D) = \int f_{\hat{\theta}_A}(Y_A|Y_C) f_{\hat{\theta}_B}(Y_B|Y_C) f_{\theta_C}(Y_C) dY_C$. The estimate $(\hat{\theta}_A, \hat{\theta}_B)$ can be obtained using the MLE method, for which Online Appendix A.2 provides an estimation algorithm. When sample size is large, the conditional approach and the fully Bayesian approach produce close fusion results. With a large sample, the integration with respect to $f_{\theta_C}(Y_C)$ can be replaced by the summation over the observations of the common variables Y_C in the datasets in order to avoid modeling Y_C. Note that as shown in Eqn (7), even if the unique variables are continuous, our DE approach assigns probability mass only on the observed values of Y_A and Y_B, and the joint distribution can be readily evaluated even if it has irregular distributional shapes. With this simple joint distribution form, one can perform various fusion analyses, such as analyses on moments, quantiles and even the distribution function itself, in a relatively straightforward manner.

(2) *LI-MI.* We next consider data fusion via multiple imputation (MI). Although MI requires additional work on storing and processing multiple imputed datasets, it can be useful for building a fused database or when the intended analysis is unknown at the time of fusion. An important benefit of our MI approaches is that the fusion model nests the GLMs commonly used to analyze fused datasets. In MI, we stack datasets A and B together to form a concatenated file with the resulting data matrix denoted as $Y = (Y_A^{obs}, Y_B^{obs}, Y_A^{mis}, Y_B^{mis}, Y_C^{obs})$, where $(Y_A^{obs}, Y_B^{obs}, Y_C^{obs})$ and (Y_A^{mis}, Y_B^{mis}) collect the observed and missing entries in the data matrix, respectively. When considered as a missing data problem, a natural approach to data fusion is to impute multiple plausible values for those unobserved entries from their predictive distribution given the observed data. The imputed datasets on the joint distribution of Y_A and Y_B can be analyzed using standard methods, and results can be pooled to produce a single inference using Rubin's combination rules. Under CIA, the posterior predictive distribution for unobserved data, $f(Y_A^{mis}, Y_B^{mis}|Y_A^{obs}, Y_B^{obs}, Y_C^{obs})$, is

$$\int \int f_{\theta_A}(Y_A^{mis}|Y_C^{obs}) f_{\theta_B}(Y_B^{mis}|Y_C^{obs}) f(\theta_A, \theta_B|Y_A^{obs}, Y_B^{obs}, Y_C^{obs}) d\theta_A d\theta_B. \tag{8}$$

To draw imputations of (Y_A^{mis}, Y_B^{mis}), first draw (θ_A, θ_B) from their posterior distributions using the MCMC algorithm in Online Appendix A.1. After burn-in period, parameter draws at iterations with sufficiently long intervals between them in the Markov chains are retained to make these draws essentially independent. Second, with each retained parameter draw

(θ_A, θ_B), we draw Y_A^{mis} and Y_B^{mis} from $f_{\theta_A}(Y_A^{mis}|Y_C^{obs})$ and $f_{\theta_B}(Y_B^{mis}|Y_C^{obs})$. Under our framework, drawing from these distributions is straightforward. For instance, if the ith observation in Y has Y_A unobserved, the imputation is drawn from $f_{\theta_A}(Y_{iA}^{mis}|Y_{iC}^{obs})$ which is a discrete distribution on the unique observed values of Y_A with the probability mass (weight) function given in Eqn (6). Draws from this distribution can be generated using existing multinomial random number generators. Similar to the traditionally popular hot-deck procedure, the imputed value in our approach is one of the values observed in the datasets, has face validity, and avoids the out-of-plausible-range problem. However, unlike hot-deck, the weights in our imputation rule are derived from probabilistic models and are thus supported by statistical theory. Repeated draws from the posterior predictive distribution can also be used to compute the posterior means of unknown quantities for prediction purpose. Imputations for Y_B^{mis} are obtained similarly.

The above two LI methods require the common variables in Y_C be fully observed. In practice, unintentional missingness often occur in common variables, as shown in Figure 1.1 (c). We therefore develop two FI extensions (FI-DE and FI-MI) that can incorporate the sampling units with missing values in Y_C into data fusion. These FI methods have the ability to control for potential selection bias due to missingness in common variables and improve estimation efficiency. These methods are described in Online Appendix A.3. In Section 5, we have also conducted extensive simulations studies to evaluate the performance of the proposed methods, which demonstrate their ability to automatically adapt to arbitrary data departures from regular distributional forms and to substantially improve fusion precision and individual consumer predictions.

4. Empirical Application

4.1 *Data Description*

This empirical application is concerned with identifying systematic patterns in consumers' counterfeit consumption behaviors. The datasets to be analyzed are from a famous branded footwear company in China, and include data from two surveys and the firm's internal consumer database. Because of a recent surge in counterfeits of popular product lines of this firm in the market, the firm conducted a survey on a sample of consumers of its stores in 2009, mainly to assess the situation and include questions about the extent (incidence and

monetary value) of counterfeit purchases. However, the survey did not contain all possible relevant questions. For example, it is of managerial interest to understand how consumers who purchase counterfeits differ from those who do not in their attitudes toward authentic product attributes, promotions, and advertisements. Such information can help managers design measures to counter counterfeits. It is also useful to know where consumers often shop for shoes, which can be useful for identifying the channels through which counterfeits reach consumers and allocating limited resources to the most affected channels.

To address this data limitation issue, data fusion is used to complement the survey data with data from another survey conducted by the same firm in the same time period on an independent sample of consumers of the firm's stores. The second survey was designed to understand the purchasing habits and attitudes of the firm's consumers and had multiple questions on them, although it did not contain questions about counterfeit purchases. Administering two surveys separately avoids the context effect of collecting sensitive questions about counterfeit purchases with other relevant data in the same survey. Such an effect can arise, for example, when a question about counterfeit purchases provides cues for respondents and thus negatively affects their responses to other questions (e.g, where they shop and motivations for purchasing authentic products such as product quality) or their response to the sensitive question is affected by other questions (e.g., questions about the authentic firm's promotion). In addition, both surveys contain multiple other questions in order to answer other marketing research questions. Administering a large number of items in one lengthy survey can also lead to survey fatigue, which degrades the quality of survey response. These considerations suggest that, to collect more faithful and cleaner data on the survey questions, it would be wise for the firm not to collect all the data in one survey.[7]

Table 1 lists the variables considered here where all the attitudinal and shopping behavior variables in Y_A are binary with an outcome of 1 (yes) or 0 (no). The data that we receive contain 1698 and 3205 observations for the first and second surveys, respectively. With the datasets currently available to the firm manager, it is impossible to relate counterfeit

[7] The biasing effects can also limit the use of the single long-survey data for other important purposes, such as the use of attitudinal questions on authentic products for product design. Of course, we are not suggesting not collecting such long-survey data. In the ideal situation where a firm could anticipate all kinds of analyses beforehand, a better design might be to obtain an additional joint sample where all relevant questions were asked for the same sample of consumers. Including this joint sample in data fusion provides opportunities to relax the CIA and at the same time control for bias in the joint sample due to context effects and survey fatigue. Extending the data fusion approaches proposed here to this situation would be very valuable.

purchases to consumers' purchasing habits and attitudes if these two datasets are analyzed separately, since neither survey contains both sets of variables. Data fusion thus provides a feasible way to overcome this challenge by examining two datasets simultaneously. Data from the two surveys are joined by the set of common variables in Table 1, including demographic, geographic, and purchase behavior variables. The three common variables for the store purchases, *Expend, PurchRate,* and *BaskSize*, are obtained from the firm's internal databases.

4.2 *Data Fusion Results*

We first conduct an LI analysis, that conditions on all the common variables and so excludes all consumers whose common variables are not fully observed. The resulting complete-case sample contains 3010 consumers. We then apply three data fusion methods: hot-deck, the parametric fusion, and our proposed FORM (Fusion via Odds Ratio Models) approach. Because the DE approach to data fusion is not available for hot-deck and difficult for the parametric fusion with continuous unique variables, we compare different fusion methods through the imputation approach. That is, FORM uses LI-MI as described in Section 3. Let Y_B denote the monetary value of counterfeits purchased over the past year. To model this unique variable that has a non-zero probability at the value of zero and a continuous distribution on the positive values, we employ the following two-tiered model with two variables: a binary variable B for consumer's decision on *whether* to purchase counterfeit and a continuous variable Y_B^* for consumer's decision on purchase amount, such that

$$B = \begin{cases} 1 & \text{if } Y_B > 0 \\ 0 & \text{if } Y_B = 0 \end{cases} \quad \text{and} \quad Y_B^* = \begin{cases} Y_B & \text{if } B = 1 \\ \text{undefined} & \text{if } B = 0. \end{cases}$$

The parametric fusion first uses logistic regression models for $f(B|Y_C)$ and $f(Y_A|Y_C)$, in which all the common variables (except *age* and three store purchasing variables) enter the model as dummy covariates, based on their unique levels in the datasets[8]. This modeling strategy helps guard against the possible nonlinear additive effects of these common variables and is used throughout for all models below. When the binary purchase decision variable B is imputed as zero, the purchase amount Y_B^* is undefined. When B is imputed as one, the parametric fusion specifies a parametric distribution of the purchase amount Y_B^*, conditioning on the set of common variables. We tested a linear regression model on the purchase amount variable and

[8] The quadratic and cubic terms of those four continuous variables are also included in all the models. The Hosmer-Lemeshow goodness of fit tests for these logistic regressions all have p-values larger than 0.2, indicating no lack of fit.

its log transformation. We found that the log transformation performs better for data fusion and thus it is used in the parametric fusion. In contrast, FORM uses odds ratio models that do not require making distributional assumptions and can automatically generate suitable distributions that match data with various nonstandard distributional shapes. Therefore, the original scale of counterfeit purchase amount is used in FORM. This circumvents the need to select a proper transformation and/or parametric distributional model. Odds ratio models are also used to model the binary purchase decision variable B and the attitudinal and shopping behavior variables in Y_A. These odds ratio models condition on Y_C in the same way as the above logistic and log-normal regression models.

Eqn (8) was used to make imputations from the posterior predictive distributions of unobserved unique variables and the LI-MI algorithm described in Section 3 was used for FORM. Standard posterior predictive sampling algorithms for logistic and log-normal regression models (Gelman et al. 2004) were used in the parametric fusion. In FORM and the parametric fusion, twenty imputations were generated[9]. We then conduct analyses on each imputed dataset and use Rubin's combination rule to pool results from multiple imputed datasets. In contrast, hot-deck does not use the posterior predictive distribution to make imputations. Instead, it uses some sort of matching algorithm to select a single "best" imputation. We use the improved hot-deck method as implemented in Gilula et al. (2006)[10].

The data fusion results are summarized in Tables 2 and 3, which for space reasons present the results only for those variables with statistically significant results for at least one method. The complete results for all variables are in Online Appendix Tables 10 and 11. Table 2 presents the log-odds-ratio estimates (standard errors) of the relationship between each of the attitudinal and shopping behavior variables in Y_A with the binary counterfeit purchase decision variable B. Note that the results from the parametric logistic regression fusion and FORM are almost identical[11]. The comparison between hot-deck and FORM shows that hot-deck tends to claim more statistical significance. This is true for two reasons. First, because

[9] Using a larger number of imputations (e.g., 50) has resulted in no change in the analysis conclusions.

[10] Because perfect matching is impossible for *age* and three store purchasing variables which are continuous matching variables, the improved hot-deck uses a metric-matching for them that defines a distance measure to find a nearest neighbor donor. Because these variables are not elliptically distributed, the distance metric is defined as the sum of absolute differences for rescaled continuous variables, scaled to be in the range of zero to one (see Gilula et al. 2006).

[11] FORM can handle various departures from binomial distributions. However, these departures do not exist for simple binary variables, which explains the similarity between the fusion results of FORM and the binary logistic approach.

of its ad hoc nature, hot-deck results in more varied estimates. Second, the standard errors are much smaller than they should be because hot-deck ignores the uncertainty due to fusion.

Table 3 reports the analysis of the relationship between each of the attitudinal and shopping behavior variables in Y_A with the counterfeit purchase amount Y_B^* for those who purchased counterfeit products. Because of its nonnormal distributional feature, we compare median differences instead of mean differences. Table 3 reports the median regression coefficient estimates and the associated standard errors. Again, we find that hot-deck results in more false positives. On the other hand, the parametric log-normal purchase amount fusion model results in more false negatives. FORM identifies variables 4, 10, 14, 16, and 19 as statistically significantly related to counterfeit purchase amount. Of these five variables, the parametric fusion can identify only three (variables 4, 14, and 16). This corresponds to an approximately 40% false-negative rate. Furthermore, the parametric fusion appears to provide much smaller parameter estimates than estimates from both hot-deck and FORM. In some cases, the underestimation is more than 50% (e.g., variable 10).

Tables 2 and 3 also report a full-information data fusion under the column "FORM-FI" that incorporates all consumers, including those with missingness in common variables.[12] The results show that "FORM-FI" substantially increases fusion efficiency, and considerably reduces standard errors compared with the complete-case analysis reported under "FORM". Notably, we find strong statistical significance for variables 15, and 19 in Table 2 and for variable 19 in Table 3. These results suggest that the consumers who shop often in open market or on street are more likely to purchase counterfeits; consumers who purchased more counterfeits are more interested in receiving promotions through catalogs. These findings are unidentified or considerably less significant in both parametric and FORM fusions.

4.3 *Managerial Implications*

Given the increasing prevalence of product counterfeiting and its significant impact on authentic prices, sales, and brand image (e.g., Qian 2008 and 2011), it is critical to examine the demand side of the counterfeit market. Designing an optimal strategy often requires knowledge about how consumers most affected by counterfeits differ from others in their

[12] The common variables subject to missingness are "Age", "Family Income" and "Number of Children at Home" which collectively cause about 40% incomplete cases that are dropped from LI fusion but incorported in the FI analysis. Our FORM-FI analysis uses the FI-MI procedure as described in Online Appendix A.3.

purchase motivations, behaviors, and attitudes toward branded products, promotions, and channels. Understanding these marketing reactions can generate insights for protecting brand and recruiting and retaining loyal customers.

The analyses based on FORM indeed reveal systematic differences in the characteristics of consumers who had different counterfeit purchase outcomes. In particular, consumers who had not purchased counterfeits in the past year had more positive attitudes toward the prices of the authentic products and tended to use the products for work and social interactions. Furthermore, consumers purchasing less counterfeits put more emphasis on the health, safety, and social interaction benefits of branded apparels. Therefore, one potentially useful strategy would be for firms to use advertisements to stress these benefits of authentic products as compared with counterfeits.

The analyses also help identify the channels through which counterfeits are brought to market. The fact that consumers who purchase more counterfeits tend to shop more often on the street or on the Internet is both intuitive and important. In particular, Internet provides firms with a convenient marketing channel. Unfortunately, it also serves counterfeiters. In order to reduce the use of the Internet by counterfeiters, an authentic company might educate consumers about the harmful effects of counterfeits and post tips on its websites for identifying fake merchandise. The firm could also work with Internet malls to reduce the amount of counterfeits through raids and/or legal actions, as Louis Vuitton has successfully sued eBay in Europe. Tightened controls over distributions, such as establishing authentic licensed stores (Qian 2008), can help separate legitimate firms from counterfeiters as well.

Promotion and advertising can also help authentic firms combat counterfeiters. As compared to legal actions, these marketing measures may more effectively entice consumers back to legitimate businesses and help reduce the damage done by counterfeits. Evidence from the data fusion results of FORM and FORM-FI suggests that incentive measures such as promotions through catalogs appeal to consumers who are attracted to counterfeits.

Our analyses also demonstrate that the fusion conclusions depend in an important manner on the fusion methods used. Below we discuss some improvements in managerial decisions from using new data fusion methods.

Improved identification of patterns in counterfeit consumption for anti-counterfeit planning As shown above, an important step in anti-counterfeit planning is understanding systematic patterns in consumer counterfeit consumption behavior. Because of its ad hoc nature, hot-deck leads to too many false-positive findings, as shown in our empirical analysis. Although FORM and the parametric fusion agree most on the analysis of counterfeit purchase incidence, which is a simple binary variable, they can have important differences for more complex variables. Our analysis shows that, in this empirical application, the parametric fusion using the standard log-normal distribution for the more complex counterfeit purchase amount variable has a low power to detect associations and a high false-negative rate (40%) because of the bias introduced into data fusion by the misspecification of its distributional form. Table 3 shows that both nonparametric procedures (FORM and hot-deck) identify significant results for variables 10 "SocialInteraction" and 19 "PromCatalog", while the parametric fusion is not able to identify them. It is important to note that these two factors are either borderline significant (SocialInteraction) or not significant at all (PromCatalog) in the parametric fusion analysis of counterfeit purchase incidence (Table 2). In addition, the variable 15 "ShopOnStreet" is not identified in the parametric fusion but is identified in FORM-FI in counterfeit purchase incidence analysis in Table 2. Therefore, the results from the parametric fusion can lead to significantly underestimating or even ignoring the importance of these three factors in anti-counterfeiting planning.[13] For example, the usefulness of emphasizing the social interaction benefits of branded shoes in anti-counterfeit advertisements, promotions through catalogs to make branded shoes more appealing to affected consumers, and counterfeit control in open market may be underestimated or ignored.

Online Appendix B reports additional analyses of this dataset that examine performance gains of FORM at different sample sizes and for detecting finer distributional differences. We observe that the improvement in power to detect association for FORM can be significantly larger for smaller sample sizes. We also find a larger performance gain of FORM for comparing tail percentiles because of its ability to match entire distributions nonparametrically.

Improved individual prediction One important use of database marketing is consumer targeting. The fusion approaches proposed here can be used to predict unobserved

[13] In this sense, both the size and statistical significance of parameter estimates are of relevance.

counterfeit purchase expenditure. Such information can be useful for selecting consumers most affected by counterfeits and applying marketing measures, such as customized promotions, to entice them back to the authentic brand. We compare different fusion methods on the performance of individual predictions. Specifically, using the observed values of counterfeit purchase expenditure, we perform leave-one-out cross-validated individual prediction and compute the RMSE for each fusion method. The results in Table 4 show that the improvement in individual prediction for FORM is 22% relative to the parametric fusion and 47% relative to the hot-deck. The cross-validation demonstrates the superior performance of our approaches in predicting consumers' counterfeit consumption.

5. Further Evaluation using Simulation Studies and Resampling Experiments.

Data fusion often involve relating non-overlapping variables with unknown complex distributional forms across complementary datasets. One of the main advantages of the proposed method is that the proposed method is nonparametric, therefore requiring no prior knowledge about how the observed data behave differently from standard distributional forms. To illustrate this point along with other benefits of our methods, we perform the following simulation studies.

5.1 *Simulation Setting*

We first describe the procedure used to generate synthetic datasets consisting of seven variables.

Simulation of Common Variables We first simulate three common demographic variables, Y_1, Y_2, Y_3, from the following distributions. Y_1 is simulated from the standard normal distribution. Given Y_1, Y_2 is a Bernoulli variable with

$$\text{logit}(P(Y_2 = 1|Y_1)) \quad = \quad \beta_{20} + \beta_{21} Y_1.$$

Given Y_1, Y_2, Y_3 is a Poisson variable

$$Y_3|Y_1, Y_2 \quad \sim \quad Poisson(\exp(\beta_{30} + \beta_{31} Y_1 + \beta_{32} Y_2)).$$

This simulates a scenario in which the common demographic variables include both continuous and discrete variables. In the simulation, we set $(\beta_{20}, \beta_{21}) = (0, 2)$, $(\beta_{30}, \beta_{31}, \beta_{32}) =$

$(0, 0, 1)$.

Simulation of Unique Variables We then simulate a unique variable Y_4 as a binary variable. One may consider this as a binary indicator for a marketing stimulus. Y_4 is simulated from a Bernoulli distribution with

$$\text{logit}(P(Y_4 = 1|Y_1, Y_2, Y3) \quad = \quad \beta_{40} + \beta_{41}Y_1 + \beta_{42}Y_2 + \beta_{43}Y_3.$$

where $(\beta_{40}, \beta_{41}, \beta_{42}, \beta_{43}) = (0, 1, 1, -1)$. We then further simulate three other unique variables with various departures from standard distributional forms. In practical applications, many forms of departures from standard distributions can arise. These include outliers, deviance from nominal dispersion, heavy-tails, skewness, multi-modality, boundedness, etc. The MLEs based on the parametric models are known to be sensitive to these departures. In practice it would be tedious, if not impossible, to check for all these different forms of departures and to account for them properly. Therefore it is important to have a data-driven nonparametric procedure that can automatically account for all these important features.

To demonstrate the benefits of our nonparametric FORM procedure in automatically accounting for departures from standard distributional forms, we generate another three unique variables, Y_5, Y_6 and Y_7. Fig 2 shows the form of distributions for these three variables. Y_5 is simulated as a count variable (e.g. number of purchases) with its error variance different from the nominal variance dictated by the Poisson model in which the variance equals the mean. Y_6 is simulated as a continuous variable (e.g., logarithm of expenditure or basket size) with its major part of error distribution being a standard normal distribution but with a heavier tail on the right side than the normal distribution. The commonly-used heavy tail distribution, t-distribution, cannot accommodate such asymmetric heavy tail properly. The variable Y_7 is a semicontinuous variable. The form of distribution for Y_7 may arise, e.g. for the price offered to a consumer. In one-to-one targeted marketing, the offered price can differ from consumer to consumer. The distribution form emulates a situation in which the majority of the offered price is the full price with gradually decreasing probability for reduced price up to around 30% reduced price, and then there are a slightly increased probability for around 50% price reduction which appear like outliers. Overall Y_7 includes important departure features such as outliers and boundedness.

These nonstandard distributions can all be accommodated by the FORM approach automatically. In fact, one can actually generate random draws from these distributions using the odds ratio models in a unified manner. In our simulation, we generate these variables in the following steps:

Step 1. Specify the density function for $f(y_j|y_{30}, y_{20}, y_{10})$, the baseline density function of $Y_j, j = 5, 6, 7$ at the reference points[14] of those common variables, as in Equation (4) where we replace y_A by y_j. Our goal is to generate data from distributional forms shown in Fig 2. For the count variable Y_5, we specify a weighted Poisson density form which is $P(y_5 = k) \propto \frac{e^{\lambda}\lambda^k}{k!}w(k)$ where k takes non-negative integer values. When $w(k) = 1 \forall\, k$, this is a standard Poisson density function. When $w(k)$ differs from the unity function, the density form leads to a count model that departs from the parametric Poisson distribution in arbitrary manner, and therefore the odds ratio model can model any departure from the parametric Poisson model[15]. In our simulation, we set $\lambda = \exp(1)$ and $w(k)$ to be Poisson density function. Fig 2 compares the density function for this weighted Poisson distribution with that of Poisson distribution with the same mean (the fitted Poisson distribution under MLE). The set of uniquely observed values as specified in Equation (4), $(y_{51}, \cdots, y_{5N_5})$, are set to be integers from 0 to 10, beyond of which the probability density function is essentially zero, as shown in Fig 2. For the variable Y_6, we specify the distribution as follows. The set of uniquely observed values are from -5 to 5 with a grid of 0.1.[16] A value x between (-5, 1.7) has a density function $dnorm(x)$, the standard normal density function; a value x between (1.7,5) has a density $dnorm(1.7)$, the standard normal density at 1.7 to simulate a heavier tail than the standard normal distribution. The values are then rescaled to have probabilities sum to one. For variable Y_7, the set of uniquely observed values are from from 6 to 12 with a grid of 0.1. As explained above, one may consider Y_7 as a price variable, where the right endpoint, 12, corresponds to full price and left-end point, 6, corresponds to half price. The density form is

[14] In our simulation these reference points are set to be at zeros.

[15] In particular, when $w(k)$ is concave (convex), it allows underdispersion (overdispersion). The parametric negative binomial model can allow for overdispersion, when the Gamma distributional assumption in the mixing distribution is satisfied. In contrast, the odds ratio model as used in FORM is nonparametric, requiring no distributional assumptions and can model arbitrary form of departures from Poisson Model. For example, the negative binomial model does not allow for underdispersion while nonparametric odds ratio models do.

[16] This corresponds to that the values of Y_5 are observed to the first decimal place. In practical situations values are almost always observed to a limited number of decimal places (such as cents for dollar spent). Our simulation results remain essentially unchanged when finer grid of values (e.g., a grid of 0.01) are used.

generated using a polynomial density form, $(x-8)^4$ for a value x taken by Y_7.[17] For variable Y_7, Y_4 can be considered as a binary consumer purchase variable (made a purchase or not). After specifying the probability mass function $p_j = (p_{j1}, \cdots, p_{jN_j}), j = 5, 6, 7$ as above, we then obtain the parameter values $\lambda_j = (\lambda_{j1}, \cdots, \lambda_{jN_j})$, where $\lambda_{jl} = \ln(p_{jl}/p_{jN_j}), l = 1, \cdots, N_j$.

Step 2. Specify the odds ratio models. The simulation uses the bilinear form as specified in Equation (5). The odds ratio parameters $(\gamma_{j1}, \gamma_{j2}, \gamma_{j3})$ that captures the relationship between $Y_j, j = 5, 6, 7$ and Y_1, Y_2, Y_3 are specified to be $(1, 0.1, -0.1)$ for Y_5 and Y_6, and $(-1, -0.1, 0.1)$ for Y_7. This simulates situations where marketing stimuli increase the number of purchases (Y_5) and the logarithm of expenditure (Y_6), and where a lower price (Y_7) increases the purchase incidence.

Step 3. For a give value of (y_1, y_2, y_3), compute the density function $f(y_j|y_1, y_2, y_3)$ as given in Equation (6) where y_A is replaced by y_j. This is a multinomial probability distribution function on the set of uniquely observed values specified above. Random draws from this multinomial probability distribution function can be straightforwardly obtained using existing random number generation routines.

We simulate data for 1000 consumers. To mimic the scenario of data fusion, we assign the first 500 consumers to Dataset A and the other 500 to Dataset B. The three common variables (Y_1, Y_2, Y_3) are set to be observed for all consumers in Datasets A and B. The variable Y_4 (Y_5, Y_6, Y_7) is (are) observed for consumers in Dataset A(B) but is (Are) set to be missing in Dataset B(A). The purpose of data fusion in the simulation study is then to measure the marginal relationship between Y_4 and each of the variables in (Y_5, Y_6, Y_7) even if there is no direct observation of the marginal relationship.

For each simulated Dataset A and B, we apply three methods for data fusion. The first one applies the proposed approach, named FORM (Fusion via Odds Ration Models), to combine the datasets; an odds ratio model is used to model the conditional distribution for each of the unique variables (Y_4, Y_5, Y_6, Y_7), conditional on the common variables (Y_1, Y_2, Y_3). An important feature of this method is that no parametric distributional assumption is imposed for any conditional distribution.

[17] A polynomial form for price variable has been observed in Qian and Xie (2011).

The second method is the hot-deck procedure. As reviewed in Section 2 of the main text, the hot-deck procedure can be considered as a matching procedure in which unobserved values for unique variables of one recipient consumer are replaced with observed values from a matching donor consumer with identical or similar values in the matching variables. Common variables often contain a mixture of discrete and continuous variables. In our analysis, we use an improved hot-deck procedure as described in Gilula et al. (2006). In this procedure, one first searches for a perfect match for a recipient. If all the common variables are categorical, one then attempts to find an exact match. If more than one respondents is identified as the perfect match, a decision must be made as to which one is to be used as the donor. For example, a randomly selected one can be used as the donor. If there is no perfect match, some ad hoc rules can be used to find an exact match. For example, one can collapse one or more categorical variables into a smaller number of categories, or even drop the categorical variables in matching. Therefore, the procedure can involve very delicate but ad hoc rules for selecting donors, and different selection rules can be applied to different recipients so every recipient can find a donor. If the common variables also contain continuous variables, perfect matching is impossible. In this case, the improved hot-deck procedure uses metric matching, which defines a measure of distance to find a nearest neighbor donor for a recipient. There can be different ways to define a distance matrix. For example, the Mahalanobis distance metric is often used if the variables are believed to be elliptically distributed. If not, one might use the sum of absolute differences for rescaled variables, scaled to be in the range of zero to one. In the data fusion of simulated datasets, one matching variable, Y_2, is taken as a truly categorical variable and the other two matching variables, Y_1 and Y_3, are taken as continuous variables in the hot-deck procedure.

For comparison purposes, we also consider a parametric approach to data fusion. This involves specifying parametric probabilistic distributions for the unique variables; we use a logistic and Poisson regression model for the binary variable Y_4 and the count variable Y_5, respectively, and linear regression models for Y_6 and Y_7, conditional on (Y_1, Y_2, Y_3). For variable Y_7, because the linear regression model can predict the outcome out of range, we have re-coded any imputed values using the linear regression model that are larger than 12 to 12 and that are smaller than 6 to 6. This represents the effort to account for boundedness by

manually setting values outside the plausible range to be at the boundary values, and we are interested in evaluating whether such manual adjustment is sufficient. These comparisons allow us to investigate the potential effects of various misspecifications of the distributions of unique variables on the results of data fusion.

Because the direct approach to data fusion under the hot-deck procedure and the parametric fusion is either unavailable or difficult, we compare different data fusion methods through imputation methods. For each simulated dataset, our analysis consists of two steps: imputation and analysis. In the imputation step, we apply different imputation methods to create imputed datasets. For both FORM and the parametric fusion, we first run 500 iterations as the burnin period. After the burnin period, twenty imputations are created with 50 iterations between two consecutive imputations. These decisions were made by visually inspecting the traceplot of the Markov chains for several test runs and using the Geweke's test statistics to test the convergence of the Markov chains. In the analysis step, we conduct various analyses using the imputed datasets and then use Rubin's combination rule (Rubin 1987) to pool results from multiple imputed datasets. We also conduct these same analyses using the complete datasets, which provide gold standard estimates. We repeat the above analysis for 100 replicated datasets and summarize the findings below.

5.2 *Assessing Distributional Effects*

We first compare the ability of different data fusion methods to identify the entire distributions of the unique variables. A better fusion method should more faithfully reproduce the distributional characteristics of the true values. Table 5 reports the Kolmogorov-Smirnov (KS) test values averaged over all the simulations for each fusion method. The two-sample KS test measures the distance between two empirical distributions, and it is able to detect differences in both the location and the shape of the empirical distributions. It is used here to compare the empirical distributions of imputed values with those of original simulated values. The larger the value of the test statistics, the more the discrepancy between the two distributions. Besides this overall test, we also examine the performance of different fusion methods to identify some important summary statistics of distributions, including mean, variance and tail percentiles (e.g., Q95). Because in simulation we have the full data,

we have run these analyses also on the full data and taken the resulting estimates as gold standards. For a particular data fusion method, we calculate the mean absolute difference (MAD) between the estimates from this data fusion method and the estimates using the full data over all the simulations as a measure of the performance of the data fusion method. The smaller the MAD, the better performance the data fusion method.

The results in Table 5 show that when the parametric model is correctly specified (i.e., for Y_4), the parametric fusion and FORM perform equally well, and both perform better than the hot-deck procedure, as evidenced by their substantially smaller KS statistics values. The summary statistics such as mean, variance and percentiles are considerably closer to the true values, as evidenced by the substantially smaller MAD values. On the other hand, when the data at hand depart from the standard distribution models (i.e., for Y_5, Y_6 and Y_7), the performance of the parametric fusion worsens, as shown by its substantially inflated KS values. It is clear that FORM performs best among all three methods because of its capability to automatically adapt to the distributional shapes of these unique variables. In particular, FORM reduces the MAD of summary statistics substantially as compared with the parametric fusion approach, with the improvement ranging from 30% to 80%.

We further investigate how the improvement in the quality of imputed values leads to an improvement in data fusion results. Table 6 reports the differences of mean and tail percentiles of Y_5, Y_6, and Y_7 between the group with $Y_4 = 1$ and the group with $Y_4 = 0$. These group differences are used to capture the marginal relationship between two sets of unique variables. Table 6 reports the MAD values of these estimates, again using the estimates from the full data as the gold standard. The results show that FORM performs best in all situations, as seen by its small MAD values, indicating that FORM generates estimates that are closest to the full data analysis.[18] The average improvement of FORM in this case as compared with parametric data fusion is about 30% for mean difference estimates and as high as 60% in the tail percentile estimates.

Overall, the simulation study demonstrates the superior performance of FORM relative

[18] This is so because the regression parameters are high-level functionals of density functions that undergo smoothing operations of integration with respect to the density function. For such parameters, a nonparametric or semiparametric procedure tends to have little estimation efficiency loss as compared to a parametric approach (Meier et al. 2004, Qian and Xie 2011). Therefore, with the sample size typically seen in marketing applications, FORM performs at least as well as the parametric data fusion approach, and can outperform it when the bias is introduced into the parametric fusion due to model misspecification.

to existing methods. Unlike the parametric approach to data fusion, FORM is more robust and minimizes the impact of distributional assumptions in data fusion. As a nonparametric procedure, FORM automatically generates suitable distributions used for data fusion, and guards against misspecifications of distributional assumptions that can occur in the parametric fusion. As a result, the quality of imputed values in FORM is at least as good as, and can be substantially better than, that of the parametric fusion. Because of this quality improvement, FORM can provide a large improvement in data fusion results relative to the parametric fusion approach. It is also interesting to note that the improvement can be much larger for the other distributional comparisons (e.g., tail percentiles) than for the mean. This is hardly surprising because the tail percentile is typically a distributional characteristic that is more sensitive to distributional misspecifications. The advantage of FORM in assessing distributional comparisons is particularly valuable because there is a growing body of literature demonstrating the importance of examining distributional effects, rather than merely average effects (e.g., Bitler et al. 2006, Firpo 2007). We believe this is important for marketing applications for at least two reasons. Marketing has long been interested in heterogeneous effects; distributional effects (e.g., quantile differences) are one instance of treatment-effect heterogeneity. Furthermore, assessing the distributional effect may provide managerially more relevant analysis. For example, the tail percentiles correspond to a segment of consumers with heavy expenditures and thus can be of more interest to managers. In such studies it becomes more critical to maintain the entire distributional shape, and nonparametric procedures are particularly important. This need for nonparametric procedures is more salient in data fusion because an important difference here, as compared with the case in which full data are available (Bitler et al. 2006, Firpo 2007), is the large proportion of unobserved information. Thus, distributional misspecifications can propagate to have a large adverse effect.

As compared with the hot-deck procedure, FORM uses the rules based on statistical models for imputation of unobserved values. Because it avoids the ad hoc rules involved in the hot-deck procedure, our procedure outperforms the hot-deck procedure.

5.3 *Individual Prediction*

Another common use of data fusion is for prediction of individual outcomes (Kamakura and Wedel 2003), instead of population distribution. Our method can also be used to guard against the misspecification of distributional assumptions in individual predictions. To compare the performance of different fusion methods for individual prediction, we compute the root mean squared errors (RMSE) of individual prediction as follows:

$$RMSE_j = \sqrt{\frac{\sum_{i=1}^{n_j}(y_{ij} - \hat{y}_{ij})^2}{n_j}},$$

where $j = 4, 5, 6, 7$ indexes the unique variable, $i = 1, \cdots, n_j$ indexes the consumers whose unique variables are set to be unobserved in synthetic datasets, y_{ij} is the original simulated value, and $\hat{y}_{ij}$ denotes the mean value computed from different fusion methods used as the individual prediction.[19] Table 7 reports the average of RMSE over all the synthetic datasets. The result shows that the hot-deck procedure performs poorly for individual predictions, which has the largest RMSE among all three methods. When a correct parametric model is posited (i.e., Y_4), the parametric and FORM perform equally well and both perform considerably better than the hot-deck procedure. However, when data at hand behave differently from standard distributional models (i.e., Y_5, Y_6 and Y_7), an appreciable improvement in individual prediction (a reduction of RMSE by 20% to 30%) is present in the FORM as compared with the parametric method. This demonstrates the value of FORM as a nonparametric model-based fusion method to guard against model misspecifications in individual predictions.

5.4 *Efficient Direct Approach Beyond Binary Unique Variables*

The proposed FORM approach provides efficient direct data fusion beyond binary unique variables. Gilula et al. (2006) develop a direct approach to data fusion that estimates the joint predictive distribution of unique variables (Y_A, Y_B) given the data D. This method is more efficient than the conventional sampling-based multiple imputation approach to data fusion. The reason is that the missing-data imputation method introduces sampling errors to the fusion result, and an unusually large number of imputations is required to control the

[19] The hot-deck procedure is not model-based. Nonetheless, we can use the imputed value from the hot-deck as the prediction.

sampling error in order to approach the result of direct data fusion.[20] As pointed out by Gilula et al. (2006), this occurs at the expense of "unnecessary computations and the creation of larger data sets."

It is important to note that the application in Gilula et al. (2006) deals with the case in which both Y_A and Y_B are binary variables. Hence the joint predictive distribution has probability mass concentrating on a total of four possible combination values of Y_A and Y_B, and can thus be conveniently evaluated. Extending the direct data fusion approach from binary variables to general cases is nontrivial, and several issues need to be addressed. **First**, as demonstrated above and in our application, more careful modeling is required for general types of data. Important features in continuous or semicontinuous variables, such as outliers, heavy tails, departure from nominal variance, skewness, multimodality, boundedness, and zero-inflation, must be carefully modeled in order not to bias data fusion results. **Second**, with conventional modeling choices, the joint predictive distribution for variables (Y_A, Y_B) that are not discrete is generally of an unknown form on a continuous (or semicontinuous) sample space. This substantially complicates the use of the predictive distribution for direct data fusion. Our proposed method simultaneously overcomes the above important roadblocks because (1) FORM imposes no distributional assumptions for any variable in data fusion and (2) the predictive distribution in FORM concentrates its probability mass on a finite number of unique observed values of (Y_A, Y_B) even when they are continuous variables, and this probability mass function can be evaluated straightforwardly. As a result, FORM represents a significant extension of efficient direct fusion to a much broader range of marketing applications.

Simulation studies as described above show that FORM outperforms the hot-deck and parametric data fusion approaches. The analysis of counterfeit purchase data in the main text also reveals several important differences in results obtained from different data fusion approaches. In this appendix, we further compare the performance of different data fusion

[20] This is not surprising in light of the typically large proportion of unobserved values in data fusion problems, because of which the commonly used 5 imputations introduce unacceptable amount of sampling errors. In order to control the sampling errors introduced by imputation, the number of imputation needs to be much larger than 5. In our case, we find that at least 20 imputations are required to obtain reasonably stabilized estimates. Other researchers have found even larger number of imputations are needed. For example, Kamakura and Wedel (1997) found the need to use 100 imputations in their applications. In general, the sufficient number of imputation required for stable data fusion is an important research question to address, and likely depends on various factors, including data configurations and models used for data fusion.

methods using datasets resampled from the counterfeit purchase data.

An ideal comparison has fully observed observations on two sets of unique variables that can be used as a gold standard to compare the performance of different data fusion methods. To achieve such a comparison, we conduct the following resampling experiments. The first set of unique variables are three purchasing variables in the stores: monthly expenditure, monthly purchase rate, and basket size. These variables are fully observed. The other set is the binary variable B for the counterfeit product purchase, which is partially observed in the dataset. We use the bootstrap method to generate new datasets that contain fully observed values for B. Specifically, we first use the common variables (excluding those three store purchasing variables) to estimate a logistic model for B. In a newly generated dataset, the values for B are replaced with the simulated values of B from the above logistic regression model. For each simulated complete data, we calculate various summary statistics as described in the paragraph below. These can be considered the gold-standard estimates since they reflect the ideal situation in which all data are available and no data fusion is required. We then randomly partition the whole dataset into two parts: Datasets A and B. In Dataset A, we set the variable B to be missing and in Dataset B we set the three purchasing variables to be missing. We then apply different data fusion approaches to combine the two datasets: the parametric fusion, hot-deck, and FORM. The two-tier model as specified in the main text is used to model the three purchasing variables. Using the fused datasets obtained from these different methods, we obtain the corresponding summary statistics under each method. These summary statistics are then compared with the corresponding gold standard estimates obtained using the complete data. We then repeat the above process 100 times.

Table 8 summarizes the results from the resampling experiments. We first compare the probability distributions of the purchasing variables between the complete dataset and the fused dataset. Specifically, for each data fusion method, we use the KS test to compare the distributions of imputed values with the distributions from the true values. The results in Table 8 report the average KS statistics values over all the resampled datasets. The results show that FORM clearly outperforms both the parametric and the hot-deck procedures in that its average KS values are smallest, indicating that it creates fused datasets that more faithfully represent the distributions of these variables. Table 8 also reports the mean ab-

solute differences (MAD) for mean and Q95 for each method, using the values from the resampled complete data as the gold standard. The results show that for "Mean" the parametric fusion and FORM perform similarly well and better than the hot-deck procedure. However, a large performance difference is observed for other distributional characteristics. In particular, as highlighted in Table 8, for the estimation of 95 percentiles, the MADs of the parametric fusion (and the hot-deck) can more than double or triple the size of the corresponding MADs of FORM. This substantial improvement by FORM is important because this segment of consumers has the largest expenditure in the store and is of particular importance to the firm.

6. Discussion

Due to the illicit and sensitive nature of counterfeits and other underground economics, relevant and detailed data can be scattered among different sources. In this study, we applied our data fusion framework to combine data from an authentic firm's internal records and surveys. The proposed approach provides a feasible way to relate sensitive data to other relevant data not collected together, and represents the first step to overcoming the important data limitation issue in the study of underground economics and counterfeit purchase behaviors. It opens the door to conducting more detailed investigations into counterfeiting phenomena by combining complementary consumer-level datasets from multiple sources.

Because the situations where relevant data can only be found in separate datasets collected from independent samples abound in marketing research, the proposed methodology has broader implications in business management and policy applications. For example, because direct fusion performs model estimation in each dataset separately without sensitive data actually having to be shared or released, our DE methods can be especially useful for addressing data privacy concerns when combining datasets from different sources. Our fusion methods can also be used for combining data for media planning, and integrating data from split-questionnaire design in lifestyle studies. In other settings, one could explore using it to combine data from government statistics, business census, industry reports, and other organizations. Importantly, with the rapid growth of information technology, databases useful for marketing researchers and managers have become increasingly available. Data fusion can make more effective use of these available databases, such as surveys, experiments, consumer

databases, and market field data, in order to inform timely managerial and policy decisions.

In many of these marketing tasks, the databases can include a mixture of highly disparate non-overlapping variables with unknown complex distributional forms. Although more empirical applications are needed, evidence so far suggests that the methods developed here are a promising set of tools for robust and efficient data fusion in these challenging situations. An important benefit of these methods is that they automate the process of generating suitable distributions, and thus their high objectivity that ensures findings from data fusion are not artifacts caused by imposed distributional assumptions. Consequently, they provide opportunities to improve the ability to properly identify consumer behavior patterns, and to make more accurate individual consumer predictions.

Although our methods do not require distributional assumptions, like others they require CIA when lacking alternative identification information. Further research is needed for new methods that relax CIA. We discuss two research directions to this end. First, our fusion approach can be extended to relax CIA and improve fusion precision by more effectively using an auxiliary sample, when available, in which all variables are jointly observed (Figures 1.1.c and d). We plan to present this extension in future work. Second, developing fusion methods that require neither CIA nor an auxiliary sample is a promising avenue for further research.

References

Blattberg, R.C., Kim, B. and Neslin, S.A. (2008) *Database Marketing.* Springer, New York.

Chen, H. Y. (2004) "Nonparametric and Semiparametric Models for Missing Covariates in Parametric Regression," *Journal of the American Statistical Association,* **99**, 1176–89.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004), "Bayesian Data Analysis", 2nd Ed. Chapman & Hall, London.

Feit, E.M., Beltramo, M.A. and Feinberg, F.M. (2010) "Reality Check: Combining Choice Experiments with Market Data to Estimate the Importance of Product Attributes" *Management Science,* 56:785-800.

Gilula, Z., McCulloch, R.E. and Rossi, P.E. (2006), "A Direct Approach to Data Fusion," *Journal of Marketing Research,* **43**: 73–83.

OECD (1998), "The Economic Impact of Counterfeiting," Paris, France.

Kamakura, W.A., Mela, C.F., Ansari, A., Fader, P., Iyengar, R., Naik, P., Neslin, S., Sun, B., Verhoef, P.C., Wedel, M., and Wilcox, R.(2005), "Choice Models and Customer Relationship Management," *Marketing Letters,* **16**: 279–291.

Kamakura, W.A. and Wedel, M (1997), "Statistical Data Fusion for Cross-Tabulation," *Journal of Marketing Research,* **34**: 485–98.

Kamakura, W.A. and Wedel, M (2000), "Factor Analysis and Missing Data," *Journal of Marketing Research,* **37**: 490–98.

Pouta, E. (2004), "Attitude and Belief Questions as a Source of Context Effect in a Contingent Valuation Survey, ", *Journal of Economic Psychology*, 25: 229–242.

Qian, Y. (2008) "Impacts of Entry by Counterfeiters", *Q.J.E.*, 123(4): 1577–1609.

Qian, Y. (2013) "Counterfeiters: Foes or Friends", *Management Sciences*, Forthcoming.

Qian, Y. and Xie, H. (2011) "No Customer Left Behind: A Distribution-free Bayesian Approach to Accounting for Missing Xs in Marketing Models", *Marketing Science*, 30: 717-736.

Rassler, S. (2002) "Statistical Matching", New York: Springer.

Rubin, D. (1974), "Characterizing the Estimation of Parameters in Incomplete-Data Problems", *Journal of the American Statistical Association*, 69: 467-474.

Winer, S Russell (2001), "A Framework for Customer Relationship Management", *California Management Review*, 43: 89-105.

Table 1

Common and Unique Variables in Data Fusion.

Common Variables (Y_C)	Summary
Gender (male/female)	35%/65%
Age (years)	38.7 (12.7)
Marital status (married/unmarried)	73%/27%
Education (elementary or less/junior/senior/college or more)[a]	12% (40%)
Family income (ten categories of income level)[a]	1.9% (16.6%)
Number of children at home (head counts)	0.66 (0.98)
Regions (eight geographic store locations)[a]	5.2% (32.5%)
Travel time to nearest store (min)	39.4 (22.7)
Time since first purchase in the stores (years)	3.3 (2.2)
Time since last purchase in the stores (years)	0.9 (1.3)
Expend: average total expenditure in store per month	7.5 (15.5)
PurchRate: average number of times visiting store per month	0.06 (0.09)
BaskSize: average expenditure in store per visit	105.9 (71.6)

[a]: Minimum (Maximum) of category percentages reported. Mean (SD) are reported for continuous variables.

Authentic Product Purchasing Motivations, Behaviors, and Attitudes Variables (Y_A).	Summary (% Yes)
Motivations for purchasing shoes with the brand	
1. The brand product is reliable.	38
2. It is comfortable.	82
3. The price is reasonable.	47
4. It is good for health.	11
5. It has a good design.	41
6. The materials are fine.	19
7. It uses high technology.	5
8. It is convenient to buy.	30
9. My friends use/recommend it.	14
10. I need for work and social interaction.	42
Places where shop often for shoes	
11. Shop often in mall.	13
12. Shop often in supermarket.	26
13. Shop often in discount store.	13
14. Shop often in licensed store.	44
15. Shop often in open market or on street.	20
16. Shop often on Internet.	24
17. Others shop for me.	2
Attitudes toward promotions	
18. Interested in promotion inside store.	41
19. Interested in receiving catalog.	23
20. Interested in getting a small gift.	28
21. Interested in getting store credits.	67
Attitudes toward advertisements	
22. Interested in advertisements in store.	17
23. Interested in advertisements on TV.	20
24. Interested in advertisements on Radio.	5
25. Interested in advertisements in magazines.	12
26. Interested in advertisements in firm-sponsored commercial activities.	7
27. Interested in advertisements in public space.	8

Counterfeit Product Purchase Variables (Y_B)	Summary
Purchased/did not purchase counterfeits of authentic brand in past year.	21.2 (% Yes)
If purchased, total monetary value paid for counterfeit products.	73.7 (29.7)

Table 2

Fusion Results with the binary variable for counterfeit purchase over the past year (B).

Y_A	Param	Hot-deck	FORM	FORM-FI
3. The price is reasonable.	-0.31 (.12)**	-0.19 (.08)**	-0.29 (.12)**	-0.21 (.09)**
4. It is good for health.	-0.85 (.23)***	-0.77 (.17)***	-0.80 (.23)***	-0.63 (.16)***
5. It has a good design.	-0.38 (.52)	-0.72 (.39)*	-0.50 (.46)	-0.24 (.37)
6. The materials are fine.	0.11 (.18)	0.22 (.13)*	0.14 (.18)	0.25 (.14)*
7. It uses high technology.	-0.31 (.48)	-0.75 (.43)*	-0.55 (.84)	-0.62 (.53)
8. It is convenient to buy.	0.04 (.13)	0.20 (.09)**	0.06 (.13)	0.05 (.10)
10. I need for work and social interaction.	-0.20 (.10)**	-0.08 (.08)	-0.20 (.12)*	-0.09 (.09)
14. Shop often in licensed store.	-0.53 (.12)***	-0.36 (.09)***	-0.57 (.16)***	-0.52 (.10)***
15. Shop often in open market or on street.	0.22 (.13)	0.18 (.10)*	0.21 (.14)	0.23 (.10)**
16. Shop often on Internet.	0.53 (.14)***	0.44 (.09)***	0.49 (.14)***	0.32 (.09)***
18. Interested in promotion inside store.	-0.14 (.15)	-0.28 (.11)**	-0.15 (.16)	-.05 (.11)
19. Interested in receiving catalog.	0.18 (.14)	0.17 (.10)*	0.16 (.17)	0.19 (.11)*
24. Interested in advertisements on radio.	-0.03 (.30)	-0.30 (.18)*	-0.16 (.28)	0.08 (.17)
26. Interested in advertisements in firm-sponsored commercial activities.	-0.21 (.23)	-0.31 (.18)*	-0.17 (.26)	-0.11 (.19)

Note: Throughout all tables, '*', '**', and '***' represent statistical significance at 0.1, 0.05, and 0.001 levels, respectively.

Table 3

Fusion Results with the monetary value of counterfeit purchases over the past year (Y_B^*).

Y_A	Param.	Hot-deck	FORM	FORM-FI
4. It is good for health.	-21.9 (10.3)**	-28.6 (11.3)**	-35.9 (11.6)**	-26.7 (9.8)**
8. It is convenient to buy.	2.1 (6.3)	14.3 (6.8)**	3.5 (9.3)	3.8 (7.3)
10. I need for work and social interaction.	-7.3 (5.2)	-14.3 (5.7)**	-16.9 (7.4)**	-17.1 (5.9)**
11. Shop often in mall.	-4.1 (8.2)	-22.9 (8.5)**	-2.1 (14.0)	-1.6 (10.0)
14. Shop often in licensed store.	-15.9 (5.5)**	-24.3 (3.8)***	-23.3 (6.8)***	-21.2 (5.5)***
16. Shop often on Internet.	14.9 (5.6)**	27.1 (5.3)***	22.9 (6.1)***	22.1 (5.7)***
19. Interested in receiving catalog.	6.8 (6.5)	15.7 (8.0)*	14.7 (8.3)*	15.1 (6.9)**

Table 4

Comparing Different Fusion Methods on Individual Prediction of $\ln(Y_B^*)$.

Method	RMSE	Improve
Parametric	0.38	0%
Hot-deck	0.57	-50%
FORM	0.30	22%

Table 5

Simulation Comparison of the Performance of Different Fusion Methods on Maintaining the Distribution of Fusion Variables.

Criteria	Method	Y_4		Y_5		Y_6		Y_7	
		MAD	Improve	MAD	Improve	MAD	Improve	MAD	Improve
KS	Parametric	0.02	0%	0.07	0%	0.16	0%	0.29	0%
	hot-deck	0.05	-154%	0.10	-43%	0.13	19%	0.10	63%
	FORM	0.02	0%	0.03	52%	0.04	75%	0.07	71%
Mean	Parametric	0.02	0%	0.06	0%	0.09	0%	0.28	0%
	hot-deck	0.05	-148%	0.35	-480%	0.55	-510%	0.40	-42%
	FORM	0.02	-1%	0.05	14%	0.06	34%	0.06	79%
Var	Parametric	0.002	0%	1.54	0%	0.31	0%	1.83	0%
	hot-deck	0.005	-263%	0.79	49%	0.54	-74%	1.04	43%
	FORM	0.002	-1%	0.24	84%	0.18	40%	0.17	89%
Q95	Parametric	-	-	0.71	0%	0.11	0%	0.10[a]	0%
	hot-deck	-	-	0.84	-18%	0.29	-163%	0.48[a]	-380%
	FORM	-	-	0.29	59%	0.03	72%	0.02[a]	76%

a: Because of the distributional shape of X_7 there is no difference in MAD for three methods at the right tail percentile (i.e., 95% percentile) of Y_7. We therefore instead examined and reported the MAD for left tail percentile (i.e. 5%).

Table 6

Simulation Comparison of the Performance of Different Fusion Methods on Group Difference (MeanDiff and Q95Diff) Estimates.

Criteria	Method	Y_5		Y_6		Y_7	
		MAD	Improve	MAD	Improve	MAD	Improve
MeanDiff	Parametric	0.08	0%	0.12	0%	0.21	0%
	hot-deck	0.14	-75%	0.16	-33%	0.20	5%
	FORM	0.07	8%	0.09	25%	0.09	57%
Q95Diff	Parametric	0.38	0%	0.29	0%	0.39[a]	0%
	hot-deck	0.54	-42%	0.15	48%	0.21[a]	46%
	FORM	0.25	35%	0.06	76%	0.05[a]	87%

a: Because of the distributional shape of X_7 there is no difference in MAD for three methods at the right tail percentile (i.e., 95% percentile) of Y_7. We therefore instead examined and reported the MAD for left tail percentile (i.e. 5%).

[hP]

Table 7

Simulation Comparison on Individual Predictions.

Method	Y_4		Y_5		Y_6		Y_7	
	RMSE	Improve	RMSE	Improve	RMSE	Improve	RMSE	Improve
Parametric	0.43	0%	1.37	0%	1.66	0%	1.69	0%
hot-deck	0.60	-39%	1.78	-29%	2.13	-28%	1.92	-14%
FORM	0.43	0%	1.11	19%	1.36	18%	1.18	30%

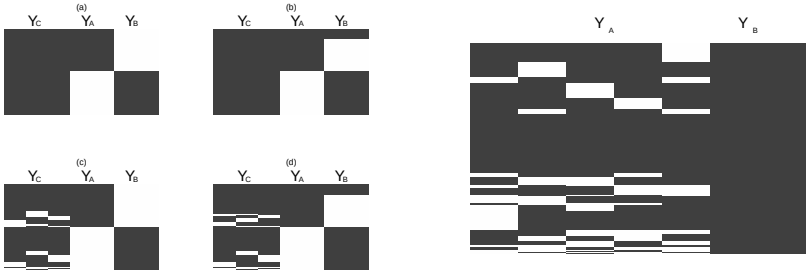
[hP]

Table 8

Resampling Experiments on Data Fusion Results Using the Counterfeit Purchase Datasets.

Parameter	B	Purchase Variables	Param	hot-deck	FORM
KS Test	0	PurchRate	0.15	0.08	0.07
		BaskSize	0.11	0.10	0.09
		Expend	0.11	0.10	0.08
	1	PurchRate	0.16	0.07	0.06
		BaskSize	0.07	0.08	0.07
		Expend	0.08	0.08	0.06
Mean	0	PurchRate	0.01	0.01	0.01
		BaskSize	2.88	5.12	2.94
		Expend	0.91	1.55	0.95
	1	PurchRate	0.00	0.01	0.00
		BaskSize	1.50	2.30	1.45
		Expend	0.68	0.61	0.39
Q95	0	PurchRate	0.07	0.04	0.02
		BaskSize	25.85	18.27	10.41
		Expend	15.12	7.14	3.81
	1	PurchRate	0.03	0.02	0.01
		BaskSize	15.95	10.81	5.78
		Expend	7.48	3.37	1.95

[h]



1 Data Fusion. 2 Missing Covariate.

Figure 1. Data Patterns.

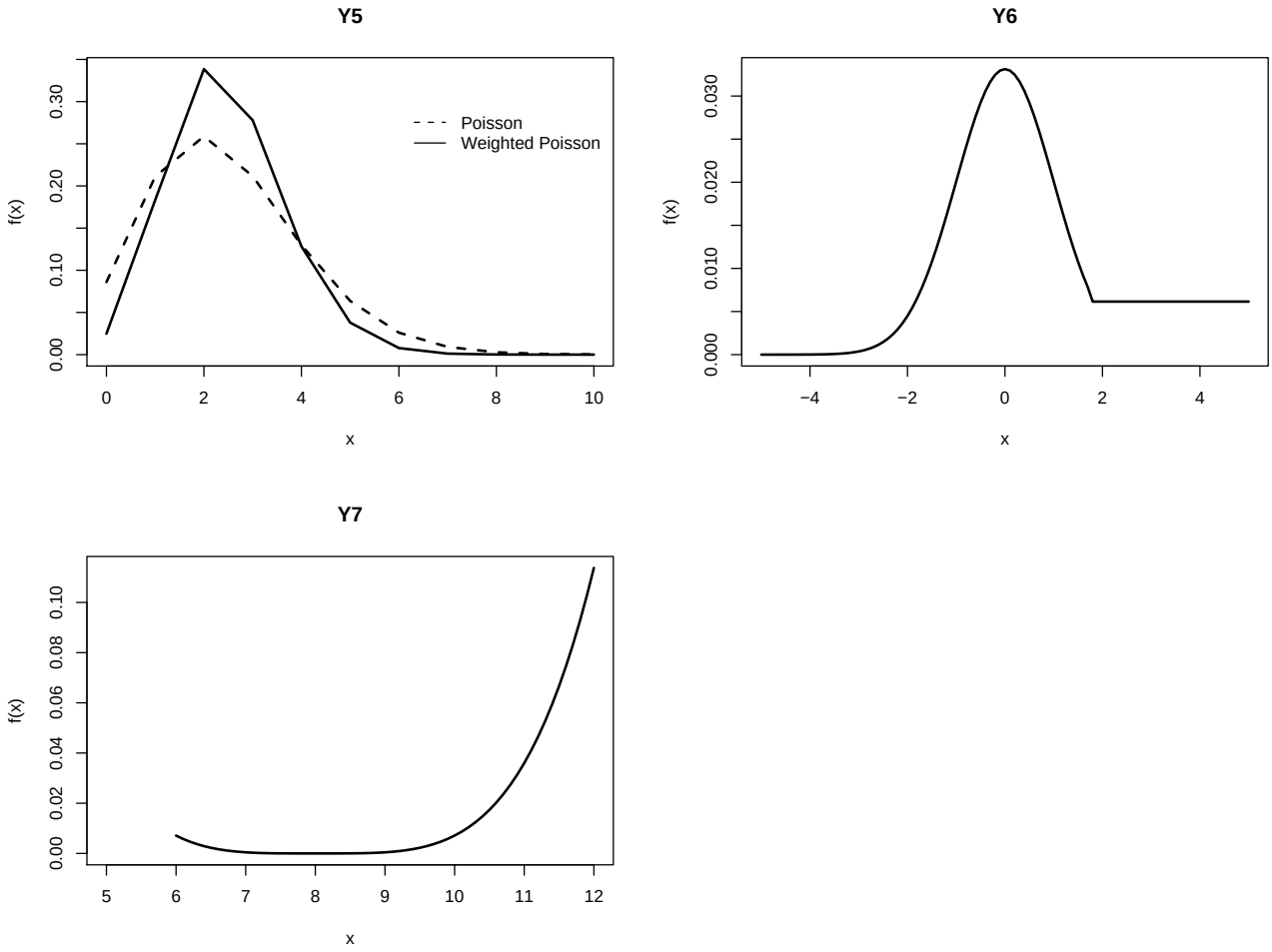


Figure 2. Different Forms of Density Functions at Reference Points of Common Variables.

Online Appendix A: Data Fusion Algorithms.

A.1: MCMC Sampling in LI Approaches to Data Fusion (Fig 1.1.(a))

In this appendix we describe an MCMC algorithm to draw samples from the posterior distribution of the odds ratio model parameters θ_A in $f_{\theta_A}(y_A|y_C)$ and θ_B in $f_{\theta_B}(y_B|y_C)$. Under the conditional independence assumption, and prior independence between θ_A and θ_B, the posterior distribution $f(\theta_A, \theta_B|D)$ factorizes as a product of posterior distributions of θ_A and θ_B which can be sampled separately. Without loss of generality, we describe how to sample θ_A below. Let $(y_{iA}, y_{iC}), i = 1, \cdots, N_A$ denotes the N_A observations on (Y_A, Y_C) in dataset A. Let $(Y_1, \cdots, Y_k)$ denote the variables in Y_C, and $f_{\theta_A}(y_A|y_C) = f_{\theta_A}(y_A|y_k, ..., y_1)$. The posterior density function for θ_A is given as follows:

$$f(\theta_A|y_A, y_1, \cdots, y_k) \propto \prod_{i=1}^{N_A} \frac{\sum_{l=1}^{L_A} 1_{(y_{iA}=y_{Al})} \eta_{\gamma_A}(y_{iA}, y_{A0}; y_{ik}, ..., y_{i1}, y_{k0}, ..., y_{10}) \exp(\lambda_{Al})}{\sum_{l=1}^{L_A} \eta_{\gamma_A}(y_{iA} = y_{Al}, y_{A0}; y_{ik}, ..., y_{i1}, y_{k0}, ..., y_{10}) \exp(\lambda_{Al})} \pi_A(\theta_A), \quad (9)$$

where $\{y_{Al}\}, l = 1, \cdots, L_A$, denotes the set of unique observed values that Y_A takes in the dataset. The model parameter $\theta_A = (\lambda_A, \gamma_A)$. Because λ_A can be viewed as the scale parameters and γ_A can be viewed as the location parameters, we assign an independent prior, $\pi_A(\theta_A) = \pi(\lambda_A)\pi(\gamma_A)$ and

$$\lambda_A \sim MVN(\mu_{\lambda_A}, \nu_{\lambda_A} I_{n_{\lambda_A}}) \quad , \quad \gamma_A \sim MVN(\mu_{\gamma_A}, \nu_{\gamma_A} I_{n_{\gamma_A}}), \quad (10)$$

where μ_{λ_A} and μ_{γ_A} are vectors of length n_{λ_A} and n_{γ_A}, respectively; n_{λ_A} and n_{γ_A} are the length of parameters λ_A and γ_A, respectively; $I_{n_{\lambda_A}}$ and $I_{n_{\gamma_A}}$ are identity matrices of size $n_{\lambda_A} \times n_{\lambda_A}$ and $n_{\gamma_A} \times n_{\gamma_A}$, respectively. In our applications, we assign μ_{λ_A} and μ_{γ_A} to be vectors of zeros and the variances in the prior, $\nu_{\lambda_A} = \nu_{\gamma_A} = 10^4$. As shown later, these constants in the priors are chosen so that they are noninformative relative to the data.

The posterior distribution of θ_A as specified in Equation (9) generally does not have a closed form. The Random-Walk Metropolis-Hasting (RW-MH) is found to be inefficient to sample from the posterior distribution since there are a relatively large number of parameters, which tend to be highly correlated. We adopt the Hybrid Monte Carlo (HMC) method to sample for this posterior distribution. The HMC method is introduced by Duane et al. (1987), described in detail in Liu (2001), and adopted in Qian and Xie (2011) for handling

missing covariates in marketing models. The HMC sampler uses the idea of Molecular Dynamic (MD) to propose new draws, which is followed by a Metropolis acceptance-rejection method to sample from a target distribution. Because the MD exploits the local dynamics of the target distribution, it suppresses the randomness of making proposal draws in the RW-MH algorithm. As a result, the HMC can substantially increase the acceptance rate of a Markov chain while maintaining a fast mixing of the chain. Recent theoretical work by Beskos et al (2011) shows that the acceptance rate under the optimal tuning of a HMC sampler is 0.651. We therefore tune the HMC sampler to have an acceptance rate at a level between 60% and 70%. To sample from the posterior distribution in Equation (9), the HMC augments the parameter θ_A with a vector of invented momentum variables p_A, which has the same dimension as θ_A, and defines the following Hamiltonian function:

$$H_A(\theta_A, p_A) = U_A(\theta_A) + \varphi_A(p_A),$$

where $U_A(\theta_A) = -\ln f(\theta_A|y_A, y_1, \cdots, y_k)$, $\varphi_A(p_A) = -\frac{1}{2}p_A^T p_A$. The HMC algorithm consists of the following steps to produce a new draw:

- Initialize the system with $(\theta_A^{old}, p_A^{old})$.

  Let θ_A^{old} be the current value of θ_A, and let the initial value $\theta_A^0 = \theta_A^{old}$. Generate p'_A from a standard Gaussian distribution and then assign to the system an initial momentum: $p_A^0 = p'_A - \frac{\delta_A}{2}U'_A(\theta_A^0)$, where $U'_A(\theta_A^0)$ is the derivative of $U_A(\cdot)$ with respect to its argument and δ_A is a user-specified stepsize.

- Generate a new state $(\theta_A^{new}, p_A^{new})$ in the phase space.

  Starting from the initial phase space (θ_A^0, p_A^0), an approximate MD algorithm, called leap-frog algorithm, is run L steps to generate a new state (θ_A^L, p_A^L) in the phase space, where

$$\theta_A^l = \theta_A^{l-1} + \delta_A p_A^{l-1},$$
$$p_A^l = p_A^{l-1} - \delta_A^l U'_A(\theta_A^l),$$

  $l = 1, ..., L$, $\delta_A^l = \delta_A$ for $l < L$ and $\delta_A^L = \frac{\delta_A}{2}$; $U'_A(\theta_A^l)$ is the derivative of $U_A(\theta_A)$ with respect to θ_A evaluated at θ_A^l, and the derivative is

$$\frac{\partial U_A(\theta_A)}{\partial \theta_A} = -\sum_i \frac{\partial \ln f_i(\theta_A)}{\partial \theta_A} - \frac{\partial \ln \pi_A(\theta_A)}{\partial \theta_A}, \tag{11}$$

where $\frac{\partial \ln f_i(\theta_A)}{\partial \theta_A}$ is given in Appendix A.2 and $\frac{\partial \ln \pi_A(\theta_A)}{\partial \theta_A}$ is given below,

$$\frac{\partial \ln \pi_A(\theta_A)}{\partial \lambda_{Al}} = -\frac{\lambda_{Al} - \mu_{\lambda_{Al}}}{\nu_{\lambda_A}}, \quad \frac{\partial \ln \pi_A(\theta_A)}{\partial \gamma_{Av}} = -\frac{\gamma_{Av} - \mu_{\gamma_{Av}}}{\nu_{\gamma_A}},$$

and $\mu_{\lambda_{Al}}$ is the corresponding element for λ_{Al} in μ_{λ_A}, and $\mu_{\gamma_{Av}}$ is the corresponding element for γ_{Av} in μ_{γ_A}. Note that the larger ν_{λ_A} and ν_{γ_A}, the smaller the contribution of the priors in the updating. In our applications, we set $\nu_{\lambda_A} = \nu_{\gamma_A} = 10^4$, which are relatively diffuse priors for these model parameters and further increase of these prior constants has little effect on the results.

- Perform accept-rejection of new draw.

  At the end of the leap-frog steps, let the candidate draw $(\theta_A^{prop}, p_A^{prop}) = (\theta_A^L, p_A^L)$. The algorithm accepts the candidate draw according to the following probability:

$$min(1, \exp\left\{-H_A(\theta_A^{prop}, p_A^{prop}) + H_A(\theta_A^{old}, p_A^{old})\right\}).$$

  If the candidate draw is accepted, θ_A^{prop} becomes the new draw; otherwise, θ_A^{old} becomes the new draw.

References in Appendix A.1

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J. and Stuart, A. (2011) "Optimal Tuning of the Hybrid Monte-Carlo Algorithm,", Forthcoming at *Bernoulli*.

Duane, S., Kennedy, A.D., Pendleton, B.J. and Roweth, D. (1987) "Hybrid Monte Carlo," *Physics Letters B,* **195**, 216-222.

Liu, J. (2001) *Monte Carlo Strategies in Scientific Computing,* Springer-Verlag, New York.

A.2: MLE in LI Approaches to Data Fusion (Fig 1.1 (a)).

Because under the CIA, the likelihood of (θ_A, θ_B) factorized to the product of the likelihood for θ_A $(f_{\theta_A}(y_A|y_C))$ and θ_B $(f_{\theta_B}(y_B|y_C))$, we can obtain their MLEs separately. The MLE of θ_A (θ_B) can be obtained through a Newton-Raphson (NR) type algorithm, which iteratively searches for the parameter values that maximize the model likelihood. Let $\hat{\theta}_A = (\hat{\lambda}_A, \hat{\gamma}_A)$ denote the MLE of $\theta_A = (\lambda_A, \gamma_A)$. Let $(y_{iA}, y_{iC}), i = 1, \cdots, N_A$, denote the N_A observations on (Y_A, Y_C) in dataset A. Then $(\hat{\lambda}_A, \hat{\gamma}_A) = \text{argmax}_{\lambda,\gamma} \sum_{i=1}^{N_A} \ln f_i(\theta_A)$, where $f_i(\theta_A) = f_{\lambda_A,\gamma_A}(y_{iA}|y_{iC})$ which has the form of Eqn (6). The NR type algorithm also requires evaluating the score functions which can be straightforwardly evaluated as follows,

$$\frac{\partial \ln f_i(\theta_A)}{\partial \lambda_{Al}} = 1_{(y_{iA}=y_{Al})} - \frac{\eta_{\gamma_A}(y_{Al}, y_{A0}; y_{ik}, ..., y_{i11}, y_{k0}, ..., y_{10}) \exp(\lambda_{Al})}{\sum_{l'=1}^{L_A} \eta_{\gamma_A}(y_{Al'}, y_{A0}; y_{ik}, ..., y_{i1}, y_{k0}, ..., y_{10}) \exp(\lambda_{Al'})},$$

$$\frac{\partial \ln f_i(\theta_A)}{\partial \gamma_{Avm}} = (y_{iA} - y_{A0})(y_{iv} - y_{v0})^m -$$
$$\frac{\sum_{l'=1}^{L_A} \eta_{\gamma_A}(y_{Al'}, y_{A0}; y_{ik}, ..., y_{i1}, y_{k0}, ..., y_{10}) \exp(\lambda_{Al'})(y_{Al'} - y_{A0})(y_{iv} - y_{v0})^m}{\sum_{l'=1}^{L_A} \eta_{\gamma_A}(y_{Al'}, y_{k0}; y_{ik}, ..., y_{i1}, y_{k0}, ..., y_{10}) \exp(\lambda_{Al'})}.$$

$$\frac{\partial \ln f_i(\theta_A)}{\partial \gamma_{Avu}} = (y_{iA} - y_{A0})(y_{iv} - y_{v0})(y_{iu} - y_{u0}) -$$
$$\frac{\sum_{l'=1}^{L_A} \eta_{\gamma_A}(y_{Al'}, y_{A0}; y_{ik}, ..., y_{i1}, y_{k0}, ..., y_{10}) \exp(\lambda_{Al'})(y_{Al'} - y_{A0})(y_{iv} - y_{v0})(y_{iu} - y_{u0})}{\sum_{l'=1}^{L_A} \eta_{\gamma_A}(y_{Al'}, y_{k0}; y_{ik}, ..., y_{i1}, y_{k0}, ..., y_{10}) \exp(\lambda_{Al'})}.$$

We have input the evaluations of model likelihood and score function values into the Quasi-Newton routine *UMING* in Fortran IMSL library to obtain the MLE. Because none of these evaluations involves integration and only summations over a finite number of points are required, the Quasi-Newton algorithm performs well and converges very quickly. There can be two situations where an alternative approach is needed. Because MLE relies on a relatively large sample, it may not perform well in small samples. Furthermore, when there is unintentional missingness in common variables in Y_C, the MLE may encounter difficulties because the likelihood can involve summations over an exploded number of combinatorial terms. In these cases, a Bayesian approach has advantages. Appendix A.1 and A.3 provide details of FORM based on MCMC algorithms.

A.3: MCMC Sampling in FI Approaches to Data Fusion (Fig 1.1.(c)).

In this appendix, we describe an MCMC algorithm to sample from the posterior distribution of model parameters in FI approaches to data fusion under the CIA. The approach provides a robust way to incorporate incomplete cases due to missingness in the common linkage variables in Y_C into data fusion, and can help correct for potential selection bias and recover efficiency loss in data fusion using only complete cases. This approach requires modeling those variables in Y_C that are subject to missingness. For the derivation below, we assume all variables in Y_C are subject to missingness. Let $Y_C = (Y_1, \cdots, Y_k)$ and we use sequential odds ratio models for modeling Y_C as follows. The joint density function for Y_C can be represented as follows:

$$f_{\theta_C}(y_1, ..., y_k) \quad = f_{\theta_1}(y_1) \prod_{j=2}^{k} f_{\theta_j}(y_j | y_{j-1}, ..., y_1),$$

where θ_1 and θ_j denote the parameters in the marginal density function for Y_1 and in the conditional density function for Y_j, respectively; the odds ration models similar to Equation (6) can be applied to model each conditional distribution in the above equation. For simpler cases in which only some variables in Y_C are subject to missingness, our approach conditions on those fully observed variables and models only variables subject to missingness. For example, in our empirical application, the common variables that have missingness are "Age", "Number of Children at Home", and "Family Income". Our analysis conditions on those fully observed common variables $(y_1, \cdots, y_{k-3})$, and then model the three common variables (y_{k-2}, y_{k-1}, y_k) as $f_{\theta_c}(y_{k-2}, y_{k-1}, y_k | y_{k-3}, \cdots, y_1) = \prod_{j=k-2}^{k} f_{\theta_j}(y_j | y_{j-1}, ..., y_1)$, and (y_{k-2}, y_{k-1}, y_k) denotes "Age", "Number of Children at Home", and "Family Income", respectively. The posterior distribution of the model parameters is

$$f(\theta_A, \theta_B, \theta_C | y_A, y_B, y_C^{obs}) \propto f(y_A, y_B, y_C^{obs} | \theta_A, \theta_B, \theta_C) \pi(\theta_A) \pi(\theta_B) \pi(\theta_C),$$

where y_A, y_B, y_C^{obs} collect all the observed data on Y_A, Y_B, Y_C, respectively, and the density

$$f(y_A, y_B, y_C^{obs} | \theta_A, \theta_B, \theta_C) = \int f_{\theta_A}(y_A | y_C^{obs}, y_C^{mis}) f_{\theta_B}(y_B | y_C^{obs}, y_C^{mis}) f_{\theta_C}(y_C^{obs}, y_C^{mis}) dy_C^{mis},$$

where y_C^{mis} denotes all the missing data on Y_C. For model estimation, we augment the model parameters with those missing data y_C^{mis}. The joint posterior distribution of model unknowns is sampled through the MCMC algorithm described below.

1. Initialize $\theta_A = (\lambda_A, \gamma_A), \theta_B = (\lambda_B, \gamma_B), \theta_C = (\lambda_1, \cdots, \lambda_k, \gamma_1, \cdots, \gamma_k)$. A set of readily available starting values can be obtained by assuming that the variables in the data matrix are independent of each other.

2. Impute Y_C^{mis}. Given draws of the model parameters, we can impute the missing values in Y_C. One strategy is to impute missing values one component at a time. Suppose that the jth variable of Y_{iC}, Y_{ij}, is missing for the ith observation in Dataset A where Y_A is observed. Then by the Bayes rule, Y_{ij}^{mis} is drawn from the following multinomial distribution on the unique observed values of this variable denoted as $(Y_{j1}, \cdots, Y_{jL_j})$:

$$Y_{ij}^{mis}|(Y_{i1}, \cdots, Y_{i(j-1)}, Y_{i(j+1)}, \cdots, Y_{ik}, Y_{iA}) \sim multinomial([P_{ij1}, \cdots, P_{ijL_j}]),$$

where the lth component in the multinomial probability vector $[P_{ij1}, \cdots, P_{ijL_j}]$, for $l = 1, \cdots, L_j$, is given as

$$\begin{aligned} P_{ijl} &= \frac{f_\theta(y_{i1}, \cdots, y_{i(j-1)}, y_{jl}, y_{i(j+1)}, \cdots, y_{ik}, y_{iA})}{\sum_{l'=1}^{L_j} f_\theta(y_{i1}, \cdots, y_{i(j-1)}, y_{jl'}, y_{i(j+1)}, \cdots, y_{ik}, y_{iA})} \\ &= \frac{f_{\theta_j}(y_{jl}|\mathcal{Y}_{ij}) \prod_{m=j+1}^{k} f_{\theta_m}(y_{im}|\mathcal{Y}_{im}(y_{jl})) f_{\theta_A}(y_{iA}|\mathcal{Y}_{ik}(y_{jl}))}{\sum_{l'=1}^{L_j} f_{\theta_j}(y_{jl'}|\mathcal{Y}_{ij}) \prod_{m=j+1}^{k} f_{\theta_m}(y_{im}|\mathcal{Y}_{im}(y_{jl'})) f_{\theta_A}(y_{iA}|\mathcal{Y}_{ik}(y_{jl'}))}, \end{aligned}$$

where $\mathcal{Y}_{ij} = (y_{i1}, \cdots, y_{i(j-1)})$ denote the set of conditioning variables for modeling y_{ij}, $\mathcal{Y}_{im}(y_{jl}) = (y_{i1}, \cdots, y_{i(j-1)}, y_{ij} = y_{jl}, y_{i(j+1)}, \cdots, y_{i(m-1)})$ in which the missing value for y_{ij} is replaced with y_{jl}. When imputing data for Y_{ij}^{mis}, all the missing values in Y_{iC}^{mis} except the jth component take the imputed values in the previous iteration. Imputing Y_C^{mis} in Dataset B where Y_B is observed proceeds similarly.

3. Draw $\theta_A = (\lambda_A, \gamma_A), \theta_B = (\lambda_B, \gamma_B), \theta_C = (\theta_1 = (\lambda_1, \gamma_1) \cdots, \theta_k = (\lambda_k, \gamma_k))$. Once missing values in Y_C^{mis} are imputed, we can make draws from the full conditional distributions of these model parameters. Note that when independent priors for $\theta_A, \theta_B, \theta_1, \cdots, \theta_k$ are assigned, their full conditional distributions are also independent. Each set of parameters in $\theta_A, \theta_B, \theta_1, \cdots, \theta_k$ can then be sampled independently from each other using the MCMC algorithm described in Appendix A.1.

4. The iteration is then repeated until convergence.

Given the draws of model parameters $(\theta_A, \theta_B, \theta_C)$ obtained using the above MCMC sampler, we describe the two FI approaches to data fusion below.

(1) *FI-DE.*

In order to incorporate all sampling units in data fusion, the FI-DE approach uses the following joint predictive distribution:

$$f(Y_A, Y_B|D) = \int\int\int \left[\int \sum_{Y_C^{mis}} f_{\theta_A}(Y_A|Y_C^{obs}, Y_C^{mis}) f_{\theta_B}(Y_B|Y_C^{obs}, Y_C^{mis}) f_{\theta_C}(Y_C^{obs}, Y_C^{mis}) dY_C^{obs} \right] f(\theta_A, \theta_B, \theta_C|D) d\theta_A d\theta_B d\theta_C, \quad (12)$$

where $D = (y_A, y_B, y_C^{obs})$, and (y_C^{obs}, y_C^{mis}) are the observed and missing entries in the Y_C matrix, $f(\theta_A, \theta_B, \theta_C|D)$ is the posterior distribution of the parameters. When computing the predictive distribution, we use the draws of model parameters $(\theta_A, \theta_B, \theta_C)$ obtained using the MCMC sampler, as described above, to evaluate the integration with respect to $f(\theta_A, \theta_B, \theta_C|D)$. To evaluate the integration with respect to $f_{\theta_C}(Y_C^{obs}, Y_C^{mis})$, note that with a large sample, the integration with respect to $f_{\theta_C}(Y_C^{obs})$ can be replaced by the summation over the observed data on the common variables Y_C in the datasets in order to avoid modeling Y_C^{obs}. Given each observation on Y_C^{obs}, the integration with respect to $f_{\theta_C}(Y_C^{mis}|Y_C^{obs})$ is replaced with a summation over a finite number of terms when using odds ratio models for modeling these missing common variables.

(2) *FI-MI.*

As in LI-MI, FI-MI stacks datasets A and B together to form a concatenated file with the resulting data matrix denoted as $Y = (Y_A^{obs}, Y_B^{obs}, Y_A^{mis}, Y_B^{mis}, Y_C^{obs}, Y_C^{mis})$, where $(Y_A^{obs}, Y_B^{obs}, Y_C^{obs})$ and $(Y_A^{mis}, Y_B^{mis}, Y_C^{mis})$ collect the observed and missing entries in the data matrix, respectively. FI-MI also consists of two steps. First, we run the MCMC sampler described above. Parameter draws at iterations with sufficiently long intervals between them in the MCMC sampling are retained to make these draws essentially independent. We also save the value of Y_C^{mis} imputed under each retained parameter draw. Given the parameter draws and imputations of Y_C^{mis}, we draw Y_A^{mis} and Y_B^{mis} from $f_{\theta_A}(Y_A^{mis}|Y_C^{obs}, Y_C^{mis})$ and $f_{\theta_B}(Y_B^{mis}|Y_C^{obs}, Y_C^{mis})$. Under our framework, drawing from these distributions is straightforward. For instance, if the ith observation in Y has Y_A unobserved, the imputation is drawn

from $f_{\theta_A}(Y_{iA}^{mis}|Y_{iC}^{obs}, Y_{iC}^{mis})$ which is a discrete distribution on the unique observed values of Y_A with the probability mass function given in Eqn (6).

Online Appendix B: Additional Analysis.

Power Analysis

Because of the relatively large sample size in our empirical application, even attenuated parameter estimates in the parametric data fusion can turn out to be statistically significant. In this case managerial decisions relying solely on testing results tend to be less affected by biased fusion results. To investigate the potential improvement of FORM when researchers have a smaller sample, we perform a power calculation. Specifically, we resample from the original dataset with different proportions of the original sample size. For example, a proportion of 0.33 means that a new sample is drawn with its size one third of the original size. For each sample size proportion, we obtain 100 bootstrap samples and then perform data fusion for each. The power is computed as the proportion of the samples that lead to rejecting the null hypothesis of no association between attitude and shopping behavior variables and counterfeit purchase amount. Fig 3 plots the power curves for the parametric fusion and FORM for those five significant attitude and shopping behavior variables. We observe that the improvement in power for FORM tends to be larger when sample size become smaller for the three variables "Health", "BuyLic" and "BuyInternet". The power can more than double or triple that of the parametric fusion at some smaller sample sizes, indicating the large performance differences in these smaller samples sizes. For "SocialInteraction" and "PromCatalog", whose association with counterfeit expenditure is not detected by the parametric fusion in the original dataset,[21] FORM provides substantially increased power as sample size increases, while the parametric fusion has almost no power because of its stronger bias in data fusion.

Distributional Comparison

Another approach to revealing useful patterns in counterfeit consumption is to examine distributional differences, which can reveal systematic patterns for certain managerially important segments of consumers. For this purpose, we investigate the performance of different methods in comparing heavy counterfeit purchasers between groups (i.e., tail percentile

[21] The variable "PromCatalog"is borderline significant in FORM. However, the more powerful FORM-FI analysis described in the main text reveals stronger and clearer significance for this variable, which increases our confidence in its significance.

comparison). Table 9 reports the 95% percentile comparisons for the above five significant variables. Both hot-deck and FORM reveal statistically significant differences at the 95% percentile for variables 4 and 16, while the parametric fusion cannot reveal significance for either one. In particular, FORM shows a larger and significant effect of the variable "Health" on heavy users of counterfeits which is not identified in parametric fusion.

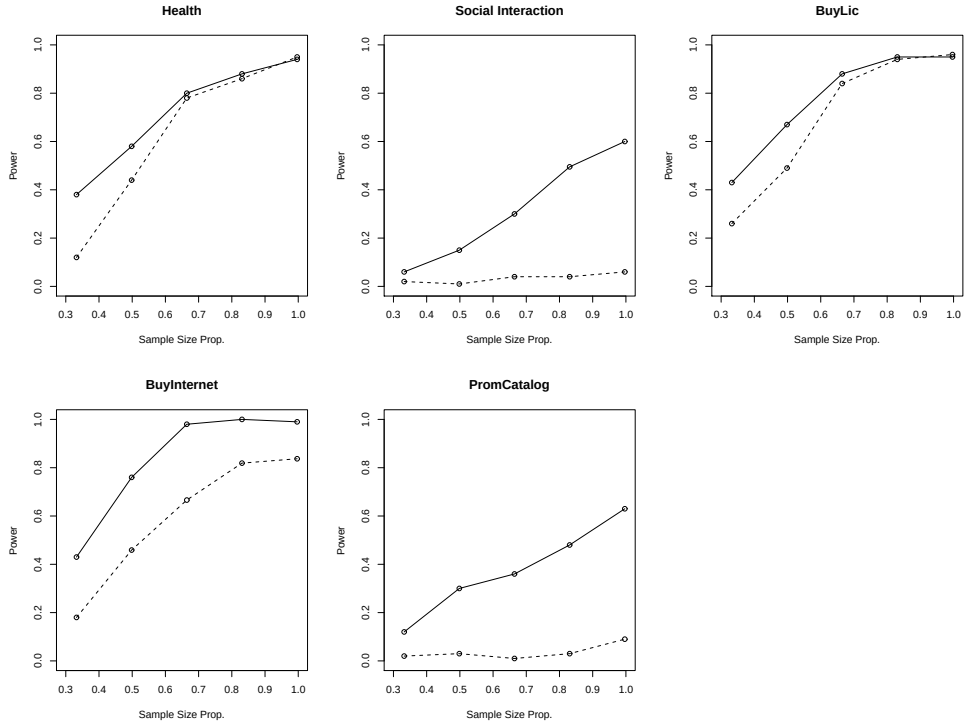


Figure 3. Power Curve for Detecting Association between Y_A and Y_B^*. The power is evaluated at the level of significance of 0.05 except for *PromCatalog*, which is evaluated at the level of significance of 0.10. The solid (dotted) line is the power curve from FORM (parametric data fusion).

Table 9

Data Fusion Results Comparing the tail 95 percentiles of monetary value of counterfeit purchases over the past year (Y_B^*).

Y_A	Param.	Hot-deck	FORM
4. It is good for health.	-26.5 (18.1)	-42.5 (7.4)**	-46.3 (9.2)**
10. I need for work and social interaction.	-12.6 (9.5)	-4.6 (6.6)	-2.3 (4.2)
14. Shop often in licensed store.	-21.1 (15.3)	14.3 (2.2)**	-8.4 (5.7)
16. Shop often on Internet.	31.1 (20.6)	5.7 (2.7)**	12.6 (5.1)**
19. Interested in receiving catalog.	11.9 (20.8)	1.43 (4.6)	0.9 (4.7)

Table 10: Data Fusion Results with the binary variable for counterfeit purchase over the past year (B). The proposed data fusion approach is abbreviated as "FORM", standing for "Fusion via Odds Ratio Models". "FORM-FI" denotes the full-information data fusion, while all the other columns use limited-information. '*', '**', and '***' represent statistical significance at 0.1, 0.05, and 0.001 levels, respectively.

Y_A	Param	Hot-deck	FORM	FORM-FI
Motivations for purchasing shoes with the brand				
1. The brand product is reliable.	-0.07	-0.11	-0.06	-0.01
	(0.12)	(0.09)	(0.12)	(0.09)
2. It is comfortable.	0.07	-0.03	0.08	0.02
	(0.18)	(0.11)	(0.17)	(0.13)
3. The price is reasonable.	-0.31**	-0.19**	-0.29**	-0.21**
	(0.12)	(0.08)	(0.12)	(0.09)
4. It is good for health.	-0.85***	-0.77***	-0.80***	-0.63***
	(0.23)	(0.17)	(0.23)	(0.16)
5. It has a good design.	-0.38	-0.72*	-0.50	-0.24
	(0.52)	(0.39)	(0.46)	(0.37)
6. The materials are fine.	0.11	0.22*	0.14	0.25*
	(0.18)	(0.13)	(0.18)	(0.14)
7. It uses high technology.	-0.31	-0.75*	-0.55	-0.62
	(0.48)	(0.43)	(0.84)	(0.53)
8. It is convenient to buy.	0.04	0.20 **	0.06	0.05
	(0.13)	(0.09)	(0.13)	(0.10)
9. My friends use/recommend it.	-0.07	0.09	-0.05	-0.02
	(0.17)	(0.12)	(0.16)	(0.12)
10. I need for work and social interaction.	-0.20**	-0.08	-0.20*	-0.09
	(0.10)	(0.08)	(0.12)	(0.09)
Places where shop often for shoes.				
11. Shop often in mall.	-0.12	-0.03	-0.13	-0.08
	(0.18)	(0.13)	(0.18)	(0.13)
12. Shop often in supermarket.	-0.04	-0.13	-0.06	0.03
	(0.13)	(0.10)	(0.15)	(0.10)
13. Shop often in discount store.	-0.04	0.03	-0.07	-0.04
	(0.12)	(0.08)	(0.13)	(0.08)
14. Shop often in licensed store.	-0.53***	-0.36 ***	-0.57***	-0.52***
	(0.12)	(0.09)	(0.16)	(0.10)
15. Shop often in open market or on street.	0.22	0.18*	0.21	0.23**
	(0.13)	(0.10)	(0.14)	(0.10)
16. Shop often on Internet.	0.53***	0.44***	0.49***	0.32***
	(0.14)	(0.09)	(0.14)	(0.09)
17. Others shop for me.	-0.05	-0.21	0.20	-0.01
	(0.52)	(0.52)	(0.71)	(0.42)
Attitudes toward promotions				
18. Interested in promotion inside store.	-0.14	-0.28**	-0.15	-.05
	(0.15)	(0.11)	(0.16)	(.11)

continued on next page

Table 10: *continued*

Y_A	Param	Hot-deck	FORM	FORM-FI
19. Interested in receiving catalog.	0.18	0.17*	0.16	0.19*
	(0.14)	(0.10)	(0.17)	(.11)
20. Interested in getting a small gift.	-0.02	0.07	-0.02	.01
	(0.13)	(0.09)	(0.14)	(.09)
21. Interested in getting store credits.	-0.08	0.08	-0.09	.03
	(0.16)	(0.09)	(0.15)	(.09)
Attitudes toward Advertisements				
22. Interested in advertisements in store.	-0.09	-0.08	-0.13	-0.05
	(0.17)	(0.12)	(0.16)	(0.12)
23. Interested in advertisements on TV.	0.02	-0.08	0.05	0.06
	(0.15)	(0.11)	(0.16)	(0.11)
24. Interested in advertisements on Radio.	-0.03	-0.30*	-.16	0.08
	(0.30)	(0.18)	(.28)	(0.17)
25. Interested in advertisements in magazines.	0.10	0.05	.07	-0.05
	(0.20)	(0.13)	(.18)	(.14)
26. Interested in advertisements in firm-sponsored commercial activities.	-0.21	-0.31 *	-.17	-0.11
	(0.23)	(0.18)	(.26))	(.19)
27. Interested in advertisements in public space.	-0.15	0.21	-.06	.12
	(0.26)	(0.21)	(.30)	(.20)

Table 11: Data Fusion Results with the monetary value of counterfeit purchases over the past year (Y_B^*). The proposed data fusion approach is abbreviated as "FORM", standing for "Fusion via Odds Ratio Models". "FORM-FI" denotes the full-information data fusion, while all the other columns use limited-information. '*', '**', and '***' represent statistical significance at 0.1, 0.05, and 0.001 levels, respectively.

Y_A	Param.	Hot-deck	FORM	FORM-FI
Motivations for purchasing shoes with the brand				
1. The brand product is reliable.	-1.9	4.3	-3.4	0.9
	(6.1)	(7.8)	(8.3)	(5.8)
2. It is comfortable.	0.5	-8.6	1.7	-0.4
	(8.7)	(9.8)	(12.0)	(9.8)
3. The price is reasonable.	-5.1	-9.4	-11.8	-9.6
	(6.2)	(7.2)	(9.5)	(7.9)
4. It is good for health.	-21.9**	-28.6**	-35.9**	-26.7**
	(10.3)	(11.3)	(11.6)	(9.8)
5. It has a good design.	-6.2	-21.4	-17.8	-8.1
	(19.4)	(40.5)	(34.2)	(23.4)
6. The materials are fine.	2.3	11.4	5.5	3.6
	(7.7)	(12.6)	(13.9)	(11.0)
7. It uses high technology.	-7.5	-17.1	-16.9	-11.8
	(19.0)	(44.5)	(37.6)	(22.8)
8. It is convenient to buy.	2.1	14.3**	3.5	3.8
	(6.3)	(6.8)	(9.3)	(7.3)
9. My friends use/recommend it.	0.3	-3.3	0.15	1.6
	(7.3)	(10.2)	(12.1)	(8.0)
10. I need for work and social interaction.	-7.3	-14.3**	-16.9**	-17.1**
	(5.2)	(5.7)	(7.4)	(5.9)
Places where shop often for shoes				
11. Shop often in mall.	-4.1	-22.9**	-2.1	-1.6
	(8.2)	(8.5)	(14.0)	(10.0)
12. Shop often in Supermarket.	-.6	-7.1	-0.6	2.1
	(6.7)	(9.7)	(11.4)	(8.6)
13. Shop often in discount store.	-1.5	-5.7	-2.1	-2.2
	(5.9)	(7.1)	(9.1)	(5.6)
14. Shop often in licensed store.	-15.9**	-24.3***	-23.3***	-21.2***
	(5.5)	(3.8)	(6.8)	(5.5)
15. Shop often in open market or on street.	8.8	12.4	11.6	12.6
	(8.1)	(8.3)	(9.1)	(7.8)
16. Shop often on Internet.	14.9**	27.1***	22.9***	22.1***
	(5.6)	(5.3)	(6.1)	(5.7)
17. Others shop for me.	3.2	24.3	5.4	4.3
	(24.9)	(54.6)	(44.8)	(27.8)
Attitudes toward promotion				

continued on next page

Table 11: *continued*

Y_A	Param	Hot-deck	FORM	FORM-FI
18. Interested in promotion inside store.	-5.4	-5.5	-9.6	-4.2
	(6.6)	(9.9)	(10.6)	(8.2)
19. Interested in receiving catalog.	6.8	15.7*	14.7*	15.1**
	(6.5)	(8.0)	(8.3)	(6.9)
20. Interested in getting small gift.	0.1	12.9	1.8	-1.1
	(6.2)	(8.4)	(9.2)	(8.1)
21. Interested in getting store credits.	0.1	10.0	1.7	-3.2
	(6.2)	(6.5)	(11.3)	(7.2)
Attitudes toward advertisements				
22. Interested in advertisements in store.	-3.6	-4.2	-4.5	-3.0
	(7.6)	(10.9)	(12.4)	(9.1)
23. Interested in advertisements on TV.	0.7	-2.9	1.9	-0.8
	(7.2)	(10.0)	(10.8)	(7.1)
24. Interested in advertisements on Radio.	10.0	18.1	6.7	6.3
	(13.4)	(15.6)	(17.2)	(12.9)
25. Interested in advertisements in magazines.	-1.5	12.9	-0.9	-4.6
	(10.6)	(10.7)	(14.5)	(12.5)
26. Interested in advertisements in firm-sponsored commercial activities.	-1.1	8.6	1.0	-1.2
	(11.1)	(18.3)	(17.4)	(15.3)
27. Interested in advertisements in public space.	-0.6	4.3	0.2	8.6
	(11.7)	(20.2)	(19.7)	(16.3)

Table 12: Data Fusion Results with the monetary value of counterfeit purchases over the past year (Y_B^*). The proposed data fusion approach is abbreviated as "FORM", standing for "Fusion via Odds Ratio Models". "FORM-FI" denotes the full-information data fusion, while all the other columns use limited-information. "Param" uses linear regression model without log tranformation. To avoid imputing negative monetary values, we specify a truncation value of zero. '*', '**', and '***' represent statistical significance at 0.1, 0.05, and 0.001 levels, respectively.

Y_A	Param.	Hot-deck	FORM	FORM-FI
Motivations for purchasing shoes with the brand				
1. The brand product is reliable.	-1.5	4.3	-3.4	0.9
	(4.7)	(7.8)	(8.3)	(5.8)
2. It is comfortable.	-1.0	-8.6	1.7	-0.4
	(7.2)	(9.8)	(12.0)	(9.8)
3. The price is reasonable.	-5.8	-9.4	-11.8	-9.6
	(4.5)	(7.2)	(9.5)	(7.9)
4. It is good for health.	-17.8*	-28.6**	-35.9**	-26.7**
	(10.1)	(11.3)	(11.6)	(9.8)
5. It has a good design.	-14.6	-21.4	-17.8	-8.1
	(16.5)	(40.5)	(34.2)	(23.4)
6. The materials are fine.	3.2	11.4	5.5	3.6
	(8.5)	(12.6)	(13.9)	(11.0)
7. It uses high technology.	-10.6	-17.1	-16.9	-11.8
	(17.6)	(44.5)	(37.6)	(22.8)
8. It is convenient to buy.	3.8	14.3**	3.5	3.8
	(6.3)	(6.8)	(9.3)	(7.3)
9. My friends use/recommend it.	0.8	-3.3	0.15	1.6
	(6.3)	(10.2)	(12.1)	(8.0)
10. I need for work and social interaction.	-6.0	-14.3**	-16.9**	-17.1**
	(5.5)	(5.7)	(7.4)	(5.9)
Places where shop often for shoes				
11. Shop often in mall.	-1.7	-22.9**	-2.1	-1.6
	(8.3)	(8.5)	(14.0)	(10.0)
12. Shop often in Supermarket.	-1.2	-7.1	-0.6	2.1
	(5.5)	(9.7)	(11.4)	(8.6)
13. Shop often in discount store.	-2.3	-5.7	-2.1	-2.2
	(4.1)	(7.1)	(9.1)	(5.6)
14. Shop often in licensed store.	-13.1**	-24.3***	-23.3***	-21.2***
	(5.1)	(3.8)	(6.8)	(5.5)
15. Shop often in open market or on street.	8.2	12.4	11.6	12.6
	(6.3)	(8.3)	(9.1)	(7.8)
16. Shop often on Internet.	15.1**	27.1***	22.9***	22.1***
	(4.6)	(5.3)	(6.1)	(5.7)
17. Others shop for me.	2.1	24.3	5.4	4.3
	(20.8)	(54.6)	(44.8)	(27.8)
Attitudes toward promotion				

continued on next page

Table 12: *continued*

Y_A	Param.	Hot-deck	FORM	FORM-FI
18. Interested in promotion inside store.	-5.9	-5.5	-9.6	-4.2
	(5.5)	(9.9)	(10.6)	(8.2)
19. Interested in receiving catalog.	5.3	15.7*	14.7*	15.1**
	(5.6)	(8.0)	(8.3)	(6.9)
20. Interested in getting small gift.	-0.7	12.9	1.8	-1.1
	(5.1)	(8.4)	(9.2)	(8.1)
21. Interested in getting store credits.	-1.1	10.0	1.7	-3.2
	(5.8)	(6.5)	(11.3)	(7.2)
Attitudes toward advertisements				
22. Interested in advertisements in store.	-3.6	-4.2	-4.5	-3.0
	(7.2)	(10.9)	(12.4)	(9.1)
23. Interested in advertisements on TV.	-0.8	-2.9	1.9	-0.8
	(5.5)	(10.0)	(10.8)	(7.1)
24. Interested in advertisements on Radio.	1.5	18.1	6.7	6.3
	(10.2)	(15.6)	(17.2)	(12.9)
25. Interested in advertisements in magazines.	-0.8	12.9	-0.9	-4.6
	(8.1)	(10.7)	(14.5)	(12.5)
26. Interested in advertisements in firm-sponsored	-1.1	8.6	1.0	-1.2
commercial activities.	(9.7)	(18.3)	(17.4)	(15.3)
27. Interested in advertisements in public space.	-7.4	4.3	0.2	8.6
	(9.3)	(20.2)	(19.7)	(16.3)

Table 13: Comparing Different Fusion Methods on Individual Prediction of $\ln(Y_B^*)$.

Method	RMSE	Improve
Parametric with log transformation	0.38	0%
Parametric without log transformation	0.39	-4%
FORM	0.30	22%

Note: Both "FORM" and "Parametric without log transformation" (as in Table 12) model the original outcome without using log transformation. The additional comparison with parametric without log transformation shows that the performance gain of FORM persists whether the parametric model takes log transformation or not.