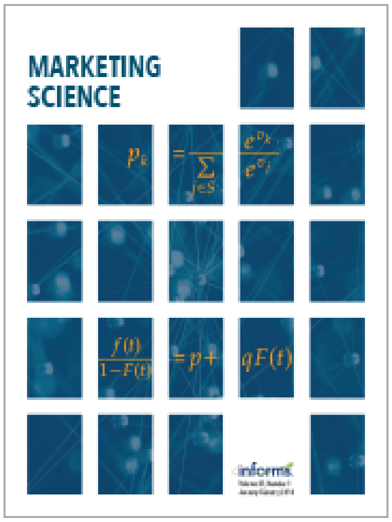


Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

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

Yi Qian, Hui Xie

To cite this article:
Yi Qian, Hui Xie (2014) Which Brand Purchasers Are Lost to Counterfeiters? An Application of New Data Fusion Approaches. Marketing Science 33(3):437-448. https://doi.org/10.1287/mksc.2013.0823

MARKETING SCIENCE
Vol. 33, No. 3, May–June 2014, pp. 437–448
ISSN 0732-2399 (print) | ISSN 1526-548X (online)

http://dx.doi.org/10.1287/mksc.2013.0823

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

Yi Qian
Department of Marketing, Kellogg School of Management, Northwestern University,
Evanston, Illinois 60208, yiqian@nber.org

Hui Xie
Division of Epidemiology and Biostatistics, School of Public Health, University of Illinois,
Chicago, Illinois 60612, huixie@uic.edu

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 they may need to link nonoverlapping 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, and they automate the process of generating suitable distributions that match data, therefore making our methods 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 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 data sets 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 data sets were analyzed separately.

Keywords: counterfeit; CRM; database marketing; nonparametric method; sensitive data; underground economics
History: Received: July 18, 2011; accepted: September 16, 2013; Preyas Desai served as the editor-in-chief and Bart Bronnenberg served as associate editor for this article. Published online in *Articles in Advance* November 26, 2013.

1. Introduction

Firms and organizations frequently need to collect sensitive data from consumers. In product marketing, the 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 looking to enhance their understanding of consumer behaviors and 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 data set—despite the purpose to study the relationship between these variables. This situation can arise for various reasons. In some cases, a single-source comprehensive data set that includes all essential variables is unavailable, prohibitively expensive to collect, or uninformative because of a small sample size. 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 data set 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 environments, there are increasingly strong 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 they have a range of important implications for database marketing (Blattberg et al. 2008). Winer (2001, p. 101) provides a concrete example:

> ... 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 data set with all essential variables is absent or even prohibited when sensitive data are involved. Instead, often available to researchers and policy makers are multiple data sets, 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 data sets; this is known as *data fusion*. The idea is to match nonoverlapping data items from *similar* consumers when matching these data from the *same* consumers is impossible. Global marketing research leaders, such as Nielsen in the United States and Gfk in Europe, have been using data fusion to link sensitive data on patient health, 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 (Rässler 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; they 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 effort 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 significant limitations of existing methods.

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 gross domestic products of 150 economies (Organisation for Economic Co-operation and Development (OECD) 1998). It has substantial effects on nearly all product sectors, ranging from apparel (Qian 2008, 2014) to software (OECD 1998). Despite the increasing prevalence and large impact of counterfeiting, it remains an underresearched area. One of the greatest challenges is the scarcity of detailed and comprehensive consumer-level data, partly as a result of the underground and sensitive nature of counterfeiting. This paper seeks to fill in the gap by combining data sets from an authentic firm's surveys and internal databases. Although none of the available data sets 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 nonparametrically match informative and complex data, the proposed fusion methods improve the identification of consumer behavior patterns in counterfeit consumption, as well as individual predictions. Consequently, they offer the opportunity for better managerial decisions, such as more effective and accurate anticounterfeit planning and targeting.

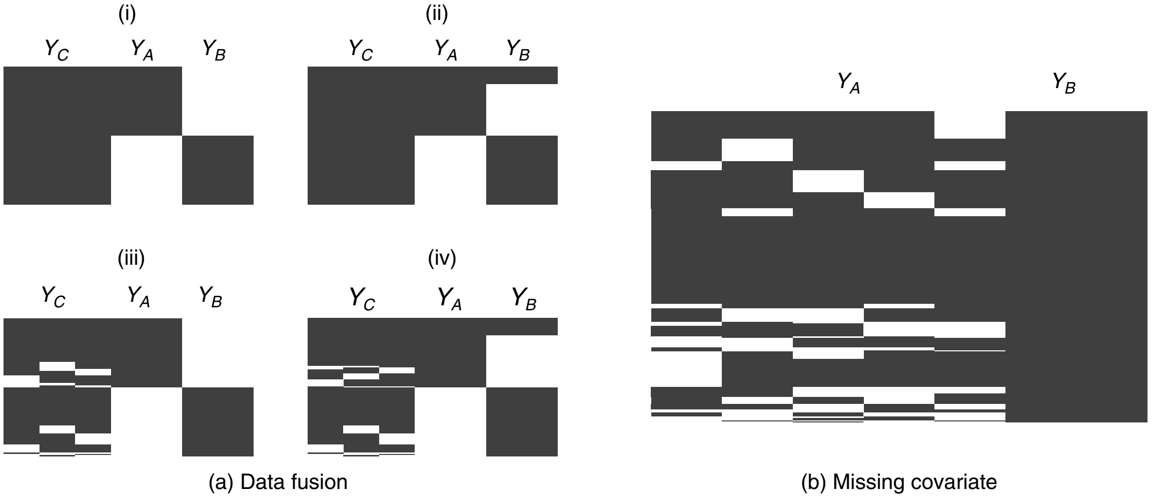
2. Data Fusion Problems and Related Work

Data fusion seeks to relate variables collected from independent samples. Figure 1, panel a(i) 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

[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 the willingness to pay for environmental goods.

[2] For example, see Nielsen (2007).

Figure 1 **Data Patterns**



(a) Data fusion

(b) Missing covariate

on their joint distribution in the concatenated data set. Data fusion is generally made possible by a set of common variables observed in both data sets, 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 is 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, exclusion and inclusion of these variables) can be used for different recipients so that each of them can 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 stem from 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 data sets, more powerful data fusion methods are needed. As noted by Kamakura et al. (2005, p. 285), "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 the following important issues:

1. *Effective and theoretically sound nonparametric solution to data fusion:*[3] Both the hot-deck procedure and our procedure are nonparametric in that neither requires specifying distributional forms for any variable in the data sets. 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 §3, the weights in our method are determined by principled and coherent rules from statistical models with theoretical justifications; thus our method can overcome major drawbacks of hot-deck, including its ad hoc nature in critical steps of data fusion, lack of theory support, invalid statistical significance level, poor

[3] This benefit is most useful for linking variables that have unknown complex distributional forms, but it is minimal for relating binary variables, which have the simplest distributional form. Furthermore, the performance gain as a result of 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.

sampling properties, and difficulty in handling missingness in matching variables. It is important to note that our method simultaneously retains some of the main merits of hot-deck that contribute 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 that outperforms 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 Figure 1, panel b, the authors' 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.

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 (see Figure 1, panels a(i) and a(iii)). We consider fusion using the commonly used CIA, under which $f(Y_A, Y_B \mid Y_C) = f(Y_A \mid Y_C)f(Y_B \mid Y_C)$. Let $(Y_1, \dots, Y_k)$ be the set of common variables in Y_C. We first consider how to model the conditional distribution $f(y_A \mid y_k, \dots, y_1)$, where y_A could be either continuous or discrete. As shown in Chen (2004), the conditional distribution can be rewritten as

$$\begin{aligned} f_{\theta_A}(y_A \mid y_k, \dots, y_1) &= (\eta(y_A, y_{A0}; y_k, \dots, y_1, y_{k0}, \dots, y_{10}) f(y_A \mid y_{k0}, \dots, y_{10})) \\ &\quad \cdot \bigg(\int \eta(y_A, y_{A0}; y_k, \dots, y_1, y_{k0}, \dots, y_{10}) \\ &\qquad \cdot f(y_A \mid y_{k0}, \dots, y_{10})\, dy_A \bigg)^{-1}, \qquad (1) \end{aligned}$$

where θ_A denotes model parameters and $(y_{10}, \dots, y_{k0}, y_{A0})$ is an arbitrarily chosen fixed point in the sample space of $(Y_1, \dots, Y_k, Y_A)$. The main point of Equation (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,

$$\eta(y_A, y_{A0}; y_k, \dots, y_1, y_{k0}, \dots, y_{10}) = \frac{f(y_A \mid y_1, \dots, y_k) f(y_{A0} \mid y_{10}, \dots, y_{k0})}{f(y_A \mid y_{10}, \dots, y_{k0}) f(y_{A0} \mid y_1, \dots, y_k)},$$

is the odds ratio function relative to the reference point $(y_{10}, \dots, y_{k0}, y_{A0})$ and captures the dependence of Y_A on $(Y_1, \dots, Y_k)$. It is a constant (equal to 1) if Y_A is independent of $(Y_1, \dots, Y_k)$. The second component function, $f(y_A \mid y_{k0}, \dots, y_{10})$, 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

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

[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 their prior (Rubin 1974), running these missing covariate methods encounters a model identifiability issue.

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

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

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

The function $f(y_A \mid y_{k0}, \dots, y_{10})$ becomes

$$\exp\left\{\frac{y_A \Psi(\beta, y_{k0}, \dots, y_{10}) - b(\Psi(\beta, y_{k0}, \dots, y_{10}))}{a(\tau)} + c(y_A, \tau)\right\},$$

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 the models' dependence on strong distributional assumptions. To overcome this drawback and enhance modeling robustness, notice that $f(y_A \mid y_{k0}, \dots, y_{10})$ 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, $f(y_A \mid y_{k0}, \dots, y_{10})$ below is modeled nonparametrically, similar to a marginal distribution. Specifically, let $(y_{A1}, \dots, y_{AL_A})$ be the unique observed values in the data set for Y_A. A nonparametric model for $f(y_A \mid y_{k0}, \dots, y_{10})$ assigns probability mass $p_A = (p_{A1}, \dots, p_{AL_A})$ on these unique data points as

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

Similar to 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 Equation (4), we reparameterize p_A as $\lambda_A = (\lambda_{A1}, \dots, \lambda_{AL_A})$ such that $\lambda_{Al} = \ln(p_{Al}/p_{AL_A})$ for $l = 1, \dots, L_A$. Thus,

$$p_{Al} = \frac{\exp(\lambda_{Al})}{\sum_{u=1}^{L_A} \exp(\lambda_{Au})}.$$

Motivated by Equation (3) from the GLMs example, the log-odds ratio function, $\ln \eta_{\gamma_A}(y_A, y_{A0}; y_k, \dots, y_1, y_{k0}, \dots, y_{10})$, is modeled in a bilinear form as

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

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 Equation (3), GLMs have a log-bilinear form of an odds ratio function, and the log-odds parameter γ_A is a reparametrization of the parameters in GLMs. Similar to 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, $f_{\lambda_A}(y_A \mid y_{k0}, \dots, y_{10})$, is modeled nonparametrically in the odds ratio model but parametrically in 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 $f_{\theta_A}(y_A \mid y_k, \dots, y_1)$ assigns point mass to the set of the observed values $(y_{A1}, \dots, y_{AL_A})$ according to the following probability mass function:

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

Prediction or imputation is simple to perform from this discrete distribution on a finite number of data points. More important, as shown above, the odds ratio model nests 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 $f_{\theta_B}(y_B \mid y_k, \dots, y_1)$.

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 below describe LI-DE and LI-MI, the two LI approaches that are designed for the classical data fusion problem as shown in Figure 1, panel a(i).

3.1. 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 $f(Y_A, Y_B \mid D) = \int f(Y_A, Y_B, \theta \mid D)\,d\theta$, where θ collects the model parameters, $D = (D_A, D_B)$, $D_A = (y_{iA}, y_{iC}^A)$, $i = 1, \ldots, N_A$, denotes the N_A observations on (Y_A, Y_C) in data set A, and $D_B = (y_{iB}, y_{iC}^B)$, $i = 1, \ldots, N_B$, denotes the N_B observations on (Y_B, Y_C) in data set B. With CIA, this joint distribution under our LI-DE is

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

and $f(\theta_A, \theta_B \mid D)$ is the posterior distribution of the parameters. Online Appendix §A.1 (available as supplemental material at http://dx.doi.org/10.1287/mksc.2013.0823) provides an Markov chain Monte Carlo (MCMC) algorithm to obtain draws from this posterior distribution, which can then be used to evaluate the integration with respect to $f(\theta_A, \theta_B \mid D)$. 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.,

$$f(Y_A, Y_B \mid D) = \int f_{\hat{\theta}_A}(Y_A \mid Y_C) f_{\hat{\theta}_B}(Y_B \mid Y_C) f_{\theta_C}(Y_C)\,dY_C.$$

The estimate $(\hat{\theta}_A, \hat{\theta}_B)$ can be obtained using the maximum likelihood estimate method, for which Online Appendix §A.2 provides an estimation algorithm. When the 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 data sets to avoid modeling Y_C. Note that as shown in Equation (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.

3.2. LI-MI

We next consider data fusion via multiple imputation (MI). Although MI requires additional work on storing and processing multiple imputed data sets, 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 data sets. In MI, we stack data sets A and B together to form a concatenated file with the resulting data matrix denoted as $Y = (Y_A^{\text{obs}}, Y_B^{\text{obs}}, Y_A^{\text{mis}}, Y_B^{\text{mis}}, Y_C^{\text{obs}})$, where $(Y_A^{\text{obs}}, Y_B^{\text{obs}}, Y_C^{\text{obs}})$ and $(Y_A^{\text{mis}}, Y_B^{\text{mis}})$ 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 data sets 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 (1987) combination rules. Under CIA, the posterior predictive distribution for unobserved data, $f(Y_A^{\text{mis}}, Y_B^{\text{mis}} \mid Y_A^{\text{obs}}, Y_B^{\text{obs}}, Y_C^{\text{obs}})$, is

$$\iint f_{\theta_A}(Y_A^{\text{mis}} \mid Y_C^{\text{obs}}) f_{\theta_B}(Y_B^{\text{mis}} \mid Y_C^{\text{obs}}) \cdot f(\theta_A, \theta_B \mid Y_A^{\text{obs}}, Y_B^{\text{obs}}, Y_C^{\text{obs}})\,d\theta_A\,d\theta_B. \tag{8}$$

To draw imputations of $(Y_A^{\text{mis}}, Y_B^{\text{mis}})$, first draw (θ_A, θ_B) from their posterior distributions using the MCMC algorithm in Online Appendix §A.1. After the 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 $f_{\theta_A}(Y_A^{\text{mis}} \mid Y_C^{\text{obs}})$ and $f_{\theta_B}(Y_B^{\text{mis}} \mid Y_C^{\text{obs}})$. 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 $f_{\theta_A}(Y_{iA}^{\text{mis}} \mid Y_{iC}^{\text{obs}})$, which is a discrete distribution on the unique observed values of Y_A with the probability mass (weight) function given in Equation (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 data sets, 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 occurs in common variables, as shown in Figure 1, panel a(iii). 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. We have also conducted extensive simulations studies to evaluate the performance of the proposed methods; these studies demonstrate the methods' ability to automatically adapt to arbitrary data departures from regular distributional forms and to substantially improve fusion precision and individual consumer predictions. These results are available from authors upon request.

4. Empirical Application

4.1. Data Description

The goal of our empirical application is to identify systematic patterns in consumers' counterfeit consumption behaviors. The data sets to be analyzed are from a famous brand-name footwear company in China; they include data from two surveys and the firm's internal consumer database. Because of a surge in counterfeits of this firm's popular product lines in the market, in 2009, the firm conducted a survey on a sample of its consumers. The main purpose of the survey was to assess the situation, and the survey included 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 for 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 the firm's consumers. The second survey was designed to increase the understanding of consumer purchasing habits and attitudes and had multiple questions addressing these issues; 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 contained multiple other questions that addressed other marketing research areas. 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 from the survey questions, the firm would be wise not to collect all the data in one survey.[7]

Table 1 lists the variables we consider, 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 1,698 and 3,205 observations for the first and second surveys, respectively. It is impossible to relate counterfeit purchases to consumers' purchasing habits and attitudes if the two data sets currently available to the store manager 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 data sets 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 thus excludes all consumers whose common variables are not fully observed. The resulting complete-case sample contains 3,010 consumers. We then apply three data fusion methods: hot-deck, the parametric fusion, and our proposed fusion via odds ratio models (FORM) 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 §3. Let Y_B denote the monetary value of counterfeits purchased over the past year. To model this unique variable that has a nonzero probability at

[7] The biasing effects can also limit the use of 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 that such long-survey data should not be collected. 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 controls for bias in the joint sample as a result of context effects and survey fatigue. Extending the data fusion approaches proposed here to this situation would be very valuable.

Table 1 Common and Unique Variables in Data Fusion

Common variables (Y_C)	Mean (SD)
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 (10 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)
Authentic product-purchasing motivations, behaviors, and attitudes variables (Y_A)	Percentage of yes responses
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 my 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 I often shop for shoes	
11. Shop often in the mall.	13
12. Shop often in the supermarket.	26
13. Shop often in a discount store.	13
14. Shop often in a licensed store.	44
15. Shop often in an open market or on the street.	20
16. Shop often on the Internet.	24
17. Others shop for me.	2
Attitudes toward promotions	
18. Interested in promotions inside the store.	41
19. Interested in receiving a 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 the 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 spaces.	8
Counterfeit product purchase variables (Y_B)	Summary
Purchased/did not purchase counterfeits of authentic brand in past year (% yes responses)	21.2
If purchased, total monetary value paid for counterfeit products ($)	73.7 (29.7)

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

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 a consumer's decision on *whether* to purchase counterfeit and a continuous variable Y_B^* for a consumer's decision on the amount to purchase 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 $f(B \mid Y_C)$ and $f(Y_A \mid Y_C)$, in which all the common variables (except *age* and the three store purchasing variables) enter the model as dummy covariates, based on their unique levels in the data sets.[8] This modeling strategy helps guard against the possible nonlinear additive effects of these common variables and is used throughout for all models. When the binary purchase decision variable B is imputed as 0, the purchase amount Y_B^* is undefined. When B is imputed as 1, 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 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 the counterfeit purchase amount is used in FORM. This circumvents the need to select a proper transformation 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.

Equation (8) was used to make imputations from the posterior predictive distributions of unobserved unique variables, and the LI-MI algorithm described in §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, 20 imputations were generated.[9] We then conducted analyses on each imputed data set and used

[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.

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

Table 2 **Fusion Results with the Binary Variable for Counterfeit Purchases Over the Past Year** (*B*)

Y_A	Parametric fusion		Hot-deck		FORM		FORM-FI	
3. The price is reasonable.	−0.31**	(0.12)	−0.19**	(0.08)	−0.29**	(0.12)	−0.21**	(0.09)
4. It is good for my health.	−0.85***	(0.23)	−0.77***	(0.17)	−0.80***	(0.23)	−0.63***	(0.16)
5. It has a good design.	−0.38	(0.52)	−0.72*	(0.39)	−0.50	(0.46)	−0.24	(0.37)
6. The materials are fine.	0.11	(0.18)	0.22*	(0.13)	0.14	(0.18)	0.25*	(0.14)
7. It uses high technology.	−0.31	(0.48)	−0.75*	(0.43)	−0.55	(0.84)	−0.62	(0.53)
8. It is convenient to buy.	0.04	(0.13)	0.20**	(0.09)	0.06	(0.13)	0.05	(0.10)
10. I need for work and social interaction.	−0.20**	(0.10)	−0.08	(0.08)	−0.20*	(0.12)	−0.09	(0.09)
14. Shop often in a licensed store.	−0.53***	(0.12)	−0.36***	(0.09)	−0.57***	(0.16)	−0.52***	(0.10)
15. Shop often in an open market or on the street.	0.22	(0.13)	0.18*	(0.10)	0.21	(0.14)	0.23**	(0.10)
16. Shop often on the Internet.	0.53***	(0.14)	0.44***	(0.09)	0.49***	(0.14)	0.32***	(0.09)
18. Interested in promotions inside the store.	−0.14	(0.15)	−0.28**	(0.11)	−0.15	(0.16)	−0.05	(0.11)
19. Interested in receiving a catalog.	0.18	(0.14)	0.17*	(0.10)	0.16	(0.17)	0.19*	(0.11)
24. Interested in advertisements on the radio.	−0.03	(0.30)	−0.30*	(0.18)	−0.16	(0.28)	0.08	(0.17)
26. Interested in advertisements in firm-sponsored commercial activities.	−0.21	(0.23)	−0.31*	(0.18)	−0.17	(0.26)	−0.11	(0.19)

*Statistical significance at the 0.1 level; **statistical significance at the 0.05 level; and ***statistical significance at the 0.001 level.

Rubin's (1987) combination rule to pool results from multiple imputed data sets. 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 2 and 3. 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 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 stemming from 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 the nonnormal distributional feature of Y_B^*, 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 the 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, 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," which incorporates all consumers, including those with missingness in common variables.[12] The results show that FORM-FI substantially increases fusion efficiency

[10] Because perfect matching is impossible for *age* and three store purchasing variables that 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 0 to 1 (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.

[12] The common variables subject to missingness are *age, family income*, and *number of children at home*, which collectively cause about 40% of the cases to be incomplete; these are dropped from LI fusion but incorporated into the FI analysis. Our FORM-FI analysis uses the FI-MI procedure as described in Online Appendix §A.3.

Table 3 **Fusion Results with the Monetary Value of Counterfeit Purchases Over the Past Year (Y_B^*)**

Y_A	Parametric fusion		Hot-deck		FORM		FORM-FI	
4. It is good for my 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 it 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 the mall.	−4.1	(8.2)	−22.9**	(8.5)	−2.1	(14.0)	−1.6	(10.0)
14. Shop often in a licensed store.	−15.9**	(5.5)	−24.3***	(3.8)	−23.3***	(6.8)	−21.2***	(5.5)
16. Shop often on the Internet.	14.9**	(5.6)	27.1***	(5.3)	22.9***	(6.1)	22.1***	(5.7)
19. Interested in receiving a catalog.	6.8	(6.5)	15.7*	(8.0)	14.7*	(8.3)	15.1**	(6.9)

*Statistical significance at the 0.1 level; **statistical significance at the 0.05 level; and ***statistical significance at the 0.001 level.

and considerably reduces standard errors compared with the complete-case analysis reported under the "FORM" column. It is noteworthy that 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 an open market or on the 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, 2014), 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 purchase motivations, behaviors, and attitudes toward branded products, promotions, and channels. Understanding these marketing reactions can generate insights for protecting a 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 fewer 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 that stress the benefits of authentic products compared with those of counterfeits.

The analyses also help identify the channels through which counterfeits are brought to market. 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, the Internet provides firms with a convenient marketing channel. Unfortunately, it also serves counterfeiters. To reduce Internet use by counterfeiters, an authentic company might educate consumers about the harmful effects of counterfeits and post tips on its website for identifying fake merchandise. The firm could also work with Internet malls to reduce the number of counterfeits through raids and/or legal actions, as in the case of Louis Vuitton successfully suing eBay in Europe (Carvajal 2008). 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. Compared with legal actions, these marketing measures may more effectively entice consumers to return 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 made from using new data fusion methods.

Improved Identification of Patterns in Counterfeit Consumption for Anticounterfeit Planning. As shown above, an important step in anticounterfeit 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*), whereas 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 (see Table 2). In addition, 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 anticounterfeiting planning.[13] For example, the usefulness of emphasizing the social interaction benefits of branded shoes in anticounterfeit advertisements, promotions through catalogs to make branded shoes more appealing to affected consumers, and counterfeit control in an open market may be underestimated or ignored.

Online Appendix B reports additional analyses of this data set 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 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 root mean square error (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.

Table 4 Comparing Different Fusion Methods on Individual Prediction of $\ln(Y_B^*)$

Method	RMSE	Improvement (%)
Parametric	0.38	0
Hot-deck	0.57	−50
FORM	0.30	22

5. Discussion

Because of 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 it 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 data sets from multiple sources.

Because the situations where relevant data can only be found in separate data sets 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 data set 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 data sets 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. It is important to note that 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, to inform timely managerial and policy decisions.

In many of these marketing tasks, the databases can include a mixture of highly disparate nonoverlapping variables with unknown complex distributional forms. Although more empirical applications

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

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; thus their high objectivity ensures that 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 (see Figure 1, panels a(ii) and a(iv)). 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.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/mksc.2013.0823.

Acknowledgments

The authors thank the editor, the associate editor, and anonymous referees whose contructive comments helped to improve earlier versions of the article. They thank Ulf Böckenholt, Chaoqun Chen, Dennis Zhang, seminar participants at Northwetern University, and the 2011 Joint Statistical Meetings for their very helpful comments on earlier versions of the paper. The first author received support from the National Natural Science Foundation of China [Grants 70972142 and 71172217]. The second author received financial support from the National Institutes of Health National Center for Research Resources [Grant UL1RR029879]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Y. Qian's current affiliation is the National Bureau of Economic Research, Cambridge, Massachusetts 02138.

References

Blattberg RC, Kim B-D, Neslin SA (2008) *Database Marketing: Analyzing and Managing Customers* (Springer, New York).

Carvajal D (2008) EBay ordered to pay $61 million in sale of counterfeit goods. *New York Times* (July 1) http://nytimes.com/2008/07/01/technology/01ebay.html.

Chen HY (2004) Nonparametric and semiparametric models for missing covariates in parametric regression. *J. Amer. Statist. Assoc.* 99(468):1176–1189.

Feit EM, Beltramo MA, Feinberg FM (2010) Reality check: Combining choice experiments with market data to estimate the importance of product attributes. *Management Sci.* 56(5): 785–800.

Gelman A, Carlin JB, Stern HS, Rubin DB (2004) *Bayesian Data Analysis*, 2nd ed. (Chapman & Hall, London).

Gilula Z, McCulloch RE, Rossi PE (2006) A direct approach to data fusion. *J. Marketing Res.* 43(1):73–83.

Kamakura WA, Wedel M (1997) Statistical data fusion for cross-tabulation. *J. Marketing Res.* 34(4):485–498.

Kamakura WA, Wedel M (2000) Factor analysis and missing data. *J. Marketing Res.* 37(4):490–498.

Kamakura WA, Mela CF, Ansari A, Bodapati A, Fader P, Iyengar R, Naik P, et al. (2005) Choice models and customer relationship management. *Marketing Lett.* 16(3/4):279–291.

Nielsen (2007) New "start with the consumer" initiatives integrate information across the company. Press release (May 14), Nielsen, New York. http://www.nielsen.com/us/en/insights/press-room/2007/Nielsen_Introduces _First_Suite_of_NielsenConnect_Services.html.

Organisation for Economic Co-operation and Development (OECD) (1998) The economic impact of counterfeiting. Report, OECD, Paris. http://www.stop-piracy.ch/documents/s10_economic_impact.pdf.

Pouta E (2004) Attitude and belief questions as a source of context effect in a contingent valuation survey. *J. Econom. Psych.* 25(2):229–242.

Qian Y (2008) Impacts of entry by counterfeiters. *Quart. J. Econom.* 123(4):1577–1609.

Qian Y (2014) Counterfeiters: Foes or friends? How counterfeits affect sales by product quality tier. *Management Sci.* Forthcoming.

Qian Y, Xie H (2011) No customer left behind: A distribution-free Bayesian approach to accounting for missing Xs in marketing models. *Marketing Sci.* 30(4):717–736.

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

Rubin DB (1974) Characterizing the estimation of parameters in incomplete-data problems. *J. Amer. Statist. Assoc.* 69(346): 467–474.

Rubin DB (1987) *Multiple Imputation for Nonresponse in Surveys* (John Wiley & Sons, New York).

Winer SR (2001) A framework for customer relationship management. *Calif. Management Rev.* 43(4):89–105.