Statistica Sinica **34** (2024), 1505-1520
doi:https://doi.org/10.5705/ss.202022.0228

ON COMBINING INDIVIDUAL-LEVEL DATA WITH SUMMARY DATA IN STATISTICAL INFERENCES

Lu Deng[1], Sheng Fu[2], Jing Qin[3] and Kai Yu[*2]

[1]*Nankai University,* [2]*National Cancer Institute, and* [3]*National Institue of Allergy and Infectious Diseases*

Abstract: Statistical models and inferences are typically based on measurements made on individual participants in a study (individual-level data). However, there is growing interest in improving statistical inference by taking advantage of aggregated summary-level data from other studies, such as statistics used in meta-analyses. Although the generalized method of moments (GMM) provides a flexible way of doing so, integrating external summary information does not always improve efficiency. Here, we provide a necessary and sufficient condition under which external summary information can be beneficial. We further extend the GMM to incorporate summary data generated from a population with a covariate distribution that is different from that of the individual-level data. Lastly, we compare the GMM with other integration procedures.

Key words and phrases: Empirical likelihood, generalized linear model, generalized method of moments, meta-analysis, summary statistics.

1. Introduction

Statistical inferences are usually conducted on detailed individual-level data observed on each participant in a study. Including relevant aggregated summary data from other studies would be preferred, although procedures for achieving such a goal might be not readily available. One exception is in the setting of meta-analysis, where estimates from comparable models established by different studies can be combined to form a more efficient estimate.

We consider a setting in which we use individual-level data (X, Y) from an internal study to investigate an underlying conditional model $f(Y \mid X; \theta)$, which specifies the conditional distribution of the outcome Y given the covariates X, with θ being the unknown parameter of interest. In addition, we assume we have summary data, represented by a set of estimates $\tilde{\beta}$, derived from external studies. The goal is to obtain a more efficient estimation of θ by combining the raw data (X, Y) from the internal study and $\tilde{\beta}$ from external studies. As in Qin (2000) and others (Imbens and Lancaster (1994); Qin et al. (2015); Chatterjee et al. (2016); Han and Lawless (2016); Cheng et al. (2018, 2019); Han and Lawless (2019); Kundu, Tang and Chatterjee (2019); Huang and Qin (2020); Zhang et al. (2020,

*Corresponding author.

2022)), we consider a broad class of summary information $\tilde{\beta}$, the true underlying value of which β satisfies a set of stochastic constraint equations $E\{u(X;\theta,\beta)\} = 0$, with the expectation taken over a fully unspecified distribution of X. For example, Imbens and Lancaster (1994) consider the case in which β is the mean value of a known function $\varphi(Y,X)$, and $\tilde{\beta}$ is the moment estimate of β based on an external study. In this case, $u(X;\theta,\beta) = E\{\varphi(Y,X) \mid X\} - \beta$, with the conditional expectation calculated over $f(Y \mid X;\theta)$. Chatterjee et al. (2016) consider a class of model-based summary data consisting of a set of coefficient estimates derived from a working parametric model different from $f(Y \mid X;\theta)$.

There are two general strategies for combining summary data with individual-level data, one based on the generalized method of moments (GMM), and the other based on the empirical likelihood framework. Imbens and Lancaster (1994) demonstrated that a GMM offers an effective way of integrating the two types of data. Kundu, Tang and Chatterjee (2019) use a GMM as a meta-analysis procedure to integrate summary statistics from different models. Their approach requires a set of reference samples that are independent of all summary data. Qin (2000) propose using the empirical likelihood approach to incorporate external summary information. A similar empirical likelihood procedure was adopted by Chatterjee et al. (2016) to synthesize general model-based summary statistics. Zhang et al. (2020) expanded the empirical likelihood approach to integrate summary data more efficiently by properly accounting for the uncertainty in $\tilde{\beta}$.

Under both the GMM and the empirical likelihood frameworks, it can be shown that adding summary data at least does not decrease the efficiency of the estimate of θ, compared with the standard maximum likelihood estimate (MLE) based on the internal study alone. However, in some cases, using external summary data does not improve the efficiency of estimates of certain components of θ. In this report, we identify a necessary and sufficient condition under which external summary information can improve efficiency. We also extend the GMM to incorporate summary data generated from a population with a different covariate distribution from that of the individual-level data. This is also called a covariate shift, a common phenomenon in practice (Sugiyama, Krauledat and Müller (2007); Moreno-Torres et al. (2012)). Finally, we show that the GMM and the empirical likelihood procedure of Zhang et al. (2020) are asymptotically equivalent.

2. Method

2.1. Notation and setup

Assume that we have an internal study consisting of random samples (X_i, Y_i), for $i = 1, \ldots, n$, from a targeted population, with Y being the outcome and X being the set of covariates. Let $f(X)$ be the distribution of X and $f(Y \mid X;\theta)$

be the underlying conditional distribution for Y given X, with θ being the set of parameters of interest. In addition to the internal study, we assume we have summary data extracted from an external study, which consists of N random samples $(X_i^{(E)}, Y_i^{(E)})$, for $i = 1, \ldots, N$, from the same or a different population.

Without loss of generality, we assume that the summary data $\tilde{\beta}$ is the solution of following estimating equations based on the external data:

$$\sum_{i=1}^{N} h(X_i^{(E)}, Y_i^{(E)}; \alpha, \beta) = 0, \tag{2.1}$$

where $h(\cdot)$ is a vector function defined by the method chosen for analyzing the external data, with the same dimension as (α, β) to ensure identifiability. Although (α, β) can be estimated from (2.1), we assume that only $\tilde{\beta}$, the estimate of β, can be used as external summary data. The vector α consists of nuisance parameters, with their estimates inaccessible to the final integrative analysis.

2.2. Integrating summary data from the same study population

Imbens and Lancaster (1994) show how to use the GMM to integrate individual-level data with information on moments of the marginal distribution of a certain variable. Here, we use their framework to integrate the summary data $\tilde{\beta}$, in the presence of the nuisance parameter α, by assuming that both the internal and the external studies are conducted in the same population.

Following the argument by White (1982), under general regularity conditions, $(\tilde{\alpha}, \tilde{\beta})$ resolved from (2.1) are consistent estimates of their population values (α_0, β_0), which jointly satisfy the stochastic constraint equation $E\{h(X, Y; \alpha_0, \beta_0)\} = 0$. Hereafter, unless specified otherwise, we use $E\{\xi(X, Y)\}$ and $\text{var}\{\xi(X, Y)\}$ to represent the mean and variance, respectively of a function $\xi(X, Y)$ under the true distribution (X, Y), which is specified as $f(X)f(Y \mid X; \theta_0)$, with θ_0 being the true value of θ. By letting

$$u(X; \theta, \alpha, \beta) = \int_Y h(X, Y; \alpha, \beta) f(Y \mid X; \theta) dY, \tag{2.2}$$

we can re-express the stochastic constraint equation $E\{h(X, Y; \alpha, \beta)\} = 0$ as

$$\int_X u(X; \theta, \alpha, \beta) dX = 0.$$

Based on the internal study, we obtain $\breve{\theta}, \breve{\alpha}$, and $\breve{\beta}$, that is, the intermediate estimates of θ, α, and β, respectively, using the following estimating equations:

$$\sum_{i=1}^{n} \psi_1(Y_i, X_i; \theta) = 0, \quad \sum_{i=1}^{n} \psi_2(X_i; \theta, \alpha, \beta) = 0, \tag{2.3}$$

with

$$\psi_1(Y, X; \theta) = \frac{\partial \log f(Y \mid X; \theta)}{\partial \theta}, \quad \psi_2(X; \theta, \alpha, \beta) = u(X; \theta, \alpha, \beta).$$

Combining these estimates with the external summary data $\tilde{\beta}$, we know

$$n^{1/2} \begin{pmatrix} \breve{\theta} - \theta_0 \\ \breve{\alpha} - \alpha_0 \\ \breve{\beta} - \beta_0 \\ \tilde{\beta} - \beta_0 \end{pmatrix} \xrightarrow{d} N \left[0, \begin{pmatrix} H & 0 \\ 0 & \Sigma/\rho \end{pmatrix} \right],$$

where

$$H = \begin{pmatrix} E\dfrac{\partial \psi_1}{\partial \theta} & 0 & 0 \\ E\dfrac{\partial \psi_2}{\partial \theta} & E\dfrac{\partial \psi_2}{\partial \alpha} & E\dfrac{\partial \psi_2}{\partial \beta} \end{pmatrix}^{-1} \begin{pmatrix} E(\psi_1 \psi_1^{\mathrm{T}}) & 0 \\ 0 & E(\psi_2 \psi_2^{\mathrm{T}}) \end{pmatrix} \begin{pmatrix} E\dfrac{\partial \psi_1}{\partial \theta^{\mathrm{T}}} & E\dfrac{\partial \psi_2}{\partial \theta^{\mathrm{T}}} \\ 0 & E\dfrac{\partial \psi_2}{\partial \alpha^{\mathrm{T}}} \\ 0 & E\dfrac{\partial \psi_2}{\partial \beta^{\mathrm{T}}} \end{pmatrix}^{-1},$$

with $(\theta, \alpha, \beta) = (\theta_0, \alpha_0, \beta_0)$ in the calculation of H, $\mathrm{cov}(\tilde{\beta}) = N^{-1}\Sigma$, and $N/n \to \rho$. We obtain the estimate of (θ, α, β) as

$$(\hat{\theta}_{\mathrm{CMD}}, \hat{\alpha}_{\mathrm{CMD}}, \hat{\beta}_{\mathrm{CMD}}) = \underset{(\theta, \alpha, \beta)}{\operatorname{argmin}} \begin{pmatrix} \breve{\theta} - \theta \\ \breve{\alpha} - \alpha \\ \breve{\beta} - \beta \\ \tilde{\beta} - \beta \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} H^{-1} & 0 \\ 0 & \rho\Sigma^{-1} \end{pmatrix} \begin{pmatrix} \breve{\theta} - \theta \\ \breve{\alpha} - \alpha \\ \breve{\beta} - \beta \\ \tilde{\beta} - \beta \end{pmatrix}. \tag{2.4}$$

This type of estimate is called the classic minimum distance (CMD) estimation (Newey and McFadden (1994)). Because of its close relationship with the GMM, we still consider it as a type of GMM estimate. In practice, we usually do not know H and Σ. Instead, we can use the standard two-step estimation procedure. First, the CMD (Newey and McFadden (1994)) estimate yields a consistent estimate of $(\theta_0, \alpha_0, \beta_0)$ by replacing H and Σ with any positive-definite matrices in (2.4). Next, we obtain consistent estimates of H and Σ using the initial estimates, and plug them into (2.4) to obtain the final efficient estimate $(\hat{\theta}_{\mathrm{CMD}}, \hat{\alpha}_{\mathrm{CMD}}, \hat{\beta}_{\mathrm{CMD}})$. Based on its asymptotic distribution, given in the Appendix, the efficiency grows as the external sample size N increases. When N is much larger than n, so that $\rho \to \infty$, its asymptotic variance becomes the same as that when the variability of the summary data is ignored.

Following Imbens and Lancaster (1994), we estimate θ and α using another type of GMM estimate. Letting $\psi(Y, X; \theta, \alpha, \beta) = (\psi_1(Y, X; \theta)^{\mathrm{T}}, \psi_2(X; \theta, \alpha, \beta)^{\mathrm{T}})^{\mathrm{T}}$, we obtain the GMM estimate as

$$(\hat{\theta}_{\mathrm{GMM}}, \hat{\alpha}_{\mathrm{GMM}}) = \underset{(\theta,\alpha)}{\operatorname{argmin}} \left[\frac{1}{n}\sum_{i=1}^{n} \psi(Y_i, X_i; \theta, \alpha, \tilde{\beta})\right]^{\mathrm{T}} C \left[\frac{1}{n}\sum_{i=1}^{n} \psi(Y_i, X_i; \theta, \alpha, \tilde{\beta})\right],$$

where $C = (\mathrm{var}\{n^{-1/2}\sum_{i=1}^{n} \psi(Y_i, X_i; \theta_0, \alpha_0, \tilde{\beta})\})^{-1}$. Again, a standard two-step estimation procedure can be used to obtain $(\hat{\theta}_{\mathrm{GMM}}, \hat{\alpha}_{\mathrm{GMM}})$ by first obtaining a consistent estimate of C.

We have the following result; the proof is given in the Appendix.

Proposition 1. *$\hat{\theta}_{\mathrm{GMM}}$ is consistent and has the same asymptotic variance-covaraince matrix as that of $\hat{\theta}_{\mathrm{CMD}}$.*

Remark 1. This result expands the conclusion of Imbens and Lancaster (1994) by integrating additional general summary data in the presence of nuisance parameters. Owing to their equivalence, we mainly consider $\hat{\theta}_{\mathrm{GMM}}$ in the following discussion.

Remark 2. Although $\hat{\theta}_{\mathrm{CMD}}$ and $\hat{\theta}_{\mathrm{GMM}}$ are asymptotically equivalent, $\hat{\theta}_{\mathrm{CMD}}$ can only be used for summary data derived from a misspecified external model that is not consistent with the true underlying model $f(Y \mid X; \theta)$. In particular, $E(uu^{\mathrm{T}})$ has to be positive definite for $\hat{\theta}_{\mathrm{CMD}}$. If (2.1) is the score equation derived from $f(Y \mid X; \theta)$, we have $u(X; \theta_0, \alpha_0, \beta_0) \equiv 0$ when $\theta_0 = (\alpha_0, \beta_0)$, because $\int_Y \partial \log f(Y \mid X; \theta)/\partial\theta f(Y \mid X; \theta) dY = 0$, leading to $E(uu^{\mathrm{T}}) = 0$. On the other hand, $\hat{\theta}_{\mathrm{GMM}}$ has no such restriction, and is equivalent to the meta-analysis estimate when the summary data are derived from the true underlying model (see the proof of Proposition 1).

Thus far, we have described the GMM and CMD estimates in the conditional likelihood setting in which the conditional distribution $f(Y \mid X; \theta)$ is specified. Similar arguments can be used to define them under a more robust quasi-likelihood framework. For example, we can consider the generalized linear model (GLM), where we specify models for the conditional mean and variance of Y given X as

$$E(Y \mid X; \theta) = \mu(X; \theta) \text{ and } \mathrm{var}(Y \mid X; \theta) = \nu(X; \theta), \text{ respectively.}$$

Under the GLM setting, we can let

$$\psi_1(Y, X; \theta) = \frac{\partial \mu(X; \theta)}{\partial \theta} \frac{1}{\nu(X; \theta)} \{Y - \mu(X; \theta)\}.$$

Assume the summary data are derived from an estimating equation based on a different GLM model (with the same link function), with $\mu^{(E)}(X; \alpha, \beta)$ and

$\nu^{(E)}(X;\alpha,\beta)$ as the conditional mean and variance models, respectively. We can define $\psi_2(X;\theta,\alpha,\beta)$ as

$$\psi_2(X;\alpha,\beta)=\frac{\partial\mu^{(E)}(X;\alpha,\beta)}{\partial(\alpha,\beta)}\frac{1}{\nu^{(E)}(X;\alpha,\beta)}\{\mu^{(E)}(X;\alpha,\beta)-\mu(X;\theta)\}.$$

It can be shown that Proposition 1 remains valid under this GLM setting. In fact, all results presented hereafter apply to both the conditional likelihood and GLM settings, unless stated otherwise.

Denote $\hat{\theta}_{\text{INT}}$ as the estimate of θ derived from the estimating equation $\sum_{i=1}^{n}\psi_1(Y_i,X_i;\theta)=0$ based on the internal study. We call it the internal estimate of θ. We can show that the GMM estimate $\hat{\theta}_{\text{GMM}}$ is at least as efficient as $\hat{\theta}_{\text{INT}}$. Therefore, $\hat{\theta}_{\text{GMM}}$ is at least as efficient as the MLE derived from the internal study if $\psi_1(Y,X;\theta)=\partial\log f(Y\mid X;\theta)/\partial\theta$.

As mentioned in the Introduction, using external summary data does not always lead to an efficiency gain. More specifically, let $\theta=(\theta_1,\theta_2)$, and denote its corresponding GMM and internal estimates as $\hat{\theta}_{\text{GMM}}=(\hat{\theta}_{\text{GMM},1},\hat{\theta}_{\text{GMM},2})$ and $\hat{\theta}_{\text{INT}}=(\hat{\theta}_{\text{INT},1},\hat{\theta}_{\text{INT},2})$, respectively. Similarly, we denote $\breve{\theta}=(\breve{\theta}_1,\breve{\theta}_2)$ as the intermediate estimate of θ based on (2.3). For certain external summary data, the variance of $\hat{\theta}_{\text{GMM},1}$ is less than that of $\hat{\theta}_{\text{INT},1}$, but $\hat{\theta}_{\text{GMM},2}$ and $\hat{\theta}_{\text{INT},2}$ have the same level of variation. The following result provides conditions under which using summary data can lead to a GMM estimate that is more efficient than the internal estimate.

Theorem 1. *If $\breve{\theta}_2$ and $\breve{\beta}$ are asymptotically independent, and $\breve{\theta}_1$ and $\breve{\beta}$ are asymptotically correlated, then $\hat{\theta}_{\text{GMM},1}$ is more efficient than $\hat{\theta}_{\text{INT},1}$, but $\hat{\theta}_{\text{GMM},2}$ and $\hat{\theta}_{\text{INT},2}$ share the same level of efficiency.*

We can use Theorem 1 to check the correlation between the intermediate estimates $\breve{\beta}$ and $\breve{\theta}$ to determine whether using summary data can lead to a more efficient GMM estimate. We derive another criterion. Note that we can obtain a consistent estimate of (α_0,β_0) using the same estimating equation (2.2) fitted from the internal study, that is,

$$\sum_{i=1}^{n}h(Y_i,X_i;\alpha,\beta)=0.$$

We denote this estimate as $(\hat{\alpha}_{\text{INT}},\hat{\beta}_{\text{INT}})$, and call it the internal estimate of (α,β).

We obtain the following result from Theorem 1.

Corollary 1. *If $\hat{\theta}_{\text{INT},2}$ and $\hat{\beta}_{\text{INT}}$ are asymptotically independent, and $\hat{\theta}_{\text{INT},1}$ and $\hat{\beta}_{\text{INT}}$ are asymptotically correlated, then $\hat{\theta}_{\text{GMM},1}$ is more efficient than $\hat{\theta}_{\text{INT},1}$, but $\hat{\theta}_{\text{GMM},2}$ and $\hat{\theta}_{\text{INT},2}$ have the same level of efficiency.*

Both Theorem 1 and Corollary 1 provide a necessary and sufficient condition under which the GMM estimate is more efficient than the internal estimate.

Although the summary data we have considered thus far include estimates of the parameters based on estimating equation (2.1), these conclusions can be expanded to summary data derived from estimating equations that do not involve the outcome Y. For example, we can derive summary data from the following estimating equation:

$$\sum_{i=1}^{N} W(X_i^{(E)}) - \beta = 0,$$

where $W(\cdot)$ is a known function of X, and the estimate of β is given by N^{-1} $\sum_{i=1}^{N} W(X_i^{(E)})$. Using Theorem 1, we can easily show that this external information on the moment of $W(X)$ does not improve the estimate of θ. This is expected, because θ is related to the conditional distribution of Y given X, whereas β contains only information about the marginal distribution of X.

Here, we provide examples to show how to use the above results.

Example 1. The internal study assumes the following underlying GLM:

$$l\{E(Y \mid X_1, X_2)\} = X_1^{\mathrm{T}}\theta_1 + X_2^{\mathrm{T}}\theta_2.$$

External summary data are derived from a nested working model given by

$$l\{E(Y \mid X_1)\} = X_1^{\mathrm{T}}\beta,$$

where $l(\cdot)$ is a known canonical link function.

Based on the result in Dai et al. (2012), we know that $\hat{\theta}_{\mathrm{INT},2}$ and $\hat{\beta}_{\mathrm{INT}}$ are asymptotically independent. Therefore, according to Corollary 1, we have the following conclusion.

Corollary 2. *Under the setting of Example* 1, $\hat{\theta}_{\mathrm{GMM},2}$ *has the same efficiency level as that of* $\hat{\theta}_{\mathrm{INT},2}$.

This result indicates that estimates from a nested external model do not improve the GMM estimates of other parameters in the full model. A direct consequence is that external summary data on main effects do not improve the estimate of the interaction effect.

Example 2. The internal study assumes the following underlying linear model:

$$Y = \theta_0 + X_1^{\mathrm{T}}\theta_1 + X_2^{\mathrm{T}}\theta_2 + \varepsilon.$$

External summary data are derived from an unnested working model given by

$$Y = \alpha + S^{\mathrm{T}}(X_1)\beta + \varepsilon'.$$

With additional assumptions on the distribution of X, we have the following result.

Corollary 3. *Under the setting of Example* 2*, if* X_1 *and* X_2 *are independent or are jointly normal,* $\hat{\theta}_{\mathrm{GMM},2}$ *has the same efficiency level as that of* $\hat{\theta}_{\mathrm{INT},2}$*.*

This conclusion can be expanded to the logistic regression model if $\mathrm{var}(Y \mid X)$ remains relatively constant over X. Because $\mathrm{var}(Y \mid X) = P(Y = 1 \mid X)\{1 - P(Y = 1 \mid X)\}$, it is often quite stable over X. In fact, the results of our extensive numerical simulations, presented later, demonstrate that Corollary 3 is (numerically) proper for the logistic regression model.

When X_1 and $S(X_1)$ are scalars, we can directly compare the contribution from the summary statistic derived from the external model $Y = \alpha + S(X_1)\beta + \varepsilon'$, with different choices of $S(X_1)$. We have the following result.

Corollary 4. *Under the setting of Example* 2 *with* X_1 *and* $S(X_1)$ *being scalars, the efficency of* $\hat{\theta}_{\mathrm{GMM},1}$ *increases as the correlation between* X_1 *and* $S(X_1)$ *increases.*

2.3. Integrating summary data from a different population

Here, we consider the setting in which the external study is conducted on a population with a distribution of X that differs from that in the internal study population.

We assume that the conditional distribution $f(Y \mid X; \theta)$ remains the same in the two populations, but that the marginal distribution of X differs. Let $f(X)$ and $f^*(X)$ be distributions of X in the internal and external populations, respectively. In addition to the summary data, we further assume that we have a set of random samples from $f^*(X)$, denoted as $\{X_i^*, i = 1, \ldots, n^*\}$.

This set of reference samples is necessary in order to characterize $f^*(X)$. Here, we focus on the setting in which the reference set is independent of that from the external study. In the Appendix, we discuss the setting in which the reference set is taken from the external study.

We change the stochastic constraint (2.2) to

$$\int_X u(X; \theta_0, \alpha_0, \beta_0) f^*(X) dX = 0.$$

The CMD estimate needs to be modified as follows. From the internal study and the set of reference samples, we obtain intermediate estimates $\breve{\theta}^*, \breve{\alpha}^*$, and $\breve{\beta}^*$ based on the following estimating equations:

$$\sum_{i=1}^{n} \psi_1(Y_i, X_i; \theta) = 0, \quad \sum_{i=1}^{n^*} \psi_2(X_i^*; \alpha, \beta) = 0.$$

We obtain the CMD estimate of (θ, α, β) by minimizing the following quadratic form:

$$(\hat{\theta}^*_{\mathrm{CMD}}, \hat{\alpha}^*_{\mathrm{CMD}}, \hat{\beta}^*_{\mathrm{CMD}}) = \underset{(\theta,\alpha,\beta)}{\operatorname{argmin}} \begin{pmatrix} \breve{\theta}^* - \theta \\ \breve{\alpha}^* - \alpha \\ \breve{\beta}^* - \beta \\ \tilde{\beta} - \beta \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} H^{*-1} & 0 \\ 0 & \rho\Sigma^{-1} \end{pmatrix} \begin{pmatrix} \breve{\theta}^* - \theta \\ \breve{\alpha}^* - \alpha \\ \breve{\beta}^* - \beta \\ \tilde{\beta} - \beta \end{pmatrix},$$

where H^* is the variance of $n^{1/2}(\breve{\theta}^*, \breve{\alpha}^*, \breve{\beta}^*)$ and is given in the Appendix.

Similarly, we can modify the GMM by directly estimating θ and α, as follows:

$$(\hat{\theta}^*_{\mathrm{GMM}}, \hat{\alpha}^*_{\mathrm{GMM}}) = \underset{(\theta,\alpha)}{\operatorname{argmin}} \begin{pmatrix} n^{-1}\sum_{i=1}^{n} \psi_1(Y_i, X_i; \theta) \\ n^{*-1}\sum_{i=1}^{n^*} \psi_2(X^*_i; \theta, \alpha, \tilde{\beta}) \end{pmatrix}^{\mathrm{T}} C^* \begin{pmatrix} n^{-1}\sum_{i=1}^{n} \psi_1(Y_i, X_i; \theta) \\ n^{*-1}\sum_{i=1}^{n^*} \psi_2(X^*_i; \theta, \alpha, \tilde{\beta}) \end{pmatrix},$$

where $C^* = (n^{1/2}\mathrm{var}^*\{n^{-1}\sum_{i=1}^{n} \psi_1^{\mathrm{T}}(Y_i, X_i; \theta_0), n^{*-1}\sum_{i=1}^{n^*} \psi_2^{\mathrm{T}}(X^*_i; \theta_0, \alpha_0, \tilde{\beta})\})^{-1}$. Again, a standard two-step estimation procedure can be used to get $(\hat{\theta}^*_{\mathrm{GMM}}, \hat{\alpha}^*_{\mathrm{GMM}})$ by first obtaining a consistent estimate of C^*.

Corresponding to Proposition 1 and Theorem 1, we have the following two results.

Proposition 2. *$\hat{\theta}^*_{\mathrm{CMD}}$ and $\hat{\theta}^*_{\mathrm{GMM}}$ have the same asymptotic variance.*

Theorem 2. *If $\breve{\theta}^*_2$ and $\breve{\beta}^*$ are asymptotically independent, and $\breve{\theta}^*_1$ and $\breve{\beta}^*$ are asymptotically correlated, then $\hat{\theta}^*_{\mathrm{GMM},1}$ is more efficient than $\hat{\theta}_{\mathrm{INT},1}$, but $\hat{\theta}^*_{\mathrm{GMM},2}$ and $\hat{\theta}_{\mathrm{INT},2}$ share the same level of efficiency.*

Example 3. Assume the same setting for the linear regression model as that in Example 2, but here, X_1 and X_2 are independent in both populations.

We have the following result.

Corollary 5. *Under the setting of Example 3, $\hat{\theta}^*_{\mathrm{GMM},2}$ has the same efficiency level as that of $\hat{\theta}_{\mathrm{INT},2}$.*

The results of simulation studies, described later, show that this conclusion still holds reasonably well under logistic regression models.

2.4. Connection with the generalized integration method

Zhang et al. (2020) recently proposed an empirical likelihood approach called the generalized integration method (GIM) for synthesizing individual-level and summary data. They considered a joint likelihood approach by treating data from both sources as observed random variables. Denote $P = (p_i \overset{\mathrm{def}}{=} dF(X_i) : i = 1, \ldots, n)$ as the empirical distribution of X supported by the internal data. The log likelihood of the internal data can be expressed as $\sum_{i=1}^{n} \log p_i + \sum_{i=1}^{n} \log f(Y_i \mid X_i; \theta)$. The summary data $\tilde{\beta}$ follows an asymptotic normal distribution, with its log likelihood function being $-N(\tilde{\beta} - \beta)^{\mathrm{T}}\Sigma^{-1}(\tilde{\beta} - \beta)/2$. Because Σ is unknown, Zhang et al. (2020) propose estimating $\mu = (\theta^{\mathrm{T}}, \alpha^{\mathrm{T}}, \beta^{\mathrm{T}})^{\mathrm{T}}$ by solving the following

optimization problem over (P, μ):

$$(\hat{P}, \hat{\mu}) = \underset{(P,\mu)}{\operatorname{argmax}} \sum_{i=1}^{n} \log p_i + \sum_{i=1}^{n} \log f(Y_i \mid X_i; \theta) - \frac{N}{2}(\tilde{\beta} - \beta)^{\mathrm{T}} V^{-1} (\tilde{\beta} - \beta), \quad (2.5)$$

subject to

$$\sum_{i=1} p_i = 1, \quad \sum_{i=1} p_i u(X_i; \theta, \alpha, \beta) = 0, \quad (2.6)$$

with $p_i \geq 0$, and V being any given positive-definite matrix with its dimension equal to that of β. Note that constraint equations (2.6) are the empirical distribution analogy of (2.2). Zhang et al. (2020) show that the estimate of θ is always consistent for any given V, and that Σ is the optimal choice of V under this empirical likelihood framework, leading to the most efficient estimate of θ. A two-step procedure can be used to obtain this most efficient estimate of θ. At the initial step, set V as the identity matrix in (2.5) to find the solution for the optimalization problem. Then, use the estimate from the initial step to obtain $\hat{\Sigma}$, a consistent estimate of Σ, and solve the optimization problem again with $V = \hat{\Sigma}$; for further details, see Zhang et al. (2020). We denote this estimate as $(\hat{\theta}_{\mathrm{EL}}, \hat{\alpha}_{\mathrm{EL}}, \hat{\beta}_{\mathrm{EL}})$. When the distributions of X are different between the two study populations, Zhang et al. (2020) modify their GIM estimate by assuming that a set of reference samples of X from the external study population are available. We denote this version of the GIM estimate as $(\hat{\theta}^*_{\mathrm{EL}}, \hat{\alpha}^*_{\mathrm{EL}}, \hat{\beta}^*_{\mathrm{EL}})$. Because the GIM adopts a likelihood approach, it requires that we specify $f(Y_i \mid X_i; \theta)$. The following result shows that the GMM and GIM are asymptotically equivalent under the conditional likelihood setting.

Theorem 3. *When two study populations have the same distribution of X, $\hat{\theta}_{\mathrm{GMM}}$ and $\hat{\theta}_{\mathrm{EL}}$ are asymptotically equivalent. When the distributions of X are different between the two study populations, $\hat{\theta}^*_{\mathrm{GMM}}$ and $\hat{\theta}^*_{\mathrm{EL}}$ are asymptotically equivalent.*

3. Simulation Study

3.1. Same study population

We first consider the setting in which the internal and external studies are carried out on the same source population. We consider an outcome Y to be either continuous or binary, and assume there are two covariates $X = (X_1, X_2)$. The true underlying model (the internal model) for the continuous outcome is given by

$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \varepsilon,$$

where $(\theta_0, \theta_1, \theta_2) = (-0.5, -0.1, 0.2)$ and ε follows the normal distribution $N(0, 2)$. For the binary outcome, the true model is specified as

Table 1. Simulation results under linear regression models, with the internal and external studies conducted in the same population.

Methods	Bias				SE.emp			
	$\hat{\theta}_{\text{INT}}$	$\hat{\theta}_{\text{GMM}}$			$\hat{\theta}_{\text{INT}}$	$\hat{\theta}_{\text{GMM}}$		
External Model		(a)	(b)	(c)		(a)	(b)	(c)
Independent normal, with $(\sigma^2, r) = (2, 0)$								
θ_1	0.0004	-0.0004	-0.0005	-0.0004	0.1277	0.0249	0.0270	0.0315
θ_2	0.0006	0.0014	0.0014	0.0013	0.1285	0.1290	0.1290	0.1291
Joint normal, with $(\sigma^2, r) = (2, 0.6)$								
θ_1	0.0002	-0.0012	-0.0011	-0.0009	0.1599	0.1001	0.1007	0.1020
θ_2	0.0006	0.0014	0.0013	0.0013	0.1607	0.1613	0.1613	0.1614
Independent uniform $U(-c, c)$, with $c = 2$								
θ_1	0.0013	0.0000	-0.0002	-0.0003	0.1107	0.0224	0.0236	0.0263
θ_2	-0.0012	-0.0004	-0.0004	-0.0004	0.1110	0.1114	0.1114	0.1114

$\hat{\theta}_{\text{INT}}$: the internal data-based MLE; $\hat{\theta}_{\text{GMM}}$: the GMM assuming that the internal and external data share the same covariate distribution; SE.emp: the empirical standard error of the estimate; External Model: nested model (a), cubic root model (b), and threshold model (c).

$$P(Y = 1 \mid X) = \frac{\exp(\theta_0 + \theta_1 X_1 + \theta_2 X_2)}{1 + \exp(\theta_0 + \theta_1 X_1 + \theta_2 X_2)},$$

with $(\theta_0, \theta_1, \theta_2) = (-0.5, -0.1, 0.2)$. For both types of outcomes, we consider the distribution of (X_1, X_2) to be either joint normal $N(0, \Sigma)$, with $\Sigma_{11} = \Sigma_{22} = \sigma^2$ and $\Sigma_{12} = \sigma^2 r$, or drawn independently from a uniform distribution $U(-c, c)$. We consider $\sigma^2 = 2$ or 100, $r = 0$ or 0.6, and $c = 2$ or 20 in the experiments. We fix the internal study sample size at $n = 250$ and the external study sample size at $N = 10{,}000$. We choose N to be much larger than n for the purpose of illustration. Other external sample sizes yield similar conclusions (results not shown). We further assume that each external working model uses the same link function (either the identity or the logit link) as the internal model, and adopts one of the following three model specifications: (a) the nested model with $l\{E(Y \mid X)\} = \alpha + X_1\beta$; (b) the cubic root model with $l\{E(Y \mid X)\} = \alpha + X_1^{1/3}\beta$; and (c) the threshold model with $l\{E(Y \mid X)\} = \alpha + I(X_1 > 0)\beta$. We generate 5,000 pairs of internal data and summary data under each scenario, and evaluate the performance of the considered methods.

The Simulation results presented in Table 1 and in Tables S1 of the Supplementary Material verify the performance of the GMM under the settings of Examples 1 and 2, with a continuous outcome. In both tables, we present the results of the GMM incorporating summary data derived from each of the three considered external models under different distributions of X. First, Table S1 of the Supplementary Material shows that the GMM estimate is consistent and its estimated standard error matches well with its empirical version. The GMM-derived confidence interval also has the proper coverage probability. Second, by comparing the empirical standard errors presented in Table 1, it is evident that

Table 2. Simulation results under logistic regression models, with the internal and external studies conducted in the same population.

Methods	Bias				SE.emp			
	$\hat{\theta}_{INT}$	$\hat{\theta}_{GMM}$			$\hat{\theta}_{INT}$	$\hat{\theta}_{GMM}$		
External Model		(a)	(b)	(c)		(a)	(b)	(c)
Independent normal, with $(\sigma^2, r) = (100, 0)$								
θ_1	-0.0024	-0.0008	-0.0009	-0.0009	0.1339	0.0261	0.0281	0.0324
θ_2	0.0059	0.0059	0.0060	0.0060	0.1367	0.1367	0.1368	0.1369
Joint normal, with $(\sigma^2, r) = (100, 0.6)$								
θ_1	-0.0026	-0.0034	-0.0034	-0.0034	0.1667	0.1043	0.1046	0.1057
θ_2	0.0057	0.0057	0.0058	0.0058	0.1697	0.1697	0.1698	0.1699
Independent uniform $U(-c, c)$, with $c = 20$								
θ_1	-0.0004	-0.0005	-0.0005	-0.0005	0.1155	0.0235	0.0245	0.0271
θ_2	0.0031	0.0031	0.0031	0.0030	0.1180	0.1180	0.1180	0.1180

$\hat{\theta}_{INT}$: the internal data-based MLE; $\hat{\theta}_{GMM}$: the GMM assuming that the internal and external data share the same covariate distribution; SE.emp: the empirical standard error of the estimate; External Model: nested model (a), cubic root model (b), and threshold model (c).

$\hat{\theta}_{\mathrm{GMM},1}$, the GMM estimate of θ_1, is more efficient than $\hat{\theta}_{\mathrm{INT},1}$, the estimate based on the internal study. However, $\hat{\theta}_{\mathrm{GMM},2}$ has the same level of efficiency as that of $\hat{\theta}_{\mathrm{INT},2}$. These observations are consistent with the conclusions from Corollaries 2 and 3. Third, a comparison of $\hat{\theta}_{\mathrm{GMM},1}$ using summary data from the three external models shows that using summary data from the nested model is more efficient than using summary data from the cubic root or threshold models. Furthermore, the GMM estimate that incorporates summary data from the cubic root model is more efficient than the one with the threshold model (Table 1). This is expected, given Corollary 4.

Table 2 and Table S2 of the Supplementary Material summarize simulation results under the logistic regression model, yielding similar conclusions to those under the linear regression model. For example, as predicted by Corollary 2, the GMM estimate of θ_2 using summary data from the nested model has the same level of efficiency as that based on the internal study (Table 2). To evaluate whether the conclusions from Corollaries 3 and 4, which are proved under the linear regression model, still hold numerically under the logistic regression model, we choose distributions of X with large variations to ensure that $\mathrm{var}(Y \mid X)$ has a relatively wide range. From Table 2, it appears that the conclusions from Corollaries 3 and 4 hold reasonably well under the logistic regression model, even when the range of $\mathrm{var}(Y \mid X)$ is large.

3.2. Different study populations

Here, we consider the setting in which the internal and external studies are performed on two different populations. Data sets from each population are generated using a similar procedure to that described in Section 3.1, with different

Table 3. Simulation results under linear regression models, with the internal and external studies conducted in two different populations.

Methods	Bias				SE.emp			
	$\hat{\theta}_{\text{INT}}$	$\hat{\theta}^*_{\text{GMM}}$			$\hat{\theta}_{\text{INT}}$	$\hat{\theta}^*_{\text{GMM}}$		
External Model		(a)	(b)	(c)		(a)	(b)	(c)
Independent normal, with $(\sigma^2, r) = (2, 0)$ vs. (1,0)								
θ_1	0.0012	0.0004	0.0003	0.0004	0.0909	0.0246	0.0266	0.0306
θ_2	-0.0014	-0.0015	-0.0015	-0.0015	0.0903	0.0903	0.0904	0.0904
Independent normal, with $(\sigma^2, r) = (100, 0)$ vs. (50,0)								
θ_1	0.0002	0.0001	0.0001	0.0001	0.0129	0.0096	0.0100	0.0106
θ_2	-0.0002	-0.0002	-0.0002	-0.0012	0.0128	0.0128	0.0128	0.0128
Independent uniform $U(-c, c)$, with $c = 2$ vs. 1								
θ_1	0.0001	0.0004	0.0004	0.0003	0.1104	0.0363	0.0377	0.0412
θ_2	0.0008	0.0007	0.0008	0.0008	0.1097	0.1098	0.1098	0.1098
Independent uniform $U(-c, c)$, with $c = 20$ vs. 10								
θ_1	0.0000	-0.0001	0.0000	0.0000	0.0110	0.0088	0.0090	0.0094
θ_2	0.0001	0.0001	0.0001	0.0001	0.0110	0.0110	0.0110	0.0110

$\hat{\theta}_{\text{INT}}$: the internal data-based MLE; $\hat{\theta}^*_{\text{GMM}}$: the GMM using a reference set of 250 samples collected from the external population; SE.emp: the empirical standard error of the estimate; External Model: nested model (a), cubic root model (b), and threshold model (c).

distributions of X chosen for the two populations. In all simulations, the sample size of the reference set is fixed at 250.

We focus on verifying Corollary 5 by considering distributions of covariates, with X_1 and X_2 independent in both populations. When X_1 and X_2 are normally distributed, we set $\sigma^2 = 2$ (or 100) and 1 (or 50) in the internal and external study populations, respectively. When X_1 and X_2 follow a uniform distribution, we set $c = 2$ (or 20) and 1 (or 10), respectively, in the two populations. Table 3 and Tables S3–S4 of the Supplementary Material present simulation results under the linear regression model. First, Table S3 shows that the GMM estimate assuming the same study population is not consistent, and its derived confidence interval does not have the correct coverage probability. On the other hand, Table S4 shows that $\hat{\theta}^*_{\text{GMM}}$, that is, the GMM estimate leveraging reference samples from the external study population, has the desired statistical properties. Second, using summary data from each of the three considered external models shows that $\hat{\theta}^*_{\text{GMM},1}$ is more efficient than $\hat{\theta}_{\text{INT},1}$. However, using the summary data does not lead to a more efficient GMM estimate of θ_2, because $\hat{\theta}^*_{\text{GMM},2}$ and $\hat{\theta}_{\text{INT},2}$ have the same level of empirical standard error (Table 3). Those observations are consistent with Corollary 5.

Table 4 and Tables S5–S6 of the Supplementary Material summarize simulation results under a logistic regression model. By comparing the empirical standard errors, it appears that $\hat{\theta}^*_{\text{GMM},2}$ has almost the same level of efficiency as $\hat{\theta}_{\text{INT},2}$, with the largest percentage difference being around 3%, which occurs

Table 4. Simulation results under logistic regression models, with the internal and external studies conducted in two different populations.

Methods	Bias				SE.emp			
	$\hat{\theta}_{INT}$	$\hat{\theta}^*_{GMM}$			$\hat{\theta}_{INT}$	$\hat{\theta}^*_{GMM}$		
External Model		(a)	(b)	(c)		(a)	(b)	(c)
Independent normal, with $(\sigma^2, r) = (2, 0)$ vs. (1,0)								
θ_1	-0.0018	0.0001	-0.0001	-0.0003	0.0966	0.0252	0.0273	0.0316
θ_2	0.0029	0.0022	0.0022	0.0022	0.0983	0.0979	0.0980	0.0981
Independent normal, with $(\sigma^2, r) = (100, 0)$ vs. (50,0)								
θ_1	-0.0027	-0.0012	-0.0014	-0.0016	0.0210	0.0143	0.0151	0.0162
θ_2	0.0049	0.0041	0.0042	0.0043	0.0281	0.0279	0.0280	0.0280
Independent uniform $U(-c, c)$, with $c = 2$ vs. 1								
θ_1	-0.0027	-0.0004	-0.0003	-0.0004	0.1157	0.0370	0.0392	0.0430
θ_2	0.0032	0.0025	0.0024	0.0025	0.1181	0.1176	0.1176	0.1177
Independent uniform $U(-c, c)$, with $c = 20$ vs. 10								
θ_1	-0.0028	-0.0017	-0.0017	-0.0019	0.0204	0.0144	0.0149	0.0157
θ_2	0.0056	0.0046	0.0047	0.0048	0.0265	0.0257	0.0257	0.0259

$\hat{\theta}_{INT}$: the internal data-based MLE; $\hat{\theta}^*_{GMM}$: the GMM using a reference set of 250 samples collected from the external population; SE.emp: the empirical standard error of the estimate; External Model: nested model (a), cubic root model (b), and threshold model (c).

when the range of $\text{var}(Y \mid X)$ is relatively large (Table 4).

4. Discussion

We have shown that the GMM can be used as a flexible procedure to effectively integrate external summary data with individual-level data. We provide a necessary and sufficient condition under which summary data improved the GMM estimate. For the purpose of illustration, we consider only summary data consisting of estimates derived from one external model. The same procedure can be applied to summary data from different external models.

When the distribution of X differs between the internal and the external study populations, we consider GMM procedures in which we assume that we have a set of samples chosen randomly from the distribution of X in the external study population. This set of reference samples is needed to estimate the empirical distribution of X. Ignoring the discrepancy in the distribution of X between the two study populations could lead to a biased estimate of θ. Recent works have propsed several strategies for dealing with this distribution shift problem, without relying on a set of reference samples (Chen et al. (2021); Zhai and Han (2022); Taylor, Choi and Han (2023)). However, these methods assume that the external summary data exhibit negligible variability. Further research is needed to develop procedures that are more robust when incorporating external summary data.

Supplementary Material

All technical details and additional numeric results are relegated to the online Supplementary Material.

Acknowledgments

This study used the computational resources of the NIH Biowulf cluster (`https://hpc.nih.gov/`). The research of Dr. Lu Deng was partially supported by the National Natural Science Foundation of China, grant #12101331. The authors also thank the associate editor and referees for their helpful and insightful comments and suggestions.

References

Chatterjee, N., Chen, Y.-H., Maas, P. and Carroll, R. J. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association* **111**, 107–117.

Chen, Z., Ning, J., Shen, Y. and Qin, J. (2021). Combining primary cohort data with external aggregate information without assuming comparability. *Biometrics* **77**, 1024–1036.

Cheng, W., Taylor, J. M., Gu, T., Tomlins, S. A. and Mukherjee, B. (2019). Informing a risk prediction model for binary outcomes with external coefficient information. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **68**, 121–139.

Cheng, W., Taylor, J. M., Vokonas, P. S., Park, S. K. and Mukherjee, B. (2018). Improving estimation and prediction in linear regression incorporating external information from an established reduced model. *Statistics in Medicine* **37**, 1515–1530.

Dai, J. Y., Kooperberg, C., Leblanc, M. and Prentice, R. L. (2012). Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. *Biometrika* **99**, 929–944.

Han, P. and Lawless, J. F. (2016). Discussion of "constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources". *Journal of the American Statistical Association* **111**, 118–121.

Han, P. and Lawless, J. F. (2019). Empirical likelihood estimation using auxiliary summary information with different covariate distributions. *Statistica Sinica* **29**, 1321–1342.

Huang, C.-Y. and Qin, J. (2020). A unified approach for synthesizing population-level covariate effect information in semiparametric estimation with survival data. *Statistics in Medicine* **39**, 1573–1590.

Imbens, G. W. and Lancaster, T. (1994). Combining micro and macro data in microeconometric models. *Review of Economic Studies* **61**, 655–680.

Kundu, P., Tang, R. and Chatterjee, N. (2019). Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika* **106**, 567–585.

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V. and Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition* **45**, 521–530.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics* **4**, 2111–2245.

Qin, J. (2000). Combining parametric and empirical likelihoods. *Biometrika* **87**, 484–490.

Qin, J., Zhang, H., Li, P., Albanes, D. and Yu, K. (2015). Using covariate-specific disease prevalence information to increase the power of case-control studies. *Biometrika* **102**, 169–180.

Sugiyama, M., Krauledat, M. and Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research* **8**, 985–1005.

Taylor, J. M. G., Choi, K. and Han, P. (2023). Data integration: Exploiting ratios of parameter estimates from a reduced external model. *Biometrika* **110**, 119–134.

White, H. L. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.

Zhai, Y. and Han, P. (2022). Data integration with oracle use of external information from heterogeneous populations. *Journal of Computational and Graphical Statistics* **31**, 1001–1012

Zhang, H., Deng, L., Schiffman, M., Qin, J. and Yu, K. (2020). Generalized integration model for improved statistical inference by leveraging external summary data. *Biometrika* **107**, 689–703.

Zhang, H., Deng, L., Wheeler, W., Qin, J. and Yu, K. (2022). Integrative analysis of multiple case-control studies. *Biometrics* **78**, 1080–1091

Lu Deng

School of Statistics and Data Science, Nankai University, Tianjin 300071, China.

E-mail: denglu014@mail.nnkai.edu.cn

Sheng Fu

Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 20892, USA.

E-mail: fu.sheng@nih.gov

Jing Qin

National Institute of Allergy and Infectious Diseases, National Institute of Allergy and Infectious Diseases, Bethesda, MD 20817, USA.

E-mail: jingqin@niaid.nih.gov

Kai Yu

Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 20892, USA.

E-mail: yuka@mail.nih.gov

(Received July 2022; accepted November 2022)