Constrained MLE for Model Calibration Using Summary-level Information from External Big Data Sources

Chatterjee N, Chen YH, Maas P, Carroll RJ.
J Am Stat Assoc. 2016 Mar

Mengqi Xu

University of Waterloo

May, 2024

Why data integration

An example: want to estimate a model in a clinical trial study with limited observations → borrow information from external data sources (e.g. observational studies, EHR data).

Limitation of a single "internal" study: limited sample size

- We want to borrow information from big "external" data source to improve estimation or prediction on our target internal data.

Challenges in data integration

1. cannot always access individual-level external data (data privacy)
2. underlying populations between internal and external studies may be different
3. various types of variables: continuous, categorical, counts data ...

Our problem setting

Now, consider building **regression models** using individual level information from an internal study while incorporating **summary level information** from an external large data source.

Existing methods

Calibration techniques: (Deville and Sarndal (1992), Wu (2003), Lumley et al. (2011) ...)

- Require access to individual-level data from the external study

"Model-based" MLE: (Scott and Wild (1997) ...)

- Assume that covariate information in the external data can be summarized into discrete strata – challenging and inefficient: external datasets often include combinations of many variables

Notations and Models

- Y: outcome of interest
- X: a set of covariates (available from both internal and external studies)
- Z: another set of covariates (only available from the internal study)
- $F(X, Z)$: c.d.f. of X, Z for the underlying population
- "Full" model $f_\beta(y|x, z)$: model built on the internal data
- "Reduced" model $g_\theta(y|x)$: model built on the external data

Assume $f_\beta(y|x, z)$ is correctly specified, but $g_\theta(y|x)$ need not be. Note: for external data, all the information we could have is only the population parameter value θ^* of θ.

Constrained Equation

Let $U(Y|X;\theta) = \partial log\{g_\theta(Y|X)\}/\partial\theta$. Then

$$E\{U(Y|X,\theta^*)\} = \int U(y|x,\theta^*)pr(y|x)pr(x)dydx = 0,$$

$f_\beta(Y|X,Z)$ correctly specified $\Rightarrow pr(y|x) = \int f_{\beta_0}(y|z,x)pr(z|x)dz$
(β_0: true value of β).
Thus the constraint can be written as

$$\int_{Z,X}\{\int_Y U(Y|X,\theta^*)f_{\beta_0}(Y|X,Z)dY\}dF(X,Z) = 0.$$

Constrained Equation

For simplicity, we rewrite the constrained equation as

$$\int_{Z,X} u_{\beta_0}(X, Z; \theta^*) dF(X, Z) = 0, \tag{1}$$

where

$$u_\beta(X, Z; \theta^*) = \int_Y U(Y|X, \theta^*) f_\beta(Y|X, Z) dY$$

Constrained Semiparametric Maximum Likelihood (SPML)

A simple case: under simple random sampling from the internal study.
The semiparametric likelihood of the **internal study**:

$$L_{\beta,F} = \prod_{i=1}^{N} f_\beta(Y_i|X_i, Z_i)dF(X_i, Z_i).$$

Goal:

- maximize $log\{L_{\beta,F}\}$ w.r.t β and $F(.)$
- with constraint Eq. (1).

Constrained Semiparametric Maximum Likelihood (SPML)

Recall:

- Maximize $L_{\beta,F} = \prod_{i=1}^{N} f_\beta(Y_i|X_i, Z_i)dF(X_i, Z_i)$.
- w/ constraint Eq. (1): $\int_{Z,X} u_{\beta_0}(X, Z; \theta^*)dF(X, Z) = 0$

Lagrange multiplier!

$$\Rightarrow \text{Maximize: } l_\lambda = log(L_{\beta,F}) + \lambda^\top \int u_\beta(X, Z; \theta^*)dF(X, Z), \quad (2)$$

Other Settings: Other Sampling Designs

This constrained SPML method could be generalized to more complex sampling designs for the internal study.

For example, case-control sampling. Suppose Y is binary. Let

- $p_1 = pr(Y = 1) = \int f_\beta(Y = 1|x, z)dF(x, z) = 1 - p_0$: the underlying marginal disease probability in the population
- N_1 and N_0: # of cases and controls sampled

Then the likelihood for the internal case-control study is given by

$$L^{cc}_{\beta,F} = \prod_{i=1}^{N_1+N_0} f_\beta(Y_i|X_i, Z_i)dF(X_i, Z_i) \times p_1^{-N_1} p_0^{-N_0},$$

Consequently our goal is to maximize

$$l^{cc}_\lambda = log(L^{cc}_{\beta,F}) + \lambda^\top \int u_\beta(X, Z; \theta^*)dF(X, Z), \qquad (3)$$

Other settings: Heterogeneity between the internal and external data

Synthetic Constrained Maximum Likelihood (SCML):

1. Assume an external reference sample $(X^\dagger, Z^\dagger)$ is available for unbiased estimation of the covariate distribution for the external population.
2. Use the empirical distribution of the external reference sample $\tilde{F}^\dagger(X, Z)$ to substitute $F(X, Z)$ in the contraint Equation (1).
3. The SCML estimator $(\tilde{\beta}, \tilde{\lambda})$ for (β, λ) can be obtained by solving $\partial l^\dagger_{\beta,\lambda}/\partial\beta = 0$ and $\partial l^\dagger_{\beta,\lambda}/\partial\lambda = 0$ – *completely ignoring* $F(.,\ .)$ *because it factors out from the likelihood of the internal study.*

$$l^\dagger_{\beta,\lambda} = log(L_{\beta,F}) + \lambda^\top \int u_\beta(X, Z; \theta^*) d\tilde{F}^\dagger(X, Z)$$

Summary

The constrained SPML method proposed by Chatterjee et al. (2016) embeds summary data from external big data sources into constraint equations that impose restrictions on the parameter space, thereby enhancing estimation in internal data.

Contributions

- Only require summary information from the external data – the regression model parameters – circumventing the need for individual-level external data. Can be applied to different types of variables instead of a discrete set of strata defined by the external covariates.
- Allow the reduced model $g_\theta(Y|X)$ to be misspecified.
- Accommodates arbitrary types of covariates and regression models: e.g. non-nested models for the internal and external data and complex sampling designs.

Subsequent studies I

- Expand to multiple external data sources: e.g. Kundu et al. (2019) and Zhang et al. (2020) extended Chatterjee et al.'s (2016) work to meta-analysis settings.

Subsequent studies II

- The internal and external study populations are not comparable (or identical): e.g. synthetic data method (Gu et al. (2023)), penalty function to identify differences (Chen et al. (2021)), and sensitivity parameter (Yang and Ding (2020)).

Subsequent studies III

The number of constraint equations or estimating equations increases linearly with the dimension of the summary data. Numeric convergence issues can occur when the dimension of the summary data becomes relatively large.

- High dimensional data integration problems, e.g. Gao and Carroll (2017), Li et al. (2022), and Yu et al. (2023).

References I

Deville JC, Sarndal CE. Calibration Estimators in Survey Sampling. *Journal of the American Statistics Association.* 1992;87:376382.

Wu C. Optimal Calibration Estimators in Survey Sampling. *Biometrika.* 2003;90:937951.

Lumley T, Shaw PA, Dai JY. Connections Between Survey Calibration Estimators and Semiparametric Models for Incomplete Data. *International Statistitical Review.* 2011;79:200220.

Chatterjee N, Chen YH, Maas P, Carroll RJ. Constrained Maximum Likelihood Estimation for Model Calibration Using Summary-level Information from External Big Data Sources. *Journal of the American Statistical Association.* 2016;111(513):107-117.

References II

Kundu P, Tang R, Chatterjee N. Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika.* 2019 Sep;106(3):567-585.

Zhang H, Deng L, Schiffman M, Qin J, Yu K. (2020) Generalized integration model for improved statistical inference by leveraging external summary data. *Biometrika.* 107, 689703.

Taylor JMG, Choi K, Han P. Data integration: exploiting ratios of parameter estimates from a reduced external model. *Biometrika.* 2022 Apr;110(1):119-134.

Gu T, Taylor JMG, Mukherjee B. A synthetic data integration framework to leverage external summary-level information from heterogeneous populations. *Biometrics.* 2023;79(4):3831-3845.

References III

Chen Z, Ning J, Shen Y, Qin J. Combining primary cohort data with external aggregate information without assuming comparability. *Biometrics.* 2021;77(3):1024-1036.

Yang S, Ding P. Combining Multiple Observational Data Sources to Estimate Causal Effects. *Journal of the American Statistical Association.* 2020;115(531):1540-1554.

Gao X, Carroll RJ. Data integration with high dimensionality. *Biometrika.* 2017;104(2):251-272.

Li S, Cai T, Li H. Transfer Learning for High-Dimensional Linear Regression: Prediction, Estimation and Minimax Optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology.* 2022 Feb;84(1):149173

References IV

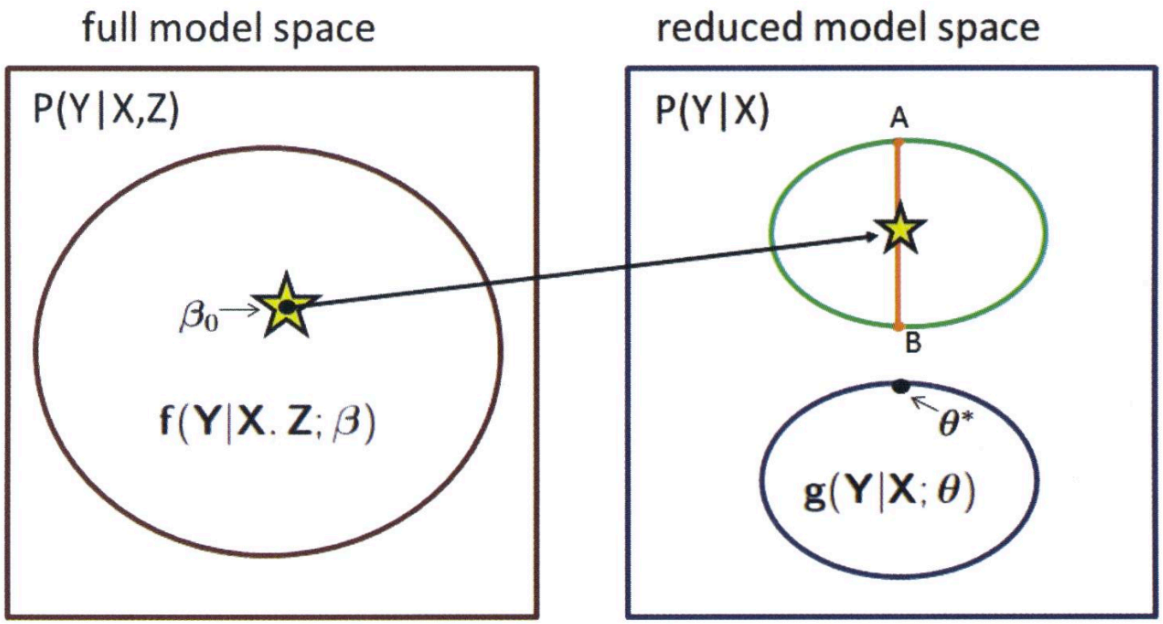
Fu S, Deng L, Zhang H, Qin J, Yu K. Integrative analysis of individual-level data and high-dimensional summary statistics. *Bioinformatics.* 2023;39(4):btad156.

Constrained Semiparametric Maximum Likelihood (SPML)

Using standard empirical likelihood (or profile likelihood) computation steps (see e.g. Qin and Lawless 1994; Scott and Wild 1997), it can be shown that the constrained SPML problem Eq. (2) is equivalent to solving the score equation associated with the "pseudo-loglikelihood":

$$l^*_{\beta,\lambda} = \sum_{i=1}^{N} log(\frac{f_\beta(Y_i|X_i, Z_i)}{1 - \lambda^\top u_\beta(X, Z; \theta^*)}) \tag{4}$$

Geographic Presentation



θ^* minimizes the Kullback-Leibler distance (Huber 1967; White 1982) between the fixed $P_0(Y|X)$ and the model space of $g_\theta(Y|X)$.