Cross Calibration Project Update

Xufei Wang, Yang Chen

Harvard University

Joint work with Meng, X.L., Vinay, K., Herman, M.

November 10, 2015

Overview

1. Explanation of Multiplicative Model
2. log-Normal Model
   - Model Description
   - Shrinkage estimators with known variance
   - Estimators with unknown variance
3. Poisson Model
4. Questions for Discussions

Explanation of Multiplicative Model

Expected Counts of instrument i source j, C_{ij}

- The effective area $A_i(E) = \mathcal{A}_i \rho_i(E)$, where only $\mathcal{A}_i$ is unknown and $\rho_i(E)$ is a fixed function estimated empirically for $E \in [E_1, E_2]$.
- The flux $F_j = \int_{E_1}^{E_2} n(E;\theta_j)dE = N_j \int_{E_1}^{E_2} q(E|\theta_j^*)dE$, where $n(E;\theta_j)$ is the spectrum of source j at energy E. $q(E|\theta_j^*)$ is known.
- The response matrix function $r_{ik}(E)$ is the probability that a photon with energy E comes to channel k through instrument i; known.
- The exposure time for instrument i source j, T_{ij}, is measured precisely.

$$C_{ij} = \sum_{\frac{E_1}{\kappa_i} \le k \le \frac{E_2}{\kappa_i}} T_{ij} \int r_{ik}(E) A_i(E) n(E;\theta_j) dE$$

$$= \mathcal{A}_i N_j \left[T_{ij} \times \int_{E_1}^{E_2} \rho_i(E) q(E|\theta_j^*) \sum_{\frac{E_1}{\kappa_i} \le k \le \frac{E_2}{\kappa_i}} r_{ik}(E) dE \right].$$

Notation Explanation

Consistently throughout the presentation, we adopt the following rules.

Upper Case: Quantity to be estimated, i.e. estimand.

Lower Case: Quantity directly obtained/calculated from the data.

Index i: Index for instrument.

Index j: Index for source.

Example:

- C_{ij} is the expected count of source j from instrument i.
- c_{ij} is the observed count of source j from instrument i.

log-Normal Model

log-Normal Model

Noting that $C_{ij} = A_i F_j$ is mathematically equivalent to

$$\log C_{ij} = \log A_i + \log F_j.$$

Define $Y_{ij} = \log C_{ij}$, $B_i = \log A_i$ and $G_j = \log F_j$. By half variance correction, we have

$$\begin{aligned}
y_{ij} &= -\frac{1}{2}\sigma_{ij}^2 + B_i + G_j + e_{ij}, \mathrm{Var}(e_{ij}) = \sigma_{ij}^2, y'_{ij} = y_{ij} + \frac{1}{2}\sigma_{ij}^2 \\
b_i &= -\frac{1}{2}\tau_i^2 + B_i + \quad + \epsilon_i, \mathrm{Var}(\epsilon_i) = \tau_i^2, b'_i = b_i + \frac{1}{2}\tau_i^2 \\
g_j &= -\frac{1}{2}\eta_j^2 + \quad + G_j + \delta_j, \mathrm{Var}(\delta_j) = \eta_j^2, g'_j = g_j + \frac{1}{2}\eta_j^2
\end{aligned}$$

Subsection 2

Shrinkage estimators with known variance

An intuitive example

For an intuitive model, suppose we know all the variances and $\sigma_{ij}^2 = \sigma_i^2$, $\eta_j^2 = 0$, we could get the MLE for B_i is

$$\begin{aligned}
\widehat{B}_i &= \omega_i b_i' + (1-\omega_i)(\bar{y}_i' - \bar{g}_i), i = 1, \dots, N \\
\bar{g}_i &= \sum_{j \in J_i} g_j / M_i, M_i = |J_i| \\
\omega_i &= \tau_i^{-2} / (\tau_i^{-2} + M_i \sigma_i^{-2})
\end{aligned}$$

The results show that $\widehat{B}_i$ is a shrinkage estimator between the observed b_i' and the estimator from the observation, $\bar{y}_{ij}' - \bar{g}_i$.

Shrinkage estimators

For a general model with known variances, we could also estimate B_i and G_j in as a shrinkage estimator.

$$\begin{aligned} \widehat{B}_i &= w_i b_i' + (1 - w_i)(\bar{y}_{i.}' - \bar{G}_i), i = 1, \ldots, N \\ \widehat{G}_j &= v_j g_j' + (1 - v_j)(\bar{y}_{.j}' - \bar{B}_j), j \in J \end{aligned}$$

$\bar{B}_i, \bar{G}_j, \bar{y}_{i.}', \bar{y}_{.j}'$ could be estimated similarly as above. The details could be found in the paper.

Variance for the estimators

We need to consider a very special case to calculate the variance of the estimators. Assume $\sigma_{ij}^2 = \sigma_i^2$,$\tau_i^2 = \tau^2$ and $J_i = \tilde{J}$, the variance are

$$\begin{aligned}
\widehat{\mathrm{Var}}(\widehat{B}_i) &= \frac{1}{M_i \sigma_i^{-2} + \tau^{-2}} + \ldots < \tau^2 \\
\widehat{\mathrm{Var}}(\widehat{G}_j) &= \frac{1}{\sum_{i \in I_j} \sigma_i^{-2} + \eta^{-2}} - \ldots < \eta^2, j \in \tilde{J} \\
\widehat{\mathrm{Var}}(\widehat{G}_j) &= \eta^2, j \notin \tilde{J}
\end{aligned}$$

The results show that with more observations, the variance of the estimands decrease.

Subsection 3

Estimators with unknown variance

Assumptions for observation error

If we have no idea about the variances, we could make some estimations of them. In this case, we make homogenous variance assumptions for σ_{ij}^2. Two major assumptions are

- The variance only depends on instrument, that is $\sigma_{ij}^2 = \sigma_i^2$;
- The impact of instrument and source on the measurement error is additive, that is $\sigma_{ij}^2 = \omega_i^2 + \nu_j^2$.

Shrinkage estimators

If the variance only depends on the instruments, we could estimate B_i and G_j as before. The only difference is that we need to estimate σ_i^2, τ^2 and η^2 from the data. In a special case, let $\tau_i^2 = \tau^2$ and $\eta_j^2 = \eta^2$, then we have

$$
\begin{aligned}
\hat{\sigma}_i{}^2 &= 2\left[\sqrt{1+S_{y,i}^2}-1\right], S_{y,i}^2 = \frac{1}{M_i}\sum_{j\in J_i}(y_{ij}-\widehat{B}_i-\widehat{G}_j)^2 \\
\hat{\tau}^2 &= 2\left[\sqrt{1+S_b^2}-1\right], S_b^2 = \frac{1}{N}\sum_{i=1}^{N}(b_i-\widehat{B}_i)^2 \\
\hat{\eta}^2 &= 2\left[\sqrt{1+S_g^2}-1\right], S_g^2 = \frac{1}{M}\sum_{j=1}^{M}(g_j-\widehat{G}_j)^2
\end{aligned}
$$

By solving the above equations, we could still get shrinkage estimators.

Variance for the estimators

To estimate the variance of the estimators, we consider a special case, that is the non-overlapping observations, which means $I_j \cap I_k = \emptyset$. Then every source is observed by one and only one instrument. We consider the following three cases:
(1) If σ^2, τ^2, η^2 as known, we have

$$\begin{aligned} \mathrm{var}(G_j) &= \left(\sum_{i \in I_j} \frac{\sigma_i^{-2} \tau^{-2}}{\sigma_i^{-2} + \tau^{-2}} + \eta^{-2} \right)^{-1} < \eta^2, |I_j| \geq 1; \\ \mathrm{var}(B_i) &= \left(\sigma_i^{-2} + \tau^{-2}\right)^{-1} + \mathrm{var}(G_j) \left(\frac{\sigma_i^{-2}}{\sigma_i^{-2} + \tau^{-2}} \right)^2 < \tau^2, i \in I_j. \end{aligned}$$

(2) If we only treat τ^2, η^2 as known, we have

$$\begin{aligned} \mathrm{var}^*(G_j) &= \left(\sum_{i \in I_j} \sigma_i^{-2} + \eta^{-2} - \sum_{i \in I_j} \frac{b_i}{a_i} \right)^{-1}; \\ \mathrm{var}^*(B_i) &= \frac{c_i}{a_i} + \mathrm{var}^*(G_j) \frac{\sigma_i^{-12}}{4a_i^2}. \end{aligned}$$

(3) If we treat all the parameters as unknown,

$$\begin{aligned} \mathrm{var}'(B_i) &= \mathrm{var}^*(B_i) + \left(d_{i,1}^2 K_{1,1} + 2 d_{i,1} d_{i,2} K_{1,2} + d_{i,2}^2 K_{2,2} \right) \\ \mathrm{var}'(G_j) &= \mathrm{var}^*(G_j) + \left(e_{j,1}^2 K_{1,1} + 2 e_{i,1} e_{j,2} K_{1,2} + e_{j,2}^2 K_{2,2} \right); \end{aligned}$$

Additive noise model

In another case, we assume $\sigma_{ij}^2 = \omega_i^2 + \nu_j^2$, we could estimate B_i, G_j, τ^2, η^2 as before. The estimator of ω_i^2 and ν_j^2 are could be solved by

$$\begin{aligned} -\frac{1}{2}\sum_{j\in J_i}\left[\frac{1}{\omega_i^2+\nu_j^2}+\frac{1}{4}-\frac{\left(y_{ij}-\widehat{B}_i-\widehat{G}_j\right)^2}{(\omega_i^2+\nu_j^2)^2}\right] &= 0 \\ -\frac{1}{2}\sum_{i\in I_j}\left[\frac{1}{\omega_i^2+\nu_j^2}+\frac{1}{4}-\frac{\left(y_{ij}-\widehat{B}_i-\widehat{G}_j\right)^2}{(\omega_i^2+\nu_j^2)^2}\right] &= 0; \end{aligned}$$

where $y'_{ij} = y_{ij} + 0.5(\omega_i^2 + \nu_j^2)$, $b'_i = b_i + 0.5\tau_i^2$, $g'_j = g_j + 0.5\eta_j^2$, and

$$B_i = \frac{b'_i/\tau_i^2 + \sum_{j\in J_i}(y'_{ij} - G_j)/(\omega_i^2+\nu_j^2)}{1/\tau_i^2 + \sum_{j\in J_i} 1/(\omega_i^2+\nu_j^2)};$$

$$G_j = \frac{g'_j/\eta_j^2 + \sum_{i\in I_j}(y'_{ij} - B_i)/(\omega_i^2+\nu_j^2)}{1/\eta_j^2 + \sum_{i\in I_j} 1/(\omega_i^2+\nu_j^2)}.$$

Poisson Model

Poisson Model

In a Poisson model, we assume c_{ij} follows a Poisson distribution with parameter as C_{ij} and make further assumptions for C_{ij}.

$$\begin{aligned} c_{i,j} &\sim \mathrm{Pois(C_{i,j}), log(C_{i,j}) = B_i + G_j} \\ b_i &= -\frac{1}{2}\tau_i^2 + B_i + \epsilon_i, \mathrm{Var}(\epsilon_i) = \tau_i^2, b_i' = \log(a_i) + \frac{1}{2}\tau_i^2 \\ g_j &= -\frac{1}{2}\eta^2 + G_j + \delta_j, \mathrm{Var}(\delta_j) = \eta_j^2, g_j' = \log(f_j) + \frac{1}{2}\eta_j^2 \end{aligned}$$

The MLE of the model should satisfies the following equations

$$
\begin{aligned}
e^{B_i} \sum_{j \in J_i} e^{G_j} - \frac{b_i - B_i}{\tau_i^2} &= \sum_{j \in J_i} c_{i,j} + \frac{1}{2} \\
e^{G_j} \sum_{i \in I_j} e^{B_i} - \frac{g_j - G_j}{\eta_j^2} &= \sum_{i \in I_j} c_{i,j} + \frac{1}{2}
\end{aligned}
$$

$$\tau_i^2 = 2\left[\sqrt{S_{b,i}^2 + 1} - 1\right] \quad , \quad S_{b,i}^2 = (b_i - B_i)^2$$

$$\eta_j^2 = 2\left[\sqrt{S_{g,j}^2 + 1} - 1\right] \quad , \quad S_{g,j}^2 = (g_j - G_j)^2$$

Questions for Discussions

Questions for Discussions

- log-Normal Model
  - Known vs unknown variance components
  - Additive noise: estimating equations
- Poisson Model
  - Model assumptions
  - Estimating equations
- Model Checking
  - Noise
  - Real data performance