



Estimating source-specific health effects

Thomas Lumley

Hao Liu, Rebecca Hubbard, Tim Larson

EPA NW Center for Particulate Air Pollution and Health

PM Symposium

Outline

- Why is it interesting?
- Why is it hard?
- How well can we do it now?

Interest in sources

Source apportionment is obviously important in understanding emission and exposure relationships. Why is it useful for epidemiology?

- Policy implications: useful to know that source X is responsible for lots of deaths?
- Regulation is applied most easily to sources, not species?
- An attempt to recognize that PM is a complex mixture (and air pollution is even more complex)?
- Collinearity makes species-specific effects harder to estimate than source-specific effects?

Why is it hard?

- We don't directly measure source-specific exposures, so we have 100% missing data.
- Except in atmospheres dominated by particular point sources, much of the day to day variation in exposure is common to all sources.

Imputing exposures

Health effects models rely on imputing source-specific exposure from the receptor models fitted by algorithms such as UNMIX and PMF.

These models are accurate enough for estimating overall features of exposure, but health effects models are more demanding

- Relative risks are related to **differences** in exposures on nearby days. Taking differences reduces the true variation and increases the uncertainty.
- Uncertainty in exposures causes bias in regression coefficients
- The uncertainties in different source-specific exposures may be highly correlated, making the size and direction of bias hard to predict

Identifiability

An additional problem is that the source apportionment models are underdetermined: there are usually multiple solutions that fit the data equally well.

In practice, expert opinion and various heuristics are used to choose a solution. In the absence of either ground truth or a formal prior distribution we cannot quantify the accuracy of these choices.

Models

Y_t Number of health events on day t

X_{st} Mass concentration of species s on day t

G_{kt} True mass concentration of source k on day t

F_{ks} True percentage concentration of species s in source k

Z_t Other variables in the health model for day t (eg temperature)

The underlying health model is

$$\log E[Y_t] = \alpha_t + \beta_1 G_{1t} + \beta_2 G_{2t} + \cdots + \beta_K G_{Kt} + \gamma Z_t$$

Models

As we do not observe G , we need a measurement model

$$X_{st} = \sum_k F_{ks} G_{kt} + \epsilon_{st}$$

or in matrix terms

$$X = F^T G + \epsilon$$

We always constrain $F > 0$, $G > 0$, $\text{var}[\epsilon_{st}] = \sigma_{st}^2$ and we may add other constraints or penalties on F or G . Since

$$F^T A A^{-1} G = F^T G$$

for any matrix A , giving **identical fit** to the data, our ability to identify differences in β s comes **entirely** from the constraints.

The exposure model gives estimates \hat{G}_{kt} and \hat{F}_{ks} .

Basic question

What happens if we just fit

$$\log E[Y_t] = \alpha_t + \beta_1 \hat{G}_{1t} + \beta_2 \hat{G}_{2t} + \cdots + \beta_K \hat{G}_{Kt} + \gamma Z_t,$$

acting as if G were known?

Simulations

- based on two real data sets: metals, ions, and carbon fractions from Spokane, WA and Phoenix, AZ. Only showing Spokane, results are better in Phoenix.
- using PMF fits or random source profiles as the 'truth'. Only showing PMF fits as truth, results worse for random profiles

The Spokane data have 852 days of 14 species: metals by XRF, ammonium, EC, OC.

Basic question

Procedure:

- Choose G and F
- Generate 1000 simulated X and Y
- Estimate 1000 \hat{G}
- Estimate 1000 β

Start the estimation of \hat{G} at the truth to mimic the impact of expert selection.

Results: bias

Bias, directly using imputed source-specific exposure to five sources simulated from a PMF analysis of data from Spokane. Relative risks are per $1\mu\text{g}/\text{m}^{-3}$.

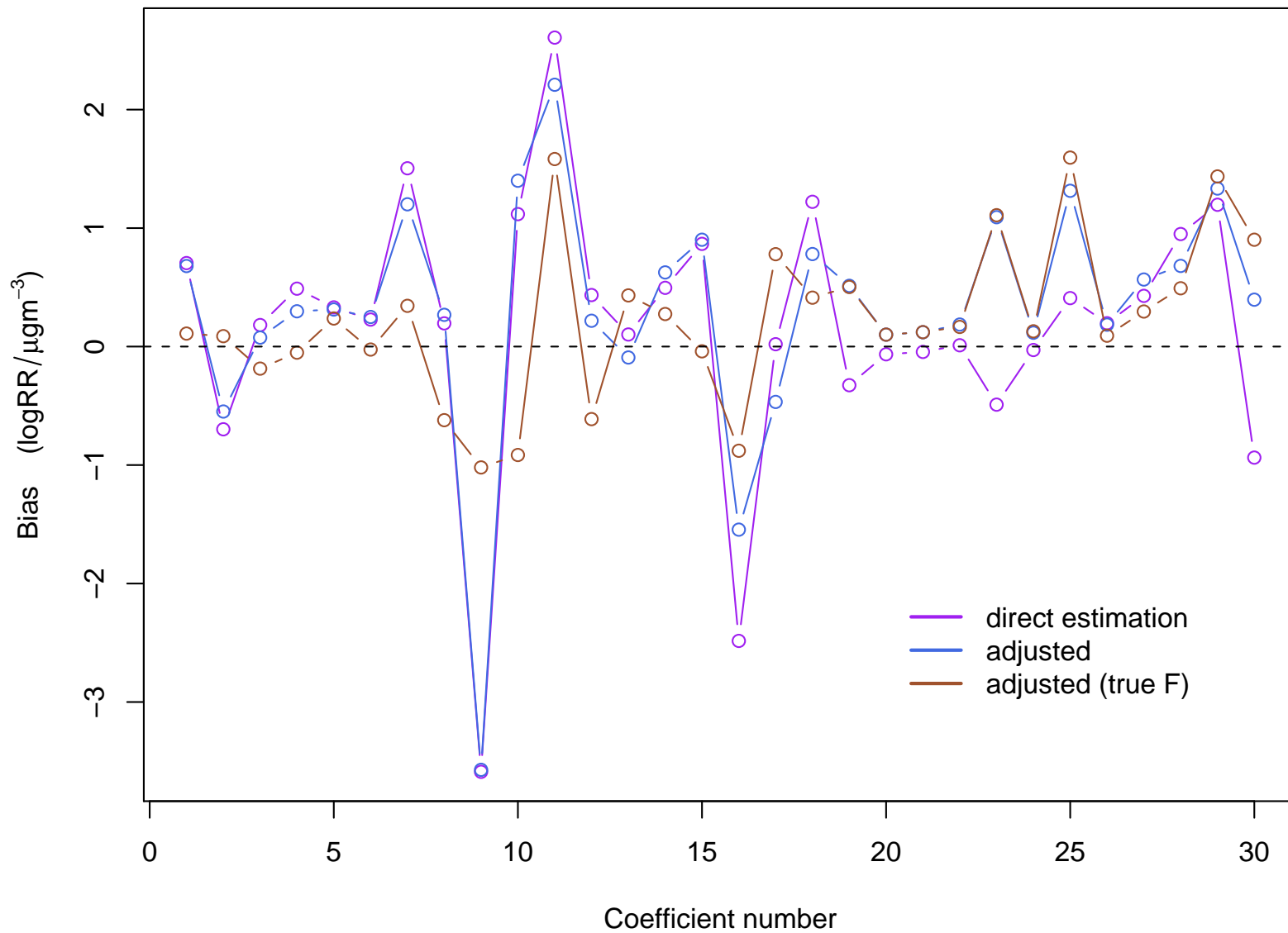
Sources	I	II	III	IV	V
mean concentration	0.70	0.08	0.73	3.25	0.05
$\log RR$	1	2	3	4	5
Direct estimation	1.7	3.5	3.1	3.7	5.4
$\log RR$	7	1	1	1	1
Direct estimation	6.3	1.2	1.5	0.9	1.2
$\log RR$	1	7	1	1	1
Direct estimation	1.1	3.4	1.9	1.0	1.4
$\log RR$	1	1	7	1	1
Direct estimation	1.5	2.1	4.5	1.0	2.0
$\log RR$	1	1	1	7	1
Direct estimation	1.3	3.6	1.0	6.5	2.2
$\log RR$	1	1	1	1	7
Direct estimation	1.2	1.4	2.2	1.0	6.1

Results: measurement error models

As the error in X is approximately known, the error in β can be computed by an extension of standard errors-in-variables regression models; however, this correction depends on F .

We tried using an estimated F and using the true simulation F .

Results: measurement error models



Summary of approaches

- Just use the imputed \hat{G} . Not very satisfactory. Confidence interval coverage is also poor.
- Apply standard measurement error correction methods (tricky: they need to be applied to X rather than G). This helps, but not enough
- Separate the impact of bias and non-identifiability by reporting all the optimal fits rather than just one. In progress
- Develop a formal Bayesian model incorporating what is known about composition, and use multiple imputations of G to incorporate the uncertainty in the health effects model
In progress