ACCOUNTING FOR MULTIPLE POLLUTANTS IN POLLUTION-MORTALITY STUDIES

Yuntae Kim [1], Dan Spitzner [2],
Zhengjun Zhang [2], Richard L. Smith [2]
and Montserrat Fuentes [1]

Paper available from:

http://www.stat.unc.edu/
postscript/rs/asa6.ps

Supported by E.P.A., in collaboration with N.R.C.S.E. and N.I.S.S.

[1] Department of Statistics, North Carolina State University, Raleigh, N.C.
[2] Department of Statistics, University of North Carolina, Chapel Hill, N.C.

1. INTRODUCTION

2. METHODOLOGY

3. PHILADELPHIA DATA

4. PHOENIX DATA

5. CONCLUSIONS

Background. Many epidemiological studies showing association between air pollution and mortality/morbidity.

Emphases: fine particulate matter ($PM_{2.5}$)

New EPA standard 1997, much controversy, legal challenge in 1999.

Current focus on particulate matter and especially on fine particles, but potential for looking at many more variables in the future (EPA Supersites program)

Key question for this talk: how to estimate the effect of many variables simultaneously.

Specific focus on

- Different lags of one pollutant

  Typical studies use combinations of lags 0, 1,...,4 days

- Different pollutants

  EPA criteria pollutants: PM_{10} (or TSP), SO_2, NO_2, CO, O_3

- Different elements

  In particular: availability of many elements in Phoenix data set

2 METHODOLOGY

$$y_t = \sum_j \beta_j x_{tj} + \epsilon_t$$

$$(y = X\beta + \epsilon)$$

y_t: Square root of daily death counts

$x_{tj},\ j \geq 1$: Covariates representing long-term trends, meteorological covariates and pollutants

Standard approaches to multicollinearity (Brown 1993):

- *Ridge regression*: for some $c > 0$,

$$\hat{\beta}^{(c)} = (X^T X + cI)^{-1} X^T y$$

- *Principal components regression*

- *Partial least squares regression*

Bayesian interpretation

Divide parameters into p "parameters of interest" and q nuisance parameters.

$$y = X_1^T \beta_1 + X_2^T \beta_2 + \epsilon,\ \epsilon \sim N(0, \sigma^2 I).$$

For OLSE $\hat{\beta}_1$,

$$\hat{\beta}_1 - \beta_1 \sim N(0, \sigma^2 V^{-1}),$$
$$V = X_1^T X_1 - X_1^T X_2 (X_2^T X_2)^{-1} X_2^T X_1.$$

Focus on β_1.

Prior distribution:

$$\beta_1 | \sigma^2 \sim N(\mu, W^{-1} \sigma^2)$$

Posterior distribution:

$$\beta_1 | \hat{\beta}_1, \sigma^2 \sim N[(V+W)^{-1}(V\hat{\beta}_1 + W\mu), (V+W)^{-1}\sigma^2].$$

Setting $W = cI$, $\mu = 0$ leads to standard ridge estimator as the posterior mean of this Bayesian calculation.

Objectives:

- Estimate each β_{1j}, $j = 1, ..., p$ with minimum MSE

- Select all β_{1j} such that $\beta_{1j} \geq t$ (*EDF problem*)

- Rank $\beta_{11}, ..., \beta_{1p}$ in order.

Possibility of *triple-goal estimates* (Shen and Louis 1998)

Key point: Ridge regression arises from one particular specification of prior and loss function, but there are many other possibilities for each.

Other aspects of methodology:

- *Scaling*

  Rescale each covariate of interest to have sample mean 1

- *Selection of c*

  Type II MLE approach

  Fully hierarchical approach (not implemented)

- *Model selection*

  Consider up to m lags but m unknown — put a prior distribution on m

Generalized ridge regression

$$\hat{\beta}_1 - \beta_1 \sim N(0, V^{-1}\sigma^2)$$
$$\beta_1 | \sigma^2 \sim N(0, W^{-1}\sigma^2)$$

$W = cI$: ridge regression

An alternative: Suppose $V = U \Lambda U^T$, U orthogonal, $\Lambda = \text{diag}\{\lambda_1, ..., \lambda_p\}, \lambda_1 \geq ... \geq \lambda_p \geq 0$.

Consider $W = UDU^T$,
$D = \text{diag}\{c_1, ..., c_p\}$.

$c_1 = ... = c_p = c > 0$: Ridge regression

$c_1 = ... = c_s = 0,\ c_{s+1} = ... = c_p = +\infty$ for some s: Principal component regression based on s PCs.

Generalized ridge regression: allow $c_1, ..., c_p$ arbitrary (≥ 0)

PHILADELPHIA DATA

Var.	Mean	SD	25%	50%	75%
Mort	31.5	6.4	26	31	38
Temp	54.3	17.8	40.0	55.3	70.3
Dew	42.3	19.1	27.8	43.5	58.8
TSP	67.3	26.9	47.5	63.0	72.0
SO_2	17.3	11.6	9.3	14.4	22.2
NO_2	39.6	12.9	30.5	37.6	46.1
CO	17.4	7.3	12.6	16.0	20.5
O_3	19.8	14.6	8.3	17.1	28.5

Table 1. Summary statistics for Philadelphia

Basic analysis:

- *Time trends* — B-splines with 180 knots

- *Meteorology* — Temperature, dewpoint with nonlinear terms, up to 4 lags. Also total sunshine.

- *Overdispersion* — about 10%

- *Autocorrelation* — not a problem

- *Pollutants* — model singly, and all together

- *Lagged effects* — suppose $\gamma_1,...,\gamma_m$ represent coefficients due to effects lagged $0,1,...,m-1$ days. Then $\sum_1^m \gamma_j$ is a measure of overall effect. Need to specify m.

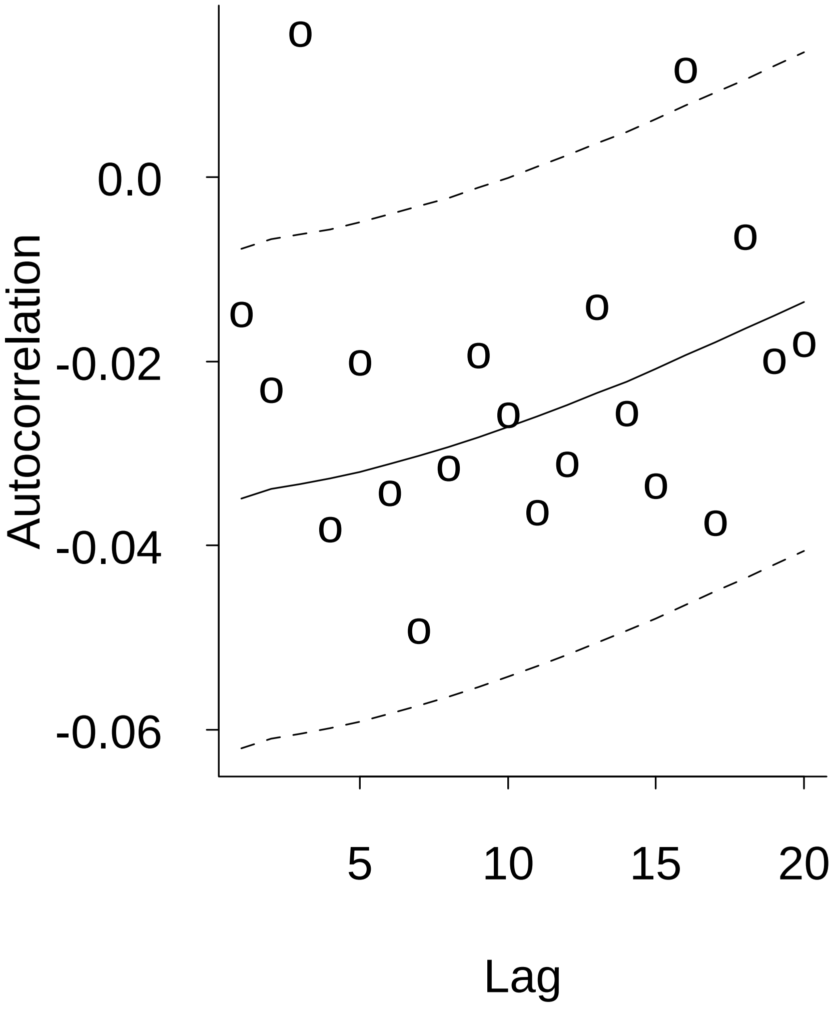


Fig. 1. Theoretical residual autocorrelations (solid curve) and 95% confidence bands (dashed curves), together with sample autocorrelations (circles).

Model	m	TSP	SO_2	NO_2	CO	O_3
Sing	1	.081	.057	.063	.049	.035
		(3.5)	(3.8)	(2.2)	(2.2)	(2.0)
Sing	2	.078	.062	.050	.044	.058
		(2.6)	(3.3)	(1.4)	(1.6)	(2.5)
Sing	3	.036	.045	.036	.036	.075
		(1.0)	(2.0)	(0.9)	(1.1)	(2.7)
All	1	.048	.042	–.050	.027	.030
		(1.4)	(2.0)	(–1.1)	(0.9)	(1.6)
All	2	.048	.054	–.082	.026	.052
		(1.1)	(2.0)	(–1.5)	(0.7)	(2.1)
All	3	–.033	.060	–.045	.030	.081
		(–0.6)	(1.9)	(–0.7)	(0.7)	(2.7)

Table 2. Parameter values and t statistics (in parentheses) when each pollutant is entered singly and when all are entered together; computed for $m = 1$, 2 and 3.

Empirical Bayes results

Calculate OLS and three forms of EB estimators (ridge, EDF estimate and generalized ridge) for the five pollutants and different values of m (explicit results shown for $m = 1, 5$).

Also, model selection for m based on prior m uniform on $1, ..., M$, for $M = 5$ or $M = 10$.

Calculate posterior probability that $\gamma_k < 0$ where γ_k represents overall effect of k'th pollutant. (Bayesian equivalent of a significance test.)

Model-averaged results based on $M = 5$:

k	Pollutant	$P(\gamma_k < 0)$
1	TSP	.29
2	SO_2	.15
3	NO_2	.59
4	CO	.30
5	O_3	.11

Some "significant" results by conditioning on individual values of m, though not for TSP.

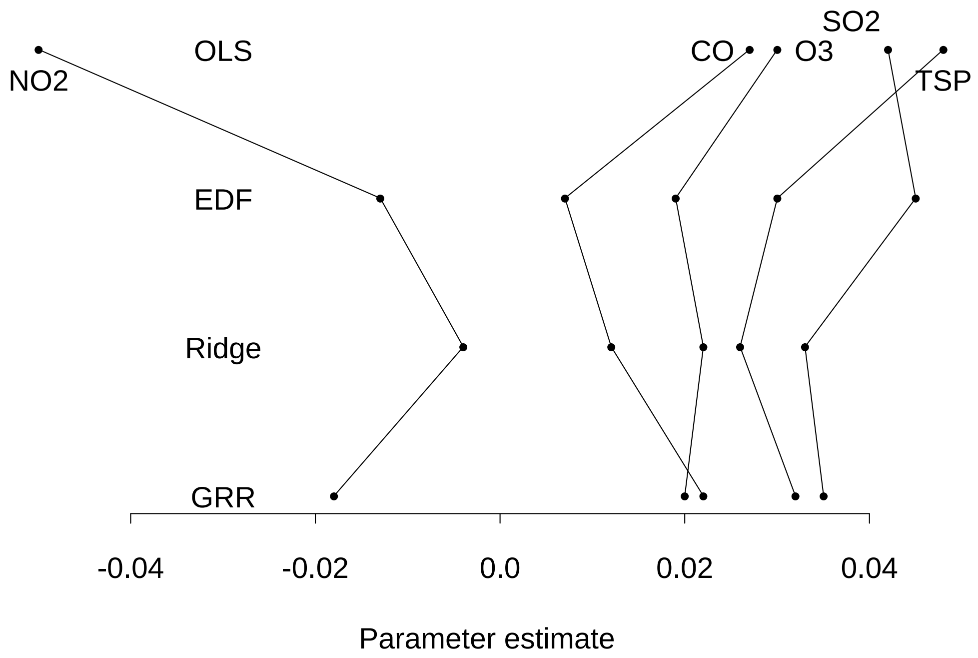


Fig. 2. Plots of the regression coefficients for five pollutants, current day's value of pollutants. Top to bottom: Least squares estimates, EDF estimates, ridge regression estimates, GRR estimates. The solid lines between the plots connect points corresponding to the same element.

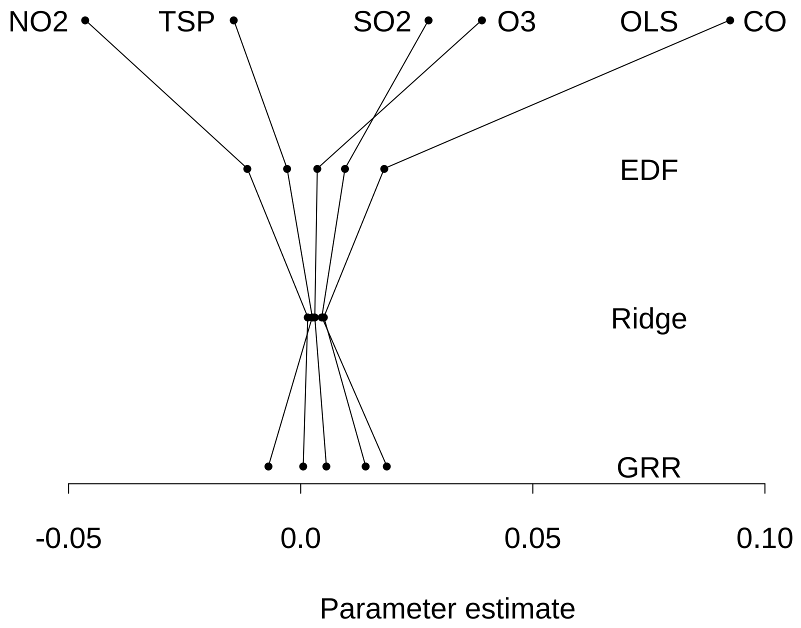


Fig. 3. Similar to Fig. 2, but based on five-day averages of pollutant variables.

PHOENIX DATA

Var.	Mean	SD	25%	50%	75%
Mort	15.4	4.4	12	15	18
MinT	17.5	8.0	10.8	16.8	24.7
MaxT	29.4	8.2	22.6	29.8	36.7
SH	8.51	4.52	5.3	7.0	10.5
PMCS	33.5	17.4	22.5	30.2	40.8
$PM_{2.5}$	13.0	7.1	8.2	11.4	16.7

Table 3. Summary statistics for Phoenix: mortality age 65+, daily minimum and maximum temperature (°C), Specific humidity (g/kg), PMCS (μg/m^3), $PM_{2.5}$ (μg/m^3).

Analysis:

- *Time trends* — B-splines with 18 knots

- *Meteorology* — Temperature, specific humidity, (wind speed)

- *Pollutants* — PM_{10}, $PM_{2.5}$, $PMCS=PM_{10}–PM_{2.5}$,

Consider variables averages over i days, lagged j days.

i	j	PMCS	t	$PM_{2.5}$	t
1	0	6	0.4	–10	–0.3
1	1	19	1.3	0	0.0
1	2	39	2.7	34	0.8
1	3	13	0.9	52	1.3
1	4	1	0.1	22	0.6
2	0	21	1.1	–9	–0.2
2	1	47	2.5	22	0.5
2	2	39	2.2	62	1.3
2	3	10	0.6	52	1.1
3	0	49	2.2	10	0.2
3	1	49	2.4	53	1.0
3	2	32	1.7	67	1.2

Table 4. PMCS and $PM_{2.5}$ coefficients for different averaging lengths i and lags j; cols. 3 and 5 give the coefficient $\times 10^4$, cols. 4 and 6 give the t values.

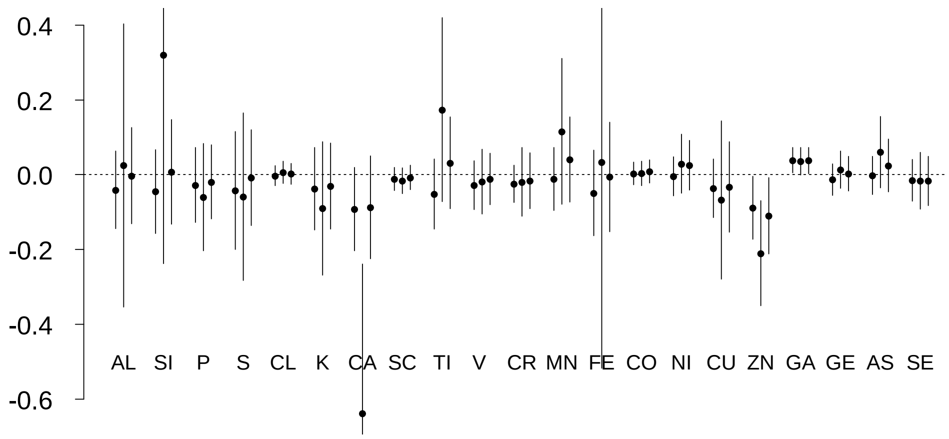
Influence of individual elements

42 chemical elements available on daily basis (300 days)

Refit meteorological model

Consider effects of individual elements singly, or 3 forms of joint estimations: OLS, ridge and EDF.

Also repeat analysis without SI and CA (outliers)



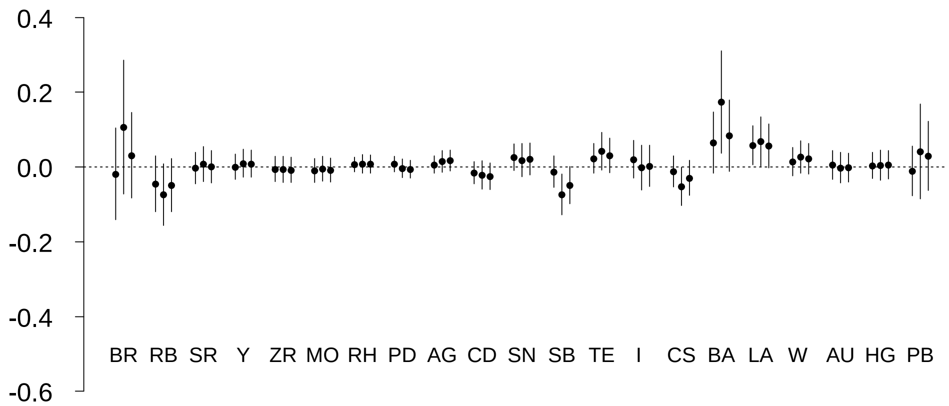


Fig. 4. Estimates and 95% confidence bands each element. Left: OLS estimate when variables added one at a time. Middle: OLS estimate when variables added all together. Right: Ridge estimates.

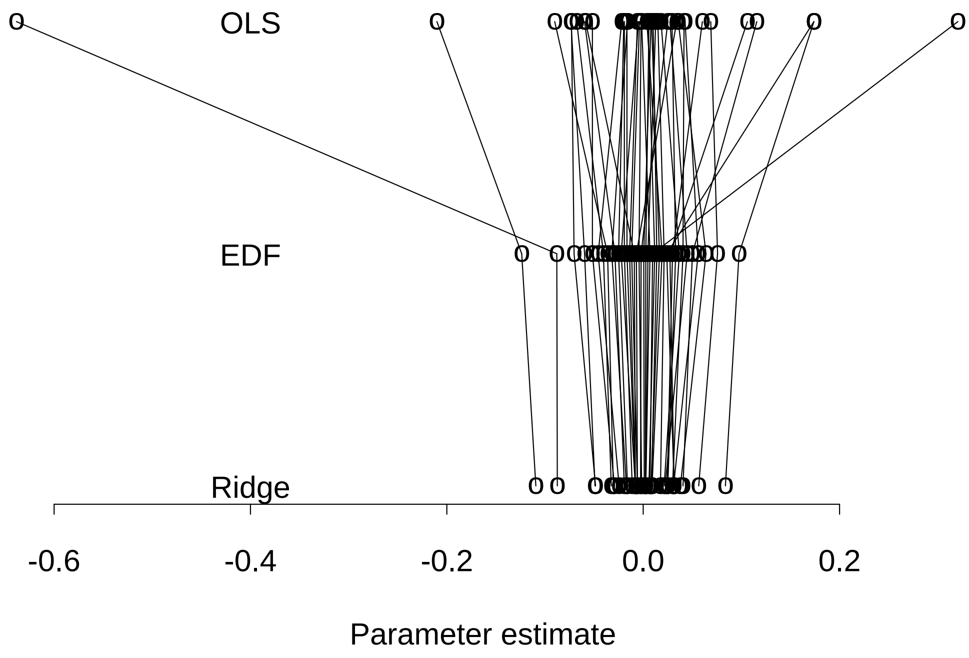


Fig. 5. Plots of the regression coefficients for 42 individual elements. Top: Least squares estimates. Bottom: Ridge regression estimates. Middle: EDF estimates. The solid lines between the plots connect points corresponding to the same element.

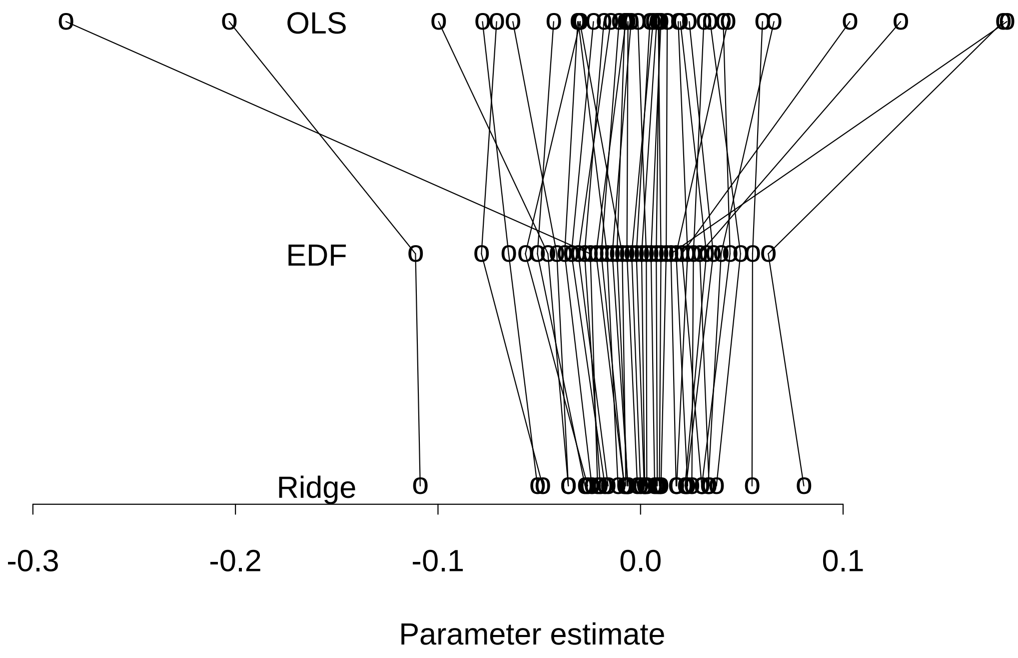


Fig. 6. Similar to Fig. 5, after repeating the analysis with SI and CA omitted.

CONCLUSIONS

Estimation of mortality effects requires consideration of multiple pollutants. EB approach is general, encompasses ridge regression, PC regression, etc. Many open questions of implementation.

For Philadelphia, we conclude that there is not enough evidence to identify effects of individual pollutants

For Phoenix, analysis suggests PMCS has stronger effect than $PM_{2.5}$, but “multiple pollutants” aspects still have to be considered

No evidence that any individual element has an especially strong effect, but data still limited.