ACCOUNTING FOR MULTIPLE POLLUTANTS IN POLLUTION-MORTALITY STUDIES

Yuntae Kim, Dan Spitzner, Zhengjun Zhang,
Richard L. Smith and Montserrat Fuentes

Key Words: Atmospheric pollution, Empirical Bayes, Health effects, Model selection, Multicollinearity, Particulate matter, Principal components regression, Ridge regression, Shrinkage, Triple-goal estimators.

Abstract:

There is extensive and continuing concern over the human health effects of atmospheric particulate matter, which is reflected in the debate over the 1997 EPA standards. Much of the evidence supporting such standards comes from statistical and epidemiological studies employing time series regression of mortality and morbidity on a variety of covariates including particulate matter. Among the statistical issues raised by such studies are multicollinearity and selection effects when large numbers of related regressors are considered simultaneously. We propose a method of dealing with such issues by viewing them in an empirical Bayes context. The relation to existing methods of dealing with multicollinearity, including ridge regression and principal components regression, is discussed, as also is the issue of shrinkage when many coefficients are estimated simultaneously. These ideas are illustrated with reference to data from Philadelphia and Phoenix. The Phoenix data set allows for direct comparison of the effects of fine and coarse particles, as well as the effects of 44 constituent chemical elements. One of our conclusions is that the effect of coarse particles appears to be stronger than that of fine particles, which is contrary to some of the thinking underlying the 1997 revision of the standard.

1. Introduction

The environmental health effects of atmospheric pollution have been the source of much political and scientific controversy over the past several years. Particular controversy has arisen over the new standards intoduced by the United States Environmental Protection Agency (EPA) in 1997, which confirmed an existing standard for PM_{10} (particulate matter of aerodynamic diameter 10 microns or less) and introduced a new standard based on $PM_{2.5}$ (particulate matter of aerodynamic diameter 2.5 microns or less). This led to many challenges including a hearing before the U.S. Court of Appeals, which has, temporarily at least, overturned the standard.

Much of the scientific debate surrounding these standards is concerned with the statistical interpretation of observed associations between daily particulate matter levels and daily mortality in observational time series. These issues have been extensively discussed in previous papers, e.g. Samet *et al.* (1995, 1997), Dominici et al. (1999a, 1999b), Zeger *et al.* (1999) and Smith *et al.* (1998, 1999a, 1999b). For a broad-based discussion of the scientific debate surrounding particulate matter, including priorities for future research, an excellent source is the NRC report (National Research Council, 1998).

The present paper examines some of the statistical issues associated with the presence of multiple pollutants and associated questions of confounding and multicollinearity. There are at least three separate issues to consider here:

- The effect on the pollution-mortality relation-

Yuntae Kim is a graduate student and Montserrat Fuentes is Assistant Professor, Department of Statistics, North Carolina State University, Raleigh, NC. Dan Spitzner and Zhengjun Zhang are graduate students and Richard L. Smith is Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260, to whom all correspondence should be addressed. Email address: rls@email.unc.edu. This research was supported in part by EPA Cooperative Agreement CR 825173-01-1 to the University of Washington, as a subcontract to the National Institute of Statistical Sciences, Research Triangle Park, NC. Richard Smith is also supported by the NSF, grant number DMS-9705166. For data sources, the authors would like to thank: Scott Zeger and Francesca Dominici of Johns Hopkins University, for the Philadelphia data; Chris Mrela of the Arizona Health Services Department, for the Phoenix mortality data, also Peter Guttorp of the University of Washington for transmitting those data to us; and the PM Research Monitoring Network Data report for Phoenix, Arizona, February 1995 - December 1997, produced by the US EPA National Exposure Research Laboratory, Research Triangle Park, NC (Phoenix air pollution data). Additional meteorological data were obtained from the web site of the National Climatic Data Center, Asheville, North Carolina. We also thank Jerome Sacks and Peter Bloomfield for comments on the work. This paper has not been subjected to the Environmental Protection Agency's internal peer review system and no endorsement by the Agency should be implied or inferred.

ship of different exposure measures created by combining different lags — for example, although there are by now numerous studies of the relationship between PM_{10} and mortality, there has been no unanimity on which lags of PM_{10} to include (present day's value, one-day lag, etc.) or on how to account for the selection of an exposure measure in subsequent tests and confidence intervals;

- The effect of different atmospheric pollutants — apart from PM_{10} and ozone (O_3), three of the EPA's other "criteria pollutants", sulfur dioxide (SO_2), nitrogen dioxide (NO_2) and carbon monoxide (CO) have been studied for possible mortality effects;

- The effect of different constituents of particulate matter — for example, the Phoenix data set discussed below includes not only PM_{10} and $PM_{2.5}$ readings, but also measurements for 44 chemical elements contained within $PM_{2.5}$, and it would be of great interest to establish which of these elements was most responsible for the observed effects.

In the remainder of the paper, we first discuss some general methodological issues and then consider two specific data sets.

2. Methodology

2.1 The problem of multicollinearity

The main model considered in this paper is a standard linear regression equation

$$y_t = \sum_j \beta_j x_{tj} + \epsilon_t, \tag{1}$$

where y_t is some transformation (e.g. log or square root) of daily mortality on day t, $\{x_{tj}\}$ are known covariates, $\{\beta_j\}$ are unknown coefficients and $\{\epsilon_t\}$ are independent, normally distributed errors with mean 0 and unknown common variance σ^2. The covariates typically fall into three categories: (a) seasonal or long-term trend effects, which we shall represent as linear combinations of cubic spline basis functions with estimated coefficients (Samet *et al* 1997, Smith *et al.* 1998), (b) meteorological variables, and (c) pollution variables. One of the difficulties in this kind of analysis is that many of the covariates are themselves highly correlated, leading to multicollinearity. In the past, this issue has surfaced even when only a single pollution variable has been considered, because of possible confounding with meteorology, but the effect may be expected to be enhanced when many pollutants are considered simultaneously. In the present paper, we concentrate on the pollutants themselves, treating the meteorology and long-term trend effects as known nuisance factors.

An excellent discussion of multicollinearity and possible remedies has been given by Brown (1993), who identified three broad strategies to deal with it:

- Ridge regression, in which the standard least squares regression estimator is replaced by

$$\hat{\beta}^{(c)} = (X^T X + cI)^{-1} X^T y. \tag{2}$$

  Here, $\hat{\beta}^{(0)}$ is the usual least squares estimator. As c increases from 0, the estimator $\hat{\beta}^{(c)}$ becomes biased, but often with a dramatic drop in mean squared error (MSE),

- Principal components regression, in which the X matrix is both orthogonalized and reduced in dimension by applying a principal components analysis to the X matrix prior to performing least squares regression,

- Partial least squares regression, which is similar to principal components regression except that the selection of suitable linear combinations of the X variables is based on maximizing the correlation with y rather than maximizing the variance among the X variables.

The ridge regression estimator is often recommended in cases where the design matrix X is far from orthogonal, but even in near-orthogonal cases, if there are many coefficients to estimate, ridge regression may improve on standard least squares, because of shrinkage effects. In modern statistics, such ideas are central to empirical Bayes methodology (see e.g. Carlin and Louis 1996), which has already been applied in a number of studies related to particulate matter (Samet *et al.* 1997, Dominici *et al.* 1999a).

However, the shrinkage effect is still only imperfectly understood: one feature, for example, is that the amount of shrinking which is appropriate depends on the purpose to which the analysis will be put (in decision-theoretic terms, on the loss function). A theoretical analysis of this phenomenon has recently been given by Shen and Louis (1998), and motivates some of the techniques which will be used in the latter part of the present paper.

2.2 Bayesian interpretation

Suppose there are $p+q$ regressors, where the first p are "parameters of interest" (e.g the PM variables) and the remaining q are nuisance parameters such as coefficients of meteorology and trend terms. Writing $\beta^T = (\beta_1^T \quad \beta_2^T)$, $X = (X_1 \quad X_2)$, equation (1) may be written

$$y = X_1^T\beta_1 + X_2^T\beta_2 + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I). \tag{3}$$

Standard theory shows that the least squares estimator of β_1, denoted $\hat{\beta}_1$, satisfies

$$\hat{\beta}_1 - \beta_1 \sim N(0, \sigma^2 V^{-1}) \tag{4}$$

where

$$V = X_1^T X_1 - X_1^T X_2 (X_2^T X_2)^{-1} X_2^T X_1. \tag{5}$$

In most cases σ^2 is unknown, but we have an independent estimator s^2, with ν degrees of freedom. The Bayesian theory is easily extended to this case, but for simplicity, we shall just assume σ^2 is known here.

Since $\hat{\beta}_1$ is sufficient for β_1, all the desired estimates, tests, etc., can be derived as functions of $\hat{\beta}_1$. For example, in the case $q = 0$, the standard ridge regression estimator may be derived from (2) as

$$\hat{\beta}_1^{(c)} = (X^T X + cI)^{-1} X^T X \hat{\beta}_1, \tag{6}$$

which can be viewed as a shrinkage operation performed directly on the least squares estimator $\hat{\beta}_1$. We adopt this viewpoint throughout the paper, i.e. take (4) as the starting point, and view all subsequent estimators as operations performed on $\hat{\beta}_1$.

Suppose $\beta_1|\sigma^2$ has prior distribution $N(\mu, W^{-1}\sigma^2)$ with μ, W known. A standard Bayesian calculation shows that given σ^2, the posterior distribution of β_1 given $\hat{\beta}_1$ is

$$N[(V+W)^{-1}(V\hat{\beta}_1 + W\mu), (V+W)^{-1}\sigma^2]. \tag{7}$$

Thus, in particular, the posterior mean of β_1 is

$$(V+W)^{-1}(V\hat{\beta}_1 + W\mu). \tag{8}$$

In the standard ridge regression setting, with $V = X^T X$, suppose we set $W = cI$, $\mu = 0$. Then (8) agrees exactly with (6). This shows how ridge regression may be derived as a special case of the Bayesian estimator.

From a decision theory perspective, it is well known that the posterior mean is also the optimal Bayes estimator under squared error loss. However, it is not necessarily optimal with respect to other loss functions. In many contexts where Bayes or empirical Bayes analysis is performed, squared error is not the appropriate loss function. Two other criteria which have been discussed are

- Select all β_j such that $\beta_j \geq t$ for fixed t (as t varies then this is equivalent to estimating the *empirical distribution function* of the $\{\beta_j\}$),
- Rank $\beta_1, ..., \beta_p$ in order,

in either case with some penalty based on the costs of misclassification.

In the context of selecting "significant" pollutants in a pollutant-mortality study, either of these would seem to be a more reasonable loss function than mean squared error.

2.3 The scaling problem

The goals of ranking the $\{\beta_j\}$, or of picking out all values above a given threshold, do not make sense if the parameters are defined on totally different scales. For example, one pollutant may be a highly toxic substance present in the environment in minute quantities, while another may be much less toxic but also much more prevalent. Evidently, the resulting β_j values are not directly comparable.

We shall resolve this problem by rescaling all the pollutants so that they have sample mean 1. The rationale for this is as follows: suppose the j'th pollutant has mean μ_j. If this is a toxic pollutant, then an ideal solution would be to reduce it to 0. In that case the benefit, as quantified through the reduction in the mean value of y, is $\beta_j\mu_j$. However if all the μ_j values are the same number, which we are arbitrarily assuming to be 1, then the benefit is just β_j. Thus on this scale, the order of the β_j values corresponds directly to the potential benefit to be derived from controlling the different pollutants.

2.4 Triple-goal estimators

Suppose $\beta_1 = (\beta_{11}, ..., \beta_{1p})$. As noted already, there are different loss functions with respect to which we might choose to estimate these parameters. Under squared error loss, the Bayes estimate for β_{1k} ($1 \leq k \leq p$) is just the posterior mean:

$$\eta_k = \mathrm{E}\{\beta_{1k} \mid y\}. \tag{9}$$

When the objective is to estimate the ranks $R_k = \mathrm{rank}(\beta_{1k})$ with minimum mean squared error, the Bayes estimator is the posterior mean ranks,

$$\bar{R}_k = \mathrm{E}\{R_k \mid y\} = \sum_{j=1}^{p} \Pr\{\beta_{1k} \geq \beta_{1j} \mid y\}.$$

In general, $\bar{R}_k$ will not be an integer. If we require that all estimated ranks be integers, the obvious estimator is based on ranking the $\bar{R}_k$,

$$\hat{R}_k = \text{rank}(\bar{R}_k). \tag{10}$$

A third objective is to estimate the empirical distribution function (EDF), $G_p(t) = \frac{1}{p}\sum_k I(\beta_{1k} \leq t)$, where $I(\cdot)$ is indicator function. Under, e.g., integrated squared error loss, the Bayes estimate for fixed t will be the posterior mean of $G_p(t)$,

$$\bar{G}_p(t) = \text{E}\{G_p(t) \mid y\} = \frac{1}{p}\sum_k \Pr\{\beta_{1k} \leq t \mid y\}. \tag{11}$$

As with $\bar{R}_k$, this is open to the criticism that the estimator is not a member of the class of objects being estimated (in this case, EDFs with at most p mass points), but an alternative estimator which achieves that objective is to put mass $1/p$ at each of the points

$$\hat{U}_j = \bar{G}_p^{-1}\left(\frac{2j-1}{2p}\right), \quad j = 1, ..., p. \tag{12}$$

The resulting estimator is denoted $\hat{G}_p$.

The formulae (9)–(12) are taken from a recent paper by Shen and Louis (1998) where they were derived in a simpler setting in which the parameters had independent posterior distributions; their derivations as Bayes estimators, however, hold in the present setting as well. Shen and Louis argued that the posterior means $\{\eta_k\}$ typically lead to too much shrinkage if the loss function is anything other than squared error, and they proposed an alternative set of *triple-goal estimators* which is a single set of estimators designed to perform reasonably under each of the three loss function. In our present setting we have not attempted to derive a single "triple-goal" estimator in the same sense as Shen and Louis, but it is important to note that different loss functions give rise to different estimates, and in particular that the $\{\eta_k\}$, which correspond to the standard ridge estimators, are typically not optimal if the loss function is other than squared error.

2.5 Specifying c

Brown (1993) discussed a number of approaches to the ridge constant c, ranging from the graphical "ridge trace" method, which is now regarded as too *ad hoc* to be of general applicability, through various forms of cross-validation, to the more formal "type II maximum likelihood" procedure, where c is chosen to maximize the likelihood based on the marginal density

$$\hat{\beta}_1 \sim N[0, \sigma^2(V^{-1} + c^{-1}I)].$$

Since type II maximum likelihood is easily implemented with modern software, we have adopted it as the method of choice in the present study, though it needs to be pointed out that the MLE does not always exist (sometimes the likelihood is maximized as $c \to \infty$).

An alternative strategy might be fully Bayesian, giving c a prior distribution and formulating the problem as a hierarchical model. Although this seems a promising possibility, it has not been considered in the present analysis.

2.6 Generalized ridge regression

So far, we have assumed that in the notation of (8), $W = cI$ with a constant c and identity matrix I. However, of course, we could consider more general specifications of the prior.

One idea is related to principal components regression (PCR), briefly mentioned at the beginning of this section. Suppose $V = U\Lambda U^T$ where $\Lambda = \text{diag}(\lambda_1, ..., \lambda_p)$, $\lambda_1 \geq ... \geq \lambda_p > 0$, and U is orthogonal. Fix s where $0 < s < p$. A PCR based on the first s principal components is equivalent to the usual least squares estimator after replacing the last $p - s$ columns of U by 0. This can also be derived as a Bayes estimator, as follows. Consider a prior based on $W = UDU^T$, in which D is a diagonal matrix with diagonal entries $c_1, ..., c_p$. Of course, the case $c_1 = ... = c_p = c$ is the same as the model for ridge regression given earlier. Consider the case $c_1 = ... = c_s = 0$ and $c_{s+1} \to \infty, ..., c_p \to \infty$. It can be shown that the Bayes estimator in this case is the same as the PCR.

Thus, this form of prior, with arbitrary $c_1, ..., c_p$, includes both ridge regression and PCR as special cases and therefore may be regarded as a generalization of both. We shall refer to it as generalized ridge regression (GRR). Type II MLEs for $c_1, ..., c_p$ are easily calculated; as with ordinary ridge regression, it can happen that some of the MLEs are infinite, but since this is equivalent to omitting the corresponding principal components, this issue is less of a problem here.

2.7 Model selection

In the analyses to follow, we take as our basic measure of pollution the average of pollutant levels for m days immediately preceding (and including) the current day. However, we still have to specify m. This

will be done via a prior distribution, with an upper bound M for m, and prior probabilities $p_1, ..., p_M$ associated with specific values of m. For example, we may consider $M = 5$ with $p_1 = ... = p_5 = \frac{1}{5}$. In general, we assume a prior density of the mixture of normals form

$$\beta_1 \sim \sum_{m=1}^{M} p_m N(0, W_m^{-1}\sigma^2),$$

where $W_m^{-1}\sigma^2$ is the prior density corresponding to model m. A technical point here is that in the typical case where the model selection corresponds to setting certain components of β_1 to 0, W_m^{-1} will be singular and so in the literal sense, W_m does not exist, but this can be dealt with formally by allowing the components in question to have positive prior variances and taking limits as those variances tend to 0.

Under this model, the marginal density of $\hat{\beta}_1$ (useful for type II maximum likelihood estimation) is

$$\hat{\beta}_1 \sim \sum_m p_m N[0, (V^{-1} + W_m^{-1})\sigma^2],$$

and the conditional density of β_1 given $\hat{\beta}_1$ (used for the Bayesian inferences) is

$$\sum_m \left(\frac{q_m}{q.}\right) N[(V + W_m)^{-1} V\hat{\beta}_1,\ (V + W_m)^{-1}\sigma^2]$$

with

$$\begin{aligned} q_m &= p_m (V^{-1} + W_m^{-1})^{-1/2} \cdot \\ &\quad \cdot \exp\left\{-\frac{1}{2\sigma^2}\hat{\beta}_1^T (V^{-1} + W_m^{-1})^{-1}\hat{\beta}_1\right\} \end{aligned}$$

and $q. = \sum q_m$. Within this framework, therefore, we may compute the posterior model selection probabilities $\{q_m/q.\}$, the type II MLE for c or $c_1, ..., c_p$ and posterior densities of $\hat{\beta}_1$, both conditionally on model m and unconditionally (i.e. averaged over models), by direct extension of the methods used for a single model. We omit the details of these calculations.

3. Analysis of Philadelphia data

3.1 Background and data

Our first analysis uses one of the most widely studied data sets, based on Philadelphia; see Samet *et al.* (1995, 1997), who give references to many other studies. The data set used by Samet *et al.* (1997) covered the period 1974–1988 included temperature and dewpoint as meteorological variables, as well five "criteria pollutants" (TSP, SO_2, O_3, NO_2, CO). Here TSP (total suspended particulates) are used as an alternative to PM_{10} as regular measurements of the latter did not begin until 1987.

Var.	Mean	SD	25%	Median	75%
Mort	31.5	6.7	27	31	36
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	82.0
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: Mortality age 65+, temperature and dewpoint (°F), TSP ($\mu g/m^3$), SO_2 (ppb), NO_2 (ppb), CO (ppb × 100), O_3 (ppb). From Samet *et al.* (1997).

Mortality data are available broken down into a number of age and cause-of-death categories, but for the present analysis, we consider only the age group 65+ and combine all non-accidental causes into a single death count. Samet *et al.* (1997) claimed an improved fit to the overall model by using different long-term trend estimates in the separate age categories 55–64, 65–74, 75+, but we have not attempted that in our analysis.

A table of summary statistics is presented in Table 1.

We focus particularly on the following questions, (a) the selection of lags of the pollution variables, (b) assessing the effects of individual pollutants within a muliple-pollutant model. Samet *et al.* (1997) compared a large number of different models with different combinations of pollutant variables, concluding that there was general evidence of an association between air pollution and mortality but that it was not possible to pinpoint this on a single pollutant based on the evidence in this data set.

Our basic analysis uses the standard linear model (1) with y_t defined as square root of daily nonaccidental mortality in the 65+ age group.

3.2 Time trends

Time trends are modeled as a linear combination of 180 B-spline basis functions: this is very similar to the method employed by Samet *et al.* (1997); see also Smith *et al.* (1998, pp. 96–98) for further discussion of B-splines.

3.3 Meteorology

We considered an initial variable set consisting of TEMP (daily mean temperature in ^{o}F), TEMPSQ (square of TEMP), HITEMP=(TEMP-80)$_+$, LOTEMP=(25-TEMP)$_+$, DEW (daily dewpoint temperature in ^{o}F), DEWSQ, HIDEW=(DEW-70)$_+$, LODEW=(10-DEW)$_+$. The variables HITEMP, LOTEMP, HIDEW, LODEW were intended to allow for possible different behavior at the extreme values of each variable, with the cutoffs 80 and 25 for TEMP, 70 and 10 for DEW, determined (arbitrarily) as the 95th and 5th percentiles of the empirical distribution for each variable. All of these variables were considered at lag 0 (i.e. today's value) as well as daily lags 1 through 4; VAR$_j$ means variable VAR lagged j days. We followed the policy that if a squared variable was included in the model then the corresponding linear variable would be included as well; otherwise, any combination of meteorological variables is permitted. Variable selection was by backward selection with a .05 significance level determining whether or not a particular variables was included in the final model. After following this strategy the following variables were included: TEMP$_1$, TEMPSQ$_1$, HITEMP$_0$, HITEMP$_1$, HITEMP$_2$, HITEMP$_4$, DEW$_0$, DEW$_1$, DEW$_2$, DEW$_4$, DEWSQ$_0$, DEWSQ$_1$.

3.4 Autocorrelation and overdispersion

Some earlier studies, notably Samet *et al.* (1995), considered the effects of autocorrelation in Philadelphia. In our analysis, based on a different model of the long-term time trend, the observed autocorrelations of the residuals are not significant when compared with their expected values under the independent-errors assumption.

Overdispersion may be said to exist when the variance of $\sqrt{Y}$ is greater than 0.25 (its approximate value when Y is Poisson). For us, the sample variance is .276, or about a 10% overdispersion. The overdispersion is somewhat greater than that reported by Samet *et al.* (1997), who suggested it was about 5%, but overall confirms that there is some, though not very strong, overdispersion in this data set.

3.5 Modeling time-dependent effects of air pollution

One issue that complicates the interpretation of an air pollution–mortality link is that the time-dependence of the effect — in other words, how the influence of a high-pollution event is spread over the several days following the event — is unknown. Past studies have employed different combinations of lagged pollution variables, either considering single-day measurements and their lagged values, or averages of between two and five daily values, again with lags. This creates difficulties over such issues as how to compensate the resulting inferences for the selection effect involved in picking out the pollution-based variable with the largest absolute value or the largest statistical significance. Most current studies have restricted attention to variables of lags 0–4 though some have suggested that effects are persistent over much longer time scales (Zeger *et al.* 1999).

Our approach is to assume that the form of the dose-response curve following a high-pollution event is unknown, but restricted to the m days following the event (including the day of the event itself). Renumbering coefficients if necessary, we may assume that $\beta_{11}, ..., \beta_{1m}$ are the coefficients of the pollution variable in the regression (1) at daily lags 0, 1, ..., $m-1$. Within such a model, if the level of a pollutant rose by an amount ϕ uniformly on all days, the net effect on E$\{y_t\}$ would be an increase of $\phi \sum_1^m \beta_{1j}$. Thus there is a reason for considering the sum of coefficients, $\sum_1^m \beta_{1j}$, as the "parameter of interest", however we model the individual β_{1j} coefficients.

In the current version of the study, we simplify this further by assuming $\beta_{11} = ... = \beta_{1m}$, though within our empirical Bayes framework, it should also be possible to consider cases in which $\beta_{11}, ..., \beta_{1m}$ are different, and we aim to consider this in future work.

3.6 Least squares results

To the meteorological and long-term trend models identified in sections 3.2 and 3.3, we now add various combinations of pollutant variables, using ordinary least squares to fit the models. The five pollutants were considered, for each of $m = 1$, 2 and 3, where for $m > 1$ the average of lagged days 0, 1,...,$m-1$ was taken to define the pollutant variable. If any of the m individual days was missing, we defined the m-day average to be missing also. The pollutants were added both one at a time, and all together. t statistics were formed by dividing each parameter estimate by its standard error. Table 2 gives the parameter estimates and t statistics for each model considered.

It can be seen that significant values have been obtained for each of the pollutants when they are added one at a time, but that the results are far less clear-cut when all the pollutants are included simultaneously. The table also shows the sensitivity

Model	TSP	SO_2	NO_2	CO	O_3
S1	0.081	0.057	0.063	0.049	0.035
	(3.5)	(3.8)	(2.2)	(2.2)	(2.0)
S2	0.078	0.062	0.050	0.044	0.058
	(2.6)	(3.3)	(1.4)	(1.6)	(2.5)
S3	0.036	0.045	0.036	0.036	0.075
	(1.0)	(2.0)	(0.9)	(1.1)	(2.7)
A1	0.048	0.042	-0.050	0.027	0.030
	(1.4)	(2.0)	(–1.1)	(0.9)	(1.6)
A2	0.048	0.054	–0.082	0.026	0.052
	(1.1)	(2.0)	(–1.5)	(0.7)	(2.1)
A3	–0.033	0.060	–0.045	0.030	0.081
	(–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 (models S1, S2, S3) and when all are entered together (A1, A2, A3); models S1, A1 use $m = 1$, S2, A2 use $m = 2$ and S3, A3 use $m = 3$.

of the results to m, the number of lags included in the model. Overall, the results are not in conflict with those of Samet *et al.* (1997), but they reinforce the sensitivity of the results to model assumptions.

3.7 Empirical Bayes results

We now consider how these results are affected by the different kinds of empirical Bayes analysis discussed in section 2. Our main purpose in the current study is to use empirical Bayes analysis to combine the five pollutants, though as already mentioned, in principle the method could be applied to more general models such as those involving variable coefficients for each lag.

Suppose all five pollutants are put into the model at the same time. We consider four variants on the analysis: (a) least squares estimates, as given in section 3.6; (b) the ridge regression estimates with parameter c determined by type II maximum likelihood; (c) the EDF estimates defined by a combination of (10) and (12) — in this case (12) is used to determine the mass points while the ordering in (10) is used to associate each variable with a corresponding mass point; (d) the GRR estimates defined in section 2.6. For the single-day values ($m = 1$ in Table 2) the results of this analysis are shown in Fig. 1. For five-day averages ($m = 5$), they are in Fig. 2.

In both cases, the different forms of empirical Bayes analysis have had the effect of shrinking the least squares estimates, but we do not feel confident in labelling any one of the four analyses as the single "best" analysis. The main message of the analysis is to highlight how much individual parameter estimates are sensitive to the choice among these different methods of estimation. However some features are persistent across different analyses: for example, in Fig. 1 the NO_2 coefficient is negative for all four analyses, while in Fig. 2 it appears that CO is the variable with the largest coefficient even though the effect is markedly reduced in all three versions of empirical Bayes analysis.

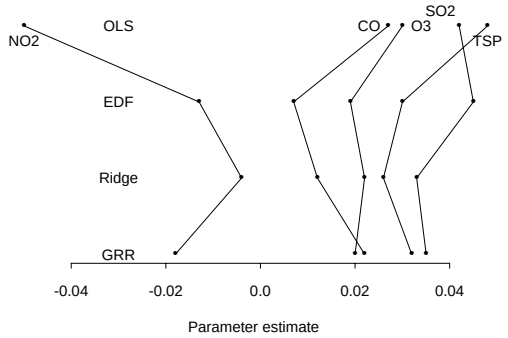

Fig. 1. Plots of the regression coefficients for five pollutants, current day's value of pollutants. From top to bottom: Least squares estimates, EDF estimates, ridge regression estimates, GRR estimates. The solid lines between the plots connect points corresponding to the same element.

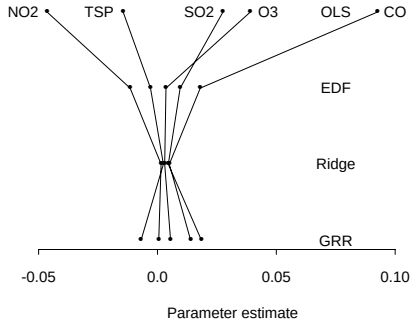

Fig. 2. Similar to Fig. 1, but based on five-day averages of pollutant variables.

Apart from the choice among different forms of empirical Bayes analysis, the preceding conclusions suggest two questions for further study. First, can we say anything about the choice of m? Second, what about the "significance" of different pollutants?

We approach the first question from the point of view outlined in section 2.7. A prior distribution

was imposed on the order m of the model by specifying $M = 5$ (the largest value of m permitted) and $p_m = \frac{1}{5}$ for $m = 1, ..., 5$. The ridge regression analysis was then rerun with c selected by type II MLE, separately for each m. As noted already in section 2.5, a more comprehensive Bayesian analysis would allow also for a prior distribution on c, but we have not implemented that yet.

For this analysis, the posterior probabilities $q_m/q.$ described in section 2.7 are .23, .17, .17, .28, .15 for $m = 1, 2, 3, 4, 5$. Evidently, the data do not provide much evidence to discriminate among the different values of m. For studying statistical significance, we have computed the posterior probability that $\beta_{1j} < 0$ for each parameter of interest. This can be considered the Bayesian equivalent of the frequentist p-value for a one-sided test of significance. These probabilities for each of the TSP, SO_2, NO_2, CO and O_3 variables are .29, .15, .59, .30, .11. There does not appear to be much evidence for "significance" of any of the five variables.

There are other versions of the analysis, which will be reported in more detail elsewhere, and some of these do show a significant effect for either SO_2 or O_3, but none of our present analyses leads to a significant result for TSP when considered in combination with other pollutants.

4. Analysis of data from Phoenix, Arizona

The data from Phoenix comprise three years' of meteorology, air pollution and mortality data covering the period February 1995–December 1997. The interest in this series is that it is one of the few currently extant series to include daily measurements of both PM_{10} and $PM_{2.5}$, and also that there are a number of breakdowns of the air pollution data including analysis of 44 chemical elements (excluding carbon) that are constituents of $PM_{2.5}$. The comparison of regression results for PM_{10} and $PM_{2.5}$ is of particular interest in the light of the EPA's decision in 1997 to create a new tightened standard based on $PM_{2.5}$ while leaving the earlier PM_{10} standard intact.

4.1 Data description

Meteorology and air pollution data were obtained from the particulate matter research monitoring platform in Phoenix, which is one of three that has been established by the EPA's National Exposure Research Laboratory at Research Triangle Park, NC. Additional meteorological data were obtained from the archives of the National Climatic Data Center.

Var.	Mean	SD	25%	Median	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
PMC	33.5	17.4	22.5	30.2	40.8
PMF	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), coarse PM (PMC) in (μg/m^3), fine PM (PMF) in (μg/m^3).

Daily readings (averaged from 24 hourly readings) were obtained of PM_{10} and $PM_{2.5}$ values measured by a Tapered Element Oscillating Microbalance (TEOM) monitor. (There are also some days with PM_1 readings — limit 1 micron — but there are too many missing values to make it possible to incorporate this variable into the analysis.) The data set also contains the results of an x-ray spectrometry analysis of $PM_{2.5}$ collected on a Teflon filter using a dual fine particle sequential sampler (DFPSS) machine. The spectrometry gives the measurements of 44 selected elements ranging in atomic number from sodium (NA) to lead (PB).

Mortality data are similar in format to those used for Philadelphia, and were obtained from the Arizona Health Services Department. We restricted attention to nonaccidental deaths in the city of Phoenix in the 65+ age group, though breakdowns into age groups, cause of death, etc., are available.

Table 3 gives selected summary statistics for this data set.

4.2 Time-trend and meteorological variables

As with Philadelphia, the basic regression analysis involves a linear model of the form (1) with y_t defined as the square root of daily death counts. Time trends were again modeled through a B-spline basis representation, using 18 knots (one for each two months of data).

The meteorological variables considered were daily temperature maxima and minima, and specific humidity. The selected model contained the following meteorological variables, in addition to the 18 B-spline terms for trend: $TMAX_2$, $TMIN_2$, $THIGH_1$, SH_3, SH_4 and $SHSQ_4$. Here TMAX and TMIN are daily maximum and minimum tempera-

i	j	PMC	t	PMF	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: PMC and PMF 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.

ture; THIGH=$(\text{TMAX}-30)_+$; SH is specific humidity; SHSQ=SH^2; and suffices denote lags, as previously.

4.3 Particulate matter effects

We now add various PM variables, one at a time, to the time trend and meteorology model. We considered both coarse PM or PMC, defined as the difference between PM_{10} and $PM_{2.5}$, and fine PM or PMF, the same as $PM_{2.5}$. For each of the two variables, we considered various exposure measures indexed by i and j where i is the number of days averaged and j is the lag. For example, $i = 3,\ j = 1$ is the three-day average lagged one day, or in other words, the average of lags 1, 2 and 3. In this case, adopting a different convention from Philadelphia, the i-day average is recorded as missing only if all i days are missing; in other cases, we average over available lags. Each of these 24 PM variables was added to the model in turn, and we computed both the coefficient and the t statistic using ordinary least squares. Results are in Table 4.

For PMC, several of the values are statistically significant — among the single-day ($i = 1$) values, the two-day lagged value is particularly significant, and among the coefficients based on averaged values, any average which includes the two-day lag is significant. On the other hand, for PMF, none of the values is statistically significant.

It is widely believed that PMF is more damaging to human health than PMC; see for example Schwartz *et al.* (1996). The results given here suggest that if there is any effect, it is more likely to be associated with PMC.

4.4 Influence of individual elements

A test has also been made for the effects of mortality corresponding to each of 42 separate elements which are constituents of $PM_{2.5}$. (Two elements were omitted because they were not available for most of the days.) Individual variables were standardized to have mean 1 as described in section 2.3. This analysis is based on about 300 days' data and the meteorological part of the model was refitted to account for the reduced length of the available series. Three analyses of the elemental variables were performed: (a) least squares analysis introducing one variable at a time, (b) least squares analysis including all variables together, (c) a ridge regression approach. In this case an *ad hoc* approach was taken to the ridge constant c since the type II MLE was infinite (recall discussion of this point in section 2.5). The three sets of parameter values are illustrated in Fig. 3. For this analysis, it can be seen that the least squares analysis based on all variables together produces occasional wild estimates (e.g. those for SI, CA), but the ridge estimates are much more stable. The analysis does not highlight any particular elements as having a strong effect, but this is a small data set and more studies of this nature are needed.

5. Conclusions

The problem of handling multiple pollution variables, whether they be different lags of the same pollutant, different constituents of particulate matter, or co-pollutants such as SO_2 and ozone, is a fundamental one in any epidemiological study of the relationship between air pollution and human health. We have argued that, like much other work in this field, an empirical Bayes approach offers a suitable statistical framework for dealing with these problems.

Our re-analysis of the Philadelphia data confirms that although there is general evidence of an association between air pollution and mortality, it is extremely difficult on the basis of this data set to pin it down to a specific variable. This conclusion is not substantially different from that of Samet *et al.* (1997), but the attempt to re-examine the question using empirical Bayes analysis has not clarified the situation, and if anything has only reinforced the uncertainty of the conclusions.

The Phoenix data set is new, and of interest because it allows for a direct comparison between the effects of PMC and PMF. The evidence presented here is that if there is any effect at all, it is due to the coarse and not the fine particles, though having

emphasized the tenuousness of conclusions about a causal effect in other data sets, it would be remiss of us not to point out that similar caution is appropriate in the interpretation of this one, given the short length of the data set.

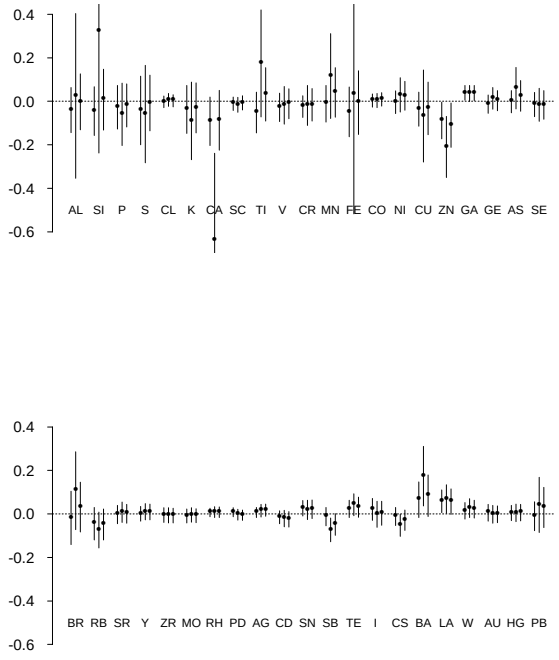

Fig. 3. Estimates and 95% confidence bands obtained for each of 42 elements. For each element, three estimates are shown. Left-hand: OLS estimate when variables are added one at a time to the model. Middle: OLS estimate when variables are added all together. Right-hand side: Ridge estimates.

Finally, our analysis of the 42 constituent elements of $PM_{2.5}$ in Phoenix has failed to yield any evidence that any single element has a significant effect, though this was based on an even shorter data set, and obviously there are many more possibilities for the study of this question.

6. References

Brown, P.J. (1993), *Measurement, Regression and Calibration.* Oxford University Press.

Carlin, B.P. and Louis, T.A. (1996), *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman and Hall, London.

Dominici, F., Samet, J.M., Xu, J. and Zeger, S.L. (1999a), Combining evidence on air pollution and daily mortality from the largest 20 US cities: a hierarchical modeling strategy. *Applied Statistics*, to appear.

Dominici, F., Zeger, S.L. and Samet, J.M. (1999b), A measurement error correction model for time-series studies of air pollution and mortality. Preprint, Johns Hopkins University.

National Research Council (1998), *Research Priorities for Airborne Particulate Matter I: Immediate Priorities and a Long-Range Research Portfolio.* National Academy Press, Washington, D.C.

Samet, J.M., Zeger, S.L. and Berhane, K. (1995), The Association of Mortality and Particulate Air Pollution. In *Particulate Air Pollution and Daily Mortality: Replication and Validation of Selected Studies. The Phase I Report of the Particle Epidemiology Evaluation Project.* Health Effects Institute, Cambridge MA, pp. 1–104.

Samet, J.M., Zeger, S.L., Kelsall, J.E., Xu, J. and Kalkstein, L.S. (1997), Air Pollution, Weather and Mortality in Philadelphia, 1973–1988. In *Particulate Air Pollution and Daily Mortality: Analyses of the Effects of Weather and Multiple Air Pollutants. The Phase IB Report of the Particle Epidemiology Evaluation Project.* Health Effects Institute, Cambridge MA, pp. 1–29.

Schwartz, J., Dockery, D.W. and Neas, L.M. (1996), Is daily mortality associated specifically with fine particles? *J. Air and Waste Manage. Assoc.* **46**, 927–939.

Shen, W. and Louis, T.A. (1998), Triple-goal estimates in two-stage hierarchical models. *J.R. Statist. Soc. B* **60**, 455–471.

Smith, R.L., Davis, J.M. and Speckman, P. (1998), Airborne particles and mortality. Chapter 6 of *Case Studies in Environmental Statistics*, edited by D. Nychka, W. Piegorsch and L.H. Cox. Springer Lecture Notes in Statistics, number 132, Springer Verlag, New York, pp. 91–120.

Smith, R.L., Davis, J.M. and Speckman, P. (1999a), Human health effects of environmental pollution in the atmosphere. Chapter 6 of *Statistics in the Environment 4: Statistical Aspects of Health and the Environment*, edited by V. Barnett, A. Stein and F. Turkman. John Wiley, Chichester, 91-115.

Smith, R.L., Davis, J.M. and Speckman, P. (1999b), Assessing the human health risk of atmospheric particles (with discussion). In *Environmental Statistics: Analysing Data for Environmental Policy.* Novartis Foundation Symposium 220. John Wiley, Chichester, 59–79.

Zeger, S.L., Dominici, F. and Samet, J.M. (1999), Harvesting-resistant estimates of air pollution and mortality. *Epidemiology* **10 (1)**, 171–175.