RESEARCH REPORT

HEALTH EFFECTS INSTITUTE

Number 175
November 2013

New Statistical Approaches to Semiparametric Regression with Application to Air Pollution Research

James M. Robins, Peng Zhang, Rajeev Ayyagari, Roger Logan, Eric Tchetgen Tchetgen, Lingling Li, Thomas Lumley, and Aad van der Vaart

New Statistical Approaches to Semiparametric Regression with Application to Air Pollution Research

James M. Robins, Peng Zhang, Rajeev Ayyagari, Roger Logan,
Eric Tchetgen Tchetgen, Lingling Li, Thomas Lumley, and Aad van der Vaart

with a Critique by the HEI Health Review Committee

Research Report 175
Health Effects Institute
Boston, Massachusetts

Trusted Science • Cleaner Air • Better Health

Publishing history: This document was posted at *www.healtheffects.org* in November 2013.

Citation for document:

Robins JM, Zhang P, Ayyagari R, Logan R, Tchetgen ET, Li L, Lumley T, van der Vaart A. 2013. New Statistical Approaches to Semiparametric Regression with Application to Air Pollution Research. Research Report 175. Health Effects Institute, Boston, MA.

Cover paper: made with at least 55% recycled content, of which at least 30% is post-consumer waste; free of acid and elemental chlorine. Text paper: made with 100% post-consumer waste recycled content; acid free; no chlorine used in processing. The book is printed with soy-based inks and is of permanent archival quality.

CONTENTS

ABOUT HEI

The Health Effects Institute is a nonprofit corporation chartered in 1980 as an independent research organization to provide high-quality, impartial, and relevant science on the effects of air pollution on health. To accomplish its mission, the institute

- Identifies the highest-priority areas for health effects research;
- Competitively funds and oversees research projects;
- Provides intensive independent review of HEI-supported studies and related research;
- Integrates HEI's research results with those of other institutions into broader evaluations; and
- Communicates the results of HEI's research and analyses to public and private decision makers.

HEI typically receives half of its core funds from the U.S. Environmental Protection Agency and half from the worldwide motor vehicle industry. Frequently, other public and private organizations in the United States and around the world also support major projects or research programs. HEI has funded more than 330 research projects in North America, Europe, Asia, and Latin America, the results of which have informed decisions regarding carbon monoxide, air toxics, nitrogen oxides, diesel exhaust, ozone, particulate matter, and other pollutants. These results have appeared in more than 260 comprehensive reports published by HEI as well as in over 1000 articles in the peer-reviewed literature.

HEI's independent Board of Directors consists of leaders in science and policy who are committed to fostering the public–private partnership that is central to the organization. The Health Research Committee solicits input from HEI sponsors and other stakeholders and works with scientific staff to develop a Five-Year Strategic Plan, select research projects for funding, and oversee their conduct. The Health Review Committee, which has no role in selecting or overseeing studies, works with staff to evaluate and interpret the results of funded studies and related research.

All project results and accompanying comments by the Health Review Committee are widely disseminated through HEI's Web site (*www.healtheffects.org*), printed reports, newsletters and other publications, annual conferences, and presentations to legislative bodies and public agencies.

ABOUT THIS REPORT

Research Report 175, *New Statistical Approaches to Semiparametric Regression with Application to Air Pollution Research,* presents a research project funded by the Health Effects Institute and conducted by Dr. James M. Robins of the Harvard School of Public Health, Boston, Massachusetts, and his colleagues. This report contains four main sections.

The HEI Statement, prepared by staff at HEI, is a brief, nontechnical summary of the study and its findings; it also briefly describes the Health Review Committee's comments on the study.

The Investigators' Report, prepared by Dr. Robins and colleagues, describes the scientific background, aims, methods, results, and conclusions of the study.

The Critique, prepared by members of the Health Review Committee with the assistance of HEI staff, places the study in a broader scientific context, points out its strengths and limitations, and discusses remaining uncertainties and implications of the study's findings for public health and future research.

An Invited Editorial, prepared by Dr. Sander Greenland, presents a perspective on the research from an expert scientist who is immersed in this particular area of statistical methods development and is familiar with research into air pollution and public health outcomes.

This report has gone through HEI's rigorous review process. When an HEI-funded study is completed, the investigators submit a draft final report presenting the background and results of the study. This draft report is first examined by outside technical reviewers and a biostatistician. The report and the reviewers' comments are then evaluated by members of the Health Review Committee, an independent panel of distinguished scientists who have no involvement in selecting or overseeing HEI studies. During the review process, the investigators have an opportunity to exchange comments with the Review Committee and, as necessary, to revise their report. The Critique reflects the information provided in the final version of the report.

HEI STATEMENT

Synopsis of Research Report 175

A Semiparametric Regression Approach for Air Pollution Research

BACKGROUND

The findings of a number of epidemiologic studies of air pollution and health have played a central role in setting air quality limits aimed at protecting public health. Since the mid-1990s, HEI has sponsored original research in this area, as well as research and review activities focused on the analytic methods used in such studies. These efforts include — among many others — the National Morbidity, Mortality, and Air Pollution Study (NMMAPS), the Reanalysis of the Harvard Six Cities Study and American Cancer Society Study of Particulate Air Pollution and Mortality, and the HEI Special Report on Revised Analyses of Time-Series Studies.

Time-series studies are commonly used to evaluate relationships between variations in short-term pollutant concentrations and acute human disease outcomes or mortality. Because time-series methods compare counts of disease events or deaths with pollutant concentrations on a specific day or other short time frame, the analyses do not need to account for subjects' smoking behavior or other risk factors that do not change from day to day. However, when evaluating the relationship between health outcomes and pollutant exposures, investigators do need to systematically adjust the data sets to control for time-dependent phenomena such as weather and seasonal trends that may influence observed disease patterns.

In 2003, HEI produced a Special Report on the Revised Analyses of Time-Series Studies of Air Pollution and Health after scientists discovered a problem with the commonly used S-Plus statistical software. The Special Report contained a number of wide-ranging recommendations for future time-series analyses, and specifically emphasized that the effect estimates derived from time-series data were shown to be sensitive to the statistical methods and parameters used to control for long-term time trends in the data.

Following publication of the Special Report, Dr. James Robins of the Harvard School of Public Health and his colleagues submitted a preliminary application to develop and apply statistical methods to address some of the issues raised by the report. They proposed to (1) develop methods that would improve the point estimates and confidence intervals for the parameters of a semiparametric regression model, (2) compare the new methods with standard methods in simulated studies, (3) develop efficient user-friendly software to implement the new methods, and (4) reanalyze critical data sets and compare the results with those from studies based on other methods.

What This Study Adds

- Robins and colleagues successfully developed semiparametric methods for epidemiologic investigations that are likely to produce risk estimates that are less biased than traditional Poisson time-series methods.
- When applied to the NMMAPS data set, the semiparametric methods produced estimates of the risk of health events relative to pollutant levels that were of similar magnitude to those obtained in HEI's Revised Analyses of Time-Series Studies, but with wider confidence intervals.
- Although the semiparametric methods are promising for future short-term studies of health events and air pollution levels, their utility and applicability could be enhanced by incorporating existing scientific understanding to control for confounding when relationships between covariates and mortality are well understood.

This Statement, prepared by the Health Effects Institute, summarizes a research project funded by HEI and conducted by Dr. James Robins at the Harvard School of Public Health, Boston, MA, and colleagues. Research Report 175 contains the detailed Investigators' Report, a Critique of the study prepared by the Institute's Health Review Committee, and an Invited Editorial by Dr. Sander Greenland.

EVALUATION

Reviews of the Investigators' Report, from committee members and selected peers, were divergent on the overall importance and utility of this work for epidemiologic analyses. The Committee commented that the research team had performed high-quality work to develop statistical methods that are complex and represent a very significant effort on their part, and that the results are technically sound. They noted that the concurrence between the current investigators' results and HEI's results from the Revised Analyses of Time-Series Studies was reassuring.

The current study, although acknowledged to be highly innovative by the broader scientific community, can be most easily understood by experts who are immersed in this particular area of statistics. Therefore the Committee invited Dr. Sander Greenland of the University of California–Los Angeles to write a short editorial to be published with the report. Dr. Greenland's editorial reflects his understanding of and views about the methods developed. His comments are provided to assist the reader in understanding and interpreting this report and its contributions to epidemiologic methods for air pollution study.

In his invited editorial, Dr. Greenland agreed with the Review Committee that the research team largely achieved their major goal of developing highly flexible semiparametric regression analysis methods and successfully applying them to the analysis of a large data set. He noted that the similarity of the results from the current analysis and those from earlier analyses of the NMMAPS data was to be expected; the earlier analyses had already employed relatively flexible methods for control of time-based confounding variables, and the data did not contain many problematic departures from linear behavior that would have been better detected and controlled by these innovative methods.

Dr. Greenland also identified some important limitations of the new methods as they apply to air pollution and health research. One of his primary scientific concerns was that the development of methods did not include the type of dose–response modeling of the exposure's effect that is desirable for exploring different ambient concentrations that might be considered for regulatory standards. He also noted that the investigators chose to test their methods without incorporating established scientific understanding of how trends in certain variables may bias a time-series study. For example, the research team directly incorporated temperature and humidity data in the models as covariates, instead of using them to adjust the mortality data according to consensus assumptions about the effects of weather on daily mortality. Because the relationships between weather variables and mortality outcomes have been well explored and are well understood, relaxing such assumptions may unnecessarily reduce the precision of the results.

Dr. Greenland was also concerned about potential residual confounding from poorly defined or poorly measured exposure or from confounding variables in the models, since the team did not investigate how the commonly encountered types of measurement error might affect the results of this semiparametric analysis. Dr. Greenland noted, however, that exploration of the impact of these limitations was either beyond the scope of the defined research project or was not undertaken due to time and funding limitations.

CONCLUSIONS

The Review Committee agreed with most of the points that Dr. Greenland made. First, even in the largest and best-conducted observational studies, errors in the measurement of pollutant concentrations and major potential confounding factors create uncertainty in the magnitude, if not the direction, of estimated effects. Second, prior expert knowledge, to the extent that it exists, can be a valuable tool to inform the design and the interpretation of research. These are points that apply in general to observational epidemiology and have been recurring themes in Dr. Greenland's work. The Committee noted, however, that environmental epidemiologists conduct research in a world of imperfect data, and that no analytic method will ever be able to perfectly adjust for the shortcomings of such observational information as that found in the NMMAPS data set and others that have been established for large-scale air pollution and health outcomes research.

Overall, the Review Committee found that the semiparametric methods developed in this study are a promising addition to current practices for short-term studies of health events and air pollution levels. Although these methods do not address all of the potentially important sources of bias or confounding, the Committee agreed with the investigators that this research could be particularly useful in investigations where the relationships between time-varying confounders and health outcomes are not clearly understood or are difficult to characterize. The Committee also agreed with Dr. Greenland's suggestion that the applicability of these new methods and the precision of the estimates of risk that they produce could be improved in practice through the use of current methods to adjust for time-varying confounders when relationships between covariates and mortality are well understood.

INVESTIGATORS' REPORT

New Statistical Approaches to Semiparametric Regression with Application to Air Pollution Research

James M. Robins, Peng Zhang, Rajeev Ayyagari, Roger Logan, Eric Tchetgen Tchetgen, Lingling Li, Thomas Lumley, and Aad van der Vaart

Department of Biostatistics (J.M.R, R.A, E.T.), and Department of Epidemiology (J.M.R., R.L., E.T.), Harvard University; Quantitative Methodology Program, University of Michigan (P.Z.); Department of Ambulatory Care and Prevention, Harvard Medical School (L.L.); Department of Biostatistics, University of Washington (T.L.); and Department of Mathematics, Vrije Universiteit, Amsterdam (A.vdV.)

1. OVERVIEW

The National Morbidity, Mortality, and Air Pollution Study (NMMAPS)[1] was an HEI-funded study of the effects of air pollution on mortality. Based on time-series data on particulate matter with an aerodynamic diameter $\leq 10\,\mu$m (PM_{10}) collected between 1987 and 1994, the NMMAPS investigators created a database including mortality, weather, and air pollution data for several US cities at least every 6 days (Peng et al., 2004). The NMMAPS investigators singled out the 20 largest cities for analysis and fit city-specific semiparametric loglinear time-series regression models to the data (Dominici et al., 2004; Samet, Dominici, et al., 2000; Samet, Zeger, et al., 2000).

The city-specific estimates were then combined via a random effects regression to obtain summary estimates (on a rate ratio scale) of the short-term effect of PM_{10} on mortality. These summary estimates have greatly influenced both scientific and policy debates. However, the magnitude of the rate ratio estimates are sufficiently small that they are at the limit of what can be reliably estimated from observational data. Thus it is critical to minimize confounding of the PM_{10} effect by predictors of daily mortality and PM_{10}, such as (calendar) time, recent temperature, and humidity. To that end, in their main city-specific analyses, the NMMAPS investigators fit a loglinear semiparametric time-series model. The model assumes that the natural logarithm of the mean number of deaths in a particular city on a particular day can be modelled as a linear function

This Investigators' Report is one part of Health Effects Institute Research Report 175, which also includes a Critique by the Health Review Committee, an Invited Editorial, and an HEI Statement about the research project. Correspondence concerning the Investigators' Report may be addressed to Professor James M. Robins, Harvard School of Public Health, 677 Huntington Avenue, HSPH1, Room 411, Boston, MA 02115; robins@hsph.harvard.edu

Although this document was produced with partial funding by the United States Environmental Protection Agency under Assistance Award CR–83234701 to the Health Effects Institute, it has not been subjected to the Agency's peer and administrative review and therefore may not necessarily reflect the views of the Agency, and no official endorsement by it should be inferred. The contents of this document also have not been reviewed by private party institutions, including those that support the Health Effects Institute; therefore, it may not reflect the views or policies of these parties, and no endorsement by them should be inferred.

[1] A list of abbreviations and other terms appears at the end of the Investigators' Report.

of the PM_{10} level on the previous day and smooth functions of time, recent temperature, and humidity. The smooth functions were modelled with various types of splines including natural splines and penalized splines. As all choices gave similar results, we shall restrict consideration to natural splines, which are piece-wise continuous polynomials. These smooth spline functions of time, temperature, and humidity are each associated with a free parameter, the number of degrees of freedom (df) that determines, for each function, how wiggly that function is allowed to be. Very wiggly functions are needed to control confounding due to very nonlinear covariate effects on mortality. Thus it is important to chose the appropriate degrees of freedom. However it is difficult to do so.

If (1) the true dose–response of temperature on mortality is a smooth function with many wiggles, and (2) PM_{10} is highly and nonlinearly correlated with temperature, then, were we to restrict the number of degrees of freedom in our temperature smooth, the nonlinear effects of temperature would be falsely ascribed to PM_{10}. Thus to decrease the potential for confounding bias, it is desirable to use many degrees of freedom for each potential confounding variable. However, when many degrees of freedom are used but not actually needed (because the true, but unknown, dose–response is not very wiggly), an inefficient estimate of and wide confidence interval for the PM_{10} effect will result. This loss of efficiency is of serious concern, because it may compromise the ability to detect a true but small PM_{10} effect. Furthermore, the empirical data often cannot determine the optimal trade-off between these conflicting needs. Current biological and meterological knowledge is also insufficient to determine the optimal trade-off.

To overcome these difficulties, at least in part, we implemented a new and theoretically better approach to semiparametric regression. This approach is based on a new theory that uses higher-order U-statistics[2] in place of the first-order statistics that are the basis for the current theory of and approach to fitting semiparametric regression models.

Our estimators offer better control of bias due to confounding by temperature and humidity in exchange for somewhat wider confidence intervals for the PM_{10} effect provided certain assumptions, discussed later, hold. Our goal was to estimate the effect of PM_{10} on mortality and to compare our estimates to those obtained by the NMMAPS investigators.

The report is organized as follows. Section 2.1 summarizes our main findings and describes the NMMAPS data and our methods. Section 2.2 describes our results and compares them with the results obtained by the NMMAPS investigators, and critically discusses the assumptions under which our results are valid. Section 2.3 explores the sensitivity of our results to our initial choice of estimators, our "base case". Section 2.4 contains a final discussion. Section 3 and the Appendices contain theoretical and mathematical results that formally justify our theoretical claims, as well as other technical material.

2. EMPIRICAL FINDINGS

2.1 SUMMARY OF SUBSTANTIVE CONCLUSIONS, DATA, AND METHODS

2.1.1 Substantive Conclusions Regarding PM_{10} and All-Cause Mortality

Reanalyses of the NMMAPS city-specific time-series data for the 22 largest NMMAPS cities (using a new approach based on higher-order influence functions) provide no evidence that the

[2] A definition section appears at the end of the Investigators' Report.

original NMMAPS estimates of the effect of PM_{10} on all-cause mortality were biased, except possibly in Minneapolis, and even there the estimated bias was small and did not change any substantive conclusions.

We obtained wider confidence intervals than did the NMMAPS investigators. This increase in confidence interval width was to be expected because our approach uses weaker assumptions than does the original NMMAPS approach concerning the smoothness of the dose–response function for the effect of temperature and humidity on daily mortality and on PM_{10}.

2.1.2 The Data

We used the same data file as the NMMAPS investigators used (Peng et al., 2004), which contains the following data.[3] For a given city,

- N is the number of days with both death and PM_{10} data available,
- Y_i is the number of deaths on the i-th such day,
- A_i is the PM_{10} level on the previous day (i.e., 1-day lag),
- D_i is the categorical day-of-week variable, and
- X_i is a vector with 4 continuous components $X_{\text{cont},i}$ and a trinary discrete variable $X_{5,i}$ of age categories.

The components of $X_{\text{cont},i}$ are

1. average temperature on day i,
2. dew-point temperature on day i,
3. adjusted 3-day lagged daily temperature, and
4. adjusted 3-day lagged dew-point temperature.

For certain cities in the NMMAPS data set, PM_{10} measurements were available only on every 6-th day. For those cities, the actual time t_i in days from the start, corresponding to observation i, is $6i$. For cities with daily PM_{10} measurements t_i is equal to i.

Throughout this report, capital letters denote random variables (i.e., variables that vary in a nondeterministic way from observation to observation) and the corresponding lowercase letters denote their possible values. The time variables i and t_i are deterministic and thus not random.

2.1.3 First-Order Estimators

The NMMAPS investigators fit city-specific loglinear semiparametric regression models

$$E_i[Y_i|A_i,X_i] = e^{\tau^* A_i + \zeta_i^*(X_i)},\ i = 1,\ldots,N. \tag{2.1}$$

The parameter τ^* is an unknown parameter encoding the loglinear effect of PM_{10} on mortality; $\zeta_i^*(\cdot)$ is an unknown function of X_i that itself can depend on the time t_i in days since the start of the study. The subscript i on $\zeta_i^*(\cdot)$ and E_i indicates the time dependence of the conditional distribution of Y_i given A_i, X_i.

Previous NMMAPS reports and journal articles authored by the NMMAPS investigators have provided comprehensive reviews of semiparametric regression models and have discussed their critical importance to adequate control of confounding by temperature, humidity, and time in time-series analyses of the effect of PM_{10} on mortality (Dominici et al., 2004; Peng et al., 2004;

[3] A notation section appears at the end of the Investigators' Report.

Samet, Dominici, et al., 2000; Samet, Zeger, et al., 2000). As discused in these publications, a useful approach to modeling the unrestricted function of X_i is, in practice, with natural splines.

To fit the models, we follow NMMAPS investigators and model $\zeta_i^*(X_i)$ using a 155-dimensional function $w_i(x)$ of t_i and x (Peng et al., 2004). The 155-dimensional variable $W_i = w_i(X_i)$ includes the categorical variables for age category and day-of-week and natural spline transformations of the time t_i and the continuous variables in X_i with

- 96 df for time, 6 df for each temperature variable, 6 df for each dew-point variable, and
- an additional 30 df for interactions between time and the age category variable.

We then refit the models without the 6-dimensional dummy variable day-of-week; in this analysis $W_i = w_i(X_i)$ was 149-dimensional. This is necessitated by the computational requirements of our higher-order influence-function estimators. As described later, the results obtained with and without adjusting for day-of-week are similar. We therefore included only 149 variables in W_i, eliminating the 6 dummy day-of-week variables.

Our notation differs slightly from earlier notation used by the NMMAPS investigators in that we represent dependence on time with the subscript i for occasion number (either every day or every 6 days) rather than as a component of the vector of other covariates X. This choice reflects the fact that time is not a random variable and, as discussed later, the ability of our second-order influence estimator to decrease bias differs depending on whether a covariate is or is not a random variable. We shall see that because time is not random, it is important to use a relatively large number of degrees of freedom to control confounding due to time.

We begin by estimating τ^* by solving an estimating equation $\widehat{\mathbb{IF}}_{1,\text{eff}}(\tau) = 0$ based on the first-order efficient influence function $\widehat{\mathbb{IF}}_{1,\text{eff}}(\tau)$ described below. This estimator $\hat{\tau}_{1,\text{eff}}$ is identical to the usual Poisson regression estimator used in previous NMMAPS analyses (see Section 3, Lemma 9).

Our reason for describing the usual Poisson regression estimator used in previous NMMAPS publications as the solution to an estimating equation is discussed further below. Technical mathematical details are presented in Section 3 and cross-referenced throughout

We also consider another first-order estimating equation, $\widehat{\mathbb{IF}}_{1,\text{new}}(\tau)$, which is similar to but differs from $\widehat{\mathbb{IF}}_{1,\text{eff}}(\tau)$ in that it is weighted by an estimate of the joint density of the 5-dimensional vector X; more precisely, by a linear transformation of X as mentioned before. For this reason, we refer to this as the density-weighted influence function, and refer to the corresponding estimator as the density-weighted first-order estimator. This estimator is discussed further in Section 3 and is precisely defined in Appendix A. This estimating equation is necessary in order to define a second-order estimating equation with the desired properties, as explained in detail in Sections 3.2.3 and 3.2.4. The corresponding estimator of τ^*, denoted by $\hat{\tau}_{1,\text{new}}$, is obtained by solving $\widehat{\mathbb{IF}}_{1,\text{new}}(\tau) = 0$.

Full and Split Estimators The contribution of the i-th observation to the estimating function $\widehat{\mathbb{IF}}_{1,\text{eff}}(\tau)$ or $\widehat{\mathbb{IF}}_{1,\text{new}}(\tau)$ depends on two nuisance functions that we estimate. For the base-case (i.e., loglinear) model, these are:

$$b_i^*(X_i) \equiv e^{\zeta_i^*(X_i)}, \text{ and}$$
$$p_i^*(X_i) \equiv \frac{E_i[A_i e^{\tau^* A_i} | X_i]}{E_i[e^{\tau^* A_i} | X_i]}.$$

The latter is a weighted mean of A_i given X_i, where the subscript i on the expectation operator indicates that the distribution of A given X may change with time and thus with i. The former equals $E_i[Y_i|A_i = 0, X_i]$ under the loglinear semiparametric model. Thus b_i^* is the expected number of deaths on occasion i if PM_{10} were absent. We define

$$\mathbb{IF}_{1,\text{eff}}(\tau) = \sum_i \left[Y_i - e^{\tau A} b_i^*(X_i) \right] \left[A - p_i^*(X_i) \right].$$

As mentioned above we also consider another estimating function, $\mathbb{IF}_{1,\text{new}}(\tau)$.

$\mathbb{IF}_{1,\text{new}}(\tau)$ differs from $\mathbb{IF}_{1,\text{eff}}(\tau)$ in that the contribution of each occasion i to $\mathbb{IF}_{1,\text{eff}}(\tau)$ is additionally multiplied by a weight $\hat{w}_i$ equal to an estimate of the joint density of X_i. Section 2.3 provides additional detail.

We consider two types of estimates of the nuisance functions:

1. an estimator that depends on all the data, and
2. an estimator that randomly splits the data into two equal parts, separately estimates both nuisance functions using half of the data, and then, when computing the contribution of an observation i in the first random half-sample to our influence functions $\mathbb{IF}_{1,\text{eff}}$ and $\mathbb{IF}_{1,\text{new}}$ (defined below), uses the estimates of the nuisance functions from the second half-sample, and vice versa, averaging the two estimates thus obtained. We indicate the two split-sample estimating functions that are averaged to give the split-sample estimator as split,(0) and split,(1). (See Appendix A for the precise formulae.)

Split-sample methods are required with higher-order influence functions because we do not assume that the nuisance functions are sufficiently smooth (i.e., differentiable) for "Donsker" conditions to hold (see Definitions section at end and van der Vaart and Wellner, 1996). When the nuisance functions are very smooth, the size (entropy) of the set of candidate nuisance functions is sufficiently small that one can estimate the nuisance functions without splitting the sample and still preserve properties (such as asymptotic normality) that would hold if the nuisance functions were known rather than estimated; in that case we say that the set of candidate nuisance functions is Donsker. When Donsker conditions hold, the correlations induced by estimating the nuisance functions from the full data can be ignored. Otherwise, these correlations must be eliminated by sample splitting.

(Because of the time-series nature of our data the two random samples of occasions are not statistically independent of one another; however, the dependence is local and thus the statistical properties of our split-sample estimator will be identical to those obtained from independent samples.)

Henceforth we use superscripts "full" and "split" to distinguish whether we used the full-data or split-sample method to estimate the nuisance functions and the parameter τ^*. The estimates $\hat{\tau}^{\text{full}}_{1,\text{eff}}$ and $\hat{\tau}^{\text{split}}_{1,\text{eff}}$ were generally within a standard error of each other (compare columns $\hat{\tau}^{\text{full}}_{1,\text{eff}}$ and $\hat{\tau}^{\text{split}}_{1,\text{eff}}$ later in Table 1); the estimates $\hat{\tau}^{\text{full}}_{1,\text{new}}$ and $\hat{\tau}^{\text{split}}_{1,\text{new}}$ are provided in Figure 1 (see later Section 2.2).

In Lemma 9 of Chapter 3, we show that $\hat{\tau}^{\text{full}}_{1,\text{eff}}$ is identical to the usual Poisson regression estimator used in previous NMMAPS analyses. This estimator $\hat{\tau}^{\text{full}}_{1,\text{eff}}$, based on solving $\widehat{\mathbb{IF}}^{\text{full}}_{1,\text{eff}}(\tau) = 0$, only depends on the estimate of $b_i^*(X_i) \equiv e^{g_i^*(X_i)}$ and not on $p_i^*(X_i)$. This is so because the contribution

from the estimated $p_i^*(X_i)$ cancels out in the estimating equation. However, the estimated $p_i^*(X_i)$ does not cancel when we use $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{eff}}(\tau)$, and thus $\hat{\tau}^{\text{split}}_{1,\text{eff}}$ does depend on $p_i^*(X_i)$.

Second-Order Estimators If the nuisance functions are very wiggly (i.e., highly nonlinear and rough), they cannot be well estimated. As a result the estimators $\hat{\tau}^{\text{full}}_{1,\text{eff}}$, $\hat{\tau}^{\text{split}}_{1,\text{eff}}$, $\hat{\tau}^{\text{full}}_{1,\text{new}}$, and $\hat{\tau}^{\text{split}}_{1,\text{new}}$ may all yield biased estimates of the PM_{10} effect, because the estimated nuisance functions are sufficiently poor estimates of the true nuisance functions such that confounding by temperature, humidity, and time is not well-controlled.

For now, suppose the nuisance functions $b_i^*(X_i)$ and $p_i^*(X_i)$ are not too wiggly (i.e., smooth enough) so that our estimates of the nuisance functions $b_i^*(X_i)$ and $p_i^*(X_i)$ are sufficiently accurate that the bias of the first-order estimators are less than their standard error.

Then the estimators $\hat{\tau}^{\text{full}}_{1,\text{eff}}$ and $\hat{\tau}^{\text{split}}_{1,\text{eff}}$ based on the estimating equations $\widehat{\mathbb{IF}}^{\text{full}}_{1,\text{eff}}(\tau)$ and $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{eff}}(\tau)$ can be semiparametric efficient and thus optimal under our semiparametric regression model. Further, 95% Wald confidence intervals (CIs) are valid and thus cover τ^* 95% of the time.

Now suppose the nuisance functions $b_i^*(X_i)$ and $p_i^*(X_i)$ are very wiggly so that our estimates of them are inaccurate. Then

1. if the errors in the estimation of $b_i^*(X_i)$ and $p_i^*(X_i)$ are correlated, we have uncontrolled confounding due to our weather variables;
2. the bias of any first-order estimator will be larger than its standard error; and
3. these estimators will not optimally trade off bias and variance; and further 95% Wald CIs will cover the true effect τ^* less than 95% of the time.

In this case, we can hope to improve on the first-order estimators by using higher-order influence functions.

Although higher-order IF estimators are defined for all integer orders ≥ 2, we report only second-order IF function estimators because of computational constraints. Specifically, as discussed below, our second-order estimator, referred to as $\hat{\tau}_2^{\text{split},(k)}$, effectively subtracts from $\hat{\tau}^{\text{split}}_{1,\text{new}}$ a U-statistic estimate $\widehat{\mathbb{IF}}_{22}^{\text{split},(k)}(\tau)$ of the bias of the first-order estimator $\hat{\tau}^{\text{split}}_{1,\text{new}}$. As a consequence, under the assumptions described below, the second-order estimator will be less biased than any first-order estimator [although it will have greater variance because of the additonal variance contributed by $\widehat{\mathbb{IF}}_{22}^{\text{split},(k)}(\tau)$]. As discussed below, a fundamental assumption required for the bias of the second-order estimator $\hat{\tau}_2^{\text{split},(k)}$ to be less than the bias of $\hat{\tau}^{\text{split}}_{1,\text{new}}$ is that we obtain an accurate estimate of the joint density of the 4-dimensional variable X_i (encoding two temperature and two humidity summary variables). In fact unless the difference between the estimated joint density and the true joint density is small, $\hat{\tau}_2^{\text{split},(k)}$ can have greater bias than $\hat{\tau}^{\text{split}}_{1,\text{new}}$. It follows that since time (encoded by the subscript i) is not random and thus does not have a density, $\hat{\tau}_2^{\text{split},(k)}$ cannot decrease the component of the bias in $\hat{\tau}^{\text{split}}_{1,\text{new}}$ attributable to residual confounding by the main effect of time.

Second-order IF estimators of τ^* are obtained as the solution to an estimating equation

$$\widehat{\mathbb{IF}}_2^{\text{split},(k)}(\tau) = 0,$$

where the estimating equation

$$\widehat{\mathbb{IF}}_2^{\text{split},(k)}(\tau) = \widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}(\tau) + \widehat{\mathbb{IF}}_{22}^{\text{split},(k)}(\tau)$$

is the sum of two components:

1. the density-weighted first-order influence function $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}(\tau)$ described above, and
2. a second-order U-statistic $\widehat{\mathbb{IF}}^{\text{split},(k)}_{22}(\tau)$ that depends on a positive integer parameter k, specified by the analyst. A second-order U-statistic is a sum of functions $u(O_i, O_j)$ of data from two different observations O_i and O_j.

Associated with each k is a different second-order IF estimator $\hat{\tau}^{\text{split},(k)}_2$ obtained by solving $\widehat{\mathbb{IF}}^{\text{split},(k)}_2(\tau) = 0$. The second-order estimators only have the desired statistical properties if the nuisance parameters are estimated using the split-sample method.

Dependence on k As k increases, the variance of $\hat{\tau}^{\text{split},(k)}_2$ increases, but the bias of $\hat{\tau}^{\text{split},(k)}_2$ generally decreases. Under certain assumptions (described informally below and formally in Sections 3.3.2 and 3.3.3) for sufficiently large k, $\hat{\tau}^{\text{split},(k)}_2$ will have a smaller bias than the usual Poisson regression estimator $\hat{\tau}^{\text{full}}_{1,\text{eff}}$.

For many different values of k beginning with $k = 1$ and ending at some $k_{\max}$, we report $\hat{\tau}^{\text{split},(k)}_2$ for 22 cities. $k_{\max}$ was generally chosen sufficiently large that either the variance of $\hat{\tau}^{\text{split},(k_{\max})}_2$ was at least 4 times the variance of $\hat{\tau}^{\text{split},(1)}_2$ (i.e., the standard error at least doubled) or we ran out of computer memory. Often we took $k_{\max} \approx 839{,}000$.

2.2 RESULTS

Our results are organized as follows. First-order estimates for 22 cities, including the 20 cities considered by the NMMAPS investigators, are shown in Table 1. This table presents results using our final models for the nuisance functions b^*_i and p^*_i; we also include $\hat{\tau}^{\text{full}}_{1,\text{eff}}(155)$, the standard Poisson regression estimate obtained using the 155 covariates used in previous NMMAPS analyses. The remainder of the results, including the second-order estimators, tests of the bias, and variance plots for the second-order estimators, are presented in the graphs in Figure 1. The analysis was done separately for each city.

Left Graphs

- On the left graphs, to the right of the vertical line at zero are the estimates $\hat{\tau}^{\text{split},(k)}_2$ and associated 95% Wald CIs for many values of k. In the figure we have suppressed the hat (ˆ) for all estimators and the split label for τ_2.
- Note the k scale on the x-axis is nonlinear to enhance viewing.
- To the left of zero, we provide estimates and associated nominal 95% CIs corresponding to the first-order estimators $\hat{\tau}^{\text{full}}_{1,\text{eff}}$ (■), $\hat{\tau}^{\text{full}}_{1,\text{new}}$ (▲), $\hat{\tau}^{\text{split}}_{1,\text{eff}}$ (□), and $\hat{\tau}^{\text{split}}_{1,\text{new}}$ (△) based on the estimating functions $\widehat{\mathbb{IF}}^{\text{full}}_{1,\text{eff}}(\tau)$, $\widehat{\mathbb{IF}}^{\text{full}}_{1,\text{new}}(\tau)$, $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{eff}}(\tau)$, and $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}(\tau)$. These equations do not depend on k and thus provide a graphical representation of some of the results from Table 1.
- For example, $\hat{\tau}^{\text{split}}_{1,\text{new}}$ is the estimator solving $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}(\tau) = 0$.
- All these estimators estimate the same quantity τ^* if the semiparametric regression model is true.
- The first-order estimates are generally close and within a standard error of one another for each of the 22 cities.
- The estimates based on $\widehat{\mathbb{IF}}^{\text{full}}_{1,\text{new}}(\tau)$ are less efficient than (have wider CIs than) those based on $\widehat{\mathbb{IF}}^{\text{full}}_{1,\text{eff}}(\tau)$, as expected, although the difference in efficiency is not large in general.

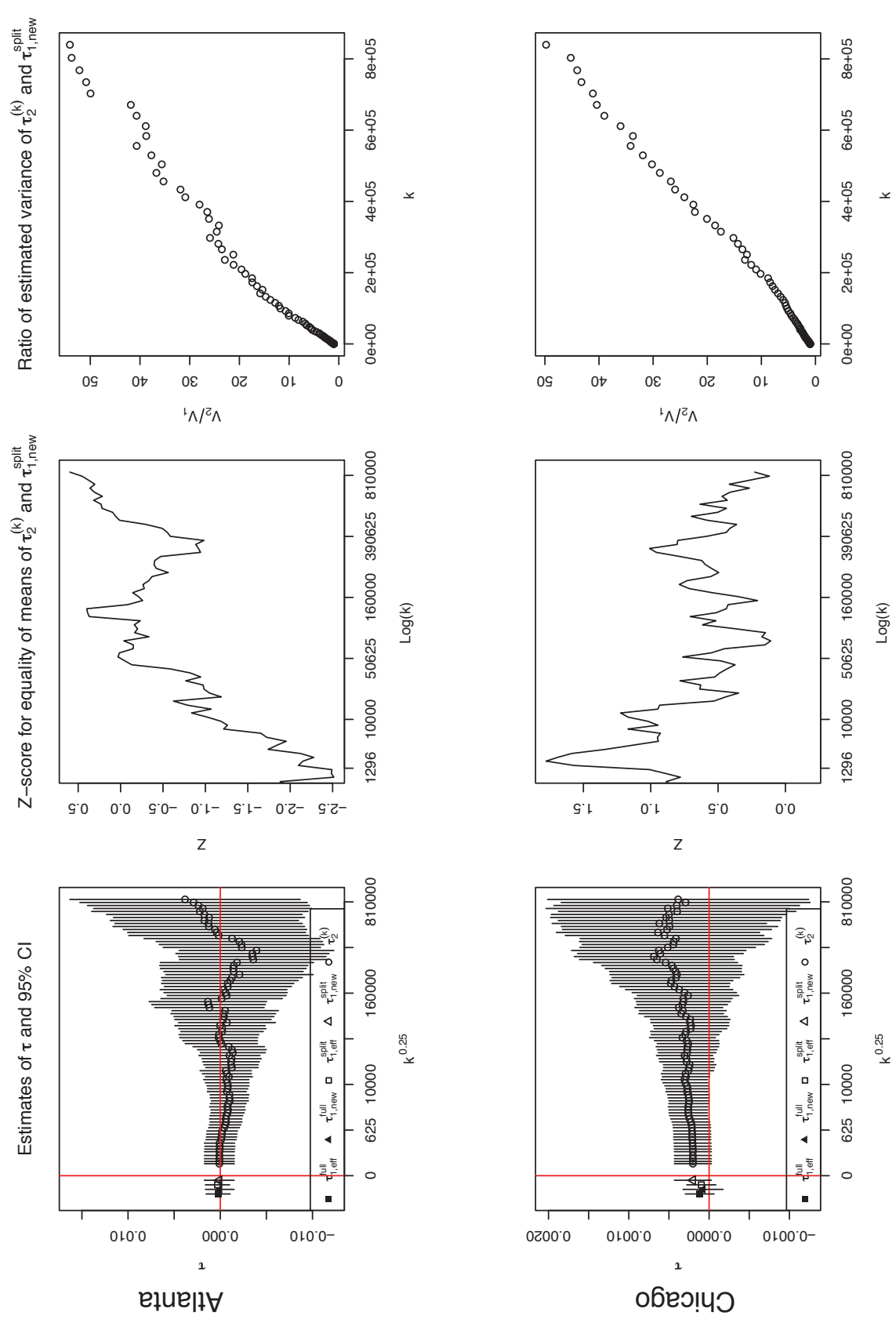

Figure 1. GLM results for 22 cities. Summary of the analysis using a loglinear model with Legendre polynomials, a density cutoff at the 80th percentile, observations between 25 and 75 days of a given day, and with k between 3 and 839523. Note that y axis scales vary among cities.

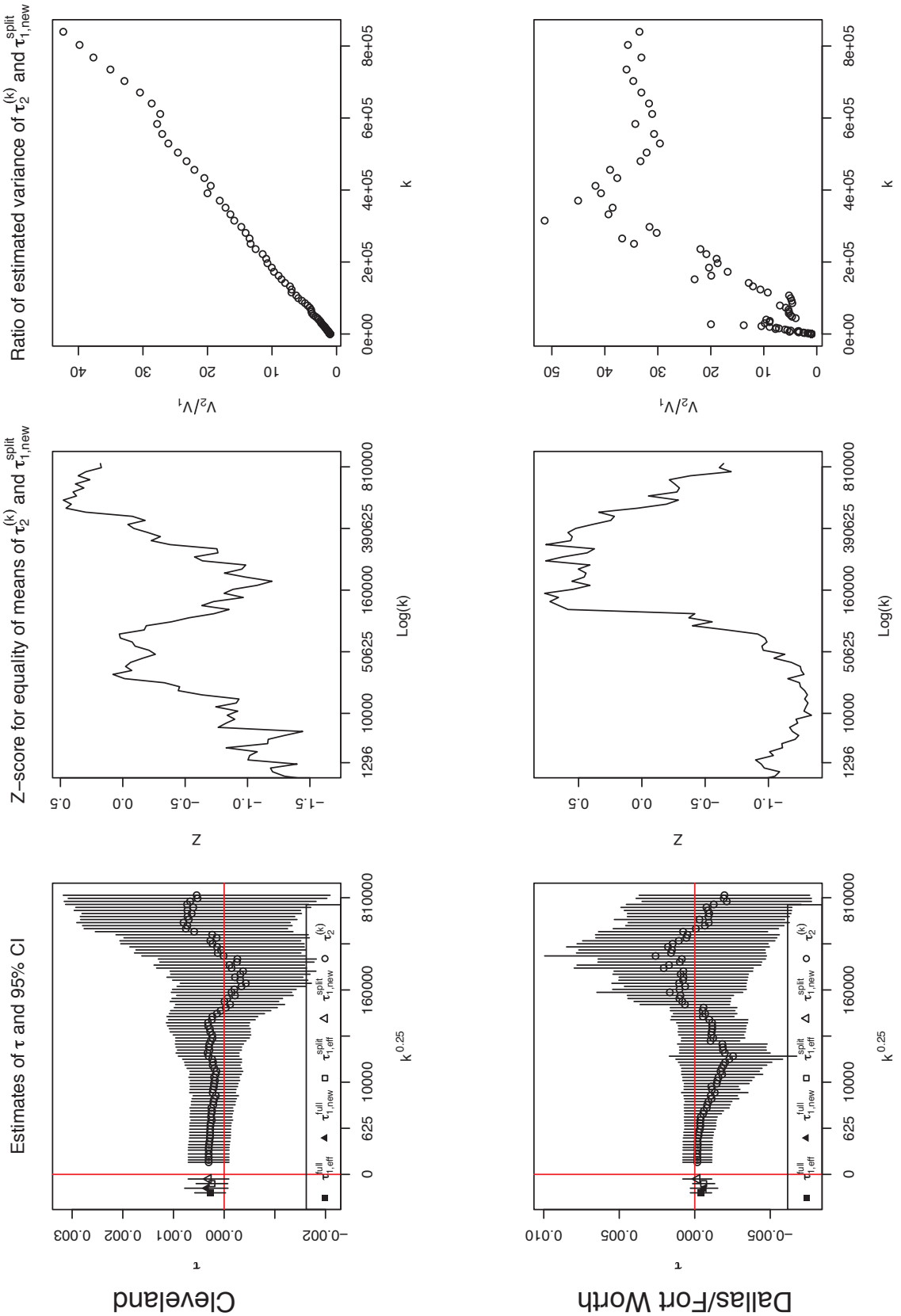

Figure 1. (continued)

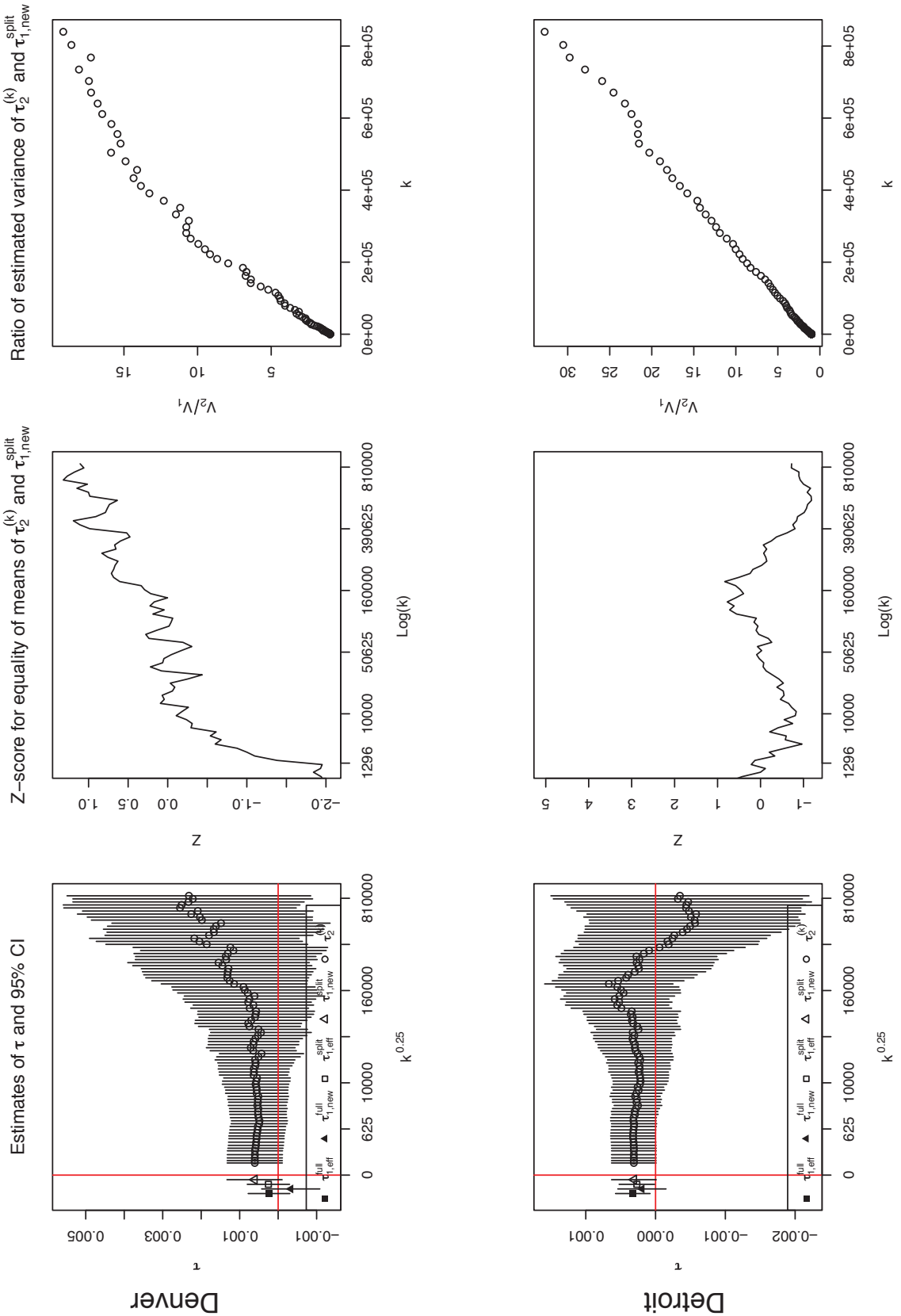

Figure 1. (continued)

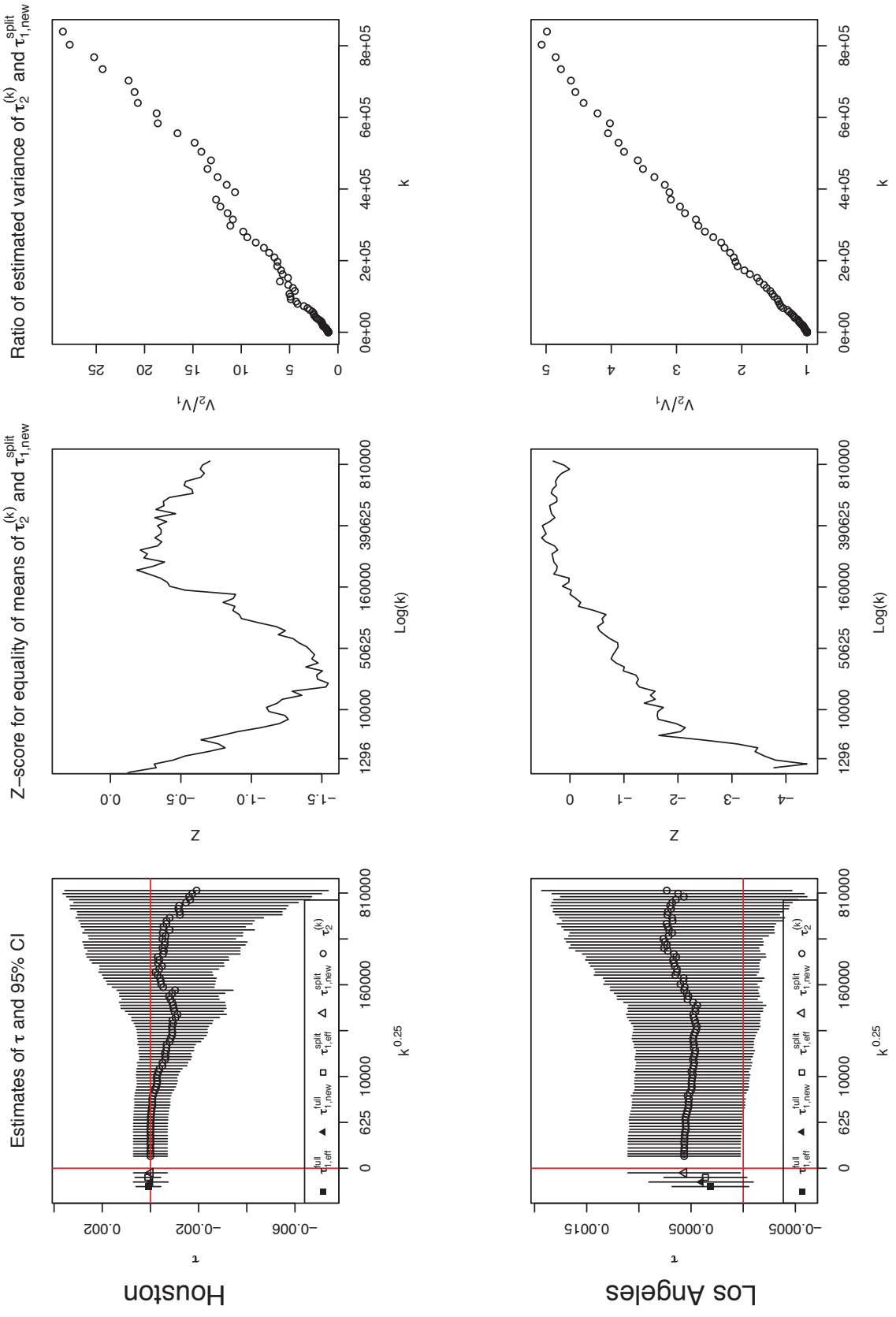

Figure 1. (continued)

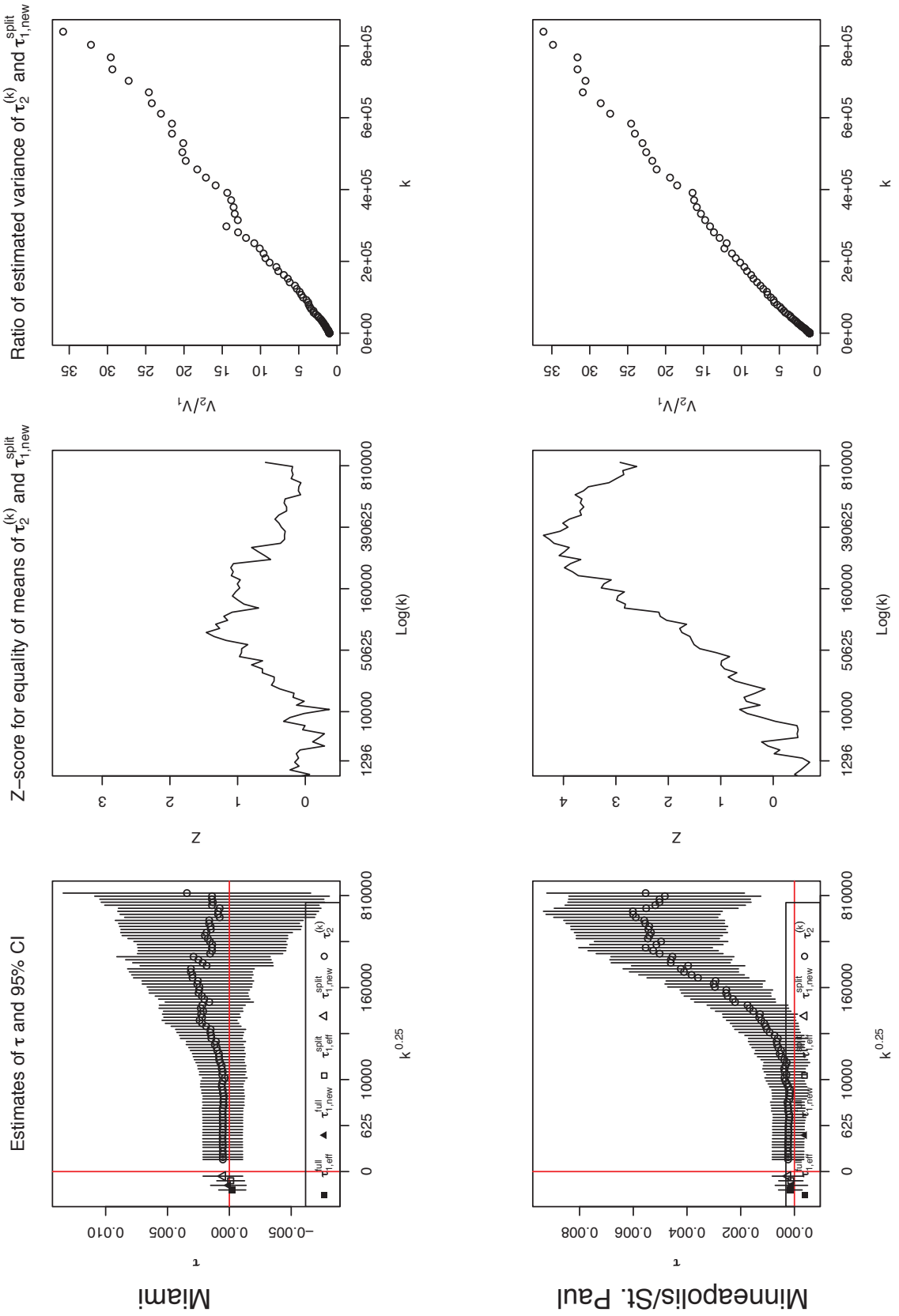

Figure 1. (continued)

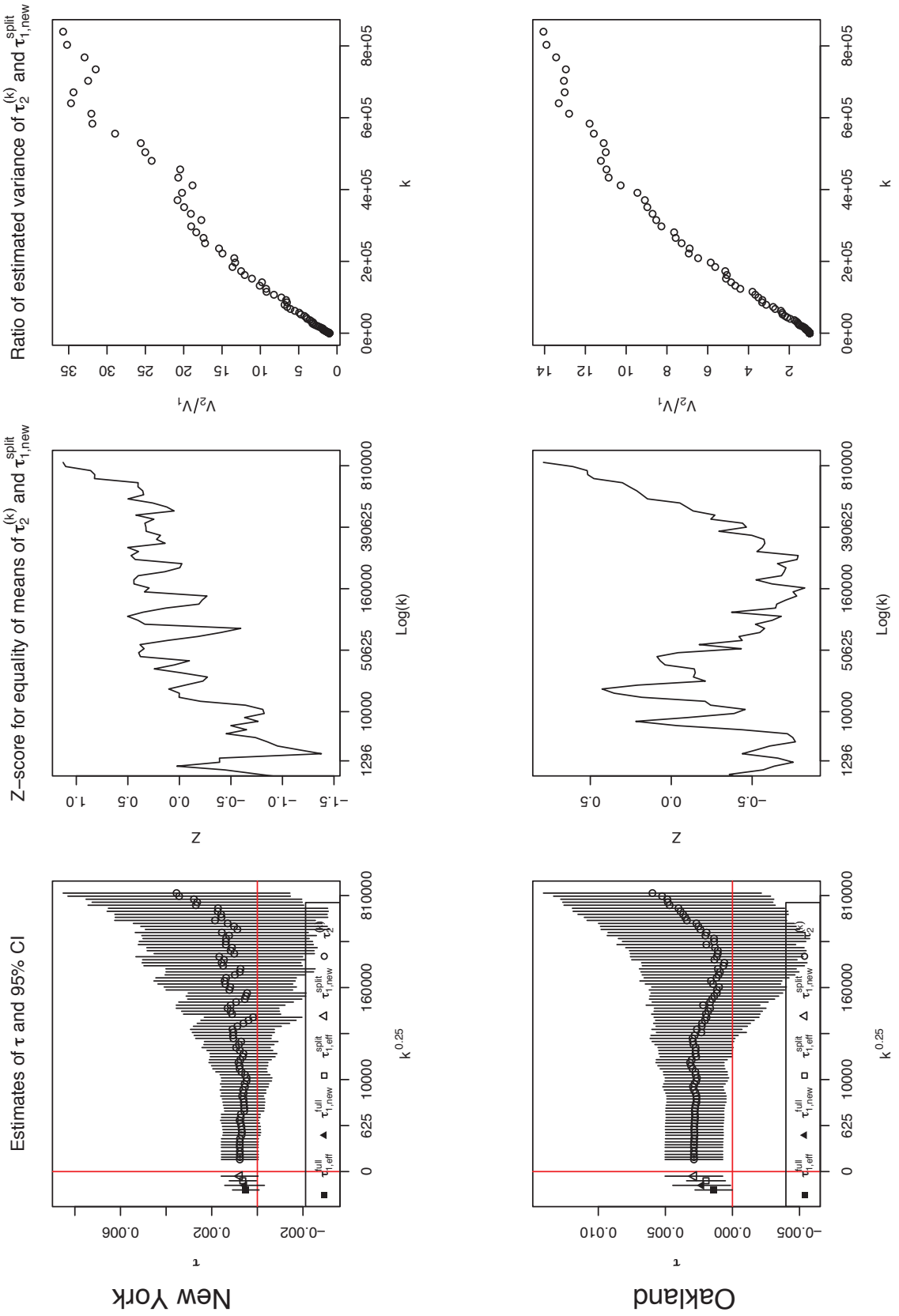

Figure 1. (continued)

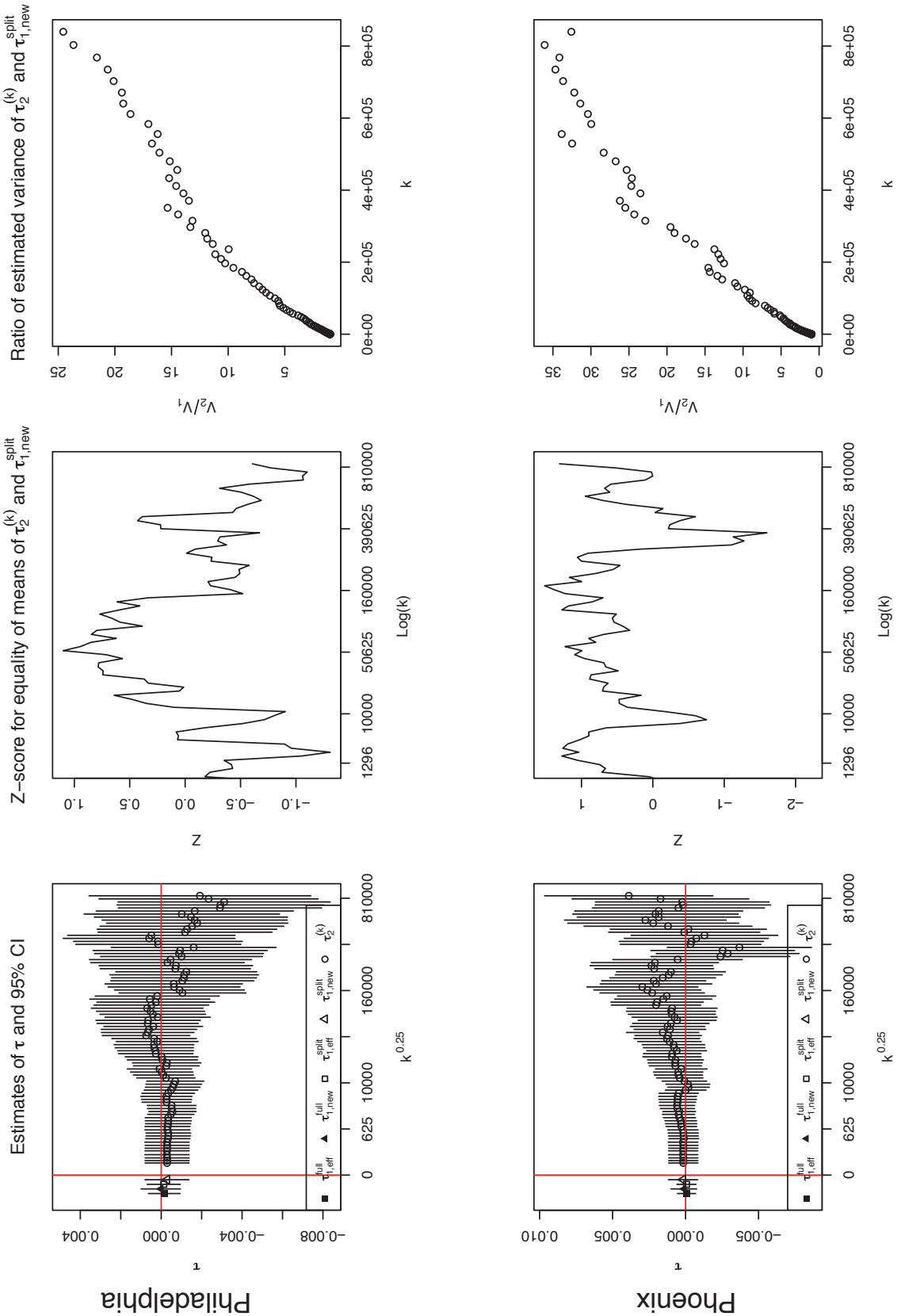

Figure 1. (continued)

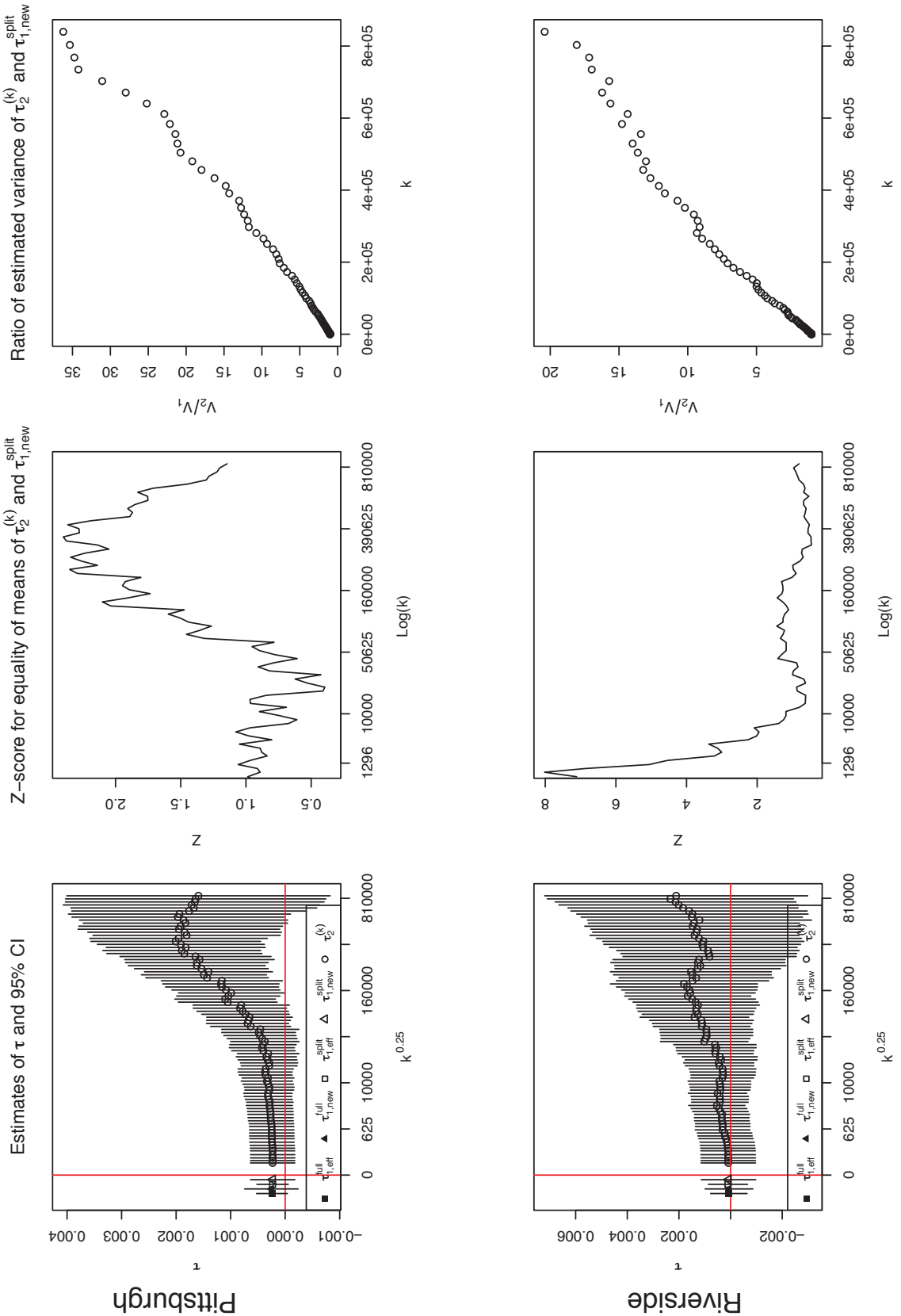

Figure 1. (continued)

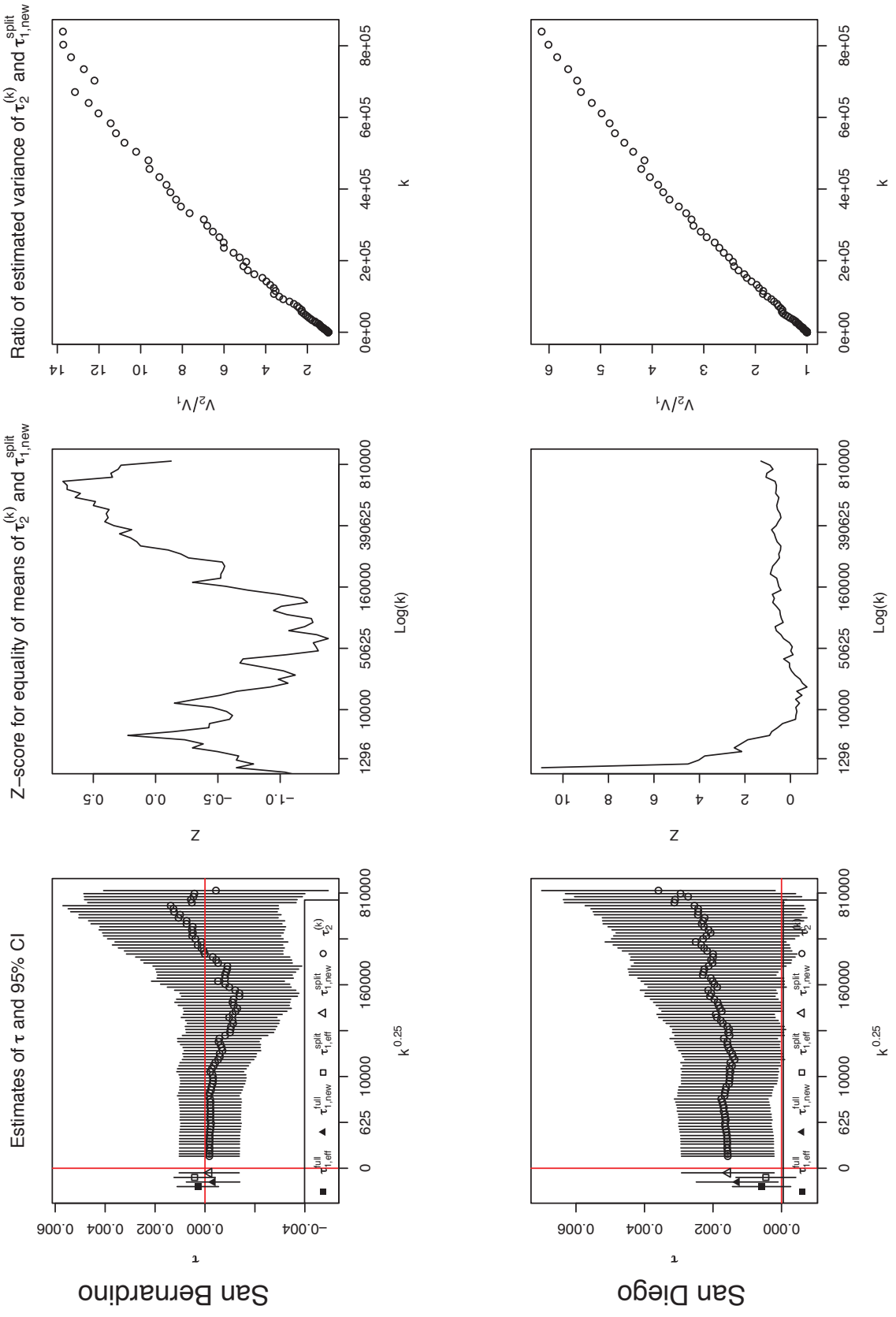

Figure 1. (continued)

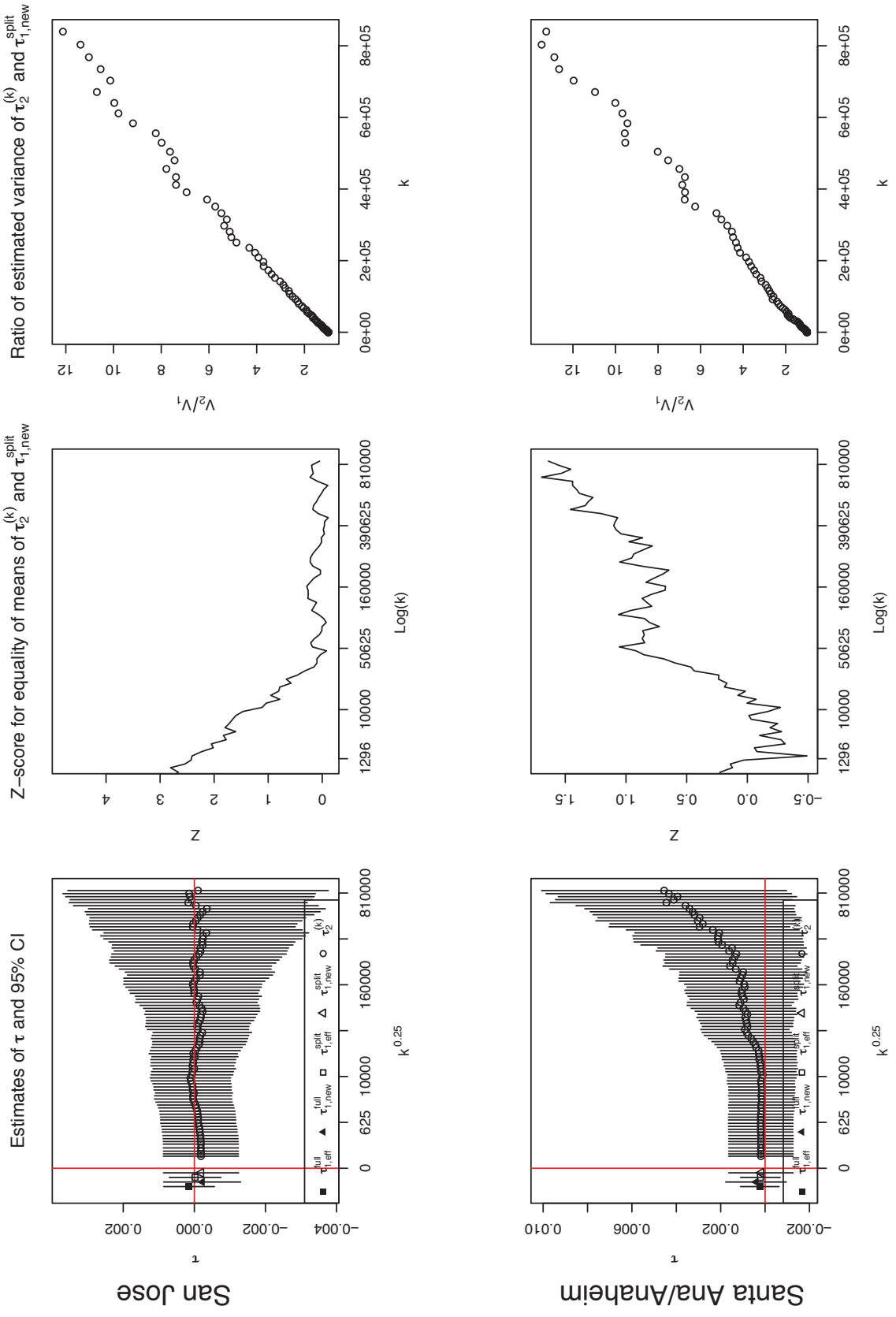

Figure 1. (continued)

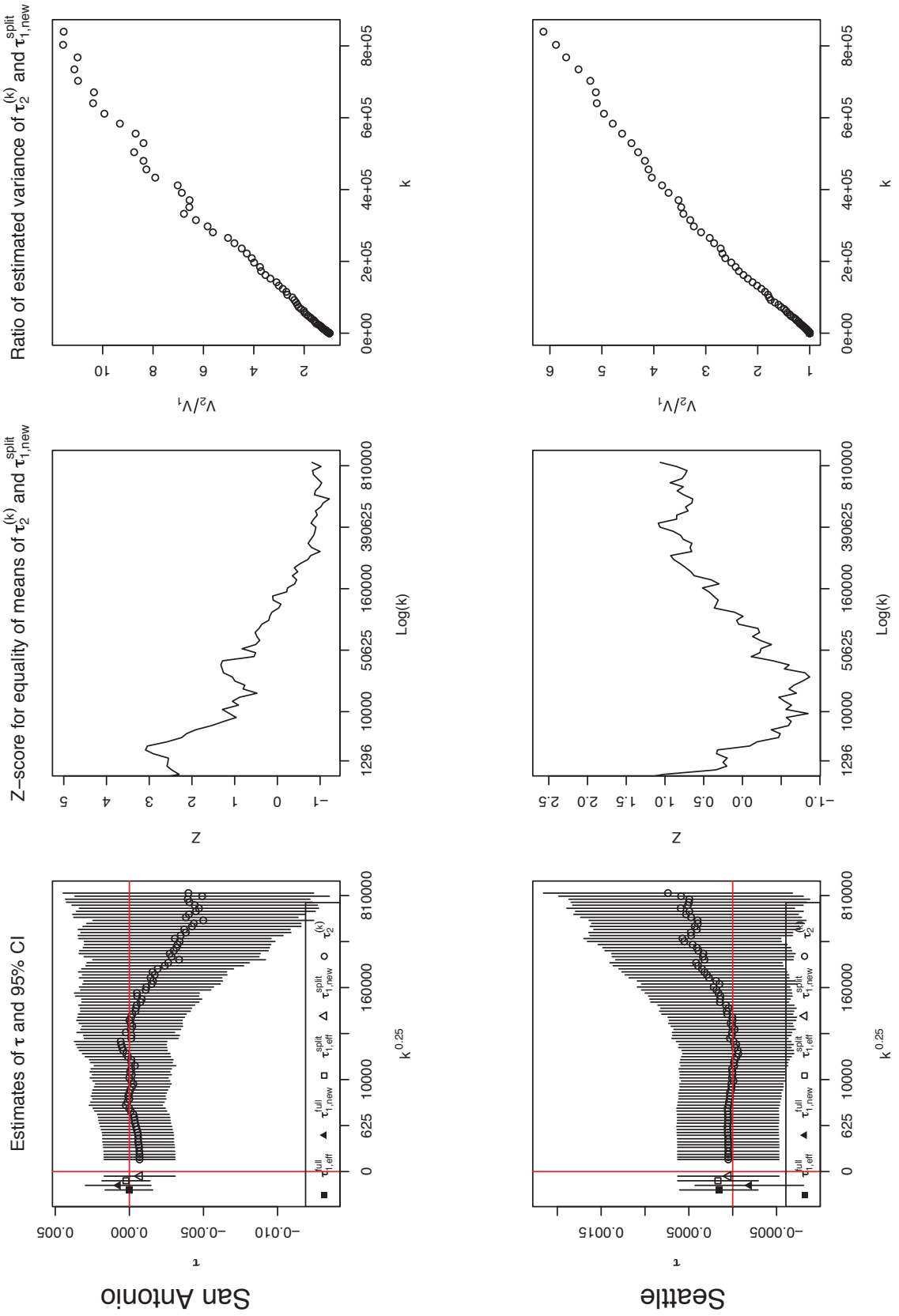

Figure 1. (continued)

Table 1. First-order estimators for 22 cities. SE estimates are in parentheses. $\hat{\tau}^{\text{full}}_{1,\text{eff}}(155)$ is the standard Poisson regression estimate obtained using the 155 covariates used by the NMMAPS investigators; the other estimators are based on 149 covariates as described in Appendix B.

City	Linear		Loglinear		
	$\hat{\tau}^{\text{full}}_{1,\text{eff}}(149)$	$\hat{\tau}^{\text{split}}_{1,\text{eff}}(149)$	$\hat{\tau}^{\text{full}}_{1,\text{eff}}(149)$	$\hat{\tau}^{\text{full}}_{1,\text{eff}}(155)$	$\hat{\tau}^{\text{split}}_{1,\text{eff}}(149)$
Atlanta	0.00169 (0.00514)	0.00185 (0.00505)	0.00023 (0.00067)	0.00082 (0.0007)	0.00029 (0.00066)
Chicago	0.00457 (0.00356)	0.00338 (0.00351)	0.00012 (0.00009)	0.00014 (0.00009)	0.0001 (0.00009)
Cleveland	0.00353 (0.00199)	0.0031 (0.002)	0.00028 (0.00015)	0.00026 (0.00016)	0.00024 (0.00016)
Dallas	−0.00718 (0.00633)	−0.0106 (0.00644)	−0.00041 (0.00037)	−0.00041 (0.00038)	−0.00059 (0.00037)
Denver	0.00161 (0.00192)	0.00163 (0.00191)	0.00024 (0.00028)	0.00027 (0.00029)	0.00025 (0.00027)
Detroit	0.00531 (0.00203)	0.00441 (0.00203)	0.00033 (0.00013)	0.00034 (0.00013)	0.00027 (0.00013)
Houston	0.00097 (0.00353)	0.00126 (0.00361)	0.00008 (0.00026)	0.0002 (0.00027)	0.00011 (0.00027)
Los Angeles	0.01719 (0.01067)	0.02258 (0.01103)	0.00031 (0.00018)	0.00037 (0.00019)	0.00036 (0.0002)
Miami	−0.00348 (0.00834)	−0.00191 (0.0083)	−0.00023 (0.00061)	−0.00012 (0.00064)	−0.00011 (0.00056)
Minneapolis	0.00148 (0.00221)	0.00122 (0.00221)	0.00015 (0.00023)	0.00026 (0.00024)	0.00013 (0.00023)
New York	0.03335 (0.02026)	0.03985 (0.02018)	0.00051 (0.00027)	0.00065 (0.00028)	0.00064 (0.0003)
Oakland	0.0112 (0.00582)	0.01653 (0.0057)	0.0014 (0.00073)	0.00125 (0.00074)	0.00197 (0.00072)
Philadelphia	−0.00214 (0.00581)	−0.00113 (0.00588)	−0.00015 (0.00042)	−0.00009 (0.00043)	−0.00013 (0.00041)
Phoenix	−0.00076 (0.00523)	−0.0003 (0.00535)	−0.00007 (0.00036)	−0.00023 (0.00036)	−0.00006 (0.00033)
Pittsburgh	0.00321 (0.00198)	0.00304 (0.00196)	0.00024 (0.00014)	0.00024 (0.00014)	0.00023 (0.00014)
Riverside	0.00075 (0.00278)	0.001 (0.00285)	0.00007 (0.00038)	0.00003 (0.00039)	0.0004 (0.00038)
San Antonio	0.00015 (0.00561)	0.0013 (0.00557)	0.00003 (0.00079)	0.0001 (0.00008)	0.00023 (0.00081)
San Bernardino	0.002 (0.00308)	0.00306 (0.00302)	0.00029 (0.00042)	0.00029 (0.00043)	0.00042 (0.00041)
San Diego	0.00895 (0.00696)	0.00553 (0.00691)	0.00058 (0.00044)	0.00057 (0.00044)	0.00046 (0.00044)
San Jose	0.00113 (0.00274)	0.00004 (0.00271)	0.00015 (0.00038)	0.00016 (0.00038)	−0.00002 (0.00036)
Santa Ana	0.00263 (0.00563)	0.00207 (0.00562)	0.00024 (0.00046)	0.00017 (0.00047)	0.00021 (0.00045)
Seattle	0.00148 (0.0021)	0.00178 (0.0021)	0.00016 (0.00023)	0.0001 (0.00024)	0.00017 (0.00023)

Middle Graphs The middle graphs plot, for many values of k, the Z-score associated with a two-sided test of the hypothesis that the mean of the estimator $\hat{\tau}^{\text{split}}_{1,\text{new}}$ and the mean of $\hat{\tau}^{\text{split},(k)}_{2}$ are equal and equal to the true parameter τ^*. Specifically, our test rejects the hypothesis if the absolute value of the Z-score $\left\{\hat{\tau}^{\text{split}}_{1,\text{new}} - \hat{\tau}^{\text{split},(k)}_{2}\right\}/\widehat{\text{se}}_k$ is sufficiently large (where $\widehat{\text{se}}_k$ is an estimator of the standard error of the numerator). A significant difference is indicative of biasedness of the first-order estimator as explained in Section 3.3.4.

Right Graphs The right graphs plot the ratio V_2/V_1 of the variance of the second-order estimator $\hat{\tau}^{\text{split},(k)}_{2}$ of τ^* to that of the first-order estimator $\hat{\tau}^{\text{split}}_{1,\text{new}}$ as a function of k. Theory suggests this ratio should increase approximately linearly with k (see Lemma 41 and Section 3.4.1), which is borne out for most cities.

2.2.1 Interpretation

If, for some k, the p-value associated with the Z-score on the middle graph implies the means $\hat{\tau}^{\text{split},(k)}_{2}$ and $\hat{\tau}^{\text{split}}_{1,\text{new}}$ differ significantly from zero (after adjusting the p-value for multiple comparisons), then, under the assumptions referred to above, we can conclude that all the first-order estimators, including the usual Poisson regression estimator, are likely biased and the associated nominal 95% Wald CIs centered on these estimators will cover the true effect τ^* less than 95% of the time.

On the other hand if, for all k, the p-value associated with the Z-score on the middle graph fails to reject the hypothesis that the means of $\hat{\tau}^{\text{split},(k)}_{2}$ and $\hat{\tau}^{\text{split}}_{1,\text{new}}$ are equal, then

1. it is likely that either

   a) the nuisance functions $b^*_i(X_i)$ and $p^*_i(X_i)$ were not very wiggly, or

b) the nuisance functions were wiggly but the nuisance functions $b_i^*(X_i)$ and $p_i^*(X_i)$ were not highly correlated, which implies that the magnitude of confounding by our weather variables is negligible;

2. we should use the original NMMAPS estimate as our estimate if our goal is to minimize mean squared error (MSE);
3. in contrast, the choice of the CI is not an empirical one: the larger k is, under more true states of nature (defined by both the wiggliness of the nuisance functions and their correlation), the more the k-specific nominal 95% Wald CI around $\hat{\tau}_2^{\text{split},(k)}$ will cover the true effect τ^*;
4. to choose which k should be used for setting CIs would require a priori knowledge of an upper bound on the wiggliness and correlation of $b_i^*(X_i)$ and $p_i^*(X_i)$. A priori knowledge is required because estimating the wiggliness of a function (e.g., the number of derivatives it has) is an ill-posed problem and it cannot be empirically estimated from the data. As a consequence we have simply displayed (in Figure 1) the CI for many different k. An interval is only valid (i.e., has actual coverage of at least 95%) under the untestable assumption that the bias of the estimator is less than its standard error.

Why Do (3) and (4) Not Contradict (1) and (2)? Our test of bias of the first-order estimators as shown on the middle graph does not have large power to detect small amounts of bias. But even a small amount of bias in a first-order estimator will lead to undercoverage by a CI centered on that estimator if the bias is the same order as the standard error. Even so, the MSE of the first-order estimator will be good, even when the bias slightly exceeds the standard error, because it has the smallest variance among our candidates.

Estimates of the variance of $\widehat{\mathbb{IF}}_2^{\text{split},(k)}(\tau)$ for $k < 10,000$ are not reliable due to numerical instability and for other theoretical reasons described in Section 3.3.2. Thus a significant Z-value on the middle graph for $k < 10,000$ is not meaningful if the Z-values become non-significant for larger k.

2.2.2 Our Assumptions and Their Consequences

Bias of the Second-Order Estimates The second-order estimates $\hat{\tau}_2^{\text{split},(k)}$ for large k are less biased than the NMMAPS estimates based on the usual Poisson regression estimator $\hat{\tau}_{1,\text{eff}}^{\text{full}}$ under certain assumptions. We now describe some of these assumptions and conditions under which they might be violated.

Need for Accurate Density Estimation $\widehat{\mathbb{IF}}_{1,\text{new}}^{\text{split}}(\tau)$ and thus $\hat{\tau}_2^{\text{split},(k)}$ depends on estimates $\hat{f}_i(x)$ of the multivariate density of X on day t_i. The bias of $\hat{\tau}_2^{\text{split},(k)}$ depends on the quality of these estimates.

To construct these estimates we used a nonparanormal density estimator smoothed over time. This is a density estimator that estimates each of the marginal densities of the 4 continuous components of X_i nonparametrically using kernel density estimation. The dependence structure of X_i is then estimated based on a parameteric model with 6 parameters. (See Section 2.3.4 for a formal definition.)

We also considered more nonparametric density estimators, based on the program Locfit by Loader (2010); however, these more nonparametric estimators performed no better than the nonparanormal density estimators. (See Section 2.3.4.)

For large k, the bias of $\hat{\tau}_2^{\text{split},(k)}$ will be less than the bias of Poisson regression $\hat{\tau}_{1,\text{eff}}^{\text{full}}$ if the error $\hat{f}_i(x) - f_i(x)$ in estimating the density is small and if the time effects are accurately modelled.

Specifically, the bias of Poisson regression depends on the average over the observations of the product of the error $\left[\hat{b}_i(X_i) - b_i^*(X_i)\right]$ in estimating $b_i^*(X_i)$ and the error $[\hat{p}_i(X_i) - p_i^*(X_i)]$ in estimating $p_i^*(X_i)$. Because the bias is the product of two error terms, we say that the bias is second order.

Further we say the Poisson regression estimator and our other first-order estimators are doubly robust because if we knew either $b_i^*(\cdot)$ or $p_i^*(\cdot)$ they would be the solutions to unbiased estimating equations.

In contrast, the bias of $\hat{\tau}_2^{\text{split},(k)}$ for large k depends on the average of the product of $\hat{f}_i(X_i) - f_i(X_i)$ with $\left[\hat{b}_i(X_i) - b_i^*(X_i)\right] \times [\hat{p}_i(X_i) - p_i^*(X_i)]$ when the time effect is accurately modelled. This bias (which holds even when k is large) is referred to as estimation bias. Thus $\hat{\tau}_2^{\text{split},(k)}$ would be triply robust (ignoring the time-related bias discussed next).

This bias, which we call third-order bias, will be less than the second-order bias if $\hat{f}_i(x) - f_i(x)$ is small. However, we cannot be certain that $\hat{f}_i(x) - f_i(x)$ is small because of the difficulty in estimating the joint density of a vector of random variables.

Need for Accurate Estimation of Time Effects The bias of $\hat{\tau}_2^{\text{split},(k)}$ for large k includes a time-series bias term due to the fact that the nuisance functions $b_i^*(x)$ and $p_i^*(x)$ can depend on the i and thus on time. This bias cannot be eliminated using higher-order influence functions because the time interval between successive observations is fixed (and thus deterministic) and not random. We refer to this term as the time-series bias because it exists only for time-series data. Thus the performance of higher-order influence functions in time-series settings will not be as good as in other settings. The key to keeping this term small (and hopefully smaller than the third-order term mentioned above) is to accurately estimate the dependence of $b_i^*(x)$ and $p_i^*(x)$ on time by using many degrees of freedom for time.

Possible Need for a Large Value of k $\widehat{\mathbb{IF}}_{22}^{\text{split},(k)}$ is a sum of k second-order U-statistics. Specifically, the joint contribution of occasions i and j is

$$\widehat{\text{IF}}_{22,ij}^{(k)} = -\left\{Y_i - \exp(\tau A_i)\,\hat{b}_i(X_i)\right\} K_k(X_i, X_j) \frac{\exp(\tau A_j)}{\hat{E}_j\left[\exp(\tau A_j)\,|X\right]} \left\{A_j - \hat{p}_j(X_j)\right\}$$

where the X_i have been rescaled to the unit cube in $\mathbb{R}^4$ using one of the transforms mentioned in Section 2.3.3,

$$K_k(X_i, X_j) = \sum_{l=1}^{k} \varphi_l(X_i)\varphi_l(X_j),$$

and $\varphi_l(\cdot)$, $l = 1, \ldots$, is an orthogonal basis for L_2 with respect to the Lebesgue measure on the unit cube in $\mathbb{R}^4$. The bias of $\hat{\tau}_{1,\text{new}}^{\text{split}}$, like the bias of Poisson regression, depends on the average over the observations of the product of the error $\hat{b}_i(X_i) - b_i^*(X_i)$ and the error $\hat{p}_i(X_i) - p_i^*(X_i)$. The mean of the sum over i and j of the $\text{IF}_{22,ij}^{(k)}$ cancels the second-order term in the bias of $\widehat{\mathbb{IF}}_{1,\text{eff}}^{\text{split}}$ more and more precisely as k increases, so that the final bias of $\widehat{\mathbb{IF}}_2^{\text{split},(k)}$ becomes third order for large k (provided the time effect is accurately modelled so that bias does not dominate).

Optimal Choice of k The variance of $\text{IF}_{22,ij}^{(k)}(\tau)$ increases approximately linearly with k. When k is small, there is an additional contribution to the bias, referred to as the truncation bias (which stems from the highest frequency components of the nuisance functions) over and above the

estimation bias referred to previously. As k increases, the truncation bias decreases although the estimation bias remains nearly the same. The optimal choice of k balances the decrease in truncation bias with the increase in variance as k increases.

The magnitude of the second-order bias term of $\hat{\tau}^{\text{split}}_{1,\text{new}}$ depends on both the wiggliness of and correlation between $b^*_i(X_i)$ and $p^*_i(X_i)$. If the magnitude is small, a small value of k can be used to control the contribution of the truncation bias to the second-order bias and thus the variance will also be small. However, since the amount of wiggliness and correlation are unknown and not directly estimable, the optimal choice of k is unknown. We can, to a limited extent, try to estimate the optimal k from the data.

The test described earlier of the equality of the means of $\hat{\tau}^{\text{split},(k)}_2$ and $\hat{\tau}^{\text{split}}_{1,\text{new}}$ can also be used to determine whether a particular value of k is too small, although not with high power. Specifically, if $\hat{\tau}^{\text{split},(k)}_2$ at some k is significantly different from $\hat{\tau}^{\text{split},(k)}_2$ at a different value of k, the lower value of k has a truncation bias that is larger than its variance. However, as a consequence of low power, we cannot exactly determine the optimal k (except possibly asymptotically as the number of observations tends to ∞). As discussed below, in all cases except Minneapolis, we had no statistical evidence that any value of k was too small. Thus, $k = 0$ is our best estimate of the optimal k. That is, our optimal choice is to simply use a first-order estimator because the insignificant Z-values imply we have no empirical evidence of significant truncation or estimation bias; among the first-order estimators, the usual Poisson regression estimator $\hat{\tau}^{\text{full}}_{1,\text{eff}}$ is the best as it should have the smallest variance.

Effect of Sample Splitting As noted above, our higher-order influence-function estimators require that we split the sample. Thus, before examining the performance of our higher-order influence-function estimator, we wished to determine what the effect of splitting the sample would be on standard first-order inference. (In Table 1 we provide, for each of the 22 cities, $\hat{\tau}^{\text{full}}_{1,\text{eff}}$ and $\hat{\tau}^{\text{split}}_{1,\text{eff}}$ for both the linear and loglinear case.)

If sample splitting resulted in a serious loss of efficiency, one would worry about the finite sample properties of any approach that uses sample splitting to estimate the nuisance functions, even including approaches, such as our higher-order influence-function approach, that have desirable asymptotic properties. However, for each of the 22 cities, the estimators $\hat{\tau}^{\text{full}}_{1,\text{eff}}$ and $\hat{\tau}^{\text{split}}_{1,\text{eff}}$ have similar estimated standard errors. Further the difference between the two estimators is always less than 1.5 estimated standard errors and usually much less than 1 standard error, suggesting that splitting the sample is rather innocuous with regard to inference based on first-order influence functions.

However, one needs to be concerned that the split-sample estimator of the standard error could be too small, since it is computed under the assumption that the two half-samples are statistically independent, which, although true under an asymptotic theory, is not true in finite samples. To rule out this possibility, we conducted a simulation study with a data generating process closely mimicking the distribution of the New York data. Details of the study are provided in Appendix C. In the simulation study, the Monte Carlo standard errors of the split-sample and full-sample estimators were very close to each other and to our analytic estimate of the standard error. Thus we conclude that the variance inflation associated with sample splitting was small.

Validity of Variance Estimators In Section 3.3.2 we give the assumptions under which our estimators of the variance of $\hat{\tau}_2^{\text{split},(k)}$ are valid. We essentially assume that the correlation between the data at different observations i and j is negligible unless the time between observations i and j is quite small. If this assumption is violated then our variance estimator will underestimate the true variance and thus the Z-values in the middle graph are inflated. As discussed below, in all cities but Minneapolis, the observed Z-values were sufficiently small that we failed to reject the null hypothesis of equality of the means of $\hat{\tau}_2^{\text{split},(k)}$ and $\hat{\tau}_{1,new}^{\text{split}}$ for any k. This conclusion would not change even if the assumption of "negligible covariance" between observations was false, since then the true Z-values should be even smaller. However, our conclusion that the null should be rejected for Minneapolis might no longer be appropriate.

2.2.3 Overall Results

For 21 of the 22 largest NMMAPS cities, our tests for a bias of the first-order estimators did not reject the hypothesis of unbiasedness. As a consequence, for these 21 cities, we obtain no evidence that the original NMMAPS estimates of the effect of PM_{10} on all-cause mortality were biased.

2.2.4 Results for Minneapolis

The only significant result based on the test reported in the middle graph is for Minneapolis. Specifically, the maximum Z-score of 4.2 for Minneapolis is significant even after Bonferroni correcting for the number of k we tested against.

Thus under our assumptions we have some evidence that the NMMAPS estimate for Minneapolis is biased. According to asymptotic theory we should not use the NMMAPS estimate as our estimate if our goal is to minimize MSE.

The question then of which $\hat{\tau}_2^{\text{split},(k)}$ we should use as our estimate has answers in certain asymptotic settings but does not have a clear answer in our finite sample setting (see Section 3.3.4 or pages 216–220 of Wasserman, 2006, for further discussion).

Even the claim of bias in the NMMAPS estimate for Minneapolis has two caveats:

1. The Z-score of 4.2 is not quite significant when we apply Bonferroni adjustment for the fact that this was the only one of 22 cities for which we found a significant effect.
2. Our test is only valid if the assumptions discussed in Section 2.2.2 above hold.

2.3 DEPENDENCE OF $\hat{\tau}_{1,\text{new}}^{\text{split}}$ AND $\hat{\tau}_2^{\text{split},(k)}$ ON USER-SPECIFIED SETTINGS

Recall that our second-order influence-function estimator $\hat{\tau}_2^{\text{split},(k)}$ is obtained as the solution to a second-order influence-function estimating equation $\widehat{\mathbb{IF}}_2^{\text{split},(k)}(\tau) = 0$, where $\widehat{\mathbb{IF}}_2^{\text{split},(k)}(\tau) = \widehat{\mathbb{IF}}_{1,\text{new}}^{\text{split}}(\tau) + \widehat{\mathbb{IF}}_{22}^{\text{split},(k)}(\tau)$. $\widehat{\mathbb{IF}}_{1,\text{new}}^{\text{split}}(\tau)$ and $\widehat{\mathbb{IF}}_{22}^{\text{split},(k)}(\tau)$ depend on various additional user-supplied quantities. Specifically,

$$\text{IF}_{1,\text{new},i} = \frac{1}{|s(i)|} \sum_{j \in s(i)} \hat{f}_j(X_i) \left[Y_i - e^{\tau A_i} \hat{b}_i(X_i) \right] \left[A_i - \hat{p}_i(X_i) \right],$$

$$\mathbb{IF}_{1,\text{new}}(\tau) = \sum_i \text{IF}_{1,\text{new},i}(\tau),$$

$$K_k(x_i, x_j) = \sum_{l=1}^{k} \varphi_l(x_i) \varphi_l(x_j),$$

$$\mathrm{IF}_{22,ij}^{(k)} = -\left[Y_i - e^{\tau A_i}\hat{b}_i(X_i)\right] K_k(X_i, X_j)\left[A_j - \hat{p}_j(X_j)\right] \frac{e^{\tau A_j}}{\widehat{E}_j[e^{\tau A_j}|X_j]},$$

$$\widehat{\mathbb{IF}}_{22}^{(k)}(\tau) = \sum_i \frac{1}{|s(i)|} \sum_{j \in s(i)} \mathrm{IF}_{22,ij}^{(k)}.$$

[Recall that $\varphi_l(x)$ is an element of an orthonormal basis for the set of all square-integrable functions on the unit cube in $\mathbb{R}^4$.] Here, the scale (linear or loglinear) of the semiparametric model, $s(i)$, $|s(i)|$, choice of transform of covariates $X_{\mathrm{cont},i}$, $\hat{f}_j(X_i)$, and $\varphi_l(x)$ are settings chosen by the user.

The results in Figures 1 and 2 use base-case choices for each of the user-supplied quantities. In this section we describe these additional quantities and our base-case choices. In the following subsections we study the sensitivity of the results in Figures 1 and 2 to the choices.

- Choice of semiparametric regression model. We chose to use the loglinear semiparametric regression model of Equation (2.1) as our base case. An alternative choice would have been to use the linear semiparametric regression model

$$E_i[Y_i|A_i, X_i] = \tau^* A_i + \zeta_i^*(X_i),\ i = 1, \ldots, N.$$

  We chose the loglinear rather than the linear form as our base case because (1) that choice is standard in the literature and (2) the number of deaths on a given day is non-negative.
- Choice of transform of covariates. When estimating τ^*, we consider three possible transforms of the continuous covariates $X_{\mathrm{cont},i}$: "id", "gs", and "gs2" (identity transform and two variations of Gram-Schmidt orthogonalization). These are defined in Section 2.3.3. Our base-case choice was gs. The transforms id, gs, and gs2 have the effect of making $\hat{f}_j(x)$ increasingly close to a joint distribution with independent uniformly distributed components. We chose gs as our base case largely because, as discussed in Section 2.3.3, if we had used the identity transformation, the peaks of $\hat{f}_j(x)$ would have been too large and sharp.
- Choice of density estimator. $\hat{f}_j(\cdot) = \hat{f}_j(x) = \hat{f}_j(x_{\mathrm{cont}})\widehat{\mathrm{pr}}[X_5 = x_5]$ is the joint density for X on day t_j, with $\widehat{\mathrm{pr}}[X_5 = x_5]$ being the empirical proportion of all subjects in age category x_5. (Since X_5 is age, it can be assumed to be independent of the continuous variables X_{cont}.) $\hat{f}_j(x_{\mathrm{cont}})$ is obtained by applying the nonparanormal density estimator algorithm for time-series data described in Section 2.3.4. We note that for $\widehat{\mathbb{IF}}_{1,\mathrm{new}}^{\mathrm{split}}(\tau)$ we still use the whole sample to estimate $\hat{f}_j(x)$, even though we use split-sample estimates of the nuisance functions $b_i^*(x)$, $p_i^*(x)$, and $q_i^*(x)$.

  As discussed earlier, we chose as our base case a nonparanormal density estimator smoothed over time. This is a density estimator that estimates each of the marginal densities of the 4 continuous components of X_i nonparametrically using kernel density estimation. The dependence structure of X_i is then estimated based on a parameteric model with 6 parameters. (See Section 2.3.4 for a formal definition.) As an alternative, we also considered more nonparametric density estimators, based on the program Locfit by Loader (2010), in Section 2.3.4; however these more nonparametric estimators performed no better than the nonparanormal density estimators.
- Choice of the other occasions associated with occasion i. The second-order U-statistic estimating function $\widehat{\mathbb{IF}}_2^{\mathrm{split},(k)}(\tau)$ associates occasion i with a set $s(i)$ of other occasions j for which a function of both their data O_i and O_j will be included in $\widehat{\mathbb{IF}}_2^{\mathrm{split},(k)}(\tau)$. We denote this set of occasions j by $s(i)$. This same set also enters our definition of $\mathrm{IF}_{1,\mathrm{new}}$. To describe our base case, we let $\hat{f}_j(x)$ be the estimated density of X at time j. Then $\hat{f}_j(X_i)$ is the density at time j evaluated

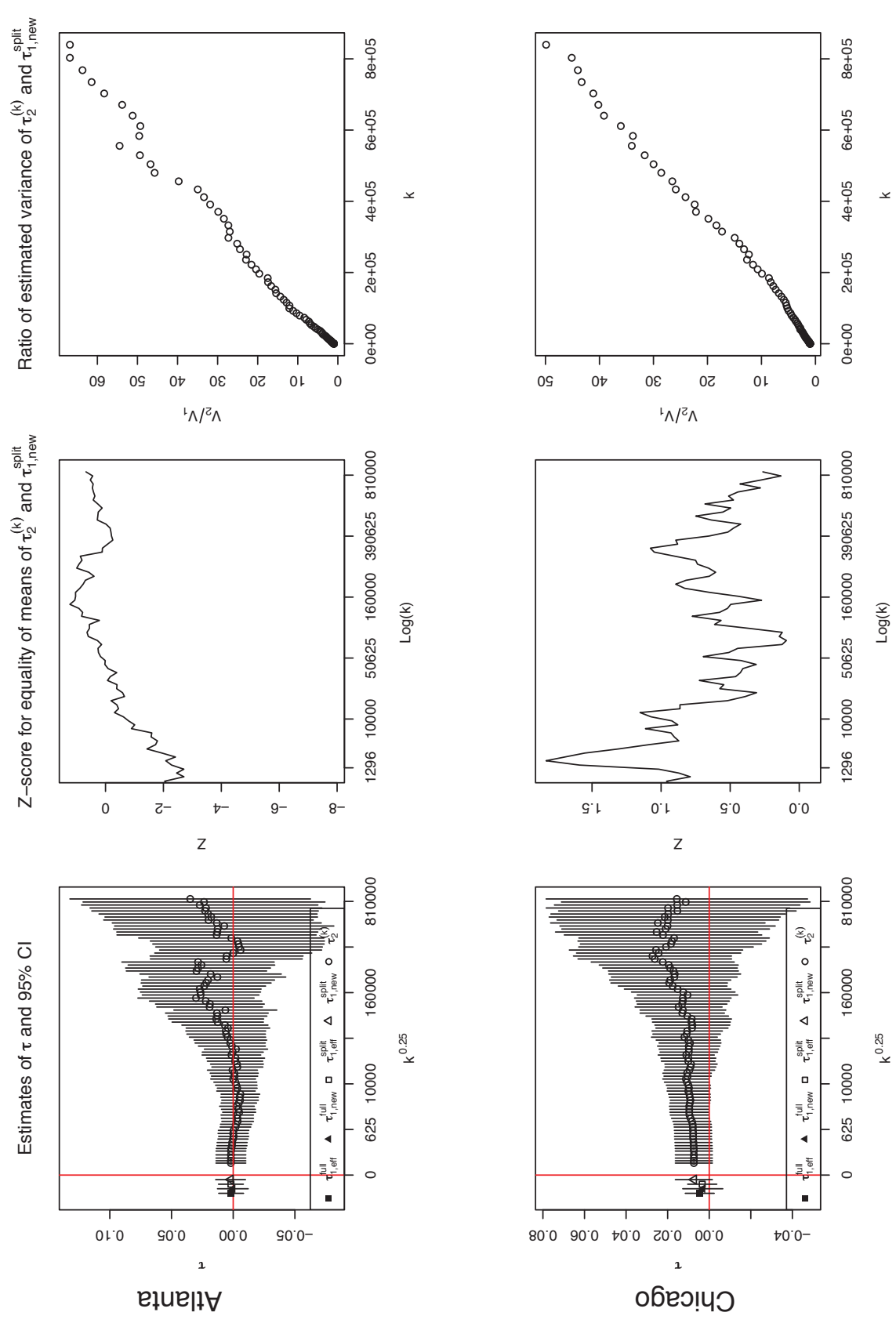

Figure 2. Linear model results for 22 cities. Summary of the analysis using a linear model with Legendre polynomials, a density cutoff at the 80th percentile, observations between 25 and 75 days of a given day, and with k between 3 and 839523. Note that y axis scales vary among cities.

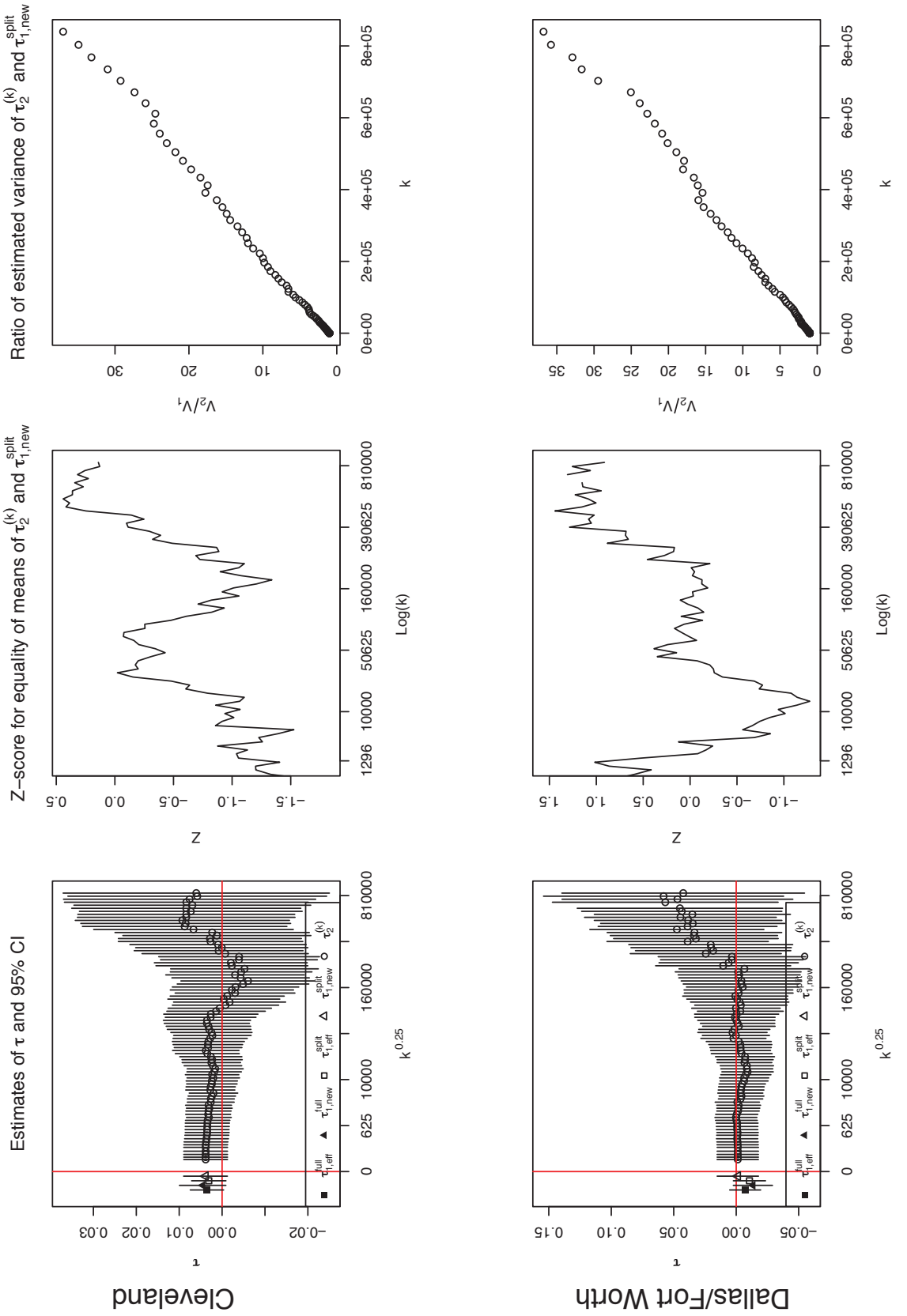

Figure 2. (continued)

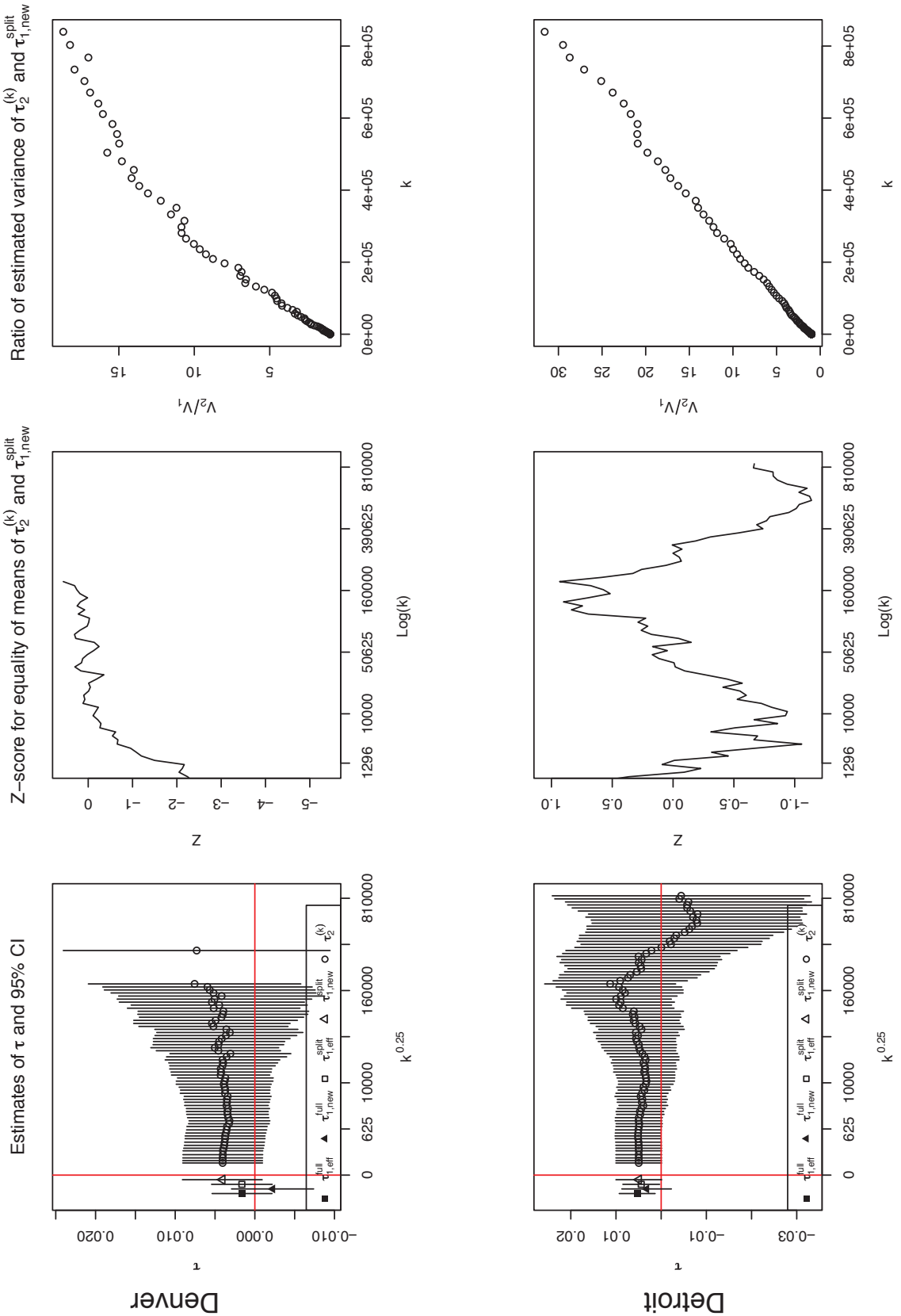

Figure 2. (continued)

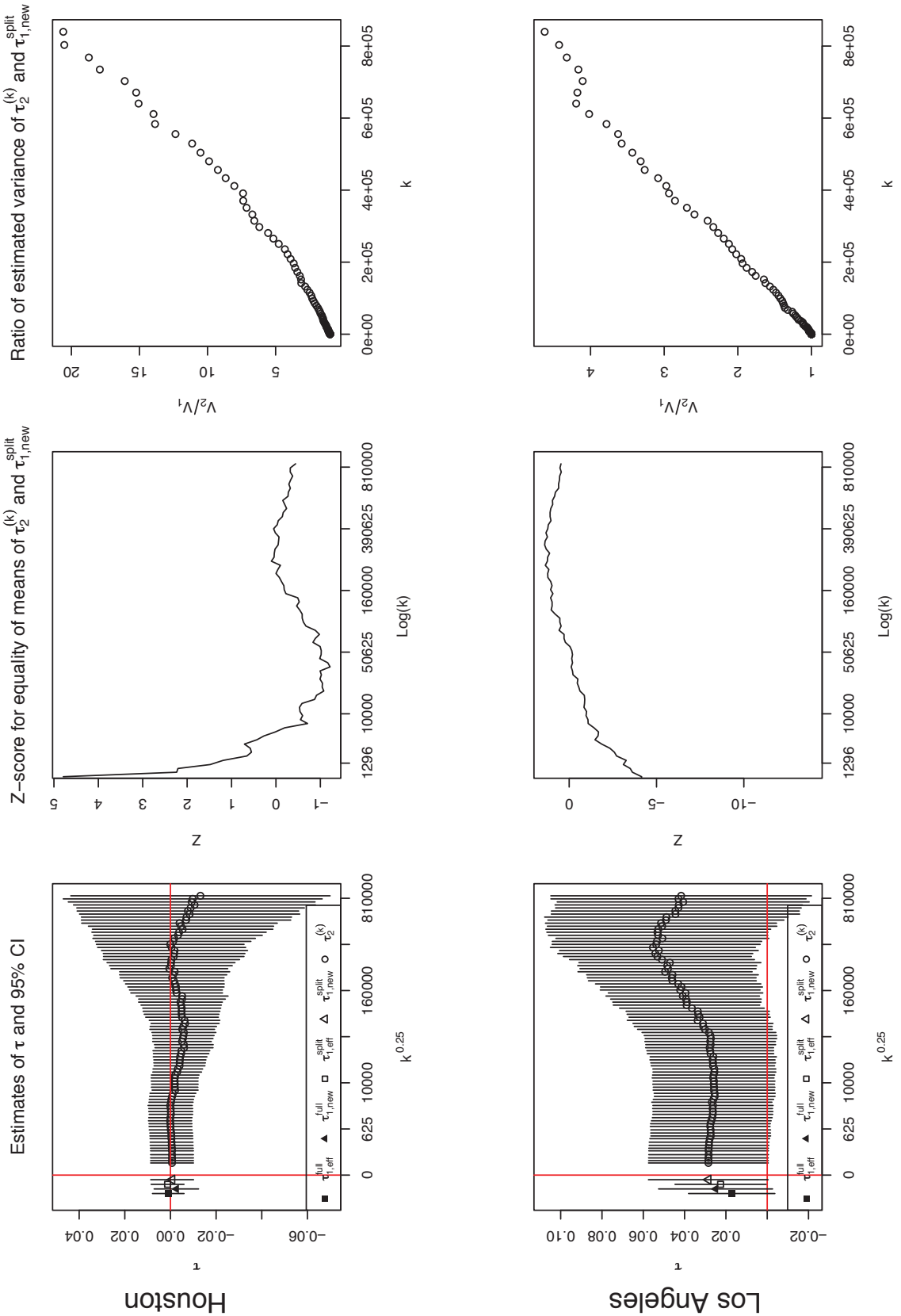

Figure 2. (continued)

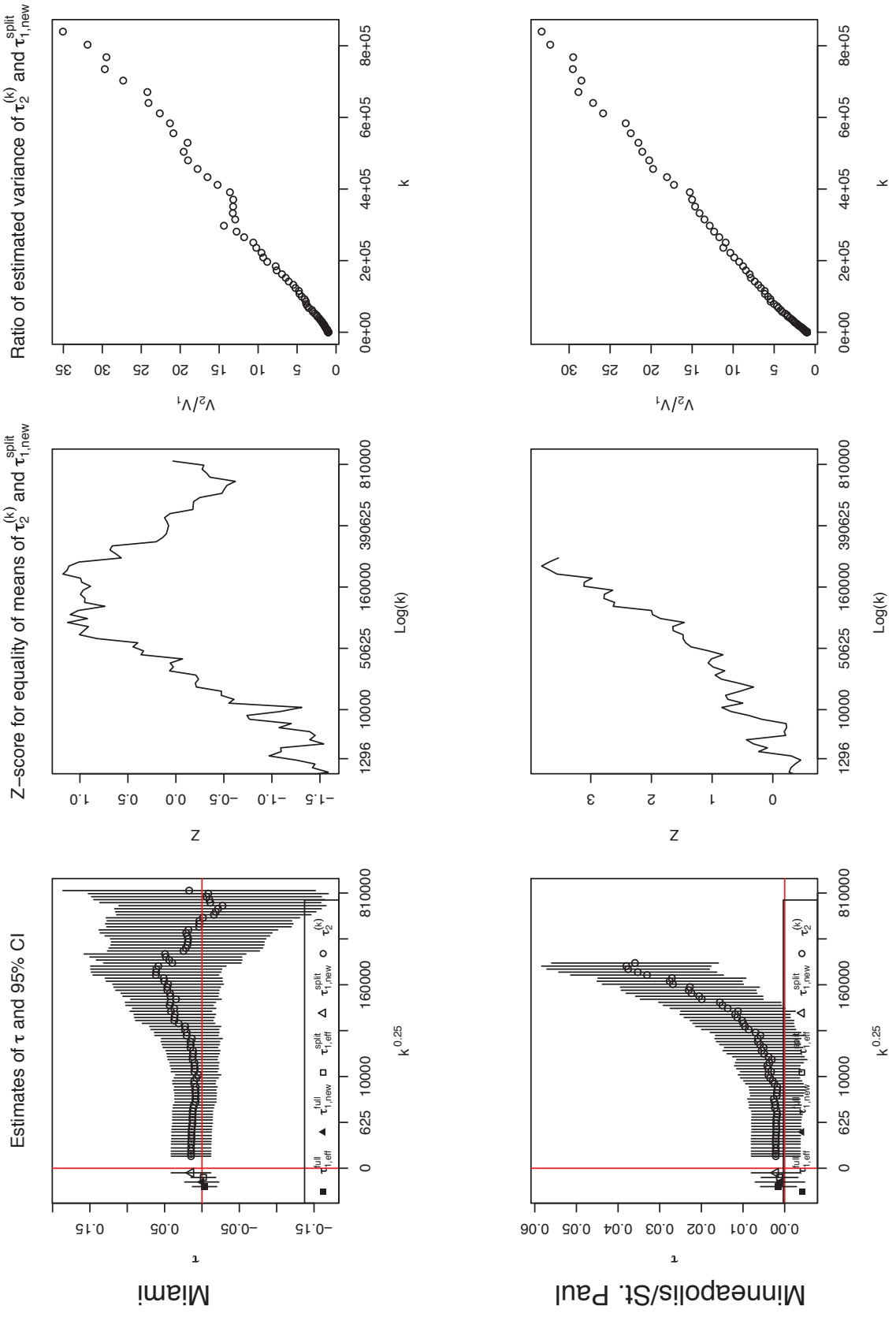

Figure 2. (continued)

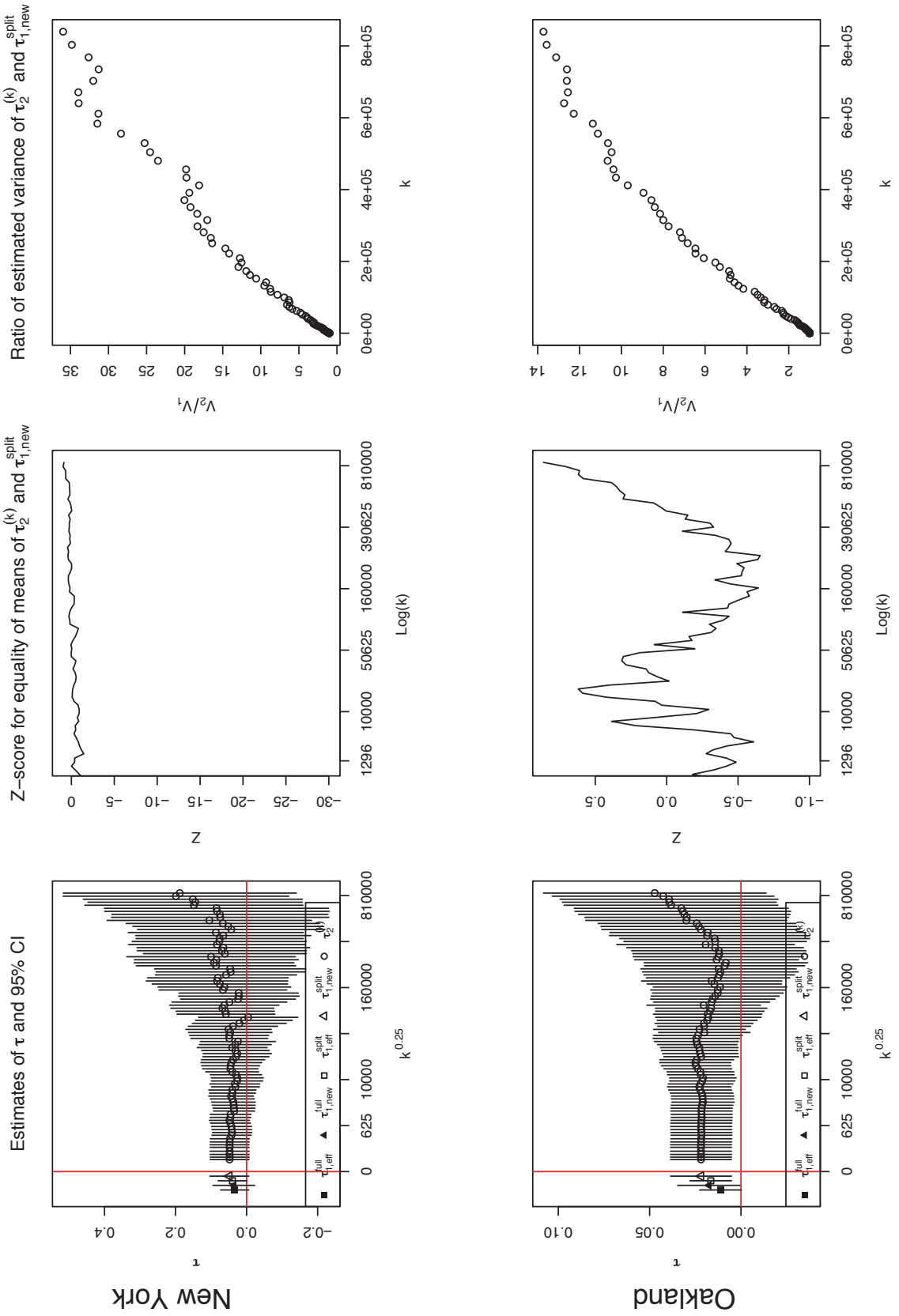

Figure 2. (continued)

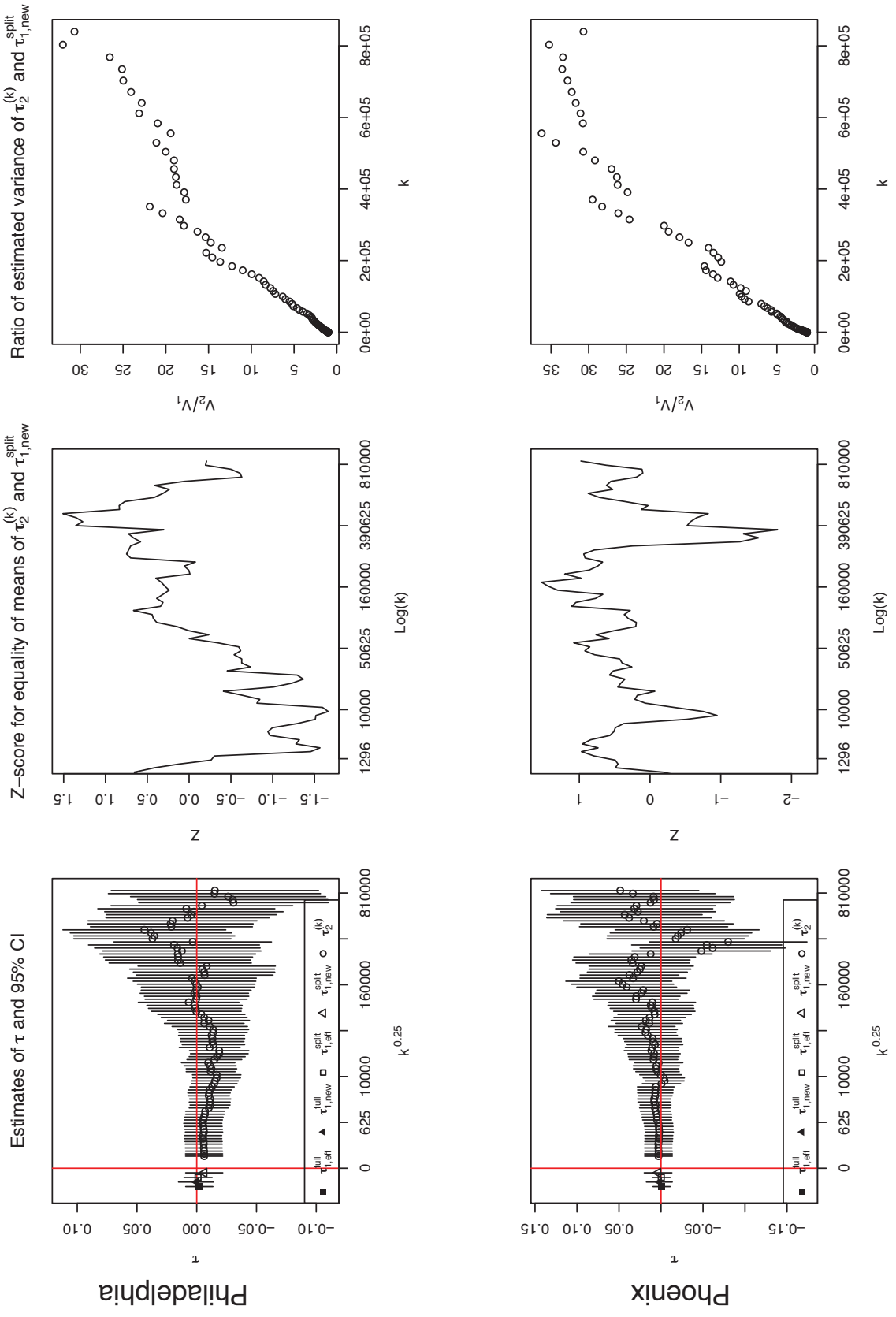

Figure 2. (continued)

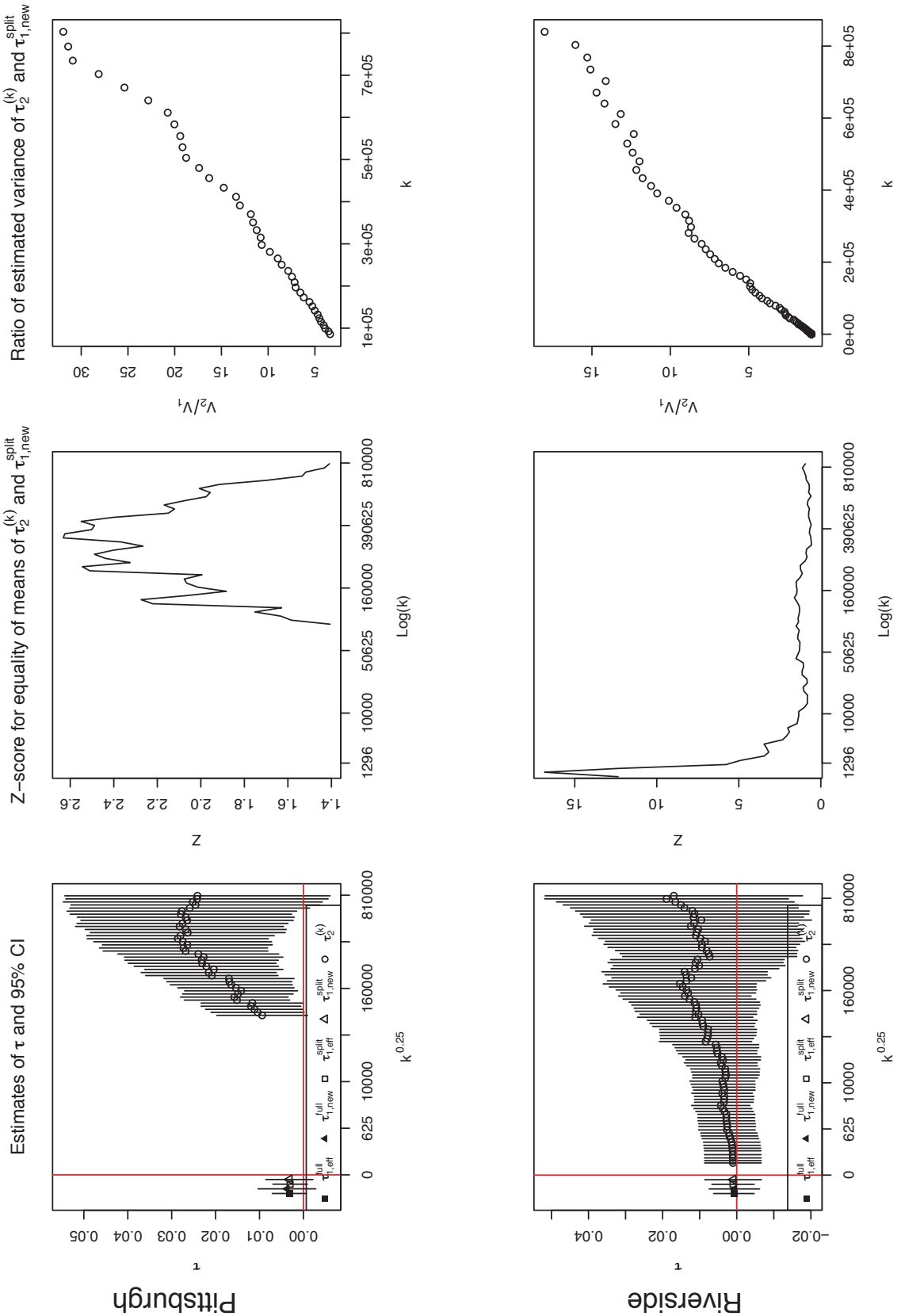

Figure 2. (continued)

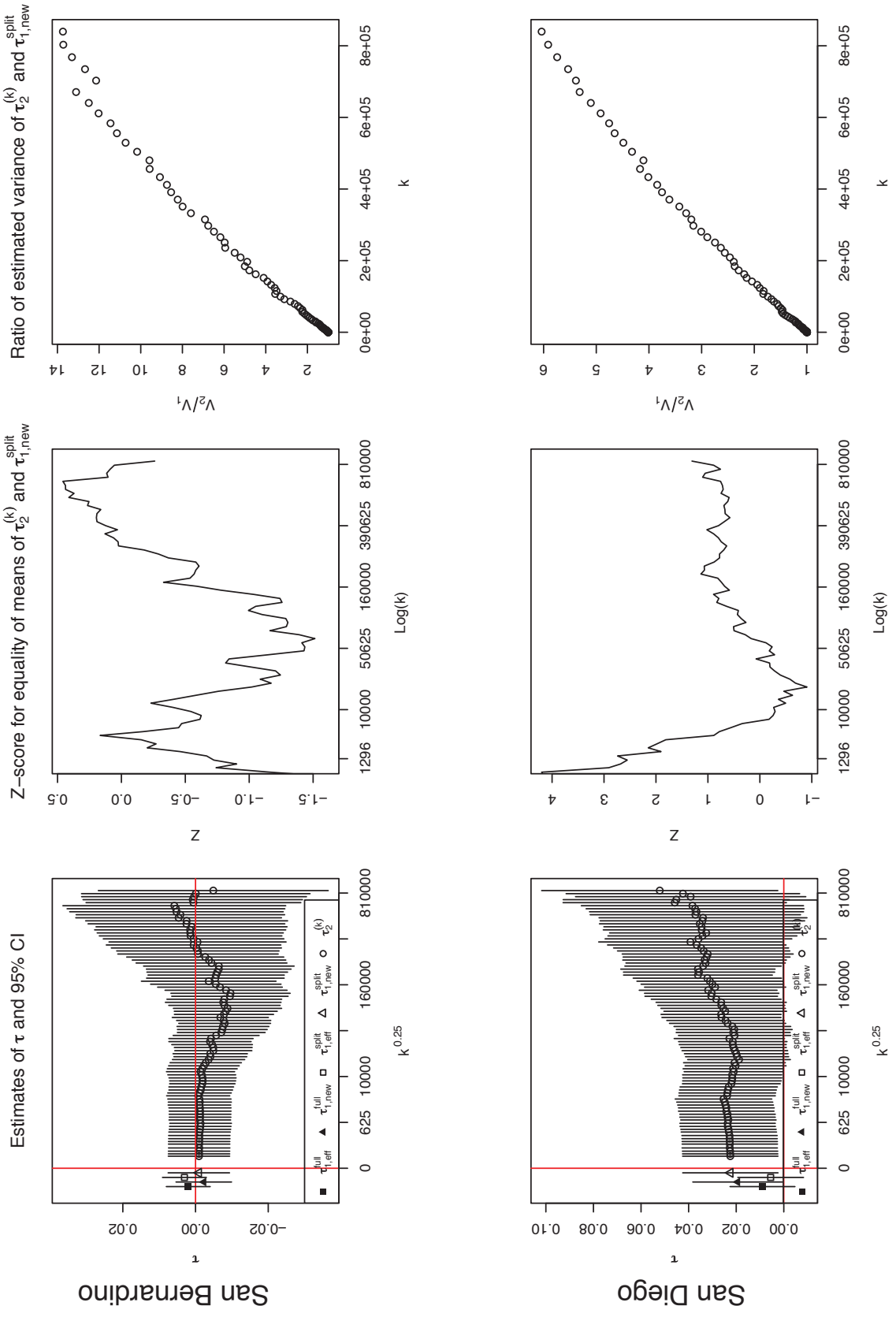

Figure 2. (continued)

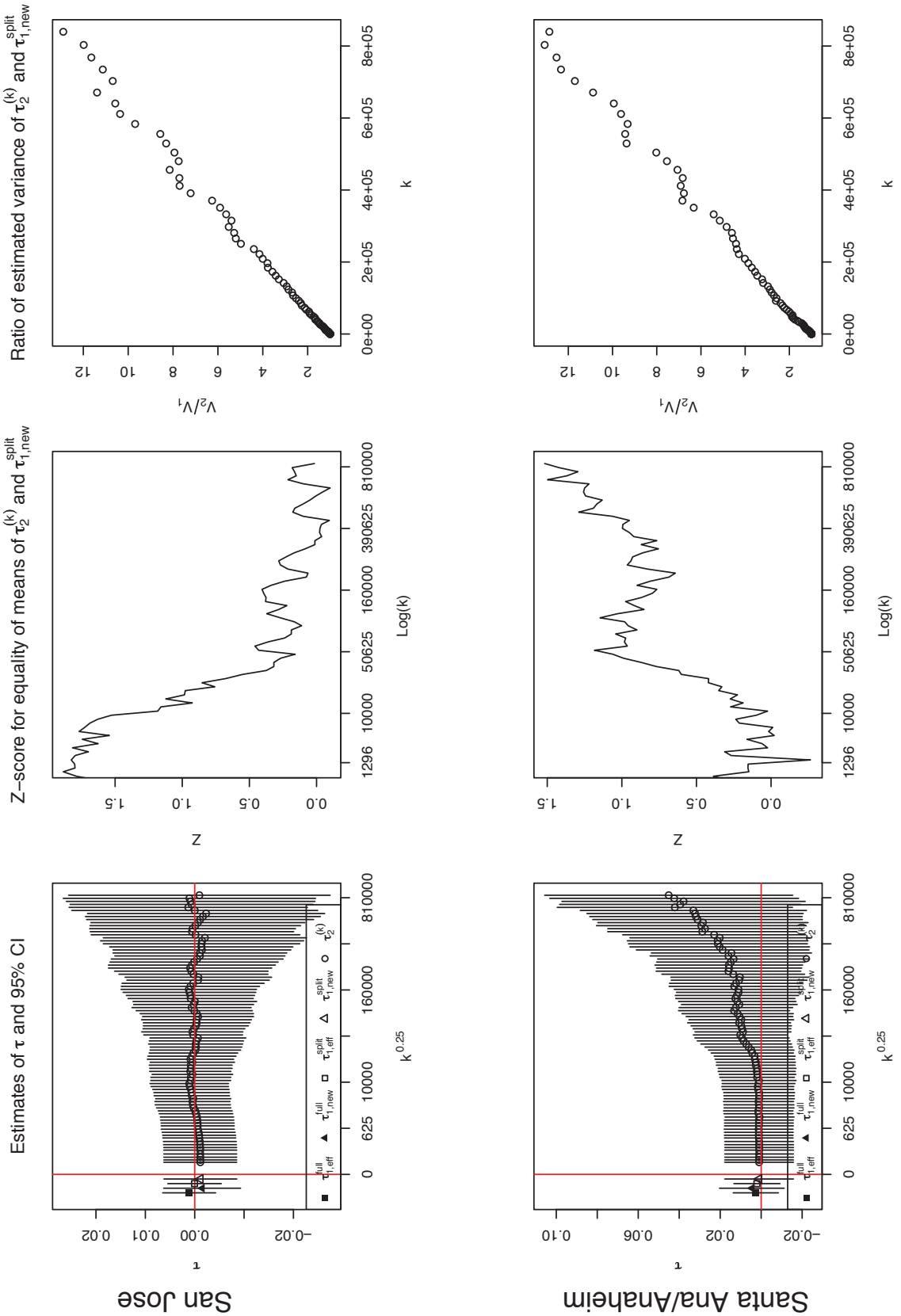

Figure 2. (continued)

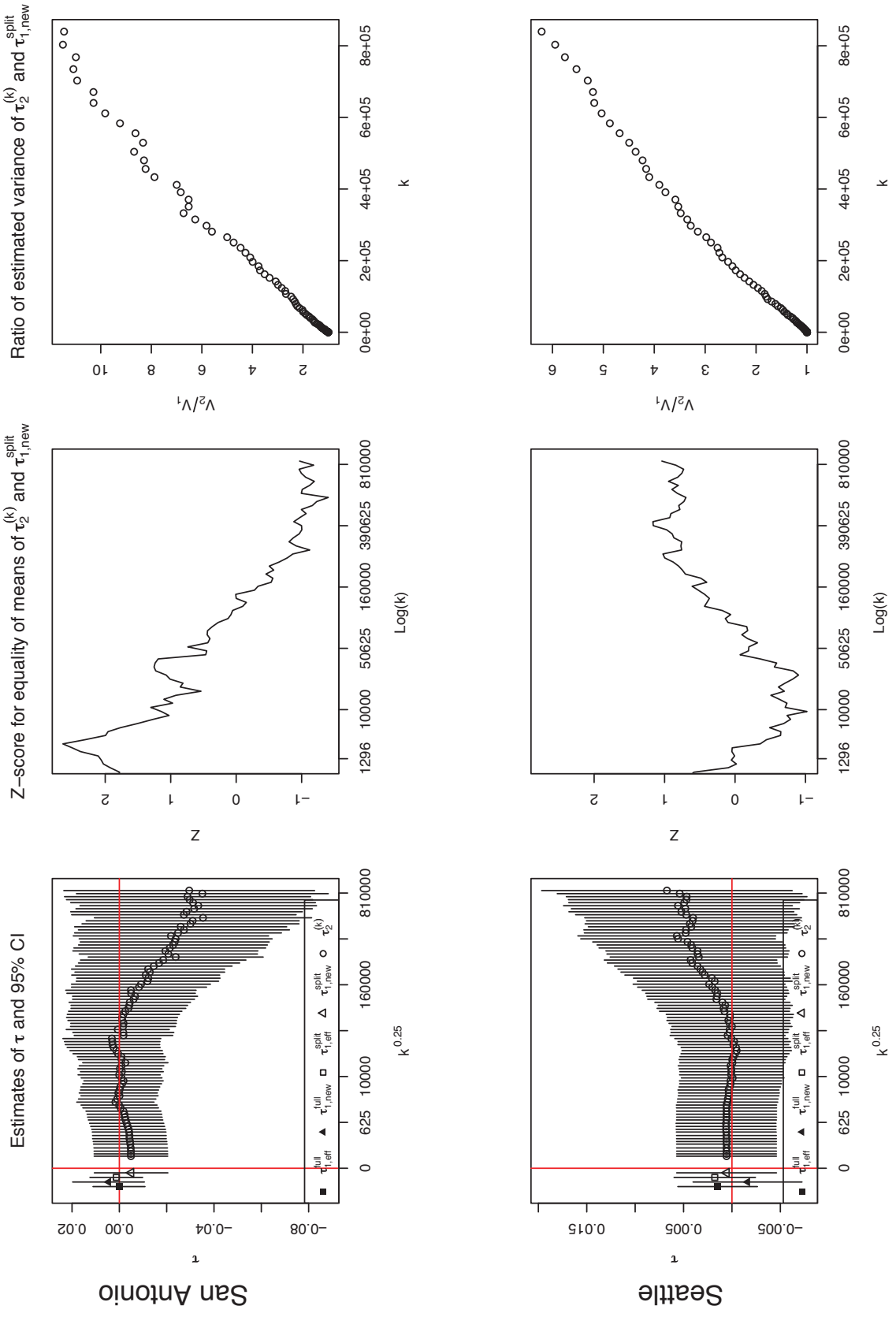

Figure 2. (continued)

at the value of X at occasion i. Next we let $\left\{\hat{f}_j(X_i); i,j = 1,\ldots,N\right\}$ be the N^2 such numbers. Our base-case choice was to include in $s(i)$ those occasions j which satisfied all the following:

1. 25 to 75 occasions either prior to or after occasion i,
2. in the same half-sample as i of the two split-samples; and
3. had estimated density $\hat{f}_j(X_i)$ less than the 80-th percentile of the N^2 values in $\left\{\hat{f}_j(X_i); i,j = 1,\ldots,N\right\}$.

Thus, for each time i, $s(i)$ is a set of time indices (i.e., a set of observations) that depends on i and on the half-sample containing i. Then $|s(i)|$ is the cardinality of the set $s(i)$.

Our base-case choice was based on the following considerations. The second-order U-statistic $\hat{\mathbb{F}}_2^{\text{split},(k)}$ requires the data O_i and O_j corresponding to occasion i and occasion j to be independently distributed when they occur together in one of its terms. We chose j to be more than 25 occasions from i to ensure this independence held at least approximately. We chose j to be within 75 occasions of i in an attempt to control the time-series bias (discussed above) by bounding at 75 occasions the time period we had to model accurately. The results of alternative choices are discussed in Section 2.3.2.

- Choice of orthonormal basis. The second-order U-statistic depends on a choice $\varphi_l(\cdot), l = 1,\ldots$, of an orthonormal basis for the set of square integrable functions on the unit cube in $\mathbb{R}^4$. For our base-case analysis, we chose the $\varphi_l(\cdot)$ to be tensor products of the univariate Legendre orthogonal polynomials. (We discuss the Haar basis and Daubechies wavelet bases as alternatives in Section 2.3.5.)

We now discuss the sensitivity to these choices in detail.

2.3.1 Choice of Linear vs. Loglinear Semiparametric Regression Model

We now demonstrate that our conclusions would be unchanged had we chosen the linear model as our base case.

Figure 2 and the linear columns in Table 1 display results for the linear model that are analogous to the results for the loglinear model shown in Figure 1 and the loglinear columns in Table 1. Remarkably, the city-specific results in Figures 1 and 2 are nearly identical, including the shape of the plot of $\hat{\tau}_2^{\text{split},(k)}$ versus k; the only discernible difference is the magnitude of the estimate $\hat{\tau}$ (see left panels of figures). Roughly, for a given city, $\hat{\tau}_{linear} = \hat{\tau}_{loglinear} \times$ (the average number of deaths per day under the loglinear model when the 1-day lagged PM_{10} level is 0), where $\hat{\tau}_{loglinear}$ denotes any of our loglinear estimators and $\hat{\tau}_{linear}$ denotes the analogous linear estimator. This is explained theoretically in Section 3.4.2.

2.3.2 Sensitivity to Choice of $s(i)$

Our base-case choice for the set $s(i)$ associated with an occasion i was defined as the intersection of two sets: the set $\{j : 25 \le |j-i| \le 75\}$ and the set $\{j : \hat{f}_j(X_i)$ is less than the 80-th percentile of the N^2 numbers $f_{j'}(X_{i'}), i', j' = 1,\ldots,N\}$. We first consider sensitivity to the choice of the range of $|j-i|$. We refer to 25–75 as the range of $|j-i|$ and to 80 as the percentile cut-off.

Range of $|j-i|$ We compared the choice $25 \le |j-i| \le 75$ to the choice $25 \le |j-i| \le 3000$ in the New York data for several choices of the percentile cut-off (80 and 70) and for several transforms from X_{cont} to X^*_{cont} (the base-case gs transform and an alternative transform gs2 described in Section 2.3.3 below). All other choices remained base case.

In all cases $\hat{\tau}^{\text{split}}_{1,\text{new}}$ and $\hat{\tau}^{\text{split},(k)}_{2}$ do not differ significantly from their base-case values for any k (data not shown). However, other statistics do depend on the choice of base case.

Consider the slope of V_2/V_1 versus k. Reading from Table 2 we see that as the range of $|i-j|$ increases, the slope decreases sharply. This is as expected since the variance of $\widehat{\mathbb{IF}}_{22}$ is inversely proportional to s, where $1/s$ is the average of the $1/|s(i)|$ values. Further the estimated variance of $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}$, obtained as the square of the estimated standard deviation (sd), varies much less than did the slope. The estimated variance of $\hat{\tau}^{\text{split}}_{1,\text{new}}$ changes by less than 10% as the range varies.

Table 2. Sensitivity to $s(i)$ range for New York data. Here $\widehat{\mathbb{IF}}$ denotes $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}$. The V_2/V_1 column denotes the slope of V_2/V_1 with respect to k and has been multiplied by a factor of 250,000.

Percent Cut-Off and Transform	Legendre Orthogonal Polynomial and Linear Models							
	$25-75$				$25-3000$			
	V_2/V_1	$\widehat{\mathbb{IF}}$	$\widehat{\text{sd}}(\widehat{\mathbb{IF}})$	$\widehat{\text{sd}}(\hat{\tau}^{\text{split}}_{1,\text{new}})$	V_2/V_1	$\widehat{\mathbb{IF}}$	$\widehat{\text{sd}}(\widehat{\mathbb{IF}})$	$\widehat{\text{sd}}(\hat{\tau}^{\text{split}}_{1,\text{new}})$
80%, gs	16.0	13.56	7.3	0.027	2.2	22.82	10.2	0.025
70%, gs	41.0	8.77	4.4	0.025	5.0	12.50	5.6	0.023
80%, gs2	7.5	48.18	42.6	0.028	1.5	56.90	40.6	0.025
70%, gs2	100.0	55.71	25.9	0.027	1.8	31.31	31.9	0.024

Percentile Cut-Off for Density We compared the percentile cut-offs 100 and 70 with our base case of 80 in the New York data and Chicago data for several choices of the range of $|j-i|$ and for several transforms (the base-case gs transform and an alternative transform gs2 described in Section 2.3.3 below). All other choices remained base case.

In all cases $\hat{\tau}^{\text{split}}_{1,\text{new}}$ and $\hat{\tau}^{\text{split},(k)}_{2}$ do not differ significantly from their base-case values for any k (data not shown). However, other statistics do depend on the choice of base case.

Reading from Table 3 we see that as the percentile cut-off decreases, the slope of $V_2/V_1 = V_2^{(k)}/V_1$ with respect to k increases, but the standard deviation of $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}$ decreases. The ratio of the squared standard deviation at a cut-off of 80% to that at 70% is greater than the ratio of the slope of V_2/V_1 at 70% to that at 80%. The estimated variance of $\hat{\tau}^{\text{split}}_{1,\text{new}}$ changes by less than 15% as the percentile cut-off varies.

Table 3. Sensitivity to percentile cutoff for New York and Chicago data. Here $\widehat{\mathbb{IF}}$ denotes $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}$; $\widehat{\text{sd}}_{\text{IF}}$ denotes $\widehat{\text{sd}}(\widehat{\mathbb{IF}})$; and $\widehat{\text{sd}}_{\tau}$ denotes $\widehat{\text{sd}}(\hat{\tau}^{\text{split}}_{1,\text{new}})$. The V_2/V_1 column denotes the slope of V_2/V_1 with respect to k and has been multiplied by a factor of 250,000.

Days in the Sample and Transform	Legendre Orthogonal Polynomial and Linear Models											
	100%				80%				70%			
	V_2/V_1	$\widehat{\mathbb{IF}}$	$\widehat{\text{sd}}_{\text{IF}}$	$\widehat{\text{sd}}_{\tau}$	V_2/V_1	$\widehat{\mathbb{IF}}$	$\widehat{\text{sd}}_{\text{IF}}$	$\widehat{\text{sd}}_{\tau}$	V_2/V_1	$\widehat{\mathbb{IF}}$	$\widehat{\text{sd}}_{\text{IF}}$	$\widehat{\text{sd}}_{\tau}$
NY, 25–75, gs					16.0	13.56	7.3	0.027	41.0	8.77	4.4	0.025
NY, 25–75, gs2					7.5	48.18	42.6	0.028	100.0	55.71	25.9	0.027
NY, 25–3000, gs	1.2	36.98	44.45	0.027	2.2	22.82	10.2	0.025	5.0	12.50	5.6	0.023
NY, 25–3000, gs2					1.5	56.90	40.6	0.025	1.8	31.31	31.9	0.024
Chicago, 25–1000, gs					1.4	7.39	8.2	0.004	1.75	3.00	5.3	0.004
Chicago, 25–1000, gs2					1.25	43.42	30.0	0.004	1.3	13.88	23.2	0.005

The decrease of the standard deviation of $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}$ is largely due to the fact that the observations with the larger $\hat{f}_j(X_i)$ are increasingly removed as the percentile cut-off decreases. However, the variance of $\hat{\tau}^{\text{split}}_{1,\text{new}}$ equals the variance of $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}$ divided by an estimate of its derivative. The decrease in the variance of $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}$ does not affect the variance of $\hat{\tau}^{\text{split}}_{1,\text{new}}$ since this denominator decreases similarly. Note that valid results based on data truncated by a cut-off require that the density estimator be reapplied to the truncated data and the analysis redone on the truncated data without further truncation. We would also expect that the effect of the cut-off on the variance of $\widehat{\mathbb{IF}}^{\text{split},(k)}_{22}$ is uncertain because of two countervailing mechanisms. First, decreasing the cut-off should tend to decrease the variance due to the fact that the observations with the larger $f_j(X_i)$ are increasingly removed because of the correlation between $f_j(X_i)$ and $\hat{f}_j(X_i)$. Second, decreasing the cut-off should increase the variance by reducing s since $s(i)$, and thus its cardinality, decreases as the percentile cut-off decreases.

In fact, one or two sets with $|s(i)|$ equal to 1 will make $\frac{1}{|s|}$ large and thus $\text{Var}(\widehat{\mathbb{IF}}^{(k)}_{22})$ large. This observation suggests that the effect of the percentile cut-off on the variance of $\widehat{\mathbb{IF}}^{(k)}_{22}$ due to this second mechanism could be quite nonlinear, particularly when the range of $|j-i|$ is small, since sets with $|s(i)|$ equal to 1 are much more likely.

To better understand these countervailing effects, we recall that the ratio of the squared standard deviation at a cut-off of 80% to that at 70% is greater than the ratio of the slope of V_2/V_1 at 70% to that at 80%, with the exception of NY 25-75, gs2. This implies that the variance of $\widehat{\mathbb{IF}}^{\text{split},(k)}_{22}$ is decreasing with a decreasing cut-off (i.e., mechanism 1 dominates), except for NY 25-75, gs2 where mechanism 2 dominates. This is consistent with expectations for two reasons.

First, for reasons discussed in Section 2.3.3, the effect of removing observations with the larger $\hat{f}_j(X_i)$ and $f_j(X_i)$ should have less effect on $\text{Var}(\widehat{\mathbb{IF}}^{\text{split},(k)}_{22})$ and $\text{Var}(\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}})$ for the gs2 transformation than for the gs transformation. This is due to the fact that, as can be seen from Table 4, truncating the right tail of the $\hat{f}_j(X_i)$ for gs will leave the distribution of the $\hat{f}_j(X_i)$ dominated by smaller values of $\hat{f}_j(X_i)$ and thus with a much smaller variance than that of the same distribution with less truncation. In contrast, for gs2, this decrease in the variance of the $\hat{f}_j(X_i)$ with increasing truncation will be less, as, for gs2, the distribution of $\hat{f}_j(X_i)$ below the third quartile is more variable than for gs. Indeed this fact can be empirically observed for $\text{Var}(\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}})$, as the ratio of the standard deviation of $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}$ comparing the 80% cut-off to the 70% cut-off in Table 3 is always less for gs2 than for gs.

Table 4. Summary for New York data of empirical distribution of the density estimates $\hat{f}_{i_j}X_i$. Estimates were pooled over all N^2 possible combinations of i and j, and were based on the id, gs, and gs2 transformations.

Transform	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
id	0.00	0.069	2.998	22.430	22.610	820.500
gs	0.00	0.211	2.809	13.240	14.680	466.900
gs2	0.00	3.290	12.670	23.190	30.630	284.400

As discussed above, the second mechanism is expected to operate more strongly when the range of $|j-i|$ is small, as it is for NY 25–75.

Table 5. Sensitivity to choice of transform for New York and Chicago data. Here $\widehat{\mathbb{IF}}$ denotes $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}$; $\widehat{\text{sd}}_{\text{IF}}$ denotes $\widehat{\text{sd}}(\widehat{\mathbb{IF}})$; and $\widehat{\text{sd}}_\tau$ denotes $\widehat{\text{sd}}(\hat{\tau}^{\text{split}}_{1,\text{new}})$. The V_2/V_1 column denotes the slope of V_2/V_1 with respect to k and has been multiplied by a factor of 250,000.

Days in the Sample and Percent Cut-Off	Legendre Orthogonal Polynomial and Linear Models											
	id				gs				gs2			
	V_2/V_1	$\widehat{\mathbb{IF}}$	$\widehat{\text{sd}}_{\text{IF}}$	$\widehat{\text{sd}}_\tau$	V_2/V_1	$\widehat{\mathbb{IF}}$	$\widehat{\text{sd}}_{\text{IF}}$	$\widehat{\text{sd}}_\tau$	V_2/V_1	$\widehat{\mathbb{IF}}$	$\widehat{\text{sd}}_{\text{IF}}$	$\widehat{\text{sd}}_\tau$
NY, 25–75, 80%					16.0	13.56	7.3	0.027	7.5	48.18	42.6	0.028
NY, 25–75, 70%					41.0	8.77	4.4	0.025	100.0	55.71	25.9	0.027
NY, 25–3000, 80%	2.0	28.26	13.7	0.024	2.2	22.82	10.2	0.025	1.5	56.90	40.6	0.025
NY, 25–3000, 70%					5.0	12.50	5.6	0.023	1.8	31.31	31.9	0.024
Chicago, 25–1000, 80%					1.4	7.39	8.2	0.004	1.25	43.42	30.0	0.004
Chicago, 25–1000, 70%					1.75	3.00	5.3	0.004	1.3	13.88	23.2	0.005

2.3.3 Sensitivity to Choice of Transform

In our base case, we first empirically orthogonalize the components of the vector X_{cont} to obtain a new vector by gs and then apply a separate scale and location shift to each component to guarantee that the empirical support for each component is the interval $[0,1]$. We considered two alternative transforms.

The first alternative, the id transform, was to apply a separate scale and location shift to each component of the original X_{cont} to guarantee that the empirical support for each component is the interval $[0,1]$. Note that in all cases the final scale and location shifts were required to insure that X_{cont} had support on the unit cube in $\mathbb{R}^4$, so that the Legendre and Haar bases would serve as orthonormal bases.

The final alternative, gs2, was to further remove seasonality and time by first fitting separately for each city and each of the 4 components of X_{cont} a natural cubic spline with 40 df for time, replacing each component of X_{cont} by its residual from this fit, and applying the above gs orthogonalization to the residuals followed by a separate scale and location shift.

From Table 5 we observe that in all cases, $\hat{\tau}^{\text{split}}_{1,\text{new}}$ and $\hat{\tau}^{\text{split},(k)}_2$ do not differ significantly from their base-case values for any k (data not shown). However, other statistics do depend on the choice of base case.

In all cases, $\text{Var}(\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}) = \left\{\text{sd}(\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}})\right\}^2$ was much greater for gs2 than for gs. Further in all cases, the ratio of $\text{Var}(\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}})$ × the slope V_2/V_1 for gs2 divided by the same quantity for gs was much greater than 1, implying that $\text{Var}\left[\widehat{\mathbb{IF}}^{\text{split},(k)}_{22}(0)\right]$ for gs2 was greater than for gs. The slope of V_2/V_1 for gs was greater than that for gs2, with the exception of the case of NY with 25–75 and 70% cut-off, implying in all cases but the exception that the relative increase in $\text{Var}(\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}})$ comparing gs and gs2 was greater than that for $\text{Var}(\widehat{\mathbb{IF}}^{\text{split},(k)}_{22})$. The standard deviation of $\hat{\tau}^{\text{split}}_{1,\text{new}}$ was slightly greater for gs2 than for gs.

Similiarly $\text{Var}(\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}})$ was much greater for gs2 than for id. Further, comparing id and gs2, the relative increase in $\text{Var}(\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}})$ was greater than that for $\text{Var}(\widehat{\mathbb{IF}}^{\text{split},(k)}_{22})$.

On the other hand, $\text{Var}(\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}})$ and $\text{Var}(\widehat{\mathbb{IF}}^{\text{split},(k)}_{22})$ were greater for id than for gs. Again, comparing id and gs, the relative decrease in $\text{Var}(\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}})$ was similar to that for $\text{Var}(\widehat{\mathbb{IF}}^{\text{split},(k)}_{22})$.

Explanation From Table 4, we observe that, for id and gs, the distribution of the $\hat{f}_j(X_i)$, especially after a cut-off, was dominated by small values compared to the distribution of gs2, which was

more nearly uniform. This represents the fact that gs2 removed the periodicity and time trends that dominated the distribution of the $\hat{f}_j(X_i)$ for gs and id; for example, without removing time trends and periodicity, $\hat{f}_j(X_i)$ with t_i in summer and t_j in winter will have a value near zero. Thus the products $\hat{f}_j(X_i)\hat{f}_{t_{j'}}(X_i)f_i(X_i)$ occurring in $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}$ have high probability of being small for gs and id compared to gs2. Further, from Table 4, the upper quantiles of the distribution of $\hat{f}_j(X_i)$ appear to be similar for all distributions, a somewhat surprising result. Thus, the variance of $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}$ should be smaller for gs and id compared to gs2.

Now the variance of $\widehat{\mathbb{IF}}^{\text{split},(k)}_{22}$ (i.e., $\widehat{\mathbb{IF}}^{\text{split},(k)}_{22}(\tau)$ evaluated at $\tau = 0$) depends on products of the $f_j(X_i)$, which for similar reasons to those just above suggests the variance of $\widehat{\mathbb{IF}}^{\text{split},(k)}_{22}$ should be greatest for gs2. However the number of $f_j(X_i)$ and/or $\hat{f}_j(X_i)$ multiplied together in computing the variance is greater for $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}$ than for $\widehat{\mathbb{IF}}^{\text{split},(k)}_{22}$ so the relative increase in variance comparing gs2 to gs or id should be greater for $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}$ than for $\widehat{\mathbb{IF}}^{\text{split},(k)}_{22}$.

However there is a second reason for a relatively larger increase in the variance of $\widehat{\mathbb{IF}}^{\text{split},(k)}_{22}$ for gs2 compared to gs or id, which is as follows.

We refer the reader to Section 3 for results and details of the arguments used in the remainder of this subsection. For the Haar basis and for $k = 2^{4J} \times 3$ for some integer J, the dependence of the variance of $\widehat{\mathbb{IF}}^{(k)}_{22}$ on k (see the proof of Lemma 41 in Appendix E) equals k times a term that increases in norm as $f_j(x)$ approaches the uniform. As a consequence the relative increase in variance of $\widehat{\mathbb{IF}}^{\text{split},(k)}_{22}$ will be greatest for gs2. Although the previous exact equality does not hold for the Legendre basis, approximate equality will hold with similar consequence for the variance of $\widehat{\mathbb{IF}}^{\text{split},(k)}_{22}$.

Although the above explanations are consistent with the empirical results, we have no explanation for (a) the decrease in the variance of $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}$ and $\widehat{\mathbb{IF}}^{\text{split},(k)}_{22}$ in comparing gs to id; or (b) that the relative increase in $\text{Var}(\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}})$ comparing gs and gs2 was less than that for $\text{Var}(\widehat{\mathbb{IF}}^{\text{split},(k)}_{22})$ for the case of NY with 25–75 and 70% cut-off.

2.3.4 Sensitivity to Type of Density Estimator

We first formally describe our base-case estimator: the nonparanormal density estimator (Liu et al., 2009) modified for time-series data. We then describe the Locfit density estimator based on Loader (1999). In both cases, we incorporate time and seasonality in our estimator of $f^*_t(x)$ by having the estimator $\hat{f}_t(x)$ depend only on observations at times close to t or times a whole number of years away from t.

The Nonparanormal Density Estimator Without the complication of time and seasonality, the nonparanormal density estimator is a multivariate density that uses standard univariate density estimators $\hat{f}_l$ for the marginal density of the component X_{li} of the 4-dimensional vector $\left(X_i \equiv X_{\text{cont},i}\right)$ and a Gaussian copula. For the l-th component, define Z_l as

$$Z_l = z_l(X) = \Phi^{-1}\left[\mathbb{F}_l(X_l)\right],$$

where $\mathbb{F}_l$ is the distribution of X_l. Then the estimated joint distribution of Z is given by

$$Z \sim N_p(0, \hat{\Xi})$$

where $\hat{\Xi}$ is the empirical covariance matrix of the $\widehat{Z}_l = \widehat{z}_l(X) = \Phi^{-1}\left[\widehat{\mathbb{F}}_l(X_l)\right]$, where $\widehat{\mathbb{F}}_l(X_l)$ is the empirical distribution of X_l. The estimated joint density of X is

$$\hat{f}(x) = \left(\prod_{l=1}^{p} \frac{\hat{f}_l(x_l)}{\phi[\widehat{z}_i(x)]}\right) \frac{1}{(2\pi)^{p/2}|\hat{\Xi}|} e^{-\widehat{z}_i(x)'\hat{\Xi}^{-1}\widehat{z}_i(x)/2}$$

with $p = 4$.

The Nonparanormal Density Estimator for Time-Series Data For the time series, the goal is to have $\hat{f}$, $\hat{\mathbb{F}}$, and $\hat{\Xi}$ estimated for an observation at time t using only data close to time t. Because of seasonality, "locally in time" includes both times s with $|s-t|$ small and times s with $|s-t|$ close to a multiple of one year. Assume that time is measured in days, let $K()$ be a smoothing kernel, and define the time-smoothing weights as

$$w_{s,t} = K\left(\frac{s-t \mod 365}{h_1}\right) K\left(\frac{s-t}{365 h_2}\right)$$

with h_1 approximately a month and $365h_2$ a few years. The default in our implementation was $h_1 = 30$ and $h_2 = 4$, with a Gaussian kernel for $K()$.

Now define the time-specific empirical cumulative distribution for the l-th covariate at time t as

$$\widehat{\mathbb{F}}_{l,t}(c) = \frac{\sum_{i=1, t_i \neq t}^{N} w_{t_i,t} I\{X_{l,i} \leq c\}}{\sum_{i=1}^{N} w_{t_i,t}}.$$

The sum is over $t_i \neq t$ because otherwise $\widehat{\mathbb{F}}_t$ has its largest jump exactly at x_t, so $\widehat{\mathbb{F}}_t(x_t)$ is too big.

The time-specific marginal density $\hat{f}_{l,t}$ is estimated by smoothing a weighted time-specific histogram of $X_{l,t}$. We used a local polynomial smoother (Wand and Jones, 1995) implemented in the `locpoly()` function in R package KernSmooth (Wand and Ripley, 2009). The histogram is estimated with the same time-smoothing weights $w_{s,t}$ as the cumulative distribution function. The local polynomial smoother for the histogram uses a Gaussian kernel and a bandwidth chosen with a plug-in estimator of the optimal bandwidth. Define

$$\widehat{Z}_{l,t} = \widehat{z}_{l,t}(X_t) = \Phi^{-1}\left[\widehat{\mathbb{F}}_{l,t}(X_{l,t})\right]$$

and estimate the time-specific empirical covariance matrix as

$$\hat{\Xi}_t = \frac{\sum_{i=1}^{N} w_{t_i,t} (\widehat{Z}_{t_i} - \overline{Z}_t)^{\otimes 2}}{\sum_{i=1}^{N} w_{t_i,t}},$$

where $\overline{Z}_t$ is the time-specific mean

$$\overline{Z}_t = \frac{\sum_{i=1}^{N} w_{t_i,t} \widehat{Z}_{t_i}}{\sum_{i=1}^{N} w_{t_i,t}}.$$

The time-specific nonparanormal density estimator is then

$$\hat{f}_t(x) = \left(\prod_{l=1}^{p} \frac{\hat{f}_{l,t}(x_i)}{\phi[\widehat{z}_{l,t}(x)]}\right) \frac{1}{(2\pi)^{p/2}|\hat{\Xi}_t|} e^{-\widehat{z}_t(x)'\hat{\Xi}_t^{-1}\widehat{z}_t(x)/2}.$$

Implementation The code takes as input a training set of times and data used to fit the time-specific nonparanormal density estimator, and a test set of times and data where the density is to be evaluated. The training and test sets can be the same. Other useful choices of the test set are a grid of data values or a uniform random sample of data values to confirm that the density estimate does in fact integrate to 1.

Density Estimator in Chicago Applying the nonparanormal density estimator to a random half of the data from Chicago (scaled and translated to have the unit cube in $\mathbb{R}^4$ as support) shows strong seasonal patterns in means, variances and covariances in $\hat{\Xi}_t$ after transformation with the time-specific quantile functions $\Phi^{-1}\left[\hat{\mathbb{F}}_{l,t}(X_{l,t})\right]$. Figure 3 shows the variances (diagonal elements of $\hat{\Xi}_t$), and some of the correlations (first column of $\hat{\Xi}_t$, scaled from covariance to correlation). Each of the curves in Figure 3 (top and bottom) represents one coordinate of the 4-dimensional vector X_j.

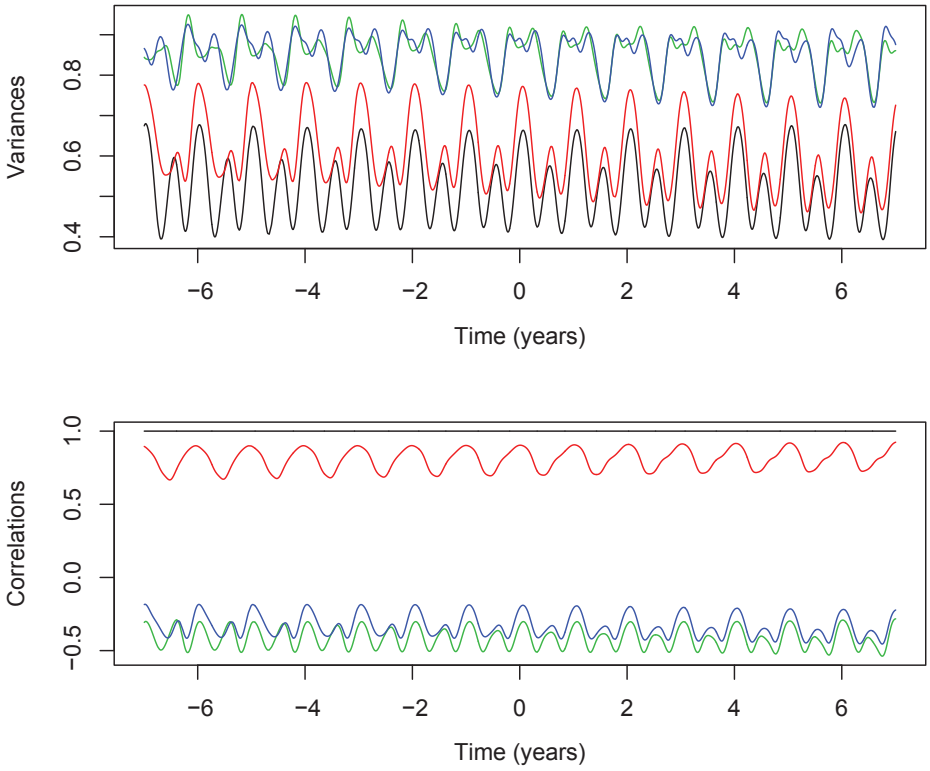

Figure 3. Latent variances and correlations for nonparanormal estimator in Chicago for 4-dimensional time series.

Checking that the density estimate gives a density, by evaluating it on uniform random samples of points, shows that although the density estimator is valid, it is not very precise. Integrating each time-specific density over the same uniformly sampled set of 2400 points (the same size as the test set used in estimation) gave results ranging from 0.6 to 4.0 when the true answer is 1.0. If $\hat{f}_t(x)$ is a density and the 4-dimensional vectors X_j are drawn from a uniform distribution, then $\frac{1}{2400}\sum_{j=1}^{2400}\hat{f}_t(X_j)$ has mean 1 and variance $\frac{1}{2400}\left[\int\left\{\hat{f}_t(x)\right\}^2 dx - 1\right]$ conditional on the training sample. Note the variance decreases as the density $\hat{f}_t(x)$ becomes closer to uniform. At the uniform the variance is 0.

To remove some of the seasonality from the variables and make the density closer to uniform we preprocessed the data matrix by the gs transformation described above (Section 2.3.3). The covariance estimator for the nonparanormal density still shows definite seasonality in Figure 4. The density is substantially less variable than without preprocessing: evaluating the densities on a uniformly sampled set of 2400 points gives integrals ranging from 0.55 to 1.16.

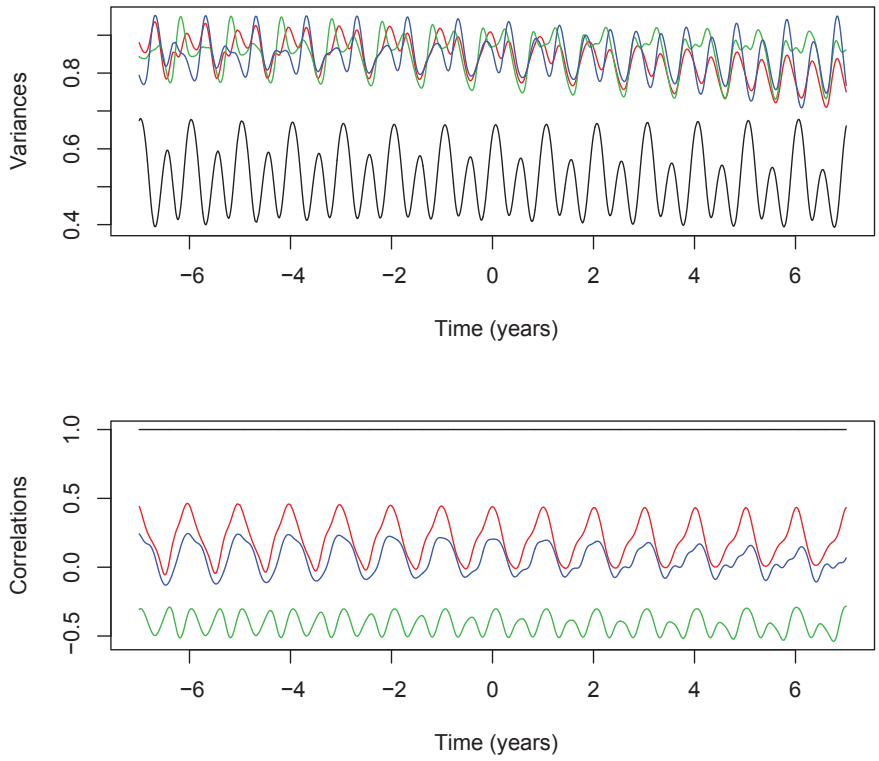

Figure 4. Latent variances and correlations for nonparanormal estimator in Chicago for 4-dimensional time series preprocessed with Gram–Schmidt orthogonalization

To further remove seasonality we applied the transformation gs2 described above. The covariance estimator for the nonparanormal density in Figure 5 shows substantially less seasonality and less correlation between the latent variables. The density is again less variable than without preprocessing: evaluating the densities on a uniformly sampled set of 2400 points gives integrals ranging from 0.96 to 1.25, with most being very close to 1.

The Locfit Density Estimator As an alternative to the nonparanormal approach, for time-series density we modified the local regression density estimator (Loader, 1999), implemented using the R package Locfit (Loader, 2010). The methods implemented there include local polynomial kernel regression for i.i.d. data. We made the following modifications to address time-series data.

As before, the variables observed include (X_1, t_1), (X_2, t_2), …, (X_N, t_N) and the goal is to estimate the density function at time t denoted by $f_t(x)$. Notice that data are observed at time t_i, but not at the time of interest t; thus, direct use of Locfit is not possible. To account for this, we weight

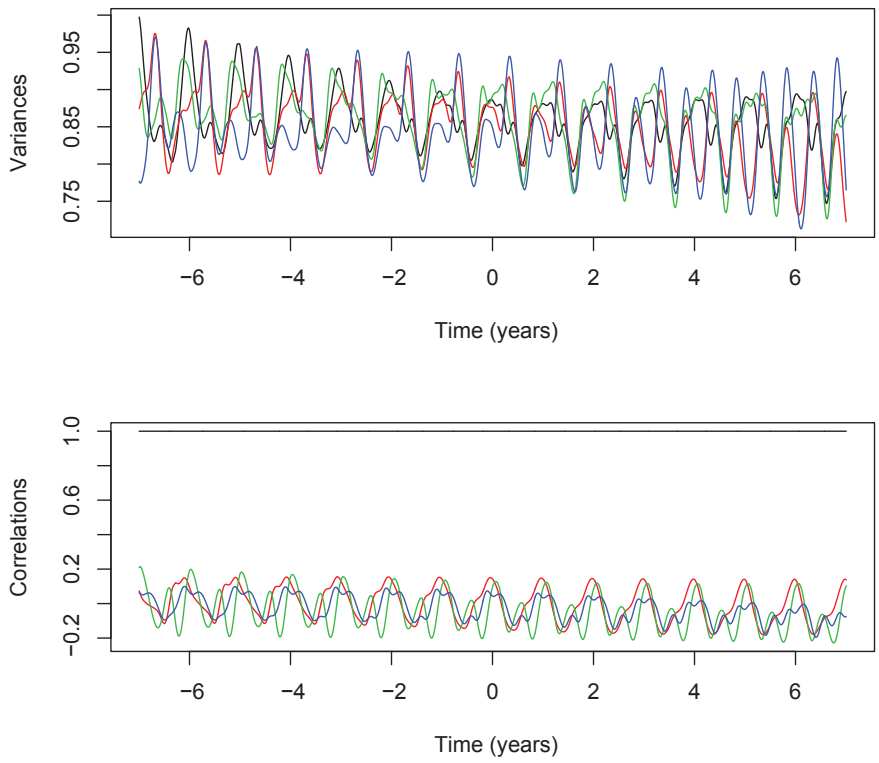

Figure 5. Latent variances and correlations for nonparanormal estimator in Chicago for 4-dimensional time series preprocessed by filtering with a 40-df cubic spline followed by Gram–Schmidt orthogonalization.

times in the same manner as in our nonparanormal approach, using the same weighting kernel as for the nonparanormal estimator:

$$w_{s,t} = K\left(\frac{s-t \bmod 365}{h_1}\right) K\left(\frac{s-t}{365 h_2}\right)$$

with $h_1 = 30$, $h_2 = 4$, and $K(\cdot)$ being the Gaussian kernel. This addresses the seasonal effect as well as the distance to the time of interest s. The idea is to utilize the local likelihood with data weighted according to the time difference. Since our goal is to obtain $f_i(X_i)$, we need to fit the Locfit model at all observed distinct times. This results in a very slow density estimation approach compared to the nonparanormal approach discussed earlier. We leave all other settings in the Locfit program at their default values. Notable ones are the following:

1. The default kernel for Locfit is

$$W(x) = (1 - |x|^3)^3 \text{ for } |x| < 1$$

and weights are calculated as

$$W\left(\frac{\|x_i - x\|}{h(x)}\right).$$

2. A local quadratic polynomial approximation is used when constructing the local likelihood function.
3. The local weight $h(x)$ is computed as follows:
   (a) $k = \lfloor n\alpha_0 \rfloor$ (by default $\alpha_0 = 0.7$);
   (b) compute $d_i = \|x - x_i\|$ and find the k-th smallest $d_{(k)}$;
   (c) $h(x) = d_{(k)}$.

Comparison of Density Estimators From Figure 6 (bottom panel) we observe that, for the gs2 transformation, results for time-series versions of Locfit and the nonparanormal were essentially identical. From Figure 6 (top panel) we observe that, for the gs transformation, the results based on Locfit differed somewhat from those based on the nonparanormal, although the substantive results were similar.

The Locfit estimator is more nonparametric than the nonparanormal estimator, as the latter allows 6 df (the 6 correlation parameters associated with the 4-dimensional vector X_{cont}). On the other hand the nonparanormal estimator is much more stable than the Locfit estimator because it less nonparametric. Thus, it is reassuring that both density estimators gave similar results.

As detailed in Section 3.4.3, a goodness-of-fit test for the density estimator can be constructed by defining

$$\widehat{\mathbb{IF}}^{\pm}_{1,\text{new}} = \frac{1}{2n}\left\{\sum_{i\in\text{split}(0)}\frac{1}{s(i)}\sum_{j\in s(i)}\hat{f}_j(X_i) + \sum_{i\in\text{split}(1)}\frac{1}{s(i)}\sum_{j\in s(i)}\hat{f}_j(X_i)\right\}$$

$$\widehat{\mathbb{IF}}^{\pm(k)}_{22} = -\frac{1}{2n}\left\{\sum_{i\in\text{split}(0)}\frac{1}{s(i)}\sum_{j\in s(i)}K_k(X_i,X_j) + \sum_{i\in\text{split}(1)}\frac{1}{s(i)}\sum_{j\in s(i)}K_k(X_i,X_j)\right\}$$

and letting $\widehat{\mathbb{IF}}^{\pm(k)}_2 = \widehat{\mathbb{IF}}^{\pm}_{1,\text{new}} + \widehat{\mathbb{IF}}^{\pm(k)}_{22}$. The goodness-of-fit statistic is then $\widehat{\mathbb{IF}}^{\pm(k)}_2 / \left\{\widehat{\mathbb{IF}}^{\pm}_{1,\text{new}} - \widehat{\mathbb{IF}}^{\pm(k)}_2\right\}$ with smaller values being better. The results are shown in Table 6.

Since we used the percentile cut-off 80 as our base case, we also provide results for that case in Table 6. Reading from Table 6, we see that for Legendre with an 80% cut-off (the base-case choices), the results for Locfit and the nonparanormal were essentially equivalent. Small absolute

Table 6. Goodness-of-fit for nonparanormal versus Locfit density estimation.

Percent Cut-Off	Basis	Transform	Density Estimator	$\widehat{\mathbb{IF}}^{\pm}_{1,\text{new}} - \widehat{\mathbb{IF}}^{(k),\pm}_2$	$\widehat{\mathbb{IF}}^{\pm}_{1,\text{new}}$	$\widehat{\mathbb{IF}}^{(k),\pm}_2$	$\frac{\widehat{\mathbb{IF}}^{(k),\pm}_2}{\widehat{\mathbb{IF}}^{\pm}_{1,\text{new}} - \widehat{\mathbb{IF}}^{(k),\pm}_2}$
100%	Legendre	gs	NPN	4.16	6.53	2.37	0.570
			Locfit	4.16	3.37	−0.79	−0.190
		gs2	NPN	17.73	20.53	2.80	0.158
			Locfit	17.73	18.19	0.47	0.026
80%	Legendre	gs	NPN	2.08	2.77	0.69	0.332
			Locfit	2.17	1.57	−0.60	−0.277
		gs2	NPN	12.75	12.73	−0.018	−0.001
			Locfit	12.14	11.95	−0.188	−0.015
	Haar	gs	NPN	1.58	2.77	1.19	0.753
			Locfit	3.16	1.57	−1.59	−0.503
		gs2	NPN	11.92	12.73	0.813	0.068
			Locfit	9.10	11.95	2.85	0.313

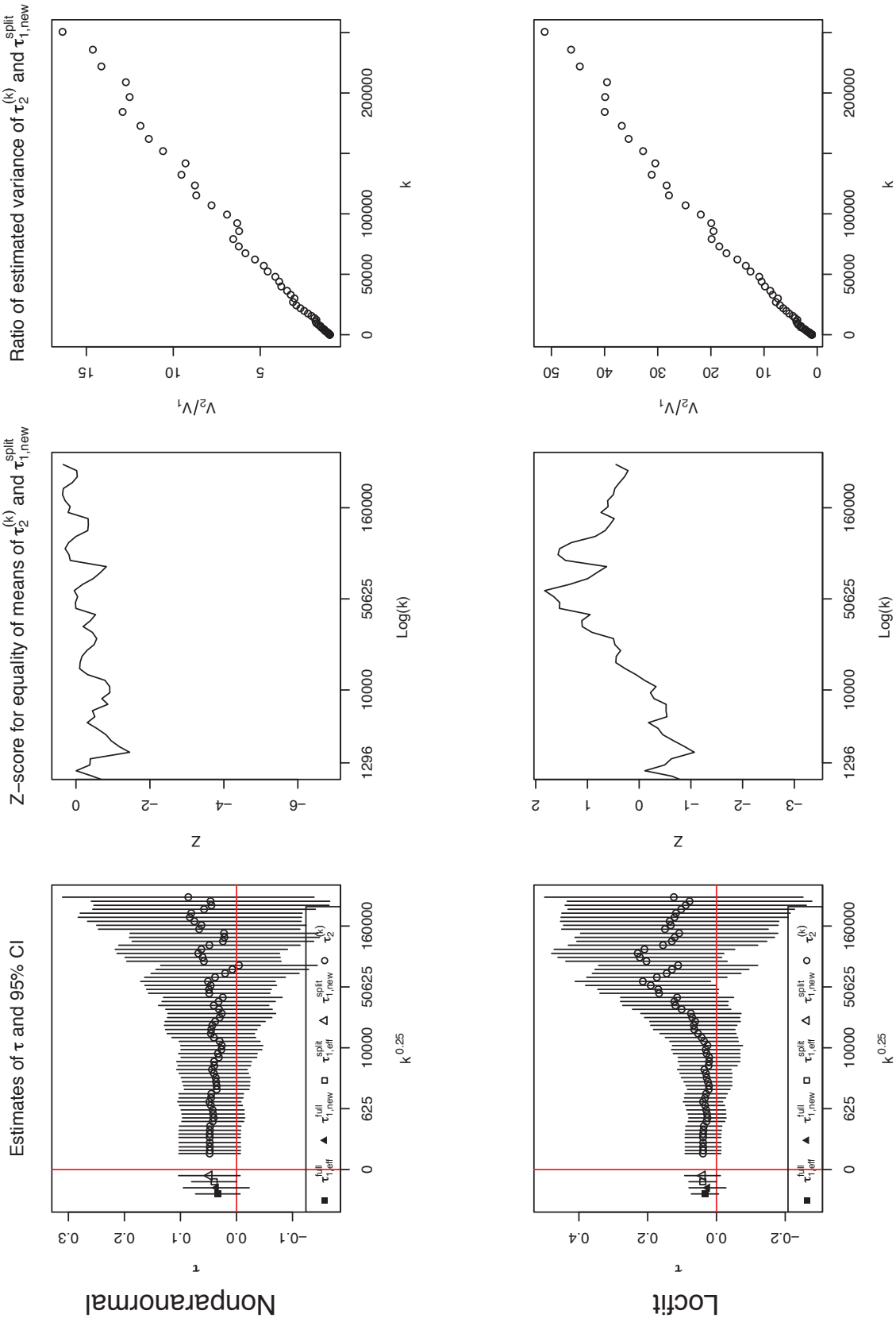

Figure 6. Local polynomial vs. nonparanormal density estimation with gs transformation. Summary of the analysis using a linear model with Legendre polynomials, data transformation using gs (top two rows of panels) or gs2 (bottom two rows), a density cutoff at the 80th percentile, observations between 25 and 75 days of a given day, and with k between 3 and 250563.

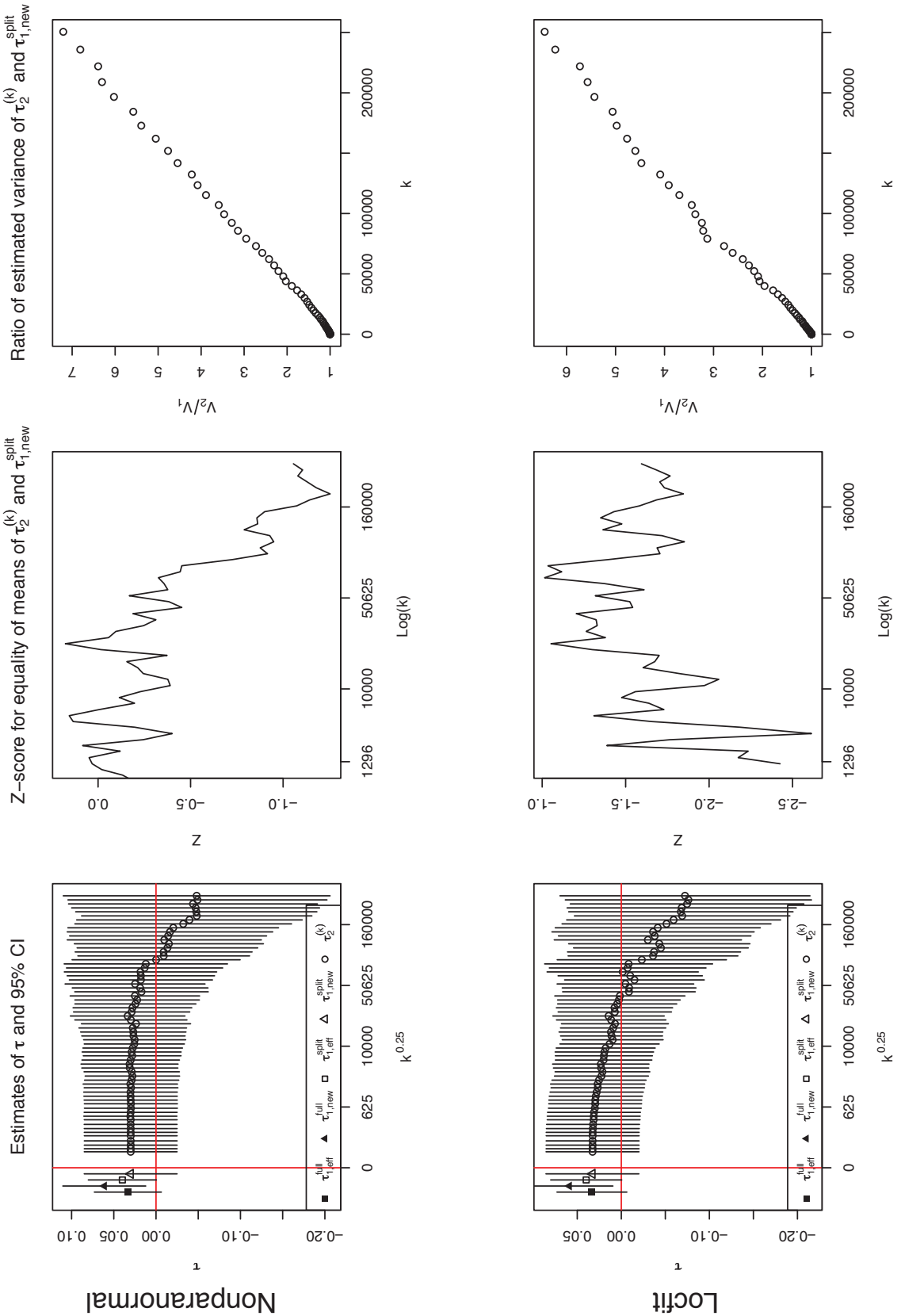

Figure 6. (continued)

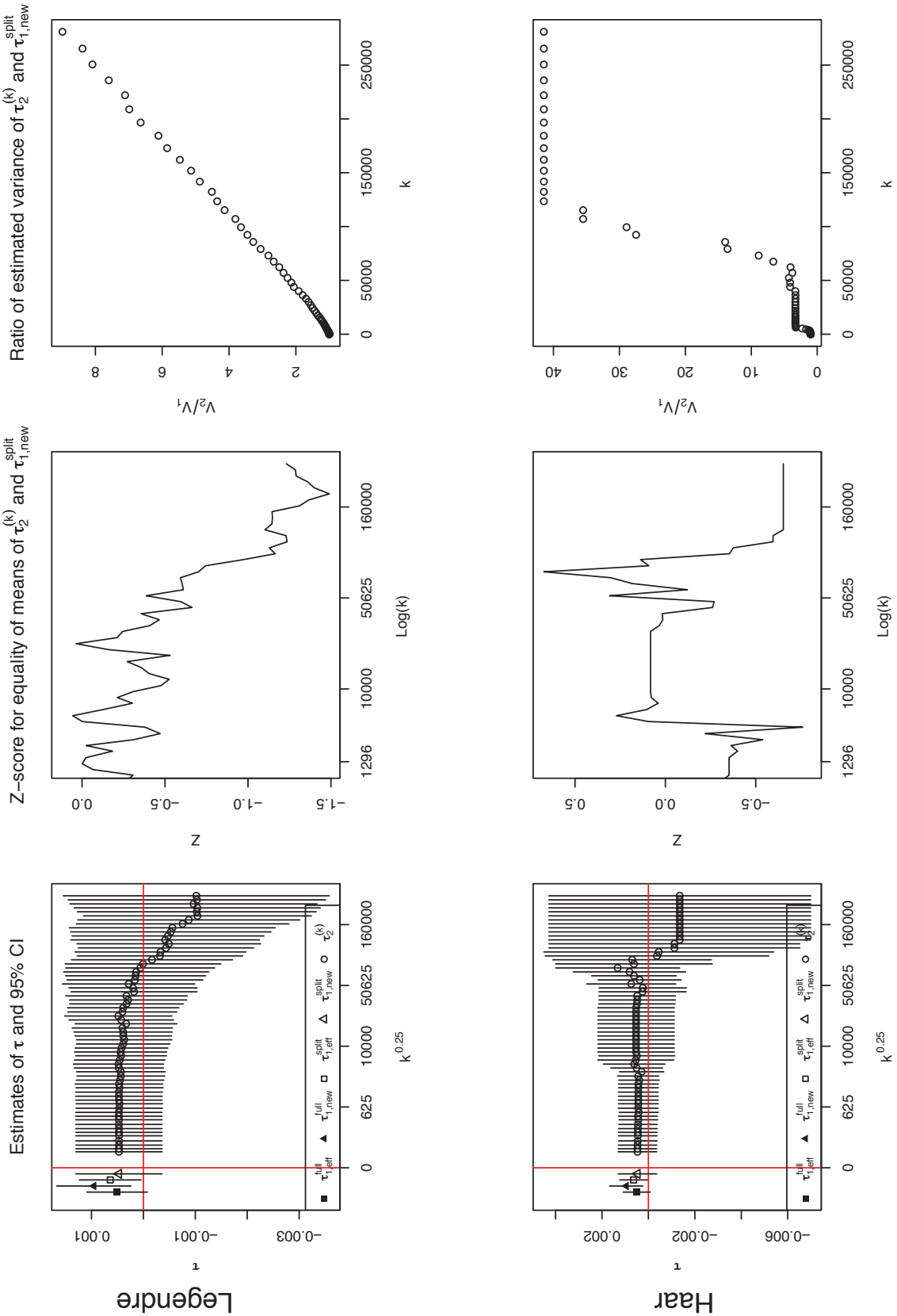

Figure 7. Harr wavelets vs. Legendre wavelets. Summary of analyses using models with both Legendre polynomials and Haar polynomials, data transformation using gs2, a density cutoff at the 80th percentile, and with k between 3 and 250563. The top two rows used a loglinear model with observations between 25 and 75 days of a given day; the bottom two rows used a linear model with observations between 25 and 3000 days of a given day.

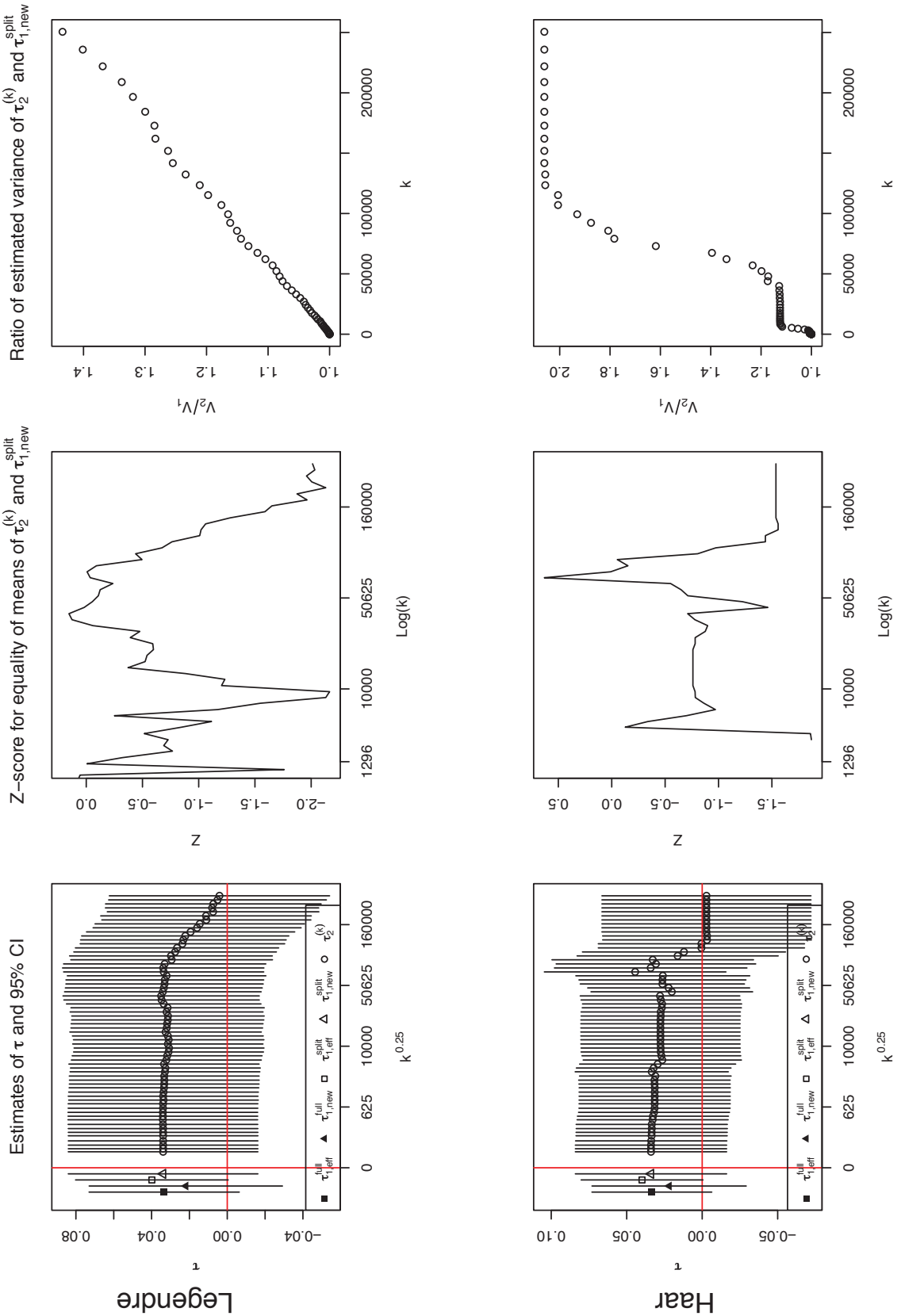

Figure 7. (continued)

values for the goodness-of-fit statistic imply a good fit. For both nonparanormal and Locfit, the fit under gs2 was excellent. For a cut-off of 100%, Locfit outperformed the nonparanormal. For the Haar basis the performances were nearly equal.

2.3.5 Sensitivity to Haar vs. Legendre vs. Daubechies Compact Wavelets

Figure 7 compares results using Haar bases to those using Legendre bases (Vidakovic, 1999). We observe for Haar, but not for Legendre, that as k increases, both the estimates in the left graphs and the variance ratios in the right graphs are constant for large stretches. Furthermore, for most values of k (especially moderate k) the variance ratio and thus the variance of $\widehat{\mathbb{IF}}_2^{\text{split},(k)}$ based on Haar exceeds that based on Legendre. In additional simulations (data not shown) we found that, for any k, the truncation bias for our second-order estimator based on Haar wavelets was much greater than that based on Legendre when one of the functions $b_i^*(x)$ and $p_i^*(x)$ had more than one derivative. Thus Legendre wavelets performed better than Haar both in terms of variance and truncation bias. Furthermore, in simulations, the Legendre truncation bias was much smaller than the truncation bias of order 20 Daubechies compact wavelets, whereas the variances were nearly identical.

The Haar estimates were constant for long stretches as k increased, because when, in the tensor product, we added a mother Haar wavelet for a particular variable, it may be that none of the pairs of observations that were in the same "bin" before the addition were both still in the new smaller bin formed by adding the new wavelet. If so, as further wavelets (corresponding to the other variables in the tensor product at that level) are added, neither the statistic nor the variance estimator will change.

Compared to Legendre, the large truncation bias of Haar when $b_i^*(x)$ or $p_i^*(x)$ is due to the well-known fact that Haar does not give optimal rates of approximation as a function of k when the function being approximated has more than one derivative. Legendre wavelets have smaller truncation bias than higher-order Daubechies because the Daubechies wavelets do not form an orthonormal basis for functions with support on the unit interval. There do exist modified Daubechies wavelets that are orthonormal, but they are difficult to compute and the required software does not seem to be easily available.

2.3.6 Sensitivity to the Choice of X_{cont} and of the Linear Spline Models

In our base case, X_{cont} was represented by a 4-dimensional vector composed of temperature and dew-point variables as follows: (1) average temperature on day t_i, (2) dew-point temperature on day t_i, (3) adjusted 3-day lagged daily temperature, and (4) adjusted 3-day lagged dew-point temperature (variables used in previous analyses of the NMMAPS data). We also considered a 6-dimensional vector of temperature and dew-point variables for X_{cont} (described below) that we believed might better control confounding due to weather.

Furthermore, in our base-case analyses, we followed previous NMMAPS analyses and used regression splines with 6 df for each of the 4 weather variables in X_{cont} (without between-variable interaction terms) to estimate the nuisance regression functions $b_i^*(x_{\text{cont}})$ and $p_i^*(x_{\text{cont}})$. As an alternative, we also considered a variety of machine learning methods to estimate $b_i^*(x_{\text{cont}})$ and $p_i^*(x_{\text{cont}})$ both when X_{cont} was a 4-dimensional and a 6-dimensional vector. These methods allow for interactions between the variables in X_{cont}. Remarkably, none of the machine learning estimates of $b_i^*(x_{\text{cont}})$ and $p_i^*(x_{\text{cont}})$ did substantially better than our base-case NMMAPS estimates on

a test set. Furthermore, we also show that the 6-dimensional variable vector version of X_{cont} did not do substantially better than the base-case 4-dimensional vector used in previous NMMAPS analyses. Thus we chose as our base case both X_{cont} and the spline models used in previous NMMAPS analyses because (1) this choice made the comparison of our estimates with those obtained in earlier NMMAPS analyses more straightforward, and (2) other choices did not perform substantially better. Details of the model selection procedure for the spline models are provided in Appendix B.

2.4 DISCUSSION AND CONCLUSIONS

Reanalyses of NMMAPS city-specific time-series data for the 22 largest NMMAPS cities using the higher-order influence-function estmiators provide no evidence that the original NMMAPS estimates of the effect of PM_{10} on all-cause mortality were biased, except possibly in Minneapolis, and even there the estimated bias was small and did not change any substantive conclusions.

We obtained wider CIs than did the NMMAPS investigators. This increase in CI width was to be expected as our approach makes weaker assumptions than does the original NMMAPS approach concerning the smoothness of the dose–response function for the effect of temperature and humidity on daily mortality and on PM_{10}.

Our estimators depend on a number of user-supplied choices. In a sensitivity analysis, we showed that our results were remarkably robust to these choices.

An important limitation of our approach, however, is that its validity depends critically on the assumptions discussed in Section 2.2.2. Another apparent limitation is that, in our main analyses, we chose the same independent variables X_i and the same number of spline degrees of freedom as did the NMMAPS investigators. We had not originally planned to do so. Rather we originally (1) included additional temperature and humidity variables in X_i to check whether the five NMMAPS variables were sufficient, and (2) used flexible machine learning approaches to find improved estimators of the regression of mortality and PM_{10} on the covariates X_i when compared to the NMMAPS spline model. However, as discussed in section 2.3.6, we suprisingly found that neither (1) nor (2) above predicted the association between mortality and PM_{10} better than the NMMAPS approach (as evaluated by cross validation). Thus we opted for a direct comparison between the first-order NMMAPS Poisson regression estimators and our second-order estimators to evaluate whether our second-order estimator alone would result in conclusions that differed from those of the NMMAPS investigators.

Our approach only provides evidence that the NMMAPS analyses were robust to confounding by temperature and humidity attributable to possible wiggliness of the regression functions of mortality and PM_{10} on the covariates X_i. As with any observational study, it is always possible that there remains important residual confounding by unmeasured covariates. We did not examine the possible confounding effect on daily mortality of temperature more than 10 days before the recorded date of death. In light of recent data on long term effects of temperature on daily mortality, it would be prudent in future analyses to include in the covariate vector X_i summary measures for temperature more than 10 days before the recorded date of death.

Although we limited our investigation to time-series analyses of the short-term effect of PM_{10} on mortality, our second-order methods could be used to decrease residual confounding bias attributable to continuous covariates such as a smoking, temperature, and income in cohort studies of the chronic effects of pollution. Indeed our methods should work better in such a setting be-

cause, in contrast to time-series studies, the uncorrectable bias due to the nonrandom nature of time would not be present. In addition, our variance estimates are not dependent on assumptions about the degree of correlation at nearby times.

In summary, we found no evidence that policy decisions based on previous NMMAPS analyses need to be reconsidered. The theoretical underpinnings of our results are further described in Section 3 and the accompanying appendices.

3. THEORY

3.1 INTRODUCTION

In recent years, time-series epidemiological studies have used linear or loglinear semiparametric models for the effect of particulate matter suspended in the air on mortality (Dominici et al., 2004). A linear or loglinear semiparametric model for this effect may be written as

$$\Phi(E_i[Y_i|A_i,X_i]) = \tau^* A_i + \zeta_i^*(X),$$

Since we assume the data have one observation for each time point, the subscript i refers to the i-th subject and also the corresponding time; Y_i is the measured number of deaths; A_i is the observed level of particulate matter in the air, for example, PM_{10}; X_i refers to covariates used to control confounding for the effect of A_i on Y_i; $\zeta_i^* : \mathbb{R}^d \to \mathbb{R}$ is an unknown function that may depend on the the time corresponding to observation i; and τ^* encodes the magnitude of the effect of pollution on mortality on the scale Φ. Here Y_i is a scalar, A_i is a nonnegative scalar, and X_i is a vector in $\mathbb{R}^d$ for some $d \geq 1$. The function Φ is a link function, equal to the identity for the linear and to the logarithm for the loglinear model. It follows that the expectation $E_i[Y_i|A_i,X_i]$ also depends on time; thus the expectation operator is also subscripted by i. A similar remark applies to variances.

Standard estimators of τ^* (1) model $\zeta_i^*(X_i)$ as a linear function of a vector of covariates $M_i = m_i(X_i)$ depending on the vector of covariates X_i and time; and (2) fit the resulting parametric model by ordinary least-squares regression in the linear case and by Poisson regression in the loglinear case. As will be shown, these approaches are equivalent to solving the efficient influence-function equation if $\mathrm{Var}_i[Y_i|A_i,X_i]$ is constant in the linear case and is proportional to $E_i[Y_i|A_i,X_i]$ in the loglinear case. To simplify notation, we suppress the subscript i until Section 3.3.

In Section 3.2 we show that the efficient influence function depends on two unknown functions $b^*(X)$ and $p^*(X)$ of the distribution of the data. Let β_b and β_p denote the maximum number of times $b^*(X)$ and $p^*(X)$ are differentiable. The statistical properties of any estimator of τ^* depend on β_b and β_p. Specifically, we show that $\sqrt{n}$-consistent estimators of τ^* exist only if $\beta_b + \beta_p > d/2$ (Donald and Newey, 1994; Robins, Tchetgen, et al., 2009). Failing this, the rate of decrease of bias with sample size will be slower than the optimal rate, and Wald CIs based on the standard estimators (Poisson and linear regression) and a consistent estimate of their standard error are invalid.

It is known that β_b and β_p cannot be estimated from the data. This raises the following questions: (1) how can we empirically determine whether nominal $1-\alpha$ Wald CIs centered on the standard estimators have a coverage probability of at least $1-\alpha$? and (2) can we find estimators

of τ^* with MSE less than that of the standard estimators when these estimators fail to converge at a $\sqrt{n}$ rate?

We will show that, under regularity conditions, estimators based on higher-order influence functions (Robins, Li, et al., 2008) both (1) can be used to test with reasonable power the hypothesis that the bias of the standard estimators is less than their standard error, and (2) have asymptotic MSE less than that of standard estimators when the standard estimators fail to converge at $\sqrt{n}$ rate.

Higher-order influence functions are higher-order U-statistics. We will consider different higher-order influence-function estimators. All three estimators require estimates of the joint density of X; the first estimator requires the estimates $\hat{f}(X)$ to be both smooth and bounded away from zero, a condition not easily achievable for multivariate density estimates. The second estimator does not require this but is often computationally intractable. The third estimator suffers from neither of these drawbacks, but may be less efficient than the second estimator.

The time-series nature of the data, which implies that observations are not independent, introduces new challenges for all three estimators. To overcome these challenges we impose conditions on the dependence between the observations, and then attempt to minimize these conditions. However, with time-series data there is a component of the bias of the standard estimators that cannot be reduced using higher-order estimators; hence we will need to assume that the functions b^* and p^* and their estimators are smooth functions of time (though they need not smooth in X).

To understand which of the modeling assumptions are critical and how far we can go without making them, we adopt the following approach. We consider the nonparametric model and redefine τ^* to be the nonparametric functional to which the standard estimators converge under possible misspecification of the semiparametric regression model. Then, as far as possible, we derive results without further modeling assumptions. With time-series data we find there are limitations to this approach that do not occur with i.i.d. data. We must therefore impose additional restrictions, including the restriction that the semiparametric model holds.

The remainder of this section is an overview of the rest of the report. In Section 3.2, we study first- and second-order inference based on the standard (first-order) estimators and on our second-order influence-function (U-statistic) estimators in the i.i.d. setting. In Section 3.3 we generalize to time-series analyses.

In Section 3.2 we define the parameter τ^* in the nonparametric model. In Section 3.2.1 we introduce a common notation that helps define efficient first-order influence functions and estimators of τ^* based on these influence functions in both the linear and loglinear cases. An important consideration is relative size of the bias and the variance of these estimators. We derive bias formulae for the first-order estimators; the variance for the first-order estimators follows from standard results for M-estimators (van der Vaart, 1998). Based on these formulae, in Section 3.2.2 we describe rates of convergence for the bias and variance of these efficient first-order estimators as a function of β_b and β_p and the smoothness of various nuisance functions. It turns out that unless β_b and β_p are sufficiently large, the squared bias is larger than the variance of the efficient first-order estimators. This leads to invalid inference, specifically with regard to CIs. The first-order estimators can be improved with the use of second-order influence functions.

We consider three second-order influence-function estimators. To do so we define an additional density-weighted first-order influence-function estimator in Section 3.2.3, since this is need-

ed to define the third second-order influence-function estimators. The second-order influence-function estimators are defined in Section 3.2.4. We derive formulae for the biases of the second-order estimators and show that these do indeed reduce the bias to third order. Of the three higher-order influence-function estimators, the second has especially attractive asymptotic properties, but was infeasible for computational reasons. The first estimator depended on the reciprocal of an estimated density and was thus numerically unstable. The third estimator, which reduces the bias of the density-weighted first-order estimator, does not suffer from these problems. In the remainder of the report we are primarily concerned with this third method. We defer discussion of the rates of convergence of these second-order estimators until after the time-series version of these second-order estimators is defined to avoid repetition.

Section 3.3 covers the time-series case, in which the observations are no longer independent or uncorrelated. Because of this complication, the definition of the parameter τ^* needs modification. The joint distribution, density, and nuisance functions are all different for each observation and must be indexed by time through the observation number i. We define both first- and second-order estimators in Sections 3.3 and 3.3.1. A precise formula for the bias and variance of the time-series second-order estimators requires additional assumptions, which we provide in Section 3.3.2. The best achievable rates for the bias of our second-order estimators are products of the optimal rates for the nuisance functions b^* and p^* in the time-series case, whereas in the i.i.d. case this is multiplied by the optimal rate for the density f. We describe these properties and their consequences for the choice of k — a truncation parameter used in the second-order influence-function estimators — in Section 3.3.3.

The choice of k depends on knowledge of the smoothnesses of various functions, or on assumptions made about these smoothnesses. Choosing k in the absence of knowledge about the smoothness of b and p is an open problem. However, we can test whether the first-order estimator has a squared bias exceeding the variance. We describe this test in Section 3.3.4.

The first- and second-order estimators depend on various user-chosen quantities, such as the choice of density estimator and choice of a set $s(i)$, to be defined later. Results of the consequences of these choices and ways of determining them are presented in Section 3.4. Section 2.4 presents conclusions and discussion.

Additional results, simulations, and proofs are presented in the appendices. Appendix C gives the results of simulations based on the NMMAPS data set comparing the relative biases and variances of the three types of estimator — full-sample, split-sample, and half-sample — considered in this report. Appendix D details an estimator for which we have a proof demonstrating that it achieves a $\sqrt{n}$ rate for the loglinear model when $\beta_b + \beta_p > d/2$; this estimator is not a time-series Poisson regression estimator. Finally, Appendix E collects the proofs of the theorems presented in the report.

3.2 ESTIMATION IN THE I.I.D. CASE

As noted above, we strive for greater generality by working in the nonparametric model, for which a definition of τ is needed. We define τ as the solution to an estimating equation chosen such that, when the semiparametric model holds, the definition of the parameter is consistent with the definition in the semiparametric model. The definitions in the i.i.d. case motivate those in the time-series case. In the i.i.d. case the model can be expressed as

$$\Phi(E[Y|A,X]) = \tau^* A + \zeta^*(X),$$

where, similar to the time-series case, Y is the measured number of deaths, A is the observed level of particulate matter in the air (say PM_{10}), X refers to covariates used to control confounding for the effect of A on Y, $\zeta^* : \mathbb{R}^d \to \mathbb{R}$ is an unknown function, and τ^* encodes the magnitude of the effect of pollution on mortality on the scale Φ. Recall that Y is a scalar, A is a nonnegative scalar, and X is a vector in $\mathbb{R}^d$ for some $d \geq 1$. The function Φ is a link function, equal to the identity for the linear model and to the logarithm for the loglinear model.

In the sequel, the linear and loglinear cases refer not to the model itself (since we are assuming the nonparametric model here) but to the definition of τ desired: one that is compatible with the linear model, or one that is compatible with the loglinear model, respectively. Either of the two definitions below is compatible with the corresponding model when the model holds.

Definition 1. *We define the parameter τ^* in two possible ways. The first is as the solution to the equation*

$$E\left[\left(Y - \tau A - E[Y - \tau A|X]\right)\left(A - E[A|X]\right)\right] = 0 \quad \textit{in the linear case, and}$$
$$E\left[\left(Y - \frac{E[Y|X]}{E[e^{\tau A}|X]}\right)\left(A - \frac{E[Ae^{\tau A}|X]}{E[e^{\tau A}|X]}\right)\right] = 0 \quad \textit{in the loglinear case.}$$

The second is

$$E\left[f(X)\left(Y - \tau A - E[Y - \tau A|X]\right)\left(A - E[A|X]\right)\right] = 0 \quad \textit{in the linear case, and}$$
$$E\left[f(X)\left(Y - \frac{E[Y|X]}{E[e^{\tau A}|X]}\right)\left(A - \frac{E[Ae^{\tau A}|X]}{E[e^{\tau A}|X]}\right)\right] = 0 \quad \textit{in the loglinear case}$$

where $f(x)$ is a fixed, known density defined on the support of X (not the true, unknown density f^* of X).

The reasons and choices for the weighting density $f(X)$ in the second definition of τ will be made clear in subsection 3.2.3. The other terms are motivated by the form of the efficient influence functions in the semiparametric models, as shown in Section 3.2.1.

3.2.1 First-Order Influence Functions and the Associated Estimators

Suppose we observe the i.i.d. random sample $(Y_i, A_i, X_i), i = 1, \ldots, N$. Following Bickel et al. (1993), consider a model $\mathcal{M} = \{P\}$. Let $P^* \in \mathcal{M}$ be the true distribution of Y, A, X. The pathwise differentiable parameter τ is a functional $\tau : \mathcal{M} \to \mathbb{R}$. In a model, an influence function for a regular parameter $\tau^* = \tau(P^*)$ in a model is an element $\mathrm{IF}_1(Y, A, X, P^*) \in \mathcal{L}_2(P^*)$ satisfying

$$\left.\frac{d\tau(\theta)}{d\theta}\right|_{\theta=\theta^*} = E_{\theta^*}\left[\mathrm{IF}_1(Y, A, X; P^*) S_\theta(Y, A, X; \theta^*)\right]$$

for every regular parametric submodel $\{P_\theta\}$, indexed by θ, of the model. Here S_θ is the score for θ in the submodel. We assume that $P_{\theta^*} = P^*$ for some θ^*, the "true" θ for the submodel. The efficient influence function is the influence function with the smallest variance.

Theorem 2. *Consider the semiparametric model*

$$E\left[\xi(Y, A, X; \tau)|A, X\right] = E\left[\xi(Y, A, X; \tau)|X\right]$$

where ξ is a known function. Let

$$\Delta\xi(Y, A, X; \tau) = \xi(Y, A, X; \tau) - E\left[\xi(Y, A, X; \tau)|X\right],$$

$$J(A,X;P^*) = \frac{d}{d\tau} E\left[\Delta\xi(Y,A,X;\tau)|A,X\right]\Big|_{\tau=\tau^*},$$
$$\tilde{J}(X;P^*) = E\Big[J(A,X;P^*)\,\mathrm{Var}^{-1}\left[\Delta\xi(Y,A,X;\tau^*)|A,X\right]|X\Big] \times E\Big[\mathrm{Var}^{-1}\left[\Delta\xi(Y,A,X;\tau^*)|A,X\right]|X\Big]^{-1}.$$

The influence functions for τ in the model are

$$\Big\{-E\left[J(A,X;P^*)\;\Delta h(A,X;P^*)\right]^{-1}\Delta\xi(Y,A,X;\tau^*)\;\Delta h(A,X,P^*) : h(A,X)\in\mathscr{L}_2(P^*) \quad E\left[J(A,X;P^*)\;\Delta h(A,X;P^*)\right]\Big|_{\theta=\theta^*}\neq 0\Big\}.$$

The efficient influence function is given by

$$\mathrm{IF}_{1,\mathrm{eff}}(Y,A,X;P^*) = \left(J(A,X;P^*) - \tilde{J}(X;P^*)\right)\;\mathrm{Var}_{\theta^*}^{-1}\left[\Delta\xi(Y,A,X;\tau^*)|A,X\right]\Delta\xi(Y,A,X;\tau^*).$$

Proofs of theorems, corollaries and lemmas are in Appendix E. The loglinear and linear semiparametric regression models are special cases of the above model with $\xi(Y,A,X;\tau) = Ye^{-\tau A}$ (loglinear) and $Y-\tau A$ (linear). Hence the theorem leads to the following corollaries.

Corollary 3. *Under the assumption of homoscedasticity,*

$$\mathrm{Var}[Y|A,X] = \sigma^2 \text{ for all } A,X,$$

the efficient influence function in the linear semiparametric model

$$E[Y|A,X] = \tau A + \zeta^*(X)$$

is

$$\mathrm{IF}_{1,\mathrm{eff,linear}}(Y,A,X;P^*) = \sigma^{-2}\left\{Y-\tau^*A - E\left[Y-\tau^*A|X\right]\right\}\left\{A-p_\tau^*(X)\right\},$$

where $p_\tau^*(X) = E[A|X]$, which does not depend on τ in the linear case.

Corollary 4. *Under the (possibly overdispersed) conditional Poisson variance assumption,*

$$\mathrm{Var}[Y|A,X] = \sigma^2 e^{\tau^*A+\zeta^*(X)} \text{ for all } A,X,$$

the efficient influence function in the loglinear semiparametric model

$$\log E[Y|A,X] = \tau^*A + \zeta^*(X)$$

is

$$\mathrm{IF}_{1,\mathrm{eff,loglinear}}(Y,A,X;P^*) = \sigma^{-2}\left[A-p_{\tau^*}^*(X)\right]\left\{Y-e^{\tau^*A+\zeta^*(X)}\right\},$$

where $p_\tau^*(X) = \frac{E[Ae^{\tau A}|X]}{E[e^{\tau A}|X]}$.

One approach to inference is to obtain estimators of τ^* by solving estimating equations based on the above first-order influence functions. Note, however, that the above influence functions themselves depend on functions of the distribution of the data, such as p_τ^* and $E[Y-\tau^*A]$. In order to use these influence functions, such functionals must be estimated.

We will need a considerable amount of notation in order to define the first-order estimators; the same notation will later help in defining the higher-order estimators. To avoid repetition we often use the same notation for the linear and loglinear cases. For brevity, we will suppress the subscripts "linear" and "loglinear" if they are clear from the context or if both are applicable. Many of our definitions and results are applicable in both cases; where necessary, we clarify which case is intended.

We start by letting $w : \mathbb{R}^d \to \mathbb{R}^{\bar{d}}$ be a fixed transform of the covariates X (typically $\bar{d} > d$). Let $W = w(X)$. W will be used to model various nuisance functions. We need the following definitions.

Definition 5. *In the loglinear case define*

$$\begin{aligned}
b^*_\tau(X) &= \frac{E[Y|X]}{E\left[e^{\tau A}|X\right]},\\
p^*_\tau(X) &= \frac{E[Ae^{\tau A}|X]}{E[e^{\tau A}|X]},\\
q^*_\tau(X) &= E[e^{\tau A}|X],\\
\varepsilon_i(\tau, b) &= Y_i - e^{\tau A_i} b(X_i),\\
q(x;\omega) &= \exp\left(\omega^T w(x)\right),\\
b(x;\eta) &= \exp\left(\eta^T w(x)\right),\\
\Delta_i(\tau, p, q) &= \left(A_i - p(X_i)\right)\frac{e^{\tau A_i}}{q(X_i)}, \text{ and}\\
S_i(\tau, \alpha) &= (A_i - \alpha^T W_i) W_i e^{\tau A_i}.
\end{aligned}$$

Definition 6. *In the linear case, define*

$$\begin{aligned}
b^*_\tau(X) &= E[Y - \tau A|X],\\
p^*_\tau(X) &= E[A|X],\\
q^*_\tau(X) &= 1,\\
\varepsilon_i(\tau, b) &= Y_i - \tau A_i - b(X_i),\\
b(x;\eta) &= \eta^T w(x),\\
q(x;\omega) &= 1,\\
S_i(\tau, \alpha) &= (A_i - \alpha^T W_i) W_i, \text{ and}\\
\Delta_i(\tau, p, q) &= A_i - p(X_i).
\end{aligned}$$

Note that ε is defined in terms of an arbitrary function $b : \mathbb{R}^d \to \mathbb{R}$, not necessarily b^*_τ. Similarly, $\Delta_i(\tau, p, q)$ is defined in terms of arbitrary functions $p, q : \mathbb{R}^d \to \mathbb{R}$. $\Delta_i(\tau, p, q)$, p^*_τ, and $S(\tau, \alpha)$ depend on τ in the loglinear case but not in the linear case. These definitions do not depend on the validity of either the linear or loglinear model. However, in the loglinear case $b^*_{\tau^*}(X) = e^{\zeta^*(X)}$ when the loglinear model holds, and in the linear case $b^*_{\tau^*}(X) = \zeta^*(X)$ when the linear model holds.

Definition 7. *Common to both the linear and loglinear models, define*

$$\Delta_i(p) = A_i - p(X_i),$$

$$
\begin{aligned}
U_{i,\text{profile}}(\tau, b) &= \varepsilon_i(\tau, b) A_i, \\
U_{i,\text{nuis}}(\tau, b) &= \varepsilon_i(\tau, b) W_i, \\
U_i(\tau, b) &= (A_i, W_i)^T \varepsilon_i(\tau, b), \text{ and} \\
\text{IF}_{1,\text{eff},i}(\tau; b, p) &= \varepsilon_i(\tau, b)(A_i - p[X_i]).
\end{aligned}
$$

Before we define the estimators using this notation, we note that our estimators can be divided into full-sample, split-sample, and half-sample estimators. In full-sample estimators, the nuisance functions as well as the estimating equation for τ are based on the full sample of N observations. The half-sample and split-sample estimators divide the sample into two halves of size n each (so $N = 2n$) and will be described shortly.

The estimators $\hat{\tau}^{\text{full}}_{1,\text{eff}}$ (for the linear and loglinear cases) are now defined as the solutions to the equations

$$
\frac{1}{N} \sum_{i=1}^{N} \text{IF}_{1,\text{eff},i}(\tau; \hat{b}^{\text{full}}_{\tau}, \hat{p}^{\text{full}}_{\tau}) = 0,
$$

where $\hat{b}^{\text{full}}_{\tau}$ and $\hat{p}^{\text{full}}_{\tau}$ are estimators of the functions b^*_τ and p^*_τ based on the parametric models $b(X_i; \eta)$ and $\alpha^T w(X_i)$, respectively.

Formally, we have the following definition for our full-sample estimators.

Definition 8. *Let $\hat{\eta}^{\text{full}}(\tau)$ and $\hat{\alpha}^{\text{full}}(\tau)$ solve*

$$
\begin{aligned}
N^{-1} \sum_{i=1}^{N} U_{i,\text{nuis}}\left[\tau, b(\eta)\right] &= 0, \text{ and} \\
N^{-1} \sum_{i=1}^{N} S_i(\tau, \alpha) &= 0,
\end{aligned}
$$

respectively, where $b(\eta)$ is the function $b(\cdot; \eta)$. The full-data estimators of b^*_τ and p^*_τ are defined as

$$
\begin{aligned}
\hat{b}^{\text{full}}_{\tau}(x) &= \begin{cases} \exp\left(\hat{\eta}^{\text{full}}(\tau)^T w(x)\right) & \textit{in the loglinear case,} \\ \hat{\eta}^{\text{full}}(\tau)^T w(x) & \textit{in the linear case;} \end{cases} \text{ and} \\
\hat{p}^{\text{full}}_{\tau}(x) &= \hat{\alpha}^{\text{full}}(\tau)^T w(x).
\end{aligned}
$$

Then the first order estimators $\hat{\tau}^{\text{full}}_{1,\text{eff,linear}}$ and $\hat{\tau}^{\text{full}}_{1,\text{eff,loglinear}}$ are the solutions to

$$
\widehat{\mathbb{IF}}^{\text{full}}_{1,\text{eff}}(\tau) \overset{def}{=} \mathbb{IF}_{1,\text{eff}}(\tau; \hat{b}^{\text{full}}_{\tau}, \hat{p}^{\text{full}}_{\tau}) = 0
$$

(for the linear and loglinear cases), where

$$
\mathbb{IF}^{\text{full}}_{1,\text{eff}}(\tau; b, p) = \frac{1}{N} \sum_{i=1}^{N} \text{IF}_{1,\text{eff},i}(\tau; b, p).
$$

(The following lemma is proved in Appendix E.)

Lemma 9. *The estimators $\hat{\tau}^{\text{full}}_{1,\text{eff}}$ are the same as the usual linear or Poisson regression estimators of the coefficient of A when regressing the outcome Y on A and W.*

Half-sample estimators are obtained by splitting the sample into disjoint "testing" and "training" parts of equal size, estimating the nuisance functions b_τ^* and p_τ^* on one half (the training half), and constructing the estimating equation for τ^* from the other half (the testing or estimation half). The half-sample estimators are not efficient in finite samples and are considered mainly because they help analyze the asymptotic properties of the split-sample estimators. In the split-sample estimators, the two half-sample estimating equations, with the roles of the training and testing halves reversed, are averaged to obtain the final estimating equation. The reason for splitting the sample in the first place is to ease analysis and is required because the functions b_τ^* and p_τ^* may not be smooth enough for Donsker conditions to hold (van der Vaart and Wellner, 1996). We use the superscripts "split" and "full" to denote the corresponding estimators. The half-sample estimators are denoted by omission of a superscript. For example, $\hat{\tau}_{1,\text{eff}}, \hat{\tau}_{1,\text{eff}}^{\text{split}}$, and $\hat{\tau}_{1,\text{eff}}^{\text{full}}$ refer to the half-sample, split-sample, and full-sample first-order efficient influence function–based estimators, respectively.

The split estimators are defined as follows.

Definition 10. *Let*

$$n = \frac{N}{2}.$$

Let $\{\text{split}(0), \text{split}(1)\}$ *be a random partition of* $\{1, \ldots, N\}$ *into two sets of equal cardinality. Define* $\hat{\eta}^{(\ell)}(\tau)$ *and* $\hat{\alpha}^{(\ell)}(\tau)$ *as the solutions to*

$$n^{-1} \sum_{i \in \text{split}(\ell)} U_{i,\text{nuis}}\left[\tau, b(\eta)\right] = 0, \text{ and}$$
$$n^{-1} \sum_{i \in \text{split}(\ell)} S_i(\tau, \alpha) = 0,$$

respectively, where $b(\eta)$ *is the function* $b(\cdot;\eta)$*. The split-sample estimators of* b_τ^* *and* p_τ^* *are defined as*

$$\hat{b}_\tau^{(\ell)}(x) = \begin{cases} \exp\left(\hat{\eta}^{(\ell)}(\tau)^T w(x)\right) & \textit{in the loglinear case,} \\ \hat{\eta}^{(\ell)}(\tau)^T w(x) & \textit{in the linear case; and} \end{cases}$$
$$\hat{p}_\tau^{(\ell)}(x) = \hat{\alpha}^{(\ell)}(\tau)^T w(x).$$

Then the first-order estimators $\hat{\tau}_{1,\text{eff,linear}}^{\text{split}}$ *and* $\hat{\tau}_{1,\text{eff,loglinear}}^{\text{split}}$ *are the solutions to*

$$\widehat{\mathbb{IF}}_{1,\text{eff}}^{\text{split}}(\tau) = \frac{1}{n} \sum_{i \in \text{split}(0)} \text{IF}_{1,\text{eff},i}(\tau; \hat{b}_\tau^{(1)}, \hat{p}_\tau^{(1)}) + \frac{1}{n} \sum_{i \in \text{split}(1)} \text{IF}_{1,\text{eff},i}(\tau; \hat{b}_\tau^{(0)}, \hat{p}_\tau^{(0)})$$

(for the linear and loglinear cases).

We next give formulae for the robust variance estimates for the full and split estimators. These are standard robust variance estimates for M-estimators.

Definition 11. *The robust variance estimates for* $\hat{\tau}_{1,\text{eff}}^{\text{full}}$ *and* $\hat{\tau}_{1,\text{eff}}^{\text{split}}$ *are*

$$\widehat{\mathbb{V}}_{1,\text{eff}}^{\text{full}} = \frac{\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}_{1,\text{eff}}^{\text{full}}(\hat{\tau}_{1,\text{eff}}^{\text{full}})\right)}{\left\{\widehat{\mathbb{DER}}\left(\widehat{\mathbb{IF}}_{1,\text{eff}}^{\text{full}}(\hat{\tau}_{1,\text{eff}}^{\text{full}})\right)\right\}^2}, \text{ and}$$

$$\widehat{\mathbb{V}}_{1,\text{eff}}^{\text{split}} = \frac{\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}_{1,\text{eff}}^{\text{split}}(\hat{\tau}_{1,\text{eff}}^{\text{split}})\right)}{\left\{\widehat{\mathbb{DER}}\left(\widehat{\mathbb{IF}}_{1,\text{eff}}^{\text{split}}(\hat{\tau}_{1,\text{eff}}^{\text{split}})\right)\right\}^2},$$

where

$$\text{DER}_{1,\text{eff},i}(\tau,b,p) = \begin{cases} \{\Delta_i(p)\}^2 & \textit{in the linear case, and} \\ \{\Delta_i(p)\}^2 e^{\tau A_i} b(X_i) & \textit{in the loglinear case} \end{cases}$$

are the derivatives of $\text{IF}_{1,\text{eff},i}(\tau;b,p)$*. Therefore,*

$$\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}_{1,\text{eff}}^{\text{full}}(\tau)\right) = N^{-2}\sum_{i=1}^{N}\left\{\text{IF}_{1,\text{eff},i}(\tau,\hat{b}_\tau^{\text{full}},\hat{p}_\tau^{\text{full}})\right\}^2,$$

$$\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}_{1,\text{eff}}^{\text{split}}(\tau)\right) = \frac{1}{4n^2}\sum_{i\in\text{split}(1)}\left\{\text{IF}_{1,\text{eff},i}(\tau,\hat{b}_\tau^{(0)},\hat{p}_\tau^{(0)})\right\}^2 + \sum_{i\in\text{split}(0)}\left\{\text{IF}_{1,\text{eff},i}(\tau,\hat{b}_\tau^{(1)},\hat{p}_\tau^{(1)})\right\}^2,$$

$$\widehat{\mathbb{DER}}\left(\widehat{\mathbb{IF}}_{1,\text{eff}}^{\text{full}}(\tau)\right) = N^{-1}\sum_{i=1}^{N}\left\{\text{DER}_{1,\text{eff},i}(\tau,\hat{b}_\tau^{\text{full}},\hat{p}_\tau^{\text{full}})\right\},\ \textit{and}$$

$$\widehat{\mathbb{DER}}\left(\widehat{\mathbb{IF}}_{1,\text{eff}}^{\text{split}}(\tau)\right) = \frac{1}{2}\left\{n^{-1}\sum_{i\in\text{split}(1)}\text{DER}_{1,\text{eff},i}(\tau,\hat{b}_\tau^{(0)},\hat{p}_\tau^{(0)}) + n^{-1}\sum_{i\in\text{split}(0)}\text{DER}_{1,\text{eff},i}(\tau,\hat{b}_\tau^{(1)},\hat{p}_\tau^{(1)})\right\}.$$

Note that $\widehat{\mathbb{DER}}\left(\widehat{\mathbb{IF}}_{1,\text{eff}}^{\text{split}}(\tau)\right)$ converges in probability to the expected derivative of $\mathbb{IF}_{1,\text{eff}}^{\text{split}}(\tau)$, and similarly for the full-sample estimator.

3.2.2 Conditional Bias and Variance of the First-Order Efficient Estimators

We now investigate the bias of the first-order half-sample estimators conditional on the training half-sample. To do so, we note that, since the estimator $\hat{\tau}_{1,\text{eff}}$ is the solution to the influence-function equation, its bias is asymptotically equivalent (up to a constant) to the bias of $\widehat{\mathbb{IF}}_{1,\text{eff}}(\tau)$ as an estimator of $E[\mathbb{IF}_{1,\text{eff}}(\tau;b_\tau^*,p_\tau^*)]$. Hence we focus on this bias in what follows.

Definition 12. *Define*

$$\mathbb{IF}_{1,\text{eff}}(\tau;b,p) = n^{-1}\sum_{i\in\text{split}(0)}\text{IF}_{1,\text{eff}}(\tau;b,p).$$

The half-sample influence function is defined as

$$\widehat{\mathbb{IF}}_{1,\text{eff}}(\tau) \stackrel{def}{=} \mathbb{IF}_{1,\text{eff}}(\tau;\hat{b}_\tau,\hat{p}_\tau) = n^{-1}\sum_{i\in\text{split}(0)}\text{IF}_{1,\text{eff}}(\tau;\hat{b}_\tau,\hat{p}_\tau),$$

where $\hat{b}_\tau$ *and* $\hat{p}_\tau$ *are estimators of* b_τ^* *and* p_τ^* *based on the other half-sample,* split(1)*. Define the conditional bias*

$$\text{Bias}_{1,\text{eff}}(\tau;b,p) = E[\mathbb{IF}_{1,\text{eff}}(\tau;b,p)] - E[\mathbb{IF}_{1,\text{eff}}(\tau;b_\tau^*,p_\tau^*)],$$

where the expectations are conditional on the observations indexed by split(1)*.*

We are interested in calculating the conditional bias conditional on observations indexed by split(1); hence $\hat{b}_\tau$ and $\hat{p}_\tau$ are deterministic functions for the purposes of this calculation. We start by giving the bias of $\mathbb{IF}_{1,\text{eff}}(\tau;b,p)$ for its expectation.

Lemma 13. *The bias of* $\mathbb{IF}_{1,\text{eff}}(\tau;b,p)$ *as an estimator of* $E[\mathbb{IF}_{1,\text{eff}}(\tau;b^*_\tau,p^*_\tau)]$ *is given by*

$$\text{Bias}_{1,\text{eff}}(\tau;b,p) = E[q^*_\tau(X)\{b^*_\tau(X)-b(X)\}\{p^*_\tau(X)-p(X)\}].$$

The rate at which this bias goes to zero can be expressed in terms of measures of the smoothness of b^*_τ and p^*_τ, specifically their Hölder exponents as defined next.

Definition 14. *We say that a function* $h: R^d \to \mathbb{R}$ *belongs to a Hölder class* $H(\beta_h, C_h)$ *with Hölder exponent* β_h *and radius* C_h *if* $h(\cdot)$ *is continuously differentiable up to order* $\lfloor\beta_h\rfloor$ *and all partial derivatives* ∂h *of order* $\lfloor\beta_h\rfloor$ *satisfy the Lipschitz condition (van der Vaart, 1998) of order* $(\beta_h - \lfloor\beta_h\rfloor)$ *with constant* C_h*:*

$$|\partial h(x) - \partial h(x')| \le C_h \|x - x'\|^{\beta_h - \lfloor\beta_h\rfloor}.$$

Suppose $b^*_\tau \in H(C_b,\beta_b)$, and $p^*_\tau \in H(C_p,\beta_p)$. We note the following.

- If $\beta_b + \beta_p \le d/2$, where d is the dimensionality of X, then under both the nonparametric and semiparametric models:
  - No estimator for τ^* exists such that the bias and standard deviation converge at the rate $n^{-1/2}$.
  - The conditional half-sample influence function $\mathbb{IF}_{1,\text{eff}}(\tau;\hat{b}_\tau,\hat{p}_\tau)$, with $\hat{b}_\tau$ and $\hat{p}_\tau$ based on the conditioning half-sample, can be analyzed.
    The best convergence rate achievable by $\hat{b}_\tau$ and $\hat{p}_\tau$ is

$$E\left[\left\{b^*_\tau(X)-\hat{b}_\tau(X)\right\}^2\right]^{1/2} \asymp n^{-\frac{\beta_b}{2\beta_b+d}}, \text{ and}$$

$$E\left[\left\{p^*_\tau(X)-\hat{p}_\tau(X)\right\}^2\right]^{1/2} \asymp n^{-\frac{\beta_p}{2\beta_p+d}},$$

    where X is in one half-sample, $\hat{b}_\tau$ and $\hat{p}_\tau$ are based on the other half-sample, and the expectation is unconditional or conditional. The notation $a_n \asymp b_n$ indicates that there are constants c_1 and c_2 such that $c_1 a_n \le b_n \le c_2 a_n$ for all n.
  - Hence the best conditional half-sample bias is

$$\text{Bias}_{1,\text{eff}}(\tau;\hat{b}_\tau,\hat{p}_\tau) \asymp n^{-\left\{\frac{\beta_b}{2\beta_b+d}+\frac{\beta_p}{2\beta_p+d}\right\}}.$$

- The asymptotic variance of $\mathbb{IF}_{1,\text{eff}}(\tau;b,p)$ is $O(1/n)$ for any nonstochastic b and p, and hence the variance, conditional on the training half-sample, is $\text{Var}\left[(\mathbb{IF}_{1,\text{eff}}(\tau;\hat{b}_\tau,\hat{p}_\tau)\right]$, is $O_P(1/n)$.
- If $\beta_b + \beta_p \le d/2$, an MSE of $O(1/n)$ is not achievable and the squared bias decreases more slowly than the variance. This implies that $\sqrt{n}(\hat{\tau}_{1,\text{eff}} - \tau^*)$ is not asymptotically unbiased, and CIs based on $\hat{\tau}_{1,\text{eff}}$ and its estimated standard error will have an asymptotic coverage probability of 0. Recalling from Lemma 9 that $\hat{\tau}_{1,\text{eff}}$ is a conditional version of the Poisson or linear regression estimator, we see that these estimators suffer from these drawbacks when $\beta_b + \beta_p < d/2$.

3.2.3 Density-Weighted First-Order Estimators

As shown above, the bias of the optimal first-order estimators is large when b^*_τ and p^*_τ are not smooth enough. In Section 3.2.4 we will describe in detail the second-order approach that subtracts an estimator of the bias of the first-order estimator in order to reduce the bias, possibly at

the cost of increasing the variance. We considered three approaches to this second-order estimator; the two approaches that use the first-order estimator $\hat{\tau}^{\text{split}}_{1,\text{eff}}$ (introduced above) are problematic to implement. The third approach, which eliminates these issues, involves a modification to the first-order estimator.

We now describe the modified first-order estimators $\hat{\tau}^{\text{full}}_{1,\text{new}}, \hat{\tau}^{\text{split}}_{1,\text{new}}$, and $\hat{\tau}_{1,\text{new}}$ required for the third approach.

Definition 15. *Define the density-weighted quantities*

$$\begin{aligned}
\text{IF}_{1,\text{new},i}(\tau; b, p, f) &= f(X_i)\varepsilon_i(\tau, b)\left[A_i - p(X_i)\right], \\
\mathbb{IF}^{\text{full}}_{1,\text{new}}(\tau; b, p, f) &= \frac{1}{N}\sum_{i=1}^{N} \text{IF}_{1,\text{new},i}(\tau; b, p, f), \\
\mathbb{IF}_{1,\text{new}}(\tau; b, p, f) &= n^{-1} \sum_{i \in \text{split}(0)} \text{IF}_{1,\text{new},i}(\tau; b, p, f), \\
\widehat{\mathbb{IF}}^{\text{full}}_{1,\text{new}}(\tau) &= \mathbb{IF}_{1,\text{new}}(\tau; \hat{b}^{\text{full}}_{\tau}, \hat{p}^{\text{full}}_{\tau}, \hat{f}), \text{ and} \\
\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}(\tau) &= \frac{1}{2n}\left\{ \sum_{i \in \text{split}(0)} \text{IF}_{1,\text{new},i}(\tau; \hat{b}^{(1)}_{\tau}, \hat{p}^{(1)}_{\tau}, \hat{f}) + \sum_{i \in \text{split}(1)} \text{IF}_{1,\text{new},i}(\tau; \hat{b}^{(0)}_{\tau}, \hat{p}^{(0)}_{\tau}, \hat{f}) \right\},
\end{aligned}$$

where $\varepsilon_i(\tau, b)$, $\hat{b}^{\text{full}}_{\tau}$, $\hat{p}^{\text{full}}_{\tau}$, $\hat{b}^{(\ell)}_{\tau}$, and $\hat{p}^{(\ell)}_{\tau}$ are as before, and $\hat{f}$ is a nonparametric or semiparametric estimate of the density f^* of X based on the same half-sample as $\hat{b}^{(\ell)}_{\tau}$. We suppress the ℓ superscript on $\hat{f}$ because we will consider estimating $\hat{f}$ from the full-sample later (Section 3.3.3). Then the first-order estimators $\hat{\tau}^{\text{full}}_{1,\text{new,linear}}$ and $\hat{\tau}^{\text{full}}_{1,\text{new,loglinear}}$ are the solutions to

$$\widehat{\mathbb{IF}}^{\text{full}}_{1,\text{new}}(\tau) = 0;$$

$\hat{\tau}^{\text{split}}_{1,\text{new,linear}}$ *and* $\hat{\tau}^{\text{split}}_{1,\text{new,loglinear}}$ *are the solutions to*

$$\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}(\tau) = 0;$$

and $\hat{\tau}_{1,\text{new,linear}}$ *and* $\hat{\tau}_{1,\text{new,loglinear}}$ *are the solutions to*

$$\widehat{\mathbb{IF}}_{1,\text{new}}(\tau) = 0.$$

The corresponding estimators of the variance follow.

Definition 16. *The robust variance estimates for $\hat{\tau}^{\text{full}}_{1,\text{new}}$ and $\hat{\tau}^{\text{split}}_{1,\text{new}}$ are*

$$\begin{aligned}
\widehat{\mathbb{V}}^{\text{full}}_{1,\text{new}} &= \frac{\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}^{\text{full}}_{1,\text{new}}(\hat{\tau}^{\text{full}}_{1,\text{new}})\right)}{\left\{\widehat{\mathbb{DER}}\left(\widehat{\mathbb{IF}}^{\text{full}}_{1,\text{new}}(\hat{\tau}^{\text{full}}_{1,\text{new}})\right)\right\}^2}, \text{ and} \\
\widehat{\mathbb{V}}^{\text{split}}_{1,\text{new}} &= \frac{\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}(\hat{\tau}^{\text{split}}_{1,\text{new}})\right)}{\left\{\widehat{\mathbb{DER}}\left(\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}(\hat{\tau}^{\text{split}}_{1,\text{new}})\right)\right\}^2},
\end{aligned}$$

where

$$\text{DER}_{1,\text{new},i}(\tau, b, p, f) = \begin{cases} f(X_i)\left\{\Delta_i(p)\right\}^2 & \text{in the linear case,} \\ f(X_i)\left\{\Delta_i(p)\right\}^2 e^{\tau A_i} b(X_i) & \text{in the loglinear case,} \end{cases}$$

$$\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}^{\text{full}}_{1,\text{new}}(\tau)\right) = N^{-2}\sum_{i=1}^{N}\left\{\text{IF}_{1,\text{new},i}(\tau,\hat{b}^{\text{full}}_{\tau},\hat{p}^{\text{full}}_{\tau},\hat{f})\right\}^2,$$

$$\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}(\tau)\right) = \frac{1}{4n^2}\left\{\sum_{i\epsilon\text{split}(1)}\left\{\text{IF}_{1,\text{new},i}(\tau,\hat{b}^{(0)}_{\tau},\hat{p}^{(0)}_{\tau},\hat{f})\right\}^2 + \sum_{i\epsilon\text{split}(0)}\left\{\text{IF}_{1,\text{new},i}(\tau,\hat{b}^{(1)}_{\tau},\hat{p}^{(1)}_{\tau},\hat{f})\right\}^2\right\},$$

$$\widehat{\mathbb{DER}}\left(\widehat{\mathbb{IF}}^{\text{full}}_{1,\text{new}}(\tau)\right) = N^{-1}\sum_{i=1}^{N}\text{DER}_{1,\text{new},i}(\tau,\hat{b}^{\text{full}}_{\tau},\hat{p}^{\text{full}}_{\tau},\hat{f}),\ \textit{and}$$

$$\widehat{\mathbb{DER}}\left(\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}(\tau)\right) = \frac{1}{2n}\left\{\sum_{i\epsilon\text{split}(1)}\text{DER}_{1,\text{new},i}(\tau,\hat{b}^{(0)}_{\tau},\hat{p}^{(0)}_{\tau},\hat{f}) + \sum_{i\epsilon\text{split}(0)}\text{DER}_{1,\text{new},i}(\tau,\hat{b}^{(1)}_{\tau},\hat{p}^{(1)}_{\tau},\hat{f})\right\}.$$

As with the estimators based on the efficient influence function, we study the bias of these estimators by deriving the bias of the corresponding half-sample influence function conditional on the other half. The conditional bias is defined in the following.

Definition 17. For any *b, p,* or *f,* the bias of $\mathbb{IF}_{1,\text{new}}(\tau;b,p,f)$ is defined by

$$\text{Bias}_{1,\text{new}}(\tau,b,p,f) = E[\mathbb{IF}_{1,\text{new}}(\tau;b,p,f)] - E[\mathbb{IF}_{1,\text{new}}(\tau;b^*_{\tau},p^*_{\tau},f)].$$

We note that $E[\mathbb{IF}_{1,\text{new}}(\tau;b^*_{\tau},p^*_{\tau},f)]$ is in terms of f, not f^*. In contrast to the bias for the unweighted first-order estimators, this bias consists of a second-order and a third-order term, as seen in the following lemma.

Lemma 18.

$$\begin{aligned}\text{Bias}_{1,\text{new}}(\tau;b,p,f) = {} & E[q^*_{\tau}(X)f^*(X)\{b^*_{\tau}(X)-b(X)\}\{p^*_{\tau}(X)-p(X)\}] \\ & + E[q^*_{\tau}(X)\{f(X)-f^*(X)\}\{b^*_{\tau}(X)-b(X)\}\{p^*_{\tau}(X)-p(X)\}].\end{aligned}$$

The first term in this expression for the bias is second order. If one uses optimal estimators $\hat{b}_{\tau},\hat{p}_{\tau}$, and $\hat{f}$ for b^*_{τ},p^*_{τ}, and f^*, then $\text{Bias}_{1,\text{new}}(\tau;\hat{b}_{\tau},\hat{p}_{\tau},\hat{f})$ (the rate for this quantity) is $O_P\left(n^{-\left\{\frac{\beta_b}{2\beta_b+d}+\frac{\beta_p}{2\beta_p+d}\right\}}\right)$. The second term is a third-order term whose bias (by the results in Section 3.2.2 on approximation using the other half-sample) is of the order

$$O_P\left(n^{-\left\{\frac{\beta_b}{2\beta_b+d}+\frac{\beta_p}{2\beta_p+d}+\frac{\beta_f}{2\beta_f+d}\right\}}\right)$$

when $\hat{f}$ is the density estimator used. As with the unweighted estimators, we can then subtract an estimate of the second-order bias to reduce overall bias, at the cost of increasing the variance, and still obtain an optimal estimator. We describe the second-order estimators for the unweighted and density-weighted estimators next.

3.2.4 Second-Order Influence-Function Estimators and Their Properties

The bias of the first-order estimators can be reduced by subtracting an estimator of the bias (or, in the case of the density-weighted estimators, the second-order component of bias) derived

above:

$$\text{Bias}_{1,\text{eff}}(\tau;b,p) = E[q_\tau^*(X)\{b_\tau^*(X) - b(X)\}\{p_\tau^*(X) - p(X)\}], \text{ and}$$
$$\text{Bias}_{1,\text{new}}(\tau;b,p,f) = E[q_\tau^*(X)f^*(X)\{b_\tau^*(X) - b(X)\}\{p_\tau^*(X) - p(X)\}].$$

In this section we define the second-order estimators and derive their bias and variance formulae. Since many of the properties of the second-order estimators in the i.i.d. case are shared by those in the time-series case, we defer a detailed discussion of the rates of convergence of the bias and variance of these estimators, and the consequences of these rates for optimal choices of k, to Section 3.3.3, which discusses rates both for the time-series and i.i.d. situations.

In order to define such an estimator, we consider an orthonormal basis $\{\varphi_l, l = 1,\ldots,\}$ of the space $\mathscr{L}_2(\lambda)$, λ being Lebesgue measure on $\mathbb{R}^d$, assuming X is absolutely continuous. Let $\bar{\varphi}_k(x) = [\varphi_1(x),\ldots,\varphi_k(x)]^T$. The three methods considered correspond to three projection kernels defined in the following.

Definition 19. *For any density f on the support of X, define $K_{f,k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ by*

$$K_{f,k}(x,y) = \sum_{l=1}^{k} \frac{\varphi_l(x)\varphi_l(y)}{f^{1/2}(x)f^{1/2}(y)}$$

and $K_{f,q,k,\text{alt}} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ by

$$K_{f,q,k,\text{alt}}(x_1,x_2) = q(x_1)^{1/2}\bar{\varphi}_k(x_1)^T E_f[q(X)\bar{\varphi}_k(X)\bar{\varphi}_k(X)^T]^{-1}\ \bar{\varphi}_k(x_2)q(x_2)^{1/2}.$$

Define $K_k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ by

$$K_k(x,y) = \sum_{l=1}^{k} \varphi_l(x)\varphi_l(y).$$

For any function $g(x) \in \mathscr{L}_2(\mu)$ expressed as $\sum_{l=1}^{\infty}\gamma_l\varphi_l(x)$, define the orthogonal projection operators with respect to Lebesgue measure

$$\prod[g(x)|\bar{\varphi}_k(x)] = \sum_{l=1}^{k}\gamma_l\varphi_l(x),\ \prod[g(x)|\bar{\varphi}_k^{\perp}(x)] = \sum_{l=k+1}^{\infty}\gamma_l\varphi_l(x).$$

Also define

$$\prod_f[g(x)|\bar{\varphi}_k(x)], \text{ and } \prod_f[g(x)|\bar{\varphi}_k^{\perp}(x)]$$

as the projections in $\mathscr{L}_2(f)$ of $g(x)$ onto $\bar{\varphi}_k(x)$ and $\bar{\varphi}_k^{\perp}(x)$, respectively; i.e., $\prod_f[g(x)|\bar{\varphi}_k(x)]$ minimizes $E_f\left[\{g(X)-h(X)\}^2\right]$ over h in the linear span of $\bar{\varphi}_k(x)$, and $\prod_f[g(x)|\bar{\varphi}_k^{\perp}(x)]$ minimizes $E_f\left[\{g(X)-h(X)\}^2\right]$ over h in the orthogonal complement of the linear span of $\bar{\varphi}_k(x)$.

Next we define second-order influence functions based on these kernels.

Definition 20. *Recall the definition of $\varepsilon_i(\tau,b)$ and $\Delta_j(\tau,p,q)$ from Definitions 5 and 6. Define $\hat{\omega}^{(\ell)}(\tau)$ as the solution to*

$$\frac{1}{n}\sum_{i\in\text{split}(\ell)}\left\{e^{\tau A_i} - e^{\omega(\tau)W_i}\right\}W_i = 0,$$

and

$$\hat{q}_\tau^{(\ell)}(x) = \begin{cases} \exp\left(\hat{\omega}^{(\ell)}(\tau)^T w(x)\right) & \textit{in the loglinear case,} \\ 1 & \textit{in the linear case.} \end{cases}$$

Next define

$$\mathrm{IF}^{(k)}_{22,i,j}(\tau;b,p,q;K) = \begin{cases} -\varepsilon_i(\tau,b)K(X_i,X_j)\Delta_j(\tau,p,q) & \textit{if } K = K_{\hat{f},k} \textit{ or } K_k, \\ -q^{-1/2}(X_i)\varepsilon_i(\tau,b)K(X_i,X_j)\Delta_j(\tau,p,q)q^{1/2}(X_j) & \textit{if } K = K_{\hat{q},\hat{f},k,\mathrm{alt}}, \end{cases}$$

$$\widehat{\mathbb{IF}}^{\mathrm{split},(k)}_{22}(\tau;K) = \frac{1}{2}\Bigg\{\frac{1}{n(n-1)}\sum_{i\neq j\in \mathrm{split}(0)} \mathrm{IF}^{(k)}_{22,ij}(\tau,\hat{b}^{(1)}_\tau,\hat{p}^{(1)}_\tau,\hat{q}^{(1)}_\tau;K) + \frac{1}{n(n-1)}\sum_{i\neq j\in \mathrm{split}(1)} \mathrm{IF}^{(k)}_{22,ij}(\tau,\hat{b}^{(0)}_\tau,\hat{p}^{(0)}_\tau,\hat{q}^{(0)}_\tau;K)\Bigg\},$$

$$\widehat{\mathbb{IF}}^{(k)}_{22}(\tau;K) = \frac{1}{n(n-1)}\sum_{i\neq j\in \mathrm{split}(1)} \mathrm{IF}^{(k)}_{22,ij}(\tau,\hat{b}^{(0)}_\tau,\hat{p}^{(0)}_\tau,\hat{q}^{(0)}_\tau;K),$$

$$\widehat{\mathbb{IF}}^{(k)}_{2,\mathrm{eff}}(\tau) = \widehat{\mathbb{IF}}_{1,\mathrm{eff}}(\tau) + \widehat{\mathbb{IF}}^{(k)}_{22}(\tau;K_{\hat{f},k}),$$

$$\widehat{\mathbb{IF}}^{(k)}_{2,\mathrm{alt}}(\tau) = \widehat{\mathbb{IF}}_{1,\mathrm{eff}}(\tau) + \widehat{\mathbb{IF}}^{(k)}_{22}(\tau;K_{\hat{f},k,\mathrm{alt}}),$$

$$\widehat{\mathbb{IF}}^{(k)}_{2,\mathrm{new}}(\tau) = \widehat{\mathbb{IF}}_{1,\mathrm{new}}(\tau) + \widehat{\mathbb{IF}}^{(k)}_{22}(\tau;K_k),$$

$$\widehat{\mathbb{IF}}^{\mathrm{split},(k)}_{2,\mathrm{eff}}(\tau) = \widehat{\mathbb{IF}}^{\mathrm{split}}_{1,\mathrm{eff}}(\tau) + \widehat{\mathbb{IF}}^{\mathrm{split},(k)}_{22}(\tau;K_{\hat{f},k}),$$

$$\widehat{\mathbb{IF}}^{\mathrm{split},(k)}_{2,\mathrm{alt}}(\tau) = \widehat{\mathbb{IF}}_{1,\mathrm{eff}}(\tau) + \widehat{\mathbb{IF}}^{\mathrm{split},(k)}_{22}(\tau;K_{\hat{f},k,\mathrm{alt}}), \textit{ and}$$

$$\widehat{\mathbb{IF}}^{\mathrm{split},(k)}_{2,\mathrm{new}}(\tau) = \widehat{\mathbb{IF}}^{\mathrm{split}}_{1,\mathrm{new}}(\tau) + \widehat{\mathbb{IF}}^{\mathrm{split},(k)}_{22}(\tau;K_k).$$

The first three definitions are in terms of the generic kernel K, and specific kernels are used to define the estimated second-order influence functions that follow. Since we focus on the "new" second-order estimators in this report, we define the following shorthand:

$$\mathrm{IF}^{(k)}_{22,i,j}(\tau;b,p,q) = \mathrm{IF}^{(k)}_{22,i,j}(\tau;b,p,q;K_k),$$

$$\widehat{\mathbb{IF}}^{\mathrm{split},(k)}_{22}(\tau) = \widehat{\mathbb{IF}}^{\mathrm{split},(k)}_{22}(\tau;K_k),$$

$$\widehat{\mathbb{IF}}^{(k)}_{22}(\tau) = \widehat{\mathbb{IF}}^{(k)}_{22}(\tau;K_k),$$

$$\widehat{\mathbb{IF}}^{(k)}_{2}(\tau) = \widehat{\mathbb{IF}}^{(k)}_{2,\mathrm{new}}(\tau),$$

$$\widehat{\mathbb{IF}}^{\mathrm{split},(k)}_{2}(\tau) = \widehat{\mathbb{IF}}^{\mathrm{split},(k)}_{2,\mathrm{new}}(\tau), \textit{ and}$$

$$\widehat{\mathbb{IF}}^{\mathrm{full},(k)}_{2}(\tau) = \widehat{\mathbb{IF}}^{\mathrm{full},(k)}_{2,\mathrm{new}}(\tau).$$

Note that the "alt" kernel uses a slightly different definition than the other kernels. This puts a factor of $1/q(X)^{1/2}$ on either side of the kernel in the conditional expectation $E[\mathrm{IF}^{(k)}_{22,i,j}(\tau;b,p,q;K)|X_i,X_j]$, rather than $1/q(X)$ on one side and no factor on the other. We shall see that this leads to a better rate of convergence for the bias than the other kernels.

As before, we use $\widehat{\mathbb{IF}}^{\mathrm{split},(k)}_{2}(\tau)$ to do estimation but analyze bias and variance properties using the half-sample estimator $\widehat{\mathbb{IF}}^{(k)}_{2}(\tau)$. In what follows, we use the notation

$$\delta b_\tau(x) = b^*_\tau(x) - \hat{b}_\tau(x), \text{ and}$$

$$\delta p_\tau(x) = p^*_\tau(x) - \hat{p}_\tau(x)$$

for brevity.

Theorem 21. *Let $\widehat{\mathbb{IF}}^{(k)}_{22}(\tau;K)$ be as in Definition 20, let f be any density on the support of X, and let K_k,$K_{f,k}$, and $K_{q,f,k,\text{alt}}$ be as in Definition 19. Then the three influence functions $\mathbb{IF}^{(k)}_{22}(\tau;K)$ in Definition 20 estimate the corresponding second-order biases up to third-order and truncation terms. Specifically (EB means estimation bias),*

$$\begin{aligned}
E[\widehat{\mathbb{IF}}^{(k)}_{22}(\tau;K_{\hat{f},k})] + \text{Bias}_{1,\text{eff}}(\tau;\hat{b}_\tau,\hat{p}_\tau) &= E\left[q^*_\tau(X)\delta b_\tau(X)\delta p_\tau(X)\left\{\frac{f^*(X)}{\hat{f}(X)} - 1\right\}\right] \\
&\quad - \int \prod\left[q^*_\tau(x)\delta b_\tau(x)\frac{f^*(x)}{\hat{f}^{1/2}(x)}\middle|\bar{\varphi}^\perp_k(x)\right]\prod\left[\delta p_\tau(x)\frac{f^*(x)}{\hat{f}^{1/2}(x)}\middle|\bar{\varphi}^\perp_k(x)\right]dx \\
&\quad + E\left[q^*_\tau(X_1)\delta b_\tau(X_1)K_{f,k}(X_1,X_2)\delta p_\tau(X_2)\left\{\frac{q^*_\tau(X_2)}{\hat{q}_\tau(X_2)} - 1\right\}\right], \\
E[\widehat{\mathbb{IF}}^{(k)}_{22}(\tau;K_{\hat{q},\hat{f},k,\text{alt}})] + \text{Bias}_{1,\text{eff}}(\tau;\hat{b}_\tau,\hat{p}_\tau) &= \text{EB}^{(3)}_{q,b,p} + \text{EB}^{(3)}_{f,b,p} \\
&\quad - E_{\hat{f}}\left\{\prod_{\hat{f}}\left[\hat{q}^{1/2}_\tau(X)\delta b_\tau(X)\middle|\{\bar{\varphi}_k(X)\hat{q}^{1/2}_\tau(X)\}^\perp\right]\prod_{\hat{f}}\left[\delta p_\tau(X)\hat{q}^{1/2}_\tau(X)\middle|\{\bar{\varphi}_k(X)\hat{q}^{1/2}_\tau(X)\}^\perp\right]\right\}, \text{ and}
\end{aligned}$$

$$\begin{aligned}
E[\widehat{\mathbb{IF}}^{(k)}_{22}(\tau;K_k)] + \text{Bias}_{1,\text{new}}(\tau;\hat{b}_\tau,\hat{p}_\tau,\hat{f}) &= E[\{f^*(X) - \hat{f}(X)\}q^*_\tau(X)\delta b_\tau(X)\delta p_\tau(X)] \\
&\quad + E\left[q^*_\tau(X_1)\delta b_\tau(X_1)K_k(X_1,X_2)\delta p_\tau(X_2)\left\{\frac{q^*_\tau(X_2)}{\hat{q}_\tau(X_2)} - 1\right\}\right] \\
&\quad - \int \prod[f^*(x)q^*_\tau(x)\delta b_\tau(x)|\bar{\varphi}^\perp_k(x)]\prod[f^*(x)\delta p_\tau(x)|\bar{\varphi}^\perp_k(x)]\,dx.
\end{aligned}$$

In the above, $\text{EB}^{(3)}_{q,b,p}$ is a sum of third- or higher-order terms whose rate of convergence to 0 depends on the differences $\hat{q}_\tau(X) - q^*_\tau(X)$ or $q^*_\tau(X)/\hat{q}_\tau(X) - 1$ and δb_τ and δp_τ. $\text{EB}^{(3)}_{f,b,p}$ is a sum of third- or higher-order terms whose rate of convergence to 0 depends on the differences $\hat{f}(X)/f^*(X) - 1$ or $f^*(X)/\hat{f}(X) - 1$ and δb_τ and δp_τ.

Consequently, the estimated influence functions $\widehat{\mathbb{IF}}^{(k)}_{2,\text{eff}}(\tau)$, $\widehat{\mathbb{IF}}^{(k)}_{2,\text{alt}}(\tau)$, and $\widehat{\mathbb{IF}}^{(k)}_{2,\text{new}}(\tau)$ have bias equal to third-order terms (whose rate does not depend on k) plus a tail or truncation term (whose rate does depend on k).

We refer to the the integral terms with projection as *truncation bias* terms since they arise from the fact that we sum only k basis functions in the kernels, which has the effect of truncating the function to its k-th term in its expansion in terms of the orthonormal basis. The truncation bias can be made arbitrarily small by choosing k large enough (at the cost of increasing the variance of the estimator). The remaining terms are called *estimation bias* terms. These are the components of bias resulting from our failure to obtain perfect estimators of the quantities $b^*_{i\tau}, p^*_{i\tau}, q^*_{i\tau}$, or f^*. Were the estimators for these quantities exactly equal to the true quantities (which would be the case if we knew the true quantities), these terms would be identically equal to zero.

The above theorem justifies the use of the second-order influence functions as estimating functions. All three second-order estimators reduce the bias to the third-order estimation bias plus truncation bias terms.

However, the first influence function $\mathbb{IF}_{22}^{(k)}(\tau; K_{\hat{f},k})$ uses $\hat{f}^{1/2}$ in its denominator and hence requires that $\hat{f}$ be bounded away from zero. Even if f is bounded away from zero, a few small values of $\hat{f}$ do occur in practice. These give the corresponding observations a disproportionate amount of influence over the corresponding estimator, resulting in a significant inflation of variance in finite samples. This was found to be a significant factor during analysis of the NMMAPS data.

The second kernel, $K_{\hat{q},\hat{f},k,\text{alt}}$, is attractive because, as explained in Section 3.3.3, the truncation term in its bias is $O_P(k^{-(\beta_b+\beta_p)/d})$, which is no worse than the truncation bias $O_P(k^{-(\min\{\beta_q,\beta_f,\beta_b\}+\min\{\beta_p,\beta_f\})/d})$ of the other two estimators. However, evaluation of this influence function requires inversion of the $k \times k$ matrix $E_{\hat{f}}[\hat{q}_\tau(X)\bar{\varphi}_k(X)\bar{\varphi}_k^T(x)]$, a computational problem that we failed to solve for large values of k.

Hence, except for a discussion of the rates of convergence of these estimators in Section 3.3.3, in the rest of this report, we restrict our attention to the density-weighted influence functions $\widehat{\mathbb{IF}}_{2,\text{new}}^{(k)}(\tau), \widehat{\mathbb{IF}}_{2,\text{new}}^{\text{split},(k)}(\tau)$. The corresponding estimators of τ^* and their variance estimators are defined in the following.

Definition 22. *Define the second-order estimators of τ^* as follows. Let $\hat{\tau}_2^{(k)}$ solve*

$$\widehat{\mathbb{IF}}_2^{(k)}(\tau) = 0,$$

and let $\hat{\tau}_2^{\text{split},(k)}$ solve

$$\widehat{\mathbb{IF}}_2^{\text{split},(k)}(\tau) = 0.$$

The variance of the above estimators is estimated by

$$\widehat{\mathbb{V}}_2^{(k)} = \frac{\widehat{\mathbb{V}}\left[\widehat{\mathbb{IF}}_{1,\text{new}}\left(\hat{\tau}_2^{(k)}\right)\right] + \widehat{\mathbb{V}}\left[\widehat{\mathbb{IF}}_{22}^{(k)}\left(\hat{\tau}_2^{(k)}\right)\right]}{\left\{\widehat{\mathbb{DER}}\left[\widehat{\mathbb{IF}}_{1,\text{new}}\left(\hat{\tau}_2^{(k)}\right)\right]\right\}^2}, \text{ and}$$

$$\widehat{\mathbb{V}}_2^{\text{split},(k)} = \frac{\widehat{\mathbb{V}}\left[\widehat{\mathbb{IF}}_{1,\text{new}}^{\text{split}}\left(\hat{\tau}_2^{\text{split},(k)}\right)\right] + \widehat{\mathbb{V}}\left[\widehat{\mathbb{IF}}_{22}^{\text{split},(k)}\left(\hat{\tau}_2^{\text{split},(k)}\right)\right]}{\left\{\widehat{\mathbb{DER}}\left[\widehat{\mathbb{IF}}_{1,\text{new}}^{\text{split}}\left(\hat{\tau}_2^{\text{split},(k)}\right)\right]\right\}^2}.$$

Here we let $O_i = (Y_i, A_i, X_i)$, and we define the symmetrized functions

$$\hat{m}_\tau^{(\ell)}(O_i, O_j) = \left\{\text{IF}_{22,ij}(\tau, \hat{b}_\tau^{(\ell)}, \hat{p}_\tau^{(\ell)}, \hat{q}_\tau^{(\ell)}; K_k) + \text{IF}_{22,ji}(\tau, \hat{b}_\tau^{(\ell)}, \hat{p}_\tau^{(\ell)}, \hat{q}_\tau^{(\ell)}; K_k)\right\}/2, \text{ and}$$

$$m_\tau(O_i, O_j) = \left\{\text{IF}_{22,ij}(\tau, b_\tau^*, p_\tau^*, q_\tau^*; K_k) + \text{IF}_{22,ji}(\tau, b_\tau^*, p_\tau^*, q_\tau^*; K_k)\right\}/2;$$

and we define

$$\widehat{\mathbb{V}}\left[\widehat{\mathbb{IF}}_{22}^{\text{split},(k)}(\tau)\right] = \frac{1}{4n^2(n-1)^2}\left\{\sum_{i<j\in\text{split}(0)}\left\{\hat{m}_\tau^{(1)}(O_i, O_j)\right\}^2 + \sum_{i<j\in\text{split}(1)}\left\{\hat{m}_\tau^{(0)}(O_i, O_j)\right\}^2\right\}, \text{ and}$$

$$\widehat{\mathbb{V}}\left[\widehat{\mathbb{IF}}_{22}^{(k)}(\tau)\right] = \frac{1}{n^2(n-1)^2}\sum_{i<j\in\text{split}(1)}\left\{\hat{m}_\tau^{(0)}(O_i, O_j)\right\}^2,$$

where the other quantities are as in Definition 16.

The expression for the variance of the influence function $\widehat{\mathbb{IF}}_2^{(k)}(\tau)$ is similar to the corresponding expression in the time-series case. The validity of this expression and the consistency of its

estimator are much more easily proved than in the time-series case. Hence we give the proof only for the time-series case and the reader is referred to Section 3.3.2.

As before, we have derived the results in this section using the half-sample estimators. That the full-sample and split-sample estimators have similar properties was established through simulations as shown in Appendix C.

The variance of $\mathbb{IF}_{22}$ is $O(k/n^2)$. The proof of this statement is similar to and simpler than the arguments in the proof for the corresponding time-series estimator in Lemma 41 of Section 3.3.2. Hence we do not prove it here and refer the reader to Lemma 41 and the comments immediately following it for an explanation.

3.3 THE TIME-SERIES ESTIMATORS

Thus far we have focused on estimation without the complication of time. As mentioned in the introduction, time introduces dependencies between observations so we are no longer in the i.i.d. case. Here we describe modifications to the influence functions that allow for time dependence. Due to the disadvantages of the other methods described before, we touch upon the other second-order estimators only briefly and focus on the density-weighted estimators in this section.

For each city in the NMMAPS data, each observation corresponds to a particular day on which measurements of deaths Y, PM_{10} in the air A, and weather variables X were recorded. Observations are equally spaced in time. We use t_i to denote the time corresponding to the i-th observed time point.

In previous analyses, time is incorporated into models such as the linear and loglinear semiparametric models by inclusion in the covariates X (Dominici et al., 2004). However, we choose to keep time distinct from the covariates X and from subscript functions depending on X by using i to indicate their dependence on time t_i. Individual observations no longer have the same distribution (since the distribution can vary with time), and are no longer independent (since observations closer in time or corresponding to the same season may be correlated). Recall that in the time-series case, the linear and loglinear semiparametric models are now written

$$\Phi(E_i[Y_i|A_i, X_i]) = \tau^* A_i + \zeta_i^*(X_i),$$

where the subscript i for the expectation specifies that the expectation is with respect to the distribution of the i-th observation, and the subscript i for the unknown function ζ_i is now necessary because it may vary from observation to observation.

The definition of τ needs to be updated for the time-series case, since the i.i.d. definition does not involve distributions that change from observation to observation.

Definition 23. *We define the parameter τ in two ways. The first is as the solution to the equation*

$$E_i\left[(Y_i - \tau A_i - E[Y_i - \tau A_i|X_i])\,(A_i - E_i[A_i|X_i])\right] = 0 \quad \textit{in the linear case, and}$$
$$E_i\left[\left(Y_i - \frac{E_i[Y_i|X_i]}{E_i[e^{\tau A_i}|X_i]}\right)\left(A_i - \frac{E_i[A_i e^{\tau A_i}|X_i]}{E_i[e^{\tau A_i}|X_i]}\right)\right] = 0 \quad \textit{in the loglinear case.}$$

The second definition of τ^* is as the solution to the equation

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|s(i)|}\sum_{j\in s(i)} E_i\left[f_j(X_i)\left(Y_i - \tau A_i - E_i[Y_i - \tau A_i|X_i]\right)\left(A_i - E[A_i|X_i]\right)\right] = 0 \quad \textit{in the linear case, and}$$

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|s(i)|}\sum_{j\in s(i)}E\left[f_j(X_i)\left(Y_i-\frac{E_i[Y_i|X_i]}{E_i[e^{\tau A_i}|X_i]}\right)\left(A_i-\frac{E[A_ie^{\tau A_i}|X_i]}{E_i[e^{\tau A_i}|X_i]}\right)\right]=0 \quad \textit{in the loglinear case,}$$

where $f=\{f_i, i=1,\ldots,N\}$, and each f_i is a fixed, known density defined on the support of X_i (not the true density f_i^*).

3.3.1 First- and Second-Order Estimators in the Time-Series Case

We modify the definitions of Section 3.2 for the time-series case in the following series of definitions. The unknown functions $b^*_{i\tau}, p^*_{i\tau}$, and $q^*_{i\tau}$ now depend on time and gain a subscript i. The transform $w_i(x)$ used to define parametric models for q_i and p_i is also time-dependent.

Definition 24. *In the loglinear case define*

$$\begin{aligned}
b^*_{i\tau}(X_i) &= \frac{E_i[Y_i|X_i]}{E_i\left[e^{\tau A_i}|X_i\right]},\\
p^*_{i\tau}(X_i) &= \frac{E_i[Ae^{\tau A_i}|X_i]}{E[e^{\tau A_i}|X_i]},\\
q^*_{i\tau}(X_i) &= E_i[e^{\tau A_i}|X_i],\\
\varepsilon_i(\tau,b) &= Y_i - e^{\tau A_i}b(X_i),\\
b_i(x;\eta) &= \exp\left[\eta^T w_i(x)\right],\\
q_i(x;\omega) &= \exp\left[\omega^T w_i(x)\right],\\
\Delta_i(\tau,p,q) &= \left[A_i - p(X_i)\right]\frac{e^{\tau A_i}}{q(X_i)}, \textit{ and}\\
S_i(\tau,\alpha) &= (A_i-\alpha^T W_i)W_i e^{\tau A_i}.
\end{aligned}$$

Definition 25. *In the linear case define*

$$\begin{aligned}
b^*_{i\tau}(X_i) &= E_i\left[Y_i-\tau A_i|X_i\right],\\
p^*_{i\tau}(X_i) &= E_i[A_i|X_i],\\
q^*_{i\tau}(X_i) &= 1,\\
\varepsilon_i(\tau,b) &= Y_i-\tau A_i - b(X_i),\\
b_i(x;\eta) &= \eta^T w_i(x),\\
q_i(x;\omega) &= 1,\\
S_i(\tau,\alpha) &= (A_i-\alpha^T W_i)W_i, \textit{ and}\\
\Delta_i(\tau,p,q) &= A_i - p(X_i).
\end{aligned}$$

Common to both, define

Definition 26.

$$\begin{aligned}
\Delta_i(p) &= A_i - p(X_i),\\
U_{i,\text{profile}}(\tau,b) &= \varepsilon_i(\tau,b)A_i,\\
U_{i,\text{nuis}}(\tau,b) &= \varepsilon_i(\tau,b)W_i,\\
U_i(\tau,b) &= (A_i,W_i)^T\varepsilon_i(\tau,b), \textit{ and}\\
\text{IF}_{1,\text{eff},i}(\tau;b,p) &= \varepsilon_i(\tau,b)\left[A_i-p(X_i)\right].
\end{aligned}$$

Next we define the time-series full-data and split-sample first-order estimators.

Definition 27. *Let* $\hat{\eta}^{\text{full}}(\tau)$ *and* $\hat{\alpha}^{\text{full}}(\tau)$ *solve*

$$N^{-1}\sum_{i=1}^{N} U_{i,\text{nuis}}\left[\tau, b_i(\eta)\right] = 0, \text{ and}$$
$$N^{-1}\sum_{i=1}^{N} S_i(\tau,\alpha) = 0,$$

respectively, where $b_i(\eta)$ *is the function* $b_i(\cdot;\eta)$*. The time-dependent full-sample estimators of* $b^*_{i\tau}$ *and* $p^*_{i\tau}$ *are defined as*

$$\hat{b}^{\text{full}}_{i\tau}(x) = \begin{cases} \exp\left[\hat{\eta}^{\text{full}}(\tau)^T w_i(x)\right] & \text{in the loglinear case,} \\ \hat{\eta}^{\text{full}}(\tau)^T w_i(x) & \text{in the linear case; and} \end{cases}$$
$$\hat{p}^{\text{full}}_{i\tau}(x) = \hat{\alpha}^{\text{full}}(\tau)^T w_i(x).$$

Then the first-order estimators $\hat{\tau}^{\text{full}}_{1,\text{eff,linear}}$ *and* $\hat{\tau}^{\text{full}}_{1,\text{eff,loglinear}}$ *are the solutions to*

$$\widehat{\mathbb{IF}}^{\text{full}}_{1,\text{eff}}(\tau) = 0$$

(for the linear and loglinear cases), where

$$\widehat{\mathbb{IF}}^{\text{full}}_{1,\text{eff}}(\tau) = \frac{1}{N}\sum_{i=1}^{N} \text{IF}_{1,\text{eff},i}(\tau; \hat{b}^{\text{full}}_{i\tau}, \hat{p}^{\text{full}}_{i\tau}).$$

Definition 28. *As in the i.i.d. case, let* $n = \frac{N}{2}$ *and let* $\{\text{split}(0), \text{split}(1)\}$ *be a partition of* $\{1,\ldots,N\}$ *into two sets of size n. For* $\ell = 0, 1$ *define* $\hat{\eta}^{(\ell)}(\tau)$ *and* $\hat{\alpha}^{(\ell)}(\tau)$ *as the solutions to*

$$n^{-1}\sum_{i\in\text{split}(\ell)} U_{i,\text{nuis}}\left[\tau, b_i(\eta)\right] = 0, \text{ and}$$
$$n^{-1}\sum_{i\in\text{split}(\ell)} S_i(\tau,\alpha) = 0,$$

respectively, where $b_i(\eta)$ *is the function* $b_i(\cdot;\eta)$*. The time-series split-sample estimators of* $b^*_{i\tau}$ *and* $p^*_{i\tau}$ *are defined as*

$$\hat{b}^{(\ell)}_{i\tau}(x) = \begin{cases} \exp\left(\hat{\eta}^{(\ell)}(\tau)^T w_i(x)\right) & \text{in the loglinear case,} \\ \hat{\eta}^{(\ell)}(\tau)^T w_i(x) & \text{in the linear case; and} \end{cases}$$
$$\hat{p}^{(\ell)}_{i\tau}(x) = \hat{\alpha}^{(\ell)}(\tau)^T w_i(x).$$

Define

$$\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{eff}}(\tau) = \frac{1}{n}\left\{\sum_{i\in\text{split}(0)} \text{IF}_{1,\text{eff},i}(\tau; \hat{b}^{(1)}_{i\tau}, \hat{p}^{(1)}_{i\tau}) + \sum_{i\in\text{split}(1)} \text{IF}_{1,\text{eff},i}(\tau; \hat{b}^{(0)}_{i\tau}, \hat{p}^{(0)}_{i\tau})\right\}.$$

The first-order estimators $\hat{\tau}^{\text{split}}_{1,\text{eff,linear}}$ *and* $\hat{\tau}^{\text{split}}_{1,\text{eff,loglinear}}$ *are the solutions to*

$$\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{eff}}(\tau) = 0$$

for both the linear and loglinear cases.

Definition 29. *The robust variance estimates for* $\hat{\tau}^{\text{full}}_{1,\text{eff}}$ *and* $\hat{\tau}^{\text{split}}_{1,\text{eff}}$ *are*

$$\widehat{\mathbb{V}}^{\text{full}}_{1,\text{eff}} = \frac{\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}^{\text{full}}_{1,\text{eff}}(\hat{\tau}^{\text{full}}_{1,\text{eff}})\right)}{\left\{\widehat{\mathbb{DER}}\left(\widehat{\mathbb{IF}}^{\text{full}}_{1,\text{eff}}(\hat{\tau}^{\text{full}}_{1,\text{eff}})\right)\right\}^2}, \text{ and}$$

$$\widehat{\mathbb{V}}^{\text{split}}_{1,\text{eff}} = \frac{\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{eff}}(\hat{\tau}^{\text{split}}_{1,\text{eff}})\right)}{\left\{\widehat{\mathbb{DER}}\left(\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{eff}}(\hat{\tau}^{\text{split}}_{1,\text{eff}})\right)\right\}^2},$$

where

$$\text{DER}_{1,\text{eff},i}(\tau,b,p) = \begin{cases} \{\Delta_i(p)\}^2 & \text{in the linear case,} \\ \{\Delta_i(p)\}^2 e^{\tau A_i} b(X_i) & \text{in the loglinear case,} \end{cases}$$

$$\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}^{\text{full}}_{1,\text{eff}}(\tau)\right) = N^{-2}\sum_{i=1}^{N}\left\{\text{IF}_{1,\text{eff},i}(\tau,\hat{b}^{\text{full}}_{i\tau},\hat{p}^{\text{full}}_{i\tau})\right\}^2,$$

$$\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{eff}}(\tau)\right) = \frac{1}{4n^2}\left\{\sum_{i\in\text{split}(1)}\left\{\text{IF}_{1,\text{eff},i}(\tau,\hat{b}^{(0)}_{i\tau},\hat{p}^{(0)}_{i\tau})\right\}^2 + \sum_{i\in\text{split}(0)}\left\{\text{IF}_{1,\text{eff},i}(\tau,\hat{b}^{(1)}_{i\tau},\hat{p}^{(1)}_{i\tau})\right\}^2\right\},$$

$$\widehat{\mathbb{DER}}\left(\widehat{\mathbb{IF}}^{\text{full}}_{1,\text{eff}}(\tau)\right) = N^{-1}\sum_{i=1}^{N}\left\{\text{DER}_{1,\text{eff},i}(\tau,\hat{b}^{\text{full}}_{i\tau},\hat{p}^{\text{full}}_{i\tau})\right\}, \text{ and}$$

$$\widehat{\mathbb{DER}}\left(\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{eff}}(\tau)\right) = \frac{1}{2n}\left\{\sum_{i\in\text{split}(1)}\text{DER}_{1,\text{eff},i}(\tau,\hat{b}^{(0)}_{i\tau},\hat{p}^{(0)}_{i\tau}) + \sum_{i\in\text{split}(0)}\text{DER}_{1,\text{eff},i}(\tau,\hat{b}^{(1)}_{i\tau},\hat{p}^{(1)}_{i\tau})\right\}.$$

We note that the variance estimators in the above definition may not be consistent except under various assumptions, including the assumption that the semiparametric model holds. This comment applies to all the variance estimates in this section. These assumptions are discussed in greater detail in Section 3.3.2. The definitions for the density-weighted estimators need additional notation in the time-series case:

- For each time i, we denote by $s(i)$ a set of time indices (i.e., a set of observations) that depends on i, defined by the intersection of

  $$\{j : n_1 \le |j-i| \le n_2\}$$

  with another set. The precise definition is provided in Equation 3.1, Section 3.4.1. The bounds n_1 and n_2 are needed to ensure that conditions necessary for bias and variance derivations hold, as will be explained in Section 3.3.2. Let $|s(i)|$ be the cardinality of the set $s(i)$. We will assume that $s(i)$ is symmetric in the sense that $j \in s(i) \Leftrightarrow i \in s(j)$.
- Next, noting that the density of the covariates changes from time to time, we use f^*_i to denote the density of X_i.
- For notational reasons, we let f denote the *sets* of functions $\{f_i : i \in \{1,\ldots,N\}\}$, where f_i is some density on the support of the covariates at time i. Also, $f^* = \{f^*_i : i \in \{1,\ldots,N\}\}$.
- $\hat{f}_i$ is an estimate of the density f^*_i, and $\hat{f} = \left\{\hat{f}_i : i \in \{1,\ldots,N\}\right\}$.

This notation is used in corresponding density-weighted estimators and their estimated variances below.

Definition 30. *Define the density-weighted quantities*

$$\begin{aligned}
\mathrm{IF}_{1,\mathrm{new},i}(\tau;b,p,f) &= \frac{1}{|s(i)|}\sum_{j\in s(i)} f_j(X_i)\epsilon_i(\tau,b)\left[A_i - p(X_i)\right],\\
\mathbb{IF}_{1,\mathrm{new}}(\tau) &= \frac{1}{n}\sum_{i\in\mathrm{split}(0)} \mathrm{IF}_{1,\mathrm{new},i}(\tau, b^*_{i\tau}, p^*_{i\tau}, \hat{f}),\\
\widehat{\mathbb{IF}}^{\mathrm{full}}_{1,\mathrm{new}}(\tau) &= \frac{1}{N}\sum_{i=1}^{N} \mathrm{IF}_{1,\mathrm{new},i}(\tau;\hat{b}^{\mathrm{full}}_{i\tau}, \hat{p}^{\mathrm{full}}_{i\tau}, \hat{f}),\\
\widehat{\mathbb{IF}}^{\mathrm{split}}_{1,\mathrm{new}}(\tau) &= \frac{1}{2n}\left\{\sum_{i\in\mathrm{split}(0)} \mathrm{IF}_{1,\mathrm{new},i}(\tau;\hat{b}^{(1)}_{i\tau}, \hat{p}^{(1)}_{i\tau}, \hat{f}) + \sum_{i\in\mathrm{split}(1)} \mathrm{IF}_{1,\mathrm{new},i}(\tau;\hat{b}^{(0)}_{i\tau}, \hat{p}^{(0)}_{i\tau}, \hat{f})\right\}, \textit{ and}\\
\widehat{\mathbb{IF}}_{1,\mathrm{new}}(\tau) &= \frac{1}{n}\sum_{i\in\mathrm{split}(0)} \mathrm{IF}_{1,\mathrm{new},i}(\tau;\hat{b}^{(1)}_{i\tau}, \hat{p}^{(1)}_{i\tau}, \hat{f}).
\end{aligned}$$

Then the first-order estimators $\hat{\tau}^{\mathrm{full}}_{1,\mathrm{new,linear}}$ *and* $\hat{\tau}^{\mathrm{full}}_{1,\mathrm{new,loglinear}}$ *are the solutions to*

$$\widehat{\mathbb{IF}}^{\mathrm{full}}_{1,\mathrm{new}}(\tau) = 0,$$

and $\hat{\tau}^{\mathrm{split}}_{1,\mathrm{new,linear}}$ *and* $\hat{\tau}^{\mathrm{split}}_{1,\mathrm{new,loglinear}}$ *are the solutions to*

$$\widehat{\mathbb{IF}}^{\mathrm{split}}_{1,\mathrm{new}}(\tau) = 0.$$

Definition 31. *The robust variance estimates for* $\hat{\tau}^{\mathrm{full}}_{1,\mathrm{new}}$ *and* $\hat{\tau}^{\mathrm{split}}_{1,\mathrm{new}}$ *are*

$$\begin{aligned}
\widehat{\mathbb{V}}^{\mathrm{full}}_{1,\mathrm{new}} &= \frac{\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}^{\mathrm{full}}_{1,\mathrm{new}}(\hat{\tau}^{\mathrm{full}}_{1,\mathrm{new}})\right)}{\left\{\widehat{\mathbb{DER}}\left(\widehat{\mathbb{IF}}^{\mathrm{full}}_{1,\mathrm{new}}(\hat{\tau}^{\mathrm{full}}_{1,\mathrm{new}})\right)\right\}^2},\\
\widehat{\mathbb{V}}^{\mathrm{split}}_{1,\mathrm{new}} &= \frac{\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}^{\mathrm{split}}_{1,\mathrm{new}}(\hat{\tau}^{\mathrm{split}}_{1,\mathrm{new}})\right)}{\left\{\widehat{\mathbb{DER}}\left(\widehat{\mathbb{IF}}^{\mathrm{split}}_{1,\mathrm{new}}(\hat{\tau}^{\mathrm{split}}_{1,\mathrm{new}})\right)\right\}^2},
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{DER}_{1,\mathrm{new},i}\left(\tau,b,p,f\right) &= \begin{cases} f_i(X_i)\left\{\Delta_i(p)\right\}^2 & \textit{in the linear case,}\\ f_i(X_i)\left\{\Delta_i(p)\right\}^2 e^{\tau A_i} b(X_i) & \textit{in the loglinear case,}\end{cases}\\
\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}^{\mathrm{full}}_{1,\mathrm{new}}(\tau)\right) &= N^{-2}\sum_{i=1}^{N}\left\{\mathrm{IF}_{1,\mathrm{new},i}(\tau,\hat{b}^{\mathrm{full}}_{i\tau},\hat{p}^{\mathrm{full}}_{i\tau},\hat{f})\right\}^2,\\
\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}^{\mathrm{split}}_{1,\mathrm{new}}(\tau)\right) &= \frac{1}{4n^2}\left\{\sum_{i\in\mathrm{split}(1)}\left\{\mathrm{IF}_{1,\mathrm{new},i}(\tau,\hat{b}^{(0)}_{i\tau},\hat{p}^{(0)}_{i\tau},\hat{f})\right\}^2\right.\\
&\qquad\left. + \sum_{i\in\mathrm{split}(0)}\left\{\mathrm{IF}_{1,\mathrm{new},i}(\tau,\hat{b}^{(1)}_{i\tau},\hat{p}^{(1)}_{i\tau},\hat{f})\right\}^2\right\},\\
\widehat{\mathbb{DER}}\left(\widehat{\mathbb{IF}}^{\mathrm{full}}_{1,\mathrm{new}}(\tau)\right) &= N^{-1}\sum_{i=1}^{N}\left\{\mathrm{DER}_{1,\mathrm{new},i}(\tau,\hat{b}^{\mathrm{full}}_{i\tau},\hat{p}^{\mathrm{full}}_{i\tau},\hat{f})\right\}, \textit{ and}\\
\widehat{\mathbb{DER}}\left(\widehat{\mathbb{IF}}^{\mathrm{split}}_{1,\mathrm{new}}(\tau)\right) &= \frac{1}{2n}\left\{\sum_{i\in\mathrm{split}(1)}\mathrm{DER}_{1,\mathrm{new},i}(\tau,\hat{b}^{(0)}_{i\tau},\hat{p}^{(0)}_{i\tau},\hat{f})\right.
\end{aligned}$$

$$+ \sum_{i \in \text{split}(0)} \text{DER}_{1,\text{new},i}(\tau, \hat{b}_{i\tau}^{(1)}, \hat{p}_{i\tau}^{(1)}, \hat{f}) \Big\}.$$

Again, the above variance estimators are consistent under the semiparametric model as described in Section 3.3.2. The second-order influence functions are defined as follows.

Definition 32. *Define* $\hat{\omega}^{(\ell)}(\tau)$ *as the solution to*

$$\frac{1}{n} \sum_{i \in \text{split}(\ell)} \left\{ e^{\tau A_i} - e^{\omega(\tau) W_i} \right\} W_i = 0$$

and

$$\hat{q}_{i\tau}^{(\ell)}(x) = \begin{cases} \exp\left(\hat{\omega}^{(\ell)}(\tau)^T w(x)\right) & \textit{in the loglinear case,} \\ 1 & \textit{in the linear case.} \end{cases}$$

Next define

$$\begin{aligned}
\text{IF}_{22,ij}^{(k)}(\tau; b, p, q) &= -\varepsilon_i(\tau, b) K_k(X_i, X_j) \Delta_j(\tau, p, q), \\
\mathbb{IF}_{22}^{(k)}(\tau) &= \frac{1}{n} \sum_{i \in \text{split}(0)} \frac{1}{|s(i)|} \sum_{j \in s(i)} \text{IF}_{22,ij}^{(k)}(\tau, b_{i\tau}^*, p_{j\tau}^*, q_{j\tau}^*), \\
\widehat{\mathbb{IF}}_{22}^{\text{split},(k)}(\tau) &= \frac{1}{2n} \Big\{ \sum_{i \in \text{split}(0)} \frac{1}{|s(i)|} \sum_{j \in s(i)} \text{IF}_{22,ij}^{(k)}(\tau, \hat{b}_{i\tau}^{(1)}, \hat{p}_{j\tau}^{(1)}, \hat{q}_{j\tau}^{(1)}) \\
&\qquad + \sum_{i \in \text{split}(1)} \frac{1}{|s(i)|} \sum_{j \in s(i)} \text{IF}_{22,ij}^{(k)}(\tau, \hat{b}_{i\tau}^{(0)}, \hat{p}_{j\tau}^{(0)}, \hat{q}_{j\tau}^{(0)}) \Big\}, \\
\widehat{\mathbb{IF}}_{22}^{(k)}(\tau) &= \frac{1}{n} \sum_{i \in \text{split}(0)} \frac{1}{|s(i)|} \sum_{j \in s(i)} \text{IF}_{22,ij}^{(k)}(\tau, \hat{b}_{i\tau}^{(1)}, \hat{p}_{j\tau}^{(1)}, \hat{q}_{j\tau}^{(1)}), \\
\mathbb{IF}_{2}^{(k)}(\tau) &= \mathbb{IF}_{1,\text{new}}(\tau) + \mathbb{IF}_{22}(\tau), \\
\widehat{\mathbb{IF}}_{2}^{\text{split},(k)}(\tau) &= \widehat{\mathbb{IF}}_{1,\text{new}}^{\text{split}}(\tau) + \widehat{\mathbb{IF}}_{22}^{\text{split},(k)}(\tau), \textit{ and} \\
\widehat{\mathbb{IF}}_{2}^{(k)}(\tau) &= \widehat{\mathbb{IF}}_{1,\text{new}}(\tau) + \widehat{\mathbb{IF}}_{22}^{(k)}(\tau).
\end{aligned}$$

These definitions are made with the kernel K_k *of Definition 19.*

Definition 33. *Define the second-order estimators of* τ^* *as follows. Let* $\hat{\tau}_2^{(k)}$ *solve*

$$\widehat{\mathbb{IF}}_{2}^{(k)}(\tau) = 0,$$

and let $\hat{\tau}_2^{\text{split},(k)}$ *solve*

$$\widehat{\mathbb{IF}}_{2}^{\text{split},(k)}(\tau) = 0.$$

The variance of the above estimators is estimated by

$$\widehat{\mathbb{V}}_2^{(k)} = \frac{\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}_{1,\text{new}}\left(\hat{\tau}_2^{(k)}\right)\right) + \widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}_{22}^{(k)}\left(\hat{\tau}_2^{(k)}\right)\right)}{\left\{\widehat{\mathbb{DER}}\left(\widehat{\mathbb{IF}}_{1,\text{new}}\left(\hat{\tau}_2^{(k)}\right)\right)\right\}^2}, \textit{ and}$$

$$\widehat{\mathbb{V}}_2^{\text{split},(k)} = \frac{\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}_{1,\text{new}}^{\text{split}}\left(\hat{\tau}_2^{\text{split},(k)}\right)\right) + \widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}_{22}^{\text{split},(k)}\left(\hat{\tau}_2^{\text{split},(k)}\right)\right)}{\left\{\widehat{\mathbb{DER}}\left(\widehat{\mathbb{IF}}_{1,\text{new}}^{\text{split}}\left(\hat{\tau}_2^{\text{split},(k)}\right)\right)\right\}^2},$$

where

$$\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}_{22}^{\text{split},(k)}(\tau)\right) = \frac{1}{4n^2}\left\{\sum_{i\in\text{split}(0)}\frac{1}{|s(i)|^2}\sum_{j\in s(i)}\left\{\text{IF}_{22,ij}^{(k)}(\tau,\hat{b}_\tau^{(1)},\hat{p}_\tau^{(1)},\hat{q}_\tau^{(1)})\right\}^2 + \sum_{i\in\text{split}(1)}\frac{1}{|s(i)|^2}\sum_{j\in s(i)}\left\{\text{IF}_{22,ij}^{(k)}(\tau,\hat{b}_\tau^{(0)},\hat{p}_\tau^{(0)},\hat{q}_\tau^{(0)})\right\}^2\right\},$$

$$\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}_{22}^{(k)}(\tau)\right) = \frac{1}{n^2}\sum_{i\in\text{split}(1)}\frac{1}{|s(i)|^2}\sum_{j\in s(i)}\left\{\text{IF}_{22,ij}^{(k)}(\tau,\hat{b}_\tau^{(0)},\hat{p}_\tau^{(0)},\hat{q}_\tau^{(0)})\right\}^2,$$

and the other quantities are as in Definition 16.

Here as for Definition 31, the variance estimates are consistent under the semiparametric model. We prove this in the next section.

3.3.2 Conditional Bias and Variance Properties of the Time-Series Estimators

The bias and variance properties of estimators in the i.i.d. case need to be modified for the time-series case. We derive these properties under various assumptions collected below.

Bias

Assumption 34. *The bias calculations will make use of the following Assumption.*

IX For any i and any $j \in s(i)$, X_i and X_j are independent if $j \in s(i)$, i.e., the joint density $f_{ij}(X_i, X_j)$ factorizes as $f_i(X_i)f_j(X_j)$.

The Assumption IX will be approximately true if the lower bound n_1 for $s(i)$ (see Equation 3.1, Section 3.4.1) is not too small.

We define the bias of the half-sample time-series first-order estimator as follows.

Definition 35. *For any b, p, and any set f, the bias of $\mathbb{IF}_{1,\text{new}}(\tau; b, p, f)$ is defined by*

$$\text{Bias}_{1,\text{new}}(\tau, b, p, f) = E[\mathbb{IF}_{1,\text{new}}(\tau; b, p, f)] - E[\mathbb{IF}_{1,\text{new}}(\tau; b_\tau^*, p_\tau^*, f)].$$

We have the following lemma.

Lemma 36.

$$\begin{aligned}&\text{Bias}_{1,\text{new}}(\tau; b, p, f)\\ &= \frac{1}{n}\sum_{i\in\text{split}(1)}\frac{1}{|s(i)|}\sum_{j\in s(i)}E_i[f_j^*(X_i)q_{i\tau}^*(X_i)\{b_{i\tau}^*(X_i) - b(X_i)\}\{p_{i\tau}^*(X_i) - p(X_i)\}]\\ &\quad + \frac{1}{n}\sum_{i\in\text{split}(1)}\frac{1}{|s(i)|}\sum_{j\in s(i)}E_i[q_{i\tau}^*(X_i)\{f_j(X_i) - f_j^*(X_i)\}\{b_{i\tau}^*(X_i) - b(X_i)\}\{p_{i\tau}^*(X_i) - p(X_i)\}].\end{aligned}$$

As before, let

$$\delta b_{i\tau}(x) = b_{i\tau}^*(x) - \hat{b}_{i\tau}^{(0)}(x),\ \delta p_{i\tau}(x) = p_{i\tau}^*(x) - \hat{p}_{i\tau}^{(0)}(x),$$

where $\hat{b}^{(0)}_{i\tau}(x)$ and $\hat{p}^{(0)}_{i\tau}(x)$ are defined in as Definition 28. The above lemma implies that

$$E[\widehat{\mathbb{IF}}_{1,\text{new}}(\tau)] = \frac{1}{n}\sum_{i\in\text{split}(1)}\frac{1}{|s(i)|}\sum_{j\in s(i)}E_i[f^*_j(X_i)q^*_{i\tau}(X_i)\delta b_{i\tau}(X_i)\delta p_{i\tau}(X_i)] + \frac{1}{n}\sum_{i\in\text{split}(1)}\frac{1}{|s(i)|}\sum_{j\in s(i)}E_i[q^*_{i\tau}(X_i)\{f_j(X_i)-f^*_j(X_i)\}\delta b_{i\tau}(X_i)\delta p_{i\tau}(X_i)],$$

which has a second-order and a third-order term as in the i.i.d. case. As before, we use a second-order U-statistic to approximate the second-order bias above. The next theorem gives the bias of the second-order estimator.

Theorem 37. *Let $\widehat{\mathbb{IF}}^{(k)}_{22}(\tau)$ be as in Definition 32 and f be any fixed set of densities on the support of X. Suppose that Assumption IX of Assumption 34 holds. Then $\mathbb{IF}^{(k)}_{22}(\tau)$ estimates the bias of the first-order estimator up to a time-series bias, third-order, and truncation terms. Specifically,*

$$\begin{aligned}
&E[\widehat{\mathbb{IF}}^{(k)}_{22}(\tau)] + E[\widehat{\mathbb{IF}}_{1,\text{new}}(\tau)]\\
&= \frac{1}{n}\sum_{i\in\text{split}(1)}\frac{1}{|s(i)|}\sum_{j\in s(i)}\Bigg\{E_i[q^*_{i\tau}(X_i)\{f_j(X_i)-f^*_j(X_i)\}\delta b_{i\tau}(X_i)\delta p_{i\tau}(X_i)]\\
&\qquad - E_{ij}\left[q^*_{i\tau}(X_i)\delta b_{i\tau}(X_i)K_k(X_i,X_j)\delta p_{j\tau}(X_j)\left\{\frac{q^*_{j\tau}(X_j)}{\hat{q}_{j\tau}(X_j)}-1\right\}\right]\\
&\qquad - E_i[f^*_j(X_i)q^*_{i\tau}(X_i)\delta b_{i\tau}(X_i)\{\delta p_{j\tau}(X_i)-\delta p_{i\tau}(X_i)\}]\\
&\qquad + \int \Pi[f^*_i(x)q^*_{i\tau}(x)\delta b_{i\tau}(x)|\bar{\varphi}^{\perp}_k]\ \Pi[f^*_j(x)\delta b_{i\tau}(x)|\bar{\varphi}^{\perp}_k]dx\Bigg\}.
\end{aligned}$$

Hence the estimated influence function $\widehat{\mathbb{IF}}^{(k)}_{2,\text{new}}(\tau)$ has bias equal to a second-order time-series term (the third term in the expression), third-order terms (whose rate of convergence to 0 does not depend on k), plus a tail or truncation term (whose rate does depend on k).

Unfortunately, no higher-order influence function can cancel any part of the third bias term in the above expression because this term represents the bias resulting from the fact that the times t_i of observation are fixed and not random. For this reason we name this term the time-series bias. Note, however, that this term equals 0 if for all (i,j) either $\delta p_{i\tau}(x) = \delta p_{j\tau}(x)$ or $\delta b_{i\tau}(x) = \delta b_{j\tau}(x)$. Thus, in the setting of i.i.d. data, the term is exactly zero, owing to the fact that the conditional distribution of $(Y_i, A_i)|X_i$ is the same as the distribution of $(Y_j, A_j)|X_j$.

With i.i.d. data, higher-order influence functions can be used to essentially eliminate estimation bias up to any order. However, the same is not true of time-series data as the time-series bias of second order will always remain. The magnitude of $\{\delta p_{i\tau^*}(x) - \delta p_{j\tau^*}(x)\} = p^*_{i\tau^*}(x) - p^*_{j\tau^*}(x) - \{\hat{p}_{i\tau}(x) - \hat{p}_{j\tau}(x)\}$ and of $\{\delta b_{i\tau^*}(X_i) - \delta b_{j\tau^*}(X_i)\}$ will generally be less when $|t_i - t_j|$ is small because $p^*_{i\tau^*}(x)$, $\hat{p}_{j\tau}(x)$, $b^*_{i\tau^*}(x)$, and $\hat{b}_{i\tau}(x)$ are smooth functions of time.

Variance

We now turn our attention to the variance of the first- and second-order estimators and estimates of this variance. As noted above, the validity of our bias calculations did not require that either of the semiparametric regression models held. Indeed the only assumption required was the independence Assumption IX in Assumption 34. In contrast, we shall find that the validity of our

expressions for the variance and thus the consistency of our variance estimators will require much stronger assumptions. Below we discuss the underlying reason for the discrepancy between the assumptions required for the validity of our expressions for (and estimates of) the variance as compared to the bias. For brevity we use the notation $O_i = (Y_i, A_i, X_i)$.

Assumption 38. *The variance calculations will assume that the semiparametric model, i.e., one of the following, holds.*

LM The linear semiparametric regression model holds.
LLM The loglinear semiparametric regression model holds.

The variance calculations will also impose the following assumptions.

I For any i and $j \in s(i)$, O_i and O_j are independent.
M For any $i \neq i'$, $E_{ii'}[\varepsilon_i(\tau^*, b^*_{i\tau^*})|A_i, X_i, O_{i'}] = E_i[\varepsilon_i(\tau^*, b^*_{i\tau^*})|A_i, X_i] = 0$.
C1 For any $i \neq i'$, $j \in s(i)$, $j' \in s(i')$, $E_{i,i',j,j'}[\varepsilon_i(\tau^*, b^*_{i\tau^*})|A_j, X_j, A_{j'}, X_{j'}, O_{i'}, X_i] = E_i[\varepsilon_i(\tau^*, b^*_{i\tau^*})|X_i] = 0$.
C2 For any $i, i', i \neq i', j \in s(i) \cap s(i')$, $E_{i,i',j}[\Delta_{j\tau^*}(p^*_{j\tau^*}, q^*_{j\tau^*})|X_j, O_i, O_{i'}] = E_j[\Delta_{j\tau^*}(p^*_{j\tau^*}, q^*_{j\tau^*})|X_j]$.

We briefly discuss these assumptions next. Assumptions I and C2 are plausible if the lower bound used in the definition of $s(i)$ (Equation 3.1 in Section 3.4.1) is not too small. The Assumption M states that the mean of the residual from Y_i, given A_i and X_i, does not furthermore depend on Y_i, A_i and X_i on any other day $t_{i'}$. This would likely be violated when $|i - i'|$ is small. The same remarks apply to Assumption C1. Our variance estimators assume that covariance terms arising from this factor are absent or very small.

We first consider the variance of the first-order estimator. Since the estimators of $b^*_{i\tau}$, $p^*_{i\tau}$, and $q^*_{i\tau}$ are consistent (although not at rate $\sqrt{n}$), we have under weak conditions

$$\mathrm{Var}[\widehat{\mathbb{IF}}_{1,\mathrm{new}}(\tau^*)] = \mathrm{Var}[\mathbb{IF}_{1,\mathrm{new}}(\tau^*)]\,[1 + o_P(1)]\,,$$

and

$$\begin{aligned}\mathrm{Var}[\mathbb{IF}_{1,\mathrm{new}}(\tau^*)] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i\in\mathrm{split}(0)} \mathrm{IF}_{1,\mathrm{new}}(\tau^*, b^*_{i\tau^*}, p^*_{i\tau^*}, \hat{f})\right] \\ &= n^{-2}\sum_{i=1}^{n} \mathrm{Var}_i\left[\mathrm{IF}_{1,\mathrm{new},i}(\tau^*, b^*_{i\tau^*}, p^*_{i\tau^*}, \hat{f})\right] \\ &\quad + 2n^{-2}\sum_{i<i'} \mathrm{Cov}_{i,i'}\left[\mathrm{IF}_{1,\mathrm{new},i}(\tau^*, b^*_{i\tau^*}, p^*_{i\tau^*}, \hat{f}), \mathrm{IF}_{1,\mathrm{new},i'}(\tau^*, b^*_{i'\tau^*}, p^*_{i'\tau^*}, \hat{f})\right],\end{aligned}$$

where

$$\begin{aligned}\mathrm{Var}_i\left[\mathrm{IF}_{1,\mathrm{new},i}(\tau^*, b^*_{i\tau^*}, p^*_{i\tau^*}, \hat{f})\right] &= E_i\left[\left(\mathrm{IF}_{1,\mathrm{new},i}(\tau^*, b^*_{\tau^*}, p^*_{\tau^*}, \hat{f})\right)^2\right] \\ &\quad - \left\{E_i\left[\mathrm{IF}_{1,\mathrm{new},i}(\tau^*, b^*_{\tau^*}, p^*_{\tau^*}, \hat{f})\right]\right\}^2.\end{aligned}$$

Since $E[\mathbb{IF}_{1,\mathrm{new}}(\tau^*, b^*_{\tau^*}, p^*_{\tau^*}, \hat{f})] = 0$ does not imply $E_i\left[\mathrm{IF}_{1,\mathrm{new},i}(\tau^*, b^*_{\tau^*}, p^*_{\tau^*}, \hat{f})\right] = 0$ in the nonparametric model, we cannot ignore the second term above. We can estimate $E_i\left[\mathrm{IF}_{1,\mathrm{new},i}(\tau^*, b^*_{\tau^*}, p^*_{\tau^*}, \hat{f})\right]$ consistently without assuming it is zero (and thus without assuming the semiparametric model holds) as follows; to keep it simple, we consider the linear case.

$$E_i\left[\mathrm{IF}_{1,\mathrm{new},i}(\tau^*, b^*_{i\tau^*}, p^*_{i\tau^*}, \hat{f})\right]$$

$$= \frac{1}{|s(i)|} \sum_{j \in s(i)} E_i \left[\hat{f}_j(X_i) \left\{ E_i \left[(Y_i - \tau^* A_i) A_i | X_i \right] - b^*_{i\tau^*}(X_i) p^*_{i\tau^*}(X_i) \right\} \right].$$

We can obtain a consistent estimator of $E_i\left[\mathrm{IF}_{1,\mathrm{new},i}(\tau^*, b^*_{i\tau^*}, p^*_{i\tau^*}, \hat{f})\right]$ if we construct an estimate, smoothed over time and X, of $E_i[(Y_i - \tau^* A_i) A_i | X_i = x]$ [in addition to estimates of $b^*_{i\tau^*}(X_i)$ and $p^*_{i\tau^*}(X_i)$] and then integrate the resulting function of X with respect to $\hat{f}_i(X)$. However doing so would create many more computational and analytic difficulties so we take the simpler, but less robust, option of assuming the semiparametric model holds.

Assuming the loglinear or linear model of Assumption 38, we get

$$\begin{aligned}
&\mathrm{Var}(\mathbb{IF}_{1,\mathrm{new}}(\tau^*, b^*_{\tau^*}, p^*_{\tau^*}, \hat{f})) \\
&= n^{-2} \sum_{i=1}^{n} E_i \left[\left(\mathrm{IF}_{1,\mathrm{new},i}(\tau^*, b^*_{i\tau^*}, p^*_{i\tau^*}, \hat{f}) \right)^2 \right] \\
&\quad + 2n^{-2} \sum_{i<i'} E_{ii'} \left[\mathrm{IF}_{1,\mathrm{new},i}(\tau^*, b^*_{i\tau^*}, p^*_{i\tau^*}, \hat{f}) \mathrm{IF}_{1,\mathrm{new},i'}(\tau^*, b^*_{i'\tau^*}, p^*_{i'\tau^*}, \hat{f}) \right]
\end{aligned}$$

If Assumption M of Assumption 38 also holds, then the covariance term above equals 0 for all $i \neq i'$. Dominici et al. (2004) impose Assumption M in their analysis of the NMMAPS data. The assumption states that the mean of the residual number of deaths at time t_i, given the 1-day lagged PM_{10} and weather on day t_i, does not furthermore depend on the number of deaths, the pollution level, or the temperature variables on any other day $t_{i'}$. This would likely be violated when $|i - i'|$ is small. Nonetheless, to keep from having to model the covariance structure, our variance estimator assumes that M holds. Thus, we obtain the following.

Theorem 39. *Under the semiparametric loglinear or linear model of Assumption 38 and Assumption M of Assumption 38, if $\hat{\tau}$ converges to τ^* in probability,*

$$\mathrm{Var}\left(\widehat{\mathbb{IF}}_{1,\mathrm{new}}(\hat{\tau})\right) = n^{-2} \sum_{i=1}^{n} E_i \left[\left(\mathrm{IF}_{1,\mathrm{new},i}(\tau^*, b^*_{i\tau^*}, p^*_{i\tau^*}, \hat{f}) \right)^2 \right] [1 + o_P(1)],$$

and

$$\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}_{1,\mathrm{new}}(\hat{\tau})\right) = n^{-2} \sum_{i=1}^{n} \left[\left(\mathrm{IF}_{1,\mathrm{new},i}(\hat{\tau}, \hat{b}_{i\hat{\tau}}, \hat{p}_{i\hat{\tau}}, \hat{f}) \right)^2 \right]$$

is a consistent estimator of the variance.

We next consider the variance at $\tau = \tau^*$ of

$$\widehat{\mathbb{IF}}_{22}(\tau) = \frac{1}{n} \sum_i \frac{1}{|s(i)|} \sum_{j \in s(i)} \mathrm{IF}_{22,ij}(\tau, \hat{b}_{i\tau}, \hat{p}_{j\tau}, \hat{q}_{j\tau}).$$

We have the following theorem.

Theorem 40. *Suppose that the semiparametric loglinear or linear model of Assumption 38 holds; Assumptions I, C1, and C2 of Assumption 38 hold; $\hat{b}_{i\tau}, \hat{p}_{i\tau}, \hat{q}_{i\tau}$ converge to $b^*_{i\tau}, p^*_{i\tau}, q^*_{i\tau}$ in probability; and $\hat{\tau}$ is a consistent estimator of τ^*. Then*

$$\mathrm{Var}\left(\widehat{\mathbb{IF}}^{(k)}_{22}(\tau)\right) = \mathrm{Var}\left(\sum_{i \in \mathrm{split}(0)} \frac{1}{|s(i)|} \sum_{j \in s(i)} \mathbb{IF}^{(k)}_{22,ij}(\tau, b^*_{i\tau}, p^*_{j\tau}, q^*_{j\tau}) \right) [1 + o_P(1)]$$

and

$$\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}_{22}(\hat{\tau})\right) = \frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{|s(i)|^2}\sum_{j\in s(i)}\left\{\mathrm{IF}_{22,ij}(\tau,\hat{b}_{i\tau},\hat{p}_{j\tau},\hat{q}_{j\tau})\right\}^2$$

is a consistent estimator of the variance of $\widehat{\mathbb{IF}}_{22}(\tau^*)$.

This theorem allows us to consistently estimate the variance of $\widehat{\mathbb{IF}}_2^{(k)}(\tau^*) = \widehat{\mathbb{IF}}_{1,\mathrm{new}}(\tau^*) + \widehat{\mathbb{IF}}_{22}^{(k)}(\tau^*)$ by $\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}_{1,\mathrm{new}}(\hat{\tau})\right) + \widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}_{22}^{(k)}(\hat{\tau})\right)$ since $\mathbb{IF}_{22}^{(k)}(\tau^*)$ and $\mathbb{IF}_{1,\mathrm{new}}(\tau^*)$ are uncorrelated under our assumptions.

We next give a result on the asymptotic rate of increase of the variance of the second-order estimator with k.

Lemma 41. *Under the assumptions of Theorem 40, and the assumption that the* ∞*-norms of* f_i *are bounded uniformly over i,*

$$\mathrm{Var}\left[\widehat{\mathbb{IF}}_{22}^{(k)}(\hat{\tau})\right] = O\left(\frac{k}{ns}\right),$$

where

$$\frac{1}{s} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|s(i)|}.$$

We note that the i.i.d. second-order U-statistic is a special case of the above theorem, with $s(i) = \mathrm{split}(0)$ or $\mathrm{split}(1)$ [depending on whether $i \in \mathrm{split}(0)$ or $i \in \mathrm{split}(1)$]. Hence $s = n$ in the i.i.d. case. This proves the claim made in the section on the variance properties of the i.i.d. estimators that the variance is $O(k/n^2)$. Robins, Li, et al. (2008) show that the variance rate is $O(k/n^2)$ even when the semiparametric model does not hold.

Note that there is a term in the variance of $\widehat{\mathbb{IF}}_{22}^{(k)}(\tau^{**})$ that becomes negligible as $k \to \infty$, but for small k can contribute to the variance through a first-order term in the Hoeffding decomposition of the U-statistic. For this reason, for small k our estimate of the variance of $\widehat{\mathbb{IF}}_{22}^{(k)}(\hat{\tau})$ may be too small.

Note that all of the above results are derived using the half-sample estimator conditional on the other half-sample. We address their generalization to the other estimators next.

3.3.3 Unconditional Bias, Variance, and Rates of Convergence for the Estimators

In our bias and variance results thus far, we evaluated the conditional bias and variance of estimators based on the estimated influence function calculated from one half-sample, given the other half-sample — that is, the bias of the half-sample estimators conditional on the training sample.

We now use the above results to examine the unconditional distributions of $\widehat{\mathbb{IF}}_{1,\mathrm{new}}^{\mathrm{split}}(\tau^*)$, $\hat{\tau}_{1,\mathrm{new}}^{\mathrm{split}}$, $\widehat{\mathbb{IF}}_2^{\mathrm{split},(k)}(\tau^*)$, and $\hat{\tau}_2^{\mathrm{split},(k)}$. We will impose the assumptions used in deriving the variances of the last subsection. In particular we will assume the semiparametric model holds. In that case $E_i[IF_{1,\mathrm{new},i}(\tau^*, b_{\tau^*}^*, p_{\tau^*}^*, \hat{f})]$ has mean zero even when, as we again assume, $\hat{f}$ is obtained based on all the data justifying our decision to use all the data to estimate the true densities f^*. We now show that whenever the variances of $\widehat{\mathbb{IF}}_{1,\mathrm{new}}^{\mathrm{split}}(\tau^*)$, $\hat{\tau}_{1,\mathrm{new}}^{\mathrm{split}}$, $\widehat{\mathbb{IF}}_2^{\mathrm{split},(k)}(\tau^*)$, and $\hat{\tau}_2^{\mathrm{split},(k)}$ exceed their respective squared bias, $\widehat{\mathbb{IF}}_{1,\mathrm{new}}^{\mathrm{split}}$, $\hat{\tau}_{1,\mathrm{new}}^{\mathrm{split}} - \tau^*$, $\widehat{\mathbb{IF}}_2^{\mathrm{split},(k)}$, and $\hat{\tau}_2^{\mathrm{split},(k)} - \tau^*$, when divided by their respective standard error estimates, will be unconditionally $N(0,1)$ in large samples, resulting in valid standard Wald CIs for τ^* and valid tests of the null hypothesis $\tau = \tau^*$.

This result follows from the fact (as discussed in Robins, Li, et al., 2008) that when the variances exceed the squared biases, the four estimation sample statistics $\widehat{\mathbb{IF}}_{1,\text{new}}(\tau^*)$, $\hat{\tau}_{1,\text{new}} - \tau^*$, $\widehat{\mathbb{IF}}_2^{(k)}(\tau^*)$, and $\hat{\tau}_2^{(k)} - \tau^*$ are conditional on the other half-sample used to estimate b_τ^*, p_τ^*, and q_τ^* (the training sample), asymptotically $N(0,1)$ when standardized by the inverse of their respective standard errors. They are also asymptotically independent of the corresponding statistics with the roles of the estimation and training samples reversed.

On the other hand, when the squared bias exceeds the variance, the unconditional distributions of $\widehat{\mathbb{IF}}_{1,\text{new}}^{\text{split}}(\tau^*)$, $\hat{\tau}_{1,\text{new}}^{\text{split}} - \tau^*$, $\widehat{\mathbb{IF}}_2^{\text{split},(k)}(\tau^*)$, and $\hat{\tau}_2^{\text{split},(k)} - \tau^*$ may have variances that exceed the conditional variances and exceed the large-sample limit of the conditional variance estimators derived in the previous section. Further the unconditional biases may differ from the conditional biases above. We now examine the rates of convergence of these estimators.

Suppose that the functions $b_{i\tau}^*$ and $\hat{b}_{i\tau}$ lie in the Hölder class $H(C_b, \beta_b)$; $p_{i\tau}^*$ and $\hat{p}_{i\tau}$ lie in $H(C_p, \beta_p)$; f_i and $f_i^* \in H(C_f, \beta_f)$; and $q_{i\tau}^*$ and $\hat{q}_{i\tau} \in H(C_q, \beta_q)$. We assume that for each of b, p, f, q, the functions at all times belong to the same Hölder class, as would be true if there were a lower bound on the Hölder exponents of the functions indexed by time. We will need the following well-known facts.

a) The Hölder exponent for the product $g(x)h(x)$ is the minimum of the Hölder exponents β_h and β_g.

b) For a function $h(x)$ with Hölder exponent β_h and support on the unit cube in $\mathbb{R}^d$, the tail norms $\left\{\int \left\{\prod \left[h(x)|\overline{\varphi}_k^{\perp}(x)\right]\right\}^2 dx\right\}^{1/2}$ and $\left\{E\left\{\prod{}^2 \left[h(X)|\overline{\varphi}_k^{\perp}(X)\right]\right\}\right\}^{1/2}$ are $O(k^{-\beta_h/d})$ as $k \to \infty$ for optimal orthonormal bases (including Legendre polynomials and natural splines) and when $\beta_h \geq 1$ decreases at rate $k^{-1/d}$ for the Haar basis. Thus for functions with more than one derivative, the Legendre basis approximates better than the Haar basis. This is relevant because the functions $f^*(x)$, $q_\tau^*(x)$, $\delta b_\tau(x)$, and $\delta p_\tau(x)$ all almost certainly have more than one derivative.

   The Cauchy-Schwartz inequality applied with the above shows that

$$\int \prod \left[h(x)|\overline{\varphi}_k^{\perp}(x)\right] \prod \left[g(x)|\overline{\varphi}_k^{\perp}(x)\right] dx = O(k^{-(\beta_g+\beta_h)/d})$$

   for optimal bases. We also have

$$E_{\hat{f}}\left\{\prod_{\hat{f}} \left[g(X)h(X)|\left\{\overline{\varphi}_k(X)g(X)\right\}^{\perp}\right]^2\right\} = O(k^{-\beta_h/d}).$$

   This depends on β_h rather than $\min\{\beta_g, \beta_h\}$ because the term $g(X)$ occurs in both the function being projected and the subspace being projected onto, and can thus be factored out of the integral. This explains why second-order estimators based on the kernel $K_{q,f,k,\text{alt}}$ have a potentially smaller truncation bias, as discussed below.

c) The optimal rate of estimation in $\mathscr{L}_2(f^*)$ norm of either the density $f(x)$ of X or of a (possibly weighted) conditional expectation given X, say $h(X)$, is of order $n^{-\frac{\beta}{2\beta+d}}$ if the density or conditional expectation has Hölder exponent β. Adaptive optimal estimators can be constructed that obtain these optimal rates and are as smooth as their target function.

We note the following.

- If $\beta_b > d/2$ and $\beta_p > d/2$, then the classes $H(C_b,\ \beta_b)$ and $H(C_p, \beta_p)$ are P^*-Donsker. Hence τ^* can be consistently estimated at $\sqrt{n}$ rates using standard results. If $\beta_b \le d/2$ or $\beta_p \le d/2$, these classes fail to be Donsker (as shown in van der Vaart and Wellner, 1996). This complicates analysis of the performance of estimators based on $\mathbb{IF}^{\text{full}}_{1,\text{eff}}(\tau; \hat{b}_\tau, \hat{p}_\tau)$.
- Suppose we are in the linear semiparametric model. Then, irrespective of the magnitude of β_b and β_p, estimators $\hat{b}_\tau$ and $\hat{p}_\tau$ that are linear combinations of the first k elements of an optimal tensor-product basis (such as products of splines or polynomials in each coordinate; see, e.g., Theorem 8, Chapter 6 of Lorentz, 1986, for polynomial bases and Theorem 12.8 of Schumaker, 1981, for splines) can be constructed based on the full-sample, such that

$$E\left[\mathbb{IF}^{\text{full}}_{1,\text{eff}}\left(\tau; \hat{b}_\tau, \hat{p}_\tau\right)\right] - E\left[\mathbb{IF}^{\text{full}}_{1,\text{eff}}\left(\tau; b^*_\tau, p^*_\tau\right)\right] = O_P(k^{-(\beta_b+\beta_p)/d}),$$

  as proved in Donald and Newey (1994). The proof does not use Donsker-type results. Using ideas similar to those in Donald and Newey (1994), we show in Appendix D that non-series Poisson regression estimators with this bias rate exist under the loglinear model. We do not have a proof that the series estimator based on Poisson regression can achieve this rate, but we conjecture that this is the case based on simulations.
- We continue with the assumption of the semiparametric model. If $\beta_b + \beta_p > d/2$, full-sample estimators $\hat{b}_\tau$ and $\hat{p}_\tau$ can be found such that $\sqrt{n}$-rate convergence is achieved using $\mathbb{IF}^{\text{full}}_{1,\text{eff}}(\tau; \hat{b}_\tau, \hat{p}_\tau)$. This holds even if $\beta_b \le d/2$ or $\beta_p \le d/2$, causing Donsker conditions to fail. This follows in the linear case from the aforementioned result in Donald and Newey (1994). Based on the above comments, we conjecture that this is true in the loglinear case as well.

Since, in estimating k coefficients of the basis representation of b^*_τ and p^*_τ from a half-sample of size n, we must have $k \le n$, it follows that $\beta_b + \beta_p \ge d/2$ is a necessary condition for $\sqrt{n}$ convergence of the root MSE of series estimators, such as the usual Poisson and linear regression estimators considered by the NMMAPS investigators Dominici et al. (2004), which are the same as those considered in Donald and Newey (1994).

- Now suppose we are in the nonparametric model, and suppose that $\beta_b + \beta_p > d/2$. Estimators based on $\mathbb{IF}^{\text{full}}_{1,\text{eff}}(\tau; \hat{b}_\tau, \hat{p}_\tau)$ will not have a bias of $O_P(k^{-(\beta_b+\beta_p)/d})$ in this case. Such estimators cannot achieve the parametric rate of $\sqrt{n}$. However, the $\sqrt{n}$ rate can still be achieved using higher-order influence-function estimators, although the order of influence functions may need to be arbitrarily large. (Such estimators are constructed in Robins, Li, et al., 2008.)
- If $\beta_b + \beta_p \le d/2$, then
  - Under the nonparametric model, no estimator for τ^* exists such that the bias and standard deviation converge at the rate $n^{-1/2}$. This is conjectured to be true in the semiparametric model as well.
  - One of $\beta_b \le d/2$ or $\beta_p \le d/2$ must be true. Hence Donsker conditions do not hold and cannot be used to analyze the performance of $\mathbb{IF}^{\text{full}}_{1,\text{eff}}(\tau; \hat{b}_\tau, \hat{p}_\tau)$. However, $\mathbb{IF}^{\text{full}}_{1,\text{eff}}(\tau; \hat{b}_\tau, \hat{p}_\tau)$ does not give optimal rates for any estimators $\hat{b}_\tau$ and $\hat{p}_\tau$ because the variance is of order $1/n$ and the squared bias is larger, leading to a non-optimal tradeoff between the two. This remark holds under both the nonparametric and semiparametric models.

Suppose that we use an optimal basis in our second-order influence functions. The above facts lead to the following consequences.

i) The third-order estimation bias terms in the bias of all three second-order i.i.d. estimators (Theorem 21), as well as the third-order estimation bias term in the bias of the second-order time-series estimator (Theorem 37), all converge to zero at rates determined by either the product of $\delta b_{i\tau^*}(X_i)$, $\delta q_{j\tau^*}(X_j)$, and $\delta p_{j\tau^*}(X_j)$, leading to the rate $O_P\left(n^{-\left(\frac{\beta_b/d}{2\beta_b/d+1}+\frac{\beta_p/d}{2\beta_p/d+1}+\frac{\beta_q/d}{2\beta_q/d+1}\right)}\right)$, or the product of $\delta b_{i\tau^*}(X_i)$, $\delta p_{j\tau^*}(X_j)$, and $\delta f_{j\tau^*}(X_j)$, leading to the rate $O_P\left(n^{-\left(\frac{\beta_b/d}{2\beta_b/d+1}+\frac{\beta_p/d}{2\beta_p/d+1}+\frac{\beta_f/d}{2\beta_f/d+1}\right)}\right)$.

ii) The time-series bias term in Theorem 37 converges to zero at a rate that is at least as fast as the product of the rates for $\delta b_{i\tau^*}(X_i)$ and $\delta p_{j\tau^*}(X_j)$, leading to an upper bound of $O_P\left(n^{-\left(\frac{\beta_b/d}{2\beta_b/d+1}+\frac{\beta_p/d}{2\beta_p/d+1}\right)}\right)$. Under the additional assumption that the residuals from the estimation of $b^*_{i\tau}$ and $p^*_{i\tau}$ do not change rapidly with time, and if we assume the upper bound on $s(i)$ is not too large, this term goes to zero at a rate as fast as the third-order terms.

iii) The truncation term for the i.i.d. second-order estimators based on $\widehat{\mathbb{IF}}^{(k)}_{2,\text{eff}}(\tau)$, $\widehat{\mathbb{IF}}^{(k)}_{2,\text{new}}(\tau)$, and the time-series second-order $\hat{\tau}^{(k)}_2$ is $O_P(k^{-(\min\{\beta_q,\beta_f,\beta_b\}+\min\{\beta_p,\beta_f\})/d})$.

iv) The truncation term for the i.i.d. second-order estimators based on $\widehat{\mathbb{IF}}^{(k)}_{2,\text{alt}}(\tau)$ yields the rate $O_P(k^{-(\beta_b+\beta_p)/d})$, which may be better than the rate for the truncation term in the other second-order estimators.

Thus, in the i.i.d. case with kernels K_k and $K_{\hat{f},k}$ the rate of convergence of the bias to zero is

$$\max\left\{n^{-\left(\frac{\beta_b/d}{2\beta_b/d+1}+\frac{\beta_p/d}{2\beta_p/d+1}+\frac{\beta_q/d}{2\beta_q/d+1}\right)}, n^{-\left(\frac{\beta_b/d}{2\beta_b/d+1}+\frac{\beta_p/d}{2\beta_p/d+1}+\frac{\beta_f/d}{2\beta_f/d+1}\right)}, k^{-(\min\{\beta_q,\beta_f,\beta_b\}+\min\{\beta_p,\beta_f\})/d}\right\}$$

and in the i.i.d. case with kernel $K_{\hat{q},\hat{f},k,\text{alt}}$,

$$\max\left\{n^{-\left(\frac{\beta_b/d}{2\beta_b/d+1}+\frac{\beta_p/d}{2\beta_p/d+1}+\frac{\beta_q/d}{2\beta_q/d+1}\right)}, n^{-\left(\frac{\beta_b/d}{2\beta_b/d+1}+\frac{\beta_p/d}{2\beta_p/d+1}+\frac{\beta_f/d}{2\beta_f/d+1}\right)}, k^{-(\beta_b+\beta_p)/d}\right\}.$$

For the time-series case, the above rates apply if the time-series bias term goes to zero at least as fast as the third-order terms.

In both the time-series and i.i.d. cases, the variance goes to zero at rate k/ns. To make sure that the variance is at least as large as the squared bias, we choose k such that k/ns exceeds the squares of the bias terms above. This is possible because, as k increases, the bias decreases and the variance increases.

The dependence of s on n is known. In the i.i.d. case, $s = n$. In the time-series case, we consider an asymptotic sequence with $s = s(n) \asymp n^\nu$ with $\nu < 1$. Hence we can use the formulae above to choose k to achieve the lowest possible rate for a given ν if the smoothnesses β_f, β_q, β_b, and β_p are known. If only some of these smoothnesses are known, additional conditions are required. For example, if we assume that

(i') the densities $f^*_i(x)$ are at least as smooth as the smoothest of $b^*_{i\tau^*}(x)$ and $p^*_{i\tau^*}(x)$, and

(ii') $q^*_{i\tau^*}(x)$ is at least as smooth as $b^*_{i\tau^*}(x)$ so that $\min\{\beta_b,\beta_q,\beta_f\} = \beta_b$ and $\min\{\beta_p,\beta_f\} = \beta_p$,

then the above rates no longer depend on β_q and β_f so knowledge of β_b and β_p is all that is required.

The truncation bias rate is better if we use a time-series version of the estimator based on $\widehat{\mathbb{IF}}^{(k)}_{2,\text{alt}}(\tau)$. However, as mentioned previously, computation of $\widehat{\mathbb{IF}}^{(k)}_{2,\text{alt}}(\tau)$ involves an inversion of a large matrix that we were unable to compute for large k.

The large-sample variance of $\widehat{\mathbb{IF}}^{(k)}_2$ is the sum of the variances of $\widehat{\mathbb{IF}}_{1,\text{new}}$ and $\widehat{\mathbb{IF}}^{(k)}_{22}$ since these are asymptotically orthogonal to each other. Hence $\text{Var}(\widehat{\mathbb{IF}}^{(k)}_2) \asymp 1/n + k/ns$ in the time-series case and $\text{Var}(\widehat{\mathbb{IF}}^{(k)}_2) \asymp 1/n + k/n^2$ in the i.i.d. case. It increases linearly in $k = k(n)$ if $n = o\big[k(n)\big]$. In order to get valid CIs, we need this variance to be larger than the square of the bias above.

3.3.4 A Biasedness Test for First-Order Estimators

In practice, we do not know the smoothnesses β_b and β_p of $b^*_{i\tau}$ and $p^*_{i\tau}$. Hence an adaptive method is required to select an optimal k to minimize the MSE. Although various adaptive methods have been considered in the literature, for example the methods in Wasserman (2006), they are not directly applicable to our case and furthermore are only asymptotic. Thus the problem of finding the optimal k remains open in general.

We can, however, test whether the first-order estimator does indeed have squared bias dominating its variance by comparing this estimator to the second-order estimators for various values of k under the following assumption: For the maximum k used, the squared bias of $\widehat{\mathbb{IF}}^{\text{split},(k)}_2$ is less than the total variance. Now, the variance is dominated by that of $\widehat{\mathbb{IF}}^{\text{split},(k)}_{22}(\tau^*) = \widehat{\mathbb{IF}}^{\text{split},(k)}_2(\tau^*) - \widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}(\tau^*)$. Further, as argued above, $\widehat{\mathbb{IF}}_{1,\text{new}}(\tau^*)$ and $\widehat{\mathbb{IF}}^{(k)}_{22}(\tau^*)$ are asymptotically uncorrelated. Thus, when $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}(\tau^*)$ has bias squared less than its variance (which is of the order $\sqrt{n}$), $\widehat{\mathbb{IF}}^{\text{split},(k)}_{22}(\tau^*) = \widehat{\mathbb{IF}}^{\text{split},(k)}_2(\tau^*) - \widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}(\tau^*)$ is asymptotically normal for all k values in the middle graph (in Figure 1) with variance equal to the variance of $\widehat{\mathbb{IF}}^{\text{split},(k)}_2(\tau^*)$ minus the variance of $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}(\tau^*)$. Thus, we can test the null hypothesis [that $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}(\tau^*)$ has bias squared less than its variance] by (1) comparing $\widehat{\mathbb{IF}}^{\text{split},(k)}_{22}(\hat{\tau}) = \widehat{\mathbb{IF}}^{\text{split},(k)}_2(\hat{\tau}) - \widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}(\hat{\tau})$ divided by the square root of the difference of the estimators of the variance of $\widehat{\mathbb{IF}}^{\text{split},(k)}_2(\tau^*)$ and $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}(\tau^*)$, and then (2) comparing the resulting Z-score to the two 2.5% tails of a $N(0,1)$ distribution. We must correct the resulting p-value for the multiple tests (one for each value of k on the middle graph in Figure 1), which we do by Bonferroni correction as we do not have an estimator of the correlation matrix of the vector of the $\widehat{\mathbb{IF}}^{\text{split},(k)}_{22}(\hat{\tau})$.

The exact same results hold when we replace $\widehat{\mathbb{IF}}^{\text{split},(k)}_2(\tau^*)$ with $\hat{\tau}^{\text{split},(k)}_2 - \tau^*$ and $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}(\tau^*)$ with $\hat{\tau}^{\text{split}}_{1,\text{new}} - \tau^*$ because (1) $\widehat{\mathbb{IF}}^{\text{split},(k)}_2(\tau^*)$ and $\hat{\tau}^{\text{split},(k)}_2 - \tau^*$ have the same asymptotic distribution, except the variance of $\hat{\tau}^{\text{split},(k)}_2 - \tau^*$ is the variance of $\widehat{\mathbb{IF}}^{\text{split},(k)}_2(\tau^*)$ divided by the limit of the square of $\widehat{\text{DER}}\Big[\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}(\hat{\tau})\Big]$, and (2) the variance of $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}(\tau^*)$ and $\hat{\tau}^{\text{split}}_{1,\text{new}} - \tau^*$ differ by the same factor.

If the test of the null hypothesis that $\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}(\tau^*)$ has bias squared less than its variance fails to reject (as in all cities except possibly Minneapolis; see Figure 1), then we take the first-order estimator $\hat{\tau}^{\text{split}}_{1,\text{new}}$ as our estimator of τ^* when our goal is to minimize MSE. This choice is consistent with the optimal procedures for adaptive estimation with the goal of minimizing MSE (discussed on pages 216–220 of Wasserman, 2006).

Equivalently, we could use $\hat{\tau}^{\text{full}}_{1,\text{eff}}$, the Poisson regression estimator of τ^* used in previous NMMAPS analyses, since this estimator has variance somewhat less than that of $\hat{\tau}^{\text{split}}_{1,\text{new}}$.

3.4 METHOD DETAILS

In this section we derive some theoretical consequences of the choice of various aspects of the estimation process that we have so far left unspecified.

3.4.1 Definition of $s(i)$

For the NMMAPS data, we let the set $s(i)$ depend on parameters n_1, n_2, and $\gamma, 0 < \gamma \leq 1$, as follows.

$$s(i) = \{j : n_1 \leq |j - i| \leq n_2\} \cap \left\{j : \hat{f}_j(X_j) \leq \gamma \max\left\{\hat{f}_{j'}(X_{i'}) : i', j' = 1, \ldots, N\right\}\right\}. \tag{3.1}$$

(For the split estimator, the above was also intersected with split(ℓ) if $i \in$ split(ℓ).) For our analysis, the most common choices (our base case) were $n_1 = 25$, $n_2 = 75$, and $\gamma = 0.8$.

Recall from Section 3.3.2 that, under our Assumptions 34 and 38,

$$\mathrm{Var}(\widehat{\mathbb{IF}}_{1,\mathrm{new}}) = \frac{1}{n} C_1 \left[1 + o(1)\right],$$

where

$$\begin{aligned} C_1 &= \frac{1}{n} \sum_{i=1}^{n} E_i[\mathrm{IF}_{1,\mathrm{new},i}(\tau^*, b^*_{i\tau^*}, p^*_{i\tau^*}, \hat{f})^2], \\ &E_i[\mathrm{IF}_{1,\mathrm{new},i}(\tau^*, b^*_{i\tau^*}, p^*_{i\tau^*}, \hat{f})^2] \\ &\quad = \frac{1}{|s(i)|^2} \sum_{j \in s(i)} \sum_{j' \in s(i)} E_i[\hat{f}_j(X_i) \hat{f}_{j'}(X_i) \varepsilon_i(\tau^*, b^*_{i\tau^*})^2 \Delta_{i\tau}(p^*_{i\tau^*}, q^*_{i\tau^*})^2], \end{aligned}$$

which is $O(1)$. Hence $C_1 = O(1)$. On the other hand

$$\mathrm{Var}(\widehat{\mathbb{IF}}^{(k)}_{22}) = \frac{1}{n^2} \sum_{i=1}^{n} \frac{1}{|s(i)|^2} \sum_{j \in s(i)} E_{ij}[\varepsilon_i(\tau^*, b^*_{i\tau^*})^2 K_k(X_i, X_j)^2 \Delta_{j\tau}(p^*_{j\tau^*}, q^*_{j\tau^*})^2] \left[1 + o(1)\right]$$

and

$$\mathrm{Var}(\widehat{\mathbb{IF}}^{(k)}_{22}) = \frac{k}{ns} C_{22} \left[1 + o_p(1)\right],$$

where $\frac{1}{s} = n^{-1} \sum_{i=1}^{n} \frac{1}{|s(i)|}$, with $|s(i)|$ being the cardinality of the set $s(i)$ as before, $C_{22} = O(1), s = O(1)$ (see proof of Lemma 41 in Appendix E).

Since, under our assumptions, $\mathrm{Var}(\widehat{\mathbb{IF}}^{(k)}_2) = \mathrm{Var}(\widehat{\mathbb{IF}}^{(k)}_{22}) + \mathrm{Var}(\widehat{\mathbb{IF}}_{1,\mathrm{new}})$, we conclude that

$$\mathrm{Var}(\widehat{\mathbb{IF}}^{(k)}_2) / \mathrm{Var}(\widehat{\mathbb{IF}}_{1,\mathrm{new}}) = 1 + \mathrm{Var}(\widehat{\mathbb{IF}}^{(k)}_{22}) / \mathrm{Var}(\widehat{\mathbb{IF}}_{1,\mathrm{new}}) = 1 + \frac{k}{s} \frac{C_{22}}{C_1}.$$

Thus we expect that $\mathrm{Var}(\widehat{\mathbb{IF}}_{1,\mathrm{new}})$ does not depend strongly on $|s|$ but that the slope of $\mathrm{Var}(\widehat{\mathbb{IF}}^{(k)}_2) / \mathrm{Var}(\widehat{\mathbb{IF}}_{1,\mathrm{new}})$ versus k is inversely proportional to $|s|$. Finally, one would expect that the relative variability of $\hat{\tau}^{\mathrm{split}}_{1,\mathrm{new}}$ would be less than even that of $\widehat{\mathbb{IF}}^{\mathrm{split}}_{1,\mathrm{new}}(0)$ because (1) the denominator $\left\{\widehat{\mathbb{DER}}\left[\widehat{\mathbb{IF}}^{\mathrm{split}}_{1,\mathrm{new}}(\hat{\tau}^{\mathrm{split}}_{1,\mathrm{new}})\right]\right\}^2$ of the estimated variance of $\hat{\tau}^{\mathrm{split}}_{1,\mathrm{new}}$ has the same dependence on $s(i)$ as does the numerator, the estimated variance of $\widehat{\mathbb{IF}}^{\mathrm{split}}_{1,\mathrm{new}}(\hat{\tau}^{\mathrm{split}}_{1,\mathrm{new}})$; and (2) the estimated variance of $\widehat{\mathbb{IF}}^{\mathrm{split}}_{1,\mathrm{new}}(\hat{\tau}^{\mathrm{split}}_{1,\mathrm{new}})$ and that of $\widehat{\mathbb{IF}}^{\mathrm{split}}_{1,\mathrm{new}}(0)$ are very close since $\hat{\tau}^{\mathrm{split}}_{1,\mathrm{new}}$ is small. These theoretical results agree qualitatively, if not quantitatively, with the empirical results in Section 2.3.2.

3.4.2 Comparison of the Linear and Loglinear Estimators

Under the loglinear model $E_i[Y_i|X_i,A_i] = e^{\tau^*_{loglinear}A_i}b^*_i(X_i)$, $b^*_i(X_i)$ has an interpretation as the expected number of deaths on occasion i when the 1-day lagged PM_{10} level is zero.

We now provide some theoretical insight into these empirical results. In the following argument we use the subscripts "linear" and "loglinear" to distinguish between the intended versions of τ^* and $b^*_{i\tau}$. For generality we focus on the nonparametric model. For concreteness, we treat the parameter corresponding to the first-order efficient estimator.

In the nonparametric model, $\tau^*_{loglinear}$ solves

$$\sum_{i=1}^{n} E_i\left[\left\{Y_i - e^{\tau A_i}b^*_{loglinear,i\tau^*}(X_i)\right\}\left(A_i - \frac{E_i\left[e^{\tau A_i}A_i|X_i\right]}{E_i\left[e_i^{\tau A_i}|X_i\right]}\right)\right] = 0,$$

which coincides with the usual definition of τ^* if the loglinear model actually holds. τ^*_{linear} solves

$$\sum_{i=1}^{n} E_i\left[\left\{Y_i - \tau A_i - b^*_{linear,i\tau^*}(X_i)\right\}(A_i - E_i[A_i|X_i])\right] = 0,$$

which coincides with the usual definition of τ^* if the linear model holds. If we assume $\tau^*_{loglinear}A_i$ is small for all A_i, and if, as in larger cities, the variability of $b^*_i(X_i)$ with i around its mean b is small compared to b, then

$$b^*_{loglinear,i\tau^*}(X_i) = E_i[Y_i|X_i]/E_i\left[e_i^{\tau A_i}|X_i\right] \approx E_i[Y_i|X_i]\left\{1 - \tau E_i[A_i|X_i]\right\} \approx b$$

for all i (the first approximation above being valid because $\tau^*_{loglinear}A_i$ is small); and $\tau^*_{loglinear}$ approximately solves

$$\begin{aligned} 0 &= \sum_{i=1}^{n} E_i\left[\left\{Y_i - (1+\tau A_i)b\right\}(A_i - E_i[A_i|X_i])\right] = \sum_{i=1}^{n} E_i\left[\left\{Y_i - \tau A_i b - b\right\}\left(A_i - E_i\left[A_i|X_i\right]\right)\right] \\ &= \sum_{i=1}^{n} E_i\left[\left\{Y_i - \tau A_i b\right\}\left(A_i - E_i[A_i|X_i]\right)\right]. \end{aligned}$$

Now τ^*_{linear} solves

$$0 = \sum_{i=1}^{n} E_i\left[\left\{Y_i - \tau A_i - b^*_{linear,i\tau}(X_i)\right\}(A_i - E_i[A_i|X_i])\right] = \sum_{i=1}^{n} E_i\left[\left\{Y_i - \tau A_i\right\}\left(A_i - E_i[A_i|X_i]\right)\right].$$

Thus $\tau^*_{linear} \approx \tau^*_{loglinear}b$. In Section 2.3.1 we show that empirical results are consistent with this prediction.

3.4.3 Goodness-of-Fit for Choice of Density Estimator

We considered two density estimators: the nonparanormal density estimator (Liu et al., 2009) and the local regression density estimator (Loader, 1999). Both are modified by weights that take into account the effect of time (including yearly seasonality), as described in Section 2.3.4.

An empirical comparison was needed to determine which of these two estimators is more suitable. It turns out that a simple modification of $\widehat{\mathbb{F}}_{22}$ and $\widehat{\mathbb{F}}_{1,\text{new}}$ could be used to produce a goodness-of-fit diagnostic for our density estimators that would allow a direct comparison of

the Locfit and nonparanormal estimators by comparing their goodness-of-fit diagnostic. Specifically we modified $\widehat{\mathbb{IF}}_{22}^{(k)}$ and $\widehat{\mathbb{IF}}_{1,\text{new}}$ by setting $\varepsilon_i(\tau,b)$, $\Delta_{i\tau}(p)$, and $\Delta^*_{i\tau}(p,q)$ to 1, and set the cut-off γ in the definition of $s(i)$ (Equation 3.1, Section 3.4.1) to 1 (so no observation was truncated based on the value of the density estimate). That is, we defined

$$\text{IF}^{\pm}_{1,\text{new}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{s(i)}\sum_{j\in\{s(i)\}}\hat{f}_j(X_i), \text{ and}$$
$$\text{IF}^{\pm(k)}_{22} = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{s(i)}\sum_{j\in\{s(i)\}}K_k(X_i,X_j).$$

Then

$$\begin{aligned}
E\left[\text{IF}^{\pm}_{1,\text{new}}\right] &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{s(i)}\sum_{j\in\{s(i)\}}\int \hat{f}_j(x)f_i(x)\,dx,\\
E\left[\text{IF}^{\pm(k)}_{22}\right] &= -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{s(i)}\sum_{j\in\{s(i)\}}\int\prod\left[f_i(x_i)|\overline{\phi}_k(x)\right]\prod\left[f_j(x_i)|\overline{\phi}_k(x)\right]dx\\
&= -\frac{1}{n}\sum_{i}\frac{1}{s(i)}\sum_{j,j\in\{s(i)\}}\left\{\int f_j(x)f_i(x)\,dx\right.\\
&\qquad\left.-\int\prod\left[f_i(x_i)|\overline{\phi}^{\perp}_k(x)\right]\prod\left[f_j(x_i)|\overline{\phi}^{\perp}_k(x)\right]dx\right\}, \text{ and}
\end{aligned}$$

$$\begin{aligned}
E\left[\text{IF}^{\pm(k)}_{2}\right] &= E\left[\text{IF}^{\pm}_{1,\text{new}}\right]+E\left[\text{IF}^{\pm(k)}_{22}\right] = \frac{1}{n}\sum_{i}\frac{1}{s(i)}\sum_{j,j\in\{s(i)\}}\int\left\{\hat{f}_j(x)-f_j(x)\right\}f_i(x)\,dx\\
&\quad-\frac{1}{n}\sum_{i}\frac{1}{s(i)}\sum_{j,j\in\{s(i)\}}\int\prod\left[f_i(x_i)|\overline{\phi}^{\perp}_k(x)\right]\prod\left[f_j(x_i)|\overline{\phi}^{\perp}_k(x)\right]dx.
\end{aligned}$$

Thus as k becomes large, $E\left[\text{IF}^{\pm(k)}_{2}\right] = E\left[\text{IF}^{\pm}_{1,\text{new}}\right]+E\left[\text{IF}^{\pm(k)}_{22}\right]$ converges to the weighted average

$$\begin{aligned}
&\frac{1}{n}\sum_{i}\frac{1}{s(i)}\sum_{j,j\in\{s(i)\}}\int\left\{\hat{f}_j(x)-f_j(x)\right\}f_i(x)\,dx\\
=\;&\frac{1}{n}\sum_{i}\frac{1}{s(i)}\sum_{j,j\in\{s(i)\}}\int\left\{\hat{f}_j(x)f_i(x)-f_j(x)f_i(x)\right\}dx
\end{aligned}$$

of $\left\{\hat{f}_j(x)-f_j(x)\right\}f_i(x)$, and $-E\left[\text{IF}^{\pm(k)}_{22}\right] = E\left[\text{IF}^{\pm}_{1,\text{new}}-\text{IF}^{\pm(k)}_{2}\right]$ converges to the average

$$\frac{1}{n}\sum_{i}\frac{1}{s(i)}\sum_{j,j\in\{s(i)\}}\int\left\{f_j(x)f_i(x)\right\}dx$$

of $f_j(x)f_i(x)$. We therefore use as our goodness-of-fit diagnostic $\text{IF}^{\pm(k)}_{2}/\left\{\text{IF}^{\pm}_{1,\text{new}}-\text{IF}^{\pm(k)}_{2}\right\}$, where we chose the value of k that minimized $\text{IF}^{\pm(k)}_{2}$ because $E\left[\text{IF}^{\pm(k)}_{2}\right]$ will generally decrease with k if $f_j(x)$ and $f_i(x)$ are not too different. Thus after $\text{IF}^{\pm(k)}_{2}$ reaches a minimum, further increases with k presumably represent sampling variability.

APPENDIX A. FORMAL DEFINITIONS

In what follows we add τ in the subscripts of the functions b_i^* and p_i^* because these functions depend on τ. This subscript was suppressed in the main text for simplicity. In the loglinear case, define

$$\begin{aligned}
b_{i\tau}^*(X_i) &= \frac{E_i[Y_i|X_i]}{E_i\left[e^{\tau A_i}|X_i\right]},\\
p_{i\tau}^*(X_i) &= \frac{E_i[Ae^{\tau A_i}|X_i]}{E[e^{\tau A_i}|X_i]},\\
q_{i\tau}^*(X_i) &= E_i[e^{\tau A_i}|X_i],\\
\varepsilon_i(\tau, b) &= Y_i - e^{\tau A_i} b(X_i),\\
b_i(x;\eta) &= \exp\left[\eta^T w_i(x)\right],\\
q_i(x;\omega) &= \exp\left[\omega^T w_i(x)\right],\\
\Delta_i(\tau, p, q) &= \left[A_i - p(X_i)\right]\frac{e^{\tau A_i}}{q(X_i)}, \text{ and}\\
S_i(\tau, \alpha) &= (A_i - \alpha^T W_i) W_i e^{\tau A_i}.
\end{aligned}$$

In the linear case, define

$$\begin{aligned}
b_{i\tau}^*(X_i) &= E_i[Y_i - \tau A_i|X_i],\\
p_{i\tau}^*(X_i) &= E_i[A_i|X_i],\\
q_{i\tau}^*(X_i) &= 1,\\
\varepsilon_i(\tau, b) &= Y_i - \tau A_i - b(X_i),\\
b_i(x;\eta) &= \eta^T w_i(x),\\
q_i(x;\omega) &= 1,\\
S_i(\tau, \alpha) &= (A_i - \alpha^T W_i) W_i, \text{ and}\\
\Delta_i(\tau, p, q) &= A_i - p(X_i).
\end{aligned}$$

Common to both, define

$$\begin{aligned}
\Delta_i(p) &= A_i - p(X_i),\\
U_{i,\text{profile}}(\tau, b) &= \varepsilon_i(\tau, b) A_i,\\
U_{i,\text{nuis}}(\tau, b) &= \varepsilon_i(\tau, b) W_i,\\
U_i(\tau, b) &= (A_i, W_i)^T \varepsilon_i(\tau, b), \text{ and}\\
\text{IF}_{1,\text{eff},i}(\tau; b, p) &= \varepsilon_i(\tau, b)\left[A_i - p(X_i)\right].
\end{aligned}$$

Before we define the estimators using this notation, we recall that our estimators can be divided into full-sample and split-sample estimators. Full-sample estimators are estimators in which the nuisance functions as well as the estimating equation for τ are based on the full sample of N observations. The split-sample estimators divide the sample into two halves of size n each (so $N = 2n$).

A.1 THE FIRST-ORDER ESTIMATORS

We now define the full-sample estimators. Let $\hat{\eta}^{\text{full}}(\tau)$ and $\hat{\alpha}^{\text{full}}(\tau)$ solve

$$N^{-1}\sum_{i=1}^{N} U_{i,\text{nuis}}\left[\tau, b_i(\eta)\right] = 0, \text{ and}$$
$$N^{-1}\sum_{i=1}^{N} S_i(\tau,\alpha) = 0,$$

respectively, where $b_i(\eta)$ is the function $b_i(\cdot;\eta)$. The time-dependent full-data estimators of $b^*_{i\tau}$ and $p^*_{i\tau}$ are defined as

$$\hat{b}^{\text{full}}_{i\tau}(x) = \begin{cases} \exp\left[\hat{\eta}^{\text{full}}(\tau)^T w_i(x)\right] & \text{in the loglinear case,} \\ \hat{\eta}^{\text{full}}(\tau)^T w_i(x) & \text{in the linear case; and} \end{cases}$$
$$\hat{p}^{\text{full}}_{i\tau}(x) = \hat{\alpha}^{\text{full}}(\tau)^T w_i(x).$$

Then the first-order estimators $\hat{\tau}^{\text{full}}_{1,\text{eff,linear}}$ and $\hat{\tau}^{\text{full}}_{1,\text{eff,loglinear}}$ are the solutions to

$$\widehat{\mathbb{IF}}^{\text{full}}_{1,\text{eff}}(\tau) = 0$$

(for the linear and loglinear cases), where

$$\widehat{\mathbb{IF}}^{\text{full}}_{1,\text{eff}}(\tau) = \frac{1}{N}\sum_{i=1}^{N} \text{IF}_{1,\text{eff},i}(\tau; \hat{b}^{\text{full}}_{i\tau}, \hat{p}^{\text{full}}_{i\tau}).$$

Next, we define the split-sample estimators. Let $n = \frac{N}{2}$ and let $\{\text{split}(0), \text{split}(1)\}$ be a partition of $\{1,\ldots,N\}$ into two sets of size n. For $\ell = 0,1$ define $\hat{\eta}^{(\ell)}(\tau)$ and $\hat{\alpha}^{(\ell)}(\tau)$ as the solutions to

$$n^{-1}\sum_{i\in\text{split}(\ell)} U_{i,\text{nuis}}\left[\tau, b_i(\eta)\right] = 0, \text{ and}$$
$$n^{-1}\sum_{i\in\text{split}(\ell)} S_i(\tau,\alpha) = 0,$$

respectively, where $b_i(\eta)$ is the function $b_i(\cdot;\eta)$. The time-series split-data estimators of $b^*_{i\tau}$ and $p^*_{i\tau}$ are defined as

$$\hat{b}^{(\ell)}_{i\tau}(x) = \begin{cases} \exp\left[\hat{\eta}^{(\ell)}(\tau)^T w_i(x)\right] & \text{in the loglinear case,} \\ \hat{\eta}^{(\ell)}(\tau)^T w_i(x) & \text{in the linear case; and} \end{cases}$$
$$\hat{p}^{(\ell)}_{i\tau}(x) = \hat{\alpha}^{(\ell)}(\tau)^T w_i(x).$$

Define

$$\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{eff}}(\tau) = \frac{1}{n}\left\{\sum_{i\in\text{split}(0)} \text{IF}_{1,\text{eff},i}\left[\tau; \hat{b}^{(1)}_{i\tau}, \hat{p}^{(1)}_{i\tau}\right] + \sum_{i\in\text{split}(1)} \text{IF}_{1,\text{eff},i}\left[\tau; \hat{b}^{(0)}_{i\tau}, \hat{p}^{(0)}_{i\tau}\right]\right\}.$$

The first-order estimators $\hat{\tau}^{\text{split}}_{1,\text{eff,linear}}$ and $\hat{\tau}^{\text{split}}_{1,\text{eff,loglinear}}$ are the solutions to

$$\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{eff}}(\tau) = 0$$

(for the linear and loglinear cases). Further, define the density-weighted quantities

$$\begin{aligned}
\mathrm{IF}_{1,\mathrm{new},i}(\tau; b, p, f) &= \frac{1}{|s(i)|} \sum_{j \in s(i)} f_j(X_i) \varepsilon_i(\tau, b) \left[A_i - p(X_i) \right], \\
\widehat{\mathbb{IF}}^{\mathrm{full}}_{1,\mathrm{new}}(\tau) &= \frac{1}{N} \sum_{i=1}^{N} \mathrm{IF}_{1,\mathrm{new},i}(\tau; \hat{b}^{\mathrm{full}}_{i\tau}, \hat{p}^{\mathrm{full}}_{i\tau}, \hat{f}), \text{ and} \\
\widehat{\mathbb{IF}}^{\mathrm{split}}_{1,\mathrm{new}}(\tau) &= \frac{1}{2n} \Bigg\{ \sum_{i \in \mathrm{split}(0)} \mathrm{IF}_{1,\mathrm{new},i}(\tau; \hat{b}^{(1)}_{i\tau}, \hat{p}^{(1)}_{i\tau}, \hat{f}) \\
&\qquad + \sum_{i \in \mathrm{split}(1)} \mathrm{IF}_{1,\mathrm{new},i}\left[\tau; \hat{b}^{(0)}_{i\tau}, \hat{p}^{(0)}_{i\tau}, \hat{f} \right] \Bigg\}.
\end{aligned}$$

Then the first-order estimators $\hat{\tau}^{\mathrm{full}}_{1,\mathrm{new,linear}}$ and $\hat{\tau}^{\mathrm{full}}_{1,\mathrm{new,loglinear}}$ are the solutions to

$$\widehat{\mathbb{IF}}^{\mathrm{full}}_{1,\mathrm{new}}(\tau) = 0,$$

and $\hat{\tau}^{\mathrm{split}}_{1,\mathrm{new,linear}}$ and $\hat{\tau}^{\mathrm{split}}_{1,\mathrm{new,loglinear}}$ are the solutions to

$$\widehat{\mathbb{IF}}^{\mathrm{split}}_{1,\mathrm{new}}(\tau) = 0.$$

Next, we define the robust variance estimators as

$$\begin{aligned}
\widehat{\mathbb{V}}^{\mathrm{full}}_{1,\mathrm{eff}} &= \frac{\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}^{\mathrm{full}}_{1,\mathrm{eff}}(\hat{\tau}^{\mathrm{full}}_{1,\mathrm{eff}}) \right)}{\left\{ \widehat{\mathbb{DER}}\left(\widehat{\mathbb{IF}}^{\mathrm{full}}_{1,\mathrm{eff}}(\hat{\tau}^{\mathrm{full}}_{1,\mathrm{eff}}) \right) \right\}^2}, \\
\widehat{\mathbb{V}}^{\mathrm{split}}_{1,\mathrm{eff}} &= \frac{\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}^{\mathrm{split}}_{1,\mathrm{eff}}(\hat{\tau}^{\mathrm{split}}_{1,\mathrm{eff}}) \right)}{\left\{ \widehat{\mathbb{DER}}\left(\widehat{\mathbb{IF}}^{\mathrm{split}}_{1,\mathrm{eff}}(\hat{\tau}^{\mathrm{split}}_{1,\mathrm{eff}}) \right) \right\}^2}, \\
\widehat{\mathbb{V}}^{\mathrm{full}}_{1,\mathrm{new}} &= \frac{\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}^{\mathrm{full}}_{1,\mathrm{new}}(\hat{\tau}^{\mathrm{full}}_{1,\mathrm{new}}) \right)}{\left\{ \widehat{\mathbb{DER}}\left(\widehat{\mathbb{IF}}^{\mathrm{full}}_{1,\mathrm{new}}(\hat{\tau}^{\mathrm{full}}_{1,\mathrm{new}}) \right) \right\}^2}, \text{ and} \\
\widehat{\mathbb{V}}^{\mathrm{split}}_{1,\mathrm{new}} &= \frac{\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}^{\mathrm{split}}_{1,\mathrm{new}}(\hat{\tau}^{\mathrm{split}}_{1,\mathrm{new}}) \right)}{\left\{ \widehat{\mathbb{DER}}\left(\widehat{\mathbb{IF}}^{\mathrm{split}}_{1,\mathrm{new}}(\hat{\tau}^{\mathrm{split}}_{1,\mathrm{new}}) \right) \right\}^2},
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{DER}_{1,\mathrm{eff},i}(\tau, b, p) &= \begin{cases} \{\Delta_i(p)\}^2 & \text{in the linear case,} \\ \{\Delta_i(p)\}^2 e^{\tau A_i} b(X_i) & \text{in the loglinear case, and} \end{cases} \\
\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}^{\mathrm{full}}_{1,\mathrm{eff}}(\tau) \right) &= N^{-2} \sum_{i=1}^{N} \left\{ \mathrm{IF}_{1,\mathrm{eff},i}(\tau, \hat{b}^{\mathrm{full}}_{i\tau}, \hat{p}^{\mathrm{full}}_{i\tau}) \right\}^2, \\
\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}^{\mathrm{split}}_{1,\mathrm{eff}}(\tau) \right) &= \frac{1}{4n^2} \Bigg\{ \sum_{i \in \mathrm{split}(1)} \left\{ \mathrm{IF}_{1,\mathrm{eff},i}(\tau, \hat{b}^{(0)}_{i\tau}, \hat{p}^{(0)}_{i\tau}) \right\}^2 \\
&\qquad + \sum_{i \in \mathrm{split}(0)} \left\{ \mathrm{IF}_{1,\mathrm{eff},i}(\tau, \hat{b}^{(1)}_{i\tau}, \hat{p}^{(1)}_{i\tau}) \right\}^2 \Bigg\},
\end{aligned}$$

$$\widehat{\mathbb{DER}}\left(\widehat{\mathbb{F}}^{\text{full}}_{1,\text{eff}}(\tau)\right) = N^{-1}\sum_{i=1}^{N}\left\{\text{DER}_{1,\text{eff},i}(\tau,\hat{b}^{\text{full}}_{i\tau},\hat{p}^{\text{full}}_{i\tau})\right\},$$

$$\widehat{\mathbb{DER}}\left(\widehat{\mathbb{F}}^{\text{split}}_{1,\text{eff}}(\tau)\right) = \frac{1}{2n}\left\{\sum_{i\in\text{split}(1)}\text{DER}_{1,\text{eff},i}(\tau,\hat{b}^{(0)}_{i\tau},\hat{p}^{(0)}_{i\tau}) + \sum_{i\in\text{split}(0)}\text{DER}_{1,\text{eff},i}(\tau,\hat{b}^{(1)}_{i\tau},\hat{p}^{(1)}_{i\tau})\right\},$$

$$\text{DER}_{1,\text{new},i}(\tau,b,p,f) = \begin{cases} f_i(X_i)\left\{\Delta_i(p)\right\}^2 & \text{in the linear case,} \\ f_i(X_i)\left\{\Delta_i(p)\right\}^2 e^{\tau A_i}b(X_i) & \text{in the loglinear case, and} \end{cases}$$

$$\widehat{\mathbb{V}}\left(\widehat{\mathbb{F}}^{\text{full}}_{1,\text{new}}(\tau)\right) = N^{-2}\sum_{i=1}^{N}\left\{\text{IF}_{1,\text{new},i}(\tau,\hat{b}^{\text{full}}_{i\tau},\hat{p}^{\text{full}}_{i\tau},\hat{f})\right\}^2,$$

$$\widehat{\mathbb{V}}\left(\widehat{\mathbb{F}}^{\text{split}}_{1,\text{new}}(\tau)\right) = \frac{1}{4n^2}\left\{\sum_{i\in\text{split}(1)}\left\{\text{IF}_{1,\text{new},i}(\tau,\hat{b}^{(0)}_{i\tau},\hat{p}^{(0)}_{i\tau},\hat{f})\right\}^2 + \sum_{i\in\text{split}(0)}\left\{\text{IF}_{1,\text{new},i}(\tau,\hat{b}^{(1)}_{i\tau},\hat{p}^{(1)}_{i\tau},\hat{f})\right\}^2\right\},$$

$$\widehat{\mathbb{DER}}\left(\widehat{\mathbb{F}}^{\text{full}}_{1,\text{new}}(\tau)\right) = N^{-1}\sum_{i=1}^{N}\left\{\text{DER}_{1,\text{new},i}(\tau,\hat{b}^{\text{full}}_{i\tau},\hat{p}^{\text{full}}_{i\tau},\hat{f})\right\}, \text{ and}$$

$$\widehat{\mathbb{DER}}\left(\widehat{\mathbb{F}}^{\text{split}}_{1,\text{new}}(\tau)\right) = \frac{1}{2n}\left\{\sum_{i\in\text{split}(1)}\text{DER}_{1,\text{new},i}(\tau,\hat{b}^{(0)}_{i\tau},\hat{p}^{(0)}_{i\tau},\hat{f}) + \sum_{i\in\text{split}(0)}\text{DER}_{1,\text{new},i}(\tau,\hat{b}^{(1)}_{i\tau},\hat{p}^{(1)}_{i\tau},\hat{f})\right\}.$$

A.2 THE SECOND-ORDER ESTIMATORS

Define $\hat{\omega}^{(\ell)}(\tau)$ as the solution to

$$\frac{1}{n}\sum_{i\in\text{split}(\ell)}\left\{e^{\tau A_i} - e^{\omega(\tau)W_i}\right\}W_i = 0,$$

and

$$\hat{q}^{(\ell)}_{i\tau}(\text{x}) = \begin{cases} \exp\left(\hat{\omega}^{(\ell)}(\tau)^T w(\text{x})\right) & \text{in the loglinear case, and} \\ 1 & \text{in the linear case.} \end{cases}$$

Define the projection kernel K_k by

$$K_k(x,y) = \sum_{l=1}^{k}\varphi_l(x)\varphi_l(y),$$

where $\varphi_k(x)$ is an orthonormal basis for the set of square-integrable functions on the support of the random vector X. Next define

$$\text{IF}^{(k)}_{22,ij}(\tau;b,p,q) = -\varepsilon_i(\tau,b)K_k(X_i,X_j)\Delta_j(\tau,p,q),$$

$$\widehat{\mathbb{IF}}^{\text{split},(k)}_{22}(\tau) = \frac{1}{2n}\left\{\sum_{i\in\text{split}(0)}\frac{1}{|s(i)|}\sum_{j\in s(i)}\text{IF}^{(k)}_{22,ij}(\tau,\hat{b}^{(1)}_{i\tau},\hat{p}^{(1)}_{j\tau},\hat{q}^{(1)}_{j\tau})\right.$$

$$+ \sum_{i \in \text{split}(1)} \frac{1}{|s(i)|} \sum_{j \in s(i)} \text{IF}^{(k)}_{22,ij}(\tau, \hat{b}^{(0)}_{i\tau}, \hat{p}^{(0)}_{j\tau}, \hat{q}^{(0)}_{j\tau}) \Big\}, \text{ and}$$

$$\widehat{\mathbb{IF}}^{\text{split},(k)}_2(\tau) = \widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}(\tau) + \widehat{\mathbb{IF}}^{\text{split},(k)}_{22}(\tau).$$

Define the second-order estimators of τ^* as follows. Let $\hat{\tau}^{\text{split},(k)}_2$ solve

$$\widehat{\mathbb{IF}}^{\text{split},(k)}_2(\tau) = 0.$$

The variance of the above estimator is estimated by

$$\widehat{\mathbb{V}}^{\text{split},(k)}_2 = \frac{\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}(\hat{\tau}^{\text{split},(k)}_2)\right) + \widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}^{\text{split},(k)}_{22}\left(\hat{\tau}^{\text{split},(k)}_2\right)\right)}{\left\{\widehat{\mathbb{DER}}\left(\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{new}}(\hat{\tau}^{\text{split},(k)}_2)\right)\right\}^2},$$

where

$$\widehat{\mathbb{V}}\left(\widehat{\mathbb{IF}}^{\text{split},(k)}_{22}(\tau)\right) = \frac{1}{4n^2}\Bigg\{ \sum_{i \in \text{split}(0)} \frac{1}{|s(i)|^2} \sum_{j \in s(i)} \left\{\text{IF}^{(k)}_{22,ij}(\tau, \hat{b}^{(1)}_{\tau}, \hat{p}^{(1)}_{\tau}, \hat{q}^{(1)}_{\tau})\right\}^2$$

$$+ \sum_{i \in \text{split}(1)} \frac{1}{|s(i)|^2} \sum_{j \in \{s(i)\}} \left\{\text{IF}^{(k)}_{22,ij}(\tau, \hat{b}^{(0)}_{\tau}, \hat{p}^{(0)}_{\tau}, \hat{q}^{(0)}_{\tau})\right\}^2 \Bigg\}.$$

APPENDIX B. SENSITIVITY TO THE CHOICE OF X_{CONT} AND OF THE LINEAR SPLINE MODELS

This appendix provides details of the model selection procedure for various nuisance parameters. Recall that the estimating equations used in both the linear and loglinear model rely on estimates of nuisance functions $b^*_{i\tau}$, $p^*_{i\tau}$, and, in the case of loglinear regression, $q^*_{i\tau}$ as well. In the case of the linear model, $b^*_{i\tau}(x) = E_i[Y_i - \tau A_i | X_i = x]$, $p^*_{i\tau}(x) = E_i[A_i | X_i = x]$, and $q^*_{i\tau}(x) \equiv 1$. In the loglinear case, $b^*_{i\tau}(x) = E[Y_i | X_i = x] / E_i\left[e^{\tau A_i} | X_i = x\right]$, $p^*_{i\tau}(x) = E_i[A_i e^{\tau} A_i | X_i = x] / E_i[e^{\tau} A_i | X_i = x]$, and $q^*_{i\tau}(x) = E_i[e^{\tau A_i} | X_i = x]$.

Our approach involves estimating these functions based on one half of the data and testing their fit on the other half. Here we describe the model selection procedure for estimation of these functions. We considered 6 of the 7 covariates used in NMMAPS, ignoring day-of-week. We dropped the adjusted 3-day lagged variables and replaced them with four new variables, which were, for each day of interest, lagged moving averages of 1–3 and 4–6 days prior, calculated from the daily temperatures and dew-point temperatures. Of these variables, day-of-week and age category are categorical; the rest are continuous variables. These variables were organized into two main sets.

The original NMMAPS variable set included:

1. average daily temperature,
2. average daily dew-point temperature,
3. adjusted 3-day lagged daily temperature,
4. adjusted 3-day lagged dew-point temperature,
5. time,
6. day-of-week, and

7. age category.

The "New" variable set:

1. average daily temperature,
2. average daily dew-point temperature,
3. 1–3 day lagged average daily temperature,
4. 4–6 day lagged average daily temperature,
5. 1–3 day lagged average dew-point temperature,
6. 4–6 day lagged average dew-point temperature,
7. time, and
8. age category.

The two sets above are denoted by X_i in our notation. Recall that the nuisance parameters are described as functions of τ, the realization x of the random vector X_i, as well as the time point i because time is not random. This dependence on time is captured by modeling these nuisance parameters on a time-dependent transform $w_i(\cdot)$ of X_i. It was not initially assumed that $b^*_{i\tau}$ and $p^*_{i\tau}$ depended on x through the same function $w_i(\cdot)$; however, after the model selection procedure was conducted we concluded that the same function could be used. In what follows we simply use the notation $w_i(\cdot)$ with the understanding that this represents a (potentially) different function for $b^*_{i\tau}$ and $p^*_{i\tau}$.

The candidate models included different functions $w_i(\cdot)$ of X_i, as well as a choice of a link function connecting $w_i(\cdot)$ to the nuisance function of interest (i.e., $b^*_{i\tau}$ or $p^*_{i\tau}$). The goodness-of-fit was assessed based on cross-validated estimates of MSE and autocorrelation for the linear model. We made the decision to use the same function $w_i(\cdot)$ (obtained for the linear model) in the loglinear model. Once it became clear the same variable set $w_i(X_i)$ could be used for $b^*_{i\tau}$ and $p^*_{i\tau}$, we also decided to use this variable set and a log link in a loglinear model for $q^*_{i\tau}$.

Several transforms of X_i (including the spline transform used in NMMAPS as well as the gs and gs2 transforms mentioned previously) formed the variable set candidates in the model selection procedure. These initial transforms created variable sets with degrees of freedom ranging from 6 to 2004. The five methods of estimation for $b^*_{i\tau}$ are these.

1. Linear outcome regression of $Y_i - \tau A_i$ on $w_i(X_i)$. This method estimates $\eta(\tau)$ by solving the equation
$$n^{-1} \sum_{i \in \text{split}(0)} \left[Y_i - \tau A_i - \eta^T w_i(X_i) \right] w_i(X_i) = 0.$$
2. Loglinear outcome regression of $Y_i - \tau A_i$ on $w_i(X_i)$. $\eta(\tau)$ is estimated by solving the equation
$$n^{-1} \sum_{i \in \text{split}(0)} (Y_i - \tau A_i - e^{\eta^T w_i(X_i)}) w_i(X_i) = 0.$$
3. Multivariate adaptive regression splines (MARS) (Friedman, 2001) regression of $Y_i - \tau A_i$ on $w_i(X_i)$. MARS uses "hinge" functions of its inputs and products of such hinge functions in a model that it selects adaptively using generalized cross validation. Thus, the function of X_i that MARS uses as an output model is different than the input $w_i(\cdot)$.
4. Multivariate adaptive polynomial regression splines (Polymars) regression of $Y_i - \tau A_i$ on $w_i(X_i)$. This algorithm is similar to MARS with different restrictions on the types of models allowed (e.g., requiring main effects to be present if interactions are present, which MARS does not require). As with MARS, using this method expands the set of candidates for $w_i(\cdot)$.

5. Generalized penalized spline (GPS) or "bridge" regression with the generalized elastic net family of penalties (Friedman, 2008) of $Y_i - \tau A_i$ on $w_i(X_i)$. This is a sparse regression method that minimizes a regularized loss function with a penalty that bridges all subsets (L_0) and ridge regression (L_2) penalties. This includes an implicit model selection procedure due to the regularization based on L_0 and L_2 penalties, and so also includes several models beyond the supplied $w_i(X_i)$.

The methods for estimation of $p^*_{i\tau}$ are these.

1. Linear outcome regression of A on $w_i(X_i)$. This method estimates $\alpha(\tau)$ by solving the equation

$$n^{-1} \sum_{i \in \text{split}(0)} \left[A_i - \alpha^T w_i(X_i) \right] w_i(X_i) = 0.$$

2. MARS (Friedman, 2001) regression of A_i on $w_i(X_i)$.
3. Polymars regression of A_i on $w_i(X_i)$.
4. GPS (Friedman, 2008) regression of A_i on $w_i(X_i)$.

The pollution outcome used in the model for $p^*_{i\tau}$ had negative values in its range, which meant loglinear regression could not be used for $p^*_{i\tau}$. Some combinations of variable set and model selection and fitting method were excluded for computational reasons.

Thus, the models used included these.

1. Models that were linear in each of the aforementioned transformed sets $w_i(X_i)$ of variables;
2. Loglinear models in each of the transformed sets of variables, where feasible;
3. Models generated internally by MARS based on the transformed sets of variables, where feasible;
4. Models generated internally by Polymars based on the transformed sets of variables; and
5. Models generated internally by GPS based on the transformed sets of variables.

The candidate variable sets for $w_i(\cdot)$ included:

1. (NMMAPS 7) The 7 NMMAPS variables: average daily temperature, average dew-point temperature, adjusted 3-day lagged versions of the preceding variables, time, day-of-week, and age category
2. (NMMAPS 155) An expanded model with 155 variables. This included a natural spline transformation based on the continuous variables in the 7 NMMAPS variables, with 12 df for temperature, 6 df for dew-point temperature, 96 df for time, 30 df for interactions between the spline functions and the age category, as well as main effects for the categorical variables day-of-week and age category.
3. (NMMAPS 6) The NMMAPS 7 variables with day-of-week excluded.
4. (NMMAPS 149) The NMMAPS 155 variables with day-of-week excluded.
5. (NMMAPS 912) The NMMAPS 6 variables with continuous variables transformed by natural splines, with 100 df for the main effects of temperature, dew-point temperature, and time variables; and 6 df for each interaction between temperature spline and age category.
6. (NMMAPS 948) The NMMAPS 6 variables with continuous variables transformed by natural splines, with 100 df for the main effects of temperature, dew-point temperature, and time variables; and 6 df for each interaction between temperature spline and age category; plus spline functions with 6 df for interactions between the adjusted lagged temperature and the dew-point temperature.

7. (gs2 NMMAPS 6) The NMMAPS 6 variables, but with the continuous variables first centered by a linear projection onto 98-df natural spline functions of time, then orthogonalized by a Gram-Schmidt process.
8. (gs2 NMMAPS 912) These are similar to NMMAPS 912, but with natural splines applied to the centered, Gram-Schmidt orthogonalized, continuous variables.
9. (gs2 NMMAPS 948) These are similar to NMMAPS 948, but with natural splines applied to the centered, Gram-Schmidt orthogonalized, continuous variables.
10. (New 8) The 8 variables obtained by eliminating day-of-week and adjusted lagged variables from the NMMAPS variables and adding the 1–3 day and 4–6 day lagged versions of average daily temperature and average dew-point temperature.
11. (New 158) Based on the New 8 variables just above, but with the continuous variables transformed using natural splines with 18 df for temperature, 9 df for dew-point temperature, 98 df for time, and 30 df for interactions between the spline functions and the age category or the main effect for age category.
12. (New 900) Natural spline functions of the New 8 variables, but with 100 df for the main effects of temperature, dew-point temperature, and time variables; and 6 df for each interaction between the temperature spline and the age category.
13. (New 2004) Natural spline functions of the New 8 variables, but with 100 df for the main effects of temperature, dew-point temperature, and time variables; and 6 df for each interaction between the temperature spline and the age category. In addition, this included up to third-order interactions between the natural spline functions of the 1–3 day and 4–6 day lagged variables with 6 df per interaction.
14. (gs2 New 8) The New 8 variables, but with the continuous variables first centered by a linear projection onto 98 df natural spline functions of time, then orthogonalized by a Gram-Schmidt process.
15. (gs2 New 158) Similar to the New 158, but with spline transforms applied to the centered and orthogonalized continuous variables instead of the original variables.
16. (gs2 New 900) Similar to the New 900, but with spline transforms applied to the centered and orthogonalized continuous variables instead of the original variables.
17. (gs2 New 2004) Similar to the New 2004, but with spline transforms applied to the centered and orthogonalized continuous variables instead of the original variables.

The criteria for model selection were MSE and autocorrelation function (ACF) of the residuals of the fit over time. For each of two cities (Chicago and New York), an initial estimate of τ was obtained by least-squares linear regression of Y on A and the NMMAPS 149 variables. The initial estimates were $\hat{\tau} = 0.033$ for New York and $\hat{\tau} = 0.005$ for Chicago. The following procedure was done once using the initial estimate of τ obtained as described above, and again with the value $\tau = 0$ for both cities. The sample was randomly split into two equal parts 20 times; models were estimated on one half and the MSE and ACF were evaluated on the other half based on the fit in the first half. The 20 estimates of MSE and ACF thus obtained were averaged to get a final estimate for each variable set and estimation method, and an empirical standard error for the estimate of the MSE was also obtained from the 20 values.

Based on these criteria, the New variable set models involving 1–3 and 4–6 day lagged variables were not found to perform significantly better than the variable sets including the adjusted 3-day lagged variables from NMMAPS, as exemplified in Appendix Tables B.1 and B.2. For example,

the MSE for τ in the NMMAPS 7 variable set is 51.25 in the linear model, and it is 50.81 in the New 8 variable set. The loglinear model yields an MSE of 50.14 in the NMMAPS 7 set and 49.59 in the New 8 set. This indicates only a minor improvement. The comparison between the NMMAPS 155 and New 158 variable sets is similarly close. These New variable sets were eliminated and consideration was focused on the variable sets with transformed versions of the NMMAPS variables.

Results for the NMMAPS variable sets are shown in Appendix Tables B.3 through B.6. Note that certain combinations of variable set and estimation method that were eliminated early, based on cross-validated MSE, are not included in these tables. Each table includes the estimate of the MSE (averaged from 20 random splits), the empirical standard deviation of the MSE estimate (over 20 splits), and autocorrelation at lags 1, 10, 50, and 250 (each averaged over 20 splits).

In the models for $b^*_{i\tau}$ for Chicago (Appendix Table B.3), the best result (MSE = 41.73) is attained with the NMMAPS 155 variables using the generalized linear model (GLM; Poisson regression). The next best result is for the NMMAPS 149 (MSE = 41.84). In New York (Appendix Table B.4), the best result is for NMMAPS 149 using GLM (MSE = 84.80) and the next best is for NMMAPS 155 using GLM (MSE = 84.91). In both New York and Chicago, the difference is not significant based on the empirical standard error of the MSE. Since we decided not to use day-of-week, the GLM using NMMAPS 149 was chosen as the model for $b^*_{i\tau}$.

We now turn our attention to the models for $p^*_{i\tau}$. For Chicago (Appendix Table B.5), MARS (MSE = 252.0) performed best in the NMMAPS 155 variable set, but the difference in MSE between MARS and the linear model (MSE = 260.4) was not significant. MARS (MSE = 261.3) also performed best for the NMMAPS 149 variable set for Chicago, and again the difference from the linear model (MSE = 269.5) was not significant. None of the four results (from linear model or MARS with NMMAPS 155 or NMMAPS 149) was significantly superior to the others.

However, in New York (Appendix Table B.6), the best result for $p^*_{i\tau}$ was from GPS (MSE = 105.1) for the gs2 NMMAPS 948 variable set. In Chicago, however, the GPS did poorly on gs2 NMMAPS 948 (MSE = 281.2). As in Chicago, the New York results for MARS and the linear models exhibited no significant differences: NMMAPS 155 MARS MSE = 106.0 and linear model MSE = 107.4; NMMAPS 149 MARS MSE = 108.6 and linear model MSE = 110.7. Because of the lack of differentiation among models, we decided to retain the NMMAPS 149 linear model for $p^*_{i\tau}$ because of its relative simplicity and to enhance the comparability of our methods with results obtained by the NMMAPS investigators.

The final model selected for $b^*_{i\tau}$ was the GLM and for $p^*_{i\tau}$ the linear model, both based on the NMMAPS 149 variable set.

The above model selection procedure was carried out for the linear model, i.e., with response $Y_i - \tau A_i$ for $b^*_{i\tau}$ and response A_i for $p^*_{i\tau}$. We decided to use the same models for $b^*_{i\tau}$ and $p^*_{i\tau}$ in the loglinear case. We also used a loglinear model for $q^*_{i\tau}$. To summarize, in the loglinear case:

$$
\begin{aligned}
b^*_{i\tau}(X_i) &= E[Y_i|X_i] \,/ E_i\left[e^{\tau A_i}|X_i\right] = e^{\alpha^T w_i(X_i)}, \\
p^*_{i\tau}(X_i) &= E_i[A_i e^{\tau} A_i|X_i] / E_i[e^{\tau} A_i|X_i] = \eta^T w_i(X_i), \text{ and} \\
q^*_{i\tau}(X_i) &= E_i[e^{\tau A_i}|X_i] = e^{\omega^T w_i(X_i)},
\end{aligned}
$$

with $w_i(\cdot)$ for $q^*_{i\tau}$, as in the linear case. The corresponding estimators for the coefficients α, η, and ω are the solutions to the equations

$$n^{-1}\sum_{i\in\text{split}(0)}\left\{(Y_i - e^{\alpha^T w_i(X_i)}e^{\tau A_i})w_i(X_i)\right\} = 0,$$
$$n^{-1}\sum_{i\in\text{split}(0)}\left\{\left[A_i - \eta^T w_i(X_i)\right]e^{\tau A_i}w_i(X_i)\right\} = 0, \text{ and}$$
$$n^{-1}\sum_{i\in\text{split}(0)}\left\{(e^{\tau A_i} - e^{\omega^T w_i(X_i)})w_i(X_i)\right\} = 0.$$

Table B.1. Comparison of goodness-of-fit for $b^*_{i\tau}$ in the linear model for the NMMAPS and New variable sets for Chicago data. $\tau = 0$. MSE and ACF are from a single split.

Models for $E[Y-\tau A\|X]$	MSE	ACF (>1)	ACF (>10)	ACF (>50)	ACF (>250)
NMMAPS 7					
LM	51.25	0.12033	0.11217	0.11217	0.09437
GLM	50.14	0.10677	0.10251	0.10251	0.08435
MARS	50.50	0.09810	0.09606	0.09538	0.08425
Polymars	43.83	0.08322	0.08322	0.08322	0.07389
GPS	51.17	0.11900	0.11328	0.11328	0.09275
NMMAPS 155					
LM	44.43	0.08772	0.08772	0.08772	0.06960
GLM	41.65	0.08702	0.08702	0.08702	0.07392
MARS	43.05	0.08384	0.08384	0.08384	0.06544
Polymars	42.53	0.08154	0.08154	0.08154	0.06620
GPS	44.41	0.08833	0.08833	0.08833	0.06914
NEW 8					
LM	50.81	0.1300	0.1215	0.1215	0.09354
GLM	49.59	0.1159	0.1118	0.1118	0.08667
MARS	49.83	0.1163	0.1067	0.1001	0.08490
Polymars	43.66	0.0895	0.0895	0.0895	0.07437
GPS	50.76	0.1283	0.1205	0.1205	0.09375
NEW 158					
LM	44.41	0.08584	0.08584	0.08584	0.06524
GLM	41.66	0.08654	0.08654	0.08654	0.06934
MARS	42.17	0.09080	0.09080	0.09080	0.08990
Polymars	43.07	0.08315	0.08315	0.08315	0.07715
GPS	44.43	0.09378	0.09378	0.08988	0.06412

Table B.2. Comparison of goodness-of-fit for $p^*_{i\tau}$ in the linear model for the NMMAPS and New variable sets for Chicago data. $\tau = 0$. MSE and ACF are from a single split.

Models for $E[A\|X]$	MSE	ACF (>1)	ACF (>10)	ACF (>50)	ACF (>250)
NMMAPS 7					
LM	310.1	0.2928	0.1030	0.10301	0.10301
MARS	295.7	0.2698	0.1027	0.09845	0.09845
Polymars	291.0	0.2643	0.1078	0.10246	0.10196
GPS	309.5	0.2922	0.1018	0.10176	0.10176
NMMAPS 155					
LM	265.5	0.2111	0.10227	0.09599	0.09599
MARS	254.2	0.1302	0.10010	0.10010	0.10010
Polymars	263.1	0.1997	0.10504	0.10504	0.10504
GPS	272.1	0.2118	0.08924	0.08924	0.08924
NEW 8					
LM	304.6	0.2860	0.09114	0.09114	0.09114
MARS	280.8	0.2329	0.09294	0.09294	0.09140
Polymars	281.1	0.2415	0.09057	0.08977	0.08977
GPS	303.2	0.2871	0.08980	0.08980	0.08861
NEW 158					
LM	263.0	0.1858	0.10308	0.10308	0.10308
MARS	253.6	0.1246	0.10332	0.09830	0.09830
Polymars	272.7	0.2212	0.09525	0.09525	0.09501
GPS	268.1	0.1967	0.09411	0.09238	0.09238

Table B.3. Goodness-of-fit for $E[Y - \tau A|X]$ for $b^*_{i\tau}$ for Chicago data. $\tau = 0.005$. MSE and ACF are averaged over 20 random splits; empirical SD is from the 20 splits.

	MSE	EMP SD	ACF (>1)	ACF (>10)	ACF (>50)	ACF (>250)
NMMAPS 7						
LM	51.14	0.7167	0.11558	0.10828	0.10825	0.09283
GLM	49.87	0.6820	0.09970	0.09667	0.09667	0.08418
MARS	50.61	0.7453	0.10191	0.10097	0.09997	0.08719
Polymars	43.74	0.6180	0.08192	0.08192	0.08192	0.07537
NMMAPS 155						
LM	44.56	0.6700	0.08757	0.08738	0.08719	0.07602
GLM	41.73	0.6108	0.08689	0.08689	0.08669	0.07583
MARS	42.37	0.6800	0.08247	0.08247	0.08077	0.06976
Polymars	43.12	1.3147	0.08451	0.08451	0.08324	0.07115
NMMAPS 6						
LM	51.25	0.7323	0.11260	0.10558	0.10531	0.09035
GLM	49.96	0.7039	0.09786	0.09513	0.09513	0.08381
MARS	50.67	0.7446	0.10311	0.10230	0.10041	0.08796
Polymars	43.81	0.5899	0.08187	0.08187	0.08187	0.08020
NMMAPS 149						
LM	44.68	0.6842	0.08683	0.08683	0.08649	0.07765
GLM	41.84	0.6254	0.08587	0.08587	0.08561	0.07715
MARS	42.40	0.6788	0.08059	0.08059	0.07947	0.07103
Polymars	43.18	1.2470	0.08581	0.08581	0.08407	0.07152
NMMAPS 912						
LM	46.54	0.687	0.08622	0.08502	0.08215	0.06506
Polymars	49.70	0.758	0.08257	0.08203	0.07988	0.07485
NMMAPS 948						
LM	46.17	0.6722	0.08315	0.08315	0.08130	0.06721
Polymars	49.73	0.7593	0.08330	0.08276	0.08061	0.07494
gs2 NMMAPS 6						
LM	56.09	0.7140	0.08196	0.08196	0.08196	0.07792
GLM	56.54	0.6981	0.08037	0.08037	0.08037	0.07896
MARS	54.69	0.7860	0.08054	0.08054	0.08054	0.07482
Polymars	50.15	0.7613	0.08043	0.08043	0.08043	0.07513
gs2 NMMAPS 912						
LM	45.94	0.7356	0.08736	0.08690	0.08586	0.07199
Polymars	50.04	0.8738	0.08231	0.08168	0.08168	0.07937
gs2 NMMAPS 948						
LM	45.73	0.6992	0.08746	0.08670	0.08501	0.07731
Polymars	50.05	0.8751	0.08222	0.08158	0.08158	0.07919

Table B.4. Goodness-of-fit for $E[Y - \tau A|X]$ for $b^*_{i\tau}$ for New York data. $\tau = 0.033$. MSE and ACF are averaged over 20 random splits; empirical SD is from the 20 splits.

	MSE	EMP SD	ACF (>1)	ACF (>10)	ACF (>50)	ACF (>250)
NMMAPS 7						
LM	105.19	5.790	0.3373	0.3044	0.2719	0.11882
GLM	108.26	5.577	0.4047	0.3746	0.3338	0.10057
MARS	103.69	5.646	0.2922	0.2659	0.2383	0.08810
Polymars	88.01	5.290	0.1192	0.1163	0.1075	0.09538
NMMAPS 155						
LM	92.00	5.227	0.1105	0.1082	0.10820	0.09195
GLM	84.91	4.186	0.1090	0.1083	0.10831	0.09262
MARS	93.97	5.404	0.1053	0.0964	0.09296	0.07416
Polymars	98.83	6.555	0.2107	0.1677	0.13768	0.06998
NMMAPS 6						
LM	104.97	5.624	0.3339	0.3026	0.2748	0.11869
GLM	107.65	5.435	0.4020	0.3723	0.3370	0.10072
MARS	103.37	5.208	0.2925	0.2655	0.2408	0.08860
Polymars	87.68	5.014	0.1183	0.1152	0.1084	0.09277
NMMAPS 149						
LM	92.08	5.171	0.11117	0.10959	0.10959	0.09161
GLM	84.80	4.131	0.11198	0.11141	0.11141	0.09106
MARS	93.65	5.173	0.09846	0.09363	0.09184	0.06907
Polymars	98.13	6.028	0.20544	0.16525	0.13698	0.06939
NMMAPS 912						
LM	469.4	225.84	0.1531	0.1341	0.1331	0.1128
Polymars	305.4	31.13	0.5602	0.4831	0.4368	0.1084
NMMAPS 948						
LM	503.3	195.66	0.1572	0.1366	0.1356	0.1078
Polymars	305.4	31.13	0.5602	0.4831	0.4368	0.1084
gs2 NMMAPS 6						
LM	119.2	5.935	0.3658	0.3028	0.2935	0.12342
GLM	125.5	5.862	0.4286	0.3681	0.3479	0.11543
MARS	117.2	5.362	0.3106	0.2569	0.2495	0.09387
Polymars	108.0	5.595	0.1741	0.1323	0.1320	0.09788
gs2 NMMAPS 912						
LM	525.5	897.26	0.1533	0.1132	0.1095	0.08393
Polymars	256.3	15.90	0.5511	0.4588	0.4155	0.10760
gs2 NMMAPS 948						
LM	571.11	019.29	0.1555	0.1142	0.1108	0.08189
Polymars	256.3	15.90	0.5511	0.4588	0.4155	0.10760

Table B.5. Goodness-of-fit for $E[A|X]$ for $p^*_{i\tau}$ for Chicago data. MSE and ACF are averaged over 20 random splits; empirical SD is from the 20 splits.

	MSE	EMP SD	ACF (>1)	ACF (>10)	ACF (>50)	ACF (>250)
NMMAPS 7						
LM	300.9	9.719	0.2895	0.1024	0.10243	0.10243
MARS	287.0	9.459	0.2656	0.1007	0.09878	0.09410
Polymars	283.2	9.412	0.2645	0.1052	0.10251	0.09622
NMMAPS 155						
LM	260.4	7.914	0.2113	0.1012	0.09378	0.09378
MARS	252.0	6.514	0.1446	0.1001	0.09892	0.09888
Polymars	262.6	10.596	0.2143	0.1001	0.09905	0.09525
NMMAPS 6						
LM	309.9	9.445	0.2979	0.0907	0.09029	0.09029
MARS	295.4	9.320	0.2736	0.1012	0.09934	0.08070
Polymars	288.8	10.673	0.2694	0.1064	0.10114	0.08590
NMMAPS 149						
LM	269.5	7.659	0.2197	0.10597	0.08380	0.07983
MARS	261.3	6.128	0.1567	0.09490	0.09377	0.09196
Polymars	265.9	10.203	0.2022	0.09646	0.09307	0.08803
NMMAPS 912						
LM	303.5	7.393	0.2147	0.1026	0.08906	0.08449
Polymars	292.2	42.846	0.2960	0.0920	0.09074	0.08135
NMMAPS 948						
LM	300.6	6.62	0.2236	0.10205	0.08587	0.07925
Polymars	285.4	26.16	0.2833	0.09502	0.09379	0.08390
gs2 NMMAPS 6						
LM	322.6	9.663	0.3468	0.1435	0.1435	0.1187
MARS	314.6	8.252	0.3325	0.1603	0.1603	0.1256
Polymars	313.5	8.909	0.3284	0.1568	0.1568	0.1198
gs2 NMMAPS 912						
LM	323.6	8.10	0.2291	0.10333	0.08873	0.08221
Polymars	293.4	16.32	0.2112	0.09726	0.09542	0.08550
gs2 NMMAPS 948						
LM	322.2	7.486	0.2269	0.1018	0.08868	0.08180
Polymars	291.3	16.758	0.2150	0.0960	0.09320	0.08488
GPS	281.2	7.520	0.2279	0.1082	0.08701	0.08232

Table B.6. Goodness-of-fit for $E[A|X]$ for p_{it}^* for New York data. MSE and ACF are averaged over 20 random splits; empirical SD is from the 20 splits.

	MSE	EMP SD	ACF (>1)	ACF (>10)	ACF (>50)	ACF (>250)
NMMAPS 7						
LM	128.7	10.84	0.13642	0.13642	0.10938	0.06493
MARS	114.9	10.45	0.09868	0.09868	0.08961	0.06115
Polymars	112.1	10.59	0.09729	0.09723	0.07989	0.06153
NMMAPS 155						
LM	107.4	9.938	0.10389	0.10192	0.09736	0.06528
MARS	106.0	8.216	0.09881	0.09754	0.09698	0.06594
Polymars	123.4	31.547	0.09554	0.09554	0.09416	0.06854
NMMAPS 6						
LM	131.2	10.52	0.1514	0.1514	0.12905	0.06535
MARS	118.4	10.84	0.1199	0.1199	0.09414	0.06140
Polymars	114.5	10.69	0.1107	0.1107	0.09337	0.05764
NMMAPS 149						
LM	110.7	9.426	0.1174	0.1145	0.09963	0.06519
MARS	108.6	8.346	0.1027	0.1027	0.09999	0.06756
Polymars	121.2	14.265	0.1067	0.1067	0.10060	0.06662
NMMAPS 912						
LM	1054.2	671.0	0.1901	0.1507	0.1486	0.12952
Polymars	161.7	116.0	0.1071	0.1066	0.1056	0.07964
NMMAPS 948						
LM	1016.7	605.53	0.1840	0.1502	0.14929	0.1266
Polymars	134.8	52.35	0.1070	0.1053	0.09971	0.0758
gs2 NMMAPS 6						
LM	139.2	11.11	0.2318	0.2030	0.2030	0.07846
MARS	137.6	10.36	0.2137	0.1872	0.1872	0.07466
Polymars	137.7	10.65	0.2140	0.1913	0.1913	0.07584
gs2 NMMAPS 912						
LM	964.4	1749.48	0.1356	0.1163	0.1112	0.08204
Polymars	125.8	23.45	0.1023	0.1018	0.1007	0.07070
gs2 NMMAPS 948						
LM	815.4	1260.85	0.1351	0.1125	0.10837	0.08332
Polymars	126.1	23.73	0.1014	0.1006	0.09855	0.06844
GPS	105.1	8.67	0.1049	0.1049	0.10249	0.06637

APPENDIX C. EFFICIENCY OF SAMPLE SPLITTING

To better understand how sample splitting affects the variance of the first-order estimators, we conducted a simulation study with a data generating process designed to yield data with some of the characteristics of actual NMMAPS data from New York (in which only about 2000 cases have complete data). These simulations have about 15,000 data points.

In NMMAPS, the continuous covariates are the average daily temperature X_1, average dew-point temperature X_2, adjusted lagged versions of these variables (X_3 and X_4), and time. Time was not simulated. X_1 and X_2 exhibit some periodicity in time and are not uncorrelated. X_3 and X_4 do not exhibit such periodicity. In order to capture some of this structure in the simulations, the following modified simulation method was adopted for these variables. First, a low-pass convolution filter (uniform, length 25) was applied to each of these variables to get a "signal" and residuals:

$$X_{1si} = \frac{1}{25}\sum_{j=i-12}^{i+12} X_{1j},$$
$$X_{1ri} = X_{1i} - X_{1si},\ i = 1,\ldots,N,$$
$$X_{2si} = \frac{1}{25}\sum_{j=i-12}^{i+12} X_{2j}, \text{ and}$$
$$X_{2ri} = X_{2i} - X_{2si},\ i = 1,\ldots,N.$$

The residuals X_{1ri}, and X_{2ri} from this filter were used as inputs to the following copula-type method in place of X_{1i}, and X_{2i}, respectively.

A single simulated data set was obtained as follows. A copula-type, or semiparametric bootstrap, method was used to approximate the observed distribution of covariates. In particular, suppose X_{cont} is the $n \times p$ matrix of observed input covariates (either the continuous covariates or their residuals), and let Σ be the empirical covariance matrix from these covariates. Let $\Sigma^{1/2}$ be a symmetric matrix (the Cholesky factorization matrix; Lawson and Hanson, 1974) satisfying $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$. Write $\Sigma^{-1/2}$ for a generalized inverse of $\Sigma^{1/2}$. Let $\tilde{X} = \Sigma^{-1/2}X_{\text{cont}}$, and suppose $\tilde{\tilde{X}}$ is an $n \times p$ matrix in which the j-th column is a simple random sample from the j-th column of $\tilde{X}$. The simulated continuous covariates equal $X^{\dagger} = \Sigma^{1/2}\tilde{\tilde{X}}$. This method has the following properties: (1) the marginal distribution of the j-th column of $X^{\dagger}$ equals the empirical distribution of the j-th column of X_{cont}, and (2) the covariance of $X^{\dagger}$ approximately equals Σ, the empirical covariance of X_{cont}. The simulated covariates, X_{sim}, are created by adding the discrete covariates to $X^{\dagger}$.

After simulation using the copula-type method, the filtered process was added back to the corresponding variables. The discrete covariates added were day-of-week and age category (with no resampling). Thus the simulated covariates were:

$$X_{\text{sim},1i} = X^{\dagger}_{1i} + X_{1si},$$
$$X_{\text{sim},2i} = X^{\dagger}_{2i} + X_{2si},$$
$$X_{\text{sim},3i} = X^{\dagger}_{3i},$$
$$X_{\text{sim},4i} = X^{\dagger}_{4i},$$
$$X_{\text{sim},5i} = DOW_i, \text{ and}$$

$$X_{\text{sim},6i} = AgeCat_i.$$

Next, $A|X$ was simulated from a Poisson distribution with mean $\exp(\hat{\zeta}^T \bar{X}_{\text{sim},5})$, where $\bar{X}_{\text{sim},5} = (1, X_{\text{sim},1i}, X_{\text{sim},2i}, X_{\text{sim},3i}, X_{\text{sim},4i}, X_{\text{sim},5i})$, and $\hat{\zeta}$ was a preliminary estimate from the original NMMAPS data based on Poisson regression for the loglinear model $E[A|X] = \exp(\zeta^T \bar{X}_5)$. Call the simulated pollution variable A_{sim}.

$Y|A,X$ was drawn from a Normal $(\hat{\mu}, \hat{\sigma}^2)$ distribution, where $\hat{\mu} = \exp\left[\hat{\xi} A_{\text{sim}} + \hat{\nu} w(X_{\text{sim}})\right]$ and $\hat{\sigma}^2 = \exp\left[\hat{\lambda}^T (1, X_1, X_6)\right]$. $\hat{\xi}$ and $\hat{\nu}$ were estimated using a Normal family GLM with log link with covariates A and $w(X)$. $\hat{\lambda}$ was estimated using the squared residuals from the Normal fit, $\left(Y - \exp\left[\hat{\xi} A + \hat{\nu} w(X)\right]\right)^2$, as the response in a Poisson regression with log link. Here, as noted above, $w(X)$ refers to the same natural spline functions of time, temperature, and dew-point temperature with the same degrees of freedom as used in NMMAPS. [Although we drop the subscript i from w_i in the interest of brevity, it is important to note that w represents a different function at each data point i because i indexes time and $w()$ is a function of time.]

We note that the $\hat{\xi}$ estimated as above is the "true" τ for the simulations. This corresponds to the smooth case where there is no misspecification for the mean of $Y|A,X$. The estimate we obtained for $\hat{\xi}$ (and hence the true τ) was 0.0544873 for New York.

The simulated data were fit using the same covariates that were used in the NMMAPS analysis. We recall here that these included natural spline functions with 12 df for temperature, 6 df for dew-point temperature, 98 df for time, categorical variables day-of-week and age category, as well as 30 df for interactions between the spline functions and the age category. In all, we used 155 df in the fit.

At the time of this analysis, the optimal (linear) model for $p^*_{i\tau}$ had not yet been determined. Hence this analysis was conducted using a loglinear model for suitably scaled and shifted PM_{10} values to ensure positivity. We expect that the conclusions would not qualitatively change by much were the analysis redone with the final linear model chosen for $p^*_{i\tau}$.

The data were fit using five different methods all based on the loglinear model.

Poisson Regression and Outcome Regression This method was a standard Poisson regression where both b and τ were estimated from the full dataset. Let w_i be the function that transforms the 6 NMMAPS variables X_i as described earlier in the report. Note that, since the models we used were not finalized at the time these simulations were run, the covariates included day of week for these runs for a total of 155 covariates. Thus $w_i(\cdot)$, though conceptually similar to the $w_i(\cdot)$ in the rest of the report (which produces 149 covariates), represents a slightly different function in this appendix. Recall that

$$\begin{aligned}
W_i &= w_i(X_i), \\
\varepsilon_i(\tau, b) &= Y_i - e^{\tau A_i} b(X_i), \\
b_i(\eta)(\cdot) &= e^{\gamma^T w_i(\cdot)}, \\
\hat{\mathbb{U}}^{\text{full}}\left[\tau, b(\eta)\right] &= N^{-1} \sum_{i=1}^{N} \varepsilon_i\left[\tau, b_i(\eta)\right] (A_i W_i)^T, \text{ and} \\
(\hat{\tau}, \hat{\gamma}) &\text{ solves } \hat{\mathbb{U}}^{\text{full}}\left[\tau, b(\eta)\right] = 0.
\end{aligned}$$

This procedure was implemented via an equivalent profile approach, in which

$$\begin{aligned}
\widehat{\mathbb{U}}^{\text{full}}_{\text{nuis}}\left[\tau, b(\eta)\right] &= N^{-1}\sum_{i=1}^{N}\varepsilon_i\left[\tau, b_i(\eta)\right]W_i = 0,\\
\hat{\eta}(\tau) \text{ solves } \widehat{\mathbb{U}}^{\text{full}}_{\text{nuis}}&\left[\tau, b(\eta)\right] = 0,\\
\hat{b}^{\text{full}}_{i\tau}(X_i) &= e^{\hat{\eta}(\tau)^T W_i},\\
\widehat{\mathbb{U}}^{\text{full}}_{\text{profile}}(\tau) &= N^{-1}\sum_{i=1}^{N}\varepsilon_i\left(\tau, \hat{b}^{\text{full}}_{i\tau}\right)A_i, \text{ and}\\
\hat{\tau} \text{ solves } \widehat{\mathbb{U}}^{\text{full}}&(\tau) = 0.
\end{aligned}$$

First-Order Efficient Influence Function on the Full Dataset This method used the empirical efficient influence function as the estimating equation. Estimation of both nuisance parameters b, and p as well as τ was based on the entire dataset. The nuisance parameters were estimated based on the models

$$\begin{aligned}
b^{*}_{i\tau}(X_i) &= \frac{E_i[Y_i|X_i]}{E_i[e^{\tau A_i}|X_i]}, \text{ and}\\
p^{*}_{i\tau}(X_i) &= \frac{E_i[e^{\tau A_i}A_i|X_i]}{E_i[e^{\tau A_i}|X_i]}.
\end{aligned}$$

Let

$$\begin{aligned}
\hat{\eta}(\tau) \text{ solve } N^{-1}\sum_{i=1}^{N}\varepsilon_i\left(\tau, b_i(\hat{\eta})\right)W_i = 0, \text{ and}\\
\hat{\alpha}(\tau) \text{ solve } N^{-1}\sum_{i=1}^{N}\left(e^{\hat{\alpha}^T W_i} - A_i\right)e^{\tau A_i}W_i = 0.
\end{aligned}$$

Define

$$\begin{aligned}
\hat{b}_{i\tau}(X_i) &= b_i\left[X_i; \hat{\eta}(\tau)\right], \text{ and}\\
\hat{p}_{i\tau}(X_i) &= e^{\hat{\alpha}(\tau)^T w_i(X_i)}.
\end{aligned}$$

Note that $\hat{p}_{i\tau}(\cdot)$ as defined here differs from previous definitions of τ in that a loglinear representation is used rather than a linear representation, as explained above. τ is estimated as the solution to

$$\widehat{\mathbb{IF}}^{\text{full}}_{1,\text{eff}}(\tau) = N^{-1}\sum_{i=1}^{N}\text{IF}_{1,\text{eff},i}(\tau, \hat{b}_{i\tau}, \hat{p}_{i\tau}) = N^{-1}\sum_{i=1}^{N}\varepsilon_i(\tau, \hat{b}_{i\tau})\Delta_i(\hat{p}_{i\tau}) = 0.$$

Note that, since the model for p is not linear in these simulations, the solution to $\widehat{\mathbb{IF}}^{\text{full}}_{1,\text{eff}}(\tau) = 0$ is not algebraically the same as the solution to $\widehat{\mathbb{U}}^{\text{full}}_{\text{profile}}(\tau) = 0$.

First-Order Efficient Influence Function on 2 Splits Swapped This method uses the empirical efficient influence function. However, the nuisance parameters are estimated from one half of the data and the estimating equation for τ is based on the other half; then the halves are reversed; finally, the estimating equations so obtained are combined.

Following the notation in the rest of the report, we let $n = N/2$, and use superscripts (0) and (1) to denote the two halves of the data. Let split(0) and split(1) denote the indices (disjoint subsets of $\{1,\ldots,N\}$) for the two halves of the data in the split. For $l = 0, 1$, let

$$\hat{\eta}^{(l)}(\tau) \text{ solve } n^{-1} \sum_{i \in \text{split}(l)} \varepsilon_i \left[\tau, b_i\left\{\hat{\eta}^{(l)}(\tau)\right\}\right] W_i = 0, \text{ and}$$

$$\hat{\alpha}^{(l)}(\tau) \text{ solve } n^{-1} \sum_{i \in \text{split}(l)} \left(e^{\hat{\alpha}^{(l)}(\tau)^T W_i} - A_i\right) e^{\tau A_i} W_i = 0.$$

Define

$$\hat{b}^{(l)}_{i\tau}(X_i) = e^{\hat{\eta}^{(l)}(\tau) w_i(X_i)},$$
$$\hat{p}^{(l)}_{i\tau}(X_i) = e^{\hat{\alpha}^{(l)}(\tau) w_i(X_i)}, \text{ and}$$

$$\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{eff}}(\tau) = n^{-1}\left\{\sum_{i \in \text{split}(1)} \text{IF}_{1,\text{eff},i}\left(\tau, \hat{b}^{(0)}_{i\tau}, \hat{p}^{(0)}_{i\tau}\right) + \sum_{i \in \text{split}(0)} \text{IF}_{1,\text{eff},i}\left(\tau, \hat{b}^{(1)}_{i\tau}, \hat{p}^{(1)}_{i\tau}\right)\right\}.$$

Finally, τ is estimated as the solution to the estimating equation

$$\widehat{\mathbb{IF}}^{\text{split}}_{1,\text{eff}}(\tau) = 0.$$

First-Order Efficient Influence Function on 10 Random Splits of the Data This method is similar to the above method based on 2 splits, swapped. Instead of just 2 splits, we generalized to Q splits (we used $Q = 10$) and dropped the swapping procedure. Thus, we randomly split the data into halves Q times.

For each split, the following procedure was adopted: One half was designated the training sample and the other as the testing sample. The nuisance functions b and p were estimated based on the data for each split. Next, the estimating equation was constructed using the empirical efficient influence function evaluated on the data in the testing sample (but with the functions b and p as estimated from the training sample).

Finally, the Q estimating equations thus obtained (one from each of the Q splits) were added together and solved to obtain the final estimate of τ.

First-Order Efficient Influence Function on 1 Split In order to understand the information loss when we do not swap the training and testing samples, we also estimated τ using sample splitting without swapping. Let $\hat{b}^{(l)}_{i\tau}$, and $\hat{p}^{(l)}_{i\tau}$ be as for the splitting + swapping method above. Here τ is estimated as the solution to

$$n^{-1} \sum_{i \in \text{split}(1)} \text{IF}_{1,\text{eff},i}\left(\tau, \hat{b}^{(0)}_{i\tau}, \hat{p}^{(0)}_{i\tau}\right) = 0.$$

The empirical squared biases and variances for the various methods based on 400 replicates are reported in Appendix Table C.1.

In order to compare the effect of smoothness on the estimation procedures, a Hölder 0.6 function of X was added to the linear predictors for $Y|A,X$ and $A|X$. The procedure for simulating X was the same as above. The simulated $A|X$ was Poisson with mean

$\exp\left[\hat{\zeta}^T \bar{X}_{\text{sim},5} + h_1(X_{\text{sim},1}, X_{\text{sim},2})\right]$, with $\hat{\zeta}$ as before. Note that h_1 is a function of average daily temperature and average dew-point temperature only. The function $h_1(X_{\text{sim},1}, X_{\text{sim},2})$ was defined as $h_1(X_{\text{sim},1}, X_{\text{sim},2}) = \varpi_{01} + \varpi_{11}\left[h(X_{\text{sim},1}) + h(X_{\text{sim},2})\right]$, where $h(x) : \mathbb{R} \to \mathbb{R}$ was a Hölder 0.6 function, and ϖ_{01} and ϖ_{11} are preselected so that $h_1(X_{\text{sim},1}, X_{\text{sim},2})$ is in the range $(-1.5, -0.6)$. These numbers were selected so that $h_1(X_{\text{sim},1}, X_{\text{sim},2})$; and has roughly the same range as $\hat{\zeta}^T \bar{X}_{\text{sim},5}$; and ϖ_{01} and ϖ_{11} do not depend on the simulated values X_{sim}. The simulated $Y|A,X$ was Normal$(\hat{\mu}, \hat{\sigma}^2)$ distribution, where $\hat{\mu} = \exp\left[\hat{\xi} A_{\text{sim}} + \hat{\nu} w(X_{\text{sim}}) + h_2(X_{\text{sim},1}, X_{\text{sim},2})\right]$. $\hat{\xi}, \hat{\nu}$ and $\hat{\sigma}^2$ are as before. $h_2(X_{\text{sim},1}, X_{\text{sim},2})$ was defined as $h_2(X_{\text{sim},1}, X_{\text{sim},2}) = \varpi_{02} + \varpi_{12}\left[h(X_{\text{sim},1}) + h\left(X_{\text{sim},2}\right)\right]$, where $h(X)$ was the same Hölder 0.6 function used for $A|X$ and $\varpi_{02} + \varpi_{12}$ are preselected so $h_2(X_{\text{sim},1}, X_{\text{sim},2})$ is in the range $(3.2, 5)$. These numbers were selected so that $h_2(X_{\text{sim},1}, X_{\text{sim},2})$ has a range comparable to that of $\hat{\xi} A_{\text{sim}} + \hat{\nu} w(X_{\text{sim}})$.

The results from 1000 replicates are summarized in Appendix Table C.2.

Next, the smoothness for each of $E[A|X]$ and $E[Y|A,X]$ was decreased to to Hölder 0.1. The method of simulation was the same as above with the exception that $h(x)$ was now Hölder 0.1. Empirical squared bias and variance estimates from 1000 replicates are given in Appendix Table C.3.

We expect from theory that the bias as well as variance increase as the smoothness of the functions $E[A|X]$ and $E[Y|A,X]$ decrease, and this is indeed observed. The method using only 1 split essentially wastes half the data in estimating the nuisance functions. That method has high bias and variance, nearly double that of the other methods. However, we do not lose much when we use the "2 splits, swapped" method, compared with using the whole dataset with the empirical efficient influence function, as indicated by the variances in the tables. By a very small margin, the "10 random splits" method seems best among those that do split the data. Although the methods based on full data are marginally better, all of the methods (except the "1 split" method) are roughly comparable in terms of both bias and variance.

Table C.1. Empirical squared bias and variance for various estimators from 400 simulation replicates. $b^*_{i\tau}$ and $p^*_{i\tau}$ were smooth functions in these simulations.

Estimator	Empirical Squared Bias	Variable
Full sample, outcome, loglinear	2.744e-06	2.749e-06
Full sample, IF1, loglinear	2.732e-06	2.739e-06
2 Splits swapped, IF1, loglinear	3.002e-06	3.009e-06
10 Random splits, IF1, loglinear	2.962e-06	2.969e-06
1 Split, IF1, loglinear	5.527e-06	5.536e-06

Table C.2. Empirical squared bias and variance for various estimators from 400 simulation replicates. $b^*_{i\tau}$ and $p^*_{i\tau}$ were in a Hölder class with exponent $\beta_b = \beta_p = 0.6$.

Estimator	Empirical Squared Bias	Variable
Full sample, outcome, loglinear	5.282e-05	1.971e-05
Full sample, IF1, loglinear	5.242e-05	1.970e-05
2 Splits swapped, IF1, loglinear	5.308e-05	2.050e-05
10 Random splits, IF1, loglinear	5.314e-05	2.023e-05
1 Split, IF1, loglinear	6.045e-05	2.735e-05

Table C.3. Empirical squared bias and variance for various estimators from 400 simulation replicates. $b^*_{i\tau}$ and $p^*_{i\tau}$ were in a Hölder class with exponent $\beta_b = \beta_p = 0.1$.

Estimator	Empirical Squared Bias	Variable
Full sample, outcome, loglinear	0.0006147	7.892e-05
Full sample, IF1, loglinear	0.0006147	7.892e-05
2 Splits swapped, IF1, loglinear	0.0006169	8.227e-05
10 Random splits, IF1, loglinear	0.0006160	8.242e-05
1 Split, IF1, loglinear	0.0006480	1.076e-04

APPENDIX D. A $\sqrt{n}$-CONSISTENT ASYMPTOTICALLY UNBIASED ESTIMATOR OF τ^* IN THE LOGLINEAR MODEL WHEN $\beta_b + \beta_p > D/2$

We assume the loglinear regression model

$$\log E[Y|A,X] = \tau A + \zeta^*(X).$$

Let

$$\begin{aligned} g(x) &= e^{\zeta^*(x)}, \\ \epsilon &= Ye^{-\tau A} - g(X), \\ h(X) &= E[A|X], \\ U &= A - E[A|X], \text{ and} \\ p_K(x) &= [p_{1K}(x) \cdots p_{KK}(x)]^T. \end{aligned}$$

$p_K(x)$ is a $K \times 1$ vector of K optimal basis functions evaluated at x. With a slight abuse of notation, let p_K also denote the matrix

$$p_K = \begin{bmatrix} p_K(X_1)^T \\ p_K(X_2)^T \\ \vdots \\ p_K(X_N)^T \end{bmatrix}_{N \times K},$$

and let

$$Q = p_K[p_K^T p_K]^{-1} p_K \text{ and } M = I_{N \times N} - Q$$

be the empirical projection matrices. Consider the estimation method:

$$\hat{\eta}(\tau) = \underset{\eta}{\operatorname{argmin}} \mathbb{P}_n[\{Y\exp(-\tau A) - p_K(X)^T\eta\}^2] \text{ or}$$
$$\hat{\eta}(\tau) \text{ solves } 0 = \mathbb{P}_n[\{Y\exp(-\tau A) - p_K(X)^T\eta\}p_K(X)],$$

and

$$\hat{\tau} = \underset{\tau}{\operatorname{argmin}} \mathbb{P}_n[\{Y\exp(-\tau A) - p_K(X)^T\hat{\eta}(\tau)\}^2] \text{ or}$$
$$\hat{\tau} \text{ solves } 0 = \mathbb{P}_n[\{Y\exp(-\tau A) - p_K(X)^T\hat{\eta}(\tau)\}A]$$

We show that this estimator is $\sqrt{n}$-consistent using a proof similar to that in Donald and Newey (1994).

In what follows, we use an underbar to denote empirical vectors. For example, $\underline{A} = [A_1, \ldots, A_N]^T$ and $\underline{Y\exp(-\tau A)} = [Y_1\exp(-\tau A_1), \ldots, Y_N\exp(-\tau A_N)]^T$. We make the following assumptions.

Assumption 42. *Assume that, in addition to the loglinear model,*

1. $E[\underline{U}^T D(\tau^*)\underline{U}/N] \to \bar{A} > 0$,
2. $K \to \infty$ as $N \to \infty$,
3. $K/N \to 0$ as $N \to \infty$,
4. $\mathrm{Var}[A_i|X_i] < \Delta < \infty$,
5. $E[h^2(X)Y^2\exp(-2\tau^* A)] < \infty$,
6. $E[U^2Y^2\exp(-\tau^* A)^2|X] < \Delta < \infty$, and
7. h can be approximated at rate $e_h(K)$ and g can be approximated at rate $e_g(K)$.

Then

$$\hat{\tau} \text{ solves } 0 = \underline{A}^T M\left\{\underline{Y\exp(-\tau A)}\right\}, \text{ so}$$
$$\hat{\tau} - \tau = -\left(\frac{d}{d\tau}\underline{A}^T M\left\{\underline{Y\exp(-\bar{\tau}A)}\right\}\right)^{-1}\left(\underline{A}^T M\underline{Y\exp(-\tau A)}\right)$$
$$= -\left(\frac{d}{d\tau}\underline{A}^T M\left\{\underline{Y\exp(-\bar{\tau}A)}\right\}\right)^{-1}\left(\underline{A}^T M\{\underline{g} + \underline{\epsilon}\}\right),$$

where $\bar{\tau}$ is a value of τ intermediate between $\hat{\tau}$ and τ^*. Note that

$$\frac{d}{d\tau}\left(\underline{A}^T_{1\times N} M_{N\times N}\left\{\underline{Y\exp(-\bar{\tau}A)}\right\}\right)_{q\times 1} = \left(\underline{A}^T M\left\{\underline{Y\exp(-\bar{\tau}A)A}\right\}_{N\times 1}\right).$$

Now, say that

$$\underline{Y\exp(-\bar{\tau}A)A}$$
$$= \begin{bmatrix} Y_1\exp(-\bar{\tau}A_1)A_1 \\ \vdots \\ Y_N\exp(-\bar{\tau}A_N)A_N \end{bmatrix} = \begin{bmatrix} Y_1\exp(-\bar{\tau}A_1) & 0 & \cdots & 0 \\ 0 & Y_2\exp(-\bar{\tau}A_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Y_N\exp(-\bar{\tau}A_N) \end{bmatrix}\begin{bmatrix} A_1 \\ \vdots \\ A_N \end{bmatrix}$$
$$= D(\bar{\tau})\underline{A}.$$

So we have

$$\hat{\tau} - \tau = -\left(\underline{A}^T M D(\bar{\tau}) \underline{A}/n\right)^{-1} \left(\underline{A}^T M\{\underline{g} + \underline{\epsilon}\}/N\right).$$

The proof proceeds in two parts. First, we show that $\underline{A}^T M D(\bar{\tau}) \underline{A}/N \xrightarrow{P} \bar{A}$. The second part, showing that $\left(\underline{A}^T M\{\underline{g} + \underline{\epsilon}\}/n\right) = O_P\left[e_h(K) e_g(K)\right]$, is identical to the proof in Donald and Newey (1994); we present only the order of each term and refer the reader to that paper for details.

- Write

$$\begin{aligned}\underline{A}^T M D(\bar{\tau}) \underline{A}/N = {} & \underline{U}^T D(\bar{\tau}) \underline{U}/N - \underline{U}^T Q D(\bar{\tau}) \underline{U}/N + \underline{U}^T M D(\bar{\tau}) \underline{h}/N \\ & + \underline{h}^T M D(\bar{\tau}) \underline{U}/N + \underline{h}^T M D(\bar{\tau}) \underline{h}/N.\end{aligned}$$

We assume that $E[\underline{U}^T D(\tau^*) \underline{U}/N] \to \bar{A} > 0$. We want to show that $\underline{U}^T D(\bar{\tau}) \underline{U}/N \xrightarrow{P} \bar{A}$ and all the other terms in the above decomposition are $o_P(1)$. In general,

$$\underline{v}_1^T M \underline{v}_2/N \le (\underline{v}_1^T M \underline{v}_1/N)^{1/2} (\underline{v}_2^T M \underline{v}_2/N)^{1/2}.$$

For the first term, note that

$$\begin{aligned}& \underline{U}^T [D(\bar{\tau}) - D(\tau^*)] \underline{U}/N \\ & = N^{-1} \sum_{i=1}^N U_i^2 Y_i \{\exp(-\bar{\tau} A_i) - \exp(-\tau A_i)\} \\ & = -N^{-1} (\bar{\tau} - \tau^*)^T \sum_{i=1}^n U_i^2 Y_i A_i \exp(-\bar{\bar{\tau}} A_i) \\ & = o_P(1) \times N^{-1} \sum_{i=1}^N U_i^2 Y_i A_i \exp(-\bar{\bar{\tau}} A_i).\end{aligned}$$

Since $\bar{\bar{\tau}} \xrightarrow{P} \tau$, $\bar{\bar{\tau}} \in \tau^* \pm \delta$ with high probability for some small δ (where $\tau^* \pm \zeta^* = [\tau^* - \delta, \tau^* + \delta]$). Hence, $\exp(-\bar{\bar{\tau}} A_i) \in \left[\min \exp(-\{\tau \pm \delta\} A_i), \max \exp(-\{\tau^* \pm \delta\} A_i)\right]$. So

$$\begin{aligned}& N^{-1} \sum_{i=1}^N U_i^2 Y_i A_i \exp(-\bar{\bar{\tau}} A_i) \\ & \in \left(\min_{2^q} N^{-1} \sum_{i=1}^N U_i^2 Y_i A_i \exp(-\{\tau^* \pm \delta\} A_i), \max_{2^q} N^{-1} \sum_{i=1}^N U_i^2 Y_i A_i \exp(-\{\tau^* \pm \delta\} A_i) \right) \\ & \xrightarrow{P} \left(\min E\left[U_i^2 Y_i A_i \exp(-\{\tau^* \pm \delta\} A_i)\right], \max E\left[U_i^2 Y_i A_i \exp(-\{\tau^* \pm \delta\} A_i)\right] \right) \\ & = O_P(1),\end{aligned}$$

so

$$\underline{U}^T [D(\bar{\tau}) - D(\tau^*)] \underline{U}/N = o_P(1) \times O_P(1) = o_P(1).$$

Hence $\underline{U}^T D(\bar{\tau}) \underline{U}/N \xrightarrow{P} \bar{A}$. We will use the above argument repeatedly.

For the second term, note that $\underline{U}^T Q D(\bar{\tau}) \underline{U}/N - \underline{U}^T Q D(\tau) \underline{U}/N = o_P(1)$. Now we use the following bound:

$$\underline{U}^T Q D(\bar{\beta}) \underline{U}/N \le (\underline{U}^T Q \underline{U}/N)^{1/2} \ (\underline{U}^T D(\bar{\beta}) Q D(\bar{\beta}) \underline{U}/N)^{1/2}.$$

Next,

$$\underline{U}^T Q\underline{U}/N = O_P(E[\underline{U}^T Q\underline{U}]/N) = O_P(E[\mathrm{tr}\,\underline{U}^T Q\underline{U}]/N) = O_P(E[\mathrm{tr}\,Q\underline{U}\underline{U}^T Q]/N)$$
$$= O_P(\mathrm{tr}\,E[Q\underline{U}\underline{U}^T Q]/N) = O_P(\mathrm{tr}\,E[QE[\underline{U}\underline{U}^T|\underline{X}]Q]/N).$$

Also,

$$\mathrm{tr}\,E[QE[\underline{U}\underline{U}^T|\underline{X}]Q]/N \le \Delta\,\mathrm{tr}\,E[QQ]/N = \Delta K/N = O(KN^{-1}) = o(1)$$

(since Q is idempotent, its trace equals its rank, which is K here). Since $\left[\underline{U}^T D(\tau)MD(\tau)\underline{U}/N\right]^{1/2} = O_P(1)$, the entire term is $o_P(1)$.

The third term is $\underline{U}^T MD(\bar{\tau})\underline{h}/N$. As before $\underline{U}^T MD(\bar{\tau})\underline{h}/N - \underline{U}^T MD(\tau^*)\underline{h}/N = o_P(1)$. We have

$$\underline{U}^T MD(\tau^*)\underline{h}/n \le \left(\underline{U}^T M\underline{U}/N\right)^{1/2}\left(\underline{h}^T D(\tau^*)MD(\tau^*)\underline{h}/N\right)^{1/2}.$$

We know that $\underline{U}^T M\underline{U}/N = o_P(1)$ from the argument above. Next,

$$\underline{h}^T D(\tau^*)MD(\tau^*)\underline{h}/N \le \underline{h}^T D(\tau^*)D(\tau^*)\underline{h}/N,$$

which is $O_P(1)$ by the weak law of large numbers as long as $E[h^2(X)Y^2\exp(-2\tau^* A)] < \infty$. Hence the product is $o_P(1)$.

The fourth term is $\underline{h}^T MD(\bar{\tau})\underline{U}/N$. Again, first note that $\underline{h}^T MD(\bar{\tau})\underline{U}/N - \underline{h}^T MD(\tau^*)\underline{U}/N = o_P(1)$ so we only need to work with the true τ^*. First note that $\underline{h}^T MD(\tau)\underline{U}/n$ is centered:

$$E[\underline{h}^T MD(\tau)\underline{U}/N] = E\left\{E[\underline{h}^T MD(\tau)\underline{U}/N|\underline{A},\underline{X}]\right\} = E\left\{\underline{h}^T ME[D(\tau)|\underline{A},\underline{X}]\underline{U}/N\right\}$$
$$= E\left\{\underline{h}^T M\mathrm{diag}[\underline{g(X)}]\underline{U}/N\right\} = E\left\{E[\underline{h}^T M\mathrm{diag}[\underline{g(X)}]\underline{U}/N|\underline{X}]\right\}$$
$$= E\left\{\underline{h}^T M\mathrm{diag}[\underline{g(X)}]E[\underline{U}^T|X]/N\right\} = E\left\{\underline{h}^T M\mathrm{diag}[\underline{g(X)}]\cdot 0/N\right\} = 0.$$

So we can use Chebychev's inequality as in Donald and Newey (1994):

$$\underline{h}^T MD(\tau)\underline{U}/N = (\underline{h} - p_K\eta)^T MD(\tau)\underline{U}/N.$$

So the variance is

$$E\left\{N^{-2}(\underline{h} - p_K\eta)^T MD(\tau)\underline{U}\underline{U}^T D(\tau)M(\underline{h} - p_K\eta)\right\}$$
$$= E\left\{N^{-2}(\underline{h} - p_K\eta)^T ME\left[D(\tau)\underline{U}\underline{U}^T D(\tau)|\underline{X}\right]M(\underline{h} - p_K\eta)\right\}$$
$$= E\left\{N^{-2}(\underline{h} - p_K\eta)^T ME\left[\left\{(Y_i\exp\left[-A_i^T\tau\right]U_i Y_j\exp\left[-A_j^T\tau\right]U_j)\right\}|\underline{X}\right]M(\underline{h} - p_K\eta)\right\}$$
$$= E\left\{N^{-2}(\underline{h} - p_K\eta)^T ME\left[\mathrm{diag}\left\{(Y^2\exp[-2\tau A]U^2)\right\}|\underline{X}\right]M(\underline{h} - p_K\eta)\right\}$$
$$\le \Delta E\left\{N^{-2}(\underline{h} - p_K\eta)^T MM(\underline{h} - p_K\eta)\right\}$$
$$\le \Delta N^{-1}e_h^2(K).$$

So $\underline{U}^T MD(\bar{\tau})\underline{h}/N = O_P\left[N^{-1/2}e_h(K)\right] = o_P(1)$.

The final term is $\underline{h}^T MD(\bar{\tau})\underline{h}/N$. Again, $\underline{h}^T MD(\bar{\tau})\underline{h}/N - \underline{h}^T MD(\tau^*)\underline{h}/N = o_P(1)$ and we bound

$$\underline{h}^T MD(\tau^*)\underline{h}/N \le \left(\underline{h}^T M\underline{h}/N\right)^{1/2}\left[\underline{h}^T D(\tau)MD(\tau^*)\underline{h}/N\right]^{1/2}.$$

As before, $\underline{h}^T M\underline{h}/N = O_P\left[e_h^2(K)\right] = o_P(1)$ if h is in the appropriate Hölder class. Next, if $E[h^2(X)Y^2\exp(-2\tau A)] < \infty$, then $\underline{h}^T D(\tau)MD(\tau)\underline{h}/N = O_P(1)$. So this term is $o_P(1)$.

- Next, $\left(\underline{A}^T M\left\{\underline{g}+\underline{\epsilon}\right\}/n\right) = O_P\left[e_h(K)e_g(K)\right]$. Here the proof is identical to that in Donald and Newey (1994). We write

$$A^T M(\underline{g}+\underline{\epsilon})/N = (\underline{U}+\underline{h})^T M(\underline{g}+\underline{\epsilon})/N = \underline{U}^T M\underline{g}/N + \underline{U}^T M\underline{\epsilon}/N + \underline{h}^T M\underline{g}/N + \underline{h}^T M\underline{\epsilon}/N$$
$$= \underline{U}^T M\underline{g}/N + \underline{U}^T \underline{\epsilon}/N - \underline{U}^T Q\underline{\epsilon}/N + \underline{h}^T M\underline{g}/N + \underline{h}^T M\underline{\epsilon}/N.$$

  The first term is $\underline{U}^T M\underline{g}/N = O_P\left[N^{-1/2}e_g(K)\right]$. The term $\underline{U}^T\underline{\epsilon}/N$ is $O_P(N^{-1/2})$. $\underline{U}^T Q\underline{\epsilon}/N = O_P(K^{1/2}N^{-1})$. The term $\underline{h}^T M\underline{g}/N = O_P\left[e_h(K)e_g(K)\right]$. Finally, $\underline{h}^T M\underline{\epsilon}/N = O_P\left[N^{-1/2}e_h(K)\right]$.

Note that all of the above terms are at least $O_P(N^{-1/2})$ except possibly for the term $O_P\left[e_h(K)e_g(K)\right]$. Now approximation rates when optimal bases are used are

$$e_h(K) = K^{-\beta_h/d}, \text{ and } e_g(K) = K^{-\beta_g/d},$$

where β_h and β_g are the Hölder exponents of $h(X) = A - E[A|X]$ and $g(X) = E[e^{\zeta^*(X)}]$, respectively (see, Theorem 8 in Chapter 6 of Lorentz, 1986, for polynomial bases and Theorem 12.8 of Schumaker, 1981, for splines). Hence, when $\beta_h + \beta_g > d/2$, we get a rate faster than $K^{-1/2}$, and can make K sufficiently close to N to get $K^{-1/2} = O_P(N^{-1/2})$ and still have $K = o(N)$. Note that the rate β_h corresponds to β_p and β_g corresponds to β_b in the rest of this report.

It is an open question whether the usual Poisson regression estimator is $\sqrt{n}$-consistent and asymptotically unbiased whenever $\beta_b + \beta_p \geq d/2$. The simulation studies in the next section suggest it is.

APPENDIX E. PROOFS

Proof of Theorem 2. The model condition is equivalent to

$$E\left[\{\xi(Y,A,X;\tau) - E[\xi(Y,A,X;\tau)|X]\}\{h(A,X) - E[h(A,X)|X]\}\right] = 0$$

for every $h(A,X) \in \mathscr{L}_2(P^*)$. For brevity, write

$$\Delta\xi(\theta) = \xi\left[Y,A,X;\tau(P_\theta)\right] - E_\theta\left[\xi\left\{Y,A,X;\tau(P_\theta)\right\}|X\right], \text{ and } \Delta h(\theta) = h(A,X) - E_\theta[h(A,X)|X].$$

Differentiating both sides of the equation and evaluating at θ^* gives

$$\begin{aligned}
0 &= \frac{d}{d\theta}E_\theta[\Delta\xi(\theta)\ \Delta h(\theta)]\Big|_{\theta=\theta^*} \\
&= \frac{d}{d\theta}E_{\theta^*}[\Delta\xi(\theta)\ \Delta h(\theta)]\Big|_{\theta=\theta^*} + \frac{d}{d\theta}E_\theta[\Delta\xi(\theta^*)\ \Delta h(\theta^*)]\Big|_{\theta=\theta^*} \\
&= \frac{d}{d\theta}\,E_{\theta^*}\left[\Delta\xi(\theta)\ \Delta h(\theta^*)\right]\big|_{\theta=\theta^*} + \frac{d}{d\theta}\,E_{\theta^*}\left[\Delta\xi(\theta^*)\ \Delta h(\theta)\right]\big|_{\theta=\theta^*} \\
&\quad + E_{\theta^*}[\Delta\xi(\theta^*)\ \Delta h(\theta^*)S_\theta(Y,A,X;\theta^*)] \\
&= \frac{d}{d\tau}\,E_{\theta^*}\left\{E_{\theta^*}[\Delta\xi(\theta)|A,X]\ \Delta h(\theta)\right\}\left(\frac{d\tau}{d\theta}\right)\Big|_{\theta=\theta^*} + E_{\theta^*}\left\{E_\theta[\xi(\theta^*)S_\theta(Y,A|X;\theta)|X]\ \Delta h(\theta)\right\}\big|_{\theta=\theta^*} \\
&\quad - E_{\theta^*}\left\{\Delta\xi(\theta)\ E_\theta[h(A,X)S_\theta(A|X;\theta)|X]\right\}\big|_{\theta=\theta^*} + E_{\theta^*}[\Delta\xi(\theta^*)\ \Delta h(\theta^*)S_\theta(Y,A,X;\theta^*)].
\end{aligned}$$

Here, $S_\theta(Y,A|X;\theta)$ and $S_\theta(A|X;\theta)$ are the conditional scores

$$S_\theta(Y,A|X;\theta) = \frac{d}{d\theta}\log f(Y,A|X;\theta), \text{ and}$$
$$S_\theta(A|X;\theta) = \frac{d}{d\theta}\log f(A|X;\theta).$$

Conditioning on X, we see that the second and third terms are zero. Writing

$$J(A,X;P^*) = \frac{d}{d\tau}\, E[\Delta\xi(Y,A,X;\tau)|A,X]|_{\tau=\tau^*},$$

we get

$$\left.\frac{d\tau}{d\theta}\right|_{\theta=\theta^*} = E_{\theta^*}\left\{-\,E_{\theta^*}\left[J(A,X;P^*)\;\Delta h(\theta)\right]\big|^{-1}_{\theta=\theta^*}\Delta\xi(\theta^*)\;\Delta h(\theta^*)S_\theta(Y,A,X;\theta^*)\right\},$$

whence the collection of influence functions can be written as

$$\Big\{-\,E_{\theta^*}\left[J(A,X;P^*)\;\Delta h(\theta)\right]\big|^{-1}_{\theta=\theta^*}\Delta\xi(\theta^*)\;\Delta h(\theta^*) : h(A,X)\in\mathscr{L}_2(P^*),$$
$$E_{\theta^*}\left[J(A,X;P^*)\;\Delta h(\theta)\right]\big|_{\theta=\theta^*}\ \text{nonzero}\Big\}.$$

To find the efficient influence function, we define

$$\tilde{J}(X;P^*) = E_{\theta^*}\left\{J(A,X;P^*)\,\mathrm{Var}^{-1}_{\theta^*}\left[\Delta\xi(\theta^*)|A,X\right]|X\right\}E_{\theta^*}\left\{\mathrm{Var}^{-1}_{\theta^*}\left[\Delta\xi(\theta^*)|A,X\right]|X\right\}^{-1},$$

and write

$$\begin{aligned}
&\mathrm{Var}^{-1}_{\theta^*}\left(E_{\theta^*}\left[J(A,X;P^*)\;\Delta h(\theta)\right]\big|^{-1}_{\theta=\theta^*}\Delta\xi(\theta^*)\;\Delta h(\theta^*)\right)\\
&= \mathrm{Var}_{\theta^*}\left(E_{\theta^*}\left[J(A,X;P^*)\;\Delta h(\theta)\right]\big|_{\theta=\theta^*}\mathrm{Var}^{-1}_{\theta^*}\left[\Delta\xi(\theta^*)\;\Delta h(\theta^*)\right]\Delta\xi(\theta^*)\;\Delta h(\theta^*)\right)\\
&= \mathrm{Var}_{\theta^*}\left(E_{\theta^*}\left[\left(J(A,X;P^*)-\tilde{J}(X;P^*)\right)\;\Delta h(\theta)\right]\big|_{\theta=\theta^*}\mathrm{Var}^{-1}_{\theta^*}\left[\Delta\xi(\theta^*)\;\Delta h(\theta^*)\right]\Delta\xi(\theta^*)\;\Delta h(\theta^*)\right)\\
&= \mathrm{Var}_{\theta^*}\Bigg(E_{\theta^*}\left[\left(J(A,X;P^*)-\tilde{J}(X;P^*)\right)\;\mathrm{Var}^{-1}_{\theta^*}\left(\Delta\xi(\theta^*)|A,X\right)\Delta\xi(\theta^*)\Delta\xi(\theta^*)\Delta h(\theta)\right]\Big|_{\theta=\theta^*}\\
&\qquad\qquad\times\mathrm{Var}^{-1}_{\theta^*}\left[\Delta\xi(\theta^*)\;\Delta h(\theta^*)\right]\Delta\xi(\theta^*)\;\Delta h(\theta^*)\Bigg)\\
&= \mathrm{Var}_{\theta^*}\left(\prod_{\theta=\theta^*}\left[\left\{J(A,X;P^*)-\tilde{J}(X;P^*)\right\}\;\mathrm{Var}^{-1}_{\theta^*}\left\{\Delta\xi(\theta^*)|A,X\right\}\Delta\xi(\theta^*)\;\middle|\;\Delta\xi(\theta^*)\Delta h(\theta)\right]\right)\\
&\le \mathrm{Var}_{\theta^*}\left\{\left[J(A,X;P^*)-\tilde{J}(X;P^*)\right]\;\mathrm{Var}^{-1}_{\theta^*}\left[\Delta\xi(\theta^*)|A,X\right]\Delta\xi(\theta^*)\right\}.
\end{aligned}$$

In the above, Π_{θ^*} refers to the $\mathscr{L}_2(P^*)$ projection of its first argument onto the closure of the linear space spanned by its second argument. Hence, the efficient influence function is

$$\left[J(A,X;P^*)-\tilde{J}(X;P^*)\right]\;\mathrm{Var}^{-1}_{\theta^*}\left[\Delta\xi(\theta^*)|A,X\right]\Delta\xi(\theta^*).$$

Proof of Corollary 3. The linear semiparametric model is a special case of the model of Theorem 2 with $\xi(Y,A,X;\tau) = Y-\tau A$. Here

$$\Delta\xi(Y,A,X;\tau) = Y-\tau A - E[Y-\tau A|X],$$

$$J(A,X;P^*) = p_\tau^*(X) - A,$$
$$\mathrm{Var}\left[\Delta\xi(Y,A,X;\tau^*)|A,X\right] = E\left(\left\{Y-\tau A-b(X)\right\}^2|A,X\right) = \sigma^2 \text{ under homoscedasticity, and}$$
$$\tilde{J}(X;P^*) = E\left\{J(A,X;P^*)\,\mathrm{Var}^{-1}\left[\Delta\xi(Y,A,X;\tau^*)|A,X\right]|X\right\}$$
$$\times E\left\{\mathrm{Var}^{-1}\left[\Delta\xi(Y,A,X;\tau^*)|A,X\right]|X\right\}^{-1} = 0.$$

The result follows directly from the formula obtained in Theorem 2.

Proof of Corollary 4. The loglinear semiparametric model is a special case of the model of Theorem 2 with $\xi(Y,A,X;\tau) = Y\exp(-\tau A)$. We assume $\mathrm{Var}[Y|A,X] = \sigma^2 e^{\tau^* A+\zeta^*(X)}$. Here

$$\Delta\xi(Y,A,X;\tau) = Ye^{-\tau A} - e^{\zeta^*(X)}E\left[e^{(\tau^*-\tau)A}|X\right],$$
$$J(A,X;P^*) = (E[A|X]-A)e^{\zeta^*(X)},$$
$$\mathrm{Var}\left[\Delta\xi(Y,A,X;\tau^*)|A,X\right] = \mathrm{Var}\left[\xi(Y,A,X;\tau^*)|A,X\right] = \mathrm{Var}(Ye^{-\tau^* A}|A,X)$$
$$= e^{-2\tau^* A}\mathrm{Var}[Y|A,X] = e^{-\tau^* A+\zeta^*(X)}\sigma^2,$$
$$\tilde{J}(X;P^*) = E\left[J(A,X;P^*)\,\mathrm{Var}^{-1}\left(\Delta\xi(Y,A,X;\tau^*)|A,X\right)|X\right]$$
$$\times E\left\{\mathrm{Var}^{-1}\left[\Delta\xi(Y,A,X;\tau^*)|A,X\right]|X\right\}^{-1}$$
$$= -e^{\zeta^*(X)}\frac{E\left[e^{\tau^* A}(A-E[A|X])|X\right]}{E\left[e^{\tau^* A}|X\right]}, \text{ and}$$
$$J-\tilde{J} = -e^{\zeta^*(X)}\left\{A-\frac{E\left[Ae^{\tau^* A}|X\right]}{E\left[e^{\tau^* A}|X\right]}\right\}.$$

So Theorem 2 implies the efficient influence function is

$$e^{\zeta^*(X)}\left\{A-\frac{E\left[Ae^{\tau^* A}|X\right]}{E\left[e^{2\tau^* A}|X\right]}\right\}\times e^{\tau^* A-\zeta^*(X)}\sigma^{-2}\times e^{-\tau^* A}\left\{Y-e^{\tau^* A+\zeta^*(X)}\right\}$$
$$= \sigma^{-2}\left\{Y-e^{\tau^* A+\zeta^*(X)}\right\}\left\{A-\frac{E\left[e^{\tau^* A}A|X\right]}{E\left[e^{\tau^* A}|X\right]}\right\}.$$

Proof of Lemma 9. It is well known that $\hat{\tau}$, the coefficient of A in the usual linear or Poisson regression of Y on A, W, also solves

$$\frac{1}{N}\sum_{i=1}^N U_{i,\mathrm{profile}}\left\{\tau, b\left[\hat{\eta}^{\mathrm{full}}(\tau)\right]\right\} = 0.$$

But

$$\frac{1}{N}\sum_{i=1}^N U_{i,\mathrm{profile}}\left\{\tau, b\left[\hat{\eta}^{\mathrm{full}}(\tau)\right]\right\} - \mathbb{IF}_{1,\mathrm{eff}}^{\mathrm{full}}(\tau) = \frac{1}{N}\sum_{i=1}^N \varepsilon_i\left\{\tau, b_i\left[\hat{\eta}^{\mathrm{full}}(\tau)\right]\right\}\ \hat{\alpha}^{\mathrm{full}}(\tau)^T W_i$$
$$= \hat{\alpha}^{\mathrm{full}}(\tau)^T\left[\frac{1}{N}\sum_{i=1}^N U_{i,\mathrm{nuis}}\left\{\tau, b_i\left[\hat{\eta}^{\mathrm{full}}(\tau)\right]\right\}\right] = 0$$

by definition of $\hat{\eta}^{\mathrm{full}}(\tau)$. Hence $\hat{\tau}$ and $\hat{\tau}_{1,\mathrm{eff}}^{\mathrm{full}}$ solve the same equation.

Proof of Lemma 13. We give the proof for the loglinear case since the linear case is similar and simpler.

$$\begin{aligned}
\mathrm{Bias}_{1,\mathrm{eff}}(\tau;b,p) &= E[\{Y-e^{\tau A}b(X)\}\{A-p(X)\}]-E[\{Y-e^{\tau A}b_\tau^*(X)\}\{A-p_\tau^*(X)\}]\\
&= E[\{Y-e^{\tau A}b_\tau^*(X)+e^{\tau A}b_\tau^*(X)-e^{\tau A}b(X)\}\{A-p_\tau^*(X)+p_\tau^*(X)-p(X)\}]\\
&\quad -E[\{Y-e^{\tau A}b_\tau^*(X)\}\{A-p_\tau^*(X)\}]\\
&= E[\{Y-e^{\tau A}b_\tau^*(X)\}\{p_\tau^*(X)-p(X)\}]\\
&\quad +E[e^{\tau A}\{b_\tau^*(X)-b(X)\}\{A-p_\tau^*(X)\}]+E[e^{\tau A}\{b_\tau^*(X)-b(X)\}\{p_\tau^*(X)-p(X)\}]\\
&= E[q_\tau^*(X)\{b_\tau^*(X)-b(X)\}\{p_\tau^*(X)-p(X)\}].
\end{aligned}$$

The last equality holds because the first and second terms are zero: the first by conditioning on X and observing that $E[Y-e^{\tau A}b_\tau^*(X)|X]=0$, and the second since $E[e^{\tau A}\{A-p_\tau^*(X)\}|X]=0$, by Definition 5.

Proof of Lemma 18. As before, we give the proof for the loglinear case.

$$\begin{aligned}
&\mathrm{Bias}_{1,\mathrm{new}}(\tau;b,p,f)\\
&= E[f(X)\{Y-e^{\tau A}b(X)\}\{A-p(X)\}]-E[f(X)\{Y-e^{\tau A}b_\tau^*(X)\}\{A-p_\tau^*(X)\}]\\
&= E[f(X)q_\tau^*(X)\{b_\tau^*(X)-b(X)\}\{p_\tau^*(X)-p(X)\}]\\
&= E[q_\tau^*(X)f^*(X)\{b_\tau^*(X)-b(X)\}\{p_\tau^*(X)-p(X)\}]\\
&\quad +E[q_\tau^*(X)\{f(X)-f^*(X)\}\{b_\tau^*(X)-b(X)\}\{p_\tau^*(X)-p(X)\}].
\end{aligned}$$

The details of the proof are similar to those for Lemma 13.

Proof of Theorem 21. We focus on the loglinear case since the linear case is similar as well as simpler. Note that

$$\int K_k(x,y)g(x)\,dx=\sum_{l=1}^{k}\gamma_l\varphi_l(x)=\prod[g(x)|\bar\varphi_k(x)];$$

that is, K_k projects $g(x)$ onto the first k components of the basis. Hence if $\xi(x),\rho(x)\in\mathscr{L}_2(\mu)$, then

$$\begin{aligned}
&\int\xi(x_1)K_k(x_1,x_2)\rho(x_2)\,dx_1\,dx_2=\int\prod\left[\xi(x)\middle|\bar\varphi_k(x)\right]\prod\left[\rho(x)\middle|\bar\varphi_k(x)\right]dx\\
&=\int\xi(x)\rho(x)\,dx-\int\prod\left[\xi(x)\middle|\bar\varphi_k^{\perp}(x)\right]\prod\left[\rho(x)\middle|\bar\varphi_k^{\perp}(x)\right]dx.
\end{aligned}$$

Next note that for the kernels $K=K_k$ or $K=K_{\hat f,k}$,

$$\begin{aligned}
E[\widehat{\mathbb{IF}}_{22}^{(k)}(\tau;K)] &= -E[\varepsilon_1(\tau,\hat b_\tau)K(X_1,X_2)\Delta_2(\tau,\hat b_\tau,\hat q_\tau)]\\
&= -E\left\{E[\varepsilon_1(\tau,\hat b_\tau)|X_1]K(X_1,X_2)E[\Delta_2(\tau,\hat b_\tau,\hat q_\tau)|X_2]\right\}\\
&= -E\left[E[e^{\tau A_1}|X_1]\left\{b_\tau^*(X_1)-\hat b_\tau(X_1)\right\}K(X_1,X_2)\left\{p_\tau^*(X_2)-\hat p_\tau(X_2)\right\}\frac{E[e^{\tau A_2}|X_2]}{\hat q_\tau(X_2)}\right]\\
&= -E\left[q_\tau^*(X_1)\delta b_\tau(X_1)K(X_1,X_2)\delta p_\tau(X_2)\frac{q_\tau^*(X_2)}{\hat q_\tau(X_2)}\right].
\end{aligned}$$

Similarly, for the kernel $K = K_{\hat{q},\hat{f},k,\text{alt}}$,

$$E[\widehat{\mathbb{IF}}_{22}^{(k)}(\tau;K)] = -E\left[\frac{q_\tau^*(X_1)}{\hat{q}_\tau^{1/2}(X_1)}\delta b_\tau(X_1)K(X_1,X_2)\delta p_\tau(X_2)\frac{q_\tau^*(X_2)}{\hat{q}_\tau^{1/2}(X_2)}\right].$$

- Proof for $E[\widehat{\mathbb{IF}}_{22}^{(k)}(\tau;K_{\hat{f},k})]$:
  Note that if X_1 and X_2 are i.i.d. $\sim f^*$, and $\xi(x), \rho(x) \in \mathscr{L}_2(\mu)$, then for any f

$$\begin{aligned}
E_{f^*}[\xi(X_1)K_{f,k}(X_1,X_2)\rho(X_2)] &= \int \prod\left[\xi(x)\frac{f^*(x)}{f^{1/2}(x)}\middle|\bar{\varphi}_k(x)\right]\prod\left[\rho(x)\frac{f^*(x)}{f^{1/2}(x)}\middle|\bar{\varphi}_k(x)\right]dx \\
&= E_{f^*}\left[\xi(X)\rho(X)\frac{f^*(X)}{f(X)}\right] - \int \prod\left[\xi(x)\frac{f^*(x)}{f^{1/2}(x)}\middle|\bar{\varphi}_k^\perp(x)\right]\prod\left[\rho(x)\frac{f^*(x)}{f^{1/2}(x)}\middle|\bar{\varphi}_k^\perp(x)\right]dx \\
&= E_{f^*}\left[\xi(X)\rho(X)\right] + E_{f^*}\left[\xi(X)\rho(X)\left\{\frac{f^*(X)}{f(X)}-1\right\}\right] \\
&\qquad - \int \prod\left[\xi(x)\frac{f^*(x)}{f^{1/2}(x)}\middle|\bar{\varphi}_k^\perp(x)\right]\prod\left[\rho(x)\frac{f^*(x)}{f^{1/2}(x)}\middle|\bar{\varphi}_k^\perp(x)\right]dx.
\end{aligned}$$

  Using this formula with $\xi(x) = q_\tau^*(x)\delta b(x)$ and $\rho(x) = p_\tau(x)$, we get

$$\begin{aligned}
-E[\widehat{\mathbb{IF}}_{22}^{(k)}(\tau;K_{\hat{f},k})] &= E[q_\tau^*(X_1)\delta b_\tau(X_1)K_{\hat{f},k}(X_1,X_2)\delta p_\tau(X_2)q_\tau^*(X_2)\hat{q}_\tau^{-1}(X_2)] \\
&= E[q_\tau^*(X_1)\delta b_\tau(X_1)K_{\hat{f},k}(X_1,X_2)\delta p_\tau(X_2)] \\
&\quad + E\left[q_\tau^*(X_1)\delta b_\tau(X_1)K_{\hat{f},k}(X_1,X_2)\delta p_\tau(X_2)\left\{\frac{q_\tau^*(X_2)}{\hat{q}_\tau(X_2)}-1\right\}\right] \\
&= E[q_\tau^*(X)\delta b_\tau(X)\delta p_\tau(X)] + E\left[q_\tau^*(X)\delta b_\tau(X)\delta p_\tau(X)\left\{\frac{f^*(X)}{\hat{f}(X)}-1\right\}\right] \\
&\qquad - \int \prod\left[q_\tau^*(x)\delta b_\tau(x)\frac{f^*(x)}{\hat{f}^{1/2}(x)}\middle|\bar{\varphi}_k^\perp(x)\right]\prod\left[\delta p_\tau(x)\frac{f^*(x)}{\hat{f}^{1/2}(x)}\middle|\bar{\varphi}_k^\perp(x)\right]dx \\
&\quad + E\left[q_\tau^*(X_1)\delta b_\tau(X_1)K_{\hat{f},k}(X_1,X_2)\delta p_\tau(X_2)\left\{\frac{q_\tau^*(X_2)}{\hat{q}_\tau(X_2)}-1\right\}\right].
\end{aligned}$$

  Note that the first term equals $\text{Bias}_{1,\text{eff}}(\tau;\hat{b}_\tau,\hat{p}_\tau)$, which is second order. The second and fourth terms are third order. The term with projections is the tail term.

- Proof for $E[\widehat{\mathbb{IF}}_{22}^{(k)}(\tau;K_{\hat{q},\hat{f},k,\text{alt}})]$:
  Note that

$$\begin{aligned}
-E[\widehat{\mathbb{IF}}_{22}^{(k)}(\tau;K_{\hat{q},\hat{f},k,\text{alt}})] &= E\left[\frac{q_\tau^*(X_1)}{\hat{q}_\tau^{1/2}(X_1)}\delta b_\tau(X_1)K_{\hat{q},\hat{f},k,\text{alt}}(X_1,X_2)\delta p_\tau(X_2)\frac{q_\tau^*(X_2)}{\hat{q}_\tau^{1/2}(X_2)}\right] \\
&= E\left[\hat{q}_\tau^{1/2}(X_1)\left\{\frac{q_\tau^*(X_1)}{\hat{q}_\tau(X_1)}-1+1\right\}\delta b_\tau(X_1)K_{\hat{q},\hat{f},k,\text{alt}}(X_1,X_2)\right. \\
&\qquad\qquad \left.\times\,\delta p_\tau(X_2)\left\{\frac{q_\tau^*(X_2)}{\hat{q}_\tau(X_2)}-1+1\right\}\hat{q}_\tau^{1/2}(X_2)\right] \\
&= E[\hat{q}_\tau^{1/2}(X_1)\delta b_\tau(X_1)K_{\hat{q},\hat{f},k,\text{alt}}(X_1,X_2)\delta p_\tau(X_2)\hat{q}_\tau^{1/2}(X_2)] \\
&\quad + E\left[\hat{q}_\tau^{1/2}(X_1)\delta b_\tau(X_1)K_{\hat{q},\hat{f},k,\text{alt}}(X_1,X_2)\delta p_\tau(X_2)\left\{\frac{q_\tau^*(X_2)}{\hat{q}_\tau(X_2)}-1\right\}\hat{q}_\tau^{1/2}(X_2)\right] \\
&\quad + E\left[\hat{q}_\tau^{1/2}(X_1)\left\{\frac{q_\tau^*(X_1)}{\hat{q}_\tau(X_1)}-1\right\}\delta b_\tau(X_1)K_{\hat{q},\hat{f},k,\text{alt}}(X_1,X_2)\delta p_\tau(X_2)\hat{q}_\tau^{1/2}(X_2)\right]
\end{aligned}$$

$$+E\left[\hat{q}_\tau^{1/2}(X_1)\left\{\frac{q_\tau^*(X_1)}{\hat{q}_\tau(X_1)}-1\right\}\delta b_\tau(X_1)K_{\hat{q},\hat{f},k,\mathrm{alt}}(X_1,X_2)\right.$$
$$\left.\times\,\delta p_\tau(X_2)\left\{\frac{q_\tau^*(X_2)}{\hat{q}_\tau(X_2)}-1\right\}\hat{q}_\tau^{1/2}(X_2)\right].$$

Note that the last three terms are third or higher order. The first term is

$$\begin{aligned}
&E[\hat{q}_\tau^{1/2}(X_1)\delta b_\tau(X_1)K_{\hat{q},\hat{f},k,\mathrm{alt}}(X_1,X_2)\delta p_\tau(X_2)\hat{q}_\tau^{1/2}(X_2)]\\
&=E_{\hat{f}}\left[\frac{f^*(X_1)}{\hat{f}(X_1)}\hat{q}_\tau^{1/2}(X_1)\delta b_\tau(X_1)K_{f,k,\mathrm{alt}}(X_1,X_2)\delta p_\tau(X_2)\hat{q}_\tau^{1/2}(X_2)\frac{f^*(X_2)}{\hat{f}(X_2)}\right]\\
&=E_{\hat{f}}\left[\hat{q}_\tau^{1/2}(X_1)\delta b_\tau(X_1)K_{\hat{q},\hat{f},k,\mathrm{alt}}(X_1,X_2)\delta p_\tau(X_2)\hat{q}_\tau^{1/2}(X_2)\right]\\
&\quad+E_{\hat{f}}\left[\hat{q}_\tau^{1/2}(X_1)\delta b_\tau(X_1)K_{\hat{q},\hat{f},k,\mathrm{alt}}(X_1,X_2)\delta p_\tau(X_2)\hat{q}_\tau^{1/2}(X_2)\left\{\frac{f^*(X_2)}{\hat{f}(X_2)}-1\right\}\right]\\
&\quad+E_{\hat{f}}\left[\left\{\frac{f^*(X_1)}{\hat{f}(X_1)}-1\right\}\hat{q}_\tau^{1/2}(X_1)\delta b_\tau(X_1)K_{\hat{q},\hat{f},k,\mathrm{alt}}(X_1,X_2)\delta p_\tau(X_2)\hat{q}_\tau^{1/2}(X_2)\right]\\
&\quad+E_{\hat{f}}\left[\left\{\frac{f^*(X_1)}{\hat{f}(X_1)}-1\right\}\hat{q}_\tau^{1/2}(X_1)\delta b_\tau(X_1)K_{\hat{q},\hat{f},k,\mathrm{alt}}(X_1,X_2)\right.\\
&\qquad\left.\times\,\delta p_\tau(X_2)\hat{q}_\tau^{1/2}(X_2)\left\{\frac{f^*(X_2)}{\hat{f}(X_2)}-1\right\}\right].
\end{aligned}$$

Again the last three terms are third or higher order and we ignore them. The first term is

$$\begin{aligned}
&E_{\hat{f}}[\hat{q}_\tau^{1/2}(X_1)\delta b_\tau(X_1)K_{\hat{q},\hat{f},k,\mathrm{alt}}(X_1,X_2)\delta p_\tau(X_2)\hat{q}_\tau^{1/2}(X_2)]\\
&=E_{\hat{f}}\left\{\prod_{\hat{f}}[\hat{q}_\tau^{1/2}(X)\delta b_\tau(X)|\bar{\varphi}_k(X)\hat{q}_\tau^{1/2}(X)]\prod_{\hat{f}}[\delta p_\tau(X)\hat{q}_\tau^{1/2}(X)|\bar{\varphi}_k(X)\hat{q}_\tau^{1/2}(X)]\right\}\\
&=E_{\hat{f}}[\hat{q}_\tau(X)\delta b_\tau(X)\delta p_\tau(X)]\\
&\quad-E_{\hat{f}}\left\{\prod_{\hat{f}}\left[\hat{q}_\tau^{1/2}(X)\delta b_\tau(X)\left|\left[\bar{\varphi}_k(X)\hat{q}_\tau^{1/2}(X)\right]^\perp\right.\right]\prod_{\hat{f}}\left[\delta p_\tau(X)\hat{q}_\tau^{1/2}(X)\left|\left[\bar{\varphi}_k(X)\hat{q}_\tau^{1/2}(X)\right]^\perp\right.\right]\right\}.
\end{aligned}$$

The first term in the above is

$$\begin{aligned}
&E_{\hat{f}}[\hat{q}_\tau(X)\delta b_\tau(X)\delta p_\tau(X)]=E\left[\frac{\hat{f}(X)}{f^*(X)}\hat{q}_\tau(X)\delta b_\tau(X)\delta p_\tau(X)\right]\\
&=E[q_\tau^*(X)\delta b_\tau(X)\delta p_\tau(X)]\\
&\quad+E[\{\hat{q}_\tau(X)-q_\tau^*(X)\}\delta b_\tau(X)\delta p_\tau(X)]\\
&\quad+E\left[\left\{\frac{\hat{f}(X)}{f^*(X)}-1\right\}\hat{q}_\tau(X)\delta b_\tau(X)\delta p_\tau(X)\right].
\end{aligned}$$

Of these three terms, the first term is the bias of the first-order estimator, and the remaining terms are higher order.

Summarizing the terms that arise, we have terms whose rate of convergence to 0 depends on the differences $\hat{q}_\tau(X)-q_\tau^*(X)$ or $q_\tau^*(X)/\hat{q}_\tau(X)-1$ and δb_τ and δp_τ. We call the sum of these

terms $\mathrm{EB}^{(3)}_{q,b,p}$. We also have terms whose rate of convergence to 0 depends on the differences $\hat{f}(X)/f^*(X)-1$ or $f^*(X)/\hat{f}(X)-1$ and δb_τ and δp_τ. We call the sum of these terms $\mathrm{EB}^{(3)}_{f,b,p}$. The truncation term

$$E_{\hat{f}}\left\{\prod_{\hat{f}}\left[\hat{q}_\tau^{1/2}(X)\delta b_\tau(X)\big|\left[\bar{\varphi}_k(X)\hat{q}_\tau^{1/2}(X)\right]^\perp\right]\prod_{\hat{f}}\left[\delta p_\tau(X)\hat{q}_\tau^{1/2}(X)\big|\left[\bar{\varphi}_k(X)\hat{q}_\tau^{1/2}(X)\right]^\perp\right]\right\}$$

is $O_P(k^{-(\beta_b+\beta_p)/d})$.

- Proof for $E[\widehat{\mathbb{IF}}^{(k)}_{22}(\tau;K_k)]$:

  For any density f on the support of X, we have

$$\begin{aligned}
&-E[\widehat{\mathbb{IF}}^{(k)}_{22}(\tau;K_k)] = E[q^*_\tau(X_1)\delta b_\tau(X_1)K_k(X_1,X_2)\delta p_\tau(X_2)q^*_\tau(X_2)\hat{q}_\tau^{-1}(X_2)]\\
&= E[q^*_\tau(X_1)\delta b_\tau(X_1)K_k(X_1,X_2)\delta p_\tau(X_2)]\\
&\quad + E\left[q^*_\tau(X_1)\delta b_\tau(X_1)K_k(X_1,X_2)\delta p_\tau(X_2)\left\{\frac{q^*_\tau(X_2)}{\hat{q}_\tau(X_2)}-1\right\}\right]\\
&= \int f^{*2}(x)q^*_\tau(x)\delta b_\tau(x)\delta p_\tau(x)dx\\
&\quad + E\left[q^*_\tau(X_1)\delta b_\tau(X_1)K_k(X_1,X_2)\delta p_\tau(X_2)\left\{\frac{q^*_\tau(X_2)}{\hat{q}_\tau(X_2)}-1\right\}\right]\\
&\quad - \int\prod\left[f^*(x)q^*_\tau(x)\delta b_\tau(x)\big|\bar{\varphi}_k^\perp(x)\right]\prod\left[f^*(x)\delta p_\tau(x)\big|\bar{\varphi}_k^\perp(x)\right]dx\\
&= E[f^*(X)q^*_\tau(X)\delta b_\tau(X)\delta p_\tau(X)] + E\left[f^*(X)q^*_\tau(X)\delta b_\tau(X)\delta p_\tau(X)\left\{\frac{q^*_\tau(X)}{\hat{q}_\tau(X)}-1\right\}\right]\\
&\quad - \int\prod\left[f^*(x)q^*_\tau(x)\delta b_\tau(x)\big|\bar{\varphi}_k^\perp(x)\right]\prod\left[f^*(x)\delta p_\tau(x)\big|\bar{\varphi}_k^\perp(x)\right]dx\\
&= E[\hat{f}(X)q^*_\tau(X)\delta b_\tau(X)\delta p_\tau(X)]\\
&\quad + E[\{f^*(X)-\hat{f}(X)\}q^*_\tau(X)\delta b_\tau(X)\delta p_\tau(X)]\\
&\quad + E\left[q^*_\tau(X_1)\delta b_\tau(X_1)K_k(X_1,X_2)\delta p_\tau(X_2)\left\{\frac{q^*_\tau(X_2)}{\hat{q}_\tau(X_2)}-1\right\}\right]\\
&\quad - \int\prod\left[f^*(x)q^*_\tau(x)\delta b_\tau(x)\big|\bar{\varphi}_k^\perp(x)\right]\prod\left[f^*(x)\delta p_\tau(x)\big|\bar{\varphi}_k^\perp(x)\right]dx.
\end{aligned}$$

  The first term equals $\mathrm{Bias}_{1,\mathrm{new}}(\tau;b,p,f)$. The second and third terms are third order. The remaining term is the tail term.

Proof of Lemma 36. The proof is similar to the proof in the i.i.d. case, i.e., Lemma 18. We focus on the loglinear case. We note that

$$E[\mathbb{IF}_{1,\mathrm{new}}(\tau;b,p,f)] = \frac{1}{n}\sum_{i\in\mathrm{split}(1)}\frac{1}{|s(i)|}\sum_{j\in s(i)}E_i[f_j(X_i)\{Y_i-e^{\tau A_i}b(X_i)\}\{A_i-p(X_i)\}]$$

and

$$E_i[f_j(X_i)\{Y_i-e^{\tau A_i}b(X_i)\}\{A_i-p(X_i)\}]$$

$$
\begin{aligned}
&= E_i[f_j(X_i)\{Y_i - e^{\tau A_i} b^*_{i\tau}(X_i) + e^{\tau A_i} b^*_{i\tau}(X_i) - e^{\tau A_i} b(X_i)\}\{A_i - p^*_{i\tau}(X_i) + p^*_{i\tau}(X_i) - p(X_i)\}] \\
&= E_i[f_j(X_i)\{Y_i - e^{\tau A_i} b^*_{i\tau}(X_i)\}\{A_i - p^*_{i\tau}(X_i)\}] + E_i[f_j(X_i) e^{\tau A_i}\{b^*_{i\tau}(X_i) - b(X_i)\}\{A_i - p^*_{i\tau}(X_i)\}] \\
&\quad + E_i[f_j(X_i)\{Y_i - e^{\tau A_i} b^*_{i\tau}(X_i)\}\{p^*_{i\tau}(X_i) - p(X_i)\}] + E_i[f_j(X_i) e^{\tau A_i}\{b^*_{i\tau}(X_i) - b(X_i)\}\{p^*_{i\tau}(X_i) - p(X_i)\}] \\
&= E_i[f_j(X_i)\{Y_i - e^{\tau A_i} b^*_{i\tau}(X_i)\}\{A_i - p^*_{i\tau}(X_i)\}] + E_i[f_j(X_i) q^*_{i\tau}(X_i)\{b^*_{i\tau}(X_i) - b(X_i)\}\{p^*_{i\tau}(X_i) - p(X_i)\}]
\end{aligned}
$$

since the second and third terms are zero. Hence,

$$
\begin{aligned}
&\mathrm{Bias}_{1,\mathrm{new}}(\tau; b, p, f) \\
&= \frac{1}{n} \sum_{i \in \mathrm{split}(1)} \frac{1}{|s(i)|} \sum_{j \in s(i)} E_i[f_j(X_i) q^*_{i\tau}(X_i)\{b^*_{i\tau}(X_i) - b(X_i)\}\{p^*_{i\tau}(X_i) - p(X_i)\}] \\
&= \frac{1}{n} \sum_{i \in \mathrm{split}(1)} \frac{1}{|s(i)|} \sum_{j \in s(i)} E_i[f^*_j(X_i) q^*_{i\tau}(X_i)\{b^*_{i\tau}(X_i) - b(X_i)\}\{p^*_{i\tau}(X_i) - p(X_i)\}] \\
&\quad + \frac{1}{n} \sum_{i \in \mathrm{split}(1)} \frac{1}{|s(i)|} \sum_{j \in s(i)} E_i[q^*_{i\tau}(X_i)\{f_j(X_i) - f^*_j(X_i)\}\{b^*_{i\tau}(X_i) - b(X_i)\}\{p^*_{i\tau}(X_i) - p(X_i)\}].
\end{aligned}
$$

Proof of Theorem 37. Since we assume that the i-th and j-th observation are independent if $j \in s(i)$,

$$
\begin{aligned}
E[\widehat{\mathbb{IF}}^{(k)}_{22}(\tau)] &= -\frac{1}{n} \sum_{i \in \mathrm{split}(1)} \frac{1}{|s(i)|} \sum_{j \in s(i)} E_{ij}[\varepsilon_i(\tau, \hat{b}_{i\tau}) K_k(X_i, X_j) \Delta_j(\tau, \hat{b}_{i\tau}, \hat{q}_{i\tau})] \\
&= -\frac{1}{n} \sum_{i \in \mathrm{split}(1)} \frac{1}{|s(i)|} \sum_{j \in s(i)} E_{ij}\left[E_i\left\{\varepsilon_i(\tau, \hat{b}_{i\tau}) | X_i\right\} K_k(X_i, X_j) E_j\left\{\Delta_j(\tau, \hat{b}_{j\tau}, \hat{q}_{j\tau}) | X_j\right\}\right] \\
&= -\frac{1}{n} \sum_{i \in \mathrm{split}(1)} \frac{1}{|s(i)|} \sum_{j \in s(i)} E_{ij}\left[q^*_{i\tau}(X_i) \delta b_{i\tau}(X_i) K_k(X_i, X_j) \delta p_{j\tau}(X_j) \frac{q^*_{j\tau}(X_j)}{\hat{q}_{j\tau}(X_j)}\right] \\
&= -\frac{1}{n} \sum_{i \in \mathrm{split}(1)} \frac{1}{|s(i)|} \sum_{j \in s(i)} E_{ij}\left[q^*_{i\tau}(X_i) \delta b_{i\tau}(X_i) K_k(X_i, X_j) \delta p_{j\tau}(X_j) \left\{\frac{q^*_{j\tau}(X_j)}{\hat{q}_{j\tau}(X_j)} - 1\right\}\right] \\
&\quad - \frac{1}{n} \sum_{i \in \mathrm{split}(1)} \frac{1}{|s(i)|} \sum_{j \in s(i)} E_{ij}\left[q^*_{i\tau}(X_i) \delta b_{i\tau}(X_i) K_k(X_i, X_j) \delta p_{j\tau}(X_j)\right].
\end{aligned}
$$

Note that for any two functions $\xi, \rho \in \mathscr{L}_2(\lambda)$,

$$
\begin{aligned}
&E_{ij}[\xi(X_i) K_k(X_i, X_j) \rho(X_j)] \\
&= \int f^*_i(x) \xi(x) \rho(x) f^*_j(x) dx - \int \prod [f^*_i(x) \xi(x) | \bar{\varphi}^\perp_k] \prod [\rho(x) f^*_j(x) | \bar{\varphi}^\perp_k] dx \\
&= E_i[f^*_j(X_i) \xi(X_i) \rho(X_i)] - \int \prod [f^*_i(x) \xi(x) | \bar{\varphi}^\perp_k] \prod [\rho(x) f^*_j(x) | \bar{\varphi}^\perp_k] dx.
\end{aligned}
$$

So

$$
\begin{aligned}
&E_{ij}\left[q^*_{i\tau}(X_i) \delta b_{i\tau}(X_i) K_k(X_i, X_j) \delta p_{j\tau}(X_j)\right] \\
&= E_i[f^*_j(X_i) q^*_{i\tau}(X_i) \delta b_{i\tau}(X_i) \delta p_{j\tau}(X_i)] \\
&\quad - \int \prod [f^*_i(x) q^*_{i\tau}(x) \delta b_{i\tau}(x) | \bar{\varphi}^\perp_k] \prod [f^*_j(x) \delta b_{i\tau}(x) | \bar{\varphi}^\perp_k] dx
\end{aligned}
$$

$$= E_i[f_j^*(X_i)q_{i\tau}^*(X_i)\delta b_{i\tau}(X_i)\delta p_{i\tau}(X_i)] + E_i[f_j^*(X_i)q_{i\tau}^*(X_i)\delta b_{i\tau}(X_i)\{\delta p_{j\tau}(X_i) - \delta p_{i\tau}(X_i)\}]$$
$$- \int \prod [f_i^*(x)q_{i\tau}^*(x)\delta b_{i\tau}(x)|\bar{\varphi}_k^{\perp}] \prod [f_j^*(x)\delta b_{i\tau}(x)|\bar{\varphi}_k^{\perp}]dx.$$

Hence

$$E[\widehat{\mathbb{IF}}_{22}^{(k)}(\tau)] + E[\widehat{\mathbb{IF}}_{1,\text{new}}(\tau)]$$
$$= \frac{1}{n}\sum_{i\in \text{split}(1)} \frac{1}{|s(i)|}\sum_{j\in s(i)} \left\{ E_i[q_{i\tau}^*(X_i)\{\hat{f}_j(X_i) - f_j^*(X_i)\}\delta b_{i\tau}(X_i)\delta p_{i\tau}^*(X_i)] \right.$$
$$- E_{ij}\left[q_{i\tau}^*(X_i)\delta b_{i\tau}(X_i)K_k(X_i,X_j)\delta p_{j\tau}(X_j)\left\{ \frac{q_{j\tau}^*(X_j)}{\hat{q}_{j\tau}(X_j)} - 1 \right\} \right]$$
$$- E_i\left[f_j^*(X_i)q_{i\tau}^*(X_i)\delta b_{i\tau}(X_i)\left\{\delta p_{j\tau}(X_i) - \delta p_{i\tau}(X_i)\right\} \right]$$
$$\left. + \int \Pi[f_i^*(x)q_{i\tau}^*(x)\delta b_{i\tau}(x)|\bar{\varphi}_k^{\perp}]\ \Pi[f_j^*(x)\delta b_{i\tau}(x)|\bar{\varphi}_k^{\perp}]dx \right\}.$$

Proof of Theorem 40. Define the symmetrized functions

$$\hat{m}_\tau(O_i,O_j) = \left\{ \text{IF}_{22,ij}^{(k)}(\tau, \hat{b}_{i\tau}, \hat{p}_{j\tau}, \hat{q}_{j\tau}) + \text{IF}_{22,ji}^{(k)}(\tau, \hat{b}_{j\tau}, \hat{p}_{i\tau}, \hat{q}_{i\tau}) \right\}/2$$
$$m_\tau(O_i,O_j) = \left\{ \text{IF}_{22,ij}^{(k)}(\tau, b_{i\tau}^*, p_{j\tau}^*, q_{j\tau}^*) + \text{IF}_{22,ji}^{(k)}(\tau, b_{j\tau}^*, p_{i\tau}^*, q_{i\tau}^*) \right\}/2$$

so

$$\widehat{\mathbb{IF}}_{22}^{(k)}(\tau) = \frac{1}{n}\sum_i \frac{2}{|s(i)|}\sum_{j\in s(i);i<j} \hat{m}_\tau(O_i,O_j), \text{ and}$$
$$\mathbb{IF}_{22}^{(k)}(\tau) = \frac{1}{n}\sum_i \frac{2}{|s(i)|}\sum_{j\in s(i);i<j} m_\tau(O_i,O_j).$$

Define

$$\sigma_b^2(X_i) = E_i[\varepsilon_i^2(\tau^*, b_{i\tau^*}^*)|X_i], \text{ and}$$
$$\sigma_p^2(X_j) = E_j\left[\Delta_{j\tau}^2(p_{j\tau^*}^*, q_{j\tau^*}^*)|X_j\right].$$

Note, under the assumption that the conditional mean model holds, since $E_i\left[\varepsilon_i\left(\tau, b_{i\tau}^*\right)|X_i\right] = E_j\left[\Delta_{j\tau}(p_{j\tau}^*, q_{j\tau}^*)|X_j\right] = 0$ when $j \in s(i)$ [and thus $i \in s(j)$], we have

$$\text{Var}\left[\mathbb{IF}_{22}^{(k)}(\tau^*)\right] = E\left[\left\{ \frac{1}{n}\sum_i \frac{2}{|s(i)|}\sum_{j\in s(i);i<j} m_{\tau^*}(O_i,O_j) \right\}^2 \right]$$
$$= \frac{4}{n^2}E\left[\sum_i \frac{1}{|s(i)|}\sum_{j,j\in\{s(i)\};i<j}\sum_{i'} \frac{1}{|s(i')|}\sum_{j',j'\in s(i');i'<j'} m_{\tau^*}(O_i,O_j)m_{\tau^*}(O_{i'},O_{j'}) \right]$$
$$= \frac{4}{n^2}E\left[\sum_i \frac{1}{|s(i)|^2}\sum_{j=j';j\in s(i);i<j} m_{\tau^*}(O_i,O_j)^2 \right]$$
$$+ \frac{1}{n^2}E\left[\sum_i \frac{1}{|s(i)|^2}\sum_{j\neq j';j,j'\in s(i);i<\min\{j,j'\}} m_{\tau^*}(O_i,O_j)m_{\tau^*}(O_i,O_{j'}) \right]$$

$$+\frac{1}{n^2}E\left[\sum_{i\neq i'}\sum_{j\in s(i);i<j}\sum_{j'\in s(i');i'<j'} m_{\tau^*}(O_i,O_j)m_{\tau^*}(O_{i'},O_{j'})\right],$$

and

$$\begin{aligned}
&\mathrm{Var}\left[\widehat{\mathbb{IF}}_{22}^{(k)}(\tau^*)\right]\\
&=E\left[\left\{\frac{1}{n}\sum_i \frac{2}{|s(i)|}\sum_{j\in s(i);i<j} m_{\tau^*}(O_i,O_j)\right\}^2\right]\\
&=\frac{4}{n^2}E\left[\sum_i \frac{1}{|s(i)|}\sum_{j;j\in s(i);i<j}\sum_{i'}\frac{1}{|s(i')|}\sum_{j';j'\in s(i');i'<j'} m_{\tau^*}\left(O_i,O_j\right)m_{\tau^*}\left(O_{i'},O_{j'}\right)\right]\\
&=\frac{4}{n^2}E\left[\sum_i \frac{1}{|s(i)|^2}\sum_{j=j';j\in\{s(i)\};i<j} m_{\tau^*}\left(O_i,O_j\right)^2\right]\\
&\quad+\frac{1}{n^2}E\left[\sum_i \frac{1}{|s(i)|^2}\sum_{j\neq j';j,j'\in\{s(i)\};i<\min(j,j')} m_{\tau^*}\left(O_i,O_j\right)m_{\tau^*}\left(O_i,O_{j'}\right)\right]\\
&\quad+\frac{1}{n^2}E\left[\sum_{i\neq i'}\sum_{j\in\{s(i)\};i<j}\sum_{j'\in s(i');i'<j'} m_{\tau^*}\left(O_i,O_j\right)m_{\tau^*}\left(O_{i'},O_{j'}\right)\right].
\end{aligned}$$

Under Assumptions C1 and C2, $E_{ij}[m_\tau(O_i,O_j)m_\tau(O_{i'},O_{j'})]=0$ unless $i=i'$ and $j=j'$. Further, $E_{ij}[\varepsilon_i(\tau^*,b^*_{i\tau^*})\Delta_{i\tau}(p^*_{i\tau^*},q^*_{i\tau^*})\Delta_{j\tau}(p^*_{j\tau^*},q^*_{j\tau^*})\varepsilon_j(\tau^*,b^*_{j\tau^*})]=0$. Hence,

$$\begin{aligned}
\mathrm{Var}\left[\mathbb{IF}_{22}^{(k)}(\tau^*)\right]&=\frac{4}{n^2}\sum_i\frac{1}{|s(i)|^2}\sum_{j;j\in s(i);i<j}E_{ij}\left[m_{\tau^*}\left(O_i,O_j\right)^2\right]\\
&=\frac{1}{n^2}\sum_{i=1}^n\frac{1}{|s(i)|^2}\sum_{j\in s(i)}E_{ij}[\varepsilon_i\left(\tau^*,b^*_{i\tau^*}\right)^2K_k(X_i^*,X_j^*)^2\Delta_{j\tau}(p^*_{j\tau^*},q^*_{j\tau^*})^2].
\end{aligned}$$

Hence under our assumptions we can consistently estimate $\mathrm{Var}\left[\widehat{\mathbb{IF}}_{22}^{(k)}(\tau^*)\right]$ by

$$\widehat{\mathbb{V}}\left[\widehat{\mathbb{IF}}_{22}(\hat\tau)\right]=\frac{1}{n^2}\sum_{i=1}^n\frac{1}{|s(i)|^2}\sum_{j\in s(i)}\left\{\mathrm{IF}_{22,ij}(\tau,\hat b_{i\tau},\hat p_{j\tau},\hat q_{j\tau})\right\}^2.$$

Proof of Lemma 41. Recall that

$$\mathrm{Var}\left[\widehat{\mathbb{IF}}_{22}^{(k)}(\tau^*)\right]=\frac{1}{n^2}\sum_{i=1}^n\frac{1}{|s(i)|^2}\sum_{j\in s(i)}E_{ij}[\varepsilon_i(\tau^*,b^*_{i\tau^*})^2K_k(X_i,X_j)^2\Delta_{j\tau}(p^*_{j\tau^*},q^*_{j\tau^*})^2]\{1+o(1)\}.$$

Now defining $\frac{1}{s}=n^{-1}\sum_{i=1}^n\frac{1}{|s(i)|}$ with $|s(i)|$ the cardinality of the set $s(i)$ as before, we can write

$$\begin{aligned}
\mathrm{Var}(\widehat{\mathbb{IF}}_{22}^{(k)})&=\frac{k}{ns}C_{22}\left[1+o_p(1)\right],\\
C_{22}&=O(1),
\end{aligned}$$

where

$$\xi_i=\frac{1}{|s(i)|}/\frac{1}{s},\qquad C_{22}=n^{-1}\sum_{i=1}^n\xi_i r_i,$$

$$r_i = \frac{1}{|s(i)|k} \sum_{j \in s(i)} E_{ij}[\varepsilon_i(\tau^*, b^*_{i\tau^*})^2 K_k(X_i, X_j)^2 \Delta_{j\tau^*}(p^*_{j\tau^*}, q^*_{j\tau^*})^2].$$

Note that $\xi_i = O(1)$. Under our assumptions,

$$\begin{aligned}
&E_{ij}[\varepsilon_i(\tau^*, b^*_{i\tau^*})^2 K_k(X_i, X_j)^2 \Delta_{j\tau^*}(p^*_{j\tau^*}, q^*_{j\tau^*})^2] \\
&= E_{ij}\left\{E_i\left[\varepsilon_i(\tau^*, b^*_{i\tau^*})^2 | A_i, X_i\right] K_k(X_i, X_j)^2 E_j\left[\Delta_{j\tau^*}(p^*_{j\tau^*}, q^*_{j\tau^*})^2 | A_j, X_j\right]\right\} \\
&\le \left\|E_i\left[\varepsilon_i(\tau^*, b^*_{i\tau^*})^2 | A_i, X_i\right]\right\|_\infty E_{ij}\left[K_k(X_i, X_j)^2\right] \left\|E_j\left[\Delta_{j\tau^*}(p^*_{j\tau^*}, q^*_{j\tau^*})^2 | A_j, X_j\right]\right\|_\infty \\
&= \left\|E_i\left[\varepsilon_i(\tau^*, b^*_{i\tau^*})^2 | A_i, X_i\right]\right\|_\infty E_{ij}\left[K_k(X_i, X_j)^2\right] \left\|E_j\left[\Delta_{j\tau^*}(p^*_{j\tau^*}, q^*_{j\tau^*})^2 | A_j, X_j\right]\right\|_\infty \\
&\le k\left\|E_i\left[\varepsilon_i(\tau^*, b^*_{i\tau^*})^2 | A_i, X_i\right]\right\|_\infty \left\|E_j\left[\Delta_{j\tau^*}(p^*_{j\tau^*}, q^*_{j\tau^*})^2 | A_j, X_j\right]\right\|_\infty \|f_i\|_\infty \|f_j\|_\infty
\end{aligned}$$

since

$$\begin{aligned}
E_{ij}\left[K_k(X_i, X_j)^2\right] &= \int\int f_i(X_i) K_k(X_i, X_j)^2 f_j(X_j) dX_i dX_j \\
&\le \|f_i\|_\infty \|f_j\|_\infty \int\int K_k(X_i, X_j)^2 dX_i dX_j
\end{aligned}$$

and, for any orthonormal basis, $\int\int K_k(X_i, X_j)^2 dX_i dX_j = k$. Thus under the assumption that all the above infinity norms are uniformly bounded, we have that $\mathrm{Var}\left[\widehat{\mathbb{IF}}^{(k)}_{2,2}(\hat{\tau})\right]$ is $O(k)$. In fact, one can show that $C'_l k \le \mathrm{Var}\left[\widehat{\mathbb{IF}}^{(k)}_{22}(\hat{\tau})\right] \le C'_u k$ for nonzero positive constants C'_l and C'_u.

Hence $r_i = O(1)$, and $C_{22} = O(1)$ and the theorem holds.

BIBLIOGRAPHY

Bickel, PJ, Klaassen, CAJ, Ritov, Y and Wellner, JA. 1993. *Efficient and Adaptive Estimation for Semiparametric Models*. The Johns Hopkins University Press, Baltimore, MD.

Dominici, F, McDermott, A and Hastie, TJ. 2004. Improved semiparametric time series models of air pollution and mortality. *J Am Stat Assoc* 99(468): 938–948.

Donald, SG and Newey, WK. 1994. Series estimation of semilinear models. *J Multivariate Anal* 50: 30–40.

Friedman, JH. 2001. Multivariate adaptive regression splines. *Annals of Statistics* 19(1): 1–67.

Friedman, JH. 2008. *Fast Sparse Regression and Classification*. Technical Report. Department of Statistics, Stanford University, Stanford, CA.

Lawson, CL and Hanson, RJ. 1974. *Solving Least Squares Problems*. Prentice-Hall, Englewood Cliffs, NJ.

Liu, H, Lafferty, J and Wasserman, L. 2009. The Nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *JMLR* 10: 2295–2328.

Loader, C. 1999. *Local Regression and Likelihood*. Springer, New York, NY.

Loader, C. 2010. *Locfit: Local Regression, Likelihood and Density Estimation.* R package version 1.5-6. Available at *http://cran.r-project.org/web/packages/locfit/locfit.pdf*.

Lorentz, GG. 1986. *Approximation of Functions*. American Mathematical Society, AMS Chelsea Publishing, Providence, RI.

Peng, RD, Welty, LJ and McDermott, A. 2004. *The National Morbidity, Mortality, and Air Pollution Study Database in R*. Technical Report. Department of Biostatistics, The Johns Hopkins University, Baltimore, MD.

Robins, JM, Li, L, Tchetgen, EJT and van der Vaart, A. 2008. Higher order influence functions and minimax estimation of nonlinear functionals. In: Probability and Statistics: Essays in Honor of David A. Freedman (Nolan D, Speed T, eds.; pp 335–421). Institute of Mathematical Statistics, Beachwood, OH.

Robins, JM, Tchetgen, EJT, Li, L and van der Vaart, A. 2009. Semiparametric minimax rates. *Electronic Journal of Statistics* 3: 1305–1321. Available at *http://imstat.org/ejs/*.

Samet, JM, Dominici, F, Zeger, SL, Schwartz, J and Dockery, DW. 2000. *The National Morbidity, Mortality, and Air Pollution Study, Part I: Methods and Methodologic Issues*. Research Report 94. Health Effects Institute, Cambridge, MA.

Samet, JM, Zeger, SL, Dominici, F, Curriero, F, Coursac, I, Dockery, DM, Schwartz, J and Zanobetti, A. 2000. *The National Morbidity, Mortality, and Air Pollution Study, Part II: Morbidity and Mortality from Air Pollution in the United States*. Research Report 94. Health Effects Institute, Cambridge MA.

Schumaker, LL. 1981. *Spline Functions: Basic Theory*. Wiley, New York.

van der Vaart, AW. 1998. *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, AW and Wellner, JA. 1996. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, NY.

Vidakovic, B. 1999. *Statistical Modeling by Wavelets*. Wiley, New York.

Wand, M and Jones, C. 1995. *Kernel Smoothing*. Chapman & Hall, CRC Press, Boca Raton, FL.

Wand, M and Ripley, B. 2009. *KernSmooth: Functions for Kernel Smoothing for Wand & Jones (1995)*. Matt Wand R port and updates by Brian Ripley. Available at *http://CRAN.R-project.org/package=KernSmooth*.

Wasserman, L. 2006. *All of Nonparametric Statistics*. Springer Science+Business Media, New York, NY.

NOTATION

N	Sample size
Y_i	Outcome; number of deaths
A_i	Exposure to airborne particulate matter (PM_{10})
X_i	Covariates, including temperature, dew point, and age category
t_i	Time (in days or multiples of 6 days)
E_i	Expectation with respect to the distribution for the i-th subject; the data are not assumed to be i.i.d.
τ^*	The true effect of exposure, on either the linear or log scale
ζ_i^*	An unknown infinite-dimensional nuisance parameter, equal to the conditional expected number of deaths given A_i and X_i when A_i equals 0
w_i	A function of time and covariates used to model ζ_i^*
b_i^*	An unknown infinite-dimensional nuisance parameter
p_i^*	An unknown infinite-dimensional nuisance parameter
$\hat{\tau}_{1,\text{eff}}$	Estimate of the effect of exposure based on the efficient influence function
$\hat{\tau}_{1,\text{new}}$	Estimate of the effect of exposure based on the density-modified influence function
$\hat{\tau}_{1,\text{eff}}^{\text{full}}$	Equal to $\hat{\tau}_{1,\text{eff}}$
$\hat{\tau}_{1,\text{eff}}^{\text{split}}$	Estimate of the effect of exposure based on the efficient influence function, but using sample splitting
$\hat{\tau}_2^{\text{split},(k)}$	Second-order estimate of the effect of exposure based on the second-order influence function, using sample splitting, and based on k basis functions
$\widehat{\mathbb{IF}}_{1,\text{eff}}$	The first-order efficient influence function, with estimates substituted for nuisance parameters
$\widehat{\mathbb{IF}}_{1,\text{new}}$	The density-weighted first-order influence function, with estimated nuisance parameters and density
$\widehat{\mathbb{IF}}_{1,\text{eff}}^{\text{full}}$	Same as $\widehat{\mathbb{IF}}_{1,\text{eff}}$
$\widehat{\mathbb{IF}}_{1,\text{eff}}^{\text{split}}$	The first-order efficient influence function based on sample splitting, with estimated nuisance parameters
$\widehat{\mathbb{IF}}_{1,\text{new}}^{\text{full}}$	Same as $\widehat{\mathbb{IF}}_{1,\text{new}}$
$\widehat{\mathbb{IF}}_{1,\text{new}}^{\text{split}}$	The density-weighted first-order efficient influence function based on sample splitting, with estimated nuisance parameters and density

$\widehat{\mathbb{IF}}_2^{\text{split},(k)}$	The second-order influence function with estimated nuisance parameters, using sample splitting, and based on k basis functions
$\widehat{\mathbb{IF}}_{22}^{\text{split},(k)}$	Equals $\widehat{\mathbb{IF}}_2^{\text{split},(k)} - \widehat{\mathbb{IF}}_{1,\text{new}}^{\text{split}}$.
k	Number of basis functions used in the second-order influence function
$k_{\max}$	The largest k considered
$\widehat{\text{se}}_k$	Estimated standard error of the second-order estimate $\hat{\tau}_2^{\text{split},(k)}$ based on k basis functions.
f_i	Joint density of the covariates X_i
$\hat{f}_i$	Estimate of joint density of the covariates X_i
φ_l	Basis functions used in calculating the second-order influence functions
K_k	Kernel used in calculating the second-order influence functions
$s(i)$	A subset of other observations corresponding to observation i, used in calculating the second-order influence functions
$\lvert s(i)\rvert$	Size of the set $s(i)$
$X_{\text{cont},i}$	The continuous components of the covariates X_i
$X_{5,i}$	Age category, the only discontinuous component of the covariates X_i
O_i	All observed variables for the i-th observation, equal to (Y_i, A_i, X_i).
V_1	Variance of $\hat{\tau}_{1,\text{eff}}^{\text{split}}$
$V_2^{(k)}$	Variance of $\hat{\tau}_2^{\text{split},(k)}$; also denoted by V_2
q_τ^*	A function arising in the definition of the first-order influence functions
$q(x;\omega)$	A function arising in the definition of the first-order influence function
$b(x;\eta)$	A function arising in the definition of the first-order influence function
$\Delta_i(\tau, p, q)$	A function arising in the definition of the first-order influence function, defined in terms of generic functions p, q
$S_i(\tau, \alpha)$	A function arising in the definition of the first-order influence function
$U_{i,\text{profile}}(\tau, b)$	A function arising in the definition of the first-order influence function, defined in terms of a generic function b
$U_{i,\text{nuis}}(\tau, b)$	A function arising in the definition of the first-order influence function, defined in terms of a generic function b
$U_i(\tau, b)$	A function arising in the definition of the first-order influence function, defined in terms of a generic function b
$\varepsilon_i(\tau, b)$	A function arising in the definition of the first-order influence function, defined in terms of a generic function b
η	A vector arising in the definition of the first-order influence function
split(0)	A subset of the data; one subset is used as a training sample, and its complement is used as a testing sample in the split-sample estimators
split(1)	A subset of the data; one subset is used as a training sample, and its complement is used as a testing sample in the split-sample estimators
$\widehat{\mathbb{V}}_{1,\text{eff}}^{\text{full}}$	Variance of $\hat{\tau}_{1,\text{eff}}^{\text{full}}$

$\widehat{\mathbb{V}}_{1,\text{eff}}^{\text{split}}$	Variance of $\hat{\tau}_{1,\text{eff}}^{\text{split}}$
$\text{DER}_{1,\text{eff},i}(\tau, b, p)$	Derivative of $\widehat{\mathbb{IF}}_{1,\text{eff}}^{\text{full}}$
$H(\beta_h, C_h)$	Hölder class of functions, a set of functions with at least $\beta_h > 0$ derivatives
C_h	Radius of the Hölder class $H(\beta_h, C_h)$
β_h	Hölder exponent of the Hölder class $H(\beta_h, C_h)$
d	Dimensionality of the argument of the nuisance parameters (equal to number of components in X_i)
$\text{Bias}_{1,\text{new}}(\tau, b, p, f)$	Bias of $\mathbb{IF}_{1,\text{new}}(\tau; b, p, f)$ as an estimator of $E[\mathbb{IF}_{1,\text{new}}(\tau; b_\tau^*, p_\tau^*, f)]$
$\text{Bias}_{1,\text{eff}}(\tau; b, p)$	Bias of $\mathbb{IF}_{1,\text{eff}}(\tau; b, p)$ as an estimator of $E[\mathbb{IF}_{1,\text{eff}}(\tau; b_\tau^*, p_\tau^*)]$
$K_{f,k}$	Density-weighted kernel used to define the density-weighted second-order influence functions
$K_{f,q,k,\text{alt}}$	Alternative kernel used to define alternative second-order influence functions
$\mathscr{L}_2(\mu)$	Hilbert space of square-integrable functions with respect to the measure μ
$\mathscr{L}_2(f)$	Hilbert space of square-integrable functions, with inner product defined by $\langle g, h \rangle = E_f[g(X)h(X)]$
Π_f	Projection operator in the Hilbert space $\mathscr{L}_2(f)$
$\text{EB}_{q,b,p}^{(3)}$	A third-order component of the bias of the second-order influence function
$\text{EB}_{f,b,p}^{(3)}$	A third-order component of the bias of the second-order influence function
$\hat{m}_\tau^{(\ell)}(O_i, O_j)$	Symmetrized function of data points i and j used to define the variance of the second-order influence function, based on split-sample estimators of the functions b, p, and q
$m_\tau(O_i, O_j)$	Symmetrized function of data points i and j used to define the variance of the second-order influence function, based on the true functions b_τ^*, p_τ^*, and q_τ^*
$E_{ii'}$	Expectation with respect to the densities at observation points i and i'
id	A transformation on the continuous covariates, scaling each separately to the interval $[0, 1]$
gs	A Gram-Schmidt transformation of the continuous covariates in addition to scaling
gs2	A Gram-Schmidt transformation of the continuous covariates after scaling and removal of seasonality and time effects

DEFINITIONS

Donsker Conditions and Classes Consider a class of functions of observed independent and identically distributed (i.i.d.) data, and consider the problem of estimating the means of all these functions simultaneously. The class is said to be Donsker if the estimated mean of the functions is consistent and asymptotically normal simultaneously for all functions in the class. Such a class is called a Donsker class of functions.

Whether or not a class of functions satisfies the Donsker condition depends on the complexity of the class of functions. Here, complexity is a measure of the size or "richness" of the class. Classes that are too rich are too demanding and cannot be estimated simultaneously (i.e., will not satisfy the Donsker conditions). For example, estimating equations in which some components belong to Hölder classes with low exponents may be too rich to be Donsker.

Further details can be found in van der Vaart and Wellner (1996).

Gram-Schmidt Orthogonalization A procedure common in linear algebra, the Gram-Schmidt orthogonalization is a method of constructing an orthogonal basis for a given vector space, starting with any basis for that vector space. Given a basis set (i.e., a minimal set of vectors whose linear combinations span the vector space), the orthogonalization begins by placing the first vector in the new basis set. From the second vector on, it proceeds by identifying the component of each vector that is perpendicular to all vectors in the new basis set and adding that component to the new basis set. After all the vectors in the original basis have been processed this way, the new basis set is now an orthogonal basis with the same linear span as the original, possibly nonorthogonal basis. That is, all the vectors in the new basis are perpendicular to each other.

Higher-Order Influence Function A higher-order influence function (a higher-order U-statistic) is a generalization of an influence function to estimate problems where no first-order U-statistic can achieve asymptotic normality. A higher-order influence function can be viewed as an adjusted first-order influence function with an estimator of the bias of the first-order influence function subtracted. Further details can be found in Robins, Li, et al. (2008).

Hölder Class A Hölder class is a way to specify a collection of real values for functions (of many variables) that satisfy a specific smoothness. The smoothness is specified by a Hölder exponent β, which is a generalization of the notion of number of derivatives; the exponent can be any positive number. A function in a Hölder class with exponent β has at least $\lfloor\beta\rfloor$ derivatives and must satisfy an additional "residual smoothness" condition for the remaining fractional part of β. For a complete definition, see Section 3.2.2.

The Hölder exponent of a class is also a measure of its complexity. Lower Hölder exponents correspond to higher complexity (more roughness) and make estimation harder.

Influence Function Many estimators of parameters in a large class of models have the property that they can be expressed asymptotically as the average of a function of the individual data points. Consider an asymptotically normal estimator $\hat{\tau}$ of a parameter τ in a model. If there exists an IF_τ depending on the data and the true distribution satisfying $\sqrt{n}(\hat{\tau}-\tau) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{IF}_\tau(O_i) + o_P(1)$, then IF_τ is called a first-order influence function, or simply an influence function.

For some parameters in some models there are no estimators that satisfy the above asymptotic display. In such models, second- or higher-order influence functions need to be considered.

U-Statistic A U-statistic is a generalization of the mean that arises naturally in a number of statistical contexts. Given a set of variables $x_1,\ldots,x_n$ and any function $f_1(x)$, the sample mean of $f_1(z)$ is defined by $\sum_{i=1}^{n} f(x_i)$. Given a function $f_2(z_1,z_2)$ of two variables rather than one, the corresponding second-order U-statistic is the average of $f_2(x_i,x_j)$ for all pairs (x_i,x_j) taken from the sample

$x_1,\ldots,x_n$. For any $k \le n$, a k-th order U-statistic is similarly an average of a function of k variables over all k-tuples from the sample.

ABBREVIATIONS AND OTHER TERMS

ACF	autocorrelation function
CI	confidence interval
df	degrees of freedom
EB	estimation bias
GLM	generalized linear model
GPS	generalized penalized spline [regression]
gs	Gram-Schmidt (orthogonalization transformation)
gs2	a variation on gs
id	identity transform
i.i.d.	independent and identically distributed [sample]
LLM	loglinear model
LM	linear model
MARS	multivariate adaptive regression splines
MSE	mean squared error
NMMAPS	National Morbidity, Mortality, and Air Pollution Study
NPN	nonparanormal (density estimator)
PM_{10}	particulate matter ≤ 10 μm in aerodynamic diameter

ACKNOWLEDGMENTS

We thank Aaron Cohen, Principal Scientist at HEI, for his patience, guidance, and support.

ABOUT THE AUTHORS

James M. Robins is Professor of Epidemiology and Biostatistics at the Harvard School of Public Health, Boston, Massachusetts.

Peng Zhang is Research Assistant Professor in the Department of Surgery at the University of Michigan, Ann Arbor, Michigan.

Rajeev Ayyagari is Associate at Analysis Group, Inc., Boston, Massachusetts.

Roger Logan is Research Scientist in the Department of Epidemiology at the Harvard School of Public Health, Boston, Massachusetts.

Eric Tchetgen Tchetgen is Associate Professor in the Departments of Biostatistics and Epidemiology at the Harvard School of Public Health, Boston, Massachusetts.

Lingling Li is an Assistant Professor and Biostatistician in the Department of Population Medicine at Harvard Medical School, Boston, Massachusetts.

Thomas Lumley is Professor in the Department of Statistics at the University of Auckland, New Zealand.

Aad van der Vaart is Professor in the Mathematical Institute, Faculty of Science, Leiden University, The Netherlands.

CRITIQUE
Health Review Committee
HEI

Research Report 175, *New Statistical Approaches to Semiparametric Regression with Application to Air Pollution Research*, J.M. Robins et al.

BACKGROUND

The findings of a number of epidemiologic studies of air pollution have played a central role in setting air quality limits aimed at protecting public health. Since the mid-1990s, HEI has sponsored original research in this area, as well as research and review activities focused on the analytic methods used in such studies. These efforts include — among many others — The National Morbidity, Mortality, and Air Pollution Study (NMMAPS; Samet et al. 2000), the Reanalysis of the Harvard Six Cities Study and American Cancer Society Study of Particulate Air Pollution and Mortality (Krewski et al. 2000), and the HEI Special Report on Revised Analyses of Time-Series Studies (HEI 2003). In addition, HEI sponsored further analyses of the American Cancer Society study data (Krewski et al. 2009) and several multi-city time-series studies such as Air Pollution and Health: A European and North American Approach (APHENA) (Katsouyanni and Samet et al. 2009); the Public Health and Air Pollution in Asia (PAPA) studies (2010); and the Multicity Study of Air Pollution and Mortality in Latin America (ESCALA) project (Romieu et al. 2012). These studies proceeded from methodologic advances or employed careful investigation of the impact of different statistical methods and parameters used to control various types of biases and confounding. Such methodologic work has helped improve scientific understanding of the relationships between air pollution exposure and important public health outcomes.

Time-series studies are commonly used to evaluate relationships between variations in short-term pollutant concentrations and acute human disease outcomes or mortality. Because time-series methods compare counts of disease events or deaths with pollutant concentrations on a specific day (or another short time frame), the analyses do not need to account for subjects' smoking behavior or other risk factors that do not change from day to day. However, investigators do need to systematically adjust the data sets to control for weather and time-dependent phenomena that might also acutely influence mortality. Of particular concern are factors that vary over time and may be related to pollutant concentrations yet are independently connected to disease or mortality (such as seasonal trends). Therefore, time-series study designs need to control for these and other sources of variation over time when evaluating the relationships between health outcomes and pollutant exposures.

Efforts to optimize control for time-related trends that influence both disease and pollutant concentrations in time-series studies have been incorporated in many HEI-sponsored time-series studies. The NMMAPS, APHENA, PAPA, and ESCALA studies all empirically investigated various methods and parameters that might control various types of time trends in the data that could confound their results. These efforts included comparisons of the effects of using different types of functional methods (e.g., natural splines) and different parameters for these methods (e.g., number of degrees of freedom) in order to optimize control while minimizing the loss of information in the data.

In 2003, HEI issued a Special Report on the Revised Analyses of Time-Series Studies of Air Pollution and Health (HEI 2003). This reanalysis was conducted because, in May, 2002, investigators at Johns Hopkins University discovered that part of the programming in the S-Plus statistical software, which they and many others had used to fit generalized additive models for time-series data, was inappropriately configured to analyze such data and may have produced spurious results (Dominici et al. 2002). Although the special report was largely focused on reanalyzing data from selected studies and from NMMAPS, a Special Panel of the HEI Health Review Committee made a number of recommendations for future research. They emphasized that the effect estimates for particulate matter (PM) that were derived from time-series data were shown, in the reanalysis project, to be sensitive to the statistical methods and parameters used to control long-term time trends in the data (HEI 2003). The Panel stated, "In general, the original PM effect estimates were more sensitive

Dr. Robins' 3-year study, "New Statistical Approaches to Semiparametric Regression with Application to Air Pollution Research", began in July 2005. Total expenditures were $686,620. The draft Investigators' Report from Robins and colleagues was received for review in August 2010. A revised report, received in June 2011, was accepted for publication at that time. During the review process, the HEI Health Review Committee and the investigators had the opportunity to exchange comments and to clarify issues in both the Investigators' Report and the Review Committee's Critique.

to the method used to account for temporal effects than to changing the convergence criteria. Further, … many estimates of effect were more sensitive to the degree of smoothing of temporal effects than either the use of stricter convergence criteria or the method used to account for temporal effects." They also noted that "Neither the appropriate degree of control for time in these time-series analyses nor the appropriate specification of the effects of weather has been determined. This awareness introduces an element of uncertainty into the time-series studies that has not been widely appreciated previously."

Following publication of that special report, Dr. James Robins, of the Harvard School of Public Health, and his colleagues submitted a preliminary application to HEI under RFPA 04-3 *Health Effects of Air Pollution*. They proposed to further develop and apply statistical methods to address some of the analytic issues with time-series data posed by the 2003 HEI Special Report. They specifically proposed to develop user-friendly software to reanalyze the NMMAPS data set using specially derived semiparametric regression models. These semiparametric models, the team suggested, would provide better control of time-varying confounding in time-series data than would splines, for example, because they operate with fewer assumptions about the form of the time trends in the data. By comparing results from these new methods to those from the earlier NMMAPS analyses, they reasoned that this work would provide some reassurance if the results were similar.

This Critique is intended to aid the sponsors of HEI and the public by highlighting both the strengths and limitations of the study and by placing the Investigators' Report into scientific and regulatory perspective.

SPECIFIC AIMS

The study by Robins and his colleagues included four major specific aims:

1. To further develop statistical methods that would provide improved point estimates and confidence intervals for the parameters of a semiparametric regression model that would encode the effect of particulates (with or without other pollutants) on mortality and morbidity based on either time-series or cohort data;
2. To compare the new methods with standard methods in simulated studies;
3. To develop efficient, user-friendly software to implement the new methods; and
4. To reanalyze critical data sets, both time-series and cohort data, using the new methods and compare the results with those from other methods based on, for example, case–crossover, generalized linear models, natural splines, generalized additive models, and penalized splines.

Because of difficulties encountered during the research, such as limits of computational resources, unanticipated obstacles with programming, and barriers to accessing the cohort data sets, the project ultimately focused on only time-series studies and the NMMAPS data set (the same data set used in the earlier investigations described above). These restrictions and the subsequent modifications to the work plans were discussed and approved by the Research Committee.

COMMENTS FROM THE HEALTH REVIEW COMMITTEE

Reviews of the Investigators' Report, from committee members and selected external peers, were divergent on the overall importance and utility of this work for epidemiologic analyses. Some reviewers believed that the statistical methods developed by the investigators were complete, and that what they presented in this report was scientifically sound even though they were unable to fully complete all of the initial objectives. Other reviewers did not see any clear advantages in using these methods over those currently in use for time-series analyses except in a limited number of circumstances.

The Committee commented that the investigators had performed high-quality work in developing statistical methods that are complex and represent a very significant effort on their part, and that the results are technically broadly sound. The Committee understood how the semiparametric approach, as developed by Robins and colleagues, could provide an informative alternative to commonly used methods for time-series analyses with regard to choosing smoothness of functions for controlling time-dependent confounding issues such as weather, seasons, and more city-specific time-related biases. However, the Committee noted that these methods did not address imperfect control for time-dependent acute risk factors other than through the smoothness of the response functions. These include the lag structure and duration effects of cold-wave and heat-wave impacts on mortality. They did note that the concurrence between the results from the current investigators' and the earlier NMMAPS results was reassuring.

INVITED EDITORIAL

The work of the Robins team, although acknowledged to be highly innovative by the broader scientific community, can be most easily understood by experts who are immersed in this particular area of statistical methods development. Therefore, the Review Committee invited Dr. Sander Greenland, Professor of Statistics and Epidemiology at the University of California–Los Angeles, to write a short editorial to be published with the report. Dr. Greenland is known for his writings and explanations of complex statistical issues.

Dr. Greenland's editorial, which follows this Critique, is provided to assist the reader with understanding and interpreting this report and its contributions to epidemiologic methods in air pollution research. Dr. Greenland's views do not necessarily reflect the opinions of the Review Committee.

The Committee did, however, agree with most of the points that Dr. Greenland made. First, even in the largest and best-conducted observational studies, errors in the measurement of pollutant concentrations and major potential confounding factors create uncertainty in the magnitude, if not the direction, of estimated effects. Second, prior expert knowledge, to the extent that it exists, can be a valuable tool to inform the design and the interpretation of research. These are points that apply in general to observational epidemiology and have been made by many authors in the literature (see, for example, Goodman and Greenland, 2007, and Goldman et al. 2011).

The Committee noted that HEI has funded extensive research that has provided some of the evidence that short-term increases in air pollutant concentrations are associated with increased daily mortality. These efforts, now augmented by Robins and his colleagues, have also demonstrated that the magnitude of such estimated effects is sensitive to alternative analytic approaches and to the effects of measurement error in pollutant concentrations and potential confounders (HEI 2003; Dominici 2004; Katsouyanni and Samet et al. 2009; Sarnat et al. 2010; Flanders et al. 2011).

The Committee also agreed with Dr. Greenland that the research team could have taken better advantage of the extensive prior knowledge to set narrower plausible limits on the complexity of potential time-varying confounding, although one could argue that removing such trends from the data before analysis would have limited the team's ability to fully test the response of their methods to these time-varying confounders. However, the Committee also noted (as did Dr. Greenland) that investigating the extent and impact of many of these limitations was beyond the scope of this research project. Therefore, these comments are provided as a guide for further research into and application of these methods. Furthermore, environmental epidemiologists conduct research in a world of imperfect data, and no analytic method will ever be able to perfectly adjust for the shortcomings of such observational information as that found in the NMMAPS data set and others that have been established for large-scale air pollution and health outcomes research.

SUMMARY AND CONCLUSIONS

Through this research project, Robins and colleagues have successfully developed semiparametric methods for epidemiologic investigations that are likely to produce risk estimates that are less biased than traditional Poisson time-series methods because they do not rely on rigid assumptions regarding the relationships between potentially confounding covariates and the outcomes of interest. These semiparametric methods, when applied to the NMMAPS data set used in previous investigations, produced estimates of the risk of health events relative to pollutant levels that were of similar magnitude to those obtained in HEI's Revised Analyses of Time-Series Studies (2003). The 95% confidence intervals were wider for estimates calculated using this team's semiparametric methods because the relaxed assumptions about the relationships between time-varying confounders and the health outcomes resulted in a greater range of uncertainty.

Overall, the Review Committee found that the semiparametric methods developed in this study are a promising addition to current practices for short-term studies of health events and air pollution levels in that they provide a means of analysis that does not rely on some important a priori assumptions about the data that may not be valid. Although these methods do not address all of the potentially important sources of bias or confounding that could complicate analyses of health and air pollution data, such as exposure measurement error, the Committee agreed with the investigators that this research could be particularly useful in studies in which the relationships between time-varying confounders and health outcomes are not clearly understood or are difficult to characterize. However, the Committee also agreed with Dr. Greenland's suggestion that the applicability of this team's methods and the precision of the estimates of risk that they produce could be improved in practice by combining them with some of the methods currently used in time-series analyses to adjust for time-varying confounders when relationships between covariates and mortality are well understood.

ACKNOWLEDGMENTS

The Health Review Committee thanks the ad hoc reviewers for their help in evaluating the scientific merit of the Investigators' Report. The Committee is also grateful to Aaron Cohen for his oversight of the study, to Dr. Sumi Mehta for her oversight of the early review process, to Dr. Kate Adams for her assistance in preparing its Critique, to Virgi Hepner for science editing this Report and its Critique, and to Carol Moyer, Ruth Shaw, Fred Howe, and William Adams for their roles in preparing this Research Report for publication.

REFERENCES

Daniels MJ, Dominici F, Zeger SL, Samet JM. 2004. Part III. PM_{10} concentration–response curves and thresholds for the 20 largest U.S. cities. In: The National Morbidity, Mortality, and Air Pollution Study. Research Report 94. Health Effects Institute, Boston, MA.

Dominici F. 2004. Time-Series Analysis of Air Pollution and Mortality: A Statistical Review. Research Report 123. Health Effects Institute, Boston, MA.

Dominici F, McDermott A, Zeger SL, Samet JM. 2002. On the use of generalized additive models in time-series studies of air pollution and health. Am J Epidemiol 156:193–203. (doi: 10.1093/aje/kwf062).

Dominici F, Zanobetti A, Zeger SL, Schwartz J, Samet J. 2005. Part IV. Hierarchical bivariate time-series models — A combined analysis of PM_{10} effects on hospitalization and mortality. In: The National Morbidity, Mortality, and Air Pollution Study. Research Report 94. Health Effects Institute, Boston, MA.

Flanders WD, Klein M, Darrow LA, Strickland MJ, Sarnat SE, Sarnat JA, Waller LA, Winquist A, Tolbert PE. 2011. A method for detection of residual confounding in time-series and other observational studies. Epidemiology 22:59–67.

Goldman GT, Mulholland JA, Russell AG, Strickland MJ, Klein M, Waller LA, Tolbert PE. 2011. Impact of exposure measurement error in air pollution epidemiology: Effect of error type in time-series studies. Environ Health 10:61. (doi: 10.1186/1476-069X-10-61).

Goodman SN, Greenland S. 2007. Why most published research findings are false: Problems in the analysis. PLoS Med 4:e168.

Health Effects Institute. 2003. Revised Analyses of Time-Series Studies of Air Pollution and Health. Special Report. Health Effects Institute, Boston, MA.

HEI Public Health and Air Pollution in Asia. 2010. Public Health and Air Pollution in Asia (PAPA): Coordinated Studies of Short-Term Exposure to Air Pollution and Daily Mortality in Four Cities. Research Report 154. Health Effects Institute, Boston, MA.

Katsouyanni K, Samet J, Anderson HR, Atkinson R, Le Tertre A, Medina S, Samoli E, Touloumi G, Burnett RT, Krewski D, Ramsay T, Dominici F, Peng RD, Schwartz J, Zanobetti A. 2009. Air Pollution and Health: A European and North American Approach (APHENA). Research Report 142. Health Effects Institute, Boston, MA.

Krewski D, Burnett RT, Goldberg MS, Hoover K, Siemiatycki J, Jerrett M, Abrahamowicz M, White WH. 2000. Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality. A Special Report of the Institute's Particle Epidemiology Reanalysis Project. Health Effects Institute, Cambridge, MA.

Krewski D, Jerrett M, Burnett RT, Ma R, Hughes E, Shi Y, Turner MC, Pope CA III, Thurston G, Calle EE, Thun MJ. 2009. Extended Follow-Up and Spatial Analysis of the American Cancer Society Study Linking Particulate Air Pollution and Mortality. Research Report 140. Health Effects Institute, Boston, MA.

Romieu I, Gouveia N, Cifuentes LA, Ponce de Leon A, Junger W, Hurtado-Díaz M, Miranda-Soberanis V, Vera J, Strappa V, Rojas-Bracho L, Carbajal-Arroyo L, Tzintzun-Cervantes G. 2012. Multicity Study of Air Pollution and Mortality in Latin America (the ESCALA Study). Research Report 171. Health Effects Institute, Boston, MA.

Samet JM, Dominici F, Zeger SL, Schwartz J, Dockery DW. 2000a. Part I. Methods and methodologic issues. In: The National Morbidity, Mortality, and Air Pollution Study. Research Report 94. Health Effects Institute, Cambridge, MA.

Samet JM, Zeger SL, Dominici F, Curriero F, Coursac I, Dockery DW, Schwartz J, Zanobetti A. 2000b. Part II. Morbidity and mortality from air pollution in the United States. In: The National Morbidity, Mortality, and Air Pollution Study. Research Report 94. Health Effects Institute, Cambridge, MA.

Sarnat SE, Klein M, Sarnat JA, Flanders WD, Waller LA, Mulholland JA, Russell AG, Tolbert PE. 2010. An examination of exposure measurement error from air pollutant spatial variability in time-series studies. J Expo Sci Environ Epidemiol 20:135–146.

INVITED EDITORIAL

Sander Greenland, D.P.H.

Research Report 175, *New Statistical Approaches to Semiparametric Regression with Application to Air Pollution Research,* J.M. Robins et al.

INTRODUCTION

In this research project, Robins and colleagues have reanalyzed the effect of PM_{10} (particulate matter ≤ 10 μg in aerodynamic diameter) on all-cause mortality using the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) city-specific time-series data for the 22 largest NMMAPS cities (Health Effects Institute 2003). The authors used a new approach based on higher-order influence-function estimators they had developed (Robins et al. 2008). These estimators may, under certain conditions, offer better control of bias due to confounding by temperature and humidity than the log-linear Poisson estimators used in earlier NMMAPS analyses. My goal here is to describe the strengths and weaknesses of the Investigators' Report and what it may or may not add to our knowledge of air pollution effects.

The methods the investigators developed use very flexible data-adaptive models to estimate the regressions of mortality and PM_{10} levels on covariates (e.g., temperature and humidity); these regressions are then incorporated into a semiparametric loglinear estimator for the effects of PM_{10} on all-cause mortality. The methods then use higher-order influence-function estimators to estimate and correct for the bias of the semiparametric estimator. We turn first to the role played by flexible regression models.

THE IMPORTANCE OF BROAD EXPERT KNOWLEDGE IN MODELING

Before going into details, a few basic concepts concerning modeling should be clarified. First, it is helpful to understand what models are doing. One perspective found in engineering is that the model is a data filter. Given a specific class of target parameters (signals) of interest, the model's performance is judged by how well it filters out noise (data components conveying no useable information about the target) and bias (spurious background signals), while capturing signals of interest (useable information) without introducing bias (spurious signals introduced by the modeling process). Creating a good filter requires some knowledge of the target. For example, a heavy rope net of the sort used to trap larger animals is bad for collecting something as small as insects, since those would go through the mesh and thus not be captured. If instead our goal was to capture and tag only large mammalian predators, a trap with a net fine enough for insects would capture many irrelevant specimens.

Similarly, a linear model, although effective for capturing single-direction (strictly monotone) trends, will capture no information about trend reversals or flat spots. The linear model is so tightly focused on detecting linear trend components that it would leave the impression that reversals and flat spots do not exist. That is alright if indeed the latter nonlinear features are negligible in contextual terms, but otherwise it is misleading. On the other hand, a more universal (less specific) filter will have a lower threshold for pattern detection and thus will capture more noise patterns; consequently (at least if its output variances are computed correctly), the less-specific filter will leave any true signal more blurred than would a restrictive model more tailored to the target.

The data set being analyzed can supply some guidance about whether a model is removing too much information or signal (through clear lack of fit), although this is far from foolproof. But without assistance from background knowledge, typical data supply little information about the opposite direction, that is, about whether the model is admitting more noise than necessary or desirable. It is here that contextual (background scientific) knowledge about what is targeted or expected may be essential for extracting data information. When that knowledge is built into a model, the model will discard with (or as) noise those data patterns that contradict the knowledge. To the extent that knowledge is correct, those data patterns will indeed be noise, and the filtration can greatly increase estimation (signal reconstruction) accuracy. To the extent the resulting model is incorrect, however, the discarded patterns may contain important information and their removal (filtration) can greatly reduce estimation accuracy by introducing bias.

The second and closely related idea is that of an "expert-consensus prior" — a set of constraints broad or relaxed enough so that experts representing various stakeholders

This Invited Editorial was written by Dr. Sander Greenland, Departments of Epidemiology and Statistics, University of California–Los Angeles *lesdomes@ucla.edu*

(e.g., industry, environmental groups, U.S. Environmental Protection Agency) would all agree that they are not only reasonable but very likely to hold in the context. As an example, consider the shape of the curve plotting mortality against dew-point temperature in a given city if all confounding were controlled (so that the curve can be taken to depict some kind of average causal effect of dew point on mortality in the city). I would expect that all experts would agree that this plot of effect would be smooth, but also that the plot would be poorly represented by most smooth curves, such as a sinusoidal curve (repeatedly up and down).

Experts might even agree that, for the effects of both PM_{10} and covariates, there is at most only one flat spot where the mortality-effect curve reverses direction from up to down or vice-versa (the sign of the slope changes). Nonetheless, they would not all be sure of — let alone agree on — where that point is, and there would be no basis for assuming the plot is symmetric around that point. By these considerations, we would find that most polynomial, trigonometric, and simple smooth functions would not capture consensus priors about how the plot should look. Indeed, some functions could introduce pure artifacts like symmetry around a minimum or maximum (e.g., simple quadratic functions) or multiple direction reversals (e.g., simple cubic functions).

Any model used to process the data needs to be flexible enough to not force such gross and easily false model-based patterns (modeling artifacts) on the output; if it did, the artifacts would be sources of bias in our estimates. Modeling artifacts would also be sources of downward bias in our uncertainty (variance) assessment, producing spurious certainty relative to the consensus by excluding plausible patterns. Thus a consensus prior that allows trend reversal should lead us to avoid models that force strictly increasing trends. Yet we do not want a model so flexible that it too easily allows noise patterns (chance artifacts) into the output, for that model would produce variances inflated to allow for highly implausible patterns (spurious uncertainty, relative to the consensus). Thus a consensus prior that ruled out multiple trend reversals should lead us to avoid models that allow multiple reversals.

GENERAL COMMENTS

The Investigators' Report by Robins and colleagues answers two comparative questions.

Comparison 1. Do the semiparametric methods developed by Robins and associates (which use very flexible data-adaptive models for the regressions of mortality and PM_{10} levels on covariates) produce qualitatively different estimates of the effect of PM_{10} on mortality than the more standard methods used by the NMMAPS investigators? The current investigators found that the use of more flexible models did not meaningfully alter any statistical conclusion compared with those from previous NMMAPS analyses.

Comparison 2. Does the use of higher-order U-statistic (higher-order flexible [HOF]) estimators produce qualitatively different estimates of the effect of PM_{10} on mortality than the more standard methods used by the NMMAPS investigators? These HOF estimators correct the bias of the flexible data adaptive GLM semiparametric estimator of the effect of PM_{10} on mortality. Robins and colleagues found that the use of HOF estimators did not meaningfully alter any statistical conclusion. The only statistical finding they highlighted was a suggestion of possible modest bias in the results from the analysis of NMMAPS data for Minneapolis.

In any further discussion of the results of the Robins study, it is important to recognize the following general points. First, time-series studies of mortality find rate ratios of 1.03–1.06 on days when PM_{10} is qualitatively elevated. Some researchers wonder whether epidemiology is capable of reliably detecting such small increases in risk. This skepticism motivates interest in reanalyzing the data using novel methods that might be better able to control for confounding by measured factors.

Second, the analysis performed by Robins and colleagues does not employ consensus-prior constraints (apart from scientifically very mild smoothness or sparsity constraints built into their data-adaptive regression estimators). This means that the methods allow such features as multiple trend reversals for underlying relations. If such complex features were excluded by prior consensus, the investigators' regressions would exhibit more uncertainty than would be exhibited by regression models using that consensus information. Interpretation of this added uncertainty depends on the analysis results. Consider the following scenarios.

Scenario 1. The investigators note that the data under study have little ability to define fine details of the functional forms for the regression of mortality on the covariates. For the semiparametric portions of the models they considered, this means that available prior consensus about these forms could play a major role in improving estimation accuracy. Suppose the additional features (e.g., "wiggles") allowed by HOF estimators (relative to the original NMMAPS approach) would have been deemed implausible by consensus. Further suppose under this scenario that there is no strong evidence to contradict the consensus. Then the small possibility of wiggles would have contributed negligibly to

uncertainty remaining after consensus prior information was incorporated into a PM_{10} health impact analysis. This means that the additional variability shown in the HOF approaches would be regarded as largely spurious uncertainty relative to the background consensus based on the original NMMAPS approach.

Scenario 2. Suppose that the HOF estimates had differed to such a degree from the previous NMMAPS estimates that sampling variability was an implausible explanation. Since the more flexible regression models used by Robins and colleagues included the more standard NMMAPS models as submodels, there would have then been a clear conflict with prior consensus and thus something to explain. Of course, one would have to further investigate the reason the estimates differed. For example, the difference could reflect simple coding or derivation errors made when deriving and implementing the HOF method. But once such errors are ruled out or judged unlikely, the difference could be due to the consensus being incorrect and, furthermore, could have been sufficiently incorrect that the estimated rate ratio of 1.06 might be wholly attributable to confounding that was not controlled in the original NMMAPS analyses but was controlled in the HOF analyses.

Scenario 2 did not come to pass but, presumably, the possibility that it might have was behind the decision to fund the Robins study.

THE IMPORTANCE OF THE PM_{10} DOSE–RESPONSE SHAPE FOR REGULATORY CUT-POINT ESTIMATION

In the report by Robins and colleagues, the flexibility issue is framed entirely in terms of modeling the added covariate function to control confounding as thoroughly as needed. In contrast to this treatment of covariates, the report focuses on single-parameter models for the dose–response relation of PM_{10} to mortality given covariates — namely, a loglinear or linear term added to the estimated covariate function.

Suppose that prior consensus holds that any PM_{10} dose–response effect on mortality (within the observed ranges) is strictly increasing. Then the restricted PM_{10} modeling is not of great concern if the only target of interest is the population-average slope for that dose–response curve (averaged over the population joint covariate distribution). This is because a misspecified model estimates the best fit of the same misspecified model to the entire population distribution, and the slope produced by that population fit is a population-averaged slope (White 1993). Furthermore, in the present setting the PM_{10} association with mortality is so small that the average slopes will appear similar on both loglinear and linear scales.

Nonetheless, restrictive PM_{10} modeling should be of concern for policy applications for at least two reasons. First, if pooling across cities is done via simple random-city effects analysis, the estimated parameter will be an average across cities of distinct city-specific parameters. Unfortunately, the implicit weighting will not correspond to any policy-derived goal and will have no biologic or epidemiologic meaning. A more meaningful pooling would incorporate features of the city-specific covariate distributions into a meta-regression to avoid such pooling distortions (Rubin 1990; Greenland and Robins 1994).

Second, if specific PM_{10} cut-points will be used for regulation, the estimated PM_{10} level at which the magnitude of adverse effects passes a threshold (with a given confidence or probability) will be sensitive to the assumed PM_{10} dose–response shape. In particular, cut-point estimates are potentially quite biased if the assumed shape of the dose–response curve cannot reasonably approximate the actual shape, for example, if the assumed shape is loglinear (which is equivalent to exponential) but the actual shape is logarithmic.

Comparing the loglinear and linear analyses performed by Robins and associates indicates that the data may provide little information to pin down fine details of the dose–response function. This lack of data information implies that any attempt to set cut-points for PM_{10} will need to take into account dose–response uncertainty, which is not accounted for in the results of the current investigators. It is a potential concern, however, that a linear model represents only a somewhat intermediate dose–response curve rather than an opposite of the exponential curve implied by using untransformed PM_{10} in a loglinear model. Thus a risk assessment suitable for informing policy might advisably examine logarithmic curves.

A related multivariate consideration is the assumption of additivity of the PM_{10} effect to the covariate effects, which is scale-dependent. Robins and colleagues cite no scientific basis for this restriction, and thus it is valuable that they demonstrate that the summary statistical conclusions do not appear sensitive to whether the additivity is assumed to take place on a loglinear or linear scale (insensitivity to choice of log vs. identity link function). As with PM_{10} dose–response, however, a potential concern for sensitivity analysis is that the loglinear and linear scales cover too small a range of possibilities.

None of the comments in this section are intended to be critical of the research conducted by Robins and colleagues nor of their Investigators' Report. Their stated goal was to compare their HOF analyses with (1) the NMMAPS analyses that had been based on single-parameter loglinear dose–response models (although the NMMAPS investigators

have reported much more complex dose–response relationships in later publications), and (2) the NMMAPS analyses that had assumed log-additivity of effects.

THE IMPORTANCE OF MEASUREMENT ERROR

The analyses by Robins and colleagues do not account for bias and uncertainty due to measurement error, which may be considerable for certain quantities. In addressing this problem carefully, it is necessary to specify precisely the target effects and, in particular, whether those are individual- or population-level effects. For individual-level targets the impact of measurement error is complex due to its convolution with aggregation artifacts (ecological biases) that can affect group-level data, such as the data used here (Greenland and Robins 1994). Thus for simplicity I will proceed as if only population-level effects are of interest.

We can illustrate the problems raised with special cases in which all PM_{10} and covariate effects are monotone (never change direction) and measurement errors are independent of one another and of true values. In those cases, random measurement errors in PM_{10} drive its slope estimate toward the null, and random measurement errors in the covariates reduce the impact of adjusting for the covariates, so that measured-covariate adjustment (no matter how fine it is) can only partially remove confounding by the actual covariates. When the assumptions are relaxed, problems multiply; for example, measurement errors can seriously distort estimates of nonlinear effect components (driving them toward the null in the independent random error case, but possibly creating spurious nonlinearities when correlated errors exist).

From these considerations I would expect that refinements in covariate adjustment beyond the early NMMAPS analyses (as in the HOF analyses) could have little practical impact in a realistic PM_{10} risk assessment, because such an assessment would have to account for the possibility of much larger residual confounding left by covariate-measurement error. This residual confounding cannot be addressed by analyses using only the data and methods developed by Robins and associates except in a very small way: There might be nonlinear artifacts induced by measurement errors that can be better adjusted for by the Robins methods than by the early NMMAPS methods. Nonetheless, I would expect the improvement from such adjustment to be an order of magnitude smaller than the remaining residual confounding that afflicts both analyses.

Even with no such artifact, measurement error is also important when interpreting estimates of the confounding problem and the impacts of adjustments. For example, with independent random errors for all variables overlaid on monotone relations, the associations among PM_{10}, the covariates, and the outcomes will all move (degrade) toward the null. With the weakened covariate–PM_{10} and covariate–mortality relations that result, we would observe much smaller impacts of adjustments or adjustment refinements. These impacts may even appear negligible under reasonable levels of error, but the adjustments or refinements might actually be important if they were applied to the underlying true variables.

Degradation due to measurement error may have contributed to the fact that Robins and colleagues did not detect any clear impact of adjustment refinement. It should thus be noted that Robins and his team considered measurement-error issues in their original grant request, but the scope of their research changed and their work concluded before they had time to develop and implement statistical procedures that, when combined with prior information on the measurement process, would adjust for measurement error in the exposure and in the covariates.

Finally, one other source of error that is not so easily captured by simple error models is the errors and ambiguities in defining the causally relevant variables. These ambiguities are illustrated by the uncertainty about which lag periods (or averaging methods) would be ideal to use.

SUMMARY

Robins and colleagues have demonstrated that HOF analyses can be operationalized and applied to the NMMAPS data set, and that doing so leads to no meaningful change in the conclusions that one could justify based on the NMMAPS data. This result may be as expected, because the NMMAPS investigators had already employed highly flexible covariate functions and no truly large qualitative nonlinearities were cited from earlier analyses. Nonetheless, given the small increase in relative risk being reported in previous studies, exploration of the robustness of the findings to more precise control for the confounding effects of the measured covariates was a scientifically reasonable strategy. The primary scientific concerns about all the results — from both the current investigators and the early NMMAPS analyses — are more in the realm of basic epidemiology and risk assessment, which were not a goal of the Robins project and thus are not addressed in the Investigators' Report: dose-response modeling of the exposure (PM_{10}) effect if regulatory cut-points are needed; exposure measurement error; and validity concerns about residual confounding from unmeasured and poorly defined or poorly measured covariates.

REFERENCES

Greenland S, Robins J. 1994. Ecologic studies: Biases, misconceptions, and counterexamples. Am J Epidemiol 139:747–760.

Health Effects Institute. 2003. Revised Analyses of Time-Series Studies of Air Pollution and Health. Special Report. Health Effects Institute, Boston, MA.

Robins J, Li L, Tchetgen E, van der Vaart A. 2008. Higher order influence functions and minimax estimation of nonlinear functionals. Probabilities and Statistics: Essays in Honor of David A. Friedman 2:335–421.

Rubin DB. 1990. A new perspective. In: The Future of Meta-Analysis (Wachter KW, Straf ML, eds.; pp. 155–165). Russell Sage Foundation, New York.

White H. 1994. Estimation, Inference, and Specification Analysis. Cambridge University Press, New York.

RELATED HEI PUBLICATIONS SHORT-TERM STUDIES AND METHODS

Number	Title	Principal Investigator	Date*
Research Reports			
176	Effect of Air Pollution Control on Mortality and Hospital Admissions in Ireland	D.W. Dockery	2013
171	Multicity Study of Air Pollution and Mortality in Latin America (The ESCALA Study)	I. Romieu	2012
170	Impact of the 1990 Hong Kong Legislation for Restriction on Sulfur Content in Fuel	C-M. Wong	2012
169	Effects of Short-Term Exposure to Air Pollution on Hospital Admissions of Young Children for Acute Lower Respiratory Infections in Ho Chi Minh City, Vietnam	HEI Collaborative Working Group	2012
161	Assessment of the Health Impacts of Particulate Matter Characteristics	M.L. Bell	2012
157	Public Health and Air Pollution in Asia (PAPA): Coordinated Studies of Short-Term Exposure to Air Pollution and Daily Mortality in Two Indian Cities		2011
	Part 1. Short-Term Effects of Air Pollution on Mortality: Results from a Time-Series Analysis in Chennai, India	K. Balakrishnan	
	Part 2. Time-Series Study on Air Pollution and Mortality in Delhi	U. Rajarathnam	
154	Public Health and Air Pollution in Asia (PAPA): Coordinated Studies of Short-Term Exposure to Air Pollution and Daily Mortality in Four Cities		2010
	Part 1. A Time-Series Study of Ambient Air Pollution and Daily Mortality in Shanghai, China	H. Kan	
	Part 2. Association of Daily Mortality with Ambient Air Pollution, and Effect Modification by Extremely High Temperature in Wuhan, China	Z. Qian	
	Part 3. Estimating the Effects of Air Pollution on Mortality in Bangkok, Thailand	N. Vichit-Vadakan	
	Part 4. Interaction Between Air Pollution and Respiratory Viruses: Time-Series Study of Daily Mortality and Hospital Admissions in Hong Kong	C-M. Wong	
	Part 5. Public Health and Air Pollution in Asia (PAPA): A Combined Analysis of Four Studies of Air Pollution and Mortality	C-M. Wong	
152	Evaluating Heterogeneity in Indoor and Outdoor Air Pollution Using Land-Use Regression and Constrained Factor Analysis	J.I. Levy	2010
148	Impact of Improved Air Quality During the 1996 Summer Olympic Games in Atlanta on Multiple Cardiovascular and Respiratory Outcomes	J.L. Peel	2010
142	Air Pollution and Health: A European and North American Approach (APHENA)	K. Katsouyanni and J.M. Samet	2009
127	Personal, Indoor, and Outdoor Exposures to $PM_{2.5}$ and Its Components for Groups of Cardiovascular Patients in Amsterdam and Helsinki	B. Brunekreef	2005

Continued

Copies of these reports can be obtained from HEI; pdf's are available for free downloading at http://pubs.healtheffects.org.

RELATED HEI PUBLICATIONS SHORT-TERM STUDIES AND METHODS

Number	Title	Principal Investigator	Date*
123	Time-Series Analysis of Air Pollution and Mortality: A Statistical Review	F. Dominici	2004
99	A Case–Crossover Analysis of Fine Particulate Matter Air Pollution and Out-of-Hospital Sudden Cardiac Arrest	H. Checkoway	2000
98	Daily Mortality and Fine and Ultrafine Particles in Erfurt, Germany	H-E. Wichmann	2000
	Part I. Role of Particle Number and Particle Mass		
97	Identifying Subgroups of the General Population That May Be Susceptible to Short-Term Increases in Particulate Air Pollution: A Time-Series Study in Montreal, Quebec	M.S. Goldberg	2000
95	Association of Particulate Matter Components with Daily Mortality and Morbidity in Urban Populations	M. Lippmann	2000
94	The National Morbidity, Mortality, and Air Pollution Study		
	Part I. Methods and Methodologic Issues	J.M. Samet	2000
	Part II. Morbidity and Mortality from Air Pollution in the United States	J.M. Samet	2000
	Part III. Concentration–Response Curves and Thresholds for the 20 Largest U.S. Cities	M.J. Daniels	2004
	Part IV. Hierarchical Bivariate Time-Series Models — A Combined Analysis of PM_{10} Effects on Hospitalization and Mortality	F. Dominici	2005
83	Daily Changes in Oxygen Saturation and Pulse Rate Associated with Particulate Air Pollution and Barometric Pressure	D.W. Dockery	1999
81	Methods Development for Epidemiologic Investigations of the Health Effects of Prolonged Ozone Exposure		
	Part I. Variability of Pulmonary Function Measures	I.B. Tager	1998
	Part II. An Approach to Retrospective Estimation of Lifetime Ozone Exposure Using a Questionnaire and Ambient Monitoring Data (California Sites)	I.B. Tager	1998
	Part III. An Approach to Retrospective Estimation of Lifetime Ozone Exposure Using a Questionnaire and Ambient Monitoring Data (U.S. Sites)	P.L. Kinney	1998
Special Reports			
18	Outdoor Air Pollution and Health in the Developing Countries of Asia: A Comprehensive Review		2010
	Revised Analyses of Time-Series Studies of Air Pollution and Health		2003
HEI Communications			
12	Internet-Based Health and Air Pollution Surveillance System	S.L. Zeger	2006

Copies of these reports can be obtained from HEI; pdf's are available for free downloading at http://pubs.healtheffects.org.

HEI BOARD, COMMITTEES, and STAFF

Board of Directors

Richard F. Celeste, Chair *President Emeritus, Colorado College*

Sherwood Boehlert *Of Counsel, Accord Group; Former Chair, U.S. House of Representatives Science Committee*

Enriqueta Bond *President Emerita, Burroughs Wellcome Fund*

Purnell W. Choppin *President Emeritus, Howard Hughes Medical Institute*

Michael T. Clegg *Professor of Biological Sciences, University of California–Irvine*

Jared L. Cohon *President Emeritus and Professor, Civil and Environmental Engineering and Engineering and Public Policy, Carnegie Mellon University*

Stephen Corman *President, Corman Enterprises*

Gowher Rizvi *Vice Provost of International Programs, University of Virginia*

Linda Rosenstock *Dean Emerita and Professor of Health Policy and Management, Environmental Health Sciences and Medicine, University of California–Los Angeles*

Henry Schacht *Managing Director, Warburg Pincus; Former Chairman and Chief Executive Officer, Lucent Technologies*

Warren M. Washington *Senior Scientist, National Center for Atmospheric Research; Former Chair, National Science Board*

Archibald Cox, Founding Chair *1980–2001*

Donald Kennedy, Vice Chair Emeritus *Editor-in-Chief Emeritus,* Science*; President Emeritus and Bing Professor of Biological Sciences, Stanford University*

Health Research Committee

David L. Eaton, Chair *Dean and Vice Provost of the Graduate School, University of Washington–Seattle*

David Christiani *Elkan Blout Professor of Environmental Genetics, Harvard School of Public Health*

Francesca Dominici *Professor of Biostatistics and Senior Associate Dean for Research, Harvard School of Public Health*

David E. Foster *Phil and Jean Myers Professor Emeritus, Department of Mechanical Engineering, Engine Research Center, University of Wisconsin–Madison*

Uwe Heinrich *Professor, Hannover Medical School; Executive Director, Fraunhofer Institute for Toxicology and Experimental Medicine, Hanover, Germany*

Grace LeMasters *Professor of Epidemiology and Environmental Health, University of Cincinnati College of Medicine*

Allen L. Robinson *Raymond J. Lane Distinguished Professor and Head, Department of Mechanical Engineering, and Professor, Department of Engineering and Public Policy, Carnegie Mellon University*

Richard L. Smith *Director, Statistical and Applied Mathematical Sciences Institute, University of North Carolina–Chapel Hill*

James A. Swenberg *Kenan Distinguished Professor of Environmental Sciences, Department of Environmental Sciences and Engineering, University of North Carolina–Chapel Hill*

HEI BOARD, COMMITTEES, and STAFF

Health Review Committee

Homer A. Boushey, Chair *Professor of Medicine, Department of Medicine, University of California–San Francisco*

Michael Brauer *Professor, School of Environmental Health, University of British Columbia, Canada*

Bert Brunekreef *Professor of Environmental Epidemiology, Institute of Risk Assessment Sciences, University of Utrecht, the Netherlands*

Mark W. Frampton *Professor of Medicine and Environmental Medicine, University of Rochester Medical Center*

Stephanie London *Senior Investigator, Epidemiology Branch, National Institute of Environmental Health Sciences*

Roger D. Peng *Associate Professor, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health*

Armistead Russell *Howard T. Tellepsen Chair of Civil and Environmental Engineering, School of Civil and Environmental Engineering, Georgia Institute of Technology*

Lianne Sheppard *Professor of Biostatistics, School of Public Health, University of Washington–Seattle*

Officers and Staff

Daniel S. Greenbaum *President*

Robert M. O'Keefe *Vice President*

Rashid Shaikh *Director of Science*

Barbara Gale *Director of Publications*

Jacqueline C. Rutledge *Director of Finance and Administration*

Helen I. Dooley *Corporate Secretary*

Kate Adams *Senior Scientist*

Hanna Boogaard *Staff Scientist*

Aaron J. Cohen *Principal Scientist*

Maria G. Costantini *Principal Scientist*

Philip J. DeMarco *Compliance Manager*

Hope Green *Editorial Assistant*

L. Virgi Hepner *Senior Science Editor*

Anny Luu *Administrative Assistant*

Francine Marmenout *Senior Executive Assistant*

Nicholas Moustakas *Policy Associate*

Hilary Selby Polk *Senior Science Editor*

Jacqueline Presedo *Research Assistant*

Margarita Shablya *Science Administrative Assistant*

Robert A. Shavers *Operations Manager*

Geoffrey H. Sunshine *Senior Scientist*

Annemoon M.M. van Erp *Managing Scientist*

Katherine Walker *Senior Scientist*

HEALTH
EFFECTS
INSTITUTE

101 Federal Street, Suite 500
Boston, MA 02110, USA
+1-617-488-2300
www.healtheffects.org

RESEARCH
REPORT

Number 175
November 2013