The Annals of Applied Statistics
2019, Vol. 13, No. 3, 1927–1956
https://doi.org/10.1214/19-AOAS1260
© Institute of Mathematical Statistics, 2019

BAYESIAN METHODS FOR MULTIPLE MEDIATORS: RELATING PRINCIPAL STRATIFICATION AND CAUSAL MEDIATION IN THE ANALYSIS OF POWER PLANT EMISSION CONTROLS[1]

BY CHANMIN KIM*, MICHAEL J. DANIELS†, JOSEPH W. HOGAN‡, CHRISTINE CHOIRAT§ AND CORWIN M. ZIGLER¶

Boston University School of Public Health, University of Florida†, Brown University School of Public Health‡, Swiss Data Science Center§, University of Texas at Austin¶*

Emission control technologies installed on power plants are a key feature of many air pollution regulations in the US. While such regulations are predicated on the presumed relationships between emissions, ambient air pollution and human health, many of these relationships have never been empirically verified. The goal of this paper is to develop new statistical methods to quantify these relationships. We frame this problem as one of mediation analysis to evaluate the extent to which the effect of a particular control technology on ambient pollution is mediated through causal effects on power plant emissions. Since power plants emit various compounds that contribute to ambient pollution, we develop new methods for multiple intermediate variables that are measured contemporaneously, may interact with one another, and may exhibit joint mediating effects. Specifically, we propose new methods leveraging two related frameworks for causal inference in the presence of mediating variables: principal stratification and causal mediation analysis. We define principal effects based on multiple mediators, and also introduce a new decomposition of the total effect of an intervention on ambient pollution into the natural direct effect and natural indirect effects for all combinations of mediators. Both approaches are anchored to the same observed-data models, which we specify with Bayesian nonparametric techniques. We provide assumptions for estimating principal causal effects, then augment these with an additional assumption required for causal mediation analysis. The two analyses, interpreted in tandem, provide the first empirical investigation of the presumed causal pathways that motivate important air quality regulatory policies.

1. Introduction. Motivated by evidence of the association between ambient air pollution and human health outcomes, the US Environmental Protection Agency (EPA) oversees a vast program for air quality management designed to limit population exposure to harmful air pollution (Dominici, Greenstone and Sunstein (2014), Pope III, Ezzati and Dockery (2009)). Fine particulate matter of diameter 2.5 micrometers or less ($PM_{2.5}$) is of particular importance with regulations

Received July 2017; revised January 2019.

[1] Supported by NIH R01ES026217, NIH R01CA183854, NIH R01GM112327, NIH R01GM111339, NCI P01 CA134294, HEI 4909, HEI 4953 and USEPA RD-83587201.

Key words and phrases. Ambient $PM_{2.5}$, Bayesian nonparametrics, Gaussian copula, multipollutants, natural indirect effect.

to limit exposure to $PM_{2.5}$ estimated to account for over half of the benefits and a substantial portion of the costs of all monetized federal regulations (Office of Management and Budget (2013)). A large contributor to ambient $PM_{2.5}$ in the US is the power generating sector, in particular coal-fired power plants. These plants emit $PM_{2.5}$ directly into the atmosphere but are also major sources of sulfur dioxide (SO_2) and nitrogen oxides (NO_x) that, once emitted into the atmosphere, contribute to secondary formation of $PM_{2.5}$ through chemical reaction, coagulation and other mechanisms. The amount of $PM_{2.5}$ formation initiated by emissions of SO_2 and NO_x depends largely on atmospheric conditions such as temperature (Hodan and Barnard (2004)). Power plants are also major sources of CO_2 emissions.

A variety of regulatory programs under the purview of the Clean Air Act (e.g., the acid rain program) are designed to reduce emissions from power plants with one goal of reducing population exposure to ambient $PM_{2.5}$. One key strategy for achieving this reduction is the installation of SO_2 control technologies such as flue-gas desulfurization scrubbers (henceforth, “scrubbers”) on power plant smokestacks to reduce SO_2 emissions and, in turn, $PM_{2.5}$. Estimates of the annualized human health benefits of regulatory polices such as the acid rain program rely heavily on presumed relationships between such control strategies, emissions, ambient $PM_{2.5}$ and human health. While the underlying physical and chemical understanding of the link between power plant emissions and $PM_{2.5}$ is well established, there remains considerable uncertainty about the effectiveness of specific strategies for reducing harmful pollution amid the realities of actual regulatory implementation. Accordingly, the EPA and other stakeholders have increasingly emphasized the need to provide evidence of which specific air pollution control strategies are most effective or efficient for reducing population exposures to $PM_{2.5}$ (HEI Accountability Working Group (2003), U. S. EPA (2013)).

The goal of this paper is to propose a statistical method to examine the causal effect of scrubbers installed at coal-fired power plants on the ambient concentration of ambient $PM_{2.5}$ using observed data on power plant emissions and ambient pollution. Physical and chemical understanding of these processes provide strong support for the expectation that scrubbers reduce ambient $PM_{2.5}$“through” reducing emissions of SO_2, but this relationship has never been empirically verified using observed data in the context of regulations that may simultaneously impact a variety of factors. A key statistical challenge to verifying this relationship derives from the fact that SO_2 emissions are highly correlated with emissions of NO_x and CO_2, and NO_x is known to play an important role in the formation of ambient $PM_{2.5}$, possibly through interactions with SO_2. Thus, the question will be formally framed as one of mediation analysis. To what extent is the causal effect of a scrubber (the “treatment”) on ambient $PM_{2.5}$ (the “outcome”) mediated through reduced emissions of SO_2, NO_x and CO_2 (the “mediators”)? Recovering a statistical answer to this question amid the problem of multiple highly correlated and possibly interacting mediators that are measured contemporaneously requires new methods development and would also serve to bolster the promise of statistical methods in studies of air pollution that have historically relied on physical and chemical knowledge and not on statistical analysis.

To answer this question, we develop new methods that draw from two frameworks for estimating causal effects in the presence of mediating variables: (1) principal stratification (Frangakis and Rubin (2002)) and (2) causal mediation analysis (Robins and Greenland (1992)). The methodological contributions of this paper come in three areas. First, we develop new methods to accommodate multivariate mediating variables that are measured *contemporaneously* (not sequentially), are correlated and may interact with each to impact the outcome (see Figure 1 for a an illustrative directed acyclic graph). This is essential for evaluating scrubbers because power plants simultaneously emit multiple pollutants that may interact through atmospheric processes to impact ambient $PM_{2.5}$. Existing methods in the literature for both principal stratification and mediation analysis have primarily focused on settings with a single mediator (e.g., Baron and Kenny (1986), Daniels et al. (2012), Frangakis and Rubin (2002), Joffe and Greene (2009), VanderWeele (2009)). Existing extensions to cases with multiple mediating variables cannot accommodate the setting of power plant emissions where mediators may simultaneously and jointly impact the outcome (Daniel et al. (2015), Imai and Yamamoto (2013), VanderWeele and Vansteelandt (2014), Wang, Nelson and Albert (2013)). Our second methodological contribution is the use of Bayesian nonparametric approaches to model the observed distribution of emissions and pollution outcomes, making use of a multivariate Gaussian copula model to link flexibly modeled marginal distributions of observed outcomes to a joint distribution of potential outcomes. Similar strategies with a single mediator have received recent attention in the principal stratification literature (Bartolucci and Grilli (2011), Conlon, Taylor and Elliott (2017), Ma, Roy and Marcus (2011), Schwartz, Li and Mealli (2011)) and are emerging for causal mediation analysis (Daniels et al. (2012), Kim et al. (2017)). These approaches are important for confronting continuous mediators and infinitely many principal strata and are deployed here in a novel way to address the problem of multiple mediators while flexibly modeling the observed-data distributions of both mediators and outcomes. Finally, we provide a unification of principal stratification and causal mediation analysis. While the mathematical relationships between these two approaches are well understood (Mattei and Mealli (2011), Mealli and Rubin (2003), VanderWeele (2011)), there has not been, to our knowledge, a comprehensive deployment of both perspectives in a complementary fashion to illuminate the scientific underpinnings of a specific problem. Baccini, Mattei and Mealli (2015) made important progress in this direction using different observed-data models to estimate principal effects and mediation effects in a problem with a single mediator. In contrast the approach developed here uses the exact same observed-data models to ground both perspectives, proposes a common set of basic assumptions for estimating both principal effects and mediating effects, modularizes an additional assumption required to augment a principal stratification analysis in order to obtain estimates of natural direct and indirect effects and considers settings with multiple mediating variables. Ultimately, we provide a new dimension of quantitative, statistical evidence for supporting air policy regulatory decisions.

2. Scrubber installation and linked data sources. Title IV of the Clean Air Act established the acid rain program (ARP) which required major emissions reductions of SO_2 (and other emissions) by 10 million tons relative to 1980 levels. This reduction was achieved mostly through cutting emissions from power plants, or more formally, electricity-generating units (EGUs). Impacts of the ARP have been evaluated extensively, and the program is generally lauded as a success due to marked national decreases in SO_2 and NO_x coming at relatively low cost. Estimates of the annualized human health benefits of the entire ARP range from $50 billion to $100 billion (Chestnut and Mills (2005)) but rely heavily on presumed relationships between power plant emissions, ambient $PM_{2.5}$ and human health.

While power plants under the ARP had latitude to elect a variety of strategies to reduce emissions, one key strategy is the installation of a scrubber to reduce SO_2 emissions. The precise extent to which installation of a scrubber reduces ambient $PM_{2.5}$ through reducing SO_2 emissions remains unknown and has never been estimated empirically amid the realities of actual regulatory implementation where pollution controls may impact a variety of factors that are also related to the formation of $PM_{2.5}$. Knowledge of these relationships is complicated by the fact that power plants emit more than just SO_2, and emissions of a variety of pollutants likely interact in the surrounding atmosphere to form ambient $PM_{2.5}$.

To provide refined evidence of the extent to which scrubbers reduce emissions and cause improvements to ambient air quality, we assembled a national database of ambient air quality measures, weather conditions and information on power plants. Specifically, we assembled data on 258 coal-fired power plants from the EPA Air Markets Program Data and the Energy Information Administration, with information on plant characteristics, emissions control technologies installed (if any) and emissions of SO_2, NO_x and CO_2 during 2005, five years after promulgation of an important phase of regulations under the ARP. For each power plant we augment the data set with annual average ambient $PM_{2.5}$ concentrations in 2005 and baseline meteorologic conditions in 2004 measured at all monitoring stations in the EPA Air Quality System that are located within 150 km. The 150 km range was chosen not only to acknowledge that atmospheric processes carry power plant emissions across distances at least this great, but also to minimize the number of monitoring stations considered within range of more than one power plant. We regard any power plant as "treated" with scrubbers in 2005 if at least 10% of the plant's total heat input was attributed to a portion of the plant equipped with a scrubber as of January 2005. Note that this proportion was nearly 0% or nearly 100% for the vast majority of

TABLE 1
Summary statistics for covariates and outcomes available for the analysis of SO_2 *scrubbers*

	Have scrubbers ($n = 59$)		Have no scrubber ($n = 190$)	
	Median	IQR	Median	IQR
Monitor Data				
Average Ambient $PM_{2.5}$ 2005 (μg/m^3)	12.4	(7.8, 14.8)	13.7	(11.8, 15.2)
Average Temperature 2004 (°C)	11.5	(10.1, 15.0)	12.8	(10.4, 16.1)
Average Barometric Pressure 2004 (mmHg)	737.8	(686.7, 752.4)	746.1	(739.1, 755.6)
Power Plant Level Data				
Total SO_2 Emission 2005 (tons)	644.3	(257.3, 1819.9)	1267.1	(504.9, 2707.6)
Total NO_x Emission 2005 (tons)	852.1	(394.2, 1531.3)	442.5	(193.7, 878.2)
Total CO_2 Emission 2005 (×1000 tons)	505.3	(232.5, 960.7)	283.6	(117.7, 559.0)
Unit Level Data				
Average Heat Input 2004 (×1000 MMBtu)	4653.3	(2266.4, 9363.9)	2783.4	(1147.6, 5448.1)
Total Operating Time 2004 (hours × # units)	7944.0	(7565.8, 8154.9)	7583.9	(7171.0, 7985.9)
Sulfur Content in Coal 2004 (lb/MMBtu)	1.0	(0.5, 2.2)	0.7	(0.3, 1.1)
Num. of NO_x Controls 2004 (# units)	1.0	(1.0, 1.5)	1.0	(0.9, 1.3)
Pct. operating Capacity 2004 (MMBtu/MMBtu × 100)	20.2	(10.0, 28.8)	16.4	(9.3, 24.6)
Heat Rate 2004 (MMBtu/MWh)	268.5	(175.5, 436.9)	254.3	(152.6, 396.8)

plants, indicating robustness to this 10% cutoff. Other power plant characteristics are listed in Table 1. The data files and programs to assemble the analysis data set are available at https://dataverse.harvard.edu/dataverse/mmediators and https://github.com/lit777/MultipleMediators respectively.

3. Causal mediation analysis and principal stratification.

3.1. *Mediation analysis with a single mediator.* To fix ideas, consider the single mediator case. Let $Z_i \in \{0, 1\}$ indicate the presence of the intervention of interest here, whether power plant i had installed scrubbers in January 2005 ($Z_i = 1$), and let $\boldsymbol{Z} = (Z_1, \dots, Z_n)$ be the vector of intervention indicators for power plants $i = 1, \dots, n$. Using potential-outcomes notation (Rubin (1974)), let $M_i(\boldsymbol{Z})$ denote the potential emissions that the ith power plant would be generated under the vector of scrubber assignments $\boldsymbol{Z}$, and let $Y_i(\boldsymbol{Z}; \boldsymbol{M})$ denote the potential ambient $PM_{2.5}$ outcome that could, in principle, be defined for any scrubber assignment vector $\boldsymbol{Z}$ and any vector of intermediate emissions values $\boldsymbol{M}$. Throughout the paper we adopt the stable unit treatment value assumption (SUTVA) which implies: (1) there is no "interference" in the sense that potential intermediate and outcome values from power plant i do not depend on scrubber treatments and emissions intermediates of other power plants (i.e, $M_i(\boldsymbol{Z}) = M_i(Z_i)$ and $Y_i(\boldsymbol{Z}; \boldsymbol{M}) =$

$Y_i(Z_i; M_i)$), and (2) there are "no multiple versions" of scrubber treatments such that whenever $Z_i = Z_i'$, $M_i(Z_i) = M_i(Z_i')$ and $Y_i(Z_i; M_i(Z_i)) = Y_i(Z_i'; M_i(Z_i'))$. For reasons that will become clear later, we augment the standard SUTVA to also assume "no multiple versions" of emissions intermediates which states, if $M_i = M_i'$, then $Y_i(Z_i; M_i) = Y_i(Z_i; M_i')$ (Forastiere et al. (2016)). We revisit possible violations of SUTVA in Section 8, but note here that the linkage of power plants to monitors within 150 km provides some justification for this assumption.

The *natural direct effect* (Pearl (2001)) is defined by NDE $= E[Y_i(1; M_i(0)) - Y_i(0; M_i(0))]$, representing the effect of the intervention obtained when setting the mediator to its 'natural' value $M_i(0)$, that is, its realization in the absence of the intervention. The *natural indirect effect* is defined as NIE $= E[Y_i(1; M_i(1)) - Y_i(1; M_i(0))]$, representing the effect of holding the intervention status fixed at $Z = 1$ but changing the value of the mediator from $M(0)$ to $M(1)$. The total causal effect of the intervention on the outcome can then be defined as TE $=$ NDE $+$ NIE $= E[Y_i(1; M_i(1)) - Y_i(0; M_i(0))]$. Similar controlled effects could also be defined to represent causal effects at specific values of M (Pearl (2001), Robins and Greenland (1992)).

Implicit in the definition of these effects is the conceptualization of hypothetical interventions that could independently manipulate values of both Z and M to, for example, "block" the effect on the mediator. Thus, it is important to note that potential outcomes of the form $Y_i(Z_i; M_i(Z_i'))$ are purely hypothetical for $Z_i \neq Z_i'$ and can never be observed for any observational unit. Such unobservable potential outcomes have been referred to as a priori counterfactuals (Robins and Greenland (1992), Rubin (2004)). We revisit conceptualization of a priori counterfactuals in the context of the power plant study in Section 4.1, but note here the distinction between a priori counterfactuals and potential outcomes of the form $Y_i(Z_i; M_i(Z_i))$ that are *observable* and actually observed for some units.

3.2. *Principal stratification.* A distinct but related framework for defining causal effects in the presence of intermediate variables is *principal stratification* (Frangakis and Rubin (2002)). Continuing with the single-mediator case, principal stratification considers only a single intervention and relies on definition of two causal effects: the effect of Z_i on M_i, defined as $M_i(1) - M_i(0)$, and the effect of Z_i on Y_i, defined as $Y_i(1; M_i(1)) - Y_i(0; M_i(0))$. The objective is to estimate *principal effects* which are average causal effects of Z_i on Y_i within *principal strata* of the population defined by $(M_i(0), M_i(1))$.

With principal stratification, *dissociative effects* are defined to quantify the extent to which the intervention causally affects outcomes when the intervention does not causally affect the mediator, for example, $E[Y_i(1; M_i(1)) - Y_i(0; M_i(0)) | M_i(1) = M_i(0)]$. Dissociative effects are similar to direct effects in a mediation analysis in that they represent causal effects of an intervention on the outcome among the subpopulation where there is no causal effect on the mediator, but they refer only to the specific subpopulation with $M(1) = M(0)$. VanderWeele

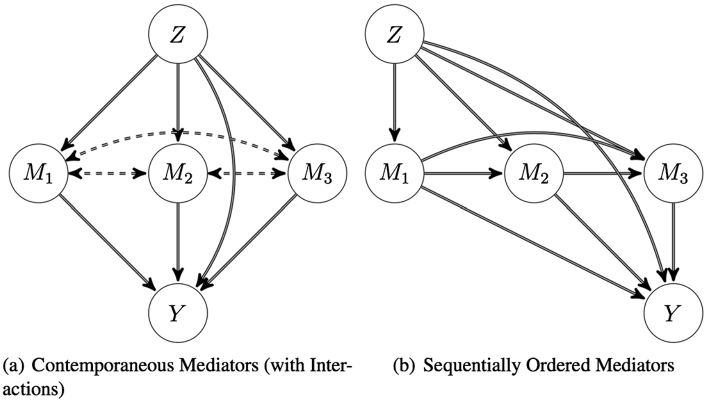
(2008) and Mealli and Mattei (2012) show that dissociative effects represent a quantity that is only one contributor to the NDE, with the amount of contribution tied to the size of the subpopulation with $M(1) = M(0)$.

Associative effects are defined to quantify the causal effect of the intervention on the outcome among those for which the intervention *does* causally affect the mediator, for example, $E[Y_i(1; M_i(1)) - Y_i(0; M_i(0))|M_i(1) < M_i(0)]$. An associative effect that is large in magnitude relative to the dissociative effect indicates that the causal effect of the intervention on the outcome is greater among those for which the mediator is causally affected, compared to those for which the mediator is not affected. This could be interpreted as suggestive of a causal pathway whereby the intervention impacts the outcome through changing the mediator, but note that associate effects are generally a combination of the NDE and NIE for a defined subpopulation.

Dissociative effects that are similar in magnitude to associative effects indicate that the intervention effect on the outcome is similar among observations that do and do not exhibit causal effects on the mediator which could be interpreted as suggestive of other causal pathways through which Z_i affects Y_i.

A primary distinction between principal stratification and causal mediation analysis is that principal effects only pertain to population subgroups comprised of observations with particular values of $(M_i(0), M_i(1))$, whereas natural direct and indirect effects are defined for the whole population (as discussed in detail in Mealli and Mattei (2012)). Importantly, note that the a priori counterfactuals of the form $Y_i(Z_i; M_i(Z_i'))$ for $Z_i \neq Z_i'$ do not appear in the definition of principal effects which rely only on the definition of *observable* potential outcomes $Y_i(Z_i; M_i(Z_i))$. Thus, there is no conception in principal stratification of a hypothetical intervention acting on M_i independently from Z_i, and there is no definition of a causal effect of Z_i on Y_i that is mediated through M_i. From a modeling perspective principal effects can be estimated when an outcome model is specified conditional on both potential mediators (intermediate outcomes), $M_i(0)$ and $M_i(1)$, while causal mediation analysis has tended to rely on an outcome model that depends on the observed mediator. The differences in modeling strategies that are typically employed in principal stratification and causal mediation analysis complicate comparisons as results of such analyses have typically been driven in part by different modeling assumptions. In Section 5 we will propose a new set of assumptions to build a common observed-data model for principal stratification and causal mediation analysis.

3.3. *Existing considerations for multiple mediators.* Extensions of the causal mediation ideas outlined in Section 3.1 to settings of multiple mediating variables are emerging. For contemporaneously observed mediators straightforward extensions of the Baron and Kenny (1986) regression-based structural equation model approach (MacKinnon (2008)) have been proposed. For each of K contemporaneous mediators $(M_1, M_2, \dots, M_K)$, a series of regression models is used to estimate



(a) Contemporaneous Mediators (with Interactions) (b) Sequentially Ordered Mediators

FIG. 1. *Directed Acyclic Graphs*: (*a*) *contemporaneous mediators with interactions* (*our case*), *and* (*b*) *sequentially ordered mediators*.

mediator-specific NIEs in a manner that implies additivity of indirect effects,

$$\text{JNIE} = \sum_{k=1}^{K} \text{NIE}_k \quad \text{and} \quad \text{TE} = \text{NDE} + \text{JNIE}, \tag{3.1}$$

where JNIE is used to denote the joint natural indirect effect due to changes in all K mediators, and $\text{NIE}_k = E[Y_i(1; M_{k,i}(1)) - Y_i(1; M_{k,i}(0))]$ represents the natural indirect effect of the kth mediator. These approaches assume that each $M_{k,i}$ mediates the treatment effect independently of the other mediators without interactions among mediators (i.e., the mediators are *causally independent* or *parallel*). Figure 1(a) without dashed lines illustrates this case. Wang, Nelson and Albert (2013) propose an alternative modeling approach under the setting of causally independent mediators. If the mediators interact with each other in terms of their impact on the outcome, then additivity of indirect effects as in the above cannot hold, and estimation of multivariate mediated effects can then be further complicated by correlations among the mediators. Dependence among mediators has been considered when M_k are observed sequentially (i.e., sequential mediators; Figure 1(b)), as in Imai and Yamamoto (2013). Albert and Nelson (2011) and Daniel et al. (2015) propose approaches for either sequentially dependent mediators or mediators that do not affect nor interact with each other. These approaches offer a decomposition of the JNIE in the case of sequential dependence and assume additivity of natural indirect effects otherwise. VanderWeele and Vansteelandt (2014) discuss an approach to decompose the JNIE further when the mediators simultaneously affect each other; however, their approach does not evaluate the impact of each individual mediator (see Section 4.3). Taguri, Featherstone and Cheng (2018) propose an approach for contemporaneous, nonordered mediators but rely on an assumption that the mediators are conditionally independent given observed covariates, which does not fully represent the possibility of contemporaneous interactions among the mediators, as may be the case with multiple emissions (in particular SO_2 and NO_x) and the formation of ambient $PM_{2.5}$. Section 6 examines the possibility of contemporaneous interactions among (possibly correlated) mediators in the context of the scrubber study.

In summary, existing methods for multiple mediators rely on either assumed causal independence of (parallel) mediators and additivity of indirect effects, sequential dependence of mediators or on restrictive assumptions of conditional independence among mediators. VanderWeele and Vansteelandt (2014) point out that, if there are interactions between the effects of (nonsequential) multiple mediators on the outcome, the joint indirect effect may not be the sum of all three indirect effects. They note that, in principle, an analysis could proceed with an outcome model, including interactions $M_j M_k$ for all $\{j, k\}$ combinations combined with models for $(E(M_j, M_k))$. However, this approach would lead to issues of model compatibility between the models for M_j and M_k and that for the product $M_j M_k$. The lack of satisfactory methods for more general settings of multiple contemporaneously measured mediators motivates the methods developed herein, where we offer a new decomposition of the joint natural indirect effect into individual indirect effects that may not affect the outcome additively.

4. New methods for causal mediation analysis and principal stratification with multiple contemporaneous mediators.

4.1. *Notation for multiple mediating variables.* Suppressing the i subscript indexing power plants, let $\{M_k(z); k = 1, \dots, K\}$ denote the potential emissions of K pollutants that would occur if a power plant were to have scrubber status $Z = z$, for $z = 0, 1$. While much of our development is general for any K, we focus on the case $K = 3$ so that $M_k(z), k = 1, 2, 3$ denotes the potential emissions of SO_2, NO_x and CO_2 respectively. The causal effect of the scrubber on emission k can then be defined as a comparison between $M_k(1)$ and $M_k(0)$. Let $\mathbf{M}(z_1, z_2, z_3) \equiv \{M_1(z_1), M_2(z_2), M_3(z_3)\}$ denote potential emissions under a set of three scrubber statuses $\{z_1, z_2, z_3\}$.

We similarly define potential $PM_{2.5}$ outcomes but extend the notation to define potential concentrations under different values of scrubber status, Z, and different possible values of emissions, $\mathbf{M}(z_1, z_2, z_3)$. Thus, in full generality each power plant has a set of $2^{K+1} = 16$ potential outcomes for $PM_{2.5}$, $Y(z; \mathbf{M}(z_1, z_2, z_3))$ which denote potential values of $PM_{2.5}$ that would be observed under intervention $Z = z$ with pollutant emissions set at values under interventions z_1, z_2, z_3. Definition of all 16 potential $PM_{2.5}$ concentrations is required for definition of natural direct and indirect effects and entails a priori counterfactuals. For example, $Y(1; \mathbf{M}(0, 0, 1))$ would represent the potential ambient $PM_{2.5}$ concentration near a plant under the hypothetical scenario where the plant installs a scrubber $(z = 1)$, but where emissions of SO_2 and NO_x are set to what they would be without the scrubber $(z_1 = z_2 = 0)$, and emissions of CO_2 are set to what they would be with the scrubber ($z_3 = 1$). This may be conceptualized as a setting where a power plant installs a scrubber but offsets the cost of the technology by burning coal with a higher sulfur content and discontinuing use of a different NO_x control, thus "blocking" the intervention and maintaining SO_2 and NO_x emissions at levels that would have occurred without the SO_2 technology. Principal stratification will only rely on potential outcomes with $z = z_1 = z_2 = z_3$ that are observable from the data, such as $\mathbf{M}(1, 1, 1)$ and $Y(1; \mathbf{M}(1, 1, 1))$ observed for any power plant that installs a scrubber. Finally, let $\boldsymbol{X}$ denote a vector of baseline covariates measured at the power plant or the surrounding area.

4.2. *Observable outcomes*: *Principal causal effects for multiple mediators*. Extending principal stratification to settings where the intermediate variable is multivariate is conceptually straightforward. Principal stratification defines a principal stratum for every combination of the joint vector $(\mathbf{M}(0, 0, 0), \mathbf{M}(1, 1, 1))$, and principal causal effects are defined as comparisons between $Y(0; \mathbf{M}(0, 0, 0))$ and $Y(1; \mathbf{M}(1, 1, 1))$ within principal strata.

For any subset $\mathcal{K} \subseteq \{1, 2, 3\}$, let $|\mathbf{M}(1, 1, 1) - \mathbf{M}(0, 0, 0)|_{\mathcal{K}}$ denote the elementwise absolute differences between emissions of the subset of pollutants in $\mathcal{K}$, for example, $|\mathbf{M}(1, 1, 1) - \mathbf{M}(0, 0, 0)|_{\mathcal{K}=\{1,3\}} = \{|M_1(1) - M_1(0)|, |M_3(1) - M_3(0)|\}$. Definitions of quantities such as average associative and dissociative effects can proceed following Zigler, Dominici and Wang (2012) by defining

$$\begin{aligned}
&\mathrm{EDE}_{\mathcal{K}}\\
&\quad = E\bigl[Y\bigl(1; \mathbf{M}(1, 1, 1)\bigr) - Y\bigl(0; \mathbf{M}(0, 0, 0)\bigr)\big|\bigl|\bigl(\mathbf{M}(1, 1, 1) - \mathbf{M}(0, 0, 0)\bigr)\bigr|_{\mathcal{K}} < C_{\mathcal{K}}^{D}\bigr],\\
&\mathrm{EAE}_{\mathcal{K}}\\
&\quad = E\bigl[Y\bigl(1; \mathbf{M}(1, 1, 1)\bigr) - Y\bigl(0; \mathbf{M}(0, 0, 0)\bigr)\big|\bigl|\bigl(\mathbf{M}(1, 1, 1) - \mathbf{M}(0, 0, 0)\bigr)\bigr|_{\mathcal{K}} > C_{\mathcal{K}}^{A}\bigr],
\end{aligned}$$

where $C_{\mathcal{K}}^{A}$ denotes a vector of thresholds beyond which a change in each emission in $\mathcal{K}$ is considered meaningful, $C_{\mathcal{K}}^{D}$ is a vector of thresholds below which changes in these emissions are considered not meaningful and $>$ and $<$ represent elementwise comparisons. Note that the dissociate effect is now defined on principal strata where potential changes (or differences) in the intermediate variables are less than some vector of thresholds $|(\mathbf{M}(1, 1, 1) - \mathbf{M}(0, 0, 0))|_{\mathcal{K}} < C_{\mathcal{K}}^{D}$ instead of principal stratum with strict equality $|(\mathbf{M}(1, 1, 1) - \mathbf{M}(0, 0, 0))|_{\mathcal{K}} = \{0, 0, 0\}_{\mathcal{K}}$ to accommodate continuous intermediate values. For example, $\mathcal{K} = \{1, 3\}$ would be used to define the associative (dissociative) effect in the subpopulation exhibiting an effect on SO_2 and CO_2 in excess of $C_{\mathcal{K}}^{A}$ (below $C_{\mathcal{K}}^{D}$) without regard to the effect on NO_x. For the data analysis in Section 7, we divide the EAE defined above into two parts. $\mathrm{EAE}_{\mathcal{K}}^{+}$ will denote the average associative effects among power plants where all emissions in $\mathcal{K}$ are causally increased in excess of $C_{\mathcal{K}}^{A}$, while $\mathrm{EAE}_{\mathcal{K}}^{-}$ will denote the average associative effect in power plants where all emissions in $\mathcal{K}$ were causally reduced in excess of $C_{\mathcal{K}}^{A}$. Note that these summary quantities only consider a subset of principal strata that may be of interest. For example, analogous average principal effects could be calculated among strata where some emissions are decreased and others are increased. We avoid burdensome notation for such summaries but will revisit estimates in additional principal strata in the context of the data analysis in Section 7.

In addition to estimating average dissociative and associative effects for different $\mathcal{K}$ as defined above, interest may lie in entire surfaces of, for example, how the causal effect on $\text{PM}_{2.5}$ varies as a function of the causal effect on each emission ("causal effect predictiveness" surface (Gilbert and Hudgens (2008))).

4.3. *Observable and a priori counterfactual outcomes*: *Natural direct and indirect effects for multiple mediators.* Extending definitions of natural direct and indirect effects to the multiple mediator setting is somewhat more complicated. The natural direct effect is defined as $\text{NDE} = E[Y(1; \mathbf{M}(0, \dots, 0)) - Y(0; \mathbf{M}(0, \dots, 0))]$, representing the causal effect of Z on Y that is "direct" in the sense that it is not attributable to changes in any of the K emissions. The joint natural indirect effect of all K mediators, $\text{JNIE}_{12\cdots K}$, is derived by subtracting the natural direct effect from the total effect, $\text{JNIE}_{12\cdots K} = \text{TE} - \text{NDE} = E[Y(1; \mathbf{M}(1, 1, \dots, 1)) - Y(1; \mathbf{M}(0, 0, \dots, 0))]$.

In addition to $\text{JNIE}_{12\cdots K}$, we introduce a decomposition into the natural indirect effects attributable to changes in different combinations of the K mediators. Maintaining focus on the case where $K = 3$, the JNIE_{123} can be decomposed into emission-specific indirect effects and the joint indirect effects of all possible pairs of emissions. See Figure 5 in the Appendix in the Supplementary Material (Kim et al. (2019)) for a graphical representation.

We define the *mediator-specific* NIE for the kth emission as a comparison between the potential $\text{PM}_{2.5}$ outcome under scrubbers and the analogous outcome with the value of the kth emission fixed to the natural potential value that would be observed without scrubbers. Specifically, for emissions of SO_2, NO_x and CO_2 define

$$\begin{aligned}
&\text{NIE}_1 = E\big[Y\big(1; \mathbf{M}(1, 1, 1)\big) - Y\big(1; \mathbf{M}(0, 1, 1)\big)\big],\\
(4.1)\qquad &\text{NIE}_2 = E\big[Y\big(1; \mathbf{M}(1, 1, 1)\big) - Y\big(1; \mathbf{M}(1, 0, 1)\big)\big],\\
&\text{NIE}_3 = E\big[Y\big(1; \mathbf{M}(1, 1, 1)\big) - Y\big(1; \mathbf{M}(1, 1, 0)\big)\big].
\end{aligned}$$

In a similar fashion we can define the joint natural indirect effect attributable to subsets of mediators j and k for $j \neq k$ as differences between the observable potential $\text{PM}_{2.5}$ outcomes under scrubbers and the analogous a priori counterfactual with values of pollutants j and k set to their natural values that would be observed without scrubbers. For example, JNIE_{12} defines the joint natural indirect effects of mediators 1 (SO_2) and 2 (NO_x) as

$$\text{JNIE}_{12} = E\big[Y\big(1; \mathbf{M}(1, 1, 1)\big) - Y\big(1; \mathbf{M}(0, 0, 1)\big)\big].$$

Values of JNIE_{jk} for other pairs of mediators can be defined analogously, and all such pairs correspond to the second row in Figure 5 in the Appendix in the Supplementary Material (Kim et al. (2019)). Note that the joint natural indirect effect of each pair of mediators is not equal to the sum of corresponding mediator-specific NIEs unless there is no overlap between mediator-specific NIEs (additivity). For example, we can represent the relationship between JNIE_{12} and the mediator-specific effects NIE_1 and NIE_2 as

$$\begin{aligned}
&(\text{NIE}_1 + \text{NIE}_2) - \text{JNIE}_{12} \\
&\quad = E\big[Y\big(1; \mathbf{M}(1,1,1)\big) - Y\big(1; \mathbf{M}(0,1,1)\big) \\
&\qquad - Y\big(1; \mathbf{M}(1,0,1)\big) + Y\big(1; \mathbf{M}(0,0,1)\big)\big].
\end{aligned}$$

Thus, if this quantity is not equal to 0, we argue that additivity of mediator-specific NIEs does not hold. Note that the above decomposition of JNIE_{123} differs from VanderWeele and Vansteelandt (2014), which considers the portion of the JNIE_{123} mediated through M_1, then sequentially considers the additional contribution of each mediator in the presence of the others. This presumed ordering of mediators precludes estimation of the effect through different pairs of mediators such as JNIE_{23} or JNIE_{13}, the availability of which is a benefit of our proposed decomposition. Our decomposition also differs from Daniel et al. (2015) who only allow interacting overlap between mediator-specific NIEs when one mediator causally affects another.

Note that alternative definitions of NIE could use contrasts of the form, $\text{NIE}_1^* = E[Y(0; \mathbf{M}(1,1,1)) - Y(0; \mathbf{M}(0,1,1))]$. Such a strategy is also considered in Daniel et al. (2015) but defining NIE_k^* in this way would rely entirely on *a priori* counterfactuals, whereas a benefit of using the definitions in (4.1) is that each definition uses the observable potential outcome $Y(1; \mathbf{M}(1,1,1))$, comparing against only one a priori counterfactual (e.g., $Y(1; \mathbf{M}(0,1,1))$).

5. Flexible Bayesian models assumptions and estimation. Under the assumptions developed in this section, Bayesian inference for the causal effects defined in Section 4 follows from specifying models for the joint distribution of all potential mediators (conditional on covariates) and the outcome model conditional on all potential mediators and covariates and prior distributions for unknown parameters. Posterior distributions cannot be computed directly from observed data because potential outcomes are never jointly observed in both the presence and absence of a scrubber and a priori counterfactuals are never observed. Our estimation strategy consists of three steps. First, we specify nonparametric models for the observed data. The marginal distribution of each observed mediator (i.e., $\mathbf{M}(0,0,0) = \{M_1(0), M_2(0), M_3(0)\}$ observed for power plants that did not install scrubbers, and $\mathbf{M}(1,1,1), = \{M_1(1), M_2(1), M_3(1)\}$ observed for those that did) is specified separately and then linked into a coherent joint distribution using a Gaussian copula model (Nelsen (1999)). The models for the potential outcomes

$Y(1; \mathbf{M}(1,1,1))$ and $Y(0; \mathbf{M}(0,0,0))$ are specified conditional on covariates and all potential mediators ($\mathbf{M}(1,1,1)$ and $\mathbf{M}(0,0,0)$) that are never observed simultaneously. Thus, the conditional outcome models are estimated via the data augmentation for unobserved potential mediators. Second, we introduce two assumptions for estimating the TE and the associative and dissociative effects. Third, we employ an additional assumption to equate the distributions of a priori counterfactuals to those of the observed potential outcomes under intervention $Z = 1$ to allow estimation of the natural direct and indirect effects. We also provide optional modeling assumptions to sharpen posterior inference for the power plant evaluation. Throughout we estimate the distribution of the covariates, $F_{\boldsymbol{X}}(\boldsymbol{x})$, using the empirical distribution.

5.1. *Models for the observed data.* We specify Dirichlet process mixtures for the marginal distribution of each mediator (Müller, Erkanli and West (1996)). For each intervention $z = 0, 1$, $k = 1, 2, 3$ and baseline covariates $\boldsymbol{X} = \boldsymbol{x}$, the conditional distribution of the kth observed mediator is specified as

$$
\begin{aligned}
M_{k,i}|Z_i = z, \boldsymbol{X}_i = \boldsymbol{x}_i &\sim N\big(\beta^z_{k0,i} + \boldsymbol{x}_i^\top \boldsymbol{\beta}^z_{k1}, \tau^z_{k,i}\big), \quad M_{k,i} \geq 0; i = 1, \dots, n_z, \\
\beta^z_{k0,i}, \tau^z_{k,i} &\sim F^z_k, \\
F^z_k &\sim DP\big(\lambda^z_k, \mathcal{F}^z_k\big),
\end{aligned}
$$

where $\{i = 1, 2, \dots, n_z\}$ denotes the observations with $Z = z$, and k indicates the kth mediator. We bound the mediator from below (0) using a truncated normal kernel (within the interval $[0, \infty)$). $\beta^z_{k0,i}$ and $\tau^z_{k,i}$ denote the intercept and precision parameters for the kth emission at the ith power plant that received intervention z. Here, DP denotes the Dirichlet process with two parameters, a mass parameter (λ^z_k) and a base measure ($\mathcal{F}^z_k$). To not overly complicate the model, we only 'mixed' over the intercept and precision parameters in the conditional distributions, $\beta^z_{k0,i}$ and $\tau^z_{k,i}$. The base distribution $\mathcal{F}^z_k$ is taken to be the normal-Gamma distribution, $N(\mu^z_k, S^z_k)G(a^z_k, b^z_k)$. Details including hyper prior specification are given in Section A of the Appendix in the Supplementary Material (Kim et al. (2019)).

The marginal distributions of each mediator under each $z = 0, 1$ are linked to model the joint distribution of $[M_1, M_2, M_3|Z = z, X = x]$ with Gaussian copula models of the form

$$
F_{\mathbf{M}(z,z,z)}(\mathbf{m}_{z,z,z}) = \Phi_3\big[\Phi_1^{-1}\{F_{M_1(z)}(m_1)\}, \Phi_1^{-1}\{F_{M_2(z)}(m_2)\}, \Phi_1^{-1}\{F_{M_3(z)}(m_3)\}\big],
$$

where $\mathbf{m}_{z,z,z}$ are values of potential mediators under intervention $Z = z$, and Φ_k is the k-variate standard normal CDF. Note that we elect to model the marginal distribution of each univariate random variable separately and then combine with the Gaussian copula model rather than directly model the joint distributions of $[M_1, M_2, M_3|Z = z, X = x]$. Thus, we allow full flexibility using DP mixtures of (truncated) normals for the marginal distributions (the fit of which can be checked empirically) and use the Gaussian copula to link them to construct the joint distribution of potential mediators. The Gaussian copula model implies some (correlation) structure to the joint distribution of all observable potential outcomes without implying any specific causal structure. Flexibility of this structure derives from the fact that each marginal distribution is modeled as nonparametric with infinite dimensional parameter spaces. The strategy is designed to coalesce with the modeling strategy in Section 5.2. Note that other potential alternatives to link the fixed marginal distributions such as mixtures of marginals (e.g., $H(x_1, x_2) = pF(x_1) + (1-p)G(x_2)$ or $H(x_1, x_2) = \sqrt{F(x_1)G(x_2)}$) do not specify the full joint distribution distribution of (x, y) (Nelsen (1999))), and our method does not limit the number of the mediators in general. While the joint distribution of all potential mediators ($\mathbf{M}(0, 0, 0)$ and $\mathbf{M}(1, 1, 1)$) is also modeled via the same Gaussian copula model, this entails modeling unobserved potential mediators and will be discussed as a part of the assumptions in Section 5.2

To model the distributions of the potential outcomes for each $z = 0, 1$ conditional on all potential mediators and covariates, we use a locally weighted mixture of normal regression models (Müller, Erkanli and West (1996)) that is induced by specifying a *DP* mixture of normals for the joint distribution of the outcome, all mediators and covariates. For each intervention $z = 0, 1$, potential values of all (counterfactual) mediators and baseline covariates $\boldsymbol{X} = \boldsymbol{x}$, the conditional distribution of the observed outcome y_i is specified as

$$\begin{aligned} & f\big(y_i|\mathbf{m}_i(0, 0, 0), \mathbf{m}_i(1, 1, 1), \boldsymbol{x}_i, Z_i = z\big) \\ & \quad = \sum_{l=1}^{\infty} \omega_l^z N\big(y_i, \mathbf{m}_i(0, 0, 0), \mathbf{m}_i(1, 1, 1), \boldsymbol{x}_i|\boldsymbol{\mu}_l^z, \Sigma_l^z\big), \end{aligned}$$

where $\omega_l^z = \gamma_l^z/(\sum_{j=1}^{\infty} \gamma_j^z N(\mathbf{m}_i(0, 0, 0), \mathbf{m}_i(1, 1, 1), \boldsymbol{x}_i|\boldsymbol{\mu}_{j,\backslash 1}^z, \Sigma_{j,(\backslash 1,\backslash 1)}^z))$ and $\boldsymbol{\mu}_{j,\backslash 1}^z$ denote all elements of mean parameters $\boldsymbol{\mu}_j^z$ except for Y_i. Similarly, $\Sigma_{j,(\backslash 1,\backslash 1)}^z$ denotes a submatrix of covariance matrix Σ_j^z formed by deleting the first row and the first column. The weight involves the parameter γ_j^z, where $\gamma_j^z = \gamma_j^{\prime,z} \prod_{h<j}(1 - \gamma_h^{\prime,z})$ and $\gamma_j^{\prime,z} \sim \mathrm{Beta}(1, \alpha^z)$. This flexible conditional model specification is a necessary feature in our case since we allow the outcome model to capture nonlinear and/or interaction effects of the mediators. Note again that this outcome model is conditional on all potential mediators $\{\mathbf{M}(0, 0, 0), \mathbf{M}(1, 1, 1)\}$ which cannot be observed at the same time. We use a similar approach to that used in Schwartz, Li and Mealli (2011) to model the observed outcome distribution conditional on partly missing potential intermediate variables by constructing *complete intermediate data*. Here, we impute unobserved potential mediators for each unit with a data-augmentation approach based on the joint distribution of all potential mediators specified above. Details about hyper prior specification and posterior computation are given in the Appendix in the Supplementary Material (Kim et al. (2019)).

5.2. *Assumptions for estimation of causal effects.* To estimate causal effects based on the model for the observed data specified in Section 5.1, we formulate assumptions relating observed quantities to both observable outcomes and a priori counterfactuals. Denote the conditional distribution $[Y(z;\mathbf{M}(z_1,z_2,z_3))|\mathbf{M}(0,0,0)=\mathbf{m}_{0,0,0},\mathbf{M}(1,1,1)=\mathbf{m}_{1,1,1},\boldsymbol{X}=\boldsymbol{x}]$ with $f_{z,\mathbf{M}(z_1,z_2,z_3)}(y|\mathbf{m}_{0,0,0},\mathbf{m}_{1,1,1},\boldsymbol{x})$ where $\mathbf{m}_{z_1,z_2,z_3}$ is a vector of hypothetical values of the mediators under the interventions z_1, z_2, z_3. The conditional distribution $[\mathbf{M}(z_1,z_2,z_3)|\boldsymbol{X}=\boldsymbol{x}]$ is denoted by $f_{\mathbf{M}(z_1,z_2,z_3)}(\mathbf{m}_{z_1,z_2,z_3}|\boldsymbol{x})$. Other conditional distributions are defined analogously, and we henceforth omit conditioning on covariates $\mathbf{X}=\boldsymbol{x}$ to simplify notation.

5.2.1. *Assumptions for principal causal effects.* We begin with an ignorability assumption stating that, conditional on covariates, "assignment" to scrubbers is unrelated to the observable potential outcomes:

ASSUMPTION 1 (Ignorable treatment assignment).

$$\{Y(z;\mathbf{M}(z,z,z)),\mathbf{M}(0,0,0),\mathbf{M}(1,1,1)\}\perp\!\!\!\perp Z|\boldsymbol{X}=\boldsymbol{x},$$

for $z=0,1$. This assumption permits estimation of the distributions of potential outcomes under intervention $Z=z$ with observed data on ambient $\mathrm{PM}_{2.5}$ and emissions under the same intervention.

We adopt a Gaussian copula model to link the distributions of $(M_1(z), M_2(z), M_3(z))$ for $z=0,1$ into a single joint distribution of observable potential outcomes.

ASSUMPTION 2. The joint distribution of all potential mediators conditional on covariates follows a Gaussian copula model (Nelsen (1999)):

$$\begin{aligned}&F_{\mathbf{M}(0,0,0),\mathbf{M}(1,1,1)}(\mathbf{m}_{0,0,0},\mathbf{m}_{1,1,1})\\&\quad=\Phi_6[\Phi_1^{-1}\{F_{M_1(0)}(m_1)\},\Phi_1^{-1}\{F_{M_2(0)}(m_2)\},\Phi_1^{-1}\{F_{M_3(0)}(m_3)\},\\&\qquad\Phi_1^{-1}\{F_{M_1(1)}(m_1)\},\Phi_1^{-1}\{F_{M_2(1)}(m_2)\},\Phi_1^{-1}\{F_{M_3(1)}(m_3)\}],\end{aligned}$$

where Φ_6 is the multivariate normal CDF with mean $\mathbf{0}$ and a correlation matrix $\boldsymbol{R}$.

Assumption 2 implies a joint distribution of all observable potential mediators in a manner consistent with the models for $[M_1, M_2, M_3|Z=z, X=x]$ described in Section 5.1. However, this entire joint distribution of potential mediators under both interventions is not fully identified from the data since potential mediators under different interventions are never jointly observed. Specifically, entries of the correlation matrix $\boldsymbol{R}$ corresponding to, for example, the correlation between $M_j(0)$ and $M_k(1)$, are not identifiable in the sense that no amount of data can estimate unique values for these parameters. Nonetheless, proper prior distributions for these parameters can still permit inference from proper posterior distributions. Such parameters are sometimes referred to as "partially identifiable" in the sense that increasing amounts of data may lead the supports of posterior distributions to converge to sets of values that are smaller than those specified in the prior distribution (Gustafson (2010), Mealli and Pacini (2013)). This can arise due to restrictions on the joint distribution implied by the models for the marginal distributions (e.g., the positive-definiteness restriction on $\boldsymbol{R}$ may exclude some possible values for its entries). We discuss two prior specifications for the partially identified parameters in $\boldsymbol{R}$, noting that further details of partial identifiability in the principal stratification context appear in Schwartz, Li and Mealli (2011).

5.2.2. *Assumptions for mediation effects.* Toward estimation of natural direct and indirect effects, we augment the assumptions of Section 5.2.1 with one relating observable outcomes to a priori counterfactual outcomes.

ASSUMPTION 3. For intervention $Z = 1$, the conditional distribution of the potential outcome given values of all potential mediators (and covariates) is the same regardless of whether the mediator values were induced by $Z = 1$ or $Z = 0$.

This assumption implies that the a priori counterfactual $Y(1; \mathbf{M}(0, 0, 0))$ and the observable potential outcomes $Y(1; \mathbf{M}(1, 1, 1))$ have the same conditional distribution,

$$\begin{aligned} &f_{1,\mathbf{M}(0,0,0)}\big(y|\mathbf{M}(0, 0, 0) = \mathbf{m}, \mathbf{M}(1, 1, 1), \boldsymbol{x}\big) \\ &\quad = f_{1,\mathbf{M}(1,1,1)}\big(y|\mathbf{M}(0, 0, 0), \mathbf{M}(1, 1, 1) = \mathbf{m}, \boldsymbol{x}\big). \end{aligned}$$

This assumption also applies to any two mediators in the absence of the intervention. For instance, the a priori counterfactual of $PM_{2.5}$, $Y(1; \mathbf{M}(0, 1, 0))$ and $Y(1; \mathbf{M}(1, 1, 1))$, have the same conditional distribution regardless of whether corresponding emissions values arose under a scrubber ($Z = 1$) or absent a scrubber ($Z = 0$),

$$\begin{aligned} &f_{1,\mathbf{M}(0,1,0)}\big(y|\mathbf{M}(0, 1, 0) = \mathbf{m}, \mathbf{M}(1, 0, 1), \boldsymbol{x}\big) \\ &\quad = f_{1,\mathbf{M}(1,1,1)}\big(y|\mathbf{M}(0, 0, 0), \mathbf{M}(1, 1, 1) = \mathbf{m}, \boldsymbol{x}\big). \end{aligned}$$

The key point is that the distribution of $PM_{2.5}$ under a given (unobservable) combination of mediators ($\mathbf{m}$) only depends on the values of the mediators and not the intervention that led to those mediators. Asserting this assumption in this case relies in part on what is known about the underlying chemistry relating SO_2, NO_x, and CO_2 emissions to $PM_{2.5}$. Note that such an assumption may be more difficult to justify in, say, a clinical study where assumptions about a priori counterfactuals might pertain to choices of study participants.

The above assumption can be cast as two homogeneity assumptions of the form proposed in Forastiere, Mealli and VanderWeele (2016). For example, one implication of Assumption 3 is that the a priori counterfactual $Y(1; M(0,0,0))$ is homogeneous across all principal strata with $M(0,0,0) = \mathbf{m}$, regardless of the value of $M(1,1,1)$. Viewing Assumption 3 in terms of the implied homogeneity across principal strata aids interpretation and justification in the context of the power plant example. Homogeneity across strata implies that the potential ambient air quality value in the area surrounding a power plant is related to (possibly counterfactual) emission levels only and not to the power plant characteristics that govern effectiveness of scrubbers for reducing emissions (i.e., the power plant characteristics that determine the exact principal stratum membership). This underscores the importance of including covariates in $\boldsymbol{X}$ that capture characteristics of the monitoring locations (e.g., temperature and barometric pressure). Appendix D provides details of the relationship between Assumption 3 and assumptions of homogeneity across principal strata. While Assumption 3 implies homogeneity assumptions, the converse is not true in the case of multiple mediators due to the connection of Assumption 3 to a priori counterfactuals defined to have mediator values induced by different interventions (e.g., $Y(1; M(0,1,0))$). We discuss a sensitivity analysis to this assumption in Appendix J in the Supplementary Material (Kim et al. (2019)).

5.2.3. *Optional modeling assumptions to sharpen posterior inference.* With the above model specification the partial identifiability of the model parameters in $\boldsymbol{R}$ warrants careful attention. Proper but noninformative prior distributions for these parameters could be specified marginally for these parameters as $\text{Unif}(-1,1)$, or equivalently, as conditionally uniform on intervals satisfying positive definiteness restrictions for the correlation matrix. In either case posterior inference may exhibit large uncertainty.

We consider in detail an alternative prior specification similar to that in Zigler, Dominici and Wang (2012) to sharpen posterior inference. Specifically, the correlations between mediators under different interventions are specified as follows:

$$\begin{aligned}\text{Cor}\big(M_j(0), M_k(1)\big) &= \frac{\text{Cor}(M_j(0), M_k(0)) + \text{Cor}(M_j(1), M_k(1))}{2} \times \rho,\\ &\text{for } j,k = 1,2,3,\end{aligned}$$

with ρ a sensitivity parameter. This strategy implies that (a) the correlation between the same mediator ($j = k$) under opposite interventions is ρ, and (b) the correlation between different mediators ($j \neq k$) under opposite interventions is an attenuated version of the correlation observed separately under each intervention. Section B of the Appendix in the Supplementary Material (Kim et al. (2019)) provides a correlation matrix implied by this assumption in the case of two mediators. We assume a single ρ and specify a uniform prior distribution, $\rho \sim \text{Unif}(0,1)$, but a different parameter could be specified for each mediator.

As an additional assumption to sharpen posterior inference, we assume that the correlations between emissions (mediators) are all positive. Support for this assumption comes from observed-data estimates of these conditional correlations that are all positive.

In summary, Assumptions 1–2 are sufficient to estimate the principal causal effects, and pertain only to observable potential outcomes. Adding Assumption 3 relating observed quantities to a priori counterfactuals permits estimation of direct and indirect effects for mediation analysis. The optional assumptions here in Section 5.2.3 are designed to sharpen posterior inference in the power plant analysis.

5.2.4. *Posterior inference.* A Markov chain Monte Carlo (MCMC) algorithm is used to sample from this posterior distribution and estimate causal effects using the following steps: (1) sampling parameters from each marginal distribution for potential mediators and conditional distribution for potential outcomes defined in Section 5.1; (2) sampling parameters from the correlation matrix R of the Gaussian copula; (3) sampling via data augmentation a priori counterfactual mediators from the joint distribution; (4) computing causal effects based on all potential mediators and outcomes including imputed a priori outcomes and mediators; (5) iterate Steps 1–4. The specifics of estimation (conditional on our specific model formulation) are based on the existing literature on Bayesian estimation of causal effects (and principal causal effects in particular), for example, in Daniels et al. (2012), Mattei and Mealli (2011), Zigler, Dominici and Wang (2012).

The Appendix in the Supplementary Material (Kim et al. (2019)) contains details of the MCMC procedure (Section F), prior specification for all other model hyper-parameters (Section A) and the procedure for computing the principal causal effects and the mediation effects from the posterior distributions of model parameters (Section C).

6. Numerical study.

We examine the performance of the proposed model under combinations of the following two data generating scenarios: (1) correlations among the mediators (Case 1: uncorrelated mediators vs. Case 2: correlated mediators), and (2) interaction terms between the mediators in the outcome model (Case A: interaction term between M_1 and M_2 vs. Case B: interaction terms between M_1 and M_2 and between M_2 and M_3). Data sets of size $n = 500$ are simulated for each of the four cases (1/A, 1/B, 2/A, 2/B), each with three continuous confounders. In all cases the three mediators are generated based on a multivariate normal distribution. See the Appendix (Section G) in the Supplementary Material (Kim et al. (2019)) for the exact data generating mechanism.

We compare our method for estimating mediation effects to a regression-based model (MacKinnon (2008)):

$$M_1 = \alpha_{01} + \alpha_{11} Z + \boldsymbol{X}^\top \boldsymbol{\alpha}_1 + \epsilon_1,$$

$$M_2 = \alpha_{02} + \alpha_{12} Z + \boldsymbol{X}^\top \boldsymbol{\alpha}_2 + \epsilon_2,$$

$$M_3 = \alpha_{03} + \alpha_{13} Z + \boldsymbol{X}^\top \boldsymbol{\alpha}_3 + \epsilon_3,$$
$$Y = \beta_0 + \beta_1 Z + \beta_2 M_1 + \beta_3 M_2 + \beta_4 M_3 + \boldsymbol{X}^\top \boldsymbol{\beta} + \epsilon_Y,$$

where ϵ_1, ϵ_2, ϵ_3 and ϵ_Y are all independently distributed as $N(0, \sigma)$.

Table 2 summarizes the results based on 400 replications for each of the four scenarios. It shows that our proposed model (BNP) performs well in terms of bias and MSE for all cases. Note that the true effects change when the mediators are correlated in the presence of interaction term(s) in the outcome model. Thus, with any interaction effects of the mediators, it is desirable to capture the correlation structure of the mediators, which our method does by flexibly modeling the joint distribution of all potential mediators. Also, the flexible Bayesian nonparametric model can capture both complex relationships/interactions among the mediators and nonadditive and nonlinear forms of mediators and/or confounders in the outcome model. In each scenario interaction terms in the outcome model introduce non-

TABLE 2

Simulation results for point estimators of causal mediation and principal causal effects over 400 *replications. The columns correspond to bias and MSE relative to the true values of the causal effects for each scenario* (*Cases* 1 *and* 2, *and Cases A and B*) *under two different models*; **Parametric**: *a regression based model for the causal mediation effects*; **BNP**: *Our Bayesian nonparametric method*

		Case 1					Case 2				
			BNP		Parametric			BNP		Parametric	
		Truth	Bias	MSE	Bias	MSE	Truth	Bias	MSE	Bias	MSE
Case A	TE	0.73	0.02	(0.09)	−0.03	(0.08)	0.92	−0.04	(0.08)	**0.20**	(0.33)
	JNIE	1.73	0.06	(0.11)	**0.21**	(0.07)	1.92	0.04	(0.08)	0.02	(0.47)
	NDE	−1	−0.04	(0.01)	**−0.25**	(0.15)	−1	−0.08	(0.01)	**−0.20**	(0.08)
	NIE$_1$	−0.16	0.00	(0.00)	−0.01	(0.01)	0.03	−0.05	(0.00)	**−0.38**	(0.26)
	NIE$_2$	2.45	0.02	(0.10)	−0.02	(0.08)	2.65	−0.05	(0.08)	**−0.39**	(0.31)
	NIE$_3$	−0.32	0.00	(0.00)	−0.01	(0.01)	−0.32	0.01	(0.00)	−0.01	(0.01)
	JNIE$_{12}$	2.05	0.05	(0.10)	**0.22**	(0.14)	2.23	0.03	(0.08)	**0.21**	(0.44)
	JNIE$_{13}$	−0.48	0.01	(0.01)	−0.01	(0.01)	−0.29	−0.04	(0.00)	**−0.38**	(0.28)
	JNIE$_{23}$	2.13	0.02	(0.10)	−0.02	(0.09)	2.33	−0.04	(0.08)	**−0.39**	(0.33)
Case B	TE	1.08	−0.02	(0.10)	−0.01	(0.08)	1.33	−0.09	(0.08)	−0.01	(0.08)
	JNIE	2.08	−0.00	(0.10)	**0.16**	(0.12)	2.33	−0.00	(0.08)	−0.08	(0.11)
	NDE	−1	−0.01	(0.00)	**−0.17**	(0.04)	−1	−0.09	(0.01)	0.08	(0.02)
	NIE$_1$	−0.16	−0.01	(0.00)	−0.01	(0.01)	0.03	−0.05	(0.01)	**−0.20**	(0.04)
	NIE$_2$	2.51	−0.02	(0.10)	0.02	(0.09)	2.78	−0.08	(0.09)	**−0.25**	(0.15)
	NIE$_3$	−0.13	0.00	(0.00)	0.01	(0.01)	−0.05	−0.02	(0.01)	−0.08	(0.01)
	JNIE$_{12}$	2.11	0.01	(0.10)	**0.25**	(0.16)	2.37	0.00	(0.08)	0.01	(0.10)
	JNIE$_{13}$	−0.29	−0.00	(0.00)	−0.01	(0.01)	−0.02	−0.07	(0.01)	**−0.27**	(0.08)
	JNIE$_{23}$	2.48	−0.04	(0.10)	−0.08	(0.09)	2.75	−0.09	(0.09)	**−0.34**	(0.21)

additivity in the joint natural indirect effect (e.g., $\mathrm{JNIE} \neq \mathrm{NIE}_1 + \mathrm{NIE}_2 + \mathrm{NIE}_3$), and the traditional regression model has larger biases (and larger MSEs) for mediation effects.

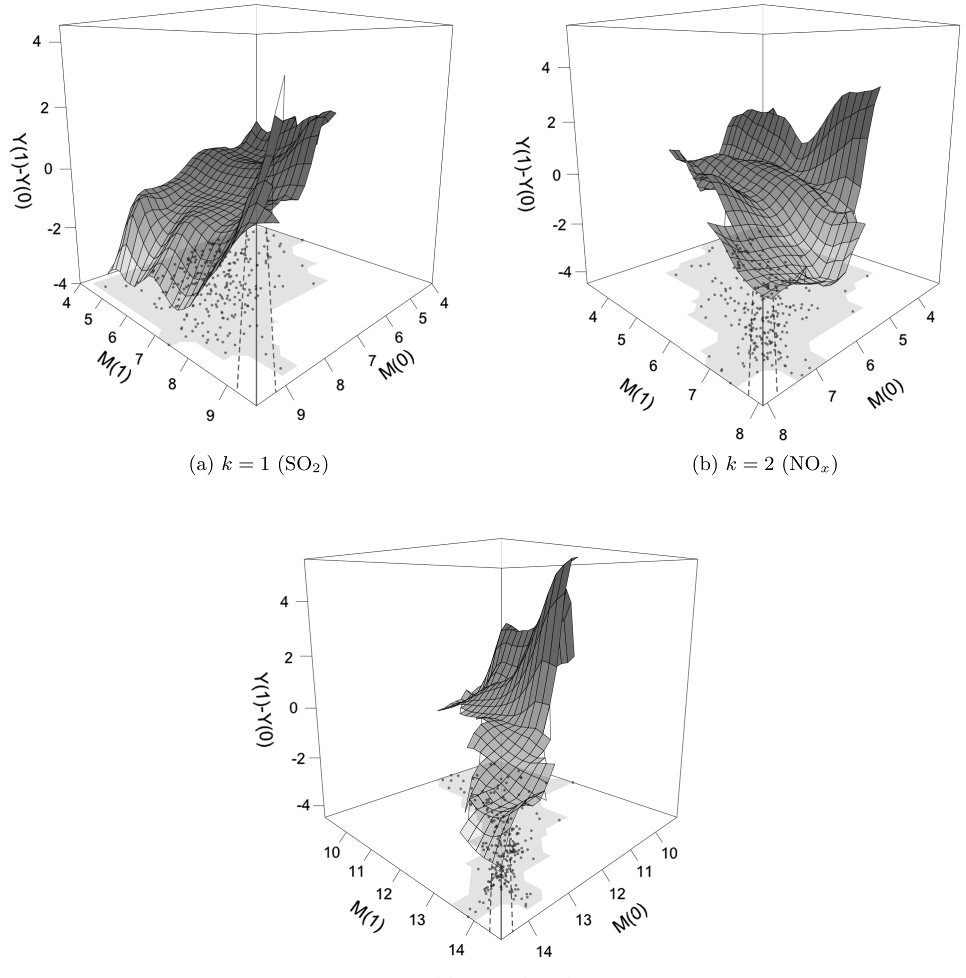
7. Analysis of power plant scrubbers in the acid rain program. Here, we estimate causal effects of having scrubbers installed in January 2005 (Z) on annual average emissions of SO_2, NO_x and CO_2 in 2005 (M_1, M_2, M_3) and on the 2005 annual average ambient $PM_{2.5}$ concentration within 150 km of a power plant (Y). Before reporting results note that basic checks of the fit of marginal nonparametric models appear in Appendix I in the Supplementary Material (Kim et al. (2019)), indicating fit that is clearly superior to simple parametric models.

A simple comparison of means indicates that the 150 km area around power plants with scrubbers installed ($Z = 1$) had average ambient $PM_{2.5}$ that was lower, on average, than the areas surrounding power plants without scrubbers (12.4 vs. 13.7 $\mu g/m^3$). Similarly, the power plants with scrubbers also emitted less SO_2, more NO_x and more CO_2 than the plants without scrubbers. Table 1 lists the covariates in X to adjust for confounding and presents summary statistics for scrubber and nonscrubber power plants.

We present an analysis with the proposed method using the constrained prior specification in Section 5.2.3. Analysis using uniform prior distributions on all elements of the correlation matrix appears in the Appendix in the Supplementary Material (Kim et al. (2019)). All reported estimates are listed as posterior means (95% posterior intervals). The analysis estimates that having scrubbers installed causes SO_2 emissions to be -1.17 (-1.86, 1.55) ×1000 tons lower, on average, than they would be without the scrubber. The analogous causal effects for NO_x and CO_2 emissions were 0.04 (0.00, 0.07) ×1000 tons and 0.001 (-0.00, 0.004) million tons respectively, indicating that scrubbers did not significantly affect these emissions on average. The total effect (TE) of having scrubbers installed on ambient $PM_{2.5}$ within 150 km is estimated to be -1.12 (-2.07, -0.29) $\mu g/m^3$, suggesting a reduction amounting to approximately 10% of the national annual regulatory standard for $PM_{2.5}$.

7.1. *Principal causal effects.* For the kth emission let σ_k denote the posterior standard deviation of the estimated individual-level causal effect of a scrubber on M_k with posterior mean estimates $\hat{\sigma}_1 = 0.24$, $\hat{\sigma}_2 = 0.42$, $\hat{\sigma}_3 = 0.02$. Let $\hat{\sigma}_{\mathcal{K}}$ denote the vector of $\hat{\sigma}_k$ for the emissions in $\mathcal{K}$. To summarize dissociative effects, we set $C^D_{\mathcal{K}} = 0.25\hat{\sigma}_{\mathcal{K}}$ to estimate $\mathrm{EDE}_{\mathcal{K}}$ among power plants where the scrubber effect on emissions in $\mathcal{K}$ is within one-fourth of a standard deviation of the effect in the population. Similarly, we summarize associative effects with $C^A_{\mathcal{K}} = 0.25\hat{\sigma}_{\mathcal{K}}$ to estimate $\mathrm{EAE}^-_{\mathcal{K}}$ ($\mathrm{EAE}^+_{\mathcal{K}}$) among power plants where the scrubber causally reduces (increases) emissions in $\mathcal{K}$ more than one-fourth of a standard deviation of the effect in the population.

Before providing estimates of specific principal effects, we first examine 3-D surface plots in Figure 2. For each emission separately ($k \in \{1, 2, 3\}$), Figure 2 depicts estimated scrubber effects on $\text{PM}_{2.5}$ across varying effects on emissions determined by values of $(M_k(0), M_k(1))$ simulated from the model. Note the pattern for all emissions that the surfaces are sloped downward in the direction of increasing $M_k(0)$ and $M_k(1)$ (sloped toward the viewer), indicating larger effects



(a) $k = 1$ (SO_2)

(b) $k = 2$ (NO_x)

(c) $k = 3$ (CO_2)

FIG. 2. *Average surface plots of the causal effect on* $\text{PM}_{2.5}$ *for different values of* $(M_k(0), M_k(1))$. *Values of* $(M_k(0), M_k(1))$ *are plotted on the* x*- and* y*-axes and determine the causal effect of a scrubber on emission* k. *The corresponding value of the causal effect of a scrubber on* $\text{PM}_{2.5}$, $Y(1) - Y(0)$, *is plotted on the* z*-axis. The cloud of points in the* xy*-plane are one MCMC draw of* 249 *pairs of* $(M_k(0), M_k(1))$. *The lines on the* xy*-plane are at* $M_k(0) = M_k(1)$ (*solid line*) *and* $+/- 0.25\hat{\sigma}_k$ (*dashed lines*).

on $PM_{2.5}$ among plants with larger emissions values under both scrubber statuses, that is, larger plants.

In Figure 2(a) for SO_2, the dots in the xy-plane lie almost entirely in the region where $M_1(1) < M_1(0)$, indicating as expected that scrubbers predominantly decrease SO_2 emissions. Associative effects for SO_2 are indicated by the downward slope of the surface in the direction of decreasing $M_1(1) - M_1(0)$ (toward the left of the viewer), indicating that larger decreases (increases) in SO_2 are associated with larger decreases (increases) in $PM_{2.5}$.

The analogous surfaces for NO_x and CO_2 appear in Figures 2(b) and 2(c) respectively. In contrast to the surface for SO_2, the dots in the xy-plane fall more closely and symmetrically around the line $M_k(1) = M_k(0)$, reflecting that scrubbers do not affect these emissions on average. The surface for NO_x exhibits some evidence of associative effects in the opposite direction of those for SO_2; there is some downward slope of the surface in the direction of increasing $M_k(1) - M_k(0)$ (toward the right of the viewer), indicating that larger increases (decreases) in these emissions are associated with larger decreases (increases) in $PM_{2.5}$.

Table 3 lists posterior mean and standard deviation of EDE, EAE^- and EAE^+ for all possible $\mathcal{K}$. Estimates of EDE for all $\mathcal{K}$ indicate little to no reduction in $PM_{2.5}$ among plants where emissions were not affected in excess of $\mathcal{C}^D_{\mathcal{K}}$, with the exception of some pronounced estimates of EDE for $\mathcal{K} = \{NO_x\}$ and $\mathcal{K} = \{CO_2\}$. Estimates of EAE^- and EAE^+ tend to be less than zero. The most pronounced estimate of $\text{EAE}^-_{\mathcal{K}} = -1.19$ (0.46) for $\mathcal{K} = \{SO_2\}$ suggests that $PM_{2.5}$ was reduced among power plants where SO_2 emissions were substantially reduced which corresponds to the contour of the surface in Figure 2(a) and is consistent with the anticipated causal pathway whereby scrubbers reduce $PM_{2.5}$ through reducing SO_2 emissions. In accordance with the opposite sloping surface in Figures 2(b), the estimate of $\text{EAE}^+_{\mathcal{K}}$ is most pronounced for $\mathcal{K} = \{NO_x\}$ and $\{NO_x, CO_2\}$, indicating that ambient $PM_{2.5}$ is decreased among plants with substantial *increases* in NO_x emissions.

Recall that the estimates in Table 3 represent average principal effects over only a subset of principal strata, in particular those where changes in multiple emissions are concordant (i.e., all decreasing, all increasing or none changing). Other strata may be of interest. Figure 3 provides estimates of principal effects in a cross-classification of strata defined by changes in CO_2 and SO_2, with changes defined as increases, decreases or no change in reference to $C^D_{\mathcal{K}}$ and $C^A_{\mathcal{K}}$. For example, the third column of Figure 3 subdivides the stratum defined by causal increases in CO_2 into three substrata: (1) those where CO_2 increases and SO_2 decreases (in excess of $C^A_{\mathcal{K}}$); (2) does not substantially change (beyond $C^D_{\mathcal{K}}$); or (3) increases (in excess of $C^A_{\mathcal{K}}$). Principal causal effect estimates for these three substrata appear along with their relative proportion among the stratum defined by CO_2 increases, indicated by the size of the plotting symbol. The light grey dot corresponds to $\text{EAE}^+_{\mathcal{K}}$ for $\mathcal{K} = \{SO_2, CO_2\}$ as reported in Table 3, but note that only 4% of the CO_2-increase stratum exhibits SO_2 increases. The dark grey dot corresponds to the principal

TABLE 3
Posterior means (standard deviations) for expected associative and dissociative effects of SO_2 *scrubbers*

		SO_2	NO_x	CO_2	SO_2 & NO_x	SO_2 & CO_2	NO_x & CO_2	SO_2 & NO_x & CO_2
EAE^-	Mean	−1.19	−0.77	−1.14	−0.84	−1.18	−0.90	−0.94
	SD	(0.46)	(0.59)	(0.56)	(0.59)	(0.57)	(0.67)	(0.68)
EDE	Mean	−0.32	−0.69	−0.82	−0.09	−0.31	−0.48	−0.15
	SD	(0.57)	(0.54)	(0.49)	(0.71)	(0.68)	(0.69)	(0.86)
EAE^+	Mean	0.60	−1.68	−1.08	0.38	1.28	−1.63	0.69
	SD	(2.52)	(0.74)	(0.75)	(3.67)	(3.78)	(1.04)	(4.68)

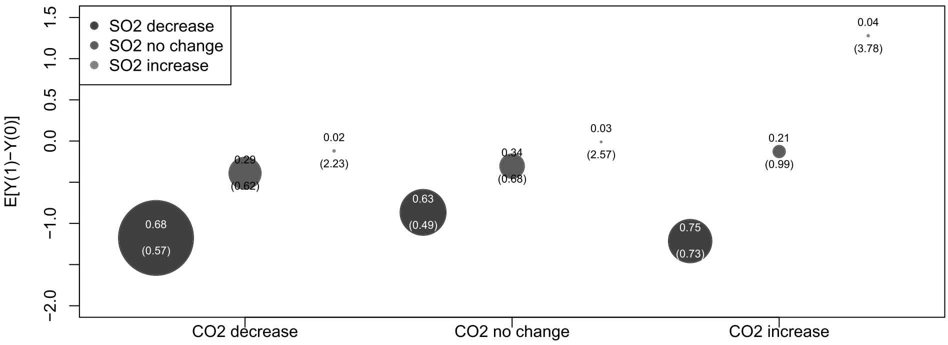


FIG. 3. *Posterior mean estimates of principal effects for strata defined by cross-classifying changes in* CO_2 *(x-axis) and changes in* SO_2 *(colored circles). Size of circle symbolizes the proportion of each* CO_2 *stratum falling in the corresponding* SO_2 *category, and number (number in parentheses) listed is posterior mean proportion (and posterior standard deviation).*

effect among the 21% of the CO_2-increase stratum in substratum (2) where SO_2 does not change, with a principal effect estimate of -0.13 (0.99). The remaining proportion (75%) of the CO_2-increase stratum belongs to substratum (3) where the plants exhibiting decreases in SO_2 and a corresponding principal effect estimate of -1.21 (0.73). Thus, for $\mathcal{K} = \{CO_2\}$, the negative estimate of $\mathrm{EAE}^{+}_{\mathcal{K}}$ from Table 3 is revealed to be generated in large part by strata where SO_2 decreases and there is a pronounced negative effect on $PM_{2.5}$. Analogously, the second column of Figure 3 considering the stratum where CO_2 emissions do not substantially change (used to estimate EDE) reveals that 63% of this strata exhibited causal reduction in SO_2 and a causal reduction in $PM_{2.5}$ of -0.87 (0.49), explaining in large part the negative estimate of $\mathrm{EDE}_{\mathcal{K}}$ for $\mathcal{K} = \{CO_2\}$ in Table 3. Analogous cross-classification of strata by changes in NO_x and SO_2 appears very similar to Figure 3 and is not presented.

The main conclusions from the principal stratification analysis are that: (1) scrubbers reduce SO_2 on average but not NO_x or CO_2; (2) there is some evidence of a nonzero dissociative effect for SO_2; (3) associative effects for SO_2 are more pronounced than dissociative effects, with $PM_{2.5}$ reduced more around plants where scrubbers cause large reductions in SO_2; (4) associative effects for NO_x and CO_2 are more pronounced than dissociative effects, with $PM_{2.5}$ reduced more around plants where scrubbers cause larger *increases* in these emissions; but that (5) strata defined by increases (or no change) in NO_x and/or CO_2 are comprised in large part by substrata where SO_2 and $PM_{2.5}$ were causally reduced. This analysis points toward (but cannot confirm) the conclusion that scrubbers affect $PM_{2.5}$ among plants where emissions are not changed and that scrubber effects on $PM_{2.5}$ are mediated in part through effects on SO_2 with less evidence of a mediating role of NO_x and CO_2.

7.2. *Mediation effects*. To estimate direct and indirect effects, we augment the principal stratification analysis with Assumption 3 in Section 5.2.2 about a priori counterfactuals. Figure 1 (top) in the Supplementary Material Appendix (Kim et al. (2019)) depicts boxplots of the posterior distributions of TE, NDE, JNIE_{123}, JNIE_{12}, JNIE_{23}, JNIE_{13} and the individual NIEs. The estimated NDE, representing the direct effect of a scrubber on ambient $\mathrm{PM}_{2.5}$ that is not mediated through any emissions changes, is -0.53 $(-1.51, 0.39)$ μg /m^3, indicating no evidence of a direct effect of scrubbers on $\mathrm{PM}_{2.5}$ that is not mediated through SO_2, NO_x or CO_2. The NIEs for NO_x (NIE_2) and CO_2 (NIE_3) are estimated to be very close to 0, -0.02 $(-0.26, 0.21)$ and -0.04 $(-0.33, 0.23)$ respectively. The estimated NIE for SO_2 (NIE_1) is -0.54 $(-1.20, -0.01)$, indicating a significant indirect effect. The joint natural indirect effects involving SO_2 are all similar in magnitude to NIE_1, with estimates of JNIE_{12}, JNIE_{13} and JNIE_{123} of -0.56 $(-1.23, -0.01)$, -0.58 $(-1.25, -0.02)$ and -0.59 $(-1.27, -0.02)$ respectively. The estimated JNIE_{23} is -0.03 $(-0.31, 0.23)$.

As discussed in Section 4.3, a benefit of the proposed approach is the accommodation of overlap between NIEs and the opportunity to examine the extent of overlap. We evaluate the relationship between the joint effects JNIE_{jk} and the mediator-specific effects NIE_1, NIE_2, NIE_3 through $(\mathrm{NIE}_1 + \mathrm{NIE}_2) - \mathrm{JNIE}_{12} = -0.01(-0.18, 0.16)$, $(\mathrm{NIE}_1 + \mathrm{NIE}_3) - \mathrm{JNIE}_{13} = 0.01(-0.22, 0.23)$ and $(\mathrm{NIE}_2 + \mathrm{NIE}_3) - \mathrm{JNIE}_{23} = 0.00(-0.19, 0.15)$ which give no evidence of overlap between NIEs. That is, the effect of a scrubber on ambient $\mathrm{PM}_{2.5}$ that is mediated through emissions changes appears to be described by indirect effects that act additively and do not exhibit any apparent synergy that would lead to overlapping effects. The lack of overlapping indirect effects, combined with the fact that: (a) all indirect effects involving SO_2 (NIE_1, JNIE_{12}, JNIE_{13} and JNIE_{123}) are similar in magnitude, and (b) all indirect effects not involving SO_2 (NIE_2, NIE_3, JNIE_{23}) are estimated to be zero, provides strong evidence that the effect of scrubbers on $\mathrm{PM}_{2.5}$ is primarily driven by effects on SO_2.

In the Appendix in the Supplementary Material (Kim et al. (2019)), we also conduct inference using flat priors on plausible values of the partially identifiable parameters, and the estimates for the effects are similar to those in the main analysis.

The conclusions of the causal mediation analysis are clear and mostly consistent with those from the principal stratification analysis; scrubber effects on ambient $\mathrm{PM}_{2.5}$ are almost entirely mediated through reductions in SO_2 emissions. Combining reductions in SO_2 with reductions of NO_x and CO_2 does not significantly change the mediated effect. In fact NO_x and CO_2 appear to play no role in the causal effect of scrubbers on $\mathrm{PM}_{2.5}$.

7.3. *Results from alternative analyses*. We conduct two simpler analyses for comparison. First, we implement separate single-mediator analyses using the methods described above with $K = 1$. Results are largely consistent with the multiple mediator analysis, as suggested by the apparent absence of overlapping effects. For SO_2 emissions the total indirect and direct effects are estimated to be

-1.28 (-2.25, -0.62), -0.70 (-1.51, -0.04) and -0.58 (-1.35, 0.37) respectively. For NO_x emissions the total indirect and direct effects are estimated to be -1.21 (-2.05, -0.40), -0.04 (-0.32, 0.28) and -1.17 (-1.99, -0.32) respectively. With CO_2 emissions the total indirect and direct effects are estimated to be -1.22 (-1.98, -0.29), 0.03 (-0.26, 0.33) and -1.25 (-2.05, -0.30) respectively. Note that significant estimated direct effects for NO_x and CO_2 suggest pathways that are not through NO_x and CO_2 (i.e., the pathway through SO_2).

For a second comparison we conduct a multiple mediator analysis using a traditional regression approach to mediation with the same model in Section 6. The mediation effects are estimated to be $\mathrm{NIE}_1 = \alpha_{11}\beta_2 = -0.39(95\%\mathrm{C.I.}\ -1.11, 0.25)$, $\mathrm{NIE}_2 = \alpha_{12}\beta_3 = -0.09(95\%\mathrm{C.I.}\ -0.44, 0.22)$, $\mathrm{NIE}_3 = \alpha_{13}\beta_4 = 0.08(95\%\mathrm{C.I.}\ -0.08, 0.35)$, $\mathrm{NDE} = \beta_1 = -0.18(95\%\mathrm{C.I.}\ -2.56, 0.11)$. Thus, while these results are on average consistent with the results from the proposed methods, the estimate of the NIE_1 is not significant. Note that this analysis explicitly assumes that the mediators do not interact with each other in the outcome model, implying an estimate of the joint indirect effect of all three mediators that is the sum of all three indirect effect (i.e., $\mathrm{JNIE}_{123} = -0.40(95\%\mathrm{C.I.}\ -1.15, 0.34)$) which is also not significant. The discrepancy between the results of the traditional regression approach and ours is due to our flexible modeling strategy using Bayesian nonparametric methods (Dirichlet process mixtures) that, even in presence of additivity, allows for nonlinearities and non-normal errors.

8. Discussion. We have developed flexible Bayesian methods for principal stratification and causal mediation analysis in the presence of multiple mediating variables. To accommodate the setting of multiple pollutants that are emitted contemporaneously and possibly interact with one another, we have developed methods to accommodate multiple contemporaneous and nonindependent mediators. Bayesian nonparametric modeling approaches provided flexible models for the observed data (marginal distribution for each mediator and conditional distribution for the outcome under each intervention $z = 0, 1$) and linked observed data distributions to joint distributions of potential mediators using explicit and transparent assumptions about both observable and a priori counterfactuals.

A key feature of our approach is the integration of principal stratification and causal mediation analysis in a manner that relies on the same models for the observed data. Deployment of these methods in the power plant analysis represents, to our knowledge, the most comprehensive consideration of these two approaches and the implications of the results in the context of a single analysis. We use Assumption 3 to relate a priori counterfactual outcomes to observed outcomes and show that this assumption implies homogeneity across principal strata which aids interpretation. This assumption also has close ties to that of sequential ignorability (Imai, Keele and Yamamoto (2010)). Benefits of formulating Assumption 3 as done here include facilitation of a sensitivity analysis to this assumption following the general approach of Daniels et al. (2012) and the aided interpretation implied by the relationship to homogeneity assumptions. While a version of sequential ignorability relevant to the setting of multiple contemporaneous mediators with interactions and that can be used to identify each mediator-specific effect has not been previously formulated, Appendix E in the Supplementary Material (Kim et al. (2019)) explores the relationship between our Assumption 3 and sequential ignorability in the case of a single mediator. In this case implications of these two assumptions are identical for the types of estimands considered here, although one assumption does not generally imply the other.

The results of the principal stratification and causal mediation analyses should be interpreted jointly and are, in this case study, largely consistent with one another. Principal stratification indicated that scrubbers tended to decrease ambient $PM_{2.5}$ around plants where scrubbers substantially reduced SO_2 emissions, a result consistent with the estimated natural indirect effects from the mediation analysis. Jointly interpreting results related to other emissions proved more subtle and highlighted the difficulty involved in interpreting principal effects as mediated effects, in particular when there are multiple mediators. A finer examination of principal strata defined by cross-classification of SO_2 changes and changes in CO_2 (or NO_x) revealed the dominating role of scrubber effects on SO_2 that was corroborated by the results of the mediation analysis. This cross-classification also reconciled the lack of evidence for a natural direct effect with the apparent evidence of dissociative effects pertaining to NO_x and CO_2 that were revealed to be driven primarily by changes in SO_2. The evidence of nonzero dissociative effects for SO_2 is likely explained by the negative expected direct effect. The relative magnitudes of principal effects and mediation effects are consistent with the well-known result that, in general, associative effects are a mixture of direct and indirect effects. Overall, these results are largely consistent with expectations: scrubbers appear to causally reduce SO_2 emissions but not those of NO_x or CO_2; scrubbers causally reduce ambient $PM_{2.5}$ (within 150 km); the effect on $PM_{2.5}$ is primarily mediated by causal reductions in SO_2 emissions and not NO_x or CO_2 emissions; and there appears to be direct effect of scrubbers on $PM_{2.5}$.

The results of this case study should be interpreted in light of several important limitations. First is the relative simplicity with which we linked power plants to monitors. Specifically, our strategy links power plants to all of the ambient monitors within 150 km. Thus, our analysis is of the causal effects of scrubbers on average $PM_{2.5}$ measured within 150 km. This likely does not reflect the full effect of emissions changes on ambient air quality which are expected to have implications at distances greater than 150 km. A related limitation is the assumption that there is no interference between observations. If the effect of a scrubber on ambient $PM_{2.5}$ extends far enough beyond 150 km so that a scrubber at a given power plant causally affects ambient $PM_{2.5}$ surrounding *other* power plants, then this assumption would be violated. More sophisticated strategies for causal inference in the presence of interference and for linking ambient monitors to power plants based on features such as atmospheric conditions and weather patterns are warranted.

Nonetheless, analysis presented here represents an important approximation that still yields valuable conclusions, especially with respect to quantifying causal pathways. Another important limitation of this analysis is that it assumes that the factors listed in Table 1 are sufficient to control for confounding which in this case would consist of differences between power plants or other features related to ambient $PM_{2.5}$ that are also associated with whether a power plant had scrubbers installed in 2005. Our approach is not readily extended to categorical mediators. We save this as potential future research. Despite these limitations we have developed new statistical methodology and leveraged an unprecedented linked data base to provide the first empirical evaluation of the presumed causal relationships that motivate a variety of regulations for improving ambient air quality and, ultimately, human health.

Acknowledgments. Contents of this work are solely the responsibility of the grantee and do not necessarily represent the official views of the USEPA. Further, USEPA does not endorse the purchase of any commercial products or services mentioned in the publication.

SUPPLEMENTARY MATERIAL

Supplement to "Bayesian methods for multiple mediators: Relating principal stratification and causal mediation in the analysis of power plant emission controls" (DOI: 10.1214/19-AOAS1260SUPP; .pdf). Appendices A–J, tables and figures are provided as supplementary materials.

REFERENCES

ALBERT, J. M. and NELSON, S. (2011). Generalized causal mediation analysis. *Biometrics* **67** 1028–1038. MR2829237

BACCINI, M., MATTEI, A. and MEALLI, F. (2015). Bayesian inference for causal mechanisms with application to a randomized study for postoperative pain control. DISIA Working Paper.

BARON, R. M. and KENNY, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* **51** 1173–1182.

BARTOLUCCI, F. and GRILLI, L. (2011). Modeling partial compliance through copulas in a principal stratification framework. *J. Amer. Statist. Assoc.* **106** 469–479. MR2866975

CHESTNUT, L. G. and MILLS, D. M. (2005). A fresh look at the benefits and costs of the US acid rain program. *J. Environ. Manag.* **77** 252–266.

CONLON, A. S. C., TAYLOR, J. M. G. and ELLIOTT, M. R. (2017). Surrogacy assessment using principal stratification and a Gaussian copula model. *Stat. Methods Med. Res.* **26** 88–107. MR3592714

DANIEL, R. M., DE STAVOLA, B. L., COUSENS, S. N. and VANSTEELANDT, S. (2015). Causal mediation analysis with multiple mediators. *Biometrics* **71** 1–14. MR3335344

DANIELS, M. J., ROY, J. A., KIM, C., HOGAN, J. W. and PERRI, M. G. (2012). Bayesian inference for the causal effect of mediation. *Biometrics* **68** 1028–1036. MR3040009

DOMINICI, F., GREENSTONE, M. and SUNSTEIN, C. R. (2014). Particulate matter matters. *Science* **344** 257–259. DOI:10.1126/science.1247348.

FORASTIERE, L., MEALLI, F. and VANDERWEELE, T. J. (2016). Identification and estimation of causal mechanisms in clustered encouragement designs: Disentangling bed nets using Bayesian principal stratification. *J. Amer. Statist. Assoc.* **111** 510–525. MR3538683

FRANGAKIS, C. E. and RUBIN, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58** 21–29. MR1891039

GILBERT, P. B. and HUDGENS, M. G. (2008). Evaluating candidate principal surrogate endpoints. *Biometrics* **64** 1146–1154. MR2522262

GUSTAFSON, P. (2010). Bayesian inference for partially identified models. *Int. J. Biostat.* **6** Art. 17, 20. MR2602560

HEI ACCOUNTABILITY WORKING GROUP (2003). *Assessing the Health Impact of Air Quality Regulations*: *Concepts and Methods for Accountability Research*. *Communication* 11. Health Effects Institute, Boston, MA.

HODAN, W. M. and BARNARD, W. R. (2004). Evaluating the contribution of PM2.5 precursor gases and re-entrained road emissions to mobile source PM2.5 particulate matter emissions. In 13*th International Emission Inventory Conference "Working for Clean Air in Clearwater."*

IMAI, K., KEELE, L. and YAMAMOTO, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statist. Sci.* **25** 51–71. MR2741814

IMAI, K. and YAMAMOTO, T. (2013). Identification and sensitivity analysis for multiple causal mechanism: Revisiting evidence from framing experiments. *Polit. Anal.* **21** 141–171.

JOFFE, M. M. and GREENE, T. (2009). Related causal frameworks for surrogate outcomes. *Biometrics* **65** 530–538. MR2751477

KIM, C., DANIELS, M. J., MARCUS, B. H. and ROY, J. A. (2017). A framework for Bayesian nonparametric inference for causal effects of mediation. *Biometrics* **73** 401–409. MR3665957

KIM, C., DANIELS, M. J., HOGAN, J. W., CHOIRAT, C. and ZIGLER, C. M. (2019). Supplement to "Bayesian methods for multiple mediators: Relating principal stratification and causal mediation in the analysis of power plant emission controls." DOI:10.1214/19-AOAS1260SUPP.

MA, Y., ROY, J. and MARCUS, B. (2011). Causal models for randomized trials with two active treatments and continuous compliance. *Stat. Med.* **30** 2349–2362. MR2829137

MACKINNON, D. P. (2008). *Introduction to Statistical Mediation Analysis*. Lawrence Earlbaum Associates, New York.

MATTEI, A. and MEALLI, F. (2011). Augmented designs to assess principal strata direct effects. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 729–752. MR2867456

MEALLI, F. and MATTEI, A. (2012). A refreshing account of principal stratification. *Int. J. Biostat.* **8** Art. 8, 19. MR2923283

MEALLI, F. and PACINI, B. (2013). Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *J. Amer. Statist. Assoc.* **108** 1120–1131. MR3174688

MEALLI, F. and RUBIN, D. B. (2003). Assumptions allowing the estimation of direct causal effects. *J. Econometrics* **112** 79–87. MR1963228

MÜLLER, P., ERKANLI, A. and WEST, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83** 67–79. MR1399156

NELSEN, R. B. (1999). *An Introduction to Copulas*. *Lecture Notes in Statistics* **139**. Springer, New York. MR1653203

OFFICE OF MANAGEMENT AND BUDGET (2013). 2013 Draft report to Congress on the benefits and costs of federal regulation and unfunded mandates on state, local, and tribal entities. Technical report, OMB, Washington, DC.

PEARL, J. (2001). Direct and indirect effects. In *Proceedings of the* 17*th Conference on Uncertainty in Artificial Intelligence* 411–420. Morgan Kaufman, San Francisco, CA.

POPE III, C. A., EZZATI, M. and DOCKERY, D. W. (2009). Fine-particulate air pollution and life expectancy in the United States. *N. Engl. J. Med.* **360** 376–386.

ROBINS, J. M. and GREENLAND, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3** 143–155.

RUBIN, D. B. (1974). Estimating causaleffects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **58** 688–701.

RUBIN, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scand. J. Stat.* **31** 161–170. MR2066246

SCHWARTZ, S. L., LI, F. and MEALLI, F. (2011). A Bayesian semiparametric approach to intermediate variables in causal inference. *J. Amer. Statist. Assoc.* **106** 1331–1344. MR2896839

TAGURI, M., FEATHERSTONE, J. and CHENG, J. (2018). Causal mediation analysis with multiple causally non-ordered mediators. *Stat. Methods Med. Res.* **27** 3–19. MR3745651

U. S. EPA (2013). Workshop on designing research to assess air quality and health outcomes from air pollution regulations. In *Designing Research to Assess Air QuAlity and Health Outcomes from Air Pollution Regulations*.

VANDERWEELE, T. J. (2008). Simple relations between principal stratification and direct and indirect effects. *Statist. Probab. Lett.* **78** 2957–2962. MR2516810

VANDERWEELE, T. J. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology* **20** 18–26.

VANDERWEELE, T. J. (2011). Principal stratification—uses and limitations. *Int. J. Biostat.* **7** Art. 28, 14. MR2843532

VANDERWEELE, T. J. and VANSTEELANDT, S. (2014). Mediation analysis with multiple mediators. *Epidemiol Methods* **2** 95–115.

WANG, W., NELSON, S. and ALBERT, J. M. (2013). Estimation of causal mediation effects for a dichotomous outcome in multiple-mediator models using the mediation formula. *Stat. Med.* **32** 4211–4228. MR3118350

ZIGLER, C. M., DOMINICI, F. and WANG, Y. (2012). Estimating causal effects of air quality regulations using principal stratification for spatially correlated multivariate intermediate outcomes. *Biostatistics* (*Oxford*, *England*) **13** 289–302. DOI:10.1093/biostatistics/kxr052.

C. KIM
DEPARTMENT OF BIOSTATISTICS
BOSTON UNIVERSITY SCHOOL OF PUBLIC HEALTH
801 MASSACHUSETTS AVENUE
BOSTON, MASSACHUSETTS 0211
USA
E-MAIL: chanmink@bu.edu

M. J. DANIELS
DEPARTMENT OF STATISTICS
UNIVERSITY OF FLORIDA
GAINESVILLE, FLORIDA 32611
USA
E-MAIL: mdaniels@stat.ufl.edu

J. W. HOGAN
DEPARTMENT OF BIOSTATISTICS
BROWN UNIVERSITY SCHOOL OF PUBLIC HEALTH
PROVIDENCE, RHODE ISLAND 02912
USA
E-MAIL: jhogan@stat.brown.edu

C. CHOIRAT
SWISS DATA SCIENCE CENTER
UNIVERSITÄTSTRASSE 25
8006 ZÜRICH
SWITZERLAND
E-MAIL: cchoirat@datascience.ch

C. M. ZIGLER
DEPARTMENT OF STATISTICS AND DATA SCIENCES
DEPARTMENT OF WOMEN'S HEALTH
DELL SHOOL OF MEDICINE
UNIVERSITY OF TEXAS AT AUSTIN
2317 SPEEDWAY D9800
AUSTIN, TEXAS 78712
USA
E-MAIL: cory.zigler@austin.utexas.edu