Journal of Econometrics 162 (2011) 345–361

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Bayesian inference in a correlated random coefficients model: Modeling causal effect heterogeneity with an application to heterogeneous returns to schooling

Mingliang Li [a,*], Justin L. Tobias [b]

[a] Department of Economics, SUNY-Buffalo, United States

[b] Department of Economics, Purdue University, United States

ARTICLE INFO

Article history:
Received 4 July 2009
Received in revised form
9 September 2010
Accepted 8 February 2011
Available online 25 February 2011

ABSTRACT

We consider the problem of causal effect heterogeneity from a Bayesian point of view. This is accomplished by introducing a three-equation system, similar in spirit to the work of Heckman and Vytlacil (1998), describing the joint determination of a scalar outcome, an endogenous "treatment" variable, and an individual-specific causal return to that treatment. We describe a Bayesian posterior simulator for fitting this model which recovers far more than the average causal effect in the population, the object which has been the focus of most previous work. Parameter identification and generalized methods for flexibly modeling the outcome and return heterogeneity distributions are also discussed.

Combining data sets from High School and Beyond (HSB) and the 1980 Census, we illustrate our methods in practice and investigate heterogeneity in returns to education. Our analysis decomposes the impact of key HSB covariates on log wages into three parts: a "direct" effect and two separate indirect effects through educational attainment and returns to education. Our results strongly suggest that the quantity of schooling attained is determined, at least in part, by the individual's own return to education. Specifically, a one percentage point increase in the return to schooling parameter is associated with the receipt of (approximately) 0.14 more years of schooling. Furthermore, when we control for variation in returns to education across individuals, we find no difference in predicted schooling levels for men and women. However, women are predicted to attain approximately 1/4 of a year more schooling than men on average as a result of higher rates of return to investments in education.

1. Introduction

A substantial volume of work in economics and statistics has been devoted to the issue of identifying and estimating the causal impact of an endogenous or "treatment" variable. The endogeneity issue in these models arises when the level or intensity of treatment is not randomly assigned, but, instead, is selected by the individual. This self-selection problem is the defining feature of observational data, as those agents observed to choose high levels of treatment are likely to possess observed and unobserved characteristics that differ from those choosing lower levels of treatment. This aspect of the problem generates a well-known bias and inconsistency in standard estimators that fail to account for the endogeneity of treatment levels.[1]

Recent work has also emphasized the importance of addressing *causal effect heterogeneity* – that individual agents may have different returns to treatment – and providing the proper interpretation of traditional estimators in the presence of such heterogeneity. For example, the important work of Imbens and Angrist (1994) shows that, under suitable conditions in a binary treatment context, the standard instrumental variable (IV) technique recovers the local average treatment effect (LATE), a treatment impact for a subpopulation of "compliers" whose behavior can be manipulated by an instrument. Whether or not this LATE parameter is inherently an object of interest is necessarily

* Corresponding author. Tel.: +1 716 645 2121; fax: +1 716 645 2127.
E-mail addresses: mli3@buffalo.edu (M. Li), jltobias@purdue.edu (J.L. Tobias).

[1] In our paper, we continue to use the word "treatment" and use it in reference to the endogenous right-hand side outcome variable. In most cases in the literature, however, "treatment" refers to variables that are binary, or perhaps discrete, while here we will consider a continuous treatment outcome. Furthermore, "treatment effects" are typically defined in the context of a binary treatment outcome in a potential outcomes framework that explicitly models both the treated and untreated states. We work in this paper with a simplified model of observed outcomes only, which, arguably, is the most common model used in applied work. We will refer to the individual-specific return to treatment as a causal effect, causal impact, or "return", and derive procedures for characterizing many properties of the distribution of causal effects in the population. Conventional treatment effects, such as ATE, TT, and LATE, are based upon a more general representation of the model with a binary treatment, and thus are not directly comparable to those considered here. Our framework, instead, is based upon the observed outcomes representation as discussed in Wooldridge (1997), Heckman and Vytlacil (1998), and Wooldridge (2003), among others.

0304-4076/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2011.02.006

an application and instrument-specific question, and, as such, researchers sometimes focus on methods that enable recovery of the average treatment effect (ATE), or average causal effect, in the presence of treatment effect heterogeneity. Notable studies in this regard, based upon observed outcomes models like the one we consider here, include those of Wooldridge (1997), Heckman and Vytlacil (1998), and Wooldridge (2003), who introduce assumptions and procedures under which the average causal effect can be consistently estimated when treatment returns are heterogeneous and potentially correlated with treatment levels. In a similar vein, Heckman and Vytlacil (1999, 2005) and others describe the marginal treatment effect (MTE) (e.g., Björklund and Moffitt, 1987, Heckman, 1997, and Heckman and Smith, 1999) as a type of unifying parameter which, when properly integrated, can be used to calculate all of the conventional mean binary treatment effect parameters, including ATE, LATE, and the effect of treatment on the treated (TT), thereby capturing important aspects of treatment effect heterogeneity.

In this paper, we take up the issue of causal effect heterogeneity in a similar spirit to the work mentioned above, but choose to address this issue via a Bayesian approach. Specifically, we consider Bayesian estimation in a variant of the correlated random coefficient (CRC) model, similar to that considered by Wooldridge (1997), Heckman and Vytlacil (1998) and Wooldridge (2003). To be sure, our study is certainly not the first such effort, and, indeed, a sizeable Bayesian literature has evolved for the estimation of treatment–response models with observational data.[2] Early efforts in this regard primarily focused on the Markov chain Monte Carlo (MCMC) implementation (e.g., Koop and Poirier, 1997 and Chib and Hamilton, 2000) and included some discussion of recovering individual-level treatment impacts within a potential outcomes framework.[3] More recent work has focused on problems associated with weak instruments generally, has discussed priors that yield posteriors similar to sampling distributions for the two-stage least squares (2SLS) and limited information maximum likelihood (LIML) estimators (e.g., Kleibergen and Zivot, 2003), has introduced a non-parametric modeling of outcomes via a Dirichlet process prior (Conley et al., 2008), and has obtained new results associated with the seminal Angrist and Krueger (1991) study (e.g., Hoogerheide et al., 2007).[4]

Our model of interest consists of a three-equation system describing the joint determination of an observed scalar outcome variable, an endogenous "treatment" variable, and an individual-specific causal effect parameter. The novelty of our approach is that we directly model the process generating the individual-specific causal effect and thus can calculate any statistic of interest (such as return percentiles or the probability of a positive treatment impact) associated with the causal effect heterogeneity distribution. Of course, our ability to do this stems from particular parametric assumptions made regarding the heterogeneity distribution, and, to this end, we describe methods that enable a flexible representation of this distribution. In addition, and similar in spirit to the income maximization presumption of the Roy (1951) model, agents can potentially choose the amount of treatment based on their own knowledge of the return to such treatment. Provided sufficient sources of exogenous variation are available, we show that this presumption becomes empirically testable.

Within the Bayesian literature, the structure of our three-equation system seems rather similar in spirit to the innovative work of Manchanda et al. (2004).[5] In this paper, the authors are interested in providing a joint description of "detailing" efforts made by drug companies and the number of physician prescriptions, noting that the decision to detail particular physicians may depend, at least in part, on the responsiveness of that physician to the detailing effort (i.e., how many more prescriptions he or she will write as a consequence of being detailed). Our model seeks to address a similar problem to that considered by Manchanda et al. (2004), though in our case the standard treatment–response framework, which must contend with problems such as confounding on unobservables, is generalized to allow treatment levels to be selected based on the "returns" to treatment. Similarly, Conley et al. (2008, section 2.5) discuss the possibility of employing a Dirichlet process prior to simultaneously allow for a nonparametric distribution of outcomes and heterogeneous treatment impacts. Despite its flexibility, this specification does not explicitly model the potential structural dependence of the endogenous treatment variable on the return to treatment, and thus differs from the specification considered here.

In some sense, one might regard our efforts in this endeavor as a step back relative to the existing classical literature, in light of the fact that we need to make specific distributional assumptions whereas others (e.g., Wooldridge, 2003) only require a few moment conditions to be satisfied. While our assumptions are clearly stronger than those typically made in these types of analysis, we argue that the benefits afforded by such assumptions may warrant their adoption: we are able to identify all parameters of our model (provided satisfactory exclusion restrictions exist) and expand our focus to directly model the entire treatment effect distribution.

We develop an efficient posterior simulator for fitting our model and illustrate in generated data experiments that it mixes well and performs adequately in recovering parameters of the data-generating process. In addition, we carefully discuss the conditions required for parameter identification and methods for relaxing normality, thereby allowing for a more flexible modeling of the outcome and return heterogeneity distributions. Finally, we employ our methods in a real application and investigate the issue of heterogeneity in the economic returns to education, following the influential work of Card (2001). To this end, we combine data sources from the sophomore cohort of the High School and Beyond (1992) Survey and the 1980 Census. We show in our paper that successful identification of our model's parameters requires the availability of some variable that has a structural impact on individual-level returns to treatment, but remains conditionally uncorrelated with our outcome variable and the level of treatment received. In this regard we first use 1980 Census data to calculate county-level average returns to education. The lagged county-level returns to schooling are then used as exogenous sources of variation which should correlate positively with the individual's (1991) private returns to education (and we find strong evidence in support of this), but are assumed to be conditionally uncorrelated with educational attainment and log wage outcomes.

[2] Important examples of this work include Koop and Poirier (1997), Li (1998), Chib and Hamilton (2000, 2002), Poirier and Tobias (2003), and Chib (2007). Li et al. (2003), Munkin and Trivedi (2003), and Deb et al. (2006) provide applications of these methods.

[3] That is, these models explicitly consider outcomes in the treated and untreated states together with the treatment decision. Chib and Hamilton (2002), for example, point out the possibility of learning about individual-level treatment effects, while Koop and Poirier (1997) and Poirier and Tobias (2003) discuss the potential for learning about outcome gain distributions and the cross-regime correlation parameter. In recent work, Chib (2007) argues in favor of avoiding explicit modeling of the counterfactual, owing to concerns associated with modeling the non-identifiable cross-regime correlation parameter. We follow in a similar spirit of working with observed outcomes in the present paper.

[4] The Bayesian approach to this model has also received considerable attention in recent textbooks, including Lancaster (2004, Chapter 8), Rossi et al. (2005, Chapter 7), and Koop et al. (2007, pp. 223–236).

[5] This paper is also described in Rossi et al. (2005).

Our results suggest strong evidence that the amount of schooling attained is determined, in part, by the individual's own return to education. Specifically, a one percentage point increase in the return to schooling parameter is associated with the receipt of (approximately) 0.14 more years of education. Further, we find evidence of heterogeneity in returns to education, with females, blacks and Hispanics possessing higher returns to schooling than males and whites.

The outline of our paper is as follows. Section 2 briefly introduces the model while Section 3 discusses identification, strategies for posterior simulation, and how learning takes place regarding the causal effect heterogeneity parameters. Section 4 conducts generated data experiments while Section 5 describes the data sets involved with our application. Results of that application are presented in Section 6, and the paper concludes with a summary in Section 7. The Appendix provides technical details regarding identification.

2. The model

The model we consider is a three-equation system as described below[6]:

$$y_i = \beta_0 + \boldsymbol{x_i}\boldsymbol{\beta} + s_i\theta_i + u_i \tag{1}$$

$$s_i = \delta_0 + \boldsymbol{x_i}\boldsymbol{\delta} + \boldsymbol{z_i}\boldsymbol{\gamma} + \theta_i\rho + v_i \tag{2}$$

$$\theta_i = \eta_0 + \boldsymbol{x_i}\boldsymbol{\eta} + \boldsymbol{w_i}\boldsymbol{\lambda} + \epsilon_i. \tag{3}$$

In Eq. (1), y_i denotes the (continuous) outcome of interest and s_i is a continuous and (potentially) endogenous treatment variable. The covariates in $\boldsymbol{x}_i$ are common to all equations, while we also allow instrumental variables $\boldsymbol{z}_i$ and $\boldsymbol{w}_i$ to appear in (2) and (3), respectively.[7]

Consistent with a majority of recent work in this area, we do not wish to impose identical treatment returns for each agent, and explicitly allow for heterogeneous causal impacts. The private return to s_i in our model is denoted as θ_i, and the notation makes clear that the return can differ across individuals. Eq. (3) then relates the (unobserved) return θ_i to observables $\boldsymbol{w_i}$ and $\boldsymbol{x_i}$. That is, individual-specific returns to education may potentially depend on things like the ability (i.e., test score) of the agent as well as other demographic characteristics. Finally, in (2), we allow the quantity of the endogenous variable s_i to depend, at least in part, on the return to treatment θ_i. That is, economic agents who may have knowledge of their private return to education, for example, may potentially choose the amount of schooling based on this knowledge. A related assumption appears frequently in close variants of this model; the Roy (1951) model, for example, is based on income maximization and posits that individuals select into binary "treatment" based on their economic gain from doing so.

3. Identification and model generality

It may not be immediately clear that the parameters of the system above are identifiable, or what conditions might be required in order to achieve identification. To begin our discussion of these issues, we first consider the likelihood $p(\boldsymbol{y}, \boldsymbol{s}|\boldsymbol{\Gamma}_{-\boldsymbol{\theta}})$ under the assumption of jointly normal errors with unrestricted covariance matrix $\boldsymbol{\Sigma}$[8]:

$$\begin{bmatrix} u_i \\ v_i \\ \epsilon_i \end{bmatrix} \Big| \boldsymbol{x_i}, \boldsymbol{w_i}, \boldsymbol{z_i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \sigma_{ys} & \sigma_{y\theta} \\ \sigma_{ys} & \sigma_s^2 & \sigma_{s\theta} \\ \sigma_{y\theta} & \sigma_{s\theta} & \sigma_\theta^2 \end{pmatrix} \right]$$
$$\equiv \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}). \tag{4}$$

Details regarding parameter identification under this assumption are described in the Appendix, yet it is worth noting here the main results of this exercise. In a sense, the most important exclusion restriction for identification purposes turns out to be the set of variables $\boldsymbol{w}$, as their presence enables full identification of the model parameters. Specifically, if no "traditional" instrument $\boldsymbol{z}$ is available (i.e., $\boldsymbol{\gamma} = 0$), then the system in (1)–(3) is still identified, provided that $\rho \neq 0$ and $\boldsymbol{w}$ appears only in (3). If $\boldsymbol{z}$ is additionally available in (2), as previously written in Eqs. (1)–(3), then the variables $\boldsymbol{w}$ could actually be included in both (1) and (3), and the model would remain fully identified. On the other hand, if $\boldsymbol{\lambda} = 0$ so that $\boldsymbol{z}$, the traditional instrument, is the only excluded variable, then the model is no longer fully identified.

It may not be clear why $\boldsymbol{w}$ is so important for identification purposes, and in what follows we seek to provide an intuitive explanation for this. The presence of $\boldsymbol{w}$ in (3) enables exogenous shifts in the distribution of θ_i which, in turn, permits the identification of ρ in (2) and subsequently all of the remaining structural parameters. In the more difficult case where $\boldsymbol{\gamma} = 0$ so that $\boldsymbol{z}$ is absent from the model, we obtain, upon integrating out the heterogeneity terms,

$$y_i = \beta_0 + \boldsymbol{x_i}\boldsymbol{\beta} + s_i\eta_0 + s_i\boldsymbol{x_i}\boldsymbol{\eta} + s_i\boldsymbol{w_i}\boldsymbol{\lambda} + (s_i\epsilon_i + u_i)$$
$$s_i = (\delta_0 + \rho\eta_0) + \boldsymbol{x_i}(\boldsymbol{\delta} + \rho\boldsymbol{\eta}) + \boldsymbol{w_i}\rho\boldsymbol{\lambda} + (\rho\epsilon_i + v_i).$$

Loosely (a more formal treatment is provided in the Appendix), the coefficient on $s_i\boldsymbol{w_i}$ enables us to recover an estimate of $\boldsymbol{\lambda}$ and the reduced form expression for s_i enables us to estimate the product $\rho\boldsymbol{\lambda}$, thus providing a means to recover ρ.[9] When $\boldsymbol{z}$ is available, but $\boldsymbol{w}$ is not present, the interaction $s_i\boldsymbol{w_i}$ disappears and ρ is no longer separately identifiable, although η_0, interpretable as the average causal impact in the population (when $\boldsymbol{\lambda} = \boldsymbol{0}$ and $\boldsymbol{x}$ is standardized to be mean zero), does remain estimable.

Of course, whether or not exclusion restrictions such as $\boldsymbol{w}$ and $\boldsymbol{z}$ are available in practice is inevitably an application-specific question, and the degree to which such restrictions can be credibly maintained will depend, in part, on adequate conditioning data and, to no small degree, on the persuasiveness of the researcher. What emerges from our model, however, is a primary requirement that is somewhat different from the traditional instrumental variables assumption: what is most necessary for identification purposes is the existence of some variable affecting the return to treatment that is also (conditionally) uncorrelated with the outcomes of interest, and the endogenous treatment variable in particular. In what follows, we illustrate application of this model to a widely studied question in the labor economics literature: estimating the return to education and characterizing heterogeneity in schooling returns.

It is also worthwhile to pause and place our study in the context of the previous literature. From the classical perspective, numerous papers have addressed various aspects of causal effect heterogeneity, and previous efforts that seem most similar to ours

[6] Bold script is used to denote vectors and matrices, while capitals are used for matrices.

[7] We consider only a scalar treatment variable in this analysis, noting that others, such as Wooldridge (2003), allow for multiple endogenous variables. Extension to the multivariate case is possible and reasonably straightforward, but is not considered here.

[8] Here, $\boldsymbol{\Gamma}_{-\boldsymbol{\theta}}$ denotes all parameters other than $\boldsymbol{\theta} = [\theta_1 \quad \theta_2 \quad \cdots \quad \theta_n]$ as the heterogeneity terms are to be integrated out.

[9] The first equation of this system is very similar to that described by Wooldridge (2003), who notes that IV/2SLS can be used to estimate the average causal effect η_0 using $\boldsymbol{w}$ and its interactions as instruments. What is required here is that the conditional covariance between s_i and θ_i does not depend on w. This is true given the assumptions of our model. Note, however, that if treatment effect homogeneity were assumed, and standard IV were applied to (1) directly using z (or w) to instrument for s, then one will not recover in general an estimate of the average causal impact. Further details regarding this issue are available upon request.

include those of Heckman and Vytlacil (1998) and Wooldridge (1997, 2003). In the most recent of these, Wooldridge (2003) introduces three assumptions that enable consistent estimation of the average treatment effect. His representation of the model is more general than ours, as multiple treatments are considered, no specific distributional assumptions are employed, and no explicit modeling of the treatment variable s is necessary. In this sense, our model might seem to offer a step in the wrong direction, as Eqs. (1)–(3) impose considerable structure beyond what has been used in past work. In our view, however, the added structure may be a worthy investment, as it enables us to expand our focus beyond the average causal effect and learn about all aspects of the heterogeneity distribution, including learning about individual-level treatment impacts. With respect to distributional concerns, these seem decidedly more minor, as we will now replace trivariate normality with a finite mixture of Gaussian distributions,[10] which provides a very flexible way to model the joint distribution of the outcome, endogenous variable, and causal effect parameter.

3.1. Posterior analysis

The normality assumption made in (4) is potentially inappropriate and often controversial, and, to this end, it is important to recognize that it can be significantly generalized. We pursue such a generalization in this and the following sections via a finite Gaussian mixture representation. We begin by writing this model in a somewhat non-traditional way as

$$y_i = \beta_{0i} + \boldsymbol{x_i}\boldsymbol{\beta_i} + s_i\theta_i + u_i \tag{5}$$

$$s_i = \delta_{0i} + \boldsymbol{x_i}\boldsymbol{\delta_i} + \boldsymbol{z_i}\boldsymbol{\gamma_i} + \rho_i\theta_i + v_i \tag{6}$$

$$\theta_i = \eta_{0i} + \boldsymbol{x_i}\boldsymbol{\eta_i} + \boldsymbol{w_i}\boldsymbol{\lambda_i} + \epsilon_i \tag{7}$$

and

$$[u_i \quad v_i \quad \epsilon_i]'|\boldsymbol{x_i}, \boldsymbol{z_i}, \boldsymbol{w_i} \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma_i}),$$

whence the density for $y_i, s_i, \theta_i|\cdot$ is readily available, as the associated Jacobian of the transformation from the error vector to $[y_i \quad s_i \quad \theta_i]'$is unity.

The notation in (5)–(7) is quite general, as it allows for individual-specific slope and intercept parameters. The finite mixture formulation of the model adds structure to this by proposing that there are, say, G distinct groups with identical parameters within each group yet different parameters across groups. Whether or not these "groups" have any intrinsic meaning as a discrete partitioning of the population of interest is mostly irrelevant, though, if such an interpretation can be convincingly given in a particular application, the mixture components may be afforded a specific interpretation. In most instances, mixture models are commonly employed as a flexible computational tool to allow for skew, multimodality, and heavy tails in the outcome distributions, and no specific meaning need be ascribed to the various mixture components.[11] With an eye toward the implementation of our posterior simulator, we first define the parameter sets

$$\begin{aligned}\boldsymbol{\phi_i} &= [\beta_{0i} \quad \boldsymbol{\beta_i}' \quad \delta_{0i} \quad \boldsymbol{\delta_i}' \quad \boldsymbol{\gamma_i} \quad \eta_{0i} \quad \boldsymbol{\eta_i}' \quad \boldsymbol{\lambda_i}]' \in \overline{\boldsymbol{\phi}} \\ &= \{\overline{\boldsymbol{\phi}}_{\boldsymbol{1}}, \overline{\boldsymbol{\phi}}_{\boldsymbol{2}}, \ldots, \overline{\boldsymbol{\phi}}_{\boldsymbol{G}}\} \\ \rho_i \in \overline{\boldsymbol{\rho}} &= \{\overline{\rho}_1, \overline{\rho}_2, \ldots, \overline{\rho}_G\}\end{aligned}$$

and

$$\boldsymbol{\Sigma_i} \in \overline{\boldsymbol{\Sigma}} = \{\overline{\boldsymbol{\Sigma}}_{\boldsymbol{1}}, \overline{\boldsymbol{\Sigma}}_{\boldsymbol{2}}, \ldots, \overline{\boldsymbol{\Sigma}}_{\boldsymbol{G}}\}.$$

All parameters other than ρ_i, θ_i, and $\boldsymbol{\Sigma_i}$ are lumped into the vector $\boldsymbol{\phi_i}$. The above equations imply that each of the individual-specific parameters will be assigned a hierarchical prior where, once the component of the mixture is known, the distribution for ρ_i, $\boldsymbol{\phi}_i$ and $\boldsymbol{\Sigma_i}$ is degenerate around one of the G values in the parameter sets given above. The allocation of agents to the appropriate component of the mixture is achieved via the addition of component indicator variables. Specifically, we will let $c_{ig} = 1$ denote that individual i "belongs to" the gth component of the mixture. Formally, we let

$$\boldsymbol{c_i} = [c_{i1} \quad c_{i2} \quad \cdots \quad c_{iG}]'$$

be the component label vector for individual i, and specify priors of the form

$$\begin{aligned}&p(\boldsymbol{\phi_i}, \rho_i, \boldsymbol{\Sigma_i}|c_{ig} = 1, \overline{\boldsymbol{\phi}}, \overline{\boldsymbol{\rho}}, \overline{\boldsymbol{\Sigma}}) \\ &\quad = I\left(\boldsymbol{\phi_i} = \overline{\boldsymbol{\phi}}_{\boldsymbol{g}}, \rho_i = \overline{\rho}_g, \boldsymbol{\Sigma_i} = \overline{\boldsymbol{\Sigma}}_{\boldsymbol{g}}\right)\end{aligned} \tag{8}$$

$$\boldsymbol{c_i} \overset{\text{i.i.d.}}{\sim} \text{Mult}(1, \overline{\boldsymbol{\pi}}) \tag{9}$$

$$\overline{\boldsymbol{\pi}} \sim \text{Dirichlet}(\boldsymbol{\alpha}), \tag{10}$$

with

$$\overline{\boldsymbol{\pi}} = [\pi_1 \quad \pi_2 \quad \cdots \quad \pi_G]', \qquad \boldsymbol{\alpha} = [\alpha_1 \quad \alpha_2 \quad \cdots \quad \alpha_G]',$$

$I(\cdot)$ denoting the standard indicator function, Mult$(\cdot)$ denoting the multinomial distribution and *Dirichlet*$(\cdot)$ denoting the Dirichlet distribution (see, e.g., Koop et al., 2007, pp. 340). Thus, given $c_{ig} = 1$ and the set of component-specific parameters and covariance matrices, the parameter vector for individual i is known, and $\boldsymbol{\phi_i}$, ρ_i and $\boldsymbol{\Sigma_i}$ merely serve as "place-holders" which are convenient for simplifying the exposition. The multinomial prior on $\boldsymbol{c_i}$ and the corresponding Dirichlet prior on the component probability vector $\overline{\boldsymbol{\pi}}$ imply that, unconditionally,

$$p(\boldsymbol{\phi_i}, \rho_i, \boldsymbol{\Sigma_i}|\overline{\boldsymbol{\pi}}, \overline{\boldsymbol{\phi}}, \overline{\boldsymbol{\rho}}, \overline{\boldsymbol{\Sigma}}) = \sum_{g=1}^{G} \pi_g I\left(\boldsymbol{\phi_i} = \overline{\boldsymbol{\phi}}_{\boldsymbol{g}} \rho_i = \overline{\rho}_g, \boldsymbol{\Sigma_i} = \overline{\boldsymbol{\Sigma}}_{\boldsymbol{g}}\right).$$

Likewise, the trivariate distribution for y_i, s_i, θ_i (not conditioned on $\boldsymbol{c}_i$) is

$$\begin{aligned}&p(y_i, s_i, \theta_i|\overline{\boldsymbol{\pi}}, \overline{\boldsymbol{\phi}}, \overline{\boldsymbol{\Sigma}}, \overline{\boldsymbol{\rho}}) \\ &\quad = \sum_{g=1}^{G} \pi_g p(y_i, s_i, \theta_i|\boldsymbol{\phi_i} = \overline{\boldsymbol{\phi}}_{\boldsymbol{g}}, \rho_i = \overline{\rho}_g, \boldsymbol{\Sigma_i} = \overline{\boldsymbol{\Sigma}}_{\boldsymbol{g}}),\end{aligned}$$

where the distribution within each component of the mixture is obtained by a change of variables from (5)–(7). This illustrates the finite mixture representation of the likelihood.

We complete the specification of our model with the following priors:

$$\overline{\boldsymbol{\phi}}_{\boldsymbol{g}} \overset{\text{i.i.d.}}{\sim} N(\boldsymbol{\phi_0}, \boldsymbol{V_\phi}), \quad g = 1, 2, \ldots, G \tag{11}$$

$$\begin{aligned}&p(\overline{\boldsymbol{\Sigma}}_{\boldsymbol{1}}, \overline{\boldsymbol{\Sigma}}_{\boldsymbol{2}}, \ldots, \overline{\boldsymbol{\Sigma}}_{\boldsymbol{G}}) \\ &\quad \propto I(\overline{\boldsymbol{\Sigma}}_{ss1} < \overline{\boldsymbol{\Sigma}}_{ss2} < \cdots < \overline{\boldsymbol{\Sigma}}_{ssG}) \prod_{g=1}^{G} p_{IW}(\overline{\boldsymbol{\Sigma}}_{\boldsymbol{g}}|p, pR)\end{aligned} \tag{12}$$

$$\overline{\rho}_g \overset{\text{i.i.d.}}{\sim} N(\rho_0, V_\rho) \quad g = 1, 2, \ldots, G. \tag{13}$$

The prior on the set of covariance matrices serves to identify the mixture components, and it does so by providing an ordering restriction on variances in the schooling equation (i.e., Eq. (2)). To derive the joint posterior distribution for the unobservables in our

[10] A proof regarding parameter identification in the finite mixture framework is omitted here, but is available upon request. The proof follows a similar strategy to that given in the Appendix, where $y|s, \boldsymbol{\Gamma}_{-\boldsymbol{\theta}}$ and $s|\boldsymbol{\Gamma}_{-\boldsymbol{\theta}}$ are obtained and their moments are characterized, revealing parameter identification. Such a strategy dates back to at least Pearson (1894), who used a moment-based approach to estimate the parameters of a two-component Gaussian mixture.

[11] See, e.g., Geweke and Keane (2007) for use of a related methodology, the smoothly mixing regression model.

model, we first define

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{s} = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{pmatrix}, \quad \boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{pmatrix}, \quad \boldsymbol{\phi} = \begin{pmatrix} \boldsymbol{\phi_1} \\ \boldsymbol{\phi_2} \\ \vdots \\ \boldsymbol{\phi_n} \end{pmatrix},$$

$$\boldsymbol{\rho} = \begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_n \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma_1} \\ \boldsymbol{\Sigma_2} \\ \vdots \\ \boldsymbol{\Sigma_n} \end{pmatrix}, \quad \text{and} \quad \boldsymbol{c} = \begin{pmatrix} \boldsymbol{c_1'} \\ \boldsymbol{c_2'} \\ \vdots \\ \boldsymbol{c_n'} \end{pmatrix}.$$

With this notation in hand, Bayes' theorem gives the joint posterior distribution up to proportionality:

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{\Sigma}, \overline{\boldsymbol{\phi}}, \overline{\boldsymbol{\rho}}, \overline{\boldsymbol{\Sigma}}, \overline{\boldsymbol{\pi}}, \boldsymbol{c} | y, s) &\propto p(\overline{\boldsymbol{\pi}}) p(\overline{\boldsymbol{\phi}}) p(\overline{\boldsymbol{\rho}}) p(\overline{\boldsymbol{\Sigma}}) \\ &\times \prod_{i=1}^{n} p(\boldsymbol{\phi_i}, \rho_i, \boldsymbol{\Sigma_i} | \boldsymbol{c_i}, \overline{\boldsymbol{\phi}}, \overline{\boldsymbol{\rho}}, \overline{\boldsymbol{\Sigma}}) p(\boldsymbol{c_i} | \overline{\boldsymbol{\pi}}) |\boldsymbol{\Sigma_i}|^{-\frac{1}{2}} \\ &\times \exp\left(-\frac{1}{2} \boldsymbol{h_i'} \boldsymbol{\Sigma_i^{-1}} \boldsymbol{h_i}\right), \end{aligned} \tag{14}$$

where

$$\boldsymbol{h_i} = \begin{pmatrix} y_i - \beta_{0i} - \boldsymbol{x_i \beta_i} - s_i \theta_i \\ s_i - \delta_{0i} - \boldsymbol{x_i \delta_i} - \boldsymbol{z_i \gamma_i} - \rho_i \theta_i \\ \theta_i - \eta_{0i} - \boldsymbol{x_i \eta_i} - \boldsymbol{w_i \lambda_i} \end{pmatrix}$$

and $p(\boldsymbol{\phi_i}, \rho_i, \boldsymbol{\Sigma_i} | \boldsymbol{c_i}, \overline{\boldsymbol{\phi}}, \overline{\boldsymbol{\rho}}, \overline{\boldsymbol{\Sigma}})$ has been given in (8).

3.2. The Gibbs algorithm

In principle, a standard Gibbs sampler can be applied to fit this model, drawing, in turn, from each of the complete posterior conditionals as implied by (14). However, this standard Gibbs sampler turns out to suffer from poor mixing properties. To this end, we employ a blocking (grouping) step where the parameters $\overline{\boldsymbol{\phi}}$ are drawn from their conditional distribution, marginalized over the heterogeneity terms $\boldsymbol{\theta}$, and then the remaining parameters are drawn from their complete conditional posterior distributions. This blocking procedure samples $\overline{\boldsymbol{\phi}}$ and $\boldsymbol{\theta}$ together in a single step, which substantially improves the mixing of the posterior simulations, as will be shown in the following section.

To implement this blocking procedure, it is useful to write the augmented likelihood in a different, though equivalent, way. Completing the square on θ_i in (14) enables us to express the augmented joint posterior as

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{\Sigma}, \overline{\boldsymbol{\phi}}, \overline{\boldsymbol{\rho}}, \overline{\boldsymbol{\Sigma}}, \overline{\boldsymbol{\pi}}, \boldsymbol{c} | y, s) &\propto p(\overline{\boldsymbol{\pi}}) p(\overline{\boldsymbol{\phi}}) p(\overline{\boldsymbol{\rho}}) p(\overline{\boldsymbol{\Sigma}}) \\ &\times \prod_{i=1}^{n} [p(\boldsymbol{\phi_i}, \rho_i, \boldsymbol{\Sigma_i} | \boldsymbol{c_i}, \overline{\boldsymbol{\phi}}, \overline{\boldsymbol{\rho}}, \overline{\boldsymbol{\Sigma}}) p(\boldsymbol{c_i} | \overline{\boldsymbol{\pi}}) \phi_N(\theta_i; qq_i^{-1} tq_i, qq_i^{-1}) | \\ &\times \boldsymbol{\Sigma}|^{-1/2} |qq_i|^{-1/2} \exp(-[1/2] tt_i)], \end{aligned} \tag{15}$$

where we have defined

$$\boldsymbol{t_i} \equiv \begin{pmatrix} y_i - \beta_{0i} - \boldsymbol{x_i \beta_i} \\ s_i - \delta_{0i} - \boldsymbol{x_i \delta_i} - \boldsymbol{z_i \gamma_i} \\ -\eta_{0i} - \boldsymbol{x_i \eta_i} - \boldsymbol{w_i \lambda_i} \end{pmatrix}, \quad \boldsymbol{q_i} \equiv \begin{pmatrix} s_i \\ \rho_i \\ -1 \end{pmatrix}, \tag{16}$$

$$\begin{aligned} &qq_i \equiv \boldsymbol{q_i'} \boldsymbol{\Sigma_i^{-1}} \boldsymbol{q_i}, \\ &tq_i \equiv \boldsymbol{t_i'} \boldsymbol{\Sigma_i^{-1}} \boldsymbol{q_i}, \quad \boldsymbol{\sigma qq_i} \equiv \boldsymbol{\Sigma_i^{-1}} - \boldsymbol{\Sigma_i^{-1}} \boldsymbol{q_i} [qq_i]^{-1} \boldsymbol{q_i'} \boldsymbol{\Sigma_i^{-1}} \quad \text{and} \\ &tt_i \equiv \boldsymbol{t_i'} [\boldsymbol{\sigma qq_i}] \boldsymbol{t_i}. \end{aligned} \tag{17}$$

To account for the ordering constraint $\overline{\Sigma}_{ss1} < \overline{\Sigma}_{ss2} < \cdots < \overline{\Sigma}_{ssG}$, we generate simulations by first ignoring the constraint and making use of the "unconstrained" sampler described below, and then permuting the labels at the end of the simulation period as necessary to achieve agreement with the ordering restriction. More discussion of this issue can be found in Frühwirth-Schnatter (2001, particularly Sections 3.3 and 3.4) and Geweke (2007, particularly Section 3). For the empirical application of Section 6, we also emphasize that the parameters of interest reported and the posterior predictive analyses conducted are not affected by the labeling issue, and as such the need to permute the labels or impose component identification through the prior is irrelevant for these pursuits. The Gibbs sampler, apart from the label permutation, then proceeds in six steps, which we enumerate below.

Step 1: $\boldsymbol{\phi}, \overline{\boldsymbol{\phi}} | \cdot, \boldsymbol{y}, \boldsymbol{s}$.

To sample the regression parameters $\boldsymbol{\phi}$ and $\overline{\boldsymbol{\phi}}$ marginalized over the random effects $\boldsymbol{\theta}$, first let

$$\begin{aligned} &\boldsymbol{r_i} \equiv (y_i \quad s_i \quad 0)', \\ &\boldsymbol{X_i} \equiv \begin{pmatrix} 1 & \boldsymbol{x_i} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \boldsymbol{x_i} & \boldsymbol{z_i} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & \boldsymbol{x_i} & \boldsymbol{w_i} \end{pmatrix}, \\ &\boldsymbol{XX_g} \equiv \sum_{\{i: c_{ig}=1\}} \boldsymbol{X_i'} [\boldsymbol{\sigma qq_i}] \boldsymbol{X_i}, \quad \text{and} \quad \boldsymbol{Xr_g} \equiv \sum_{\{i: c_{ig}=1\}} \boldsymbol{X_i'} [\boldsymbol{\sigma qq_i}] \boldsymbol{r_i}. \end{aligned}$$

With some algebra, we obtain

$$\begin{aligned} p(\boldsymbol{\phi}, \overline{\boldsymbol{\phi}} | \cdot, \boldsymbol{y}, \boldsymbol{s}) \propto \prod_{g=1}^{G} \phi_{\mathcal{N}} [\overline{\boldsymbol{\phi}}_g | (\boldsymbol{V_\phi^{-1}} &+ \boldsymbol{XX_g})^{-1} (\boldsymbol{V_\phi^{-1}} \boldsymbol{\phi_0} + \boldsymbol{Xr_g}), (\boldsymbol{V_\phi^{-1}} \\ &+ \boldsymbol{XX_g})^{-1}] \prod_{\{i: c_{ig}=1\}} I(\boldsymbol{\phi_i} = \overline{\boldsymbol{\phi}}_g), \end{aligned}$$

where it is to be understood that "·" in the conditioning in this case denotes all parameters other than $\boldsymbol{\phi}$, $\overline{\boldsymbol{\phi}}$, and $\boldsymbol{\theta}$. This result implies that $\boldsymbol{\phi}$ and $\overline{\boldsymbol{\phi}}$ can be sampled by first drawing independently, for $g = 1, 2, \ldots, G$, from

$$\begin{aligned} \overline{\boldsymbol{\phi}}_g | \cdot, \boldsymbol{y}, \boldsymbol{s} \sim \mathcal{N}[&(\boldsymbol{V_\phi^{-1}} + \boldsymbol{XX_g})^{-1} (\boldsymbol{V_\phi^{-1}} \boldsymbol{\phi_0} + \boldsymbol{Xr_g}), \\ &(\boldsymbol{V_\phi^{-1}} + \boldsymbol{XX_g})] \end{aligned} \tag{18}$$

and then setting, for $i = 1, 2, \ldots, n$,

$$\boldsymbol{\phi_i} = \sum_{g=1}^{G} c_{ig} \overline{\boldsymbol{\phi}}_g. \tag{19}$$

The "sampling" of $\boldsymbol{\phi_i}$ in (19) reiterates that these quantities are largely incidental to the problem, and merely simplify the exposition of the model and the algorithm.

Step 2: $\theta_i | \cdot, \boldsymbol{y}, \boldsymbol{s}$.

The conditional posterior density for the heterogeneity terms θ_i can be deduced directly from the form of the joint posterior in (15). Specifically,

$$\theta_i | \cdot, \boldsymbol{y}, \boldsymbol{s} \overset{\text{ind}}{\sim} \mathcal{N}([qq_i]^{-1} tq_i, [qq_i]^{-1}), \quad i = 1, 2, \ldots, n, \tag{20}$$

where the terms in this conditional have been defined just prior to step (1).

Step 3: $\boldsymbol{\rho}, \overline{\boldsymbol{\rho}} | \cdot, \boldsymbol{y}, \boldsymbol{s}$.

To sample the parameters $\overline{\boldsymbol{\rho}}$, we again need to introduce some additional notation. Let

$$\boldsymbol{f_i} = \begin{pmatrix} y_i - \beta_{0i} - \boldsymbol{x_i \beta}_i - s_i \theta_i \\ s_i - \delta_{0i} - \boldsymbol{x_i \delta_i} - \boldsymbol{z_i \gamma_i} \\ \theta_i - \eta_{0i} - \boldsymbol{x_i \eta_i} - \boldsymbol{w_i \lambda_i} \end{pmatrix} \quad \text{and} \quad \boldsymbol{p_i} = \begin{pmatrix} 0 \\ \theta_i \\ 0 \end{pmatrix}.$$

The conditional posterior distribution of $\boldsymbol{\rho}$ and $\overline{\boldsymbol{\rho}}$ can be shown to be, up to proportionality,

$$p(\boldsymbol{\rho}, \overline{\boldsymbol{\rho}}|\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\Sigma}, \boldsymbol{c}, \overline{\boldsymbol{\phi}}, \overline{\boldsymbol{\Sigma}}, \overline{\boldsymbol{\pi}}, \boldsymbol{y}, \boldsymbol{s}) \propto \prod_{g=1}^{G} \phi_N \left[\overline{\rho}_g | \left(V_\rho^{-1} + \sum_{i:c_{ig}=1} \boldsymbol{p}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{p}_i \right)^{-1} \times \left(V_\rho^{-1} \rho_0 + \sum_{i:c_{ig}=1} \boldsymbol{p}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{f}_i \right), \left(V_\rho^{-1} + \sum_{i:c_{ig}=1} \boldsymbol{p}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{p}_i \right)^{-1} \right] \prod_{\{i:c_{ig}=1\}} I(\rho_i = \overline{\rho}_g).$$

This implies that the sampling of $\boldsymbol{\rho}$ and $\overline{\boldsymbol{\rho}}$ can proceed by first independently drawing, for $g = 1, 2, \ldots, G$, from

$$\overline{\rho}_g|\cdot, \boldsymbol{y}, \boldsymbol{s} \overset{\text{ind}}{\sim} \mathcal{N} \left[\left(V_\rho^{-1} + \sum_{i:c_{ig}=1} \boldsymbol{p}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{p}_i \right)^{-1} \times \left(V_\rho^{-1} \rho_0 + \sum_{i:c_{ig}=1} \boldsymbol{p}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{f}_i \right), \left(V_\rho^{-1} + \sum_{i:c_{ig}=1} \boldsymbol{p}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{p}_i \right)^{-1} \right], \tag{21}$$

and then setting, for $i = 1, 2, \ldots, n$,

$$\rho_i = \sum_{g=1}^{G} c_{ig} \overline{\rho}_g. \tag{22}$$

Step 4: $\boldsymbol{\Sigma}, \overline{\boldsymbol{\Sigma}}|\cdot, \boldsymbol{y}, \boldsymbol{s}$.

From (14), the conditional posterior distribution of $\boldsymbol{\Sigma}$ and $\overline{\boldsymbol{\Sigma}}$ is, up to proportionality, given as

$$p(\boldsymbol{\Sigma}, \overline{\boldsymbol{\Sigma}}|\cdot, \boldsymbol{y}, \boldsymbol{s}) \propto p(\overline{\boldsymbol{\Sigma}}) \prod_{i=1}^{n} p(\boldsymbol{\Sigma}_i|\overline{\boldsymbol{\Sigma}}, \boldsymbol{c}_i) |\boldsymbol{\Sigma}_i|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \boldsymbol{h}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{h}_i \right) \propto \prod_{g=1}^{G} p_{IW} \left[\overline{\boldsymbol{\Sigma}}_g | p + \sum_{i=1}^{n} c_{ig}, p\boldsymbol{R} + \sum_{\{i:c_{ig}=1\}} \boldsymbol{h}_i \boldsymbol{h}_i' \right] \times \prod_{\{i:c_{ig}=1\}} I(\boldsymbol{\Sigma}_i = \overline{\boldsymbol{\Sigma}}_g).$$

This implies that the sampling of $\boldsymbol{\Sigma}$ and $\overline{\boldsymbol{\Sigma}}$ can proceed by first drawing, independently, from

$$\overline{\boldsymbol{\Sigma}}_g|\cdot, \boldsymbol{y}, \boldsymbol{s} \overset{\text{ind}}{\sim} IW \left[p + \sum_{i=1}^{n} c_{ig}, p\boldsymbol{R} + \sum_{\{i:c_{ig}=1\}} \boldsymbol{h}_i \boldsymbol{h}_i' \right], \quad g = 1, 2, \ldots, G \tag{23}$$

and then setting, for $i = 1, 2, \ldots, n$:

$$\boldsymbol{\Sigma}_i = \sum_{g=1}^{G} c_{ig} \overline{\boldsymbol{\Sigma}}_g. \tag{24}$$

Step 5: $\boldsymbol{c}_i|\cdot, \boldsymbol{y}, \boldsymbol{s}$.

To describe the sampling from the component label vector $\boldsymbol{c}_i$, we first define terms similar to those defined just prior to step 1. In this case, we make the dependence of the objects in (16) and (17) on specific parameters explicit, to avoid possible confusion when constructing the component probabilities. To this end, let

$$\boldsymbol{t}_i(\overline{\boldsymbol{\phi}}_g) \equiv \begin{pmatrix} y_i - \overline{\beta}_{0g} - \boldsymbol{x}_i \overline{\boldsymbol{\beta}}_g \\ s_i - \overline{\delta}_{0g} - \boldsymbol{x}_i \overline{\boldsymbol{\delta}}_g - \boldsymbol{z}_i \overline{\boldsymbol{\gamma}}_g \\ -\overline{\eta}_{0g} - \boldsymbol{x}_i \overline{\boldsymbol{\eta}}_g - \boldsymbol{w}_i \overline{\boldsymbol{\lambda}}_g \end{pmatrix}, \qquad \boldsymbol{q}_i(\overline{\rho}_g) \equiv \begin{pmatrix} s_i \\ \overline{\rho}_g \\ -1 \end{pmatrix},$$

$$qq_i(\overline{\rho}_g, \overline{\boldsymbol{\Sigma}}_g) \equiv q_i(\overline{\rho}_g)' \overline{\boldsymbol{\Sigma}}_g^{-1} q_i(\overline{\rho}_g),$$

$$\boldsymbol{\sigma qq}_i(\overline{\rho}_g, \overline{\boldsymbol{\Sigma}}_g) \equiv \overline{\boldsymbol{\Sigma}}_g^{-1} - \overline{\boldsymbol{\Sigma}}_g^{-1} \boldsymbol{q}_i(\overline{\rho}_g) [qq_i(\overline{\rho}_g, \overline{\boldsymbol{\Sigma}}_g)]^{-1} \boldsymbol{q}_i(\overline{\rho}_g)' \overline{\boldsymbol{\Sigma}}_g^{-1},$$

$$tt_i(\overline{\boldsymbol{\phi}}_g, \overline{\rho}_g, \overline{\boldsymbol{\Sigma}}_g) \equiv \boldsymbol{t}_i(\overline{\boldsymbol{\phi}}_g)' \boldsymbol{\sigma qq}_i(\overline{\rho}_g, \overline{\boldsymbol{\Sigma}}_g) \boldsymbol{t}_i(\overline{\boldsymbol{\phi}}_g),$$

$$\tilde{\pi}_{ig} \equiv \frac{\pi_g |\overline{\boldsymbol{\Sigma}}_g|^{-\frac{1}{2}} |qq_i(\overline{\rho}_g, \overline{\boldsymbol{\Sigma}}_g)|^{-\frac{1}{2}} \exp\left[-\frac{1}{2} tt_i(\overline{\boldsymbol{\phi}}_g, \overline{\rho}_g, \overline{\boldsymbol{\Sigma}}_g) \right]}{\sum_{h=1}^{G} \pi_h |\overline{\boldsymbol{\Sigma}}_h|^{-\frac{1}{2}} |qq_i(\overline{\rho}_h, \overline{\boldsymbol{\Sigma}}_h)|^{-\frac{1}{2}} \exp\left[-\frac{1}{2} tt_i(\overline{\boldsymbol{\phi}}_h, \overline{\rho}_h, \overline{\boldsymbol{\Sigma}}_h) \right]},$$

and $\tilde{\boldsymbol{\pi}}_i \equiv (\tilde{\pi}_{i1} \ \tilde{\pi}_{i2} \ \cdots \ \tilde{\pi}_{iG})'$. Straightforward algebra produces that

$$\boldsymbol{c}_i|\cdot, \boldsymbol{y}, \boldsymbol{s} \sim \text{Mult}(1, \tilde{\boldsymbol{\pi}}_i). \tag{25}$$

Step 6: $\overline{\boldsymbol{\pi}}|\cdot, \boldsymbol{y}, \boldsymbol{s}$.

Finally, let

$$\tilde{\boldsymbol{\alpha}} = \begin{pmatrix} \alpha_1 + \sum_{i=1}^{n} c_{i1} \\ \alpha_2 + \sum_{i=1}^{n} c_{i2} \\ \vdots \\ \alpha_G + \sum_{i=1}^{n} c_{iG} \end{pmatrix}.$$

The conditional posterior distribution of $\overline{\pi}$ can be shown to be

$$\overline{\boldsymbol{\pi}}|\cdot, \boldsymbol{y}, \boldsymbol{s} \sim \text{Dirichlet}(\tilde{\boldsymbol{\alpha}}). \tag{26}$$

A posterior simulator proceeds by sampling from (18)–(26).

3.3. Causal effect heterogeneity

Having discussed issues of model flexibility and methods for posterior simulation, we now turn our attention to key parameters of interest. Of course, a primary focus of our model concerns the causal effect heterogeneity terms, θ_i. In particular, we would like to make use of our analysis to answer questions such as the following. What have we learned about the overall distribution of such impacts in the population? How can we use our model to make predictive statements regarding the effect of future, out-of-sample treatments?

We separately consider the cases of in-sample and out-of-sample prediction. To this end, we first note from (5)–(10) that, marginally,

$$p(\theta_i|\cdot) \sim \sum_{g=1}^{G} \pi_g \phi(\theta_i; \eta_{0g} + \boldsymbol{x}_i \boldsymbol{\eta}_g + \boldsymbol{w}_i \boldsymbol{\lambda}_g, \sigma_{\theta g}^2),$$

where the subscript g denotes parameters associated with the gth component of the mixture, the "$\cdot$" in the conditioning explicitly reflects that we are conditioning on the model's parameters, and $\phi(x; \mu, \sigma^2)$ denotes a normal density for x with mean μ and variance σ^2. Thus, our model assumes that the distribution of treatment effect heterogeneity can be adequately represented as a finite mixture of Gaussian distributions, which seems unlikely to be a controversial assumption in practice.

If interest centers on summarizing the overall shape of the heterogeneity distribution, or in making predictions about

treatment returns for a future sample, such questions can be addressed by deriving and calculating the appropriate posterior predictive density. For example, adding a subscript f to denote "future" outcomes, and considering the case of a particular agent with known characteristics $\boldsymbol{x_f}$ and $\boldsymbol{w_f}$, the desired posterior predictive distribution can be obtained via "Rao–Blackwellization". Specifically,

$$p(\theta_f|\boldsymbol{x_f}, \boldsymbol{w_f}, \boldsymbol{y}, \boldsymbol{s}) \approx \frac{1}{M}\sum_{m=1}^{M}\sum_{g=1}^{G}\pi_g^{(m)}\phi(\theta_f; \eta_{0g}^{(m)} + \boldsymbol{x_f}\boldsymbol{\eta_g^{(m)}} + \boldsymbol{w_f}\boldsymbol{\lambda_g^{(m)}}, \sigma_{\theta g}^{2\ (m)}),$$

where M denotes the total number of post-convergence simulations and m indexes these simulations. A single, "representative" posterior predictive density could be calculated by fixing $\boldsymbol{x_f}$ and $\boldsymbol{w_f}$ to their sample means or rounded integer values, as appropriate. Perhaps more desirably, an additional step can be added to average these over the empirical distribution of the $\boldsymbol{x_f}$ and $\boldsymbol{w_f}$ characteristics to effectively drop the conditioning on $\boldsymbol{x_f}$ and $\boldsymbol{w_f}$.

In this out-of-sample exercise, we focus on the θ_f marginal density since the future treatment level s_f and future outcome y_f are not observed. Within the sample, however, this is not the case, and the mechanism for learning about θ_i is slightly different. In particular, the structure of our model suggests that the observed y_i and s_i convey information about θ_i beyond what is learned from (7) only. To shed some light on this issue more formally, let $\boldsymbol{\Gamma}$ denote all parameters other than θ_i, and note that

$$\begin{aligned} p(\theta_i|\boldsymbol{y}, \boldsymbol{s}) &= \int p(\theta_i, \boldsymbol{\Gamma}|\boldsymbol{y}, \boldsymbol{s})\mathrm{d}\boldsymbol{\Gamma} \\ &= \int p(\theta_i|\boldsymbol{\Gamma}, \boldsymbol{y}, \boldsymbol{s})p(\boldsymbol{\Gamma}|\boldsymbol{y}, \boldsymbol{s})\mathrm{d}\boldsymbol{\Gamma} \\ &= \int p(\theta_i|\boldsymbol{\Gamma}, y_i, s_i)p(\boldsymbol{\Gamma}|\boldsymbol{y}, \boldsymbol{s})\mathrm{d}\boldsymbol{\Gamma}, \end{aligned}$$

where the last line follows from the fact that, given $\boldsymbol{\Gamma}$, θ_i is independent of outcomes other than y_i and s_i.

To fix ideas and make progress in understanding how we learn about specific within-sample treatment impacts, let us consider the single-component Gaussian model.[12] The conditional posterior $p(\theta_i|\boldsymbol{\Gamma}, y_i, s_i)$, which is to be averaged over $p(\boldsymbol{\Gamma}|\boldsymbol{y}, \boldsymbol{s})$ to obtain $p(\theta_i|\boldsymbol{y}, \boldsymbol{s})$, can be obtained from (1)–(3). Importantly, this derivation shows that the conditional posterior distribution of θ_i is not just the marginal density in (3), but, instead, the conditional distribution $\theta_i|\boldsymbol{\Gamma}, y_i, s_i$ is normal with mean $E(\theta_i|\boldsymbol{\Gamma}, y_i, s_i)$, given in Box I, where we have defined $\mu_{\theta i} \equiv \eta_0 + \boldsymbol{x_i}\boldsymbol{\eta} + \boldsymbol{w_i}\boldsymbol{\lambda}$, $\tilde{y}_i \equiv y_i - \beta_0 - \boldsymbol{x_i}\boldsymbol{\beta}$, $\tilde{s}_i \equiv s_i - \delta_0 - \boldsymbol{x_i}\boldsymbol{\delta} - \boldsymbol{z_i}\boldsymbol{\gamma}$, and σ^{ij} denotes the (i, j) element of Σ^{-1}.[13] This mean, of course, is different from $\mu_{\theta i}$, the mean of (3) that is used for out-of-sample prediction purposes. The law of iterated expectations then implies that

$$E(\theta_i|\boldsymbol{y}, \boldsymbol{s}) = E_{\boldsymbol{\Gamma}|\boldsymbol{y},\boldsymbol{s}}\big[E(\theta_i|\boldsymbol{\Gamma}, \boldsymbol{y}, \boldsymbol{s})\big] = E_{\boldsymbol{\Gamma}|\boldsymbol{y},\boldsymbol{s}}\big[E(\theta_i|\boldsymbol{\Gamma}, y_i, s_i)\big],$$

so that the posterior expected return to treatment for agent i is the conditional posterior mean $E(\theta_i|\boldsymbol{\Gamma}, y_i, s_i)$ averaged over the posterior distribution of $\boldsymbol{\Gamma}$.

With a little rearranging, the conditional posterior mean $E(\theta_i|\boldsymbol{\Gamma}, y_i, s_i)$ above can be represented as a type of weighted average of three pieces: $\tilde{y}_i/s_i$, $\tilde{s}_i/\rho$, and $\mu_{\theta i}$. These three pieces emerge quite naturally as "estimators" of θ_i from (1)–(3), as each of these involves "solving" for θ_i in the respective equations. In the limiting case where $\boldsymbol{\Sigma}$ is diagonal, it is straightforward to show, for example, (holding all else constant in each case) that $\theta_i|\boldsymbol{\Gamma}, y_i, s_i$ collapses around $\tilde{y}_i/s_i$ as $\sigma_y \to 0$, collapses around $\tilde{s}_i/\rho$ as $\sigma_s \to 0$, and collapses around $\mu_{\theta i}$ as $\sigma_\theta \to 0$. Thus, in-sample predictions regarding individual-level treatment impacts use information from all three equations of our system, and therefore more precise estimates of our outcome and treatment equations in (1) and (2) can lead to better learning about individual-level causal effect parameters.

4. Generated data experiments

We illustrate the performance of our algorithm via two generated data experiments. In the first experiment, we simulate a large sample size of $n = 10{,}000$ observations from a correctly specified two-component mixture version of the model in (1)–(3). For this case, the exogenous variables are generated as follows:

$$\begin{pmatrix} x_i' \\ z_i \\ w_i \end{pmatrix} = \begin{pmatrix} x_{1,i} \\ x_{2,i} \\ z_i \\ w_i \end{pmatrix} \sim N_4\left[\begin{pmatrix}0\\0\\0\\0\end{pmatrix}, \begin{pmatrix} 1 & 0.1 & -0.2 & 0.1 \\ 0.1 & 0.25 & -0.2 & 0.05 \\ -0.2 & -0.2 & 4 & -0.2 \\ 0.1 & 0.05 & -0.2 & 1 \end{pmatrix}\right], \quad (27)$$

and the following hyperparameters are selected: $\phi_0 = 0_{k\times 1}$, $V_\phi = 10^6 \times I_k$, $\rho_0 = 0$, $V_\rho = 10^6$, $p = 8$, $R = \mathrm{diag}\{1, 10^2, 0.1^2\}$, $\alpha_1 = 1, \alpha_2 = 1$. These yield reasonably "diffuse" marginal prior distributions, so the information contained in the prior is small relative to the information contained in the data.

Table 1 reports the actual parameters of the data-generating process as well as their posterior means and posterior standard deviations from the experiment. The results of this table reveal that our algorithm successfully recovers the parameters of the data-generation process and that our code for fitting such models is likely to be free of errors.[14] Experiments with fewer observations provided similar results, and, finally, experiments based upon more than two mixture components also revealed that the code and posterior simulator performed adequately.[15]

In Table 2, we illustrate the performance of our method in a different way. The table presents inefficiency factors associated with the sampling of three different parameters: $\overline{\beta}_{01}$, $\overline{\rho}_1$, and $\overline{\Sigma}_{yy1}$. The first column presents such factors using our posterior simulator outlined in the previous section, while the second column, for the sake of comparison, presents analogous results for a sampler that fails to block ϕ and θ together. As the table clearly illustrates, the gains to blocking are substantial, and, moreover, the mixing of the simulations in our sampler is adequate, though still falling somewhat short of the numerical efficiency obtained under i.i.d. sampling.

A second generated data experiment was also conducted to illustrate the performance of the mixture method under misspecification. Since the return heterogeneity distribution is of

[12] Alternatively, what follows applies to a particular component of the mixture, so that this assumption is made essentially without loss of generality.

[13] It is also worth mentioning that $\mathrm{Var}(\theta_i|\boldsymbol{\Gamma}, y_i, s_i)$ is simply the inverse of the denominator in the expression for $E(\theta_i|\boldsymbol{\Gamma}, y_i, s_i)$ above.

[14] A more formal diagnosis of the code was obtained by performing some of the checks suggested by Geweke (2004). Though not reported here, these tests did not provide any evidence that a mistake had been made.

[15] It is worth noting that the ability of our algorithm to accurately estimate the true parameter values clearly depended on the quality of the instruments (i.e., the magnitude of γ and λ) and the degree of confounding (i.e., ρ_{ys}, $\rho_{y\theta}$, $\rho_{s\theta}$). We do not attempt to further characterize these relationships in the present study, as doing so thoroughly will lead us well beyond the scope and goals of this paper. Whether or not such issues are relevant for the applied researcher will inevitably depend on the application at hand and the data available.

$$E(\theta_i|\boldsymbol{\Gamma}, y_i, s_i) = \frac{\sigma^{11}\tilde{y}_i s_i + \sigma^{22}\tilde{s}_i\rho + \sigma^{33}\mu_{\theta i} + \sigma^{12}[s_i\tilde{s}_i + \rho\tilde{y}_i] - \sigma^{13}[\tilde{y}_i + s_i\mu_{\theta i}] - \sigma^{23}[\tilde{s}_i + \rho\mu_{\theta i}]}{\sigma^{11}s_i^2 + \sigma^{22}\rho^2 + \sigma^{33} + 2\sigma^{12}s_i\rho - 2\sigma^{13}s_i - 2\sigma^{23}\rho}.$$

Box I.

Table 1
Parameter posterior means, standard deviations, and true values.

Parameter	Component 1			Component 2		
	True value	E(β|D)	Std(β|D)	True value	E(β|D)	Std(β|D)
Component probability						
$\overline{\pi}_1$	0.7	0.689	0.006	0.3	0.311	0.006
Equation for y						
$\overline{\beta}_0$	0.5	0.523	0.033	1	1.07	0.084
$\overline{\beta}_1$	−3	−2.94	0.039	−1.5	−1.4	0.111
$\overline{\beta}_2$	−1	−1.06	0.054	−3	−3.0	0.161
Equation for s						
$\overline{\delta}_0$	1.5	1.45	0.044	0.5	0.351	0.075
$\overline{\delta}_1$	−2	−1.89	0.051	2	2.07	0.081
$\overline{\delta}_2$	1	0.943	0.056	−1	−1.14	0.135
$\overline{\gamma}$	−1.5	−1.49	0.014	1.5	1.51	0.031
$\overline{\rho}$	2.5	2.47	0.013	1.5	1.48	0.021
Equation for θ						
$\overline{\eta}_0$	−2.5	−2.52	0.008	−2	−1.96	0.025
$\overline{\eta}_1$	3	3	0.007	2.5	2.48	0.021
$\overline{\eta}_2$	−0.5	−0.495	0.015	−2.5	−2.54	0.044
$\overline{\lambda}$	2	2	0.007	3.0	3.0	0.021
Covariance matrix						
$\overline{\sigma}_y$	1	0.975	0.027	2	1.98	0.071
$\overline{\sigma}_s$	2	2	0.022	3	3.03	0.046
$\overline{\sigma}_\theta$	0.5	0.495	0.006	1	0.98	0.017
$\overline{\rho}_{ys}$	0.2	0.16	0.030	0.1	0.08	0.044
$\overline{\rho}_{y\theta}$	−0.1	−0.136	0.030	−0.2	−0.15	0.042
$\overline{\rho}_{s\theta}$	0.1	0.0941	0.016	0.2	0.24	0.024

Table 2
Inefficiency factors for a selection of three parameters.

Parameter	Blocking ϕ and θ together	Without blocking ϕ and θ together
$\overline{\beta}_{01}$	2.72	929
$\overline{\rho}_1$	22.84	3554
$\overline{\Sigma}_{yy1}$	34.99	388

Table 3
Bayes factors supporting the four-component model from the second generated data experiment.

Bayes factor	$j=2$	$j=3$	$j=4$	$j=5$
$p(y,s\|\mathcal{M}_4)/p(y,s\|\mathcal{M}_j)$	2.62×10^{814}	1.05×10^{155}	1	1.60

primary interest in our study, we introduced a departure from normality in the generation of these returns by first sampling ϵ_i from a lognormal distribution with reasonable skew, recentered to have mean zero:

$$\epsilon_i \overset{\text{i.i.d.}}{\sim} \text{lognormal}\big(0, 0.25\big) - \exp(0.125).$$

The joint distribution of u_i and v_i was then sampled as a (conditionally) bivariate normal:

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \Bigg| \epsilon_i \overset{\text{ind}}{\sim} \mathcal{N}\left[\begin{pmatrix} -0.2\log[\epsilon_i + \exp(0.125)] \\ 0.4\log[\epsilon_i + \exp(0.125)] \end{pmatrix}, \begin{pmatrix} 0.99 & 0.42 \\ 0.42 & 3.96 \end{pmatrix}\right],$$

and the parameter values and process for generating the covariates were then identical to those employed for the first component of the first generated data experiment. Unconditional moments of the joint distribution of u, v, and ϵ can be derived for this experiment, but we omit these details here for the sake of brevity. Instead, we focus on the most important issue of assessing how well our mixture model fares at picking up this departure from normality when it is present.[16]

We fit a variety of different mixture models to this data, focusing on models with 2–5 different mixture components. For the sake of parsimony, we only allow the intercepts and covariance matrices of the mixture components to differ and restrict all the slope coefficients to be the same across components. A summary of the different model performances in terms of the return distribution is presented graphically in Fig. 1. To fix ideas, we used the simulations produced from the algorithm of Section 3.2 to obtain a posterior predictive return heterogeneity distribution for an individual of average characteristics (i.e., setting all covariates to zero). We repeated this exercise four different times, considering models with 2–5 mixture components.[17] We then compared the posterior predictive densities for these cases to the actual heterogeneity distribution for this "average" individual.

As is evident from Fig. 1, the mixture models perform well. Even the two-component model is able to capture the most salient features of the lognormal heterogeneity distribution, while the four-component and five-component models are able to capture its shape almost exactly. Table 3 shows the calculated Bayes factors[18] for the competing mixture models, using the notation $\mathcal{M}_j$ to denote the specification employing j mixture components. The results in the table are reported as Bayes factors in support of the four-component model. These clearly reveal that the four-component specification is favored relative to those with two or three components, and also reveals near indifference between the

[16] The slope coefficients and remaining model parameters were also well estimated in this exercise. Details of these results are available upon request.

[17] The one-component Gaussian model produced a symmetric predictive return distribution and was clearly inferior to those with more mixture components.

[18] These were computed using the method described in Gelfand and Dey (1994) and Geweke (1999).

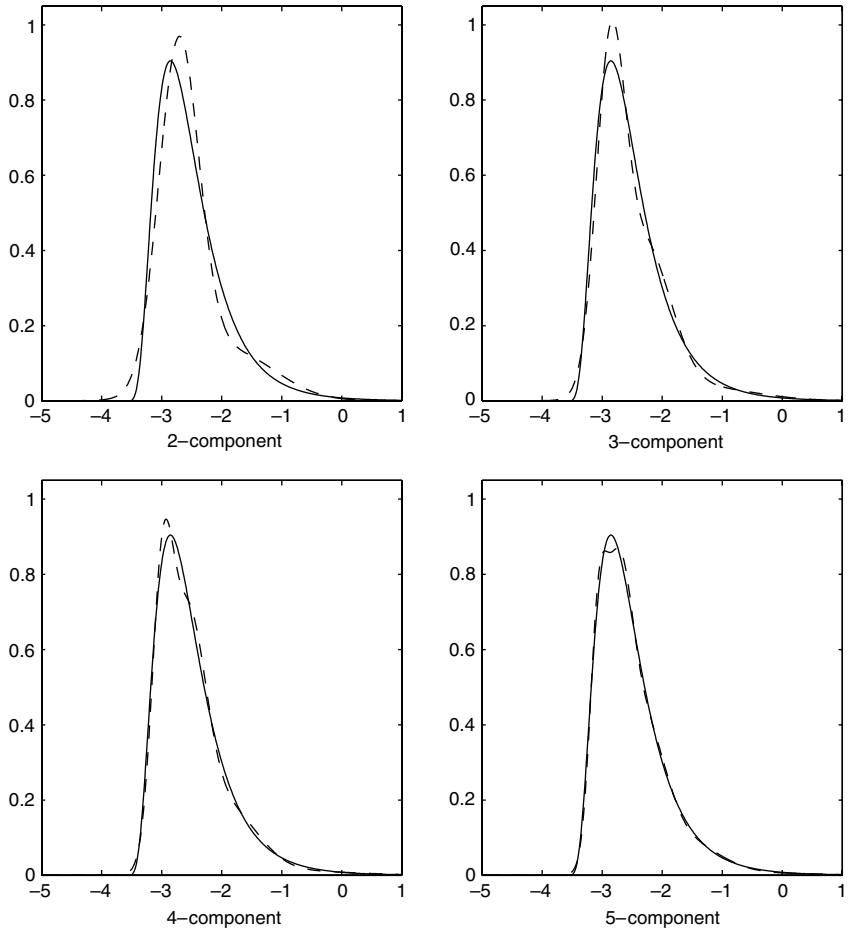

Fig. 1. True return heterogeneity distribution for an individual of average characteristics (solid lines) compared to posterior predictive densities estimated from models with 2–5 mixture components (dashed lines).

use of four and five mixture components. As Fig. 2 suggested, both the four-component and five-component models performed well in terms of capturing features of the return heterogeneity distribution, and our marginal likelihood calculations weakly support the more parsimonious four-component model over the five-component alternative.

The results of this exercise are encouraging, and they suggest that the mixture models can fare well in picking up departures from normality. Of course, this exercise has not investigated other sources of misspecification, such as measurement error or incorrect specification of the conditional mean functions. We should not expect our method to be immune to such problems, and indeed it will not be; what we have documented here is simply a degree of robustness to distributional assumptions in the absence of other confounding problems.

5. The data

For our empirical application, we make use of data from two distinct sources. The first and primary data set, which has been widely employed in the applied literature, is the High School and Beyond (HSB) survey. HSB is a survey conducted on behalf of the National Center for Education Statistics (NCES), and it was designed with the intent of yielding a sample of students representative of American high school students. HSB is a biennial survey starting in 1980, and we focus attention on the sophomore subsample of the HSB data. For our earnings outcome measure, we employ the most recent data available to us, the 1992 survey, from which 1991 earnings can be obtained. In practice, we restrict the HSB sample to individuals who have worked for at least nine months during 1991, and whose monthly earnings were between $500 and $6000. Finally, it is worth mentioning that the HSB data set employed here, like several other widely used micro data sets, contains a wealth of demographic information on the sample respondents, such as family background characteristics and individual test scores, making it an attractive data source for our application. Descriptive statistics associated with key variables in the model are provided in Table 4.

Table 4
Descriptive statistics.

Variable	Mean	Std. Dev.
Log monthly earnings	7.51	0.465
Schooling	13.6	2.06
Father's education	12.5	3.28
Mother's education	12.3	2.81
Base year family income ($10,000)	2.09	0.999
Base year test score	0	1
Number of siblings	2.92	1.61
Female	0.463	0.499
Age as of 1 January 1991	26.8	0.536
Hispanic	0.154	0.361
Native American	0.0191	0.137
Asian/Pacific	0.0316	0.175
Black	0.123	0.328
Other/Missing race	0.00371	0.0608
1980 county grp. avg. log hourly wage	1.53	0.109
1980 county grp. avg. schooling	12.7	0.434
1980 county grp. avg. return to schooling	0.0785	0.012

As discussed earlier in the paper, for identification purposes, we require an instrument or set of instruments. The most important

of these is some characteristic or set of characteristics that are conditionally correlated with individual-level returns to schooling, but can be excluded from the earnings and schooling equations. As shown in the Appendix, the model parameters are fully identifiable with such an exclusion restriction, provided that $\rho \neq 0$. Our choice in this regard is to obtain a county-level return to schooling estimate based on 1980 Census data and use this lagged return as a right-hand side variable in Eq. (3).

The lagged county-level returns are constructed using data from the public use 5% sample of the 1980 Census. We restrict the Census sample to those individuals who are between 16 and 28 years of age[19] working for at least 40 hours a week in 1979 with an hourly wage between \$1 and \$100. For each race and gender cell within a given county group, we calculate the corresponding county-level average log hourly wage, highest grade completed, and return to schooling.[20] This county-level return to schooling is obtained by running a regression of individual log hourly wages in the given county on highest grade completed, potential labor market experience (age minus schooling minus 6), and potential experience squared. We consider the schooling coefficient estimated from this regression as the 1980 county-level average return to schooling for that group.[21]

The lagged county-level average returns are then matched with the HSB data. This, of course, requires county identifiers for the HSB sample. However, this matching is not trivially performed, as there are no directly available county-level indicators in HSB, and the NCES does not publicize this information. To this end, we follow and expand upon the approach of Hanushek and Taylor (1990), Rivkin (1991), Ganderton (1992), Grogger (1996a,b), and Li (2006, 2007), who are able to match individuals to *states* of residence via other information provided in HSB. Specifically, a school survey component of the HSB data provides a variety of information on local labor market conditions associated with each school represented in the survey. In practice, we implement our identification strategy in two steps. First, we utilize the available state-level geographically related HSB variables and match them to publicly available state-level data to uncover the state associated with each HSB individual. In the second step, once the state has been identified, we repeat the same procedure at the county level by making use of available county-level labor market conditions to identify the county of residence for each HSB participant during their high school years. In this way, we are able to match lagged county-level returns to schooling to each individual in the HSB sample.[22]

The validity of lagged returns to schooling as an instrumental variable rests on two assumptions. First, we must assume that lagged county-level returns to education are correlated with the (contemporaneous) private return to education for the given individual. Recognizing that many individuals will choose to work in the same county group as their high school was located, this correlation may result from a type of autoregressive process in county-level returns to education. Unlike traditional IV analyses, this first identification assumption, however, is not "directly" empirically testable, as θ_i in (3) is not observed.[23]

The second (and surely more controversial) assumption is that the lagged county-level returns can be excluded from, most importantly, the schooling equation in (2) and, to a lesser extent, the log wage equation in (1). That is, conditioned on a variety of individual-level controls, 1980 average returns to education in a county are uncorrelated with unobservables affecting wages and educational attainment observed in 1991.

There are certainly a few reasons to think that this assumption is suspect. For example, educational attainment decisions may be based, in part, on the contemporaneous return to schooling observed by the agent. That is, sophomores in 1980 (who generally become seniors in 1982) may make decisions about college entry based on currently available information regarding the return to a college degree. If this is the case, then lagged returns may have some non-ignorable role in explaining 1991 schooling outcomes, which would undermine our identification strategy. Our assumption in this regard, however, is that the agent makes educational attainment decisions based on his or her *own* return to education parameter θ_i, and conditioned on this parameter (and other characteristics), lagged county-level information is superfluous. Again, this assumption is probably not without controversy, but we maintain it in the current analysis. As a way to partially mitigate some of these effects, we also include in the log earnings equation the 1980 average log earnings for that county group, and likewise, in the schooling equation, we include the 1980 average level of schooling for that particular county group. Thus our argument is that, conditioned on the individual return to schooling parameter as well as lagged average levels of schooling and earnings, lagged county-level returns to education do not play an independent structural role in schooling decisions and earnings determination.

Finally, in addition to the instruments described above, we also control for parental education and income, family size, sophomore year test scores,[24] age, gender, and a variety of racial indicator variables. This produced an HSB sample of 8886 individuals from 471 county groups (or 536 counties).

6. Empirical results

Before discussing results from any particular model, we first consider the general issue of model selection. To this end, we estimate the single-component normal model along with two-component and three-component normal mixture models as competing specifications. For each case, we run our posterior simulator for 200,000 iterations and discard the first 20% (40,000) as the burn-in period. For the mixture models, we restrict all slope coefficients to be the same across components, yet allow the intercepts for each equation and covariance matrices to differ. These slope restrictions were imposed in order to minimize added parameterization while still being able to accommodate skew, heavy tails, or other departures from normality.

[19] The HSB sophomores were around 16 in 1980 and 28 in 1992. Therefore, labor market outcomes and educational attainments of individuals from the Census aged between 16 and 28 can be considered most relevant to our HSB sample.

[20] In the 1980 Census, county groups are typically defined as contiguous areas with an aggregate population of at least 100,000. They may be actual county groups or single counties. In some cases, a county is split up into several "county groups". In such situations, we create a larger county group by encompassing all "county groups" belonging to the same county.

[21] In some cases, a subgroup (i.e., a particular race and gender category within a particular county group) is found to have fewer than 15,000 observations. In such cases, we first enlarge the sample by including all persons who are from the same county group and of the same sex, regardless of their ethnic background, and add a set of race dummies to the log hourly wage regression. If the resulting pooled sample size is still below 15,000 observations, we combine all people from the same county group, irrespective of their gender and ethnic origin, and then include dummies for gender and race in the log hourly wage regression.

[22] Specific (and tedious) details regarding how this is done are available upon request.

[23] We do, however, find strong evidence supporting a role for the lagged returns in (3), as documented in the following section.

[24] In 1980, the sophomores who participated in the HSB survey also took a battery of seven tests that were designed to measure the cognitive abilities of these individuals. We add together the number of questions answered correctly in the seven tests and rescale this variable so that it has a mean of 0 and a standard deviation of 1.

Under the same priors employed for the generated data experiments of Section 4 (and equal prior probabilities over each of these three models), we find evidence against normality for our application. Specifically, the two-component specification is favored over the "textbook" Gaussian model by an overwhelming factor of 2.58×10^{15} and it is also favored over the more general three-component model by a factor of 1043. This support for the more parameter rich two-component model over the Gaussian specification occurs despite the fact that our priors, though proper, are still quite uninformative.[25] Given these results, we focus in the remainder of this discussion on results obtained from the two-component mixture model, as it is strongly preferred over these competitors, and model-averaged posteriors would essentially reduce to those same posteriors obtained under the two-component specification.

We first consider the log monthly earnings equation from the two-component model, as reported in Table 5.[26] From this first portion of the table it is evident that, holding all else constant, males earn more, on average, than women, as do workers who come from smaller families (in terms of fewer siblings) with high annual incomes.[27] In addition, whites, the excluded category, generally earn more than other racial/ethnic groups, and the lagged county-level average log wage also has an important role in describing current monthly earnings.[28]

Our second equation explains the variation in the quantity of schooling attained by individuals in our sample. There is overwhelming evidence from Table 5 supporting the assertions that, holding all else constant, the quantity of education attained increases with student-level test scores and also increases with parental education and income, as no posterior simulations associated with these parameters were negative. Similarly, children of larger families attain less schooling, on average, than those from smaller families, while those of Asian/Pacific descent attain about one more year of education than whites.

We also find an important role for the individual-level return to schooling parameter θ_i in the schooling equation. Specifically, a one percentage point increase in the return to education will lead the individual to acquire 0.135 more years of schooling, on average. If we consider a one-standard-deviation increase in the conditional distribution of the return to education parameter (which corresponds, approximately, to increasing θ_i by 0.05), this leads to an expected increase in the number of years of education equal to 2/3 of a year. These results suggest both statistical and economic significance of the return to education variable in the schooling equation, as the associated marginal effect described above is clearly meaningfully large, and more than 99% of the posterior simulations associated with this coefficient were positive. The results here are quite interesting, as they clearly reveal that agents with higher returns to education do, in fact, acquire more schooling.

The final equation of our system, with estimates reported in Table 6, explains the individual-level variation in returns to education. In terms of coefficient point estimates, the results of the table generally suggest that those family background characteristics leading individuals to acquire more schooling and receive higher earnings, such as parental education and family income, also tend to *lower* an individual's return to an added year of education.[29] This interesting result makes some intuitive sense, since we can certainly imagine that a college degree, for example, may significantly alter the earnings profile for someone coming from a low-income family. At the same time, a college degree for an individual from a high-income family will also be valued in the labor market, but it is seemingly likely that the high-income child, owing to family connections or other social networks, would fare better in the absence of the college degree than the low-income individual. Finally, we also note that returns to schooling do not seem to vary in any systematic way with test scores, as the posterior probability that this parameter was positive was 0.68. This is not to say that test scores do not matter in the production of wages and education—indeed previous portions of the table clearly point to an important role for test scores in the production of both variables. Instead, we find little evidence that returns to education vary in a systematic way with cognitive ability; Koop and Tobias (2004) also document a similar result, albeit with a very different model and data set.

In addition, we find modest evidence supporting the notion that minority groups – blacks in particular and Hispanics to a lesser extent – have higher returns to education than whites, and that females have higher returns to education than men.[30] Lagged county-level returns to education were also clearly important in explaining the individual-level variation in returns to schooling, which is critical for identification purposes. Specifically, a one percentage point increase in the 1980 return to education in the county is associated with a 0.16 percentage point increase in the individual's 1991 private return, and no posterior simulations associated with this parameter were negative.

6.1. Decomposing a covariate's effect on log wages

The foregoing discussion clearly illustrates that a covariate in our model has many channels through which it impacts log wage outcomes. Our previous discussion of results has, in fact, focused primarily on directional impacts and brief discussions akin to the "significance" of particular variables in light of the multifaceted nature of their influences. Variables such as family income and parental education, for example, have direct effects on schooling levels and returns to schooling, and each of these filter through

[25] Bartlett's paradox, for example, illustrates that the adoption of such priors results in the Bayes factor lending support for the restricted model.

[26] Although not reported in this set of tables, calculations of the coefficient prior means and prior standard deviations reveal that a substantial amount of learning has taken place regarding all slope, intercept, and covariance matrix parameters. The priors used are the same as those employed in Section 4. For the sake of space, we do not report in the table posterior statistics for the intercept and covariance matrix parameters, although these are available upon request. The results do, however, show a strong, positive correlation between schooling and log earnings unobservables, and strong negative correlations between these unobservables and those associated with returns to education. A similar pattern is found in terms of observed characteristics, a point we discuss later in this section.

[27] Values of the third column are reported as one (zero) when all (none) of the posterior simulations were positive.

[28] In the HSB data information on family income, parental education, number of siblings, and base year test score are often missing. We do not, in this paper, take up the issue of how best to model the missing data, or whether these observations are missing at random, or not missing randomly. Instead, in the case where these variables are absent, we set the corresponding variables equal to their sample mean values and add a dummy variable to the regression equation denoting whether or not the given covariate is missing or observed in the sample. The posterior means and standard deviations of the parameters associated with the missing indicators are included in the analysis but not reported in the tables for the sake of brevity.

[29] As the reader can see, there is considerable uncertainty associated with many of the parameters at this stage of the model. Formal Bayes factors, computed via the Savage–Dickey density ratio, were found to support the inclusion of only the black, female, and lagged county-level returns to education variables in Eq. (3), although there is rather considerable support based on the marginal posterior distributions for retaining the mother's education and family income as well. Of course, the priors employed for these parameters were quite flat, lending substantial prior support to the restricted variants of the model (e.g., Bartlett's paradox). The results of the table are clearly suggestive that females and blacks have higher returns to education while individuals from wealthy families have lower returns to schooling.

[30] Henderson et al. (2009) recently document a similar result, as have previous studies in the literature.

Table 5
Posterior means, standard deviations, probabilities of being positive, and numerical standard error (NSE) values from the two-component mixture model.

Variable	E($\beta|D$)	Std($\beta|D$)	Pr($\beta > 0|D$)	NSE
Log monthly earnings equation				
Father's education	0.00828	0.0116	0.761	4.17e−005
Mother's education	0.0208	0.013	0.945	5.43e−005
Base year family income ($10,000)	0.15	0.033	1	9.7e−005
Base year test score	0.0185	0.0372	0.691	0.000606
Number of siblings	−0.0214	0.0193	0.133	5.59e−005
Female	−0.527	0.0624	0	0.000723
Age as of 1 January 1991	−0.0437	0.0589	0.21	0.00996
Hispanic	−0.0634	0.0914	0.244	0.00026
Native American	−0.299	0.266	0.13	0.000777
Asian/Pacific	0.0286	0.179	0.562	0.000543
Black	−0.281	0.102	0.00286	0.000293
Other/Missing race	−0.682	0.539	0.102	0.00148
1980 county group average log hourly wage	0.47	0.041	1	0.00194
Schooling equation				
Father's education	0.117	0.0142	1	0.00023
Mother's education	0.0894	0.0171	1	0.000667
Base year family income ($10,000)	0.236	0.0595	1	0.00483
Base year test score	0.724	0.0437	1	0.00103
Number of siblings	−0.0824	0.0242	0.000687	0.000442
Female	−0.0807	0.148	0.307	0.0141
Age as of 1 January 1991	−0.379	0.0745	0	0.0119
Hispanic	0.218	0.121	0.952	0.00435
Native American	−0.401	0.341	0.098	0.00618
Asian/Pacific	0.918	0.206	1	0.000617
Black	0.24	0.154	0.928	0.0101
Other/Missing race	−0.509	0.738	0.248	0.0293
1980 county group average schooling	0.149	0.0425	1	0.00326
Return to schooling	13.5	5.2	0.997	0.569

Table 6
Posterior means, standard deviations, probabilities of being positive, and numerical standard error (NSE) values from the two-component mixture model.

Variable	E($\beta|D$)	Std($\beta|D$)	Pr($\beta > 0|D$)	NSE
Return to schooling equation				
Father's education	−0.00057	0.000831	0.247	2.41e−006
Mother's education	−0.00128	0.000945	0.0879	3.07e−006
Base year family income ($10,000)	−0.00868	0.00236	0.00015	6.8e−006
Base year test score	0.00119	0.0026	0.678	3.77e−005
Number of siblings	0.00111	0.00141	0.785	4.09e−006
Female	0.0242	0.00452	1	5.45e−005
Age as of 1 January 1991	0.00137	0.00442	0.626	0.000784
Hispanic	0.00846	0.00678	0.894	1.93e−005
Native American	0.0155	0.0208	0.771	6.09e−005
Asian/Pacific	0.00121	0.0121	0.541	3.55e−005
Black	0.0174	0.00748	0.99	2.15e−005
Other/Missing race	0.0548	0.0419	0.906	0.000115
1980 county grp. avg. return to schooling	0.162	0.0261	1	0.000125

the model to define that variable's "total" impact on earnings. In attempt to identify this total impact as well as the component pieces that define it, in this section we look into the posterior predictive distribution, as discussed in Section 3.3.

Specifically, let y_f, s_f, and θ_f denote the log monthly earnings, the level of schooling, and the return to schooling parameter for some hypothetical or "future" individual f. The posterior predictive distribution of these outcomes can be obtained as

$$p(y_f, s_f, \theta_f|\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{x_f}, \boldsymbol{w_f}, \boldsymbol{z_f}) = \int p(y_f, s_f, \theta_f|\boldsymbol{\Gamma}_{-\theta}, \boldsymbol{x_f}, \boldsymbol{w_f}, \boldsymbol{z_f}) p(\boldsymbol{\Gamma}_{-\theta}|\boldsymbol{y}, \boldsymbol{s}) \mathrm{d}\boldsymbol{\Gamma}_{-\theta}. \quad (28)$$

Samples from this trivariate posterior predictive distribution can therefore be drawn, given a set of simulations from the posterior distribution $p(\boldsymbol{\Gamma}_{-\theta}|\boldsymbol{y}, \boldsymbol{s})$, the maintained model in (1)–(3), and values of the covariates $\boldsymbol{x_f}$, $\boldsymbol{w_f}$, and $\boldsymbol{z_f}$.

We generate a series of simulations from this posterior predictive distribution and use these to summarize the effects of various covariate changes on each outcome. Specifically, we consider the effects of (a) attaining a BA degree of both parents (as opposed to both being high school graduates only), (b) increasing family income by $10,000 (which corresponds almost exactly to a one-standard-deviation increase in family income), (c) increasing the baseline achievement scores by one standard deviation, (d) having two additional siblings, and (e) being female.

Table 7 shows the posterior mean and posterior standard deviation of the impacts of such covariate changes on each of the three outcomes of interest. In reading the table, recognize, for example, that the "schooling" column summarizes the direct impact of the covariate change on educational attainment (as read directly from Table 6) *plus* any indirect effect that such a change may also have on returns to education and, consequently, schooling levels. The monthly earnings figure in the first column of the table offers a complete summary of how the given covariate change filters through all channels and affects earnings. To evaluate all of these effects, we generate draws from the posterior predictive distribution, as described in (28), with each exercise requiring appropriate definitions of the covariates $\boldsymbol{x_f}$.

The rightmost column of Table 7, which describes the effect of the stated change on returns to schooling, can simply be read from Table 6. Again, these results reveal that females have a much higher return to education (approximately 2.4 percentage points

Table 7
Posterior means (and standard deviations) from posterior predictive exercises.

	Monthly earnings ($)	Schooling	Return to education
Both parents have a BA	103.8	0.727	−0.007
	(103.5)	(0.031)	(0.004)
Increase family income $10,000	81.7	0.119	−0.009
	(47.1)	(0.020)	(0.002)
One-standard-deviation increase in test scores	160.4	0.740	0.001
	(124.3)	(0.020)	(0.002)
Adding two siblings	−41.5	−0.135	0.002
	(28.13)	(0.023)	(0.003)
Female	−360.2	0.245	0.024
	(171.2)	(0.047)	(0.004)

higher) than men. Student achievement scores do not appear to play a role in explaining variation in returns to schooling, while parental education and family income lower returns to education. Of these, family income plays the largest, though still reasonably minor, role, as a one-standard-deviation increase in family income lowers returns to schooling by less than 1 percentage point on average.

The second column summarizes the impacts of the considered covariate changes on the quantity of schooling attained. A one-standard-deviation increase in student achievement scores, or attaining a four-year degree of both parents, produces a large change in the quantity of education attained, as both effects are found to increase educational attainment by approximately 3/4 of a year. Furthermore, these effects are estimated rather precisely, as the posterior standard deviation of the schooling impact in either case is very small relative to the mean. For these two particular exercises, the schooling increase primarily arises from the "direct effect", as revealed in Table 6, as increases in parental education and test scores were not strongly linked to private rates of return to education.

In contrast to this, Table 7 also shows that females acquire more schooling on average than males. This result appears, at first glance, at odds with Table 5, as the coefficient on the female indicator in the schooling equation is actually negative, though with considerable mass placed on both sides of zero. What Table 5 does not directly summarize, however, is the fact that females have much higher returns to education, and given that $\rho > 0$, tend to acquire more schooling on average as a result. Higher levels of educational attainment for women is also a feature of our data: women receive, on average, 0.33 years more education than men in our HSB sample. Our model is able to reproduce this feature of the data, as it predicts women to receive approximately 0.25 years more education than men, with the observed outcome of 0.33 falling within two posterior standard deviations of this point estimate. In our view, this result is quite interesting and, perhaps, new to the literature: once the variation in rates of return to education has been accounted for, there is no discernable difference in the predicted quantity of schooling attained by men and women. However, women attain more schooling, on average, then men because of a comparably high rate of return on such an investment.

The first column of Table 7 aggregates all of these channels and provides overall estimates of the various impacts on monthly earnings. In terms of posterior means, females earn approximately $360 less per month than men, even though they tend to acquire more schooling, on average. A one-standard-deviation increase in student achievement scores increases monthly earnings by about $160 on average, with the bulk of this increase explained by increased levels of educational attainment for those of higher ability. Similarly, graduation from college of both parents increases the monthly (child) earnings by about $104, which, again, results primarily from higher educational attainment by such children. The impacts of family income and number of siblings also operate in the directions we might expect, although the magnitude of these changes is smaller: a one-standard-deviation increase in parental income is associated with an average increase in (child) monthly earnings equal to $81.2, while the addition of two siblings tends to lower monthly earnings by about $42. At this stage of the model, there are also reasonably large amounts of uncertainty surrounding these mean impacts, as their estimation involves an aggregation of effects at each level of the system.

In Fig. 2, we again use our posterior simulations to characterize the differences in rates of return to education, educational attainment, and monthly earnings, and this time, calculate such quantities for a representative white male and a representative black female. When performing these calculations, we fix the covariates at group-specific sample averages rather than restricting the covariate vectors to be equal for both groups.

As shown in the leftmost columns of Fig. 2, the posterior distribution of the return to education parameter for black females is shifted to the right relative to that of a white male, with the posterior mean of the former being 0.093 and the latter approximately 0.05. Returns to education are, however, rather variable and difficult to completely characterize through observables, as summarized by the calculation: $\Pr(\theta_{bf} > \theta_{wm}|\boldsymbol{y}, \boldsymbol{s}) \approx 0.73$.[31]

Unlike the return to schooling distributions, the educational attainment posterior predictive distributions are quite similar for both groups. Higher rates of return for black females lead them to acquire more education than white males, although this increase is offset by the fact that black females tend to come from less educated, larger and less wealthy families on average than those of white males, and black females also have lower average test scores in our data. These offsetting effects culminate in very similar predictive distributions for educational attainment for both groups.

The rightmost columns of Fig. 2 plot the posterior predictive monthly earnings distributions for both representative individuals. The expected monthly earnings of a white male were approximately $2300 and those of a black female were approximately $1700.[32] White males are far less likely to be characterized as low income as, for example, $\Pr(\text{MonthEarn}_{\text{wm}} > \$1100|\boldsymbol{y}, \boldsymbol{s}) \approx 0.95$, while $\Pr(\text{MonthEarn}_{\text{bf}} > \$1100|\boldsymbol{y}, \boldsymbol{s}) \approx 0.80$.[33] Taken together, these calculations illustrate how complete outcome summaries can be obtained within the framework of our model while still identifying the separate individual channels through which particular covariates, or changes in them, filter to affect earnings.

[31] Here, the subscript "bf" refers to black female while the subscript "wm" denotes white male.

[32] Although these numbers might seem small, keep in mind that these are 1991 outcomes for a sample of young workers whose average age (and standard deviation) is 26.8 (0.54).

[33] The choice of $1100 as a threshold is simply to fix ideas, yet is partially guided by policy. The 1991 HHS poverty threshold for a family of four, for example, was $13,400, motivating the monthly figure of $1100 as a choice with some interest.

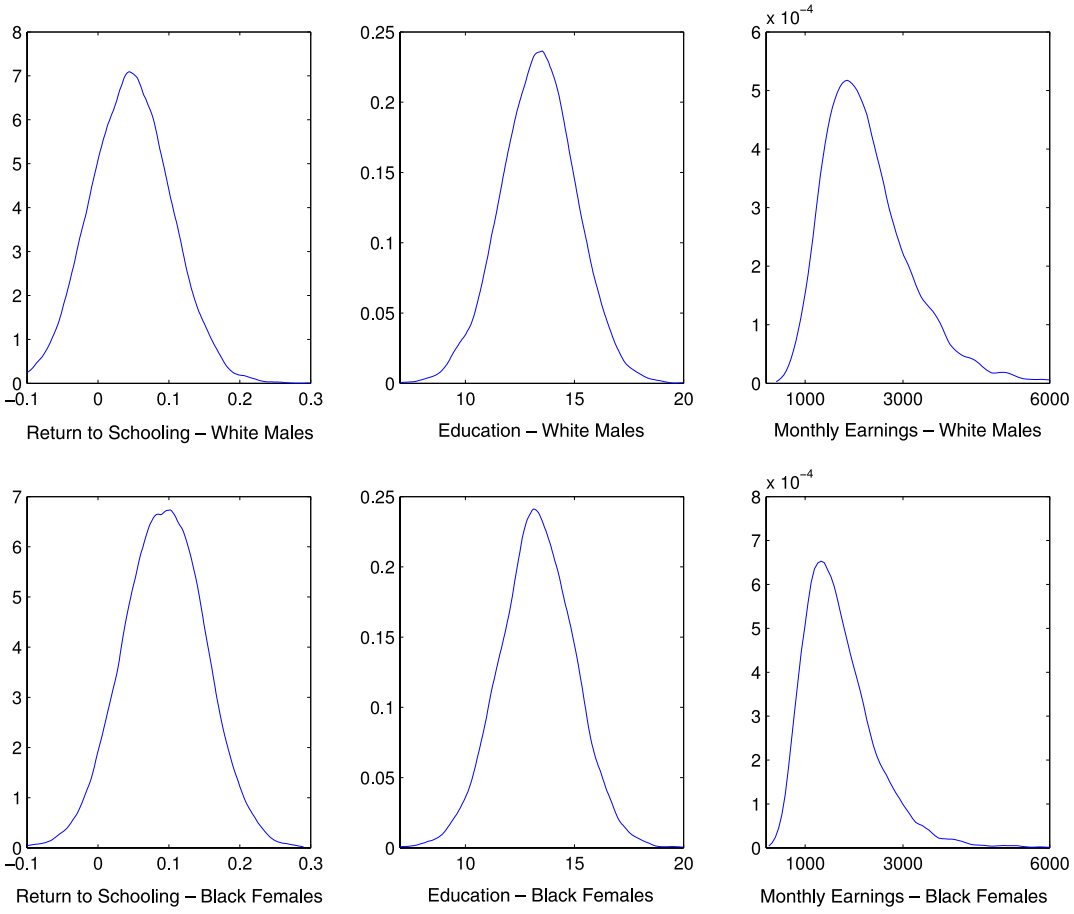

Fig. 2. Posterior predictive outcome distributions for white males (top row) and black females (bottom row).

We conclude this discussion of our findings by offering a few comments regarding how our results fit within the context of the rather vast literature on this topic. To this end, we consider as a benchmark what we term the "standard IV model". This model is obtained by imposing homogeneity in causal impacts (i.e., setting $\theta_i = \theta$ in (1), and thus Eq. (3) becomes irrelevant) and dropping the term $\theta_i\rho$ from (2). The resulting two-equation system is then fit using standard MCMC methods. When doing so, we obtain an estimate (posterior mean) of the common schooling effect equal to 0.059. This result is not terribly out of line with the findings of previous IV-based studies, although the majority of such studies tend to report larger impacts.[34]

In Table 8, we also report estimates of the overall average return to education as well as estimates of this effect that are broken down by racial and gender groups using our two-component version of the correlated random coefficient model. On the whole, we can see that our estimate of the average causal effect and that from the homogeneous effect standard IV model are rather similar, differing by about 0.7 percentage point, and with a fair degree of overlap between the marginal posterior distributions of these quantities. Point estimates of the average return to education for most racial and gender groups exceed the common effect IV estimate, while the point estimate of returns to education for white males is lower than the IV estimate. Despite the similarity of results for the average causal effects from both models, it remains important to note that we should not expect these two estimates to converge to the same parameter; the standard IV procedure will not consistently estimate the average causal effect in the population in general when treatment effect heterogeneity, as described by our model, is present.[35] Our analysis can, however, recover this causal effect and much more, including characterizing the distribution of heterogenous returns, describing if and to what extent agents act upon knowledge of their private returns, and clarifying the various channels through which covariates influence the outcomes of interest.

Table 8
Posterior estimates of average return to education across groups and models.

Predictive return to schooling	$E(\cdot\|D)$	$Std(\cdot\|D)$	$Pr(\cdot > 0\|D)$
Standard IV model			
θ	0.0591	0.0148	1
CRC model			
$\bar{x}\eta + \bar{w}\lambda + \sum_{g=1}^{2} \bar{\pi}_g \bar{\eta}_{0g}$	0.0663	0.0133	1
White male	0.0511	0.0136	1
White female	0.0753	0.0136	1
Black male	0.0686	0.0150	1
Black female	0.0927	0.0149	1

[34] Card (2001), for example, provides a review of a number of influential IV studies on returns to education. Most of these report IV estimates substantially exceeding their OLS estimates, and often in excess of 10%. Our homogenous "IV-type" estimate tends to be closer to the consensus OLS estimate of these studies rather than their IV counterparts. As a partial explanation for this difference, it is important to recognize that we focus on a sample of young, reasonably well-educated workers in the HSB data for whom returns to education are likely to be smaller, on average.

[35] Further details on this issue are available on request, although most of these will repeat the arguments of Wooldridge (2003), who establishes conditions under which a properly implemented IV procedure will consistently recover the average causal effect.

7. Conclusion

In this paper, we have taken up the issue of Bayesian estimation of a correlated random coefficients model. In the past, estimation in these types of models has focused almost exclusively on the estimation of the average causal effect in the population. Our model, though decidedly more parameterized than these previous studies, enables the estimation of far more parameters of interest, including the variability and other features of the causal effect distribution in addition to learning how individuals make treatment decisions on the basis of their gain from receipt of that treatment.

We applied our method in practice to a widely studied problem in labor economics: estimation of the private return to education. Using data combined from High School and Beyond and the 1980 Census, we find evidence of heterogeneity in returns to education. Specifically, we find that some characteristics of agents typically associated with higher levels of schooling (such as family income) are, at the same time, associated with lower returns to schooling. This finding supports the idea that those who benefit most from education are not necessarily the ones who are observed to acquire the most education. In addition, individuals can be viewed to make their schooling decisions based, at least in part, on their return to education. Specifically, a one percentage point increase in the return to education is associated with an increase in schooling quantity equal to approximately 0.135 years.

Appendix. Identification

To fix ideas, we focus on one observation's contribution to the likelihood, denoted as $p(y, s|\boldsymbol{\Gamma}_{-\boldsymbol{\theta}})$, where the subscript i is dropped for simplicity and $\boldsymbol{\Gamma}_{-\boldsymbol{\theta}}$ denotes all parameters other than the return θ, which is to be integrated out of (1)–(3). In this regard, we first note that the marginal density $s|\boldsymbol{\Gamma}_{-\boldsymbol{\theta}}$ is obtained as

$$s = [\delta_0 + \rho\eta_0] + \boldsymbol{x}(\boldsymbol{\delta} + \rho\boldsymbol{\eta}) + \boldsymbol{z}\boldsymbol{\gamma} + \boldsymbol{w}\rho\boldsymbol{\lambda} + \tilde{u}_s, \tag{29}$$

where $\tilde{u}_s \equiv \rho\epsilon + v$.

We now seek to derive the conditional density $p(y|s, \boldsymbol{\Gamma}_{-\boldsymbol{\theta}})$. To this end, we note that

$$p(y|s, \boldsymbol{\Gamma}_{-\boldsymbol{\theta}}) = \int_{-\infty}^{\infty} p(y, \theta|s, \boldsymbol{\Gamma}_{-\boldsymbol{\theta}})\mathrm{d}\theta \tag{30}$$

$$= \int_{-\infty}^{\infty} p(y|\boldsymbol{\Gamma}, s)p(\theta|\boldsymbol{\Gamma}_{-\boldsymbol{\theta}}, s)\mathrm{d}\theta, \tag{31}$$

where $\boldsymbol{\Gamma}$ denotes all parameters in the model. The assumptions of (1)–(3) imply that

$$y|\boldsymbol{\Gamma}, s \sim N(\beta_0 + \boldsymbol{x}\boldsymbol{\beta} + s\theta + r_1[s - \delta_0 - \boldsymbol{x}\boldsymbol{\delta} - \boldsymbol{z}\boldsymbol{\gamma} - \theta\rho] + r_2[\theta - \eta_0 - \boldsymbol{x}\boldsymbol{\eta} - \boldsymbol{w}\boldsymbol{\lambda}], V_y), \tag{32}$$

with

$$V_y \equiv \sigma_y^2 - [\sigma_{ys} \quad \sigma_{y\theta}]\begin{bmatrix}\sigma_s^2 & \sigma_{s\theta}\\ \sigma_{s\theta} & \sigma_\theta^2\end{bmatrix}^{-1}\begin{bmatrix}\sigma_{ys}\\ \sigma_{y\theta}\end{bmatrix}$$

and

$$r_1 \equiv \frac{\sigma_\theta^2\sigma_{ys} - \sigma_{y\theta}\sigma_{s\theta}}{\sigma_\theta^2\sigma_s^2 - \sigma_{s\theta}^2}, \qquad r_2 \equiv \frac{\sigma_s^2\sigma_{y\theta} - \sigma_{s\theta}\sigma_{ys}}{\sigma_\theta^2\sigma_s^2 - \sigma_{s\theta}^2}. \tag{33}$$

The integration in (31) also requires $p(\theta|\boldsymbol{\Gamma}_{-\boldsymbol{\theta}}, s)$. Given the bivariate normality of $(s, \theta|\boldsymbol{\Gamma}_{-\boldsymbol{\theta}})$ implied from (2) and (3), we obtain

$$\theta|\boldsymbol{\Gamma}_{-\boldsymbol{\theta}}, s \sim N\left(\eta_0 + \boldsymbol{x}\boldsymbol{\eta} + \boldsymbol{w}\boldsymbol{\lambda} + \frac{b}{c}(s - \delta_0 - \boldsymbol{x}\boldsymbol{\delta} - \boldsymbol{z}\boldsymbol{\gamma} - \rho[\eta_0 + \boldsymbol{x}\boldsymbol{\eta} + \boldsymbol{w}\boldsymbol{\lambda}]), V_\theta\right), \tag{34}$$

Table 9
Coefficients on terms in $E(y|s, \boldsymbol{\Gamma}_{-\boldsymbol{\theta}})$ with $b \equiv \sigma_{s\theta} + \rho\sigma_\theta^2$ and $c \equiv \sigma_s^2 + 2\rho\sigma_{s\theta} + \rho^2\sigma_\theta^2$.

Variable	Coefficient
Constant	$\beta_0 - c^{-1}(\delta_0 + \rho\eta_0)\left(\sigma_{ys} + \rho\sigma_{y\theta}\right)$
$\boldsymbol{x}$	$\boldsymbol{\beta} - c^{-1}(\boldsymbol{\delta} + \rho\boldsymbol{\eta})\left(\sigma_{ys} + \rho\sigma_{y\theta}\right)$
w	$-\lambda\rho c^{-1}\left(\sigma_{ys} + \rho\sigma_{y\theta}\right)$
s	$\eta_0 + c^{-1}\left(\sigma_{ys} + \rho\sigma_{y\theta}\right) - [b/c](\delta_0 + \rho\eta_0)$
z	$-\gamma c^{-1}\left(\sigma_{ys} + \rho\sigma_{y\theta}\right)$
s^2	bc^{-1}
$\boldsymbol{sx}$	$\boldsymbol{\eta} - bc^{-1}(\boldsymbol{\delta} + \boldsymbol{\eta}\rho)$
sw	$\lambda - bc^{-1}\rho\lambda$
sz	$-\gamma bc^{-1}$

where

$$b \equiv \sigma_{s\theta} + \rho\sigma_\theta^2, \qquad c \equiv \sigma_s^2 + 2\rho\sigma_{s\theta} + \rho^2\sigma_\theta^2 \quad \text{and} \quad V_\theta \equiv \sigma_\theta^2 - (b^2/c). \tag{35}$$

Given (32) and (34), the integration involves completing the square in θ, recognizing a portion of the integrand as the kernel of a Gaussian distribution, and then accounting for all remaining terms. The result of this calculation shows that $y|s, \boldsymbol{\Gamma}_{-\boldsymbol{\theta}}$ is also normal; its regression function contains a constant and additive terms involving $\boldsymbol{x}$, w, s, z, s^2, $\boldsymbol{sx}$, sw, and sz. The parameters multiplying each of these terms in the regression function for $y|s$ are given in Table 9.

Similar algebra also reveals that

$$\mathrm{Var}(y|s, \boldsymbol{\Gamma}_{-\boldsymbol{\theta}}) = \left(\sigma_y^2 - \frac{[\sigma_{ys} + \rho\sigma_{y\theta}]^2}{c}\right) + 2s_i\left(\sigma_{y\theta} - \frac{(\sigma_{ys} + \rho\sigma_{y\theta})b}{c}\right) + s_i^2\left(\sigma_\theta^2 - \frac{b^2}{c}\right). \tag{36}$$

Some quick accounting suggests that, without any further restrictions placed on the model, there are 15 sets of unknowns and 17 sets of equations from $s|\boldsymbol{\Gamma}_{-\boldsymbol{\theta}}$ and $y|s, \boldsymbol{\Gamma}_{-\boldsymbol{\theta}}$ that can be used to recover these "structural" parameters. Henceforth, we restrict ourselves to establishing identification in the more difficult (and perhaps more realistic) case where $\boldsymbol{\gamma} = \mathbf{0}$, $\rho \neq 0$ and w is a scalar. In other words, we have one exclusion restriction in (3), no exclusion restrictions in (2), and make the assumption that $\rho \neq 0$. In this case, we have one less set of parameters ($\boldsymbol{\gamma}$) to estimate, but setting $\boldsymbol{\gamma} = \mathbf{0}$ eliminates three of our estimating equations. Under these restrictions, the parameter vector can be broken down into eight sets of regression parameters

$$[\beta_0 \quad \boldsymbol{\beta} \quad \delta_o \quad \boldsymbol{\delta} \quad \rho \quad \eta_0 \quad \boldsymbol{\eta} \quad \lambda]$$

and six parameters of the covariance matrix:

$$[\sigma_y^2 \quad \sigma_s^2 \quad \sigma_\theta^2 \quad \sigma_{ys} \quad \sigma_{y\theta} \quad \sigma_{s\theta}].$$

In terms of equations that can be used to identify the above values, let $\boldsymbol{a}_{\boldsymbol{r}}^{\boldsymbol{j}}$ denote the (estimable) coefficient on variable $\boldsymbol{r}$ in equation $j, j \in \{s, y\}$ and $\boldsymbol{r} \in \{co, \boldsymbol{x}, w, s, s^2, \boldsymbol{sx}, sw\}$. It is understood that $j = s$ refers to $s|\boldsymbol{\Gamma}_{-\boldsymbol{\theta}}$ in Eq. (29), $j = y$ refers to the equation for $y|s, \boldsymbol{\Gamma}_{-\boldsymbol{\theta}}$ in Table 5, and $\boldsymbol{r} = co$ denotes the constant term in each equation. From the s marginal density, we estimate three coefficients and a variance parameter which, in the above notation, provides

$$[a_{co}^s \quad \boldsymbol{a}_{\boldsymbol{x}}^{\boldsymbol{s}} \quad a_w^s \quad \hat{c}].$$

Similarly, the $y|s, \boldsymbol{\Gamma}_{-\boldsymbol{\theta}}$ equation gives

$$[a_{co}^y \quad \boldsymbol{a}_{\boldsymbol{x}}^{\boldsymbol{y}} \quad a_w^y \quad a_s^y \quad a_{s^2}^y \quad \boldsymbol{a}_{\boldsymbol{sx}}^{\boldsymbol{y}} \quad a_{sw}^y \quad V_{co}^y \quad V_s^y \quad V_{s^2}^y],$$

where, for the final three terms, V_r^y denotes the (estimable) coefficient multiplying the variable r in the expression for $\mathrm{Var}(y|s, \boldsymbol{\Gamma}_{-\theta})$ in (36). Thus, we have 14 sets of equations to use in recovering the 14 sets of structural parameters.

Note that

$$a_{co}^s = \widehat{\delta_0 + \rho\eta_0} \tag{37}$$

$$\boldsymbol{a_x^s} = \widehat{\boldsymbol{\delta} + \rho\boldsymbol{\eta}} \tag{38}$$

$$a_w^s = \widehat{\rho\lambda}. \tag{39}$$

The relationships in Table 5 can be stacked together to produce

$$\begin{bmatrix} 1 & \mathbf{0} & -a_{co}^s\hat{c}^{-1} & 0 & 0 & \mathbf{0} & 0 \\ \mathbf{0} & \boldsymbol{I} & -\boldsymbol{a_x^s}\hat{c}^{-1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ 0 & \mathbf{0} & -a_w^s\hat{c}^{-1} & 0 & 0 & \mathbf{0} & 0 \\ 0 & \mathbf{0} & \hat{c}^{-1} & 1 & -a_{co}^s\hat{c}^{-1} & \mathbf{0} & 0 \\ 0 & \mathbf{0} & 0 & 0 & \hat{c}^{-1} & \mathbf{0} & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & -\boldsymbol{a_x^s}\hat{c}^{-1} & \boldsymbol{I} & \mathbf{0} \\ 0 & \mathbf{0} & 0 & 0 & -a_w^s\hat{c}^{-1} & \mathbf{0} & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}} \\ \widehat{\sigma_{ys} + \rho\sigma_{y\theta}} \\ \hat{\eta}_0 \\ \hat{b} \\ \hat{\boldsymbol{\eta}} \\ \hat{\lambda} \end{bmatrix}$$

$$= \begin{bmatrix} a_{co}^y \\ \boldsymbol{a_x^y} \\ a_w^y \\ a_s^y \\ a_{s^2}^y \\ \boldsymbol{a_{sx}^y} \\ a_{sw}^y \end{bmatrix}$$

or, succinctly,

$$\boldsymbol{H_y\Gamma_y} = \boldsymbol{a_y},$$

where $\boldsymbol{I}$ denotes an identity matrix with an appropriate size. The matrix $\boldsymbol{H_y}$ is full rank; hence, the terms in $\boldsymbol{\Gamma_y}$ are identified and could be estimated as $\boldsymbol{\Gamma_y} = \boldsymbol{H_y^{-1}a_y}$. The remaining parameters ρ, $\boldsymbol{\delta}$, and δ_0 can then be obtained from (37)–(39) as

$$\hat{\rho} = a_w^s\hat{\lambda}^{-1} \tag{40}$$

$$\hat{\boldsymbol{\delta}} = \boldsymbol{a_x^s} - \hat{\rho}\hat{\boldsymbol{\eta}} \tag{41}$$

$$\hat{\delta_0} = a_{co}^s - \hat{\rho}\hat{\eta}_0. \tag{42}$$

It remains to discuss the parameters of $\boldsymbol{\Sigma}$. Note that $\hat{c}$, $\hat{b}$, $[\widehat{\sigma_{ys} + \rho\sigma_{y\theta}}]$, and (36) can be employed to recover their values. Specifically,

$$\begin{bmatrix} \rho & 0 & 0 & 0 & 1 & 0 \\ \rho^2 & 1 & 0 & 0 & 2\rho & 0 \\ 0 & 0 & 0 & \rho & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \sigma_\theta^2 \\ \sigma_s^2 \\ \sigma_y^2 \\ \sigma_{y\theta} \\ \sigma_{s\theta} \\ \sigma_{ys} \end{bmatrix}$$

$$= \begin{bmatrix} \hat{b} \\ \hat{c} \\ [\widehat{\sigma_{ys} + \rho\sigma_{y\theta}}] \\ V_{co}^y + [\widehat{\sigma_{ys} + \rho\sigma_{y\theta}}]^2\hat{c}^{-1} \\ V_s^y/2 + [\widehat{\sigma_{ys} + \rho\sigma_{y\theta}}]\hat{b}\hat{c}^{-1} \\ V_{s^2}^y + \hat{b}^2\hat{c}^{-1} \end{bmatrix},$$

or, succinctly,

$$\boldsymbol{H_\sigma\Gamma_\sigma} = \boldsymbol{V_\sigma}.$$

The matrix $\boldsymbol{H_\sigma}$ is, again, full rank; hence, the parameters of the covariance matrix are identifiable and could be estimated as $\boldsymbol{\Gamma_\sigma} = \boldsymbol{H_\sigma^{-1}V_\sigma}$.

References

Angrist, J.D., Krueger, A.B., 1991. Does compulsory school attendance affect schooling and earnings? Quarterly Journal of Economics 106, 979–1014.

Björklund, A., Moffitt, R., 1987. The estimation of wage gains and welfare gains in self-selection models. Review of Economics and Statistics 69 (1), 42–49.

Card, D., 2001. Estimating the return to schooling: progress on some persistent econometric problems. Econometrica 69, 1127–1160.

Chib, S., 2007. Analysis of treatment response data without the joint distribution of potential outcomes. Journal of Econometrics 140 (2), 401–412.

Chib, S., Hamilton, B.H., 2000. Bayesian analysis of cross-section and clustered data treatment models. Journal of Econometrics 97 (1), 25–50.

Chib, S., Hamilton, B.H., 2002. Semiparametric Bayes analysis of longitudinal data treatment models. Journal of Econometrics 110, 67–89.

Conley, T., Hansen, C., McCulloch, R., Rossi, P.E., 2008. A semi-parametric Bayesian approach to the instrumental variable problem. Journal of Econometrics 144 (1), 276–305.

Deb, P., Munkin, M.K., Trivedi, P., 2006. Bayesian analysis of the two-part model with endogeneity: application to health care expenditure. Journal of Applied Econometrics 21 (7), 1081–1099.

Frühwirth-Schnatter, S., 2001. Markov chain Monte Carlo estimation of classical and dynamic switching mixture models. Journal of the American Statistical Association 96 (453), 194–209.

Ganderton, P.T., 1992. The effect of subsidies in kind on the choice of college. Journal of Public Economics 48, 269–292.

Gelfand, A.E., Dey, D.K., 1994. Bayesian model choice: asymptotics and exact calculations. Journal of the Royal Statistical Society. Series B 501–514.

Geweke, J., 1999. Using simulation methods for Bayesian econometric models: inference, development and communication. Econometric Reviews 18 (1), 1–73.

Geweke, J., 2004. Getting it right: joint distribution tests of posterior simulators. Journal of the American Statistical Association 99, 799–804.

Geweke, J., 2007. Interpretation and inference in mixture models: simple MCMC works. Computational Statistics and Data Analysis 51, 3529–3550.

Geweke, J., Keane, M., 2007. Smoothly mixing regressions. Journal of Econometrics 138, 252–290.

Grogger, J., 1996a. Does school quality explain the recent black/white wage trend? Journal of Labor Economics 14 (2), 231–253.

Grogger, J., 1996b. School expenditures and post-schooling earnings: evidence from high school and beyond. Review of Economics and Statistics 78 (4), 628–637.

Hanushek, E.A., Taylor, L.L., 1990. Alternative assessments of the performance of schools: measurement of state variation in achievement. Journal of Human Resources 25 (2), 179–201.

Heckman, J.J., 1997. Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations. Journal of Human Resources 32 (3), 441–462.

Heckman, J.J., Smith, J., 1999. Evaluating the welfare state. In: Strom, S. (Ed.), Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial Symposium. In: Econometric Society Monographs, Cambridge University Press, Cambridge.

Heckman, J.J., Vytlacil, E., 1998. Instrumental variables methods for the correlated random coefficient model: estimating the average rate of return to schooling when the return is correlated with schooling. Journal of Human Resources 33, 974–987.

Heckman, J.J., Vytlacil, E., 1999. Local instrumental variables and latent variable models for identifying and bounding treatment effects. Proceedings of the National Academy of Sciences 96, 4730–4734.

Heckman, J.J., Vytlacil, E., 2005. Structural equations, treatment effects and econometric policy evaluation. Econometrica 73 (3), 669–738.

Henderson, D.J., Polachek, S.W., Wang, L., 2009. Heterogeneity in schooling rates of return. Working Paper. Department of Economics, SUNY-Binghamton.

Hoogerheide, L., Kleibergen, F., van Dijk, H.K., 2007. Natural conjugate priors for the instrumental variables regression model applied to the Angrist–Krueger data. Journal of Econometrics 138 (1), 63–103.

Imbens, G., Angrist, J., 1994. Identification and estimation of local average treatment effects. Econometrica 62 (2), 467–475.

Kleibergen, F.R., Zivot, E., 2003. Bayesian and classical approaches to instrumental variable regression. Journal of Econometrics 114, 29–72.

Koop, G., Poirier, D.J., 1997. Learning about the across-regime correlation in switching regression models. Journal of Econometrics 78, 217–227.

Koop, G., Poirier, D.J., Tobias, J.L., 2007. Bayesian Econometric Methods. Cambridge University Press.

Koop, G., Tobias, J.L., 2004. Learning about heterogeneity in returns to schooling. Journal of Applied Econometrics 19 (7), 827–849.

Lancaster, T., 2004. An Introduction to Modern Bayesian Econometrics. Blackwell.

Li, K., 1998. Bayesian inference in a simultaneous equation model with limited dependent variables. Journal of Econometrics 85 (2), 387–400.

Li, M., 2006. High school completion and future youth unemployment: new evidence from high school and beyond. Journal of Applied Econometrics 21, 23–53.

Li, M., 2007. Bayesian proportional hazard analysis of the timing of high school dropout decisions. Econometric Reviews 26, 529–556.

Li, M., Poirier, D.J., Tobias, J.L., 2003. Do dropouts suffer from dropping out? Estimation and prediction of outcome gains in generalized selection models. Journal of Applied Econometrics 19 (2), 203–225.

Manchanda, P., Rossi, P.E., Chintagunta, P.K., 2004. Response modeling with non-random marketing-mix variables. Journal of Marketing Research XLI, 467–478.
Munkin, M., Trivedi, P., 2003. Bayesian analysis of a self-selection model with multiple outcomes using simulation-based estimation: an application to the demand for healthcare. Journal of Econometrics 114, 197–220.
Pearson, K., 1894. Contributions to the mathematical theory of evolution. Philosophical Transactions of the Royal Society of London, Series A 185, 71–110.
Poirier, D.J., Tobias, J.L., 2003. On the predictive distributions of outcome gains in the presence of an unidentified parameter. Journal of Business and Economic Statistics 21 (2), 258–268.
Rivkin, S.G., 1991. Schooling and employment in the 1980's: who succeeds? Ph.D. Dissertation. UCLA, Department of Economics.
Rossi, P.E., Allenby, G.M., McCulloch, R., 2005. Bayesian Statistics and Marketing. Wiley.
Roy, A.D., 1951. Some thoughts on the distribution of earnings. Oxford Economic Papers. New Series, vol. 3, pp. 135–146.
Wooldridge, J.M., 1997. On two stage least squares estimation of the average treatment effect in a random coefficient model. Economics Letters 56, 129–133.
Wooldridge, J.M., 2003. Further results on instrumental variables estimation of average treatment effects in the correlated random coefficient model. Economics Letters 79, 185–191.