ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2020)

A Bayesian multivariate factor analysis model for evaluating an intervention by using observational time series data on multiple outcomes

Pantelis Samartsidis and Shaun R. Seaman
University of Cambridge, UK

Silvia Montagna,
University of Turin, Italy

André Charlett,
Public Health England, London, UK

Matthew Hickman
University of Bristol, UK

and Daniela De Angelis
University of Cambridge, UK

[Received October 2018. Revised February 2020]

Summary. A problem that is frequently encountered in many areas of scientific research is that of estimating the effect of a non-randomized binary intervention on an outcome of interest by using time series data on units that received the intervention ('treated') and units that did not ('controls'). One popular estimation method in this setting is based on the factor analysis (FA) model. The FA model is fitted to the preintervention outcome data on treated units and all the outcome data on control units, and the counterfactual treatment-free post-intervention outcomes of the former are predicted from the fitted model. Intervention effects are estimated as the observed outcomes minus these predicted counterfactual outcomes. We propose a model that extends the FA model for estimating intervention effects by jointly modelling the multiple outcomes to exploit shared variability, and assuming an auto-regressive structure on factors to account for temporal correlations in the outcome. Using simulation studies, we show that the method proposed can improve the precision of the intervention effect estimates and achieve better control of the type I error rate (compared with the FA model), especially when either the number of preintervention measurements or the number of control units is small. We apply our method to estimate the effect of stricter alcohol licensing policies on alcohol-related harms.

Keywords: Causal inference; Factor analysis; Intervention evaluation; Panel data

1. Introduction

In this work, we consider the problem of estimating the causal effect of an intervention on an outcome of interest in the setting where

Address for correspondence: Pantelis Samartsidis, Medical Research Council Biostatistics Unit, University of Cambridge, University Forvie Site, Robinson Way, Cambridge, CB2 0SR, UK.
E-mail: pantelis.samartsidis@mrc-bsu.cam.ac.uk

(a) the intervention is binary,
(b) assignment of the sample units to the intervention is non-randomized,
(c) only a small number of units are treated and
(d) there are multiple measurements of the outcome both before and after the intervention occurs.

This problem is frequently encountered in various fields of scientific research, including econometrics, epidemiology, marketing, public health and political science. For example, Card (1990) studied the effect that the mass migration in 1980 of Cubans to Miami had on Miami's labour market, by treating Miami as having received the 'intervention' of mass Cuban migration and comparing it with other US states that were not subject to such migrations; Cavallo *et al.* (2013) assessed the effect that large-scale natural disasters, such as earthquakes and storms, had on the gross domestic product of a country by comparing countries that are subject to such natural disasters (the 'intervention') with countries not experiencing such disasters; de Vocht (2016) investigated whether the increased use of mobile phones (the 'intervention') led to an increase in incidences of certain types of brain cancer in England.

A general difficulty when estimating the causal effect of an intervention from observational data is the potential existence of confounding variables. These are variables which affect both the outcome of interest and the probability of being assigned to the intervention. Failure to account for confounding can lead to biased estimation of causal effects. When the number of units receiving the intervention is large, propensity score methods (Robins *et al.*, 2000) can be used. However, when few units receive the intervention, there is not enough information to fit propensity score models. For this reason, several new methodologies for causal inference in the setting where conditions (a)–(d) apply have recently been proposed. For a recent review, see Samartsidis *et al.* (2019).

Many of these methods, including those of Abadie *et al.* (2010), Hsiao *et al.* (2012), Gobillon and Magnac (2016), Chan and Kwok (2016) and Xu (2017), build on the *factor analysis* (FA) model. FA is a natural way to adjust for confounding. The model allows for unobserved confounders that remain constant over time but have a time-varying effect on the outcome. However, current methodologies based on the FA model have shortcomings. Firstly, they can be applied to only a single outcome at a time. When there is more than one correlated outcome, it might be more efficient to model them jointly. Secondly, none of the aforementioned methods explicitly models the temporal correlation between the multiple measurements of the outcome. Modelling auto-correlation may improve efficiency. Thirdly, when the total number of units is small, it is difficult to perform inference for the causal effects by using the existing methods. Finally, some of the approaches above require specifying the number of factors in the model. Although guidance is provided for how to use the data to choose this number, inference using these methods does not account for this data-dependent choice and hence tends to be anticonservative.

In this paper, we attempt to address these shortcomings. We consider extensions of the FA model that can exploit the correlation between different outcomes and the temporal correlation within each outcome, leading to more efficient estimates. Also, by taking a Bayesian approach, we can obtain credible intervals for the causal effects that account for the uncertainty in the number of factors. We contribute to the literature on causal inference in the setting where conditions (a)–(d) apply, in three ways. Firstly, we develop a novel approach that uses multivariate outcomes. An alternative multivariate model was suggested by Robbins *et al.* (2017). However, their method is designed for high dimensional data and its utility in a small data setting is unclear. Secondly, our method is one of the few that model temporal correlation within each

outcome. Brodersen *et al.* (2015) recently proposed the *causal impact* method to account for such correlation. However, causal impact can be applied to only a single treated unit and single outcome at a time. Thirdly, our use of the Bayesian approach enables more inference on the causal effects of interest in comparison with other FA-based approaches.

Our method has connections with various applications of FA in contexts other than causal inference. More specifically, this is not the first time that a multivariate factor model has been used in practice. De Vito *et al.* (2018a, b) and Avalos-Pacheco *et al.* (2018) demonstrated the benefits of taking a multivariate approach in genomic applications when dealing with multiple studies rather than multiple outcomes. Assuming a temporal structure in the FA model is also not uncommon; see, for example, McAlinn *et al.* (2019) for a recent application in macroeconomics. However, to our knowledge, this is the first time that either of these extensions (joint outcome modelling and explicit modelling of the temporal correlation) to the standard FA model has been implemented in a causal inference problem and the benefits of using them demonstrated. Our methodology, together with other causal inference methodologies based on FA, is related to the *latent class analysis* (LCA) causal inference approach (Lanza *et al.*, 2013; Bartolucci *et al.*, 2016; Tullio and Bartolucci, 2019). In LCA, there is a fixed number of classes (the analogue of factors in FA) and the distribution of the outcomes on each unit and at each time point depends on the unobserved class of that unit at that time point. Hence, both LCA and FA attempt to model the variability in the outcome by using latent variables. A difference between LCA and FA is that classes are discrete whereas factors are continuous. Despite being similar in spirit to FA approaches, LCA-based causal inference methods cannot be used in the setting where conditions (a)–(d) apply, mainly because they require estimation of the propensity score, which is problematic when the number of treated units is small. Moreover, these methods focus on the causal effect of the intervention on the probability that an individual belongs to a certain class, whereas in our problem the interest is in the effect of the intervention directly on the outcomes.

The paper is structured as follows. Section 2 introduces our motivating example. Section 3.1 introduces the notation and causal framework. The standard FA model is formulated in Section 3.2. Section 3.3 presents the methodology proposed. Section 3.4 describes how our method accounts for the uncertainty regarding the true number of factors. Prior distributions and posterior sampling are discussed in Section 3.5. Section 3.6 describes how point estimates and inferences are obtained for the causal effects of interest. In Section 4 we perform a series of simulation studies to evaluate the utility of the methodology, using the standard FA model as our benchmark. Section 5 describes the application of methods to our motivating data set. Finally, Section 6 contains a discussion and suggests some possible directions for future research.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from `https://osf.io/4d7c6/`.

2. Motivating example

Alcohol consumption has an adverse effect on society, being responsible for some harmful health conditions and behaviours. National policy makers have long focused on the development of effective strategies to limit these negative effects. For example, the 2003 Licensing Act (`http://www.legislation.gov.uk/ukpga/2003/17/contents`) in England and Wales enables local authorities to develop *cumulative impact policies* (CIPs) i.e. to reject automatically new licensing applications unless these are supported by evidence that granting will not negatively impact on surrounding premises.

In a recent study, de Vocht *et al.* (2017) assessed the effect that CIPs had on alcohol-related

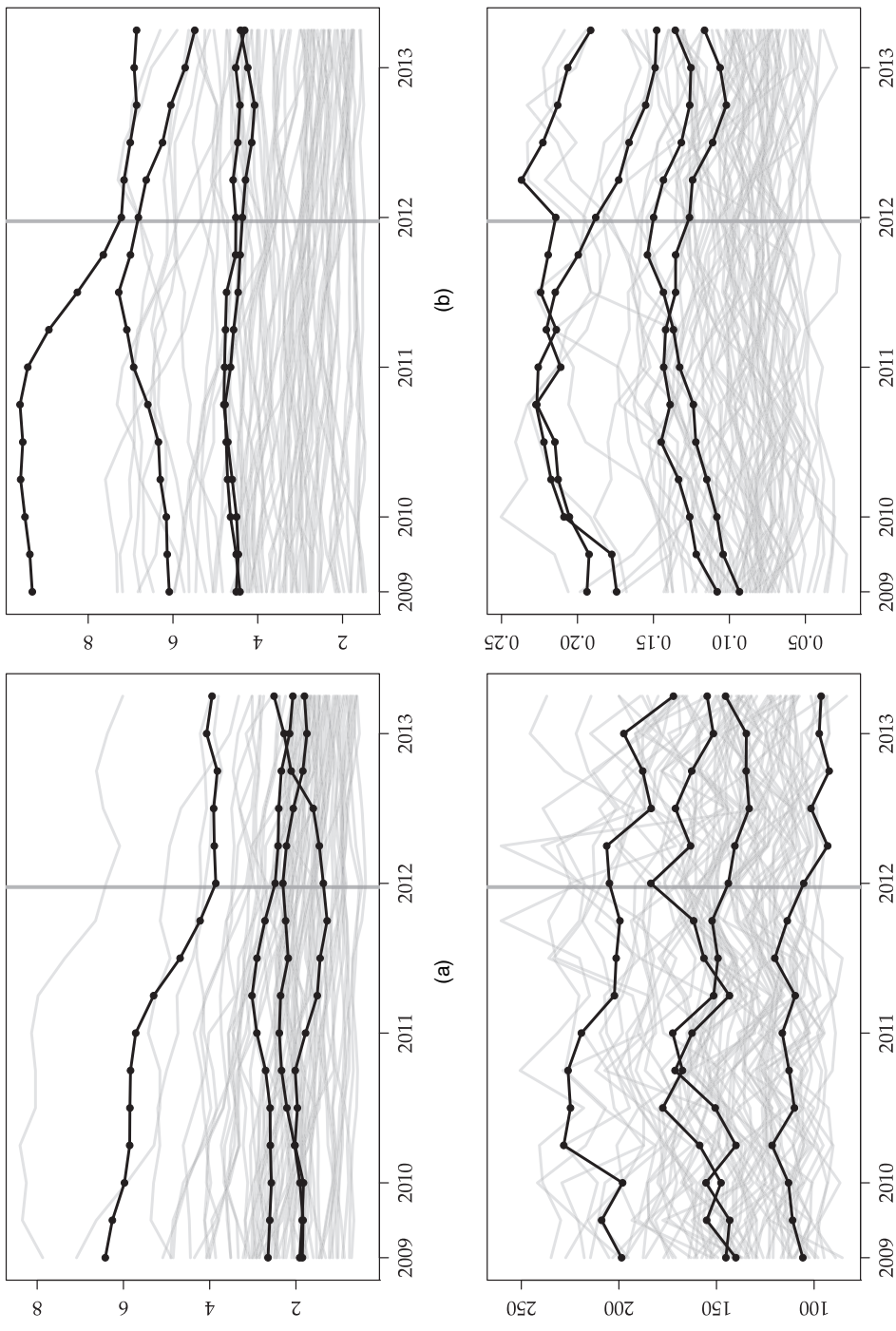

Fig. 1. The alcohol licensing data set: each plot shows the time series for each unit for one of the four outcomes (a) antisocial behaviour, (b) violent crimes, (c) hospital admissions and (d) sexual crimes: ———, controls; ●, treated; |, introduction of the intervention

harms. They collected data on four alcohol-related outcomes: hospital admission rate per 10000 people, violent crimes rate per 1000 people, sexual crime rate per 1000 people and antisocial behaviour incidence rate per 1000 people. The data on each outcome were collected quarterly for the period from mid-2009 to 2015. de Vocht *et al.* (2017) defined intervention sites as local councils implementing a CIP in 2012, and control sites as local councils that did not adopt a CIP at any time during the study period. They identified five treated and 86 control sites in England and Wales.

In Section 5, we demonstrate our proposed methodology by using a subset of data in de Vocht *et al.* (2017). We exclude data from one treated site (Tyneside) because the intervention was implemented earlier in this site and from nine control sites because of missing values in some of the outcomes. Finally, we use data only up to mid-2013, because trends in the following months might be due to changes in the way that crimes were reported (de Vocht *et al.*, 2017). Fig. 1 shows the data.

3. Model specifications

3.1. Notation and causal framework

We have observations y_{itk}, where $i = 1, \ldots, n$ indexes the units, $t = 1, \ldots, T$ indexes the time points and $k = 1, \ldots, K$ indexes the outcomes. The units are ordered so that the first n_1 are the controls, i.e. units that do not receive the intervention during the course of the study. For the remaining $n_2 = n - n_1$ units, there is a time point T_1 after which they all receive the intervention. We refer to these units as the *treated* units. Let r_i be a binary indicator of whether unit i is treated. The study can be split into two periods: the preintervention period consisting of the first T_1 time points when none of the n units has the intervention, and the post-intervention period consisting of the remaining $T_2 = T - T_1$ time points when the intervention is in place for the n_1 treated units.

In this paper, we adopt the *Rubin causal model* (Rubin, 1974; Holland, 1986). This means that for each treated unit i ($i > n_1$), time t after intervention (i.e. $t > T_1$) and outcome k there are two potential outcomes $y_{itk}^{(0)}$ and $y_{itk}^{(1)}$; $y_{itk}^{(0)}$ represents the outcome that would have been observed if the intervention had not been applied and $y_{itk}^{(1)}$ is the outcome that would be observed if the intervention were applied. Hence, the causal effect of the intervention for any $i > n_1$, $t > T_1$ and k is given by

$$\theta_{itk} = y_{itk}^{(1)} - y_{itk}^{(0)}. \tag{1}$$

We are further interested in the average treatment effect on outcome k at time t in the treated units, ϑ_{tk}, defined as

$$\vartheta_{tk} = \frac{1}{n_2} \sum_{i=n_1+1}^{n} \theta_{itk}. \tag{2}$$

For treated units before intervention (i.e. $i > n_1$ and $t \leqslant T_1$) and for control units at all times (i.e. $i \leqslant n_1$ and all t), $y_{itk}^{(0)} = y_{itk}$ for all k, and so is observed. We do not observe $y_{itk}^{(0)}$ for treated units after intervention, and so causal effects (1) and (2) are not observed. Our approach is to assume a model for $y_{itk}^{(0)}$, to use this to obtain predictions $\hat{y}_{itk}^{(0)}$ for the counterfactuals $y_{itk}^{(0)}$ for $i > n_1$ and $t > T_1$, and then to estimate equations (1) and (2) as $\hat{\theta}_{itk} = y_{itk}^{(1)} - \hat{y}_{itk}^{(0)}$ and $\hat{\vartheta}_{tk} = (1/n_2)\Sigma_{i=n_1+1}^{n}$ $\hat{\theta}_{itk}$ respectively.

3.2. Factor analysis model for a single outcome

For time series observational data, a model that is frequently used for $y_{itk}^{(0)}$ is the FA model,

which is also known as the *interactive fixed effects* model (Bai, 2009). Gobillon and Magnac (2016), Chan and Kwok (2016) and Xu (2017) used the FA model for causal inference in the setting that we are investigating, i.e. that where conditions (a)–(d) apply. Abadie *et al.* (2010) and Hsiao *et al.* (2012) showed that their proposed estimators of the counterfactuals $y_{itk}^{(0)}$ ($i > n_1$ and $t > T_1$) are unbiased when the FA model is the data-generating mechanism.

The FA model for the kth outcome assumes that

$$y_{itk}^{(0)} = \boldsymbol{\gamma}_{ik}^{\mathrm{T}} \mathbf{s}_{tk} + \varepsilon_{itk}, \tag{3}$$

where $\boldsymbol{\gamma}_{ik} = (\gamma_{ik1}, \ldots, \gamma_{ikp_1})^{\mathrm{T}}$ is the p_1-vector of unobserved unit-specific loadings for outcome k, $\mathbf{s}_{tk} = (s_{tk1}, \ldots, s_{tkp_1})^{\mathrm{T}} \sim N_{p_1}(\mathbf{0}, \mathbf{I})$ is the p_1-vector of unobserved time-specific factors for outcome k and $\varepsilon_{itk} \sim N(0, \psi_{ik}^2)$ is the error term. One can view the loadings $\boldsymbol{\gamma}_{ik}$ as unit characteristics that remain constant over time and the factors $\mathbf{s}_{tk}$ as their time-varying effect on the potential outcome. Xu (2017) refered to $\mathbf{s}_{tk}$ as 'shocks'; $\boldsymbol{\gamma}_{ik}$ describe the magnitude of the effect that these shocks have on unit i's outcome k. Variables that are predictive of $y_{itk}^{(0)}$ but are not affected by the intervention can be incorporated as covariates $\mathbf{x}_{it}$ in the FA model by replacing equation (3) with

$$y_{itk}^{(0)} = \boldsymbol{\gamma}_{ik}^{\mathrm{T}} \mathbf{s}_{tk} + \boldsymbol{\beta}_k^{\mathrm{T}} \mathbf{x}_{it} + \varepsilon_{itk}. \tag{4}$$

Such variables may include observed confounders measured before time T_1. For simplicity, we shall omit such covariates until Section 3.5.

We note in passing that a special case of the FA model is the *difference-in-differences* model (Angrist and Pischke, 2009; Jones and Rice, 2011). This is the FA model with $p_1 = 2$, $s_{tk1} = 1$ and $\gamma_{ik2} = 1$, i.e. fixed effects for units and time points. The difference-in-differences model assumes that the 'shocks' at each time point affect all the units in the same way. This model is frequently used for causal inference with time series observational data.

Recall that the potential outcome $y_{itk}^{(0)}$ is observed at all times (i.e. $t = 1, \ldots, T$) for control units ($r_i = 0$) but at only the preintervention times (i.e. $t = 1, \ldots, T_1$) for treated units ($r_i = 1$). Model (3) is fitted to these observed data, considering the post-intervention outcomes $y_{itk}^{(0)}$ on treated units as missing. The resulting estimator $\hat{\theta}_{itk}$ of the intervention effect on unit i and outcome k at time t is asymptotically unbiased as $n_1 \to \infty$ and $T_1 \to \infty$, provided that r_i is independent of $\varepsilon_{i1k}, \ldots, \varepsilon_{iTk}$ (and assuming regularity conditions) (Xu, 2017).

There are two intuitive ways to understand why this asymptotic unbiasedness holds. First, as n_1 and T_1 become larger (assuming fixed $n - n_1$ and $T - T_1$), the amount of data for learning about the factors $\mathbf{s}_{tk}$ and the loadings $\boldsymbol{\gamma}_{ik}$ increases, so that the factors and loadings (and hence the expectation of $y_{itk}^{(0)}$) are increasingly accurately estimated. Second, by letting $\mathbf{y}_{tk}^{(0)} = (y_{1tk}^{(0)}, \ldots, y_{ntk}^{(0)})^{\mathrm{T}}$, defining $\boldsymbol{\Gamma}_k$ as the $n \times p_1$ matrix with ith row $\boldsymbol{\gamma}_{ik}$, and letting $\boldsymbol{\epsilon}_{tk} = (\epsilon_{1tk}, \ldots, \epsilon_{Ntk})^{\mathrm{T}}$, equation (3) implies that

$$\mathbf{y}_{tk}^{(0)} = \boldsymbol{\Gamma}_k \mathbf{s}_{tk} + \boldsymbol{\epsilon}_{tk}, \tag{5}$$

for all t and k. From this, marginally, i.e. integrating out the factors and error terms, we have that

$$\mathrm{cov}(\mathbf{y}_{tk}^{(0)}) = \boldsymbol{\Gamma}_k \boldsymbol{\Gamma}_k^{\mathrm{T}} + \boldsymbol{\Psi}_k, \tag{6}$$

where $\boldsymbol{\Psi}_k = \mathrm{diag}(\psi_{1k}^2, \ldots, \psi_{nk}^2)$. Hence, the FA model assumes that the covariance of the potential (treatment-free) outcomes of the n units is the same at all time points. The preintervention data are used to learn about this covariance which is then used to predict the (counterfactual) potential outcomes of the treated units after intervention from the (observed) potential outcomes of the control units after intervention. The larger are n_1 and T_1, the more information is available to

estimate $\mathbf{\Gamma}_k$ and $\mathbf{\Psi}_k$, and hence the more accurately we can estimate them (and, from them, the expectation of $y_{itk}^{(0)}$).

It is worth noting that the FA model allows for a certain form of unmeasured confounding. This is because the aforementioned asymptotic unbiasedness property of $\hat{\theta}_{itk}$ does not require r_i to be independent of $\boldsymbol{\gamma}_{ik}$. If $\boldsymbol{\gamma}_{ik}$ is indeed associated with r_i then, because it is also associated with $y_{itk}^{(0)}$ (see equation (3)), it is an (unobserved) confounder.

3.3. Extending the factor analysis model

Our proposed model involves two extensions to the FA model: joint outcome modelling and temporal dependence. We present these two extensions separately, although the model that we finally propose, 'MVFA+AR', includes both extensions.

3.3.1. Joint outcome modelling

The classical FA model considers each of the K different outcomes independently; it makes no assumptions about correlations between outcomes k and k' ($k' \neq k$). In situations where the different outcomes are measures of, or are influenced by, a common underlying process, this may be an inefficient way to estimate intervention effects. For example, the outcomes gross domestic product and employment rate can be considered to be two measures of the underlying health of an economy; and rates of hospital admission, violent crime, sexual crime and antisocial behaviour are all influenced by problematic alcohol use. In these situations, part of the variability of the different outcomes is shared. Such shared variability can be modelled by using a multivariate FA model. As we explain below, the multivariate FA model enables the counterfactual post-intervention kth outcomes of the treated units to be estimated by using the data on all K outcomes, rather than (as in the FA model) just the data on the kth outcome. This makes it possible to estimate these counterfactual outcomes—and hence the intervention effects—more precisely.

The multivariate FA model assumes that

$$y_{itk}^{(0)} = \boldsymbol{\gamma}_{ik}^{\mathrm{T}} \mathbf{s}_{tk} + \boldsymbol{\lambda}_i^{\mathrm{T}} \mathbf{f}_{tk} + \varepsilon_{itk}, \tag{7}$$

where $\boldsymbol{\gamma}_{ik}$ and $\mathbf{s}_{tk}$ are as defined earlier (i.e. they are unit-specific loadings and time-specific factors, both of which are specific to the kth outcome), $\boldsymbol{\lambda}_i$ is the p_2-vector of unit-specific loadings that are shared across outcomes, $\mathbf{f}_{tk} \sim N_{p_2}(\mathbf{0}, \mathbf{I})$ is a p_2-vector of time-specific factors for $\boldsymbol{\lambda}_i$, and $\varepsilon_{itk} \sim N(0, \psi_{ik}^2)$ is the error term. Again, covariates can be included in the model by adding the term $\boldsymbol{\beta}_k^{\mathrm{T}} \mathbf{x}_{it}$ to the right-hand side of equation (7).

The interpretation of the multivariate FA model follows that of the FA model. More specifically, as well as $\boldsymbol{\gamma}_{ik}$, we now have $\boldsymbol{\lambda}_i$, which can be thought of as unit-specific unobserved variables that affect all outcomes; their effect on outcome k at time t is quantified by the factor $\mathbf{f}_{tk}$. One way to think about the benefit of jointly modelling the outcomes when estimating the counterfactual outcomes is by appreciating that the joint model learns about $\boldsymbol{\lambda}_i$, the unit-specific loadings that are common to all the outcomes, from the data on all the outcomes. This means that $\boldsymbol{\gamma}_{ik}^{\mathrm{T}} \mathbf{s}_{tk} + \boldsymbol{\lambda}_i^{\mathrm{T}} \mathbf{f}_{tk}$, the expectation of the counterfactual kth outcome, is more accurately estimated than in the case when modelling the outcomes independently. An alternative way to think about this benefit is to consider the covariance matrix for $\mathbf{y}_{ik}^{(0)}$. For the FA model, this is given by equation (6). For the multivariate FA model, we have that

$$\mathrm{cov}(\mathbf{y}_{tk}^{(0)}) = \mathbf{\Gamma}_k \mathbf{\Gamma}_k^{\mathrm{T}} + \mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}} + \mathbf{\Psi}_k,$$

for each k and t, where $\mathbf{\Lambda}$ is the $n \times p_2$ matrix with ith row $\boldsymbol{\lambda}_i$. By modelling the outcomes jointly, this covariance can be estimated more accurately, because the part that is attributable to the shared factors, i.e. $\mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}}$, is estimated by using data from all the K outcomes. Since this covariance is used to predict the (counterfactual) potential outcomes of the treated units after the intervention, estimating it more accurately should lead to more accurate estimation of those outcomes.

We expect the benefit of jointly modelling the outcomes to be greatest when T_1 is small. In this situation, there are little data to learn the unit-specific loadings, and so the gain from learning about some of them (specifically $\boldsymbol{\lambda}_i$) by using all the outcomes is likely to be most marked. Also, the greater is the proportion of factors that are common, i.e. the larger the p_2/p_1, the greater is likely to be the benefit from using the multivariate FA model. Note that joint modelling of multiple outcomes should be beneficial in terms of improving the precision of the estimate of the causal effect even when the effect of intervention on only one of the K outcomes is of interest. Also note that time-dependent variables that are predictive of $y_{itk}^{(0)}$ but which are affected by the intervention cannot be included as covariates in the multivariate FA model (as in equation (4)). However, they can be used as additional outcomes in the multivariate FA model.

3.3.2. *Modelling temporal dependence*

The effect of the unit-specific loading $\boldsymbol{\gamma}_{ik}$ (or $\boldsymbol{\lambda}_i$) on the outcomes at times t and t' is represented by $\mathbf{s}_{tk}$ and $\mathbf{s}_{t'k}$ (or $\mathbf{f}_{tk}$ and $\mathbf{f}_{t'k}$). It may be reasonable to believe that this effect is likely to be more similar at two nearby times than at two distant times; for example $\mathbf{s}_{tk}$ and $\mathbf{s}_{t+1,k}$ are likely to be more similar than $\mathbf{s}_{tk}$ and $\mathbf{s}_{t+10,k}$. Neither the FA nor the multivariate FA model described above takes this time ordering into account. We can take into account the time ordering by assuming that, for each outcome k, the factors are generated by an auto-regressive AR(1) process. Specifically, we assume that, for each k and $j = 1, \ldots, p_1 + p_2$, we have that

$$s_{tkj} = \rho_{kj} s_{t-1,kj} + \eta_{tkj}, \tag{8}$$

where $s_{tkj} = f_{tk,j-p_1}$ for $p_1 < j \leqslant p_1 + p_2$, $\rho_{kj} \in (-1, 1)$ are persistent parameters and $\eta_{tkj} \sim N(0, 1)$.

Assuming that factors are generated by an AR(1) process may improve prediction of the counterfactual outcomes $y_{itk}^{(0)}$ $(i > n_1, t > T_1)$ and hence increase the precision of the intervention effect estimates. This can become clear as follows. By integrating out the factors and error terms, we find that, for $t' \neq t$,

$$\operatorname{cov}(\mathbf{y}_{tk}^{(0)}, \mathbf{y}_{t'k}^{(0)}) = \mathbf{\Gamma}_k \operatorname{cov}(\mathbf{s}_{tk}, \mathbf{s}_{t'k}) \mathbf{\Gamma}_k^{\mathrm{T}} + \mathbf{\Lambda} \operatorname{cov}(\mathbf{f}_{tk}, \mathbf{f}_{t'k}) \mathbf{\Lambda}^{\mathrm{T}}. \tag{9}$$

Equation (9) shows that, by assuming an AR(1) prior for the factors, an *a priori* correlation both between y_{itk} and $y_{it'k}$ is allowed, as well as between y_{itk} and $y_{i't'k}$, where $i \neq i'$. If these correlations are strong, the sharing of information across time points can lead to more accurate estimates of the counterfactuals. This does not happen in the standard FA model which assumes that $\rho_{kj} = 0$, and therefore the right-hand side of equation (9) reduces to $\mathbf{0}$.

We expect that assuming an AR structure for the factors will increase efficiency in settings where n_1 is small and T_1 large. In these settings, there are few observations per time point and therefore factors cannot be estimated accurately. By assuming an AR(1) structure, we allow for the sharing of information between nearby time points. When T_1 is small, there may be less advantage, because there is then less information to estimate ρ_{kj}.

We call the FA model with this AR(1) structure 'FA+AR' and the multivariate FA model with AR(1) structure MVFA+AR.

3.4. Choosing the number of factors

One of the challenges when implementing FA is choosing the total number of factors in the model. Many researchers have proposed solutions for this problem; for example, Bai and Ng (2002) proposed some criteria to choose the number of factors; Lopes and West (2004) developed a reversible jump Markov chain Monte Carlo (MCMC) algorithm to estimate the number of factors; Carvalho *et al.* (2008) took an evolutionary stochastic model search approach; Srivastava *et al.* (2017) used a continuous shrinkage prior on the loadings. In this work, we account for the uncertainty in p_1 and p_2 by assuming a *multiplicative gamma process shrinkage* (MGPS) prior (Bhattacharya and Dunson, 2011) on the loadings.

We use the MGPS prior for both the outcome-specific loadings $\boldsymbol{\gamma}_{ik}$ and shared loadings $\boldsymbol{\lambda}_i$. We shall describe how this prior works for the outcome-specific loadings; for the shared loadings, the specifications are analogous. Let the loading vector $\boldsymbol{\gamma}_{ik}$ be of dimension $p_1 = \infty$. MGPS assumes that for each $j = 1, \ldots, \infty$ we have

$$\gamma_{ikj} \sim N\left(0, \frac{1}{\phi_{ikj}\tau_{kj}}\right), \tag{10}$$

where ϕ_{ikj} and τ_{kj} (both greater than 0) are the local and global shrinkage parameters respectively, such that

$$\tau_{kj} = \prod_{l=1}^{j} \delta_{kj}. \tag{11}$$

For appropriately chosen priors on ϕ_{ikj} and δ_{kj}, the product $\phi_{ikj}\tau_{kj}$ increases, thus encouraging the magnitude of the elements of $\boldsymbol{\gamma}_{ik}$ to decrease progressively towards 0. Hence, although the number of columns in each matrix $\boldsymbol{\Gamma}_k$ is infinite there will be a column such that all columns after this column have an L_1-norm of almost 0, indicating that no more factors are required for the data set under consideration.

In practice, it is not possible to carry out computations when loadings are infinite dimensional. So, we let $\boldsymbol{\gamma}_{ik}$ be of dimension k_1 (and k_2 for the shared loadings), where k_1 is sufficiently large. This approach can be computationally wasteful when p_1 is much smaller than the specified k_1. However, one can easily detect this through a pilot run of the algorithm that is used to simulate from the posterior; if most of the columns of $\boldsymbol{\Gamma}_{ik}$ have an L_1-norm that is very low, then it is recommended to decrease k_1 in the final run. Alternatively, an adaptive way to determine k_1 was discussed by Bhattacharya and Dunson (2011).

The MGPS prior can be used to perform inference on the number of factors. At iteration l of the MCMC algorithm (which we use to draw samples from the posterior), let $d^{(l)}$ denote the total number of columns in $\boldsymbol{\Gamma}_k$ whose absolute elements $|\gamma_{1kj}|, \ldots, |\gamma_{nkj}|$ are all below a prespecified threshold m. The effective number of factors at iteration l is $k_1 - d^{(l)}$. Therefore, one can use the posterior distribution of $k_1 - d^{(l)}$ to estimate the total number of factors in the model (e.g. as the posterior median of this distribution) and to construct credible intervals (by using the quantiles). This approach is sensitive to the choice of threshold m.

The reasons that we use the MGPS prior instead of the other methods that we mention are twofold. Firstly, MGPS allows for a conjugate formulation of the model, which simplifies posterior sampling. Secondly, it has been shown that this method performs well in a wide range of applications (e.g. Montagna *et al.* (2012), Montagna, Irincheeva and Tokdar (2018) and Montagna, Wager, Barrett, Johnson and Nichols (2018)).

3.5. Prior distributions and Markov chain Monte Carlo algorithm

The prior distributions are as follows. For all i and k, we let the variance parameters $\psi_{ik}^2 \sim$

InverseGamma(0.001, 0.001). For the AR parameters, $\rho_{kj} \sim \text{uniform}(-1, 1)$ for all k and j. For the shrinkage parameters, we follow recommendations by Bhattacharya and Dunson (2011) and let $\phi_{ikj} \sim \text{gamma}(\frac{3}{2}, \frac{3}{2})$ for all i, k and j, $\delta_{k1} \sim \text{gamma}(2.1, 1)$ for all k, and $\delta_{kj} \sim \text{gamma}(3.1, 1)$ for $j > 1$. If covariates $\mathbf{x}_{it}$ are included in MVFA+AR, we let the regression coefficients $\boldsymbol{\beta}_k \sim N(0, 10^3\mathbf{I})$ for all k.

The posterior distribution resulting from the MVFA+AR model of equations (7), (8), (10) and (11) and the prior distributions that are stated in this section is analytically intractable. We therefore use MCMC sampling to draw samples from it. In particular, we propose a hybrid Gibbs sampler where each parameter (or block of parameters) is sampled from its full conditional given the remaining parameters, using either Gibbs or Metropolis–Hastings steps. The main challenge is to simulate from high dimensional normal full conditionals. More specifically, the vector of factors $\mathbf{f}_k = (\mathbf{s}_{1k}^{\mathrm{T}}, \mathbf{f}_{1k}^{\mathrm{T}}, \ldots, \mathbf{s}_{Tk}^{\mathrm{T}}, \mathbf{f}_{Tk}^{\mathrm{T}})^{\mathrm{T}}$ for each outcome is drawn from a $T(k_1 + k_2)$-dimensional normal distribution, and the vector of loadings $\tilde{\boldsymbol{\lambda}}_i = (\boldsymbol{\lambda}_i^{\mathrm{T}}, \boldsymbol{\gamma}_{i1}^{\mathrm{T}}, \ldots, \boldsymbol{\gamma}_{iK}^{\mathrm{T}})^{\mathrm{T}}$ for each unit is drawn from a $(Kk_1 + k_2)$-dimensional normal distribution. We perform both these updates with good computational efficiency by using the method of Rue (2001). The update of AR hyperparameters ρ_{jk} is also challenging, because these parameters have bounded support and it is not possible to simulate them directly from their full conditionals. To overcome this issue, we update ρ_{jk} with Metropolis–Hastings steps by using the proposals that were developed by Kastner and Frühwirth-Schnatter (2014). The remaining model parameters $\boldsymbol{\beta}_k$, ϕ_{ikj}, δ_{kj} and ψ_{ik}^2 can be easily drawn from their full conditionals. For full details of the MCMC algorithm, see section A of the web-based supplementary material, where we also provide a sketch of the sampler. Similar MCMC algorithms can be used to draw from the posterior distribution of the FA, FA+AR and MVFA models.

We emphasize that the factors and loadings are not identifiable. Since we are not interested in interpreting these parameters but only in the counterfactuals (which are identifiable), we choose not to impose any identifiability constraints. Users who are interested in interpreting these parameters can resort to one of the existing approaches for ensuring identifiability; see for example section 12.1.3 of Murphy (2012) for a fairly recent overview. One method is to restrict the loading matrix to the class of lower diagonal matrices (Geweke and Zhou, 1996).

3.6. Point estimation and inference

Samples from the posterior distribution are used to obtain samples from the posterior distribution of the causal effect θ_{itk}. First, we simulate from the posterior predictive distribution of the counterfactual outcomes

$$y_{itk}^{(0,l)} = (\boldsymbol{\gamma}_{ik}^{(l)})^{\mathrm{T}} \mathbf{s}_{tk}^{(l)} + (\boldsymbol{\lambda}_i^{(l)})^{\mathrm{T}} \mathbf{f}_{tk}^{(l)} + \varepsilon_{itk}^{(l)} \qquad (i > n_1, t > T_1), \tag{12}$$

where $\varepsilon_{itk}^{(l)} \sim N\{0, (\psi_{ik}^2)^{(l)}\}$ and l indexes the MCMC draw. Then, samples $\theta_{itk}^{(l)}$ from the posterior of the individual effect θ_{itk} are readily available as $\theta_{itk}^{(l)} = y_{itk} - y_{itk}^{(0,l)}$. Similarly, samples $\vartheta_{tk}^{(l)}$ from the posterior of the average treatment effect ϑ_{tk} are obtained as $\vartheta_{tk}^{(l)} = (1/n_2)\Sigma_{i=n_1+1}^{n} \theta_{itk}^{(l)}$. We can use these to calculate point estimates and to perform inference. For instance, the point estimate of ϑ_{tk} will be $(1/L)\Sigma_{l=1}^{L} \vartheta_{tk}^{(l)}$, where L is the number of MCMC samples and the 95% credible interval for ϑ_{tk} will be given by the 2.5% and 97.5% percentiles of the $\vartheta_{tk}^{(l)}$. To test for a positive intervention effect, one can estimate the posterior probability that $\vartheta_{tk} > 0$ as $(1/L)\Sigma_{l=1}^{L} \mathbb{I}(\vartheta_{tk}^{(l)} > 0)$, where $\mathbb{I}(\cdot)$ is the indicator function.

4. Simulation studies

4.1. Setting

We performed a series of simulation studies to answer the question of whether we can obtain

estimates of ϑ_{tk} that are more precise than those obtained from the standard FA model by

(a) modelling multiple outcomes jointly,
(b) assuming an AR(1) structure for the factors and
(c) doing both simultaneously.

Each data set (from a total of 10000) was simulated as follows. We used MVFA+AR to generate data on $n=35$ units with $T_1=40$ preintervention and $T_2=5$ post-intervention time points. There were $K=3$ outcomes, $p_1=2$ outcome-specific loadings and $p_2=4$ shared loadings, and the persistent parameters of the factors were $\rho_{kj}=0.9$ for all k and j. (We set the variance of the error terms η_{tkj} in equation (8) to $1/(1-\rho_{jk})$ for all k and j, so that $s_{tkj}\sim N(0,1)$ for all t, j and k.) s_{0kj} were drawn from an $N(0,1)$ distribution. For each k, i and j, we drew the loadings from an $N(0,1)$ distribution. Finally, for each k and i, we set $\psi_{ik}^2=\frac{1}{3}$.

We randomly chose $n_2=5$ treated units from these 35 units by using the expected values on the first outcome. To introduce unobserved confounding, each unit had selection probability proportional to

$$\operatorname{expit}\left\{\kappa\sum_{t=41}^{45}(\boldsymbol{\gamma}_{i1}^{\mathrm{T}}\mathbf{s}_{t1}+\boldsymbol{\lambda}_i^{\mathrm{T}}\mathbf{f}_{t1})\right\}$$

of being selected to be a treated unit, where $\operatorname{expit}(\cdot)=\exp(\cdot)/\{1+\exp(\cdot)\}$. The value of κ controls the degree of unobserved confounding; $\kappa=0$ means no unobserved confounding, and $\kappa>0$ means that units with larger expected values of the post-intervention (possibly counterfactual) treatment-free first outcome are more likely to be treated. We chose $\kappa=0.75$ because we found that a simple t-test comparing $\mathcal{Y}_1=\{y_{1,41,1}^{(0)},\ldots,y_{30,41,1}^{(0)}\}$ and $\mathcal{Y}_2=\{y_{31,41,1}^{(0)},\ldots,y_{35,41,1}^{(0)}\}$ had a roughly 17.5% rejection rate (this would be around 5% if the elements of $\mathcal{Y}_1$ and $\mathcal{Y}_2$ were exchangeable). This procedure gave us data sets of $n_1=30$ control units and $n_2=5$ treated units (set-up I), and $T_1=40$ and $T_2=5$.

We expected the answers to questions (a)–(c) to depend on T_1 and n_1. To obtain data sets with fewer than 40 preintervention time points and/or fewer than 30 control units, we discarded the data in the first $40-T_1$ preintervention time points and/or randomly discarded $30-n_1$ control units. The values of (T_1,n_1) in set-ups II–IX are $(40,30)$, $(40,15)$, $(40,5)$, $(20,30)$, $(20,15)$, $(20,5)$, $(10,30)$, $(10,15)$ and $(10,5)$ respectively. The total number of treated units ($n_2=5$) and post-intervention observations ($T_2=5$) were common to all set-ups.

The point estimate of the ϑ_{tk} is

$$\hat{\vartheta}_{tk}=\frac{1}{n_2}\sum_{i=n_1+1}^{n}(y_{itk}^{(1)}-\hat{y}_{itk}^{(0)})=\vartheta_{tk}+\frac{1}{n_2}\sum_{i=n_1+1}^{n}(y_{itk}^{(0)}-\hat{y}_{itk}^{(0)}).$$

So, if the average (over simulated data sets and over treated units) value of $\hat{y}_{itk}^{(0)}$ ($i>n_1$ and $t>T_1$) is equal to the average (over simulated data sets and over treated units) of $y_{itk}^{(0)}$, then $\hat{\vartheta}_{tk}$ will be unbiased for any value of θ_{itk}. Similarly, if the credible interval for $\Sigma_{i=n_1+1}^{n}y_{itk}^{(0)}$ contains the true value of this sum, then the credible interval for ϑ_{tk} will also contain the true value of the ϑ_{tk} for any θ_{itk}. Thus, it sufficed to study the case where $\theta_{itk}=0$.

We fit the following models to all data sets: FA, FA+AR, MVFA and MVFA+AR. Note that all these models were correctly specified for the data that we generated. For all methods, we ran the MCMC algorithm for 31250 iterations, applied a thinning factor of 25 to obtain a total of 1250 posterior draws, discarded the first 250 as a burn-in and used the remaining 1000 for inference. FA and FA+AR are designed for univariate outcomes and hence we applied these to each of the outcomes in turn. For all models, we used the MGPS prior, setting $k_1=k_2=12$.

Table 1. Results of the simulation study for the *first outcome* $k = 1$ and post-intervention time points $t = T_1 + 1$ and $t = T$†

Model	*Results for the following set-ups:*								
	I	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*	*IX*
					T_1				
	40	*40*	*40*	*20*	*20*	*20*	*10*	*10*	*10*
					n_1				
	30	*15*	*5*	*30*	*15*	*5*	*30*	*15*	*5*
Results for $k = 1$ and $t = T_1 + 1$									
Bias									
MVFA+AR	0.017	0.030	0.119	0.034	0.056	0.155	0.072	0.103	0.195
MVFA	0.035	0.075	0.327	0.055	0.105	0.332	0.100	0.158	0.357
FA+AR	0.026	0.043	0.126	0.065	0.090	0.166	0.126	0.146	0.209
FA	0.044	0.087	0.324	0.082	0.131	0.330	0.144	0.186	0.351
Standard error									
MVFA+AR	0.314	0.354	0.480	0.347	0.388	0.508	0.396	0.440	0.539
MVFA	0.322	0.383	0.684	0.355	0.418	0.669	0.406	0.470	0.666
FA+AR	0.330	0.371	0.492	0.387	0.424	0.526	0.460	0.494	0.569
FA	0.335	0.398	0.691	0.392	0.449	0.677	0.466	0.517	0.680
Mean credible interval width									
MVFA+AR	1.245	1.394	1.904	1.370	1.548	2.055	1.628	1.833	2.286
MVFA	1.262	1.471	2.429	1.386	1.618	2.450	1.653	1.912	2.614
FA+AR	1.309	1.465	1.934	1.539	1.707	2.104	1.902	2.045	2.357
FA	1.317	1.517	2.425	1.532	1.736	2.462	1.898	2.076	2.634
False positive rate									
MVFA+AR	0.050	0.047	0.052	0.053	0.051	0.051	0.048	0.046	0.050
MVFA	0.054	0.055	0.085	0.059	0.058	0.088	0.051	0.056	0.074
FA+AR	0.048	0.049	0.052	0.051	0.049	0.057	0.049	0.050	0.057
FA	0.052	0.059	0.090	0.058	0.064	0.089	0.054	0.062	0.081
Results for $k = 1$ and $t = T$									
Bias									
MVFA+AR	0.014	0.032	0.208	0.043	0.074	0.263	0.113	0.170	0.333
MVFA	0.035	0.084	0.370	0.069	0.133	0.409	0.152	0.241	0.481
FA+AR	0.033	0.057	0.217	0.100	0.140	0.282	0.225	0.259	0.357
FA	0.054	0.107	0.372	0.121	0.186	0.420	0.246	0.308	0.490
Standard error									
MVFA+AR	0.326	0.375	0.676	0.374	0.437	0.737	0.484	0.569	0.800
MVFA	0.331	0.398	0.780	0.383	0.467	0.825	0.496	0.606	0.877
FA+AR	0.355	0.414	0.703	0.465	0.538	0.784	0.640	0.703	0.861
FA	0.358	0.433	0.800	0.469	0.554	0.858	0.642	0.717	0.913
Mean credible interval width									
MVFA+AR	1.284	1.485	2.483	1.485	1.746	2.660	1.913	2.216	2.931
MVFA	1.283	1.498	2.456	1.459	1.714	2.533	1.874	2.163	2.789
FA+AR	1.392	1.607	2.518	1.812	2.052	2.722	2.454	2.626	3.023
FA	1.373	1.574	2.462	1.728	1.920	2.570	2.341	2.471	2.863

(*continued*)

Table 1 (*continued*)

Model	*Results for the following set-ups:*								
	I	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*	*IX*
					T_1				
	40	*40*	*40*	*20*	*20*	*20*	*10*	*10*	*10*
					n_1				
	30	*15*	*5*	*30*	*15*	*5*	*30*	*15*	*5*
False positive rate									
MVFA+AR	0.050	0.050	0.066	0.049	0.049	0.078	0.052	0.058	0.084
MVFA	0.055	0.062	0.125	0.060	0.076	0.147	0.067	0.086	0.141
FA+AR	0.052	0.054	0.076	0.053	0.063	0.089	0.065	0.072	0.099
FA	0.057	0.074	0.133	0.070	0.091	0.159	0.082	0.102	0.155

†The table presents the bias of the point estimates of ϑ_{t1}, the standard error of the point estimates, the mean width of the 95% credible intervals and the false positive rate. All results are based on 10000 simulated data sets from MVFA+AR.

We used the posterior mean as the point estimate of the ϑ_{tk}; credible intervals were obtained by using the 2.5% and 97.5% quantiles of the posterior distribution.

We compared the performance of the models in terms of the bias and standard error of the point estimates for the ϑ_{tk}s, and mean width and false positive rate of the 95% credible intervals of the ϑ_{tk}s. As a measure of 'power', we used the probability of detecting an intervention effect when the true ϑ_{tk} was not equal to 0. Here, we defined 'detection' as a credible interval that excludes zero. For convenience, we assumed that θ_{itk} was the same for all units, times and outcomes.

4.2. Results

The results for the first outcome ($k = 1$) are summarized in Table 1 and Figs 2 and 3. Table 1 presents the bias, standard error, mean credible interval width and false positive rate for $(k, t) = (1, T_1 + 1)$ and $(k, t) = (1, T)$. Figs 2 and 3 show power for $(k, t) = (1, T_1 + 1)$ and $(k, t) = (1, T)$ respectively. In section B of the web-based supplementary material, we present results for $(k, t) = (2, T_1 + 1)$ (Table 1 and Fig. 1), and for $(k, t) = (3, T)$ (Table 1 and Fig. 2).

To answer question (a), we compare the results that were obtained from MVFA+AR and MVFA with the results that were obtained from FA+AR and the FA model respectively. In settings where $T_1 < 40$ and $n_1 \geqslant 15$ (i.e. set-ups IV, V, VII and VIII), we see that joint modelling of outcomes leads to considerable gains in precision: MVFA+AR and MVFA decrease the standard error of the point estimates and the mean credible interval width in these settings (see Table 1 and Table 1 of the web-based supplementary material section B). The gains in efficiency are also apparent from the power: Figs 2 and 3 and 1 and 2 of web-based supplementary material section B show that the use of a multivariate model instead of a univariate model can substantially improve the chance of detecting an intervention effect. For example, for $(k, t) =$

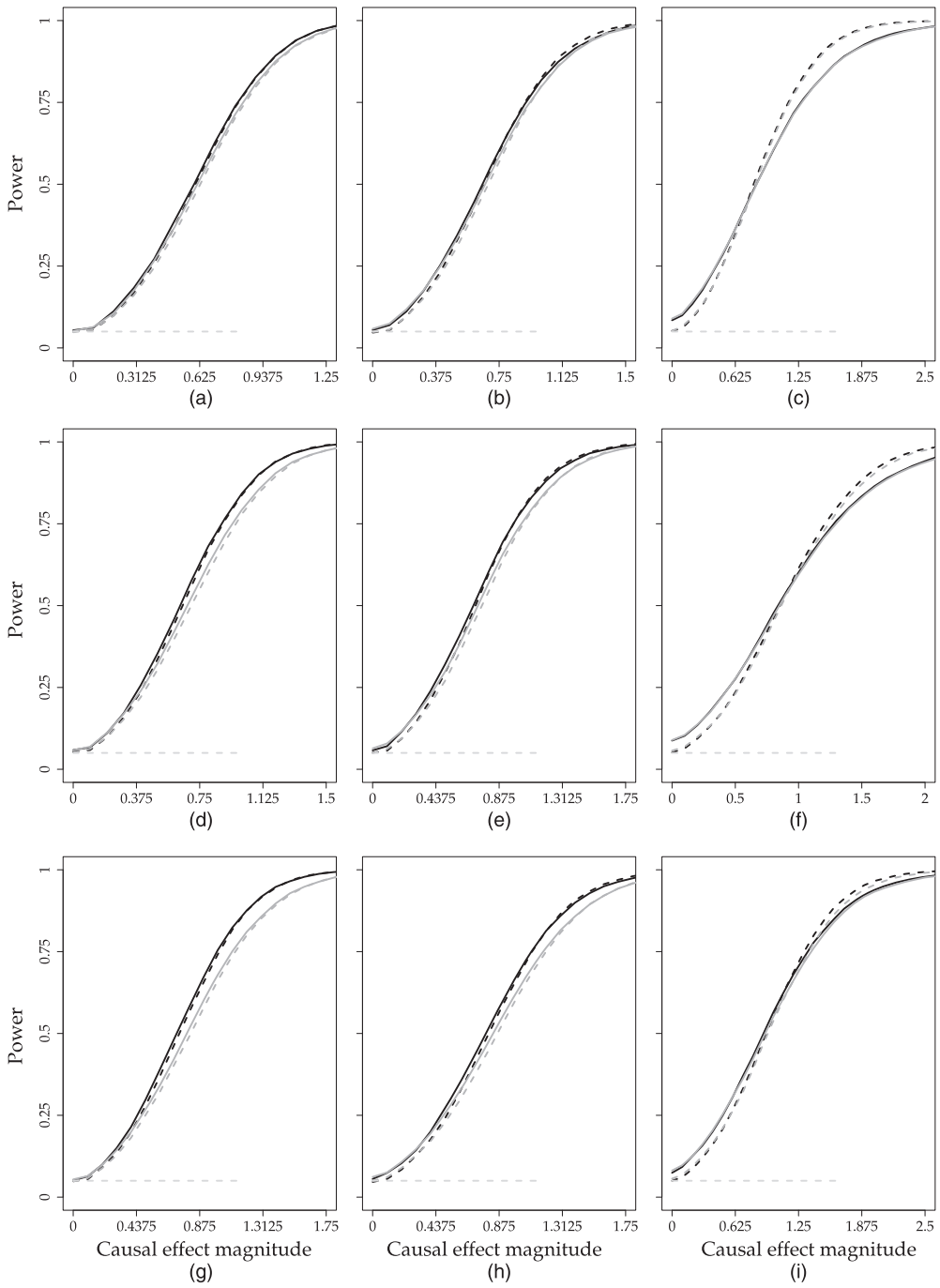

Fig. 2. Results of the simulation study for the *first outcome* $k = 1$ and *first post-intervention time point* $t = T_1 + 1$ (the figure presents the probability of detecting an intervention effect (y-axis) as a function of $\vartheta_{T_1+1,1}$ (x-axis) in set-ups I–IX; all results are based on 10000 data sets simulated from MVFA+AR (– – – –)) (– – – –, 5%, the desired detection rate when the intervention $\vartheta_{T_1+1,1} = 0$; ———, MVFA; – – – –, FA+AR; ———, FA): (a) set-up I, $T_1 = 40$, $n_1 = 30$; (b) set-up II, $T_1 = 40$, $n_1 = 15$; (c) set-up III, $T_1 = 40$, $n_1 = 5$; (d) set-up IV, $T_1 = 20$, $n_1 = 30$; (e) set-up V, $T_1 = 20$, $n_1 = 15$; (f) set-up VI, $T_1 = 20$, $n_1 = 5$; (g) set-up VII, $T_1 = 10$, $n_1 = 30$; (h) set-up VIII, $T_1 = 10$, $n_1 = 15$; (i) set-up IX, $T_1 = 10$, $n_1 = 5$

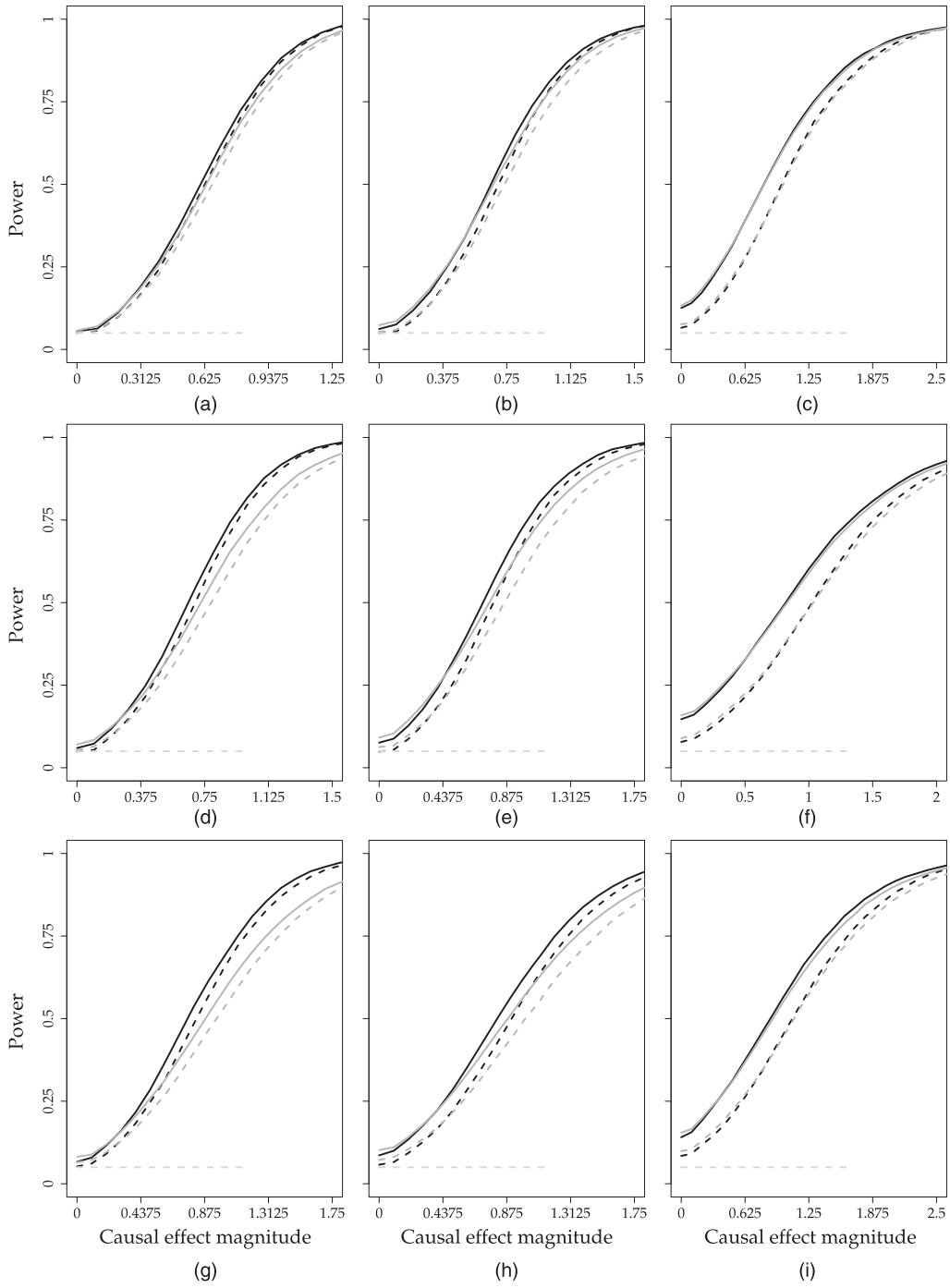

Fig. 3. Results of the simulation study for the *first outcome* $k = 1$ and *last post-intervention time point* $t = T$ (the figure presents the probability of detecting an intervention effect (y-axis) as a function of $\vartheta_{T,1}$ (x-axis) in set-ups I–IX; all results are based on 10000 data sets simulated from MVFA+AR (– – –)) (– – –, 5%, the desired detection rate when the intervention $\vartheta_{T,1} = 0$; ———, MVFA; – – –, FA+AR; ———, FA): (a) set-up I, $T_1 = 40$, $n_1 = 30$; (b) set-up II, $T_1 = 40$, $n_1 = 15$; (c) set-up III, $T_1 = 40$, $n_1 = 5$; (d) set-up IV, $T_1 = 20$, $n_1 = 30$; (e) set-up V, $T_1 = 20$, $n_1 = 15$; (f) set-up VI, $T_1 = 20$, $n_1 = 5$; (g) set-up VII, $T_1 = 10$, $n_1 = 30$; (h) set-up VIII, $T_1 = 10$, $n_1 = 15$; (i) set-up IX, $T_1 = 10$, $n_1 = 5$

$(1, T)$ and set-up VII, we find that, when $\vartheta_{T_1+1,1} = 1.2$, the intervention effect is detected by MVFA with probability 78% whereas it is detected by the FA model with probability 66%. The power curves for $(k, T) = (2, T_1 + 1)$ and $(k, t) = (3, T)$ (web-based supplementary material section B) reveal a similar pattern. We find no settings in which the univariate models outperform the corresponding multivariate models for any of the performance measures that we consider.

To answer question (b), we compare the results that were obtained from MVFA+AR and FA+AR with those obtained from MVFA and the FA model respectively. We find that the inclusion of the AR component leads to either improved or unchanged performance. The improvements occur mainly in settings where $n_1 = 5$ (i.e. set-ups III, VI and IX). In these settings with few control units, the FA and MVFA models perform very poorly in terms of bias and false positive rate for outcome $k = 1$ (see Table 1). FA+AR and MVFA+AR offer significant improvements in terms of both bias and false positive rate compared with the FA and MVFA models. For example, for $(k, t) = (1, T)$ and set-up VI, the false positive rate is 8.9% for the FA model whereas it is 5.7% for MVFA+AR. Note that, for outcomes $k = 2$ and $k = 3$, the bias and false positive rate in set-ups III and VI are not as high as for outcome $k = 1$ (see Table 1 of web-based supplementary material section B). The reason is that we have chosen the treated units by using the expected outcomes on $k = 1$ and therefore the effect of confounding is greater for $k = 1$. The inclusion of the AR component also leads to gains in efficiency in set-ups III and VI, as it reduces both the standard error of the point estimates and the mean credible interval width (Table 1). The gains in power can be moderate. For example, for $(k, t) = (1, T)$ and set-up III, the intervention effect is detected with probability 90% by FA+AR and 83% by the FA model when $\vartheta_{T_1+1,1} = 1.5$ (Fig. 2). The improvement in power is more prominent for outcome $k = 2$ (because the bias of all methods is close to 0 for this outcome). For instance, in set-up III, a $\vartheta_{T_1+1,3} = 1.5$ is detected with probability 86% by FA+AR and 72% by the FA model (Fig. 1 of web-based supplementary material section B).

We find no set-ups in which MVFA+AR performs better than both MVFA and FA+AR. The reason is that, as we explained earlier in Section 3.3, the two proposed extensions (joint outcome modelling and the AR(1) prior on factors) improve on FA in very different settings: the former when T_1 is small and n_1 is large; the latter when T_1 is large and n_1 is small. In contrast, we find no settings where either FA+AR or MVFA outperforms MVFA+AR. Therefore, the advantage of MVFA+AR is that it can be used in all settings. For this reason, we suggest that this is the model that should be used in practice.

The gains in efficiency that are obtained by using either FA+AR, MVFA and MVFA+AR will depend not only on T_1 and n_1 but also on the total number of outcomes K, the number of factors $p_1 + p_2$ and the ratio p_1/p_2. As K increases, $\boldsymbol{\lambda}_i$ will be estimated with higher precision. As the total number of factors $p_1 + p_2$ increases, the total number of model parameters that need to be estimated increases. Hence, sharing of information (either between outcomes by using MVFA or MVFA+AR or between time points by using FA+AR or MVFA+AR) is more important for larger values of $p_1 + p_2$. Finally, MVFA and MVFA+AR are well suited to applications where the ratio p_1/p_2 is low, i.e. where the number of shared loadings is large compared with the number of outcome-specific loadings.

5. Application to the motivating data set

5.1. Model details

In this section we apply the proposed methodology to the alcohol licensing data set that was introduced in Section 2. The data set consists of data on $K = 4$ outcomes relating to the harms of alcohol consumption in society. For each outcome, there are $n_1 = 72$ control and $n_2 = 4$

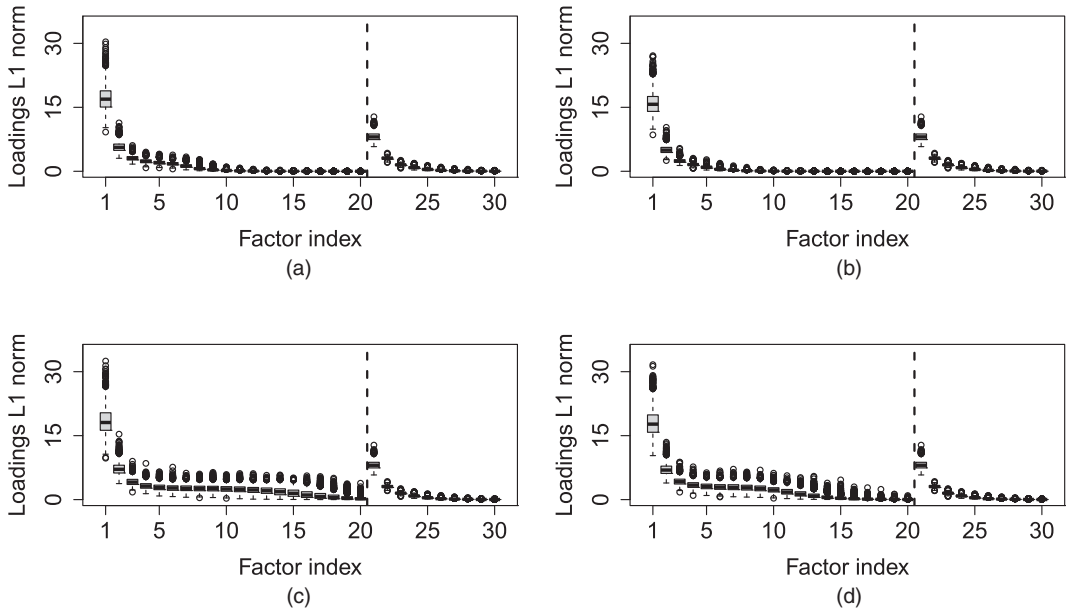

Fig. 4. Analysis of the alcohol licensing data set by using the proposed MVFA model (the figure presents posterior boxplots of $\Sigma_{i=1}^{n}|\gamma_{ikj}|$, the L_1-norm of the jth column of the loadings matrix; the boxplots are based on an MCMC sample of size 2500; factors 1–20 are outcome specific whereas factors 21–30 are shared across outcomes): (a) antisocial behaviour; (b) violent crimes; (c) hospital admissions; (d) sexual crimes

treated units. There are $T = 16$ observations per unit per outcome, $T_1 = 10$ of which are in the preintervention period. The objectives of the analysis are threefold. Firstly, we are interested in assessing the evidence for the existence of common factors underlying the four outcomes. Secondly, we are aiming to investigate the effect of the intervention on each of the four treated units (Derby, Enfield, Kingston upon Thames and Southwark) individually. Thirdly, we wish to assess the evidence for a non-zero average intervention effect ϑ_{tk} $(t > 10)$.

To achieve these goals, we fit our proposed model MVFA+AR. We set $k_1 = 20$ and $k_2 = 10$. We run the MCMC algorithm for 1500000 iterations (this took approximately 9 h on a Linux machine with an Intel i7-6700 3.4-GHz central processor unit); the first 250000 samples are discarded as a burn-in and we apply a thinning factor of 500 to the remaining draws. Therefore our MCMC sample consists of 2500 draws from the joint posterior of the model parameters. Convergence is assessed by visual inspection of posterior trace plots for some randomly chosen shrinkage parameters δ_{jk}, variance parameters ψ_{ik}^2 and the counterfactuals $y_{itk}^{(0)}$ $(i > 72$ and $t > 10)$. These indicate that the chain has reached its stationary distribution. Further, we run an additional nine chains and compare the posterior densities of these parameters with those obtained from the first chain. No major differences are found. Therefore, we conclude that the chain has converged.

For each outcome, we also fit the univariate FA model with the MGPS prior and $k_1 = 20$. However, the conclusions that we reached regarding the effect of the intervention are very similar (except for minor losses in precision of the causal effect estimates) to those obtained from MVFA+AR. Thus, the results from the FA model are not discussed further.

5.2. Results

Fig. 4 summarizes the posterior distribution of $\Sigma_{i=1}^{n_1+n_2}|\gamma_{ikj}|$, i.e. the L_1-norm of the jth column

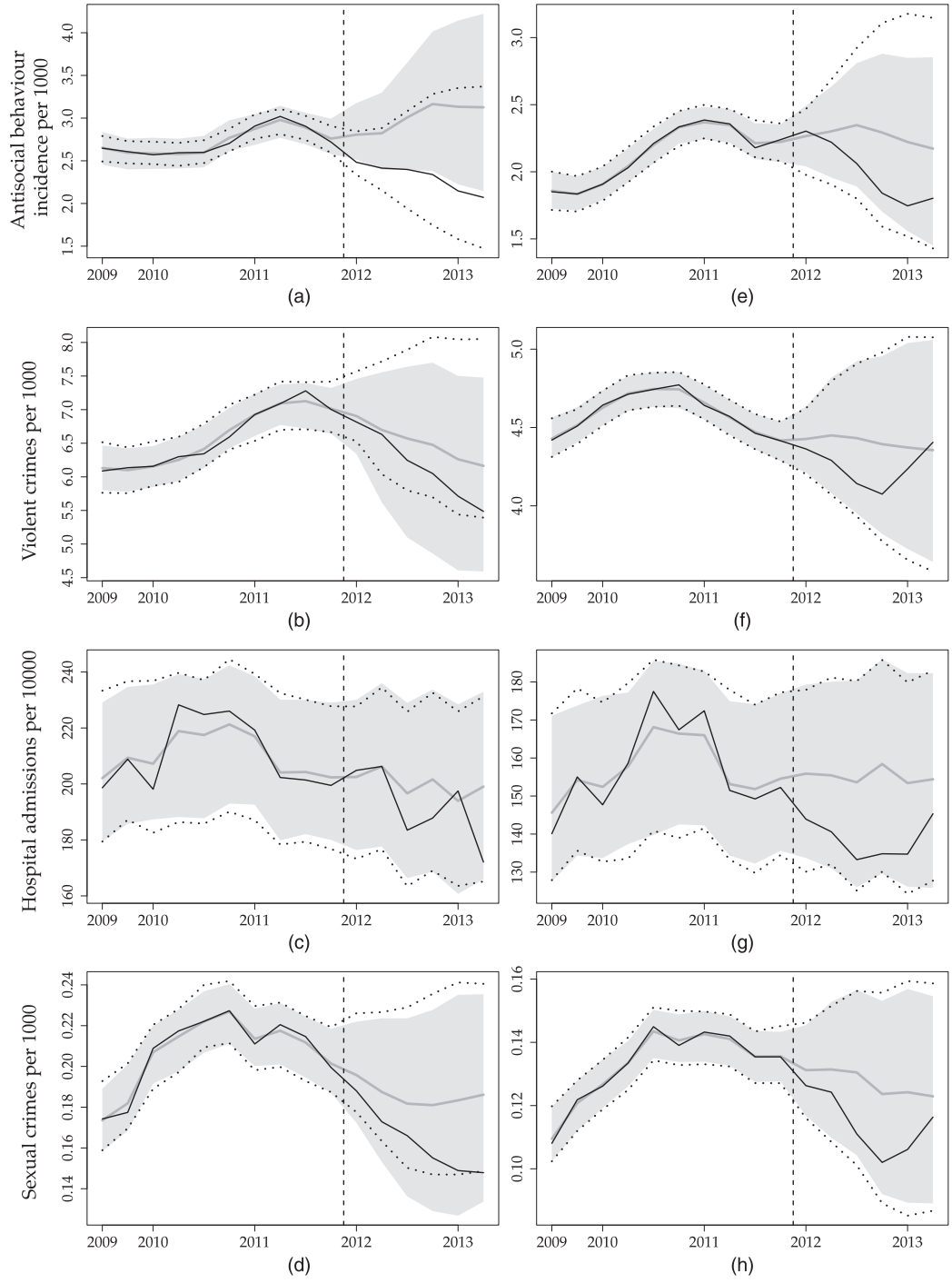

Fig. 5. Results of the real data analysis for (a)–(d) Derby and (e)–(h) Enfield (———, observed data; ———, posterior mean of $y_{itk}^{(0)}$ obtained by fitting MVFA+AR; ■, 95% credible intervals of $y_{itk}^{(0)}$ obtained from the same model; · · · · · ·, 95% credible intervals for $y_{itk}^{(0)}$ obtained by analysing each outcome in turn with the FA model): (a), (e) antisocial behaviour; (b), (f) violent crimes; (c), (g) hospital admissions; (d), (h) sexual crimes

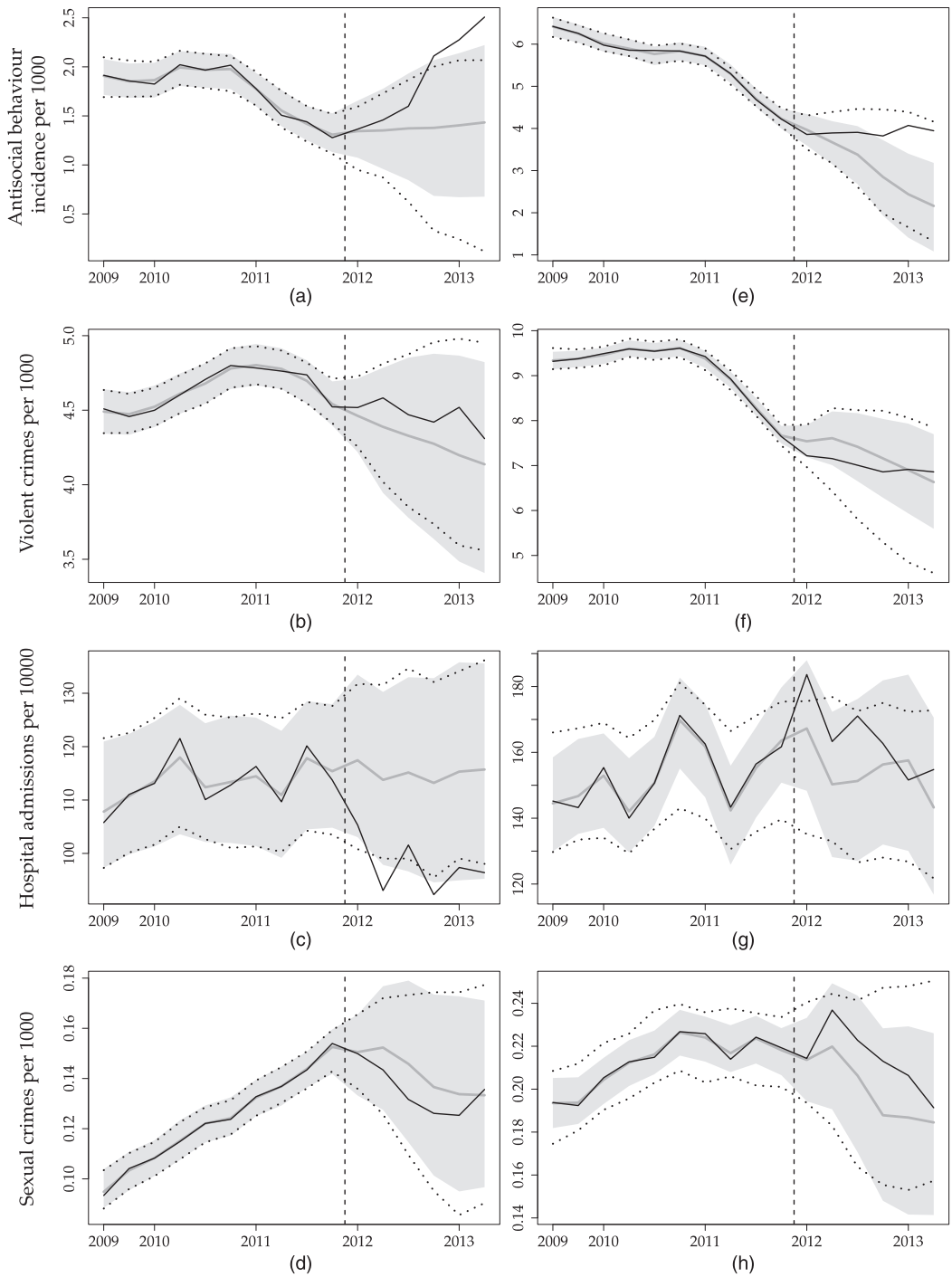

Fig. 6. Results of the real data analysis for (a)–(d) Kingston and (e)–(h) Southwark (———, observed data; ———, posterior mean of $y_{itk}^{(0)}$ obtained by fitting MVFA+AR; ■, 95% credible intervals of $y_{itk}^{(0)}$ obtained from the same model; · · · · · · , 95% credible intervals for $y_{itk}^{(0)}$ obtained by analysing each outcome in turn with the FA model): (a), (e) antisocial behaviour; (b), (f) violent crimes; (c), (g) hospital admissions; (d), (h) sexual crimes

Table 2. Estimated average (over units) treatment effect for each outcome and post-intervention time point (with 95% posterior credible intervals in parentheses) and average (over units) observed values for each outcome and post-intervention time point

t	*Results for antisocial behaviour*	*Results for violent crimes*	*Results for hospital admissions*	*Results for sexual crimes*
Estimates of ϑ_{tk}				
11	−0.09 [−0.26, 0.07]	−0.1 [−0.29, 0.09]	−1.2 [−13.1, 9.8]	−0.003 [−0.013, 0.007]
12	−0.04 [−0.27, 0.19]	−0.12 [−0.44, 0.21]	−5.6 [−18.6, 6.7]	−0.003 [−0.018, 0.011]
13	−0.04 [−0.36, 0.28]	−0.22 [−0.63, 0.23]	−6.9 [−20.9, 7.2]	−0.008 [−0.026, 0.011]
14	0.11 [−0.32, 0.52]	−0.22 [−0.7, 0.27]	−12.9 [−27.9, 1.2]	−0.008 [−0.028, 0.012]
15	0.26 [−0.19, 0.73]	−0.09 [−0.6, 0.45]	−9.8 [−25.1, 5]	−0.01 [−0.031, 0.012]
16	0.36 [−0.12, 0.86]	−0.06 [−0.58, 0.49]	−11.2 [−26.6, 3.2]	−0.009 [−0.029, 0.013]
Mean observed values				
11	2.50	5.73	159.4	0.170
12	2.50	5.66	150.8	0.169
13	2.49	5.47	147.3	0.158
14	2.53	5.35	144.4	0.149
15	2.56	5.35	145.3	0.147
16	2.58	5.26	142.2	0.148

of the loadings matrix, where $\gamma_{ikj} = \lambda_{ik,j-k_1}$ for $j > k_1$. We see that the norm quickly decreases with j for both outcome-specific and shared factors, demonstrating the shrinkage effects of the MGPS prior. Inference on the number of non-negligible factors can be done as described in Section 3.4. If we use $m = 0.1$, then the median posterior number of non-negligible shared factors is 2 (95% credible interval [2,4]) and the median posterior number of factors specific to outcomes 1–4 is 6 (95% credible interval [4,7]), 4 (95% credible interval [3,5]), 14 (95% credible interval [10,19]) and 11 (95% credible interval [9,14]) respectively.

There is not enough evidence in the data to support a significant intervention effect in each unit individually. This is evident in Figs 5 and 6 which show estimated counterfactuals along with their 95% credible intervals for Derby and Enfield, and Kingston and Southwark respectively. We see that, for all treated units and outcomes, the estimated counterfactuals do not appear to be systematically higher than observed values. Further, the 95% credible intervals of the counterfactuals contain the observed values y_{itk} for most of the combinations of i ($i > 72$), t ($t > 10$) and k. This suggests that the data are compatible with what would have been observed if intervention had not taken place.

The point estimates of ϑ_{tk}, the average (over units) intervention effect, for all post-intervention time points 11–16 and outcomes, along with their 95% credible intervals, are shown in Table 2. We see that the credible intervals for antisocial behaviour, violent crimes and sexual crimes are nearly symmetrical about zero. Therefore we conclude that there is no evidence for an effect of the intervention on these outcomes. For hospital admissions, the point estimates are all negative (a negative value means that admissions would be higher with no intervention). One of the advantages of the Bayesian approach is that it enables us to estimate many interesting causal quantities directly from the posterior distribution of the counterfactuals. Here we focus on the probability that $\vartheta_{tk} > 0$, which for hospital admissions and time points 11–16 is 0.41, 0.18, 0.16, 0.04, 0.10 and 0.07 respectively. Some of these values are low, suggesting that the intervention succeeded in reducing the rate of hospital admissions. However, most of them are higher than 5% and therefore the evidence is inconclusive.

6. Discussion

In this work, we have introduced the model MVFA+AR for evaluating the effect of a dichotomous intervention from time series observational data. Our model extends in two ways the FA model that is frequently used for causal inference in this setting. First, it models multiple correlated outcomes jointly. Second, it accounts for auto-correlation within each of the outcomes. Both of these extensions enable more efficient estimation of the effect of an intervention on all, or any one, of the multiple outcomes. An important facet of the model proposed is that it provides posterior credible intervals for the causal effects of interest that account for the uncertainty about the number of factors.

The ability to make inference is inherent in the Bayesian framework and therefore in our method. This gives it an advantage over many existing approaches for causal inference using time series observational data, including frequentist approaches based on the FA model (Gobillon and Magnac, 2016; Chan and Kwok, 2016; Xu, 2017; Li, 2018) and synthetic control-type approaches (Abadie *et al.*, 2010; Hsiao *et al.*, 2012; Doudchenko and Imbens, 2016; Ben-Michael *et al.*, 2018). The reason is that, to allow for inference, these methods require assumptions that might be unlikely to hold in some applications and therefore may yield confidence intervals that do not reflect the true uncertainty in the estimates of the causal effect. For example, the parametric bootstrap approach of Xu (2017) relies on the assumption that the error terms in the FA model are homoscedastic at each time point. For the approaches of Abadie *et al.* (2010) and Hsiao *et al.* (2012), inference is typically done with a 'placebo test': a procedure that is akin to a permutation test. However, the validity of this test is debatable unless we are willing to assume that the unit that received the intervention was chosen at random (Ben-Michael *et al.*, 2018). These assumptions are not essential for our method, suggesting that it can be a useful alternative in applications where they are unlikely to hold.

Our simulation studies indicate that the estimates of intervention effects obtained from MVFA+AR are more precise than those obtained from the standard FA model. This can lead to considerable gains in power for detecting an intervention effect. Further, we found that MVFA+AR has better type I error rate control compared with the standard FA model. Both these gains occur when either the total number of preintervention time points or the total number of control units is relatively small. This is so in many practical problems.

We applied our methodology to estimate the effect of CIPs on alcohol-related harms. We found evidence for the existence of common factors driving the outcomes that we considered. We identified no major effect of CIPs on the rate of antisocial behaviour incidents, violent crimes or sexual crimes. The analysis provides some evidence that the intervention has led to a decrease in the rate of alcohol-related hospital admissions. However, the effect is not significant, i.e. the 95% credible intervals contain zero.

There are limitations to the method proposed. Our model allows for loadings that are shared across all outcomes. However, with $K > 2$ outcomes there is the possibility that there are loadings which are shared between only k of them, where $2 \leqslant k \leqslant K - 1$. There may be a benefit in extending the model to allow for loadings that are common only to a subset of the outcomes. Another extension that may be useful would replace the AR(1) structure that we assume with a more general auto-regressive moving average process. However, this may not be feasible, given the length of the time series that is encountered in many practical applications.

Several possible directions for future research exist. The model proposed does not make use of the geographical location of the units. Such information may be of value, since we expect the outcomes of units with spatial proximity to be correlated. It may be worth extending the model to account for this. Lopes *et al.* (2008) achieved this by assuming a spatial model for

the loadings. Finally, although our model should perform well when the normality assumption holds approximately, it cannot be used when the data drastically deviate from this assumption, e.g. when the outcomes are dichotomous. Therefore, it is important to develop a model for mixed outcomes. We shall consider such extensions in our future research.

Acknowledgements

The authors thank Frank de Vocht for useful discussions regarding the real data application. This work is funded by the National Institute for Health Research Health Protection Unit on Evaluation of Interventions (to PS and MH), Medical Research Council grants MC_UU_00002/10 (to SRS) and MC_UU_00002/11 (to DDeA), and by Public Health England (to DDeA). Dr Montagna was partially supported by grant MONS_RILO_18_02 from the University of Turin. This study is further funded by the National Institute for Health Research programme grants for applied research programme (grant RP-PG-0616-20008). The views expressed are those of the author(s) and not necessarily those of the National Institute for Health Research or the Department of Health and Social Care.

References

Abadie, A., Diamond, A. and Hainmueller, J. (2010) Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. *J. Am. Statist. Ass.*, **105**, 493–505.

Angrist, J. D. and Pischke, J.-S. (2009) *Mostly Harmless Econometrics: an Empiricist's Companion*. Princeton: Princeton University Press.

Avalos-Pacheco, A., Rossell, D. and Savage, R. S. (2018) Heterogeneous large datasets integration using Bayesian factor regression. *Preprint arXiv:1810.09894*. Harvard University, Boston.

Bai, J. (2009) Panel data models with interactive fixed effects. *Econometrica*, **77**, 1229–1279.

Bai, J. and Ng, S. (2002) Determining the number of factors in approximate factor models. *Econometrica*, **70**, 191–221.

Bartolucci, F., Pennoni, F. and Vittadini, G. (2016) Causal latent Markov model for the comparison of multiple treatments in observational longitudinal studies. *J. Educ. Behav. Statist.*, **41**, 146–179.

Ben-Michael, E., Feller, A. and Rothstein, J. (2018) The augmented synthetic control method. *Preprint arXiv:1811.04170*. University of California at Berkeley, Berkeley.

Bhattacharya, A. and Dunson, D. B. (2011) Sparse Bayesian infinite factor models. *Biometrika*, **98**, 291–306.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N. and Scott, S. L. (2015) Inferring causal impact using Bayesian structural time-series models. *Ann. Appl. Statist.*, **9**, 247–274.

Card, D. (1990) The impact of the Mariel boatlift on the Miami labor market. *Industrl Lab. Reln*, **43**, 245–257.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q. and West, M. (2008) High-dimensional sparse factor modeling: applications in gene expression genomics. *J. Am. Statist. Ass.*, **103**, 1438–1456.

Cavallo, E., Galiani, S., Noy, I. and Pantano, J. (2013) Catastrophic natural disasters and economic growth. *Rev. Econ. Statist.*, **95**, 1549–1561.

Chan, M. and Kwok, S. (2016) Policy evaluation with interactive fixed effects. University of Sydney, Sydney.

De Vito, R., Bellio, R., Trippa, L. and Parmigiani, G. (2018a) Multi-study factor analysis. *Biometrics*, **75**, 337–346.

De Vito, R., Bellio, R., Trippa, L. and Parmigiani, G. (2018b) Bayesian multi-study factor analysis for high-throughput biological data. *Preprint arXiv:1806.09896*. Brown University, Providence.

Doudchenko, N. and Imbens, G. W. (2016) Balancing, regression, difference-in-differences and synthetic control methods: a synthesis. *Preprint arXiv:1610.07748*. University of Stanford, Stanford.

Geweke, J. and Zhou, G. (1996) Measuring the pricing error of the arbitrage pricing theory. *Rev. Finan. Stud.*, **9**, 557–587.

Gobillon, L. and Magnac, T. (2016) Regional policy evaluation: interactive fixed effects and synthetic controls. *Rev. Econ. Statist.*, **98**, 535–551.

Holland, P. W. (1986) Statistics and causal inference. *J. Am. Statist. Ass.*, **81**, 945–960.

Hsiao, C., Ching, S. H. and Wan, S. K. (2012) A panel data approach for program evaluation: measuring the benefits of political and economic integration of Hong Kong with mainland China. *J. Appl. Econmetr.*, **27**, 705–740.

Jones, A. M. and Rice, N. (2011) Econometric evaluation of health policies. In *The Oxford Handbook of Health Economics* (eds S. Glied and P. Smith), pp. 890–923. Oxford: Oxford University Press.

Kastner, G. and Frühwirth-Schnatter, S. (2014) Ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC estimation of stochastic volatility models. *Computnl Statist. Data Anal.* C, **76**, 408–423.
Lanza, S. T., Coffman, D. L. and Xu, S. (2013) Causal inference in latent class analysis. *Struct. Equn Modlng*, **20**, 361–383.
Li, K. (2018) Inference for factor model based average treatment effects. *Preprint*. University of Pennsylvania, Philadelphia. (Available from SSRN 3112775.)
Lopes, H. F. and West, M. (2004) Bayesian model assessment in factor analysis. *Statist. Sin.*, **14**, 41–67.
Lopes, H. F., Salazar, E. and Gamerman, D. (2008) Spatial dynamic factor analysis. *Baysn Anal.*, **3**, 759–792.
McAlinn, K., Aastveit, K. A., Nakajima, J. and West, M. (2019) Multivariate Bayesian predictive synthesis in macroeconomic forecasting. *J. Am. Statist. Ass.*, to be published.
Montagna, S., Irincheeva, I. and Tokdar, S. T. (2018) High-dimensional Bayesian Fourier analysis for detecting circadian gene expressions. *Preprint arXiv:1809.04347*. University of Toronto, Toronto.
Montagna, S., Tokdar, S. T., Neelon, B. and Dunson, D. B. (2012) Bayesian latent factor regression for functional and longitudinal data. *Biometrics*, **68**, 1064–1073.
Montagna, S., Wager, T., Barrett, L. F., Johnson, T. D. and Nichols, T. E. (2018) Spatial Bayesian latent factor regression modeling of coordinate-based meta-analysis data. *Biometrics*, **74**, 342–353.
Murphy, K. P. (2012) *Machine Learning: a Probabilistic Perspective*. Cambridge: MIT Press.
Robbins, M. W., Saunders, J. and Kilmer, B. (2017) A framework for synthetic control methods with high-dimensional, micro-level data: evaluating a neighbourhood-specific crime intervention. *J. Am. Statist. Ass.*, **112**, 109–126.
Robins, J. M., Hernan, M. A. and Brumback, B. (2000) Marginal structural models and causal inference in epidemiology. *Epidemiology*, **11**, 550–560.
Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, **66**, 688–701.
Rue, H. (2001) Fast sampling of Gaussian Markov random fields. *J. R. Statist. Soc.* B, **63**, 325–338.
Samartsidis, P., Seaman, S. R., Presanis, A. M., Hickman, M. and De Angelis, D. (2019) Assessing the causal effect of binary interventions from observational panel data with few treated units. *Statist. Sci.*, **34**, 486–503.
Srivastava, S., Engelhardt, B. E. and Dunson, D. B. (2017) Expandable factor analysis. *Biometrika*, **104**, 649–663.
Tullio, F. and Bartolucci, F. (2019) Evaluating time-varying treatment effects in latent Markov models: an application to the effect of remittances on poverty dynamics. *Paper 91459*. University of Munich, Munich.
de Vocht, F. (2016) Inferring the 1985–2014 impact of mobile phone use on selected brain cancer subtypes using Bayesian structural time series and synthetic controls. *Environ. Int.*, **97**, 100–107.
de Vocht, F., Tilling, K., Pliakas, T., Angus, C., Egan, M., Brennan, A., Campbell, R. and Hickman, M. (2017) The intervention effect of local alcohol licensing policies on hospital admission and crime: a natural experiment using a novel Bayesian synthetic time-series method. *J. Epidem. Commty Hlth*, **71**, 912–918.
Xu, Y. (2017) Generalized synthetic control method: causal inference with interactive fixed effects models. *Polit. Anal.*, **25**, 57–76.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Web-based supplementary materials for "A Bayesian multivariate factor analysis model using observational time-series data on multiple outcomes"'.