Intervention, underdetermination, and theory
generation

Jan-Willem Romeijn
Department of Philosophy, University of Groningen
Jon Williamson
Department of Philosophy, University of Kent

Abstract
We consider the use of intervention data for eliminating the indeterminacy of statistical modelling, and for guiding extensions of the statistical models. The leading example is factor analysis, a major statistical tool in the social sciences. We first relate indeterminacy in factor analysis to the wider problem of underdetermination. Then we draw a parallel between factor analysis models and Bayesian networks with hidden nodes, which allows us to elucidate the use of intervention data for dealing with underdetermination. It will be shown that in some cases, the underdetermination can be resolved by an intervention. In the other cases, the intervention data suggest specific extensions of the model. Intervention data can thus play a role in resolving the problem of underdetermination and in guiding the generation of theory, replacing theoretical criteria often used for these purposes.

1. Introduction. It has often been argued that the problem of underdetermination is a logical curiosity rather than a real life possibility. Indeed, in the natural sciences it is not easy to find empirically equivalent rivals to our best candidate theories. However, we will argue that underdetermination of a particular type is widespread in the social sciences, in particular where these sciences employ statistical modelling. We substantiate this claim with examples from exploratory factor analysis, a widely used statistical modeling tool. Within a given experimental setup or population study, it may very well happen that the statistical model allows for distinctions between
hypotheses that do not correspond to a difference in the likelihood function of the hypotheses. Here the problem of underdetermination appears as the problem that the best fitting hypothesis has a number of equally well fitting rivals. The standard response to this, at least in factor analysis, is to look for theoretical criteria that force a choice between the rivals, such as simplicity or explanatory force. The underdetermination is then resolved by an appeal to theoretical considerations.

In this paper we investigate a different response to underdetermination in statistical modelling. Our main claim is that the resolution of underdetermination may also be driven by further empirical criteria, specifically by intervention data. The particular underdetermination involved is thus relative to what is taken as observable in the given experimental setup. Relative to one experimental setup, the background theory may generate statistical hypotheses that have exactly the same likelihood functions, and thus perform equally well on the observation data. The model of these hypotheses is then called unidentifiable. However, the hypotheses in the model need not be altogether empirically equivalent. We can consider specific changes to the experimental setup, or interventions for short, such that the background theory generates different likelihood functions for the results of the altered experiments. The statistical hypotheses associated with the theory then come apart.

This paper is concerned with the use of interventions for telling apart identical statistical hypotheses. But next to this application of interventions, we also consider the use of interventions for changes to the background theory. It may happen that the data resulting from an experimental intervention dismiss all the statistical hypotheses under consideration. In that case, the background theory fails to generate a likelihood function that can accommodate the data. In the following we show that in such cases, the intervention data suggest a change to the theory, and thereby an extension of the statistical model, to accommodate the experimental data. In other words, whenever the intervention leaves us empty-handed, it directs us towards a revision of the underlying background theory.

The upshot of all this is that aspects of scientific method that are typically associated with theoretical considerations, such as resolving underdetermination and generating new theory, are here seen to be driven by em-
pirical facts. The case study for illustrating these ideas involves exploratory factor analysis. Coincidentally, factor analysis has already made an appearance in the philosophy of science. In Haig (2005) and Schurz (2008), for example, factor analysis is proposed as a model for abductive inference, and thus as a tool for generating new theory. It must be emphasised that in this paper we take an entirely different perspective. We employ exploratory factor analysis as an illustration of the more general problem concerning statistical underdetermination, and we focus on the role of interventions in the resolution of statistical underdetermination. In addition, we show the use of interventions in the generation of new theory. It is very different to take exploratory factor analysis as a model for abductive inference and hence as a tool for theory choice.

The paper is set up in the following way. In section 2 we describe three distinct problems of underdetermination in factor analysis. We show that these problems are typically resolved by imposing further theoretical criteria. Seeing that factor analysis is essentially identical to estimating parameters in a Bayesian network with hidden nodes, we show in section 3 that Bayesian networks, and thus models in factor analysis, allow for incorporating intervention data, and we argue that in specific cases, intervention data can be used to resolve the underdetermination problem. Interventions may therefore take the place of theoretical criteria. In section 4 we argue that in certain other cases, the intervention data guide us towards adding a hidden node of a latent variable in the statistical model. In other words, interventions may guide theory generation. In the section 5, finally, we suggest how the model for intervention and theory generation may prove of some use, both to the philosophy of experiment and to scientific methodology itself.

2. Underdetermination in statistics. In this section we sketch the problem of underdetermination for statistics. This problem is made precise for factor analysis, a well-known statistical technique in psychometrics. We also suggest how theoretical notions like simplicity, causality, and the like can be used to break the underdetermination.

2.1. Underdetermination in statistics. Consider a simple statistical problem, in which we estimate the chance of some event in independent and
identical trials. An observation at time $t + 1$ is denoted by the variable $Q_{t+1}$, with possible values $q^0_{t+1}$ or $q^1_{t+1}$. We denote a series of $t$ observations by $E_t$, and the event that earlier results were some ordered series $\langle q^1_0 q^0_2 q^1_3 \ldots q^1_t \rangle$ by $e^{(010\ldots1)}$, or $e_t$ for short. Denoting the hypothesis that the chance of finding $q^1_t$ is $\theta$ with $h_\theta$, we have

$$ p(q^1_{t+1}|h_\theta \cap e_t) = \theta $$

for each trial $t + 1$, an expression often called the likelihood function of $h_\theta$. We may assume that the chance $\theta$ of the event $q^1_{t+1}$ may be any value in $[0, 1]$. Then on the basis of some series of events $E_t$, we can provide an estimation of $\theta$. We can do so either by defining a prior $p(h_\theta)$ and then computing a posterior by Bayesian conditioning, or by defining an estimator function over the event space, typically $\hat{\theta} = \sum_{i=1}^t I^1_1(Q_i)/t$, in which the indicator $I_1(Q_i) = 1$ if $Q_i$ takes the value $q^1_i$ and 0 otherwise.

Now we may wonder whether there is any sense in which the above estimation problem suffers from an underdetermination problem. Of course, we are dealing with probabilistic relations between observations and hypotheses. It may very well be that the binary results summarised in $E_t$ are really the tosses of a fair coin, so that the true hypothesis is $h_{1/2}$, and that nevertheless, by some unfortunate coincidence, we find seven consecutive tails, $e^{0000000}$, indicating $h_0$ as the best estimation. Even stronger, it is logically possible that the fair coin keeps landing tails until the end of time, and similarly that a coin biased according to $H_{1/3}$ yields an infinitely long series of heads. In fact any infinitely long series of results is in principle consistent with any of the hypotheses $h_\theta$, and in that sense we are encountering an underdetermination problem in the estimation.

However, in another and to our mind more important sense, the above estimation problem is completely unproblematic. It may be noted that the observations have a different bearing on each of the hypotheses in the model, i.e. the set of statistical hypotheses. If there is indeed a true hypothesis in the set, then according to well-known convergence theorems, the probability of assigning a probability 1 to this hypothesis will tend to one. In the limit, we can therefore almost always, in the technical sense of this expression, tell the statistical hypotheses apart.
This situation is entirely different in the following statistical problem. Imagine that some researcher defined a slightly different set of statistical hypotheses, characterised as follows:

$$p(q_{t+1}^1|h_{\xi} \cap e_t) = \xi^2 \quad \xi \in [-1, 1].$$

The set of hypotheses considered in the statistical problem is essentially the same. It is only labelled in a funny way. The hypotheses $h_{1/2}$ and $h_{-1/2}$ are indistinguishable, because they both assign exactly the same probability to all the observations. More generally, we have $p(q_{t+1}^1|h_{\xi} \cap e_t) = p(q_{t+1}^1|h_{-\xi} \cap e_t)$. So how are we ever able to tell apart the pairs of statistical hypotheses $h_{\xi}$ and $h_{-\xi}$? Whereas previously we were able to tell apart the statistical hypotheses because the observations did have a different bearing on them, we are now stuck with pairs of hypotheses that react exactly identically to the observations. In such a case, we speak of an unidentifiable model. Unidentifiable models constitute statistical underdetermination by the observations, different from the underdetermination associated with the standard estimation problem.

2.2. Factor analysis. The above example of statistical underdetermination is rather nonsensical. No reason is given for distinguishing between the regions $\xi > 0$ and $\xi < 0$. However, there are other and more complex cases in which it makes perfect sense to introduce distinctions between hypotheses that do not differ in their likelihood functions. This subsection is devoted to presenting one of these cases, involving so-called exploratory factor analysis. The exposition is partly borrowed from Romeijn (2008).

Exploratory factor analysis is a technique that posits a specific statistical model of latent random variables on the basis of an analysis of the correlational structure of observed random variables. See Lawley and Maxwell (1971) for a classical statistical overview, Mulaik (1985) for a philosophically-minded discussion, and Bartholomew and Knott (1999) for a very insightful introduction from a Bayesian perspective. In fact all these treatises introduce exploratory factor analysis next to the much less problematic statistical tool of confirmatory factor analysis. In most of the following we concentrate on the former, and simply call it factor analysis. Confirmatory factor analysis makes a modest reappearance in Section 4.
Say that in some experimental setting we observe the levels of fear $F$ and loathing $L$ in a number of individuals indexed $i$, and we find a positive correlation between these two variables, $p(F_i, L_i) \neq p(F_i)p(L_i)$. One way of accounting for the correlation is by positing a statistical model over the variables in which fear and loathing may be correlated directly, and then estimate the parameters in the model. But we may feel that this model is does not capture the causal or mechanistic details of the experimental setup. It may be that it is neither the loathing that instills fear in people, nor the fear that invites loathing, but rather that both these feelings are caused by a drug $D_i$ that is administered to the subjects in the experiment. The correct statistical model, we may argue, posits a correlation between the drug and the fear, and similarly a correlation between the drug and the loathing, while conditional on a certain drug dosage, fear and loathing are uncorrelated: $p(F_i, L_i, D_i) = p(D_i)p(F_i|D_i)p(L_i|D_i)$. We then say that the drug dosage is the common factor to the observed variables of fear and loathing. The correlations between drug dosage and fear and loathing respectively we call the factor loadings.

Factor analysis has a number of standard applications, which are usually subdivided according to whether the observed and latent variables are categorical or continuous. In this paper we discuss one of the most straightforward applications of factor analysis, in which both the observed and latent variables are binary. In the example, the drug is either present in subject $i,$ $d^1_i,$ or absent, $d^0_i,$ and similarly for fear and loathing. We assume that the probabilistic relations between the variables are independent and identically distributed. Out of the many possible probabilistic dependencies between $F_i,$ $L_i$ and $D_i,$ we thus confine ourselves to

$$p(f^1_i|d^j_i) = \phi_j,$$  \hspace{1cm} (2)

$$p(l^1_i|d^j_i) = \lambda_j,$$  \hspace{1cm} (3)

for $j = 0, 1,$ a conditional version of the Bernoulli model of Equation (1). Similarly for the variables $D_i,$

$$p(d^j_i) = \delta$$  \hspace{1cm} (4)

The probability over the variables $D_i,$ $L_i$ and $F_i$ is thus given by five Bernoulli distributions, each characterised by a single chance parameter.
We want to emphasise that this model is much simpler than what is typical in factor analysis. In many applications the variables are not binary but continuous, the probabilistic relations between the variables are linear regressions with normal errors, and the variable $D$ is assumed to be normally distributed as well. Writing $F_i = x$ for the event that the level of fear is $x \in \mathbb{R}$, and similarly for the drug dosage $D_i = y$, the relation between $F_i$ and $D_i$, for example, is

$$p(F_i = y|D_i = x) = \mathcal{N}(\lambda_F x, \epsilon_F)$$

in which $\mathcal{N}(\lambda x, \epsilon)$ is a distribution over the values $y$ of $F_i$. Thus the relation between the variables $D_i$ and $F_i$ is characterised by a richer family of distributions, parameterised by a regression parameter $\lambda_F$ and an error $\epsilon_F$. However, the purpose of this paper is to illustrate the underdetermination problems in factor analysis. For this aim, the simpler factor model of Equations (2) to and (4) suffices. The crucial characteristic in all of what follows is that there are latent variables explaining the correlational structure among the observed variables.

Now in a standard experimental setting, we can observe the common factor of whether the drug has been administered. But in situations in which the causal or mechanistic story is unknown, we may nevertheless want to posit such an underlying story. For example, recurring feelings of fear and loathing may be two of a large number of variables on negative emotions, used to describe individuals in a general population. Now if all these variables are strongly positively correlated, it may be that we can account for all the correlations in a statistical model positing a fairly small number of common factors, or even a single common factor such as depression. Exploratory factor analysis is a technique for doing the latter in a systematic way. When given a set of correlations among observed variables, it produces a statistical model of latent common factors that can account for these correlations and which, given specific values of the latent common factors, leaves the observed variables uncorrelated.

It will not be surprising that applications of factor analysis suffer from problems of underdetermination. After all, it is positing a theoretical structure, namely an unobservable common cause, over and above the observational facts, namely the correlations between observable variables. For
one thing, when explaining more complex correlational structures there will generally be a large number of latent common factor models of variable complexity which will fit the data to variable degrees, and so there will have to be a trade-off between goodness of fit and model simplicity. In other words, exploratory factor analysis suffers from the type of underdetermination associated with model selection. But this need not surprise us too much: all statistical modeling must at some point address this worry. We will argue in section 3 that this type of underdetermination, insofar as it concerns causal structure, can be dealt with by turning to interventions as well.

However, it turns out that even if the modelling choices have been made and the common factor model is given, underdetermination problems may appear. These underdetermination problems are associated with unidentifiable models, as discussed in Section 2.1.

2.3. Underdetermination in factor analysis. The statistical underdetermination problems inherent to a given factor model, such as the models introduced above, come in two different types. The second problem is mentioned here because it has been hotly debated in psychological methodology. But the first is much more important to our present concerns.

This first underdetermination is essentially the problem discussed in the foregoing. It is based on the fact that the model contains sets of statistical hypotheses that share the same likelihood function. Specifically, consider the factor model of Equations (2) to Equations (4). For a quick understanding of the problem, focus on the dimensions of the model. In total we count a number of 5 parameters, namely $\delta$, and $\phi_j$ and $\lambda_j$ for $j = 0, 1$.

On the other hand, we have observations $F_i$ and $L_i$ that can be used to determine these parameters. But because we are using Bernoulli hypotheses, only the observed relative frequencies of the possible combinations of $F_i$ and $L_i$ matter, irrespectively of how many subjects $i$ have been investigated. And because we have 4 possible combinations of $F_i$ and $L_i$, whose relative frequencies must add up to 1, we have only 3 frequencies to determine the 5 parameters in the model. After having used the observations in the determination of the parameters, therefore, we still have 2 degrees of freedom left. Hence the values of the parameters in the model cannot be determined by the observations uniquely.
We state this problem more mathematically by looking at the likelihoods for the observations of possible combinations of $F_i$ and $L_i$. Abbreviating $\theta = \langle \delta, \phi_0, \phi_1, \lambda_0, \lambda_1 \rangle$, we have

\begin{align*}
p(f_0^i \land l_1^i | h_\theta) &\equiv \eta_{01} = \delta(1 - \phi_1)\lambda_1 + (1 - \delta)(1 - \phi_0)\lambda_0, \quad (6) \\
p(f_1^i \land l_0^i | h_\theta) &\equiv \eta_{10} = \delta\phi_1(1 - \lambda_1) + (1 - \delta)\phi_0(1 - \lambda_0), \quad (7) \\
p(f_1^i \land l_1^i | h_\theta) &\equiv \eta_{11} = \delta\phi_1\lambda_1 + (1 - \delta)\phi_0\lambda_0, \quad (8)
\end{align*}

where the $\eta_{jk}$ are shorthands. The fourth likelihood, $p(f_0^i \land l_0^i | h_\theta)$, can be derived from these expressions. The salient point is that the system of equations resulting from filling in particular values for the likelihoods $\eta_{jk}$ has infinitely many solutions in terms of the components of $\theta$: for any value of the likelihoods $\eta_{jk}$, the space of solutions in $\theta$ has 2 dimensions. Conversely, different hypotheses $h_\theta$ will have the same set of likelihoods $\eta_{jk}$ for the observations. In a Bayesian analysis, the hypotheses $h_\theta$ that are associated with the same likelihoods $\eta_{jk}$ cannot be told apart, in the same way as that the hypotheses $h_\xi$ and $h_{-\xi}$ cannot be told apart by the observations.

The fact that hypotheses cannot be told apart shows up in the shape of the posterior distribution over the hypotheses. Say that we have observed the relative frequencies

$$r_{jk}(e_t) = \frac{1}{7} \sum_{i=1}^{t} I^i(F_i)I^k(L_i),$$

where the indicators $I^i(F_i) = 1$ if $e_t \subset F_i^1$ and 0 otherwise, and $I^k(L_i)$ analogously. By means of the likelihoods given in Equations (6) to (8) we can then determine a posterior probability for the hypotheses in the model by means of Bayesian conditioning:

$$p(h_\theta | e_t) \propto p(h_\theta)p(e_t | h_\theta) = p(h_\theta) \prod_{jk} r_{jk}(e_t).$$

We can choose the likelihoods $\eta_{jk}$ such that the overall likelihood $p(e_t | h_\theta)$ is maximal, namely by setting $\eta_{jk} = r_{jk}(e_t)$. But there are infinitely many hypotheses $h_\theta$ that have these likelihoods $\eta_{jk}$. Consequently, there is no unique hypothesis $h_\theta$ that has maximal overall likelihood $p(e_t | h_\theta)$, and within the
set of hypotheses with maximal likelihood, the shape of the posterior is simply proportional of the shape of the prior.

From the posterior distribution over the hypotheses we can generate estimations of the parameters in $\theta$, according to

$$\hat{\theta} = \int_{[0,1]^5} \theta p(h_\theta | e_t) d\theta$$

These estimations will also suffer from the fact that the hypotheses cannot be told apart. The results of the estimations will depend on the prior probability over the hypotheses. Of course, this is usually the case in a Bayesian analysis. What is more troublesome is that this dependence can be completely misleading, and that no amount of additional data can eliminate this dependence of the estimations on the prior.

The same kind of underdetermination also occurs in more complicated models using linear regressions as in Equation (5). But in such models it takes a slightly different shape. Note first that we can extend factor models to include any number of common factors. But once a model includes more than one common factor, we find that the factor loadings are not completely determined. Say, for example, that we analyse fear $F$, loathing $L$, and sleeplessness $S$ in terms of two common factors, depression $D$ and manic disposition $M$. Every individual is supposed to occupy a specific position in the $D \times M$ surface. However, we might feel that a more natural way of understanding the surface of latent variables is by labelling the states in this surface differently, for example by introducing a linear combination of $D$ and $M$, calling it bi-polarity, and further introducing another coordinate that is perpendicular to it, perhaps calling it a neurotic disposition. More generally, the factors in a model may be linearly combined or, in more spatial terms, rotated to form any new pair of factors.

The underdetermination problem with this is that, if we allow the latent factors to be correlated, any rotation of factors will perform equally well on the estimation criterion, be it maximum likelihood, generalized least squares, or similar. This problem is thus known as the problem of the rotation of factor scores. The estimation criterion or Bayesian conditioning on the data do not lead to a single best hypothesis in the factor model, but to a collection of them, meaning that the factor model is again unidentifiable, with all the attached problems as listed above. However, in this paper we will
not elaborate the mathematical details of underdetermination in these more complicated models. The simpler, discrete model suffices for the purpose of this paper.

Instead we aim for a philosophical understanding of the underdetermination at stake. A standard reaction to the problem with rotation is to adopt the theoretical criterion that the latent variables must be independent. In that case, we cannot freely rotate the axes in the space of latent variables anymore. The parameterisation of the space must be such that there are no correlations between the latent variables. There are, however, alternative theoretical criteria for choosing the parameterisation of the space of latent variables. For example, it may be interesting to have maximal variation among the factor loadings which, intuitively, comes down to coupling each latent variable with a distinct subset of observable variables. The thing to note is that, from the point of view of statistics, the choice for how to parameterise the space of latent variables is completely underdetermined: we cannot decide between these parameterisations on the basis of the observations. The causal and nomic structure of the factors underlying the correlations, and the conceptual structure that goes with it, therefore remain obscure.

2.4. Factor score indeterminacy. Quite apart from the foregoing, there is another problem with factor analysis that can be framed as underdetermination. See Steiger (1979) for some historical context, Maraun (1996) for a philosophical evaluation, McDonald (1974) for an excellent classical statistical discussion, and Bartholomew and Knott (1999) for a Bayesian account of it.

Say that we have rotated the factors to meet the theoretical criterion of our choice, for instance by assuming a single common factor or by fixing the independence of the latent factors. Can we then reconstruct the latent variable itself, that is, can we provide a labeling in which each individual, i.e. each valuation of the observable variables, is assigned a determinate expected latent score? Sadly, the classical statistical answer here is negative. We still have to deal with the so-called indeterminacy of factor scores, meaning that there is a variety of ways in which we can organize the allocation of the individuals on the latent scores, all of them perfectly consistent with
the estimations. There are strong restrictions to this allocation, most importantly that the ordinal structure of the factor scores must be preserved. However, quite apart from scaling conventions, the metrical structure of factor scores cannot be determined.

The type of underdetermination presented by factor score indeterminacy depends on what we take to be the statistical inference underlying factor analysis. In the context of this paper, we take the factor analysis model to specify a complete probability assignment over the latent and observed variables, including a prior probability over the latent variables. As explained in Bartholomew and Knott (1999), factor score indeterminacy is thereby eliminated, as long as there are sufficiently many observed variables that are related to the latent variables according to distributions of a suitable form, namely from the exponential family. In this paper we will employ Bayesian statistics, and we will therefore ignore most of the discussion on factor score indeterminacy.

There is, however, a point at which the problem of factor score indeterminacy enters the present discussion. Depending on the specifics of the model, a transformation in the space of latent variables involves transforming the prior probability over the latent variables. The problem concerning the parameterisation of the latent variables is thus related to the problem of choosing a prior probability, and therefore also to the problem of factor score indeterminacy. More specifically, we will argue in the following that the use of intervention data, which resolves the problem of underdetermination discussed above, provides a new perspective on the problem of the indeterminacy of factor scores as well.

This completes the illustration of statistical underdetermination in terms of factor analysis. Perhaps the main reason for the illustration is to show that underdetermination is not merely an academic problem: factor analysis is routinely used to interpret psychological test data, and it is a live problem that the data do not allow for a full determination of the structure of the underlying factors. In the next section we will try to get to the heart of the statistical underdetermination, as exhibited in factor analysis, by discussing it in the setting of Bayesian networks. This will suggest how we might resolve underdetermination by means of intervention data.
3. Interventions and Bayesian networks. In the foregoing we have argued that factor analysis suffers from statistical underdetermination. In this section we will explain the underdetermination inherent to factor analysis by identifying analogous problems in the estimation of parameters in Bayesian networks. This leads us to consider a specific solution to the underdetermination problem, namely by means of intervention data. We first introduce Bayesian networks, the notion of intervention, and its use in estimating Bayesian networks, after which we discuss its application to factor analysis.

3.1. Bayesian networks and factor analysis. In general, a Bayesian network consists of a directed acyclic graph on the variables of interest $V_1, \ldots, V_n$, together with the probability distribution $p(V_i | Par_i)$ of each variable conditional on its parents in the graph. The graph is related to probability by an assumption known as the Markov Condition: each variable is probabilistically independent of its non-descendants in the graph, conditional on its parents, written $V_i \perp \perp ND_i | Par_i$. Under this assumption the network suffices to determine the joint probability distribution over the variables, via the identity

$$p(v_{i1}^j \cdots v_{in}^j) = \prod_{i=1}^{n} p(v_{i}^j | par_i)$$

(10)

where $par_i$ is the assignment of values to the parents of $V_i$ that is induced by the assignment $v_{i1}^j \cdots v_{in}^j$ of values to the whole domain.

Effectively the introduction of the factor analysis for binary variables was already an introduction to a specific class of Bayesian networks. First, we assume that the same probability assignment describes all subjects,

$$p(F_i, L_i, D_i) = p(F'_i, L'_i, D'_i),$$

(11)

so that for convenience we can omit the subscript $i$. For each subject $i$ the factor analysis determines a probability function $p(F, L, D)$ that observes a specific symmetry: conditional on the latent variable $D$ there is no correlation between the observed variables $F$ and $L$,

$$p(F, L, D) = p(D)p(F|D)p(L|D).$$

(12)
Figure 1: The graphical structure representing the independence relations in a factor analysis of drugs, fear and loathing.

On the basis of this we build a network, with the variables $F$, $L$ and $D$ as nodes. Quite apart from the exact probability values, the probability function determined by factor analysis can thus be represented in the Bayesian network depicted in figure 1.

There are also differences between the theory of Bayesian networks and factor analysis. For one, factor analysis entails a rather specific network structure: there are latent parent nodes, observable child nodes, there are typically fewer parents than children, any child can be connected to any parent, and vice versa. Furthermore, applications of the former are usually restricted to probability functions over finite or at least countable domains. Nodes with continuous domains are not that commonly discussed, although they have been studied in the context of structural equations models, for example in Pearl (2000) and, from the side of latent variable modeling, in von Eye and Clogg (1994). A related difference is that in most applications of factor analysis the probability functions that are considered are restricted to normal distributions over latent nodes, and to linear regressions with normal errors between latent and observable nodes. Applications of Bayesian networks are typically (but not necessarily) restricted to Bernoulli distributions.

In this paper we approach factor analysis more from the angle of Bayesian networks, using the framework for Bayesian inference over Bayesian networks presented in Romeijn et al. (2006). Hence the statistical underdetermination presented in Section 2.3 is framed as a problem to do with determining the posterior probability distribution over the parameters that characterise the Bayesian network of Figure 1. As announced, we are going to resolve this
3.2. Interventions. A causally interpreted Bayesian network, or causal net for short, is a Bayesian network where the graph is interpreted as a causal graph. That is, each arrow in the graph is interpreted as denoting a direct causal relationship from the parent variable to the child variable. Under this interpretation, the Markov Condition is called the Causal Markov Condition and says that each variable is probabilistically independent of its non-effects conditional on its direct causes. It is often assumed that the Causal Markov Condition is bound to hold if the graph in the net is correct and is closed under common causes (any common causes of variables in the net are also included in the net). While there are situations in which the condition is implausible, it can be justified as a default assumption (Williamson, 2005), and we shall take it for granted here.

Causal nets are helpful for predicting the effects of interventions. When an experimenter intervenes to fix the value of a variable, she interrupts the normal course of affairs and sets the variable exogeneously. The usual mechanisms, according to which the variable is determined, are thereby replaced with new mechanisms, according to which the variable is determined only by the experimenter. An ideal or divine intervention is one in which the intervention only changes the intended variable, without changing other variables under consideration and without changing other causal relationships under consideration. We write \( p(v'_i \| v'_k) \) to signify the probability that variable \( V_i \) takes its \( j \)'th value after an ideal intervention has been performed that sets \( V_k \) to its \( l \)'th value. We then have the following connections between probability and a causal net. For assignments \( x, y, z \) of values to distinct sets \( X, Y, Z \) of variables \( V_i \),

- \( p(x) \) is determined from the causal net via Equation 10,
- \( p(x|y) \) is determined from the causal net by setting \( p(x|y) = p(xy)/p(y) \) where the numerator and denominator are obtained via Equation 10,
- \( p(x||z) \) is determined from the causal net by first forming a new net by deleting arrows into variables in \( z \), and then calculating \( p(x|z) \) in the new net.
• $p(x|y;z)$ is determined by a combination of these methods.$^1$

In fact causal nets can handle a rather more general notion of intervention. We can write $p(v^j_i) = s \parallel p(v^t_k) = t$ to say that there is probability $s$ that variable $V_i$ takes its $j$'th value when an ideal intervention has been performed to set the probability of $v^t_k$ to be $t$. This is the case iff, when the causal net is transformed by eliminating arrows into $V_k$ and setting its unconditional distribution to $p(v^t_k) = t$, the new causal net deems that $p(v^j_i) = s$. This kind of intervention is sometimes called an imperfect intervention or a stochastic intervention, to distinguish it from the divine interventions considered above.$^2$ A stochastic intervention is itself a special case of another kind of intervention—called a parametric intervention—where, instead of intervening to fix the effect variable, one intervenes to change how the causes impact on the effect variable. Thus $p(v^j_i) = s \parallel p'(V_k|Par_k)$ says that there is probability $s$ that variable $V_i$ takes its $j$'th value after intervening to change the distribution of $V_k$ conditional on its direct causes $Par_k$ to $p'$. The probability $s$ is calculated from the causal net after substituting $p'(V_k|Par_k)$ for $p(V_k|Par_k)$. See Korb et al. (2004) and Eberhardt and Scheines (2007) for discussion of these kinds of intervention.

Interventions can help with underdetermination in two ways. First, they can help with underdetermination of causal structure. If more than one causal structure is compatible with evidence, one can intervene and collect more evidence: this new evidence may decide between the causal structures. To take the example presented in the foregoing, suppose variables $F$, $L$ and $D$ are all measured, and that the resulting data shows that $F$ and $L$ are probabilistically independent conditional on $D$, $F \perp \perp L \mid D$. This evidence is compatible with the causal graph of Figure 1, but equally with Figures 2 and 3. The evidence can be used to fill in the conditional probability distributions on these causal models, but can not decide between them. An

$^1$In fact there are more efficient algorithms for calculating these quantities, but the above methods suffice to show the link between probabilities, conditional probabilities and interventional probabilities.

$^2$Note that the causal net and the transformed causal net determine different probability distributions, so that the function $p$ on the left-hand side of $p(v^j_i) = s \parallel p(v^t_k) = t$ is different to the function $p$ on the right-hand side.
intervention can decide between them, however. If, after intervening to change the distribution of $D$, the distribution of $F$ and $L$ is changed, then that favours Figure 1. Otherwise if only the distribution of $L$ is changed after intervention, then Figure 2 is supported, and if only the distribution of $F$ is changed then Figure 3 is supported.

While resolving underdetermination of causal structure is the main application of interventions in the literature, interventions can also be used to resolve underdetermination of the parameters in a causal net. In this case, suppose that the causal structure is known and evidence is collected which determines the probability distributions of some variables conditional on their parents, but which does not fully determine conditional distributions that attach to other variables. By carrying out an ideal intervention, an experimenter effectively changes the conditional distribution of one variable without changing the distributions of other variables. The data obtained after the intervention can then be used in conjunction with the old data to further constrain the values of the underdetermined distributions.

3.3. Interventions and underdetermination. Now that we have discussed interventions in Bayesian networks, we show how exactly interventions can be used to further constrain the underdetermined distributions, and thus solve the underdetermination introduced in section 2.3. We first consider the example of depression, fear, and loathing, after which we sketch how the idea can be extended to factor analysis more generally.

Let us first explain the basic idea of using interventions for the purpose of solving underdetermination. First, an intervention on the subjects is assumed to change the distribution over the latent variables of the subjects,
and not the probabilistic relations between the latent and the observed variables. Second, note that after the intervention we obtain an entirely new estimation problem for the parameters in the Bayesian network. But because the data are obtained by intervention, we can assume that the parameters associated with the relations between latent and observed variables do not change. To accommodate the intervention data, we therefore have a smaller space of parameters available. In the following we show that, depending on the model, intervention data can thus be used to select a unique best estimate for the parameter values in the factor model.

Consider again the model characterised by Equations (2) to (4), (11) and (12). As explained in the foregoing, an intervention is an external shift to the probability assignment. In this particular case, we intervene on the node $D$, meaning that we change the probability of depression, $p(d^1) = \delta$, to a new value, $p'(d^1) = \delta'$. The relations of the depression variable to the variables of fear and loathing, however, are not changed by the intervention, $p'(f^1|d^1) = \phi$ and $p(l^1|d^1) = \lambda$. Finally, after the intervention we record the observations $e'_t$, and in particular the relative frequencies of $f_j^i$ and $l_k^i$, the fractions $r'_{jk}(e'_t)$ defined in analogy to $r_{jk}(e_t)$.

To get the point of this across quickly, we focus again on the dimensions of the model. This time we count a number of six parameters, namely $\delta, \phi_j$ and $\lambda_j$ for $j = 0, 1$, and finally $\delta'$. On the other hand, we have a richer set of observations that can be used to determine these parameters. Specifically, we have three observed relative frequencies of $f_j^i \land l_k^i$ before intervention, $r_{jk}(e_t)$, and three of them after intervention, $r'_{jk}(e'_t)$, so six in total. It thus appears that we can determine a unique solution. Whereas previously we had two degrees of freedom left after the incorporation of the data, together with the intervention data we can fill in all the parameter values of the factor model.

We can make this precise mathematically by once again looking at the likelihoods. As before, we have the likelihoods of Equations (6) to (8). But to these expressions we now add the likelihoods of the hypotheses after the
The point to note is that the system of equations resulting from filling in any set of values for the likelihoods $\eta_{jk}$ and $\eta'_{jk}$ has a unique solution in terms of the components of $\theta$ and the additional parameter $\delta'$. Conversely, every hypothesis $h_\theta$ is associated with a unique set of likelihoods $\eta_{jk}$ and $\eta'_{jk}$. By adding the intervention data, therefore, we can tell apart all the hypotheses in the factor model, in the same way as we were able to tell apart $h_\theta$ in the model of Equation (1).

Intervention data can thus be used to resolve statistical underdetermination. The posterior distribution will generally have a unique maximum after Bayesian conditioning on the normal and the intervention data. And accordingly, the parameter estimations will in the long run be independent of the prior distribution over the hypotheses. Empirical criteria for theory evaluation, based on the targeted acquisition of intervention data, can in this way take the place of the theoretical criteria that guide theory choice in the face of underdetermination.

The latter point is of particular interest in more advanced uses of factor analysis. Recall the problem of underdetermination due to the rotation of latent variables, as discussed in Section 2.3. The standard response to this underdetermination is to employ theoretical criteria on the latent variables, for example by supposing that they are independent, or by choosing the latent variables such that the regression parameters show maximal variation. The idea here is that these theoretical criteria can be replaced by intervention data. However, we leave the mathematical details of this to another paper, because it addresses an audience of psychologists and statisticians rather than philosophers. For present purposes we want to emphasise that within statistics, the problem of underdetermination seems to have fuzzy edges: it can be resolved by an appeal to theoretical criteria, but it can also be resolved by extending the realm of observations with intervention data.
Figure 4: The probabilistic independence relations between drugs, fear, loathing and sleeplessness.

4. Interventions and invention. In the foregoing we have shown how interventions can be used to resolve underdetermination in factor models. In the present section we go one step further. It may so happen that the intervention data overdetermine the factor model at hand. In that case the model may start to look inadequate not because it leaves parameters free, but rather because it fails to accommodate, or explain, the correlational structures in the data. We show that the overdetermination and poor fit of a factor model after intervention may lead to controlled changes to the model. In other words, interventions may guide theory change.

4.1. Model change. We first need a model in which the overdetermination after intervention can occur. To this aim, consider the factor model alluded to earlier, which analyses the observed variables fear $F$, loathing $L$, and sleeplessness $S$ in terms of one common factor, depression $D$. Figure 4 shows the network associated with the factor model. In total we count a number of 7 free parameters in the model, and also 7 correlations between the observable variables, so the model is fully determined. Now imagine that we want to see the effect of a treatment on depression. We first observe a number of subjects on the variables $F$, $L$, and $S$, after which we treat the subjects and observe them again.

The idea of the treatment is of course that it only interferes with the probability of depression. After the intervention we have $p'(d_1^j) = \delta'$, while $p'$ is unchanged for all the other variables. Note, however, that after the intervention we have a total of 14 relative frequencies to accommodate, for which we have only 8 variables. This means that the model is overdeter-
mined. Naturally, it is possible that all the 14 frequencies can be accounted for by the parameters of the model, or that there is a unique set of parameter values that approximates the relative frequencies fairly well. In that case the factor model is used in a confirmatory role, thereby entering the domain of confirmatory factor analysis alluded to in Section 2.2. There are more data than can be accommodated by the model, and the fact that the model nevertheless shows a good fit with the data confirms the factor structure it imposes on the data.

On the other hand, it may happen that the fit with the intervention data is poor, according to some model selection or fit criterion. For example, we may estimate the parameters as indicated and then check the predictive performance of the estimations on the data set at hand, by the percentage success of the predictions, or some other notion of distance to perfect or optimal prediction. If the estimations in the given model fall short of the criterion set for predictive performance, then we deem the fit poor. For a recent discussion of model fit, see Waldorp and Wagenmakers (2006).

The aim of the present section is to provide a procedure for improving a model after intervention data have led to poor model fit. In other words, we will sketch a procedure for changing the model that is guided by intervention data. We will do so by considering ways of adding latent variables to the model, or in terms of Bayesian networks, adding hidden nodes and edges to observed variables to the network. We thereby effectively enlarge the parameter space, and thus the space of probability assignments over the variables before and after the intervention. As will be argued, the intervention data suggest specific ways of extending the space of probability assignments, thereby suggesting specific ways of changing the Bayesian network that represents the model.

4.2. Network dynamics. In the case in which overdetermination disconfirms a particular causal net, there are various ways of modifying the network in order to reconcile the disconfirming data. Various strategies are available for generating a new causal hypothesis, including starting from scratch, adding arrows, adding values to variables and adding common causes.

Clearly, when faced with an overdetermined model we can discard the old net and start from scratch, taking the new causal net to be the Bayesian
net on the same set of variables that fits the entirety of the data best. This is the typical machine learning approach (see, e.g., Neapolitan, 2003; Spirtes et al., 1993) and it is guaranteed to work because, given a set of variables, there will always be some net on those variables that determines the distribution of the data involving those variables. However, often there is evidence that supports the relations posited in the current net, such as evidence of mechanisms linking causes and effects, so this approach is liable to throw the baby out with the bath water.

An alternative approach involves adding arrows to the existing network to better fit the data. This approach is also guaranteed to work (Williamson, 2005, Chapter 3). But in this case causal relations between the measured variables may simply not be plausible. There may be no plausible physical mechanism that makes fear responsible for loathing, for instance. In fact, in the factor models, the observable variables are almost always indicators that cannot themselves play any causal role. Hence, additional arrows will have to obtain between the latent and the observed variables.

There are methods for handling overdetermination that retain the existing causal relations and avoid positing causal relations from one measured variable to another. One such method is the method of increasing the range of values that a common cause can take. Consider the case of overdetermination sketched in the foregoing: there are three measured binary variables \( F, L, S \), and an unmeasured common cause of these variables, the binary variable \( D \), as depicted in Figure 4. Now suppose that the data involving \( F, L, S \) before and after intervention do not fit well with the structure of the network. One way of increasing the dimensionality of the space of parameters associated with the network then is to increase the number of values \( D \) can take. Specifically, if \( D \) becomes a three-valued variable then the parameter space has 15 dimensions, and can thus be made to fit the data.

This method must be used with caution though. It is rarely plausible to suppose that the range of values of a variable can be extended in an unconstrained way. Rather, changing a variable from a binary variable to an \( n \)-ary variable for \( n > 2 \) would indicate a change from categorical measurement, \( D = \text{present} / \text{absent} \), to numerical measurement (\( D = 0, 1, 2, 3 \)), and it would usually be expected that the effects of the common cause would
monotonically increase or decrease in line with its value; increased depression would lead to increased fear, loathing and sleeplessness in our example. One can handle these kinds of presumptions within the causal net framework as inequality constraints on the space of parameters of the network. Setting these constraints requires an interpretation of the common cause variable, if only to know whether an effect increases or decreases with an increase in the value of the cause. Moreover, while constraints of the latter kind do not necessarily reduce the dimensionality of the space of parameters, they may be such that the data are still overdetermined by the correlational structure. See Klugkist et al. (2005) for more on inequality-constrained modelling.

There is a second method for handling overdetermination while keeping the existing common factors and their causal relations to observed variables, and avoiding the addition of arrows between the measured variables: the method of introducing new common causes, or new common factors, to the factor model (Kwoh and Gillies, 1996; Binder et al., 1997). This method is completely in line with exploratory factor analysis, as is illustrated by the example on fear and loathing. Suppose the existing model posits that the variables \( F \) and \( L \) are not causally related and have no common factor, as depicted in Figure 5. In this model fear and loathing are probabilistically independent, \( F \perp \perp L \), so that the model has two independent parameters \( \phi \) and \( \lambda \). Now say that the data shows a dependence between \( F \) and \( L \), leaving these two parameters overdetermined. The whole idea of factor analysis is that we then posit an unmeasured common cause, the binary variable \( D \), as in Figure 1, extending the parameter space to account for the dependence. In terms of dimensionality, the data have three dimensions, the old model has two dimensions, and the new model will have five dimensions. It can thus be made to fit the data.
Figure 6: A new common cause as a cause of $D$.

Figure 7: A template for the common cause strategy.

Note that there are various ways of adding new hidden causes, or common factors, and thereby extending the space of parameters, but that not all of them will render the model compatible with the data. Suppose the current causal net is based around the graph of Figure 4. One can add a new common cause $C$ as a cause of $D$, as depicted in Figure 6. However, if the current net is incompatible with data the new net will be too.

Instead we have in mind a common factor addition that yields a subgraph of the template in Figure 7: the nodes and arrows from Figure 4 are held fixed, the node $C$ is added, as well as a selection of the arrows running from $C$. The undirected edge in the template indicates that there may be an arrow from $C$ to $D$, or vice versa, or even a further common cause $E$ of the two variables. In part, the new model will be dictated by the theory that describes the physical mechanisms underlying the various causal relations. But the point here is that it will also be determined by the interventional data itself. On the basis of the specific intervention data that presents the overdetermination of the model, we can decide what subgraph of the template is the best new model.
4.3. Determining the new network. Having settled on changing the network by adding hidden nodes, we now give a global algorithm for determining the new network. Generally, evidence of the physical mechanisms will impose constraints on the structure of the causal graph, as will the evidence of the observed relative frequencies, including the frequencies after intervention. For present concerns, the key point is that the decision to adapt the causal network in a particular way can be determined in part by empirical data. The modification of a causal network, such as a factor model, thus exemplifies a theory change that is motivated by empirical fact.

Recall that there are two problems discussed in this paper. First, there is the statistical problem discussed in previous sections: given observations of relative frequencies, both before and after interventions, and a causal graph, we can determine the parameter values that yields a distribution that fits the data. Second, there is the structural problem discussed in this section: given an initial causal graph whose corresponding net does not fit the observed frequencies sufficiently well, we can modify that graph to yield a network that does fit the data. Assuming that a solution is available to the first problem, the following algorithm applies that solution to solve the second problem:

Algorithm 4.1 (Network Dynamics)

**Input:** A set $\chi$ of structural constraints imposed by mechanistic evidence; an initial causal graph $\mathcal{G}$; and frequency data, both before and after interventions.

- While the model generated by the graph $\mathcal{G}$ does not adequately fit the data:
  - Determine structural constraints $\chi'$ imposed by interventional data.
  - If possible, add the arrow to $\mathcal{G}$—from those compatible with $\chi, \chi'$ and acyclicity constraints—which yields the parameters that fits the data best.$^3$

---

$^3$If there is more than one optimal arrow, one can spawn a new graph for each such
– Otherwise add a new (common cause) node to \( G \); add to \( \chi \) constraints to ensure that future graphs fit the template of Figure 7.

**Output:** A modified causal graph \( G \).

Note that this algorithm uses a greedy search—it adds arrows and nodes one at a time—to find a network that fits the data sufficiently well. The output network is not guaranteed to be the smallest net that fits the data adequately. Implicitly, though, simplicity considerations underlie this approach, because smaller nets are searched first.

Some remarks on this algorithm are in order. Most importantly, we have not yet filled in how to choose the arrow that yields the best fit with the data. In order to choose, we might consider a set of extended graphs, each resulting from adding some arrow to the network, construct the models that are associated with these extended graphs, and then see which of the models has the highest marginal likelihood on the data. Alternatively, for each model we might select the best estimate within that models, and then compare the likelihoods of these estimates. Yet another approach to the problem of choosing the addition of arrows takes into consideration sets of possible arrows, and then works through a model selection procedure. This latter approach is called for if we want to compare changes to the network that are associated with extended models that differ in dimensionality.

Note that the addition of an arrow will most likely also lead to changes in estimation of parameter values that were already included in the old model. Next to the criteria for choosing among the possible additional arrows mentioned above, we might consider the criterion that the new parameter values must not deviate too far from the old ones. In particular, we can introduce, for each extension of the old model with a certain arrow, a penalty term that is proportional to the relative entropy distance between the old and the new best estimate. Of course, other distance measures may be chosen for this purpose as well. By including such a penalty, we express that the optimal arrow. The resulting set of graphs can be pruned at a later stage by eliminating those that are no longer optimal after further structural changes.
old parameter estimations still have a pull on what we take to be the correct value for the new model.

Independently of what criterion is chosen when deciding which of the arrows to include in the new causal network, it is important to get clear on the exact status of the intervention. Going back to the example of Figure 7, once the factor $C$ is added, it is not clearcut that the intervention was setting the value of $D$. Instead it may have set the value of $C$, thereby leading to changes in the probabilities of the observed variables. In addition, we must get clear on the relation between the variables $C$ and $D$ before we can frame the results of the intervention on the observed variables. If we intervene on $D$ and take $C$ to be causally dependent on $D$, then fitting the intervention data on some extended model, arrived at by the algorithm above, leads to different constraints on the model parameters than if we take the intervention on $D$ to leave $C$ unaffected.

Because of all these uncertainties on the level of the latent variables, it might seem necessary to reconsider the notion of intervention that we employ in framing the intervention data. If, for instance, we provide the subjects with a treatment in the hope of alleviating their depression, we cannot be sure that this intervention does not intervene on both variables $C$ and $D$, and this severely restricts the application of the foregoing. On the other hand, to obtain intervention data one does not have to know beforehand what the results of the intervention on the variables $C$ and $D$ is. To some extend, we can decide on how to frame the intervention after the results of it have been recorded.

Depending on what criterion is chosen to decide on the addition of arrows, and on how the issues raised here are resolved, we have devised a way to adapt a causal graph in the light of intervention data, when these data overdetermine the statistical model at hand. By adapting the graph in this way, we are making controlled changes to the statistical model, introducing new variables and adding causal relations. Thereby we are changing the theory underlying the statistical model, based on intervention data. Admittedly, we have only provided a very rough sketch of this procedure in the foregoing. Much is left for future research. But we hope to have shown, minimally, that intervention data can play an important role in theory change.
5. Directions for further research. In this paper we have investigated the use of interventions for two separate problems in the methodology of science. The first of these is the problem of statistical underdetermination: if two statistical hypotheses have exactly the same optimal likelihoods for all the possible observations, then how do we choose between them? While an answer to this question often invokes theoretical criteria such as simplicity and explanatory considerations, we have attempted to provide an answer in terms of empirical criteria. The idea is to use the background theory that generates the hypotheses, namely the causal picture. This theory provides us with a recipe for how to deal with interventions, and the resulting intervention data enable us to tell the indistinguishable hypotheses apart.

The second problem may be termed the problem of overdetermination after intervention: if none of the statistical hypotheses has a high enough likelihood for the observations after the intervention, then how do we adapt the set of hypotheses, or the model for short, in order to accommodate the poor fit? Here again the background theory or the causal picture provides us with the answer: we may adapt the causal picture in a controlled way, and thereby we extend the statistical model, or the set of statistical hypotheses. The extension of the model is simply determined by which extended model allows for the best fit improvement.

Both these problems in the methodology of science have been illustrated by means of factor analysis. For the problem of underdetermination, we have worked out how additional intervention data can be framed in terms of alterations to the theory behind the factor model, and then employed to resolve the underdetermined factor loadings. Unfortunately, we have not been able to apply the same ideas to the more practical setting of factor analysis with normal distributions over continuous variables. We have suggested that the underdetermination problem identified in the setting of discrete Bayesian networks is essentially identical to the underdetermination associated with the rotation of factors in the continuous setting. We are confident that in future work we can present a resolution of this problem of rotation on the basis of intervention data.

We must admit that our treatment of the second problem is much more sketchy. We hope to have showed that relative to a given causal picture linking latent and observable variables, extensions of the statistical model
can be guided by intervention data. But the precise methods and algorithms for putting this idea to work have not been provided, and we are therefore doubly removed from giving a practical application of the idea of controlled model change in, for instance, factor analysis. On the other hand, we see many potential applications of the idea. Once we have provided a concretisation of the algorithm of Section 4.2 along the lines of Section 4.3, we think the resulting tool can be of use to experimental scientists, but also to computer scientists working on the automated search of network structures.

All these applications lie within the realm of scientific methodology. However, there may also be a rather different application of the present ideas, within more traditional philosophy of science. The confirmatory practice of scientists has received a lot of attention from formally oriented philosophers of science, often with the aim to explain or rationalise scientists, or to provide them with norms. The experimental practice, on the other hand, has not been subject to the same critical and constructive scrutiny. Experiments have been the subject of science studies, but formal philosophers of science have by and large avoided the subject. But we think that the time is ripe to include experiments among the topics of formal philosophy of science, especially because the tools to describe interventions formally are available. We hope that with the present study, we are making the beginnings of a formal philosophy of experiment.

References


