Intervention, underdetermination, and theory generation

Abstract

We consider the use of intervention data for eliminating the underdetermination in 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 problem of underdetermination. Then we draw a parallel between factor analysis models and Bayesian networks with hidden nodes, which allows us to clarify the use of intervention data for dealing with indeterminacy. It will be shown that in some cases, the indeterminacy can be resolved by an intervention. In the other cases, the intervention data suggest specific extensions of the model. The upshot is that intervention data can replace theoretical criteria that are typically employed in resolving underdetermination and theory change.

1. Introduction. It has been argued that the problem of underdetermination, according to which several theories all fit the empirical facts equally well, is a logical curiosity rather than a real life possibility; see Laudan and Leplin (1991) and Douven (2008). Indeed, it is not easy to find empirically equivalent rivals to our best candidate theories. But we show that a particular form of underdetermination 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, not because of any lack of data, but rather because the hypotheses predict the same about the data and are therefore indistinguishable in principle. The standard response to this, at least in statistical model selection, is to look for theoretical criteria that force a choice between the rivals, such as

simplicity or explanatory force.[1] The underdetermination is then resolved by an appeal to theoretical considerations.

In this paper we investigate a different response to such underdetermination in statistical modelling, in line with what Laudan and Leplin suggest: the resolution of the underdetermination may be driven by changing the range of empirical data, specifically by making use of intervention data. The underdetermination involved is relative to what is taken as observable in a 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 consisting of these hypotheses is then called unidentifiable. However, the hypotheses in the model need not be altogether equivalent. We can consider specific changes to the experimental setup, or interventions for short, such that the background theory generates different likelihood functions over the results. The hypotheses then come apart.

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 background theory.

The objective of this paper is threefold. First, we illustrate that aspects of scientific method that are typically associated with theoretical considerations, such as resolving underdetermination and generating new theory, can also be driven, at least partly, by empirical fact. This enriches our understanding of underdetermination. Second, our discussion brings to the fore an important and undervalued aspect of scientific confirmation, namely the use of intervention data following experimentation. Thus we show that the views of Hacking (1980) and, much later, Pearl (2000) can come to fruition in confirmation theory. Third, we aim to develop the ideas in this paper into working tools for the social scientist. Up until now there have been few methodological tools available for exploiting intervention data. Scientists are often not aware of the potential of intervention data over and above the use of data simpliciter.

---

[1] In factor analysis, in particular, researchers use theoretical criteria pertaining to the variation among the estimations of the statistical parameters, such as "varimax". See, e.g., Lawley and Maxwell (1971).

The setting for illustrating these ideas is exploratory factor analysis. As it happens, factor analysis has already made an appearance in the philosophy of science in another context. 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 emphasized that in this paper we take a 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 the technique of exploratory factor analysis itself as a model for abductive inference and hence as a tool for theory choice.

The paper is set up in the following way. In §2 we describe two distinct problems of indeterminacy in factor analysis. We show in §3 that factor analysis is essentially identical to estimating parameters in a Bayesian network with hidden nodes. Like Bayesian networks, models in factor analysis therefore allow for incorporating intervention data. We argue that in specific cases, intervention data can be used to resolve the underdetermination problem. In §4 we argue that in certain other cases, the intervention data guide us towards adding a latent variable to the statistical model. In §5, finally, we suggest how the model for intervention and theory generation may prove useful to the philosophy of experiment, and more importantly, to scientific methodology.

2. Underdetermination in statistics. In §2.1 we characterize the problem of statistical underdetermination. Subsequently 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 S_t, and the event that earlier results were some ordered series $\langle q^0_1 q^1_2 q^0_3 \dots q^1_t \rangle$ by $s^{\langle 010\dots 1\rangle}$, or s_t for short. Denoting the hypothesis variable that the chance

of finding q_t^1 is θ with H_θ, with valuations h_θ, we have[2]

$$P(q_{t+1}^1 | h_\theta \cap s_t) = \theta \tag{1}$$

for each trial $t+1$, an expression often called the likelihood function of h_θ. We may assume that the chance θ of the event q_{t+1}^1 may be any value in $[0, 1]$. Then on the basis of some series of events S_t, we can provide an estimation of θ. 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 the so-called observed relative frequency

$$\hat{\theta}(S_t) = \sum_{i=1}^{t} I^1(Q_i)/t,$$

in which the indicator $I^1(Q_i) = 1$ if Q_i takes the value q_i^1 and 0 otherwise.

The above estimation problem is completely unproblematic. The observations have a different bearing on each of the hypotheses in the model, i.e. each member of the set of 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.[3]

This situation is different in the following statistical problem. Imagine that some researcher defined a slightly different set of statistical hypotheses G_ξ, characterized as follows:

$$P(q_{t+1}^1 | g_\xi \cap s_t) = \xi^2 \qquad \xi \in [-1, 1].$$

The set of hypotheses considered in the statistical problem is essentially the same. It is only labeled in a funny way. The hypotheses $g_{1/2}$ and $g_{-1/2}$ are

[2]We are writing the probability of data according to a particular hypothesis as $P(\cdot|h_\theta)$, and not as $p_{h_\theta}(\cdot)$, following the convention of Bayesian statistics: the hypothesis h_θ can serve as an argument of the probability function.

[3]Strictly speaking, there is a sense in which the estimation problem already suffers from an underdetermination problem. Note that we are dealing with probabilistic relations between observations and hypotheses. It may very well be that the binary results summarized in s_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, $s^{0000000}$, 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_θ, and in that sense we are encountering an underdetermination problem in the estimation. However, in this paper we will not consider this type of underdetermination.

indistinguishable, because they both assign exactly the same probability to all the observations. More generally, we have $P(q^1_{t+1}|g_\xi \cap s_t) = P(q^1_{t+1}|g_{-\xi} \cap s_t)$. We are 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. This is the notion of underdetermination at issue in this paper. Importantly, statistical underdetermination is not definitive: it is not ruled out that there are experiments or additional observations that allow us to disentangle the statistical hypotheses that are, in the light of the available data, indistinguishable. In fact, in this paper we will indicate what additional experiments can achieve this. On the other hand, if we take the statistical hypotheses involved in the foregoing simply as probability functions over the data, then these hypotheses are exactly identical; and identical things cannot be distinguished, period. It is only because the hypotheses are associated with different causal structures, as will be made clear in the following, that we can design experiments to distinguish them.

2.2. Factor analysis. The above example of statistical underdetermination is rather contrived. 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).

The technique of exploratory factor analysis posits a specific statistical model of hidden, or latent, random variables on the basis of an analysis of the correlational structure of observed, or manifest, 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. 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 §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, $P(F_i, L_i) > P(F_i)P(L_i)$. One way of accounting for the correlation is by positing a statistical model over the variables in which fear and loathing may be related directly, and then estimate the parameters in the model. But we may feel that this model 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 the presence or absence of a drug E. The correct statistical model, we may argue, posits a correlation between the drug and the fear, positive or as the case may be, negative, and similarly a correlation between the drug and the loathing, while conditional on a certain drug dosage, fear and loathing are uncorrelated: $P(E_i, F_i, L_i) = P(E_i)P(F_i|E_i)P(L_i|E_i)$. We then say that the drug dosage is the common factor to the manifest 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 manifest 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 manifest and latent variables are binary. In the example, the drug is either present in subject i, e_i^1, or absent, e_i^0, 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 E_i, we thus confine ourselves to

$$P(f_i^1|e_i^j) = \phi_j\,, \tag{2}$$

$$P(l_i^1|e_i^j) = \lambda_j\,, \tag{3}$$

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

$$P(e_i^1) = \epsilon \tag{4}$$

The probability over the variables E_i, L_i and F_i is thus given by five Bernoulli distributions, each characterized independently by a single chance parameter.

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 behind the correlations 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 negative emotions used to describe individuals in a general population whose constitution is otherwise unknown. If these variables are strongly positively correlated, it may be that we can account for the correlations in a statistical model positing a fairly small number of common factors, or even a single common factor, for example, the common factor depression, denoted D. Exploratory factor analysis is a technique for arriving at such common factors in a systematic way. When given a set of correlations among manifest 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 manifest variables uncorrelated.[4]

It will not be surprising that applications of factor analysis suffer from problems of underdetermination. After all, factor analysis posits 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. But this need not surprise us too much: almost all statistical modeling must at some point address this worry.

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 §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 problem 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 (5) to Equations (7), but replace the drug variable E with the depression variable D:

$$P(f_i^1|d_i^j) = \phi_j\,, \tag{5}$$

$$P(l_i^1|d_i^j) = \lambda_j\,, \tag{6}$$

$$P(d_i^1) = \delta \tag{7}$$

For a quick understanding of the problem, focus on the dimensions of the model. In total we count a number of 5 parameters, namely δ, and ϕ_j and λ_j for $j = 0, 1$. On the other hand, we have the binary 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

[4]Seeing that exploratory factor analysis generates a structure that explains the observed correlations, it is rather natural that Haig (2005) and Schurz (2008) present it as a formal model of abduction.

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. We write $\theta = \langle \delta, \phi_0, \phi_1, \lambda_0, \lambda_1 \rangle$. Further, the observations of individuals i are $f_i^j \wedge l_i^k$, which may be summarized as q_i^u with $u = 2j + k$. The sequences s_t are again observations of individuals $s^{u_1 u_2 \ldots u_t}$. Finally, we abbreviate $\eta_{jk} = P(q_i^{2j+k}|h_\theta) = P(f_i^j \wedge l_i^k|h_\theta)$:

$$P(f_i^0 \wedge l_i^1|h_\theta) \equiv \eta_{01} = \delta(1-\phi_1)\lambda_1 + (1-\delta)(1-\phi_0)\lambda_0\,, \tag{8}$$

$$P(f_i^1 \wedge l_i^0|h_\theta) \equiv \eta_{10} = \delta\phi_1(1-\lambda_1) + (1-\delta)\phi_0(1-\lambda_0)\,, \tag{9}$$

$$P(f_i^1 \wedge l_i^1|h_\theta) \equiv \eta_{11} = \delta\phi_1\lambda_1 + (1-\delta)\phi_0\lambda_0\,, \tag{10}$$

The fourth likelihood, $P(f_i^0 \wedge l_i^0|h_\theta)$, can be derived from these expressions. The salient point is that the system of equations resulting from filling in particular values for the likelihoods η_{jk} has infinitely many solutions in terms of the components of θ: for any value of the likelihoods η_{jk}, the space of solutions in θ has 2 dimensions. Conversely, different hypotheses h_θ will have the same set of likelihoods η_{jk} for the observations. In a Bayesian analysis, the hypotheses h_θ that are associated with the same likelihoods η_{jk} cannot be told apart, in the same way as that the hypotheses h_ξ and $h_{-\xi}$ cannot be told apart by the observations.

The fact that hypotheses cannot be told apart means that classical maximum-likelihood estimation does not lead to a unique best hypothesis, and the same problem shows up in the shape of the posterior distribution over the hypotheses.[5] We observe the relative frequencies

$$r_{jk}(s_t) = \frac{1}{t}\sum_{i=1}^{t} I^j(F_i, s_t) I^k(L_i, s_t), \tag{11}$$

where the indicators $I^j(F_i, s_t) = 1$ if $s_t \subset f_i^1$ and 0 otherwise, and $I^k(L_i)$ analogously. By means of the likelihoods given in Equations (8) to (10) we can then determine a posterior probability for the hypotheses in the model

[5]We will not devote much time to classical estimations here. Most of the paper is cast in Bayesian terms, so in the following we determine the expectation value of parameters instead. For increasing size of the data set, the maximum likelihood estimation tends towards this expectation value.

by means of Bayesian conditioning:

$$\begin{aligned} P(h_\theta|s_t) &\propto P(h_\theta)P(s_t|h_\theta) \\ &= P(h_\theta)\prod_{jk}\eta_{jk}^{t\,r_{jk}(s_t)}. \end{aligned} \tag{12}$$

The overall likelihood $P(s_t|h_\theta)$ is maximal if we set $\eta_{jk} = r_{jk}(s_t)$. But there are infinitely many hypotheses h_θ that have these particular values for the likelihoods. Consequently, there is no unique hypothesis h_θ that has maximal overall likelihood $P(s_t|h_\theta)$. Within the set of hypotheses with maximal likelihood, the shape of the posterior is simply proportional of the shape of the prior, no matter how large s_t is.

From the posterior distribution over the hypotheses we can generate the estimation, or rather the expectation value, of the parameters in θ, according to

$$\mathrm{E}[\theta] = \int_{[0,1]^5} \theta P(h_\theta|s_t)d\theta \tag{13}$$

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.[6] What is more troublesome is that no amount of additional data can eliminate this dependence of the estimations on the prior. That is, the estimations are statistically underdetermined by the data.

One reaction is to downplay this problem, and to say that it is merely a problem for the statisticians involved in the research, since it only concerns the values of parameters. But because the estimations and expectations are not fully determined, the causal, nomic and conceptual structure of the factors underlying the observed variables is not determined either. Different values for the parameters ϕ_j and λ_j entail different systematic relations between depression, fear and loathing, and ultimately this reflects back on our understanding of the posited notion of depression itself. In the statistical underdetermination exemplified here we therefore find back the classical underdetermination of theory by data.

2.4. Underdetermination in multivariate linear regression We are well aware that the statistical model considered in the foregoing 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 E is assumed to be governed by some continuous distribution as well. Writing $F_i = y$ for the

[6]The dependence on the prior was already briefly considered in footnote 3.

event that the level of fear is $y \in \mathbb{R}$, and similarly for depression $D_i = x$, the relation between F_i and D_i, for example, is

$$P(F_i = y|D_i = x) = N(\lambda_F x, \sigma_F) \tag{14}$$

in which $N(\lambda x, \sigma)$ is a normal distribution over the values y of F_i. So the relation between the variables D_i and F_i is characterized by a richer family of distributions, parameterized by a regression parameter λ_F and an error of size σ_F.

Despite these differences, the same kind of underdetermination also occurs in the more complicated statistical models. But in such models it takes a slightly different shape. Note first that we can extend factor models like the one above to include any number of common factors. However, once a model includes more than one common factor, we find that the factor loadings are not completely determined. Say, for example, that we analyze 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 labeling the states in this surface differently, for example by introducing a linear combination of D and M, calling it bipolarity, and further introducing another coordinate that is perpendicular to it, perhaps calling it a neurotic disposition. The factors in a model may be linearly combined or, in more spatial terms, rotated to form any new pair of factors.[7]

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 appropriately known as the problem of the *rotation of factor scores*. Neither the estimation criteria, often maximum likelihood, nor Bayesian methods of incorporating the data lead to a single best hypothesis in the factor model. The result is rather a collection of such models, meaning that the factor model is again unidentifiable, with all the attached problems listed above.

A standard reaction to the rotation problem 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, because the parameterization of the space must be such that there are no correlations between the latent variables. There are, however, alternative theoretical criteria for choosing the parameterization of the space of latent

[7]For readers familiar with linear algebra: the space of latent variables can be characterized in terms of different bases.

variables. For example, it may be interesting to have maximal variation among the regression coefficients which, intuitively, comes down to coupling each latent variable with a distinct subset of manifest variables. The thing to note is that, from the point of view of statistics, the choice for how to parameterize the space of latent variables is underdetermined: we cannot decide between these parameterizations on the basis of the observations alone.

In this paper we will not elaborate the mathematical details of underdetermination in these more complicated models. For present purposes, it suffices to use the simpler factor model of Equations (2) to (4). The crucial characteristic in all of what follows is that there are latent variables explaining the correlational structure among the manifest variables, and that these structures are not fully determined by the correlations among the observed variables.

2.5. 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 simply 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 some restrictions to this allocation, however. For example, as worked out in Ellis and Juncker (1997), if we let the number of manifest variables increase and assume that the latent variable is tail-measurable in terms of these manifest variables, then the factor scores are determined up to a functional transformation.

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 to specify a complete probability assignment over the latent and manifest 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 manifest variables that are related to the latent variables according to distributions of a suitable,

namely exponential, form. In this paper we will therefore ignore most of the discussion on factor score indeterminacy. However, there is one point at which the problem of factor score indeterminacy enters the present discussion. We will show in the following that intervention data can also be used to choose among a class of priors. But as indicated, the problem of choosing a prior probability is related to the problem of factor score indeterminacy. Therefore 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 to psychologists working on personality tests and the like that the data do not allow for a full determination of the structure of the underlying factors.

3. Interventions to resolve underdetermination. In the foregoing we have shown that factor analysis suffers from statistical underdetermination. 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 in §3.1, then the notion of intervention in §3.2, and finally its use in resolving underdetermination in §3.3. After this we discuss the implications in §3.4, and we briefly revisit the indeterminacy of factor scores in §3.5.

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 $P(V_i \mid Par_i)$ 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 $V_i \perp\!\!\!\perp ND_i \mid Par_i$; see Pearl (2000). Under this assumption the network suffices to determine the joint probability distribution over the variables, via the identity

$$P(v_1^{j_1} \cdots v_n^{j_n}) = \prod_{i=1}^{n} P(v_i^{j_i} \mid par_i) \tag{15}$$

where par_i is the assignment of values to the parents of V_i that is induced by the assignment $v_1^{j_1} \cdots v_n^{j_n}$ of values to the whole domain.

It is well-known that Bayesian networks, structural equations modeling, and factor analysis are closely related; see Pearl (2000). Effectively, the introduction of 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 on depression, fear and loathing,

$$P(F_i, L_i, D_i) = P(F_{i'}, L_{i'}, D_{i'}), \tag{16}$$

so that 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 depression D there is no correlation between the manifest fear F and loathing L,

$$P(F, L, D) = P(D)P(F|D)P(L|D). \tag{17}$$

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.

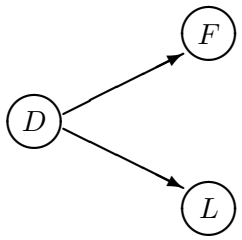

Figure 1: The graphical structure representing the independence relations in a factor analysis of depression, fear and loathing.

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, and any child can be connected to any parent. On the other hand, 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 inference over Bayesian networks presented in Romeijn et al. (2009). Hence the statistical underdetermination presented in §2.3 is framed as a problem to do with determining the posterior probability distribution over the parameters that characterize the Bayesian network of Figure 1. As announced, we are going to resolve this statistical underdetermination by means of intervention data. To this aim we first introduce interventions in the context of Bayesian networks.

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*; it 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 exogenously. 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 $P(v_i^j \| v_k^l)$ 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 the probability assignments and the 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 15,
- $P(x|y)$ is determined from the causal net by setting

$$P(x|y) = P(xy)/P(y)$$

  where the numerator and denominator are obtained via Equation 15,

- $P(x\|z)$ is determined from the causal net by first forming a new net by deleting arrows pointing towards 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.[8]

In fact causal nets can handle a rather more general notion of intervention. We can write

$$P(v_i^j) = s \parallel P(v_k^l) = t$$

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_k^l 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 $P(v_k^l) = t$, the new causal net deems that $P(v_i^j) = s$. This kind of intervention is sometimes called an *imperfect* intervention or a *stochastic* intervention, to distinguish it from the divine interventions considered above.[9] 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

$$P(v_i^j) = s \parallel P'(V_k|Par_k)$$

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, collect more evidence, and use this new evidence to 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, written $F \perp\!\!\!\perp L \mid D$. This evidence is compatible with the causal graph of Figure 1, but equally with

[8]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.

[9]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 $P(v_i^j) = s \parallel P(v_k^l) = t$ is different to the function p on the right-hand side.

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 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.

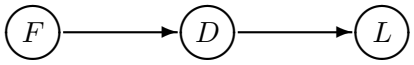

Figure 2: A chain of fear F causing depression D, which causes loathing L.

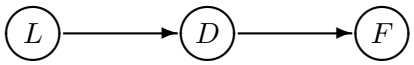

Figure 3: A chain of loathing L causing depression D, which causes fear F.

While resolving underdetermination of causal structure is the main application of interventions in the literature, interventions can also be used to resolve the statistical underdetermination of the parameters in a causal net. In this case, suppose that the causal structure is known and that 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 we show how exactly interventions can be used as further constraints, and thus solve the underdetermination introduced in §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 in factor analysis. Before anything else, we need to assume that the factor model is not simply a convenient way of representing the probability functions involved. The arrows in the factor model need to be interpreted causally, that is, the common factors must be taken as the cause of the correlational structure among the observed variables. With this causal assumption in place, 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 manifest 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 manifest 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 characterized by Equations (5) to (7), (16) and (17). 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, giving all the subjects a treatment intended to change the probability for depression. In terms of the foregoing, we change the probability of depression, $P(d^1) = \delta$, to a new value,

$$P'(d_i^1) = \delta',$$

which is supposed to be lower than δ. The relations of the depression variable to the variables of fear and loathing, given by $P'(f_i^1|d_i^j) = \phi_j$ and $P'(l_i^1|d_i^j) = \lambda_j$, are not changed by the intervention: the treatment is supposed to change the probability for depression but not how the depression, when absent of present, affects feelings of fear and loathing. Finally, after the intervention we record the observations s'_t. In particular, we observe the fractions $r'_{jk}(s'_t)$, or r'_{jk} for short, the relative frequencies of f_i^j and l_i^k after the intervention.

To get the point of this across quickly, we focus again on the dimensions of the model. This time we count a number of 6 parameters, namely δ, ϕ_j and λ_j for $j = 0, 1$, and finally δ'. On the other hand, we have a richer set of observations that can be used to determine these parameters. Specifically, we have 3 observed relative frequencies of $f_i^j \wedge l_i^k$ before intervention, $r_{jk}(s_t)$, and 3 of them after intervention, $r_{jk}(s'_t)$, so six in total. Whereas previously we had two degrees of freedom left after the incorporation of the data, it seems that we can now fill in all the parameter values of the factor model.

Let us make this more precise. As before, we have the likelihoods of Equations (8) to (10). But to these expressions we now add the likelihoods of the hypotheses after the intervention:

$$P'(f_i^0 \wedge l_i^1|h_\theta) \equiv \eta'_{01} = \delta'(1-\phi_1)\lambda_1 + (1-\delta')(1-\phi_0)\lambda_0\,, \quad (18)$$

$$P'(f_i^1 \wedge l_i^0|h_\theta) \equiv \eta'_{10} = \delta'\phi_1(1-\lambda_1) + (1-\delta')\phi_0(1-\lambda_0)\,, \quad (19)$$

$$P'(f_i^1 \wedge l_i^1|h_\theta) \equiv \eta'_{11} = \delta'\phi_1\lambda_1 + (1-\delta')\phi_0\lambda_0\,. \quad (20)$$

The system of equations that results from equating likelihoods and observed relative frequencies

$$\eta_{jk} = r_{jk}, \qquad \eta'_{jk} = r'_{jk},$$

has an unique solution in terms of the components of θ and the additional parameter δ', up to a transformation of the values for D. That is, the solutions always come in mirror-image pairs, differing in the interpretation of the values for the variable D. Fixing d^1 to mean the depressed state, the solution is unique.[10]

Apart from this symmetry in the solutions, every hypothesis h_θ in the model is thus associated with a unique set of values for the likelihoods η_{jk} and η'_{jk}. Conversely, if the data are generated by a chance process associated with one such hypotheses h_θ, then we can identify this hypothesis, in the same way as we were able to identify the true h_θ in the model of Equation (1). However, this does not hold for the entire range of possible values for the observed frequencies. For extremal values there is still an infinity of solutions, and more importantly, some combinations of frequencies simply do not correspond to any of the statistical hypotheses within the model, leading to a bad model fit. In the next section, we discuss a case in which the intervention data necessitate a revision of the model in question.

The intermediate conclusion of this section is that intervention data can indeed be used to resolve the statistical underdetermination, as it was introduced in §2.1. Modulo a role swap for d^0 and d^1, the maximum likelihood estimation will return a unique maximum. Similarly, in a Bayesian statistical inference the posterior distribution will generally have a unique maximum after Bayesian conditioning on the normal and the intervention data. Accordingly, the expectation value for the parameter will in the long run be independent of the prior distribution over the hypotheses.

3.4. Philosophical and practical implications The philosophical upshot of this result is that empirical criteria for theory evaluation, based on the targeted acquisition of intervention data, can take the place of the theoretical criteria that normally guide theory choice in the face of underdetermination. Where we had otherwise used a theoretical criterion to choose among the equally well fitting alternative hypotheses, we can now decide on the basis of additional data, obtained after intervention. Within statistics, one might say, the problem of underdetermination has 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.

[10]We have checked this using the solver in Mathematica. Thanks to David Atkinson, we also have an analytic derivation of the solutions, but we need not burden the reader with it here.

It is remarkable that we do not need to know anything about the exact impact of the intervention. That is, we do not need to know the exact value of δ'. The mere fact that we have changed something to the probability of the latent variable suffices. However, this is not to say that the use of intervention data requires no assumptions whatsoever. As indicated in the foregoing, the new data can only be taken as pertaining to the same parameters if we assume that the causal structure of the latent and observed variables is roughly correct. More specifically, we need to assume that the probabilistic relations between the latent and the observed variables, expressed in ϕ_i and λ_i, remain invariant under intervention. So in order to employ the intervention data for a resolution of the statistical underdetermination, we have to make particular causal assumptions. Nevertheless we think that the resolution of underdetermination by causal assumptions and further empirical data is to be preferred over a resolution that employs a theoretical criterion directly.

We want to emphasize that the results of this paper are not only philosophical. The foregoing may also be of particular interest in the practical application of factor analysis. Recall the problem of underdetermination due to the rotation of latent variables, as discussed in §2.3. This rotation problem is particularly pressing for the design of clinical and personality tests: how do we relate clusters of tests to specific personality traits? And what traits should we distinguish in the first place? The fact that we can opt for a multitude of latent structures, each associated with a different causal story on how the correlations between observed variables has emerged, presents researchers with a genuine problem. The standard response to this problem is to employ theoretical criteria on the latent variables, for example by supposing that the traits are independent, or by choosing the latent variables such that the regression parameters show maximal variation, thus associating each test with a minimal number of traits.

The idea of the present paper is that these theoretical criteria can be replaced by intervention data. For example, for clinical psychologists working with factor analysis, interventions may constrain the latent structure behind their tests, thereby providing a clearer view of what the tests are measuring. However, it leads us to far away from the line of this paper to explicate an application here. We leave the details of this to another paper, because it addresses an audience of psychologists and statisticians rather than philosophers.

3.5. Interventions and the indeterminacy of factor scores. Finally, we want to remark on the problem of the indeterminacy of factor scores, as discussed in 2.5. Insofar as there is a problem with factor scores in the Bayesian treatment, intervention data can play an interesting part.

Recall that the expected value E[θ], given in Equation (13), depends on the posterior probability over the parameter $P(h_\theta|s_t)$, and that according to Equation (12), this posterior depends on the prior probability $P(h_\theta)$. As shown in Bartholomew and Knott (1999), the indeterminacy of factor scores in classical factor analysis derives directly from the fact that a prior probability is not provided. And because in a Bayesian treatment such a prior is assumed, we can say that Bayesian factor analysis is not pestered by factor score indeterminacy. However, the prior is assumed, not derived, so a classical statistician may well ask for a motivation of the prior probability assignment.

With the above ideas in place, the prior probability may be given an after the fact motivation by means of intervention data. Instead of choosing a single prior, we might consider a whole collection of possible priors over the parameter values. For example, we might consider as priors all so-called Beta-distributions:

$$P(h_\theta) = \frac{(2n-1)!}{((n-1)!)^2}\delta^{n-1}(1-\delta)^{n-1},$$

parameterized by the natural numbers $n > 0$. Effectively, we thereby increase the dimension of the parameter space by one. However, we might know from a different study that the chance of being depressed after the treatment, δ', is has some particular value, or is deterministically related to the chance on depression before treatment. This reduces the number of parameters with one again, because δ' is then fixed, or every δ' is coupled to a unique value δ. The net effect is that we can again make an estimation of all the parameters, namely δ, ϕ_j and λ_j for $j = 0, 1$, and finally the parameter n.

The point to note is that the last parameter to be estimated, n, determines the prior over δ. In other words, just like we can estimate the effects of an intervention, δ', we can estimate the prior probability assignment that best suits the factor model. In this way we can provide an independent, after the fact justification for choosing one of the priors, namely one of the values of n, as the correct starting point for the factor model.[11]

Now this is of course just a toy example. We have not said anything about the more realistic continuous case, in which we typically assume a normal distribution over the continuous variable D_i as prior. Moreover, it is highly unrealistic to suppose that there is a clear and deterministic relation between the parameters governing the distribution over the variables

[11] In the statistical literature, the idea that we can confirm or disconfirm probability distributions over statistical parameters has become known as hierarchical Bayesian modelling. See, for instance, Chapter 5 of Gelman (2004).

D_i before and after the intervention. Nevertheless, the foregoing gives an indication of how intervention data can be of use in dealing with the heir of the problem of factor score indeterminacy in Bayesian factor analysis, namely the problem of how to choose a prior.

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 the resulting poor fit of a factor model after intervention may lead to controlled changes to the model. In other words, interventions may guide model change.

There are many ways in which interventions may guide model change and in this section we can do no more than scratch the surface. In §4.1 we will explain the idea of overdetermination. In §4.2 we will sketch some of the ways in which overdetermination after intervention may lead to model change. Then in §4.3 we will give a high-level overview of one kind of model change: adding new common causes and causal relations. Finally in §4.4 we will present a worked example of this kind of model change.

4.1. Model change. We first need a model in which the overdetermination after intervention occurs. To this aim, consider the factor model alluded to earlier, which analyzes 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 independent relative frequencies, associated with the combinations of observable variables. Hence the model can fully determined by the observations. Of course we do not claim that this model has a perfect fit for any set of observed relative frequencies.[12] However, for present purposes we assume that the relative frequencies are such that the parameters can all be determined, and fitting the frequencies exactly.

Now imagine that we want to see the effect of a certain therapy. We first observe a number of subjects on the variables F, L, and S, obtaining a

[12]As in the foregoing, when we were considering two observed variables and a single latent variable, we actually obtain two mirror-image solutions for frequencies within a particular range, no solutions outside that range, and a continuum of solutions for frequencies with extremal values.

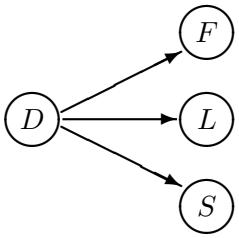

Figure 4: The probabilistic independence relations between drugs, fear, loathing and sleeplessness.

solution for all the parameters, after which we treat the subjects and observe them again. The idea of the treatment is that it only interferes with the probability of depression and not with the relation between depression and its observable consequences. After the intervention we have $P'(d_i^1) = \delta'$, while the rest of the model remains unchanged. Note, however, that after the intervention we have a total of 14 relative frequencies to accommodate, for which we have only 8 free parameters. 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 observed 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 §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.

It may also 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 success percentage of the predictions, or by some other notion of distance to perfect or optimal prediction. If the estimations in the given model fall short of the criterion for predictive performance, then we deem the fit poor and say that the data *overdetermines* the model.[13]

To properly understand the notion of overdetermination intended here, recall that we distinguish two kinds of evidence. First there is the *data*, i.e., empirical observations of values of the variables in the model, possibly after intervention of one or more variables. This kind of evidence bears on the model through the model's consequences: the data fit the model well just

[13]For a number of relatively recent discussions of model selection and fit, see Waldorp and Wagenmakers (2006).

when they accord with the probabilistic predictions of the model. Second, there is evidence of *mechanisms*, or *causal relations*; a theoretical understanding of the domain can tell one about the kinds of causal or mechanistic connections between the variables. This kind of evidence is often qualitative and bears on the model through the model's structure: the mechanisms accord with the model just when they are reflected in the relations posited by the graph of the model.

Overdetermination only concerns the fit between model and *data*. A model may be overdetermined by data yet accord well with one's knowledge of underlying mechanisms. Conversely, a model may poorly reflect knowledge of qualitative connections between the variables yet its predictions may fit the observed data well, in which case there is no overdetermination.

4.2. Network dynamics. In the case in which overdetermination disconfirms a particular factor model or causal network, there are numerous 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. Here we shall discuss these strategies in turn.

First, we can discard the old network 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 approach of machine learning (see, e.g., Neapolitan, 2003; Spirtes et al., 1993) and it is guaranteed to work in the sense that, given a set of variables, there will always be some network on those variables that fits the observational data involving the variables best. However, there is no guarantee that one network fits both the interventional data and the observational data, and it is certainly not guaranteed that the network arrived at in this manner complies to the format enforced by factor analysis, in which there are no edges between the observable variables. Moreover, often there is evidence that supports the qualitative relations posited in the current net, such as evidence of mechanisms linking causes and effects, and in such cases 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 lead to a good fit (Williamson, 2005, Chapter 3). But as indicated before, posited causal relations may simply not be plausible. There may be no plausible physical mechanism that makes fear responsible for loathing, for instance. In the factor models the observable variables are indicators that cannot themselves play any causal role. Hence, additional arrows will have to obtain between the latent and the observed variables, and there is no guarantee that this will lead to a good fit.

There are other methods for handling overdetermination that retain the existing causal relations and avoid positing causal relations from one measured variable to another. One such is the method of increasing the range of values that a common cause can take. Consider the case of overdetermination sketched earlier: 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 may then 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, the depression being present or absent, to numerical measurement, the depression being on one of a number of levels. 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. But 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.[14]

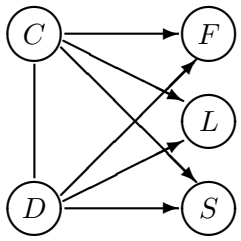

Figure 5: A template for the common cause strategy.

[14]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.

This leaves us with a final method for handling overdetermination while keeping the existing causal relations: the method of introducing new common causes (also called *latent factors* or *hidden nodes*) to the factor model (Kwoh and Gillies, 1996; Binder et al., 1997). This procedure yields a subgraph of the template in Figure 5: 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.

We call model changes of the last two kinds—adding new causal relations or adding new common causes—*conservative* model changes because they conserve (and augment) the causal structure of the previous model. Conservative changes are appropriate when the data overdetermine the model but when existing causal relations remain plausible given mechanistic evidence. We now describe such model changes in more detail.

4.3. Conservatively changing the network. The aim of the present section is to sketch a procedure for conservatively changing a model after intervention data have led to poor model fit. We will consider ways of adding latent variables to the model, or in terms of Bayesian networks, adding to the network hidden nodes and edges towards observed variables. We thereby effectively enlarge the parameter space, and thus the family 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.

We first give a high-level 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 facts.

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 the intervention, and a causal

graph, we can determine the parameter values that yield a distribution fitting 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 χ 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 χ' imposed by interventional data.
  - If possible, add the arrow to $\mathcal{G}$—from those compatible with χ, χ' and acyclicity constraints—which yields the parameters that fits the data best.[15]
  - Otherwise add a new (common cause) node to $\mathcal{G}$; add to χ constraints to ensure that future graphs fit the template of Figure 5.

Output: A modified causal graph $\mathcal{G}$.

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. See Bang et al. (2003) for an application of this kind of algorithm to metabolic network prediction.

Some remarks on this algorithm are in order. First, we have left open the question of how to measure the best fit with the data. 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

[15]If there is more than one optimal arrow, one can spawn a new graph for each such 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.

on the data. Alternatively, for each model we might select the best estimate within that model, 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, thereby departing from the idea of a greedy search of one arrow at a time. This latter approach is applicable if we want to compare changes to the network that are associated with extended models that differ in dimensionality.

Second, it is important to be aware that the effects of original intervention vary according to the output of the above algorithm. Suppose, for example, that some instance of the template of Figure 5 is output, and that mechanistic considerations lead one to suppose that D was intervened upon. Now if D is a cause of C in the output graph, then intervening on D also changes C. However, if D is not a cause of C in the output graph the intervention on D fails to change C. It is this dependence of the intervention on the structure of the graph that enables one to infer the structure of the graph from the intervention.

4.4. Worked example. Recall the example on fear, loathing, and sleeplessness of §4.1. As indicated, after the intervention we have a total number of 8 independent parameter values to accommodate 14 observed relative frequencies. In such a case it can very well happen that the best estimate for the parameters does not fit the data well enough according to the chosen fit criterion. In that case, we may decide to expand the factor model by the above template.

The data determine the relative frequencies r_{jkm} and r'_{jkm} for the combinations $f^j \wedge l^k \wedge s^m$ of observed variables before and after the intervention. We have the set of parameters θ, now extended with σ_0 and σ_1, from which we can construct the likelihoods η_{jkm} and η'_{jkm}, as before. Now say that we adopt the criterion that the quadratic distance between the observed relative frequencies and the parameter values must be minimal, and that deviations of more than 0.025, i.e.

$$|r_{jkm} - \eta_{jkm}| > 0.025, \quad |r'_{jkm} - \eta'_{jkm}| > 0.025,$$

are considered to be a poor fit. Of course, this notion of fit and its associated criterion are very primitive. A typical application will involve fit by maximum likelihood and a fit criterion in which the complexity of the model, usually measured by the number of parameters in the model, is represented as well. However, for the illustrative purpose of the present paper, the simplified version suffices.

Now imagine that we obtain the following set of observed relative frequencies before and after the intervention:

ijk	000	001	010	011	100	101	110	111
r_{ijk}	0.370	0.036	0.018	0.092	0.011	0.065	0.041	0.367
r'_{ijk}	0.615	0.309	0.013	0.014	0.007	0.009	0.004	0.031

We now estimate, by least squares, the values of the parameters that correspond to the original causal structure, as expressed in Figure 4. We find the values expressed in the table below.

Parameter	δ	δ'	ϕ_0	ϕ_1	λ_0	λ_1	σ_0	σ_1
Value	0.526	0.054	0.021	0.833	0.028	0.908	0.288	0.912

However, this estimation does not have a sufficiently good fit. In the table below the estimations that show more than a size 0.025 deviation are presented in boldface.

ijk	000	001	010	011	100	101	110	111
η_{ijk}	**0.322**	**0.137**	0.016	0.077	0.010	**0.040**	0.035	0.363
η'_{ijk}	**0.641**	**0.260**	0.019	0.015	0.014	0.009	0.004	0.038

Since the factor model used for the estimation cannot be improved upon by adding edges, the conclusion is that we must look for a factor model with more latent variables.

Following the algorithm of §4.3 we now determine which of the edges in the template of Figure 5 must be added. We follow a greedy search, that is, we add edges until we find a fit that satisfies our criterion. However, not all additions of edges lead to a more versatile model, or to a model that is better able to accommodate the data. Moreover, whether a certain addition of edges indeed leads to a better overall performance on the data also depends on the model selection criterion used to compare the different causal structures. To keep matters as simple as possible, we will rank the additions of causal arrows according to the number of additional degrees of freedom they introduce, and we will measure the fit merely by the binary judgement of sufficient proximity to the target frequencies.

The minimal addition of degrees of freedom consists in a single edge from C to D, and a further edge from C to the observed variables S, L, or F. The first of these three new networks is depicted below.
We may now calculate the estimations for each of the new networks and then compare the resulting likelihoods η_{ijk} and η'_{ijk}, belonging to the observations of $f^i \wedge l^j \wedge s^k$ before and after the intervention, to the relative frequencies r_{ijk} and r'_{ijk} provided in the foregoing. As it turns out, the likelihoods that result from the network structure given in Figure 6 can be matched exactly to these frequencies. In other words, the extension of the latent causal structure underlying the correlations among the observed variables allows for a perfect fit on the assumption of the revised network. Hence

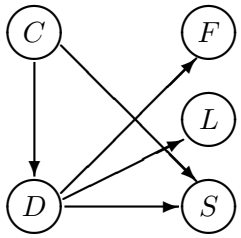

Figure 6: A new latent causal structure, involving a second latent variable C next to the depression variable D.

we conclude, from the intervention data and the graph, that another latent factor is at play in the correlation between fear, loathing, and sleeplessness, to do specifically with this latter indicator.

The exact alteration needed for fitting the relative frequencies[16] is merely a change in the value of the parameter σ_0. For the new estimations we find

Parameter	δ	δ'	ϕ_0	ϕ_1	λ_0	λ_1	σ_0	σ_1	σ'_0
Value	0.600	0.050	0.010	0.800	0.020	0.850	0.050	0.900	0.333

So before intervention we estimate $\sigma_0 = 0.050$, and after the intervention $\sigma'_0 = 0.333$. As indicated in the network, this made possible by the introduction of the latent variable C as a parent to D. Writing $P(c^1) = \gamma$, $P(d^1|c^0) = \delta_0$ and $P(d^1|c^1) = \delta_1$, we have that $\delta = \gamma\delta_1 + (1 - \gamma\delta_0)$. For ease of exposition, define an alternative parameterization

$$P(c^1|d^1) = \frac{\gamma\delta_1}{\delta} = \gamma_1,$$

and similarly for $P(c^1|d^0) = \gamma_0$. Before the intervention we have that $\sigma_0 = \gamma_0\sigma_{01}+(1-\gamma_0)\sigma_{00}$, while after the intervention we have $\sigma'_0 = \gamma\sigma_{01}+(1-\gamma)\sigma_{00}$, and similarly for σ'_1. In other words, where the parameters σ_k were previously left invariant under the intervention, they are now a mixture of σ_{0k} and σ_{1k}, and thereby subject to changes that derive from the intervention.

The fact that due to the intervention only σ_0 undergoes a change drops out of the estimation: the estimations for σ_1 and σ'_1 are simply identical. However, the new causal structure is not determined by the data entirely. We are left with some degrees of freedom introduced by the presence of the new latent variable C: different values of the parameters γ, δ_0, δ_1, δ', and

[16]Naturally, these frequencies are the result of reverse engineering. The table was constructed on the basis of the parameter values, but these parameter values were then found back by means of an estimation procedure.

σ_{00} to σ_{11} all match the restrictions imposed by the relative frequencies. Specifically, in terms of the parameters defined above, the 2 restrictions for δ and δ' and the 4 restrictions for σ_k and σ'_k with $k = 0, 1$ leave us with 2 degrees of freedom. So the intervention data direct the development of new concepts, as in the introduction of the variable C, but in turn this development may create a new underdetermination problem.

To some extent we can restrict the values of the parameters in the new latent structure by means of a causal story. Instead of just depression causing the sleeplessness and the feelings fear and loathing, we imagine that there is some other cause associated with sleeplessness in particular. For instance, subjects might simply be prone to worrying and therefore lie awake at night without being depressed. The specific way in which the disposition to worrying influences the sleeplessness experienced by the subjects is naturally aligned with the estimations: among the depressed people there will be relatively more people prone to worrying than among those who are not depressed, so that the number of people suffering from sleeplessness among those not depressed is relatively small. But if we successfully treat depression, then a larger portion of those who are not depressed, now also consisting of people cured from depression, will be prone to worrying, increasing the number of people suffering from sleeplessness among the not depressed.

This may be illustrative of the way in which intervention data can induce conceptual change, but we are well aware that the story is far from finished. We need a much more elaborate study on a whole range of topics: the selection of models on the basis of intervention data, the exact relation between changes in a causal network and extensions of a statistical model, and the interplay between causal networks and causal stories, and so on. But here is not the place for that. We do hope that with the foregoing we have given a perspective on an exciting new research area on the intersection of statistics, causal modelling, and the philosophy of science.

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 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. Together with some assumptions

on the causal structure of the latent and observed variables, the intervention data enable us to tell the seemingly 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 thereby determined by the data.

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 interventions can be framed in terms of alterations to the factor model, and how the intervention data can then be employed to resolve the underdetermination of the 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. But we believe that the underdetermination problem identified in discrete Bayesian networks is essentially identical to the underdetermination associated with the rotation of factors in the continuous setting, and we are confident that in future work we can present a resolution of this problem of rotation on the basis of intervention data.

Our treatment of the second problem is much more sketchy. We showed that, relative to a given causal picture that links 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 think there are many potential applications of the idea. Once we have provided a concretization of the algorithm of §4.2 along the lines of §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 of explaining or rationalizing scientists, or of providing them with norms on how to make a justified step from data to theory. The experimental practice, on the other hand, has not been subject to the same scrutiny from the side of formal modelling. 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

Bang, J.-W., Chaleil, R., and Williamson, J. (2003). Two-stage Bayesian networks for metabolic network prediction. In Lucas, P., editor, *Proceedings of the Workshop on Qualitative and Model-Based Reasoning in Biomedicine at the 9th European Conference on Artificial Intelligence in Medicine (AIME'03)*, pages 19–23. www.cs.ru.nl/˜peterl/mbqr-aime03.pdf.

Barnett, V. (1999). *Comparative Statistical Inference.* John Wiley, New York.

Bartholomew, D.J. and Knott, M. (1999). *Latent variable models and factor analysis.* Oxford University Press, New York.

Binder, J., Koller, D., Russell, S., and Kanazawa, K. (1997). Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29:213–244.

Douven, I. (2008). Underdetermination. In S. Psillos and M. Curd (eds.), *The Routledge Companion to the Philosophy of Science*, London, Routledge, pp. 292–301.

Eberhardt, F. and Scheines, R. (2007). Interventions and causal inference. *Philosophy of Science*, 74:981–995.

Ellis, J. L. and Juncker, B. W. (1997). Tail-measurability in monotone latent variable models. *Psychometrika*, 62(4): 495–523.

von Eye, A. and Clogg, C. C. (1994). *Latent variables analysis: applications for developmental research.* Sage, Thousand Oaks (CA).

Gelman, A. *et al.* (2004). *Bayesian Data Analysis, Second Edition.* Boca Raton: Chapman and Hall.

Glymour, C. (1998) What Went Wrong? Reflections on Science by Observation and the Bell Curve. *Philosophy of Science* 65:1–32.

Hacking, I. (1983) *Representing and intervening* Cambridge University Press, Cambridge

Haig, B.D. (2005). An abductive theory of scientific method. *Psychological Methods* 10:371–388.

Korb, K., Hope, L., Nicholson, A., and Axnick, K. (2004). Varieties of causal intervention. In *Proceedings of the Pacific Rim International Conference on AI*, New York. Springer.

Kwoh, C.-K. and Gillies, D. F. (1996). Using hidden nodes in Bayesian networks. *Artificial Intelligence*, 88:1–38.

Klugkist, I., Laudy, O. and Hoijtink, H. (2005) Inequality Constrained Analysis of Variance: A Bayesian approach. *Psychological Methods* 10:477–493.

Laudan, L. and Leplin, J. (1991) Empirical Equivalence and Underdetermination. *Journal of Philosophy* 88:449–472.

Lawley, D.N. and Maxwell, A.E. (1971). Factor analysis as a statistical method. Butterworths, London.

Maraun, M.D. (1996). Metaphor Taken as Math: Indeterminancy in the Factor Analysis Model. *Multivariate Behavioral Research* 31:517–538.

McDonald, R.P. (1974). The measurement of factor indeterminacy. *Psychometrika* 39:203–222.

Mulaik, S.M. (1985). Factor analysis and Psychometrika: Major developments. *Psychometrika* 51:23-33

Neapolitan, R. E. (2003). *Learning Bayesian networks.* Pearson / Prentice Hall, Upper Saddle River NJ.

Pearl, J. (2000). *Causality.* MIT press, New York.

Romeijn, J.W. (2008). The all-too-felxible abductive method. *J*ournal of Clinical Psychology

Romeijn, J.W., Haenni, R., Wheeler, G., and J. Williamson (2009). Logical relations in a statistical problem. In Loewe et al. *P*roceedings of Foundations of the Formal Sciences VI, College Publications, London.

Schurz, G. (2008). Common Cause Abduction and the Formation of Theoretical Concepts. TPD preprints, No.2.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search.* MIT Press, Cambridge MA, second (2000) edition.

Steiger, J.H. (1979). Factor indeterminacy in the 1930's and the 1970's some interesting parallels. *Psychometrika* 44:157–167

Waldorp, L. and Wagenmakers, E.J. (2006) Model selection: Theoretical developments and applications (Special issue). *Journal of Mathematical Psychology* 50.

Williamson, J. (2005). *Bayesian nets and causality: philosophical and computational foundations.* Oxford University Press, Oxford.