Comparison Between Suitable Priors for Additive Bayesian Networks

Kratzer, Gilles ; Furrer, Reinhard ; Pittavino, Marta

DOI: https://doi.org/10.1007/978-3-030-30611-3_10

Comparison Between Suitable Priors for Additive Bayesian Networks

Gilles Kratzer, Reinhard Furrer and Marta Pittavino

Abstract Additive Bayesian networks (ABN) are types of graphical models that extend the usual Bayesian-generalised linear model to multiple dependent variables through the factorisation of the joint probability distribution of the underlying variables. When fitting an ABN model, the choice of the prior for the parameters is of crucial importance. If an inadequate prior—like a not sufficiently informative one—is used, data separation and data sparsity may lead to issues in the model selection process. In this work we present a simulation study to compare two weakly informative priors with a strongly informative one. For the weakly informative prior, we use a zero mean Gaussian prior with a large variance, currently implemented in the R package *abn*. The candidate prior belongs to the Student's t-distribution. It is specifically designed for logistic regressions. Finally, the strongly informative prior is Gaussian with a mean equal to the true parameter value and a small variance. We compare the impact of these priors on the accuracy of the learned additive Bayesian network as function of different parameters. We create a simulation study to illustrate Lindley's paradox based on the prior choice. We then conclude by highlighting the good performance of the informative Student's t-prior and the limited impact of Lindley's paradox. Finally, suggestions for further developments are provided.

Keywords Graph theory · Structural search · Binomial regression

G. Kratzer (✉)
Department of Mathematics, University of Zurich, Zurich, Switzerland
e-mail: gilles.kratzer@math.uzh.ch

R. Furrer
Department of Mathematics and Department of Computational Science, University of Zurich, Zurich, Switzerland
e-mail: reinhard.furrer@math.uzh.ch

M. Pittavino
Geneva School of Economics and Management, Research Center for Statistics, University of Geneva, Geneva, Switzerland
e-mail: marta.pittavino@unige.ch

R. Argiento et al. (eds.), *Bayesian Statistics and New Generations*,
Springer Proceedings in Mathematics & Statistics 296,
https://doi.org/10.1007/978-3-030-30611-3_10

1 Introduction to ABN

Additive Bayesian network (ABN) models are types of graphical models that extend the usual generalised regression framework to multiple dependent variables through the representation of the joint probability distribution. ABNs are a special type of Bayesian network (BN) model in that each node in the graph comprises a generalised linear model (GLM). This property is used to compute additively the global score of the structure. All types of BN models consist of two reciprocally dependent parts: a qualitative one (the structure) and a quantitative one (the model parameters). BN models are statistical models that derive a directed acyclic graph (DAG) from empirical data, describing the dependency structure of the random variables. The DAG is the graphical representation of the joint probability distribution of all random variables represented by the data. The model parameters stem from the local probability distribution of all the variables in the network.

In the last few decades, BN modelling has been widely used in biomedical science and in systems biology to analyse multi-dimensional data [3, 4, 12, 13, 25]. Recently, ABN modelling approaches have been successfully applied to the field of veterinary epidemiology [10, 24, 28]. A very appealing feature of ABN is its ability to generalise standard regression methodology. A general introduction to BN and ABN modelling in veterinary epidemiology is provided by [18]. Further applications of BN to veterinary studies were described by [29]. Graphical modelling techniques based on ABN used to analyse epidemiological data were used by [17, 19, 20], resulting in several publications.

The paper is structured as follows. Relevant technical details of ABN models are presented in Sect. 2. Section 3 explains the issue of data separation and Lindley's paradox and highlights the importance of appropriate prior choice. Section 4 reports the results of a simulation study underpinning the necessity of careful prior selection with respect to data separation and Lindley's paradox. We conclude the article in Sect. 5 with future research directions.

The main original contribution of this paper is to compare the impact of different priors on the fitting performance of ABN models using synthetic datasets in a realistic epidemiological setting. The underlying idea is to render the ABN approach more robust in practice when using an adequate prior. This is of high importance, as a typical systems epidemiology dataset has a relatively small sample size. Our findings show that none of the currently proposed alternative priors are fully satisfactory and further studies need to be conducted.

2 Additive Bayesian Networks in a Nutshell

A BN model $\mathcal{B}$ for a set of random variables $X = \{X_1, \ldots, X_n\}$ consists of:

– A DAG structure $\mathcal{S} = (V, E)$, where V is a finite set of nodes and E is a finite set of directed edges between the nodes. A DAG is *acyclic*; hence, the edges in

E do not form directed cycles. A random variable X_j corresponds to each node $j \in V = \{1, \ldots, n\}$ in the graph. We do not distinguish between a variable X_j and the corresponding node j.
- A node k is said to be a *parent* of a node j if the edge set E contains an edge from k to j. A set of parents for a node j is denoted by $\mathbf{Pa}_j$. P_j indicates the total number of parents for a node j, i.e., $\dim(\mathbf{Pa}_j) = P_j \geq 0$ and $P_j = 0$ for orphan nodes.
- A set of local probability distributions for all variables in the network is encoded by $\boldsymbol{\theta}_{\mathcal{B}}$. Each node j, with parent set $\mathbf{Pa}_j$, is parametrised by a local probability distribution: $P(X_j \mid \mathbf{Pa}_j)$.

We denote a BN model $\mathcal{B}$ by the pair $\mathcal{B} = (\mathcal{S}, \boldsymbol{\theta}_{\mathcal{B}})$, representing the DAG *structure* $\mathcal{S}$ and the model *parametrization* $\boldsymbol{\theta}_{\mathcal{B}}$, respectively.

Edges represent both *marginal* and *conditional dependencies*. The main role of the network structure is to express the conditional independence relationships among the variables in the model through graphical separation, thus specifying the factorisation of the global probability distribution:

$$P(\boldsymbol{X}) = \prod_{j=1}^{n} P(X_j \mid \mathbf{Pa}_j). \tag{1}$$

The left panel of Fig. 1 shows a simple BN example for four nodes.

A general BN has an arbitrary distribution for each of the factors in (1), whereas an ABN defines each of these factors through a GLM. For example, assuming a discrete distribution for node 3 of Fig. 1, using classical notation for the exponential family parametrisation [22] its probability mass function writes

$$P(X_3 = x \mid X_1 = x_1, X_2 = x_2) = \exp\big(\eta(\theta_3)T(x) - A(\theta_3)\big)H(x), \tag{2}$$

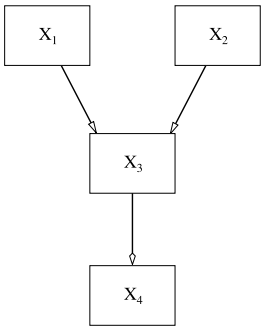
where the functions η, T, A, H may be node-dependent (we have omitted the indices to simplify the notation) and where θ_3 incorporates the configuration of the parents' node

$$\eta(\theta_3) = \beta_{3,0} + \beta_{3,1}x_1 + \beta_{3,2}x_2. \tag{3}$$

For nodes with continuous distributions or with different number of parents, (2) and (3) can be adapted accordingly.

Using only binary variables, i.e., $X_j \in \{0, 1\}$, simplifies the previous equations. More specifically, each node equation (2) collapses to θ_j and represents 2^{P_j} nonzero probabilities. In (2) we use the logit link function, resulting in classical logistic regression models for all nodes.

In the Bayesian framework, the priors are placed upon the parameters $\beta_{j,k}$, $j = 1, \ldots, n$ and $k = 0, \ldots, P_j$.



X_1 is (conditionally) independent of X_2:

$$P(X_1 = 1) = \theta_1, \quad \log\left(\frac{\theta_1}{1-\theta_1}\right) = \beta_{1,0}$$

X_2 is (conditionally) independent of X_1:

$$P(X_2 = 1) = \theta_2, \quad \log\left(\frac{\theta_2}{1-\theta_2}\right) = \beta_{2,0}$$

X_3 is conditionally dependent jointly upon X_1 and X_2:

$$P(X_3 = 1 \mid x_1, x_2) = \theta_3, \quad \log\left(\frac{\theta_3}{1-\theta_3}\right) = \beta_{3,0} + \beta_{3,1}x_1 + \beta_{3,2}x_2$$

X_4 is conditionally dependent upon X_3:

$$P(X_4 = 1 \mid x_3) = \theta_4, \quad \log\left(\frac{\theta_4}{1-\theta_4}\right) = \beta_{4,0} + \beta_{4,1}x_3$$

Fig. 1 Left: A four node Bayesian network with structure $\mathcal{S} = (V, E)$, where $V = \{1, \ldots, 4\}$, $E = \{(1, 3), (2, 3), (3, 4)\}$, e.g., $\mathbf{Pa}_1 = \mathbf{Pa}_2 = \varnothing$, $\mathbf{Pa}_3 = \{1, 2\}$ and $\mathbf{Pa}_4 = \{3\}$. Right: ABN parametrisation for binary nodes based on the logit link function. Note that the probabilities θ_3 and θ_4 depend on the configuration of the parents

3 Potential Limitation When Fitting ABN Models

From an applied perspective when fitting ABN models, we need to be aware of two possible issues: data separation and Lindley's paradox. The former is linked to the data and the latter is inherently linked to the fitting procedure. An adequate prior could potentially control both of their impacts on the fitted structure.

3.1 Data Separation

The data separation arises when a linear combination of predictors perfectly predicts the outcome. This is surprisingly common in applied logistic regression. Data separation induces estimation problems for the entire model, not only for the parameters directly involved.

Due to the large number of models necessary to evaluate (for each node a GLM for each parent configuration), data separation is a serious concern when modelling discrete data with an ABN model. The separation occurs when the dataset is too small to observe events with low probabilities. Therefore, the smaller the sample size, the higher the probability of not observing given instances which have a low probability. The issue is intensified with increasing complexity of the model. A popular solution is to remove predictors until the design matrix becomes fully ranked. However, this often leads to the deletion of the strongest predictors, which is not desirable, especially in the context of ABN [30]. Alternatively, the natural "Bayesian" solution is to use a prior that will drive the posterior whenever data separation arises. Multiple prior distributions have been proposed to tackle this issue. A notable one is the Jeffreys prior [5] which is, however, hard to interpret as prior information. Indeed, the Jeffreys prior is not parametrised on the scale of the parameter. Moreover, when applied to

sparse data the prior may lead to poor numerical results. As a result, dedicated priors have been developed which are weakly informative enough to be used in a general context and which can still drive the posterior if separation arises [8]. They have been designed to produce stable and regularised estimates. These priors are based on the Student's *t*-distribution. This paper compares the effect of the currently implemented prior in the *abn* R package, which is non-informative with the weakly informative prior promoted by [8].

3.2 Lindley's Paradox

A common approach to fitting an ABN model is using a score-based method. A popular choice is to take a score that is decomposable, i.e., each node of the network has a score contribution which is additive. The score of the total network is the sum of each individual atomic network. Typically a cache of scores is pre-computed and an optimisation algorithm is applied to select the structure that contains all nodes and maximises the score, i.e., choosing a global structure. The cache-building essentially implies fitting all possible combinations of the random variables. Even for very limited numbers of random variables, approximations are required to render the computation tractable.

Under the viewpoint of building a global structure based on the nodes' individual models, ABN can be viewed as a model selection technique. It is known that when a weakly informative prior is used, Bayesian model selection will asymptotically always prefer the simpler model, regardless of the data. This is called Lindley's paradox [21]. Using a weakly informative prior for the parameters leads to reasonable parameter estimates compared to a pure maximum likelihood estimation for a given network. But the main objective of ABN analysis is performing structural inference, which is precisely negatively affected by weakly informative priors. So, Lindley's paradox is potentially a massive threat for an ABN analysis. Indeed, the quality of the inferred structure relies on the paradigm that the score should be representative (as a proxy) of the ability of a given structure to fit the data. If a systematic asymptotic error is made regarding the scoring of the atomic structures, then the final model selected will tend to be too simple compared to an ABN selection using frequentist computed scores. Controlling the final complexity of the model through the adequate choice of the parameter prior is a major challenge when using Bayesian model selection approach applied to ABN models.

4 Implementation and Simulation Study

From a practical perspective, computational speed is the major concern in an ABN context. Indeed the number of models to be evaluated is a function of the number of nodes. This function is super-exponential with the number of random variables [27].

The estimation of Bayesian regression coefficients using Gibbs or Metropolis algorithms is usually not fast enough, especially because the model selection approach is based on a point estimate of the posterior rather than on the full network information. An appealingly fast and reliable procedure to fit the model and return an approximate posterior mode is described in [8] and a detailed explanation is given in the Sect. 14.8 of [7]. The procedure is an alteration of the classical iterative reweighted least squares algorithm that uses an approximate expectation-maximisation algorithm to update the regression coefficients at each step. The prior information is taken into account through augmented data. This procedure is used to estimate the posterior mode for every possible combination of all the variables. The output of this procedure is a comprehensive list of scores. Further details for this first step are given in [16]. In a second step, an exact search is performed to select the network with the highest possible global score [14]. The simulation study has been carried out using the package *abn* [15] in the R software environment [26].

4.1 Data Separation

In order to illustrate the influence of the prior on an ABN analysis, we randomly simulate BNs consisting of 10 binary random variables with 80% of the possible edges expressed. Each edge represents the same regression coefficient (β coefficients in the right panel of Fig. 1) set to 0.99 on the logit scale, i.e., =expit(5). For sample sizes $N = 100, 500, 1000$ and 10,000 we randomly generate 50 distributions of the selected network. The two priors used are a weakly informative prior (WI) which is a normal distribution with mean zero and variance 1000 and a Student's t-prior (ST) with one degree of freedom (i.e., Cauchy) and scale parameter 2.5. The simulation results are not very sensitive to the scale parameter within the range of 1.5–3. Then the true positive rate (TPR) and the false positive rate (FPR) are used to measure the accuracy of the selected networks. Every selected network is transformed to an essential graph, as two networks of the same Markov class of equivalence could differ substantially in terms of structure but have the same score because they represent the same assertions of conditional independence [6]. Indeed, the implemented scoring approach can differentiate networks up to the Markov class of equivalence only.

Figure 2 shows the TPR and FPR as a function of the sample size for two different priors and illustrates that both priors exhibit a proper "asymptotic" behaviour when sample size increases: TPR and FPR tend to 100% and 0%, respectively. The chosen coefficients (0.99) of the edges in each BN leads almost surely to data separation for most of the possible variables' combinations. Not surprisingly, the Student's t-prior has a better accuracy for network scoring for selecting both positive and negative edges.

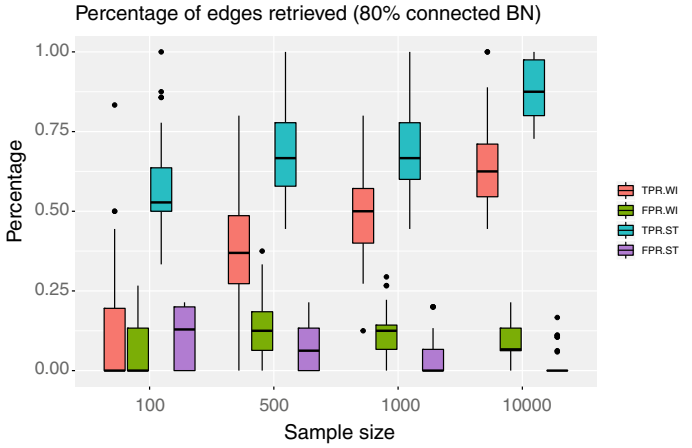


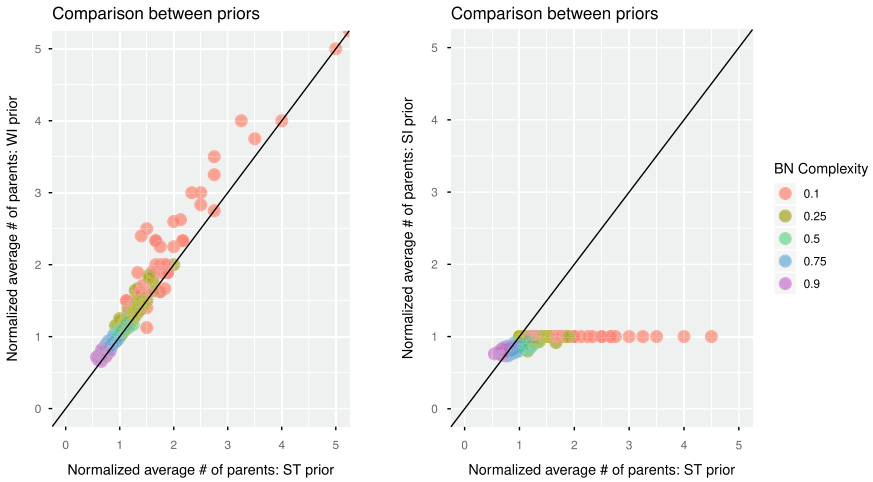
Fig. 2 Accuracy measures for retrieved edges for 80%-connected ten nodes ($n = 10$) simulated Bayesian networks as a function of data sample size ($N = 100$, 500, 1000 and 10,000). The boxplots (each based on 50 simulations) show the true positive rate (TPR) and the false positive rate (FPR) outcome of the weakly informative prior (WI) and Student's t-prior (ST)

4.2 Lindley's Paradox

An ABN modelling approach is based on multiple approximations. The network score is a proxy for selecting the best network. This score is conveniently chosen as decomposable, i.e., the marginal likelihood in this case. The procedure to compute the marginal likelihood is subject to numerical approximations. So even if Lindley's paradox is a known theoretical concern, it could potentially have a limited impact in practice.

In order to illustrate Lindley's paradox in a plausible situation, we randomly simulate BNs of $n = 10$ nodes with a range of different edge densities. Each edge has a known regression coefficient (β coefficients in the right panel of Fig. 1). Then, we simulate 50 synthetic datasets of 1000 observations per network density. For this simulation study three priors have been used: the two priors described above and a strongly informative prior (SI), which is a normal distribution with its mean set to the true regression coefficient for each edge and variance 0.1. Of course, this last prior is not realistic in practice but it is added here to illustrate the "asymptotic" behaviour. The average normalised number of parents is used to illustrate Lindley's paradox. For this illustration, we divide the average number of a simulated network by the true number of parents of the original network. Then, BNs are fitted using binomial regression with different priors, and the essential graphs are extracted.

Figure 3 summarises the simulation result and compares the (normalised) average number of parents of the fitted BN under different priors. If the selected DAGs are subjected to Lindley's paradox due to the weakness of prior information, one should see a scatter plot deviating from the diagonal. Indeed, the Lindley's paradox implies



(a) Student's t-prior (ST) and the weakly informative prior (WI).

(b) Student's t-prior (ST) and the strongly informative prior (SI).

Fig. 3 Comparison of different priors for different network complexities (edge densities varying between 0.1 and 0.9)

that the selected model tends to be too simple. Then the normalised average number of parents tend to be higher than one and should depend on the prior used. As seen in Fig. 3, sparse networks, i.e. low network complexity, are more impacted by than highly connected ones. The marginal posterior likelihood seems to overfit the sparse network structure and to underfit dense networks. In Fig. 3a one can see that a weakly informative prior performs comparably as Student's t-prior, whereas the effect of highly informative prior is clearly visible in Fig. 3b. The selected networks almost never exceed the true networks in term of complexity. Surprisingly, even such a prior does not allow the scoring procedure to optimally select dense networks. Indeed, one could expect that such amounts of information would lead to perfect scoring and thus quasi no sampling error. Again, with 1000 observations, the simulation results are not very sensitive to the scale parameter.

5 Future Developments

In Sect. 4 we showed that parameter priors play a major role in ABN modelling by (i) comparing the effect of different priors on data separation when dealing with discrete data and (ii) selecting networks depending on the prior information to address the Lindley's paradox conditional to the structure complexity. The simulation study highlights the need to further study suitable priors for ABN modelling. Indeed, the presented priors are not fully satisfactory. They partially improve the situation in

the discrete case, but as mentioned in the introduction, an ABN modelling usually involves a mixture of distributions.

A conjugate prior that contains enough information regarding the data would lead to evident benefits. For example, a closed-form distribution for the posterior might be available. This result would lead to huge advantages in terms of marginal likelihood computation by reducing the time for the structure selection process. Similarly, the parameters estimates will also benefit from this choice. Another common issue of the BN literature is the so-called score equivalence problem [11, 23] that could be solved using an appropriate prior.

In order to achieve this goal, we consider the link between ABN models and GLMs and exploit features of the exponential family. A good candidate for this purpose is the conjugate prior distribution that belongs to a flexible family of priors called the Diaconis–Ylvisaker conjugate priors [2]. This prior distribution was introduced by [1]. A change of variables and the resulting properties need to be checked (as in [9]) in order to apply this distribution to our specific case. Further work will be conducted in this direction in order to formally verify all the desirable assumptions. Additionally, the R package *abn* [15] should be equipped with further priors for practical usage and availability for the statistical community.

References

1. Chen, M., Ibrahim, J.G.: Conjugate priors for generalized linear models. Statistica Sinica **13**, 461–476 (2003)
2. Diaconis, P., Ylvisaker, D.: Conjugate priors for exponential families. Ann. Stat. **7**(2), 269–281 (1979)
3. Djebbari, A., Quackenbush, J.: Seeded Bayesian networks: constructing genetic networks from microarray data. BMC Syst. Biol. **2**(1), 57 (2008)
4. Dojer, N., Gambin, A., Mizera, A., Wilczyński, B., Tiuryn, J.: Applying dynamic Bayesian networks to perturbed gene expression data. BMC Bioinform. **7**(1), 249 (2006)
5. Firth, D.: Bias reduction of maximum likelihood estimates. Biometrika **80**(1), 27–38 (1993)
6. Flesch, I., Lucas, P.J.: Markov equivalence in Bayesian networks. In: Lucas, P., Gámez, J.A., Salmerón, A. (eds.) Advances in Probabilistic Graphical Models, pp. 3–38. Springer, Berlin, Heidelberg (2007)
7. Gelman, A., Stern, H.S., Carlin, J.B., Dunson, D.B., Vehtari, A., Rubin, D.B.: Bayesian Data Analysis. Chapman and Hall/CRC (2013)
8. Gelman, A., Jakulin, A., Pittau, M.G., Su, Y.S.: A weakly informative default prior distribution for logistic and other regression models. Ann. Appl. Stat. **2**(4), 1360–1383 (2008)
9. Gutiérrez-Peña, E., Smith, A.F.M.: Conjugate parameterizations for natural exponential families. J. Am. Stat. Assoc. **90**(432), 1347–1356 (1995)
10. Hartnack, S., Springer, S., Pittavino, M., Grimm, H.: Attitudes of Austrian veterinarians towards euthanasia in small animal practice: impacts of age and gender on views on euthanasia. BMC Vet. Res. **12**(1), 26 (2016)
11. Heckerman, D., Geiger, D., Chickering, D.M.: Learning Bayesian networks: the combination of knowledge and statistical data. Mach. Learn. **20**(3), 197–243 (1995)
12. Hodges, A.P., Dai, D., Xiang, Z., Woolf, P., Xi, C., He, Y.: Bayesian network expansion identifies new ROS and biofilm regulators. PLOS One **5**(3), e9513 (2010)

13. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., Gerstein, M.: A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science **302**(5644), 449–453 (2003)
14. Koivisto, M., Sood, K.: Exact Bayesian structure discovery in Bayesian networks. J. Mach. Learn. Res. **5**(May), 549–573 (2004)
15. Kratzer, G., Pittavino, M., Ian, L.F., Lewis, I.F.: abn: an R package for modelling multivariate data using additive Bayesian networks, R package version 1.3 (2018). https://CRAN.R-project.org/package=abn
16. Kratzer G, Furrer R (2018) Information-Theoretic Scoring Rules to Learn Additive Bayesian Network Applied to Epidemiology. arXiv:1808.01126
17. Lewis, F.I.: Bayesian networks as a tool for epidemiological systems analysis. In: AIP Conference Proceedings vol. 1493, pp. 610–617 (2012)
18. Lewis, F.I., Brülisauer, F., Gunn, G.J.: Structure discovery in Bayesian networks: an analytical tool for analysing complex animal health data. Prev. Vet. Med. **100**(2), 109–115 (2011)
19. Lewis, F.I., McCormick, B.J.: Revealing the complexity of health determinants in resource-poor settings. Am. J. Epidemiol. **176**(11), 1051–1059 (2012)
20. Lewis, F.I., Ward, M.P.: Improving epidemiologic data analyses through multivariate regression modelling. Emerg. Themes Epidemiol. **10**(1), 4 (2013)
21. Lindley, D.V.: A statistical paradox. Biometrika **44**(1/2), 187–192 (1957)
22. Pitman, E.J.G.: Sufficient statistics and intrinsic accuracy. Math. Proc. Camb. Philos. Soc. **32**(4), 567–579 (1936)
23. Pittavino, M.: Additive Bayesian networks for multivariate data: parameter learning, model fitting and applications in veterinary epidemiology. Ph.D. thesis, University of Zurich (2016)
24. Pittavino, M., Dreyfus, A., Heuer, C., Benschop, J., Wilson, P., Collins-Emerson, J., Torgerson, P.R., Furrer, R.: Comparison between generalized linear modelling and additive Bayesian network; identification of factors associated with the incidence of antibodies against Leptospira interrogans sv Pomona in meat workers in New Zealand. Acta Trop. **173**, 191–199 (2017)
25. Poon, A.F.Y., Lewis, F.I., Pond, S.L.K., Frost, S.D.W.: Evolutionary interactions between N-linked glycosylation sites in the HIV-1 envelope. PLOS Comput. Biol. **3**(1), e11 (2007)
26. R Core Team: R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2017)
27. Robinson, R.W.: Counting unlabeled acyclic digraphs. In: Little, C.H.C. (ed.) Combinatorial Mathematics V, pp. 28–43. Springer, Berlin, Heidelberg (1977)
28. Sanchez-Vazquez, M.J., Nielen, M., Edwards, S.A., Gunn, G.J., Lewis, F.I.: Identifying associations between pig pathologies using a multi-dimensional machine learning methodology. BMC Vet. Res. **8**(1), 151 (2012)
29. Ward, M.P., Lewis, F.I.: Bayesian graphical modelling: applications in veterinary epidemiology. Prev. Vet. Med. **110**(1), 1–3 (2013)
30. Zorn, C.: A solution to separation in binary response models. Polit. Anal. **13**(2), 157–170 (2005)