Additive Bayesian networks for an epidemiological analysis of swine diseases

Reti Bayesiane additive per un'analisi epidemiologica di malattie suine

Marta Pittavino and Reinhard Furrer

Abstract Additive Bayesian networks (ABNs) are types of graphical models that extend the usual generalized linear model (GLM) to multiple dependent variables through the representation of joint probability distribution. Thanks to their flexible properties, ABNs have been widely used in epidemiological analyses. In this work we present a veterinary case study where ABNs are used to explore multivariate swine diseases data of medical relevance. We then compare the results with a classical methodology. Finally, we highlight the key difference between a multivariable standard (GLM) and a multivariate (ABN) approach: the latter attempts not only to identify statistically associated variables, but also to additionally separate these into those directly and indirectly dependent with one or more outcome variables.

Abstract *Le reti Bayesiane additive (ABNs) sono tipi di modelli grafici che estendono l'usuale modello lineare generalizzato (GLM) a variabili multiple dipendenti attraverso la rappresentazione della distribuzione di probabilità congiunta. Grazie alle loro proprietà flessibili, le reti ABNs sono state ampiamente utilizzate nelle analisi epidemiologiche. In questo lavoro presentiamo un caso di studio veterinario in cui il metodo ABN viene utilizzato per esplorare dati multivariati su malattie suine di rilevanza medica. In seguito confrontiamo i risultati con una metodologia classica. Infine, evidenziamo la differenza chiave tra un approccio standard multivariabile (GLM) e uno multivariato (ABN): quest'ultimo tenta non solo di identificare le variabili associate statisticamente, ma anche di separarle ulteriormente in quelle direttamente e indirettamente dipendenti con una o più variabili d'esito.*

Key words: abn R package; data mining; machine learning; statistical modeling; structure discovery; zoonosis.

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

Reinhard Furrer
Department of Mathematics (I-MATH) and Department of Computational Science, University of Zurich (UZH), Zurich, Switzerland e-mail: reinhard.furrer@math.uzh.ch

1 Introduction

A primary objective of many epidemiological studies is to investigate hypothesized relationships between covariates of interest, and one and more outcome variables, through analyses of appropriate data. From a data analysis perspective, this is often far from being trivial. Diseases and health conditions, which are a priority for control or eradication in humans and animals, are increasingly recognized to have highly complex determinants. Typically, the unknown stochastic processes, which generated these data, are highly complex, resulting in multiple correlation/dependencies between covariates and between outcome variables. Standard epidemiological and statistical approaches cannot adequately describe such inter-dependent multifactorial relationships. ABN modelling is a data mining/machine learning methodology, which has demonstrated to be ideally suited for such analyses [1, 2].

2 Material and methods

2.1 The data

We present data on disease occurrence in pigs provided by the industry body 'British Pig Health Scheme' (BPHS). The main objective of BPHS is to improve the productivity of pig in the UK, and reducing disease occurrence is a significant part of this process. The data we consider here comprise of a randomly chosen batch of 50 pigs from each of 500 randomly chosen pig producers in the UK. In total we deal with 25'000 observations, i.e. animals entering the human food chain at an abattoir: 'finishing pigs'. Each animal is assessed for the presence of a range of different disease conditions by a specialist swine veterinarian.

Then, the resulting variables are binary due to the presence or the absence of a specific disease. We consider here the following ten disease conditions, all abbreviated to ease the notation and described in [3]: enzootic-pneumonia (EP); pleurisy (PL); milk spots (MS); hepatic scarring (HS); pericarditis (PC); peritonitis (PT); lung abscess (AB); tail damage (TD); pyaemia (PY) and papular dermatitis (PD).

The presence of any of these conditions results in an economic loss to the producer. Either directly due to the relevant infected part of the animal being removed from the food chain, or indirectly in cases such as enzootic pneumonia, which may potentially indicate poor herd health and efficiency losses on the farm. An additional loss, though not directly monetary, is the presence of tail damage which may be suggestive of welfare concerns and linked to sub-optimal production efficiency. Milk spots and hepatic scarring result from infestation with Ascaris suum, which is particularly important as this is a zoonotic helminth parasite.

2.2 Additive Bayesian networks

A Bayesian network for a set of random variables $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ consists of:

- A *directed acyclic graph* (DAG) structure $\mathscr{S} = (V, E)$, where V is a finite set of vertices or nodes and E is a finite set of directed edges between the vertices. A

DAG is *acyclic*; hence, the edges in E do not form directed cycles. A random variable X_j corresponds to each node $j \in V = \{1, \dots, n\}$ in the graph. We do not distinguish between a variable X_j and the corresponding node j.

- A set of parents for a node j is denoted by $\mathbf{Pa}_j$. A node j is said to be a *parent* of a node k if the edge set E contains an edge from j to k. P_j indicates the total number of parents for a node $j:\ \dim(\mathbf{Pa}_j) = P_j \geq 0$. $P_j = \emptyset$ for orphan nodes.
- A set of local probability distributions for all variables in the network called $\boldsymbol{\theta}_{\mathscr{B}}$. Each node j, with parent set $\mathbf{Pa}_j$, is parametrized by a local probability distribution: $P(X_j|\mathbf{Pa}_j)$.

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 factorization of the global probability distribution: $P(\boldsymbol{X}) = \prod_{j=1}^{n} P(X_j|\mathbf{Pa}_j)$.

We denote a Bayesian network (BN) model $\mathscr{B}$ for a set of random variables $\boldsymbol{X}$ by a pair $\mathscr{B} = (\mathscr{S}, \boldsymbol{\theta}_{\mathscr{B}})$. The DAG $\mathscr{S}$ defines the *structure*, and $\boldsymbol{\theta}_{\mathscr{B}}$ the *parametrization* of the model. In order to specify a $\mathscr{B}$ for $\boldsymbol{X}$, we must therefore specify a DAG structure and a set of local probability distributions.

An additive Bayesian network $\mathscr{A}$ consists of a Bayesian network $\mathscr{B}$ that generalizes the multinomial logistic regression model $\mathscr{M}$. The multinomial logistic regression model $\mathscr{M}$ can be integrated into a BN $\mathscr{B}$ by modelling each of its conditional probability table $P(X_j = s \mid \mathbf{Pa}_j = c) = \theta_{jcs}$ with a multinomial logistic regression model, where X_j is progressively the outcome variable and the resulting regression design matrix is constructed from $\mathbf{Pa}_j$, as showed in [4] and in detail in Figure 1.

X_1 is independent: $logit(\theta_1) = \beta_{1,0}$

X_2 is independent: $logit(\theta_2) = \beta_{2,0}$

X_3 is jointly dependent upon X_1, andX_2: $logit(\theta_3) = \beta_{3,0} + \beta_{3,1}X_1 + \beta_{3,2}X_2$

X_4 is conditionally dependent upon X_3: $logit(\theta_4) = \beta_{4,0} + \beta_{4,1}X_3$

X_5 is conditionally dependent upon X_3: $logit(\theta_5) = \beta_{5,0} + \beta_{5,1}X_3$

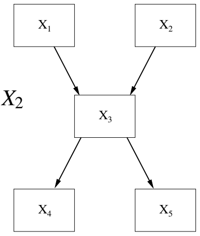

Fig. 1 A binary additive Bayesian network model $\mathscr{A}$ for five random variables.

2.3 Analysis with ABN

All analyses were conducted using the software R [5] and specifically the "abn" R package [6] which is available from CRAN "cran.r-project.org" with additional documentation and case studies at "http://www.r-bayesian-networks.org".

Prior distributions were defined. All DAG structures were equally supported a priori with a uniform, i.e., uninformative, prior. It is possible to construct informative structural priors, i.e. penalizing models with more structural complexity, but as noted in [7] these are problematic to specify or lead to undesirable properties as in

[9]. Uninformative Gaussian priors were applied for the additive parameters at each node: specifically, independent Gaussian priors with mean zero and variance 1000.

As we are searching across DAGs - to identify optimally fitting structures - there is also the need for a prior on structures. The default being that each structure is equally supported a priori. It is possible to construct informative structural priors, for example to penalize models with more structural complexity, e.g. more arcs, but as noted in [6] these are problematic to specify in practice. In [8] an informative structural prior on the number of parents within an individual node is used, where this assumes that parent combinations with the same cardinality are equally likely. This prior gives equal weighting to a parent combination with cardinality zero and cardinality m1 which may not be entirely desirable. In the subsequent case study analyses an uninformative - flat - structural prior is used.

A two-steps procedure was used to identify a robust model.

The first step was to find an optimal ABN model $\mathscr{A}_1$. The process of identifying an optimal ABN is referred to in the literature as *structure learning* [8]. This was found with an order based exact search method [9]. The best goodness of fit to the available data was computed using the marginal likelihood (ML), equivalent to Bayes factors for models with equal structural priors and the standard Bayesian score function used in BN literature [7, 8]. The ML includes an implicit penalty for model complexity and in a binary additive Bayesian network for a node j is:

$$P(\mathscr{D}_j|\mathscr{S}) = \int_{-\infty}^{+\infty} \prod_{i=1}^{m} \left(\frac{e^{z_{ij}^T \boldsymbol{\beta}_j}}{1+e^{z_{ij}^T \boldsymbol{\beta}_j}} \right)^{x_{ij}} \left(\frac{1}{1+e^{z_{ij}^T \boldsymbol{\beta}_j}} \right)^{1-x_{ij}} \times \prod_{c=1}^{C_j} \frac{1}{\sqrt{2\pi}\sigma_c} e^{-\frac{(\beta_c-\mu_c)^2}{2\sigma_c^2}} \, d\boldsymbol{\beta}_j$$

where $\mathscr{D}_j$ are the observed data at node j, and consist of tuples of $[x_{ij}, z_{ij}^T]$. The parameter vector at node j is represented by $\boldsymbol{\beta}_j$, and has the same length as the possible parents configuration, denoted by C_j, then $\dim(\boldsymbol{\beta}_j) = C_j$. The marginal likelihood was estimated using the Laplace approximation at each node. To find the best model, the maximum number of parents allowed per node (number of covariates in each regression model at each node) was increased until the goodness of fit remained constant and thereby identified the same globally optimal ABN. The model selection procedure started from one possible parent per node and then the parent limit increased gradually until five possible parents per node [6].

In the second step, the model $\mathscr{A}_1$ was adjusted by checking it for overfitting using Markov chain Monte Carlo (MCMC) simulation implemented in JAGS (just another Gibbs sampler) [10]. A parametric bootstrapping approach was suggested in [8] which uses simulation to assess whether a chosen model comprises more complexity than could reasonably be justified given the observed data. Simulated datasets were generated with MCMC as iterations of an identical size as the original one, from the optimal model found in step one. An identical exact search for an optimal model structure was then performed exactly as in the first step, but applied to the bootstrapped data rather than original data. It was repeated 10240 times [6], a large enough number to get robust results, using the same parent limit per node as the one found in the initial search. Arcs present in less than 50% of the globally optimal ABNs - estimated from the bootstrapped data - were considered not to be robust and removed from the DAG generated in the first step. A most robust ABN model $\mathscr{A}_2$ fully adjusted for over-fitting was identified at the end of this second step, equivalent to a multivariate GLM.

2.4 Analysis with GLM

Data were analysed using the software R [5] and the "glm" function, available in the "stats" R package . As many different generalized linear models (GLMs), in particular multivariable logistic regression following the data structure, as the different number of variables were performed. Only two models based on the most significant variables and with the highest AIC score have been selected and shown.

3 Results with ABN and GLM

The resulting best fitting ABN comprised 12 arcs and a maximum number of three parents (Fig. 8 in [6]), for the variable PD (papular dermatitis) and PL (pleurisy). After the bootstrap analysis, four of the arcs in the globally optimal ABN were only weakly supported. Therefore the number of arcs was reduced from 12 to 8 (Fig. 7 in [6]). The final globally optimal additive Bayesian network model after adjustment for over-fitting is shown in Fig. 2 on the left. Three different epidemiological pathways can be identified resulting in variables indirectly linked together. The goodness of fit for this model is -44245.73. For example, forcing an additional arc connecting PC and AB gives a poorer log marginal likelihood of -44249.58.

The two GLM models with the most significant variables and the highest AIC score have PC (pericarditis) and AB (lung abscess) as response variables. Figure 2, in the middle and on the right, shows the two corresponding models. A GLM is simply a DAG where arcs are only allowed directly between the covariates and response variable. In each case we find that an arc is identified between PC and AB.

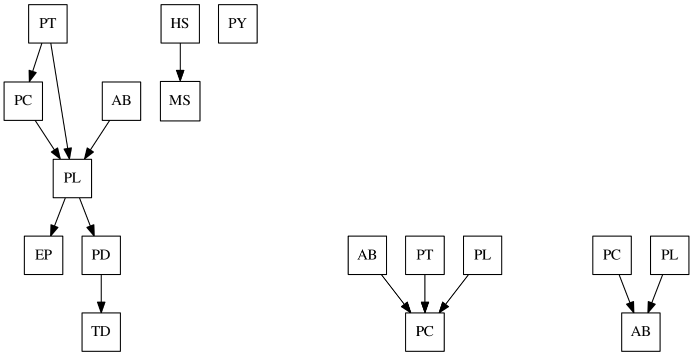

Fig. 2 Final ABN model of swine diseases data with 8 arcs after bootstrapping adjustment (left). Two globally optimal GLMs - one with PC as the dependent variable (middle), and a second with Abscess as the dependent variable (right).

4 Conclusion

In summary, we find that while the GLM analyses identifies a strongly supported statistical association between presence of pericarditis (PC) and lung abscess (AB); the ABN model does not support a direct statistical dependency between PC and AB. In the ABN model there is no arc connecting these variables, this relationship is via the intermediate variable pleurisy (PL).

This highlights the key difference between a multivariable GLM and a multivariate GLM (ABN). The former identifies variables which may be associated with the response (dependent) variable within a very restrictive model space: arcs are only allowed from covariates direct to the response variable. When considering the same data within a larger model space, which incorporates other relationships within the underlying epidemiological system which generated the observed data, then such variables may then only be supported as indirectly, rather than directly, related to the response variable. In [2, 11] there are further similar GLM and ABN examples.

These results provides a conceptual justification of the Yule-Simpson paradox, which states that an apparent relationship between variables may disappear or even be reversed when others are taken into account.

In conclusion, data analyses using ABNs have the potential to offer new insights into complex epidemiological systems.

References

1. Pittavino, M.: Additive Bayesian Networks for Multivariate Data: Parameter Learning, Model Fitting and Applications in Veterinary Epidemiology. PhD thesis. Universität Zürich (2016)
2. Pittavino, M., Dreyfus, A., Heuer, C., Benschop, J., Wilson, P., Torgerson, P., 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 Tropica **173**, 191-199, ELSEVIER (2017)
3. Sanchez-Vazquez, M.J., Nielen, M., Edwards, S.A., Gunn, G.J. and Lewis F.I.: Identifying associations between pig pathologies using a multidimensional machine learning methodology. BMC Vet. Res. **8**:151, 1-11 (2012)
4. Rijmen, F.: Bayesian networks with a logistic regression model for the conditional probabilities, Int. Jour. of Appr. Reas. **48**:2, 659-666 (2008)
5. R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org (2017)
6. Kratzer, G., Pittavino, M., Lewis, F., Furrer, R.: abn: an R package for modelling multivariate data using additive Bayesian networks. The Comprehensive R Archive Network, 1-37 (2017)
7. Heckerman, D., Geiger, D. and Chickering, D. M.: Learning Bayesian networks: the combination of knowledge and statistical data. Machine Learning. **20**:3, 197-243 (1995)
8. Friedman, N., Goldszmidt, M., Wyner, A.: Data analysis with Bayesian networks: A Bootstrap approach, Proc. 15th Conf. on Uncert. in Artif. Intell. (UAI'99), San Francisco: Morgan Kaufmann, 196-205 (1999)
9. Koivisto, M., Sood, K. Exact Bayesian structure discovery in Bayesian networks. Jour. of Mach. Lear. Res. **5**, 549-573 (2004)
10. Plummer, M.: JAGS: a program for analysis of Bayesian graphical models using Gibbs 701 sampling. Proc. 3rd Int. Work. Dist. Stat. Comp. (DSC 2003), Vienna, Austria, 1-10 (2003)
11. Pittavino, M., Dreyfus, A., Heuer, C., Benschop, J., Wilson, P., Torgerson, P., Furrer, R.: Data on Leptospira interrogans sv Pomona in Meat Workers in New Zealand. Data in Brief **13**, 587-596, ELSEVIER (2017)