Network Model with Application to Allergy Diseases

Konrad Furmańczyk$^{1,5[0000-0002-7683-4787]}$,
Wojciech Niemiro$^{2,3[0000-0002-7076-8838]}$,
Mariola Chrzanowska$^{4,5[0000-0002-8743-7437]}$, and
Marta Zalewska$^{5[0000-0002-8163-961X]}$

[1] Institute of Information Technology, Warsaw University of Life Sciences, Warsaw, Poland konrad_furmanczyk@sggw.edu.pl
[2] Faculty of Mathematics, Informatics and Mechanics University of Warsaw, Poland
[3] Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Poland wniemiro@gmail.com
[4] Institute of Economics and Finance, Warsaw University of Life Sciences, Poland mariola_chrzanowska@sggw.edu.pl
[5] Department of Prevention of Environmental Hazards, Allergology and Immunology, Medical University of Warsaw, Poland marta.zalewska@wum.edu.pl

Abstract. We propose a new graphical model to describe the comorbidity of allergic diseases. We present our model in two versions. First, we introduce a generative model that reflects the variables' causal relationships. Then, we propose an approximation of the generative model by a misspecified model, which is computationally more efficient and easily interpretable. In both versions of our model, we consider typical allergic disease symptoms and covariates. We consider two directed acyclic graphs (DAGs). The first one describes information about the coexistence of certain allergic diseases (binary variables). The second graph describes the relationships between particular symptoms and the occurrence of these diseases. In the generative model, the edges lead from diseases to symptoms, corresponding to causal relations. In the misspecified model, we reverse the direction of edges: they lead from symptoms to diseases. The proposed model is evaluated on a cross-sectional multicentre study in Poland (www.ecap.pl). An assessment of the stability of the proposed model is obtained using the bootstrap and jackknife techniques. Our results show that the misspecified model is a good approximation of the generative model and helps predict the incidence of allergic diseases.

Keywords: Network Model · Bayesian Network · Logistic Regression · Allergy Diseases.

1 Introduction

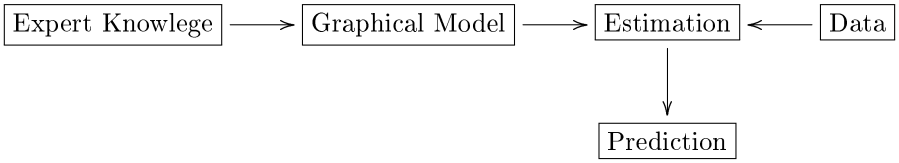
Modelling dependence between different binary variables is an essential statistical task with many applications in medicine, life sciences, economics, and sociology. The basic statistical tools used in such situations are the autologistic (AL)

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_9

model [2] and graphical network modelling [15], [1], [4]. General information on graphical models for discrete data can be found in [13] and [12]. The classical AL model [2] has been applied in epidemiology, marketing, agriculture, ecology, forestry, geography, and image analysis [6], [5], [17], [7], [11]. The most common approach to estimation of the model parameters is the pseudo-likelihood [3] method. Zalewska et al. [21] recommended a heuristic estimation method. Recently, Shin et al. [17] invented and applied an AL network model for a disease progression study using pseudo-likelihood to estimate the model parameters.

Our paper proposes a new graphical model that is related to but different from the AL model. We aim to describe the interdependence of allergic diseases in contrast to most studies that do not consider dependences between allergies [9], [20], [8].

We present our model in two versions. First, we introduce a generative model that reflects the variables' causal relationships. Then, we propose an approximation of the generative model by a misspecified model, which is computationally more efficient and easily interpretable. We focus on the misspecified version, which we consider more practical. In both versions of our model, we consider typical allergic disease symptoms, family history of allergic disease, and control variables as covariates. We consider two directed acyclic graphs (DAGs), both based on experts' knowledge. The first one describes information about the coexistence of certain allergic diseases (binary variables). The second graph describes the relationships between particular symptoms and the occurrence of these diseases. In the generative model, the edges lead from diseases to symptoms, corresponding to causal relations. In the misspecified model, we reverse the direction of edges: they lead from symptoms to diseases. This trick significantly reduces computational costs. Our model was naturally divided into separate logistic models for individual allergy diseases. Each logistic regression is estimated by the standard generalized linear model (GLM) procedure. Our general approach is very flexible and can be applied to any dependence model for binary variables. We comapre predictions based on two versions of our model. We argue that the misspecified model is a good approximation for the more logically consistent but computationally expensive generative model. The paper is organized as follows. Section 2 introduces the proposed methodology. In Section 3, we apply this methodology to construct a new model of comorbidity of allergic diseases, based on a big epidemiological data set (ECAP) [16]. At the end of this section, we present an evaluation of the proposed model. Section 4 and 5 contain discussion and conclusions. All computations are carried out with the R package (www.r-project.org). Below we provide a graphical user guide illustrating our methodology.



2 Hierarchical Logistic Network Models

2.1 Genarative model

Our proposed model contains four groups of variables. In the first group, we consider a random vector $\mathbf{Y} = (Y_1, \ldots, Y_p)^T$ with binary components. Each of these variables determines presence or absence of a given allergic disease for a patient. In our application we describe p allergic diseases. Taking into account the known co-occurrence of diseases, the relationships between them are described by a directed graph with the adjacency matrix $\mathbf{A} = (a_{ki})$ as follows: $a_{ki} = 1$ if Y_i is affected by Y_k and otherwise $a_{ki} = 0$.

In the second group, we have a random vector of symptoms of our diseases $\mathbf{S} = (S_1, \ldots, S_m)^T$. The remaining two groups consist of common factors $\mathbf{F} = (F_1, \ldots, F_l)^T$, which can affect all considered diseases (for example genetic features) and a vector of additional covariates $\mathbf{X} = (X_1, \ldots, X_r)^T$ such as gender, age, residence of a patient, etc. Symptoms S_i can be continuous or discrete random variables. It is usually known which symptoms are characteristic for each disease. This knowledge can be represented by a directed graph with adjacency matrix $\mathbf{B} = (b_{kj})$ such that: $b_{kj} = 1$ if Y_k causes S_j and otherwise $b_{kj} = 0$.

The full generative model includes diseases $\mathbf{Y}$, symptoms $\mathbf{S}$, common factors $\mathbf{F}$ and additional covariates $\mathbf{X}$. This graph has edges among $\mathbf{Y}, \mathbf{S}$ variables given by matrices $\mathbf{A}, \mathbf{B}$, and all edges leading from $\mathbf{F}, \mathbf{X}$ variables to all components of $\mathbf{Y}, \mathbf{S}$. We assume that the graph corresponding to the adjacency matrix $\mathbf{A}$ is acyclic. Consequently, the whole graph is a directed acyclic graph (DAG). The conditional probability distribution of $\mathbf{Y}, \mathbf{S}$ is given by

$$\begin{aligned} P(\mathbf{Y} = \mathbf{y}, \mathbf{S} = \mathbf{s} | \mathbf{F} = \mathbf{f}, \mathbf{X} = \mathbf{x}) = & \prod_{i=1}^{p} P(Y_i = y_i | \mathbf{Y}_{pa}(Y_i), \mathbf{F} = \mathbf{f}, \mathbf{X} = \mathbf{x}) \\ & \times \prod_{j=1}^{m} P(S_j = s_j | \mathbf{Y}_{pa}(S_j), \mathbf{F} = \mathbf{f}, \mathbf{X} = \mathbf{x}), \end{aligned} \tag{1}$$

where $\mathbf{Y}_{pa}(Y_i) = \{Y_k : Y_k \to Y_i\}$ is a set of diseases which affect the occurrence of disease Y_i, $\mathbf{Y}_{pa}(S_j) = \{Y_k : Y_k \to S_j\}$ is a set of diseases which cause symptom S_j. We assume the following parametric form of conditional distributions:

$$\log \frac{P(Y_i = 1 | \mathbf{Y}_{pa}(Y_i), \mathbf{F} = \mathbf{f}, \mathbf{X} = \mathbf{x})}{P(Y_i = 0 | \mathbf{Y}_{pa}(Y_i), \mathbf{F} = \mathbf{f}, \mathbf{X} = \mathbf{x})} = \omega_{0i} + \sum_{k=1}^{p} a_{ki} \omega_{ki} Y_k + \mathbf{x}^T \alpha_i + \mathbf{f}^T \beta_i, \tag{2}$$

$$\log \frac{P(S_j = 1 | \mathbf{Y}_{pa}(S_j), \mathbf{F} = \mathbf{f}, \mathbf{X} = \mathbf{x})}{P(S_j = 0 | \mathbf{Y}_{pa}(S_j), \mathbf{F} = \mathbf{f}, \mathbf{X} = \mathbf{x})} = \gamma_{0j} + \sum_{k=1}^{p} b_{kj} \gamma_{kj} Y_k + \mathbf{x}^T \delta_j + \mathbf{f}^T \epsilon_j. \tag{3}$$

We thus have the following model parameters: $\omega_{0i} \in R, \omega_{ki} \in R, \alpha_i \in R^r, \beta_i \in R^l, \gamma_{0j} \in R, \gamma_{kj} \in R, \delta_j \in R^r, \epsilon_j \in R^l$. Since the conditional probability (1) consists of the product of $p + m$ probabilities, the parameters of each factor can

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_9

be estimated separately by a standard logistic regression procedure. To improve prediction accuracy we also applied weighted logistic regression. However, the results obtained by both methods were almost identical (Supplement [19]: C3-C4).

2.2 Misspecified model

Unfortunately, the model presented in the previous subsection is computationally demanding, and its parameters are difficult to interpret. We propose using another, misspecified model that does not reflect causal relations between variables but is computationally more accessible for a big network and has parameters with simple, intuitive meaning. We change the direction of edges joining symptoms and diseases. Entries of adjacency matrix **B** will now be interpreted as follows: $b_{ij} = 1$ indicates the presence of arrow $Y_i \leftarrow S_j$. We assume that the remaining edges of the graph are the same as in the generative model. In the misspecified model, equation (1) is replaced by equation (4), and equations (2)-(3) are replaced by equation (5) as follows:

$$P(\mathbf{Y} = \mathbf{y}|\mathbf{S}, \mathbf{F}, \mathbf{X}) = \prod_{i=1}^{p} P(Y_i = y_i|\mathbf{Y}_{pa}(Y_i), \mathbf{S}_{pa}(Y_i), \mathbf{F}, \mathbf{X}), \tag{4}$$

where $\mathbf{S}_{pa}(Y_i) = \{S_j : Y_i \leftarrow S_j\}$ is a set of symptoms related to occurrence of disease Y_i. Similarly as in generative model, we assume a log-linear form of conditional distributions. To simplify notation, we use the same symbols for the parameters for both models.

$$\begin{aligned}\log \frac{P(Y_i = 1|\mathbf{Y}_{pa}(Y_i), \mathbf{S}_{pa}(Y_i), \mathbf{F} = \mathbf{f}, \mathbf{X} = \mathbf{x})}{P(Y_i = 0|\mathbf{Y}_{pa}(Y_i), \mathbf{S}_{pa}(Y_i), \mathbf{F} = \mathbf{f}, \mathbf{X} = \mathbf{x})} &= \omega_{0i} + \sum_{k=1}^{p} a_{ki}\omega_{ki}Y_k \\ &+ \sum_{j=1}^{m} b_{ij}\gamma_{ij}S_j + \mathbf{x}^T\alpha_i + \mathbf{f}^T\beta_i.\end{aligned} \tag{5}$$

3 Application to Modelling Allergic Diseases

In this section we apply the proposed approach to investigate the prevalence of allergic diseases and their interdependences. Our model is based on a big epidemiological study in Poland (ECAP) [16]. More details can be found in the Supplement [18]-Section A.

3.1 The structure of the model

The first group of variables consists of 5 selected allergic diseases Y_1, Y_2, Y_3, Y_4, Y_5. The left panel of Figure 1 illustrates the dependences between them, based on the literature and on discussions with medical doctors [10], [14]. The second group

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_9

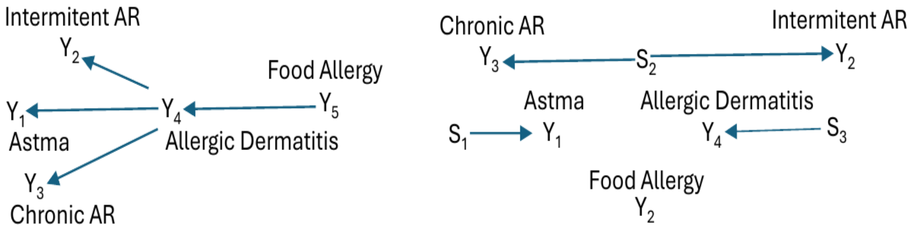


Fig. 1. The Graphs with adjacency matrices $\mathbf{A}, \mathbf{B}$

of variables consists of typical symptoms of those diseases: S_1, S_2, S_3. Additionally we consider history of allergy diseases in the family: F_1, F_2, F_3, F_4, F_5. The right panel of Figure 1 shows dependences between allergic diseases and their symptoms. The direction of arrows in Figure 1 lead from symptoms to diseases which corresponds to the misspecified model. In the last group of variables, we consider control covariates: X_1, X_2, X_3, X_4 (they decsribe age, gender, residence of patients). More detailed description of all the variables can be found in the Supplement [18]-Section B).

3.2 Generative and misspecified models of allergy diseases

We recall the generative model in which diseases cause symptoms. Taking into account the structure of the graph with adjacency matrices $\mathbf{A}, \mathbf{B}$, we see that, conditionally on covariates $\mathbf{F}$ and $\mathbf{X}$, the conditional distribution of $\mathbf{Y}$ given symptoms $\mathbf{S}$ has the form

$$P(Y_1|Y_2, Y_3, Y_4)P(Y_2|Y_4)P(Y_3|Y_4)P(Y_4|Y_5)P(Y_5)P(S_1|Y_1)P(S_2|Y_2, Y_3)P(S_3|Y_4).$$

(We omitted $\mathbf{F}$ and $\mathbf{X}$ in this formula).

Now we turn to the misspecified model. Conditionally on covariates $\mathbf{F}$ and $\mathbf{X}$, the joint probability $P(\mathbf{Y}, \mathbf{S})$ is determined as:

$$P(\mathbf{Y}|\mathbf{S}) = P(Y_1|Y_2, Y_3, Y_4, S_1)P(Y_2|Y_4, S_2)P(Y_3|Y_4, S_2)P(Y_4|Y_5, S_3)P(Y_5).$$

We now formulate specific equations restricting attention to the misspecified model only. We assume the logistic form of the conditional probabilities (formulas (4)-(5)). We estimate each of them separately using standard R function 'glm'. The subsequent equations concern the logits for asthma Y_1, intermittent allergic rhinitis Y_2, chronic allergic rhinitis Y_3, allergic dermatitis Y_4. The equations are:

$$logit_1 = \omega_{01} + \sum_{j=1}^{4} \alpha_{j1} X_j + \sum_{j=1}^{5} \beta_{j1} F_j + \gamma_{11} S_1 + \sum_{j=2}^{4} \omega_{j1} Y_j.$$

$$logit_2 = \omega_{02} + \sum_{j=1}^{4} \alpha_{j2} X_j + \sum_{j=1}^{5} \beta_{j2} F_j + \gamma_{22} S_2 + \omega_{42} Y_4.$$

$$logit_3 = \omega_{03} + \sum_{j=1}^{4} \alpha_{j3} X_j + \sum_{j=1}^{5} \beta_{j3} F_j + \gamma_{32} S_2 + \omega_{43} Y_4.$$

$$logit_4 = \omega_{04} + \sum_{j=1}^{4} \alpha_{j4} X_j + \sum_{j=1}^{5} \beta_{j4} F_j + \gamma_{43} S_3 + \omega_{54} Y_5.$$

3.3 Comparision of two versions of our model

We compute the 'diagnostic' probabilities of diseases given symptoms for the generative and the misspecified model. It is worth noting that in the case of a large network, it would not be possible to calculate $P(\mathbf{Y}|\mathbf{S},\mathbf{F},\mathbf{X})$ or $P(Y_i|\mathbf{S},\mathbf{F},\mathbf{X})$ exactly in the generative model. In this situation, the misspecified model has an advantage over the generative model. The two models can be compared in the case of a small network as that considered here.

We consider five scenarios (different values of of covariates $\mathbf{X},\mathbf{F}$, symptoms $\mathbf{S}$ and coexistent deseases Y_i). Let $p_1 = P(Y_1 = 1|Y_2 = 0, Y_3 = 0, Y_4 = 0, S_1), q_1 = P(Y_1 = 1|Y_2 = 1, Y_3 = 1, Y_4 = 1, S_1), p_2 = P(Y_2 = 1|Y_4 = 0, S_2), q_2 = P(Y_2 = 1|Y_4 = 1, S_2), p_3 = P(Y_3 = 1|Y_4 = 0, S_2), q_3 = P(Y_3 = 1|Y_4 = 1, S_2), p_4 = P(Y_4 = 1|Y_5 = 0, S_3), q_4 = P(Y_4 = 1|Y_5 = 1, S_3)$. The results for the first two scenarios are presented in Tables 1 (all 5 scenarios are given in the Supplement [18]-Section C). The difference between the two models obtained is negligible.

Table 1. Comparison between the generative model and misspecified model

Scenario	Model	p_1	p_2	p_3	p_4	q_1	q_2	q_3	q_4
1	generative	0.021	0.081	0.077	0.024	0.597	0.104	0.134	0.024
	misspecified	0.023	0.085	0.080	0.015	0.566	0.097	0.125	0.044
2	generative	0.103	0.282	0.322	0.208	0.886	0.326	0.461	0.524
	misspecified	0.088	0.270	0.307	0.081	0.842	0.299	0.421	0.216

3.4 Estimation of parameters and evaluation of the model

We report the estimated coefficents of logistic regression, their standard errors, the odds ratios with confidence intervals (CI) in the Supplement [18]-Section C. The accuracy of estimators and robustness of our model is evaluated using the bootstrap and jackknife techniques. The dataset is divided into a learning and testing sample to assess if the proposed model is adequate. The ROC curve and average AUC on the testing sample are determined from 20 repetitions. Table 2 shows the AUC values for the averaged AUC values for bootstrap and jackknife. The ROC curves (Fig1-Fig16) are collected in the Supplement [18]-Section D as well as interpretation of the results from the medical point of view. Our results show good stability of the model.

Table 2. AUC for each logit

$logit_i$	$i = 1$	$i = 2$	$i = 3$	$i = 4$
bootstrap	0.8470	0.6986	0.7201	0.7931
jackknife	0.8165	0.6857	0.7215	0.7921

4 Discussion

Previous studies of multimorbidity in allergy [9], [20], [8], [14], [10] were based on fitting of single logistic models that did not take into account the correlations between the studied diseases. Our graphical model uses two DAGs to describe such dependences. The proposed model can be used in studies of associations of other diseases and, in general, in the study of correlations in complex systems.

5 Conclusions

Both versions of our model (generative and misspecified) produced similar results. The latter is computationally more efficient and easily interpretable. Evaluation of the model using bootstrap and jackknife techniques yielded average AUCs ranging from 0.67 to 0.84 (Table 2), indicating relatively high stability of the results. Both bootstrap and jackknife methods could be used to construct confidence intervals for the model parameters and classifacation metrics. Our model can help predict the incidence of allergic diseases and will allow for a better understanding of the complex co-occurrence of these diseases. It also sheds light on the impact of such covariates as gender, age, family history, etc. on allergic diseases. The proposed model can be easily extended by adding other potential factors influencing the occurrence of the diseases. Due to the nature of our task, we considered the low-dimensional case where the number of observations n is greater than the number of features p. Naturally, the proposed approach can be generalized to the high-dimensional case $p > n$ by adding the Lasso [19] or Ridge penalty for log-likelihood for each logit model separately. This will be the topic of further research.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abeyasinghe, P.M. et al.: Consciousness and the dimensionality of DOC patients via the generalized Ising model. J. Clin. Med., **9**(5): 1332 (2020)
2. Besag J., E.: Nearest-Neighbour Systems and the Auto-Logistic Model for Binary Data. J. R. Stat. B: Stat. Methodol, **34**(1), 75–83 (1972)
3. Besag J., E.: Statistical analysis of non-lattice data. The Statistician, **24**(3), 179—195 (1975)
4. Briganti, G., Linkowski, P.: Exploring network structure and central items of the Narcissistic Personality Inventory. Inventory. Int J Methods Psychiatr Res., 2020 Mar;29(1):e1810 (2000)
5. Caragea, P.C., Kaiser, M.S.: Autologistic models with interpretable parameters. JABES, **14**, 281—300 (2009)
6. Gégout-Petit A., Guérin-Dubrana L., Li S.: A new centered spatio-temporal autologistic regression model with an application to local spread of plant diseases. Spat. Stat. **31** 100361 (2019)
7. He, F., Zhou, J., Zhu, H.: Autologistic regression model for the distribution of vegetation. JABES **8**(2), 205–222 (2003)

8. Jung, S. et al.: Risk Factors and Comorbidities Associated With the Allergic Rhinitis Phenotype in Children According to the ARIA Classification. Allergy Asthma Immunol Res. **12**(1): 72–85 (2020)
9. Kim, H. Y. et al.: Prevalence and comorbidity of allergic diseases in preschool children. Korean J. Pediatr., **56**(8), 338—342 (2013)
10. Krzych-Fałta E., Furmańczyk K., Piekarska B., Tomaszewska A., Sybilski A., Samoliński BK.: Allergies in urban versus countryside settings in Poland as part of the Epidemiology of the Allergic Diseases in Poland (ECAP) study—challenge the early differential diagnosis. Adv Dermatol Allergol., **33**(5), 359-–368 (2016)
11. Koutsias N.: An autologistic regression model for increasing the accuracy of burned surface mapping using Landsat Thematic Mapper data. Int. J. Remote Sens., **24**(10), 2199–2204 (2003)
12. Maathuis, M., Drton, M, Lauritzen, S., Wainwright, M., eds.: Handbook of Graphical Models. Chapman & Hall/CRC Press (2019)
13. Madigan D., York J., Allard D.: Bayesian graphical models for discrete data. Int Stat Rev **63**(2), 215–232 (1995)
14. Raciborski F. et al.: Dissociating polysensitization and multimorbidity in children and adults from a Polish general population cohort. Clin. transl. allergy, 9:4 (2019)
15. Ravikumar, P., Wainwright, M. J., Lafferty, J.: High-dimensional Ising model selection using l1-regularized logistic regression. Ann. Statist. **38**, 1287—1319 (2010)
16. Samoliński B., Raciborski F., Lipiec A. et al.: Epidemiologia Chorób Alergicznych w Polsce (ECAP). Alergol Pol. **1**(1): 10–18 (2014)
17. Shin Y.E., et al.: Autologistic network model on binary data for disease progression study. Biometrics **75**(4), 1310–1320 (2019)
18. Supplement Furmańczyk K., Niemiro W., Chrzanowska M., Zalewska M.: Supplementary Material to the paper 'Network Model with Application to Allergy Diseases' (2024)
    https://github.com/kfurmanczyk/Network-_Allergy/blob/main/Supplement1.pdf
19. Tibshirani, R.: Regression shrinkage and selection via the lasso. J.R. Stat., **58**(1), 267—288 (1996)
20. Westman M. et al.: Natural course and comorbidities of allergic and nonallergic rhinitis in children. J Allergy Clin Immunol **129**(2), 403–408 (2012)
21. Zalewska M., Niemiro W., Samoliński B.: MCMC imputation in autologistic model. Monte Carlo Methods and Applications, De Gruyter, vol. 16(3-4), 421–438 (2010)