Graficzny model współwystępowania chorób alergicznych

Konrad Furmańczyk [1,3]
Wojciech Niemiro [2]
Mariola Chrzanowska [1,3]
Marta Zalewska [3]

Szkoła Główna Gospodarsta Wiejskiego w Warszawie

Uniwersytet Warszawski

Warszawski Uniwersytet Medyczny

V Kongres Statystyki Polskiej, Warszawa 2025

Outline

- Generative Model.
- Misspecified Model.
- Modelling Allergy Diseases (Dataset).
- Comparision of Two Version of Our Model.
- Evaluation of the Model.
- Conclusions.

Generative Model-part 1

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.

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 $a_{ki} = 0$.

Generative Model-part 2

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 $b_{kj} = 0$.

Generative Model-part 3

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).

Generative Model-part 4

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.

Generative Model-part 5

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: $\omega_{0i} \in R, \omega_{ki} \in R, \alpha_i \in R^r, \beta_i \in R^l, \gamma_{0j} \in R, \gamma_{kj} \in R, \delta_j \in R^r, \epsilon_j \in R^l$.

Misspecified Model -part 1

Unfortunately, generative model 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.

Misspecified Model -part 2

In the misspecified model, equation (1) is replaced by equation (4), and equations (2)-(3) are replaced by equation (5) as follows:

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

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.

Misspecified Model -part 3

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

Since the conditional probability (1) consists of the product of $p + m$ probabilities, the parameters of each factor can be estimated separately by a standard logistic regression procedure.

Modeling Allergy Diseases-Dataset

In our work we used data from the big epidemiological study in Poland (ECAP). This project was conducted in one rural and eight urban areas. The study had two stages; the first stage involved grouping the 22,500 respondents based on their questionnaire responses using a Personal Digital Assistant (PDA); the second stage involved complementary examination (4,783 patients) of a subgroup of the first stage respondents who underwent a medical examination. The final data set contains 18,617 units (cases, response) and 1,225 variables (mostly binary). Our model used data in the second stage of this survey.

Modeling Allergy Diseases-Dataset

We consider 5 allergic diseases $Y_1 - Y_5$ and their symptoms $S_1 - S_3$:
Y_1 -atopic asthma;
Y_2 -intermittent allergic rhinitis;
Y_3 -chronic allergic rhinitis;
Y_4 -allergic dermatitis;
Y_5 -food allergy;
S_1 -Have you had wheezing or whistling in your chest at any time in the last 12 months?;
S_2 - Have you ever had a problem with sneezing or a runny or blocked nose when you did not have fever, a cold, or the flu?;
S_3 -Have you ever had eczema or any other form of skin allergy?

Modeling Allergy Diseases-Dataset

History of allergy diseases in the family $F_1 - F_5$:

F_1-Does anyone in your immediate family suffer from allergies? - mother;

F_2-Does anyone in your immediate family suffer from allergies? - father;

F_3-Does anyone in your immediate family suffer from allergies? - siblings of the child being tested;

F_4-Does anyone in your immediate family suffer from allergies? -grandparents on mother's side;

F_5-Does anyone in your immediate family suffer from allergies? - grandparents on father's side.

Modeling Allergy Diseases-Dataset

Controls variables $X_1 - X_4$:
Age of patients with three age group: children 6-7 y.o., children 13-14 y.o., and adults (20-44 y.o.). This variable we replaced by new two binary variables:
X_1- for children 13-14 y.o. and X_2- for adults.,
X_3- urban area (binary variables with 1 for urban area),
X_4 -gender (binary variable, 1 for male).

Figure

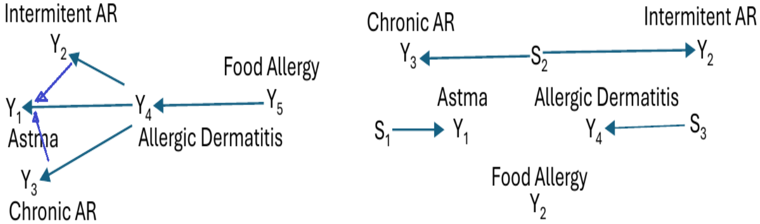

Figure: The Graphs with adjacency matrices **A**, **B**

Modeling Allergy Diseases

The left panel of Figure 1 illustrates the dependences between allergy diseases, based on the literature and on discussions with medical doctors.
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.

Generative and Misspecified Models

We recall the generative model in which diseases cause symptoms. We see that, conditionally on covariates **F** and **X**, the conditional distribution of **Y** given symptoms **S** has the form

$P(\mathbf{Y}|\mathbf{S}) =$
$P(Y_1|Y_2, Y_3, Y_4)P(Y_2|Y_4)P(Y_3|Y_4)P(Y_4|Y_5)P(Y_5)P(S_1|Y_1)$
$\mathrm{P}(\mathrm{S}_2|Y_2, Y_3)P(S_3|Y_4).$

Now we turn to the misspecified model. Conditionally on covariates **F** and **X**

$P(\mathbf{Y}|\mathbf{S}) = P(Y_1|Y_2, Y_3, Y_4, S_1)P(Y_2|Y_4, S_2)P(Y_3|Y_4, S_2)$
$\mathrm{P}(\mathrm{Y}_4|Y_5, S_3)P(Y_5).$

Misspecified Model

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:

$$logit_1 =$$

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

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

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

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

Comparision of Two Versions of Our Model

We compute the 'diagnostic' probabilities of diseases given symptoms for the generative and the misspecified model.

We consider five scenarios (different values of covariates $\mathbf{X}, \mathbf{F}$, symptoms $\mathbf{S}$ and coexistent deseases Y_i).

Let

$$p_1 = P(Y_1 = 1 | Y_2 = 0, Y_3 = 0, Y_4 = 0, S_1),$$
$$q_1 = P(Y_1 = 1 | Y_2 = 1, Y_3 = 1, Y_4 = 1, S_1),$$
$$p_2 = P(Y_2 = 1 | Y_4 = 0, S_2),\ q_2 = P(Y_2 = 1 | Y_4 = 1, S_2),$$
$$p_3 = P(Y_3 = 1 | Y_4 = 0, S_2),\ q_3 = P(Y_3 = 1 | Y_4 = 1, S_2),$$
$$p_4 = P(Y_4 = 1 | Y_5 = 0, S_3),\ q_4 = P(Y_4 = 1 | Y_5 = 1, S_3).$$

Comparision of Two Versions

We consider five scenarios of covariates $\mathbf{X}, \mathbf{F}$:

- Scen. 1: rural area, children 13-14 y.o., male, without allergy history in family $F_1 = \ldots = F_5 = 0$ and without symptoms $S_1 = S_2 = S_3 = 0$;
- Scen. 2: rural area, children 13-14 y.o., male, without allergy history in family $F_1 = \ldots = F_5 = 0$ and with symptoms $S_1 = S_2 = S_3 = 1$;
- Scen. 3: urban area, children 13-14 y.o., male, without allergy history in family $F_1 = \ldots = F_5 = 0$ and with symptoms $S_1 = S_2 = S_3 = 1$;
- Scen. 4: urban area, children 13-14 y.o., male, without allergy history in family $F_1 = \ldots = F_5 = 0$ and without symptoms $S_1 = S_2 = S_3 = 0$;
- Scen. 5: urban area, children 13-14 y.o., male, with allergy history in family $F_1 = \ldots = F_5 = 1$ and with symptoms $S_1 = S_2 = S_3 = 1$.

Results of Comparision

Table: Comparison between the generative model and misspecified model

Sc.	Mod.	p_1	p_2	p_3	p_4	q_1	q_2	q_3	q_4
1	gen.	0.021	0.081	0.077	0.024	0.597	0.104	0.134	0.024
	miss.	0.023	0.085	0.080	0.015	0.566	0.097	0.125	0.044
2	gen.	0.103	0.282	0.322	0.208	0.886	0.326	0.461	0.524
	miss.	0.088	0.270	0.307	0.081	0.842	0.299	0.421	0.216
3	gen.	0.074	0.212	0.321	0.248	0.845	0.250	0.463	0.581
	miss.	0.064	0.202	0.306	0.116	0.793	0.226	0.420	0.290
4	gen.	0.015	0.056	0.074	0.007	0.509	0.073	0.130	0.029
	miss.	0.016	0.060	0.080	0.022	0.482	0.068	0.124	0.064
5	gen.	0.120	0.332	0.359	0.233	0.902	0.398	0.511	0.561
	miss.	0.130	0.295	0.314	0.210	0.893	0.326	0.429	0.452

Evaluation of the Model

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.

We draw 20 ordinary non-parametric bootstrap samples, calculate regression coefficients on each bootstrap sample, treat the whole real sample as a test sample, draw the ROC for it, and calculate the AUC.

We also used the jackknife (10-fold cross-validation) method: we drew 10% of the sample for testing and treated the other 90% as a training sample. We repeated the experiment 20 times.

Evaluation of the Model

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. Our results show good stability of the model.

Table: 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

ROC

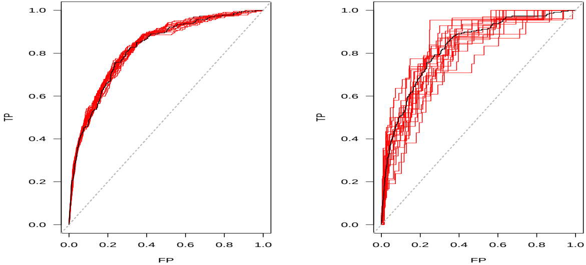

Figure: ROC for *$logit_1$* for bootstrap (left) and jackknife (right).

Conclusions

The proposed model can be used in studies of associations of other diseases and, in general, in the study of correlations in complex systems.

Both versions of our model (generative and misspecified) produced similar results. The latter is computationally more efficient and easily interpretable.

Conclusions

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.

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.

References

- 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)
- Kim, H. Y. et al.: Prevalence and comorbidity of allergic diseases in preschool children. Korean J. Pediatr., **56**(8), 338—342 (2013)
- 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)

References

- Raciborski F. et al.: Dissociating polysensitization and multimorbidity in children and adults from a Polish general population cohort. Clin. transl. allergy, 9:4 (2019)
- Ravikumar, P., Wainwright, M. J., Lafferty, J.: High-dimensional Ising model selection using l1-regularized logistic regression. Ann. Statist. **38**, 1287—1319 (2010)
- Furmańczyk K., Niemiro W., Chrzanowska M., Zalewska M.: Supplementary Material to the paper 'Network Model with Application to Allergy Diseases' International Conference on Computational Science (ICCS), Lecture Notes In Computer Science, vol. 14835, s.105-112 (2024)

References

- 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

- Westman M. et al.: Natural course and comorbidities of allergic and nonallergic rhinitis in children. J Allergy Clin Immunol {129}(2), 403–408 (2012)

- 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)

Dziękuję za uwagę.