

Methodology Using Gröbner basis for Generalizing Confounding Relations in Linear Combination Models

Yuichi Nishioka

y_n@naramed-u.ac.jp

Nara Medical University

Research Article

Keywords: Real-World Data, Administrative Claims Database, Grobner basis, Groebner basis, Grbner basis, Statistical Design, Confoundings, Linear Combination Model

Posted Date: September 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10834410/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: Competing interest reported. YN received consultation fees from Novo Nordisk, and speaker fees from Sanofi, Daiichi Sankyo, and DeSC Healthcare.

1 Research Article

2 **Methodology Using Gröbner basis for Generalizing Confounding Relations in Linear Combination**

3 **Models**

4

5 Yuichi Nishioka^{1,2}

6 1 Department of Public Health, Health Management and Policy, Nara Medical University, Nara, Japan

7 2 Department of Diabetes and Endocrinology, Nara Medical University, Nara, Japan

8

9 Correspondence to: Yuichi Nishioka, M.D., Ph.D.

10 Department of Public Health, Health Management and Policy, Nara Medical University

11 840 Shijyo-cho, Kashihara City, Nara 634-8521, JAPAN

12 E-mail: y_n@naramed-u.ac.jp

13 Phone: +81-744-22-3051

14 Fax: +81-744-22-0037

15

ABSTRACT

The study explores the application of Gröbner bases theory to medical statistics, specifically in the context of confounding relationships in epidemiological studies. Gröbner bases provide a method for analyzing whether certain effects in experimental designs are confounded. The methodology involves constructing a linear combination model within a polynomial ring, where the confounding relationships can be analyzed using Gröbner bases. The study illustrates this with examples, showing how confounded interactions can be identified and how certain parameters cannot be determined simultaneously due to these confounding effects. The discussion emphasizes the importance of considering confounding in study design and analysis to accurately estimate causal effects. The study concludes by suggesting that the generalized approach to confounding relationships could be applied to real-world data, potentially improving the understanding and application of these models in medical statistics.

Keywords: Real-World Data, Administrative Claims Database, Grobner basis, Groebner basis, **Gröbner basis**, Statistical Design, Confoundings, Linear Combination Model

1 Introduction

Application of Gröbner bases theory to experiments is an topic in medical statistics, which is called computational algebraic statistics.(Cox et al., 1997; Aoki et al., 2012) The confounding relation can be generalized designs and expressed concisely using the Gröbner bases theory. In fact, the problem of judging whether two effects are confounded or not is equivalent to the problem of ideal membership. The ideal membership problem is solved by calculation of a reduced Gröbner basis of the design ideal. (Aoki and Takemura, 2012)

Gröbner bases theory is applied to such as polynomial equation solving and algebraic relations. (Bose, 1995). The purpose of this study is to generalize confounding relations of epidemiology in linear combination models.

2 Methodology

This study was approved by the Ethics Committee of Nara Medical University (approval no. 1123-7, October 8th, 2015). The need for consent was waived by the ethics committee. In polynomial ring $K[x_1, \dots, x_n]$, we constructed a linear combination model as follows. (x_k lying in $\{0, 1\}$, $a_k \in \mathbb{R}$ $y \in \mathbb{R}$)

$$y = \sum_{k=0}^{2^{(2^n)}-1} \left\{ a_k \prod_{i=1}^n (b_i)_k x_i \right\}$$

where $(b_i)_k$ is defined

$$\begin{pmatrix} (b_1)_0 & \cdots & (b_n)_0 \\ \vdots & \ddots & \vdots \\ (b_1)_{2^{(2^n)-1}} & \cdots & (b_n)_{2^{(2^n)-1}} \end{pmatrix} = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \ddots & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & \cdots & 0 & 1 \\ 1 & 1 & \cdots & 1 & 0 \\ 1 & 1 & \cdots & 1 & 1 \end{pmatrix}$$

Here, consider 2^n possible combinations of values of $x_1, \dots, x_n$. Thus, we have

$$\begin{pmatrix} (x_1)_1 & \cdots & (x_n)_1 \\ \vdots & \ddots & \vdots \\ (x_1)_{2^{(2^n)}} & \cdots & (x_n)_{2^{(2^n)}} \end{pmatrix} = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \ddots & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & \cdots & 0 & 1 \\ 1 & 1 & \cdots & 1 & 0 \\ 1 & 1 & \cdots & 1 & 1 \end{pmatrix}$$

Suppose that s elements of the $j_1, \dots, j_s$ lines are unobservable and the other $(2^{(2^n)} - s)$ elements are

observable. In this case, $x_1, \dots, x_n$ satisfies the following relation $f_1, \dots, f_n$.

$$f_1 = \prod_{i=1}^n \{x_i + (x_i)_{j_1} - 1\} = 0$$

$$\vdots$$

$$f_s = \prod_{i=1}^n \{x_i + (x_i)_{j_s} - 1\} = 0$$

For a set of polynomials $F = [f_1, \dots, f_s, x_1(x_1 - 1), \dots, x_n(x_n - 1)]$, Gröbner basis $G = \text{sp.groebner}(F)$

based on any monomial ordering describes exactly the confounding relationship in the model

$$y = \sum_{k=0}^{2^{(2^n)}-1} \left\{ a_k \prod_{i=1}^n (b_i)_k x_i \right\}$$

3 Example

(1) Example 1

In the model

66
$$y = a_0 + a_1x_3 + a_2x_2 + a_3x_2x_3 + a_4x_1 + a_5x_1x_3 + a_6x_1x_2 + a_7x_1x_2x_3$$

67 if $(x_1, x_2, x_3) = (0,0,0)(0,0,1)(0,1,0)(0,1,1)(1,0,0)(1,0,1)(1,1,1)$ are observed, and $(x_1, x_2, x_3) =$
 68 $(1,1,0)$

69 is not observed, then we have

70
$$x_1x_2(x_3 - 1) = 0$$

71 For $F = [x_1x_2(x_3 - 1), x_1(x_1 - 1), x_2(x_2 - 1), x_3(x_3 - 1)]$, and $G = \text{sp.groebner}(F)$, we have the
 72 Boolean Gröbner basis G based on the lexicographic ordering as follows.

73
$$G = [x_1(x_1 - 1), x_1x_2x_3 - x_1x_2, x_2(x_2 - 1), x_3(x_3 - 1)]$$

74 and if we focus on $x_1x_2x_3 - x_1x_2$, we can see that the third-order interaction $x_1x_2x_3$ and the second-order
 75 interaction x_1x_2 are confounded in this model. Therefore, a_6 and a_7 cannot be obtained simultaneously
 76 in the model

77
$$y = a_0 + a_1x_3 + a_2x_2 + a_3x_2x_3 + a_4x_1 + a_5x_1x_3 + a_6x_1x_2 + a_7x_1x_2x_3$$

78

79 (2) Example 2

80 In the model

81
$$y = a_0 + a_1x_3 + a_2x_2 + a_3x_2x_3 + a_4x_1 + a_5x_1x_3 + a_6x_1x_2 + a_7x_1x_2x_3$$

82 if $(x_1, x_2, x_3) = (0,0,0)(0,1,0)(0,1,1)(1,0,0)(1,0,1)(1,1,1)$ are observed, and $(x_1, x_2, x_3) =$
 83 $(1,1,0)(0,0,1)$ are not observed, then we have

$$x_1x_2(x_3 - 1) = 0$$

$$(x_1 - 1)(x_2 - 1)x_3 = 0$$

For $F = [x_1x_2(x_3 - 1), (x_1 - 1)(x_2 - 1)x_3, x_1(x_1 - 1), x_2(x_2 - 1), x_3(x_3 - 1)]$, and

$G = \text{sp.groebner}(F)$, we have the Boolean Gröbner basis G based on the lexicographic ordering as follows.

$$G = [x_1(x_1 - 1), x_1x_2 - x_1x_3 - x_2x_3 + x_3, x_2(x_2 - 1), x_3(x_3 - 1)]$$

and if we focus on $x_1x_2 - x_1x_3 - x_2x_3 + x_3$, we can see that $(x_3 + x_1x_2)$, and $(x_2x_3 + x_1x_3)$ are

confounded in this model.

In the model,

$$y = a_0 + a_1x_3 + a_2x_2 + a_3x_2x_3 + a_4x_1 + a_5x_1x_3 + a_6x_1x_2 + a_7x_1x_2x_3$$

$$y = a_0 + a_1(x_3 + x_1x_2) + a_2x_2 + a_3(x_2x_3 + x_1x_3) + a_4x_1 + (a_5 - a_3)x_1x_3 + (a_6 - a_1)x_1x_2$$

$$+ a_7x_1x_2x_3$$

a_1 , and a_3 cannot be obtained simultaneously.

4 Discussion

Consideration of confoundings is fundamental to the design, analysis, and interpretation of

studies in order to estimate causal effects. (Greenland and Morgenstern, 2001) In this study, I generalized

confounding relations of epidemiology in linear combination models. This allows confounding in a binary

variable model to be expressed concretely, as illustrated by the above examples. Furthermore, by using the

dummy variable concept, confounding in models with more than three values can also be expressed concretely.

5 Conclusions

In this study, I generalized confounding relations of epidemiology in linear combination models.

Future application of this method to real-world data is expected to enhance understanding of the model.

List of Abbreviations

RCT; Randomized Controlled Trial

Declarations

Ethics Approval

All methods were conducted in accordance with relevant guidelines and regulations. All experimental protocols were approved by the Ethics Committee of Nara Medical University (approval no. 1123-7, October 8th, 2015). The need for informed consent was waived in view of the study design; all patient data were anonymized before analysis.

Consent to Participate

120 Not applicable

121

122 *Availability of Data and Materials*

123 Not applicable

124

125 *Competing Interests*

126 YN received consultation fees from Novo Nordisk, and speaker fees from Sanofi, Daiichi Sankyo, and

127 DeSC Healthcare.

128

129 *Funding*

130 This study was supported by the Japan Society for the Promotion of Science KAKENHI (grant numbers:

131 22H03355, 23K24613, and 24K22341).

132

133 *Authors' Contributions*

134 YN led the study and wrote the draft of the paper.

135

Acknowledgment

N/A

References

Aoki, S., Hara, H., & Takemura, A. (2012). Markov bases in algebraic statistics. Springer Science & Business Media.

Aoki, S., & Takemura, A. (2012). Design and analysis of fractional factorial experiments from the viewpoint of computational algebraic statistics. *Journal of Statistical Theory and Practice*, 6(1), 147-161.

Bose, N., & Bose, N. (1995). Gröbner bases: An algorithmic method in polynomial ideal theory. Springer.

Cox, D., Little, J., O'shea, D., & Sweedler, M. (1997). Ideals, varieties, and algorithms. Springer.

Greenland, S., & Morgenstern, H. (2001). Confounding in health research. *Annu Rev Public Health*, 22, 189-212. doi:10.1146/annurev.publhealth.22.1.189.