**A simultaneous spatial autoregressive model for compositional data**

**Analyzing the impacts of socio-economic factors on French departmental elections with CoDa methods**

DEFINITIONS AND NOTATIONS OF THE COMPOSITIONAL DATA ANALYSIS

Một D-thành phần **u** là một vector của D các phần (parts) của một tổng thể mang thông tin tương đối và do đó có thể được biểu diễn trong cái gọi là một không gian đơn giản S^D định nghĩa bởi Aichison (1968) như:

ở đó ^T là toán tử chuyển vị(hoán vị hay gì đó). Với bất kỳ vector w thuộc R^+D, phép đóng được định nghĩa bởi…

nhắc lại các phép toán thông thường dung để xác định cấu trúc vector trên không gian đơn hình:  
(cái gì đó về phép cộng vt và phép nhân vt trong đơn hình đã đc ghi rồi)

hơn nữa tích thành phần của ma trận vector ký hiệu là ⊡ được định nghĩa như sau:  
  
(1)  
  
ở đó u thuộc không gian S^D, B = b\_lm) với l=1,…,L; m = 1,…,D là một ma trận LxD thỏa mãn B\_j\_D = 0\_D và B^T\_j\_D = 0\_D, ở đó j\_D (tương ứng với 0\_D) chứng tỏ D-chiều cột vector của 1’s(tương ứng với 0’s)

không gian S^D cũng có thể được trang bị sản phẩm bên trong thành phần/Aitchison (…) để xác định khoảng cách

Phân tích dữ liệu đa hợp sử dụng các biến đổi tỷ số log để ánh xạ không gian đơn giản S^D*SD* vào \mathbb{R}^qR*q* (thường là q = D - 1*q*=*D*−1) do tính đồng nhất bậc 0 của chúng (tính không đổi tỷ lệ). Các biến đổi kinh điển bao gồm biến đổi tỷ số log cộng (alr), biến đổi tỷ số log tâm (clr) và biến đổi tỷ số log đồng đều (ilr). Trong bài báo này, chúng tôi sẽ chủ yếu sử dụng một số biến đổi ilr do tính chất bảo toàn khoảng cách của chúng. Để biết thêm chi tiết về biến đổi alr, xem Thomas-Agnan et al. (2020). Vì chúng ta cần nó để định nghĩa ilr, hãy nhớ lại định nghĩa của biến đổi clr của một vector u [ S^D*SD* ]:

ở đó g(u) = căn bậc D của gì đó là giá trị trung bình hình học của các thành phần.   
Hãy để V^D*VD* là một ma trận tương phản kích thước D \times (D - 1)*D*×(*D*−1) (ví dụ, Pawlowsky-Glahn (et al., 2015) liên quan đến một cơ sở chéo bằng ( e\_1, \dots, e\_{D-1}*e*1​,…,*eD*−1​ ) đã cho của S^D*SD* bởi:

(tự giải thích: ma trận tương phản là một ma trận mà mỗi cột tương ứng với một vector trong cơ sở chéo bằng, và mỗi hàng tương ứng với một phần tử trong không gian vector.)

ở đây clr được hiểu theo từng cột. với mỗi ma trận V\_D như vậy một phép biến đổi tr số log đẳng cự(ilr) sau đó được xác định bởi:

ở đó logarit của u thuộc S^D được hiểu theo từng thành phần. Phép biến đổi nghịch đảo là:

Lưu ý rằng ma trận tương phản \( V^D \) kích thước \( D \times (D - 1) \) thỏa mãn các tính chất sau:

\[

V^D(V^D)^T = I\_D - \frac{{j\_D j\_D^T}}{{D}}

\]

trong đó \( I\_D \) là ma trận đơn vị kích thước \( D \times D \) và \( j\_D \) là vector cột kích thước \( D \times 1 \) của các số một.

\[

(V^D)^T V^D = I\_{D-1}

\]

\[

(V^D)^T j\_D = \mathbf{0}\_{D-1}

\]

trong đó \( \mathbf{0}\_{D-1} \) là vector cột kích thước \( (D - 1) \times 1 \) của các số không.

Ma trận \( V^D \) sau đây có kích thước \( D \times (D - 1) \) được định nghĩa trong [5] là một ví dụ về ma trận tương phản cho \( D = 3 \).

Ma trận cụ thể này định nghĩa các tọa độ ilr sau:

\[

\text{ilr}\_1(x) = \frac{1}{\sqrt{6}} \left( 2 \ln x\_1 - \ln x\_2 - \ln x\_3 \right) = \frac{2}{\sqrt{6}} \ln x\_1 \sqrt{\frac{x\_2}{x\_3}}

\]

Tọa độ ilr thứ hai được định nghĩa như sau:

\[

\text{ilr}\_2(x) = \frac{1}{\sqrt{2}} \left( \ln x\_2 - \ln x\_3 \right) = \frac{1}{\sqrt{2}} \ln \frac{x\_2}{x\_3}

\]

Tọa độ ilr đầu tiên chứa thông tin về mức độ quan trọng tương đối của thành phần đầu tiên \( x\_1 \) so với trung bình hình học của các thành phần thứ hai và thứ ba \( \sqrt{x\_2 x\_3} \). Tọa độ ilr thứ hai chứa thông tin về mức độ quan trọng tương đối của thành phần thứ hai \( x\_2 \) so với thành phần thứ ba \( x\_3 \).

Vì dữ liệu của chúng tôi được tạo thành từ các mẫu của các vector hợp thành, chúng tôi lưu trữ chúng trong một ma trận kích thước $n \times D$ $Y = (Y\_{il})$, với $i = 1, \ldots , n$, $l = 1, \ldots , D$, trong đó $n$ là số lượng đơn vị không gian. Mỗi hàng của ma trận này, được ký hiệu bởi $Y\_{i \cdot}$, là một vector cột hợp thành của $S^D$. $Y\_{\cdot l}$, $l = 1, \ldots , D$, ký hiệu cột thứ $l$ của $Y$ và chúng ta có:

Xác định phần mở rộng của phép biến đổi ilr của ma trận Y bằng cách:

Lưu ý rằng ilr(Y) là một ma trận n x (D-1)

Tương tự, với các điều kiện giống như trước đó đối với ma trận $ B $, chúng ta có thể định nghĩa một sự mở rộng của phép nhân ma trận bằng cách cho $ Y \circledast B $ là ma trận kích thước $ n \times D $ mà các hàng là các vector $ B^T \circledast Y\_{i\cdot} $. Sau đó, ta dễ dàng nhận thấy rằng điều này tương đương với:

where ilr(Y ⊡ B) is n × (D − 1) matrix.

Trong các mô hình kinh tế không gian, các ma trận trọng số không gian được sử dụng để chỉ định cấu trúc hàng xóm. Đối với \( n \) địa điểm không gian, các phần tử \( w\_{ij} \) của ma trận \( n \times n \) \( W \) là các đơn vị đo của sự gần gũi giữa các địa điểm \( i \) và \( j \) (ví dụ, xem Bivand et al., 2008, để biết các đặc điểm khác nhau). Các ma trận này xác định một mô hình hiệp phương sai cho vector dữ liệu và đóng một vai trò tương tự như biến đổi không gian ẩn trong địa lý thống kê. Đối với một ma trận như vậy và một vector dữ liệu \( \mathbf{Z} \) cố định có kích thước \( n \), vector trễ \( W\mathbf{Z} \) chứa các trung bình của biến \( \mathbf{Z} \) tại các địa điểm hàng xóm khi \( W \) được chuẩn hóa hàng. Trong trường hợp của chúng tôi, chúng tôi cần áp dụng thao tác này cho mỗi cột của ma trận dữ liệu \( \mathbf{Y} \) và chúng tôi muốn kết quả này là một ma trận trong cùng không gian như ban đầu \( (SD)^n \). Như thường lệ trong phân tích dữ liệu hợp thành, chúng tôi sử dụng các tọa độ tỷ số log (Mateu-Figueras et al., 2011) và biểu diễn kết quả trong không gian đơn giản. Do đó, chúng tôi định nghĩa thao tác sau đây.

\textbf{Định nghĩa 1.} Hãy để \( W \) là một ma trận \( n \times n \). Phép \( \triangle \cdot \) là một ánh xạ từ tích cartesian của các không gian đơn giản \( (SD)^n \) đến chính nó được xác định bởi:

trong đó \( V^D \) là một ma trận tương phản kích thước \( D \times (D - 1) \).

Lưu ý rằng \( (W \triangle \cdot Y) \) là một ma trận \( n \times D \) và \( \text{Wilr}(Y) \) là một ma trận \( n \times (D - 1) \).

Thao tác này thỏa mãn các tính chất sau:

Mệnh đề 1. Hãy để \( Y \) là một ma trận \( n \times D \) sao cho mỗi hàng \( (Y\_{i\cdot}, i = 1, \ldots , n) \) là một vector hợp thành trong \( S^D \). Hãy để \( W = (W\_{ij}), (i, j = 1, \ldots , n) \) là một ma trận \( n \times n \) và \( a \in \mathbb{R} \) là một số vô hướng bất kỳ. Chúng ta có:

\[

W \triangle \cdot (a \odot Y) = a \odot (W \triangle \cdot Y)

\]

\[

\text{ilr} (W \triangle \cdot (a \odot Y)) = a \cdot W \text{ilr} (Y) = a \text{ilr}(W \triangle \cdot Y)

\]

\[

(W \triangle \cdot Y)\_{i\cdot} = \begin{bmatrix} \frac{\sum\_{j=1}^{n} Y\_{ij} W\_{ij1}}{\sum\_{j=1}^{n} Y\_{ij}} ; \frac{\sum\_{j=1}^{n} Y\_{ij} W\_{ij2}}{\sum\_{j=1}^{n} Y\_{ij}} ; \ldots ; \frac{\sum\_{j=1}^{n} Y\_{ij} W\_{ijD}}{\sum\_{j=1}^{n} Y\_{ij}} \end{bmatrix} \text{ cho } i = 1, \ldots , n

\]

Hãy để \( Y\_1, Y\_2 \in (SD)^n \), sau đó \( W \triangle \cdot (Y\_1 \oplus Y\_2) = (W \triangle \cdot Y\_1) \oplus (W \triangle \cdot Y\_2) \).

Đối với bằng chứng của Mệnh đề 1, xem Phụ lục A trong dữ liệu bổ sung trực tuyến. Lưu ý rằng tính chất 3 ngụ ý rằng đối với mỗi cá nhân, mỗi thành phần của \(W \triangle \cdot Y\) là một trung bình hình học có trọng số của các giá trị hàng xóm của nó, được trọng số bởi trọng số hàng xóm.

Cấu trúc không gian vector của simplex \( S^D \) được xác định bởi các phép toán nghịch đảo và mũ:

\[

\mathbf{x} \oplus \mathbf{y} = C(x\_1y\_1, \ldots, x\_Dy\_D)^T, \quad \mathbf{x}, \mathbf{y} \in S^D

\]

\[

\lambda \odot \mathbf{x} = C(x\_1^\lambda , \ldots , x\_D^\lambda)^T, \quad \lambda \text{ là một số vô hướng, } \mathbf{x} \in S^D

\]

Nguyên tắc của các mô hình hồi quy đa hợp là sử dụng một phép biến đổi để chuyển dữ liệu từ không gian đơn giản sang một không gian tọa độ nào đó và giả định rằng dữ liệu đã được biến đổi tuân theo một mô hình hồi quy Gaussian trong không gian tọa độ, như trong Egozcue et al. (2012). Mô hình sau đó có thể được chuyển trở lại không gian đơn giản bằng phép biến đổi nghịch đảo. Trong trường hợp của chúng tôi, mô hình trong không gian tọa độ phải là một mô hình hồi quy đa biến vì chúng tôi có nhiều biến phản hồi.

Để đơn giản, chúng tôi tập trung vào mô hình tự động lặp đồng thời được gọi là mô hình độ trễ không gian tự động lặp, bao gồm các biến động lặp có nguồn gốc kết quả trên phía bên phải của các phương trình mô hình. Thuật ngữ độ trễ đến từ toán tử độ trễ được gọi là toán tử trễ. Một mở rộng cho công thức của mô hình Durbin sẽ được thực hiện một cách dễ dàng. Vì mô hình của chúng tôi sẽ được viết trong không gian tọa độ, chúng tôi chọn đánh dấu tất cả các biến và tham số trong phần thứ ba bằng dấu hoa thị. Lưu ý rằng mô hình mà chúng tôi mô tả dưới đây không cụ thể cho dữ liệu đa hợp.

**ở đây bắt đầu khó nè, vì 2 tài liệu đang có cách trình bày và mục tiêu khác nhau, từ đoạn này cũng khá khó hiểu nên đang chưa biết làm như nào, các tốt nhất có lẽ là cố dịch cả 2 chăng? Vì cái mình đang dịch nó sát với cái lúc trong video mình xem hơn!**Mô hình trong không gian tọa độ log-ratio(tỷ lệ log)

Chúng tôi xem xét một mẫu có kích thước n và giả định rằng chúng tôi có M biến endogenous, do đó có M phương trình hồi quy tuyến tính (M sẽ là D − 1 trong phần thứ ba tạm thời chưa hình dung ra là phần nào). Đối với một ma trận A kích thước n × M, chúng tôi sẽ sử dụng cùng một ký hiệu, như trong phần thứ hai, với A\_l là cột thứ l, và A\_i. là hàng thứ i của A dưới dạng một vector cột. Trong phần tiếp theo, khi một biến được đánh dấu, nó chỉ ra thực tế rằng nó được thu được bằng phép biến đổi log-ratio từ một biến có giá trị trên không gian simplex và khi một tham số được đánh dấu, nó chỉ ra rằng đó là một tham số của mô hình trong không gian tọa độ log-ratio.

Hãy để \( Y^\* \) là một ma trận \( n \times M \) của các biến phụ thuộc và \( X \) là một ma trận \( n \times K \) của các biến giải thích. Chúng tôi sẽ cho phép sử dụng một tập hợp khác nhau của các biến giải thích trong mỗi phương trình. Vì lí do này, chúng tôi ký hiệu bằng \( \text{SY}^\*\_l \), \( \text{SX}\_l \), \( \text{SWY}^\*\_l \) các tập hợp chỉ số của các biến xuất hiện trong phương trình thứ \( l \) cho \( Y^\* \), \( X \), \( \text{WY}^\* \), tương ứng. Tương tự, \( Y^\*\_{\text{SY}^\*\_l} \), \( X\_{\text{SX}\_l} \), \( Y^\*\_{\text{SWY}^\*\_l} \) sẽ ký hiệu các cột của \( Y^\* \), \( X \), \( \text{WY}^\* \) xuất hiện trong phương trình thứ \( l \).

(**chú ý giải thích cho biến phụ thuộc là gì, biến giải thích là gì Y\* là nxM và X là nxK)**

Hãy để \( G^\* = (G^\*\_{ml}) \) và \( R^\* = (R^\*\_{ml}) \), (\( m, l = 1, \ldots , M \)) là các ma trận kích thước \( M \times M \) của các tham số. \( R^\* \) chứa các tham số liên quan đến các biến endogenous đã trễ ở phía bên phải của phương trình mô hình. Tương tự như trong văn học về phương trình đồng thời trong kinh tế lượng, mỗi biến endogenous cũng có thể xuất hiện trong mỗi phương trình mô hình để \( G^\* \) chứa các tham số tương ứng.

Cuối cùng, \( \mathbf{b}^\* \) là một ma trận \( K \times M \) của các tham số cho các biến giải thích và \( \mathbf{e}^\* \) là một ma trận sai số kích thước \( n \times M \).

Giống trong Kelejian and Prucha (2004), chúng ta xem xét theo mô hình bên dưới:

Các mô hình đa biến có thể được biểu diễn bằng một biểu thức ma trận sử dụng một phương trình duy nhất hoặc với một hệ thống các phương trình đồng thời cho mỗi hàng hoặc mỗi cột. Mô hình (5) được viết với một hệ thống các phương trình đồng thời, một cho mỗi cột của \( Y^\* \) và các phương trình này được liên kết bởi cấu trúc hiệp phương sai của các sai số. Thực sự, chúng ta giả định rằng các sai số được tập trung \( \mathbb{E}(\mathbf{e}^\*) = \mathbf{0}\_M \) và \( \mathbb{E}(\mathbf{e}^\*\_i \cdot (\mathbf{e}^\*\_j)^T) = \mathbf{S}^\* \) nếu các cá nhân \( i \) và \( j \) là giống nhau, và \( \mathbf{0} \) nếu chúng khác nhau, nơi \( \mathbf{S}^\* \) là ma trận hiệp phương sai kích thước \( (D - 1) \times (D - 1) \). Điều này có nghĩa là các cá nhân là độc lập nhau nhưng các thành phần của một cá nhân cụ thể có một cấu trúc hiệp phương sai \( \mathbf{S}^\* \).

Kelejian và Prucha (2004) đề xuất và nghiên cứu tính chất của một phương pháp S2SLS cũng như một phương pháp S3SLS để ước lượng mô hình (5). Tuy nhiên, Gibbons và Overman (2012) chứng minh rằng nếu các thuật ngữ Durbin không gian được bao gồm, thì một số công cụ có thể trở nên yếu, làm cho bộ ước lượng biến công cụ không được khuyến khích. Đối với mô hình LAG không gian của phương trình (5), theo gợi ý của Kelejian và Prucha (2004), chúng ta xem xét \( H \) là một tập con của các cột độc lập tuyến tính của ma trận \( n \times 3K \) (\( X, WX, W^2X \)). Hãy để \( PH = H(H^TH)^{-1}HT \) là ma trận chiếu vào không gian được tạo ra bởi các cột của \( H \). Đối với phương trình thứ \( l \), chúng ta nhóm các biến và tham số của phía bên phải thành một ma trận \( Z\_l \) các biến và một vector \( \mathbf{d}^\*\_l \) các tham số:

Phương pháp ước lượng S2SLS cho mô hình này sau đó được thực hiện như sau cho mỗi phương trình (tức là, mỗi thành phần) một cách riêng biệt:  
\* Thực hiện một phương trình hồi quy một biến của mỗi cột của $Z$ trên $H$ và tính toán các giá trị đã điều chỉnh $Z\_l$:

Thực hiện một phương trình hồi quy một biến của \( Y^\*\_l \) trên \( \widetilde{Z}\_l \):

\[ \mathbf{d}^\*\_l = (\widetilde{Z}^T\_l \widetilde{Z}\_l)^{-1}\widetilde{Z}^T\_l \mathbf{Y}^\*\_l \]

ở cuối cùng của bước 2, hung ta có thể tính phần dư bằng:

và nhận được ước lượng của ma trận hiệp phương sai \( S^\* \) với

Cho đến nay, cấu trúc hiệp phương sai giữa các phương trình chưa được tính toán và phương pháp S3SLS được giả định để sửa chữa điều này. Để viết biểu thức của bộ ước lượng S3SLS, chúng ta cần biến đổi vector của \( Y^\* \) (bằng cách xếp chồng các cột của Y) kết quả là \( \mathbf{y}^\* = \text{vec}(\mathbf{Y}^\*) \) và viết các ma trận giải thích như sau:

Chúng ta sau đó nhận được một bộ ước lượng đã được sửa đổi \( \hat{d}^\* \) của \( d^\* \):

Trong ứng dụng của phần thứ năm, chúng tôi xem xét một mô hình ít phổ biến hơn một chút trong đó chúng tôi bao gồm các biến hợp thành trong số các biến giải thích (ví dụ, Filzmoser và cộng sự, 2018). Trong trường hợp đó, được hiểu là các biến hợp thành được chuyển đổi trước đó bằng phương pháp biến đổi ilr thích hợp trước khi được bao gồm trong ma trận X và mỗi biến như vậy tương ứng với một số cột của X

Viết mô hình hồi quy LAG trong không gian đơn hình

Bây giờ chúng ta bắt đầu với một mẫu có kích thước \( n \) của các vector hợp thành của \( S^D \), được lưu trữ trong một ma trận \( \mathbf{Y} \) có kích thước \( n \times D \) và với một biến đổi ilr được ký hiệu là \( \text{ilrY} \), tương ứng với một ma trận đối sánh \( \mathbf{VY} \) có kích thước \( D \times (D - 1) \). Chúng tôi giả định rằng trong số các biến giải thích, có \( K1 \) trong số chúng là cổ điển (không hợp thành) và chúng được thu thập vào một ma trận \( \mathbf{X1} \) có kích thước \( n \times K1 \). Phần còn lại của các biến giải thích (\( K2 \) trong số chúng) mang tính hợp thành, mỗi biến có thể có một số thành phần khác nhau \( L\_j \). Chúng tôi ký hiệu chúng bằng \( (\mathbf{x}\_j)^2 \) ( \( j = 1, \dots, K2 \) ).

Viết mô hình hồi quy liên kết Y với các biến giải thích bằng công thức ma trận như sau:

Ở đây, \( R \) và \( G \) là các ma trận \( D \times D \) thỏa mãn \( R\_j^D = G\_j^D = 0\_D \) và \( R\_j^T = G\_j^T = 0\_D \); \( \mathbf{b}\_1 \) là một vector \( K1 \times D \) của các tham số liên quan đến các biến giải thích cổ điển; và mỗi \( \mathbf{B}\_j \) là một ma trận các tham số liên quan đến biến giải thích hợp thành \( \mathbf{X2}\_j \) (số hàng của \( \mathbf{B}\_j \) là số \( L\_j \) thành phần của \( \mathbf{X2}\_j \) và số cột của nó là số \( D \) thành phần của \( \mathbf{Y} \)).

Bây giờ mục tiêu là trước hết chứng minh rằng nếu chúng ta chuyển mô hình này sang không gian tọa độ với một biến đổi ilr ilrY, chúng ta sẽ tìm thấy một mô hình như (5) và sau đó ánh xạ các tham số của một không gian sang các tham số của không gian khác. Để viết mô hình tương ứng với (7) trong không gian tọa độ, chúng ta cần xác định một lựa chọn của các biến đổi ilr cho mỗi biến hợp thành tương ứng với một ma trận đối sánh \( \mathbf{VX2}\_j \) cho biến \( \mathbf{X2}\_j \).

Áp dụng biến đổi ilrY tương ứng với \( \mathbf{V\_Y} \) vào phía bên phải của phương trình (7) và sử dụng phương trình (1), ta dễ dàng nhận thấy rằng:

Theo dõi Filzmoser, Chen và cộng sự, nguyễn và cộng sự, khi đó dễ dàng để chúng ta thấy:

Chúng ta thu được một phần của các thuật ngữ ở phía bên phải của (5). Các thuật ngữ duy nhất đòi hỏi một xử lý mới là các thuật ngữ tương ứng với các thuật ngữ kết quả trễ. Hãy trước tiên thiết lập một công thức thay thế cho tích ma trận (3). Giống như trong (9), hãy giới thiệu \( \mathbf{R}^\* \) sao cho:

Điều này cũng tương đương với \( \mathbf{RV\_Y} = \mathbf{V\_YR}^\* \) và với \( \mathbf{R}^\* = (\mathbf{V\_Y}^T \mathbf{Y}) \mathbf{RV\_Y} \). Khi đó, chúng ta có:

Áp dụng (13) à sử dụng (4) khi đó mang lại

Đó chính xác là công thức ma trận của số hạng nội sinh có độ trễ trong (5). Cuối cùng chúng tôi nhận được:

Xem xét từng cột của \( \mathbf{Y}^\* \) trong phương trình trên và so sánh với (5), chúng ta thấy rằng chúng trùng khớp miễn là \( \mathbf{b}^\* \) trong (5) bao gồm các tham số \( b1^\* \) và các tham số \( L\_j - 1 \) trong các cột của mỗi \( \mathbf{B}^\*\_j \), \( j = 1, \dots, K2 \) và tương ứng với đó, \( \mathbf{X} \) trong (5) bao gồm \( \mathbf{X1} \) và tất cả các cột của \( \mathbf{X2}^\*\_j \) cho \( j = 1, \dots, K2 \).

Các phương trình (9) đến (12) cũng cho phép chúng ta thiết lập mối liên hệ giữa các tham số trong không gian tọa độ cho các biến đổi ilr khác nhau. Quay trở lại không gian đơn giản sau khi đặt mô hình cho phép loại bỏ sự lựa chọn tùy ý có thể có của một biến đổi ilr. Những mối quan hệ này đúng cho các tham số dân số, và câu hỏi tiếp theo là liệu chúng vẫn còn đúng cho các tham số ước lượng hay không. Thực tế, kết quả vẫn đúng vì hai bước của phương pháp S2SLS dựa trên phương pháp OLS và chúng ta biết rằng phương pháp này bảo toàn mối quan hệ giữa các tham số sau khi ước lượng (ví dụ, Nguyen, 2019).

Liệu mô hình có nên được xây dựng trong không gian đơn giản hay không gian tọa độ? Trong quá trình xây dựng, thường đặt ra một số hạn chế về các tham số như việc một số trong số chúng có giá trị bằng không. Nếu công thức được xây dựng trong không gian tọa độ, các phương trình (9) và (12) chỉ ra rằng giả định phụ thuộc vào một lựa chọn cụ thể của tập hợp các biến đổi ilr và điều này không phù hợp với một phương pháp hợp thành. Do đó, người ta nên xây dựng mô hình trong không gian đơn giản và tránh việc áp đặt hạn chế giá trị bằng không cho các tham số của không gian tọa độ. Điều này có ba ảnh hưởng quan trọng. Thứ nhất, điều này cho biết rằng một giả định như \( \mathbf{R}^\* \) là đường chéo không mong muốn. Thứ hai, việc bao gồm một thuật ngữ \( \mathbf{Y} \otimes \mathbf{G} \) sẽ yêu cầu, vì lý do xác định, phải đặt một số tham số của \( \mathbf{G}^\* \) bằng không trong không gian tọa độ, điều này không phù hợp với một phương pháp hợp thành. Cuối cùng, việc không đặt hạn chế giá trị bằng không cho các tham số biến giải thích trong không gian tọa độ và cho các tham số ma trận \( \mathbf{R}^\* \) ngụ ý rằng các ước lượng S2SLS \( \tilde{d}^\* \) và S3SLS \( \hat{d}^\* \) là giống nhau (Greene, 2003, tr. 488).

Mô hình hồi quy thành phần

Bây giờ chúng ta mô tả mô hình hồi quy CoDa. \( \mathbf{Y}\_i \in S^L \) đề cập đến giá trị phản ứng hợp thành của quan sát thứ i, và \( \mathbf{X}^{(q)}\_i \in S^{D\_q} \), \( q = 1, \ldots , Q \), đề cập đến giá trị của biến đổi hợp thành thứ q cho quan sát thứ i, \( \mathbf{Z}\_{ki} \), \( k = 1, \ldots , K \), đề cập đến biến đổi cổ điển thứ k của quan sát thứ i. Như trong [3], hãy cho \(\otimes\) là tích ma trận hợp thành, tương ứng với tích ma trận trong không gian tọa độ thông qua biến đổi ilr.

Ở đây, \( \mathbf{x} \in S^D \), \( \mathbf{B} = (b\_{lj}) \), \( l = 1, \ldots , L \), \( j = 1, \ldots , D \), là một ma trận tham số sao cho \( \mathbf{j}^T \mathbf{LB} = \mathbf{0}^T\_D \) và \( \mathbf{B}\_{jD} = \mathbf{0}\_L \), và \( \mathbf{B}^\* \) liên quan đến \( \mathbf{B} \) bởi \( \mathbf{B}^\* = \mathbf{V}\_T \mathbf{LB} \mathbf{V}\_D \). Các điều kiện về \( \mathbf{B} \) đảm bảo rằng phép thao tác này là tuyến tính.

Các ký hiệu được sử dụng trong bài viết này được tóm tắt trong Bảng 1 (trong trường hợp của chúng tôi là n = 95, Q =3, K =5).  
Sau khi chọn các phép biến đổi ilr cho từng biến thành phần của mô hình, trước tiên chúng ta giới thiệu mô hình hồi quy CoDA trong không gian tọa độ ilr như sau:

Ở đây, \text{ilr}(\mathbf{Y}\_i), \text{ilr}(\mathbf{X}^{(q)}\_i) là các tọa độ ilr của \( \mathbf{Y}\_i \), \( \mathbf{X}^{(q)}\_i \) (q = 1, \ldots , Q) tương ứng; \( \mathbf{b}^\*\_0 \), \( \mathbf{B}^{(q)\*} \), \( \mathbf{c}^\*\_k \) là các tham số trong không gian tọa độ, và \text{ilr}(\boldsymbol{\varepsilon}\_i) là các dư thừa. Giả định về phân phối là \text{ilr}(\boldsymbol{\varepsilon}) tuân theo phân phối đa biến Gaussian với trung bình bằng không và ma trận hiệp phương sai \( \boldsymbol{\Sigma} \).

Mô hình hồi quy này cũng có thể được viết tương đương trong không gian đơn giản như sau:

Ở đây, \( b\_0 \), \( \mathbf{B}^{(1)} \), ... , \( \mathbf{B}^{(Q)} \), \( \mathbf{c}\_1 \), ... , \( \mathbf{c}\_K \) là các tham số thỏa mãn \( b\_0 \in S^L \), \( \mathbf{B}^{(q)} \) là một ma trận \( L \times D\_q \), \( q = 1, \ldots , Q \), \( \mathbf{c}\_k \in S^L \), \( k = 1, \ldots , K \), \( \mathbf{j}^T \mathbf{LB}^{(q)} = \mathbf{0}^T\_{D\_q} \), \( \mathbf{B}^{(q)}\_{jD\_q} = \mathbf{0}\_L \). Giả định về phân phối là \( \boldsymbol{\varepsilon}\_i \in S^L \) tuân theo phân phối Gaussian trên không gian đơn giản với trung bình bằng không và ma trận hiệp phương sai \( \Sigma \) liên quan đến lựa chọn biến đổi ilr cho biến phụ thuộc (xem, ví dụ, [1]).

Thường thì ta ước lượng mô hình (2) bằng phương pháp bình phương ít nhất riêng biệt cho mỗi tọa độ, do đó giả định độc lập giữa các tọa độ ilr (xem, ví dụ, [3,4]). Tuy nhiên, cũng có thể tưởng tượng về phương pháp cực đại ước lượng. Vì mật độ của các biến chuẩn hóa Gaussian được biểu diễn thông qua các tọa độ ilr (xem, ví dụ, [4]), cực đại ước lượng của các tham số trên đa giác tương đương với cực đại ước lượng của các tham số trong không gian tọa độ. Vì các dư thừa Gaussian trên đa giác tương ứng với các dư thừa Gaussian trong không gian tọa độ, việc áp dụng các phương pháp hồi quy đa biến tiêu chuẩn trong không gian tọa độ là đủ. Lưu ý rằng việc thực hiện các ước lượng bình phương ít nhất riêng biệt dẫn đến một trường hợp cụ thể của hồi quy đa biến với giả định bổ sung về sự độc lập giữa các tọa độ. Mối liên kết giữa các tham số (tương ứng với các ước lượng cực đại của tham số) trên đa giác và trong các tọa độ ilr được cho trong Đề xuất 2.1 (tương ứng với Đề xuất 2.2). Chen và cộng sự [3] cung cấp một số công thức liên quan đến các tham số trên đa giác và trong không gian tọa độ. (các tính chất 2.1 và 2.3). Chúng tôi mở rộng kết quả của họ cho trường hợp một biến đổi không hợp thành bổ sung và cho trường hợp bất kỳ biến đổi ilr nào trong Đề xuất 2.1. Chúng tôi sau đó chứng minh rằng các mối quan hệ này cũng đúng cho các ước lượng cực đại của tham số trong Đề xuất 2.2.

**Mệnh đề 2.1:** Trong mô hình (1)-(2), mối quan hệ giữa các tham số trong đơn hình và đối tượng của chúng trong không gian tọa độ được cho bởi:

Ở đây, \( \mathbf{V}\_L \) và \( \mathbf{V}\_{D\_q} \), \( q = 1, \ldots , Q \), là các ma trận tương phản liên quan đến các biến đổi ilr được lựa chọn.

Chứng minh: Các mối quan hệ giữa \( b\_0 \) và \( \mathbf{b}^\*\_0 \) và giữa \( \mathbf{c}\_k \) và \( \mathbf{c}^\*\_k \) được cung cấp trong [18].

Chúng ta có:

Theo địnhnghĩa của phép biến dổi ilr, chúng ta nhận được

Nhìn vào hai biểu thức của mô hình trong không gian tọa độ, chúng ta có thể nhận ra rằng \(\text{ilr}(\mathbf{B}^{(q)} \mathbf{x})\) tương đương với \(\mathbf{B}^{(q)\*} \mathbf{x}^\*\), sau đó chúng ta có:

Vì điều này đúng cho tất cả các \(\mathbf{x}\), ta có \( \mathbf{V}^T \mathbf{L} \mathbf{B}^{(q)} = \mathbf{B}^{(q)\*} \mathbf{V}^T\_{D\_q} \). Do đó, vì \( \mathbf{j}^T \mathbf{L} \mathbf{B}^{(q)} = \mathbf{0}^T\_{D\_q} \),

Bây giờ chúng ta hãy chú ý đến mối quan hệ giữa các ước lượng khả năng tối đa tương ứng.

Đề xuất 2.2: Đặt \( \hat{b}\_0 \) (tương ứng với \( \hat{\mathbf{B}}^{(q)} \) và \( \hat{c}\_k \)) là các ước lượng cực đại hợp lý của \( b\_0 \) (tương ứng với \( \mathbf{B}^{(q)} \) và \( c\_k \)) trong mô hình đa giác (2). Đặt \( \hat{b}^\*\_0 \) (tương ứng với \( \hat{\mathbf{B}}^{(q)\*} \) và \( \hat{c}^\*\_k \)) là các ước lượng cực đại hợp lý của \( b^\*\_0 \) (tương ứng với \( \mathbf{B}^{(q)\*} \) và \( c^\*\_k \)) trong mô hình không gian tọa độ ilr (1). Chúng ta có:

Chứng minh: Từ mối quan hệ giữa mật độ trong đa giác và mật độ trong không gian tọa độ, chúng ta thấy rằng việc tối đa hóa hàm hợp lý của \( Y \) đối với các tham số trong đa giác tương đương với việc tối đa hóa hàm hợp lý của \( Y^\* \) đối với các tham số trong không gian tọa độ. Chỉ cần nhìn thấy rằng Phương trình (3) xác định một mối quan hệ một một giữa \( b\_0 \), \( \mathbf{B}^{(q)} \), \( q = 1, \ldots, Q \), \( c\_k \), \( k = 1, \ldots, K \) và \( \hat{b}^\*\_0 \), \( \hat{\mathbf{B}}^{(q)\*} \), \( \hat{c}^\*\_k \) và sử dụng các tính chất của ước lượng cực đại hợp lý đối với biến đổi tham số.

Tác động của các biến giải thích thành phần và cổ điển

Trong phần 3.4. chúng tôi lập luận rằng việc giải thích các tham số của các mô hinh này không đơn giản như vậy [15] và đề xuất sử dụng đồ thị để minh họa mối quan hệ giữa tỷ lệ phiếu bầu dự đoán và các biến giải thích. Giá trị dự đoán của biến phụ thuộc cho các mô hình trên được cho bởi:

trong đó \( \hat{b}\_0 \), \( \hat{\mathbf{B}}^{(q)} \) và \( \hat{c}\_k \) là các tham số ước lượng. Chúng ta có thể viết lại (5) như sau:

Để minh họa các công thức này, chúng ta sẽ tập trung vào việc vẽ đồ thị các giá trị dự đoán của biến phụ thuộc dưới dạng hàm của một biến quan tâm cụ thể: hai trường hợp phải được xem xét tùy thuộc vào việc biến đó là cổ điển hay biến thành phần. Trong cả hai trường hợp, chúng tôi sẽ tạo một lưới các giá trị tiềm năng của giải thích cụ thể và sửa các biến giải thích khác theo các giá trị mà chúng lấy cho một điểm đã chọn của tập dữ liệu (chúng tôi lặp lại cho một số điểm đã chọn)

Đối với trường hợp khi biến cụ thể là một biến giải thích cổ điển \( Z\_i \), từ (6) có tồn tại \( \mathbf{a}^\wedge = ( \hat{a}\_1, \ldots, \hat{a}\_L)^T \in SL \) (điều này chứa các tác động của tất cả các biến giải thích khác nhưng là hằng số khi \( Z\_i \) biến đổi một mình) sao cho:

Với

Bây giờ, đối với trường hợp khi biến cụ thể là một biến hợp thành \( X\_i \), hãy lấy để đơn giản \( D\_q = 3 \) và \( L = 3 \). Như trước đó, từ (6), có tồn tại \( \mathbf{a}^\wedge \in SL \) (điều này chứa các tác động của tất cả các biến giải thích khác nhưng là hằng số khi \( X\_i \) biến đổi một mình) sao cho

Dữ liệu lượm them:  
Performance of the hotelling T 2 control chart for compositional data in the presence of measurement errors

[Complex thinking and social entrepreneurship. An approach from the methodology of compositional data analysis (sciencedirectassets.com)](https://pdf.sciencedirectassets.com/313379/1-s2.0-S2405844023X00036/1-s2.0-S2405844023006229/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEML%2F%2F%2F%2F%2F%2F%2F%2F%2F%2FwEaCXVzLWVhc3QtMSJHMEUCIDsLObmQGL1kr%2F5tf1GmA5ag104ecxfCgRiiiP1NPRCbAiEAyQ3uzVFnMc9LNZaOeesEfF7zaGL6FN4NioCcvZoa4qoqvAUI6%2F%2F%2F%2F%2F%2F%2F%2F%2F%2F%2FARAFGgwwNTkwMDM1NDY4NjUiDCVBPw15wOerpPHBHCqQBcBs%2BOyJ7aJThK79hf9PvUMcn8%2FyxAN1yhSawBwqi2u1gA%2B974ORlqtlnrFy0uKi0dg3LDQdLPaWEDQ3cCjHY%2FX0S0EZqutbutfgc7a2B5gtwN28j78h7rWSDXBHJ1eQjJjcVz9buMhuSIeymilA%2FyDyAZ6WNtE6mf9SacOtgW056W3woFHuwV8kbhNfjPGI5E4QL9xR8aYT7pPOz2sL6U5Om%2BF29Mgz44DClpNzr%2Boh7rT3bxYRqgLD22vlBVY4aIQIKLuiHqd3GDMLa5m%2BKoDuru1cEIIbLVbxm45pV1YRlFJhnq0R4hEDo0TWRCkFkgk4zki1W8KLJ%2Bgfp%2BDl6qsiwRM8DJ%2FK%2F56vNgG76UErB9U6Ke5TocT5Y7TxmDzL4xRyKL3tcudMWUw%2B6Uej8rbyd%2BKWDiWCWWp9YTNrOJAOIfSoa6wr2OESWekKqK5XJV3%2FbNGnz0fHDOjlHdzNPLe9ohvWYjx3C8JRQDDuxMW2IGKmetdK8zl1oA90Nn0mdxFnCCwjOYtfWWvsCpVsK01rhaYbr2NVBPmVGJnVad8Fnycgq7viYDXN3qisT5SXx0uJZg7WMpJ57R96kkZ7AL%2F1YK%2Bqq5l1k1tBvMwQmvNd8pI9u0gHD5ZPuWIdRCjSRTmDGtRPeefJC%2BinVU9iXyGi3QohHQ88wZgED5LEDlob5%2FvgRJs45AXM2utVPY%2B5vUyYmWPJNBLsizU959sgMAmh5r0J1AUJsn73VswXwWMtQcucwyJIZNtR0TiLTotNPcKY6Oi6bZJ4RNdscnKP5yLJXFyiJ67VeraswpgKTir9ZPH1%2FnOAZ94xgdNayKryiFBKSPlGDc0yUyZ1imfzF4zyg8O6hYDPg1VzKXOBbzr6MM38zrAGOrEBvrJ3Ai8W1Tkugzx%2F5Oe20kpGdqvS5pbxJ%2FtxNSpzAEl%2Bu7EjWmMeSzE%2FKcRwEH3ZmvhAJ3TzuHgLcw4MOcpAUlQWliDWmR8qMA%2FsXcKzL2YpWkctV5v%2FPmr3moplLxvcjuqi9ek0v1KIFJcVW7NHrN%2F%2FHlWWHW8UcSk18QPzyJ6%2BcKPrGBK44N8XaQFCH4ryTmLZwZ1lVUgUC15Z8Ood3E%2BWSeiag2BLw9flzG0O7rHX&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20240408T102955Z&X-Amz-SignedHeaders=host&X-Amz-Expires=300&X-Amz-Credential=ASIAQ3PHCVTYRBKXP6NN%2F20240408%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=a3b975842ba0408746d860445b7e5b6c2c2161b738ea2f1f1d5e692bdbed23ec&hash=b7c40652ce407ad087afb547bb6c1600bcd2b203f1efaf819479016e44d729c6&host=68042c943591013ac2b2430a89b270f6af2c76d8dfd086a07176afe7c76c2c61&pii=S2405844023006229&tid=spdf-7279c79b-4564-4ff9-896c-cd6d1b394ef8&sid=80b4badb495f524ae4598a49c0ea2740a7a8gxrqa&type=c)

<https://www.sciencedirect.com/science/article/pii/S2405844023006229>