# CÁC KỸ THUẬT HIỆU CHUẨN HÓA HỌC

* 1. **Tổng quan về hiệu chuẩn hóa học**

Hiệu chuẩn hoá học là một công cụ được sử dụng để trích xuất thông tin hữu ích và có liên quan từ dữ liệu hóa lý đo hoặc đã biết thô. Hiệu chuẩn hoá học dựa trên việc xây dựng, sau đó là khai thác một mô hình hành vi sử dụng các công cụ thống kê hay học máy. Hiệu chuẩn hoá học nằm trong phạm trù các hệ thống phức tạp và do đó thường là các hệ đa biến. Hiệu chuẩn hóa học thường được trình bày như một ngành nghiên cứu ứng dụng và những khó khăn gặp phải khá giống với những khó khăn của các bộ môn thực nghiệm. Vượt ra ngoài phạm vi hóa lý của các hệ thống được nghiên cứu, hiệu chỉnh hóa học bao gồm lĩnh vực toán học ứng dụng hơn là thiết bị đo đạc.

Mục tiêu chính trên hết của hiệu chuẩn hóa học là tìm kiếm ý nghĩa hóa học qua các thông tin về chúng. Chính ở góc độ này mà các phương pháp hướng dữ liệu, tích hợp thông tin hóa lý vào việc mô hình hóa hoặc vào giải pháp, phải được ưu tiên. Cách tiếp cận toán học là cần thiết để xây dựng các mô hình và giải thích kết quả thu được. Mục tiêu ban đầu cũng có thể mang tính lý thuyết nhiều hơn, việc phân tích các dữ liệu thực nghiệm sau đó dẫn đến đề xuất giả thuyết và kiểm tra chúng thông qua các mô hình vật lý. Từ một phương pháp tiếp cận hàn lâm, hiệu chỉnh hóa học kết hợp nhiều nguồn thông tin cho phép chúng ta có một cái nhìn khác về các vấn đề hóa lý được nghiên cứu. Bất chấp sự tồn tại của nhiều phần mềm ứng dụng, hiệu chỉnh hóa học vẫn là một lĩnh vực nghiên cứu của các chuyên gia có khả năng diễn giải số và hóa học các giải pháp được đề xuất.

* 1. **Các kĩ thuật tiền xử lý**

Tiền xử lý dữ liệu quang phổ cận hồng ngoại (NIR) đã trở thành một phần không thể thiếu trong mô hình hiệu chuẩn hóa học. Mục tiêu của quá trình tiền xử lý là loại bỏ các hiện tượng vật lý trong quang phổ để cải thiện hồi quy đa biến tiếp theo, mô hình phân loại hoặc phân tích thăm dò. Các kỹ thuật tiền xử lý được sử dụng rộng rãi nhất trong quang phổ cận hồng ngoại có thể được chia thành hai loại: phương pháp hiệu chỉnh tán xạ và đạo hàm phổ.

Các phương pháp tiền xử lý hiệu chỉnh tán xạ đầu tiên bao gồm hiệu chỉnh tán xạ nhân (Multiplicative Scatter Correction – MSC), hiệu chỉnh tán xạ nhân nghịch đảo (Inverse MSC), hiệu chỉnh tán xạ nhân mở rộng (Extended MSC – EMSC), hiệu chỉnh tán xạ nhân nghịch đảo mở rộng, biến chuẩn (Standard Normal Variate – SNV) và chuẩn hóa.

Đạo hàm phổ được đại diện bởi hai kĩ thuật: đạo hàm Norris-Williams (NW) và bộ lọc hàm đa thức Savitzky-Golay (SG). Cả hai phương pháp đều làm mịn phổ trước khi tính đạo hàm để giảm tác động bất lợi đến tỉ lệ tín hiệu trên nhiễu. Mục tiêu của bước tiền xử lý có thể là một trong ba mục tiêu sau:

## Cải thiện một phân tích thăm dò tiếp theo

## Cải thiện một mô hình hiệu chỉnh song tuyến tính tiếp theo (buộc dữ liệu phải tuân theo định luật Lambert-Beer)

* Cải thiện một mô hình phân lớp tiếp theo.

Các kĩ thuật tiền xử lý đều có mục tiêu là giảm độ biến thiên không được mô hình hóa trong dữ liệu để tăng đặc trưng trong phổ. Điều này có thể đạt được bằng cách sử dụng một kỹ thuật tiền xử lý phù hợp, nhưng luôn có nguy cơ áp dụng sai loại hoặc áp dụng quá trình tiền xử lý quá nghiêm ngặt sẽ loại bỏ thông tin có giá trị. Rất khó để đánh giá lựa chọn tiền xử lý phù hợp trước khi xác thực mô hình, nhưng nói chung, không nên thực hiện một số bước tiền xử lý và theo yêu cầu tối thiểu, tiền xử lý phải duy trì hoặc giảm độ phức tạp của mô hình hiệu quả.

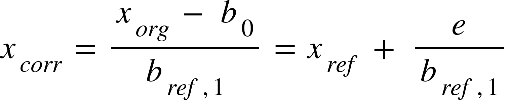
* + 1. **Hiệu chỉnh tán xạ**
       - 1. **Hiệu chỉnh tán xạ nhân**

Hiệu chỉnh tán xạ nhân (MSC) có thể là kĩ thuật tiền xử lý được sử dụng rộng rãi nhất cho phổ cận hồng ngoại (NIR). MSC lần đầu được giới thiệu bởi Martens và cộng sự vào năm 1983 và sau đó được đóng góp thêm bởi Geladi và cộng sự năm 1985. Ý tưởng đằng sau MSC là loại bỏ các hiệu ứng tán xạ không tốt khỏi ma trận dữ liệu trước khi mô hình hóa dữ liệu. MSC bao gồm hai bước:

1. Ước tính các hệ số hiệu chỉnh

x subscript o r g end subscript space equals space b subscript 0 space end subscript plus space b subscript r e f comma 1 end subscript space cross times space x subscript r e f end subscript space plus space e (1)

1. Hiệu chỉnh phổ

 (2)

Trong đó xorg là một mẫu phổ ban đầu được đo bằng thiết bị NIR, xref là phổ tham chiếu được sử dụng để tiền xử lý toàn bộ tập dữ liệu, e là phần không được mô hình hóa của xorg, xcorr là phổ đã được hiệu chỉnh, b0 và bref,1 là các tham số vô hướng, khác nhau đối với mỗi mẫu.

MSC cơ bản đã được mở rộng thành các phần phức tạp hơn, thường được gọi là EMSC (Expand MSC). Sự mở rộng này bao gổm cả việc đa thức bậc hai phù hợp với phổ tham chiếu, đường cơ sở phù hợp với trục bước sóng và sừ dụng kiến thức của phổ giao thoa trước đó. Phương trình của MSC:

blankblankError converting from MathML to accessible text. (3)

Trong đó lambda là vector hiệu chỉnh cho sự phụ thuộc trục bước sóng, bold italic x subscript bold k bold n bold o bold w bold n bold comma bold i end subscript là sự bao hàm của kiến thức tiên nghiệm đối với thông tin phổ mong muốn / không mong muốn (ví dụ: một phổ giao thoa đã biết), b là tập hợp các vô hướng (hệ số hiệu chỉnh) được cho bởi phương trình (4)

bold italic b bold space bold equals bold space stretchy left square bracket b subscript 0 space space space b subscript r e f comma 1 end subscript space space space b subscript r e f comma 2 end subscript space space space b subscript lambda comma 1 end subscript space space space b subscript lambda comma 2 end subscript space space space b subscript k n o w n comma 1 end subscript space space space b subscript k n o w n comma 2 end subscript space space space... stretchy right square bracket (4)

Trong đó b0 là hiệu chỉnh bù, bold italic b subscript bold r bold e bold f bold comma bold 1 end subscript là hiệu chỉnh theo thứ tự I của tham chiếu, bold italic b subscript bold lambda bold comma bold i end subscript là hiệu chỉnh của sự phụ thuộc bước sóng thứ i và bold italic b subscript bold k bold n bold o bold w bold n bold comma bold i end subscript là hiệu chỉnh của thông tin đã biết thứ i. Sự phụ thuộc vào trục bước sóng thường được bao gồm dưới dạng đa thức bậc hai phù hợp trên trục bước sóng với phổ. Khi không bao gồm hiệu chỉnh tham chiếu, bước sóng đơn phù hợp này cũng có tên là phổ xu thế giảm, nó có thể được xem như một hiệu chỉnh đường cơ sở. Điều quan trọng cần lưu ý là việc bao gồm sự phụ thuộc bước sóng trong Phương trình hiệu chỉnh đầy đủ (3) thay vì sử dụng nó như một bước riêng biệt sẽ dẫn đến hiệu ứng hiệu chỉnh nhỏ hơn. Điều này là do hoạt động đảo ngược ma trận được thực hiện đồng thời cho tất cả các tham số hiệu chỉnh trong MSC, trong đó các hiệu chỉnh khác nhau sẽ ảnh hưởng lẫn nhau trong tiêu chuẩn phù hợp bình phương nhỏ nhất. Khi sự phụ thuộc bước sóng được xác định một cách độc lập, chỉ có trục bước sóng (chứ không phải phổ tham chiếu) ảnh hưởng đến quá trình hiệu chỉnh, dẫn đến làm phẳng phổ được xử lý.

Theo ghi nhận của Pedersen và cộng sự, có một quy trình khá đơn giản để áp dụng phiên bản nghịch đảo của MSC, được gọi là Hiệu chỉnh tín hiệu nghịch đảo (Inverse Signal Correction - ISC). Việc ước tính các tham số hiệu chỉnh, hệ số b, được tìm thấy theo cách tương tự như MSC thông thường:

Error converting from MathML to accessible text. (5)

Có thể thấy là xorg và xref đã đổi chỗ cho nhau so với phương trình (3). Một lợi thế của ISC (Extended ISC - EISC) là sự đơn giản của phương trình hiệu chỉnh:blank

Error converting from MathML to accessible text. (6)

Trong ISC và EISC, cả việc ước tính các hệ số hiệu chỉnh và việc tự hiệu chỉnh đều được thực hiện theo phương pháp chuyển tiếp, thuận tiện cho việc đưa vào các thuật ngữ bổ sung và/hoặc tín hiệu tham chiếu. Hoạt động đảo ngược ma trận được đề cập trước đây cần thiết cho ước tính tham số trong MSC có thể dễ dàng trở thành điều kiện số nếu nó bao gồm các hiệu chỉnh tham chiếu đa thức bậc cao hơn. Đây là một lập luận có lợi cho ISC. Tuy nhiên, ISC giả định, theo phương pháp bình phương nhỏ nhất, rằng sai số trong phổ ghi được (cần hiệu chỉnh) nhỏ hơn sai số đối với phổ tham chiếu. Trong hầu hết các ứng dụng thực tế, tham chiếu là phổ trung bình được tính từ n mẫu trong bộ dữ liệu (ví dụ: bộ hiệu chuẩn). Mức nhiễu dự kiến cho tham chiếu này có cường độ square root of n nhỏ hơn phổ riêng lẻ (bỏ qua độ lệch do sự khác biệt về tán xạ trong tập hợp). Đây là lập luận chống lại ISC, vì một lỗi nhỏ trong quang phổ sẽ ảnh hưởng đến quá trình tiền xử lý ở mức độ lớn hơn so với MSC ban đầu.

Thách thức chính trong MSC là xác định phổ tham chiếu thích hợp.Giá trị này thường được đặt ở mức trung bình của phổ hiệu chuẩn. Gallagher và cộng sự đã trình bày một biến thiên tự nhiên đối với MSC bằng cách đưa sơ đồ trọng số vào trong bước xử lý trước. Hai lựa chọn thay thế đã được trình bày:

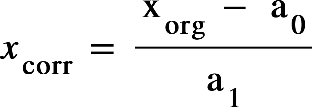
* Sử dụng một vector trọng số được xác định trước cho trục độ dài sóng
* Một vòng lặp tìm kiếm véc-tơ trọng số tối ưu.

Giải pháp dùng vòng lặp được tìm thấy bằng cách đặt trọng số thấp hơn cho các biến hoặc bước sóng có chênh lệch cao giữa dữ liệu thô và giải pháp hiệu chỉnh.  Việc tính toán các trọng số tiếp tục cho đến khi sự khác biệt giữa phổ đã hiệu chỉnh cho hai lần lặp tiếp theo nhỏ hơn mức nhiễu giả định trong dữ liệu. Thật không may, phương pháp khá đơn giản này không phải lúc nào cũng hoạt động tốt với dữ liệu NIR, vì độ mở rộng trong phạm vi bước sóng cao hơn thường biểu thị sự phân tán nhiều hơn và cần được hiệu chỉnh thay vì giảm trọng lượng.

Một giải pháp khác để tìm hiệu chỉnh tham chiếu trong MSC là Loopy MSC. Phương pháp này tìm phổ trung bình từ tập dữ liệu MSC đã được hiệu chỉnh. Sau đó, MSC được lặp lại nhiều lần để cập nhập phổ tham chiếu làm trung bình của tập dữ liệu được hiệu chỉnh sau mỗi bước lặp.

* + - * 1. **Chuẩn biến**

Tiền xử lý SNV có lẽ là phương pháp được áp dụng nhiều thứ hai để hiệu chỉnh tán xạ dữ liệu NIR. Định dạng cơ bản của SNV và chuẩn hóa hiệu chỉnh giống như định dạng cho MSC truyền thống:

 (7)

Với SNV, a0 là giá trị trung bình của phỗ mẫu cần hiệu chỉnh, a1 là độ lệch chuẩn của phổ mẫu. Với chuẩn hóa a0 được đặt bằng 0.

Các khái niệm hiệu chỉnh tín hiệu đằng sau SNV và chuẩn hóa giống như đối với MSC ngoại trừ tín hiệu tham chung là không bắt buộc. Thay vào đó, mỗi quan sát được xử lý riêng

* 1. **Cà**