# Mục lục

# Danh sách bảng

# Danh sách hình vẽ

# Danh sách ký hiệu

| | |
|---|---|
| $\mathbb{R}^{m\times n}$ | Tập các ma trận kích thước $m \times n$ trên trường số thực |
| $I$ | Ma trận đơn vị |
| $H^T$ | Chuyển vị của ma trận $H$ |
| $A^{-1}$ | Nghịch đảo của ma trận vuông, không suy biến $A$ |
| $\mathrm{rank}(H)$ | Hạng của ma trận $H$ |
| $\det(A), \lvert A\rvert$ | Định thức của ma trận vuông $A$ |
| $\mathrm{Span}(H)$ | Không gian tuyến tính sinh bởi các cột của ma trận $H$ |
| $\mathcal{N}(H)$ | Không gian nghiệm (*null space*) sinh bởi ma trận $H$ |
| $\dim(S)$ | Số chiều của không gian tuyến tính $S$ |
| $\mathrm{diag}(a_1, \ldots, a_n)$ | Ma trận đường chéo kích thước $n \times n$ |
| $\mathrm{diag}(D_1, D_2)$ | Ma trận đường chéo khối |
| $\mathrm{Tr}(A)$ | Vết (*trace*) của ma trận vuông $A$ |
| $E(.)$ | Phép lấy kỳ vọng toán học |
| $H^+$ | Giả nghịch đảo Moore-Penrose của ma trận $H$ |
| $\mathcal{L}_1 - \mathcal{L}_2$ | Tập các vector thuộc $\mathcal{L}_1$ và trực giao với $\mathcal{L}_2$ |
| $\mathcal{L}^\perp$ | Phần bù trực giao của đa tạp tuyến tính $\mathcal{L}$ |
| $\mathcal{N}(\mu, \Sigma)$ | Phân phối xác suất chuẩn (phân phối Gaussian) với kỳ vọng $\mu$ và ma trận hiệp phương sai $\Sigma$ |

# Lời nói đầu

Mô hình hồi quy tuyến tính là mô hình đơn giản và phổ biến nhất trong thực tế, đặc biệt là trong các lĩnh vực kinh tế, tài chính và môi trường. Việc phát triển các thuật toán ước lượng tham số cho mô hình hồi quy tuyến tính đã đạt được nhiều tiến bộ to lớn trong những thập kỷ qua. Tuy nhiên, trong thời đại bùng nổ dữ liệu như hiện nay, nhiều vấn đề thực tế đặt ra cho bài toán ước lượng tham số vẫn còn là thách thức cho các nhà nghiên cứu. Kích thước dữ liệu thu thập quá lớn không chỉ yêu cầu các công nghệ tính toán mạnh mẽ mà còn đòi hỏi các thuật toán xử lý hiệu quả hơn. Do đó, trong đồ án này, em sẽ xây dựng một thuật toán truy hồi cho ước lượng tham số trong mô hình hồi quy tuyến tính. Thuật toán này nhằm làm giảm độ phức tạp tính toán của các mô hình quan sát có số chiều lớn, đồng thời đưa ra một phương pháp chính quy hóa cho bài toán không đặt chỉnh nhằm ổn định ước lượng thu được.

Nội dung đồ án được trình bày trong 3 chương:

i. Chương 1: nhắc lại các kiến thức toán học về ma trận và mô hình hồi quy tuyến tính sẽ được sử dụng trong đồ án.

ii. Chương 2: trình bày thuật toán truy hồi cho ước lượng tham số trong mô hình hồi quy tuyến tính và các tính chất toán học của nó.

iii. Chương 3: giới thiệu hai ví dụ mô phỏng số nhằm minh họa cho các kết quả lý thuyết của thuật toán.

Tính khoa học và tính sáng tạo của đồ án nằm ở quá trình xây dựng một thuật toán truy hồi nhằm làm giảm độ phức tạp tính toán cho ước lượng tham số trong mô hình hồi quy tuyến tính khi dữ liệu quan sát và tập trạng thái cần ước lượng có kích thước lớn (trong thực tế cỡ $10^4 - 10^7$). Các tính chất toán học và thống kê quan trọng của phương pháp ước lượng này cũng được chứng minh chặt chẽ. Các ví dụ mô phỏng số dựa trên phương pháp Monte-Carlo được thực hiện nhằm kiểm tra các tính chất và sự đúng đắn của thuật toán. Phương pháp truy hồi này có ý nghĩa quan trọng trong nhiều bài toán thực tế, đặc biệt được áp dụng trong khai phá dữ liệu khí tượng và hải dương học.

Để đạt được những kết quả này, em xin gửi lời cảm ơn sâu sắc nhất tới **PGS. TS. Tống Đình Quỳ**, người thầy đã tận tình hướng dẫn em hoàn thành nội dung đồ án. Em cũng xin gửi những lời cảm ơn chân thành tới **TS. Tạ Thị Thanh Mai** và tập thể lớp **KSTN Toán Tin K57** đã luôn ủng hộ và giúp đỡ em trong quá trình thực hiện đồ án. Em xin cảm ơn các thầy cô trong Viện Toán ứng dụng và Tin học, Đại học Bách Khoa Hà Nội đã dành sự quan tâm, chỉ bảo cũng như tạo mọi điều kiện thuận lợi nhất cho em trong suốt quá trình 5 năm học tập tại trường.

Mặc dù đồ án được hoàn thành với nhiều nỗ lực cố gắng, tuy nhiên không thể tránh khỏi những thiếu sót. Em mong nhận được sự đóng góp ý kiến của thầy cô và các bạn để nội dung đồ án được hoàn thiện hơn.

Em xin chân thành cảm ơn!

Hà Nội, ngày 6 tháng 6 năm 2017

Sinh viên thực hiện .

**Lê Văn Chiến**

# Giới thiệu

Thống kê và các ngành khoa học dữ liệu ngày càng có vai trò quan trọng trong thực tế, xuất phát từ sự bùng nổ dữ liệu trong những năm gần đây cũng như nhu cầu tìm hiểu thông tin ngày càng cao của con người. Trong các bài toán dữ liệu, người ta tìm cách ước lượng các thông tin chưa biết từ các thông tin thu thập được, bằng cách xem xét mối quan hệ phụ thuộc giữa chúng. Rất nhiều mô hình toán học đã được ra đời nhằm mô tả các mối quan hệ đó, trong đó phổ biến nhất là các mô hình thống kê, mô hình chuỗi thời gian cũng như áp dụng trí tuệ nhân tạo trong khai phá dữ liệu.

Mô hình hồi quy tuyến tính là mô hình thống kê đơn giản và phổ biến nhất trong mọi lĩnh vực của đời sống xã hội, đặc biệt trong kinh tế, tài chính và môi trường. Trong mô hình này, các thông tin chưa biết được ước lượng thông qua các dữ liệu thu thập được, dựa trên ràng buộc tuyến tính giữa chúng. Trong những thập kỷ qua, việc phát triển các thuật toán ước lượng tham số cho mô hình hồi quy tuyến tính đã đạt được nhiều tiến bộ to lớn, thể hiện qua nguồn tài liệu, nghiên cứu dồi dào cũng như các ứng dụng thực tế của nó.

Tuy nhiên, khi khối lượng dữ liệu thu thập được ngày càng lớn và phức tạp, các bài toán thực tế đặt ra cho ước lượng tham số vẫn còn là thách thức không hề nhỏ đối với các nhà nghiên cứu. Vấn đề cơ bản của bài toán ước lượng là xây dựng một phương pháp hiệu quả nhằm đảm bảo tính ổn định của ước lượng, đồng thời có khả năng mở rộng để giải quyết các bài toán phức tạp với kích thước dữ liệu lớn. Một cách tổng quát, trong thực tế,

các thông tin đã biết của mô hình hồi quy tuyến tính thường chỉ được xác định bởi các giá trị xấp xỉ của chúng. Do đó, các hệ tuyến tính không đặt chỉnh có thể dẫn đến các ước lượng không ổn định và có sai số ước lượng lớn. Có nhiều phương pháp khác nhau đã được đề xuất để giải quyết một số trường hợp của hệ không đặt chỉnh, trong đó, tiêu biểu có thể kể đến phương pháp chính quy hóa phụ thuộc vào hai tham số $\alpha$ và $\beta$ cho hệ tuyến tính với ma trận hiệp phương sai của nhiễu là ma trận suy biến đã được đề xuất trong [HB13].

Một khó khăn khác đặt ra cho bài toán ước lượng khi khối lượng dữ liệu thu thập được ngày càng khổng lồ. Kích thước dữ liệu quá lớn dẫn đến khối lượng tính toán lớn, không chỉ yêu cầu các nền tảng tính toán, công nghệ xử lý mạnh mẽ mà còn đòi hỏi các thuật toán hiệu quả hơn. Một ví dụ trong khai phá dữ liệu khí tượng và hải dương học, hệ quan sát tuyến tính thường có kích thước cỡ $10^6 - 10^7$ [Dal91]. Điều này xảy ra do các thông tin cần ước lượng nằm trên lưới trong không gian ba chiều. Những thuật toán ước lượng mẫu cho hệ hồi quy với số chiều lớn đầu tiên đã được đề xuất bởi các nghiên cứu gần đây [HBT01, HB11b, HB14a].

Trong nội dung đồ án này, em sẽ trình bày một thuật toán ước lượng tham số cho mô hình hồi quy tuyến tính sử dụng công thức truy hồi nhằm giảm độ phức tạp tính toán cho các hệ quan sát có kích thước lớn. Đồ án cũng chỉ ra rằng, thuật toán truy hồi cho phép chính quy hóa các ước lượng thu được từ các hệ quan sát không đặt chỉnh, bằng cách lựa chọn ma trận hiệp phương sai ban đầu của vector tham số cần ước lượng. Các ví dụ mô phỏng số được thực hiện nhằm kiểm tra các kết quả lý thuyết của thuật toán. Ứng dụng thực tế của thuật toán trong khai phá dữ liệu khí tượng và hải dương học có thể xem tại [HB14b]. Chi tiết mã nguồn cài đặt của thuật toán có thể tham khảo tại https://github.com/lvchien/thesis.

# Chương 1

# Các kiến thức cần chuẩn bị

## 1.1 Mô hình hồi quy tuyến tính

Trong phần này, khái niệm và các tính chất của mô hình hồi quy tuyến tính tổng quát sẽ được giới thiệu. Mô hình này sẽ được sử dụng xuyên suốt trong toàn bộ nội dung đồ án.

**Định nghĩa 1.1** (Mô hình hồi quy tuyến tính)**.** *Mô hình hồi quy tuyến tính tổng quát được cho dưới dạng:*

$$z = Hx + v, \tag{1.1}$$

*trong đó,* $x \in \mathbb{R}^n$ *là biến độc lập,* $z \in \mathbb{R}^p$ *là biến phụ thuộc,* $H \in \mathbb{R}^{p\times n}$ *là ma trận hệ số,* $v \in \mathbb{R}^p$ *là nhiễu mô hình (*$n, p$ *là hai số nguyên dương bất kỳ).*

Mô hình hồi quy tuyến tính là mô hình đơn giản và phổ biến nhất trong thống kê, đặc biệt trong các lĩnh vực kinh tế, tài chính, môi trường, .... Trong giới hạn của đồ án này, mô hình hồi quy tuyến tính được sử dụng để mô tả mối quan hệ tuyến tính giữa tập các biến trạng thái và tập các quan sát của một hệ quan sát ngẫu nhiên. Khi đó, $z$ là tập các giá trị quan sát được của hệ, $x$ là tập các trạng thái chưa biết cần ước lượng và $H$ là ma trận hệ số quan sát đã biết.

Ta giả sử rằng, mô hình quan sát (1.1) thỏa mãn các điều kiện sau:

$$E(v) = 0,\ E(vv^T) = V, \tag{1.2}$$

$$E(x) = \bar{x},\ E(ee^T) = M,\ E(ev^T) = N,\ e := x - \bar{x}. \tag{1.3}$$

Trong nhiều bài toán, ma trận hiệp phương sai $V$ của nhiễu $v$ thường có dạng ma trận đường chéo $V = \alpha I$, với $\alpha$ là một hằng số thực không âm. Nếu $\alpha = 0$ thì phương trình (1.1) được gọi là mô hình quan sát không nhiễu. Ngược lại nếu $\alpha > 0$, điều kiện (1.2) cho ta một mô hình quan sát với chuỗi nhiễu ngẫu nhiên $\{v_i\}$ không tương quan có cùng phân phối.

Đặt $\tilde{v} = \left(e^T, v^T\right)^T$. Khi đó, ma trận hiệp phương sai của $\tilde{v}$ có thể biểu diễn dưới dạng ma trận khối:

$$E\left(\tilde{v}\tilde{v}^T\right) = \left[\begin{array}{c|c} M & N \\ \hline N^T & V \end{array}\right]$$

Các ma trận hiệp phương sai $M$ và $V$ là đối xứng và nửa xác định dương, nhưng có thể là ma trận suy biến (sẽ được trình bày trong phần sau). Do đó, hệ (1.1, 1.2, 1.3) được gọi là mô hình hồi quy tuyến tính với hiệp phương sai không âm bất kỳ. Như vậy, không có giả thiết nào khác được đặt ra với hệ quan sát (1.1) ngoại trừ phân phối xác suất của $\tilde{v}$ và các số nguyên dương $p, n$ cho trước.

## 1.2 Ma trận và các tính chất liên quan

Trong phần này, ta sẽ nhắc lại một số định nghĩa và các tính chất của ma trận sẽ được sử dụng trong đồ án. Trước hết, ta định nghĩa chuẩn của vector $x \in \mathbb{R}^n$:

$$\|x\| = \left[\sum_{i=1}^{n} x_i^2\right]^{\frac{1}{2}} = \left(x^T x\right)^{\frac{1}{2}}.$$

### 1.2.1 Không gian trực giao và các tính chất

Một số khái niệm cơ bản của đại số tuyến tính và các tính chất của chúng sẽ được nhắc lại sau đây: đa tạp tuyến tính, vector trực giao với đa tạp tuyến tính và phần bù trực giao.

**Định nghĩa 1.2** (Đa tạp tuyến tính)**.** *Đa tạp tuyến tính $\mathcal{L}$ trên không gian Euclidean $\mathbb{R}^n$ là một tập con khác rỗng của $\mathbb{R}^n$ đóng với các phép cộng và nhân vô hướng, nghĩa là nếu $x, y$ là hai phần tử của $\mathcal{L}$, $\alpha$ và $\beta$ là hai vô hướng thì $\alpha x + \beta y \in \mathcal{L}$.*

**Định nghĩa 1.3** (Vector trực giao)**.** *Vector $x \in \mathbb{R}^n$ được gọi là trực giao với đa tạp tuyến tính $\mathcal{L}$, ký hiệu là $x \perp \mathcal{L}$, nếu $x$ trực giao với mọi phần tử của $\mathcal{L}$, nghĩa là $x^T y = 0,\ \forall y \in \mathcal{L}$.*

**Tính chất 1.3.1.** *Cho $\mathcal{L}_1, \mathcal{L}_2$ là hai đa tạp tuyến tính trên không gian $\mathbb{R}^n$. Nếu $\mathcal{L}_1 \subseteq \mathcal{L}_2$ thì $\mathcal{L}_2 - \mathcal{L}_1$, được định nghĩa là tập tất cả các vector thuộc $\mathcal{L}_2$ và trực giao với $\mathcal{L}_1$, cũng là một đa tạp tuyến tính.*

**Định nghĩa 1.4** (Phần bù trực giao)**.** *Cho $\mathcal{L}$ là một đa tạp tuyến tính trên $\mathbb{R}^n$. Phần bù trực giao của $\mathcal{L}$, ký hiệu là $\mathcal{L}^\perp$, là tập hợp tất cả các vector trực giao với $\mathcal{L}$.*

Dễ dàng chứng minh được rằng $\mathcal{L}^\perp = \mathbb{R}^n - \mathcal{L}$ và do đó $\left(\mathcal{L}^\perp\right)^\perp = \mathcal{L}$, với $\mathcal{L}$ là một đa tạp tuyến tính bất kỳ trên $\mathbb{R}^n$. Tiếp theo, ta sẽ nhắc lại phép chiếu trực giao một vector lên đa tạp tuyến tính và các tính chất liên quan.

**Định nghĩa 1.5** (Phép chiếu trực giao)**.** *Cho $x \in \mathbb{R}^n$ và $\mathcal{L}$ là một đa tạp tuyến tính trên $\mathbb{R}^n$. Khi đó, tồn tại duy nhất một vector $\hat{x} \in \mathcal{L}$ sao cho $x - \hat{x} \perp \mathcal{L}$, nghĩa là tồn tại duy nhất một phép phân tích của $x$ :*

$$x = \hat{x} + \tilde{x}, \tag{1.4}$$

*trong đó, $\hat{x} \in \mathcal{L}$ và $\tilde{x} \in \mathcal{L}^\perp$. Vector $\hat{x}$ được gọi là chiếu trực giao của $x$ trên $\mathcal{L}$.*

**Tính chất 1.5.1.** *Cho $x \in \mathbb{R}^n$, $\mathcal{L}$ là một đa tạp tuyến tính trên $\mathbb{R}^n$, $\hat{x}$ là chiếu trực giao của $x$ trên $\mathcal{L}$. Khi đó, với mọi $y \in \mathcal{L}$, $y \neq \hat{x}$, ta có:*

$$\|x - y\| > \|x - \hat{x}\| .$$

Tính chất 1.5.1 nói rằng, chiếu trực giao của $x$ là vector "gần nhất" của $x$ trên $\mathcal{L}$. Sau đây, ta sẽ định nghĩa hai đa tạp tuyến tính thường được sử dụng trong thực hành.

**Định nghĩa 1.6** (Không gian sinh)**.** *Cho ma trận $H \in \mathbb{R}^{m \times n}$. Không gian sinh bởi các cột của ma trận $H$, ký hiệu là $\mathrm{Span}(H)$, là không gian tuyến tính được xác định bởi:*

$$Span(H) = \{y \in \mathbb{R}^m | \exists x \in \mathbb{R}^n : y = Hx\} . \tag{1.5}$$

**Định nghĩa 1.7** (Không gian nghiệm)**.** *Cho ma trận $H \in \mathbb{R}^{m \times n}$. Không gian nghiệm (null space) sinh bởi ma trận $H$, ký hiệu là $\mathcal{N}(H)$, là không gian tuyến tính được xác định bởi:*

$$\mathcal{N}(H) = \{x \in \mathbb{R}^n | Hx = 0\} . \tag{1.6}$$

**Tính chất 1.7.1.** $\mathrm{Span}(H^T) = \mathcal{N}^{\perp}(H)$.

**Tính chất 1.7.2.** *Cho ma trận $H \in R^{m \times n}$. Khi đó, tồn tại duy nhất một phép phân tích của $x \in \mathbb{R}^n$ có dạng:*

$$x = \hat{x} + \tilde{x}, \tag{1.7}$$

*trong đó, $\hat{x} \in \mathrm{Span}(H^T)$ và $\tilde{x} \in \mathcal{N}(H)$.*

**Định nghĩa 1.8** (Ma trận không suy biến)**.** *Ma trận vuông $A$ được gọi là ma trận không suy biến nếu không gian nghiệm của nó chỉ chứa vector 0. Ngược lại, $A$ được gọi là ma trận suy biến.*

Nếu $A$ là ma trận không suy biến thì định thức $\det(A) > 0$, đồng thời tồn tại ma trận nghịch đảo $A^{-1}$ của $A$.

### 1.2.2 Vết của ma trận và đạo hàm vết

**Định nghĩa 1.9** (Vết của ma trận)**.** *Cho ma trận vuông $A = (a_{ij}) \in \mathbb{R}^{n\times n}$. Vết của $A$, ký hiệu là $\mathrm{Tr}(A)$, là tổng các phần tử trên đường chéo chính của $A$:*

$$\mathrm{Tr}(A) = \sum_{i=1}^{n} a_{ii}. \tag{1.8}$$

Các tính chất của vết của ma trận:

i. $\mathrm{Tr}(A+B) = \mathrm{Tr}(A) + \mathrm{Tr}(B)$, $\mathrm{Tr}(\lambda A) = \lambda\,\mathrm{Tr}(A)$, với $\lambda \in \mathbb{R}$ bất kỳ,

ii. $\mathrm{Tr}(A^T) = \mathrm{Tr}(A)$,

iii. $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$, với $AB$ và $BA$ là các ma trận vuông.

Để tính đạo hàm vết, ta nhắc lại định nghĩa đạo hàm hàm vô hướng của ma trận:

**Định nghĩa 1.10** (Đạo hàm ma trận)**.** *[MN07] Cho $\phi$ là hàm vô hướng, khả vi theo mọi phần tử của ma trận $X \in \mathbb{R}^{m\times n}$. Khi đó, đạo hàm ma trận $\dfrac{\partial\phi(X)}{\partial X}$ được xác định bởi:*

$$\frac{\partial\phi(X)}{\partial X} = \left(\frac{d\phi(x_{ij})}{dx_{ij}}\right)_{m\times n} \tag{1.9}$$

Đạo hàm vết của một số hàm ma trận thông dụng được cho bởi bảng 1.1.

Bảng 1.1: Đạo hàm vết của một số hàm ma trận thông dụng

| $F(X)$ | $\partial_X \mathrm{Tr}(F)$ | $F(X)$ | $\partial_X \mathrm{Tr}(F)$ | $F(X)$ | $\partial_X \mathrm{Tr}(F)$ |
|---|---|---|---|---|---|
| $X$ | $I$ | $AX^T$ | $A$ | $XX^TB$ | $BX + B^TX$ |
| $XA$ | $A^T$ | $X^2$ | $2X^T$ | $XBX^T$ | $XB^T + XB$ |
| $AXB$ | $A^TB^T$ | $X^2B$ | $XB + BX$ | $BX^TX$ | $XB^T + XB$ |
| $AX^TB$ | $BA$ | $X^TBX$ | $BX + B^TX$ | $X^TXB$ | $XB^T + XB$ |
| $X^TA$ | $A$ | $BXX^T$ | $BX + B^TX$ | $X^TX$ | $2X$ |

### 1.2.3 Trị riêng, vector riêng và phép phân tích phổ

**Định nghĩa 1.11** (Trị riêng và vector riêng)**.** *Cho ma trận vuông $A \in \mathbb{R}^{n\times n}$. Các giá trị riêng của $A$ được định nghĩa là nghiệm của phương trình đặc trưng:*

$$|\lambda I - A| = 0. \tag{1.10}$$

*Cho $\lambda$ là một trị riêng của $A$. Khi đó vector $x \neq 0$ thỏa mãn:*

$$Ax = \lambda x \tag{1.11}$$

*được gọi là một vector riêng của $A$ tương ứng với giá trị riêng $\lambda$.*

Sau đây là một số tính chất của trị riêng và vector riêng. Các chứng minh có thể xem chi tiết tại [MN07].

**Tính chất 1.11.1.** *Ma trận thực, đối xứng chỉ có các giá trị riêng thực.*

**Tính chất 1.11.2.** *Giả sử $A$ và $G$ là các ma trận vuông kích thước $n \times n$, đồng thời $G$ không suy biến. Khi đó, $A$ và $G^{-1}AG$ có cùng tập các giá trị riêng.*

**Tính chất 1.11.3.** *Ma trận suy biến có ít nhất một giá trị riêng bằng không.*

**Tính chất 1.11.4.** *Một ma trận là xác định dương (nửa xác định dương) nếu và chỉ nếu tất cả các trị riêng của nó đều dương (không âm).*

**Tính chất 1.11.5.** *Cho ma trận vuông $A \in \mathbb{R}^{n\times n}$ với các trị riêng $\lambda_1, \lambda_2, \ldots, \lambda_n$. Khi đó, ta có:*

$$\mathrm{Tr}(A) = \sum_{i=1}^{n} \lambda_i, \qquad |A| = \prod_{i=1}^{n} \lambda_i. \tag{1.12}$$

**Tính chất 1.11.6.** *Các vector riêng tương ứng với các trị riêng phân biệt thì độc lập tuyến tính với nhau.*

Xét ma trận đối xứng $A \in \mathbb{R}^{n\times n}$. Ta có phép phân tích phổ (*eigendecomposition* hoặc *spectral decomposition*) của ma trận $A$:

$$A = UDU^T, \tag{1.13}$$

trong đó,

$$D = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix}$$

và $U = (x_1, x_2, \ldots, x_n)$, với $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ là các trị riêng thực của $A$ và $x_1, x_2, \ldots, x_n$ là các vector riêng tương ứng.

Xét ma trận đối xứng nửa xác định dương $M \in \mathbb{R}^{n\times n}$. Khi đó, $M$ có $n$ giá trị riêng thực không âm $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$. Xét ma trận:

$$S = \operatorname{diag}\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_n}\right).$$

Khi đó, ta có $S^2 = SS^T = D$. Đặt $L = US$, ta có phân tích của ma trận $M$:

$$M = UDU^T = USS^TU^T = LL^T. \tag{1.14}$$

Trường hợp đặc biệt, $M \in \mathbb{R}^{n\times n}$ là ma trận đối xứng xác định dương, nghĩa là $M$ có $n$ trị riêng dương $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0$. Khi đó, $D$ và $M$ là các ma trận khả nghịch:

$$D^{-1} = \operatorname{diag}\left(\lambda_1^{-1}, \lambda_2^{-1}, \ldots, \lambda_n^{-1}\right) = \begin{bmatrix} \lambda_1^{-1} & 0 & \ldots & 0 \\ 0 & \lambda_2^{-1} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n^{-1} \end{bmatrix}$$

Ta có phân tích phổ của ma trận $M$ và ma trận nghịch đảo $M^{-1}$:

$$M = UDU^T, \qquad M^{-1} = UD^{-1}U^T. \tag{1.15}$$

### 1.2.4 Ma trận giả nghịch đảo Moore-Penrose

Ma trận nghịch đảo chỉ được định nghĩa cho ma trận vuông không suy biến. Tuy nhiên, trong nhiều bài toán, ta quan tâm đến khái niệm "nghịch đảo" của ma trận không vuông hoặc ma trận vuông suy biến. Khi đó, khái niệm ma trận giả nghịch đảo *Moore-Penrose* thường được sử dụng vì tính duy nhất của nó.

**Định nghĩa 1.12** (Giả nghịch đảo Moore-Penrose)**.** *[Alb72] Cho ma trận $H \in \mathbb{R}^{m\times n}$. Khi đó, giới hạn:*

$$H^{+} = \lim_{\delta\to 0} \left(H^T H + \delta^2 I\right)^{-1} H^T = \lim_{\delta\to 0} H^T \left(HH^T + \delta^2 I\right)^{-1} \tag{1.16}$$

*luôn tồn tại và được gọi là giả nghịch đảo Moore-Penrose của ma trận $H$.*

Một định nghĩa khác tương đương cũng thường được sử dụng:

**Định nghĩa 1.13** (Giả nghịch đảo Moore-Penrose)**.** *Ma trận $H^{+} \in \mathbb{R}^{n\times m}$ được gọi là giả nghịch đảo Moore-Penrose của ma trận $H \in \mathbb{R}^{m\times n}$ nếu thỏa mãn đồng thời các điều kiện sau:*

*i.* $HH^{+}H = H$

*ii.* $H^{+}HH^{+} = H^{+}$

*iii.* $(HH^{+})^T = HH^{+}$

*iv.* $(H^{+}H)^T = H^{+}H$

*v.* $0^{+} = 0$.

Sau đây là một số tính chất của ma trận giả nghịch đảo Moore-Penrose. Các chứng minh chi tiết có thể xem tại [Alb72, MN07].

**Tính chất 1.13.1.** *Với mọi ma trận $H$, giả nghịch đảo Moore-Penrose $H^+$ của $H$ đều tồn tại và duy nhất, đồng thời $\left(H^+\right)^+ = H$.*

**Tính chất 1.13.2.** *Nếu $H$ là ma trận vuông không suy biến thì $H^+ = H^{-1}$.*

**Tính chất 1.13.3.** *Với mọi ma trận $H$, ta có:*

$$\left(H^T\right)^+ = \left(H^+\right)^T, \qquad H^+ = \left(H^T H\right)^+ H^T = H^T \left(HH^T\right)^+.$$

**Tính chất 1.13.4.** *Nếu $H$ là ma trận $1 \times 1$ (vô hướng) thì:*

$$H^+ = \lim_{\delta \to 0} \left(H^2 + \delta^2 I\right)^{-1} H = \begin{cases} 0 & \text{nếu } H = 0 \\ \dfrac{1}{H} & \text{nếu } H \neq 0. \end{cases}$$

**Tính chất 1.13.5.** *Nếu $H$ là ma trận đường chéo*

$$H = \operatorname{diag}(a_1, a_2, \ldots, a_n)$$

*thì*

$$H^+ = \operatorname{diag}(a_1^+, a_2^+, \ldots, a_n^+).$$

Xét ma trận đối xứng $H$, ta có phân tích phổ của $H$:

$$H = UDU^T,$$

trong đó, $U$ là ma trận trực giao, $D$ là ma trận đường chéo. Khi đó, áp dụng định nghĩa 1.12, ta có giả nghịch đảo Moore-Penrose của $H$:

$$\begin{aligned} H^+ &= \lim_{\delta \to 0} U \left(D^2 + \delta^2 I\right)^{-1} DU^T \\ &= U \left[\lim_{\delta \to 0} \left(D^2 + \delta^2 I\right)^{-1} D\right] U^T = UD^+U^T. \end{aligned} \tag{1.17}$$

Trường hợp đặc biệt, $H$ là ma trận đối xứng xác định dương, nghĩa là $H$ có $n$ giá trị riêng dương $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0$. Khi đó, ta có:

$$\begin{aligned} D^+ &= \operatorname{diag}\left(\lambda_1^+, \lambda_2^+, \ldots, \lambda_n^+\right) = \operatorname{diag}\left(\lambda_1^{-1}, \lambda_2^{-1}, \ldots, \lambda_n^{-1}\right) = D^{-1}, \\ H^+ &= UD^+U^T = UD^{-1}U^T = H^{-1}. \end{aligned}$$

## 1.3 Ước lượng tham số cho mô hình hồi quy tuyến tính

Xét bài toán ước lượng vector tham số chưa biết $x$ trong mô hình hồi quy tuyến tính (1.1, 1.2, 1.3). Do $E(v) = 0$ nên cách đơn giản nhất để ước lượng $x$ là bỏ qua nhiễu $v$ và giải hệ phương trình $z \approx Hx$. Nhắc lại rằng, hệ phương trình tuyến tính $z = Hx$ là tương thích nếu $z \in \mathrm{Span}(H)$, nghĩa là tồn tại $x^0 \in \mathbb{R}^n$ sao cho $Hx^0 = z$. Sau đây, ta định nghĩa khái niệm bài toán đặt chỉnh và không đặt chỉnh:

**Định nghĩa 1.14** (Bài toán đặt chỉnh)**.** *[TA77] Xét bài toán giải hệ phương trình tuyến tính:*

$$Hx = z. \tag{1.18}$$

*Bài toán* (1.18) *được gọi là đặt chỉnh (well-posed) nếu thỏa mãn tất cả các điều kiện sau:*

*i. Với mọi $z \in \mathbb{R}^p$ đều tồn tại nghiệm $x \in \mathbb{R}^n$.*

*ii. Nghiệm thu được là duy nhất.*

*iii. Bài toán là ổn định trên cặp không gian $(\mathbb{R}^n, \mathbb{R}^p)$, nghĩa là với mọi $\epsilon > 0$ đều tồn tại $\delta(\epsilon) > 0$ sao cho $\|z_1 - z_2\| \leq \delta(\epsilon)$ thì $\|x_1 - x_2\| \leq \epsilon$.*

*Ngược lại,* (1.18) *được gọi là bài toán không đặt chỉnh (ill-posed).*

Bài toán (1.18) thường là không đặt chỉnh. Để giải quyết điều kiện không tương thích của hệ $z = Hx$, cách đơn giản nhất là sử dụng phương pháp bình phương cực tiểu. Vector ước lượng $x$ là nghiệm bài toán tối ưu $\hat{x} = \mathrm{argmin}\, J(x)$ với $J(x) = \frac{1}{2}(z - Hx)^T(z - Hx)$. Khi đó, $Hx$ là chiếu trực giao của $z$ trên không gian $\mathrm{Span}(H)$. Nghiệm của bài toán tồn tại duy nhất khi và chỉ khi $\mathrm{rank}(H) = n$. Với ma trận $H$ bất kỳ, ta quan tâm tìm nghiệm của bài toán tối ưu có chuẩn $\|x\|$ nhỏ nhất. Khi đó, ước lượng $x$ có thể biểu diễn dưới dạng giả nghiệm thông qua giả nghịch đảo Moore-Penrose $H^+$ của ma trận $H$ [Alb72]:

$$\hat{x} = H^+ z. \tag{1.19}$$

Trong thực tế, tất cả các ma trận $H, M, N, V$ và vector quan sát $z$ chỉ được cho bởi các xấp xỉ $H^\delta, M^\delta, N^\delta, V^\delta, z^\delta$, với $\delta$ đặc trưng cho sai số của xấp xỉ. Khi đó, ta sẽ thu được ước lượng xấp xỉ $\hat{x}^\delta$ thay vì ước lượng đúng $\hat{x}$. Các nghiên cứu [Alb72, TA77] đã chỉ ra rằng, nếu hệ $z = Hx$ là không đặt chỉnh thì nghiệm $\hat{x}^\delta = H^{\delta,+} z^\delta$ có thể có sai số lớn tùy ý với một xấp xỉ $\delta$ đủ nhỏ. Để giải quyết bài toán không đặt chỉnh này, một thuật toán ổn định chính quy đã được đề xuất trong [HB13] nhằm tìm ước lượng $\hat{x}^\delta$ đủ gần $\hat{x}$ với xấp xỉ $\delta$ nhỏ.

Phương pháp chính quy hóa được cho bởi phương trình (1.19) thường không có nhiều ý nghĩa trong thực tế do không làm cực tiểu phương sai và bỏ qua sự tồn tại của nhiễu $v$. Khi $V$ là ma trận không suy biến, xét ước lượng được xác định bởi:

$$\hat{x} = \left[H^T V^{-1} H\right]^+ H^T V^{-1} z. \tag{1.20}$$

Khi đó, (1.20) là ước lượng cực tiểu phương sai trong tất cả các ước lượng tuyến tính không chệch của $x$ [HB13].

Cuối cùng, ta trình bày định nghĩa ước lượng tuyến tính tối ưu, được xây dựng dựa trên các kết quả về lý thuyết ma trận giả nghịch đảo của [Alb72].

**Định nghĩa 1.15** (Ước lượng tuyến tính tối ưu)**.** *[HB13] Xét mô hình hồi quy tuyến tính (1.1, 1.2, 1.3) với các số nguyên dương $n, p$ và các ma trận $H, V$ bất kỳ. Ước lượng tuyến tính tối ưu của vector tham số chưa biết $x$ được định nghĩa bởi:*

$$\hat{x} = Gz,\ G = H^+ \left\{ I - V_1 \left[ \left( I - HH^+ \right) V_1 \right]^+ \right\}, \tag{1.21}$$

*trong đó, ma trận $V = V_1 V_1^T$.*

Dễ dàng thấy rằng, khi ma trận hiệp phương sai $V$ có dạng đường chéo $V = \alpha I (\alpha \geq 0)$, ước lượng (1.21) có dạng như trong phương trình (1.19). Tương tự, ước lượng (1.20) có thể thu được từ (1.21) khi $V$ là ma trận không suy biến.

# Chương 2

# Phương pháp truy hồi cho ước lượng tham số

## 2.1 Mô hình hồi quy tuyến tính không nhiễu

Xét mô hình tuyến tính (1.1) với nhiễu $v = 0$. Ta có mô hình quan sát không nhiễu:

$$z = Hx, \tag{2.1}$$

với các giả thiết:

$$E(x) = \bar{x}, \quad E(ee^T) = M, \quad e := x - \bar{x}, \tag{2.2}$$

trong đó, $z \in \mathbb{R}^p$ là vector quan sát không nhiễu, $x \in \mathbb{R}^n$ là vector trạng thái chưa biết cần ước lượng, $H \in \mathbb{R}^{p\times n}$ là ma trận hệ số quan sát đã biết.

Đặt $z_i$ là thành phần thứ $i$ của vector quan sát $z$, $h_i$ là vector hàng thứ $i$ của ma trận $H$, với giả sử $h_i \neq 0$, $\forall i = 1, 2, \ldots, p$. Khi đó, $z = (z_1, z_2, \ldots, z_p)^T$ và $H = \left(h_1^T, h_2^T, \ldots, h_p^T\right)^T$.

Trong phần này, ta sẽ xây dựng một thuật toán truy hồi tìm ước lượng chính quy và cực tiểu sai số $\hat{x}$ cho hệ quan sát (2.1) khi kích thước $p$ của vector quan sát $z$ lớn.

### 2.1.1 Thuật toán truy hồi

Xét các ước lượng $x_i$ được cho bởi hệ phương trình truy hồi [HB14b]:

$$x_{i+1} = x_i + K_{i+1}\left[z_{i+1} - h_{i+1}x_i\right], \tag{2.3a}$$

$$K_{i+1} = M_i h_{i+1}^T \left[h_{i+1}M_i h_{i+1}^T\right]^+, \tag{2.3b}$$

$$M_{i+1} = M_i - K_{i+1}h_{i+1}M_i, \tag{2.3c}$$

$$i = 0, 1, \ldots, p-1,$$

trong đó, $x_0 = \bar{x}$ và $M_0$ là ma trận đối xứng nửa xác định dương cho trước.

Sau đây là các định lý và hệ quả về một số tính chất của hệ truy hồi (2.3).

**Định lý 2.1.** *Giả sử $M_0$ là ma trận đối xứng nửa xác định dương. Khi đó, mọi ma trận $M_j$, $j = 0, 1, \ldots, p$ đều là ma trận đối xứng nửa xác định dương.*

*Chứng minh.* Sử dụng phương pháp quy nạp.

Theo giả thiết, $M_0$ là ma trận đối xứng nửa xác định dương.

Giả sử mệnh đề đúng với $j = k$. Ta sẽ chứng minh mệnh đề cũng đúng với $j = k+1$.

i. *Trường hợp 1:* $h_{k+1}M_k h_{k+1}^T = 0$

Theo định nghĩa giả nghịch đảo Moore-Penrose của $0$, suy ra $M_{k+1} = M_k$.

ii. *Trường hợp 2:* $h_{k+1}M_k h_{k+1}^T \neq 0$

Ký hiệu:

$$C = \left[h_{k+1}M_k h_{k+1}^T\right]^+ = \left[h_{k+1}M_k h_{k+1}^T\right]^{-1} > 0.$$

Viết lại $M_{k+1}$ dưới dạng:

$$\begin{aligned} M_{k+1} &= M_k - M_k h_{k+1}^T \left[h_{k+1}M_k h_{k+1}^T\right]^+ h_{k+1}M_k \\ &= M_k - C\left(h_{k+1}M_k\right)^T h_{k+1}M_k. \end{aligned}$$

Do $M_k$ là ma trận đối xứng nên dễ thấy $M_{k+1}$ cũng là ma trận đối xứng.

Ta sẽ chứng minh $M_{k+1}$ là ma trận nửa xác định dương, nghĩa là:

$$u^T M_{k+1} u = u^T M_k u - C\left(h_{k+1} M_k u\right)^T h_{k+1} M_k u \geq 0, \quad \forall u \in \mathbb{R}^n.$$

Do $M_k$ là ma trận đối xứng nửa xác định dương nên theo tính chất (1.14), tồn tại ma trận $L$ sao cho $L^T L = M_k$. Khi đó, theo bất đẳng thức Cauchy-Schwarz, ta có:

$$\begin{aligned} C^{-1} u^T M_k u &= \left(h_{k+1} L^T L h_{k+1}^T\right)\left(u^T L^T L u\right) = \left\|L h_{k+1}^T\right\|^2 . \|Lu\|^2 \\ &\geq \left\|h_{k+1} L^T L u\right\|^2 = \|h_{k+1} M_k u\|^2 = \left(h_{k+1} M_k u\right)^T h_{k+1} M_k u. \end{aligned}$$

Do đó, $u^T M_k u \geq C\left(h_{k+1} M_k u\right)^T h_{k+1} M_k u$ hay $u^T M_{k+1} u \geq 0$.

Từ hai trường hợp trên ta có điều phải chứng minh. □

**Định lý 2.2.** *[HB14b] Giả sử tập các vector $\left\{h_1^T, h_2^T, \ldots, h_p^T\right\}$ độc lập tuyến tính, đồng thời hệ phương trình (2.1) là tương thích. Khi đó, với mỗi vector $\bar{x}$ hữu hạn và ma trận $M_0$ đối xứng xác định dương, ta có: $Hx_p = z$.*

Để chứng minh định lý 2.2, ta cần sử dụng các kết quả sau:

**Bổ đề 1.** *[HB14b] Với mọi ma trận $M_j$ được định nghĩa như trong (2.3c), ta có:*

$$h_i M_j = 0, \qquad \forall i = 1, 2, \ldots, j. \tag{2.4}$$

*Chứng minh.* Sử dụng phương pháp quy nạp.

Với mỗi $j = 1, 2, \ldots, p$, từ phương trình (2.3b) và (2.3c), ta có:

$$h_j M_j = h_j\left(M_{j-1} - K_j h_j M_{j-1}\right) = h_j M_{j-1} - h_j M_{j-1} h_j^T \left[h_j M_{j-1} h_j^T\right]^+ h_j M_{j-1}.$$

i. *Trường hợp 1:* $h_j M_{j-1} h_j^T \neq 0$

Theo định nghĩa giả nghịch đảo Moore-Penrose của một số, ta có:

$$h_j M_{j-1} h_j^T \left[h_j M_{j-1} h_j^T\right]^+ = 1 \quad \text{hay} \quad h_j M_j = h_j M_{j-1} - h_j M_{j-1} = 0.$$

ii. *Trường hợp 2:* $h_j M_{j-1} h_j^T = 0$

Theo định lý 2.1, tồn tại ma trận $L$ sao cho $M_{j-1} = L^T L$. Khi đó, ta có:

$$h_j M_{j-1} h_j^T = h_j L^T L h_j^T = \left\| L h_j^T \right\|^2 = 0 \Leftrightarrow L h_j^T = 0 \Rightarrow h_j M_{j-1} = 0 \Rightarrow h_j M_j = 0.$$

Vậy $h_j M_j = 0,\ \forall j = 1, 2, \ldots, p$. Do đó, phương trình (2.4) đúng với $j = 1$.

Giả sử (2.4) đúng với mọi $1 \leq j \leq l$. Ta chứng minh (2.4) đúng với $j = l+1$, nghĩa là:

$$h_i M_{l+1} = 0, \qquad \forall i = 1, 2, \ldots, l+1$$

Thật vậy, với mỗi $1 \leq i \leq l$, ta có:

$$h_i M_{l+1} = h_i \left( M_l - K_{l+1} h_{l+1} M_l \right) = h_i M_l - h_i M_l h_{l+1}^T \left[ h_{l+1} M_l h_{l+1}^T \right]^+ h_{l+1} M_l$$

Do $h_i M_l = 0$ nên $h_i M_{l+1} = 0,\ \forall i = 1, 2, \ldots, l$.

Kết hợp với $h_{l+1} M_{l+1} = 0$, ta có: $h_i M_{l+1} = 0,\ \forall i = 1, 2, \ldots, l+1$. □

**Hệ quả 2.2.1.** *[HB14b] Giả sử $M_0$ là ma trận đối xứng xác định dương và $l$ vector $\left\{ h_1^T, h_2^T \ldots, h_l^T \right\}$ độc lập tuyến tính. Khi đó* $\mathrm{rank}[M_l] = n - l$.

*Chứng minh.* Sử dụng phương pháp quy nạp

Với mỗi $j = 1, 2, \ldots, p$, xét ma trận:

$$\begin{aligned} \Delta M_j := M_{j-1} - M_j &= M_{j-1} h_j^T h_j M_{j-1} \left[ h_j M_{j-1} h_j^T \right]^+ \\ &= (h_j M_{j-1})^T h_j M_{j-1} \left[ h_j M_{j-1} h_j^T \right]^+ . \end{aligned}$$

Dễ dàng thấy rằng $\mathrm{rank}(\Delta M_j) \leq 1,\ \forall j = 1, 2, \ldots, p$, nghĩa là mọi phần tử của không gian $\mathrm{Span}(\Delta M_j)$ đều có thể biểu diễn thông qua 1 vector $b_j = M_{j-1} h_j^T$.

Với $l = 1$, từ bổ đề 1 ta có: $h_1 M_1 = 0$. Do $h_1 \neq 0$ nên $\dim(\mathrm{Span}(M_1)) \leq n - 1$.

Giả sử $\dim(\mathrm{Span}(M_1)) = n - n',\ n' = 2, 3, \ldots$, nghĩa là tồn tại $n - n'$ vector độc lập tuyến tính $\{a_1, a_2, \ldots, a_{n-n'}\}$ sao cho mọi phần tử của không gian $\mathrm{Span}(M_1)$ đều có thể biểu diễn thông qua $n - n'$ vector trên.

Do $M_0 = M_1 + \Delta M_1$ nên mọi phần tử của không gian $\mathrm{Span}(M_0)$ có thể biểu diễn thông qua tổ hợp tuyến tính của $n - n' + 1$ vector $\{a_1, a_2, \ldots, a_{n-n'}, b_1\}$. Điều này mâu thuẫn với giả thiết ma trận $M_0$ là không suy biến. Vậy, $\dim(\mathrm{Span}(M_1)) = n - 1$.

Giả sử hệ quả 2.2.1 đúng với mọi $l \geq 1$. Ta sẽ chứng minh hệ quả này cũng đúng với $l := l + 1$.

Thật vậy, theo bổ đề 1, ta có: $h_i M_{l+1} = 0, \forall i = 1, 2, \ldots, l+1$. Do $l+1$ vector $h_1^T, h_2^T, \ldots, h_{l+1}^T$ độc lập tuyến tính nên $\dim(\mathcal{N}(M_{l+1})) \geq l + 1$. Mặt khác, $\mathrm{Span}(M_{l+1})$ là phần bù trực giao của $\mathcal{N}(M_{l+1})$, do đó $\dim(\mathrm{Span}(M_{l+1})) \leq n - l - 1$.

Chứng minh tương tự trường hợp $l = 1$, suy ra giả thiết $\dim(\mathrm{Span}(M_{l+1})) < n - l - 1$ mâu thuẫn với $\dim(\mathrm{Span}(M_l)) = n - l$. Như vậy, $\dim(\mathrm{Span}(M_{l+1})) = n - l - 1$. □

**Hệ quả 2.2.2.** *[HB14b] Giả sử $h_{l+1}^T$ phụ thuộc tuyến tính vào tập $l$ vector $\left\{h_1^T, h_2^T, \ldots, h_l^T\right\}$. Khi đó $M_{l+1} = M_l$.*

*Chứng minh.* Viết lại $h_{l+1}$ dưới dạng:

$$h_{l+1} = \sum_{i=1}^{l} \alpha_i h_i \quad (\alpha_i \in \mathbb{R}).$$

Thay biểu thức này vào phương trình (2.3c), ta có:

$$M_{l+1} = M_l - K_{l+1}\left(\sum_{i=1}^{l} \alpha_i h_i\right) M_l = M_l - K_{l+1} \sum_{i=1}^{l} \alpha_i h_i M_l$$

Theo bổ đề 1, $h_i M_l = 0, \forall i = 1, 2, \ldots, l$. Vậy $M_{l+1} = M_l$. □

**Hệ quả 2.2.3.** *[HB14b] Giả sử $M_0$ là ma trận đối xứng xác định dương, $h_1^T, h_2^T, \ldots, h_l^T$ là các vector độc lập tuyến tính và $h_{l+1}^T$ phụ thuộc tuyến tính vào tập $l$ vector $\left\{h_1^T, h_2^T, \ldots, h_l^T\right\}$. Khi đó $\mathrm{rank}[M_{l+1}] = n - l$.*

*Chứng minh.* Từ hệ quả 2.2.2, ta có $M_{l+1} = M_l$. Do đó, theo hệ quả 2.2.1, $\mathrm{rank}[M_{l+1}] = n - l$. □

**Định lý 2.3.** *Giả sử $M_0$ là ma trận đối xứng xác định dương, đồng thời các vector $\left\{h_1^T, h_2^T, \ldots, h_{l+1}^T\right\}$ độc lập tuyến tính. Khi đó, ta có:*

$$h_{l+1}M_l \neq 0 \quad \textit{và} \quad h_{l+1}K_{l+1} = 1. \tag{2.5}$$

*Chứng minh.* Theo bổ đề 1 và hệ quả 2.2.1, ta có:

$$\mathcal{N}(M_l) = \text{Span}\left(h_1^T, h_2^T, \ldots, h_l^T\right).$$

Do giả thiết $\left\{h_1^T, h_2^T, \ldots, h_{l+1}^T\right\}$ độc lập tuyến tính nên $h_{l+1}^T \notin \text{Span}\left(h_1^T, h_2^T, \ldots, h_l^T\right)$ hay $h_{l+1}^T \notin \mathcal{N}(M_l)$. Do đó, $h_{l+1}M_l \neq 0$.

Theo định lý 2.1, tồn tại ma trận $L$ sao cho $LL^T = M_l$. Do $h_{l+1}M_l \neq 0$ nên $h_{l+1}L \neq 0$. Do đó:

$$h_{l+1}M_l h_{l+1}^T = h_{l+1}LL^T h_{l+1}^T \neq 0.$$

Từ định nghĩa giả nghịch đảo Moore-Penrose của một số, ta thu được:

$$h_{l+1}K_{l+1} = h_{l+1}M_l h_{l+1}^T \left[h_{l+1}M_l h_{l+1}^T\right]^+ = 1.$$

Ta có điều phải chứng minh. □

**Bổ đề 2.** *[HB14b] Giả sử $M_0$ là ma trận đối xứng xác định dương, đồng thời các vector $\left\{h_1^T, h_2^T, \ldots, h_p^T\right\}$ độc lập tuyến tính. Với mọi vector $x_j$ được định nghĩa như trong (2.3a), ta có:*

$$h_i x_j = z_i, \qquad \forall i = 1, 2, \ldots, j. \tag{2.6}$$

*Chứng minh.* Sử dụng phương pháp quy nạp.

Với mỗi $j = 1, 2, \ldots, p$, áp dụng định lý 2.3 ta có:

$$\begin{aligned} h_j x_j &= h_j\left(x_{j-1} + K_j\left[z_j - h_j x_{j-1}\right]\right) = h_j x_{j-1} + h_j K_j\left[z_j - h_j x_{j-1}\right] \\ &= h_j x_{j-1} + z_j - h_j x_{j-1} = z_j. \end{aligned}$$

Do đó, phương trình (2.6) đúng với $j = 1$.

Giả sử phương trình (2.6) đúng với mọi $1 \leq j \leq l$, nghĩa là:

$$h_i x_j = z_i, \qquad \forall i = 1, 2, \ldots, j, \quad \forall j = 1, 2, \ldots, l.$$

Ta sẽ chứng minh rằng (2.6) cũng đúng với $j = l + 1$, nghĩa là:

$$h_i x_{l+1} = z_i, \qquad \forall i = 1, 2, \ldots, l + 1.$$

Thật vậy, với mỗi $i = 1, 2, \ldots, l$, từ định nghĩa của $x_{l+1}$ ta có:

$$h_i x_{l+1} = h_i \left[x_l + K_{l+1} \left(z_{l+1} - h_{l+1} x_l\right)\right] = z_i + h_i K_{l+1} \left(z_{l+1} - h_{l+1} x_l\right).$$

Từ định nghĩa của $K_{l+1}$, áp dụng bổ đề 1, ta có:

$$h_i K_{l+1} = h_i M_l h_{l+1}^T \left[h_{l+1} M_l h_{l+1}^T\right]^+ = 0, \qquad \forall i \leq l.$$

Kết hợp với $h_{l+1} x_{l+1} = z_{l+1}$ ta có điều phải chứng minh. □

*Chứng minh định lý 2.2.* Áp dụng kết quả bổ đề 2 với $j = p$, ta có:

$$h_i x_p = z_i, \ \forall i = 1, 2, \ldots, p \quad \text{hay} \quad H x_p = z.$$

Do đó ta có điều phải chứng minh. □

Định lý 2.2 nói rằng, với $x_0 = \bar{x}$ và ma trận đối xứng xác định dương $M_0$ cho trước bất kỳ, thuật toán truy hồi (2.3) cho ta một nghiệm của hệ phương trình tuyến tính (2.1) sau $p$ bước lặp. Trong khi đó, các hệ quả 2.2.1, 2.2.2 và 2.2.3 cho phép kiểm tra quá trình tính toán thông qua hạng của các ma trận $M_i$. Nếu $H$ là ma trận vuông không suy biến kích thước $n \times n$ thì sau $n$ bước lặp ta sẽ thu được $M_n = 0$.

### 2.1.2 Tính chính quy của ước lượng

Trong phần này, ta sẽ kiểm tra xem liệu nghiệm của hệ truy hồi (2.3a, 2.3b, 2.3c) có phải là ước lượng tối ưu của (2.1) hay không?

Viết lại ước lượng tối ưu (1.21) dưới dạng [HB14b]:

$$\hat{x} = \bar{x} + M_0 H^T \left[ H M_0 H^T \right]^+ (z - H\bar{x}). \tag{2.7}$$

Đặt $M_0 = I$, ta thu được ước lượng

$$\hat{x} = \bar{x} + H^T \left[ H H^T \right]^+ (z - Hx) \in \text{Span}(H^T). \tag{2.8}$$

Khi số lượng quan sát $p$ rất nhỏ so với kích thước vector tham số cần ước lượng $n$:

$$\dim\left(\text{Span}(H^T)\right) \leq p \ll n = \dim\left(\mathbb{R}^n\right).$$

Do đó, phương pháp ước lượng bằng cách đặt $M_0 = I$ như trong [Alb72] không giải quyết được bài toán không đặt chỉnh (2.1) khi $p < n$.

**Định lý 2.4.** *[HB14b] Giả sử $\bar{x} \in \text{Span}(\overline{M})$ và $\text{Span}(M_0) \subseteq \text{Span}(\overline{M})$. Khi đó, tất cả các ước lượng $x_i\,(i = 1, 2, \ldots, p)$ đều thuộc không gian sinh bởi các cột của $\overline{M}$.*

*Chứng minh.* Sử dụng phương pháp quy nạp.

Viết lại ước lượng $x_1$ dưới dạng:

$$x_1 = \bar{x} + M_0 h_1^T \left( h_1 M_0 h_1^T \right)^+ \left[ z_1 - h_1 \bar{x} \right].$$

Do $\bar{x} \in \text{Span}(\overline{M})$ và $\text{Span}(M_0) \subseteq \text{Span}(\overline{M})$ nên ta có $x_1 \in \text{Span}(\overline{M})$.

Giả sử mệnh đề đúng với $i \geq 1$, nghĩa là $x_j \in \text{Span}(\overline{M}),\ \forall j \leq i$. Theo (2.3), ta có:

$$x_i = x_{i-1} + M_{i-1} h_i^T \left( h_i M_{i-1} h_i^T \right)^+ \left[ z_i - h_i x_{i-1} \right].$$

Do $x_{i-1} \in \text{Span}(\overline{M})$ nên $\text{Span}(M_{i-1}) \subseteq \text{Span}(\overline{M})$.

Ta sẽ chứng minh mệnh đề cũng đúng với $i := i + 1$. Thật vậy, từ định nghĩa của $M_i$, ta có:

$$M_i = M_{i-1} - M_{i-1}h_i^T \left(h_i M_{i-1} h_i^T\right)^+ h_i M_{i-1}$$

Do $\mathrm{Span}(M_{i-1}) \subseteq \mathrm{Span}(\overline{M})$ nên $\mathrm{Span}(M_i) \subseteq \mathrm{Span}(\overline{M})$. Kết hợp với $x_i \in \mathrm{Span}(\overline{M})$, ta được:

$$x_{i+1} = x_i + M_i h_{i+1}^T \left(h_{i+1} M_i h_{i+1}^T\right)^+ [z_i - h_{i+1}x_i] \in \mathrm{Span}(\overline{M}).$$

Do đó ta có điều phải chứng minh. □

Định lý 2.4 nói rằng tất cả các ước lượng $x_i$, $i = 1, 2, \ldots, p$ là hình chiếu của $x$ trên không gian $\mathrm{Span}(\overline{M})$. Kết quả này rất quan trọng khi số lượng quan sát $p$ rất nhỏ so với số lượng tham số cần ước lượng $n$. Việc lựa chọn ma trận $\overline{M}$ là bài toán quan trọng nhất nếu muốn thuật toán truy hồi (2.3) cho các ước lượng tốt. Do đó, định lý 2.4 cho ta một phương pháp chính quy hóa bài toán không đặt chỉnh (2.1) khi số lượng quan sát nhỏ hơn số lượng tham số cần ước lượng, bằng cách lựa chọn ma trận hiệp phương sai cho trước $M_0 = \overline{M}$.

### 2.1.3 Ước lượng cực tiểu sai số

Giả sử $x$ là vector ngẫu nhiên có kỳ vọng $\bar{x}$ và ma trận hiệp phương sai $\overline{M}$.

**Định lý 2.5.** *[HB14b] Các vector $x_i$ được xác định bởi hệ truy hồi (2.3) là ước lượng không chệch và cực tiểu sai số của $x$ trong lớp tất cả các ước lượng không chệch phụ thuộc tuyến tính vào $\bar{x}$ và $z_1, z_2, \ldots, z_i$.*

*Chứng minh.* Ký hiệu:

$$z^i := (z_1, z_2, \ldots, z_i)^T, \qquad H^i := \left(h_1^T, h_2^T, \ldots, h_i^T\right)^T. \tag{2.9}$$

Lớp tất cả các ước lượng $x_i'$ của $x$ phụ thuộc tuyến tính vào $\bar{x}$ và $z_1, z_2, \ldots, z_i$ có dạng:

$$x_i'(A, B) = A\bar{x} + Bz^i. \tag{2.10}$$

Lấy kỳ vọng của ước lượng $x_i'(A, B)$:

$$E\left[x_i'(A, B)\right] = E\left(A\bar{x} + Bz^i\right) = A\bar{x} + E\left(BH^i x\right) = \left(A + BH^i\right)\bar{x}.$$

Do tính không chệch của ước lượng $x_i'$ nên ta có:

$$E\left[x_i'(A, B)\right] = \left(A + BH^i\right)\bar{x} = \bar{x},\ \forall \bar{x} \in \mathbb{R}^n.$$

Do đó:

$$A + BH^i = I \quad \text{hay} \quad A = I - BH^i. \tag{2.11}$$

Thay biểu thức (2.11) vào phương trình (2.10), ta thu được ước lượng không chệch phụ thuộc vào một ma trận tham số $B$:

$$x_i'(B) = \bar{x} + B\left[z^i - H^i\bar{x}\right]. \tag{2.12}$$

Xét bài toán tối ưu cực tiểu sai số:

$$\min_{B \in \mathbb{R}^{n \times i}} J(B) = \mathrm{Tr}\left(E\left(ee^T\right)\right) = E\left[\|e\|^2\right], \qquad e := x - x_i'(B). \tag{2.13}$$

Từ phương trình của $x_i'(B)$, ta có:

$$\begin{aligned}
E\left(ee^T\right) &= E\left[x - \bar{x} - B\left(z^i - H^i\bar{x}\right)\right]\left[x - \bar{x} - B\left(z^i - H^i\bar{x}\right)\right]^T \\
&= E\left[\left(I - BH^i\right)(x - \bar{x})\right]\left[\left(I - BH^i\right)(x - \bar{x})\right]^T \\
&= E\left[\left(I - BH^i\right)(x - \bar{x})(x - \bar{x})^T\left(I - BH^i\right)^T\right] \\
&= \left(I - BH^i\right)E\left[(x - \bar{x})(x - \bar{x})^T\right]\left(I - BH^i\right)^T \\
&= \left(I - BH^i\right)\overline{M}\left(I - BH^i\right)^T \\
&= \overline{M} - BH^i\overline{M} - \overline{M}H^{i,T}B^T + BH^i\overline{M}H^{i,T}B^T.
\end{aligned}$$

Đạo hàm hàm mục tiêu $J(B) = \text{Tr}\left(E\left(ee^T\right)\right)$ theo ma trận $B$, ta thu được:

$$\begin{aligned}\frac{\partial}{\partial B}\text{Tr}\left(E\left(ee^T\right)\right) &= \frac{\partial}{\partial B}\text{Tr}\left(\overline{M} - BH^i\overline{M} - \overline{M}H^{i,T}B^T + BH^i\overline{M}H^{i,T}B^T\right)\\ &= -\overline{M}^T H^{i,T} - \overline{M}H^{i,T} + BH^i\overline{M}H^{i,T} + BH^i\overline{M}^T H^{i,T}.\end{aligned}$$

Do ma trận hiệp phương sai $\overline{M}$ đối xứng nên $\overline{M}^T = \overline{M}$. Khi đó:

$$\frac{\partial}{\partial B}\text{Tr}\left(E\left(ee^T\right)\right) = -2\overline{M}H^{i,T} + 2BH^i\overline{M}H^{i,T}.$$

Giải phương trình $\frac{\partial}{\partial B}\text{Tr}\left(E\left(ee^T\right)\right) = 0$, ta thu được nghiệm tối ưu:

$$B_0 = \overline{M}H^{i,T}\left[H^i\overline{M}H^{i,T}\right]^+. \tag{2.14}$$

Thay nghiệm (2.14) vào phương trình (2.12) ta thu được điều phải chứng minh. □

## 2.2 Mô hình hồi quy tuyến tính với nhiễu ngẫu nhiên

Xét mô hình hồi quy tuyến tính (1.1) với nhiễu ngẫu nhiên $v$. Ký hiệu $v^i = (v_1, v_2, \ldots, v_i)^T$, $z^i$ và $H^i$ được xác định như trong (2.9). Khi đó, (1.1) có thể viết dưới dạng:

$$z^i = H^i x + v^i.$$

Giả thiết chuỗi nhiễu ngẫu nhiên $\{v_j\}_{j=1,2,\ldots,i}$ không tương quan và có cùng phân phối xác suất, nghĩa là:

$$E\left[v^i\left(x - \bar{x}\right)^T\right] = 0, \quad E\left(v^i v^{i,T}\right) = \alpha I \quad (\alpha \in \mathbb{R}^+). \tag{2.15}$$

Ta tìm ước lượng không chệch cực tiểu sai số $\hat{x}$ của $x$ phụ thuộc vào một tham số $\alpha$ thể hiện phương sai của nhiễu quan sát.

Xét tất cả các ước lượng $x_i'$ của $x$ phụ thuộc tuyến tính vào $\bar{x}$ và $z_1, z_2, \ldots, z_i$:

$$x_i'(A, B) = A\bar{x} + Bz^i. \tag{2.16}$$

Lấy kỳ vọng của ước lượng $x_i'(A,B)$, kết hợp với điều kiện $E\left(v^i\right)=0$, ta được:

$$E\left[x_i'(A,B)\right]=E\left(A\bar{x}+Bz^i\right)=A\bar{x}+E\left(BH^ix+Bv^i\right)=\left(A+BH^i\right)\bar{x}.$$

Để $x_i'$ là ước lượng không chệch của $x$ thì

$$E\left[x_i'(A,B)\right]=\left(A+BH^i\right)\bar{x}=\bar{x},\ \forall\bar{x}\in\mathbb{R}^n.$$

hay $A=I-BH^i$.

Khi đó, ta thu được ước lượng không chệch $x_i'(B)$ của $x$ phụ thuộc tuyến tính vào $\bar{x}$ và $z_1,z_2,\ldots,z_i$ có dạng như (2.12).

Xét bài toán tối ưu cực tiểu sai số:

$$\min_{B\in\mathbb{R}^{n\times i}}J(B)=\mathrm{Tr}\left(E\left(ee^T\right)\right),\qquad e:=x-x_i'(B)$$

với ma trận hiệp phương sai:

$$\begin{aligned}E\left(ee^T\right)&=E\left[x-\bar{x}-B\left(z^i-H^i\bar{x}\right)\right]\left[x-\bar{x}-B\left(z^i-H^i\bar{x}\right)\right]^T\\&=E\left[\left(I-BH^i\right)(x-\bar{x})-Bv^i\right]\left[\left(I-BH^i\right)(x-\bar{x})-Bv^i\right]^T\\&=E\left[\left(I-BH^i\right)(x-\bar{x})(x-\bar{x})^T\left(I-BH^i\right)^T\right]+E\left[Bv^iv^{i,T}B^T\right]\\&\quad-E\left[\left(I-BH^i\right)(x-\bar{x})v^{i,T}B^T\right]-E\left[Bv^i(x-\bar{x})^T\left(I-BH^i\right)^T\right].\end{aligned}$$

Kết hợp giả thiết (2.15) của nhiễu ngẫu nhiên, ta có:

$$E\left(ee^T\right)=\left(I-BH^i\right)\overline{M}\left(I-BH^i\right)^T+\alpha BB^T.$$

Đạo hàm hàm mục tiêu $J(B)=\mathrm{Tr}\left(E\left(ee^T\right)\right)$ theo ma trận $B$, ta thu được:

$$\begin{aligned}\frac{\partial}{\partial B}\mathrm{Tr}\left(E\left(ee^T\right)\right)&=\frac{\partial}{\partial B}\mathrm{Tr}\left(\left(I-BH^i\right)\overline{M}\left(I-BH^i\right)^T+\alpha BB^T\right)\\&=-2\overline{M}H^{i,T}+2BH^i\overline{M}H^{i,T}+2\alpha B.\end{aligned}$$

Giải phương trình $\frac{\partial}{\partial B}\operatorname{Tr}\left(E\left(ee^T\right)\right)=0$, ta thu được nghiệm tối ưu:

$$B_0(\alpha)=\overline{M}H^{i,T}\left[H^i\overline{M}H^{i,T}+\alpha I\right]^{-1}. \tag{2.17}$$

Do $\overline{M}$ là ma trận đối xứng nửa xác định dương nên theo tính chất 1.14, tồn tại ma trận $L$ sao cho $LL^T=\overline{M}$. Đặt $A=H^iL$, ta có:

$$\lim_{\alpha\to 0^+}B_0(\alpha)=L\lim_{\delta\to 0}A^T\left(AA^T+\delta^2 I\right)^{-1}=LA^+=LA^T\left[AA^T\right]^+=B_0. \tag{2.18}$$

Thay $B_0(\alpha)$ vào phương trình (2.12) ta thu được ước lượng không chệch cực tiểu sai số của $x$:

$$x_i=\bar{x}+\overline{M}H^{i,T}\left[H^i\overline{M}H^{i,T}+\alpha I\right]^{-1}\left[z^i-H^i\bar{x}\right]. \tag{2.19}$$

Áp dụng bổ đề 1 trong [HB11a], ta có thuật toán truy hồi cho bài toán ước lượng tham số trong mô hình hồi quy tuyến tính với nhiễu ngẫu nhiên không tương quan cùng phân phối:

$$x_{i+1}(\alpha)=x_i(\alpha)+K_{i+1}(\alpha)\left[z_{i+1}-h_{i+1}x_i(\alpha)\right], \tag{2.20a}$$

$$K_{i+1}(\alpha)=M_i(\alpha)h_{i+1}^T\left[h_{i+1}M_i(\alpha)h_{i+1}^T+\alpha\right]^{-1}, \tag{2.20b}$$

$$M_{i+1}(\alpha)=M_i(\alpha)-K_{i+1}(\alpha)h_{i+1}M_i(\alpha), \tag{2.20c}$$

$$i=0,1,\ldots,p-1,$$

trong đó, $\alpha$ là tham số đặc trưng cho phương sai của nhiễu ngẫu nhiên, $x_0(\alpha)=\bar{x}, M_0(\alpha)=\overline{M}$ cho trước.

Như vậy, hệ phương trình (2.20) xác định các ước lượng không chệch cực tiểu sai số của $x$, phụ thuộc vào một tham số $\alpha$ thể hiện phương sai của nhiễu quan sát $v$. Dễ dàng thấy rằng, khi $\alpha\to 0$ thì $x_i(\alpha)\to x_i$ được định nghĩa như trong hệ truy hồi (2.3).

## 2.3 Mô hình hồi quy tuyến tính với số chiều lớn

Trong thực tế, nhiều bài toán dẫn đến hệ quan sát mà vector trạng thái có kích thước lớn, đặc biệt trong khai phá dữ liệu và dữ liệu lớn. Trong lĩnh vực khí tượng và hải dương học, các vector tham số cần ước lượng $x$ thường có kích thước cỡ $10^6 - 10^7$ [Dal91]. Điều này xảy ra do $x$ là tập hợp biến trạng thái xác định trên các lưới ba chiều. Khi đó, thuật toán (2.3) dẫn đến bài toán tính toán ma trận với $10^{12} - 10^{14}$ phần tử. Nếu phương pháp truy hồi (2.3) cho phép giải quyết bài toán khi số chiều của vector quan sát $z$ lớn thì trong phần này, ta sẽ xây dựng một thuật toán nhằm giảm khối lượng tính toán khi vector tham số cần ước lượng có kích thước rất lớn.

### 2.3.1 Các kết quả lý thuyết

Xét ma trận hiệp phương sai $M$ có $n$ trị riêng $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$ và $n$ vector riêng tương ứng $y_1, y_2, \ldots, y_n$. Khi đó, ta có phép phân tích phổ của $M$:

$$M = UDU^T, \tag{2.21}$$

trong đó, $D = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ và $U = (y_1, y_2, \ldots, y_n)$.

Với $m \leq n$, đặt $D(m) = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m) \in \mathbb{R}^{m\times m}, U(m) = (y_1, y_2, \ldots, y_m)$ và $M_0 = U(m)D(m)U^T(m)$. Khi đó, hệ truy hồi (2.3) cho ta ước lượng cực tiểu sai số của $x$ trên không gian con $\mathrm{Span}(U(m))$.

**Định lý 2.6.** *[HB14b] Cho $p \leq m_1 \leq m_2 \leq n$. Xét thuật toán truy hồi (2.3) với hai ma trận:*

$$M_0^i = U(m_i)D(m_i)U^T(m_i), \qquad i = 1, 2.$$

*Khi đó ta có bất đẳng thức:*

$$E\left[\|e_2\|^2\right] \leq E\left[\|e_1\|^2\right], \tag{2.22}$$

*trong đó,* $e_i = x - \hat{x}(m_i)$, $\hat{x}(m_i)$ *là ước lượng thu được bởi thuật toán* (2.3) *ứng với ma trận hiệp phương sai* $M_0 = M_0^i$, $i = 1, 2$.

*Chứng minh.* Ký hiệu các không gian tuyến tính:

$$\mathcal{L}_1 := \text{Span}\left(U(m_1)\right),$$
$$\mathcal{L}_2 := \text{Span}\left(U(m_2)\right) - \text{Span}\left(U(m_1)\right),$$
$$\mathcal{L}_3 := \text{Span}^{\perp}\left(U(m_2)\right).$$

Xét phép phân tích trực giao của vector $x$:

$$x = x_1 + x_2 + \tilde{x}, \tag{2.23}$$

trong đó, $x_1 \in \mathcal{L}_1, x_2 \in \mathcal{L}_2$ và $\tilde{x} \in \mathcal{L}_3$.

Theo định lý 2.5, $\hat{x}(m_1), \hat{x}(m_2)$ lần lượt là các ước lượng cực tiểu sai số của $x$ trên không gian $\mathcal{L}_1$ và $\text{Span}\left(U(m_2)\right)$. Do $\mathcal{L}_1, \mathcal{L}_2$ và $\mathcal{L}_3$ là các không gian trực giao nên $\hat{x}(m_2)$ có thể được phân tích dưới dạng:

$$\hat{x}(m_2) = \hat{x}(m_1) + \hat{x}_r(m_2)$$

trong đó, $\hat{x}_r(m_2) \in \mathcal{L}_2$, là ước lượng cực tiểu sai số của $x$ trên không gian $\mathcal{L}_2$.

Thay vào công thức sai số ước lượng, ta được:

$$e_1 = x - \hat{x}(m_1) = (x_1 - \hat{x}(m_1)) + x_2 + \tilde{x},$$
$$e_2 = x - \hat{x}(m_2) = (x_1 - \hat{x}(m_1)) + (x_2 - \hat{x}_r(m_2)) + \tilde{x}.$$

Do đó:

$$E\left[\|e_1\|^2\right] = E\left[\|x_1 - \hat{x}(m_1)\|^2\right] + E\left[\|x_2\|^2\right] + E\left[\|\tilde{x}\|^2\right],$$
$$E\left[\|e_2\|^2\right] = E\left[\|x_1 - \hat{x}(m_1)\|^2\right] + E\left[\|x_2 - \hat{x}_r(m_2)\|^2\right] + E\left[\|\tilde{x}\|^2\right].$$

Do $\hat{x}_{2r}$ là ước lượng cực tiểu sai số của $x$ trên không gian $\mathcal{L}_2$ nên

$$E\left[\|x_2 - \hat{x}_{2r}\|^2\right] \leq E\left[\|x_2\|^2\right] \quad \text{hay} \quad E\left[\|e_2\|^2\right] \leq E\left[\|e_1\|^2\right].$$

□

### 2.3.2 Thuật toán truy hồi giản lược

Đặt $x_e = U^T(m)x \in \mathbb{R}^m$ và $H_e = HU(m) \in \mathbb{R}^{p\times m}\,(m \le n)$. Khi đó, từ (2.1) ta có mô hình quan sát tuyến tính không nhiễu giản lược:

$$z = H_e x_e, \tag{2.24}$$

trong đó:

$$E(x_e) = E\left[U^T(m)x\right] = U^T(m)\bar{x}, \quad e := x_e - U^T(m)\bar{x},$$
$$E(ee^T) = E\left[x_e - U^T(m)\bar{x}\right]\left[x_e - U^T(m)\bar{x}\right]^T = U^T(m)M(m)U(m) = D(m).$$

Khi đó, thuật toán truy hồi (2.3) có thể viết lại dưới dạng giản lược:

$$x(i+1) = U(m)x_e(i+1), \tag{2.25a}$$
$$x_e(i+1) = x_e(i) + K_e(i+1)\left[z_{i+1} - h_e(i+1)x_e(i)\right], \tag{2.25b}$$
$$K_e(i+1) = M_e(i)h_e^T(i+1)\left[h_e(i+1)M_e(i)h_e^T(i+1)\right]^+, \tag{2.25c}$$
$$M_e(i+1) = M_e(i) - K_e(i+1)h_e(i+1)M_e(i), \tag{2.25d}$$
$$h_e(i+1) = h(i+1)U(m),\ M_e(0) = D(m),\ x_e(0) = U^T(m)\bar{x}, \tag{2.25e}$$
$$i = 0, 1, \ldots, p-1.$$

Thuật toán (2.25) tính toán trên hệ quan sát kích thước $m \times p$ thay vì $n \times p$ như trong (2.3). Do đó, (2.25) cho các ước lượng $x_e(i)$ thuộc không gian tuyến tính $\mathbb{R}^m$, đồng thời khối lượng tính toán của hệ giảm $\left(\frac{n}{m}\right)^2$ lần. Trong thực tế, đặc biệt trong các bài toán khai phá dữ liệu, thuật toán này có ý nghĩa rất lớn. Khi số chiều của vector trạng thái quá lớn, ta chỉ cần quan tâm đến các hướng mà theo đó độ biến thiên của tập dữ liệu là lớn nhất. Nói cách khác, các vector riêng ứng với các trị riêng lớn nhất của $M$ thể hiện các hướng mà theo đó sai số ước lượng của thuật toán tăng nhanh nhất. Do đó, theo thuật toán truy hồi giản lược (2.25), bằng cách chọn giá trị $m < n$, ta có thể giảm đáng kể khối lượng tính toán nhưng vẫn đảm bảo mức độ sai số của ước lượng cho phép.

# Chương 3

# Các ví dụ mô phỏng số

Trong chương này, hai ví dụ mô phỏng số cho ước lượng vector tham số chưa biết với kích thước $n$ lần lượt là 3 và 100 chiều sẽ được thực hiện nhằm minh họa cho thuật toán đã trình bày. Phương pháp mô phỏng Monte-Carlo được sử dụng để sinh mẫu ngẫu nhiên tuân theo phân phối xác suất chuẩn nhiều chiều $\mathcal{N}(\mu, \Sigma)$. Chi tiết thuật toán Monte-Carlo có thể tham khảo tại [Fis96]. Do giới hạn về thời gian và năng lực tính toán, các mô phỏng thực hiện trong phạm vi đồ án chỉ nhằm kiểm tra các kết quả lý thuyết của thuật toán. Ứng dụng thực tế của phương pháp truy hồi trong khai phá dữ liệu khí tượng, hải dương học với kích thước hệ quan sát lớn ($p \approx 10^4 - 10^5, n \approx 10^6 - 10^7$) có thể tham khảo tại [HB14b]. Các mô phỏng số được tiến hành với phần mềm MATLAB 2013a trên máy tính cá nhân chạy hệ điều hành Windows 10 Pro 64 bit, bộ vi xử lý Inter(R) Core(TM) i3-3120M CPU @ 2.50GHz, RAM 4.00GB. Chi tiết mã nguồn cài đặt của thuật toán có thể tham khảo tại `https://github.com/lvchien/thesis`.

## 3.1 Mô hình quan sát không nhiễu ba chiều

Xét hệ quan sát không nhiễu (2.1) với giả thiết vector kỳ vọng:

$$\bar{x} = (0, 0, 0)^T$$

và ma trận hiệp phương sai:

$$M = \begin{bmatrix} 0.70365 & 0.22044 & -0.14413 \\ 0.22044 & 0.17497 & -0.27906 \\ -0.14413 & -0.27906 & 1.55398 \end{bmatrix}.$$

Giả sử thành phần thứ nhất và thứ ba của vector trạng thái $x$ là quan sát được. Khi đó ta có mô hình hồi quy tuyến tính không nhiễu:

$$z = Hx, \quad H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

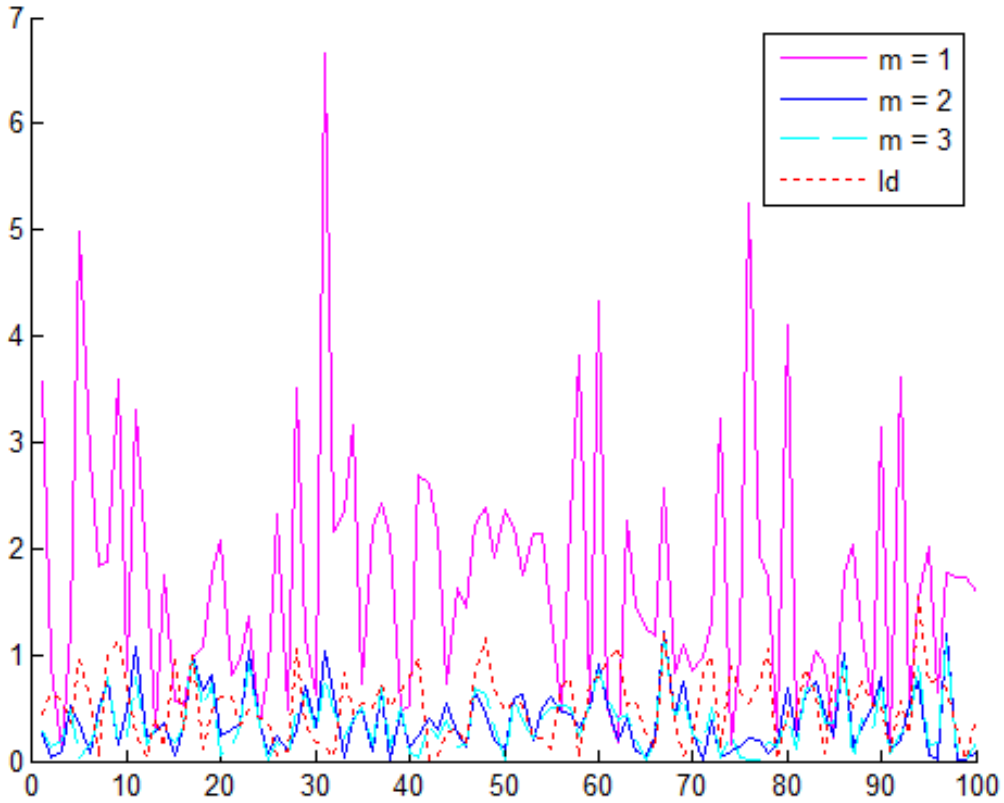
Sử dụng phân tích phổ cho ma trận hiệp phương sai $M = UDU^T$, trong đó:

$$U = [u_1, u_2, u_3] = \begin{bmatrix} -0.19606 & 0.936205 & 0.29169 \\ -0.21118 & 0.25017 & -0.94489 \\ 0.95758 & 0.24685 & -0.14866 \end{bmatrix},$$

$$D = \text{diag}(\lambda_1, \lambda_2, \lambda_3) = \text{diag}(1.64503, 0.72455, 0.06302).$$

Áp dụng thuật toán truy hồi giản lược (2.25) với bốn ma trận hiệp phương sai $M_0 = M(m) = U(m)D(m)U^T(m)$, $m = 1, 2, 3$ và $M_0 = I$. Các kết quả lần lượt được ký hiệu là $ALG(1), ALG(2), ALG(3)$ và $ALG(0)$.

Thuật toán Monte-Carlo được sử dụng để sinh 100 mẫu ngẫu nhiên tuân theo phân phối xác suất chuẩn $\mathcal{N}(\bar{x}, M)$ mô phỏng giá trị đúng của vector $x$. Sai số trung bình phương (*root mean square - rms*) của các ước lượng được minh họa như trong hình 3.1 và bảng 3.1.




Hình 3.1: Sai số trung bình phương của hệ quan sát không nhiễu 3 chiều

Có thể thấy từ hình 3.1 và bảng 3.1, lỗi ước lượng lớn nhất trong $ALG(1)$ và $ALG(0)$, trong khi đó, các thuật toán $ALG(2)$ và $ALG(3)$ không có sự khác nhau nhiều về sai số ước lượng. Các kết quả của $ALG(1)$ và $ALG(0)$ thể hiện rằng, xuất phát từ các thông tin đã biết không đủ tốt sẽ dẫn đến các ước lượng của hệ quan sát kém hiệu quả. Trong ví dụ này, vector riêng $u_2$ cho phép lưu trữ các "thông tin" quan trọng chứa trong quan sát thứ nhất $z_1$. Bỏ qua $u_2$ (tương đương với bỏ qua quan sát $z_1$) sẽ dẫn đến sai số ước lượng lớn, như trong thuật toán $ALG(1)$. Mặt khác, vector riêng $u_3$ tương ứng với giá trị riêng $\lambda_3$ nhỏ có ảnh hưởng yếu đối với tham số cần ước lượng $x$. Điều này giải thích tại sao $ALG(2)$ và $ALG(3)$ cho các kết quả gần giống nhau. Thuật toán $ALG(0)$ tương ứng với

Bảng 3.1: Kỳ vọng mẫu sai số trung bình phương của các ước lượng tham số 3 chiều.

| m = 1 | m = 2 | m = 3 | Id |
|---|---|---|---|
| 1.71385 | 0.38376 | 0.35892 | 0.51929 |

$M_0 = I$ cho ta một ước lượng thuộc không gian $\mathrm{Span}(H^T)$ có $\dim\left(\mathrm{Span}(H^T)\right) = 2$. Do đó, $ALG(0)$ có sai số nằm giữa $ALG(1)$ và $ALG(2)$. Mô phỏng này đã xác nhận lại các kết quả lý thuyết, đồng thời chứng minh rằng thuật toán truy hồi có thể cải thiện chất lượng của ước lượng thu được thông qua ma trận chính quy hóa $M$.

## 3.2 Mô hình quan sát không nhiễu nhiều chiều

Xét hệ quan sát không nhiễu (2.1) với kích thước $n = 100, p = 10$. Giả thiết kỳ vọng của vector tham số cần ước lượng:

$$\bar{x} = (0, 0, \ldots, 0)^T \in \mathbb{R}^{100}.$$

Ma trận $M$ được sinh ngẫu nhiên bởi thuật toán sinh ma trận hiệp phương sai (xem thêm chi tiết tại phụ lục A).

Giả sử các thành phần thứ $1, 10, 20, 25, 40, 50, 75, 80, 90, 100$ của vector trạng thái $x$ là quan sát được. Khi đó ta cần ước lượng vector tham số $x \in \mathbb{R}^{100}$ trong mô hình quan sát tuyến tính:

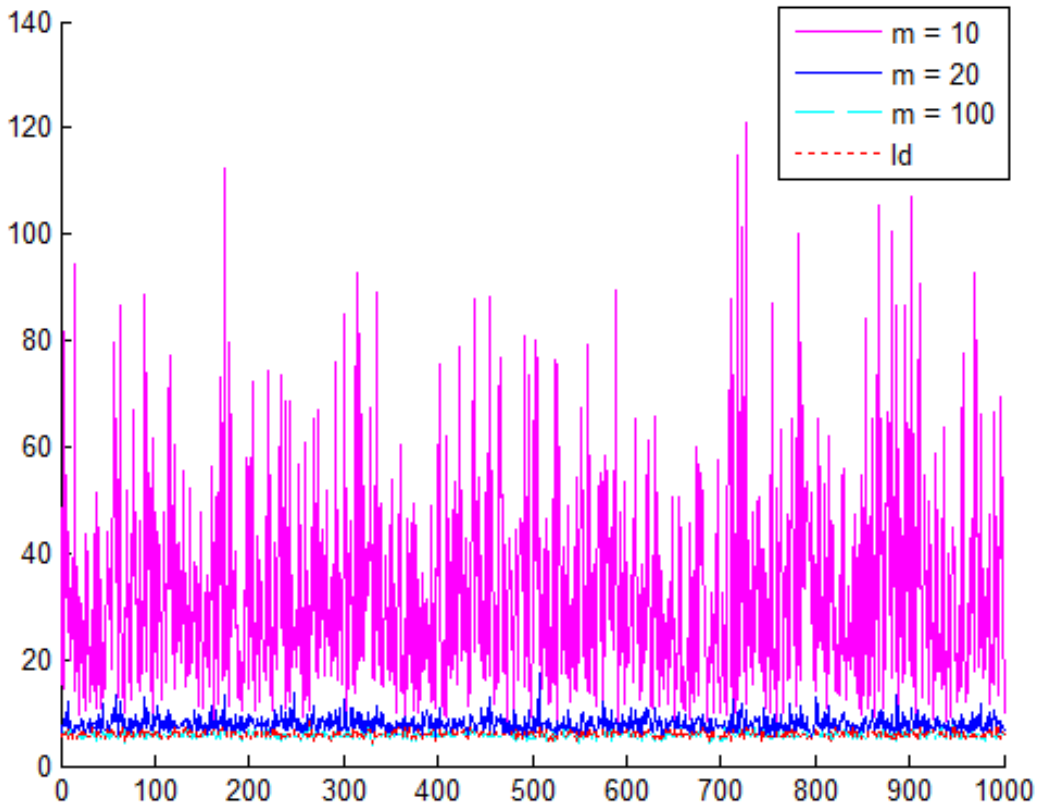
$$z = Hx,$$

trong đó $H \in \mathbb{R}^{10\times 100}$ là ma trận thưa, chỉ có giá trị bằng 1 tại 10 phần tử.

Thuật toán Monte-Carlo được sử dụng để sinh 1000 mẫu ngẫu nhiên mô phỏng giá trị đúng của $x$ theo phân phối xác suất chuẩn $\mathcal{N}(\bar{x}, M)$. Xét thuật toán (2.25) với các giá trị khác nhau của $m$ và với ma trận $M_0 = I$. Sai số trung bình phương của một số ước lượng được minh họa bởi hình 3.2 và bảng 3.2

Bảng 3.2: Kỳ vọng mẫu sai số trung bình phương của một số ước lượng tham số nhiều chiều.

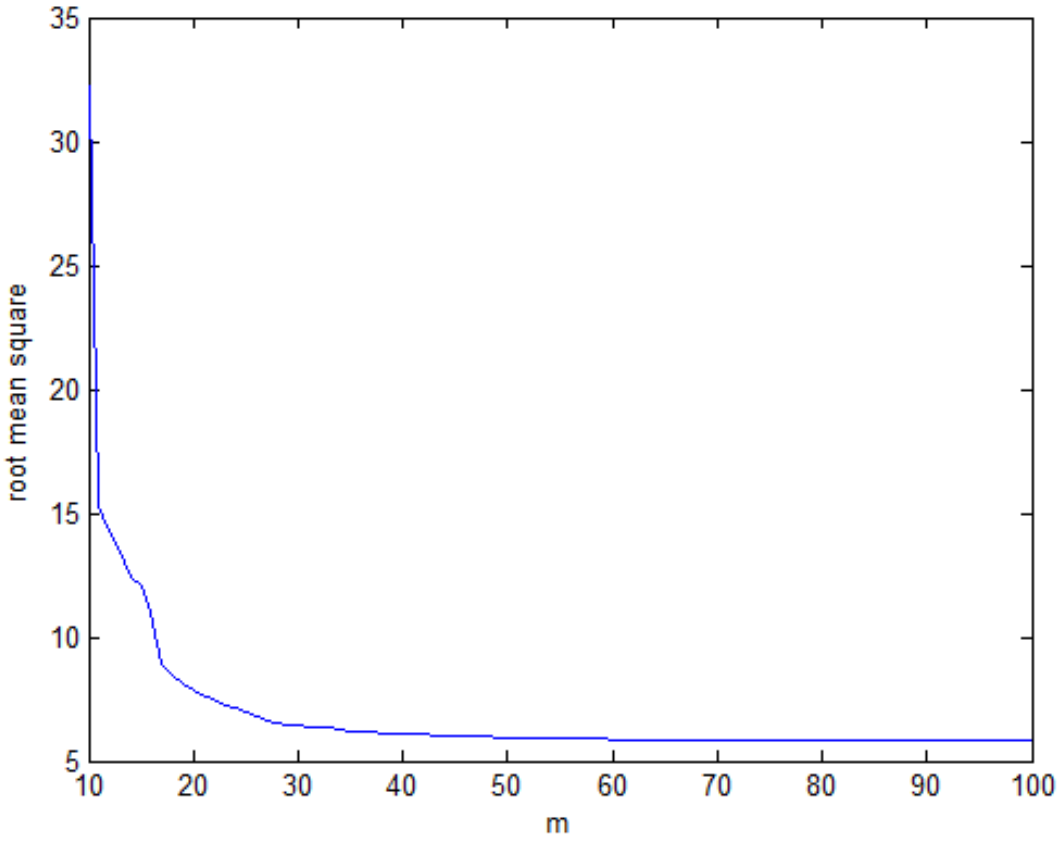
| m = 10 | m = 20 | m = 30 | m = 40 | m = 50 | m = 100 | Id |
|---|---|---|---|---|---|---|
| 32.2405 | 7.8102 | 6.4086 | 6.0921 | 5.9540 | 5.8036 | 5.9394 |




Hình 3.2: Sai số trung bình phương của hệ quan sát không nhiễu nhiều chiều

Hình 3.3 thể hiện sự phụ thuộc của sai số ước lượng vào giá trị của $m$. Sai số trung bình phương của các ước lượng giảm nhanh khi giá trị $m$ tăng và hội tụ đến sai số nhỏ nhất ứng với $m = 100$. Kết quả này hoàn toàn phù hợp với kết quả lý thuyết được phát biểu trong định lý 2.6, đồng thời thể hiện tính hiệu quả của thuật toán (2.25).

Trong ví dụ này, bằng cách chọn $m = 30$, ta có thể giảm khối lượng tính toán của hệ xuống khoảng $9$ lần mà vẫn đảm bảo sai số của ước lượng đủ nhỏ. Như vậy, ví dụ mô phỏng này đã xác nhận ưu điểm của phương pháp truy hồi giản lược. Bằng cách chọn giá trị $m < n$ thích hợp, thuật toán (2.25) cho phép giảm đáng kể khối lượng tính toán của bài toán, đồng thời vẫn đảm bảo sai số cho phép của ước lượng.




Hình 3.3: Sự phụ thuộc của sai số ước lượng vào giá trị của $m$.

# Kết luận

Trong nội dung đồ án này, em đã giới thiệu một phương pháp truy hồi cho ước lượng tham số trong mô hình hồi quy tuyến tính với cấu trúc ma trận hiệp phương sai không âm bất kỳ. Mục tiêu chính của đồ án là xây dựng một thuật toán hiệu quả nhằm giải quyết các khó khăn tính toán cho ước lượng thống kê trong các hệ quan sát có kích thước lớn của cả vector quan sát cũng như vector tham số cần ước lượng. Các tính chất tối ưu của ước lượng thu được được chứng minh chặt chẽ về mặt lý thuyết. Nội dung đồ án cũng chỉ ra rằng, bằng việc lựa chọn ma trận hiệp phương sai ban đầu, thuật toán truy hồi cho phép chính quy hóa các ước lượng thu được từ các hệ quan sát tuyến tính không đặt chỉnh.

Tính hiệu quả của thuật toán truy hồi được thể hiện qua các ví dụ mô phỏng số, với hệ quan sát không nhiễu ba chiều và nhiều chiều. Trong giới hạn thời gian và năng lực tính toán của đồ án, các ví dụ mô phỏng chỉ nhằm minh họa cho các kết quả lý thuyết. Ứng dụng thực tế của thuật toán cho hệ quan sát trong khai phá dữ liệu khí tượng và hải dương học với vector quan sát và vector tham số có kích thước rất lớn (cỡ $10^4 - 10^7$) có thể tham khảo tại [HB14b]. Các nghiên cứu trong tương lai nên tập trung kết hợp thuật toán truy hồi đã xây dựng với các mô hình dữ liệu và chuỗi thời gian khác nhau để giải quyết các bài toán thống kê trong thực tế, đặc biệt trong các ngành khoa học dữ liệu, khai phá dữ liệu cũng như dữ liệu lớn.

# Tài liệu tham khảo


[Alb72] A. Albert. *Regression and the Moore-Penrose Pseudo-Inverse*. Academy Press, New York, 1972.

[CH96] M. Cooper and K. Haines. Altimetric assimilation with water property conservation. *Journal of Geophysical Research*, 101:1059–1077, 1996.

[Dal91] R. Daley. *Atmospheric Data Analysis*. Cambridge University Press, New York, 1991.

[Fis96] George Fishman. *Monte Carlo: concepts, algorithms, and applications*. Springer-Verlag, New York, 1996.

[Fra00] Joel N. Franklin. *Matrix theory*. Dover publications, New York, dover edition, 2000.

[GL96] G.H. Golub and C.F.V. Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 3rd edition, 1996.

[HB11a] H.S. Hoang and R. Baraille. Approximate approach to linear filtering problem with correlated noise. *Engineering and Technology*, 5:11–23, 2011.

[HB11b] H.S. Hoang and R. Baraille. Prediction error sampling procedure based on dominant schur decomposition. application to state estimation in high dimen-

sional oceanic model. *Applied Mathematics and Computation*, 218:3689–3709, 2011.

[HB13] H.S. Hoang and R. Baraille. A regularized estimator for linear regression model with possibly singular co-variance. *IEEE Transactions on Automatic Control*, 58:236–241, 2013.

[HB14a] H.S. Hoang and R. Baraille. A low cost filter design for state and parameter estimation in very high dimensional systems. In *Proceedings of the 19th IFAC Congress*, pages 3156–3161, Cape Town, 24-29 August 2014.

[HB14b] H.S. Hoang and R. Baraille. Some properties of a recursive procedure for high dimensional parameter estimation in linear model with regularization. *Open Journal of Statistics*, 4(11):921–932, 2014.

[HBT01] H.S. Hoang, R. Baraille, and O. Talagrand. On the design of a stable adaptive filter for high dimensional systems. *Automatica*, 37:341–359, 2001.

[Kal60] R.E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:35–45, 1960.

[MN07] J.R. Magnus and H. Neudecker. *Matrix Differential Calculus with Application in Statistics and Econometrics*. Johns Wiley & Sons, 3rd edition, 2007.

[SL03] G.A.F. Seber and A.J. Lee. *Linear Regression Analysis*. Johns Wiley & Sons, 2nd edition, 2003.

[Spa00] J.C. Spall. Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE Transactions on Automatic Control*, 45:1839–1853, 2000.

[TA77] A. N. Tychonoff and V. Y. Arsenin. *Solution of Ill-Posed Problems.* DC: Winston and Sons, Washington, 1977.

# Phụ lục A

# Thuật toán sinh ngẫu nhiên ma trận hiệp phương sai

## A.1 Kết quả lý thuyết

**Định lý A.1.** *Cho ma trận đối xứng nửa xác định dương $M = (m_{ij}) \in \mathbb{R}^{n\times n}$ bất kỳ. Khi đó, tồn tại một vector ngẫu nhiên $x \in \mathbb{R}^n$ nhận $M$ là ma trận hiệp phương sai.*

*Chứng minh.* Do $M$ là ma trận đối xứng nửa xác định dương nên ta có phân tích phổ:

$$M = UDU^T, \tag{A.1}$$

trong đó, $D = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n), U = (a_1, a_2, \ldots, a_n)$ với $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$ là các trị riêng của $M$ và $a_1, a_2, \ldots, a_n$ là các vector riêng tương ứng.

Đầu tiên, ta chứng minh $m_{ii} \geq 0, \forall i = 1, 2, \ldots, n$. Thật vậy, từ công thức (A.1) ta có:

$$m_{ii} = \sum_{k=1}^{n} \lambda_k a_{ik}^2 \geq 0, \tag{A.2}$$

trong đó, $a_i = (a_{i1}, a_{i2}, \ldots, a_{in})^T$.

Tiếp theo, ta sẽ chứng minh $m_{ij}^2 \leq m_{ii}m_{jj}$. Thật vậy, ta có:

$$m_{ij} = \sum_{k=1}^{n} \lambda_k a_{ik} a_{kj}.$$

Do đó, áp dụng bất đẳng thức Cauchy-Schwarts, ta có:

$$m_{ij}^2 = \left(\sum_{k=1}^{n} \lambda_k a_{ik} a_{kj}\right)^2 \leq \left(\sum_{k=1}^{n} \lambda_k a_{ik}^2\right)\left(\sum_{k=1}^{n} \lambda_k a_{kj}^2\right) = m_{ii}m_{jj}. \tag{A.3}$$

Kết hợp (A.2) và (A.3) ta có điều phải chứng minh. □

## A.2 Thuật toán sinh ma trận hiệp phương sai

Từ kết quả lý thuyết trong phần A.1, ta xây dựng thuật toán sinh ngẫu nhiên ma trận hiệp phương sai dựa trên tính chất của ma trận đối xứng nửa xác định dương. Chi tiết mã nguồn thuật toán có thể tham khảo tại https://github.com/lvchien/thesis.

**Thuật toán 1:** Thuật toán sinh ngẫu nhiên ma trận hiệp phương sai

B1. Sinh ngẫu nhiên $A \in \mathbb{R}^{n\times n}$;

B2. $B := A + A^T$;

B3. Phân tích phổ $B = UDU^T$;

B4. Sinh ngẫu nhiên $u \in \mathbb{R}^n$;

B5. $v := \text{abs}(u)$;

B6. $D' := \text{diag}(v)$;

B6. $M := UD'U^T$;

B7. Trả về $M$;