Chọn mô hình tốt nhất trong thống kê Bayes mờ và ứng dụng trong phân tích tài chính

- **Phạm Hoàng Uyên**
- **Lê Thanh Hoa**
- **Nguyễn Đình Thiên**

Trường Đại học Kinh tế - Luật, ĐHQG HCM - *Email: hoalt@uel.edu.vn*

(Bài nhận ngày 22 tháng 12 năm 2016, hoàn chỉnh sửa chữa ngày 9 tháng 02 năm 2017)

TÓM TẮT

Trong phân tích tài chính, thông thường người ta chỉ sử dụng giá đóng cửa và lựa chọn phân phối của mô hình là phân phối chuẩn. Tuy nhiên, chứng khoán biến động được ghi nhận thông qua bộ bốn giá trị đó là các giá trị giá mở cửa, giá cao nhất, giá thấp nhất và giá đóng cửa. Do đó, chúng tôi sử dụng thêm giá cao nhất và giá thấp nhất nhằm cung cấp thêm thông tin với hy vọng đưa ra kết quả chính xác hơn. Như vậy, bộ dữ liệu sẽ dao động trong một khoảng biến động chứ không phải là một giá trị, tức là dữ liệu dưới dạng số mờ. Và hơn nữa, giả định một bộ dữ liệu tuân theo phân phối chuẩn không phải lúc nào cũng thỏa mãn. Mặt khác, việc kiểm định một dữ liệu có tuân theo phân phối chuẩn hay không thông thường theo kiểm định Jarque Bera hoặc kiểm định Chi bình phương. Để thực hiện các kiểm đinh này cần phải dựa vào giá trị p-value, nhưng hiện nay có rất nhiều tranh cãi xung quanh việc sử dụng giá trị p-value. Do đó, trong bài báo này chúng tôi sử dụng ước lượng điểm Bayes mờ cho dự báo nhằm lựa chọn phân phối phù hợp nhất. Kết quả khi phân tích 9 mã cổ phiếu có giá trị vốn hóa lớn tại thị trường chứng khoán Việt Nam trong khoảng thời gian từ thời điểm niêm yết đến ngày 06/11/2015 thấy rằng có một số mã có các phân phối khác phù hợp hơn phân phối chuẩn, một số mã cổ phiếu phù hợp với phân phối chuẩn.

Từ khóa: *Kiểm tra mô hình Bayes, dữ liệu mờ, ước lượng điểm Bayes mờ, ứng dụng trong phân tích tài chính*

1. GIỚI THIỆU

Việc thu thập dữ liệu không phải lúc nào cũng thu được dữ liệu rõ, các dữ liệu có thể không chính xác do sai số của máy móc cũng như của con người. Do đó, trên thực tế dữ liệu thu thập được trình bày dưới dạng số mờ. Các tính toán thống kê mô tả đối với số mờ như trung bình mẫu mờ, phương sai mẫu mờ, phân phối thực nghiệm của mẫu mờ... được trình bày chi tiết trong (Frühwirth - Schnatter, 1992) . Tương tự như vậy, bài toán kiểm định giả thuyết cho dữ liệu mờ được chỉ ra trong bài (Römer and Kandel, 1995). Thêm vào đó, trong bài (Römer and Kandel, 1995), các tác giả đã trình bày không mức ý nghĩa cho kiểm định phân phối xác suất mờ và kiểm định tham số mờ. Việc kết hợp giữa phương pháp thống kê và lý thuyết tập mờ là một xu hướng cần thiết của thời đại đã được chứng minh trong bài báo (Taheri, 2003). Chính vì vậy, sự mở rộng của lý

thuyết mờ trong thống kê Bayes là một vấn đề quan trọng không chỉ trong lý thuyết mà còn trong thực hành, đặc biệt là trong phân tích tài chính.

Thật sự, thống kê Bayes là rất hữu ích khi cỡ mẫu nhỏ. Không chỉ vậy thống kê Bayes còn thể hiện ưu điểm khi kết hợp giữa định lý Bayes và dữ liệu mờ (Viertl and Hule, 1991). Trong bài báo này, các tác giả đã phân tích phân phối hậu nghiệm mờ, miền biến thiên hậu nghiệm nhỏ nhất cũng như hàm mật độ dự báo mờ. Chẳng hạn như, nếu dữ liệu được chọn tuân theo phân phối mũ, nghiên cứu chọn phân phối tiên nghiệm dạng liên hợp là phân phối gamma thì phân phối hậu nghiệm là phân phối gamma. Việc tính toán miền biến thiên hậu nghiệm nhỏ nhất có thể được tính toán qua chương trình máy tính, nhằm ước lượng tham số θ cần ước lượng. Ngoài ra, phương pháp Bayes về kiểm định giả thuyết mờ được trình bày trong (Taheri and Behboodian, 2001), đồ thị mờ, phân phối xác suất mờ, miền ước lượng mờ, kiểm định giả thuyết mờ... được trình bày trong (Wu, 2005), dự báo mờ và quyết định thống kê được tính toán trong (Viertl, 2006).

Trong suy luận Bayes mờ của dữ liệu không chỉ từ dữ liệu mờ, mà nó còn có thể thông qua phân phối tiên nghiệm mờ, cụ thể là qua tham số tiên nghiệm mờ được chỉ ra trong bài báo (Frühwirth-Schnatter, 1993) . Bởi vậy, có hai loại thông tin mờ đó là dữ liệu mờ $x_1^*, x_2^*, ..., x_n^*$ thông qua hàm hợp lý $l(\theta; x_1^*, x_2^*, ..., x_n^*)$ và thông tin tiên nghiệm mờ $\pi^*(\theta)$ trong không gian tham Θ, cũng được chỉ ra như (Viertl, 2006).

Hầu hết các nghiên cứu trước đây hạn chế trong một tham số, xem (Wu, 2004a). Giả sử rằng ta có n thành phần, mỗi thành phần i được trình bày như một biến ngẫu nhiên Bernoulli Y_i, với xác suất xuất hiện tính chất cần xét là p. Khi đó, tổng của các biến ngẫu nhiên Y_i độc n lập thỏa mãn tính chất cần xét ký hiệu là $X = \sum_{i=1} Y_i$. Với phân phối xác suất của X là phân phối nhị thức. Thông thường, người ta sử dụng phân phối tiên nghiệm liên hợp của p là phân phối beta. Khi đó, phân phối hậu nghiệm của p cũng là phân phối beta. Vì vậy, ước lượng điểm Bayes $\hat{p}$ với hàm tổn thất sai số bình phương phụ thuộc vào cận trên và cận dưới của tham số tại mức $\delta - cut$.

Do đó, trường hợp mở rộng cho nhiều tham số với phân phối chuẩn hay phân phối Weibull được chỉ ra trong (Huang et al., 2006). Với dữ liệu mẫu $D = (x_1, x_2, ..., x_n)$, hàm phân phối mật độ xác suất với dữ liệu thực tế đã xác định $f(x \mid \theta)$. Trong không gian tham số Θ, giả sử phân phối tiên nghiệm là $\pi(\theta)$ thì phân phối hậu nghiệm của tham số θ được xác định như sau

$$\pi(\theta \mid D) = \pi(\theta \mid x_1, x_2, ..., x_n) \propto \pi(\theta) \times l(\theta; x_1, x_2, ..., x_n). \quad (1)$$

Người ta thường sử dụng phân phối tiên nghiệm Jeffrey cho hai tham số của phân phối chuẩn. Còn đối với phân phối Weibull thì người ta sử dụng trường hợp phân phối tiên nghiệm đều. Tổng quát, trong bài báo (Huang et al., 2006), các tác giả hệ thống một phương pháp xác định hàm thành viên cho phân phối nhiều tham số bởi giải thuật di truyền và mạng nhân tạo. Mặc dù vậy, đây là một phương pháp khó để xác định khoảng ước lượng hoặc miền mật độ hậu nghiệm nhỏ nhất...

Dữ liệu thực tế có thể được giả sử tuân theo một số phân phối, như phân phối mũ, phân phối Weibull, phân phối gamma và phân phối log chuẩn... Tương ứng với các phân phối trên các hàm mật độ xác suất, ước lượng tham số, tỷ lệ thành công, tỷ lệ thất bại đã được trình bày trong bài (Shafiq and Viertl, 2016).

Thông thường, trong thống kê tần suất chúng ta thường giả định rằng dữ liệu xấp xỉ

phân phối chuẩn cho bài toán ước lượng hoặc kiểm định giả thuyết. Ngược lại, đối với thống kê Bayes, các nghiên cứu (Jha et al., 2009), (Carlin and Chib, 1995), (Rigoux et al., 2014) đã chỉ ra rằng việc kiểm định dạng phân phối của dữ liệu là hết sức quan trọng bởi vì, chỉ khi có dạng phân phối của dữ liệu, ta mới định ra được phân phối tiên nghiệm cho tham số ước lượng; làm cơ sở tìm phân phối hậu nghiệm để sử dụng cho các tính toán tiếp theo.

Khi đó, chúng ta sẽ sử dụng kiểm định phi tham số để kiểm tra dạng phân phối của dữ liệu. Việc kiểm tra phân phối của dữ liệu thông thường dựa vào giá trị p - value của thuật toán kiểm tra mô hình, hoặc sử dụng phương pháp mô phỏng Monte Carlo (simulated Monte Carlo hoặc Markov chain Monte Carlo). Nhưng hiện nay, đang có rất nhiều tranh cãi về việc sử dụng p-value có thể dẫn đến sai lầm trong việc đưa ra quyết định đối với bài toán kiểm định giả thuyết (Goodman, 2008), (van Helden, 2016)... Bên cạnh đó, khi sử dụng phương pháp mô phỏng Monte Carlo (Markov chain Monte Carlo), cỡ mẫu và tính ổn định của mô phỏng cũng cần được quan tâm đúng mức tạo nên giá trị của kết quả thu được. Do đó, chúng ta rất cần một phương pháp để tìm phân phối tốt nhất xấp xỉ bộ dữ liệu.

Trong bài nghiên cứu này, chúng tôi dựa vào kết quả dự báo đúng cho từng dạng phân phối thông dụng, nếu phân phối nào có kết quả dự báo đúng cao nhất thì dữ liệu phù hợp với phân phối đó nhất. Sau đó, chúng tôi đưa ra một danh sách các phân phối thích hợp cho dữ liệu tài chính khi mà đặc thù của dữ liệu giá chứng khoán nhận giá trị dương và không ổn định và trình bày công thức Bayes tương ứng trong phần 2 của bài báo.

Trong phần 3 của bài báo, chúng tôi trình bày các công thức ước lượng điểm Bayes cho dữ liệu mờ.Và cuối cùng trong phần 4, chúng tôi sử dụng dữ liệu thực tế về giá chứng khoán nhằm ước lượng cho các quan sát tiếp theo. Với mỗi trường hợp, chúng ta có thể kết luận phân phối tốt nhất phù hợp với các dữ liệu thực tế. Phần cuối cùng của bài báo là kết luận và hướng mở rộng.

2. DANH SÁCH CÁC PHÂN PHỐI XÁC SUẤT SỬ DỤNG TRONG THỐNG KÊ BAYES VỚI DỮ LIỆU TÀI CHÍNH

Đối với dữ liệu tài chính, cụ thể là giá chứng khoán, mỗi phiên khung thời gian quan sát luôn có 4 thông tin về giá: mở cửa, thấp nhất, cao nhất và đóng cửa. Trong bốn loại giá trên, giá đóng cửa là quan trọng nhất. Do đó, thông thường chúng ta chỉ sử dụng giá đóng cửa để phân tích cũng như dự báo cho giá đóng cửa phiên tiếp theo.

Như vậy, chúng ta đã mất khá nhiều thông tin về giá cao nhất và giá thấp nhất, ví dụ như giá đóng cửa gần giá thấp nhất thì nhiều khả năng giá đóng cửa của phiên tiếp theo có thể có xu hướng giảm... Trong bài báo này, chúng tôi cố gắng sử dụng thêm thông tin từ các bộ giá chứng khoán này.

Như đã đề cập ở phần trước, dữ liệu trong tài chính thường không ổn định do đó chúng ta sẽ chuyển hóa dữ liệu giữa giá thấp nhất và giá đóng cửa tại thời điểm (ngày) t có dạng như sau

$$low_1(t) = \frac{\text{The lowest price } (t)}{\text{Closing price}(t)}; \qquad (2)$$

trong đó

$low_1(t)$: là giá thấp nhất chuyển hóa tại thời điểm t;

$\text{The lowest price } (t)$: là giá thấp nhất tại thời điểm t;

$\text{Closing price}(t)$: là giá đóng cửa tại thời điểm t.

Và

$$high_1(t) = \frac{\text{The highest price } (t)}{\text{Closing price } (t)}, \quad (3)$$

trong đó

$high_1(t)$: là giá thấp nhất chuyển hóa tại thời điểm *t*;

$\text{The highest price } (t)$: là giá cao nhất tại thời điểm *t*;

$\text{Closing price}(t)$: là giá đóng cửa tại thời điểm *t*.

Rõ ràng, giá trị $low_1(t)$ nằm trong khoảng $(0,1]$ và giá trị $high_1(t)$ nằm trong khoảng $[1,c]$ với hằng số c. Đối với dữ liệu trong tài chính, hằng số c thường không quá lớn, đối với thị trường chứng khoán Việt Nam, trong giai đoạn quan sát, hằng số c lớn nhất nhận giá trị 1.4196.

Suy ra giá trị thấp nhất chuyển hóa $low_1(t)$ và giá cao nhất chuyển hóa $high_1(t)$ của dữ liệu phụ thuộc vào thời gian là ổn định. Vì vậy, chúng ta có hai bộ dữ liệu về giá thấp nhất chuyển hóa low_1 và giá cao nhất chuyển hóa $high_1$, như là một số mờ tại $\delta - cut$ với $\delta = 0$. Ta dễ dàng nhận thấy, số mờ này luôn chứa giá trị 1.

Giả sử rằng mẫu ngẫu nhiên $x_1, x_2, ..., x_n$ bao gồm các quan sát độc lập và cùng phân phối. Tuy nhiên, trong thống kê Bayes, chúng ta chỉ cần các quan sát là thay đổi vị trí được và ổn định. Như vậy, các dữ liệu giá chuyển hóa chứng khoán theo thời gian thỏa mãn điều kiện và nhận giá trị dương nên chúng ta sẽ liệt kê một số phân phối phù hợp dưới đây:

2.1. Phân phối chuẩn và đã biết phương sai σ^2 của tổng thể

Giả sử hàm hợp lý là phân phối chuẩn $N(\mu, \sigma^2)$. Khi đó, chúng ta chọn phân phối tiên nghiệm liên hợp cho trung bình μ là phân phối chuẩn $\pi(\mu) \sim N(\mu_0, \sigma_0^2)$. Phân phối hậu nghiệm cho trung bình cũng là phân phối chuẩn $\pi(\mu \mid x_1, x_2, ..., x_n) \sim N(\mu', \sigma'^2)$ xem (Bolstad, 2013) và (Gelman et al., 2014), được xác định bởi công thức

$$\mu' = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{\mu}{\frac{\sigma^2}{n}}}{\frac{1}{\sigma_0^2} + \frac{1}{\frac{\sigma^2}{n}}}; \frac{1}{\sigma'^2} = \frac{1}{\sigma_0^2} + \frac{1}{\frac{\sigma^2}{n}}. \quad (4)$$

Khi đó, trung bình của phân phối hậu nghiệm là:

$$\mu_{N'} = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{\mu}{\frac{\sigma^2}{n}}}{\frac{1}{\sigma_0^2} + \frac{1}{\frac{\sigma^2}{n}}}. \quad (5)$$

2.2. Phân phối đều

Giả sử hàm hợp lý là phân phối đều $\mathrm{U}(0, \theta)$, khi đó chúng ta chọn phân phối tiên nghiệm liên hợp cho tham số θ là phân phối Pareto $\pi(\theta) \sim \mathrm{P}(x_m, k)$, với $x_1, x_2, ..., x_n$ sao cho $x_i \geq x_m, \ \forall i = \overline{1,n}$ và $k > 1$.

Do đó, phân phối hậu nghiệm cho tham số θ là phân phối Pareto

$$\pi(\theta \mid x_1, x_2, ..., x_n) \sim \mathrm{P}(x_{m'} = max\{x_1, x_2, ..., x_n, x_m\}, k' = k + n) \quad (6)$$

Khi đó, trung bình của phân phối hậu nghiệm cho $k > 1$ là

$$\mu'_{\mathrm{U}} = \frac{k' \times x_{m'}}{k' - 1} = \frac{(k+n) \times (max\{x_1, x_2, ..., x_n, x_m\})}{k + n - 1}. \quad (7)$$

2.3. Phân phối Pareto với trường hợp đã biết giá trị nhỏ nhất x_m

Giả sử hàm hợp lý là hàm Pareto $\mathrm{P}(x_m, k)$, thì chúng ta chọn hàm phân phối tiên nghiệm liên hợp cho tham số hình dạng k là phân phối gamma $\pi(k) \sim \mathrm{G}(\alpha, \beta)$ Chúng ta có phân phối hậu nghiệm cho tham số hình dạng k là phân phối gamma

$$\pi(k \mid x_1, x_2, ..., x_n) \sim \mathrm{G}\left(\alpha' = \alpha + n, \beta' = \beta + \sum_{i=1}^{n} ln\left(\frac{x_i}{x_m}\right)\right). \quad (8)$$

Khi đó, trung bình của phân phối hậu nghiệm được xác đinh bởi công thức

$$\mu'_{\mathrm{P}} = \frac{\alpha'}{\beta'} = \frac{\alpha + n}{\beta + \sum_{i=1}^{n} ln\left(\frac{x_i}{x_m}\right)}. \quad (9)$$

2.4. Phân phối Weibull với đã biết tham số hình dạng β

Giả sử hàm hợp lý tuân theo phân phối Weibull $\mathrm{W}(\theta, \beta)$, khi đó chúng ta chọn phân phối tiên nghiệm liên hợp cho tham số tỷ lệ θ là hàm gamma ngược $\pi(\theta) \sim \mathrm{I\,G}(a, b)$. Do đó, chúng ta sẽ có phân phối hậu nghiệm cho tham số tỷ lệ θ là phân phối gamma ngược

$$\pi(\theta \mid x_1, x_2, ..., x_n) \sim \mathrm{I\,G}(a' = a + n, b' = b + \sum_{i=1}^{n} x_i^{\beta}) \quad (10)$$

Trung bình của phân phối hậu nghiệm được xác định bởi công thức

$$\mu'_{\mathrm{W}} = \frac{b'}{a' - 1} = \frac{b + \sum_{i=1}^{n} x_i^{\beta}}{a + n - 1}. \quad (11)$$

2.5. Phân phối log chuẩn với trường hợp đã biết độ chính xác τ

Giả sử hàm hợp lý có dạng log chuẩn $\mathrm{LN}(\mu, 1/\tau)$. Chúng ta chọn phân phối tiên nghiệm liên hợp cho tham số μ là phân phối chuẩn $\pi(\mu) \sim N(\mu_0, 1/\tau_0)$. Khi đó, phân phối hậu nghiệm cho μ là phân phối chuẩn

$$\pi(\mu \mid x_1, x_2, ..., x_n) \sim N\left(\mu' = \frac{\tau_0\mu_0 + \tau\sum_{i=1}^{n} ln(x_i)}{\tau_0 + n\tau}, \frac{1}{\tau'} = \frac{1}{\tau_0 + n\tau}\right). \quad (12)$$

Trung bình của phân phối hậu nghiệm được xác định bởi công thức

$$\mu'_{\mathrm{LN}} = \mu' = \frac{\tau_0\mu_0 + \tau\sum_{i=1}^{n} ln(x_i)}{\tau_0 + n\tau}. \quad (13)$$

2.6. Phân phối mũ

Giả sử rằng hàm hợp lý có dạng phân phối mũ $\mathrm{E}(\lambda)$, chúng ta chon hàm phân phối tiên nghiệm liên hợp cho tham số λ là phân phối gamma $\pi(\lambda) \sim \mathrm{G}(\alpha, \beta)$. Do đó, chúng ta có phân phối hậu nghiệm cho tham số λ cũng là phân phối gamma

$$\pi(\lambda \mid x_1, x_2, ..., x_n) \sim \mathrm{G}\left(\alpha' = \alpha + n, \beta' = \beta + \sum_{i=1}^{n} x_i\right) \quad (14)$$

Trung bình của phân phối hậu nghiệm được xác định bởi công thức

$$\mu'_{\mathrm{E}} = \frac{\alpha'}{\beta'} = \frac{\alpha + n}{\beta + \sum_{i=1}^{n} x_i}. \quad (15)$$

2.7. Phân phối gamma với điều kiện đã biết tham số hình dạng $\propto$

Nếu dữ liệu tuân theo phân phối gamma $\mathrm{G}(\alpha, \beta)$, chúng ta sẽ chọn phân phối tiên nghiệm liên hợp cho tham số tỷ lệ β là phân phối gamma $\pi(\beta) \sim \mathrm{G}(\alpha_0, \beta_0)$. Khi đó, phân phối hậu nghiệm cho tham số tỷ lệ β cũng là phân phối gamma

$$\pi(\beta \mid x_1, x_2, ..., x_n) \sim \mathrm{G}\left(\alpha' = \alpha_0 + n\alpha, \beta' = \beta_0 + \sum_{i=1}^{n} x_i\right) \quad (16)$$

Trung bình của phân phối hậu nghiệm được xác định bởi công thức

$$\mu'_{\mathrm{G}} = \frac{\alpha'}{\beta'} = \frac{\alpha_0 + n\alpha}{\beta_0 + \sum_{i=1}^{n} x_i}. \quad (17)$$

2.8. Phân phối gamma ngược với điều kiện đã biết tham số hình dạng a

Giả sử hàm hợp lý có dạng phân phối gamma ngược $\mathrm{I\,G}(a, b)$, chúng ta chọn hàm

phân phối tiện nghiệm liên hợp cho tham số hình dạng ngược β là phân phối gamma $\pi(\beta) \sim G(\alpha_0, \beta_0)$. Khi đó, phân phối hậu nghiệm cho tham số hình dạng ngược có dạng

$$\pi(\beta \mid x_1, x_2, ..., x_n) \sim G\left(\alpha' = \alpha_0 + na, \beta' = \beta_0 + \sum_{i=1}^{n} \frac{1}{x_i}\right). \quad (18)$$

Trung bình của phân phối hậu nghiệm được xác định bởi công thức

$$\mu'_{1G} = \frac{\alpha'}{\beta'} = \frac{\alpha_0 + na}{\beta_0 + \sum_{i=1}^{n} \frac{1}{x_i}}. \quad (19)$$

3. CÔNG THỨC ƯỚC LƯỢNG ĐIỂM BAYES CHO DỮ LIỆU MỜ

Trước hết, chúng ta tìm hiểu định nghĩa số mờ và $\delta - cut$, xem (Viertl, 2011).

Định nghĩa 1. Một số mờ x^* được xác định bởi hàm đặc trưng tương ứng $\xi(.)$ thỏa mãn các tính chất sau:

Hàm thực $\xi : \mathbb{R} \to [0; 1]$

Với mọi $\delta \in [0; 1]$ tương ứng với $\delta - cut$ được xác định:

$C_\delta(x^*) = \{x \in \mathbb{R}: \xi(x) \geq \delta\}$. $\delta - cut$ là hợp hữu hạn của các khoảng bị chặn $[a_{\delta,j}; b_{\delta;j}]$, tức là:

$$C_\delta(x^*) = \bigcup_{j=1}^{k_j} [a_{\delta,j}; b_{\delta;j}] \neq \varnothing.$$

Tập hỗ trợ của $\xi(.)$, định nghĩa bởi $supp[\xi(.)] = \{x \in \mathbb{R} : \xi(x) > 0\}$ là bị chặn.

Trong bài báo này, chúng tôi sử dụng mẫu ngẫu nhiên mờ dạng liên tục và chỉ có một đỉnh nên $\delta - cut$ tương ứng với các quan sát sẽ chỉ là một khoảng bị chặn.

Giả sử, ta có mẫu ngẫu nhiên mờ $x_1^*, x_2^*, ..., x_n^*$. Khi đó theo nguyên lý mở rộng Zadeh, thì mỗi quan sát có cận dưới $\underline{x_i}$ và cận trên $\overline{x_i}$. Tương tự như vậy, cận dưới và cận trên tương ứng cho các tham số của hàm hợp lý, hàm tiên nghiệm và hàm hậu nghiệm.

Sử dụng $\delta - cut$ của các giá trị mờ $\pi^*(\theta), \theta \in \Theta$ được biểu thị bởi $[\underline{\pi}_\delta(\theta), \overline{\pi}_\delta(\theta)]$. Tương tự như vậy, $\delta - cut$ của hàm hợp lý $l(\theta; x_1^*, x_2^*, ..., x_n^*)$ với các giá trị tương ứng là $[\underline{l}_\delta(\theta; x_1^*, x_2^*, ..., x_n^*), \overline{l}_\delta(\theta; x_1^*, x_2^*, ..., x_n^*)]$.

Khi đó, hàm phân phối hậu nghiệm mờ $\pi^*(\theta \mid x_1^*, x_2^*, ..., x_n^*)$ được xác định bởi công thức $[\underline{\pi}_\delta(\theta \mid x_1^*, x_2^*, ..., x_n^*), \overline{\pi}_\delta(\theta \mid x_1^*, x_2^*, ..., x_n^*)]$ thông qua định nghĩa sau:

$$\underline{\pi}_\delta(\theta \mid x_1^*, x_2^*, ..., x_n^*) = \frac{\underline{\pi}_\delta(\theta) \times \underline{l}_\delta(\theta; x_1^*, x_2^*, ..., x_n^*)}{\int_\Theta \frac{1}{2}\left[\underline{\pi}_\delta(\theta) \times \underline{l}_\delta(\theta; x_1^*, x_2^*, ..., x_n^*) + \overline{\pi}_\delta(\theta) \times \overline{l}_\delta(\theta; x_1^*, x_2^*, ..., x_n^*)\right]};$$

$$\overline{\pi}_\delta(\theta \mid x_1^*, x_2^*, ..., x_n^*) = \frac{\overline{\pi}_\delta(\theta) \times \overline{l}_\delta(\theta; x_1^*, x_2^*, ..., x_n^*)}{\int_\Theta \frac{1}{2}\left[\underline{\pi}_\delta(\theta) \times \underline{l}_\delta(\theta; x_1^*, x_2^*, ..., x_n^*) + \overline{\pi}_\delta(\theta) \times \overline{l}_\delta(\theta; x_1^*, x_2^*, ..., x_n^*)\right]};$$

$$\forall \theta \in \Theta, \delta \in [0,1].$$

Áp dụng những kết quả trên vào từng dạng phân phối, chúng ta tìm ước lượng điểm Bayes mờ cho trung bình hậu nghiệm. Sau đó, chúng ta sử dụng khoảng ước lượng này cho quan sát tiếp theo. Nếu giá trị thật của quan sát tiếp theo rơi vào đúng khoảng dự báo thì chúng ta kết luận dự báo đúng, và ngược lại thì dự báo sai.

Trong bài báo này, chúng tôi muốn kiểm tra một dữ liệu tuân theo phân phối nào là tốt nhất. Phân phối nào tốt nhất thì có nhiều giá trị quan sát thật rơi vào khoảng dự báo. Chúng tôi cố gắng minh họa bằng dữ liệu thực nghiệm.

4. ỨNG DỤNG ƯỚC LƯỢNG ĐIỂM BAYES CHO DỮ LIỆU MỜ TẠI MỨC $\delta - cut$ =0

Chúng ta sử dụng tập dữ liệu $low_1(t)$ và $high_1(t)$ tương ứng với cận dưới và cận trên tại mức δ_{cut}, δ =0. Sử dụng kỹ thuật tương tự trong (Wu, 2004b) cho ước lượng điểm Bayes mờ thích hợp với mỗi phân phối.

4.1. Dữ liệu thực nghiệm

Dữ liệu thực nghiệm được sử dụng là dữ liệu giá chứng khoán của sàn giao dịch chứng khoán Hà Nội, Việt Nam bao gồm 9 mã cổ phiếu. Các mã cổ phiếu này từ thời điểm bắt đầu lên sàn đến ngày 06/11/2015. Chúng tôi chọn 9 mã cổ phiếu này dựa vào giá trị của các mã cổ phiếu tại ngày 06/11/2015 theo bảng 1. Các cổ phiếu này có tính thanh khoản cao, điều này giúp cho giá cổ phiếu khó bị "làm giá" và dữ liệu sẽ tốt hơn.

Bảng 1. Các mã cổ phiếu quan tâm

Mã cổ phiếu	'DXP'	'HAT'	'MAS'	'NTP'	'SLS'	'TCT'	'VCS'	'VNF'	'WCS'
Ngày niêm yết (Ngày/	26	29	10	11	16	06	17	01	17
Tháng/	12	10	9	12	10	12	12	12	9
Năm)	2005	2010	2009	2006	2012	2006	2007	2010	2010
Tổng số quan sát dự báo	2222	711	707	2096	406	2126	1801	934	1004

4.2. Phân tích dữ liệu

Trong bảng 2 thể hiện kết quả dự báo dựa trên danh sách các phân phối và tính toán của tác giả.

Bảng 2. Tỷ lệ dự báo đúng dựa trên ước lượng điểm Bayes cho dữ liệu mờ

Phân phối và mã cổ phiếu	'DXP'	'HAT'	'MAS'	'NTP'	'SLS'	'TCT'	'VCS'	'VNF'	'WCS'
Chuẩn	0.9743	**0.9789**	**0.9929**	**0.9690**	**0.9926**	**0.9708**	0.9611	**0.9636**	**0.9751**
Đều	0.9167	0.8636	0.8571	0.8726	0.9704	0.8960	0.8978	0.9111	0.8337
Pareto	0.9770	0.8833	0.9321	0.9380	0.9803	0.9600	**0.9672**	0.9550	0.8815
Weibull	0.9721	0.8861	0.9321	0.9380	0.9828	0.9633	0.9645	0.9540	0.8855
Log chuẩn	**0.9779**	0.8790	0.9321	0.9399	0.9852	0.9610	0.9622	0.9529	0.8865
Mũ	**0.9779**	0.8833	0.9321	0.9389	0.9803	0.9610	0.9656	0.9550	0.8825
Gamma	0.3240	0.8270	0.8416	0.2171	0.6995	0.2855	0.3037	0.4989	0.4303
Gamma ngược	0.3240	0.8270	0.8416	0.2166	0.6995	0.2855	0.3032	0.4989	0.4303

Dựa vào bảng 2, chúng ta thấy rằng có một điều đặc biệt là các mã cổ phiếu HAT, MAS và SLS hầu như xấp xỉ đối với phân phối nào cũng đều cho kết quả dự báo tốt, mặc dù phân phối chuẩn vẫn là phân phối tốt nhất. Cụ thể là các mức dự báo đúng trên 80 phần trăm cho HAT và MAS, đúng trên 70 phần trăm cho mã cổ phiếu SLS. Còn đối với dự báo tốt nhất cho phân phối chuẩn tương ứng với ba mã cổ phiếu này có tỷ lệ dự báo đúng lần lượt là mã cổ phiếu HAT là 0.978, mã cổ phiếu MAS là 0.993 và mã cổ phiếu SLS là 0.993.

Tiếp theo đó, chúng ta thấy rằng các mã cổ phiếu DXP, NTP, TCT, VCS, VNF và WCS phù hợp với các phân phối chuẩn, đều, Pareto, Weibull, log chuẩn và phân phối mũ hơn phân phối gamma và gamma ngược, do tỷ lệ đúng cao hơn. Cụ thể là với mã cổ phiếu DXP có phân phối đúng tốt nhất là phân phối mũ và phân phôi log chuẩn với tỷ lệ dự báo đúng xấp xỉ 0.978. Các phân phối xấp xỉ đúng tiếp theo phù hợp với mã cổ phiếu DXP này là phân phối Pareto với tỷ lệ dự báo đúng là 0.977, phân phối chuẩn với tỷ lệ dự báo đúng là 0.974, phân phối Weibull với tỷ lệ dự báo đúng là 0.972 và phân phối đều với tỷ lệ dự báo đúng là 0.917. Tuy nhiên, khi chuyển qua xấp xỉ mã cổ phiếu DXP dưới dạng phân phối gamma hay phân phối gamma ngược thì tỷ lệ dự báo đúng chỉ xuống còn 0.324.

Còn đối với các mã cổ phiếu NTP, TCT, VNF và WCS thì phân phối tốt nhất là phân phối chuẩn. Điều này phù hợp với hầu hết các nghiên cứu về giá chứng khoán hiện nay, khi họ coi phân phối xấp xỉ tốt nhất cho dữ liệu giá chứng khoán.

Vậy có một câu hỏi đặt ra rằng, phải chăng vì khoảng dự báo quá rộng nên dự báo thì chắc chắn đúng. Do đó, chúng tôi sẽ hiệu chỉnh lại độ dài khoảng dự báo đúng.

4.3. Hiệu chỉnh khoảng dự báo

Trong thị trường chứng khoán Việt Nam, biên độ dao động đến 20 phần trăm cho hầu hết các mã cổ phiếu (trừ hai mã cổ phiếu 'VCS' dao động đến 35.29 phần trăm và 'VNF' dao động đến 25.74 phần trăm). Do đó, đầu tiên chúng ta thử thu hẹp miền dự báo trong khoảng 10 phần trăm. Kết quả dự báo đúng cho phiên giao dịch tiếp theo với miền dự báo có độ dài 10 phần trăm được tác giả thể hiện trong bảng 3.

Bảng 3. Miền dự báo 10 phần trăm

Phân phối và các mã cổ phiếu	**'DXP'**	**'HAT'**	**'MAS'**	**'NTP'**	**'SLS'**	**'TCT'**	**'VCS'**	**'VNF'**	**'WCS'**
Chuẩn	0.9001	**0.5809**	0.5827	**0.8698**	0.7931	0.8791	0.7512	0.7334	**0.7610**
Đều	0.7912	0.4501	0.5573	0.7228	0.6650	0.8043	0.6219	0.5557	0.5000
Pareto	0.9181	0.5724	**0.5997**	0.8440	**0.8227**	0.9280	**0.7640**	0.7430	0.6922
Weibull	0.9181	0.5724	**0.5997**	0.8440	**0.8227**	**0.9285**	0.7618	**0.7420**	0.6873
Log chuẩn	**0.9190**	0.5724	**0.5997**	0.8445	**0.8227**	0.9280	0.7618	**0.7420**	0.6892
Mũ	0.9185	0.5724	**0.5997**	0.8449	**0.8227**	0.9276	0.7607	0.7388	0.6902
Gamma	0.1566	0.4613	0.3607	0.0654	0.3079	0.1317	0.1321	0.1991	0.2151
Gamma ngược	0.1566	0.4613	0.3607	0.0654	0.3079	0.1317	0.1321	0.1991	0.2151

Theo kết quả của bảng 3, nếu chúng ta thu hẹp miền dự báo xuống còn 10 phần trăm thì các mã cổ phiếu DXP, NTP, SLS, TCT và VCS hầu như có tỷ lệ dự báo đúng không giảm nhiều so với khoảng dự báo gốc ban đầu. Tuy nhiên, hai mã cổ phiếu HAT và MAS có giảm tỷ lệ dự báo đúng một cách tương đối lớn, với mức giảm khoảng 40 phần trăm. Điều này có nghĩa là khoảng tin cậy của hai mã cổ phiếu HAT và MAS lớn, vì vậy khoảng biến động này dài nên ít có ý nghĩa trong thực tế.

Trong khi đó các mã cổ phiếu DXP, SLS, TCT, VCS và VNF thích hợp với phân phối Pareto, Weibull, log chuẩn, mũ hơn phân phối chuẩn thì hai mã cổ phiếu NTP và WSS xấp xỉ

phân phối chuẩn tốt hơn các phân phối khác.

Dựa vào tỷ lệ dự báo đúng trong bảng 3, ta thấy với miền dự báo với khoảng sai lêch 10 phần trăm vẫn còn ở mức xác suất tương đối cao, khoảng 70 đến 80 phần trăm.

Như vậy, đây là một tín hiệu tốt cho ứng dụng của thống kê Bayes mờ trong phân tích tài chính.

Bảng 4. Miền dự báo 5 phần trăm

Phân phối và các mã cổ phiếu	'DXP'	'HAT'	'MAS'	'NTP'	'SLS'	'TCT'	'VCS'	'VNF'	'WCS'
Chuẩn	0.6571	0.3235	0.4286	**0.6398**	0.5419	0.6308	0.4770	0.4722	**0.4811**
Đều	0.4982	0.2293	0.3479	0.4046	0.3300	0.4581	0.3137	0.3062	0.2580
Pareto	**0.6760**	**0.3882**	**0.4668**	0.6307	**0.6502**	0.6458	0.5097	0.5300	0.4771
Weibull	0.6751	**0.3882**	**0.4668**	0.6312	**0.6502**	0.6468	**0.5108**	0.5268	0.4771
Log chuẩn	0.6742	**0.3882**	**0.4668**	0.6360	0.6478	0.6491	0.5097	0.5321	**0.4811**
Mũ	0.6742	**0.3882**	**0.4668**	0.6369	0.6478	**0.6496**	0.5092	**0.5332**	**0.4811**
Gamma	0.1071	0.2968	0.2702	0.0344	0.2365	0.0626	0.0772	0.1413	0.1434
Gamma ngược	0.1071	0.2968	0.2702	0.0344	0.2365	0.0626	0.0772	0.1413	0.1434

Nguồn: Kết quả nghiên cứu

Nếu chúng ta thu hẹp miền dự báo với khoảng biến động 5 phần trăm, kết quả được xác định trong bảng 4. Kết quả bây giờ không còn cao nữa. Tuy nhiên với khoảng biến động quá bé, miền dự báo chỉ còn khoảng $1/3$ hoặc $1/4$ so với khoảng biến động cho phép. Do đó, chỉ các mã cổ phiếu DXP, NTP, SLS và TCT có tỷ lệ dự báo đúng là chấp nhận được, tức là ở khoảng trên 60 phần trăm. Tức là, các mã cổ phiếu này có xấp xỉ theo các phân phối Pareto, Weibull, log chuẩn, mũ thích hợp hơn so với phân phối chuẩn, cũng như phân phối đều, gamma và gamma ngược. Kết quả tương tự đối với các mã cổ phiếu TCT và SLS. Tuy nhiên, mã cổ phiếu NTP phù hợp với phân phối chuẩn hơn các phân phối khác.

5. KẾT LUẬN

Trong thực hành về phân tích dữ liệu theo thống kê Bayes, việc kiểm tra xem dữ liệu phù hợp với phân phối nào nhất là một vấn đề hết sức quan trọng. Có một số cách để kiểm tra mô hình tương tự như kiểm định chi square trong thống kê tần suất hoặc mô phỏng Monte Carlo. Tuy nhiên, cách kiểm tra mô hình này lại dựa vào giá trị p-value. Trong khi việc sử dụng giá trị p-value đang gây nhiều tranh cãi, nhóm tác giả cũng đã có một nghiên cứu liên quan đến vấn đề này trong bài báo (Nguyen et al., 2016). Còn nếu phương pháp sử dụng mô phỏng Monte Carlo cho phân phối hậu nghiệm, thì câu hỏi đặt ra là số lượng mô phỏng là bao nhiêu, đến khi nào thì ổn định... nhất là khi áp dụng trong tài chính với nhiều bộ dữ liệu, mỗi bộ dữ liệu bao gồm cả ngàn quan sát theo thời gian.

Đặc biệt, trong trường hợp dữ liệu mờ việc kiểm tra mô hình của dữ liệu lại càng quan trọng. Do đó, trong bài báo này chúng tôi muốn lấy đúng thực tiễn để chứng minh cho vấn đề đưa ra. Tức là, chúng tôi giả định một số dạng phân phối thường gặp cho dữ liệu giá chứng

khoán. Sau đó, sử dụng công thức Bayes cho từng dạng phân phối nhằm dự báo cho giá đóng cửa của phiên kế tiếp. Tỷ lệ dự báo tuân theo phân phối nào lớn hơn thì chứng tỏ dữ liệu tuân theo phân phối đó tốt hơn.

Phương pháp sử dụng trong bài báo thông qua ước lượng điểm thống kê Bayes mờ, có hiệu chỉnh cho phù hợp trong phân tích tài chính. Kết quả dự báo với 9 mã cổ phiếu cho thấy tỷ lệ dự báo tương đối tốt ở mức 70 đến 90 phần trăm khi sử dụng toàn bộ miền ước lượng điểm hoặc thu hẹp biên độ 10 phần trăm. Còn khi thu hẹp biên độ dao động là 5 phần trăm thì mức độ dự báo đúng khoảng 60 phần trăm. Hơn nữa, thông qua kết quả dự báo đúng, chúng tôi cũng đã chứng tỏ sự phù hợp của mô hình. Cách đánh giá này khác với cách đánh giá kết quả truyền thống khi mà độ phù hợp của mô hình được ẩn sau xác suất dự báo đúng.

Với kết quả tương đối khả quan của bài báo, chúng tôi hy vọng ứng dụng của thống kê Bayes mờ áp dụng sâu rộng hơn vào trong phân tích tài chính với không chỉ sử dụng giá đóng cửa mà còn sử dụng thêm thông tin giá cao nhất và giá thấp nhất để dự báo. Đây là một kết quả hoàn toàn mới của chúng tôi khi chưa có ai sử dụng cách xử lý dữ liệu mới là thống kê Bayes mờ vào bộ dữ liêu theo cách hiệu chỉnh như vậy.

Chúng tôi xin chân thành cảm ơn Giáo sư Nguyễn Trung Hưng, Trường Đại học New Mexico và Đại học Chiang Mai vì sự giúp đỡ tận tâm của ông đối với nghiên cứu của chúng tôi thông qua các Hội nghị, Hội thảo và các cuộc thảo luận. Bên cạnh đó, chúng tôi cũng cảm ơn Trường Đại học Kinh tế - Luật đã tài trợ cho chúng tôi trong khuôn khổ đề tài, với mã số CS 2016-13.

Choosing the best model in fuzzy Bayesian statistics and its application in financial analysis

- **Pham Hoang Uyen**
- **Le Thanh Hoa**
- **Nguyen Dinh Thien**

University of Economics and Law, VNU HCM - *Email: hoalt@uel.edu.vn*

ABSTRACT

Analysts generally use closing price and normal distribution assumption for a model's distribution in financial analysis. However, stock price fluctuation is reflected by a set of four values, namely opening, highest, lowest and closing prices. We therefore include the highest and the lowest prices to take into account more information in the hope of ending up with a more exact result as data contains a ranges of values instead of one only (i.e. the data is a form of fuzzy number). Moreover, the assumption that data is normally distributed is not always satisfied and Jacque Bera or Chi square tests are often employed to test the data's normality. The tests require the use of p-value which is quite controversial at present. This paper employs fuzzy Bayes point estimator to choose the most suitable distribution. On a sample of 9 stocks with large capitalization in Vietnam from their listed dates until November 06, 2015, we found that some stocks have prices distributed more reasonably than normal distribution and some are not.

Key word: *Testing Bayes model, fuzzy data, the estimate of fuzzy Bayes point, application in financial analysis.*

TÀI LIỆU THAM KHẢO

[1]. Bolstad, W.M. (2013), Introduction to Bayesian statistics. John Wiley & Sons.

[2]. Carlin, B.P., Chib, S. (1995), B*ayesian model choice via Markov chain Monte Carlo methods.* J. R. Stat. Soc. Ser. B Methodol. 473–484.

[3]. Frühwirth-Schnatter, S., O*n fuzzy Bayesian inference.* Fuzzy Sets Syst. 60, 41–58 (1993).

[4]. Frühwirth-Schnatter, S. (1992), *On statistical inference for fuzzy data with applications to descriptive statistics.* Fuzzy Sets Syst. 50, 143–165.

[5]. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B. (2014), *Bayesian data analysis.* Chapman & Hall/CRC Boca Raton, FL, USA.

[6]. Goodman, S. (2008), *A dirty dozen: twelve p-value misconceptions, in: Seminars in Hematology*. Elsevier, pp. 135–140.

[7]. Huang, H.-Z., Zuo, M.J., Sun, Z.-Q. (2006), Bayesian reliability analysis for fuzzy

lifetime data. Fuzzy Sets Syst. 157, 1674–1686.

[8]. Jha, S.K., Clarke, E.M., Langmead, C.J. (2009), Legay, A., Platzer, A., Zuliani, P., *A bayesian approach to model checking biological systems, in: International Conference on Computational Methods in Systems Biology*. Springer, pp. 218–234.

[9]. Nguyen, S.P., Pham, U.H., Nguyen, T.D., Le, H.T. (2016), *A New Method for Hypothesis Testing Using Inferential Models with an Application to the Changepoint Problem, in: Integrated Uncertainty in Knowledge Modelling and Decision Making: 5th International Symposium*, IUKM 2016, Da Nang, Vietnam, November 30-December 2, 2016, Proceedings. Springer, pp. 532–541.

[10]. Rigoux, L., Stephan, K.E., Friston, K.J., Daunizeau, J. (2014), *Bayesian model selection for group studies—revisited*. Neuroimage 84, 971–985.

[11]. Römer, C., Kandel, A. (1995), Statistical tests for fuzzy data. Fuzzy Sets Syst. 72, 1–26.

[12]. Shafiq, M., Viertl, R. (2016), On the Estimation of Parameters, Survival Functions, and Hazard Rates Based on Fuzzy Life Time Data. Commun. Stat.-Theory Methods.

[13]. Taheri, S.M. (2003), *Trends in fuzzy statistics*. Austrian J. Stat. 32, 239–257.

[14]. Taheri, S.M., Behboodian, J. (2001), *A Bayesian approach to fuzzy hypotheses testing*. Fuzzy Sets Syst. 123, 39–48.

[15]. Helden, J. (2016), *Confidence intervals are no salvation from the alleged fickleness of the P value*. Nat. Methods 13, 605–606.

[16]. Viertl, R. (2011), *Statistical methods for fuzzy data*. John Wiley & Sons.

[17]. Viertl, R. (2006), *Univariate statistical analysis with fuzzy data*. Comput. Stat. Data Anal. 51, 133–147.

[18]. Viertl, R., Hule, H. (1991), *On Bayes' theorem for fuzzy data*. Stat. Pap. 32, 115–122.

[19]. Wu, H.-C. (2005), *Statistical hypotheses testing for fuzzy data*. Inf. Sci. 175, 30–56.

[20]. Wu, H.-C. (2004a), *Fuzzy reliability estimation using Bayesian approach. Comput*. Ind. Eng. 46, 467–493.

[21]. Wu, H.-C. (2004b), *Fuzzy Bayesian estimation on lifetime data*. Comput. Stat. 19, 613–633.