VNUHCM Journal of Advanced Research and Natural Sciences 2026, 10(2): 3654-3672

Open Access Full Text Article

Bài báo nghiên cứu

GIẢI CHẬP CÁC HÀM PHÂN PHỐI TÍCH LŨY TRONG TRƯỜNG HỢP PHƯƠNG SAI NHIỄU GAUSS CHƯA BIẾT.

Thai Phuc Hung [1,2,3], Nguyen Hoang Thanh [1,2,*]

Use your smartphone to scan this QR code and download this article

[1]*Faculty of Mathematics and Computer Science, University of Science, Ho Chi Minh City, Vietnam*

[2]*Vietnam National University, Ho Chi Minh City, Vietnam*

[3]*Faculty of Basics, Soc Trang Community College, Can Tho, Vietnam*

Liên hệ

Nguyen Hoang Thanh, Faculty of Mathematics and Computer Science, University of Science, Ho Chi Minh City, Vietnam

Vietnam National University, Ho Chi Minh City, Vietnam

Email: nguyenhoangthanh1010@gmail.com

Lịch sử

- Ngày nhận: 09-12-2025
- Ngày sửa đổi: 27-02-2026
- Ngày chấp nhận: 24-06-2026
- Ngày đăng: 28-06-2026

DOI : https://doi.org/10.32508/vnuhcmj-arns.v10i2.1504

Check for updates

Bản quyền

© Tạp chí ĐHQG Tp.HCM. Đây là bài báo công bố mở được phát hành theo các điều khoản của the Creative Commons Attribution 4.0 International license.

TÓM TẮT

Bài toán giải chập trong mô hình sai số đo lường, nơi các quan sát bị nhiễu được mô tả bởi $Y_j = X_j + \varepsilon_j$ với $\varepsilon_j \sim N(0, \sigma^2)$) có phương sai chưa biết, đã được trình bày với mục tiêu là khôi phục hàm phân phối tích lũy F_X của biến ngẫu nhiên mục tiêu X chỉ từ một mẫu quan sát nhiễu duy nhất. Khác với phần lớn các nghiên cứu tập trung vào ước lượng mật độ, suy luận trực tiếp F_X, một đại lượng giàu thông tin và có nhiều ứng dụng trong đánh giá rủi ro, kiểm định và phân loại được khảo sát.

Thủ tục gồm ước lượng bán tham số gồm hai bước: (i) ước lượng phương sai nhiễu σ^2 trực tiếp từ dữ liệu quan sát; và (ii) thay thế giá trị ước lượng này vào phương pháp giải chập để thu được ước lượng cho F_X. Khi phân phối F_X thuộc lớp ordinary smooth, đặc trưng bởi điều kiện suy giảm đa thức của hàm đặc trưng trong không gian Sobolev, chúng tôi chứng minh bộ ước lượng đạt tốc độ hội tụ $(lnn)^{-(\alpha+1/2)}$. Tốc độ này trùng với chuẩn minimax trong trường hợp phương sai nhiễu đã biết, cho thấy khả năng thích nghi hoàn toàn với mức nhiễu chưa biết mà không làm giảm hiệu quả thống kê.

Phương pháp này không yêu cầu bất kỳ mẫu bổ sung nào như dữ liệu thuần nhiễu hay phép đo lặp lại, nhờ đó phù hợp với nhiều bối cảnh thực nghiệm nơi dữ liệu hạn chế. Đây là công trình đầu tiên cung cấp đảm bảo lý thuyết đầy đủ cho ước lượng hàm phân phối trong mô hình giải chập Gaussian với phương sai chưa biết trong miền ordinary smooth.

Từ khoá: Giải chập, Hàm phân phối, Nhiễu Gaussian, Ước lượng bán tham số

MỞ ĐẦU

Bài toán khôi phục hàm phân phối của một biến ngẫu nhiên mục tiêu không quan sát được dưới dạng mô hình cộng bị nhiễu bởi nhiễu Gauss đã được nghiên cứu, với việc xét mô hình

$$Y_j = X_j + \varepsilon_j, j = 1,2,\dots,n \quad (1)$$

trong đó các quan sát $Y_1,\dots,Y_n$ là độc lập và có cùng phân phối (i.i.d.), $X_j \sim X$ là các giá trị của biến ngẫu nhiên mục tiêu cần nghiên cứu, và các nhiễu $\varepsilon_j \sim N(0,\sigma^2)$ có phân phối Gauss với phương sai chưa biết và độc lập với X. Các mô hình sai số đo lường dạng này xuất hiện trong nhiều lĩnh vực, bao gồm kinh tế và tài chính, nơi giá tài sản chịu tác động của nhiễu vi mô thị trường; trong kỹ thuật và xử lý tín hiệu, nơi dữ liệu cảm biến bị ảnh hưởng bởi các sai lệch thiết bị; và trong khoa học y sinh, nơi các tín hiệu bị nhiễu bởi các yếu tố biến thiên từ môi trường hoặc thí nghiệm.

Kí hiệu f_X, f_Y, và f_ε lần lượt là các hàm mật độ của X, Y và ε. Ta có phép tích chập,

$$f_Y(x) = \int_{-\infty}^{+\infty} f_X(x-u) f_\varepsilon(u) du = (f_X * f_\varepsilon)(x) \quad (2)$$

Trích dẫn bài báo này: Thai P H, Nguyen H T. **GIẢI CHẬP CÁC HÀM PHÂN PHỐI TÍCH LŨY TRONG TRƯỜNG HỢP PHƯƠNG SAI NHIỄU GAUSS CHƯA BIẾT.**. *VNUHCM J. Adv. Res. Nat. Sci. 2026; 10(2):3654-3672.*

Tương đương, xét theo hàm phân phối tích lũy (CDF),

$$F_X(x) = (F_X * f_\varepsilon)(x)$$

trong đó F_X và F_Y lần lượt là CDF của X và Y. Vì vậy, mục tiêu khôi phục F_X từ một mẫu dữ liệu nhiễu, là một trường hợp điển hình của bài toán giải chập vốn đã được nghiên cứu sâu rộng trong cả ước lượng hàm mật độ và hàm phân phối (xem [1–4, 6–10, 13, 14]).

Hầu hết các nghiên cứu hiện nay tập trung vào ước lượng hàm mật độ, bởi việc biết f_X cho phép khai thác các đặc trưng phân phối cục bộ, hỗ trợ mô hình hóa phi tham số và tính toán mômen. Tuy nhiên, trong thực tiễn, hàm phân phối F_X lại thường trực quan và giàu thông tin hơn, vì cung cấp cái nhìn tổng quan về quy luật nền và cho phép đánh giá trực tiếp các xác suất như $P(X < Y)$ một đại lượng quan trọng trong phân tích rủi ro, kiểm định chẩn đoán và phân loại (xem [11, 12]). Dù có ý nghĩa như vậy, các phương pháp ước lượng CDF trong mô hình giải chập vẫn còn hạn chế. Đặc biệt, trường hợp phân phối nhiễu chỉ được xác định một phần mới chỉ được khảo sát trong vài công trình gần đây (xem [8, 9]) do phát sinh nhiều thách thức kỹ thuật bổ sung.

Một thách thức lớn xuất phát từ việc xử lý phân phối sai số. Lý thuyết giải chập thường giả định phân phối nhiễu đã biết hoàn toàn, qua đó đơn giản hóa đáng kể quá trình suy luận. Tuy nhiên, giả định này hiếm khi phù hợp với thực tế. Để khắc phục, một số nghiên cứu đề xuất khai thác dữ liệu bổ trợ nhằm nới lỏng giả định đó: chẳng hạn, sử dụng một mẫu thuần nhiễu để ước lượng trực tiếp phân phối sai số (xem [8]), hoặc tận dụng các phép đo lặp lại của cùng một biến tiềm ẩn.

$$Y_{j,k} = X_{j,k} + \varepsilon_{j,k}\,, j = 1,2,\dots,n, k = 1,2,$$

có thể được sử dụng để suy ra luật nhiễu (xem [9]). Mặc dù các cách tiếp cận này có thể rất hiệu quả, chúng lại phụ thuộc vào dữ liệu bổ sung vốn thường không sẵn có và có thể làm tăng đáng kể chi phí nghiên cứu. Điều này thúc đẩy việc phát triển các phương pháp chỉ dựa trên một mẫu dữ liệu nhiễu duy nhất.

Phát biểu bài toán

Giả định rằng nhiễu tuân theo phân phối Gauss với phương sai chưa biết, còn phân phối F_X thuộc một lớp con trơn của các phân phối Sobolev. Cụ thể, hàm đặc trưng của X thỏa mãn điều kiện trơn Sobolev:

$$\varphi_X(t) = \kappa\,|t|^{-\lambda}\,[1 + \psi(t)], \quad t \in R, \tag{3}$$

với hằng số $\kappa > 0, \lambda > 0$, và hàm $\psi: R \to C$ thỏa mãn điều kiện Sobolev. Điều kiện này đặc trưng cho tốc độ suy giảm đa thức của φ_X, vốn xác định miền trơn

thường, trái ngược với tốc độ suy giảm mũ xuất hiện ở các phân phối siêu trơn (supersmooth).

Trong khuôn khổ bán tham số, một bộ ước lượng mới cho hàm phân phối F_X được đề xuất. Phương pháp gồm hai bước: (i) ước lượng trực tiếp phương sai nhiễu σ^2 từ mẫu quan sát nhiễu; và (ii) thay thế giá trị ước lượng này vào sơ đồ giải chập để thu được ước lượng cho F_X. Chúng tôi chứng minh rằng, đối với lớp phân phối $\mathcal{F}_{\alpha,\mathcal{L}}$ được mô tả trong (3), bộ ước lượng này đạt tốc độ hội tụ

$$(\ln n)^{-(\alpha+1/2)},$$

tương ứng với chuẩn minimax đã được thiết lập cho bài toán giải chập Gaussian với phương sai biết trước. Kết quả này cho thấy có thể thích nghi với mức nhiễu chưa biết mà không làm suy giảm hiệu quả thống kê. Khác với các phương pháp trước, cách tiếp cận này ước lượng mức nhiễu hoàn toàn từ mẫu quan sát nhiễu, không đòi hỏi bất kỳ phép đo bổ sung hay lặp lại nào. Đây là công trình đầu tiên cung cấp đảm bảo lý thuyết đầy đủ cho ước lượng CDF trong mô hình nhiễu Gaussian với phương sai chưa biết trong miền ordinary smooth.

Cấu trúc bài báo

Mục 3 trình bày các kết quả lý thuyết chính, bao gồm các tốc độ hội tụ. Mục 4 báo cáo nghiên cứu mô phỏng minh họa hiệu suất thực nghiệm của bộ ước lượng. Mục 5 tổng hợp các chứng minh cho các kết quả chính.

PHƯƠNG PHÁP NGHIÊN CỨU

Khảo sát bài toán dựa vào mô hình bán tham số.

KẾT QUẢ VÀ THẢO LUẬN

Kí hiệu

Với $a, b \in R$, kí hiệu $L^2(a,b)$ là không gian các hàm bình phương khả tích $f: (a,b) \to C$ được trang bị chuẩn

$$|f|_{L^2(a,b)} = \left(\int_a^b |f(t)|^2 \, dt \right)^{1/2}.$$

Khi $a = -\infty, b = +\infty$, ta viết gọn $L^2(R)$. Với $f, g \in L^2(R)$, tích vô hướng được kí hiệu bởi

$$\langle f, g \rangle = \int_R f(t) \, \overline{g(t)} \, dt.$$

Với $f \in L^2(R)$, biến đổi Fourier của f được định nghĩa theo công thức

$$f^{ft}(t) = \int_R e^{itx} f(x) \, dx, \qquad t \in R.$$

Khi X là biến ngẫu nhiên có hàm mật độ f_X, khi đó hàm đặc trưng φ_X được cho bởi công thức

$$\varphi_X(t) = E\left(e^{itX}\right) = f_X^{ft}(t), \qquad t \in R.$$

Tiếp theo, chúng tôi giới thiệu lớp phân phối $\mathcal{F}_{\alpha,\mathcal{L}}$, lớp đóng vai trò khuôn khổ chính để thiết lập tốc độ hội tụ. Kí hiệu F là CDF của một biến ngẫu nhiên với hàm đặc trưng φ. Với $\alpha > -\frac{1}{2}$ và $0 < L < \infty$, định nghĩa

$$\mathcal{F}_{\alpha,\mathcal{L}} := \Big\{ F:\ \varphi(t) = \kappa\,|t|^{-\lambda}[1+\psi(t)],\ \text{với}\ \ \kappa,\lambda > 0,$$
$$\text{và hàm } \psi \text{ thoả mãn } \int_R (1+|t|^{2\alpha})\ |\psi(t)|^2\,dt < L \Big\}.$$

Nhận xét 1. Ở đây, λ đặc trưng cho tốc độ suy giảm đa thức của φ hay mức độ trơn của f_X trong khi ψ mô tả một nhiễu Sobolev bậc α. Trong thực tế, các hàm mật độ trơn thường (ordinary smooth) thường có thể biểu diễn dưới dạng này: thừa số chi phối $|t|^{-\lambda}$ phản ánh hành vi suy giảm chủ đạo của φ, còn phần dư ψ, thỏa mãn điều kiện Sobolev, đóng vai trò hấp thụ các sai khác còn lại trong biểu diễn.

Nhận xét 2. Với $\lambda > \alpha + \frac{1}{2}$, nếu $F \in \mathcal{F}_{\alpha,\mathcal{L}}$, thì hàm đặc trưng φ của nó thỏa mãn một ràng buộc kiểu Sobolev

$$\int_R (1+|t|^{2\alpha})\ |\varphi(t)|^2\,dt < L',$$

với hằng số $L' > 0$ phụ thuộc vào κ, λ , và L. Do đó, lớp $\mathcal{F}_{\alpha,\mathcal{L}}$ có thể được xem như một lớp con của họ phân phối tiêu chuẩn kiểu Sobolev.

$$\mathcal{G}_{\alpha,\mathcal{L}'} := \left\{ F: \int_R (1+|t|^{2\alpha})\ |\varphi(t)|^2\,dt < L' \right\}.$$

Tính duy nhất và khả năng ước lượng của bài toán

Từ (2), hàm đặc trưng của Y thỏa mãn

$$\varphi_Y(t) = \varphi_X(t)\,\varphi_\varepsilon(t), \quad t \in R. \qquad (4)$$

Vậy, bài toán giải chập là ước lượng được nếu tồn tại các bộ ước lượng $\hat{\sigma}$ và $\widehat{F_X}$ hội tụ đến σ và F_X, tương ứng. Để làm rõ tính duy nhất, giả sử tồn tại một cặp (X', ε') khác sao cho

$$Y = X' + \varepsilon', \quad \varepsilon' \sim N\left(0, \sigma'^2\right), \quad F_Y = F_{X'} * f_{\varepsilon'}.$$

Mô hình được gọi là xác định duy nhất nếu từ sự trùng nhau về phân phối của Y có thể suy ra duy nhất cả phân phối tiềm ẩn của X và phương sai của nhiễu. Cụ thể, điều này có nghĩa là

$$F_{X'} * f_{\varepsilon'} = F_X * f_\varepsilon \Rightarrow (F_{X'}, \sigma') = (F_X, \sigma).$$

Vì $\varphi_Y(t)$ có thể được xấp xỉ bởi hàm đặc trưng thực nghiệm $\frac{1}{n}\sum_{j=1}^{n} e^{itY_j}$ khi n đủ lớn, nên về nguyên tắc cả tính duy nhất và khả năng ước lượng đều có thể được phân tích thông qua φ_Y.

Để ước lượng σ^2, nhận thấy từ (4) rằng, với mọi $t \in R$,

$$|\varphi_Y(t)| = |\varphi_X(t)|\,|\varphi_\varepsilon(t)|.$$

Lấy logarit hai vế, sẽ được,

$$\sigma^2 = \frac{2}{t^2}[\ln|\varphi_X(t)| - \ln|\varphi_Y(t)|] \tag{5}$$

Các điều kiện đủ nhằm bảo đảm rằng phương sai nhiễu σ^2 có thể được ước lượng, và cặp (σ, F_X) có thể duy nhất được trong mô hình đã đề xuất.

Định lý 1. Giả sử $F_X, F_{X'} \in \mathcal{F}_{\alpha,\mathcal{L}}$. Khi đó:

(Tính duy nhất của (σ, F_X)). Nếu $F_{X'} * f_{\varepsilon'} = F_X * f_\varepsilon$, thì $(F_{X'}, \sigma') = (F_X, \sigma)$.

(a) (Tính ước lượng được của σ^2). Dưới điều kiện (4), phương sai nhiễu có dạng biểu diễn như sau.

$$\sigma^2 = \lim_{t\to\infty} \frac{8}{9t^2}[2\ln|\varphi_Y(t)| - \ln|\varphi_Y(t/2)| - \ln|\varphi_Y(2t)|].$$

Ước lượng và tốc độ hội tụ

Vì X là một biến ngẫu nhiên liên tục, nên hàm phân phối tích luỹ F_X của nó có thể được biểu diễn dưới dạng (xem ([5])).

$$F_X(x) = \frac{1}{2} - \frac{1}{\pi}\int_0^\infty \frac{1}{t}\,\mathrm{Im}\left(e^{-itx}\varphi_X(t)\right)dt. \tag{6}$$

Trong (6), hàm đặc trưng φ_X là không biết và do đó không thể được sử dụng trực tiếp để tính F_X. Tuy nhiên, bằng cách thay thế φ_X bằng một bộ ước lượng thực nghiệm thích hợp, ta có thể thu được một xấp xỉ khả thi của F_X. Do vậy, bắt đầu bằng việc ước lượng φ_X.

Từ (4), ta có

$$\varphi_X(t) = \frac{\varphi_Y(t)}{\varphi_\varepsilon(t)}, t \in R.$$

Vì $\varepsilon \sim N(0, \sigma^2)$ với phương sai chưa biết $\sigma^2 > 0$, suy ra $\varphi_\varepsilon(t) = \exp\left(-\frac{1}{2}\sigma^2 t^2\right) \neq 0$ với mọi $t \in R$.

Áp dụng khuôn khổ bán tham số, trong đó dạng hàm của phân phối nhiễu (Gaussian) là đã biết, trong khi phương sai σ^2 vẫn chưa biết. Bên cạnh việc ước lượng hàm phân phối F_X, hướng đến ước lượng phương sai nhiễu σ^2. Trước hết, xây dựng một bộ ước lượng cho σ^2, sau đó được đưa vào quá trình ước lượng F_X. Để đảm bảo tính duy nhất và ổn định số, áp đặt điều kiện bị chặn sau đây.

Giả thiết 2.1. Tồn tại σ_*, σ^* sao cho

$$0 < \sigma_* \leq \sigma \leq \sigma^*.$$

Dưới Giả thiết 2.1 và lưu ý rằng $\lim_{|t|\to\infty} \varphi_X(t) = 0$, với $|t|$ đủ lớn và từ (4) suy ra,

$$-\sigma^{*2}t^2 \leq \ln|\varphi_Y(t)| \leq 0.$$

Với $v \in R$, định nghĩa hàm chặt cụt (cut-off) như sau

$$H_t(v) := \max\{e^{-\sigma^{*2}t^2/2}, v\},$$

giá trị này là dương. Các hàm chặt cụt tương tự cũng đã được đề cập trong [14]. Dựa trên đẳng thức chính xác của σ^2 trong [5] và thực tế rằng $f_X \in L^2(R)$, xấp xỉ σ^2 bằng một trung bình tích phân, qua đó làm cơ sở cho việc ước lượng thực nghiệm của nó.

Định nghĩa 1 (Ước lượng phương sai của nhiễu). Đặt

$$\widehat{\varphi_Y}(t) = \frac{1}{n}\sum_{j=1}^{n} e^{itY_j},$$

là hàm phân phối thực nghiệm của Y. Cho $\omega_n > 0$ là một dãy thỏa mãn $\lim_{n\to\infty} \omega_n = +\infty$. Sử dụng hàm cut-off $H_t(.)$ được định nghĩa ở trên, xây dựng bộ ước lượng cho σ^2 như sau

$$\hat{\sigma}^2_{\omega_n} = \frac{1}{\omega_n}\int_{\omega_n}^{2\omega_n} \frac{8}{9t^2}[2\ln H_t(|\widehat{\varphi_Y}(t)|) - \ln H_t(|\widehat{\varphi_Y}(t/2)|) - \ln H_t(|\widehat{\varphi_Y}(2t)|)]dt.$$

Hàm $H_t(v)$ bảo đảm cho tất cả các đối số của hàm logarit luôn dương. Tiếp theo, xây dựng bộ ước lượng cho F_X. Lưu ý rằng

$$\varphi_X(t) = \frac{\varphi_Y(t)}{\varphi_\varepsilon(t)} = e^{\sigma^2t^2/2}\varphi_Y(t).$$

Để bảo đảm rằng phương sai ước lượng nằm trong miền cho phép $[\sigma_*, \sigma^*]$, xây dựng một phiên bản cắt ngọn (truncated) của $\hat{\sigma}_{\omega_n}$ như sau:

$$S_n = S(\hat{\sigma}_{\omega_n}) = \begin{cases} \hat{\sigma}_{\omega_n}, & \text{nếu } \sigma_* \leq \hat{\sigma}_{\omega_n} \leq \sigma^* \\ \sigma^*, & \text{nếu } \hat{\sigma}_{\omega_n} > \sigma^* \\ \sigma_*, & \text{nếu } \hat{\sigma}_{\omega_n} < \sigma_* \end{cases}$$

Thay $\varphi_X(t)$ trong (6) bằng ước lượng thay thế (plug-in estimator)

$$\widehat{\varphi_X}(t) = e^{S_n^2 t^2/2}\widehat{\varphi_Y}(t) = e^{S_n^2 t^2/2}\frac{1}{n}\sum_{j=1}^{n} e^{itY_j},$$

có được công thức ước lượng của F_X.

Định nghĩa 2. (Ước lượng hàm phân phối). Cho $a_n > 0$ là dãy số dương sao cho $\lim_{n\to\infty} a_n = \infty$. Với $x \in R$, có định nghĩa

$$\hat{F}_{X,\omega_n,a_n}(x) = \frac{1}{2} - \frac{1}{n\pi}\sum_{j=1}^{n}\int_0^{a_n} e^{S_n^2 t^2/2}\frac{\sin t\,(Y_j - x)}{t}\,dt.$$

Định lý sau đây giúp chặn trên của các ước lượng của $\hat{\sigma}^2_{\omega_n}$ và $\hat{F}_{X,\omega_n,a_n}(x)$.

Định lý 2. Xét mô hình (1) dưới Giả thiết 2.1, giả sử $\lambda > \alpha + 1/2$. Định nghĩa

$$\mathcal{A} = [\sigma_*, \sigma^*] \times \mathcal{F}_{\alpha,\mathcal{L}}.$$

Khi đó, các ước lượng được đề xuất thỏa mãn các tốc độ hội tụ sau đây.

(a) Ước lượng phương sai

Với mọi hằng số $0 < c_\omega < 1/(2\sigma^*)$ và $\omega_n = c_\omega\sqrt{\ln n}$, tồn tại hằng số $C > 0$, độc lập với n, sao cho

$$\sup_{(\sigma,F_X)\in\mathcal{A}} E\left[\left(\hat{\sigma}^2_{\omega_n} - \sigma^2\right)^2\right] \le C(\ln n)^{-(\alpha+5/2)}.$$

(b) Ước lượng hàm phân phối

Với mọi giá trị cố định $x \in R$, let $0 < c_{\sigma^*} < 1/\left(\sqrt{2}\sigma^*\right)$ và $a_n = c_{\sigma^*}\sqrt{\ln n}$. Tồn tại hằng số $C > 0$, độc lập với n, sao cho

$$\sup_{(\sigma,F_X)\in\mathcal{A}} E\left[\left(\hat{F}_{X,\omega_n,a_n}(x) - F_X(x)\right)^2\right] \le C(\ln n)^{-(\alpha+1/2)}.$$

Nhận xét 3. Tốc độ hội tụ của ước lượng phương sai trong Định lý 2(a) trùng với tốc độ minimax được thiết lập trong [14, Định lý 2.2]. Tuy nhiên, phân tích này được thực hiện trên không gian tham số hẹp hơn $\mathcal{A} \subset [\sigma_*, \sigma^*] \times \mathcal{G}_{\alpha,\mathcal{L}'}$, vốn đặt ra các ràng buộc chặt chẽ hơn về độ trơn của lớp phân phối mục tiêu. Kết quả này cho thấy rằng, ngay cả trong một bối cảnh hạn chế hơn, ước lượng được đề xuất vẫn đạt cùng bậc tối ưu như trong khuôn khổ minimax tổng quát.

Nhận xét 4. Tốc độ hội tụ dạng logarit thu được trong Định lý 2(b) phù hợp với các kết quả đã biết trước đây trong bài toán giải chập:

- Khi phân phối nhiễu đã biết, [2] thiết lập tốc độ $(\ln n)^{-(2\alpha+1)/\gamma}$ đối với trường hợp sai số siêu trơn; trong trường hợp nhiễu Gauss ($\gamma = 2$), tốc độ này trở thành $(\ln n)^{-(\alpha+1/2)}$), trùng khớp với kết quả của chúng tôi.

- Khi phân phối nhiễu chưa biết nhưng có mẫu phụ trợ, [8] thu được tốc độ $(\ln(\min\{n,m\}))^{-(2\alpha+1)/\gamma}$, và tốc độ này cũng rút gọn thành $(\ln n)^{-(\alpha+1/2)}$ khi $m = n$ và $\gamma = 2$.

Ngược lại, trong thiết lập này, không giả định phân phối nhiễu đã biết, cũng không sử dụng mẫu phụ trợ. Mặc dù làm việc trên lớp hàm hẹp hơn $\mathcal{F}\alpha, \mathcal{L} \subset \mathcal{G}\alpha, \mathcal{L}'$, Định lý 2(b) vẫn đạt được cùng tốc độ hội tụ dạng logarit, cho thấy tính ổn định và hiệu quả của thủ tục ước lượng được đề xuất.

NGHIÊN CỨU MÔ PHỎNG

Các thí nghiệm số trong mô hình giải chập cộng tiêu chuẩn, theo tinh thần của [2]. Trong toàn bộ phần này, mật độ mục tiêu f_X được giả định thuộc lớp trơn thông thường $\mathcal{F}_{\alpha,\mathcal{L}}$, và xét

$$\text{M1} \quad X \sim \text{Laplace}\left(0, 2^{-1/2}\right),$$

Có hàm đặc trưng tương ứng là $\varphi_X(t) = (1 + t^2/2)^{-1}$. Ký hiệu σ_X là độ lệch chuẩn của biến ngẫu nhiên X. Để khảo sát ảnh hưởng của cường độ nhiễu, xét hai phân phối nhiễu Gauss với tỷ lệ nhiễu–tín hiệu $\sigma^2/\sigma_X^2 \in \{0.2, 0.5\}$:

$$\text{(N1)} \quad \varepsilon \sim \mathcal{N}(0, 1/5), \qquad \text{(N2)} \quad \varepsilon \sim \mathcal{N}(0, 1/2).$$

Ước lượng hàm phân phối

Theo Giả thiết 2.1, đặt

$$\sigma_* = 10^{-6}, \qquad \sigma^* = 0.7\,\text{sd}(Y),$$

trong đó $\text{sd}(Y)$ là độ lệch chuẩn mẫu của dữ liệu quan sát. Cận dưới σ_* giúp ngăn ngừa mất ổn định số khi mức nhiễu nhỏ, trong khi cận trên σ^* phản ánh thực tế rằng mức nhiễu hữu hạn thường ở mức vừa phải so với tổng độ biến thiên của Y.

Để triển khai ước lượng chặt cụt S_n cho phương sai nhiễu, lựa chọn

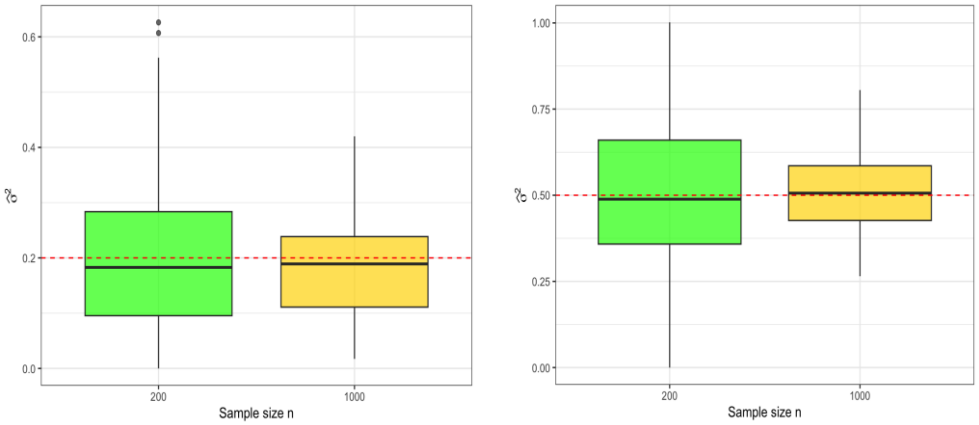
$$c_\omega = \frac{1}{(\max\{2.01,\ 1.85\,\text{Var}(Y)\})\sigma^*},$$

phù hợp với Định lý 2 và được điều chỉnh dựa trên nhiều mô phỏng sơ bộ. Việc lựa chọn c_ω có ảnh hưởng đáng kể đến tốc độ hội tụ của ước lượng phương sai; việc xây dựng một quy tắc lựa chọn hoàn toàn dựa trên dữ liệu vẫn là một hướng nghiên cứu cho tương lai.

Các mô phỏng được tiến hành với kích thước mẫu $n \in \{200, 1000\}$ và 100 lần lặp Monte Carlo. Trong mỗi lần lặp, chúng tôi sinh mẫu i.i.d.

$$X_i \sim (M1), \qquad \varepsilon_i \sim (N1 \text{ or } N2), \qquad Y_i = X_i + \varepsilon_i, \quad i = 1, \dots, n.$$

Các biểu đồ hộp (box plot) của ước lượng phương sai nhiễu $\hat{\sigma}^2$ được trình bày trong Hình 1.



Hình 1. Biểu đồ hộp của ước lượng phương sai nhiễu σ^2 với kích thước mẫu $n \in \{200,1000\}$ trong hai kịch bản: $\sigma^2 = 0.2$ (trái) và $\sigma^2 = 0.5$ (phải).

Việc ước lượng chính xác σ^2 là hết sức quan trọng, bởi nó được sử dụng trong bộ ước lượng thay thế có cắt cụt đối với $\varphi_X(t)$, và từ đó quyết định hiệu quả của bộ ước lượng hàm phân phối tích luỹ $\widehat{F_X}$.

Ước lượng hàm phân phối

Tiếp theo, đánh giá độ chính xác của bộ ước lượng hàm CDF $\widehat{F_X}$ đề xuất cho mô hình Laplace (M1).

Các mô phỏng được thực hiện với kích thước mẫu $n \in 200,1000$ và hai trường hợp nhiễu $\sigma^2 = 0{,}2$ và $\sigma^2 = 0{,}5$. Dữ liệu được tạo ra hoàn toàn tương tự như mô tả trong tiểu mục trước.

Trước hết, phương sai nhiễu được ước lượng bằng bộ ước lượng cắt cụt S_n (Định nghĩa 1). Giá trị ước lượng này sau đó được sử dụng để xây dựng bộ ước lượng thế chỗ của $\varphi_X(t)$, từ đó suy ra bộ ước lượng $\widehat{F_X}$.

$$\widehat{\varphi_X}(t) = \exp(S_n^2 t^2/2)\,\widehat{\varphi_Y}(t),$$

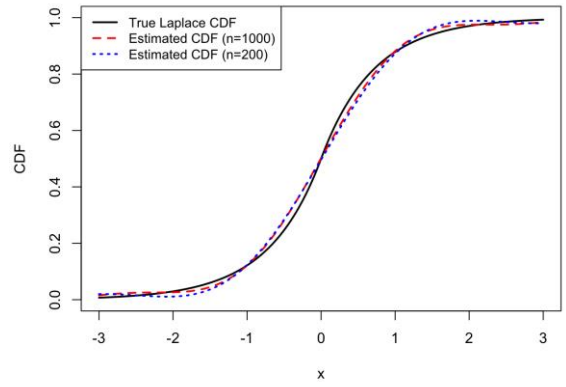
giá trị này là cơ sở để xây dựng bộ ước lượng CDF $\widehat{F_X}$.

Để triển khai ngưỡng cắt (cut-off) trong Định lý 2, thiết lập

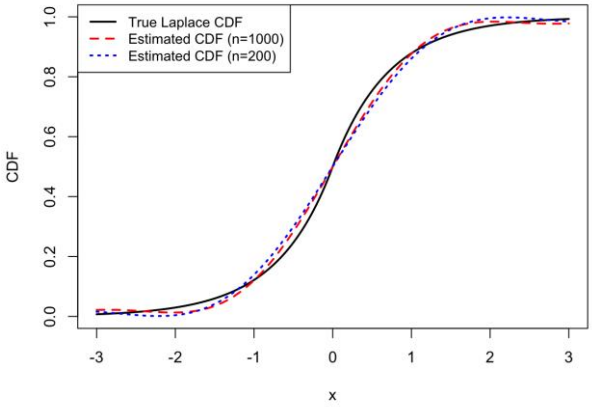
$$c_{\sigma^*} = \frac{1}{1.42\,\sigma^*}, \qquad a_n = c_{\sigma^*}\sqrt{\ln n}.$$

Tương tự như với c_ω, việc lựa chọn c_{σ^*} một cách cẩn trọng là yếu tố then chốt để thu được các ước lượng ổn định. Trong nghiên cứu này, c_{σ^*} được lựa chọn dựa trên định hướng từ định lý và kinh nghiệm mô phỏng phong phú. Việc phát triển một quy tắc lựa chọn hoàn toàn dựa trên dữ liệu và có khả năng thích ứng cho c_{σ^*} vẫn là một hướng nghiên cứu mở trong tương lai.

Hình 2 và Hình 3 so sánh CDF thực (đường liền) với các CDF ước lượng (đường đứt đoạn) trong 100 lần lặp Monte Carlo.



Hình 2. So sánh hàm CDF của mô hình Laplace thực (đường liền) với phương sai nhiễu $\sigma^2 = 0.2$ và các CDF ước lượng (đường đứt đoạn) tương ứng với hai kích thước mẫu $n \in \{200,1000\}$.



Hình 3. So sánh hàm CDF của mô hình Laplace thực (đường liền) với phương sai nhiễu $\sigma^2 = 0.5$ và các CDF ước lượng (đường đứt đoạn) tương ứng với hai kích thước mẫu $n \in \{200,1000\}$.

Độ chính xác điểm tại các phân vị tiêu biểu

Để bổ sung cho các kết quả đồ thị trong Hình 2 và Hình 3, tiếp tục khảo sát độ chính xác theo từng điểm của bộ ước lượng hàm CDF đề xuất $\widehat{F_X}$ tại một số phân vị đại diện.

$$F_X(x_0) \in \{0.1, 0.25, 0.5, 0.75, 0.9\}.$$

Các phân vị này bao phủ cả vùng trung tâm lẫn vùng đuôi của phân phối Laplace, nơi bài toán giải chập thường gặp nhiều thách thức hơn do tỷ lệ tín hiệu trên nhiễu thay đổi.

Đối với mỗi thiết lập, sai số bình phương trung bình (MSE) theo kinh nghiệm được tính trên 100 lần lặp Monte Carlo cho cả hai cỡ mẫu (n = 200) và (n = 1000), với hai mức nhiễu $\sigma^2 = 0.2$ và $\sigma^2 = 0.5$. Để thuận tiện cho việc so sánh, các giá trị MSE báo cáo được nhân với 100.

Các kết quả được tóm tắt trong Bảng 1. Như dự đoán, cỡ mẫu lớn hơn dẫn đến độ chính xác cao hơn, trong khi mức nhiễu lớn hơn làm gia tăng MSE, đặc biệt tại các phân vị thuộc vùng đuôi phía trên, nơi tỷ lệ tín hiệu trên nhiễu thấp hơn.

Bảng 1. Sai số bình phương trung bình theo kinh nghiệm của $\hat{F}_{X,\omega_n,a_n}(x)$ tại các phân vị được chọn đối với $X \sim \text{Laplace}(0,2^{-1/2})$ dưới hai mức nhiễu và hai cỡ mẫu, dựa trên 100 lần chạy Monte Carlo.

	$F_X(x_0)$	MSE (× 100)				
		0.1	0.25	0.5	0.75	0.9
$X \sim \text{Laplace}(0, 2^{-1/2})$ Noise: $\varepsilon \sim \mathcal{N}(0, \sigma^2)$	Case 1: $\sigma^2 = 0.2$					
	n=200	0.04736	0.29943	0.09251	0.31102	0.05223
	n=1000	0.03103	0.15434	0.02349	0.15973	0.02679
	Case 1: $\sigma^2 = 0.5$					
	n=200	0.07278	0.37647	0.13719	0.48934	0.06608
	n=1000	0.03559	0.19864	0.02542	0.20909	0.03918

PHẦN CHỨNG MINH

Chứng minh Định lý 1

(a) Tính duy nhất (Identifiability)

Từ công thức

$$\varphi_X(t) = \kappa|t|^{-\lambda}[1 + \psi(t)], \quad \text{và} \quad \varphi_Y(t) = \varphi_X(t)e^{-\sigma^2t^2/2},$$

thu được

$$\ln|\varphi_Y(t)| = \ln\kappa - \lambda\ln|t| + \ln|1 + \psi(t)| - \frac{\sigma^2t^2}{2}.$$

Thay t bởi $2t$ và $t/2$, tương ứng, được công thức

$$\ln|\varphi_Y(2t)| = \ln\kappa - \lambda\ln|2t| + \ln|1 + \psi(2t)| - 2\sigma^2t^2,$$

$$\ln|\varphi_Y(t/2)| = \ln\kappa - \lambda\ln\left|\frac{t}{2}\right| + \ln|1 + \psi(t/2)| - \frac{\sigma^2t^2}{8}.$$

Việc trừ hai biểu thức này cho

$$2\ln|\varphi_Y(t)| - \ln|\varphi_Y(t/2)| - \ln|\varphi_Y(2t)| = \frac{9\sigma^2t^2}{8} + 2\ln|1 + \psi(t)| - \ln|1 + \psi(t/2)| - \ln|1 + \psi(2t)|,$$

$$\ln|\varphi_Y(t)| - \ln|\varphi_Y(2t)| = \lambda\ln 2 + \frac{3\sigma^2t^2}{2} + \ln|1 + \psi(t)| - \ln|1 + \psi(2t)|.$$

Vì $\psi(t) \to 0$ khi $t \to \infty$, sẽ có

$$\begin{aligned}
\sigma^2 &= \lim_{t\to\infty}\frac{8}{9t^2}[2\ln|\varphi_Y(t)| - \ln|\varphi_Y(t/2)| - \ln|\varphi_Y(2t)|], \\
\lambda &= \lim_{t\to\infty}\frac{1}{\ln 2}\left[\ln|\varphi_Y(t)| - \ln|\varphi_Y(2t)| - \frac{3\sigma^2t^2}{2}\right], \\
\ln\kappa &= \lim_{t\to\infty}\left[\ln|\varphi_Y(t)| + \lambda\ln|t| + \frac{\sigma^2t^2}{2}\right].
\end{aligned}$$

Do đó, bộ ba $(\kappa, \lambda, \sigma^2)$ được xác định duy nhất bởi φ_Y. Khi các tham số này đã được biết, phần dư $\psi(t)$ cũng được xác định duy nhất dưới dạng

$$\psi(t) = \frac{\varphi_Y(t)}{\kappa|t|^{-\lambda}e^{-\sigma^2t^2/2}} - 1.$$

Do đó, hàm đặc trưng $\varphi_X(t)$, và vì thế cả phân phối F_X, đều được xác định duy nhất. Đặc biệt, nếu một phép phân rã khác $Y = X' + \varepsilon'$ thỏa mãn $\varphi_{Y'} = \varphi_Y$, thì buộc phải có

$$(\kappa', \lambda', \sigma'^2) = (\kappa, \lambda, \sigma^2), \quad \psi'(t) = \psi(t),$$

Do đó $F_{X'} = F_X$. Điều này chứng minh tính xác định (identifiability) của cặp (F_X, σ).

(b) Tính ước lượng được

Biểu thức thu được trong phần (a) dẫn trực tiếp đến

$$\sigma^2 = \lim_{t\to\infty} \frac{8}{9t^2}[2\ln|\varphi_Y(t)| - \ln|\varphi_Y(t/2)| - \ln|\varphi_Y(2t)|],$$

qua đó cung cấp một dạng hàm có thể ước lượng được của phương sai nhiễu. Điều này hoàn tất chứng minh.

Chứng minh Định lý 2

Chứng minh phần (a)

Ta có,

$$\begin{aligned}\hat{\sigma}^2_{\omega_n} - \sigma^2 &= \frac{1}{\omega_n}\int_{\omega_n}^{2\omega_n} \frac{8}{9t^2}[2\ln H_t(|\widehat{\varphi_Y}(t)|) - \ln H_t(|\widehat{\varphi_Y}(t/2)|) \\ &\qquad - \ln H_t(|\widehat{\varphi_Y}(2t)|)]dt - \sigma^2 \\ &= \frac{1}{\omega_n}\int_{\omega_n}^{2\omega_n}[Q_1(t) + Q_2(t) + Q_3(t) + Q_4(t)]dt,\end{aligned}$$

trong đó,

$$\begin{aligned}Q_1(t) &= \frac{8}{9t^2}[2\ln H_t(|\varphi_Y(t)|) - \ln H_t(|\varphi_Y(t/2)|) - \ln_t(|\varphi_Y(2t)|)] - \sigma^2,\\ Q_2(t) &= \frac{16}{9t^2}[\ln H_t(|\widehat{\varphi_Y}(t)|) - \ln H_t(|\varphi_Y(t)|)],\\ Q_3(t) &= \frac{8}{9t^2}[\ln H_t(|\varphi_Y(t/2)|) - \ln H_t(|\widehat{\varphi_Y}(t/2)|)],\\ Q_4(t) &= \frac{8}{9t^2}[\ln H_t(|\varphi_Y(2t)|) - \ln H_t(|\widehat{\varphi_Y}(2t)|)].\end{aligned}$$

Bước 1. Chặn cho $Q_1(t)$

Đối với t đủ lớn, tính trực tiếp sẽ cho

$$\begin{aligned}Q_1(t) &= \tfrac{8}{9t^2}[2\ln H_t(|\varphi_Y(t)|) - \ln H_t(|\varphi_Y(t/2)|) - \ln_t(|\varphi_Y(2t)|)] - \sigma^2\\ &= \tfrac{8}{9t^2}\left[2\ln|\varphi_X(t)| - \ln|\varphi_X(t/2)| - \ln|\varphi_X(2t)| + \tfrac{9t^2}{8}\sigma^2\right] - \sigma^2\\ &= \tfrac{8}{9t^2}[2\ln|1+\psi(t)| - \ln|1+\psi(t/2)| - \ln|1+\psi(2t)|].\end{aligned}$$

Với n lớn, sẽ có

$$\left|\frac{1}{\omega_n}\int_{\omega_n}^{2\omega_n} Q_1(t)dt\right|^2 \le \left|\frac{8}{9\omega_n}\int_{\omega_n}^{2\omega_n}\frac{1}{t^2}[2|\psi(t)| - |\psi(t/2)| - |\psi(2t)|]dt\right|^2$$

$$\leq \frac{64}{81\omega_n^2}\int_{\omega_n}^{2\omega_n} t^{2\alpha} \cdot 3[4|\psi(t)|^2 + |\psi(t/2)|^2 + |\psi(2t)|^2]dt \int_{\omega_n}^{2\omega_n} \frac{1}{t^{2\alpha+4}}dt \leq \frac{C_1}{\omega_n^{2\alpha+5}}. \quad (7)$$

Bước 2. Chặn cho $Q_2(t)$.

Do $e^{-\sigma^{*2}t^2/2} \leq H_t(v)$, ta có

$$|\ln H_t(u) - \ln H_t(v)| \leq e^{\sigma^{*2}t^2/2}|H_t(u) - H_t(v)| \leq e^{\sigma^{*2}t^2/2}|u - v|, \quad \forall u, v \in R.$$

Do đó,

$$\left|\frac{1}{\omega_n}\int_{\omega_n}^{2\omega_n} Q_2(t)dt\right| = \left|\frac{16}{9\omega_n}\int_{\omega_n}^{2\omega_n} \frac{1}{t^2}[\ln H_t(|\widehat{\varphi_Y}(t)|) - \ln H_t(|E\widehat{\varphi_Y}(t)|)]dt\right|$$

$$\leq \frac{16}{9\omega_n}\int_{\omega_n}^{2\omega_n} \frac{1}{t^2} e^{\sigma^{*2}t^2/2}|\widehat{\varphi_Y}(t) - E\widehat{\varphi_Y}(t)|dt \leq \frac{16}{9\omega_n}\left(\int_{\omega_n}^{2\omega_n} \frac{1}{t^4} e^{\sigma^{*2}t^2}dt\right)^{1/2}\left(\int_{\omega_n}^{2\omega_n} |\widehat{\varphi_Y}(t) - E\widehat{\varphi_Y}(t)|^2 dt\right)^{1/2}.$$

Hơn nữa, vì

$$\mathrm{Var}\widehat{\varphi_Y}(t) = \frac{1}{n}\mathrm{Var}\left(e^{itY_1}\right) \leq \frac{1}{n}, \quad \forall t \in R,$$

suy ra rằng

$$E\left|\frac{1}{\omega_n}\int_{\omega_n}^{2\omega_n} Q_2(t)dt\right|^2 \leq \frac{C_2}{\omega_n^2}\int_{\omega_n}^{2\omega_n} \frac{1}{t^4} e^{\sigma^{*2}t^2}dt \int_{\omega_n}^{2\omega_n} \mathrm{Var}\,\widehat{\varphi_Y}(t)dt \leq \frac{C_2 e^{4\sigma^{*2}\omega_n^2}}{n\omega_n^4}. \quad (8)$$

Bước 3. Chặn cho $Q_3(t)$ và $Q_4(t)$.

Bằng các lập luận tương tự, chúng tôi suy ra rằng

$$E\left|\frac{1}{\omega_n}\int_{\omega_n}^{2\omega_n} Q_3(t)dt\right|^2 \leq \frac{C_3\, e^{4\sigma^{*2}}\,\omega_n^2}{n\,\omega_n^4}, \quad E\left|\frac{1}{\omega_n}\int_{\omega_n}^{2\omega_n} Q_4(t)dt\right|^2 \leq \frac{C_4 e^{4\sigma^{*2}\omega_n^2}}{n\omega_n^4}. \quad (9)$$

Bước 4. Kết hợp tất cả các chặn lại.

Kết hợp (7), (8) và (9), thu được

$$\mathrm{E}\left(\hat{\sigma}^2_{\omega_n} - \sigma^2\right)^2 \le 4\sum_{k=1}^{4}\mathrm{E}\left(\left|\frac{1}{\omega_n}\int_{\omega_n}^{2\omega_n} Q_k(t)\,dt\right|^2\right) \le \frac{4C_1}{\omega_n^{2\alpha+5}} + 4(C_2+C_3+C_4)\frac{e^{4\sigma^{*2}\omega_n^2}}{n\omega_n^4}.$$

Sử dụng giả thiết $0 < c_\omega < 1/(2\sigma^*)$, sauy ra $1 - 4\sigma^{*2}c_\omega^2 > 0$. Đặt $\omega_n = c_{\omega\sqrt{\ln n}}$, sẽ có

$$\frac{e^{4\sigma^{*2}\omega_n^2}}{n} = \frac{1}{n^{1-4\sigma^{*2}c_\omega^2}},$$

và do đó,

$$\mathrm{E}\left[\left(\hat{\sigma}^2_{\omega_n} - \sigma^2\right)^2\right] \le C(\ln n)^{-(\alpha+5/2)}.$$

Chứng minh (b)

Từ công thức

$$\hat{F}_{X,\omega_n,a_n}(x) = \frac{1}{2} - \frac{1}{n\pi}\sum_{j=1}^{n}\int_0^{a_n} e^{s_n^2 t^2/2}\frac{\sin t\left(Y_j - x\right)}{t}\,dt,$$

Sẽ có

$$\mathrm{E}\left(\hat{F}_{X,\omega_n,a_n}(x) - F_X(x)\right)^2 = \mathrm{Bias}^2\hat{F}_{X,\omega_n,a_n}(x) + \mathrm{Var}\hat{F}_{X,\omega_n,a_n}(x),$$

trong đó,

$$\mathrm{Bias}\,\hat{F}_{X,\omega_n,a_n}(x) = -\frac{1}{\pi}\int_{a_n}^{\infty}\frac{1}{t}\mathrm{Im}\left[e^{-itx}\varphi_X(t)\right]dt,$$

$$\mathrm{Var}\,\hat{F}_{X,\omega_n,a_n}(x) = \frac{1}{n}\mathrm{Var}\left(\frac{1}{\pi}\int_0^{a_n} e^{s_n^2 t^2/2}\frac{\sin t\left(Y_1 - x\right)}{t}\,dt\right).$$

Áp dụng bất đẳng thức Cauchy-Schwarz,

$$\left|\mathrm{Bias}\,\hat{F}_{X,\omega_n,a_n}(x)\right| \le \frac{1}{\pi}\int_{a_n}^{\infty}\left|\frac{\varphi_X(t)}{t}\right|dt \le \frac{1}{\pi}\left(\int_{a_n}^{\infty}\frac{1}{t^{2\alpha+2}}dt\right)^{1/2}\left(\int_{a_n}^{\infty}t^{2\alpha}|\varphi_X(t)|^2\,dt\right)^{1/2} \le \frac{(L')^{1/2}}{\pi\sqrt{2\alpha+1}\,a_n^{\alpha+1/2}}.$$

Mặt khác,

$$\frac{1}{n}\mathrm{Var}\left(\frac{1}{\pi}\int_0^{a_n} e^{S_n^2 t^2/2}\frac{\sin t\,(Y_1 - x)}{t}\,dt\right) \le \frac{1}{n}E\left(\frac{1}{\pi}\int_0^{a_n} e^{S_n^2 t^2/2}\frac{\sin t\,(Y_1 - x)}{t}\,dt\right)^2$$

$$\le \frac{1}{n2\pi}E\left(\frac{1}{\pi}\int_0^{a_n}\frac{e^{S_n^2 t^2}}{S_n}\int_{-\infty}^{+\infty}\frac{\sin[t(Y_1 - x + p)]\,e^{-p^2/(2S_n^2)}}{t}\,dpdt\right)^2.$$

Áp dụng định lý Fubini,

$$\frac{1}{\pi}\int_0^{a_n}\frac{e^{S_n^2 t^2}}{S_n}\int_{-\infty}^{+\infty}\frac{\sin[t(Y_1 - x + p)]\,e^{-p^2/(2S_n^2)}}{t}\,dpdt$$

$$\le \frac{1}{\pi S_n}\int_{-\infty}^{+\infty} e^{-p^2/(2S_n^2)}\left(\int_0^{a_n}\frac{\sin[t(Y_1 - x + p)]\,e^{S_n^2 t^2}}{t}\,dt\right)dp$$

$$\le \frac{1}{\pi S_n}e^{S_n^2 a_n^2}\int_{-\infty}^{+\infty} e^{-p^2/(2S_n^2)}\left(\int_0^{a_n}\frac{\sin[t(Y_1 - x + p)]}{t}\,dt\right)dp \le e^{S_n^2 a_n^2}\cdot\frac{\sqrt{2\pi}}{2}.$$

Suy ra

$$E\left(\frac{1}{\pi}\int_0^{a_n}\frac{e^{S_n^2 t^2}}{S_n}\int_{-\infty}^{+\infty}\frac{\sin[t(Y_1 - x - p)]\,e^{-p^2/(2S_n^2)}}{t}\,dpdt\right)^2 \le \frac{\sqrt{2\pi}}{2}e^{2\sigma^{*2}a_n^2}.$$

Do đó,

$$\mathrm{Var}\,\hat{F}_{X,\omega_n,a_n}(x) \le \frac{e^{2\sigma^{*2}a_n^2}}{n\,2\sqrt{2\pi}}.$$

Kết hợp bất đẳng thức cho Bias $\hat{F}_{X,\omega_n,a_n}(x)$ và $\mathrm{Var}\,\hat{F}_{X,\omega_n,a_n}(x)$, ta có

$$E\left(\hat{F}_{X,\omega_n,a_n}(x) - F_X(x)\right)^2 \le C_1 a_n^{-(2\alpha+1)} + C_2\frac{e^{2\sigma^{*2}a_n^2}}{n}.$$

Sử dụng giả thiết $0 < c_{\sigma^*} < \frac{1}{\sqrt{2}\sigma^*}$, ta có $1 - 2\sigma^{*2}c_{\sigma^*}^2 > 0$. Chọn $a_n = c_{\sigma^*\sqrt{\ln n}}$ khi đó

$$\frac{e^{2\sigma^{*2}a_n^2}}{n} = n^{-\left(1-2\sigma^{*2}c_{\sigma^*}^2\right)}.$$

Suy ra

$$E\left(\hat{F}_{X,\omega_n,a_n}(x) - F_X(x)\right)^2 \le C(\ln n)^{-(\alpha+1/2)}.$$

TUYÊN BỐ XUNG ĐỘT LỢI ÍCH

Các tác giả đồng ý không có bất kỳ xung đột lợi ích nào liên quan đến các kết quả đã công bố.

ĐÓNG GÓP CỦA CÁC TÁC GIẢ

Nguyễn Hoàng Thanh viết bản thảo.

Thái Phúc Hưng chạy số mô phỏng và hoàn chỉnh bản thảo.

TÀI LIỆU THAM KHẢO

[1] Dattner, I., Goldenshluger, A., Juditsky, A. (2011). On deconvolution of distribution functions. *The Annals of Statistics*, 39(5), 2477–2501.

[2] Dattner, I., Reiser, B. (2013). Estimation of distribution functions in measurement error models. *Journal of Statistical Planning and Inference*, 143 (3), 479–493.

[3] Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, 19(3), 1257–1272.

[4] Gaffey, R. (1959). A consistent estimator of a component of a convolution. *The Annals of Mathematical Statistics*, 30(1), 198–205.

[5] Gil-Peláez, J. (1951). Note on the inversion theorem. *Biometrika*, 38(3-4), 481-482.

[6] Lee, M. , Hall, P. (2013). Deconvolution estimation of mixture distributions with boundaries. *Electronic Journal of Statistics*, 7, 323–341.

[7] Matias, C. (2002). Semiparametric deconvolution with unknown noise variance. ESAIM: *Probability and Statistics*, 6, 271–292.

[8] Phuong, C. X. (2020). Deconvolution of cumulative distribution function with unknown noise distribution. *Acta Applicandae Mathematicae*, 170(1), 483–514.

[9] Trang, B. T., Thuy, L. H.,, Phuong, C. X. (2023). Nonparametric deconvolution of cumulative distribution function from repeated observation with unknown noise distribution. *Communications in Statistics - Theory and Methods*, 53(24), 8787–8818.

[10] Trong, D. D., Hung, T. P. (2023). Parameter estimation for diffusion process from perturbed discrete observations. *Communications in Statistics - Simulation and Computation*, 52(3), 925–944.

[11] Trong, D. D., ,Hung, T. P. (2024). Deconvolution of $P(X_t < Y_t)$ for stationary processes with supersmooth error distributions. *Statistics*, 58(6), 1463–1487.

[12] Trong, D. D., Nguyen, T. T. Q., Phuong, C. X. (2017). Deconvolution of $P(X < Y)$ with compactly supported error densities. *Statistics & Probability Letters*, 123, 171–176.

[13] Trong, D. D., , Phuong, C. X. (2019). Deconvolution of a cumulative distribution function with some non-standard noise densities. *Vietnam Journal of Mathematics*, 47(2), 327–353.

[14] Trong, D. D., Thanh, N. H., Minh, N. D., ,Lan, N. N. (2021). Density estimation of a mixture distribution with unknown point-mass and normal error. *Journal of Statistical Planning and Inference*, 215, 268–288.

VNUHCM Journal of Advanced Research and Natural Sciences 2026, 10(2): 3654-3672

Open Access Full Text Article

Research article

A CUMULATIVE DISTRIBUTION FUNCTION OF A MIXTURE MODEL WITH NORMAL ERROR

Thai Phuc Hung [1,2,3], Nguyen Hoang Thanh [1,2,*]

Use your smartphone to scan this QR code and download this article

ABSTRACT

The deconvolution problem in a classical measurement error model was investigated where contaminated observations were given by $Y_j = X_j + \varepsilon_j,\ \varepsilon_j \sim N(0, \sigma^2)$, with unknown noise variance σ^2. The objective was to recover the cumulative distribution function F_X of the latent target variable X from a single sample of noisy observations. In contrast to the majority of existing literature which primarily focused on density estimation, a direct inference on F_X, a functionally rich object with broad applications in risk assessment, hypothesis testing was aimed and classified. A two-step semiparametric estimation procedure was proposed. First, the noise variance σ^2 was directly estimated from the contaminated data. Second, the resulting estimator was plugged into a deconvolution object with broad applications in risk assessment, hypothesis testing was testetd. A two-based reconstruction method to obtain an estimator of F_X. when F_X belonged to an ordinary smooth class, characterized by the polynomial decay of the characteristic function within a Sobolev-type regularity framework, namely $(\varphi_X(t)| \asymp (1+(t|)^{-\alpha}$, the proposed estimator attaining the convergence rate $(lnn)^{-(\alpha+1/2)}$*was established* . This rate coincided with the minimax-optimal rate achieved in the case of known noise variance, demonstrating a full adaptivity with respect to the unknown noise level without any loss of statistical efficiency. Importantly, this approach did not require any auxiliary sample, such as pure noise observations or repeated measurements, making it particularly suitable for experimental settings with limited data availability. To the best of our knowledge, this was the first work providing a complete theoretical guarantee for distribution function estimation in the Gaussian deconvolution model with unknown variance in the ordinary smooth regime.

Key words: Deconvolution, Cumulative distribution function, Gaussian noise, Semiparametric estimation

[1]*Faculty of Mathematics and Computer Science, University of Science, Ho Chi Minh City, Vietnam*

[2]*Vietnam National University, Ho Chi Minh City, Vietnam*

[3]*Faculty of Basics, Soc Trang Community College, Can Tho, Vietnam*

Correspondence

Nguyen Hoang Thanh , Faculty of Mathematics and Computer Science, University of Science, Ho Chi Minh City, Vietnam

Vietnam National University, Ho Chi Minh City, Vietnam

Email: nguyenhoangthanh1010@gmail.com

History

- Received: 09-12-2025
- Revised: 27-02-2026
- Accepted: 24-06-2026
- Published Online:28-06-2026

DOI : https://doi.org/10.32508/vnuhcmj-arns.v10i2.1504

Check for updates

Copyright

© VNUHCM Journal. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Cite this article : Thai P H, Nguyen H T. **A CUMULATIVE DISTRIBUTION FUNCTION OF A MIXTURE MODEL WITHNORMAL ERROR**. *VNUHCM J. Adv. Res. Nat. Sci.2026;10(2):3654-3672.*