Statystyczne problemy odwrotne

Zbigniew Szkutnik, AGH Kraków

Notatki do wykładu podczas
XXXII Konferencji Statystyka Matematyczna, Wisła 2006
(drobne zmiany: V 2015, VII 2016, X 2017)

1 Wstęp

Przez statystyczny (lub stochastyczny) problem odwrotny będziemy rozumieć problem estymacji (zwykle nieparametrycznej) na podstawie obserwacji zniekształconych i zaszumionych. Mechanizm zniekształcania, lub transformacji danych będziemy opisywać pewnym operatorem, a probabilistyczny model szumu będzie zależał od sposobu obserwacji danych.

1.1 Niestochastyczny problem odwrotny

Niech $\mathcal{X}$ i $\mathcal{Y}$ będą przestrzeniami Banacha, a $\mathcal{K} : \mathcal{X} \longrightarrow \mathcal{Y}$ operatorem liniowym. Ponadto, niech $D \subset \mathcal{X}$ oraz $R \subset \mathcal{Y}$ będą ustalonymi podzbiorami - niekoniecznie podprzestrzeniami liniowymi. Problemem odwrotnym nazywamy problem sformułowany następująco:

Dla $y \in R$ znaleźć $x \in D$ taki, że

$$\mathcal{K}x = y, \tag{1}$$

czyli problem rozwiązania równania operatorowego z określoną dziedziną i przeciwdziedziną. Mówimy, że taki problem jest *dobrze postawiony w sensie Hadamarda* gdy (Hadamard, 1932):

1. dla każdego $y \in R$ ma jednoznaczne rozwiązanie w D
2. rozwiązanie zależy w sposób ciągły od prawej strony, tzn. dla pewnej funkcji $w(\varepsilon)$ takiej, że $w(\varepsilon) \to 0$ gdy $\varepsilon \to 0$, mamy

$$\forall \varepsilon > 0 \; \forall x_1, x_2 \in D \;\; \|\mathcal{K}x_1 - \mathcal{K}x_2\| \leq \varepsilon \implies \|x_1 - x_2\| \leq w(\varepsilon).$$

W przeciwnym razie mówimy, że problem jest *źle postawiony* (ang. ill-posed inverse problem). Niedokładna znajomość y, a także ograniczona dokładność rachunków numerycznych rodzą wtedy problemy z wyznaczeniem x i konieczne są specjalne zabiegi stabilizujące rozwiązanie - tzw. *regularyzacja*. Literatura dotycząca tych technik jest ogromna, por. np. Tichonow i Arsenin (1977), Engl i inni (1996), Kaipio i Somersalo (2005).

Od tego momentu będziemy zakładać, że jednoznaczne rozwiązanie problemu istnieje (można to często zapewnić przez odpowiedni dobór zbiorów D i R), a źródłem złego postawienia problemu jest brak ciągłości $\mathcal{K}^{-1}$, rozumianego jako odwrócenie bijekcji $\mathcal{K}$ między R a odpowiednim podzbiorem zbioru D. Intensywne i systematyczne badania różnych technik rozwiązywania źle postawionych problemów odwrotnych rozpoczęły się na początku lat 60 XX wieku.

W 1962 r. Ivanov zaproponował konstrukcję tzw. pseudo-rozwiązań, która poźniej rozwinęła się w tzw. metodę sit. W podejściu tym rozpatruje się ciąg wypukłych zbiorów zwartych $D_1 \subset D_2 \subset \ldots$ takich, że $\overline{\bigcup_i D_i} = D$ i konstruuje się ciąg pseudo-rozwiązań x_i przez minimalizację $\|\mathcal{K}x - y\|$ względem $x \in D_i$. Przy pewnych ogólnych założeniach Ivanov pokazał, że $x_i \to x = \mathcal{K}^{-1}y$.

W tym samym czasie Tichonow(1963) i Phillips (1962) zaproponowali metodę regularyzacji, która jest równoważna metodzie Ivanova. Rozwiązanie równania (1) minimalizuje (do zera) $\|\mathcal{K}x - y\|^2$, ale jest niestabilne. Wprowadza się więc funkcjonał kary $\Omega(x)$ określony na D, ograniczony, nieujemny, półciągły z dołu i taki, że zbiory $M_d = \{x : \ \Omega(x) \leq d\}$ są zwarte i konstruuje się pseudo-rozwiązania x_γ przez minimalizację względem $x \in D$ wyrażenia

$$\|\mathcal{K}x - y\|^2 + \gamma\Omega(x).$$

Można pokazać, że $x_\gamma \to x = \mathcal{K}^{-1}y$, gdy $\gamma \to 0$. Tichonow rozpatrywał także przypadek, gdy prawa strona równania (1) jest znana tylko z dokładnością do ε, tzn. znamy $\tilde{y}$ takie, że $\|\tilde{y} - \mathcal{K}x\| \leq \varepsilon$, ale nie znamy y. Tichonow pokazał, że minimalizując $\|\mathcal{K}x - \tilde{y}\|^2 + \gamma(\varepsilon)\Omega(x)$ względem $x \in D$ otrzymamy pseudo-rozwiązania x_ε zbieżne do $x = \mathcal{K}^{-1}y$ gdy $\varepsilon \to 0$, o ile $\gamma(\varepsilon) \to 0$ i $\varepsilon^2/\gamma(\varepsilon) \to 0$.

Innym sposobem podejścia do tego ostatniego problemu jest modyfikacja samego problemu przez dalsze zawężenie dziedziny D tak, aby uzyskać ciągłość $\mathcal{K}^{-1}$ i poszukiwać rozwiązań o "najlepszym tempie zbieżności" gdy $\varepsilon \to 0$. Stopień trudności tak zmodyfikowanego problemu można opisać modułem ciągłości w_D odwzorowania $\mathcal{K}^{-1}$ z dziedziną $\mathcal{K}D$:

$$w_D(\varepsilon) = \sup\left\{\|x_1 - x_2\| : \ \|\mathcal{K}x_1 - \mathcal{K}x_2\| \leq \varepsilon, \ x_1, x_2 \in D\right\}.$$

Zakładamy, jak wyżej, że $\|\tilde{y} - \mathcal{K}x\| \leq \varepsilon$ i szukamy rozwiązania $\hat{x}(\tilde{y})$. Jakość tego rozwiązania mierzy się jego "deterministycznym ryzykiem":

$$R(\hat{x}, x) = \sup_{y:\, \|y - \mathcal{K}x\| \leq \varepsilon} \|\hat{x}(y) - x\|.$$

W podejściu minimaksowym chcemy zminimalizować ryzyko w przypadku najgorszym. Można pokazać, że jeżeli D jest symetryczny względem zera i wypukły, to

$$\inf_{\hat{x}} \sup_{x \in D} R(\hat{x}, x) \geq w_D(\varepsilon)$$

i że istnieje rozwiązanie $\tilde{x}$ spełniające

$$\sup_{x \in D} R(\tilde{x}, x) \leq 2w_D(\varepsilon).$$

Warto podkreślić, że o takich "tempach zbieżności" można mówić dopiero po uciągleniu $\mathcal{K}^{-1}$. Mamy wtedy już do czynienia z problemem dobrze postawionym, choć zwykle źle uwarunkowanym, a stopień złego uwarunkowania zależy od kształtu modułu ciągłości $w_D(\varepsilon)$ w prawostronnym otoczeniu zera. Uciąglanie $\mathcal{K}^{-1}$ przez zawężenie dziedziny D można interpretować jako wykorzystywanie informacji a priori do stabilizacji rozwiązania.

Oczywiście, ciągłość lub jej brak zależą od norm w przestrzeniach $\mathcal{X}$ i $\mathcal{Y}$ i mogłoby się wydawać, że przez wybór odpowiedniej normy można pozbyć się problemu uciąglając $\mathcal{K}^{-1}$. Sprawa nie jest jednak taka prosta.

Przykład: Niech $D = \mathcal{X} = C([0,1])$ i $\mathcal{Y} = \mathcal{K}(\mathcal{X}) \subset C([0,1])$ z

$$(\mathcal{K}x)(t) = y(t) = \int_0^t x(s)\mathrm{d}s$$

i z normami maksimum w $\mathcal{X}$ i $\mathcal{Y}$. Jednoznaczne rozwiązanie równania (1) ma postać $x(t) = \dot{y}(t)$, ale operator różniczkowania jest nieciągły przy przyjętych normach, co widać natychmiast, jeżeli weźmiemy pewien ustalony $x \in \mathcal{X}$ i przyjmiemy $y = \mathcal{K}x$ oraz $x_n(s) = x(s) + \cos(ns)$, $n \in \mathbb{N}$. Wtedy $y_n(t) = (\mathcal{K}x_n)(t) = y(t) + \sin(nt)/n$ i $\|y_n - y\| = 1/n \to 0$ gdy $n \to \infty$ ale $\|x_n - x\| = 1$. Operator różniczkowania można by uciąglić zmieniając normę w $\mathcal{Y}$ na przykład na $\|y\|_1 = \max_t |y(t)| + \max_t |\dot{y}(t)|$. (Wtedy $\|y_n - y\|_1 = 1 + 1/n \nrightarrow 0$ a $\mathcal{K}^{-1}$ jest oczywiście ciągły.) Przy konstrukcji modelu norma nie może być jednak całkiem dowolna i musi odzwierciedlać sens bliskości "mierzonego" $\tilde{y}$ i prawdziwego y. W przykładzie "mierzymy" wartości y, a problem odwrotny polega na wyznaczeniu pochodnej. Norma maksimum jest wtedy naturalna. Gdybyśmy przyjęli drugą normę, to znaczyłoby to, że mierzymy także wartości pochodnej i problem odwrotny przestałby być interesujący.

1.2 Stochastyczny problem odwrotny

O stochastycznym problemie odwrotnym mówimy wtedy, gdy prawa strona równania (1) jest "mierzona z pewnym losowym błędem". Rodzi to natychmiast problemy, bo błąd pomiarowy może (podobnie jak w przypadku niestochastycznym) wyprowadzić zmierzony $\tilde{y}$ poza R nawet gdy "prawdziwy" y jest w R. Ten aspekt na razie pominiemy. Ponadto, nawet gdy $y \in R$, błąd propaguje do $x = \mathcal{K}^{-1}y$ i może być dowolnie duży, gdy $\mathcal{K}^{-1}$ nie jest ciągły.

Od tego momentu rozpatrywane przestrzenie będą najczęściej, choć nie zawsze, (rzeczywistymi) przestrzeniami Hilberta, zwykle przestrzeniami funkcyjnymi. Będzie to podkreślone zmianą notacji. Będziemy pisać $\mathcal{K} : H_1 \longrightarrow H_2$, a równanie (1) będziemy zapisywać w postaci

$$\mathcal{K}f = g. \tag{2}$$

Typowe modele szumu obserwacyjnego, a więc statystycznej komponenty modelu, są następujące:

- Obserwacje "zgodne w normie": $\|\tilde{g}_n - g\| \xrightarrow{p} 0$ gdy $n \to \infty$ i n opisuje "wielkość eksperymentu". Vapnik i Stefanyuk (1978) (patrz też Vapnik, 1995) pokazali, że dla

  $$\tilde{f}_n = \text{argmin}_f \, \|\mathcal{K}f - \tilde{g}_n\|^2 + \gamma_n \|f\|^2,$$

  a więc dla rozwiązania otrzymanego przez regularyzację Tichonowa z $\Omega(f) = \|f\|^2$, mamy

  $$\forall \varepsilon > 0 \; \exists \gamma_0 = \gamma_0(\varepsilon) \; \forall \{\gamma_n\} \leq \gamma_0 \; P\left(\|f - \tilde{f}_n\| > \varepsilon\right) \leq 2P\left(\|g - \tilde{g}_n\|^2 > \gamma_n \varepsilon\right)$$

  Aby można było stosować tę nierówność dla każdego ε, np. w dowodach zgodności, trzeba przyjąć, że $\gamma_n \to 0$, ale nie zbyt szybko, aby prawa strona nierówności mogła dążyć do zera. W dowodzie powyższej nierówności istotny jest fakt, że H_1 jest przestrzenią Hilberta, bo wtedy kula w H_1 jest słabo zwarta i możliwa jest odpowiednia modyfikacja twierdzenia Tichonowa. Ogólnie, warunek zwartości zbioru M_d w regularyzacji Tichonowa jest mocny i nie jest spełniony z $\Omega(f) = \|f\|^2$ w nieskończenie wymiarowych przestrzeniach Banacha.

- Obserwacje w "białym szumie gaussowskim": $\tilde{g} = g + \varepsilon\xi$, gdzie $\varepsilon > 0$ jest poziomem szumu, a ξ jest "uogólnionym białym szumem w H_2". Oznacza to, że dla każdego $h \in H_2$ można obserwować zmienną losową

  $$< \tilde{g}, h > = < g, h > + \varepsilon < \xi, h >$$

  gdzie $< \xi, h > \sim N(0, \|h\|^2)$, a dla $h_1, h_2 \in H_2$

  $$\mathrm{E}\left(< \xi, h_1 >< \xi, h_2 >\right) = < h_1, h_2 >$$

  Donoho (1995) pisze, że nie jest mu znany żaden rzeczywisty problem naukowy, w którym dane byłyby zbierane zgodnie z modelem białego szumu. Atrakcyjność tego modelu wynika z faktu, że jest on graniczną postacią innych, bardziej realistycznych modeli, a wyniki asymptotyczne uzyskane dla tego modelu są często stosowalne do modeli bardziej realistycznych, zgodnie z tak zwaną zasadą równoważności (Brown i Low, 1996: Efromovich, 1999, Rozdz. 7, Brown i in. 2004). Reprezentatywne dla tego nurtu są np. prace Donoho (1995), Cavalier i Tsybakov (2002), Cavalier (2006).

- Obserwacje typu regresyjnego: Dla pewnego zbioru punktów $\{y_i\}$, mierzymy $g(y_i)$ z niezależnymi błędami, zwykle gaussowskimi. Problemy tego typu zaczęli badać Wahba (1977) oraz Nychka i Cox (1989).

- "Odwrotna estymacja gęstości": Obserwujemy próbę prostą z rozkładu o gęstości g (por. np. Mair i Ruymgaart, 1996 i van Rooij i Ruymgaart, 1996).

- Obserwacje w szumie poissonowskim: Obserwujemy niejednorodny proces Poissona o funkcji intensywności g względem pewnej miary μ (por. np. Johnstone i Silverman, 1990, 1991, Szkutnik, 2000, 2003, 2005, Antoniadis i Bigot, 2006). Ten model można związać z poprzednim przez spostrzeżenie, że gęstość rozkładu procesu $\mathcal{N}_g$ odpowiadającego g względem rozkładu procesu $\mathcal{N}_{g_0}$ odpowiadającego pewnej ustalonej funkcji g_0 ma postać (por. Reiss, 1993)

$$\frac{\mathrm{d}\mathcal{L}(\mathcal{N}_g)}{\mathrm{d}\mathcal{L}(\mathcal{N}_{g_0})} = \exp\left[\int (g_0 - g)\mathrm{d}\mu + \int \log(g/g_0)\mathrm{d}\mathcal{N}_g\right].$$

Drugim podstawowym składnikiem problemu odwrotnego, od którego zależą jego własności, jest operator opisujący mechanizm zniekształcania lub transformacji danych. Szczególnie ważna i dobrze zbadana jest klasa operatorów całkowych postaci

$$(\mathcal{K}f)(y) = \int_\Omega k(y,x)f(x)\mathrm{d}x, \quad x, y \in \Omega \subset \mathbb{R}^k. \tag{3}$$

Naturę złego postawienia problemu widać bardzo dobrze w przypadku operatorów zwartych, którym będzie poświęcony następny rozdział. Czasami formalizm przestrzeni Hilberta nie jest specjalnie pomocny. Tak jest np. w przypadku operatorów opisujących różne mechanizmy cenzurowania (por. Groeneboom, 1996). Interesujące są też problemy, w których operator jest znany tylko w przybliżeniu (por. Szkutnik, 2000 dla szumu poissonowskiego, oraz Efromovich i Koltchinskii, 2001 dla obserwacji w białym szumie).

Uwaga terminologiczna: Groeneboom (1996), głównie w kontekście problemów cenzurowania, nazywa problemem odwrotnym taki problem estymacji, w którym:

1. dysponujemy tylko pośrednią informacją o interesujących nas zmiennych

2. tempo zbieżności każdego (punktowego) estymatora dystrybuanty zmiennych nas interesujących jest wolniejsze niż parametryczne $n^{-1/2}$ (w przypadku próby prostej).

Taka typologia wydaje się jednak dyskusyjna, bo nie pozwala na przykład potraktować "problemów prostych" z parametrycznym tempem zbieżności jako problemów odwrotnych z operatorem tożsamościowym.

1.3 Przykłady

Estymacja gęstości: Obserwujemy próbę prostą z rozkładu o dystrybuancie g na $\mathbb{R}$ i gęstości $f = g'$. Wtedy

$$g(y) = \int_{-\infty}^{y} f(x)\mathrm{d}x.$$

Dla dystrybuanty empirycznej g_n mamy nierówność Dvoretzky'ego-Kiefera-Wolfowitza

$$P(\|g_n - g\|_\infty > \varepsilon) \leq 2\exp(-2n\varepsilon^2)$$

a więc mamy do czynienia z przypadkiem obserwacji "zgodnych w normie". Stosując nierówność Vapnika-Stefanyuka dla regularyzacji Tichonowa otrzymamy

$$P(\|f - \tilde{f}_n\| > \varepsilon) \leq 4\exp(-2\varepsilon n\gamma_n)$$

co daje zgodność $\tilde{f}_n$ w normie przestrzeni Hilberta H_1, gdy $\gamma_n \to 0$ i $n\gamma_n \to \infty$.

Dekonwolucja na okręgu: f jest interesującą nas gęstością na $[0, 2\pi]$ spełniającą "okresowe warunki brzegowe", k jest gęstością rozkładu błędu na $[0, 2\pi]$ i obserwujemy próbę prostą z gęstości

$$g(y) = \int_0^{2\pi} k(y - x)f(x)\mathrm{d}x$$

gdzie argumenty funkcji są dodawane modulo 2π (splot funkcji na grupie addytywnej $[0, 2\pi]$ z dodawaniem modulo 2π). Należy więc wyestymować f na podstawie danych z addytywnym błędem.

Dekonwolucja na prostej: Wersja powyższego ze zwykłą arytmetyką argumentów, z gęstościami f i k na $\mathbb{R}$ oraz

$$g(y) = \int_{-\infty}^{\infty} k(y - x)f(x)\mathrm{d}x$$

Problemy stereologiczne: Jednym z takich problemów jest tzw. problem Wicksella. W nieprzezroczystym ośrodku rozmieszczone są kule o losowych promieniach o rozkładzie Q na $[0, 1]$, których środki tworzą jednorodny proces Poissona o intensywności c. Obserwujemy koła będące przekrojami tych kul z przekrojem ośrodka płaszczyzną o losowej orientacji. Jeżeli przez n oznaczymy parametr opisujący "wielkość eksperymentu", to zmierzone promienie kół tworzą niejednorodny proces Poissona na $[0, 1]$ o funkcji intensywności

$$g(y) = ncy \int_y^1 \frac{\mathrm{d}Q(x)}{\sqrt{x^2 - y^2}}$$

Zadanie polega na odwikłaniu stałej c i rozkładu Q na podstawie obserwacji procesu promieni kół.

Ornitologiczny problem Hampela: Niech F oznacza dystrybuantę rozkładu czasu pobytu ptaków określonego gatunku w pewnym rejonie. Ptaki są łapane i obrączkowane, co

pozwala rejestrować czasy między kolejnymi schwytaniami tego samego ptaka. Oznaczmy przez g gęstość rozkładu tej ostatniej zmiennej. Hampel (1987) pokazał (patrz też Groeneboom, 1996), że

$$g(y) = \frac{2}{c}\int_y^\infty (x-y)\mathrm{d}F(x)$$

gdzie $0 < c = \int_0^\infty x^2 \mathrm{d}F(x) < \infty$. Wynika z tego, że $g'(y) = -(2/c)(1-F(y))$, co pokazuje, że g musi być ograniczoną, malejącą i wypukłą gęstością na $[0,\infty)$. Po skonstruowaniu odpowiedniego estymatora $\hat{g}_n$ takiej gęstości, można by dalej estymować F przez $\hat{F}_n(x) = 1 - \hat{g}_n'(x)/\hat{g}_n'(0)$.

Odwracanie transformacji Laplace'a: W analizie czasów przeżycia rozpatruje się (Jewell, 1982) ważone mieszanki rozkładów wykładniczych o dystrybuantach postaci:

$$g(t) = \int_0^\infty \left(1 - \mathrm{e}^{-tx}\right) f(x)\mathrm{d}x.$$

Zadanie polega na odwikłaniu gęstości mieszającej f na podstawie obserwacji próby prostej z rozkładu o dystrybuancie g (Chauveau i inni, 1994).

1.4 Podejście minimaksowe i dolne ograniczenia dla temp zbieżności

Będziemy szukać rozwiązań z optymalnym tempem zbieżności ryzyka $R(\hat{f}_n, f) = \mathbb{E}_f\|\hat{f}_n - f\|^2$ na danej klasie funkcji. Rozwiązanie tak postawionego zadania polega na znalezieniu dolnego ograniczenia dla tempa zbieżności ryzyka dowolnego estymatora i pokazaniu, że jest ono osiągalne. Niech $\mathcal{F}$ będzie klasą funkcji, w której szukamy rozwiązania, a n parametrem opisującym wielkość eksperymentu, np. licznością próby prostej.

Definicja: *Mówimy, że ciąg $\{\alpha_n\}$ jest dolnym ograniczeniem dla tempa zbieżności ryzyka w problemie estymacji f z klasy $\mathcal{F}$, jeżeli dla pewnej dodatniej stałej C_0 mamy*

$$\liminf_{n\to\infty} \frac{\inf_{\hat{f}_n}\sup_{f\in\mathcal{F}}\mathbb{E}_f\|\hat{f}_n - f\|^2}{\alpha_n} \geq C_0. \tag{4}$$

Jeżeli ponadto dla pewnego C_1 i pewnego estymatora $\tilde{f}_n$ mamy

$$\limsup_{n\to\infty} \frac{\sup_{f\in\mathcal{F}}\mathbb{E}_f\|\tilde{f}_n - f\|^2}{\alpha_n} \leq C_1, \tag{5}$$

to mówimy, że $\{\alpha_n\}$ jest minimaksowym tempem zbieżności, oraz że estymator $\tilde{f}_n$ jest minimaksowy w sensie rzędu zbieżności na klasie $\mathcal{F}$.

Zauważmy, że zamiast (4) i (5) można napisać, dla dużych n i z pewnymi stałymi $0 < C_0' < C_0$ i $C_1 < C_1' < \infty$

$$\forall \hat{f}_n \quad \sup_{f \in \mathcal{F}} \mathbb{E}_f \|\hat{f}_n - f\|^2 \geq C_0' \alpha_n$$

oraz

$$\sup_{f \in \mathcal{F}} \mathbb{E}_f \|\tilde{f}_n - f\|^2 \leq C_1' \alpha_n,$$

co oznacza, że

$$\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}_f \|\hat{f}_n - f\|^2 \asymp \alpha_n,$$

a więc, minimaksowe tempo zbieżności wyznacza rząd wielkości ryzyka minimaksowego. Obszerniejszą dyskusję wprowadzonych tu pojęć można znaleźć w książce Korosteleva i Tsybakova (1993, Rozdz. 1.7 i 2.1).

Załóżmy, że $\mathcal{K}^{-1}$ jest ciągły na $\mathcal{KF}$. Dla niestochastycznych problemów odwrotnych istnieje elegancki opis dolnych i górnych ograniczeń "tempa zbieżności" w terminach modułu ciągłości $\mathcal{K}^{-1}$ (por. Rozdz. 1.1). Nie ma analogicznych ogólnych wyrażeń tego typu w przypadku stochastycznym. Istnieje wprawdzie praca Solo (2000), w której podjęto próbę wyrażenia dolnych ograniczeń dla temp zbieżności w terminach modułu ciągłości i tzw. pojemności Kołmogorowa klasy $\mathcal{F}$, ale dowód podanego tam bardzo ogólnego twierdzenia, obejmującego także przypadek operatorów nieliniowych, jest błędny.

Do wyznaczenia dolnych ograniczeń można, idąc śladami Ibragimova i Chasminskiego (1981), wykorzystać pewien fakt z teorii kodowania, znany jako lemat Fano (por. Fano, 1961). Przytoczymy go w zmodyfikowanej wersji Birgégo (por. Birgé, 1983), który zmodyfikował oryginalny lemat dodając oszacowanie tzw. entropii warunkowej przy pomocy odległości Kullbacka-Leiblera $KL(\cdot, \cdot)$.

Lemat Fano: *Niech na przestrzeni mierzalnej $(\mathcal{X}, \mathcal{F})$ będą określone: r miar probabilistycznych $P_1, \ldots, P_r$ oraz mierzalna funkcja $\psi : \mathcal{X} \longrightarrow \{1, \ldots, r\}$, $r \geq 2$. Wtedy*

$$\max_{1 \leq i \leq r} P_i(\psi(x) \neq i) \geq 1 - \frac{\max_{i,j} KL(P_i, P_j) + \log 2}{\log(r-1)}.$$

W teorii kodowania elementy przestrzeni $\mathcal{X}$ są sygnałami obserwowanymi na wyjściu kanału transmisji, P_i jest rozkładem pojawiającym się przy transmisji i-tego elementu alfabetu wejściowego, a ψ jest regułą decyzyjną "identyfikującą" przesyłany element alfabetu na podstawie sygnału wyjściowego. Lemat Fano podaje więc dolne oszacowanie dla maksymalnego względem znaków alfabetu prawdopodobieństwa błędnej transmisji. Szukając dolnych ograniczeń dla tempa zbieżności estymatora można wykorzystać fakt, że mając estymator o określonej precyzji można rozwiązać problem wielodecyzyjny analogiczny do rozpatrywanego w lemacie Fano.

Niech $\mathcal{F}^0 = \{f_1, \ldots, f_r\}$ będzie układem elementów $\mathcal{F}$ takich, że $\|f_i - f_j\| > 2\varepsilon$, gdy $i \neq j$, a P_i niech oznacza rozkład obserwacji, zależny od $g_i = \mathcal{K}f_i$ (i od indeksu n). Dla

danego estymatora $\hat{f}_n$, niech ψ będzie regułą decyzyjną, która na podstawie obserwacji Y_n wybiera z $\mathcal{F}^0$ element f_i najbliższy $\hat{f}_n$ w normie przestrzeni H_1. Korzystając z nierówności Markowa, z implikacji $\|\hat{f}_n - f_i\| < \varepsilon \Rightarrow \psi(Y_n) = f_i$ i z lematu Fano, otrzymujemy

$$\begin{aligned}\sup_{f\in\mathcal{F}} \mathbb{E}_f\|\hat{f}_n - f\|^2 &\geq \max_i \mathbb{E}_f\|\hat{f}_n - f_i\|^2 \geq \varepsilon^2 \max_i P_i\left(\|\hat{f}_n - f_i\| > \varepsilon\right)\\ &\geq \varepsilon^2 \max_i P_i\left(\psi(Y_n) \neq f_i\right)\\ &\geq \varepsilon^2\left(1 - \frac{\max\limits_{i,j} KL(P_i, P_j) + \log 2}{\log(r-1)}\right).\end{aligned}$$

Dobre dolne ograniczenie dla błędu estymatora otrzymamy szukając możliwie wielu elementów f_i odległych co najmniej o 2ε, które jednocześnie prowadzą do "bliskich" w sensie Kullbacka-Leiblera rozkładów obserwacji P_i. Bywa to technicznie trudne. Dla problemów z operatorem zwartym, program ten został zrealizowany przez Johnstone'a i Silvermana (1990). Alternatywne podejścia do wyznaczania dolnych ograniczeń są opisane np. przez van Rooija i Ruymgaarta (1996), Korosteleva i Tsybakova (1993) i Tsybakova (2009).

2 Operatory zwarte i SVD

Mówimy, że (być może nieliniowy) operator $\mathcal{K} : H_1 \longrightarrow H_2$ jest zwarty, jeżeli obraz $\mathcal{K}(Z)$ dowolnego zbioru ograniczonego $Z \subset H_1$ jest względnie zwarty w H_2. Jeżeli H_1 jest przestrzenią metryczną, równoważnym warunkiem zwartości operatora jest by obraz dowolnego ciągu ograniczonego zawierał podciąg zbieżny. Gdy $\mathcal{K}$ jest nieliniowy, zwartość $\mathcal{K}$ nie musi implikować jego ciągłości. Operator zwarty i ciągły nazywamy pełnociągłym. Każdy zwarty operator liniowy jest ciągły, więc dla operatorów liniowych zwartość i pełnociągłość są równoważne (por. Krasnoselskij, 1963). Wiadomo, że jeżeli $\dim H_1 = \infty$ i $\mathcal{K}$ jest zwarty, to $\mathcal{K}^{-1}$ jest nieograniczony, w istocie jako konsekwencja braku zwartości kuli domkniętej w przestrzeni nieskończenie wymiarowej. Zwartość, a w konsekwencji złe postawienie problemu odwrotnego są bardzo częste w zastosowaniach. Przyjrzyjmy się, w szczególności, operatorom całkowym postaci (3) i przypomnijmy, że jądro $k(y,x)$ jest nazywane słabo osobliwym, gdy zbiór Ω jest mierzalny i ograniczony w $\mathbb{R}^k$, oraz dla $x \neq y$ $k(y,x)$ daje się przedstawić w postaci $k(y,x) = a(y,x)/|x-y|^\alpha$, gdzie $\alpha \in (0,k)$ a $a(y,x)$ jest funkcją mierzalną i ograniczoną. Dla operatorów całkowych wiadomo, że (patrz np. Kołodziej, 1970):

1. Jeżeli jądro $k(y,x)$ jest ciągłe albo słabo osobliwe z ciągłą funkcją $a(\cdot,\cdot)$ i zwartym zbiorem Ω, to $\mathcal{K}$ jest zwartym operatorem w $C(\Omega)$ (z normą maksimum)

2. Jeżeli jądro $k(y,x)$ jest całkowalne z kwadratem na $\Omega \times \Omega$, albo jest słabo osobliwe, to $\mathcal{K}$ jest zwartym operatorem w $L^2(\Omega)$.

Gdy na przykład

$$(\mathcal{K}f)(y) = g(y) = \int_0^y f(x)\mathrm{d}x$$

z $f \in C([0,1])$, to $k(y,x) = \mathbf{1}_{(x<y)}$ z $(y,x) \in [0,1]^2$ i jądro jest całkowalne z kwadratem i słabo osobliwe, bo można je np. przedstawić w postaci

$$k(y,x) = \frac{|y-x|^{1/2}\mathbf{1}_{(x<y)}}{|y-x|^{1/2}}.$$

W konsekwencji operator całkowania jest zwarty, a odwrotny operator różniczkowania jest nieograniczony zarówno w normie maksimum, jak i w normie $L^2([0,1])$. To wyjaśnia statystyczną trudność problemu estymacji gęstości: choć potrafimy estymować dystrybuantę z parametrycznym tempem zbieżności (np. w normie supremum), to tempo zbieżności estymatora gęstości jest zwykle wolniejsze. Zwarte są też operatory splotu dla dekonwolucji na okręgu (ale nie na prostej!) i operator z problemu Wicksella.

Klasa problemów z operatorem zwartym jest szczególnie dobrze zbadana i dobrze ilustruje źródło trudności dzięki następującemu twierdzeniu o istnieniu tzw. dekompozycji operatora zwartego wg wartości singularnych (por. np. Kaipio, Somersalo, 2005):

TWIERDZENIE: *Dla liniowego i zwartego operatora $\mathcal{K} : H_1 \to H_2$ istnieją*

- *skończony lub zbieżny monotonicznie do zera ciąg liczb dodatnich $\{s_n\}_{n\in I}$ i*
- *ciągi ortonormalne $\{\phi_n\}_{n\in I} \subset H_1$ i $\{\psi_n\}_{n\in I} \subset H_2$*

takie że

1. $\overline{\mathrm{Span}}\{\phi_n : n \in I\} = \mathrm{Ker}^{\perp}(\mathcal{K})$
2. $\overline{\mathrm{Span}}\{\psi_n : n \in I\} = \overline{\mathrm{Range}(\mathcal{K})}$
3. $\mathcal{K}f = \sum_n s_n < f, \phi_n > \psi_n$ *oraz* $\mathcal{K}^* g = \sum_n s_n < g, \psi_n > \phi_n$

Ponadto, $g \in \mathrm{Range}(\mathcal{K})$ wtedy i tylko wtedy, gdy spełniony jest tzw. warunek Picarda

$$\sum_n \frac{1}{s_n^2} | < g, \psi_n > |^2 < \infty$$

Wtedy rozwiązania równania $\mathcal{K}f = g$ mają postać

$$f = f_0 + \sum_n \frac{1}{s_n} < g, \psi_n > \phi_n \tag{6}$$

gdzie $f_0 \in \mathrm{Ker}(\mathcal{K})$ jest dowolne.

Dowód: Operator $\mathcal{K}^*\mathcal{K}$ jest samosprzężony i zwarty, więc $\mathcal{K}^*\mathcal{K}\phi_n = s_n^2\phi_n$ dla pewnych ortonormalnych elementów $\phi_n \in H_1$ oraz $s_1^2 \geq s_2^2 \geq \ldots \geq 0$. Niech $I := \{n : s_n > 0\}$ oraz $\psi_n := s_n^{-1}\mathcal{K}\phi_n$, dla $n \in I$. Wtedy także $\psi_n,\ n \in I$ tworzą układ ortonormalny, bo $< \psi_k, \psi_l >= s_n^{-2} < \mathcal{K}\phi_k, \mathcal{K}\phi_l >= s_n^{-2} < \phi_k, \mathcal{K}^*\mathcal{K}\phi_l >= \delta_{kl}$. Przypomnijmy, że $\mathrm{Ker}(\mathcal{K}) = \mathrm{Ker}(\mathcal{K}^*\mathcal{K})$, bo jeżeli $u \in \mathrm{Ker}(\mathcal{K}^*\mathcal{K})$, to $\|\mathcal{K}u\|^2 =< \mathcal{K}u, \mathcal{K}u >=< u, \mathcal{K}^*\mathcal{K}u >$ $= 0$ i w konsekwencji $\mathcal{K}u = 0$. Dlatego, wykorzystując lemat A6 z Kaipio i Somersalo (2005), mamy

$$\mathrm{Ker}^{\perp}(\mathcal{K}) = \mathrm{Ker}^{\perp}(\mathcal{K}^*\mathcal{K}) = \left(\mathrm{Range}^{\perp}(\mathcal{K}^*\mathcal{K})\right)^{\perp} = \overline{\mathrm{Range}(\mathcal{K}^*\mathcal{K})} = \overline{\mathrm{Span}}\{\phi_n : n \in I\},$$

co dowodzi punktu 1. (Wykorzystaliśmy przy tym fakt, że podwójne dopełnienie ortogonalne przestrzeni jest jej domknięciem, por. np. Mlak, 1972, str. 80) Dla dowodu 2 zauważmy, że zwarty i samosprzężony jest też operator $\mathcal{K}\mathcal{K}^*$ i że $\mathcal{K}\mathcal{K}^*\psi_n = s_n^2\psi_n$, $n \in I$, jest jego rozkładem spektralnym, więc rozumując jak wyżej, ale z zamianą $\mathcal{K}$ na $\mathcal{K}^*$, dostajemy $\mathrm{Ker}^{\perp}(\mathcal{K}^*) = \overline{\mathrm{Span}}\{\psi_n : n \in I\}$, co dowodzi 2, bo $\mathrm{Ker}^{\perp}(\mathcal{K}^*) =$ $(\mathrm{Range}^{\perp}\mathcal{K})^{\perp} = \overline{\mathrm{Range}(\mathcal{K})}$, po ponownym skorzystaniu z lematu A6 z Kaipio i Somersalo (2005). Zależności w 3 są oczywiste, a warunek Picarda wynika stąd, że $g = \mathcal{K}f$ z pewnym $f \in H_1$ wtedy i tylko wtedy, gdy

$$\sum_n < g, \psi_n > \psi_n = g = \mathcal{K}f = \sum_n s_n < f, \phi_n > \psi_n$$

z $\sum_n | < f, \phi_n > |^2 < \infty$, czyli z $\sum_n s_n^{-2} | < g, \psi_n > |^2 < \infty$. To kończy dowód.

Ciąg $\{s_n\}$ nazywamy ciągiem wartości singularnych operatora, elementy ϕ_n prawymi, a ψ_n lewymi elementami singularnymi. Reprezentację operatora z punktu 3 twierdzenia nazywamy jego dekompozycją według wartości osobliwych (SVD).

Uwaga: Nie należy mylić wartości własnych z wartościami singularnymi, które są czymś innym, nawet w przypadku automorfizmów. Jeżeli np. $H_1 = H_2$, a $\{e_k\}$ jest układem ortonormalnym i zupełnym w H_1, oraz

$$\mathcal{K}f = \sum_{k=1}^{\infty} \frac{1}{k} < f, e_k > e_{k+1},$$

to $\mathcal{K}$ jest zwarty jako granica (zwartych) operatorów skończenie wymiarowych

$$\mathcal{K}_n f = \sum_{k=1}^{n} \frac{1}{k} < f, e_k > e_{k+1}$$

i łatwo sprawdzić, że $\mathcal{K}$ nie ma wartości własnych. Jednocześnie jest oczywiste, że $s_k = 1/k$, $\phi_k = e_k$, $\psi_k = e_{k+1}$, $k = 1, 2, \ldots$ tworzą SVD operatora $\mathcal{K}$.

W dalszych rozważaniach przyjmujemy dla uproszczenia, że $\mathcal{K}$ jest odwracalny. Jeżeli g jest wielkością mierzoną, to oczywiście nie ma żadnej gwarancji, że $g \in \mathrm{Range}(\mathcal{K})$. Kuszącym pomysłem jest więc zrzutowanie pomiarowego g na $\overline{\mathrm{Range}(\mathcal{K})}$,

czyli zastąpienie prawej strony równania przez $\sum_n < g, \psi_n > \psi_n$. Ponieważ jednak Range($\mathcal{K}$) nie jest domkniętą podprzestrzenią H_2 (gdyby był, to $\mathcal{K}^{-1}$ byłby ograniczony na Range($\mathcal{K}$) zgodnie z twierdzeniem Banacha), więc zwykle nawet zrzutowane g nie będzie należało do Range($\mathcal{K}$). Ponadto, z postaci (6) rozwiązania widać, że rosnące do nieskończoności współczynniki $1/s_n$ wzmacniają szum pomiarowy związany z wyznaczaniem $< g, \psi_n >$, wyjaśniając naturę złego postawienia problemu.

Zwróćmy uwagę, że SVD wprowadza nowe ortonormalne bazy, w których operator się diagonalizuje i problem estymacji można sformułować i rozwiązywać w tej nowej parametryzacji, co jest zwykle prostsze. Wiadomo, że s_n^2 i ϕ_n są, odpowiednio, wartościami i wektorami własnymi operatora samosprzężonego $\mathcal{K}^*\mathcal{K}$, co czasami pozwala stosunkowo łatwo wyznaczyć SVD operatora, zwłaszcza operatora całkowego.

Przykład: Dla operatora całkowania w $L^2[0,1]$ mamy

$$(\mathcal{K}^*\mathcal{K}f)(x) = (1-x)\int_0^x f(u)\mathrm{d}u + \int_x^1 (1-u)f(u)\mathrm{d}u$$

i problem własny $\mathcal{K}^*\mathcal{K}f = \mu f$ można łatwo przedstawić w równoważnej postaci różniczkowej $f'' + \mu^{-1}f = 0$ z warunkami brzegowymi $f(1) = 0$, $f'(0) = 0$, co daje w efekcie:

$$s_n = \frac{1}{\pi(n+1/2)}, \quad n = 0, 1, \ldots$$

$$\phi_n(x) = \sqrt{2}\cos\left[\left(n+\frac{1}{2}\right)\pi x\right] \quad \psi_n(y) = \sqrt{2}\sin\left[\left(n+\frac{1}{2}\right)\pi y\right]$$

3 Podejście minimaksowe z operatorem zwartym.

Przeprowadzona w poprzednim rozdziale analiza źródeł złego uwarunkowania problemu sugeruje próbę konstrukcji "dobrych" estymatorów przez modyfikację wag s_n^{-1} we wzorze (6), tzn. konstrukcję estymatora postaci

$$\hat{f}_n = \sum_i w(s_i) < g, \psi_i > \phi_i. \tag{7}$$

Często funkcja $w(\cdot)$ jest konstruowana w ten sposób, że suma w (7) jest skończona. Do rozwiązań tego typu prowadzi też np. regularyzacja typu Tichonowa-Phillipsa, która konstruuje rozwiązanie przez minimalizację względem f wielkości

$$\|\mathcal{K}f - g\|^2 + \delta\|f\|^2$$

i daje jako prostą konsekwencję twierdzenia o SVD (por. też Kaipio, Somersalo, 2005)

$$\hat{f}_{n,\delta} = (\mathcal{K}^*\mathcal{K} + \delta I)^{-1}\mathcal{K}^* g = \sum_i \frac{s_i}{s_i^2 + \delta} < g, \psi_i > \phi_i. \tag{8}$$

Okazuje się, że estymatory postaci (7) rzeczywiście mogą osiągać minimaksowe tempa zbieżności, gdy klasa $\mathcal{F}$ jest zdefiniowana jako klasa typu Sobolewa względem układu $\{\phi_i\}$, co czasami daje się wyrazić w terminach jawnych założeń o gładkości funkcji f.

Naszkicowane tu podejście było szczegółowo badane np. przez Johnstona i Silvermana (1990, 1991), Korosteleva i Tsybakova (1993) oraz van Rooija i Ruymgaarta (1996). Za tymi autorami będziemy zakładać, że klasa $\mathcal{F}$ jest postaci

$$\mathcal{F}_{\alpha,C} = \left\{ f = \sum_i f_i \phi_i \in H_1 \ : \ \sum_i a_i^2 |f_i|^2 \leq C^2 \right\}$$

gdzie C jest stałą, oraz $a_i \asymp i^\alpha$ z pewnym $\alpha > 0$. Im większe jest α, tym szybciej maleją do zera współczynniki Fouriera funkcji f względem układu $\{\phi_i\}$ i tym gładsza jest funkcja f. Biorąc np. $H_1 = L^2(0,1)$ oraz $\phi_i(x) = \sqrt{2}\cos[(i+1/2)\pi x]$, jak w przykładzie z operatorem całkowania, oraz $a_i = [(i+1/2)\pi]^k$ z pewnym całkowitym k, można pokazać (patrz Mair i Ruymgaart, 1996), że $\mathcal{F}$ można scharakteryzować jako klasę funkcji z $L^2(0,1)$, które mają $k-1$ absolutnie ciągłych pochodnych, których k-ta pochodna ma kwadrat normy L^2 ograniczony przez C^2 i których pochodne rzędów parzystych zerują się na końcach przedziału $[0,1]$.

Stopień złego uwarunkowania problemu zależy z kolei od tempa, w jakim maleją do zera wartości singularne. Jeżeli $s_i \asymp i^{-\beta}$ to, przy pewnych dodatkowych technicznych założeniach, minimaksowe tempo zbieżności ma postać

$$\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}_{\alpha,C}} \mathbb{E}_f \|\hat{f}_n - f\|_{L^2}^2 \asymp n^{-\frac{2\alpha}{2\alpha+2\beta+1}}, \tag{9}$$

gdy dysponujemy próbą prostą z rozkładu o gęstości $g = \mathcal{K}f$. Dodatkowe techniczne założenie potrzebne do uzyskania dolnego ograniczenia może mieć np. postać

$$KL(g_a, g_b) \leq c\|g_a - g_b\|^2 \tag{10}$$

gdzie $KL(\cdot,\cdot)$ jest odległością Kullbacka-Leiblera a c pewną stałą. Wynik ten był nieformalnie sformułowany przez Johnstona i Silvermana (1990, str. 276), którzy rozpatrywali szczególny przypadek tomografii emisyjnej, ale nietrudno zmodyfikować ich dowód do przypadku ogólniejszego. Łatwo też pokazać, że warunek (10) będzie spełniony, jeżeli np. wszystkie gęstości g ze zbioru $\mathcal{KF}$ są ograniczone i odcięte od zera. W pewnych ważnych przypadkach nie jest to jednak prawdą i trzeba wtedy korzystać z nieco innych metod (np. van Rooij i Ruymgaart, 1996).

Przykład: W problemie z dekonwolucją na okręgu, niech $k(t) \sim \sum_\nu b_\nu \exp(i\nu t)$ będzie formalnym rozwinięciem gęstości błedu w zespolony szereg Fouriera. Przy założeniu $b_\nu = (1+|\nu|)^{-\beta}$ z pewnym $\beta > 0$ można pokazać, że jest to istotnie rozwinięcie gęstości probabilistycznej (można wykorzystać kryterium Polya), oraz że b_ν są wartościami singularnymi a funkcje singularne mają postać $\phi_\nu(t) = \psi_\nu(t) = (2\pi)^{-1/2}\exp(i\nu t)$. Biorąc $f_0 = (2\pi)^{-1/2}$ zapewniamy, że $f \in \mathcal{F}$ jest gęstością probabilistyczną. Z $\alpha > 1/2$ można wtedy pokazać (Johnstone, Silverman, 1991), że spełnione są potrzebne

warunki techniczne i minimaksowe tempo zbieżności jest takie jak w (9). Zauważmy, że duże β oznacza bardziej gładki rozkład błedu i prowadzi do trudniejszego problemu dekonwolucji z wolniejszym tempem zbieżności. Klasa $\mathcal{F}$ z całkowitym α zawiera gęstości, które mają α pochodnych, a pochodna rzędu α ma ograniczoną normę L^2. Im α jest większe, tym klasa $\mathcal{F}$ jest mniejsza i tempo zbieżności większe.

4 Operatory nie zwarte i twierdzenie spektralne w wersji Halmosa

Typowe problemy dekonwolucji z rozkładami o nośnikach nieograniczonych w $\mathbb{R}$ wyprowadzają nas poza klasę operatorów zwartych, gdy np. zanurzamy zbiór całkowalnych z kwadratem gęstości probabilistycznych w $L^2(\mathbb{R})$. Gdy obserwujemy $Y = X + e$, gdzie X ma nieznaną gęstość f, a błąd e ma znaną gęstość h, to Y ma gęstość

$$g(y) = \int h(y-x)f(x)\mathrm{d}x,$$

jądro $k(y,x) = h(y-x)$ nie jest ani słabo osobliwe, ani całkowalne z kwadratem, a operator splotu przestaje być zwarty. Analiza problemów z operatorami nie zwartymi jest nieco trudniejsza, bo takie operatory mają bardziej skomplikowaną strukturę spektralną. Rozpatrzmy na początek przypadek operatora samosprzężonego działającego w H_1, jaki otrzymamy w problemie dekonwolucji, gdy błąd ma rozkład symetryczny względem zera. W takich przypadkach rolę SVD przejmuje twierdzenie spektralne w wersji Halmosa mówiące, że taki operator jest unitarnie równoważny operatorowi mnożenia w pewnej przestrzeni funkcyjnej (Halmos, 1963).

Twierdzenie: *Niech $\mathcal{K}$ będzie operatorem samosprzężonym i injektywnym działającym w pewnej ośrodkowej przestrzeni Hilberta H. Istnieją: przestrzeń $(S, \mathcal{F}, \mu)$ z σ-skończoną miarą μ, nieujemna funkcja rzeczywista $b \in L^\infty(S, \mathcal{F}, \mu)$ i operator unitarny $U : H \to L^2(S, \mathcal{F}, \mu)$ taki, że $\mathcal{K} = U^{-1} M_b U$, a $M_b\eta = b \cdot \eta$ jest operatorem mnożenia przez $b \in L^\infty(S, \mathcal{F}, \mu)$ z $\eta \in L^2(S, \mathcal{F}, \mu)$.*

Uwaga: Przypadek operatorów zwartych i SVD otrzymujemy jako przypadek szczególny z $S = \mathbb{N}$, miarą liczącą μ oraz $L^2(S, \mathcal{F}, \mu) = \ell^2(\mathbb{N})$. Operator U przekształca obiekt f na nieskończony ciąg jego współczynników Fouriera względem bazy $\{\phi_n\}$.

Przykład: W problemie dekonwolucji z symetryczną gęstością błędu, oznaczmy

$$\tilde{h}(t) = \int_{-\infty}^{\infty} \mathrm{e}^{ixt} h(x)\mathrm{d}x = \int_{-\infty}^{\infty} \cos(xt)h(x)\mathrm{d}x$$

i zdefiniujmy transformatę Fouriera jako operator unitarny w $L^2(\mathbb{R})$ zadając

$$(Uf)(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{ixt} f(x)\mathrm{d}x \quad \text{dla } f \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$$

i rozszerzając w sposób ciągły na $L^2(\mathbb{R})$. Wtedy oczywiście $U(h * f) = \tilde{h} \cdot Uf$ i $\mathcal{K} = U^{-1}M_{\tilde{h}}U$ oraz $\mathcal{K}^{-1} = U^{-1}M_{1/\tilde{h}}U$.

Zauważmy, że $\tilde{h} \in L^2(\mathbb{R})$, a więc $\tilde{h}(t) \to 0$ gdy $|t| \to \infty$, co pokazuje, że problem jest źle postawiony w sensie Hadamarda. Wiadomo z analizy harmonicznej, że transformata Fouriera funkcji tym szybciej zbiega do zera, im ta funkcja jest gładsza. Tak jak w przypadku dekonwolucji na okręgu, im gładsza jest gęstość błędu, tym trudniejszy jest problem dekonwolucji.

Wyznaczanie minimaksowych temp zbieżności jest w tym przypadku technicznie trudniejsze niż w przypadku operatorów zwartych, ale wiele wyników jest dostępnych (np. Mair i Ruymgaart, 1996). Gdy np.

$$\mathcal{F} = \left\{ f \in L^2(\mathbb{R}) : \ f - \text{g. probab. i } \int_{-\infty}^{\infty} (1+t^2)^{\beta} |\tilde{f}(t)|^2 \mathrm{d}t \leq C \right\}$$

oraz $h(x) = \exp(-|x|)/2$ jest gęstością rozkładu Laplace'a, to minimaksowe tempo zbieżności ma postać

$$\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}_f \|\hat{f}_n - f\|_{L^2}^2 \asymp n^{-(2\beta-4)/(2\beta+5)}$$

a gdy błąd ma standardowy rozkład normalny z nieskończenie gładką gęstością, to

$$\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}_f \|\hat{f}_n - f\|_{L^2}^2 \asymp (\log n)^{-(\beta-2)}.$$

Estymatory o minimaksowych tempach zbieżności mogą być w tym przypadku skonstruowane przy wykorzystaniu tzw. pseudo-odwrotności operatora, w których $\mathcal{K}^{-1} = U^{-1}M_{1/\tilde{h}}U$ zastępuje się przez $\mathcal{K}_r^{-1} = U^{-1}M_{r(1/\tilde{h})}U$, a rolą funkcji $r(\cdot)$ jest zmodyfikowanie $1/\tilde{h}$ tak, aby była ograniczona.

W przypadku operatora, który nie jest samosprzężony, przekształca się model do postaci $\mathcal{K}^*g = \mathcal{K}^*\mathcal{K}f$ i stosuje twierdzenie spektralne do $\mathcal{K}^*\mathcal{K}$. Estymacja $\mathcal{K}^*g$ jest wtedy zwykle możliwa z parametrycznym tempem zbieżności (por. van Rooij i Ruymgaart, 1999 oraz Mair i Ruymgaart, 1996). Jeżeli np. gęstość $h(\cdot)$ rozkładu błędu w problemie dekonwolucji nie jest symetryczna, to

$$(\mathcal{K}^*g)(x) = \int h(y-x)f(y)\mathrm{d}y$$

i $\sqrt{n}$-zgodnym estymatorem $(\mathcal{K}^*g)(x)$ jest

$$\frac{1}{n}\sum_i h(Y_i - x) = \int h(y-x)\mathrm{d}P_n(y),$$

gdy Y_i są obserwacjami z gęstości g.

5 Procedury-wyrocznie i adaptacyjność

Estymatory o minimaksowych tempach zbieżności zależą zwykle od pewnych parametrów, np. współczynnika w funkcjonale Tichonowa-Phillipsa. Wartości tych parametrów, które minimalizują ryzyko estymatora zależą zwykle od nieznanego estymowanego obiektu. Wartości zapewniające minimaksowe tempo zbieżności są zwykle wyznaczone z dokładnością do mnożnika i zależą od rozważanej klasy obiektów $\mathcal{F}$. W praktyce dobór tych parametrów odbywa się na podstawie danych. Estymator z pewną ustaloną procedurą doboru parametrów na podstawie danych nazywa się adaptacyjnym, jeżeli osiąga on minimaksowe tempo zbieżności nie tylko na jednej, ustalonej i wąskiej klasie $\mathcal{F}$, ale na całej rodzinie takich klas, np. rodzinie klas Sobolewa $\mathcal{F}_{\alpha,C}$ z parametrami α i C zmieniającymi się w pewnym zbiorze. Estymator taki ma więc zdolność adaptacji np. do nieznanej a priori gładkości estymowanej funkcji.

Technicznym narzędziem do konstrukcji estymatorów adaptacyjnych mogą być tzw. procedury-wyrocznie, zdolne (asymptotycznie) wybierać najlepszy spośród skończonego zbioru estymatorów. Niech $\{\hat{f}_\lambda : \lambda \in \Lambda\}$ będzie skończonym zbiorem alternatywnych estymatorów dla f (dla uproszczenia notacji opuszczamy indeks wskazujący na wielkość próby lub poziom szumu). Niech λ^* będzie wartością λ wybieraną na podstawie danych, a $\hat{f}^* = \hat{f}_{\lambda^*}$ będzie odpowiednim estymatorem. Niech $R(\hat{f}, f)$ oznacza ryzyko estymatora $\hat{f}$. Procedurę wyboru λ^* nazwiemy wyrocznią (por. Cavalier i inni, 2002), jeżeli dla wszystkich f z pewnej klasy

$$R(\hat{f}^*, f) \leq (1 + o(1)) \min_{\lambda \in \Lambda} R(\hat{f}_\lambda, f), \tag{11}$$

gdy poziom szumu dąży do zera, gdzie $o(1)$ nie zależy od f, ale zależy od rodziny Λ. Wyrocznia wybiera więc na podstawie danych ten estymator ze skończonego zbioru, który jest najlepszy dla prawdziwej ale nieznanej funkcji.

Prześledzimy szczegóły tego podejścia na przykładzie problemu z operatorem zwartym o SVD postaci $\{s_n, \phi_n, \psi_n\}_{n \in \mathbb{N}}$. Gdy $f = \sum_i \theta_i \phi_i$, to $g = \mathcal{K}f = \sum_i s_i \theta_i \psi_i$ i $< g, \psi_k > = s_k \theta_k$. Załóżmy, że potrafimy estymować/mierzyć $< g, \psi_k >$ z addytywnym błędem $\varepsilon \xi_k$, gdzie ε jest poziomem szumu, a $\xi_k \sim N(0,1)$ są niezależne. Obserwujemy więc $Y_k = s_k \theta_k + \varepsilon \xi_k$, $k \in \mathbb{N}$ lub, po podzieleniu przez s_k, obserwujemy $X_k = \theta_k + \varepsilon \sigma_k \xi_k$, gdzie $\sigma_k = s_k^{-1}$. Rozważmy estymatory liniowe postaci $\hat{f}_\lambda = \sum_k \hat{\theta}_k \phi_k$ z $\hat{\theta}_k = \lambda_k X_k$. Estymator jest więc wyznaczony przez ciąg (w praktyce skończony) $\lambda = (\lambda_1 . \lambda_2, \ldots)$. Niech ryzyko ma postać $R(\hat{f}, f) = \mathbb{E}_f \|\hat{f} - f\|^2$. Wtedy

$$R(\hat{f}_\lambda, f) = \mathbb{E}_f \|\hat{\theta} - \theta\|^2 = \sum_k (1 - \lambda_k)^2 \theta_k^2 + \varepsilon^2 \sum_k \sigma_k^2 \lambda_k^2.$$

Widać z tego, że gdy choć jeden $\lambda_k \notin [0,1]$, to estymator jest niedopuszczalny. Załóżmy więc, że

$$\max_{\lambda \in \Lambda} \sup_k |\lambda_k| \leq 1 \quad \text{i} \quad 0 < \sum_{k=1}^{\infty} \sigma_k^2 \lambda_k^2 < \infty. \tag{12}$$

Adaptacyjny wybór λ ze zbioru Λ może być oparty na heurystycznej zasadzie minimalizacji nieobciążonego estymatora ryzyka. Dla nieznanego θ_k^2 występującego w wyrażeniu na ryzyko, mamy nieobciążony estymator postaci $X_k^2 - \varepsilon^2\sigma_k^2$, co implikuje, że wyrażenie

$$U(\lambda, X) = \sum_{k=1}^{\infty}(\lambda_k^2 - 2\lambda_k)X_k^2 + 2\varepsilon^2\sum_{k=1}^{\infty}\sigma_k^2\lambda_k \tag{13}$$

z $X = (X_1, X_2, \ldots)$ jest nieobciążonym estymatorem $R(\hat{f}_\lambda, f) - \sum_k \theta_k^2$ i wybieramy

$$\lambda^* = \mathrm{argmin}_{\lambda\in\Lambda}U(\lambda, X). \tag{14}$$

Dodatkowo do (12) załóżmy, że istnieje stała $C > 0$ taka, że dla wszystkich $\lambda \in \Lambda$

$$\sum_{i=1}^{\infty}\sigma_i^4\lambda_i^2 \le C\sum_{i=1}^{\infty}\sigma_i^4\lambda_i^4. \tag{15}$$

Wobec $|\lambda_i| \le 1$ oznacza to, że obie sumy są wielkościami tego samego rzędu. Oznaczmy:

$$\rho(\lambda) = \sup_k \sigma_k^2|\lambda_k|\left\{\sum_{i=1}^{\infty}\sigma_i^4\lambda_i^4\right\}^{-1/2}$$

$$\rho = \max_{\lambda\in\Lambda}\rho(\lambda)$$

$$S = \max_{\lambda\in\Lambda}\sup_i \sigma_i^2\lambda_i^2 / \min_{\lambda\in\Lambda}\sup_i \sigma_i^2\lambda_i^2$$

i pozwólmy, aby klasa Λ, a w szczególności jej liczność, mogła zmieniać się z ε. Zauważmy, że w typowych zastosowaniach, gdy poziom szumu ε maleje do zera, to liczność Λ rośnie, a niezerowe mnożniki λ_i zbliżają się do jedynki, co implikuje, że zwykle w takich sytuacjach ρ dąży do zera. Można wtedy pokazać, że procedura (14) ma własności wyroczni (Cavalier i inni, 2002).

Twierdzenie: *Przy założeniach (12) i (15), jeżeli* $\#\Lambda = N$ *i*

$$\lim_{\varepsilon\to 0}\rho^2\log(NS) = 0$$

to (11) zachodzi dla wszystkich $f = \sum_k \theta_k\phi_k$ *takich, że* $\sum_k \theta_k^2 < \infty$.

Przykład: Rozpatrzmy estymatory typu projekcji. Niech $\Lambda = \{\lambda^1, \ldots, \lambda^N\}$

$$\lambda_i^1 = \mathbf{1}(i \le w_1), \lambda_i^2 = \mathbf{1}(i \le w_2), \ldots, \lambda_i^N = \mathbf{1}(i \le w_N), \quad i = 1, 2, \ldots$$

z pewnymi całkowitymi $1 \le w_1, \ldots < w_N$. Załóżmy, że $s_k \asymp k^{-\beta}$. Można wtedy pokazać, że jeżeli $N = N(\varepsilon)$ oraz $w_j = w_j(\varepsilon)$, $j = 1, \ldots, N$ są takie, że

$$\lim_{\varepsilon\to 0}\frac{\log(Nw_N/w_1)}{w_1} = 0$$

to (11) zachodzi dla wszystkich $f = \sum_k \theta_k \phi_k$ takich, że $\sum_k \theta_k^2 < \infty$.

Przykład: Rozpatrzmy estymator Tichonowa-Phillipsa (por. (8)) w przypadku, gdy $s_k \asymp k^{-\beta}$ i, zastępując $< g, \psi_k > / s_k$ przez θ_k, zapiszmy go w postaci

$$\hat{f}_{n,\delta} = (\mathcal{K}^*\mathcal{K} + \delta I)^{-1}\mathcal{K}^* g = \sum_k \frac{s_k}{s_k^2 + \delta} < g, \psi_k > \phi_k = \sum_k \frac{1}{1 + (k/w)^{2\beta}} \theta_k \phi_k.$$

Dostarcza to motywacji do rozpatrywania klasy estymatorów postaci

$$\Lambda = \left\{ \lambda = \{\lambda_k\} : \lambda_k = \frac{1}{1 + (k/w)^{\alpha}},\ w \in \mathcal{W},\ \alpha \in \mathcal{A} \right\}$$

gdzie $\mathcal{A}$ i $\mathcal{W}$ są zbiorami skończonymi. Optymalne wartości parametrów w i α zależą od nieznanego obiektu f i w praktyce muszą być wyznaczane adaptacyjnie. Należy oczekiwać, że $w \to \infty$, gdy $\varepsilon \to 0$.

Gdy $\mathcal{W}$ jest skończonym podzbiorem przedziału $[w_1, w_{max}]$ z $0 < w_1 < w_{max} = \mathcal{O}(\varepsilon^{-b})$, $\#\mathcal{W} = \mathcal{O}(\varepsilon^{-a})$, z ustalonymi, dodatnimi a, b, to jeżeli $\alpha_{min} > 2\beta + 1/2$ i $w_1 / \log^2(1/\varepsilon) \to \infty$, to (11) zachodzi dla wszystkich $f = \sum_k \theta_k \phi_k$ takich, że $\sum_k \theta_k^2 < \infty$.

Zauważmy, że nie ma żadnych ograniczeń na wykładnik a, więc zbiór $\mathcal{W}$ może być dowolnie "gęsty" w rozpatrywanym przedziale, co otwiera drogę do pokazania, że procedura pozostaje wyrocznią także w przypadku, gdy w zmienia się w sposób ciągły w przedziale.

Literatura

Antoniadis A., Bigot J. (2006) Poisson inverse problems. *Ann. Statist.* **34**, 2132-2158.

Birgé L. (1983) Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeit und verwandte Gebiete* **65**, 181-237.

Brown L.D., Low M.L. (1996) Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24**, 2384-2398.

Brown L.D., Carter A.V., Low M.L., Zhang C.H. (2004) Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift. *Ann. Statist.* **32**, 2074-2097.

Chauveau D.E., van Rooij A.C.M., Ruymgaart F.H. (1994) Regularized inversion of noisy Laplace transforms. *Advances in Applied Mathematics* **15**, 186-201.

Cavalier L. (2006) Inverse problems with non-compact operators. *J. Statist. Plann. Inference* **136**, 390-400.

Cavalier L., Golubev G.K., Picard D., Tsybakov A.B. (2002) Oracle inequalities for inverse problems. *Ann.Statist.* **30** , 843-874 .

Cavalier L.,Tsybakov A.B. (2002) Sharp adaptation for inverse problems with random noise. *Probab. Theory Related Fields* **123**, 323-354.

Donoho D.L. (1995) Nonlinear solution of linear inverse problems by wavelet-vaguelet decomposition. *Appl. and Comput. Harmonic Anal.* **2**, 101-126.

Efromovich S. (1999) Nonparametric Curve Estimation. Springer, New York.

Efromovich S., Koltchinskii V. (2001) On inverse problems with unknown operators. *IEEE Trans. Inform. Theory* **47**, 2876-2894.

Engl H.W., Hanke M., Neubauer A. (1996) Regularization of Inverse Problems. Kluver, Dordrecht.

Fano R.M. (1961) Transmission of Information. MIT Press, Massachusetts.

Groeneboom P. (1996) Lectures on Inverse Problems. w: Lectures on Probab. Theory and Statistics, Ecole d'Eté de Probabilités de Saint-Flour XXIV-1994, Springer, Berlin.

Hadamard J. (1932) Le problème de Cauchy et les équations aux dérivées partielles linéaires hyperboliques. Herman, Paris.

Halmos P.R. (1963) What does the spectral theorem say? *Amer. Math. Monthly* **70**, 241-247.

Hampel F.R. (1987) Design, modelling and analysis of some biological datasets., w: Mallows C.L. (ed.) Design, Data and Analysis, by Some Friends of Cuthbert Daniel, Wiley, New York, 111-115.

Ibragimov I.A., Chasminskij R.Z. (1981) Statistical Estimation. Asymptotic Theory. Springer, Berlin.

Ivanov V.V. (1962) On linear problems which are not well-posed. *Soviet Math. Docl.* **3**, 981-983.

Jewell N.P. (1982) Mixtures of exponential distributions. *Ann. Statist.* **10**, 479-484.

Johnstone I. M., Silverman B. W. (1990) Speed of estimation in positron emission tomography and related inverse problems. *Ann. Statist.* **18**, 251-280.

Johnstone I. M., Silverman B. W. (1991) Discretization effects in statistical inverse problems. *J. Complexity* **7**, 1-34.

Kaipio J., Somersalo E. (2005) Statistical and Computational Inverse Problems. Springer, Berlin.

Kołodziej W. (1970) Wybrane Rozdziały Analizy Matematycznej. PWN, Warszawa.

Korostelev A.P., Tsybakov A.B. (1993) Minimax Theory of Image Reconstruction. Lecture Notes in Statistics, v.82, Springer, New York.

Krasnoselskij M.A. (1963) Topological Methods in the Theory of Nonlinear Integral Equations. Pergamon Press, Oxford.

Mair B. A., Ruymgaart F. H. (1996) Statistical inverse estimation in Hilbert scales. *SIAM J. Appl. Math.* **56**, 1424-1444.

Mlak W. (1972) Wstęp do teorii przestrzeni Hilberta, PWN, Warszawa.

Nychka D., Cox D.D. (1989) Convergence rates for regularized solutions of integral equations. *Ann. Statist.* **17**, 556-572.

Reiss, R. D. (1993) A Course on Point Processes. Springer, New York.

Phillips D.Z. (1962) A technique for numerical solution of certain integral equation of the first kind. *J. Assoc. Comput. Mach.* **9**, 85-96.

van Rooij A. C. M., Ruymgaart F. H. (1996) Asymptotic minimax rates for abstract linear estimators. *J. Statist. Plann. Inference* **53**, 389-402.

Solo V. (2000) Limits to estimation in stochastic ill-conditioned inverse problems. *IEEE Transactions on Information Theory* **46**, 1872-1880.

Szkutnik Z. (2000) Unfolding intensity function of a Poisson process in models with approximately specified folding operator. *Metrika* **52**, 1-26.

Szkutnik Z. (2003) Doubly smoothed EM algorithm for statistical inverse problems. *J. Am Statist. Assoc.* **98**, 178-190.

Szkutnik Z. (2005) B-splines and discretization in an inverse problem for Poisson processes. *J. Multiv. Anal.* **93**, 198-221.

Tikhonov A.N. (1963) On solving ill-posed problem and method of regularization. *Doklady Akademii Nauk USSR* **153**, 501-503.

Tikhonov A.N., Arsenin V.Y. (1977) Solution of Ill-Posed Problems. Winston, Washington.

Tsybakov A.B. (2009) Introduction to Nonparametric Estimation. Springer, New York.

Vapnik V.N. (1995) The Nature of Statistical Learning Theory. Springer, New York.

Vapnik V.N., Stefanyuk A.R. (1978) Nonparametric methods for restoring probability densities. *Avtomatika i Telemechanika* **8**, 38-52.

Wahba G. (1977) Spline Models for Observational Data. SIAM, Philadelphia.