Робастно статистическо моделиране чрез тримиране

Нейко М. Нейков
Национален Институт по Метеорология и Хидрология
Българска Академия на Науките

АВТОРЕФЕРАТ

на дисертационен труд
за присъждане на научна степен Доктор на науките

гр. София, 2016

Съдържание

Общо въведение

Статистиката е както наука така и изкуство за извличане на полезна информация от данни. Един ефективен начин за получаване на тази информация е свързан с използването на параметричните стохастични модели. Методите на най-малките квадрати (МНК) и максималното правдоподобие (ММП) са основни методи за получаване на точкови и доверителни оценки, и проверка на хипотези за неизвестните параметри. В основата на тези методи са залегнали строги предположения, които идеализират действителността. При анализ на реални данни възникват противоречия, свързани с качеството на данните. Качеството на данните се характеризира с наличието на несъгласувани (discordant, outlier) наблюдения в тях, вследствие от повреда на апаратурата в процеса на измерване, събиране, въвеждане, предаване, презапис и съхраняване на информационни носители. Като правило многомерните масиви от данни съдържат несъгласувани наблюдения, процентът на които е право пропорционален на размерността на данните. Несъгласуваните наблюдения се отклоняват от мажоритарната, преобладаващата част от наблюденията в данните, но те не винаги могат да бъдат идентифицирани чрез стандартна визуална проверка и едномерни статистически техники. Ефектът от наличието на несъгласуваните наблюдения в данните може да има непредсказуеми последствия върху крайните изводи при анализ на данните с класическите многомерни статистически процедури. Причината за това е, че класическите многомерни статистически процедури се основават на оценките на многомерната средна, ковариационна матрица и параметри на множествената линейна регресия, които не са робастни (устойчиви), спрямо наличието на несъгласувани наблюдения в данните. Понякога, наличието на едно несъгласувано наблюдение, отклоняващо се значително от мажоритарната част на данните, може да промени значително тези оценки, както показват Huber (1981) и Rousseeuw and Leroy (1987). Това е причината за създаването и развитието на робастната статистика като алтернатива на класическата статистика основана на МНК и ММП. Основни монографии по робастна статистика са Huber (1981), Hampel et al. (1986), Rousseeuw and Leroy (1987), Staudte and Sheather (1990),

Atkinson and Riani (2000), Dutter et al. (2003), Atkinson et al. (2004), Hubert et al. (2004), Maronna et al. (2006), Heritier et al. (2009), Huber and Ronchetti (2009) и Farcomeni and Greco (2015). Робастните методи се развиват интензивно, създават се ефективни изчислителни алгоритми и програмни процедури за целите на приложната статистика. Така например, библиотеката robustbase от програмната среда R съдържа голям брой процедури, реализиращи различни типове робастни оценки. Използването им изисква определен професионализъм и знания по робастна статистика.

Според общата теория на робастната статистика Huber (1981), Hampel et al. (1986) и Maronna et al. (2006) една статистика е робастна, ако притежава:

- ограничена функция на влияние;
- ограничена функция на изменение на асимптотичната дисперсия;
- положителна прагова стойност (точка);

Първите две характеристики са локални мерки, докато третата е глобална мярка за робастност на дадена статистика.

Настоящитият дисертационен труд е посветен на статистически оценки с висока прагова точка. Най-общо казано, под прагова точка на една статистика се разбира максималният процент наблюдения, чиито стойности могат да бъдат заместени с произволни стойности, без това да влоши съществено качеството на статистиката. Извадка от наблюдения, в която част от наблюденията са заместени с произволни наблюдения ще наричаме замърсена. Пример за статистика с висока прагова точка е медианата на непрекъснатите едномерни разпределения, тъй като почти 50% от данните могат да бъдат заместени с произволни стойности, без това да влияе на стойността на тази статистика съществено.

0.1 Дефиниция на прагова точка

Нека $\Omega = \{\omega_i \in \mathbb{R}^p, \quad i = 1, \ldots, n\}$ е извадка с обем от n наблюдения, $\tilde{\Omega}_m$ е извадката, получена чрез заместване на кои да е m наблюдения от Ω с произволни стойности. Максималната изместеност (bias) на статистиката T се дефинира като

$$\mathrm{bias}(\mathrm{m}, \mathrm{T}, \Omega) = \sup_{\tilde{\Omega}_m} \|T(X) - T(\tilde{\Omega}_m)\|,$$

където супремумът е взет по всевъзможните замърсени извадаки $\tilde{\Omega}_m$, $\|.\|$ е евклидовата норма.

Definition 0.1 *Donoho and Huber (1983). Праговата точка на статистиката T за крайната извадка* Ω *се дефинира като*

$$\varepsilon_n^*(T,\Omega) = \max\{\frac{m}{n} : \sup_{\tilde{\Omega}_m} \|\mathrm{bias(m,T},\Omega)\| < \infty\}.$$

Тъй като за статистиките T се предполага, че са крайни върху извадката Ω, в литературата по робастностна статистика се използва представянето

$$\varepsilon_n^*(T,\Omega) = \max\{\frac{m}{n} : \sup_{\tilde{\Omega}_m} \|\mathrm{T}(\widetilde{\Omega}_\mathrm{m})\| < \infty\},$$

Това означава, че стойностите на статистиката са подмножество на ограничено множество, когато кои да е m наблюдения на извадката Ω бъдат заместени с произволни.

0.2 Тримирани регресионни оценки

Нека е даден класическият модел на линейна множествена регресия

$$y_i = x_i^T\theta + \varepsilon_i \quad \text{за} \quad i = 1,\ldots,n,$$

където $y_i \in R^1$ и $x_i \in R^p$ са наблюденията на зависимата и предикторните променливи, $\theta \in R^p$ е вектор от неизвестни параметри, ε_i са н.е.р. сл. грешки с очакване $\mathrm{E}(\varepsilon_i) = 0$ и константна дисперсия $\mathrm{var}(\varepsilon_i)$=$\sigma^2 > 0$.

Регресионните остатъци се дефинират като

$$r_i(\theta) := y_i - x_i^T\theta \quad \text{за} \quad i = 1,\ldots,n.$$

Definition 0.2 *Оценката по метод на най-малките квадрати (МНК, LSE) се дефинира като*

$$\widehat{\theta}_{LSE} := \arg\min_\theta \sum_{i=1}^n r_i^2(\theta).$$

Definition 0.3 *Оценката по метод на най-малките модули (МНМ, LAD) се дефинира като*

$$\widehat{\theta}_{LAD} := \arg\min_\theta \sum_{i=1}^n |r_i(\theta)|,$$

Definition 0.4 *(Rousseeuw, 1984)Оценките по метод на най-малката медиана на квадратите (LMS), по метод на най-малкия q–ти квадарат (LQS) и по метод на най-малките тримирани квадрати се дефинират като:*

$$\widehat{\theta}_{MED} := \arg\min_{\theta} \operatorname*{med}_{i} r_i^2(\theta), \tag{1}$$

$$\widehat{\theta}_{LQS} := \arg\min_{\theta} r_{\nu(k)}^2(\theta), \tag{2}$$

$$\widehat{\theta}_{LTS} := \arg\min_{\theta} \sum_{i=1}^{k} r_{\nu(i)}^2(\theta), \tag{3}$$

където $r_{\nu(1)}^2(\theta) \leq r_{\nu(2)}^2(\theta) \leq \ldots \leq r_{\nu(n)}^2(\theta)$ са наредените стойности на $r_i^2(\theta)$ в θ, $\nu = (\nu(1), \ldots, \nu(n))$ е пермутацията на индексите на наблюденията, зависеща от θ, k е параматър на тримиране, удовлетворяващ условието $\lfloor \frac{n+p+1}{2} \rfloor \leq k \leq n$ ако всеки p наблюдения са линейно независими, т.е. всеки $p+1$ са в общо положение.

Definition 0.5 *(Rousseeuw and Leroy, 1987) Оценките по метод на най-малките тримирани модули (LTAD) се дефинира като:*

$$\widehat{\theta}_{LTAD} := \arg\min_{\theta} \sum_{i=1}^{k} f_{\nu(i)}(\theta),$$

където $f_{\nu(1)}(\theta) \leq f_{\nu(2)}(\theta) \leq \ldots \leq f_{\nu(n)}(\theta)$ са наредените стойности на $f_i(\theta) = |r_i(\theta)|$ в θ, k е параматър на тримиране, удовлетворяващ условието $\lfloor \frac{n+p+1}{2} \rfloor \leq k \leq n$ ако всеки p наблюдения са линейно независими, т.е. всеки $p+1$ са в общо положение.

От дефинициите на LMS, LQS и LTS оценките следва, че минимумът се достига за някоя подизвадка с обем k. Целевите функции, дефиниращи тези оценки са непрекъснати, но не са диференцируеми и не притежават единствено решение. Следното представяне, дадено от Krivulin (1992), изяснява комбинаторния характер на тези оценки и връзката им с класическите оценки за регресия, понеже се основава на функциите min и max

$$\min_{\theta} r_{\nu(k)}^2(\theta) = \min_{\theta} \min_{I \in I_k} \max_{i \in I} r_i^2(\theta),$$

където $I = \{i_1, \ldots, i_k\} \in I_k$, а I_k е множеството от всевъзможните подмножества от индекси на $\{1, \ldots, n\}$ с обем k. Понеже обемът на извадката е краен, то за съответната оптимазационна задача е в сила представянето

$$\min_{\theta} r_{\nu(k)}^2(\theta) = \min_{\theta} \min_{I \in I_k} \max_{i \in I} r_i^2(\theta) = \min_{I \in I_k} \min_{\theta} \max_{i \in I} r_i^2(\theta).$$

Аналочен резултат е в сила за $f_{\nu(k)}(\theta)$, където $f_i(\theta) = |r_i(\theta)|$

$$\min_\theta f_{\nu(k)}(\theta) = \min_\theta \min_{I\in I_k} \max_{i\in I} f_i(\theta) = \min_\theta \min_{I\in I_k} \max_{i\in I} |r_i(\theta)| = \min_{I\in I_k} \min_\theta \max_{i\in I} |r_i(\theta)|,$$

което е добре известната оптимзационна задача - минимазиция на Чебишевата норма по всевъзможните подизвадки от k наблюдения.

Подобни представяния са в сила за LTS и LTAD оценките

$$\begin{aligned}
\min_\theta \sum_{i=1}^{k} r^2_{\nu(i)}(\theta) &= \min_\theta \min_{I\in I_k} \sum_{i\in I} r_i^2(\theta) = \min_{I\in I_k} \min_\theta \sum_{i\in I} r_i^2(\theta), \\
\min_\theta \sum_{i=1}^{k} f_{\nu(k)}(\theta) &= \min_\theta \min_{I\in I_k} \sum_{i\in I} f_i(\theta) = \min_\theta \min_{I\in I_k} \sum_{i\in I} |r_i(\theta)| = \min_{I\in I_k} \min_\theta \sum_{i\in I} |r_i(\theta)|.
\end{aligned}$$

Следователно, определянето на тези оценки се свежда до провеждането на $\binom{n}{k}$ минимизационни задачи по МНК, МНМ, Чебишева норма на квадратите и модули на регресионните остатъци. Характерното за тези методи за оценяване е, че изключват информацията на онези $n-k$ наблюдения, които не следват модела. Ясно е, че при голям обем на извадката n, пресмятането на тези оценки е невъзможно. Rousseeuw and van Driessen (2000a) предлагат приближен алгоритъм за определяне на LTS оценките. LTS оценките са предочитани, поради $\sqrt{n}$ състоятелност и асимптотична нормалност пред LMS и LSQ оценките.

Праговите точки на тези оценки са характеризирани в работите на Rousseeuw (1984), Rousseeuw and Leroy (1987) и Hössjer (1994), при предположение за линейна независимост на всеки p наблюдения, т.е., всеки $p+1$ наблюдения са в общо положение. Müller (1995, 1997) и Mili and Coakley (1996) показват, че праговата точка на LTS оценките, дефинирани чрез (3), се определя от параметъра

$$\mathcal{N}(X) := \max_{0\neq\beta\in\mathbb{R}^p} \operatorname{card}\left\{n\in\{1,\dots,N\};\ x_n^\top\beta = 0\right\},$$

където $X = (x_1,\dots,x_N)^\top \in \mathbb{R}^{N\times p}$ е матрицата от наблюденията на предикторните променливи в линеен модел. $\mathcal{N}(X)$ представлява максималния брой повторени наблюдения, лежащи в подпространство на предикторните променливи. Ако наблюденията са в общо положение, т.е., всеки p са линейно независими, тогава $\mathcal{N}(X) = p-1$, което е минималната стойност на $\mathcal{N}(X)$. В други случаи стойността на $\mathcal{N}(X)$ е много по-голяма, например в случаите, когато предикторните променливи са дискретни (категорийни).

Ще отбележим, че праговата точка на тези оценки е характеризирана от Нейков (1995) и Vandev and Neykov (1998), чрез техниката на d–пълнота, предложена от Vandev (1993), при предположение за линейна независимост на наблюденията.

0.3 Тримирани максимално правдоподобни оценки

Нека $x_i \in R^p$ за $i = 1, \ldots, n$ са независими наблюдения с плътност на разпределение $\psi(x, \theta)$, $\theta \subseteq \Theta^q$, където θ е неизвестен параметър и $l_i(\theta) = l(x_i, \theta) = -\log \psi(x_i, \theta)$. Neykov and Neytchev (1990) предлагат регресионните остатъци $r_i^2(\theta)$ в LMS и LTS оценките, дефинирани от Rousseeuw (1984), да бъдат заместени с отрицателните логаритми на правдоподобие $l_i(x_i, \theta)$. Като следствие от това са въведени следните два класа статистически оценки:

Definition 0.6 *(Neykov and Neytchev, 1990) Оценките по метода на минималната медиана от отрицателни логаритми на правдоподобия (LME(k)) и минималната тримирана сума от отрицателни логаритми на правдоподобия (TLE(k)) се дефинират като*

$$\widehat{\theta}_{LME} := \arg\min_{\theta \in \Theta} l(x_{\nu(k)}, \theta) \qquad \text{and} \qquad \widehat{\theta}_{TLE} := \arg\min_{\theta \in \Theta} \sum_{i=1}^{k} l(x_{\nu(i)}, \theta),$$

където $l(x_{\nu(1)}, \theta) \leq l(x_{\nu(2)}, \theta) \leq \ldots \leq l(x_{\nu(n)}, \theta)$ са наредените стойности на $l(x_i, \theta)$ за $i = 1, \ldots, n$ в θ, $\nu = (\nu(1), \ldots, \nu(n))$ е съответната пермутация от индекси на наблюденията, която зависи от θ и k е параметър на тримиране.

Основната идея за тези оценки се състои в изключването на онези $n - k$ наблюдения, които е малко вероятно да бъдат наблюдавани, при условие че избрания модел е истинския. TLE съвпада с оценките по метода на максималното правдоподобие (МПО), когато $k = n$. От комбинаторния характер на дефиницията следва представянето

$$\min_{\theta \in \Theta} \sum_{i=1}^{k} l(x_{\nu(i)}, \theta) = \min_{\theta \in \Theta} \min_{I \in I_k} \sum_{i \in I} l(x_i, \theta) = \min_{I \in I_k} \min_{\theta \in \Theta} \sum_{i \in I} l(x_i, \theta)$$

където $I = \{i_1, \ldots, i_k\} \in I_k$, а I_k е множеството от всевъзможните подмножества от индекси на $\{1, \ldots, n\}$ с обем k. Следователно TLE се дефинира като МПО по някоя подизвадка от всевъзможните $\binom{n}{k}$ на брой подизвадки.

Vandev (1993) разглежда по-широк клас от функции, частен случай на които са регресионните остатъци по модул, в квадрат или (отрицателни) логаритми от плътности на разпределение. Нека $f : X \times \Theta \to \mathbb{R}^+$, където $\Theta \subseteq \mathbb{R}^q$ е отворено множество $F = \{f_i(\theta) = f(x_i, \theta), \text{ за } i = 1, \ldots, n\}$.

Definition 0.7 *(Vandev, 1993) Обобщена Медианна Оценка (GMedE(k)) и Обобщена Тримирана Оценка (GTE(k)) се дефинират като*

$$\widehat{\theta}^{\mathrm{k}}_{\mathrm{GMedE}} := \arg\min_{\theta\in\Theta^q} f_{\nu(k)}(\theta) \qquad и \qquad \widehat{\theta}^{\mathrm{k}}_{\mathrm{GTE}} := \arg\min_{\theta\in\Theta} \sum_{i=1}^{k} f_{\nu(i)}(\theta)$$

където $f_{\nu(1)}(\theta) \leq f_{\nu(2)}(\theta) \leq \ldots \leq f_{\nu(n)}(\theta)$ са наредените във възходящ ред стойности на $f_i(\theta)$ в θ, $\nu = (\nu(1), \ldots, \nu(n))$ е съответната пермутация на индексите на наблюденията, която зависи от θ, k е параметър на тримиране.

За да характеризира праговата точка на тези оценки, Vandev (1993) предлага техниката на d–пълнота, основана на следните две дефиниции:

Definition 0.8 *(Vandev, 1993) Множеството F се нарича d–пълно ако за всяко подмножество $J \subset \{1, \ldots, n\}$ с кардиналност d ($|J| = d$) функцията $\varphi(\theta) = \max_{j\in J} f_j(\theta)$, $\theta \in \Theta$ е субкомпактна.*

Definition 0.9 *(Vandev, 1993) Реалната функция $\varphi : \Theta \to \mathbb{R}$, $\Theta \subseteq \mathbb{R}^q$ се нарича субкомпактна ако Лебеговото множество $L_{\varphi(\theta)}(C) = \{\theta : \varphi(\theta) \leq C\}$ е компактно множество за всяка константа C.*

Proposition 0.1 *(Vandev, 1993) Ако множеството $F = \{f_1, \ldots, f_n\}$ е d–пълно, то F е $(d+1)$–пълно.*

Theorem 0.1 *(Vandev, 1993) Праговата точка на $GTE(k)$ оценките $\varepsilon^*_n(GTE(k)) \geq \frac{n-k}{n}$ ако $F = \{f_1, \ldots, f_n\}$ е d–пълно множество от неотрицателни непрекъснати функции, $n \geq 3d$ и $\frac{n+d}{2} \leq k \leq n - d$.*

Neykov (1995) въвежда следния клас оценки и характеризира праговата им точка с техниката на d–пълнота

Definition 0.10 *(Neykov, 1995; Vandev and Neykov, 1998). Претеглени Обобщени Тримирани Оценки $wGTE(k)$ се дефинират като:*

$$\widehat{\theta}_{wGTE} := \arg\min_{\theta\in\Theta} \sum_{i=1}^{k} w_{\nu(i)} f_{\nu(i)}(\theta), \tag{4}$$

където $f_{\nu(1)}(\theta) \leq f_{\nu(2)}(\theta) \leq \ldots \leq f_{\nu(n)}(\theta)$ са наредените стойности на $f_i(\theta)$ в θ, във възходящ ред, $\nu = (\nu(1), \ldots, \nu(n))$ е съответната пермутация на индексите на наблюденията, която зависи от θ, k е параметър на тримиране, теглото $w_i \geq 0$ е свързано с f_i, за $i = 1, \ldots, n$ и $w_{\nu(k)} > 0$.

От комбинаторния характер на дефиницията (4) следва представянето

$$\widehat{\theta}_{wGTE} := \arg \min_{\theta \in \Theta^p} \min_{I \in I_k} \sum_{i \in I} w_i f_i(\theta), \tag{5}$$

където $I = \{i_1, \ldots, i_k\} \in I_k$, а I_k е множеството от всевъзможните подмножества от индекси на $\{1, \ldots, n\}$ с обем k. Следователно $wGTE(k)$ оценката се дефинира като претеглена МПО по някоя подизвадка от всевъзможните $\binom{n}{k}$ на брой подизвадки.

Proposition 0.2 *(Neykov, 1995) Ако множеството $F = \{f_1, \ldots, f_n\}$ от неотрицателни непрекъснати функции е d–пълно, то (5) е непразно компактно множество за всяко $k \geq d$.*

Theorem 0.2 *(Neykov, 1995; Vandev and Neykov, 1998) Праговата точка на $wGTE(k)$ оценките $\varepsilon_n^*(wGTE(k)) \geq \frac{n-k}{n}$ ако $F = \{f_1, \ldots, f_n\}$ е d–пълно множество от неотрицателни непрекъснати функции, $n \geq 3d$ и $\frac{n+d}{2} \leq k \leq n - d$.*

Proposition 0.3 *(Neykov, 1995) Нека $f, g : \Theta \to \mathbb{R}$, $\Theta \subseteq \mathbb{R}^q$, f е субкомпактна функция и $f(\theta) \leq g(\theta)$ за всяко $\theta \in \Theta$. Тогава $g(\theta)$ е субкомпактна функция.*

Частен случай на wGTE(k) оценките са WTLE(k) оценките за $f_i(\theta) = f(x_i, \theta) = -\log \psi(x_i, \theta)$. Основните резултати за праговата точка на wGTE(k), а така също за праговата им точка в линейни регресионни модели и модел на логистична регресия при линейно независими наблюдения са разгледани от Vandev and Neykov (1998).

ЗАБЕЛЕЖКА 1: В дисертационния труд се използват следните означения: GTE(k) вместо S(k), въведено от Vandev (1993); wGTE(k) вместо $R(k)$ и W_k, въведени от Neykov (1995), Vandev and Neykov (1998) и Dimova and Neykov (2004).

0.4 Структура на дисертационния труд

Дисертационният труд се състои от 8 глави, следващи съдържанието на статиите с номера [1], [2], [4], [5], [6], [7] и [8] от приложения списък с публикации по-долу. Глава 3 се основава на резултатите в статиите с номера [3], [9] и [10]. Публикацията [11] представлява първоначален вариант на статията [5]. В апендиксите към Глави 4 и 5 са формулирани и дадени доказателствата на две твърдения, характеризиращи праговите точки на WTLE(k) оценките за вероятностни модели, които са предмет на изследване в тези глави.

[1] Müller, Ch. and Neykov, N. M. (2003). Breakdown Points of the Trimmed Likelihood and Related Estimators in Generalized Linear Models. *J. Statist. Plann. and Inference*, **116**, 503-519. **IF: 0.307**.

[2] Neykov, N. M. and Müller, Ch. (2003). Breakdown Point and Computation of Trimmed Likelihood Estimators in Generalized Linear Models. In: *Developments in Robust Statistics*, Dutter, R., Filzmoser, P., Gather, U., and Rousseeuw, P. (eds.), Physica-Verlag, Heidelberg, 277-286.

[3] Dimova, R. and Neykov, N. M. (2004a). Generalized d-fullness Technique for Breakdown Point Study of the Trimmed Likelihood Estimator with Applications. In: *Theory and Applications of Recent Robust Methods*, M. Hubert, G. Pison, A. Struyf and S. Van Aelst (eds.), Birkhauser, Basel, 83-92.

[4] Neykov, N.M., Dimova, R. and Neytchev, P.N. (2005). Trimmed Likelihood Estimation of the Parameters of the Generalized Extreme Value Distribution: A Monte-Carlo Study. *Pliska Stud. Math. Bulgar.*, **17**, 187-200.

[5] Neykov, N. M., Filzmoser, P., Dimova, R. and Neytchev, P. N. (2007). Robust fitting of mixtures using the Trimmed Likelihood Estimator. *Comput. Statist. Data Anal.*, **52**, 299-308. **IF: 1.029**.

[6] Neykov, N. M., Filzmoser, P. and Neytchev, P. N. (2012). Robust joint modeling of mean and dispersion through trimming. *Comput. Statist. Data Anal.* **56**, 34-48. **IF: 1.304**.

[7] Neykov, N. M., Čížek, P., Filzmoser, P. and Neytchev, P.N. (2012). The least trimmed quantile regression. *Comput. Statist. Data Anal.* **56**, 1757-1770. **IF: 1.304**.

[8] Neykov, N. M., Filzmoser, P. and Neytchev, P. N. (2014). Ultrahigh dimensional variable selection through the penalized maximum trimmed likelihood estimator. *Stat. Papers*, **55**, 187-207. **IF: 0.813**.

[9] Dimova, R. and Neykov, N.M. (2003). Generalized d-fullness Technique for Breakdown Point Study of the Trimmed Likelihood Estimator. *Compt. rend. Acad. Bulg. Sci.*, Tome **56**, No 5, 7-12.

[10] Dimova, R. and Neykov, N.M. (2004b). Application of the d-fullness Technique for Breakdown Point Study of the Trimmed Likelihood Estimator to a generalized Logistic Model. *Pliska Stud. Math. Bulgar.*, **16**, 35-41.

[11] Neykov, N.M., Filzmoser, P., Dimova, R. and Neytchev, P.N. (2004). Mixture of Generalized Linear Models and the Trimmed Likelihood Methodology. In: *Proceedings in Computational Statistics*, J. Antoch (ed.), Physica-Verlag, 1585-1592.

0.5 Предмет и цели на дисертационния труд

Основен предмет на настоящия дисертационен труд са робастните методи с висока прагова точка, основани на тримиране. Въведени са класове робастни оценки, като: S–оценките, основани на минориране - можориране от k-ти нареден елемент на d–пълно множество от неотрицателни функции, тримирани квази-правдоподобни оценки, тримирани квантилни оценки, тримирани (правдоподбни) оценки с пенализация в задачи с висока размерност. Предложени са обобщения на техниката на d–пълнота на Vandev (1993), за да бъдат характеризирани праговите точки на някои от изброените класове оценки, основани на тримиране. Предложени са FAST-TLE алгоритъм за приближено пресмятане на тримирани правдоподобни оценки и FAST-GTE алгоритъм за приближено пресмятане на обобщени тримирани оценки.

Крайните цели на дисертацията са:

- създаване на унифицирана методология за робастно оценяване с висока прагова точка, за разкриване и редуциране на несъгласувани наблюдения в данните, на основата на тримиране на най-широко използваните класически методи за статистическо моделиране:

  а) с обобщени линейни модели със случайна компонента от дисперсионната фамилия от разпределения, частен случай на която е линейната експоненциална фамилия от разпределения;

  б) с множествена линейна квантилна регресия;

  в) със смеси от разпределения от линейната експоненциална фамилия от разпределения;

  г) с разпределения на екстремалните стойности;

  д) в задачи с големи размерности (броя на предикторите е по-голям от наблюденията) за селектиране на значими предиктори с методите на пенализацията (регуляризация) за разпределения от линейната експоненциална фамилия от разпределения;

- изследване поведението на изброените класове оценки и сравняването им с класическите оценки върху крайни извадки от данни с и без наличие на несъгласувани наблюдения в данните чрез методите на имитационното моделиране.

Обем на дисертацията

Дисертацията съдържа 200 машинописни страници, включващи увод, 8 глави и цитирана литература. Написана е на английски език.

В изложението е използвана двойна номерация. Първата цифра означава номер на глава, а втората пореден номер на релация, определение, лема, теорема, следствие или забележка.

Благодарности

- Изказвам благодарност на ръководствата на НИМХ-БАН в лицето на чл. кор. В. Андреев, проф. дфн Д. Сираков, проф. д-р В. Спиридонов, доц. д-р Г. Корчев и проф. д-р Хр. Брънзов, а също така на доц. д-р Пламен Н. Нейчев в качеството му на Директор на департамент "Прогнози на времето" за предоставяне на възможности и съдействие в научно - изследователската ми дейност по робастна статистика, както и на всички, които искаха да видят успешен завършек на моя труд;

- Изказвам благодарност на доц. д-р Пламен Н. Нейчев, НИМХ-БАН и Prof. Dr. Peter Filzmoser, Institute of Statistics, Vienna University of Technology, за дългосрочното и ползотворно сътрудничество в областта на изчислителните аспекти на робастната статистика, основана на тримиране. Без техните знания в областта на числените методи на статистиката, опит и умения в създаването на статистически софтуер за научни изследвания, резултатите от симулационните експерименти не биха били на необходимата висота;

- Изказвам благодарност на Prof. Dr. Christine Müller, Fakultät Statistik, TU Dortmund, за ползотворно съвместно сътрудничество, относно развитието на техниката на d-пълнота за характеризиране на праговата точки на оценките, основани на тримиране;

- Изказвам благодарност на Dr. Росица Димова, докторант през 2002-2005г. във ФМИ, СУ Св. "Кл. Охридски" за ползотворното сътрудничество в областта на оценяването на параметрите на смеси от разпределения с тримираните правдоподобни оценки, създаването и развитието на обобщената техниката на d-пълнота за характеризиране на праговата на оценките, основани на тримиране;

- Изказвам благодарност на Prof. Dr. Pavel Čížek, Tilburg School of Economics and Management Econometrics and Operations Research, Tilburg University, за ползот-

ворното сътрудничество, свързано с асимптотичното поведение на тримираните квантилни линейни регресионни оценки;

- Бих искал да изкажа голямата си признателност и благодарност към доц. д-р Димитър Л. Вындев, който предложи техниката на d–пълнота за характеризиране на праговата точка на статистическите оценки, основани на тримиране. Този дисертационен труд не би изглеждал по този начин без неговият значим принос в робастната статистика.

- Накрая, но не на последно място, бих искал да изкажа благодарност и признателност към родителите ми Мария Александрова и Матей Нейков Марков, на тях посвещавам този дисертационен труд.

Глава 1

Прагова точка на тримирани правдоподобни оценки и свързаните с тях обобщени линейни модели

1.1 Въведение

В секция 1.2 е даден най-общият резултат за праговата стойност на обобщените тримирани оценки wGTE, дефинирани за d–пълно множество от функции. Характеризирана е праговата точка на TLE оценките. Определена е връзката на индекса на d– пълнота, дефиниран от Vandev (1993), чре параметъра $\mathcal{N}(X)$, дефиниран от Müller (1997). Като следствие от това са характеризирани праговата точка на линейния регресионен модел, линейната логистична регресия и лог-линейната (Поасонова) регресия, съответно в секциии 1.4 и 1.5. Секция 1.6 е посветена на линейните модели с разпределение на Лаплас от ред q. На основата на техниката на d– пълнота е характеризирана праговата точка на регресионните S оценки, дефинирани от Rousseeuw and Yohai (1984) и Rousseeuw and Leroy (1987). Дадени ни са опростени доказателства на резултатите на Vandev (1993) и Vandev and Neykov (1998).

1.2 Прагова точка на wGTE(k) оценките

Нека $y = (y_1, \ldots, y_N)^\top$ са независими наблюдения на сл. вел Y с плътност на разпределение $f_n(y_n, \theta)$, където θ е неизвестен параметър. Нека $l_n(y, \theta) = -\log f_n(y_n, \theta)$ и $l(y, \theta) = (l_1(y, \theta), \ldots, l_N(y, \theta))^\top$.

Нека Θ е топологично пространство и $\mathrm{int}(\Theta)$ е множеството от вътрешните точки

на Θ, $\mathcal{Y}_M(y) := \left\{\overline{y} \in \mathcal{Y}^N;\ \text{card}\{n;\ y_n \neq \overline{y}_n\} \leq M\right\}$ е множеството от всевъзможните замърсени извадки с не повече от M наблюдения.

Definition 1.1 *Праговата точка на оценката* $\widehat{\theta} : \mathcal{Y}^N \to \Theta$ *в* $y \in \mathcal{Y}^N$ *се дефинира като*

$$\epsilon^*(\widehat{\theta}, y) := \quad \frac{1}{N}\min\{M;\ \nexists \text{ компактно множество } \Theta_0 \subset \text{int}(\Theta),\ \text{за което } \{\widehat{\theta}(\overline{y});\ \overline{y} \in \mathcal{Y}_M(y)\} \subset \Theta_0\}.$$

Ако $\Theta = I\!R^p$, то $N \cdot \epsilon^*(\widehat{\theta}, y)$ е минималният брой M от замърсени наблюдения, за който $\{\widehat{\theta}(\overline{y});\ \overline{y} \in \mathcal{Y}_M(y)\}$ е неограничено множество. В определени случаи праговата точка е $\epsilon^*(\widehat{\theta}, y) = 0$, когато не съществува единствена оценка $\widehat{\theta}(y)$. Така например МПО не съществуват ако максимумът на функцията на правдоподобие $\prod_{n=1}^N f_n(y_n, \theta)$ се достига за няколко различни стойности на θ. За подобни случаи казваме, че ако $\gamma(\theta) = \sum_{n=1}^N l_n(y, \theta)$, то праговата точка е нула, което означава че $\{\theta \in \Theta;\ \gamma(\theta) \leq C\}$ не се съдържа в компактно подмножество на $\text{int}(\Theta)$ за всяка константа $C \geq \min_\theta \gamma(\theta)$. Поради тази причина се предлага едно по-слабо условие за субкомпактност на една функция.

Definition 1.2 *Функцията* $\gamma : \Theta \to I\!R$ *се нарича субкомпактна ако множеството* $\{\theta \in \Theta;\ \gamma(\theta) \leq C\}$ *се съдържа в компактно множество* $\Theta_C \subset int(\Theta)$ *за всяка константа* $C \in I\!R$.

В дефиницията за субкомпактност, дадена от Vandev and Neykov (1993) се предполага компактност на множеството $\{\theta;\ \gamma(\theta) \leq C\}$, което е ограничение за задачите, които са предмет на изследване в тази глава. Следната лема характеризира праговата точка на МПО оценка $\widehat{\theta}$ на θ.

Lemma 1.1 *Ако* $\gamma(\theta) = \sum_{n=1}^N l_n(y, \theta)$ *е субкомпактна функция, тогава* $\epsilon^*(\widehat{\theta}, y) > 0$.

Понеже $\max_{n=1,\ldots,N} l_n(y, \theta) \leq \sum_{n=1}^N l_n(y, \theta) \leq N \max_{n=1,\ldots,N} l_n(y, \theta)$ то Лема 1.1 ще бъде вярна ако $\gamma(\theta) = \max_{n=1,\ldots,N} l_n(y, \theta)$. За да характеризираме праговата точка на МПО върху подизвадки, ще използваме Дефиниция 0.8 за d–пълнота, предложена от Vandev (1993). Ще отбележим, че d–пълнотата на $\{l_n(y, \cdot);\ n = 1, \ldots, N\}$ осигурява положителна прагова точка на МПО за всяка подизвадка от d наблюдения.

Ще въведем следния клас $S(y)$ оценки, основани на минориране - мажориране и ще характеризираме праговата им точка.

Definition 1.3 *$S(y)$ оценката се дефинира като*

$$S(y) := \arg\min_{\theta\in\Theta} s(y,\theta),$$

където $s : \mathcal{Y}^N \times \Theta \to I\!R$, съществуват константи $\alpha,\ \beta \in I\!R$, $\alpha \neq 0$ такива, че

$$\alpha\, l_{(h)}(y,\theta) \le s(y,\theta) \le \beta\, l_{(h)}(y,\theta) \tag{1.1}$$

за всяко $y \in \mathcal{Y}^N$, $\theta \in \Theta$ и $h \in \{1,\dots,N\}$.

Theorem 1.1 *Ако множеството $\{l_n(y,\cdot);\ n = 1,\dots,N\}$ е d–пълно и условието (1.1) е удовлетворено, тогава за праговта точка на S оценките е в сила неравенството*

$$\epsilon^*(S,y) \ge \frac{1}{N}\,\min\{N-h+1, h-d+1\}.$$

Доказателството е основано на следната

Lemma 1.2 *Ако $\{l_n(y,\cdot);\ n = 1,\dots,N\}$ е d–пълно, $M \le N-h$ и $M \le h-d$, тогава $l_{(d)}(y,\theta) \le l_{(h)}(\overline{y},\theta) \le l_{(N)}(y,\theta)\ \ \forall\ \ \overline{y} \in \mathcal{Y}_M(y)$ и $\theta \in \Theta$.*

Ще отбележим, че Теорема 1.1 е обобщение на Theorem 1 на Vandev and Neykov (1998) за праговата точка на wGTE оценките, без налагане на допълнителното условие $N \ge 3\,d$ и $(N+d)/2 \le h \le N-d$. Тази прагова точка достига максимум, когато параметъра h на тримиране удовлетвоярва условието $\left\lfloor \frac{N+d}{2} \right\rfloor \le h \le \left\lfloor \frac{N+d+1}{2} \right\rfloor$, където $\lfloor z \rfloor := \max\{n \in I\!N;\ n \le z\}$.

Theorem 1.2 *Нека $\{l_n(y,\cdot);\ n = 1,\dots,N\}$ е d–пълно и $\left\lfloor \frac{N+d}{2} \right\rfloor \le h \le \left\lfloor \frac{N+d+1}{2} \right\rfloor$. Тогава праговата точка на $wGTE(h)$ оценките удовлетворяя неравенството*

$$\epsilon^*(wGTE(h),y) \ge \frac{1}{N}\left\lfloor \frac{N-d+2}{2} \right\rfloor.$$

ЗАБЕЛЕЖКА: Резултатите от тази секция са в сила за произволно d–пълно множество от положителни функции, не само за $l_n(y,\theta) = -\log f_n(y_n,\theta)$, поради което Теорема 1.2 е формулирана в термините на $wGTE(h)$ оценките. В статията Müller and Neykov (2003) теоремата е формулирана в термините на $WTLE(k)$ оценките, което не отразява степента на обобщеност.

1.3 Приложение в обобщените линейни модели от линейната експоненциална фамилия от разпределения

В тази секция е определен индексът на d–пълнота за линейната експоненциална фамилия от разпределения с константен дисперсионен параметър. Нека Y_n е сл. вел. с разпределение от линейната експоненциална фамилия

$$f(y_n, x_n, \beta) = \exp\{T(y_n)^\top g(x_n^\top \beta) + c(x_n^\top \beta) + b(y_n)\},$$

където $T : \mathcal{Y} \to I\!R^r$, $g : I\!R \to I\!R^r$, $c : I\!R \to I\!R$, and $b : \mathcal{Y} \to I\!R$ са известни функции с $\mathcal{Y} \subset I\!R^q$, $x_n \in \mathcal{X} \subset I\!R^p$, $n = 1, \ldots, N$, са известни предикторни променливи и $\beta \in I\!R^p$ е неизвестен параметър. В този случай за $-\log f(y_n, x_n, \beta)$ получаваме

$$l_n(y, X, \beta) = -T(y_n)^\top g(x_n^\top \beta) - c(x_n^\top \beta) - b(y_n),$$

където $X = (x_1, \ldots, x_n)^\top$.

Показано е, че индексът на пълнота на тези функции $\{l_1(y, X, \cdot), \ldots, l_N(y, X, \cdot)\}$. се дефинира чрез параметъра $\mathcal{N}(X)$.

Lemma 1.3 *Нека $X \in I\!R^{N\times p}$, $I \subset \{1, \ldots, N\}$ и* card(I) $= \mathcal{N}(X) + 1$. *Тогава $\{\beta \in I\!R^p;\ \max_{i\in I} |x_i^\top \beta| \le D\}$ е ограничено за всяка константа $D \in I\!R$.*

Theorem 1.3 *Ако функцията γ_z зададена чрез $\gamma_z(\theta) = -T(z)^\top g(\theta) - c(\theta) - b(z)$ е субкомпактна за всяко $z \in \mathcal{Y}$ тогава множеството $\{l_n(y, X, \cdot);\ n = 1, \ldots, N\}$ е $\mathcal{N}(X)+1$ пълно за всяко $y \in \mathcal{Y}^N$ и всяко $X \in \mathcal{X}^N$.*

Като следствие от тези резултати е разгледан линейният модел с нормално разпределение на грешката. Показано е, че съответното множество от отрицателни логаритми на правдоподобие е субкомпактна функция и са удовлетворени условията на Теорема 1.3. Следователно съгласно Теорема 1.1 праговата точка на WTLE(h) оценките на класическия линеен модел с нормално разпределение на грешката е $\frac{1}{N}\left\lfloor \frac{N-\mathcal{N}(X)+1}{2} \right\rfloor$, понеже тези оценки са афиинно еквивариантни и $\frac{1}{N}\left\lfloor \frac{N-\mathcal{N}(X)+1}{2} \right\rfloor$ е една горна граница за праговата точка.

Ще отбележим, че Vandev and Neykov (1998) характеризират праговата точка на WTLE(h) оценките, но при предположение, че $x_1, \ldots, x_N$ са в общо положение.

Теорема 1.3 заедно с Теорема 1.1 дават само долна граница за праговата точка. Следващата Лема характеризира праговата точка на МП оценки за линейната експоненциална фамилия.

Lemma 1.4 *Ако праговата точка на МП оценки удовлетворява условието*

$$\epsilon^*(ML, y, X) \leq \frac{1}{N}$$

$\forall\ y \in \mathcal{Y}^N$ *и* $X \in \mathcal{X}^N$*, тогава праговата точка на* $WTLE(h)$ *удовлетворява неравенството*

$$\epsilon^*(WTLE(h), y, X) \leq \frac{1}{N}(N - h + 1).$$

Условието $\epsilon^*(ML, y, X) \leq \frac{1}{N}$ на Лема 1.4 трябва да бъде доказвано за всеки модел от линейната експоненциална фамилия от разпределения.

1.4 Линейна логистична регресия

Следната теорема характеризира праговите точки на WTLE(h) оценките за линейна логистична (бинарна) регресия.

Theorem 1.4 *Праговата точка на* $WTLE(h)$ *оценките на линейната логистична регресия е*

$$\min_{y \in \mathcal{Y}^*} \epsilon^*(WTLE(h), y, X) = \frac{1}{N}\ \min\{N - h + 1, h - \mathcal{N}(X)\}.$$

Ще отбележим, че $\mathcal{N}(X) = p - 1$, когато $x_1, \ldots, x_N$ са в общо положение, долната граница на праговата точка е намерена от Vandev and Neykov (1998) при допълнителното ограничение $N \geq 3(p + 1)$.

1.5 Лог-линеен регресионен модел

Следната теорема характеризира праговата точка на WTLE(h) оценките за лог-линеен регресионен модел (линейна Поасонова регресия) .

Нека Y_n е честотно разпределена сл. вел. с разпределение на Поасон с параметър $\lambda_n = \exp(x_n^\top \beta)$, чиито логаритъм на функцията на правдоподобие е

$$l_n(y, X, \beta) = -y_n\, x_n^\top \beta + \exp(x_n^\top \beta) + \log(y_n!).$$

Theorem 1.5 *Праговата точка на $WTLE(h)$ оценките на лог-линеен модел е*

$$\min_{y\in\mathcal{Y}^*}\epsilon^*(WTLE(h),y,X)=\frac{1}{N}\,\min\{N-h+1,h-\mathcal{N}(X)\}.$$

Резултатите от тази секция са илюстрирани с два примера за логистична и Поасонова регресия с дискретни предиктори, т.е. наблюденията не са в обща позиция.

1.6 Разпределение на Лаплас от ред q с дисперсионен параметър

Нека наблюденията $Y_n,\ n=1,\ldots,N$ са Лапласово разпределени от ред q с плътност

$$f(y_n,x_n,\beta,\sigma)=\frac{q\,(1/2)^{(1+1/q)}}{\sigma\,\Gamma(1/2)}\,\exp\left(-\frac{1}{2}\left|\frac{y_n-x_n^\top\beta}{\sigma}\right|^q\right\},$$

където $\Gamma(.)$ е гама функцията, $\beta\in I\!R^p$ и $\sigma\in I\!R^+$.

Частни случаи на това разпределение са нормалното ($q=2$), стандартното Лапласово ($q=1$), двойното експоненциално ($0<q<2$), leptokurtic ($1<q<2$), platikurtic ($q>2$) и правоъгълното ($q\to\infty$) разпределения. Следната Лема характеризира параметъра на d–пълнота на множеството

$$\{l_n(y,X,\beta,\sigma)=-\log f(y_n,x_n,\beta,\sigma);\ n=1,\ldots,N\} \tag{1.2}$$

Lemma 1.5 *За всяко $q>0$ параметърът на d–пълнота на множеството от функции (1.2) е $\mathcal{N}(X)+1$.*

Vandev and Neykov (1998) показват, че $d=\mathcal{N}(X)=p-1$ при ограничителното предположение всеки p наблюдения $x_1,\ldots,x_N$ са линейно независими. От Теорема 1.1 и Лема 1.5 следва, че $WTLE(h)$ оценката за (β,σ) има прагова точка не по-малка от $\frac{1}{N}\min\{N-h+1,h-\mathcal{N}(X)\}$, която достига максимална стойност при $\frac{1}{N}\left\lfloor\frac{N-\mathcal{N}(X)+1}{2}\right\rfloor$ ако $\left\lfloor\frac{N+\mathcal{N}(X)+1}{2}\right\rfloor\le h\le\left\lfloor\frac{N+\mathcal{N}(X)+2}{2}\right\rfloor$.

Друг основен резултат в тази секция се отнася за праговата точка на линейните регресионни S_c оценки, предложени от Rousseeuw and Yohai (1984), които се дефинират като

$$S_c(y,X):=\arg\min_\beta s_c(y,X,\beta),$$

където $s_c(y, X, \beta)$ е неявно решение на

$$\frac{1}{N}\sum_{n=1}^{N}\rho_c\left(\frac{|y_n - x_n^\top\beta|}{s_c(y, X, \beta)}\right) = K.$$

Rousseeuw and Leroy (1987) твърдят, че S_c оценката достига максимална прагова стойност при следните условия: ρ_c е строго монотонна в интервала $[0, c]$ и константа в интервала $[c, \infty)$, а $x_1, \ldots, x_N$ са в общо положение. Изводите на тези автори се основават на неравенството

$$\alpha_1\, l_{(h)}(y, X, \beta) \le s_c(y, X, \beta) \le \alpha_2\, l_{(h)}(y, X, \beta), \tag{1.3}$$

от което заключават, че праговите точки на $LQS(h)$ оценката, дефинирана чрез $l_{(h)}(y, X, \beta)$ и $s_c(y, X, \beta)$ оценката съвпадат.

Доказателството на Теорема 1.1 показва, че са необходими допълнителни аргументи, основани на техниката на d–пълнота, а именно че множеството от регресионни остатъци $|y_n - x_n^\top\beta|$ за $n = 1, \ldots, N$ е $(\mathcal{N}(X)+1)$–пълно, за да бъде вярно твърдението на тези автори. В сила е следната

Theorem 1.6 *Праговата точка на S_c оценката, удовлетворява неравенството $\epsilon^*(S_c, y, X) \ge \frac{1}{N}\min\{N - h + 1, h - \mathcal{N}(X)\}$, като равенство се достига при $\lfloor (N + \mathcal{N}(X) + 1)/2 \rfloor \le h \le \lfloor (N + \mathcal{N}(X) + 2)/2 \rfloor$.*

1.6.1 Резюме

В глава 1 е дефиниран клас S статистически оценки, основани на минориране - мажориране на k-тия нареден елемент на d–пълно множество от функции. Основен резултат в тази глава е Теорема 1.1, характеризираща праговата точка на S оценките. Теорема 1 на Vandev and Neykov (1998), характеризираща праговата точка на $wGTE(k)$ оценките, представлява частен случай на Theorem 1.1, понеже $wGTE(k)$ оценките са частен случай на S оценките. Без налагане на допълнителни условия за обема на извадката и параметъра на тримиране k е намерена по-ниска граница за праговата точка на $wGTE(k)$ оценките. Определена е връзката на индекса на d– пълнота, дефиниран от Vandev (1993), с параметъра $\mathcal{N}(X)$, дефиниран от Müller (1997). Характеризирана е праговата точка на линейния регресионен модел, линейната логистична регресия и лог-линейната (Поасонова) регресиия и линейни регресионни модели с Лапласово разпределение на грешката от ред q. С техниката на d–пълнота е характеризирана праговата точка на линейните регресионни S_c–оценки на Rousseeuw and Leroy (1987). Предложени са опростени доказателства на резултатите на Vandev (1993) и Vandev and Neykov (1998).

Глава 2

Прагова точка и приближено пресмятане на TLE(k) оценките за обобщени линейни модели

Основен резултат в глава 2 е FAST-TLE алгоритъмът за приближено определяне на $WTL(k)$ оценката. Наименованието FAST-TLE алгоритъм е свързано с FAST-LTS алгоритъма, предложен от Rousseeuw and Van Driessen (1999a), понеже са еквивалентни над линейния регресионен модел с Гаусово разпределение на грешката.

2.1 FAST-TLE алгоритъм за приближено пресмятане на TLE оценки

Нека множеството $F = \{f(y_i, \theta) \quad \text{за} \quad i = 1, \ldots, n\}$ е d–пълно и $k \geq d$.

1. Нека $H^{old} = \{y_{j_1}, \ldots, y_{j_m}\} \subset \{y_1, \ldots, y_n\}$.
2. Нека $\widehat{\theta}^{old}$ бъде МПО на θ, основана на H^{old} и $Q^{old} := \sum_{i=1}^{m} f(y_{j_i}, \widehat{\theta}^{old})$;
3. Нека $f(y_{\nu(1)}, \widehat{\theta}^{old}) \leq \cdots \leq f(y_{\nu(k)}, \widehat{\theta}^{old}) \leq \cdots \leq f(y_{\nu(n)}, \widehat{\theta}^{old})$ са наредените във възходящ ред стойности на $f(y_i, \widehat{\theta}^{old})$ за $i = 1, \ldots, n$ и $\nu = (\nu(1), \ldots, \nu(n))$ е съответната пермутация;
4. Нека $\widehat{\theta}^{new}$ е МПО на θ основана на $H^{new} := \{y_{\nu(1)}, \ldots, y_{\nu(k)}\}$;
5. Нека $Q^{new} := \sum_{i=1}^{k} f(y_{\nu(i)}, \widehat{\theta}^{new})$ и $H^{old} := H^{new}$

6. Да формираме цикъл по точки 2-5;

При тези предположения е в сила следното твърдение

Proposition 2.1 $Q^{new} \leq Q^{old}$.

Стъпките 1-6 на FAST-TLE алгоритъма се наричат *C–step*, където C означава концентрация, тъй като H^{new} се основава на наблюдения, които са по-концентрирани около локалния максимум на целевата функция. Чрез цикъла на *C–step* се дефинира итерационен процес. Когато $Q^{new} = Q^{old}$ този процес се прекратява. Дискутирани са начините за избор на наблюденията и обемът на H^{old} извадката, които се основават на резултатите от Müller and Neykov (2003).

Разгледани са примери на обобщени линейни модели като - бинарна линейна логистична и лог-линейна (Поасонова) регресия, илюстриращи теоритичните резултати от глава 1 за $TLE(k)$, включително определянето на индекса на d–пълнота
$\mathcal{N}(X) := \max_{0 \neq \beta \in I\!R^p} \text{card}\left\{n \in \{1, \ldots, N\};\ x_n^\top \beta = 0\right\}$.

Програмната реализация на FAST-TLE алгоритъма се основава на стандартните програмни процедури за оценяване, достъпни в широко разпространените програмни статистически пакети, тъй като е върху подизвадки с различни обеми на извадката.

2.1.1 Резюме

В глава 2 е предложен FAST-TLE алгоритъм за приближено пресмятане на $TLE(k)$ оценките. Стъпката на концентрация на този алгоритъм е еквивалентна на стъпките на концентрация на FAST-LTS и FAST-MCD алгоритмите, предложени от Rousseeuw and Van Driessen (1999a) и Rousseeuw and Van Driessen (1999b) за приближено пресмятане на параметрите на линейната множественна регресия с нормално разпределена грешка, многомерната средна и ковариационна матрица на многомерното нормално разпределение, съответно.

Глава 3

Обобщена техника на d–пълнота за изследване на праговата точка на wGTE оценките и приложения

В тази глава е дадено обобщение на техниката на d–пълнота, предложена от Vandev (1993) и разширена от Müller and Neykov (2003) за изучаване на праговата точка на $wGTE(k)$ оценките в ситуации на комплексни вероятностни модели.

3.1 Въведение

Условието в дефиницията за субкомпактност на една функция *за всяка реална константа* C е твърде ограничително и не винаги е изпълнено. Пример за това е множеството F от отрицателните логаритми от крайна смес от вероятностни разпределения.

Нека функцията $g : \Theta \to \mathbb{R}$, $\partial\Theta$ е множеството от гранични точки на Θ, $\Theta_\infty = \{\{\theta_k\}_{k=1}^\infty : \theta_k \in \Theta, \|\theta_k\| \to \infty\}$ е множеството на всички неограничени по норма редици, $\underline{g}$ е дефинирана като

$$\underline{g} = \begin{cases} \inf\limits_{\theta^* \in \partial\Theta} \liminf\limits_{\theta_k \to \theta^*} g(\theta_k), & \text{ако } \Theta \text{ е ограничено, или} \\ \\ \inf\limits_{\theta^* \in \partial\Theta} \liminf\limits_{\substack{\theta_k \to \theta^* \\ \{\theta_k\} \in \Theta_\infty}} g(\theta_k), & \text{ако } \Theta \text{ е неограничено.} \end{cases} \tag{3.1}$$

Да въведем следните условия:

A1. $F = \{f_i(\theta) \geq 0,\ i = 1, \ldots, n,\ \text{for } \theta \in \Theta\}$ е множеството от непрекъснати неотрицателни функции;

A2. Съществува $\theta_0 \in \Theta$, че за всяко подмножество $J \subset \{1, \ldots, n\}$ с кардиналност d, $c^{**} g_J(\theta_0) < C$, където $g_J(\theta) = \max_{j \in J} f_j(\theta)$ и $C = \inf_J \underline{g_J}$, $c^{**} = \frac{c^*}{w^*}$, $c^* = \sum_{i=1}^{k} w_{\nu(i)}$, и $w^* = \min\{w_j > 0,\ j = 1, \ldots, n\}$.

Забележка: Класът на d–пълните множества от функции е частен случай на класът от множества от функции, удовлетворяващи условията A1 и A2, тъй като $\underline{g} = \infty$, откъдето следва че g е субкомпактна функция съгласно

Lemma 3.1 *Нека $g : \Theta \to \mathbb{R}$ е непрекъсната функция, $\Theta \subseteq \mathbb{R}^q$ е отворено множество. Ако съществува $\theta_0 \in \Theta$ и реална константа $a \geq 1$, такава че $ag(\theta_0) < C$, където $C \leq \underline{g}$, тогава множеството $S = \{\theta : g(\theta) < C\}$ е ограничено и непразно.*

Следващото предложение 3.1 представлява необходимо условие за съществуване на решение на wGTE(k) оптимизационната задача (4).

Proposition 3.1 *Ако $k \geq d$, A1 и A2 са в сила, тогава $\widehat{\theta}_{wGTE(k)}$ е непразно компактно множество.*

Праговата точка на wGTE(k) оценките за множеството от функции, удовлетворяващи условията A1 и A2, е характеризирана от предложение 3.2, което представлява обобщение на Теорема 1 на Vandev and Neykov (1998), които изискват d–пълнота на множеството F.

Proposition 3.2 *Ако A1 и A2 са в сила, тогава праговата точка на* wGTE(k) *оценките е не по-малка от $\frac{1}{n}\min(n-k, k-d)$.*

Твърдението е в сила за $WTLE(k)$ оценките, които са частен случай на wGTE(k) оценките. Следователно изучаването на праговата точка на $WTLE(k)$ оценките за конкретно вероятностно разпределение $\phi(x_i, \theta)$ се свежда до проверка за валидност на условията A1 и A2 на съответното множество от функции $f_i(\theta) = -\log \phi(x_i, \theta)$ за $i = 1, \ldots, n$. По този начин праговата точка може да бъде изследвана за интервал от стойности на $k \in [d, n]$.

Приложение към модел на обобщена логистична линейна регресия С помощта на обобщената техника за d–пълнота е определена праговата точка на WTLE оценките за модел на групова бинарна линейна логистична регресия с обобщена свързваща функция. Данните са от следния тип (y_i, x_i^T) за $i = 1, \ldots, N$, където y_i е биномно разпределена, $b(y_i \mid n_i, \pi_i)$, n_i е броят на наблюденията с вероятността за успех π_i, x_i е p–мерен вектор от предиктори. Общият брой на наблюденията е $n = n_1 + n_2 + \cdots + n_N$.

Ще предполагаме, че $0 < y_i < n_i$ за всяко i, и π_i е дефинирана като обобщено логистично разпределение на Prentice (1976)

$$\pi_i = (1 + \exp(-\eta_i))^{-a},$$

където $a > 0$, $\eta_i = x_i^T \beta$ е линейния предиктор и β е p–мерен вектор от неизвестни параметри. Частен случай на това разпределение при a=1 е разгледан в глава 1, Müller and Neykov (2003).

Множеството от функции $F = \{f(y_i, \eta_i, a) = -\log \binom{n_i}{y_i} + y_i a \log(1 + e^{-\eta_i}) - (n_i - y_i) \log(1 - (1 + e^{-\eta_i})^{-a}),\ i = 1, \ldots, N\}$ удовлетворва условията A1 и A2. Като следствие от това и Предложение 3.2 получаваме

Proposition 3.3 *Множеството* $\{f(y_i, x_i, \beta, a), i = 1, \ldots, N\}$ *е* $\mathcal{N}(X) + 1$*–пълно.*

Corollary 3.1 $WTLE(k)$ *оценките на групов бинарен линеен модел с обобщена логистична свързваща функция са непразно компактно множество, ако* $k \geq \mathcal{N}(X) + 1$.

От тези резултати и Теорема 1.2 на Müller and Neykov (2003) следва

Corollary 3.2 *Ако* $\lfloor (N + \mathcal{N}(X) + 1)/2 \rfloor \leq k \leq \lfloor (N + \mathcal{N}(X) + 2)/2 \rfloor$*, тогва праговата точка на* $WTLE(k)$ *оценките на групов бинарен линеен модел с обобщена логистична свързваща функция е*

$$\varepsilon_N^*(WTL_k) \geq \frac{1}{N} \left\lfloor \frac{N - \mathcal{N}(X) + 1}{2} \right\rfloor.$$

3.1.1 Резюме

Глава 3 е посветена на обобщената техника на d–пълнота за определяне на праговата точка, понеже условието в дефиницията за субкомпактност на една функция *за всяка реална константа* C е твърде силно изискване, което не винаги е изпълнено. За преодоляване на този проблем Dimova and Neykov (2003), Dimova and Neykov (2004a), and Dimova and Neykov (2004b) предлагат обобщена техника на d–пълнота за изучаване на праговата точка на $wGTE(k)$ оценките в ситуации на комплексни вероятностни модели. Основни резултати са Proposion 3.1 и 3.2 , които дават необходими условия за съществуване на решение на съответните оптимизационни проблеми и долна граница на праговата точка на $wGTE(k)$ за множество от функции, удовлетворяващи условията A1 и A2. Тези резултати представляват обобщение на Теорема 1 на Vandev and Neykov (1998). Proposion 3.3 характеризира праговата точка за линейната логистична (бинарна) регресия с обобщена свързваща функция, с което са обобщени резултатите на Vandev and Neykov (1998) и Müller and Neykov (2003), отнасящи се до линейната логистична (бинарна) регресия.

Глава 4

TLE оценяване на параметрите на GEV разпределението чрез тримиране: Изследване с Монте-Карло методи

4.1 Въведение

Тази глава е посветена на TLE(k) оценяване на параметрите на регресионни модели със зависима променлива, която има обобщено разпределение на екстремалните стойности (GEV). Характеризирана е праговата точка на $TLE(k)$ оценките за разпределението на Гумбел, което е частен случай на GEV разпределението. С методите Монте-Карло е изследвано поведението на МПО и TLE(k) оценките по данни с и без замърсяване с несъгласувани наблюдения в зависимата и предикторните променливи с различни проценти на тримиране. Необходимите пресмятания са проведени в програмнта среда R, на основата на FAST-TLE алгоритъма с библиотеката *ismev*, създадена от Coles (2001) и адаптирана за R от Stephenson (2002).

Твърде малък брой статии третират робастно оценяване на параметрите на разпределенията на екстремалните стойности. Dupuis and Field (1998), Dupuis and Tawn (2001) и Dupuis and Morgenthaler (2002) третират този въпрос чрез робастните M-оценки с ограничена функция на влияние, предложени от Hampel et al. (1986). Праговата точка на тези оценки е $1/p$ и клони към нула с нарастването на броя на предикторните променливи p, което прави този подход неизползваем за целите на приложната статистика.

4.2 Основни дефиниции

Обобщеното разпределение на екстремалните стойности се дефинира като

$$G(x;\mu,\sigma,\xi) = \begin{cases} \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} & \text{ако} \quad \xi \neq 0, \\ \exp\left\{-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right\} & \text{ако} \quad \xi = 0. \end{cases}$$

където $\{x : 1+\xi(x-\mu)/\sigma > 0\}$, $\sigma > 0$, и μ, σ, ξ са параметри на локацията (положението), мащаба и формата, Coles (2001). Разпределенията на Фреше и Вайбул се получават като частни случаи при $\xi < 0$ и $\xi > 0$, съответно, докато $\xi = 0$ се интерпретира като гранично при $\xi \to 0$ и е известно като разпределние на Гумбел. МПО са регулярни при $\xi > -0.5$, оценката съществува, но не е регулярна за $-1 < \xi < -0.5$ и не съществува за $\xi < -1$, според Smith (1985).

Индекса на пълнота d на множеството от отрицателни логаритми от плътността на разпределението на Гумбел е 2. Когато, локационния параметър μ_i на разпределението на Гумбел е монотонна функция от линеен предиктор $\mu_i = h(z_i^T\beta)$, където $\beta \in R^p$ е вектор от неизвестни параметри, $Z := (z_i^\top)$ е матрицата от наблюденията $z_i \in R^p$ на предикторните променливи и всеки p наблюдения са линейно независими, тогава индекса на пълнота е равен на $d = p$.

4.3 Симулационен експеримент

Поведението на МПО и TLE(k) оценките е изследвано с методите Монте Карло в експерименти с регресионен модел, чиято зависима променлива е GEV разпределена. Регулярните и несъгласуваните наблюдения следват моделите, съответно

$$y_i \sim \text{GEV}(\mu_i = 1 + x_i, \sigma = 1, \xi = 0.0), \text{ където } \quad x \sim N(0,7),$$

$$y_i \sim \begin{cases} \text{U}(y_{max}+\mu_i, y_{max}+(y_{max}-y_{min})+\mu_i) & \text{ако} \quad x_i \geq \bar{x}, \\ \text{U}(y_{min}-(y_{max}-y_{min})-\mu_i, y_{min}-\mu_i) & \text{ако} \quad x_i < \bar{x}. \end{cases}$$

В проведените експерименти е използвана извадка с обем $n = 100$, като нивата на замърсяване с несъгласувани наблюдения са 0%, 10%, 20%, 30% и 40% от обема на извадката, докато процентите на тримиране $\frac{n-k}{n}100\%$ е 0%, 5%,10%, . . . ,40%, 45%. Поведението на МПО и TLE(k) оценките е изследвано при всевъзможните комбинации на извадки без и със замърсяване и процент на тримиране. Разпределението на оценките се основава на 400 независими повторения на всеки симулационен експеримент и са представени с бокс-плотове на фиг.4.1-4.5.

4.4 Резултати от симулационните експерименти

На панелните плотове на фиг.4.1-4.4 са представени резултатите от някои експерименти. Забелязва се, че МПО са безсмислени при наличие на несъгласувани наблюдения в данните, които не следват модела. TLE(k) оценките се надеждни и се стабилизират, когато $\frac{n-k}{n}100\% \geq \alpha$, където α е процента на замърсяване.

Групата от бокс-плотове на фиг. 4.5 дава представа за разпределението на МПО и TLE оценките на параметрите на GEV разпредлението: локация (μ_i, свободен член и наклон, мащаб (σ, scale) и форма (ξ, shape) при различни проценти на замърсяване и тримиране, получени в 400 независими симулационни експеримента. Вариациите в отделните бокс-плотове е по-голяма при голям процент на замърсяване и малък процент на тримиране и обратно. Наблюдава се стабилизиране на оценките за определени проценти на тримиране, след което се забелазва значително по-голяма вариабилност.

4.5 Резюме

Глава 4 е посветена на TLE(k) оценяване на параметрите на регресионни модели със зависима променлива, която има обобщено разпределение на екстремалните стойности (GEV). Характеризирана е праговата точка на TLE(k) оценките за разпределението на Гумбел, частен случай на GEV разпределението. С методите Монте-Карло е изследвано поведението на МПО и TLE(k) оценките по данни с и без замърсяване с несъгласувани наблюдения в зависимата и предикторните променливи при различни проценти на тримиране. Използван е FAST-TLE алгоритъмът за приближено пресмятане на неизвестните регресионни параметри.

4.6 Апендикс към глава 4:

Плътността на разпределение на Гумбел е

$$\phi(y;\mu,\sigma) = \frac{1}{\sigma}\exp\left[-\left(\frac{y-\mu}{\sigma}\right)\right]\exp\left\{-\exp\left[-\left(\frac{y-\mu}{\sigma}\right)\right]\right\}$$

Нека $\mu_i = \eta_i = z_i^T\beta$, където $Z_{n\times p} = (z_i^T)_{i=1}^n$ е матрицата от наблюденията на предикторите с ранг p. Ясно е, че всеки p наблюдения на матрицата $Z_{n\times p}$ са линейно независими, понеже са свързани с екстремалните стойноисти y_i, които се случват твърде рядко. Ще покажем, че множеството $F = \{f(y_i;\eta_i,\sigma) = -\log(\phi(y_i;\eta_i,\sigma)$ за $i = 1,\ldots,n\}$ удов-

летворява условията A1 и A2 от глава 3. От $\lim_{\eta_i \to \pm} f(y_i; \eta_i, \sigma) = +\infty$, $\lim_{\sigma \to \infty} f(y_i; \mu, \sigma) = +\infty$ и $\lim_{\sigma \to 0} f(y_i; \mu, \sigma) = +\infty$ следва, че $f(y_i; \eta_i, \sigma)$ е субкомпактна функция, понеже $\underline{f} = +\infty$

Proposition 4.1 *Множеството* $\{f(y_i, \beta, \sigma), i = 1, \ldots, n\}$ *е* p*–пълно.*

Прилагайки Theorem 2 на Müller and Neykov (2003) получаваме

Corollary 4.1 *Ако* $\lfloor (n + p + 1)/2 \rfloor \leq k \leq \lfloor (n + p + 2)/2 \rfloor$*, тогава праговта точка на TLE(k) оценките на регресионните параметри на локационния параметър с линк функция идентитет на разпределението на Гумбел удовлетворява неравенството*

$$\varepsilon_n^*(TLE(k)) \geq \frac{1}{n} \left\lfloor \frac{n - p + 1}{2} \right\rfloor .$$

Глава 5

Робастно оценяване на смеси от разпределения чрез TLE оценките

5.1 Въведение

Смесите от вероятностни разпределения се използват за моделиране на нееднородни данни. Стандартен метод за оценяване на неизвестните параметри на сместа от разпределения е ММП чрез ЕМ алгоритъма, McLachlan and Peel (2000). ММП не е робастен спрямо наличието на несъгласувани наблюдения в данните. За преодоляване на този пробем са предлагани различни параметрични алтернативи, на основата на робастните М-оценки, например Huber (1981) и Hampel et al. (1986).

Директното използване на робастните оценки в условията на нееднородни данни в общия случай е ограничено. Причината за това е, че тези оценки са създадени за оценяване на неизвестните параметри, използвайки наблюденията от мажоритарна част на данните, която е не по-малка от 50% от данните, докато останалите наблюдения, които не следват модела биват третирани като несъгласувани наблюдения. Като правило при реални данни с комплексна структура мажоритарната част е много по-малка от 50%.

Благодарение на ЕМ алгоритъма, оценяването по ММП на параметрите на крайна смес от m вероятностни разпределения за моделиране на комплексни структури в данните се редуцира до оценяване по претеглен ММП на параметрите на отделните m вероятностни разпределения по данните. Заместването на процедурите за МП оценяване с робастни процедури, основани на М-оценките за параметрите на локационно-скалната фамилия от разпределния, в средата на ЕМ алгоритъма, даде възможност за редуциране влиянието на несъгласуваните наблюдения в данните. Подробности са

дадени в Campbell (1984), Kharin (1996), Davé and Krishnapuram (1997), Medasani and Krishnapuram (1998), McLachlan and Peel (2000), Hennig (2003). Недостатък на M-оценки е, че не са робастни спрямо наличието на несъгласувани наблюдения в предикторните променливи, което ги прави неизползваеми за целите на приложнатата статистика и анализ на данни.

По този начин, след години на паралелно развитие на моделирането на нееднородни данни със смеси от разпределения, клъстерни методи за групиране на нееднородни данни, методи за разкриване на несъгласувани наблюдения в данни и робастно оценяване за моделиране на мажоритарната част на данните, се породи необходимостта от синтез на някои от тези методи за анализ на данни с разпределения извън локационно-скалната фамилия от разпределния. Такъв синтез би могъл да бъде гъвкав и мощен инструмент за ефективно моделиране и анализиране на нееднородни данни.

Основната цел на тази глава е да направи стъпка към постигането на тази цел, предлагайки единен подход на основата на TLE оценяването на неизвестните параметри при моделирането на нееднородни данни със смес от вероятностни разпределения от линейната експоненциална фамилия от разпределения. Предимствата на този подход в сравнение с ММП са илюстрирани чрез симулационни експерименти с Монте Карло методите.

В секция 2 са дадени дефинициите на wGTE(k) и WTLE(k) оценките и техните свойства. В секция 3 е даден кратък обзор на моделирането със смеси от разпределения, използвайки EM алгоритъма и са дискутирани възможностите за привличане на робастни методи за оценяване. Храктеризирана е праговата точка на WTLE(k) в средата на смеси от разпределения с техниката на d-пълнота. Дискутирани са принципите, на които трябва да бъде подчинен софтуерът, реализиращ FAST-TLE алгоритъма за смеси от разпределения, чрез процедурите за оценяване от библиотеката FlexMix, реализираща EM алгоритъма. В секция 4 по резултатите от симулационни експерименти е направен сравнителен анализ на ММП и TLE оценяването на модели, които са смеси от линейни регресионни модели с нормално разпределение на грешката, смеси от лог-линейни модели с Поасоново разпределение и смес от двумерни нормални разпределения с нееднородни ковариационни матрици.

5.2 Методология на тримираната функция на правдоподобие

В тази секция е изложен EM алгоритъмът за оценяване на неизвестните параметри на смес от вероятностни разпределения. Подробно разглеждане е дадено от McLachlan and Peel (2000).

МПО и EM алгоритъм за смеси от разпределения. Нека (y_i, x_i^T) за $i = 1, \ldots, n$ е извадка от н.е.р. наблюдения, такива че y_i принадлежи на смес от условни вероятностни разпределения $\psi_1(y_i; x_i, \theta_1), \ldots, \psi_g(y_i; x_i, \theta_g)$ на $x_i \in I\!R^p$ в пропорции $\pi_1, \ldots, \pi_g$, дефинирана като

$$\varphi(y_i; x_i, \Psi) = \sum_{j=1}^{g} \pi_j \psi_j(y_i; x_i, \theta_j), \tag{5.1}$$

където $\Psi = (\pi_1, \ldots, \pi_{g-1}, \theta_1, \ldots, \theta_g)^T$ е вектор от неизвестни параметри. Пропорциите удовлетворяват съотношението $\pi_j > 0$ за $j = 1, \ldots, g$, и $\sum_{j=1}^{g} \pi_j = 1$. МПО на Ψ се дефинира като максимум на логаритъма на правдоподобие

$$\log L(\Psi) = \sum_{i=1}^{n} \log\Big\{\sum_{j=1}^{g} \pi_j \psi_j(y_i; x_i, \theta_j)\Big\}. \tag{5.2}$$

МПО на Ψ съществува и принадлежи на компактно множестово при определени предположения за $\psi_j(y_i; x_i, \theta_j)$ за $j = 1, \ldots, g$. EM алгоритъмът е стандартната техника за намирането на МПО на Ψ, тъй като директното максимизиране не е уместно, поради вида на функцията (5.2). Предполага се, че всяко наблюдение (y_i, x_i^T) е свързано с ненаблюдаемо състояние $z_i = (z_{i1}, z_{i2}, \ldots, z_{ig})^T$ за $i = 1, \ldots, n$, където z_{ij} е 1 или 0, в зависимост от това дали y_i принадлежи или не принадлежи на j-та компонента на сместа. Третирайки вектора на пълните данни (y_i, x_i^T, z_i^T) като известен (наблюдаван), неговата функция на правдоподобие се дефинира като $P(y_i, x_i, z_i) = P(y_i, x_i|z_i)P(z_i) = \prod_{j=1}^{g} \psi_j(y_i; x_i, \theta_j)^{z_{ij}} \pi_j^{z_{ij}}$. Следователно за пълния набор от данни функцията на правдоподобие се дефнира като

$$\log L_c(\Psi) = \sum_{i=1}^{n} \sum_{j=1}^{g} z_{ij} \{\log \pi_j + \log \psi_j(y_i; x_i, \theta_j)\}. \tag{5.3}$$

Разглеждайки z_{ij} като "липсващо" наблюдение, EM алгоритъмът свежда задачата до провеждането на следните две стъпки, наречени E-стъка и M-стъпка, съответно от очакване и максимизиране. На $(l+1)$-та итерация на E-стъпката се пресмята условното очакване на логаритъма на правдоподобие на пълните данни, при дадени (y_i, x_i^T) и

текущата оценка на $\Psi^{(l)}$ of

$$Q(\Psi;\Psi^{(l)}) = \sum_{i=1}^{n}\sum_{j=1}^{g}\tau_j(y_i;x_i,\Psi^{(l)})\{\log\pi_j + \log\psi_j(y_i;x_i,\theta_j)\}, \tag{5.4}$$

където $\tau_j(y_i;x_i,\Psi^{(l)}) = \pi_j^{(l)}\psi_j(y_i;x_i,\theta_j^{(l)})/\sum_{h=1}^{g}\pi_h^{(l)}\psi_h(y_i;x_i,\theta_h^{(l)})$ е текущата оценка на апостериорните вероятности за принадлежност на y_i на jта компонента на сместа. Функцията $Q(\Psi;\Psi^{(l)})$ минорира $\log L(\Psi)$, т.е., $Q(\Psi;\Psi^{(l)}) \leq \log L(\Psi)$ и $Q(\Psi^{(l)};\Psi^{(l)}) = \log L(\Psi^{(l)})$. М-стъпката на $(l+1)$-та итерация максимизира $Q(\Psi;\Psi^{(l)})$ по Ψ, вследствие на което се получава текущата оценка $\Psi^{(l+1)}$. Тези две стъпки се провеждат в последователен ред в хода на итерационния процес до достигане на сходимост.

Оптимизационната задача (5.4) се опростява, понеже се състои от две функции. Първата зависи само от $\pi_1,\ldots,\pi_{g-1}$, докато втората зависи само от $\theta_1,\ldots,\theta_g$. Като следствие от това, априорните вероятности π_j се обновяват както следва

$$\pi_j^{(l+1)} = \frac{1}{n}\sum_{i=1}^{n}\tau_j(y_i;x_i,\Psi^{(l)}), \tag{5.5}$$

докато функцията

$$\max_{\theta_1,\ldots,\theta_g}\sum_{i=1}^{n}\sum_{j=1}^{g}\tau_j(y_i;x_i,\Psi^{(l)})\log\psi_j(y_i;x_i,\theta_j), \tag{5.6}$$

се максимизира по θ_j, използвайки апостериорните вероятности $\tau_j(y_i;x_i,\Psi^{(l)})$ като априорни тегла.

При предположение, че θ_j за $j=1,\ldots,g$ нямат общи координати, изразът (5.6) се максимизира по всяка компонента,

$$\max_{\theta_j}\sum_{i=1}^{n}\tau_j(y_i;x_i,\Psi^{(l)})\log\psi_j(y_i;x_i,\theta_j), \quad \text{за } j=1,\ldots,g. \tag{5.7}$$

Когато векторите θ_j имат общи координати, тогава се използват различни техники за превръщането на двойното сумиране в израза (5.6) в единична сума. Подробности са дадени в McLachlan and Peel (2000).

Класификационен EM алгоритъм. Този подход се състои в присвояване на наблюдението (y_i, x_i^T) към h-та компонента, ако $\tau_h(y_i;x_i,\Psi^{(l)}) \geq \tau_j(y_i;x_i,\Psi^{(l)})$ за $j=1,\ldots,g$. При равни апостериорни вероятности, дадено наблюдение се присвоява случайно към една от компонентите на сместа. Следователно вместо (5.7) се максимизира изразът

$$\max_{\theta_j}\sum_{i=1}^{n_j}\log\psi_j(y_i;x_i,\theta_j) \qquad \text{за } j=1,\ldots,g, \tag{5.8}$$

където n_j е обемът на j-тия клъстер и $n_1 + n_2 + \ldots + n_g = n$. Това е добре известен алгоритъм за класификация от тип "k–means" , който е сходящ за краен брой итерации, но получените оценки не са състоятелни, и не са МПО (McLachlan and Peel (2000)). Въпреки това те се използват за начални оценки в ЕМ алгоритъма.

Изразите (5.7) и (5.8), представляват стандартни МП задачи. По този начин ЕМ алгоритъмът декомпозира комплексни МП задачи в стандартни, значително опростени задачи, за които съществуват стандартни програмни средства за оценяване в широко разпространените статистически пакети.

Прагова точка на WTLE оценките за смес от разпределения. Като следствие на ЕМ алгоритъма, праговата точка на WTLE(k) оценките може да бъде характеризирана чрез праговата точка на тримираното условно очакване на пълната функция на правдоподобие (5.4), дефинирана като

$$\min_{\Psi} \min_{I \in I_k} \sum_{i \in I} \sum_{j=1}^{g} -\tau_j(y_i; x_i, \Psi^{(l)}) \{\log \pi_j + \log \psi_j(y_i; x_i, \theta_j)\}. \tag{5.9}$$

При предположение, че θ_j за $j = 1, \ldots, g$ нямат общи координати, това е еквивалентно на минимизирането на отделните g компоненти по всевъзможните подизвадки с обем k от n наблюдения:

$$\min_{I \in I_k} \begin{cases} \pi_j^{(l+1)} = \frac{1}{k} \sum\limits_{i \in I} \tau_j(y_i; x_i, \Psi^{(l)}) \quad \text{for} \quad j = 1, \ldots, g \\ \min\limits_{\theta_1} \sum\limits_{i \in I} -\tau_1(y_i; x_i, \Psi^{(l)}) \log \psi_1(y_i; x_i, \theta_1) \\ \ldots \\ \min\limits_{\theta_g} \sum\limits_{i \in I} -\tau_g(y_i; x_i, \Psi^{(l)}) \log \psi_g(y_i; x_i, \theta_g) \end{cases} \tag{5.10}$$

В тази секция е разгледана само праговата точка на WTLE(k) оценките на векторните параметри θ_j за $j = 1, \ldots, g$ при предположение, че нямат общи координати. Индексът на пълнота на множеството $F_{\theta_j} = \{-\log \psi_j(y_i; x_i, \theta_j) \quad i = 1, \ldots, n\}$ е d_j за $j = 1, \ldots, g$. Определянето на този индекс е стандартна задача. За да съществува оценка $\widehat{\theta}_j$ на θ_j за $j = 1, \ldots, g$, на всяка една компонента на сместа са необходими поне d_j наблюдения. Тогава поне gd наблюдения гарантират съществуването на решение на задачата (5.10), където $d = \max(d_1, \ldots, d_g)$. Следователно, задачата (5.9) ще има решение, ако k^* и k бъдат избирани в интервала $[gd, n]$, където k^* е броя на наблюденията в опитната стъпка на FAST-TLE алгоритъма. Ако k удовлетворява $\lfloor (n + gd)/2 \rfloor \leq k \leq \lfloor (n + gd + 1)/2 \rfloor$, тогава праговата точка ще достига максималната си стойност, равна на $\frac{1}{n} \lfloor (n - gd)/2 \rfloor$.

Индексите d_j са равни ако $\psi_j(y_i; x_i, \theta_j)$ за $j = 1, \ldots, g$ са от една вероятностна фамилия от разпределения, например, $d_j = p$ за p-мерното нормално разпределение, според Vandev and Neykov (1993). Праговата точка на смес от p-мерни нормални разпределения с нееднородни ковариационни матрици е равна на $\frac{1}{n}\lfloor(n - gp))/2\rfloor$, докато индексът на d–пълнота за p-мерни нормални разпределения с еднородни ковариационни матрици е gp, от което следва че праговата точка на WTLE(k) оценките е равна на $\frac{1}{n}\lfloor(n - gp)/2\rfloor$. Според Vandev and Neykov (1993), WTLE(k) оценките се редуцират до MCD (Rousseeuw, 1985) оценките на ковариационната матрица при $g = 1$. В тези случаи праговата точка на WTLE(k) оценките съвпада с тази на MCD оценките $\frac{1}{n}\lfloor(n - p)/2\rfloor$. Този извод е верен за смес от линейни регресионни модели с нормално разпределение на грешката и лог-линейни регресионни модели с Поасоново разпределени данни, в които наблюденията на предикторните променливи са в общо положение. Ако наблюденията не са в общо положение, което е по-често срещаният случай в средата на обобщените линейни модели, то индексът на d–пълнота е $g(\mathtt{N}(X) + 2)$, където $\mathtt{N}(X)$ е дефинирано в Müller and Neykov (2003).

Робастно оценяване на смеси от разпределения. За приближено намиране на WTLE(k) оценките се използва $FAST - TLE$ алгоритъмът.

Програмни настройки за смеси от разпределения и FAST-TLE алгоритъма. Понеже двете стъпки на FAST-TLE алгоритъма са стандартни МП процедури за оценяване, то реализирането на FAST-TLE алгоритъма не изисква специални средства за програмното му реализиране.

WTLE методологията е илюстрирана с три примера на смеси от линейни регресионни модели с нормално разпределение на грешката, от двумерни нормални разпределения и смес от две Поасонови линейни регресии. Използвана е програмната библиотека FlexMix в програмната среда R, предложена от Leisch (2004). Чрез FlexMix могат да бъдат оценявани по ММП параметрите на произволна крайна смес от разпределения от линейната експоненциална фамилия, чрез EM алгоритъма.

За смес от g разпределения, обемът на опитната извадка k^* задължително трябва да бъде поне gp, за получаване на неизродено начално решение, за което функцията на правдоподбие е ограничена. Ето защо препоръчваме извличане на начална подизвадка с по-голям обем, с надеждата че поне p наблюдения ще бъдат използвани за оценяване на параметрите на всяка компонента на сместа, за да бъдат използвани като начални оценки в итерационния процес. Ако възникнат проблеми с оценяването на параметрите на някоя компонента, се извлича нова подизвадка. Този процес на проба и грешка продължава, до тогава докато не бъде получена начална оценка на параметрите на всяка една компонента на сместа. Обемът на извадката в стъпката

на подобрение е k, което трябва да удовлетворява условието за максимална прагова точка на TLE $\lfloor (n + gp))/2 \rfloor$.

Голям брой програмни процедури за оценяване на параметрите на смеси от разпределения, в частност FlexMix, максимизират израза (5.7) или (5.8) в зависимост от избора на теглата. Така например, ако бъдат използвани тегла само от 0 или 1, тогава се използва класификационният EM алгоритъм от FlexMix. Използването на програмата показва, че тази възможност дава надеждни резултати в опитната стъпка на FAST-TLE алгориъма.

Стойностите на функциите $\{-\log \varphi(y_i; x_i, \Psi)\}$ за $i = 1, \ldots, n$, дефинирани чрез (5.1) се пресмятат за текущата стойност на оценката $\widehat{\Psi}$ и сортират във възходящ ред, за да бъдат извлечени индексите на онези k наблюдения, които съответестват на най-малки отрицателни логаритми на правдоподобие, стартирайки с опитната оценка Ψ^* на Ψ на първата итерация в стъпката на подобрение. На практика са необходми 4-5 итерации в дадена стъпка на подобрение, за достигане на локална сходимост.

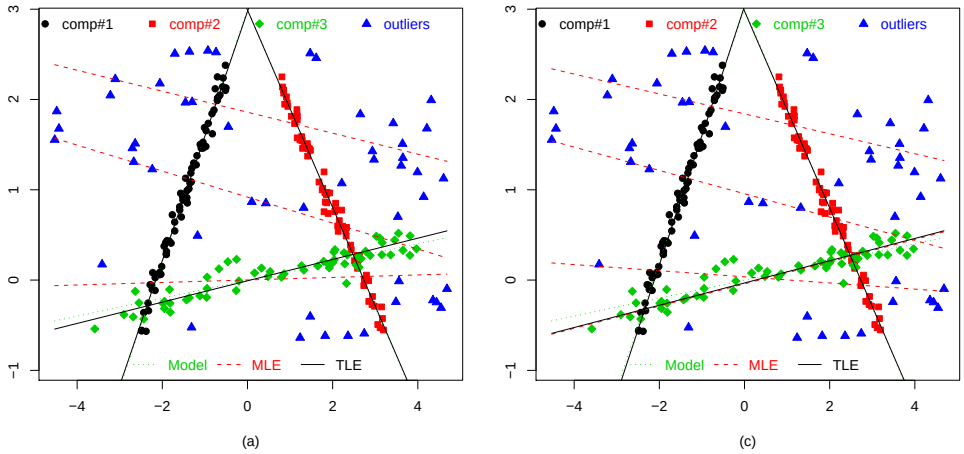
5.3 Примери

Смес от три линейни регресионни прави със замърсяване

В този експеримент е разгледана смес от три прости линейни регресионни прави с нормално разпределена грешка със замърсяване в данните. Регресионните линии са генерирани по моделите $y_{1i} = 3 + 1.4x_i + \epsilon_i$ (70 наблюдения), $y_{2i} = 3 - 1.1x_i + \epsilon_i$ (70 наблюдения), и $y_{3i} = 0.1x_i + \epsilon_i$ (60 наблюдения), където x_i са равномерно разпределени в интервала [-3,-1] и [1,3], съответно, ϵ_i е стандартно нормално разпределен със стандартно отклонение $\sigma = 0.1$. Към тези данни са добавени 50 несъгласувани наблюдения в правоъгълника $[-4.5, 4.5] \times [-0.8, 2.8]$. Наблюденията, които следват модела са означени с ромб, докато несъгласуваните наблюдения с триъгълник. На плота на фиг. 5.1 са дадени типичните резултати, получени от процедурата за МПО и FAST-TLE. Линиите, маркирани с точки, тирета и непрекъсната линия, съответстват на истинската линия по модела, оценените параметри по ММП и FAST-TLE, съответно. Стартирайки с различни проценти на тримиране от 20% до 45% и различен брой компоненти на сместа от 2 до 5 се забелязва, че при FAST-TLE алгоритъма се наблюдава сходимост към трите компоненти на модела на сместа в почти всички опити, докато чрез ММП не се наблюдава сходимост.

Смес от три двумерни нормални модела със замърсяване

С този пример е изследвано поведението на FAST-TLE алгоритъма върху симулира-



Фигура 5.1: Смес от три прости регресионни модела (линнии от точки), оценен модел по ММП (пунктирни линии) и FAST-TLE (плътни линии) с (a) 20% тримиране с 3 компоненти, (b) 40% тримиране и 4 компоненти.

ни данни, анализирани от McLachlan and Peel (2000). Тези данни се състоят от 100 наблюдения генерирани от смес от три двумерно нормални разпределения с равни пропорции и параметри както следва

$$\mu_1 = (0 \quad 3)^T, \quad \mu_2 = (3 \quad 0)^T, \quad \mu_3 = (-3 \quad 0)^T,$$

$$\Sigma_1 = \begin{pmatrix} 2 & 0.5 \\ 0.5 & .5 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} .1 & 0 \\ 0 & .1 \end{pmatrix}, \quad \Sigma_3 = \begin{pmatrix} 2 & -0.5 \\ -0.5 & .5 \end{pmatrix}.$$

Към тези данни са добавени 50 несъгласувани равномерно разпределени наблюдения в правоъгълника $[-10, 10] \times [-10, 10]$. Така дефинираната извадка е с обем от 150 наблюдения. McLachlan and Peel (2000) моделират тези данни със смес от три двумерни t–разпределения, за да редуцират влиянието на несъгласуваните наблюдения.

Оригиналните данни, несъгласуваните наблюдения и трите компоненти на сместа, оценени по ММП и FAST-TLE алгоритъма с 15%, 25%, 35% и 45% на тримиране са представени на фиг. 5.2 (a)–(d). Наблюденията, които следват модела са означени с ромбове, квадрати и кръгове, докато несъгласуваните наблюдения са означени с триъгълници. Контурите на елипсите от точки съответстват на модела, докато 99% доверителни елипси с пунктирните и непрекъснати контури, съответстват на оценките по ММП и FAST-TLE. Забелязва се, че при по-нисък или по-висок процент на

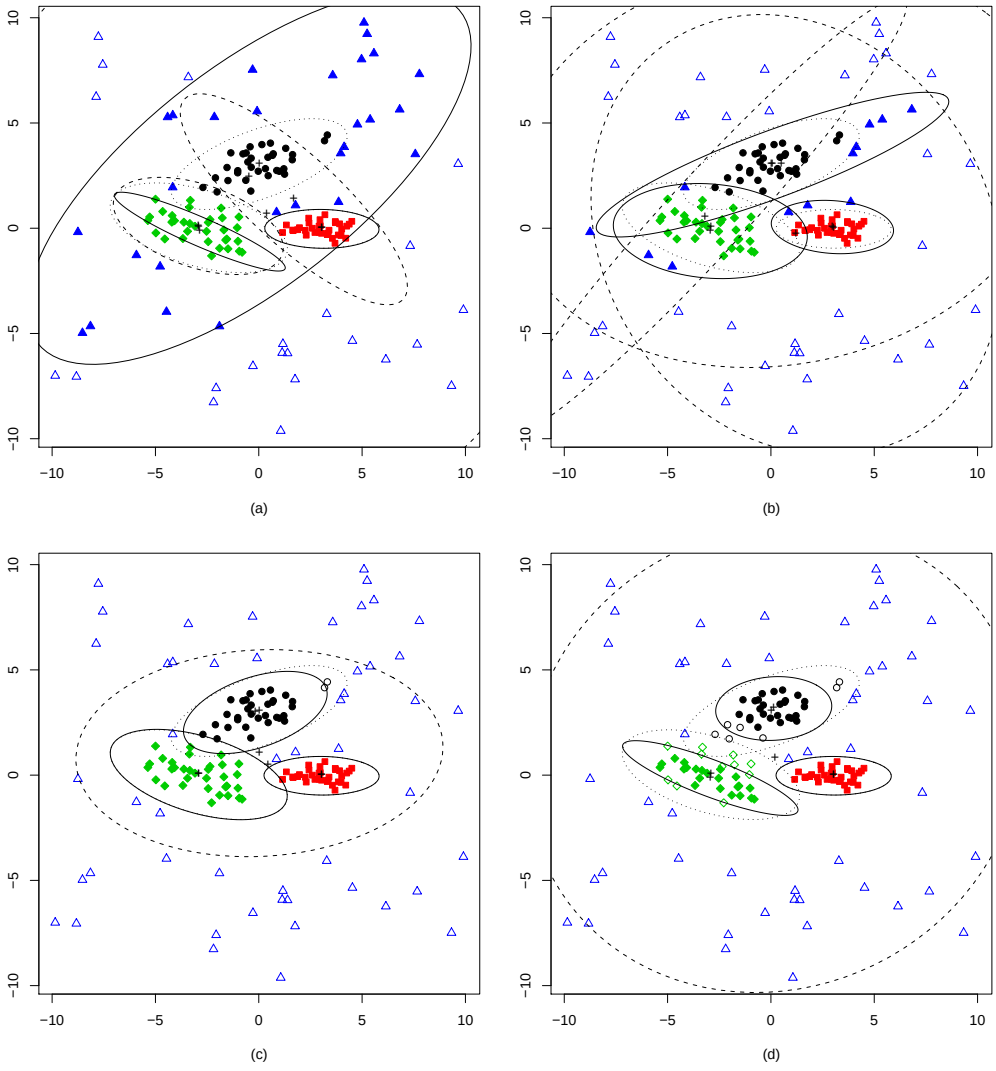
Таблица 5.1: Симулационни експерименти за данните на McLachlan and Peel (2000): медиани на TBIC (закръгкени) стойностите, основани на различен брой компоненти на сместа (редове) (rows) и различен процент на тримиране (колони).

	0%	5%	10%	15%	20%	25%	30%	35%	40%	45%
1	1672	1510	1382	1253	1119	1003	915	837	749	650
2	1654	1494	1338	1202	1054	920	822	734	643	559
3	1585	1436	1313	1190	*1047*	*902*	*795*	*709*	*620*	*538*
4	1595	*1429*	*1304*	*1178*	1040	908	807	720	631	549
5	*1594*	1430	1309	1184	1051	922	822	736	647	566

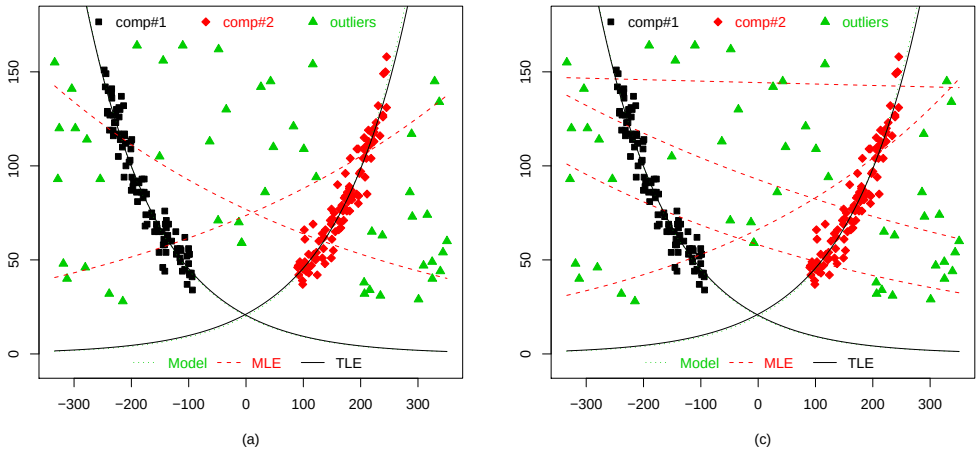
тримиране, спрямо процента на замърсяване, TLE оценките вярно оценяват центровете на елипсоидите, докато съответните ковариационни матрици са надценени или подценени. Оценките по ММП са силно повлияни от наличието на несъгласуваните наблюдения в данните, когато използваме 3 или по-голям брой компоненти на сместа.

В реалните данни, броят на смесите не е известен, поради което се използва инфомационният критерий на Бейс (BIC), за идентифициране на оптимален брой компоненти. Тримирания аналог на BIC се дефинира като $\text{TBIC} = -2\log(TL_k(\tilde{\Psi})) + m\log(k)$, където $TL_k(\tilde{\Psi})$ е максимумът на тримираната функция на правдоподобие, k е параметърът на тримиране, а m броя на параметрите в модела на сместа. TBIC се редуцира до BIC ако $k = n$. За да придобием представа за емпиричното разпределение на тези величини, за този пример беше проведено ограничено симулационно изследване с методите Монте Карло. Данните бяха анализирани с 1, 2, 3, 4 и 5 компоненти при различни проценти на тримиране от три 0% до 45% със стъпка 5%. Експериментът беше повторен независимо 500 пъти за всяка комбинация. Получените медиани на стойностите на TBIC (закръглени) са дадени Table 5.1. Най-малките стойности във всяка колона са означени с курсив. Забелязва се, че тези стойности се стабилизират за модел с 3 компоненти, който е верният модел. Провеждането на двуфазова линейна регресия (проста сплайн регресия) на стойностите от 3тия ред срещу процента на тримиране, разкрива точката на промяната, която е между 25% и 30% процент на тримиране и би могла да се интерпретира като оценка на процента на замърсяване.

От този и други симулационни експерименти заключаваме, че TBIC може да бъде използван за робастно оценяване на броя на компонентите на сместа и процента на



Фигура 5.2: Данни на McLachlan and Peel (2000) - смес от 3 двумерно нормални разпределения с добавен шум: истински модел (линии с точки) и оценени по модела на сместа от 3 компоненти по метода на МП (пунктирни линии) и FAST-TLE (непрекъснати линиии) с (a) 15%, (b) 25%, (c) 35%, и (d) 45% на тримиране.



Фигура 5.3: Смес от две Поасонови регресионни компоненти: истински модел (точкови линии), оценки по ММП (пунктирни линии) и FAST-TLE (непрекъснати линии) с (a) 20% тримиране и 2 компоненти, и (b) 40% тримиране и 4 компоненти.

несъгласуваните наблюдения в данните.

Смес от два Поасонови регресионни модела с шум

В този пример се разглежда смес от два Поасонови линейни регресионни модела с равни пропорции на сместа. За всеки от двете компоненти на сместа са генерирани по 100 наблюдения с разпределение на Поасон със средни $\log \lambda_1 = 3 - 0.008x$ и $\log \lambda_2 = 3 + 0.008x$, където x е равномерно разпределена в интервала [-225,-25] и [25,225], съответно. Към тези данни са добавени 50 наблюдения с равномерно разпределение в горните два интервала. На плотовете на Figure 5.3 са дадени резутати от оценяването по ММП и FAST-TLE алгоритъма. Наблюденията, които следват модела са означени с квадрати и ромбове, докато несъгласуваните с модела наблюдения са означени с триъгълници. Точковата, пунктирната и непрекъсната линия съответстват на истинската моделна линия и оценките по ММП и TLE, съответно. FAST-TLE алгоритъмът идентифицира правилно двете компоненти на сместа в повечето от експериментите, докато ММП не успява да идентифицира структурата, стартирайки с нарастващ брой компоненти от 2 до 5, както е показано на Фигура 5.3.

За да придобием по-добра представа за този тип оценяване ние генерирахме 100 независими извадки, според описания модел. Всяка от извадките беше анализирана с 2, 3, 4 и 5 компоненти и 20% тримиране. Подобно на предишните примери. броят на оценените компоненти, които FlexMix предоставя на изхода може да бъде по-малък

от зададения. За всеки разглеждан модел със зададен максимален брой компненти на сместа ние преброихме успешно оценените компоненти във всички симулирани извадки. Резултатите за оценките по ММП и FAST-TLE са дадени в Table 5.2. Броят на заявените компоненти на входа на FlexMix е зададен в първата колона на всеки ред на таблицата, докато броят на изхода от програмата FlexMix е даден в съответните колони за оценките по ММП и FAST-TLE. В допълнение, към честотните в последния ред на таблицата е даден броят на успешно идентифицираните две компоненти на сместа. Вижда се, че шансовете за успешно идентифициране на двете компоненти по ММП нараства със задаването на по-голям брой компоненти на входа на FlexMix. Като цяло, поведението на ММП при наличието на шум в данните е проблематично. Двете копмпоненти на сместа са успешно идентифициране само в 37 случая от 400 опита. Задаването на по-голям брой компоненти на входа на FlexMix не оказва влияние на резултатите от FAST-TLE, тъй като модел с две компоненти е оптимален и е идентифициран в повече от 90% от опитите. Общият брой на успешно идентифицираните две компоненти на сместа е 392 от 400 експеримента.

Таблица 5.2: Симулационни резултати за смес от две Поасонови регресии. Използвани са модели с 2, 3, 4, и 5 компоненти за анализ на извадка с обем 100. От 400 експеримента, 37 са успешни оценени по ММП и 392 по FAST-TLE.

	MLE returned components					FAST-TLE returned components				
started	2	3	4	5	Total	2	3	4	5	Total
2	100				100	100				100
	1				*1*	*98*				*98*
3		100			100	93	7			100
		2			*2*	*92*	*7*			*99*
4		94	6		100	96	4	0		100
		4	*4*		*8*	*94*	*4*			*98*
5		19	15	66	100	94	6		0	100
		3	*7*	*16*	*26*	*91*	*6*			*97*
Total	100	213	21	66	400	383	7	0	0	400
	1	*9*	*11*	*16*	*37*	*375*	*17*			*392*

5.4 Резюме

Глава 5 е посветена на робастно оценяване на параметрите на крайна смес от вероятностни разпределения с $TLE(k)$ оценките. Предимствата на този подход, в сравнение с оценките по ММП, са илюстрирани чрез примери и симулационно изследване с Монте Карло методите. Характеризирана е праговата точка на $WTLE(k)$ оценките за смеси от разпределения. Предложен е адаптивен подход за определяне на параметъра на тримиране k, основан на тримирания информационен критерий на Бейс (BIC). Като частен случай на $TLE(k)$ оценките в модели на смеси от разпределения е дефинирана тримираната класификационна функция на правдоподобие за целите на робастното клъстериране на данни. За приближено пресмятане на неизвестните параметри на смесите от разпределния е използван FAST-TLE алгоритъмът в средата на EM алгоритъма. Резултатите от тази глава са публикувани в Neykov et al. (2004), and Neykov et al. (2007). Забелязани са 100 цитирания на статията на Neykov et al. (2007), докато статията на Neykov et al. (2004) е цитирана 11 пъти.

Ще отбележа, че следните автори от приложения към дисертационния труд списък с цитирания, работещи в областта по разпознаване на образи в медицината, като ядрено магнитен резонанс и анализ на елктро-енцефелограми за целите на компютърна диагностика на множествена склероза и тумори в мозъка, показват че идентифицирането на съответните структури, характеризиращи заболяването, е успешно чрез използването на TLE оценките на параметрите на смеси от двумерно нормални разпределения за разлика от оценяването по ММП, следвайки Neykov et al. (2004) и Neykov et al. (2007): Ait-Ali et al. (2005), Ait-Ali et al. (2006), Ait-Ali (2006), Herrera (2006), Lecoeur and Barillot (2007), Bricq (2008a), Bricq et al. (2008b), Bricq et al. (2008c), Bricq et al. (2008d), Bricq et al. (2010), Tomas-Fernandez and Warfield (2012), Mortazavi et al. (2012), Wang et al. (2012a), Barillot et al. (2014), Wang et al. (2014a), Wang et al. (2014b), Wang et al. (2014c), Galimzianova et al. (2015), Tomas-Fernandez and Warfield (2015), Karpate (2015) Galimzianova et al. (2016), Jerman et al. (2016).

Ще отбележа също така, че групата от статистици от департамента по статистика на Университета във Валядолид, Испания, публикува серия от статии, основани на тримираната класификационна функция на правдоподобие за моделиране и клъстериране на нееднородни данни с комплексни вероятностни модели. Цитирането на статията на Neykov et al. (2007) и развитието на този подход са дадени в работите на Garcia-Escudero et al. (2010a), Garcia-Escudero et al. (2010b), Garcia-Escudero et al. (2011), Garcia-Escudero et al. (2013), Garcia-Escudero et al. (2014), Garcia-Escudero et al. (2015) и Garcia-Escudero et al. (2016). Разработената от този колектив програмна

библиотека tclust в средата на R, описана в работите на Fritz et al. (2013a), Fritz et al. (2013b) и Ruwet et al. (2013), се основава на модификация на FAST-TLE алгоритъма, предложен от Neykov and Müller (2003).

Апендикс към глава 5

Proposition 5.1 *Ако множеството* $F_{\theta_j} = \{-\log \psi(y_1, x_1, \theta_j), \ldots, -\log \psi(y_n, x_n, \theta_j)\}$ *е* d_j*– пълно за всяко* $j = 1, \ldots, g$*, то праговата точка на WTLE(k) оценките за модела на смес от разпределения (5.1) удовлетворява* $\varepsilon_n^*(WTLE(k)) \geq \frac{1}{n}\min(n-k, k-d)$*, където* $d = \max(d_1, d_2, \ldots, d_g)$*.*

Доказателство: Нека $\tilde{Z}_m = \{(\tilde{y}_1, \tilde{x}_1), \ldots, (\tilde{y}_n, \tilde{x}_n)\}$ се получава от $Z = \{(y_1, x_1), \ldots, (y_n, x_n)\}$ чрез замяна на $m = \min(n-k, k-d_j)$ наблюдения с произволни. Следователно, ако $m = n - k \leq k - d_j$, то $k - m \geq d_j$, и ако $m = k - d_j$, то $k - m = d_j$, т.е. за всеки k наблюдения от $\tilde{Z}_m$ съществуват поне d_j, които са от Z. Нека $\tilde{\theta}^{(l)}$ е стойността на параметъра на l–та итерация и $\tilde{\tau}_{ij}^{(l)}$ е съответната стойност на $\tau_{ij}^{(l)} = \tau_j(y_i, x_i, \theta_j^{(l)})$ върху $\tilde{Z}_m$. Без ограничение на общността можем да предположим, че на $(l+1)$-та итерация сме получили множеството $I^{(l+1)} = \{1, \ldots, k\}$. Нека $J \subset I^{(l+1)}$ е с кардиналност $|J| = d_j$ и за $i \in J$, $(\tilde{y}_i \tilde{x}_i) \in Z$. Тогава за всяко θ_j за $j = 1, \ldots, g$ е изпълнено

$$-\sum_{i=1}^{k} \tilde{\tau}_{ij}^{(l)} \log \psi(\tilde{y}_i, \tilde{x}_i, \theta_j) \geq -\sum_{i \in J} \tau_{ij}^{(l)} \log \psi(\tilde{y}_i, \tilde{x}_i, \theta_j) \geq \tau_{iJ}^{(l)} g_J(\theta_j),$$

където $\tau_{iJ}^{(l)} = \min_{j \in J} \tau_{ij}^{(l)}$ и $g_J(\theta_j) = \max_{i \in J}\{-\log \psi(y_i, x_i, \theta_j)\}$. Функцията от дясната страна на неравенството е субкомпактна, понеже F_{θ_j} е d_j–пълно, от което следва, че сумата от лявата страна на неравенството е субкомпактна. Следователно, множеството $\{\theta_j : -\sum_{i=1}^{k} \tilde{\tau}_{ij}^{(l)} \log \psi(\tilde{y}_i, \tilde{x}_i, \theta_j) \leq C\}$ е компактно и $\tilde{\theta}_j^{(l+1)}$ се съдържа в него.

Множествата F_{θ_j} ще бъдат едновременно d-пълни, ако $d = \max(d_1, d_2, \ldots, d_g)$ (Vandev, 1993: ако едно множество е q-пълно, то е $q+1$-пълно също така). □

Глава 6

Робастно моделиране на очакването и дисперсионния параматър чрез TLE оценките

6.1 Въведение

Нека y_i за $i = 1, \ldots, n$ са наблюдения на сл. вел. Y с разпределение от експоненциалната дисперсионна фамилия от разпределения (ЕДФР), дефинирана чрез

$$f(y_i, \theta_i, \phi_i) = \exp\left\{\frac{y_i\theta_i - \kappa(\theta_i)}{\phi_i} + c(y_i, \phi_i)\right\}$$

където θ_i е неизвестен параметър, ϕ_i е дисперсионен параметър, $\kappa(\theta)$ е кумулантата (семиинварианта), докато аналитичният вид на функцията $c(y_i, \phi)$ е известен само в някои частни случаи. Очакването и дисперсията на ЕДФР са $E(y_i) = \mu_i = \kappa'(\theta_i)$ и $\mathrm{var(y_i)} = \phi_\mathrm{i}\kappa''(\theta_\mathrm{i}) = \phi_\mathrm{i}\mathrm{V}(\mu_\mathrm{i})$ (Jørgensen, 1997), където функцията $V(.)$ се нарича дисперсионна функция. Линейната експоненциална фамилия от разпределения с известен дисперсионен параметър $\phi_i = \phi$ е частен случай на ЕДФР.

Систематичната компонента на обобщен линеен модел със случайна компонента с разпределение от ЕДФР се дефинира като

$$g(\mu_i) = g(\kappa'(\theta_i)) = \eta_i = x_i^T\beta, \qquad h(\phi_i) = z_i^T\lambda \qquad \text{и} \qquad \mathrm{var}(y_i) = \phi_i V(\mu_i), \tag{6.1}$$

където g и h са известни монотонно свързващи функции, x_i и z_i са векторни предикторни променливи с размерност p и q, β и λ са вектори от неизвестни регресионни параметри, съответно.

В общия случай функцията на правдоподобие за модели от ЕДФР не притежава явно аналитично представяне поради неявния вид на функцията на $c(y_i, \phi_i)$. Поради тази причина, Nelder and Pregibon (1987) предлагат максимизирането на една нейна апроксимация (saddlepoint approximation), така наречената модифицирана функция на квази-правдободобие (extended quasi-log-likelihood, EQL)

$$Q^+(\beta, \lambda) = \sum_{i=1}^{n} -\frac{1}{2}\left\{\log\left[2\pi\phi_i(\lambda)V(y_i)\right] + \frac{d_i(\beta)}{\phi_i(\lambda)}\right\} \tag{6.2}$$

$$= \sum_{i=1}^{n} q^+(y_i; \mu_i(\beta), \phi_i(\lambda)) = \sum_{i=1}^{n} q^+(y_i; \mu_i, \phi_i), \tag{6.3}$$

където $d_i \equiv d(y_i; \mu_i) = -2\int_{y_i}^{\mu_i} \frac{y_i - u}{V(u)} du$ е функцията на отклоненията на i-то наблюдение.

Ще отбележим, че EQL апроксимацията съвпада с логаритъма на функцията на правдоподобие на нормалното разпределение, понеже $V(y_i) = 1$, $\phi = \sigma^2$ и $d(y_i, \mu_i) = (y_i - \mu_i)^2$, Smyth (1989). EQL се дефинира чрез първите два момента на разпределението, което позволява по-голяма свобода и гъвкавост за целите на статистическото моделиране.

Целевата функция (6.3) показва, че квази-правдоподобната оценка $\widehat{\beta}$ на β се определя от минимизирането на функцията на отклоненията $\sum_{i=1}^{n} d_i$ вместо да максимизираме директно $Q^+(\beta, \lambda)$, докато квази-правдоподобната оценка $\widehat{\lambda}$ на λ може да бъде получена чрез $\widehat{\mu}_i = \mu_i(\widehat{\beta})$. Това е така, тъй като от $E\left(\partial^2 Q^+ / \partial\mu_i \partial\phi_i\right) = 0$ следва ортогоналност между β и λ. Следователно оптимизационната задача с размерност $p + q$ се редуцира до две отделни оптимизационни задачи с размерности p и q. В резултат на това, неизвестните параметри β и λ могат да бъдат оценени чрез последователно оценяване с два обобщени линейни модела, първият е стандартен, а вторият е гама обобщен линеен модел, в който отклоненията d_i за $i = 1, \ldots, n$ на първия обобщен линеен модел, представляват стойностите на зависима променлива на втория обобщен линеен модел

$$E(y_i) = \mu_i \qquad g(\mu_i) = \eta_i = x_i^T\beta \qquad \mathrm{var}(y_i) = \phi_i V(\mu_i) \tag{6.4}$$

$$E(d_i) = \phi_i \qquad h(\phi_i) = \xi_i = z_i^T\lambda \qquad \mathrm{var}(d_i) = 2\phi_i^2. \tag{6.5}$$

От изчислителна гледна точка, следвайки Green (1984), оценяването на параметрите чрез тези два алтернириращи обобщени линейни модела се свежда до две еквивалентни итерационни линейни регресионни задачи с изкуствени зависими променливи и тегла. Оценяването на неизвестните параметри в тези регресионни модели е по метода на

най-малките квадрати:

$$\min_{\beta} \left(u_m - X\beta\right)^T W_m \left(u_m - X\beta\right) \tag{6.6}$$

$$\min_{\lambda} \left(u_{d^*} - Z\lambda\right)^T W_{d^*} \left(u_{d^*} - Z\lambda\right), \tag{6.7}$$

където X и Z са $n \times p$ и $n \times q$ матрици на предикторите, u_m и u_{d^*} са изкуствените зависими променливи на модела на очакването и дисперсията с елементи $u_{m,i} = x_i^T\beta + \frac{\partial \eta_i}{\partial \mu_i}(y_i - \mu_i)$ и $u_{d^*,i} = z_i^T\lambda + \frac{\partial \xi_i}{\partial \phi_i}(y_i - \phi_i)$, а $W_m = \mathrm{diag}\left((\phi_i(\partial\eta_i/\partial\mu_i)^2 V(\mu_i))^{-1}\right)$ и $W_{d^*} = \mathrm{diag}\left((2(1-\rho_i)\phi_i^2(\partial\xi_i/\partial\phi_i)^2)^{-1}\right)$ са матриците на теглата, ρ_{ii} е i-ти диагонален елемент на $W_m^{1/2}X(X^TW_mX)X^TW_m^{1/2}$, чиито стойности са пресметнати чрез текущите оценки на β и λ.

Използването на EQL предоставя голяма гъвкавост при моделирането с обобщени линейни модели, което се подпомага от широко разпространените програмни системи като библиотеките *dglm*, *JointModeling*, *statmod* от програмната среда R, виж Smyth (2009a), Ribatet and Iooss (2009), and Smyth (2009b).

Известно е, че оценките по ММП и квази-правдоподобните оценки не са устойчиви спрямо наличието на несъгласувани наблюдения в данните. Неробастността на МПО и квази-правдоподобните оценки спрямо наличието на несъгласувани наблюдения в рамките на един обобщен линеен модел е обект на интензивни изследвания в литературата, например, Markatou et al. (1997), Cantoni and Ronchetti (2001), Müller and Neykov (2003), Maronna et al. (2006).

В тази глава се разглеждат робастно оценяване на съвместното моделиране на модела на очакването и дисперсията чрез тримиране, за да бъде редуцирано влиянието на несъгласувани наблюдения в данните.

6.2 Тримирани квази-правдоподобни оценки

Нека $\theta = (\beta, \lambda)$ и да заместим $f_i(\theta) := f_i(\beta, \lambda) = -q^+(y_i; \mu_i(\beta), \phi_i(\lambda))$ в *Definition 0.10*. По този начин дефинираме частен случай на wGTE оценките, за които ще използваме означението ETQL от Extended Trimmed Quasi-Likelihood.

Definition 5.2 ETQL оценката $(\widehat{\beta}, \widehat{\lambda})$ на (β, λ) се дефинира като

$$\begin{aligned} \max_{\beta,\lambda} Q^+_{\mathrm{trim}}(\beta, \lambda) &= \max_{\beta,\lambda} \max_{I \in I_k} \sum_{i \in I} q^+(y_i; \mu_i, \phi_i) = \max_{I \in I_k} \max_{\beta,\lambda} \sum_{i \in I} q^+(y_i; \mu_i, \phi_i) \\ &= \min_{I \in I_k} \min_{\beta,\lambda} \sum_{i \in I} - q^+(y_i; \mu_i, \phi_i) \end{aligned}$$

Оценката ETQL е EQL оценка, пресметната за някоя подизвадка с обем k на оригиналната извадка с обем от n наблюдения. Следователно за всички подизвадки с обем k оценяването се осъществява по двата обобщени линейни модела, дефинирани с (6.4) и (6.5). Вследствие на това, праговата точка на ETQL оценките се характеризира чрез по-малката от двете прагови точки на тези свързани обобщени линейни модели. Следователно е необходимо определянето на индекса на d–пълнота на съответните множества от отрицателните логаритми на правдоподобие на обобщените линейни модела (Müller and Neykov, 2003). Определянето на индекса на d–пълнота на множествата от отрицателните логаритми на правдоподобие на двата обобщени линейни модела (6.4) и (6.5) е еквивалентно на определянето на индексите на d–пълнота на съответните им функции на отклонения, поради пропорционалност с точност до константа.

Theorem 6.1 *Праговата точка на оценката ETQL е равна на*

$$\frac{1}{n}\min\left\{n-k, k-\max[\mathcal{N}(X),\mathcal{N}(Z)]-1\right\}$$

и достига максимум за

$$\left\lfloor\left\{n+\max[\mathcal{N}(X),\mathcal{N}(Z)]+1\right\}/2\right\rfloor \le k \le \left\lfloor\left\{n+\max[\mathcal{N}(X),\mathcal{N}(Z)]+2\right\}/2\right\rfloor,$$

който е равен на $\frac{1}{n}\left\lfloor\left\{n-\max[\mathcal{N}(X),\mathcal{N}(Z)]-1\right\}/2\right\rfloor$.

6.3 Алгоритъм за пресмятане на GTE оценките

В тази секция е предложен FAST-GTE алгоритъм за приближено пресмятане на $GTE(k)$ оценката за d–пълно множество от неотрицателни функции. Стъпката на концентрация е като в алгоритъма FAST-TLE от глава 2.

За да съществува решение с положителна прагова точка на оптимизационната задача (4), *Definition 0.10*, ще предполагаме, че множеството F е d–пълно и $k \ge d$. При тези условия алгоритъмът се състои в провеждането на много двустъпкови процедури, състоящи се от опитна, изпробваща стъпка, последвана от стъпка на подобрение:

Опитна стъпка:

1. Нека $F^{old} = \{f_{i_1}(\theta), \ldots, f_{i_l}(\theta)\} \subset F = \{f_i(\theta) \ge 0 \text{ for } i = 1, \ldots, n\}$ е d–пълно, където $l \ge d$;
2. Нека $\widehat{\theta}^{old}$ бъде произволна стойност или минимум на $\sum_{j=1}^{l} f_{i_j}(\theta)$;

Стъпка на подобрение:

3. Нека $F^{new} = \{f_{\nu(1)}(\theta), \ldots, f_{\nu(k)}(\theta)\} \subset F$ където $f_{\nu(1)}(\widehat{\theta}^{old}) \leq \ldots \leq f_{\nu(n)}(\widehat{\theta}^{old})$ са наредените стойности във възходящ ред на $f_i(\widehat{\theta}^{old})$ за $i = 1, \ldots, n$;
4. Нека за $\widehat{\theta}^{new}$ се достига минимум на $S(\theta) = \sum_{i=1}^{k} f_{\nu(i)}(\theta)$, където $f_{\nu(i)} \in F^{new}$ за $i = 1, \ldots, k$;
5. Нека $\widehat{\theta}^{old} := \widehat{\theta}^{new}$;
6. Формиране на цикъл по стъпки 3 - 5 до достигане на сходимост или е достигнат определен брой итерации.

Proposition 6.1 $S(\widehat{\theta}^{new}) \leq S(\widehat{\theta}^{old})$.

На основата на Теорема 6.1 са дискутирани изборът на параметъра k на тримиране, както и на подизвадки с минимален обем не съдържащи несъгласувани наблюдения с максимална вероятност, а така също и за разбиване на данните на непресичащи се подгрупи при голям стойности на n, броя на наблюденията, върху всяка от които да бъде прилаган, предложения алгоритъм.

Ще отбележим, че частни случай на "стъпката на концентрация"се явяват съответните стъпки в работите на Visek (1996), Rousseeuw and van Driessen(1999a), и Hawkins and Olive (2002), съотвтено за LTS и LTAD линейни регресии, както и на Hawkins and Khan (2009) за LTS нелинейниа регресия; на Rousseeuw and van Driessen(1999b) и Herwindiati et al. (2007) за намиране на MCD и MVVE оценките на ковариационни матрици; на Neykov and Müller (2003), Gallegos and Ritter (2005), Neykov et al. (2007), Garcia-Escudero et al. (2008), Cuesta-Albertos et al. (2008), и Gallegos and Ritter (2010) за оценяване с TLE на параметрите на смеси от регресионни разпределения и с TLE версията за класификация и клъстериране.

В секции 5.4 и 5.5 е изследвано поведението на ETQL и EQL оценките с методите Монте Карло със замърсяване на данните с различни проценти на несъгласувани наблюдения както в зависимата така и в предикторните променливи в моделите на очакването и дисперсионната функция с различни проценти на тримиране. в модели на класическа линейна регресия с Гаусова грешка с нееднородна дисперсия. Разпределенията на тези оценки за различните комбинации на замърсяване и тримиране както и стойностите на съответните целевите функции са представени графично. Поведението на тези оценки е изследвано в следните 2=3 модела

6.4 Пример

Поведението на EQL и ETQL оценките е изследвано върху известните литературни данни, анализирани от Zuliani et al. (1983) и Smyth and Verbyla (1999) чрез замърсяване на оригиналните данни с несъгласувани наблюдения. Показано е, че ETQL оценките са стабилни при замърсените и незамърсени данни, за разлика от EQL оценките, които са твърде ненадеждни.

6.5 Симулационни експерименти

В тази секция е изследвано поведението на EQL и ETQL оценките в ситуации на вярно (правилно) и невярно (неправилно) зададен дисперсионен модел, при данни без наличие на несъгласувани и със замърсени данни, при различни проценти на тримиране. Пресмятанията са проведени с предложения алгоритъм, реализаран със стандартни програмни библиотеки за съвместно моделиране на очакването и дисперсионната функция от програмната среда R, понеже двете стъпки на алгоритъма са стандартни EQL процедури, използващи подизвадки с различни обеми.

6.5.1 Симулационен експеримент

Поведението на ETQL и EQL е изследвано с методите на имитационното моделиране за три статистически модела.

1ви експеримент: Този експеримент се отнася до класическата линейна регресия с нормално разпределена грешка и нееднородна дисперсия. Генерирани са данни са

$$\begin{aligned} y_i &= 1 + x_{i1} + x_{i2} + \sqrt{\phi_i}\epsilon_i \quad \text{за} \quad i = 1, \ldots, 40 \\ \log(\phi_i) &= -4 - 4x_{i3}, \end{aligned}$$

където x_{i1}, x_{i2} и x_{i3} са равномерно разпределени в интервала [0,1] а ϵ_i са стандартно нормално разпределени. Данните са замърсени чрез модифициране на 4 от генерираните стойности на следните наблюдения както следва: $x_{37,3} := x_{37,3} - 5$, $x_{38,2} := x_{38,2} - 5$, $x_{39,1} := x_{39,1} + 5$, and $y_{40} := y_{40} - 10$. По този начин три от несъгласуваните наблюдения са в предикторните променливи, а четвъртото в зависимата променлива. Резултатите от двата пакета са почти еднакви.

2ри експеримент: В този експеримент е разгледан модел с очакване гама разпределение, т.е. моделите на очакването и дисперсионната функция са гама обобщени

линейни модели. Обемът на извадката е 40 и данните са генерирани с както следва

$$\begin{aligned} \log(\mu_i) &= 1 + x_{i1} + x_{i2} \quad \text{за} \quad i = 1, \ldots, 40 \\ \log(\phi_i) &= -2 - 2x_{i3}, \end{aligned}$$

където предикторите x_{i1}, x_{i2} и x_{i3} са равномерно разпределени в интервала $[-1, 1]$. Следователно наблюденията y_i са $Gamma(\phi_i\mu_i, \phi_i^{-1})$ разпределени с параметри на мащаба и формата $\phi_i\mu_i$ и ϕ_i^{-1}, съответно. Четири от стойностите на генерираните наблюдения са заменени чрез следните генерирани стойности : $x_{37,1} := x_{37,1} \pm 14$, $x_{38,2} := x_{38,2} \pm 20$, $x_{39,3} := x_{39,3} \pm 20$ и $y_{40} := y_{40} \pm 14$, където $\pm$ означава знакът плюс или минус е случайно генериран. Три от несъгласуваните наблюдения са в предикторните променливи, а четвъртото в зависимата променлива. Ще отбележим, че вместо гама сме използвали digamma дисперсионен обобщен линеен модел (6.5) както препоръчват Smyth (1989), и Lee et al (2005). Поради тази причина за необходимите пресмятания е използван пакета *dglm* на Smyth (2009).

Зти експеримент: В този експеримент данните са генерирани според модел от фамилията на разпределението на Tweedie с модел на дисперсионната функция $var(y_i) = \phi_i\mu_i^\theta$, параметър $\theta = 1$, очакване μ_i и дисперсионен параметър ϕ_i, дефинирани чрез

$$\begin{aligned} \log(\mu_i) &= 1 + x_{i1} + x_{i2} \quad \text{за} \quad i = 1, \ldots, 40 \\ \log(\phi_i) &= -4 - 4x_{i3}. \end{aligned}$$

Предикторите x_{i1}, x_{i2} и x_{i3} са равномерно разпределени в интервала $[-1, 1]$. Данните са замърсени чрез модифициране на 4 от генерираните стойности на следните наблюдения както следва: $x_{37,3} := x_{37,3} - 5$, $x_{38,2} := x_{38,2} \pm 5$, $x_{39,1} := x_{39,1} \pm 5$, и $y_{40} := y_{40} + 10$. Подобно на горните два модела три от несъгласуваните наблюдения са в предикторите, а четвъртото в зависимата променлива. За генерациите сме използвали библиотеката *tweedie* от R, разработена от Dunn (2009) докато необходимите пресмятания са проведени с библиотеката *dglm* на Smyth (2009).

Симулационните експерименти са повторени 1000 пъти, в резултат на което са получени редици от оценки, чиито разпределения (бокс-плотове) са дадени на съответните фигури за всеки от параметрите на двата модела. Симулационните експерименти в трите експеримента показват, че при подходящ избор на стойността на параметър на тримиране, по-голям от процента на несъгласуваните наблюдения в данните, поведението на ETQL оценките в замърсените данни е сходно с поведението на EQL оценките, пресметнати по незамърсените данни, докато поведението на EQL оценките, пресметнати по замърсените данни е ненадеждно.

Обикновено процентът на несъгласуваните наблюдения в реални данни е неизвестен. Една техника за избор на параметъра на тримиране $\frac{n-k}{n}100\%$ би могла да се основават на последователно оценяване на модела за различни проценти на тримиране и където се наблюдава стабилизиране на оценките спрямо процента на тримиране. Това предлага плотиране на оценките на параметрите срещу процента на тримиране $\frac{n-k}{n}100\%$, където k варира в интервала $[(n+\max[\mathcal{N}(X),\mathcal{N}(Z)]+1)/2, n]$, и селектиране на подходящата стойност на k, за която оценките на параметрите остават едновременно стабилни, което да гарантира положителна прагова точка и висока ефективност на оценката. Така например, една стратегия би могла да бъде на постепенно намаляване на параметъра k на тримиране, стартирайки от $k = n$. По този начин, не само параметрите, но също така и процента на несъгласуваните наблюдения в данните може да бъде оценен робастно.

6.6 Резюме

Глава 6 третира оценяването на неизвестните параметри на обобщени линейни модели с разпределения от дисперсионната фамилия от разпределения, Jørgensen (1997), частен случай на която са разпределенията от линейната експоненциална фамилия. Оценяването на параметрите се основава на максимизирането на модифицираната функция на квази-прадоподобие (Extended Quasi-Likelihood, EQL). EQL функцията апроксимира логаритъма на функцията на правдоподобие, докато за обобщените линейни регресионни модели с нормално, инверсно нормално и гама разпределения, тези две функции са еквивалентни, Smyth (1989). В глава 6 са предложени $ETQL(k)$ оценките, основани на тримираната версия на EQL оценките, за да бъде преодолян ефектът от влиянието на несъгласуваните наблюдения в данните върху EQL оценките. Теорема 6.1 характеризира праговата точка на тези оценки. За приближено пресмятане на $ETQL(k)$ оценките на неизвестните параметри е предложено обобщение на FAST-TLE алгоритъма. Предимствата на $ETQL(k)$ оценките, в сравнение с оценките по ММП и EQL, е илюстрирано чрез примери и разширено симулационно изследване с методите Монте Карло. Предложена е модификация на FAST-TLE алгоритъма.

Глава 7

Тримирана квантилна регресия

7.1 Въведение

Тази глава е посветена на линейната квантилна регресия с тримиране. Да разгледаме модела на линейна множествена регресия

$$y_i = x_i^T\theta + \varepsilon_i \quad \text{за} \quad i = 1, \ldots, n, \tag{7.1}$$

където y_i е i-то наблюдение на зависимата променлива, $x_i^T = (x_{i1}, \ldots, x_{ip})$ е вектор от предикторни променливи, θ е $p \times 1$ вектор от неизвестни параматри, ε_i, $i = 1, \ldots, n$ са независими еднакво разпределени. Нека $r_i(\theta) = y_i - x_i^T\theta$ са регресионните остатъци. Koenker and Bassett (1978) дефинират квантилната регресионна (QR) оценка като всеки вектор $\widehat{\theta}_n(\tau)$ такъв че

$$\widehat{\theta}_n(\tau) := \arg\min_{\theta \in R^p} \sum_{i=1}^{n} \rho_\tau(r_i(\theta)), \tag{7.2}$$

където

$$\begin{aligned} \rho_\tau(r(\theta)) &= |r(\theta)| \left[\tau 1_{\{r(\theta)\geq 0\}} + (1-\tau) 1_{\{r(\theta)<0\}}\right] \\ &= \begin{cases} (\tau - 1)r(\theta) & \text{ако} \quad r(\theta) < 0, \\ \tau r(\theta) & \text{ако} \quad r(\theta) \geq 0, \end{cases} \end{aligned}$$

$0 < \tau < 1$, $1_{\{A\}}$ е индикаторната(характеристичната) функция на множеството A, която приема стойност 1 ако A истина и 0 в противен случай.

Различни квантилни регресионни оценки $\widehat{\theta}_n(\tau)$ могат да бъдат дефинирани за различни стойности на τ. Това предоставя на анализатора един по-пълен статистически

модел отколкото класическия регресионен модел на очакването на зависимата променлива.

Квантилните регресионни оценки са робастни спрямо несъгласуваните наблюдения в зависимата променлива. При твърде общи предположения, оценките на параметрите са асимптотично многомерно нормално разпределени, което позволява провеждането на стандартни статистически изводи.

Праговата точка на квантилната линейна регресия е разгледана от Jurečková (2010). Алгоритмите за пресмятане на оценките на квантилната линейна регресия са основани на линейното програмиране (Koenker, 2005a), или на техниките на максимизиране-минимизиране от Hunter and Lange (2000) и Chen (2004). Koenker (2005b) е разработил библиотеката *quantreg* от програмната среда на R (http://www.R-project.org), което улеснява рутинното използване на квантилната регресия. Повече подробности са дадени в монографията на Koenker (2005a).

Квантилната линейна регресия не е робастна спрямо наличието на несъгласувани наблюдения в предикторните променливи (He et al., 1990, Jurečková, 2010). Предлагани са различни подходи за претегляне на този тип наблюдения, на основата на разстоянията в пространството на предикторите, например Hubert and Rousseeuw (1998), Giloni et al., (2006). Показано е, че процедури от този род са успешни в твърде специални случай с малък брой предиктори и несъгласувани наблюдения, които са равномерно разпределени, което е нереалистично и твърде ограничително условие.

Алтернативни оценки на квантилните регресионни оценки са предложени от Rousseeuw and Hubert (1999), и Adrover et al., (2004), чиято праговата точка не зависи от комплексността на регресионния модел и е пропорционална на $\min\{\tau, 1-\tau\}$, където τ е съответния квантил, представляващ интерес. Недостатък на тези методи са от изчислителен характер и нестандартни асимптотични разпределения на оценките.

В тази глава е разгледан алтернативен подход за робастно оценяване на параметрите на квантилната регресия (the least trimmed quantile regression, LTQR), основан на техниката на тримиране, за да бъде редуцирано влиянието на несъгласуваниет наблюдения в предикторните променливи.

Предолагания метод обобщава работите на Koenker and Bassett (1978), Tableman (1994a,b) за оценяване по метод на най-малките тримирани по модул остатъци (Least Trimmed Absolute deviation, LTAD) на параметъра на положението (локацията), а така също на параметрите на линейния регресионен модел (7.1) с LTAD оценките от Hawkins and Olive (1999).

7.2 Регресионни квантилни оценки, основани на тримиране

Методът на тримираната линейна квантилна регресия (LTQR estimator) е частен случай на wGTE оценките (4).

Definition 7.1 *Оценката по метода на тримираната линейна квантилна регресия се дефинират като*

$$\widehat{\theta}_n^k(\tau) := \arg\min_{\theta} \left\{ Q_{n,k}(\theta) = \min_{I \in I_k} \sum_{i \in I} \rho_\tau(r_i(\theta)) \right\}, \tag{7.3}$$

където I_k е множеството от всевъзможните подмножества на k–наблюдения от n, $\rho_\tau(r_i(\theta))$ е зададена чрез (7.2) и $0 < \tau < 1$.

От дефиницията следва, че оценката по метода на тримираната линейна квантилна регресия се дефинира като класическата оценка на линейна квантилна регресия, но върху някоя подизвадка от k наблюдения. Частни случаи на този клас оценки са класическата квантилна регресия, която се дефинира при $k = n$, и класическата линейна регресия по метода на минималните модули за $\tau = 0.5$.

Основен резултат в тази глава е следната теорема, характеризираща праговата точка на LTQR оценките. Понеже LTQR оценките са частен случай на wGTE оценките на Vandev and Neykov (1998), то праговата им точка е следствие от Теорема 1 на Müller and Neykov (2003), понеже множеството $F = \{\rho_\tau(r_i(\beta)); i = 1, \ldots, n\}$ е $(\mathcal{N}(X) + 1)$–пълно, където $\mathcal{N}(X) = \max_{0 \neq \theta \in \mathbb{R}^p} \text{card}\{i \in \{1, \ldots, n\}; x_i^T \theta = 0\}$

Theorem 7.1 *Праговата точка на LTQR оценките е равна на*
$\frac{1}{n} \min\{n - k, k - \mathcal{N}(X) - 1\}$ *и достига максималната си стойност*
$\frac{1}{n} \lfloor \{n - \mathcal{N}(X) - 1\}/2 \rfloor$ *при* $\lfloor \{n + \mathcal{N}(X) + 1\}/2 \rfloor \leq k \leq \lfloor \{n + \mathcal{N}(X) + 2\}/2 \rfloor$.

Понеже $\mathcal{N}(X)$ е ограничено и не зависи от n, то за стойността на тримиране $k = \lfloor \{n + \mathcal{N}(X) + 1\}/2 \rfloor$ се гарантира максимална прагова точка, която асимптотично е равна на $1/2$ и не зависи от τ. От доказателството на тази теорема се вижда, че размерът на компактното множество съдържащо оценката зависи от $\{\tau, 1 - \tau\}^{-1}$ и въпреки, че за всяко τ праговата точка може да достигне $1/2$, а съответното максимално изместване при замърсяване на наблюденията с произволни да бъде минимално, то при $\tau \to 0$ или $\tau \to 1$ максималното изместване расте и може да приеме неограничено големи стойности.

7.3 Състоятелност на тримираните квантилни регресионни оценки

В тази секция е показано, че LTQR оценката (7.3) е състоятелна, относно параметрите на наклона на модела (7.1).

Без ограничение на общността, можем да предположим, че функцията на разпределение F на ε_i в (7.1) е дефинирана в $(-\infty, \infty)$. За тримиращия параметър k, дефиниращ LTQR оценките, ще използваме означението k_n, понеже зависи от обема на извадката n и ще предполагаме, че $\lim_{n\to\infty} k_n/n = \lambda \in \langle 0, 1\rangle$ съществува. За локационния модел (7.1), който не съдържа предиктори, Tableman (1994a) показва, че LTQR за $\tau = 0.5$ идентифицира медианата на интервал с минимална дължина Δ, за който $P(y_i \in \Delta) = \lambda$. За да формализираме това в общия случай с предиктори, да разгледаме следните предположения за процеса генериращ данните.

Предположение D. Векторът (x_i, ε_i) формира редица от независими и еднакво разпределени сл. векторни величини с краен $(1+\delta)$-ти момент за някое $\delta > 0$.

Предположение F. Нека функцията на разпределение F е непрекъсната и строго монотонна, имаща едномодална, ограничена и диференцируема плътност f в дефиниционната си област.

Да припомним, че за всеки интервал $\Delta(a, \lambda) = \langle F^{-1}(a), F^{-1}(a+\lambda)\rangle$, $a \in (0, 1-\lambda)$, и фиксирано $\tau \in (0, 1)$, Tableman (1994a, p. 390) доказва, че оценката по метода на минималните тримирани модули за локационния параметричен модел на едномерни данни с функция на разпределение F е състоятелна

$$\mu^*(\tau) = F^{-1}(a^*(\tau) + \tau\lambda), \tag{7.4}$$

където $a^*(\tau) = \arg\min_{a\in(0,1)} \int_{\Delta(a,\lambda)} \rho_\tau(\varepsilon - F^{-1}(a+\tau\lambda))dF(\varepsilon)$.

Нека свободният коефициент (intercept) е първата координата на вектора θ. LTQR оценката е състоятелна оценка за вектора $\theta^*(\tau) = (\mu^*(\tau), 0, \ldots, 0)^T + \theta$ за регресионния модел (7.1), където координатите на параметъра $\theta^*(\tau)$ съвпадат с координатите на θ с изключение на първата координата (свободния коефициент). Това означава, че свободният коефициент, получен чрез LTQR оценката съответства на $\theta_1 + \mu^*(\tau)$, където в общия случай $\mu^*(\tau) \neq 0$ ($\mu^*(\tau) = 0$ ако F е симетрично и $\tau = 1/2$, например).

Theorem 7.2 *Нека D и F са удовлетворени, а $\tau \in (0,1)$ е фиксиран квантил. Да предположим, че $\theta \in B$ и $\theta^*(\tau) = (\mu^*(\tau), 0, \ldots, 0)^T + \theta$, където $\mu^*(\tau)$ е дефиниран чрез (7.4), B е компактно параметрично пространство, Тогава LTQR оценката,*

дефинирана чрез $k_n = [\lambda n]$, $\lambda \in (0, 1\rangle$ *е състоятелна оценка на* $\theta^*(\tau)$, $\widehat{\theta}_n^{k_n}(\tau) \to \theta^*(\tau)$, *когато* $n \to \infty$.

Теоремата показва, че при условие F, LTQR оценката вярно идентифицира коефициентите пред предикторните променливи, но дава различни оценки за свободния коефициент. За да получим оценка за свободния коефициент, представляващ класическия τ квантил, е необходимо да използваме остатъците $r_i(\widehat{\theta}_n^{k_n}(\tau))$, получени чрез LTQR оценката, да пресметнем техния емпириричен квантил q_τ, който да добавим към LTQR оценката на свободния коефициент. Възражението срещу една такава процедура са нейните робастни свойства: тази нова оценка на свободния коефициент (асимптотично) има ограничена прагова точка от $\min\{\tau, 1-\tau\}$, която е подходяща за стойности на τ близки до 0.5 и твърде неуместна за квантилите близки до 0 и 1.

7.4 Примери

Понеже опитната и подобрената стъпка на FAST-GTE алгоритъма са стандартни квантилни регресионни процедури, то алгоритъмът лесно може да бъде реализиран, използвайки широко разпространен софтуер. Ние сме илюстрирали това на основата на библиотеката *quantreg*, предложена от Koenker (2005). В частност, първо сме сравнили оценките от класическата квантилна регресия с тези на LTQR оценките върху известни литературни данни и с разширено симулационно изследване, накрая робастните свойства на LTQR оценките са сравнени с други съществуващи методи за робастно оценяване на параметрите на квантилни регресионни модели.

7.4.1 Пример - клъстер CYB OB1

В тази секция са анализирани данните за клъстер CYB OB1 от звезден, състоящ се от 47 наблюдения и анализирани от Adrover et al. (2004) and Rousseeuw and Leroy (1987).

На всички плотове на Фигура 7.1 горе вдясно се виждат 4 несъгласувани наблюдения в предикторната променлива, които не следват тренда в мажоритарната част на данните. Данните бяха анализирани с класическата квантилна линейни регресии и LTQR оценките с различни проценити на тримиране и $\tau = 0.25, 0.50, 0.75$.

На плотовете на горния панел са дадени квантилните регресионни линиии - вляво на плота по всички наблюдения, докато на вдясно на плота по редуцираните без 4те наблюдения. Вижда се силното влияние на 4-те наблюдения върху класическата квантилна оценка.

Останалите плотове на Фигура 7.1 показват резултатите от LTQR оценяването по всички данни с 4%, 9%, 11% и 17% тримиране. Това съответства на отделяне (тримиране) на 2, 4, 5 и 8 наблюдения. Тримираните чрез LTQR процедура наблюдения са означени със следните символи: малки квадрати за $\tau = 0.25$, обратен триъгълник за $\tau = 0.50$ и триъгълник за $\tau = 0.75$. В случая на 4% процента на тримиране, съответната LTQR регресионна линии е повлияна от несъгласуваните наблюдения. Несъгласуваните наблюдения са идентифицирани при 9% на тримиране чрез LTQR процедурата и на практика резултатите са еквивалентни на тези, получени чрез класическата квантина регресия по редуцираните данни.

При по-висок процент на тримиране (11%, 17%) допълнителни наблюдения биват идентифицирани като несъгласувани, но въпреки това съответните регресионни линии са стабилни. Интересно е да си види, че идентифицирането на несъгласуваните наблюдения зависи от избора на τ, т.е., това са наблюденията които не следват модела и те са различни за различните квантили. Забелязва се, че LTQR регресионните линии за $\tau = 0.75$ с 11% и 17% процент на тримиране не се влияят от наличието на несъгласуваните наблюдения, както оценките на Adrover et al. (2004, Figure 2).

.

7.4.2 Симулационен експеримент

Поведението на QR и LTQR оценките е изследвано чрез симулационни Монте Карло експерименти. Сравнението е проведено чрез генериране на данни по модел на множествена линейна регресия с нееднородна грешка

$$\begin{aligned} y_i &= \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \sigma_i \varepsilon_i \quad \text{for} \quad i = 1, \ldots, 100, \\ \sigma_i &= \sqrt{\exp(0.11(x_{i1} + x_{i2}))}, \end{aligned}$$

където $\theta_0 = \theta_1 = \theta_2 = 0$, което не ограничение на общността, поради еквиваринтността на LTQR оценките, $\varepsilon_i \sim \mathrm{N}(z_\alpha, 1)$, z_α е α–квантил на N(0,1).

Разгледани са два типа разпределение на предикторите: в *1ви експеримент*, предикторите x_{i1} и x_{i2} са равномерно разпределени в интервала $[-10, 10]$, т.е., $x_{ij} \sim \mathrm{U}[-10, 10]$ за $j = 1, 2$; във *2ри експеримент*, предикторите x_{i1} и x_{i2} са нормално разпределени, т.е., $x_{ij} \sim \mathrm{N}(0, 1)$ за $j = 1, 2$. Замърсяването е внесено с модифициране на първите $m = \lfloor 100\epsilon \rfloor$ наблюдения за $\epsilon = 0.1, 0.2, 0.3$ както следва ($r = 2, 3, 4$): в *1ви експеримент*, $x_{ij} \sim \mathrm{U}[-30, -20]$ за $j = 1, 2$ и $y_i \sim \mathrm{U}[-10r, -10r + 10]$; във *2ри експеримент*, $(x_{i1}, x_{i2}, y_i)^T \sim \mathrm{N}_3(\mu, \Sigma)$ където $\mu = (-10, -10, -10r)^T$ и $\Sigma = 3I_3$ за $i = 1, \ldots, m$.

По този начин всички модифицирани наблюдения са несъгласувани в пространството на предикторните променливи с различна степен на влияние. Понеже резултатите са подобни за различен избор на r, в тази секция са дадепи само резултатите за *1ви експеримент* за $r = 2$ и *2ри експеримент* с $r = 3$.

Всички симуалционни експерименти са проведени 1000 пъти, за да изследваме поведението на малката извадка върху класическата QR и LTQR оценки при $\tau = (0.5, 0.75, 0.90)$, различни проценти на тримиране за данни без и със замърсяване с несъгласувани наблюдения.

Разпределението на тези оценки от резултатите на тези симулационни експеримент са дадени в бокс-плотове на Фугури 7.2-7.9. Панела от плотове на най-горния ред на тези фигури показва резултатите за свободния коефициент θ_0, докато останалите два панела показват резултатите от коефициентите на наклона θ_1 и θ_2, съответно. На всички плотове, "коректния" процент на тримиране, съответен на процента на замърсяване, е означен с точкови вертикални прави, докато истинските стойности на параметрие в симулациите са дадени с хоризонтални пунктирни линии.

Плотовете на фиг.7.2-7.3 демонстрират представянето на оценките в двата експеримента с равномерно и нормално разпределение на предикторите (QR съотвтетства на 0% тримиране). С увеличаване на процента на тримиране, оценките на свободния коефициент са неизместени за $\tau = 0.5$ (разпределението на грешката е симетрично), но изместването за $\tau = 0.75$ и $\tau = 0.90$ е право пропорционален на процента на тримиране. Причината за този ефект е дадена в секция 4, където е отбелязано че LTQR идентифицира сумата от свободния коефициент и $\mu^*(\tau) = F^{-1}(a^*(\tau) + \tau\lambda)$, според уравнение (7.4), което зависи от квантила τ и процента на тримиране $\lambda = \lim_{n\to\infty} k_n/n$.

QR и LTQR оценките на наклоните са дадени на плотовете на останалите панели. Тези оценки са неизместени, което е в съгласие с теорията, въпреки че се забелязва по-голяма вариабилност с нарастването на процента на тримиране, което се дължи на по-малкия брой наблюдения в извадките.

На плотовете на фигури 7.4-7.9 са представени резултатите за оценките от *1ви* и *2ри* експеримент, съответстващи на нарастване на процентът на несъгласуваните наблюдения $\epsilon = 0.1, 0.2, 0.3$. Резултати за LTQR оценките са близки до истинските, когато процентът на тримиране съвпада с процента на замърсяване на данните. Подобни са резутатите когато процента на тримиране е по-висок от процента на замърсяване на данните, т.е., $(1 - \lambda \geq \epsilon)$, наблюдаваме почти същите плотове като при незамърсените данни.

Забелязваме, че когато процентът на тримиране е по-малък от процента на замърсяване, т.е. $1 - \lambda < \epsilon$, то качеството на оценките са влошава . В подобни ситуации,

изместеността и разсейването в разпределението на оценките нараства значително, поради недостатъчната робстност на процедурата.

Процентът на несъгласувани наблюдения в реални данни е неизвестен. Една техника за автоматичен избор на $k_n = [\lambda n]$ или на процента $[(1-\lambda)n]100\%$ на тримиране на LTQR оценките по подобие на Čížek (2010), предложена за LTS оценките, която е адаптация на подхода на Gervini and Yohai (2002).

7.4.3 Сравнителен анализ с други робастни регресионни квантилни оценки

За да изучим поведението на малката извадка върху LTQR оценките, а така също да ги сравним с други робастни регресионни квантилни оценки, предложени от Rousseeuw and Hubert (1999) и Adrover et al. (2004), ние оценихме максималния ефект на замърсяване в едно наблюдение чрез LTQR оценките в термините на средно квадратичната грешка. За целта е възпроизведен симуалционния експеримент на Adrover et al. (2004). Всеки експеримент се състои от $n = 50$ наблюдения. Данните са генерирани чрез модел на множествена линейна регресия $y_i = \theta_0^T \mathbf{x}_i + u_i$ за $i = 1, \ldots, n$, където $\mathbf{x}_i^T := (1, x_{i1}, \ldots, x_{i,p-1})$, като всички $p-1$ предиктори са $N(0,1)$ разпределени, $u_i \sim N(z_\alpha, 1)$, z_α е α–квантила на $N(0,1)$, и $\theta_0 = 0$, който може да изберем по този начин, понеже LTQR оценките са регресионно еквивариантни. Замърсяването в една точка е въведено чрез заместване на последните $m = \lfloor \epsilon n \rfloor$ наблюдения чрез следните несъгласувани наблюдения: $y_i := y_0 = 5b$ and $\mathbf{x}_i^T := \mathbf{x}_0^T = (1, 5\mathbf{e}_1^T) \in R^p$ за $i = n - m + 1, \ldots, n$ за различни стойности на ϵ (виж Table 7.1), където $\mathbf{e}_1$ е първият елемент на каноничния базис на R^{p-1}. Следвайки Adrover et al. (2004), замърсяването в наклона b варира в грид от 0 до 10 със стъпка 0.1, за да бъде намерена най-неблагоприятната ситуация. Експериментът е проведен за различен брой регресионни предиктори $p = 2$ и $p = 5$, като за всяко p е повторен 500 пъти. Процентът на тримиране k_n на LTQR оценката беше избран равен на $\lfloor (1-\epsilon)n \rfloor$, а също така и за една по-малка стойност $\lfloor (1 - \epsilon - 0.1 * (1-\tau))n \rfloor$.

Максималната средно квадратична грешка $\|\widehat{\theta}_n^{k_n} - \theta_0\|^2$ (MaxMSE) е използвана като мярка за качеството на оценката и нейните стойности за различни квантили са дадени в Table 7.1. MaxMSE е пресметната за коефициентите на наклона и за свободния коефициент. Понеже LTQR оценката на свободния коефициент не е състоятелна оценка, то MaxMSE е пресметната за вектора на параметрите и по необходимост отразява изместеността на свободния коефициент, който не е свързан директно с изместването, причинено от замърсяването. Като допълнение в таблицата са дадени резултатите

за LTQR при по-голям процент на тримиране $1 - \lambda = \epsilon + 0.1 * (1 - \tau)$ отколкото е процентът на замърсяване. Причината за това е, че на практика процентът на несъгласуваните наблюдения в данните е неизвестен, поради което една по-реалистична процедура е да използва по-висок процент на тримиране, отколкото е очакванияъ процент на несъгласувани наблюдения в данните. Тези стойности на LTQR оценките са сравнени с резултатите от класическата Koenker–Bassett (K-B) и робастифицираната от Adrover et al. (2004) оценка на Koenker–Bassett (RobKB), както и други алтернативи като "maximum depth estimator" (MaxDep), предложена от Rousseeuw and Hubert (1999). За сравнението са използвани оригиналните резултати от работата на Adrover et al. (2004), които използват медианата вместо средно квадратичната грешка. Необходимите пресмятания за MaxDep оценките бяха проведени с FORTRAN програмата на Van Aelst et al. (2002).

LTQR оценката при $1 - \lambda = 0$, т.е., за $k_n = n$ се редуцира до класическата оценка на Koenker–Bassett и нейната стойност за данните без замърсяване при $\epsilon = 0$ са използвани като контролна стойност на средно квадратичната грешка на QR оценката, дадена в Table 7.1. За положителни нива на замърсяване, качеството на представяне на LTQR оценките пропорционално на нивата на замърсяване ϵ, но не зависи значимо от оценявания квантил. Това се вижда от резултатите за нивата на замърсяване $\epsilon \in (0.10, 0.12)$: MaxMSE на LTQR нараства с 50% ако $\tau = 0.50$ и достига 0.90. От друга страна, за най-робастната алтернативна оценка RobKB, предложени от Adrover et al. (2004), се забелязва увеличаване на изместеността $\epsilon \in (0.10, 0.12)$ с 50–100% ако $\tau = 0.50$ достига 0.75 и продължава да нараства неограничено ако $\tau = 0.90$.

В общия случай, LTQR оценката превъзхожда другите методи за оценяване за $\tau > 0.5$ както и за по-високи проценти на замърсяване, като следствие от независмостта на нейната прагова точка от квантила. В допълнение, LTQR оценката се представя по-добре в сравнение с другите методи при високите квантили като $\tau = 0.90$, независимо от процента на замърсяване на данните. Оценките RobKB и MaxDep се представят по-добре за квантили близко до $\tau = 0.50$ и по-малък процент на замърсяванеане. Това е очаквано за RobKB оценките, понеже този клас оценки при даден квантил се основава на всевъзможните разлики на остатъците, според Stromberg et al. (2000).

7.5 Резюме и изводи

Глава 7 е посветена на квантилната линейна множествена регресия (Koenker, 2005). Популярността на тази регресия се дължи на възможността за разкриване на линейни

структури в целия спектър (за всеки квантил $\tau \in (0,1)$) на наблюденията на зависимата променлива, за разлика от класическата линейна множествена регресия, чрез която се дефинира само моделът на средната стойност на наблюденията на зависимата променлива. Квантилната регресия е робастна спрямо наличието на насъгласувани наблюдения в зависимата променлива, за $\tau = 0.5$ е еквивалентна на метода на минималните модули за оценяване на параматрите на линейния регресионен модел. За оценяване на неизвестните параметри на тази регресия са предложени ефективни алгоритми на основата на линейното програмиране. Недостатък на квантилните регресионни оценки е, че не са робастни спрямо наличието на несъгласувани наблюдения в пространството на предикторните променливи. За отстраняване на този недостатък, в тази глава са предложени тримираните квантилни регресионни оценки $LTQR(k)$. Теореми 7.1 и 7.2 характеризират праговата им точка, асимптотична състоятелност и нормалност, съотвтено. За приближено пресмятане на $LTQR(k)$ оценките на неизвестните параметри е използван обобщения FAST-TLE алгоритъм. Поведението на $LTQR(k)$ оценките, в сравнение с оценките на класическата квантилна регресия, е илюстрирано чрез примери и разширено симулационно изследване с методите Монте Карло.

Изборът на параметъра на тримиране за този клас оценки е изключително важен и от решаващо значение за качествен анализ на данните. Числените експерименти подсказват, че ако процента на тримиране е по-ниско от нивото на замърсяване, това може да доведе до лоши оценки, но всеки по-висок процент на тримиране дава стабилни резултати. Поради това, едно правило при анализ на данни е да се работи с един консервативен избор на процента на тримиране или процента на тримиране да бъде оценен по данните както предлага Čížek (2010) или Gervini и Yohai (2003).

Глава 8

Робастен избор на предиктори в регресионен тип задачи с големи размерности чрез тримирани правдоподобни оценки с пенализация

Активна област на изследване в статистиката през последните години са задачите с висока размерност. Характерното за тези задачи е, че броят на променливите p е съизмерим или по-голям от броя на наблюденията на извадката n. За преодоляване на проблема с размерността се създават нови методи на основата на регуляризацията по Тихонов. Това означава, че оптимизирането на класически статистически функционал за получаване на оценки се трансформира в оптимизиране на класическия статистически функционал с ограничения на подходяща норма на неизвестните параметри. Първата L_1 норма или свързани с нея се използват широко в статистиката за получаване на оценки на неизвестните параметри в обобщените линейни модели. Причината за това е, че решението на съответната оптимизационна задача се достига върху връх на многомерен симплекс, което означава че голяма част от неизвестните параметри се оценяват като нула. По този начин се осъществява автоматичен избор (скрининг) на значими променливи, което представлява изключителен интерес за приложната статистика. Ясно е, че за редуциране влиянието на несъгласуваните наблюдения в данни с висока размерност са необходими робастни методи, основани на тези принципи. В тази глава се разглеждат GTE оценки с регуляризация по Тихонов за анализ на данни с висока размерност. Въведен е клас $wGTE(k)$ оценки с пенализация, ограничение върху нормата на вектора от неизвестни параметри. Чрез подходящ скрининг

на предикторните променливи, на основата на условните маргинални разпредления на зависимата променлива чрез TLE оценките, се селектира робастно определен брой $q < \min(n, p)$ потенциални предиктори.

Използването на регуляризация (пенализация) в задачи с ултра високи размерности е ограничено. Според Fan and Lv (2008), задачите с ултра високи размерности са от порядъка $\log(p) = O(n^{\alpha})$ за $0 < \alpha < 1$ при $p >> n$. Fan and Lv (2010) предлагат скрининг на предикторните променливи чрез SIS и $ISIS$ процедури за селектиране на най-значимите предиктори, за да редуцират драстично броя на предикторите от p до q, където $q << \min(n, p)$. Селекцията се основава на ранжирането на стойностите на статистическия функционал, като минимална остатъчна сума от квадрати или максимум на функцията на правдоподобие, които са пресмятат за всеки предиктор по отделно. По този начин се избират потенциално q предиктора. SIS и $ISIS$ процедурите не са робастни спрямо наличието на несъгласувани наблюдения в предикторните променливи, понеже се основават на класически статистически методи за оценяване. Нещо повече, дори SIS процедурите, основани на робастните M-оценки с пенализация (регуляризация), не са робастни срямо този тип наблюдения, Bühlmann and van der Geer (2011). За преодоляването на този проблем в тази глава са въведени робастни оценки по ММП с пенализация, едностъпкови и двустъпкови SIS и $ISIS$ процедури, основани на тримиране чрез $wGTE$ оценките. Характеризирането на праговите точки на тези тримирани процедури се свежда до известни резултати от глави 1-7, отнасящи се до праговата точка на $wGTE$ оценките или $WTLE(k)$ оценките за обобщените линейни модели от дисперсионната фамилия от разпределения. Изследвано е поведението на тези оценки с методите Монте Карло.

8.1 Въведение

Нека (y_i, x_i^T), за $i = 1, \ldots, n$, са независими наблюдени, y_i е i-то наблюдение на зависимата променлива Y и $x_i^T = (x_{11}, x_{12}, \ldots, x_{1p})$ е i-ти ред на матрицата на предикторните променливиX. Предполагаме, че y_i зависи x_i чрез линейния предиктор $\eta_i(\theta) = x_i^T\theta$ на целевата функция $L(\eta_i(\theta), y_i)$. Без ограничение на общността нека $L(\eta_i(\theta), y_i)$ е логаритъма на правдоподобието.

Definition 8.1

$$\widehat{\theta}_{n,MLE} := \arg\max_{\theta} \left\{ \ell_n(\theta) = \sum_{i=1}^{n} L(\eta_i(\theta), y_i) \right\}. \tag{8.1}$$

Definition 8.2 *МПО с ограничение върху нормата на вектора от неизвестни параметри (Penalized MLE) се дефинира като*

$$\widehat{\theta}_{n,PMLE} := \arg\max_{\theta} \left\{ \ell_n(\theta) - n \sum_{j=1}^{p} p_\lambda(|\theta_j|) \right\}, \tag{8.2}$$

където, $p_\lambda(.)$ е функцията на ограниченията, а $\lambda \geq 0$ е регулязиращ параметър.

Благодарение на пенализиращата функция, някои от координатите на θ могат да се редуцират до нула автоматически, т.е да бъде получено разредена (sparse) оценка. Това е еквивалентно на избор на значими предикторни променливи, което е актуална задача на статистиката. При големи стойности на λ оптимизационната процедура автоматично избира опростен модел и обратно, чрез малки стойности на λ и модел с голям брой параметри се избират комплексни модели с много предиктори. В реалните задачи параметърът λ е неизвестен. Различни техники за оценяването на λ е по данните, чрез формирането на обучаваща и валидиращи иазвадки, подробности са дадени в Bühlmann and van der Geer (2011).

Често използвани функции на ограниченията са: 1) L_1 нормата $p_\lambda(|\theta_j|) = \lambda |\theta_j|$ известна в статистическата литература като LASSO (least absolute shrinkage and selection operator), предложена от Tibshirani (1996); L_q-нормата $p_\lambda(|\theta|) = \lambda |\theta_j|^q$ за $0 < q \leq 2$, предложена от Frank and Friedman (1993); SCAD, предложена от Fan and Li (2001)

$$p_\lambda(|\theta|) = \begin{cases} \lambda|\theta| & \text{if } |\theta| < \lambda, \\ \frac{(a^2-1)\lambda^2-(|\theta|-a\lambda)^2}{2(a-1)} & \text{if } \lambda \leq |\theta| < a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\theta| \geq a\lambda, \end{cases} \tag{8.3}$$

където $a = 3.7$, или MCP функцията $p'_\lambda(|\theta_j|) = (\lambda - |\theta|/a)_+$ предложена от Zhang (2008).

SCAD и MCP са неизпъкнали функции, които притежават свойството "oracle". Това означава, че съществените променливи могат да бъдат идентифицирани с голяма вероятнотност, докато останалите променливи биват изключвани от модела. Други функции на ограниченията върху параметрите са дискутирани от Antoniadis et al. (2011), както и методи за избор на оптимална стойност на регуляризиращия параметър в обобщените линейни модели.

Оптимазионната задача (8.2) е изпъкнала ако $\ell_n(\theta)$ е вдлъбната и е използвана L_1 нормата. В общия случай, за фиксирана стойност на параметъра λ, PMLE със SCAD функция на ограничения върху параметрите не е изпъкнала оптимизационна задача. Поради това са разработени специфични изчислителни алгоритми. Така например,

Zou and Li (2008) предлагат един ефективна локална линейна апроксимация (LLA) за оптимизаране на (8.2) със SCAD функция на пенализиране. Идеята е да се приближи (мажорира) функцията SCAD чрез линейна функция на m-та итерация

$$p_\lambda(|\theta|) \approx p_\lambda(|\theta^{(m)}|) + p'_\lambda(|\theta^{(m)}|)(|\theta| - |\theta^{(m)}|). \tag{8.4}$$

Вследстие на това пенализираната функция на правдоподобие (8.2) се редуцира до

$$\ell_n(\theta) - n \sum_{j=1}^{p} w_j^{(m)} |\theta_j| , \tag{8.5}$$

където $w_j^{(m)} = p'_\lambda(|\theta_j^{(m)}|)$. Чрез квадратичната апроксимация на $\ell_n(\theta)$ в $\theta^{(m)}$ оптимзационния проблем се превръща в итерационен претеглен МНК с L_1 пенализация, тясно свързан с адаптивната LASSO процедура за оценяване (Zou, 2006), която дава разредено решение, което води до автоматично селектиране на предиктори. Библиотеката SIS, разработена от Fan et al. (2009) реализира LLA алгоритъма в средата на R (R Development Core Team, 2012). Zou and Li (2008) дискутират също така други итеративни алгоритми като претеглен МНК с пенализираща L_1 чрез LARS алгоритъма, прдложен от Efron et al. (2004).

Добре известно е, че МНК и ММП не са робастни спрямо малък процент несъгласувани наблюдения в данните (Huber, 1981; Hampel et al., 1986; Maronna et al., 2006). Следователно МНК и ММП с пенализация също не са робастни спрямо несъгласувани наблюдения в данните. За преодоляването на този проблем са използвани робастните М-оценки с пенализация (Fan and Li, 2001; Fan and Lv, 2010). Известно е, че М-оценките не са робастни спямо несъгласувани наблюдения в предикторните променливи на линейните регресионни модели и следователно М-оценките с пенализация не са робастни също.

Твърде малко са робастни алтернативи на ММП, които са едновременно робастни спрямо несъгласувани наблюдения в зависимата и предикторни променливи, например претегления ММП, предложен от Markatou et al. (1997) и метода на тримираното правдоподобие (TLE) от Neykov and Neytchev (1990). Нито една от тези алтернативи на ММП не е използвана в задачите за статистическо моделиране с висока размерност. Ето защо, целта на тази глава е да предложи алтернатива на ММП с пенализция, основана на TLE оценките с пенализация, което ще доведе до редуциране на влиянието на несъгласувани наблюдения в предикторните променливи, а така също до робастно селектиране на предикторни променливи. От дефиницията на TLE оценките следва, че съответната TLE оптимизационна задача с пенализация се редуцира до МП задача с пенализация, но върху някоя подизвадки с обем от $k > n/2$ наблюдения от n.

По този начин методологията за селектиране на значими предикторни променливи в задачи с големи размерности се редуцира до стандартна МП задача с пенализация, комбинаторна по набюденията. Предимствата на този подход пред класическите МПО са предмет на настоящата глава.

8.2 Тримирани правдоподобни оценки с пенализация

В тази секция се разглеждат GTE оценки (4) с пенализация, регуляризация по Тихонов, за анализ на данни с висока размерност.

Definition 8.3 *GTE оценките с ограничения на нормата от неизвестните параметри се дефинира като*

$$\min_{\theta\in\Theta^p} S^P_{n,k}(\theta) = \min_{\theta\in\Theta^p}\left\{\min_{I\in I_k}\sum_{i\in I} f_i(\theta) + k\sum_{j=1}^{p} p_\lambda(|\theta_j|)\right\} \tag{8.6}$$

$$= \min_{I\in I_k}\left\{\min_{\theta\in\Theta^p}\left[\sum_{i\in I} f_i(\theta) + k\sum_{j=1}^{p} p_\lambda(|\theta_j|)\right]\right\} \tag{8.7}$$

$$= \min_{I\in I_k}\left\{\min_{\theta\in\Theta^p}\sum_{i\in I}\left[f_i(\theta) + \sum_{j=1}^{p} p_\lambda(|\theta_j|)\right]\right\}. \tag{8.8}$$

От дефницията следва, че GTE с пенализация е класическа МПО с пенализация, но дефинирана върху всевъзможните подизвадки с обем k от n наблюдения. Това означава, че FAST-GTE алгоритъмът може да бъде използван за намирането на приближено решение на оптимизационната задача. За фиксирано λ, праговата точка GTE с пенализация може да бъде характеризирана чрез индекса на d–пълнота на множеството от функции $F_\lambda = \{f_i(\theta) + \sum_{j=1}^{p} p_\lambda(|\theta_j|),\ i = 1,\ldots,n\}$. Нека F е d–пълно. От включването

$$\{\theta\in R^p : \max_{j\in J}(f_j(\theta) + \sum_{l=1}^{p} p_\lambda(|\theta_l|)) \le C\} \subseteq \{\theta\in R^p : \max_{j\in J} f_j(\theta) \le C\}$$

следва, че F_λ е d–пълно, понежее F е d–пълно. В действителност F_λ е 1–пълно при условие, че $\{\theta\in R^p : \sum_{l=1}^{p} p_\lambda(|\theta_l|) \le C\}$ се съдържа в компактно множество. Това е очевидно за изпъкнали функции с пенализация като L_1, докато за неизпъкнали като SCAD е необходимо привличането на обобщената техника на d–пълнота, предложена от Dimova and Neykov (2004). От изчислителна гледна точка, LLA оценките, дефинирани чрез (8.4) могат да бъдат използвани за получаване на приближено решение

на оптимизационната GTE задача със SCAD пенализараща функция. Понеже LLA е изпъкнала мажоранта на SCAD функцията, това осигурява d–пълнота на съответното множество от функции F_λ. Следователно винаги съществува решение на GTE задачата с пенализация, ако множеството F_λ е d–пълно. Ще отбележим, че е възможно решението да не бъде единствено, поради което е необходимо да бъдат наложени допълнителни ограничения.

При $k = n$, подхядящ избор на $f_i(\theta)$ и $p_\lambda(.)$ GTE с пенализация се редуцира до различни класически статистически оценки с пенализация като LASSO, предложена от Tibshirani (1996), МПО с L_1 пенализация на Tibshirani (1997), ММО с пенализираща функция SCAD, въведени от Fan and Li (2001), LAD-LASSO от Wang et al. (2007), и робастни M-оценки с пенализация от Fan and Li (2001).

За съжаление, тези оценки не са робастни спрямо наличието на несъгласувани наблюдения в предикторните променливи в средата на обобщените линейни модели. Едно изключение са MCD оценките с пенализация, разгледани от Croux and Haesbroeck (2010) и LTS оценките с пенализация, предложени от Alfons et al. (2012), които оценки са дефинирани върху подизвадки от k наблюдения от извадка с обем n. Тези два класа робастни оценки с висока прагова точка могат да бъдат получени като часни случаи на GTE оценките с пенализация ако заместим $f_i(.)$ с разстоянието на Mahalanobis и остатъци на квадрат на линейния регресионен модел, съответно. По този начин се дефинира нов клас статистически оценки, който ще наричаме тримирани правдоподобни осенки с пенализация (PMTLE, Penalized Maximum Trimmed Likelihood Estimator)

Definition 8.4 *PMTLE се дефинира като частен случай на GTE оценките с пенализация, когато $f_i(\theta)$ в (8.6) бъдат заместени с отрицателния логаритъм на правдоподобие на i-то наблюдение.*

PMTLE оценката може да достигне най-висока прагова точка при условие, че множеството F_λ е d–пълно. Понеже F_λ наследява индекса на d–пълнота на F, поради това е достатъчно да бъде определен индексът на d–пълнота на F. Индексите на пълнота на множества F_λ с изпъкнали функции на пенализация са известни, понеже индексът на d–пълнота за случаите, когато $p < n$ е определен за различни множества от функции от Vandev and Neykov (1993), Müller and Neykov (2003), Dimova and Neykov (2004b), Dimova and Neykov (2005), Neykov et al. (2012a), Neykov et al. (2012b). Като следствие от това, праговата точка на PMTLE е равна на $\frac{1}{n}\min\{n-k, k-\mathcal{N}(X)-1\}$.

Праговата точка на GTE оценката за данни с висока размерност $p >> n$ не може да бъда характеризирана директно. За тази ситуация ще напомним Bühlmann and

van der Geer (2011), чиято монография е посветена на тази тематика: "The philosophy that will generally rescue us, is to 'believe' that in fact only a few coordinates of the θ are non-zero". Следвайки тази философия, в следващите секции ще бъдат разгледани два подхода за редуциране на размерността на този тип задачи с подходящ скрининг на предикторите.

За да бъде редуцирано влиянието на несъгласуваните наблюдения върху избора на значими параметри, ние препоръчваме използването на BIC с пенализация, основан на тримирне, който е дефиниран чрез $\text{PTBIC}(\lambda) = -2\log(S^P_{n,k}(\widehat{\theta})) + df(\lambda)\log(k)$, където $S^P_{n,k}(\widehat{\theta})$ е PMTLE оценката с $df(\lambda)$ степени на свобода, определени по броя на ненулевите координати на $\widehat{\theta}$. PTBIC се редуцира до BIC ако $k = n$ и $\lambda = 0$.

8.3 Робастни SIS и ISIS процедури, основани на тримиране

Използването на пенализацията в задачи с ултра високи размерности е ограничено. Според Fan and Lv (2008), задачите с ултра високи размерности са от порядъка $\log(p) = O(n^\alpha)$ за $0 < \alpha < 1$ при $p >> n$. Fan and Lv (2010) предлагат скрининг процедура (SIS) за селектиране на най-значимите предиктори, която да редуцира драстично броя на предикторите от p до q, където $q << \min(n, p)$. Селекцията се основава на ранжирането на целевата функция, която може да бъде остатъчна сума от квадрати или функция на правдоподобие, пресметнати за всеки предиктор поотделно. По този начин се избират оптималните q предиктора.

SIS процедурите не са робастни, понеже се основават на класически статистически методи за оценяване. За преодоляването на този проблем сме въвели SIS процедури, оновани на тримиране.

8.3.1 Ранжиране на променливи

Без ограничение на общността, нека $L(.)$ е отрицателния логаритъм на функцията на правдоподобие. Всяка друга целева функция като остатъчна сума от квадрати или квази правдоподобие би могла да бъде $L(.)$. Нека да дефинираме маргиналните

функции на правдоподобие на предикторната променлива X_j, за $j = 1, \ldots, p$, чрез

$$L_0 = \min_{\theta_0} \frac{1}{n} \sum_{i=1}^{n} L(y_i, \theta_0), \tag{8.9}$$

$$L_j = \min_{\theta_0, \theta_j} \frac{1}{n} \sum_{i=1}^{n} L(y_i, \theta_0 + x_{ij}\theta_j), \tag{8.10}$$

където L_j е целевата функция с линеен предиктор $\theta_0 + x_{ij}\theta_j$ за модела на y_i.

SIS процедурата се състои в следните стъпки: пресмятане стойностите на $L_1, \ldots, L_p$, ранжирането им $L_{\nu(1)} \leq \ldots \leq L_{\nu(q)} \leq \ldots \leq L_{\nu(p)}$, където $(\nu(1), \ldots, \nu(p))$ е пермутацията на индексите $(1, \ldots, p)$, селектиране на вектора от предиктори $(X_{\nu(1)}, X_{\nu(2)}, \ldots, X_{\nu(q)})$ за същински анализ на следващ етап. По този начин SIS селектира най-значимите маргинални предиктори за $j = 1, \ldots, q$. Необходимите пресмятания за L_j не изискват специални изчислителни ресурси, понеже задачата се свежда до оценяване на два параметъра в едномерна задача.

Fan and Lv (2008) препоръчват за избор на $q = \lfloor n/\log n \rfloor$ в модели на множествена линейна регресия и $q = \lfloor n/(2\log n) \rfloor$ за Поасонова регресия. Параметъра q се избира достатъчно голям, но $q < n$, за да осигури надежден скрининг. Данните трябва да бъдат задължително стандартизиране.

8.3.2 Псевдо-правдоподобни оценки с пенализация

Твърде вероятно е някои от селектираните q предикторни променливи да са маловажни. За да отстранят подобни ситуации Fan and Song (2010) препоръчват използването на ММП с пенализация, за отделяне на незначимите предиктори от q. Нека $X_1, \ldots, X_q$ са избраните предиктори чрез процедурата SIS $x_{i,q} = (x_{i1}, \ldots, x_{iq})^T$ и да предифинираме $\theta = (\theta_1, \ldots, \theta_q)^T$.

Минимизирането на целевата функция с пенализация

$$\frac{1}{n} \sum_{i=1}^{n} L(y_i, \theta_0 + x_{i,q}^T \theta) + \sum_{j=1}^{q} p_\lambda(|\theta_j|), \tag{8.11}$$

ще даде така наречените разредени (sparse) регресионни оценки, с много нули в координатите на θ, докато параметъра на регуляризация може да бъде определен чрез процедура за кросвалидация. Нека броят на ненулевите координати на θ означим с $\widehat{\mathcal{M}}$.

8.3.3 Робастни SIS-SCAD процедури, основани на тримиране

Двустъпковата SIS процедура се основава на ММП или на ММП с пенализация, които не са робастни спрямо наличието на несъгласувани наблюдения в данните. Поради тази причина в тази секция е въведена SIS процедура, основана на тримиране. TSIS-SCAD се дефинира като:

$$\min_{I\in I_k}\begin{cases} \text{SIS процедура} \\ \begin{cases} L_0^{trim} := \min\limits_{\theta_0} \frac{1}{k}\sum\limits_{i\in I} L(y_i,\theta_0) \\ L_j^{trim} := \min\limits_{\theta_0,\theta_j} \frac{1}{k}\sum\limits_{i\in I} L(y_i,\theta_0+x_{ij}\theta_j) \end{cases} \\ (X_1,\ldots,X_q) := (X_{\nu(1)},X_{\nu(2)},\ldots,X_{\nu(q)}) \\ \text{SIS}-\text{SCAD процедура} \\ S_{k,n}^{P,trim} := \min\limits_{\theta_0,\theta}\left(\frac{1}{k}\sum\limits_{i\in I} L(y_i,\theta_0+x_{i,q}^T\theta)+\sum_{j=1}^q p_\lambda(|\theta_j|)\right). \end{cases} \tag{8.12}$$

Следователно за всички k-подмножества свързаните SIS оптимизационни задачи (8.12) трябва да бъдат последователно решени и оценката TSIS-SCAD с пенализация се дефинра като оценка за някоя подизвадка с обем k за която $S_{k,n}^{P,trim}$ е минимална.

Предикторните променливи трябва да бъдат стандартизирани чрез тяхните средни и дисперсии, за да бъде редуцирано влиянието на мащаба на данните.

За линейната множествена и Поасонова регресия индекса на d-пълнота на $F_j = \{L(y_i,\theta_0+x_{ij}\theta_j) \text{ for } i=1,\ldots,n\}$ е $\mathcal{N}(X_j)+1$ for $j=1,\ldots,p$, според Müller and Neykov (2003). Ето защо, праговата точка TLE в едномерния случай на TSIS процедурата е равна на $\frac{1}{n}\min\{n-k,k-\mathcal{N}(X_j)-1\}$, докато за TSIS-SCAD оценката е $\frac{1}{n}\min\{n-k,k-\mathcal{N}(X_{n\times q})-1\}$, според Müller and Neykov (2003). Следователно праговата точка на двустъпковата TSIS-SCAD процедура е равна на $\frac{1}{n}\min\{n-k,k-D-1\}$, където $D=\max[\max_j \mathcal{N}(X_j),\mathcal{N}(X_{n\times q})]$.

Тази прагова точка се максимизира за $\lfloor\{n+D+1\}/2\rfloor \le k \le \lfloor\{n+D+2\}/2\rfloor$ и е равна на $\frac{1}{n}\lfloor\{n-D-1\}/2\rfloor$.

Вместо да задаваме минимална стойност на k, за да постигнем максимална прагова точка понякога е удачно да избираме този параметър както следва $k=\lfloor\alpha n\rfloor$ за $\alpha\in(0.5,1]$, при условие че предикторите са непрекъснати. Например, изборът $\alpha=0.80$

осигурява едновременно робастност срещу 20% несъгкасувани наблюдения в данните, но води до по-висока ефективност на оценката.

8.3.4 Итеративен избор на предикторни променливи

Възможно е резултатите от SIS процедурата да не са удовлетворителни, понеже някои от предикторите могат да не са зависими маргинално със зависимата променлива, но участието им в линейна комбинация с останалите предиктори да минимизира съответната целева функция на зависимата променлива. Поради тази причина Fan and Lv (2008), Fan et al. (2009), и Fan and Song (2010) предлагат итеративна SIS (ISIS) процедура.

В първата стъпка на ISIS процедурата се провежда двустъпковата SIS процедура, за да селектира подмножеството $\widehat{\mathcal{M}}_1$ от предиктори. Тогава Fan et al. (2009) предлагат да бъдат пресметнати следните оптимизационни задачи, чрез които преценява важността на предиктора X_j, който не е бил селектиран от скрининга със SIS-SCAD процедурата:

$$L_j^{(2)} = \min_{\theta_0, \theta_{\widehat{\mathcal{M}}_1}, \theta_j} n^{-1} \sum_{i=1}^{n} L(y_i, \theta_0 + x_{i,\widehat{\mathcal{M}}_1}^T \theta_{\widehat{\mathcal{M}}_1} + x_{ij}\theta_j), \tag{8.13}$$

за $j \in \widehat{\mathcal{M}}_1^c = \{1, \ldots, p\} \setminus \widehat{\mathcal{M}}_1$, където $x_{i,\widehat{\mathcal{M}}_1}$ е подвектор на x_i, състоящ се от онези елементи в $\widehat{\mathcal{M}}_1$.

Оптимизационната задача (8.13) е с ниска размерност. Допълнителният принос на предиктора X_j при дадено $\widehat{\mathcal{M}}_1$ може да бъде оценен чрез маргиналния критерий на отношението на правдоподобие (разликата на две функции на отклоненията в рамките на обобщените линейни модели):

$$L_j^{LR} = \min_{\theta_0, \theta_{\widehat{\mathcal{M}}_1}} n^{-1} \sum_{i=1}^{n} L(y_i, \theta_0 + x_{i,\widehat{\mathcal{M}}_1}^T \theta_{\widehat{\mathcal{M}}_1}) - L_j^{(2)}. \tag{8.14}$$

След нареждане на L_j^{LR} във възходящ ред за $j \in \widehat{\mathcal{M}}_1^c$ се селектират индексите, съответни на най-малките m_2 елемента, които формират множеството $\widehat{\mathcal{A}}_2$.

След тази предварителна скринингова стъпка се провежда оценявне по ММП с пенализиция, за да бъде получена една разредена оценка

$$\theta_2 = \underset{\theta_0, \theta_{\widehat{\mathcal{M}}_1}, \theta_{\widehat{\mathcal{A}}_2}}{\arg\min} \left(n^{-1} \sum_{i=1}^{n} L(y_i, \theta_0 + x_{i,\widehat{\mathcal{M}}_1}^T \theta_{\widehat{\mathcal{M}}_1} + x_{i,\widehat{\mathcal{A}}_2}^T \theta_{\widehat{\mathcal{A}}_2}) + \sum_{j \in \widehat{\mathcal{M}}_1 \cup \widehat{\mathcal{A}}_2} p_\lambda(|\theta_j|) \right). \tag{8.15}$$

В резултат на това получаваме множеството $\widehat{\mathcal{M}}_2$ от активни индекси от ненулевите координати на θ_2. По този начин процедурата дава възможност за отделяне на предиктори, които са били селектирани на предходна стъпка с индекси от $\widehat{\mathcal{M}}_1$. Процесът, който итеративно включва и изключва предиктори може да бъде повторен неколкократно, докато не получим множество от индекси $\widehat{\mathcal{M}}_l$ с размерност q или $\widehat{\mathcal{M}}_l = \widehat{\mathcal{M}}_{l-1}$. По този начин се определя крайната оценка на параметрите θ_l.

8.3.5 Робастно итеративно селектиране на предиктори, основано на тримиране

Подобно на двустъпковата SIS-SCAD процедура за оценяване, ние можем да заменим оптимизационния проблем (8.13) и (8.15) с тяхната тримирана версия и да решим задачата за всевъзможните подизвадки с обем k от n, за да намерим подмножеството, за което (8.15) е минимална. По този начин двустъпковата тримирана ISIS-SCAD (TISIS-SCAD) процедура за оценяване се дефинира като

$$\min_{I\in I_k}\begin{cases} \text{ISIS процедура} \\ \begin{cases} L_{0,\widehat{\mathcal{M}}_1}^{trim} = \min\limits_{\theta_0,\theta_{\widehat{\mathcal{M}}_1}} k^{-1}\sum\limits_{i\in I} L(y_i,\theta_0 + x_{i,\widehat{\mathcal{M}}_1}^T\theta_{\widehat{\mathcal{M}}_1}) \\ L_j^{(2,trim)} = \min\limits_{\theta_0,\theta_{\widehat{\mathcal{M}}_1},\theta_j} k^{-1}\sum\limits_{i\in I} L(y_i,\theta_0 + x_{i,\widehat{\mathcal{M}}_1}^T\theta_{\widehat{\mathcal{M}}_1} + x_{ij}\theta_j) \end{cases} \\ \text{ISIS}-\text{SCAD процедура} \\ \tilde{S}_{k,n}^{P,trim} = \min\limits_{\theta_0,\theta_{\widehat{\mathcal{M}}_1},\theta_{\widehat{\mathcal{A}}_2}} \left(k^{-1}\sum\limits_{i\in I} L(y_i,\theta_0 + x_{i,\widehat{\mathcal{M}}_1}^T\theta_{\widehat{\mathcal{M}}_1} + x_{i,\widehat{\mathcal{A}}_2}^T\theta_{\widehat{\mathcal{A}}_2}) \right. \\ \qquad\qquad \left. + \sum\limits_{j\in\widehat{\mathcal{M}}_1\cup\widehat{\mathcal{A}}_2} p_\lambda(|\theta_j|) \right) \end{cases} \tag{8.16}$$

Нека $r = |\widehat{\mathcal{M}}_1 \cup \widehat{\mathcal{A}}_2|$ е с кардиналност $\widehat{\mathcal{M}}_1 \cup \widehat{\mathcal{A}}_2$ и $\widehat{\mathcal{M}}_1^* = \widehat{\mathcal{M}}_1 + 1$. Подобно на предишната секция можем да заключим, че множествата $(\mathcal{N}(X_{n\times r})+1)$ и $(\mathcal{N}(X_{n\times\widehat{\mathcal{M}}_1^*})+1)$ са пълни, и това е минималния брой наблюдения, който гарантира идентифицируемост на θ (Müller and Neykov, 2003). Следователно праговата точка на TLE оценките, дефинрани чрез (8.16) е равна на $\frac{1}{n}\min\left\{n-k, k-\mathcal{N}(X_{n\times\widehat{\mathcal{M}}_1^*})-1\right\}$, докато за TISIS-SCAD оценката с пенализация е равна на $\frac{1}{n}\min\{n-k, k-\mathcal{N}(X_{n\times r})-1\}$. Използвайки означенията $\tilde{D} = \max[\mathcal{N}(X_{n\times\widehat{\mathcal{M}}_1^*}), \mathcal{N}(X_{n\times r})]$, получаваме че праговата точка на

двустъпковата TISIS-SCAD процедура (8.16) е равна на $\frac{1}{n}\min\left\{n-k, k-\tilde{D}-1\right\}$. Тази прагова точка се максимизира за $\left\lfloor\left\{n+\tilde{D}+1\right\}/2\right\rfloor \leq k \leq \left\lfloor\left\{n+\tilde{D}+2\right\}/2\right\rfloor$ и е равна на $\frac{1}{n}\left\lfloor\left\{n-\tilde{D}-1\right\}/2\right\rfloor$.

8.4 Симулационно изследване

В тази секция е изследвано поведението на SIS-SCAD, ISIS-SCAD и техните тримирани версии в симулационни експерименти на линейна множествена и лог-линейна (Поасонова) регресия. Две различни конфигурации от данни са симулирани и резултатите от тях са дискутирани.

8.4.1 Мерки за поведение на оценките

Според симулационния експеримент, описан в следващата секция, обучаващата извадка от данни е генерирана със и без замърсяване с несъгласувани наблюдения и параметрите на моделите θ са оценени с различни методи. В допълнение, n тестови наблюдения са генерирани със същата схема, но без наличие на несъгласувани наблюдения. Да означим предикторните променливи в тестовите данни с $\tilde{x}_i$, а зависимата променлива с $\tilde{y}_i$, за $i = 1, \ldots, n$. За качеството на линейните предиктори (свързващата функция в термините на обобщените линейни модели) $\tilde{\eta}_i = \tilde{x}_i^T\widehat{\theta}$ за линейната множествена регресия и $\log(\tilde{\eta}_i) = \tilde{x}_i^T\widehat{\theta}$ за лог-линейния (Поасонов) регресионен модел се съди по средно квадратичната грешка на предсказваните стойности (RMSEP), която се дефинира като

$$\mathrm{RMSEP}(\widehat{\theta}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - \tilde{\eta}_i)^2}.$$

RMSEP се пресмята за всяка оценка по генерираните данни на тестовата извадка, в резултат на което са представени усреднените стойности и медианите от получената редица от оценки за проведените симулации. Пресметнати са още oracle оценките, където истинските стойности на параметрите θ са използвани като мярка за надеждността на изчислителните процедури. За надеждността и качеството на методите на оценяване се съди по възможостите им да селектират правилно предикторните променливи, използвайки FPR и FNR критериите, дефинирани както следва

$$FPR(\widehat{\theta}) = \frac{|\{j \in \{1,\ldots,p\} : \widehat{\theta}_j \neq 0 \wedge \theta_j = 0\}|}{\{j \in \{1,\ldots,p\} : \theta_j = 0\}} \tag{8.17}$$

$$FNR(\widehat{\theta}) = \frac{|\{j \in \{1,\ldots,p\} : \widehat{\theta}_j = 0 \wedge \theta_j \neq 0\}|}{\{j \in \{1,\ldots,p\} : \theta_j \neq 0\}} \tag{8.18}$$

Грешният положителен процент (false positive rate, FPR) се отнася за предикторните променливи, които са селектирани според съответния метод като значими, въпреки че в симулационния експеримент, параметрите пред тези променливи са нули. За разлика от FPR, грешният негативен процент (false negative rate, FNR) дава представа за значимите предикторни променливи, оценени като незначими, т.е., коефициентите в симулациония модел са ненулеви пред предикторната променлива, но са оценени като нулеви от съответния метод.

За сравнение на симулационните резултати с тези на Fan et al. (2009) за Поасоновия регресионен модел са дадени също така и стойностите на медианата като мярка за оценяване $||\theta - \widehat{\theta}||_1 = \sum_{i=0}^{p} |\theta_j - \widehat{\theta}_j|$ и $||\theta - \widehat{\theta}||_2 = (\sum_{i=0}^{p} (\theta_j - \widehat{\theta}_j)^2)^{1/2}$, AIC - информационния критерий на Акайке и BIC - информационния критерий на Бейс.

8.4.2 Симулационен експеримент - множествена линейна регресия)

Използвана е 3$^{\text{та}}$ симулационна схема, разгледана от Alfons et al. (2012), които дефинират разредена LTS оценка с L_1 ограничение на параметите (LTS с LASSO пенализация). Техните оценки са сравнени със SIS-SCAD процедурата и нейната тримирана версия TSIS-SCAD. Разгледали сме само SIS-SCAD, понеже нейното поведение е по-надеждно от SIS-LASSO, според симулационните експерименти (без замърсяване) на Fan et. al. (2009). Ще отбележим, че Fan et al. (2009) използват следните означения Van-SIS и Van-ISIS вместо SIS-SCAD и ISIS-SCAD, съответно.

Генерирани са $n = 100$ наблюдения от p-мерно нормално разпеделение $N_p(0, \Sigma)$, $p = 1000$. Ковариационната матрица $\Sigma = (\Sigma_{ij})_{1 \le i,j \le p}$ е зададена чрез $\Sigma_{ij} = 0.5^{|i-j|}$, за да корелират предикторните променливи. Разреденият вектор от параматри $\theta = (\theta_1, \ldots, \theta_p)^T$ е със следните координати $\theta_1 = \theta_7 = 1.5$, $\theta_2 = 0.5$, $\theta_4 = \theta_{11} = 1$, и $\theta_j = 0$ за $j \in \{1, \ldots, p\} \backslash \{1, 2, 4, 7, 11\}$.

Стойностите на зависимата променлива са генерирани според модела на линейна множествена регресия $y_i = x_i^T \theta + \varepsilon_i$, където случайната грешка ε_i следва нормално разпределение с $\mu = 0$ и $\sigma = 0.5$. Приложена е схемата на замърсяване с несъгласувани наблюдения на Alfons et al. (2012), предложена от Khan et al. (2007):

1. **Без замърсяване**

2. **Вертикални аутлайери: 10% от данните на сл. грешка в модела следват $N(20, \sigma^2)$, вместо $N(0, \sigma^2)$.**

3. **Несъгласувани наблюдения в предикторите: както в 2., но 10% от замърсените данни съдържат несъгласувани наблюдения в предикторите чрез генериране на независими предиктори с $N(50, 1)$ разпределение.**

Резултатите от симулационния експеримент са дадени в Табалица 8.1. Първите два реда са взети от Таблица 3 на Alfons et al. (2012) за сравнение. *L1-LTSraw* **е резултатът от LTS процедурата с L1- пенализация, докато** *L1-LTS* **претеглена версия на оценките, според Alfons et al., 2012. Средните стойности (***mean***) и медиани (***med***), съответно на RMSEP, FPR и FNR са получени за 500 симулационни експеримента за всеки от методите, ISIS-SCAD е означен чрез** *ISIS*, **а неговата тримирана версия чрез** *TISIS-XX*, **където** *XX* **показва процента на тримиране - 10, 20, 25.**

Резултатите, основани на средните и медианите са почти еднакви в нашите симулационни експерименти. Надеждността на ISIS-SCAD за сценария без замърсяване е висока, докато RMSEP е близка до оракъл оценката. Поведението на ISIS-SCAD при наличие на несъгласувани наблюдения в зависимата и предикторни променливи е лошо, както би трябвало да се очаква, докато робастните L1-LTS и TISIS са стабилни. TISIS показва отлично представяне: RMSEP е близка до oracle оценките, а стойностите на FPR и FNR критериите са много малки. Нещо повече, резултата от различните проценти на тримиране са сходни.

8.4.2 Симулационен експеримент - Поасонова регресия регресия)

Симулационните конфигурации в тази секция са същите както на Fan et al. (2009). Следните три типа предиктори $X_1, \ldots, X_p$ и регресионни коефициенти $\theta_0, \theta_1, \ldots, \theta_p$, за $p = 1000$ и обеми на извадката $n = 200$ са генерирани:

1. $X_1, \ldots, X_p$ **са независими и еднакво разпределени** $N(0, 1)$ **сл. величини;** $\theta_0 = 5$, $\theta_1 = -0.5423$, $\theta_2 = -0.5314$, $\theta_3 = -0.5012$, $\theta_4 = -0.4850$, $\theta_5 = -0.4133$, $\theta_6 = -0.5234$, **и** $\theta_j = 0$ **за** $j > 6$;
2. $X_1, \ldots, X_p$ **са съвместно нормално, индивидуално (маргинално)** $N(0, 1)$, **с** $cor(X_i, X_4) = 1/\sqrt{2}$ **за всички** $i \neq 4$ **и** $cor(X_i, X_j) = 1/2$ **ако** i **и** j **са различни елементи на** $\{1, \ldots, p\}\backslash\{4\}$; $\theta_0 = 5$, $\theta_1 = \theta_2 = \theta_2 = 0.6$, $\theta_4 = -0.9\sqrt{2}$; **и** $\theta_j = 0$ **за** $j > 4$;
3. $X_1, \ldots, X_p$ **са съвместно нормално, индивидуално (маргинално)** $N(0, 1)$, **и с** $cor(X_i, X_5) = 0$ **за всяко** $i \neq 5$, $corr(X_i, X_4) = 1/\sqrt{2}$ **за всяко** $i \notin \{4, 5\}$, **и**

$cor(X_i, X_j) = 1/2$ **ако** i **и** j **са различни елементи на** $\{1, \ldots, p\}\backslash\{4, 5\}$**;** $\theta_0 = 5$**,** $\theta_1 = \theta_2 = \theta_2 = 0.6$**,** $\theta_4 = -0.9\sqrt{2}$**,** $\theta_5 = 0.15$**, и** $\theta_j = 0$ **за** $j > 5$**.**

Случаят с независимите предиктори е най-простата ситуация за селектиране на променливи. В тази схема, коефициентите $\theta_1, \ldots, \theta_6$ **бяха генерирани като** $\left(\frac{\log n}{\sqrt{n}} + |Z|/8\right) U$ **с** $Z \sim N(0, 1)$ **и** $U = 1$ **с вероятност 0.5 и** $U = -1$ **с вероятност 0.5, независимо от** Z**.**

Вторият и третият случаи са усложнени, поради серийната корелация. Нещо повече, въпреки че $\theta_4 \neq 0$**, изборът на другите регресионни коефициенти за случаите 2 и 3 осигурява** $\mathbf{cor}(X_4, Y) = 0$**, което прави селектирането на предикторни променливи много по-трудно. Коефициентът** $\theta_0 = 5$ **е използван за контрола като подходящ за отношението на сигнал към шум.**

Данните (x_i^T, y_i) **за** $i = 1, \ldots, 200$ **са независими повторения, където** y_i **е условно по** x_i **Поасоново** $(\mu(x_i))$ **разпределена, където** $\log(\mu(x_i)) = \theta_0 + x_i^T\theta$**. Използвали сме следната схема на замърсяване а наблюденията:**

1. **Без замърсяване**
2. **Вертикални несъгласувани наблюдения: 10% и 20% замърсяване с данни чрез замяна, съотвтено на първите 20 и 40 наблюдения** $y_i := y_i + \exp(7)$**, за** $i = 1, \ldots, 20$**, съответно 40.**
3. **Несъгласувани наблюдения в предикторите: 10% и 20% замърсяване с данни е внесено чрез модификация съответно на първите 20 и 40 реда на матрицата на предикторите както следва** $x_{ij} := -3B_j\mathrm{sign}(x_{ij})$ **за** $i = 1, \ldots, 20$**, където** $B_j = \max_{1 \leq i \leq n}(|x_{ij}|)$ **for** $j = 1, \ldots, p$**.**

Следвайки препоръките на Fan et al. (2009), пресмятанията бяха проведени с ISIS-SCAD и TISIS-SCAD с $q = \left\lfloor \frac{n}{2 \log n} \right\rfloor = 18$ **като избор, основан на асимптотични резултати. Определянето на крайната стойност на параматъра на регуляризация със SCAD пенализираща функция беше избран чрез кросвалидация, както препоръчват Fan et al. (2009). Критерия BIC е използван, за избор на регуляризиращия параметър на всяка междинна стъпка на ISIS процедурата в тези 3 случая.**

С различните методи за оценяване бяха анализирани обучаващите данни и тестващите данни с обем от $n = 200$ **наблюдения, генерирани по схемата на експериментите без замърсяване. Резултатите за различните проценти на замърсяване на TISIS-SCAD процедурата. В таблиците са дадени няколко мерки за качеството на методите за оценяване, които са пресметнати в 100 независими Монте Карло повторения.**

Таблиците съдържат медианите на тези мерки. Първите два реда дават представа за оценката на грешката $||\theta - \widehat{\theta}||_1$ и $||\theta - \widehat{\theta}||_2$, съответно оценени по обучаващата извадка. В 3^{rd} и 4^{th} редове са дадени FPR и FNR, съответно за обучаващата извадка. В редовете с номера 5-8 са дадени информационните критерии AIC (Akaike, 1974) и BIC (Schwartz, 1978), пресметнати за обучаващата и тестова (с добавено t) извадка от данни. В последните два реда са дадени RMSEP за тестовите данни (RMSEP.t) и за истинските регресионни параметри (RMSEP.o). Символът "*"в таблиците се отнася за твърде голяма стойност, по-голяма от 250000,

В двете последователни таблици са дадени резултатите от експеримента с вертикалните несъгласувани наблюдения в зависимата променлива, докато втората таблица за несъгласуваните наблюдения в предикторните променливи.

За симулациооните експерименти без замърсяване, резултати за ISIS-SCAD, дадени в Таблица 8.2-8.7 са близки до тези за Van-ISIS, дадени в Tables 5-7 на Fan et al. (2009). Резултатите за ISIS-SCAD в 1ви симулационен експеримент, в случаите на замърсяване с вертикални несъгласувани наблюдения са дадени в Таблица 8.2, докато с несъгласувани наблюдения в предикторните променливи са дадени в Таблица 8.3. Вижда се, че ISIS-SCAD оценките са ненадеждни. Аналогични са резултатите за ISIS-SCAD във 2ри и 3ти симулационен експеримент, реултатите са дадени съответно в Таблици 8.4-8.5 и Таблици 8.6-8.7 Поведението на робастната версия TISIS-SCAD е надеждно във всички симулационни схеми с и без замърсяване. Резултатите са близки до ISIS-SCAD оценките върху данните без замърсяване. Същото важи за FPR и FNR, получени по TISIS-SCAD процедурата, тъй като не превишават 1% във всички сценарии.

8.4.2 Резюме и изводи.

Въведена е робастна версия на ММП с пенализация, основана на тримиране. Характеризирана е праговата точка с техниката на d-пълнотата. В разширено симулационно изследване е изучено поведението на тези оценки за крайната извадка в средата на данни с висока размерност за линейна множествена и лог-линейна (Поасонова) регресия. Представянето на предложените оценки е удовлетворително, което се потвърждава от симулационните експерименти. Необходимите изчисления са проведени с процедурите SIS/ISIS на Fan et al. (2009). Всяка друга процедура, която провежда регуляризиращи техники, може да бъде използвана.

Библиография

Adrover, J., Maronna, R.A. and Yohai, V.J., 2004. Robust regression quantiles. J. Statist. Plann. Infer., vol. 122, pp. 187 – 202.

Akaike, H. 1974. A new look at the statistical model identification. IEEE Trans. Auto. Control, vol. 19, pp. 716 – 723.

Alfons, A., Croux, C. and Gelper, S. 2013. Sparse least trimmed squares regression. Ann. Appl. Stat., vol. 7, pp. 226 – 248.

Antoniadis, A., Gijbels, I. and Nikolova, M. 2011. Penalized likelihood regression for generalized linear models with non-quadratic penalties. Ann. Inst. Stat. Math. vol. 63, 585 – 615.

Atanasov, D. V. 1998. About the finite sample breakdown point of the WTL estimators, MSc. Thesis, Faculty of Mathematics, Sofia Univ., (in Bulgarian).

Atanasov, D. V. and Neykov, N. M. 2001. On the finite sample breakdown point of the weighted trimmed likelihood estimators and the d–fullness of a set of continous functions. In: *Proceedings of the CDAM Confrence*, 10-14 September 2001, Minsk, Belarus, Aivazian, S., Yu. Kharin and H. Reider (eds.), vol. 1, pp. 52 – 57.

Atkinson, A. C. and Riani, M. 2000. *Robust diagnostic regression analysis.* Springer, NY.

Atkinson, A. C. Riani, M. and Cerioli,A. 2004. *Exploring Multivariate Data with the Forward Search.* Springer, NY.

Barão, M. I. and Tawn, J. A. 1999. Extremal analysis of short series with outliers: Sea-levels and athletic records. Appl. Statist. vol. 48, 469 – 487.

Bednarski, T. and Clarke, B. R. 1993. Trimmed likelihood estimation of location and scale of the normal distribution. Austral. J. Statist. vol. 35, pp. 141 – 153.

Beran, R. 1982. Robust estimation in models for independent nonindentically distributed data. Ann. Statist. vol. 10, pp. 415 – 428.

Breheny, P. and Huang, J., 2011. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. Ann. Appl. Stat., vol. 5, pp. 232 – 253.

Bühlmann, P. and van der Geer, S. 2011. *Statistics for High Dimensional Data: Methods Theory and Applications*. Springer. New York.

Campbell N. A. 1984. Mixture models and atypical values. Math. Geology 16, pp. 465 – 477.

Cantoni, E. and Ronchetti, E. 2001. Robust inference for generalized linear models. J. Amer. Statist. Assoc. vol. 96, pp. 1022 – 1030.

Carroll, R. J. and Pederson, S. 1993. On robustness in the logistic regression model. J. R. Statist. Soc. B. vol. 55, pp. 693 – 706.

Chen, C. (2004). An adaptive algorithm for quantile regression. In: *Theory and Applications of Recent Robust Methods*. Hubert, M., Pison, G., Struyf, A., Van Aelst, S. (Eds.), Birkhäuser, Basel, pp. 39 – 48.

Cheng, T.-C. 2011. Robust diagnostics for the heteroscedastic regression model. Comput. Statist. and Data Anal. vol. 55, pp. 1845 – 1866.

Christmann, A. 1994. Least median of weighted squares in logistic regression with large strata. Biometrika. vol. 81, pp. 413 – 417.

Christmann, A. and Rousseeuw, P. J. 2001. Measuring overlap in logistic regression. Comput. Statist. and Data Anal.. vol. 28, pp. 65 – 75.

Choi, E. P., Hall, P. and Presnell, B. 2000. Rendering parametric procedures more robust by empirically tilting the model. Biometrika vol. 87, pp. 453 – 465.

Čížek, P. 2002. *Robust estimation in nonlinear regression and limited dependent variable models.* http://econpapers.hhs.se/paper/wpawuwpem/0203003.htm

Čížek, P. 2004. General trimmed estimation: robust approach to nonlinear and limited dependent variable models. (Discussion Paper No. 130), Tilburg University, Center for Economic Research.

Čížek, P. 2006. Least trimmed squares in nonlinear regression under dependence. J. Statist. Plann. Infer., vol. 136, pp. 3967 – 3988.

Čížek, P., 2008. Robust and Efficient Adaptive Estimation of Binary-Choice Regression Models, J. Amer. Statist. Assoc., vol. 103, pp. 685 – 698.

Čížek, P. 2008. General trimmed estimation: Robust approach to nonlinear and limited dependent variable models. Econometric Theory, vol. 24, pp. 1500 – 1529.

Čížek, P. 2010. Reweighted least trimmed squares: an alternative to one-step estimators. CentER Discussion Paper 2010/91, Tilburg University, The Netherlands.

Čížek, P. 2011. Semiparametrically weighted robust estimation of regression models. Comput. Statist. and Data Anal., vol. 55, pp. 774 – 786.

Čížek, P. 2013. Reweighted least trimmed squares: an alternative to one-step estimators. TEST, vol. 22, pp. 514 – 533.

Coles, S. G. 2001. An introduction to statistical modeling of extreme values. Springer-Verlag, London.

Copas, J. B. 1988. Binary regression models for contaminated data (with discussion). J. R. Statist. Soc. B. vol. 50, pp. 225 – 265.

Cuesta-Albertos,J.A., Matrán, C. and Mayo-Iscar, A. 2008. Robust estimation in the normal mixture model based on robust clustering. J. R. Statist. Soc. B vol. 70, pp. 779 – 802.

Davé, R., Krishnapuram, R. 1997. Robust clustering methods: a unified view. IEEE Transactions on Fuzzy Systems vol. 5, pp. 270 – 293.

Demidenko, E. Z. 1989 *Optimization and regression.*(in Russian) Nauka, Moscow.

Dimova, R. and Neykov, N.M. (2003). Generalized d-fullness Technique for Breakdown Point Study of the Trimmed Likelihood Estimator. Compt. rend. Acad. Bulg. Sci., Tome 56, 5, 7 – 12.

Dimova, R., Neykov, N. M. 2004. Generalized d-fullness technique for breakdown point study of the trimmed likelihood estimator with applications. In: *Theory and Applications of Recent Robust Methods*. Hubert, M., Pison, G., Struyf, A., Van Aelst, S. (Eds.), Birkhäuser, Basel. pp. 83 – 91.

Dimova, R. and Neykov, N.M. (2004b). Application of the d-fullness Technique for Breakdown Point Study of the Trimmed Likelihood Estimator to a generalized Logistic Model. Pliska Stud. Math. Bulgar., vol. 16, 35 – 41.

Donoho, D. L. and Huber, P. J. 1983. *The notion of breakdown point.* In: A festschrift for Eric Lehmann. P.J. Bickel, K. A. Doksum and J. L. Hodges (eds.). Belmont, CA: Wadsworth, 157 – 184.

Dupuis, D. J. and Field, C. A. 1998. Robust estimation of extremes. The Canadian J. of Statistics, vol. 26, 199 – 215.

Dupuis, D. J. and Morgenthaler, S. 2002. Robust weighted likelihood estimators with an application to bivariate extreme value problems. The Canadian J. of Statistics, vol. 30, 19 – 31.

Dupuis, D. J. and Tawn, J. A. 2001. Efects of misspecification in bivariate extreme value problems. Extremes, vol. 4, 315 – 330.

Dunn, P. 2009. Tweedie exponential family models.http://cran.R-project.org/doc/ packages/tweedie.pdf

Dutter, R., Filzmoser, P., Gather, U. and Rousseeuw, P. J. (eds.) 2003. *Developments in robust statistics*. Physica-Verlag, Heidelberg

Efron, B. 1986. Double exponential families and their use in generalized linear regression. J. Amer. Statist. Ass. vol. 81, pp. 709 – 721.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. 2004. Least angle regression. Ann. Statist., vol. 32, pp. 407 – 499.

Fan, J. and Li, R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc., vol. 96, pp. 1348 – 1360.

Fan, J. and Lv, J. 2008. Sure independence screening for ultrahigh dimensional feature space (with discussion), J. R. Statist. Soc. B, vol. 70, pp. 849 – 911.

Fan, J. and Lv, J. 2010. A selective overview of variable selection in high dimensional feature space. Statistica Sinica, vol. 20, pp. 101 – 148.

Fan, J., Samworth, R. and Wu, Y. 2009. Ultrahigh dimensional variable selection: beyond the linear model. J. Mach. Learn. Res., vol. 10, pp. 1829 – 1853.

Fan, J. and Song, R. 2010. Sure independence screening in generalized linear models with NP-dimensionality. Ann. Statist., vol. 38, pp. 3567 – 3604.

Farcomeni, A. and Greco,L. 2015. *Robust methods for data reduction*. CRC Press, New York.

Field, C. and Smith, B. (1994). Robust estimation - a weighted maximum likelihood approach. Int. Statist. Rev. vol. 62, pp. 405 – 424.

Frank, I.E. and Friedman, J.H. 1993. A statistical view of some chemometrics regression tools (with discussion). Technometrics, vol. 35, pp. 109 – 148.

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. 2007. Pathwise coordinate optimization. Ann Appl Statist., vol. 1, pp. 302 – 332.

Friedman, J., Hastie, T. and Tibshirani, R. 2010. Regularization paths for generalized linear models via coordinate descent. J. Statist. Software, vol. 33, pp. 1 – 22. URL http://www.jstatsoft.org/v33/i01/.

Fritz, H., Garc?a-Escudero, L.A., and Mayo-Iscar, A. 2013. Robust constrained fuzzy clustering. Information Sciences, vol. 245, pp. 38 – 52.

Fritz, H., Garc?a-Escudero, L.A. and Mayo-Iscar, A. 2013. A fast algorithm for robust constrained clustering. Comput. Statist. and Data Anal. vol. 61, pp. 124 – 136.

Gallegos, M. T., Ritter, G. 2005. A robust methods for cluster analysis. Ann. Statist. vol. 33, pp. 347 – 380.

Gallegos, M. T. and Ritter, G. 2010. Using combinatorial optimization in model-based trimmed clustering with cardinality constraints. Comput. Statist. and Data Anal. vol. 54, pp. 637 – 654.

Garcia-Escudero, L. A., Gordaliza, A., Matran, C. 2003. Trimming tools in exploratory data analysis. Journal of Computational and Graphical Statistics, vol. 12, pp. 434 – 449.

Garcia-Escudero, L. A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A. 2008. A general trimming approach to robust cluster analysis. Ann. Statist. vol. 36, pp. 1324 – 1345.

Garcia-Escudero, L. A., Gordaliza, A., Martin R. S., and Mayo-Iscar, A. 2010a Robust Clusterwise Linear Regression Through Trimming, submitted to Comput. Statist. and Data Analysis. vol. 54, pp. 3057 – 3069.

Garcia-Escudero, L. A., Gordaliza, A., Matran, C. and Mayo-Iscar, A. 2010b. A review of robust clustering methods. Adv. Data Anal. Classif. vol 4, pp. 89 – 109.

Garcia-Escudero, L. A., Gordaliza, A., Matr?n, C., Mayo-Iscar, A. 2011. Exploring the number of groups in robust model-based clustering. Statistics and Computing, vol. 2, pp. 585 – 599.

Garcia-Escudero, L. A., Gordaliza, A. and Mayo-Iscar, A. 2013. Comments on: model-based clustering and classification with non-normal mixture distributions. Statistical Methods and Applications, vol. 22, 459 – 461.

Garcia-Escudero, L. A., Gordaliza, A., and Mayo-Iscar, A. 2014. A constrained robust proposal for mixture modeling avoiding spurious solutions. Advances in Data Analysis and Classification, vol. 8, pp. 27 – 43.

Garcia-Escudero, L. A., Gordaliza, A., Martin R. S., and Mayo-Iscar, A. 2015. Avoiding Spurious Local Maximizers in Mixture Modeling. Stat. Computing. vol. 25, pp. 619 – 633.

Garcia-Escudlero, L. A., Gordaliza, A., Greselin, F., Ingrassia, S., and Mayo-Iscar, A. 2016. The joint role of trimming and constraints in robust estimation for mixtures of Gaussian factor analyzers. Comput. Statist. and Data Anal. vol. 99, pp. 131 – 147.

Gervini, D. and Yohai, V.J. 2002. A class of robust and fully efficient regression estimators. Ann. Statist. vol. 30, 583 – 616.

Giloni, A., Simonoff, J.S. and Sengupta, B. 2006. Robust weighted LAD regression. Comput. Statist. and Data Anal., vol. 50, pp. 3124 – 3140.

Green,P. J. 1984. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives, J. Roy. Statist. Soc. Ser. B vol. 46, pp. 149 – 192

Hadi, A. S. and Luceño, A. 1997. Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms. Comput. Statist. and Data Anal., vol. 25, pp. 251 – 272.

Hand, D. J., Daly, F., Lunn, A.D., Mc Conway, K.J. and Ostrowski, E. 1994. *A Handbook of Small Data Sets* (Chapman & Hall, London

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P.J. and Stahel, W.A. 1986. Robust statistics. The approach based on influence functions. Wiley, New York.

Hardin, J., Rocke, D. M. 2004. Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. Comput. Statist. and Data Anal., vol. 44, pp. 625 – 638.

Hawkins, D. M. and Khan, D. M. 2009. A procedure for robust fitting in nonlinear regression. Comput. Statist. and Data Anal. vol. 53, pp. 4500 – 4507.

Hawkins, D. M. and Olive, D. J. 1999. Applications and algorithms for least trimmed sum of absolute deviations regression. Comput. Statist. and Data Anal.. 32, pp. 119 – 134.

Hawkins, D. M. and Olive, D. J. 2002. Inconsistency of resampling algorithms for high-breakdown regression estimators and a new algorithm (with discussions). J. Amer. Statist. Assoc. vol. 97, pp. 136 – 159.

He, X., Jurečkova, J., Koenker, R. and Portnoy, S., 1990. Tail behavior of regression estimators and their breakdown points. Econometrics, vol. 58, pp. 1195 – 1214.

Hennig, C. 2003. Clusters, outliers, and regression: fixed point clusters. J. of Multivariate Analysis, vol. 86, pp. 183 – 212.

Herwindiati, D. E., Djauhari, M. A. and Mashuri, M. 2009. Robust multivariate outlier labeling. Communications in Statistics: Simulation and Computation. vol. 36, pp. 1287 – 1294.

Heritier, S., Cantoni, E., Copt, S. and Victoria-Feser, M.-P. 2009. *Robust methods in biostatistics*. Wiley, Chichester, U.K.

Hössjer, O. 1994. Rank-based estimates in the linear model with high breakdown point. J. Amer. Statist. Assoc. vol. 89, pp. 149 – 158.

Huber, P. 1981. *Robust statistics*. John Wiley & Sons, New York.

Hubert, M. 1997. The breakdown value of the L_1 estimator in contingency tables. Statistics and Probability Letters. vol. 33, pp. 419 – 425.

Huber, P. and Ronchetti, E. 1981. *Robust statistics*. John Wiley & Sons, New York.

Hubert, M. and Rousseeuw, P. J. 1998. The catline for deep regression. J. Multivariate Analysis, vol. 66, pp. 270 – 296.

Hubert, M., Pison, G., Struyf, A., Van Aelst, S. (Eds.) 2004. *Theory and Applications of Recent Robust Methods*. Birkhäuser, Basel.

Hubert, M., Rousseeuw, P. J. and Van Aelst, S. 2005. Multivariate Outlier Detection and Robustness. In: Handbook of Statistics, vol. 23: *Data Mining and Computation in Statistics*, C.R. Rao, E. Wegman, and J.L. Solka (eds.), Amsterdam: Elsevier North-Holland, pp. 263 – 302.

Hubert, M., Rousseeuw, P. J. and Van Aelst, S. 2008. High-breakdown robust multivariate methods. Statistical Science, vol. 23, pp. 92 – 119.

Hunter, D. and Lange, K. (2000) Quantile regression via an MM. J. of Computational and Graphical Statistics, vol. 9, pp. 60 – 77.

Jennrich, R. I. and Moore, R. H. 1975. Maximum likelihood estimation by means of nonlinear least squares. Proc. of the Statistical Computing Section of the Amer. Statist. Assoc. pp. 57 – 65.

Jørgensen, B., 1997. *The theory of dispersion models*. London: Chapman & Hall.

Jurečková, J. 2010. Finite-sample distribution of regression quantiles. Statistics and Probability Letters, vol. 80, pp. 1940 – 1946.

Khan, J. A., Van Aelst, S. and Zamar, R. H. 2007. Robust linear model selection based on least angle regression. J. Amer. Statist. Assoc., vol. 102, 1289 – 1299.

Kharin, Yu. S. 1996. *Robustness in Statistical Pattern Recognition.* Kluwer Academic Publishers, Dordrecht, London.

Koenker, R. W. 2005a. *Quantile Regression.* Cambridge University Press, Cambridge.

Koenker, R. W. 2005b. Quantile Regression in R. http://cran.R-project.org/doc/packages/quantreg/quantreg.pdf

Koenker, R. W. and Bassett, G. Jr. 1978. Regression quantiles. Econometrica, vol. 84, pp. 33 – 50.

Koenker, R. and Machado, J. 1999. Goodness of fit and related inference processes for quantile regression. J. Amer. Statist. Assoc. vol. 94, pp. 1296 – 1309.

Krivulin, N. 1992. An analysis of the least median of squares regression problem. In: Proceedings in Comput. Statist., Y. Dodge and J. Whittaker (eds.), Heidelberg, Physica-Verlag, pp. 471 – 476.

Künsch, H. R., Stefanski, L. A. and Carroll, R. J. 1989. Conditionally unbiased bounded influence estimation in general regression models, with applications to generalized linear models. J. Amer. Statist. Assoc. vol. 84, pp. 460 – 466.

Lee, Y. and Nelder, J.A. 1998. Generalized linear models for the analysis of quality-improvement experiments. Can. J. Statist. vol. 26, pp. 95 – 105

Lee, Y. and Nelder, J.A. 2000. The relationship between double-exponential families and extended quasi-likelihood families, with application to modelling Geissler's human sex ration data. Appl. Statist. vol. 49, pp. 413 – 419.

Lee, Y., Nelder, J.A. and Pawitan, Y. 2006. Generalized Linear Models with Random Effects: Unified analysis via h-likelihood. London: Chapman & Hall/CRC.

Leisch, F. 2004. FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. J. of Statist. Soft. 11, http://www.jstatsoft.org/

Marazzi, A. and Yohai, V. 2004. Adaptively truncated maximum likelihood regression with asymmetric errors. J. Statist. Plann. Inference, vol. 122, pp. 271 – 291.

Markatou, M., Basu, A. and Lindsay, B. 1997. Weighted likelihood estimating equations: The discrete case with applications to logistic regression. J. Statist. Plann. Inference. vol. 57, pp. 215 – 232.

Markatou, M. 2000. Mixture models, robustness, and the weighted likelihood methodology. Biometrics, vol. 56, pp. 483 – 486.

Maronna,R. A., Martin, R. D. and Yohai, V. J. 2006. *Robust Statistics: Theory and Methods*, John Wiley and Sons, New York.

Medasani, S., Krishnapuram, R. 1998. Robust mixture decomposition via maximization of trimmed log-likelihood with application to image database organization. In: Proceedings of the North American Fuzzy Information Society Workshop, Pensacola, August 1998, pp. 237 – 241.

McCullagh, P. and Nelder, J.A. 1989. *Generalized linear models*. London: Chapman & Hall.

McLachlan, G. J., Peel, D. 2000. *Finite mixture models*. Wiley, New York.

Mili, L. and Coakley, C. W. 1996. Robust estimation in structured linear regression. Ann. Statist. vol. 15, 2593 – 2607.

Mizera, I. and Müller, C. H. 1999. Breakdown points and variation exponents of robust M-estimators in linear models. Ann. Statist., vol. 27, 1164 – 1177.

Müller, Ch. H. 1995. Breakdown points for designed experiments. J. Statist. Plann. Inference. vol. 45, pp. 413 – 427.

Müller, Ch. H. 1997. *Robust Planning and Analysis of Experiments* (Shpringer, New York

Müller, C. H., Neykov, N. M. 2003. Breakdown points of the trimmed likelihood and related estimators in generalized linear models. J. Statist. Plann. Inference, vol. 116, pp. 503 – 519.

Nelder, J.A. and Pregibon, D. 1987. An extended quasi-likelihood function. Biometrika vol. 74, pp. 221 – 232.

Neykov, N. M. and Neytchev, P. N. 1990. A robust alternative of the maximum likelihood estimator. Short communications of COMPSTAT, Dubrovnik , pp. 99 – 100.

Neykov, N. M. (1995). *Robust methods with high breakdown point in the multivariate statistical analysis* Ph.D. Thesis, Faculty of Mathematics, Sofia University, (in Bulgarian).

Neykov, N. M. and Müller, C. H. 2003. Breakdown point and computation of trimmed likelihood estimators in generalized linear models. In: Dutter, R., Filzmoser, P., Gather, U., Rousseeuw, P.J. (Eds.), Developments in robust statistics. Physica-Verlag, Heidelberg, pp. 277 – 286.

Neykov, N. M., Filzmoser, P., Dimova, R., and Neytchev, P. N. 2004. Mixture of generalized linear models and the Trimmed Likelihood methodology. In: Antoch (Ed.), Proceedings in Computational Statistics. Physica-Verlag, pp. 1585 – 1592.

Neykov, N. M., Dimova, R. and Neytchev, P. N. 2005. Trimmed Likelihood Estimation of the Parameters of the Generalized Extreme Value Distribution: A Monte-Carlo Study. Pliska Stud. Math. Bulgar. vol. 17, 187 – 200.

Neykov, N. M., Filzmoser, P., Dimova, R. and Neytchev, P. N. 2007. Robust fitting of mixtures using the Trimmed Likelihood Estimator. Comput. Statist. and Data Anal., vol. 52, pp. 299 – 308.

Neykov, N. M., Filzmoser, P. and Neytchev, P. N. 2012. Robust joint modeling of mean and dispersion through trimming. Comput. Statist. and Data Anal. 56, 34 – 48.

Neykov, N. M., Cizek, P., Filzmoser, P. and Neytchev, P. N. 2012. The least trimmed quantile regression. Comput. Statist. and Data Anal., vol. 56, 1757 – 1770.

Neykov, N. M., Filzmoser, P. and Neytchev, P. N. 2014. Ultrahigh dimensional variable selection through the penalized maximum trimmed likelihood estimator. Stat. Papers, vol. 55, 187 – 207.

Neytchev, P. N. Neykov,N. M. and Todorov,V. K. 1994. User's manual of REGRESS PC program system for fitting models to data. TR of NIMH, Sofia

O'Hara Hines, R. J. and Carter, E. M. 1993. Improved added variable and partial residual plots for the detection of influential observations in generalized linear models. Appl. Statist. vol. 42, pp. 3 – 20.

Prentice, R. L. 1976. A generalization of the probit and logit methods for dose response curves. Biometrics vol. 32, pp. 761 – 768.

Ribatet, M. and Iooss, B. 2009. Joint modeling of mean and dispersion package. http://cran.R-project.org/ doc/ packages/ JointModeling.pdf

Ritter, G. 2010. *Robust Cluster Analysis and Variable Selection.* Chapman & Hall / CRC Press.

Rousseeuw, P. J. (1984). Least median of squares regression. J. Amer. Statist. Assoc. vol. 79, pp. 851 – 857.

Rousseeuw, P. J. 1986. Multivariate Estimation with High Breakdown Point. In: Mathematical Statistics and Applications Vol I B, W. Grossmann, G. Pflug, I. Vincze, and W. Wertz (eds.), Dordrecht: Reidel Publishing Company , pp. 283 – 297.

Rousseeuw, P. J. and Hubert, M. 1999. Regression depth. J. Amer. Statist. Assoc., vol. 94, pp. 388 – 402.

Rousseeuw, P. J. and Leroy, A. M. 1987. *Robust regression and outlier detection.* Wiley, New York.

Rousseeuw, P. J. and Van Driessen, K. 1999a. Computing least trimmed of squares regression for large data sets. Estadistica, vol. 54, pp. 163 – 190.

Rousseeuw, P. J. and Van Driessen, K. 1999b. A fast algorithm for the minimum covariance determinant estimator. Technometrics, vol. 41, pp. 212 – 223.

Ruwet, C., Garc?a-Escudero, L. A., Gordaliza, A., and Mayo-Iscar, A. 2013. On the breakdown behavior of the TCLUST clustering procedure. TEST, vol. 22, pp. 466 – 487.

Shane, K. V. and Simonoff, J. S. 2001. A Robust approach to categorical data analysis. J. Computational and Graphical Statistics, vol. 10, 135 – 157.

Schwartz, G. 1978. Estimating the dimension of a model. Ann. Statist., vol. 6, 461 – 464.

Smyth, G. K. 1989. Generalized linear models with varying dispersion. J. R. Statist. Soc. B vol. 51, pp. 47 – 60.

Smyth, G. K. 2009a. Double generalized linear models. http://cran.R-project.org/ doc/ packages/ dglm.pdf

Smyth, G. K. 2009b. Statistical Modeling. http://cran.R-project.org/ doc/ packages/ statmod.pdf

Smyth, G. K. and Verbyla, A. P. 1999. Adjusted likelihood methods for modelling dispersion in generalized linear models. Environmetrics vol. 10, pp. 696 – 709.

Smith, R. L. 1985. Maximum Likelihood estimation in a class of non regular cases. Biometrika, vol. 72, 67 – 90.

Stephenson, A. G. 2002. EVD: Extreme Value Distributions. R-News, 2, 31 – 32, URL http://CRAN.R-project.org/doc/Rnews/

Stromberg, A. J. and Ruppert, D. 1992. Breakdown in nonlinear regression. J. Amer. Statist. Assoc. vol. 87, pp. 991 – 997.

Stromberg, A. J., Hössjer, O. and Hawkins, D. M. 2000. The Least Trimmed Differences Regression Estimator and Alternatives. J. Amer. Statist. Assoc., vol. 95, pp. 853 – 864.

Tableman, M. 1994a. The asymptotics of the least trimmed absolute deviations (LTAD) estimator. Statistics and Probability Letters, vol. 19, pp. 387 – 398.

Tableman, M. 1994b. The influence functions for the least trimmed squares and the least trimmed absolute deviations estimator. Statistics and Probability Letters, vol. 19, pp. 329 – 337.

Tibshirani, R. 1996. Regression shrinkage and selection via the LASSO. J. Roy. Statist. Soc. Ser. B, vol. 58, 267 – 288.

Tibshirani, R. 1997. The lasso method for variable selection in the Cox model. Statist. Med., vol. 16, 385 – 95.

Van Aelst, S., Rousseeuw, P. J., Hubert, M. and Struyf, A. 2002. The deepest regression method. J. Multivariate Analysis, vol. 81, pp. 138 – 166.

Van der Vaart, A. W. and Wellner, J. A. 1996. *Weak Convergence and Empirical Processes With Applications to Statistics*. Springer, New York.

Vandev, D. L. 1993. A note on breakdown point of the least median squares and least trimmed squares. Statistics and Probability Letters vol. 16, pp. 117 – 119.

Vandev, D. L. and Marincheva, M. 1996. The BP of the WLT estimators in the general elliptic family of distribution. In: *Proc. of Statistical Dada Analysis*, Vandev, D.L. (ed.) Varna, pp. 25 – 31.

Vandev, D. L. and Neykov, N. M. 1993. Robust maximum likelihood in the Gaussian case, In: *New Directions in Data Analysis and Robustness*, Morgenthaler, S., Ronchetti, E. and Stahel, W. A. (eds.), Birkhäuser Verlag, Basel, pp. 259 – 264.

Vandev, D. L. and Neykov, N. M. 1998. About regression estimators with high breakdown point. Statistics, vol. 32, pp. 111 – 129.

Wang, H., Li, G. and Jiang, G. 2007. Robust regression shrinkage and consistent variable selection through the LAD-lasso. J. of Business & Economic Statist. vol. 25, 347 – 355.

Varmuza, K. and Filzmoser, P. 2008. *Introduction to multivariate statistical analysis in chemometrics*. CRC press, New York.

Windham, M. P. 1995. Robustifying model fitting. J. Roy. Statist. Soc. Ser. B vol. 57, 599 – 609.

Zhang, C. H. 2008. Discussion of One-step sparse estimates in nonconcave penalized likelihood models by H. Zou and R. Li. Ann. Statist., vol. 36, 1553 – 1560.

Zou, H. 2006. The Adaptive lasso and its oracle properties. J. Amer. Statist. Assoc., vol. 101, 1418 – 1429.

Zou, H. and Li, R. 2008. One-step sparse estimates in nonconcave penalized likelihood models (with discussion). Ann. Statist., vol. 36, 1509 – 1533.