Известия НАН Армении, Математика, том 54, н. 5, 2019, стр. 53 – 69

МЕТОД ЧЕРЕДУЮЩИХСЯ НАИМЕНЬШИХ КВАДРАТОВ ДЛЯ ОБОБЩЕННЫХ ЛИНЕЙНЫХ МОДЕЛЕЙ

А. Г. МИНАСЯН

Ереванский Государственный Университет
E-mail: *arsh.minasyan@gmail.com*

Аннотация. В статье доказывается сходимость последовательной процедуры известной как покоординатный спуск к оценке максимального правдоподобия для обобщенных линейных моделей. Покоординатный спуск для линейной регрессии известен как метод чередующихся наименьших квадратов. Оптимизационная задача в случае экспоненциального семейства остается вогнутой и свойство концентрации вокруг истинного параметра позволяет использовать разложение Тейлора до второго порядка. Численные примеры иллюстрируют доказанную сходимость с последующим обсуждением начального значения.

MSC2010 number: 62L12, 62F12, 49M05.

Ключевые слова: обобщенная линейная модель; экспоненциальное семейство; чередующая максимизация.

1. Введение

Многие статистические задачи можно рассматривать как задачи полу параметрического оценивания, когда неизвестное распределение данных описывается параметром высокой или бесконечной размерности, в то время как параметр, который нас интересует имеет низкую размерность. Типичными примерами являются функциональная оценка, оценка функции в точке или просто оценка данного подвектора вектора параметров. Классическая статистическая теория обеспечивает общее решение этой задачи: оценивать полный вектор параметра методом максимального правдоподобия и проецировать полученную оценку на целевое подпространство. Этот подход известен как профильный метод максимального правдоподобия и является эффективным при достаточно общих условиях, которые в случае обобщенных линейных моделей выполнены. Для более общего случая, например, М-оценок, эти технические условия следует вводить отдельно и проверять, выполнены ли они или нет. Мы ссылаемся на [8], [6] и [10] для подробного изложения теории и дальнейших ссылок.

В этом исследовании рассмотрена задача полупараметрической оценки профильного параметра (см. [4], [5] и ссылки внутри этих статей). Одной из таких задач, о которой стоит упомянуть, является задача выбора модели. В большинстве случаев практических задач нереалистично ожидать, что модельные предположения будут выполнены, даже если используются богатые непараметрические модели. Это означает, что истинное распределение данных $\mathbb{P}$ не принадлежит к рассматриваемому параметрическому семейству, в нашем случае – экспоненциальному семейству. Применимость общей полупараметрической теории в таких случаях сомнительна. Важной особенностью представленного подхода является то, что он в равной степени применим при неправильной спецификации модели.

Пусть $\mathcal{Y}$ это распределение, из которого идут наблюдаемые данные и статистическая модель предполагает, что неизвестное распределение данных $\mathbb{P}$ принадлежит заданному параметрическому семейству $(\mathbb{P}_v)$:

$$\mathcal{Y} \sim \mathbb{P} = \mathbb{P}_{v^*} \in (\mathbb{P}_{\mathbf{v}}, v \in \Theta), \tag{1.1}$$

где Θ некое параметрическое пространство.

Метод максимального правдоподобия в параметрической оценке позволяет оценить весь вектор параметров v путем максимизации соответствующего логарифмического правдоподобия

$$L(v) = \log \frac{d\mathbb{P}_v}{d\mu_0}$$

для некоторой доминирующей меры μ_0. Определим оценку максимального правдоподобия $\tilde{v}$ векторного параметра v следующим образом

$$\tilde{v} \stackrel{\text{def}}{=} \arg\max_{v \in \Theta} L(v). \tag{1.2}$$

Неправильная спецификация модели означает $\mathbb{P} \notin (\mathbb{P}_v, v \in \Theta)$. Другими словами, $L(v)$ есть функция квази максимального правдоподобия на Θ. Истинный параметр v^* определяется следующим образом

$$v^* \stackrel{\text{def}}{=} \arg\max_{v \in \Theta} \mathbb{E}L(v). \tag{1.3}$$

В случае допущения, что истинная модель не принадлежит нашему параметрическому семейству v^* определяет наилучшее параметрическое соответствие $\mathbb{P}$ рассматриваемым семейством. Относительно аналогичных результатов см. [1]. Сначала Кнайп начал работу в этом направлении, введя упорядоченные линейные функционалы (см. [9]). Для общих результатов чередующейся максимизации (минимизации) см. [2].

Ключевой момент данной работы заключается в том, что покоординатный спуск дает лишь небольшой выигрыш, или вовсе даже не дает, в сложности вычисления оптимальной точки для линейных моделей (см. [12]) при некоторых условиях на размерность параметров. В случае же с нелинейными моделями выигрыш ощутимый. В нелинейных моделях в большинстве случаев решения в явной форме нет, в некоторых случаях даже численные решения условий первого порядка могут быть очень сложными для реализации в полной размерности параметра. Метод, известный метод как покоординатного спуска (максимизации или минимизации) [3], помогает в таких ситуациях и эффективно оценивет вектор параметров.

Рассматриваемая модель имеет параметр v, размерность которого $p+q$, где p - размерность интересующегося нами параметра, а q - размерность остального вектора. Обычно p невелика, потому что мы также заботимся о пригодности и интерпретируемости нашей модели, но q может быть и очень большим, хотя оценивание этого параметра является второстепенной задачей, но для полноты модели мы не можем его исключить. Основные сложности с вычислениями происходят для случая больших размерностей, т. е. в случаях, когда $p+q$ достаточно велика, то обратить матрицу $(p+q)\times(p+q)$ становиться вычислительно невозможным.

Метод покоординатного спуска является частным случаем EM-алгоритма. EM-алгоритм является популярным алгоритмом, который впервые был получен в [7]. В [7] также описано, как EM-алгоритм можно реализовать в разных облостях. Мы ссылаемся на [11] за краткое введение в разработку EM-алгоритма и ограничиваемся ссылкой на известный результат сходимости [13], который по-прежнему является самым современным в большинстве случаев. К сожалению, результат описанный в [13], как и большинство результатов сходимости по этим итеративным процедурам, обеспечивает только локальную сходимость. В этой работе рассматривается один из особых случаев, когда можно доказать фактическую сходимость метода.

Остальная часть статьи имеет следующую структуру. Параграф 2 содержит предварительные сведения об обобщенных линейных моделях в классе случайных величин, называемых экспоненциальным семейством. Параграф 3 содержит

основные результаты о сходимости покоординатного спуска для обобщенных линейных моделей. Параграф 4 иллюстрирует численную работу алгоритма, что подтверждает результат теоремы из параграфа 3.

2. Введение в обобщенные линейные модели

В этом разделе мы введем класс обобщенно линейных моделей с немного другой точки зрения. Этот параграф является подготовительным для параграфа 3.

Пусть Y_i независимые случайные величины, $X_i \in \mathbb{R}^p$ и $Y_i \sim P_i \in (\mathcal{P}_v)$, что означает $\exists v_i : P_i = P_{v_i}$, где $(\mathcal{P}_v)$ предполагается экспоненциальным семейством распределений с каноническим параметром. Экспоненциальное семейство будет обсуждено далее в этом параграфе. Обобщенные линейные модели могут быть записаны следующим образом $Y_i \sim P_{v(X_i)}$. В случае гауссовского распределения мы получаем $Y_i = v(X_i) + \varepsilon_i$ с произвольной функцией $v(\cdot)$.

Функция $v(x)$ может быть записана как

$$v(x) = \sum_{j=1}^{\infty} \theta_j \psi_j(x).$$

Тогда, линейное параметрическое предположение дает

$$v(x) = \sum_{j=1}^{p+q} \theta_j \psi_j(x) = \sum_{j=1}^{p} \theta_j \psi_j(x) + \sum_{j=p+1}^{p+q} \theta_j \psi_j(x) \tag{2.1}$$

для заданного базиса $\psi_j(\cdot)$. Обозначим $\eta_i = \theta_{p+i}$ и вектор столбец $\eta = (\eta_1, \ldots, \eta_q)^T \in \mathbb{R}^q$.

$$Y_i \sim P_{v_i}, \;\; v_i = \Psi_i^T \theta + \Phi_i^T \eta, \tag{2.2}$$

где $\Psi_i = (\psi_1(x), \ldots, \psi_p(x))^T \in \mathbb{R}^p$, $\Phi_i = (\psi_{p+1}(x), \ldots, \psi_{p+q}(x))^T \in \mathbb{R}^q$ и $\theta \in \mathbb{R}^p$, $\eta \in \mathbb{R}^q$.

Логарифм правдоподобия в данном случае равен

$$\log \frac{dP_\theta}{d\mu_0^n}(\mathcal{Y}) = \sum_{i=1}^{n} (v_i Y_i - g(\nu_i)) = \sum_{i=1}^{n} \left(\Psi_i^T \theta Y_i + \Phi_i^T \eta Y_i - g(\Psi_i^T \theta + \Phi_i^T \eta) \right),$$

что следует из эквивалентности семейства (2.10), а функция $g(\cdot)$ может быть выведена из (2.8). Эквивалентно, имеем

$$L(\theta, \eta) \stackrel{\text{def}}{=} S^T \theta + R^T \eta - A(\theta, \eta), \tag{2.3}$$

где

$$S \stackrel{\text{def}}{=} \sum_{i=1}^{n} Y_i \Psi_i \in \mathbb{R}^p, \;\; R \stackrel{\text{def}}{=} \sum_{i=1}^{n} Y_i \Phi_i \in \mathbb{R}^q, \;\; A(\theta, \eta) \stackrel{\text{def}}{=} \sum_{i=1}^{n} g(\Psi_i^T \theta + \Phi_i^T \eta).$$

Обозначим логарифм правдоподобия через $L(\theta, \eta)$ или $L(v)$, где $v \stackrel{\text{def}}{=} (\theta, \eta)$.

Тогда, в терминах v получаем

$$L(v) = \Upsilon^T v - A(v),$$

где $\Upsilon = \begin{pmatrix} S \\ R \end{pmatrix}^T \in \mathbb{R}^{p+q}$. Матрица информации Фишера определяется как $\mathcal{D}^2 \stackrel{\text{def}}{=} -\nabla^2 \mathbb{E} L(v^*) = F(v^*)$, где $v = \begin{pmatrix} \theta \\ \eta \end{pmatrix}^T \in \mathbb{R}^{p+q}$, а ∇ это оператор дифференцирования. Матрица Гессе в точке v записывается следующим образом

$$\mathbb{F}(v) = -\nabla^2 \mathbb{E} L(v) = \begin{pmatrix} \mathbb{F}_{\theta\theta}(v) & \mathbb{F}_{\theta\eta}(v) \\ \mathbb{F}_{\eta\theta}(v) & \mathbb{F}_{\eta\eta}(v) \end{pmatrix}.$$

Далее мы докажем, что функция $g(\cdot)$ выпукла, откуда будет следовать положительная определенность матрицы информации Фишера $\mathbb{F}$.

Определим вектор $\nabla \stackrel{\text{def}}{=} \nabla L(v^*)$, а так же стандартизированную версию этого вектора $\breve{\xi}$ следующим образом

$$\breve{\xi} = \mathcal{D}^{-1} \nabla.$$

Параметры $\tilde{v} = (\tilde{\theta}, \tilde{\eta})$ зависят от данных, следовательно являются случайными, в то время как $v^* = (\theta^*, \eta^*)$ истинное значение параметра, которе не является случайной величиной. В реальности истинное распределение Y неизвестно, но мы делаем параметрическое предположение на класс распределений.

Запишем определения (1.2) и (1.3) таким образом

$$\tilde{v} = (\tilde{\theta}, \tilde{\eta}) = \arg\max_{\theta, \eta} L(\theta, \eta), \;\; v^* = (\theta^*, \eta^*) = \arg\max_{\theta, \eta} \mathbb{E} L(\theta, \eta). \tag{2.4}$$

Из определения v^* следует, что $\nabla \mathbb{E} L(v^*) = 0$ откуда вытекает

$$\mathbb{E} \begin{pmatrix} S \\ R \end{pmatrix}^T = \nabla A(\theta^*, \eta^*)$$

или

$$\mathbb{E} \Upsilon = \nabla A(v^*).$$

Важное свойство экспоненциального семейства заключается в том, что стохастическая компонента $\zeta(\theta, \eta)$ логарифма правдоподобия линейна по θ и η. Пусть

$\varepsilon_i = Y_i - \mathbb{E}Y_i$ и $\zeta = L - \mathbb{E}L$ тогда

$$\zeta(\theta,\eta) = (S^T - \mathbb{E}S^T)\theta + (R^T - \mathbb{E}R^T)\eta = \sum_{i=1}^{n} \varepsilon_i(\Psi_i^T\theta + \Phi_i^T\eta),$$

$$\nabla\zeta(\theta,\eta) = \begin{pmatrix} S - \mathbb{E}S & R - \mathbb{E}R \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} \varepsilon_i\Psi_i & \sum_{i=1}^{n} \varepsilon_i\Phi_i \end{pmatrix}.$$

Теперь рассмотрим следующее эллиптическое множество

$$\Omega_\circ(r) \stackrel{\text{def}}{=} \{v : \|\mathcal{D}(v - v^*)\| \le r\}. \tag{2.5}$$

Множество $\Omega_\circ(r)$ называется локальной окрестностью v^* для $\mathcal{D}^2 = -\nabla^2\mathbb{E}L(v^*) = \mathbb{F}(v^*)$ и $\mathcal{V}^2 \stackrel{\text{def}}{=} \mathbb{C}ov(\nabla L(v^*))$.

Запишем ковариационную матрицу в блочной форме

$$\mathcal{V}^2 = \begin{pmatrix} V^2 & E \\ E^T & Q^2 \end{pmatrix}. \tag{2.6}$$

Матрица информации Фишера $\mathbb{F}(v^*) = -\nabla^2\mathbb{E}L(v^*)$ в блочной форме

$$\mathbb{F}(v) = \begin{pmatrix} \mathbb{F}_{\theta\theta}(v) & \mathbb{F}_{\theta\eta}(v) \\ \mathbb{F}_{\eta\theta}(v) & \mathbb{F}_{\eta\eta}(v) \end{pmatrix}. \tag{2.7}$$

Для центральной точки v^* разложение в блочной форме

$$\mathcal{D}^2 = \mathbb{F}(v^*) = \begin{pmatrix} D^2 & A \\ A^T & H^2 \end{pmatrix},$$

где $D^2 = \mathbb{F}_{\theta\theta}(v^*)$, $A = \mathbb{F}_{\theta\eta}(v^*)$ и $H^2 = \mathbb{F}_{\eta\eta}(v^*)$. Разложим так же вектор $\nabla \stackrel{\text{def}}{=} \nabla L(v^*)$ следующим образом

$$\nabla = \begin{pmatrix} \nabla_\theta \\ \nabla_\eta \end{pmatrix}.$$

2.1. **Экспоненциальное семейство с каноническим параметром.** В этой части мы формально определяем экспоненциальное семейство распределений. Стоит отметить, что экспоненциальное семейство представляет собой довольно широкий класс распределений. Этот класс содержит такие распределения как нормальное, биномиальное, пуассоновское, гамма, мультиномиальное и другие. Простейшими примерами распределений, не принадлежащими к экспоненциальному семейству, являются распределения Стьюдента и равномерное. Красота и удобство экспоненциального семейства состоит в том, что логарифмическая функция правдоподобия имеет простой вид и может быть записана явно. В общем случае мы говорим, что случайная величина с функцией плотности вероятности $f(\cdot)$ принадлежит экспоненциальному классу, если функция плотности

вероятности может быть выражена следующим образом

$$f(x|\eta) = h(x)\exp\{\eta^T T(x) - A(\eta)\} \tag{2.8}$$

где η вектор параметров, $T(X)$ *достаточная статистика*, а $A(\cdot)$ *функция ссылки (link function).*

Как уже упомянулось выше, существует огромное число известных распределений, функции плотности вероятности которых могут быть выражены в виде (2.8). Чтобы убедиться в этом можно выразить функцию плотности нормального распределения в форме (2.8) или функции распределения Бернулли, Пуассона задав заранее соответсвующие области определения параметра.

2.1.1. *Обобщенные линейные модели.* Пусть Y является зависимым вектором от двух множеств независимых переменных Ψ и Φ. Рассмотрим модель

$$Y \sim P \in (\mathcal{P}_v)_{v\in\mathbb{R}} \ll \mu_0, \tag{2.9}$$

где $(\mathcal{P}_v)_{v\in\mathbb{R}}$ экспоненциальное семейство распределений с каноническим параметром, а μ_0 некоторая доминирующая мера. Итак,

$$\log\frac{dP_v}{d\mu_0}(y) = y\cdot v - g(v) \text{ для некоторой функции } g(\cdot). \tag{2.10}$$

Лемма 2.1. *Пусть $(\mathcal{P}_\nu)$ экпоненциальное семейство рапределений. Тогда,*

$$\mathbb{E}_\nu Y = g'(\nu),\ \ \mathbb{V}ar_\nu(Y) = \mathbb{E}_\nu[Y - g'(\nu)]^2 = g''(\nu) \tag{2.11}$$

Следовательно, функция $g(\cdot)$ выпукла вверх.

Замечание 2.1. Для простаты доказательство приведено для одномерной функции $g(\cdot)$, но данный результат верен и в случае многомерных функций и переменных. Одним из различий это то, что вместо дисперсии будет ковариационная матрица, которая в свою очередь является положительно определенной.

2.2. **Концентрация меры.** Напомним, что из свойств обобщенных линейных моделей следует, что стохастическая компонента логарифмической функции правдоподобия $\zeta(\theta,\eta)$ является линейной по θ и η, а детерминированя часть $\mathbb{E}L(v)$ — вогнутой по v.

Рассмотрим эллиптическое множество определенным в (2.5). В [12] доказано, что существует такое эллиптическое множество $\Omega_\circ(\mathrm{r})$ вокруг v такое что $\tilde{v}$ принадлежит этому множеству с большой вероятностью. Далее мы предполагаем, что x фиксированная и достаточно большая величина, чем и определяется

уровень доминирующей вероятности. Мы называем случайное множество $\Omega_0(x)$ доминирующей вероятностью, если

$$\mathbb{P}\left(\Omega_0(x)\right) \geq 1 - \mathcal{C}e^{-x}.$$

Значение x может зависеть от n и стремиться к бесконечности с ростом n. Возможные значения x это $x \asymp n^{1/2}$ и $x \asymp \log n$, которые дают $\mathbb{P}\left(\Omega_0(x)\right) \geq 1 - \mathcal{C}/n$. Единственное требование к последовательности $\{x_n\}$ это, чтобы она не росла слишком быстро, формально, $x \leq x_c \overset{\text{def}}{\asymp} n^{1/2}$.

Все результаты, полученные ниже, считаются верными на случайном множестве $\Omega_\circ(x)$ и, поскольку это множество доминирующей вероятности, тогда все результаты верны с большой вероятностью. Мы всегда помним об этом факте, но для удобства и простоты обозначений мы исключаем его из формулировки теорем.

2.3. **Локальная квадратичная аппроксимация функции логарифмического правдоподобия.** Напомним, что функция $L(\theta, \eta)$ может быть переписана в терминах v как $L(v)$. В этой части мы покажем, что аппроксимация функции $L(v)$ до второго порядка с помощью разложения Тейлора является корректной в окрестности v^*. Положим $L(v_1, v_2) = L(v_1) - L(v_2)$ и напомним, что $\tilde{v}$ случайная оценка зависищая от данных. Формально,

$$\tilde{v} = \arg\max_v L(v), \quad v^* = \arg\max_v \mathbb{E}L(v).$$

Далее имеем,

$$L(v, v^*) = \nabla L(v^*)(v - v^*) - \frac{1}{2}\|\mathcal{D}^2(v - v^*)\|^2 + \alpha'(v, v^*), \tag{2.12}$$

где $\alpha'(\cdot)$ определена в (2.12).

Аналогичным образом, мы аппроксимируем функцию $L(v)$ в окрестности $\tilde{v}$ используя тот факт, что $\nabla L(\tilde{v}) = 0$, следовательно

$$L(v, \tilde{v}) = -\frac{1}{2}\|\mathcal{D}^2(v - \tilde{v})\|^2 + \alpha(v, \tilde{v}). \tag{2.13}$$

Замечание 2.2. Свойство концентрации позволяет использовать разложение Тейлора до второго порядка, которая хорошо аппроксимирует изначальную функцию $L(\cdot)$. Идея заключается в том, что вогнутая функция в окрестности максимума имеет квадратичную форму, т.е. разложение Тейлора до второго порядка не влечет большие ошибки аппроксимации.

3. Метод чередующихся наименьших квадратов для обобщенных линейных моделей

Этот параграф обобщает обычный метод наименьших квадратов для линейных моделей (см. [12]) с помощью покоординатного спуска и доказывает аналогичный результат в случае обобщенных линейных моделей. Задача нетривиальна, поскольку оценки для большинства моделей нельзя получить в явной форме. Вместо этого, мы аппроксимируем функцию правдоподобия с помощью квадратичной функции на случайном множестве, где наблюдается концентрация меры.

Обобщенные линейные модели часто используются во многих моделях и имеют ряд приложений в самых разных областях. Например, в категориальном анализе данных, задачах классификации, Пуассоновской регресии, и т.д. В теории статистического обучения и оценивания плотности вероятности в первую очередь рассматриваются именно обобщенные линейные модели. Линейные модели порой слишком простые, чтобы описать всю модель, поэтому во многих случаях целесообразно использовать обобщенные линейные модели.

Далее мы обсудим покоординатный спуск для функции квази-правдоподобия, полученной в предыдущей главе. В общем случае процедура чередования максимизации (минимизации) используется в тех случаях, когда прямые вычисления полной размерности невозможны или очень трудно реализуемы.

Пусть $L(v)$ функция правдоподобия, где вектор $v = (\theta, \eta)$ может быть разложен как *целевой* параметр θ и параметр η, который нас не интересует. Метод чередующейся максимизации это итеративный алгоритм начинающийся с какого-то начального значения $v^\circ \in \mathbb{R}^{p+q}$ и правилом обновления как показано ниже

$$\begin{aligned} \tilde{v}_{k,k} &\stackrel{\text{def}}{=} (\hat{\theta}_k, \hat{\eta}_k) = \left(\hat{\theta}_k, \operatorname*{argmax}_{\eta \in \mathbb{R}^q} \mathcal{L}(\hat{\theta}_k, \eta) \right), \\ \tilde{v}_{k+1,k} &\stackrel{\text{def}}{=} (\hat{\theta}_{k+1}, \hat{\eta}_k) = \left(\operatorname*{argmax}_{\theta \in \mathbb{R}^p} \mathcal{L}(\theta, \hat{\eta}_k), \hat{\eta}_k \right). \end{aligned} \tag{3.1}$$

В этом разделе мы постараемся ответить на некоторые естественные вопросы, возникающие с описанной выше итерационной процедурой: Сходиться ли последовательность $\left(\hat{\theta}_k\right)$? Какова скорость сходимости? При каких условиях эта последовательность сходиться к оценке максимального правдоподобия $\tilde{v}$?

3.1. **Сходимость к оценке максимального правдоподобия.** Одним из основных результатов является следующая теорема про сходимость предложенной

процедуры к оценке максимального правдоподобия для обобщенных линейных моделей.

Теорема 3.1. *Пусть модель задана как в* (2.2) *и положим* $v = (\theta, \eta) \in \mathbb{R}^{p+q}$. $L(v)$ *определена в* (2.3) *и* $\mathcal{D}^2 = -\nabla^2 \mathbb{E} L(v^*)$ *матрица Гессе для логарифмической функции правдоподобия* $L(v)$ *в блочно-матричной форме* $\begin{pmatrix} D^2 & A \\ A^T & H^2 \end{pmatrix}$ *в точке* $\tilde{v}$. *Предположим, что* $\rho \stackrel{\text{def}}{=} \|D^{-1}AH^{-1}\|_{op}^2 < 1$ *и условие* $\|\mathcal{D}^{-1}\nabla^2 \mathbb{E} L(v)\mathcal{D}^{-1} - I_{p+q}\| \le \delta(r) \le \delta$ *выполнено, где* I_{p+q} *– единичная матрица.*

Тогда, последовательность оценок полученных методом чередующейся максимизации сходиться к $\tilde{\theta} = \Pi_\theta \tilde{v} \stackrel{\text{def}}{=} \Pi_\theta \arg\max_v L(v)$, *а* Π_θ *проектор вектора на свой подвектор* θ.

Замечание 3.1. Обозначение $\tilde{v}_{k(+1),k}$ используется в случаях, когда результат верен для $\tilde{v}_{k,k}$ и $\tilde{v}_{k+1,k}$.

Доказательство. Запишем (2.13) в терминах θ и η

$$L(v, \tilde{v}) =$$

$$-\frac{1}{2}\|D^2(\theta - \tilde{\theta})\|^2 - \frac{1}{2}\|H^2(\eta - \tilde{\eta})\|^2 - (\theta - \tilde{\theta})^T A(\eta - \tilde{\eta}) + \alpha(v, \tilde{v}). \tag{3.2}$$

Пусть θ° есть начальное значение и используя метод опианный в (3.1) получаем

$$\hat{\eta}_0 = \tilde{\eta}(\theta^\circ) =$$

$$\operatorname*{argmin}_\eta \left[\frac{1}{2}\|D^2(\theta^\circ - \tilde{\theta})\|^2 + \frac{1}{2}\|H^2(\eta - \tilde{\eta})\|^2 + (\theta^\circ - \tilde{\theta})^T A(\eta - \tilde{\eta}) + \alpha(v, \tilde{v})\right].$$

Тогда, условие первого порядка дает нам следующее соотношение

$$H^2(\hat{\eta}_0 - \tilde{\eta}) = A^T(\tilde{\theta} - \theta^\circ) + \nabla_\eta \alpha(\tilde{v}_{0,0}, \tilde{v}).$$

Аналогичным образом, решение $\hat{\theta}_1 \stackrel{\text{def}}{=} \tilde{\theta}(\hat{\eta}_0)$ имеет следующую форму

$$D^2(\hat{\theta}_1 - \tilde{\theta}) = A(\tilde{\eta} - \hat{\eta}_0) + \nabla_\theta \alpha(\tilde{v}_{1,0}, \tilde{v}).$$

Тогда итерационный процесс чередующейся максимизации дает нам следующую рекурсивную систему уравнений, зависящую от начального значения:

$$\begin{cases} H^2(\hat{\eta}_k - \tilde{\eta}) = A^T(\tilde{\theta} - \hat{\theta}_k) + \nabla_\eta \alpha(\tilde{v}_{k,k}, \tilde{v}) \\ D^2(\hat{\theta}_{k+1} - \tilde{\theta}) = A(\tilde{\eta} - \hat{\eta}_k) + \nabla_\theta \alpha(\tilde{v}_{k+1,k}, \tilde{v}). \end{cases}$$

или

$$\begin{cases} H^2(\hat{\eta}_k - \tilde{\eta}) = A^T(\tilde{\theta} - \hat{\theta}_k) + \nabla_\eta \alpha(\tilde{v}_{k,k}, \tilde{v}) \\ D(\hat{\theta}_{k+1} - \tilde{\theta}) = D^{-1}A(\tilde{\eta} - \hat{\eta}_k) + D^{-1}\nabla_\theta \alpha(\tilde{v}_{k+1,k}, \tilde{v}). \end{cases} \tag{3.3}$$

Далее мы выражаем $(\tilde{\eta} - \hat{\eta}_k)$ использую второе уравнение (3.3) и подставляем в первое уравнение. В итоге, получим

$$D(\hat{\theta}_{k+1} - \tilde{\theta}) = D^{-1}AH^{-2}A^T(D^{-1}D)(\hat{\theta}_k - \tilde{\theta}) + D^{-1}\left[\nabla_\theta\alpha(v,\tilde{v}) - AH^{-2}\nabla_\eta\alpha(v,\tilde{v})\right].$$

Теперь определим $M_\circ \stackrel{\text{def}}{=} D^{-1}AH^{-2}A^TD^{-1}$ и

$$\Xi(\tilde{v}_{k(+1),k}) \stackrel{\text{def}}{=} D^{-1}\left[\nabla_\theta\alpha(\tilde{v}_{k+1,k},\tilde{v}) - AH^{-2}\nabla_\eta\alpha(\tilde{v}_{k,k},\tilde{v})\right]$$

дает следующую рекурсивную формулу

$$D(\hat{\theta}_{k+1} - \tilde{\theta}) = M_\circ \cdot D(\hat{\theta}_k - \tilde{\theta}) + \Xi(\tilde{v}_{k(+1),k}). \tag{3.4}$$

Следовательно, суммируя для всех k, начиная с начального значения, взяв норму и используя неравенство треугольника, получим следующий результат

$$\|D(\hat{\theta}_{k+1} - \tilde{\theta})\| \le \|M_\circ\| \cdot \|D(\hat{\theta}_k - \tilde{\theta})\| + \|\Xi(\tilde{v}_{k(+1),k})\| \le \|M_\circ\|^k \cdot \|D(\theta^\circ - \tilde{\theta})\| + \sum_{\ell=0}^{k-1}\|M_\circ\|^\ell \cdot \|\Xi(\tilde{v}_{k(+1),k})\| = \|M_\circ\|^k \cdot \|D(\theta^\circ - \tilde{\theta})\| + \|\Xi(\tilde{v}_{k(+1),k})\| \cdot \frac{1 - \|M_\circ\|^k}{1 - \|M_\circ\|}.$$

Используя предположение $\rho \stackrel{\text{def}}{=} \|D^{-1}AH^{-1}\| < 1$ легко видеть, что первый член стремиться к нулю, когда k стремиться к бесконечности.

Далее нужно показать, что $\Xi(\tilde{v}_{k(+1),k})$ уменьшается с ростом k.

Для $D^{-1}\nabla_\theta\alpha(\tilde{v}_{k+1,k},\tilde{v})$ Теорема С.1 в [2] дает нужную верхнюю оценку, которая стремиться к 0 с ростом k. Заметим, что используя эту теорему можно построить верхние оценки для оставшихся членов $\Xi(\tilde{v}_{k(+1),k})$, которые, в свою очередь, дают верхную оценку для $\Xi(\tilde{v}_{k(+1),k})$ целиком.

Следовательно, используя Теорему С.1 из [2] и неравенство треугольника получим нужную оценку для $\Xi(\tilde{v}_{k(+1),k})$. Итого, получается

$$\|\Xi(\tilde{v}_{k(+1),k})\| \le \|D^{-1}\nabla_\theta\alpha(\tilde{v}_{k+1,k},\tilde{v})\| + \|D^{-1}AH^{-2}\nabla_\eta\alpha(\tilde{v}_{k,k},\tilde{v})\| \to 0 \text{ когда } k \to \infty.$$

Объединив все полученные свойства, получим, что

$$\|D(\hat{\theta}_{k+1} - \tilde{\theta})\| \le s_k \to 0 \text{ as } k \to \infty. \tag{3.5}$$

Используя те же выкладки ясно, как получить свойство сходимости для параметра η, что и завершает доказательство теоремы. □

Замечание 3.2. Последовательность s_k из (3.5) может быть интерпретирована как радуис эллиптического множества вокруг $\tilde{\theta}$, куда оценка $\hat{\theta}_{k+1}$ попадает с большой вероятностью.

Замечание 3.3. Результат Теоремы 3.1 говорит о том, что единственным условием для сходимости последовательности $\left(\hat{\theta}_k\right)$ является $\rho \stackrel{\text{def}}{=} \|M_\circ\| < 1$ и $\|\mathcal{D}^{-1}\nabla^2\mathbb{E}L(v)\mathcal{D}^{-1} - I_{p+q}\| \leq \delta$. Более того, наблюдается линейная сходимость к оценке максимального правдоподобия $\tilde{\theta}$, которую во многих случаях вычислительно трудно посчитать.

3.2. **Чередующаяся оценка.** Выше мы показали, что последовательность оценок, полученных с использованием метода чередующихся наименьших квадратов сходится к соответствующей оценке максимального правдоподобия. Это сильный и принципиально важный для практики результат. Как было сказано выше, есть две основные проблемы, которые делают проблему нетривиальной. Первая из них – невозможность получить явное решение в большинстве случаев, а вторая – большая размерность не интересующегося нас параметра, что делает невозможным непосредственное применение известного метода Ньютона-Рафсона. Альтернативный метод максимизации преодолел эти проблемы, а недопустимые ранее оценки поменялись на оценки "приближенные"к оценке максимального правдоподобия.

Далее в этом разделе мы покажем, что чередующаяся оценка близка к истинной оценке (θ^*, η^*). Напомним, что

$$v^* = (\theta^*, \eta^*) \stackrel{\text{def}}{=} \arg\max_v \mathbb{E}L(v).$$

Следующая теорема известна как разложение Фишера и мы формулируем ее в рамках обобщенных линейных моделей. Вспоминая определения, приведенные в предыдущем параграфе, теперь мы готовы сформулировать и доказать разложение Фишера.

Теорема 3.2. *Пусть выполнены условия Теоремы* (3.1) *и* $\breve{\xi} \stackrel{\text{def}}{=} D^{-1}\breve{\nabla}$, *где* $\breve{\nabla} = \nabla_\theta - AH^{-2}\nabla_\eta$.

Тогда

$$\|D(\tilde{\theta}_k - \theta^*) - \breve{\xi}\| \to 0, \ \textit{когда } k \to \infty.$$

Доказательство. Основано на идеях доказательства Теоремы (3.1). Здесь мы разлагаем логарифмическую функцию правдоподобия вокруг v^*.

Сначала мы используем условия первого порядка и получаем

$$D(\tilde{\theta}_k - \theta^*) = D^{-1}\nabla_\theta L(v^*) - D^{-1}A(\tilde{\eta}_k - \eta^*) + D^{-1}\nabla_\theta\alpha(\tilde{v}_{k,k}, v^*)$$

$$H(\tilde{\eta}_k - \eta^*) = H^{-1}\nabla_\eta L(v^*) - H^{-1}A^T(\tilde{\theta}_{k-1} - \theta^*) + H^{-1}\nabla_\eta\alpha(\tilde{v}_{k-1,k}, v^*) \quad (3.6)$$

Основываясь на [2] мы можем ограничить $D^{-1}\nabla_\theta\alpha(\tilde{v}_{k,k}, v^*)$ и$H^{-1}\nabla_\eta\alpha(\tilde{v}_{k-1,k}, v^*)$. Для системы уравнений (3.6) верно

$$D(\tilde{\theta}_k - \theta^*) = D^{-1}\nabla_\theta L(v^*) - D^{-1}[AH^{-2}\nabla_\eta L(v^*) - AH^{-2}A^T(\tilde{\theta}_{k-1} - \theta^*) + AH^{-2}\nabla_\eta\alpha(\tilde{v}_{k-1,k}, v^*)] + D^{-1}\nabla_\theta\alpha(\tilde{v}_{k,k}, v^*),$$

отсюда следует

$$D(\tilde{\theta}_k - \theta^*) = M_\circ D(\tilde{\theta}_{k-1} - \theta^*) + D^{-1}\left[\nabla_\theta L(v^*) - AH^{-2}\nabla_\eta L(v^*)\right] + D^{-1}\left\{\nabla_\theta\alpha(\tilde{v}_{k,k}, v^*) - AH^{-2}\nabla_\eta\alpha(\tilde{v}_{k-1,k}, v^*)\right\}. \tag{3.7}$$

Заметим, что используя определение $\breve{\xi}$, (3.7) может быть переписан следующим образом

$$D(\tilde{\theta}_k - \theta^*) - \breve{\xi} = M_\circ D(\tilde{\theta}_{k-1} - \theta^*) + D^{-1}\left\{\nabla_\theta\alpha(\tilde{v}_{k,k}, v^*) - AH^{-2}\nabla_\eta\alpha(\tilde{v}_{k-1,k}, v^*)\right\}$$

Далее суммируя по всем k начиная с начального значения, после того, как взяли нормы от обеих сторон и использую предположение $\rho = \|M_\circ\| < 1$, получаем

$$\|D(\tilde{\theta}_k - \theta^*) - \breve{\xi}\| \leq \mathcal{C}(\rho)\cdot D^{-1}\left\{\nabla_\theta\alpha(\tilde{v}_{k,k}, v^*) - AH^{-2}\nabla_\eta\alpha(\tilde{v}_{k-1,k}, v^*)\right\}, \tag{3.8}$$

где $\mathcal{C}(\rho)$ это константа зависищая только от ρ. Остальные члены можно оценить используя Теоремы C.1 из [2].

Напомним, что для того чтобы получить верхнюю границу для $D^{-1}\nabla_\theta\alpha(\tilde{v}_{k,k}, v^*)$ нужно условие $\|\mathcal{D}^{-1}\nabla^2\mathbb{E}L(v)\mathcal{D}^{-1} - I_{p+q}\| \leq \delta$ для некоторой константы $\delta > 0$.

Наконец,

$$\|D(\tilde{\theta}_k - \theta^*) - \breve{\xi}\| \leq \breve{s}_k \to 0, \text{ когда } k \to \infty. \tag{3.9}$$

□

Отметим, что математическое ожидание случайной величины $\breve{\xi}$ равна нулю, более того, $\mathbb{V}ar(\breve{\xi}) = D^{-1}V^2D^{-1}$.

Замечание 3.4. Стоит так же отметить, что разложение Фишера остается верным даже в случае конечного k. Тогда, соответсвующая норма ограничена не нулем, а последовательностью $\breve{s}_k$ стремящемуся к нулю.

Замечание 3.5. Случайный вектор $\breve{\xi} \in \mathbb{R}^p$ в случае правильной спецификации модели имеет нормальное распределение, следовательно $\|\breve{\xi}\|^2$ имеет хи-квадрат χ^2_p распределение с p степенями свободы. В случае же неверной спецификации модели распределение $\|\breve{\xi}\|^2$ неизвестно в конечномерном случае, но ассимптотически оно имеет хи-квадрат распределение с p степенями свободы.

3.3. **Выбор начального значения.** Начальное значение может сыграть решающую роль в сближении чередующегося метода, и если мы "преуспеем"с ним, тогда выигрыш будет двояким. Первое – условия Теоремы (3.1) могут быть ослаблены и второе – число итераций для сходимости может быть сильно снижено по сравнению с "плохой"начальной точкой. Хорошие начальные значения θ° относятся к первому условию Теоремы (3.1), т.е. $\rho \stackrel{\text{def}}{=} \|D^{-1}AH^{-1}\|^2 < 1$.

Обозначим через $\mathcal{V}$ векторное пространство собственных векторов матрицы $M_\circ$ соответсвующие собственным значениям которые больше 1. Формально, $\mathcal{V} \stackrel{\text{def}}{=} \operatorname{span}(v_1, v_2, \ldots, v_l)$, где

$$M_\circ v_i = \lambda_i v_i, \forall i \in \{1, \ldots, l\} \quad \text{и} \ |\lambda_i| \geq 1.$$

Лемма 3.3. *Если θ° выбран так, что*

$$u \perp \mathcal{V}, \tag{3.10}$$

где $u \stackrel{\text{def}}{=} D(\theta^\circ - \tilde{\theta})$, тогда имеет место сходимость.

Доказательство этой леммы основано на простом линейной алгебры, тем не менее, кратко объясним идею. Заметим, что предположение $\rho \stackrel{\text{def}}{=} \|D^{-1}AH^{-1}\|^2 < 1$ означает, что все собственные значения находятся внутри единичного круга, так что процесс будет сходиться. Это верно независимо от первоначального значения. Тем не менее, мы можем ослабить этот результат "хорошим"выбором начальных значений. Если матрица $M_\circ$ имеет собственные значения, находящиеся вне единичного круга, то начальное предположение могло бы помочь обратить их в нуль, будучи ортогональным пространству соответствующих собственных векторов.

Стоит также отметить, что условие Теоремы 3.1

$$\|\mathcal{D}^{-1}\nabla^2\mathbb{E}L(v)\mathcal{D}^{-1} - I_{p+q}\| \leq \delta(r) \leq \delta$$

никак не относиться к выбору начальной точки и нужно, чтобы ограничить член $\Xi(\tilde{v}_{k(+1),k})$. Следовательно, это условие не может быть ослаблено в зависимости от начальной точки θ°.

4. Численный пример

В этом параграфе мы приведем численный пример и проиллюстрируем его сходимость согласно теореме 3.2. Предположим, что на столе лежат n монет и кто-то случайным образом выбирает одну из монет и подбрасывает k раз. Пусть в

дальнейшем $n = 2$. Пусть у первой монеты вероятность появления орла p_1, а для второй — p_2. В этом случае параметр, оценить который является второстепенной задачей будет π – вероятность выбора первой монеты.

Предполагая, что при каждом выборе монета подбрасывается 10 раз и таких выборов 10, получается 50 наблюдений, 2 параметра, которые нас интересуют и один параметр π, оценка которой нам не интересна.

Пусть имеем

$$\text{Монета 2: } H, T, H, T, T, H, T, H, H, T.$$
$$\text{Монета 1: } H, H, H, H, H, H, T, H, H, H.$$
$$\text{Монета 1: } H, T, H, T, H, H, H, H, H, H.$$
$$\text{Монета 2: } H, T, H, T, T, T, T, H, H, T.$$
$$\text{Монета 1: } H, H, H, T, H, H, T, H, H, T.$$

Информация про монеты задана лишь для получения оценок максимального правдоподобия, но алгоритм этого не знает. Понятно, что

$$\tilde{p}_1 = \frac{24}{24+6} = 0.8 \qquad \tilde{p}_2 = \frac{9}{9+11} = 0.45.$$

Пусть $p_1^\circ = 0.6$ и $p_2^\circ = 0.5$ начальные значения параматров p_1, p_2. Проводя аналогию с параграфом 3 видим, что θ это вектор(p_1, p_2), а η — π. Вероятность получить k орлов в 10 подбрасываниях, где $c \in \{1, 2\}$ равна

$$p_c(k) = C_k^{10} p_c^k (1 - p_c)^{10-k}. \tag{4.1}$$

Заметим, что биномиальный коэффициент для обеих монет одинаковый, следовательно, остается только отношение таких факторов $p_c^k(1-p_c)^{10-k}$ для некоторых k. Используя начальные значения и (4.1) получаем слеудующую таблицу

Первая итерация			
π	$1-\pi$	Монета 1	Монета 2
0.45	0.55	$\approx 2.2H, 2.2T$	$\approx 2.8H, 2.8T$
0.80	0.20	$\approx 7.2H, 0.8T$	$\approx 1.8H, 0.2T$
0.73	0.27	$\approx 5.9H, 1.5T$	$\approx 2.1H, 0.5T$
0.35	0.65	$\approx 1.4H, 2.1T$	$\approx 2.6H, 3.9T$
0.65	0.35	$\approx 4.5H, 1.9T$	$\approx 2.5H, 1.1T$
		$\approx 21.3H, 8.6T$	$\approx 11.7H, 8.4T$

И тогда, в следующей итерации мы получаем

$$\hat{p}_1^{(1)} = \frac{21.3}{21.3+8.6} = 0.71 \qquad \hat{p}_2^{(1)} = \frac{11.7}{8.4+11.7} = 0.58. \tag{4.2}$$

Число итераций до сходимости зависит от начальных условий и насколько далеко мы находимся от оценок максимального правдоподобия. При заданных начальных значениях $p_1^\circ = 0.6$ и $p_2^\circ = 0.5$ получаем следующую последовательность оценок. Так же в Таблице 1 представлены оценки полученные в случае со случайными начальными значениями.

Таблица 1. Сходимость с фиксированным и случайным начальным значением

Итерация	фиксированный старт		случайный старт	
	$\hat{p}_1^{(i)}$	$\hat{p}_2^{(i)}$	$\hat{p}_1^{(i)}$	$\hat{p}_2^{(i)}$
1	0.600000000	0.500000000	0.935954410	0.0659086863
2	0.713012235	0.581339308	0.759177699	0.434881703
3	0.745292036	0.569255750	0.78052855	0.485752725
4	0.768098834	0.549535914	0.79025212	0.505705724
5	0.7831645	0.534617454	0.794205962	0.513867055
6	0.791055245	0.52628116	0.795774011	0.517227986
7	0.794532537	0.522390437	0.79639082	0.518613125
8	0.79592866	0.520729878	0.796632802	0.519183793
9	0.796465637	0.520047189	0.796727694	0.519418794
10	0.796668307	0.519770389	0.796764932	0.519515523
11	0.796744149	0.519658662	0.796779561	0.51955532
12	0.796772404	0.519613607	0.796785317	0.519571692
13	0.796782900	0.519595434		

В обоих случаях сходимость к оценке максимального правдоподобия происходит с большой точностью.

Замечание 4.1. Заметим, что распределение Бернулли (подбрасывание монет) можно легко распространить, например, до нормального распределения. Пусть $X_1, \ldots, X_\ell \sim \mathcal{N}(\mu_1, \sigma_1^2)$ и $X_{\ell+1}, \ldots, X_n \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Идея довольно общая и может быть применена и в случае n разных распределений $\mathcal{N}(\mu_i, \sigma_i^2)$, $\forall i \in \{1, \ldots, n\}$ с разными вероятностями (p_i) принадлежности классу $i \in \{1, \ldots, n\}$. Более того, вместо нормальных распределений может быть использовано любое удобное для данной задачи распределение.

Abstract. We derived a convergence result for a sequential procedure known as alternating maximization (minimization) to the maximum likelihood estimator for a pretty large family of models - Generalized Linear Models (GLMs). Alternating procedure for linear regression becomes to the well-known algorithm of Alternating

Least Squares (ALS), because of the quadraticity of log-likelihood function $L(\mathbf{v})$. In GLMs framework we lose quadraticity of $L(\mathbf{v})$, but still have concavity due to the fact that error-distribution is from exponential family (EF). Concentration property makes the Taylor approximation of $L(\mathbf{v})$ up to the second order accurate and makes possible the use of alternating minimization (maximization) technique. Examples and experiments confirm convergence result followed by the discussion of the importance of initial guess.

Список литературы

[1] A. Andresen, "Finite sample analysis of profile M-estimation in the single index model", Electronic Journal of Statistics, **9**(2): 2528 – 2641 (2015).

[2] A. Andresen, V. Spokoiny, "Two convergence results for an alternation maximization", (2015) arXiv: 1501.01525.

[3] J. Bezdek, R. Hathaway, "Convergence of alternating optimization", Neural, Parallel & Pacific Computations **11**, 351 — 368 (2003).

[4] L. Cavalier, G. K. Golubev, D. Picard, A. B. Tsybakov, "Oracle inequialities for inverse problems",. The Annals of Statistics, **30** (3), 843 — 874 (2002).

[5] X. Chen, Large Sample Sieve Estimation of Semi-Nonparametric Models, Handbook of Econometrics, (2007) 6:55495632.

[6] V. Chernozhukov, D. Chetverikov, K. Kato, "Anti-concentration and honest, adaptive confidence bands", The Annals of Statistics, **34** (4), 1653 — 1677 (2014).

[7] A. P. Dempster, N. M. Laird, D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society, Series B, **39**, 1 – 38 (1977).

[8] I. A. Ibragimov, R. Z. Khas'minskij, Statistical Estimation. Asymptotic Theory, Translation from Russian by Samuel Kotz, New York - Heidelberg - Berlin: Springer-Verlag (1981).

[9] A. Kneip, "Ordered linear smoothers", The Annals of Statistics, **22**(2), 835 – 866 (1994).

[10] M. R. Kosorok, Introduction to Empirical Processes and Semiparametric Inference, Springer in Statistics (2005).

[11] G. J. McLachlan and T. Krishman, The EM Algorithm and Extensions, Wiley, New York (1997).

[12] V. Spokoiny, T. Dickhaus, Basics of Modern Mathematical Statistics, Springer Texts in Statistics (2015).

[13] C. F. J. Wu, "On the convergence properties of the EM algorithm", Annals of Statistics, **11**, 95 – 103 (1983).

Поступила 26 октября 2018

После доработки 1 февраля 2019

Принята к публикации 25 апреля 2019