

Решение задачи линейной регрессии с помощью быстрого преобразования Хафа

Павел Безматерных
ИСА РАН
bezmpavel@gmail.com

Тимур Ханипов
ИППИ РАН
timour.khanipov@gmail.com

Дмитрий Николаев
ИППИ РАН
dimonstr@iitp.ru

Аннотация

В данной работе изучается устойчивость метода решения задачи линейной регрессии с помощью быстрого преобразования Хафа при наличии нормально распределенного аддитивного координатного и выбросового шумов, а также производится сравнение предлагаемого метода с некоторыми робастными модификациями метода наименьших квадратов (МНК) и методами Тейла и Сена. Демонстрируется, что предлагаемый метод является устойчивым, и при наличии сильного выбросового шума дает более точные результаты, нежели варианты метода наименьших квадратов с итеративным пересчетом весов и методы Тейла и Сена. Также показано, что метод БПХ наиболее устойчив в случае, когда дополнительно присутствует некоррелированный шум с круговым нормальным распределением.

1. Введение

Важной задачей статистики является нахождение оптимальных оценок параметров модели. Одной из самых часто встречающихся задач такого рода является задача линейной регрессии. В случае, когда результаты измерений искажены только аддитивным или только некоррелированным шумом, данная задача является относительно простой, но при одновременном наличии обоих типов шумов она существенно усложняется. На практике встречается множество приложений, где имеет место подобная ситуация (например, задача цветовой сегментации изображений [1] или задача отслеживания движения объекта [2]). В работе [3] авторами был предложен метод решения этой задачи, основанный на применении

БПХ и гауссовых сверток, однако устойчивость метода изучена не была. В данной работе будет показано, что предложенный способ является робастным и в некоторых случаях даже более точным, нежели распространенные варианты МНК с итеративным пересчетом весов и методы Тейла и Сена [5].

2. Применение преобразования Хафа в задаче линейной регрессии

Задача линейной регрессии состоит в восстановлении линейной зависимости между двумя переменными, заданными в виде множества пар $(x_i, y_i), i = 1, \dots, n$, где x_i — свободная переменная, y_i — зависимая, а n — число пар. Заданное множество для наглядности зачастую представляется в виде набора точек на двумерной плоскости. Также данное множество может быть проинтерпретировано как пиксели некоторого монохромного изображения. Таким образом, неформально задача может быть сведена к поиску на монохромном изображении прямой, которая наилучшим образом описывает заданное распределение точек. Для поиска такой прямой будет использоваться метод, базирующийся на вычислении преобразования Хафа.

Преобразование Хафа является дискретным аналогом преобразования Радона на плоскости и ставит в соответствие каждой прямой на изображении сумму яркостей пикселей вдоль нее. Данное преобразование активно используется в области обработки изображений для поиска прямых. Оно также может быть обобщено для поиска фигур произвольной формы.

Применение преобразования Хафа в сочетании с гауссовым фильтром в решении задачи линейной регрессии уже обсуждалось в работе [3], где были опровергнуты некоторые

заблуждения о неэффективности метода, включая его неустойчивость к нормальному распределенному аддитивному координатному шуму. Стоит отметить, что сложность вычисления преобразования Хафа — $O(n^2 \log n)$, а не $O(n^3)$, как это может показаться на первый взгляд (здесь n обозначает линейный размер изображения). Несмотря на то, что это довольно медленно для разреженных изображений, для насыщенных картинок с количеством точек $N = O(n^2)$, способ является достаточно быстрым.

3. Генерация тестовых распределений

Необходимо отметить, что в данной работе в качестве тестовых данных для простоты использовалось не строгое «ортогональное» распределение (равномерное вдоль прямой и нормальное поперек), а двумерное гауссово распределение с высоким эксцентриситетом. При достаточно большой дисперсии вдоль прямой и малой дисперсии в поперечном направлении свойства построенного распределения не сильно отличаются от «идеального».

Алгоритм генерации тестовых распределений может быть описан следующими тремя шагами:

- случайным образом выбираются «центр» и

- наклон (моделируется отрезок прямой);
- генерируется заданное количество точек с указанным распределением (и выбранным центром);
- добавляется однородный выбросовый шум.

Стандартное отклонение нормального распределения вдоль прямой составляло одну четвертую длины моделируемого отрезка (при таком выборе приблизительно 95% точек попадают внутрь), а длина отрезка бралась равной половине линейного размера квадратного изображения (составляющего 512 пикселей). Отношение дисперсий было равно 10.0, а количество точек распределения и доля выбросового шума были параметрами. Пример подобного рода распределения изображен на рисунке 1.

4. Мера ошибки прямой

Чтобы измерить, насколько хорошо был найден моделируемый отрезок, использовался следующий метод: брался «идеальный» отрезок (A, B) и находился максимум из двух расстояний от его концов до экспериментально найденной прямой. Если обозначить через $Q(L, AB)$ ошибку экспериментальной прямой L , то это

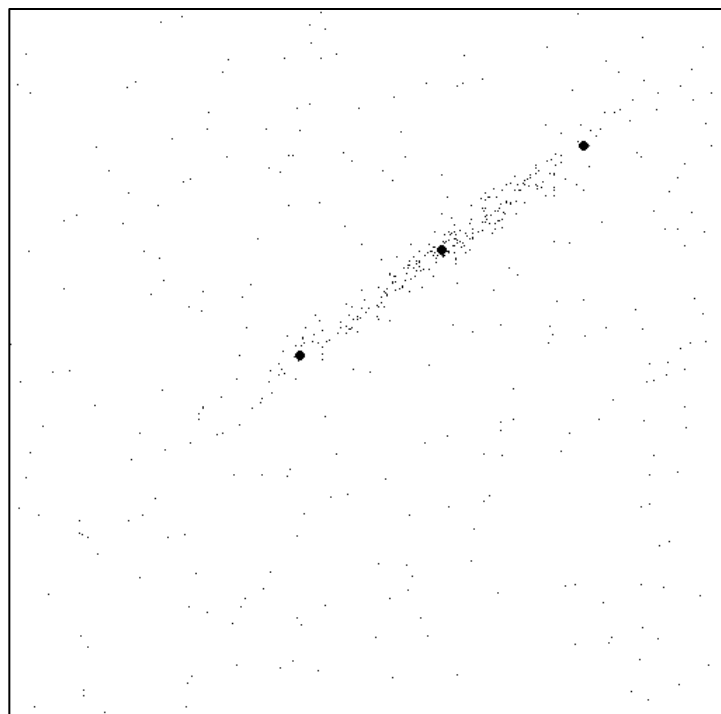


Рисунок 1 Образец распределения с аддитивным и выбросовым шумами. Отметками показаны концы и центр моделируемого отрезка

определение можно записать так:
 $Q(L, AB) = \max(\rho(L, A), \rho(L, B))$.

5. Зависимость ошибки от размера окна сглаживания

Очевидно, что существенное влияние на качество работы алгоритма с использованием БПХ оказывает размер окна, используемого для гауссова сглаживания. В первом эксперименте была исследована зависимость ошибки нахождения прямой от размера данного окна. Для каждого размера окна из некоторого набора была сгенерирована серия из ста распределений, затем для каждого распределения вычислялась прямая и фиксировалась ошибка. Полученный результат представлен на рисунке 2, где ось абсцисс задает отношение сглаживающего параметра σ_H к стандартному отклонению ортогонального аддитивного шума σ_L . Точками на рисунке обозначаются средние ошибки, а вверх и вниз от них откладываются среднеквадратичные отклонения. Можно заметить, что результат является наилучшим в случае, когда σ_H близко к σ_L , т.е. когда размер окна не сильно отличается от стандартного отклонения шума.

6. Сравнение с модификациями МНК с итеративным пересчетом весов

Было проведено экспериментальное сравнение метода БПХ с тремя другими методами: простым МНК и МНК с итеративным пересчетом весов для двух весовых функций. В данном эксперименте использовались весовая функция Хьюбера и биквадратная весовая функция [4].

Было проведено экспериментальное сравнение

метода БПХ с тремя другими методами: простым МНК и МНК с итеративным пересчетом весов для двух весовых функций. В данном эксперименте использовались весовая функция Хьюбера и биквадратная весовая функция [4]. Для указанных функций использовались канонические параметры: $k_H = 1.345$ для Хьюбера и $k_B = 4.685$ для биквадратной функции.

На рисунке 3 видно, как средняя ошибка каждого из четырех методов зависит от плотности выбросового шума μ – отношения количества выбросовых шумовых точек к общему числу пикселей на изображении (плотность выбросового шума отложена по оси абсцисс, а средняя ошибка метода представлена на оси ординат). Каждый замер проводился на ста различных сгенерированных распределениях. Из рисунка 3 очевидно, что метод, основанный на применении БПХ, оказался самым устойчивым к добавлению выбросового шума, хотя при довольно малом количестве выбросовых точек он уступает другим методам в точности. К тому же, данный метод обладает наименьшим среднеквадратичным отклонением среди всех представленных методов (вверх и вниз от точки на рисунке отложены двадцатипятипроцентные значения от среднеквадратичного отклонения ошибки).

7. Сравнение с методами Тейла-Сена

В рамках данной работы также было произведено сравнение предлагаемого подхода с широко распространенными методами Тейла-Сена. В традиционном методе Тейла-Сена наклон прямой вычисляется как медиана среди наклонов прямых, проведенных через все пары точек

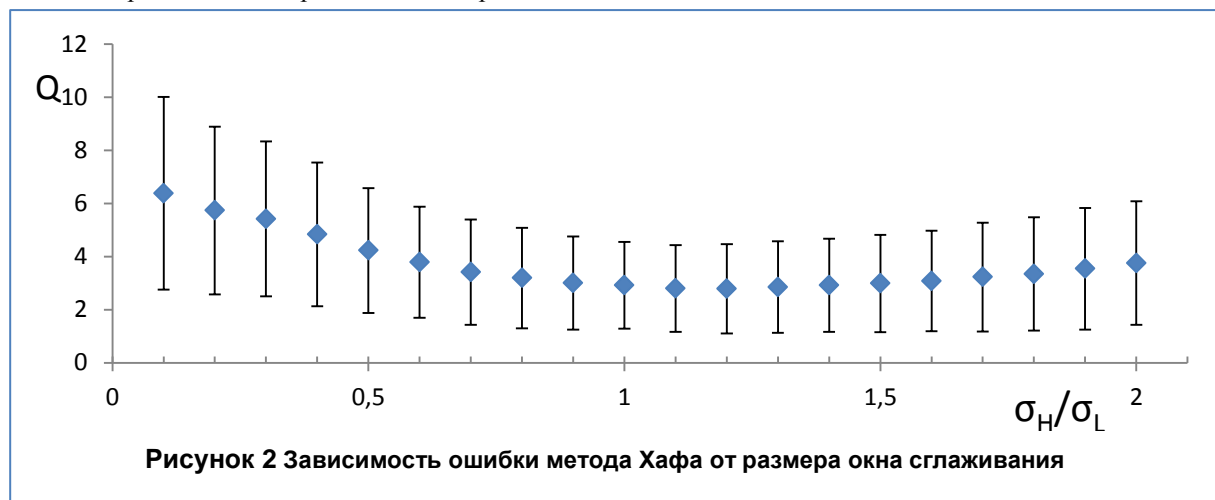


Рисунок 2 Зависимость ошибки метода Хафа от размера окна сглаживания

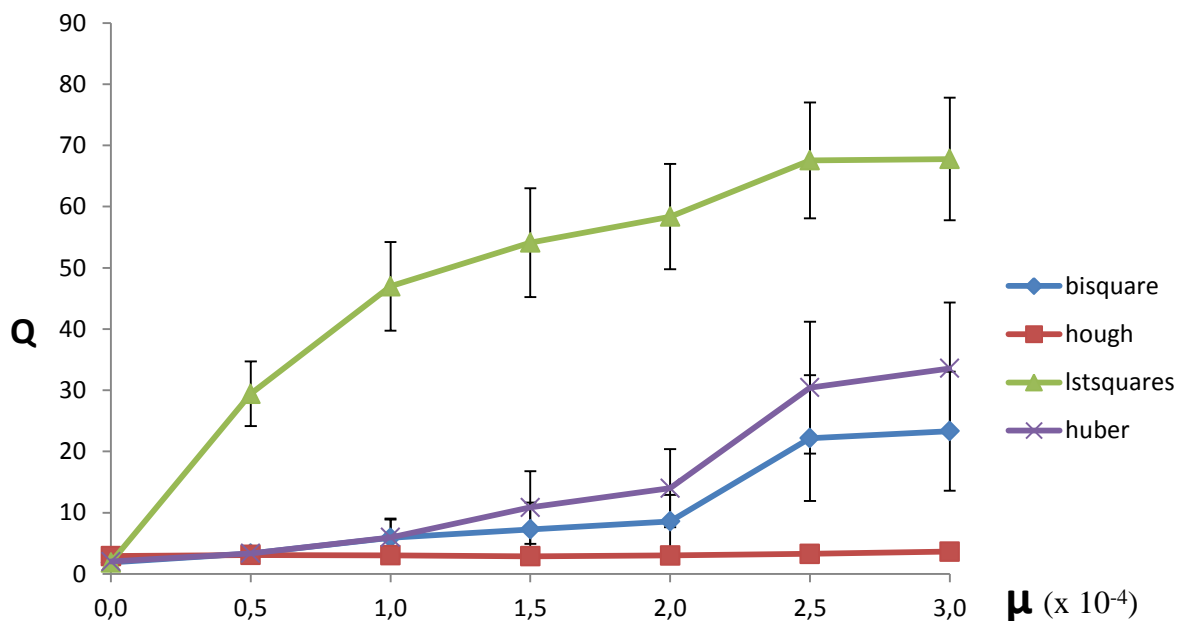


Рисунок 3 Зависимость средней ошибки методов от плотности выбросового шума

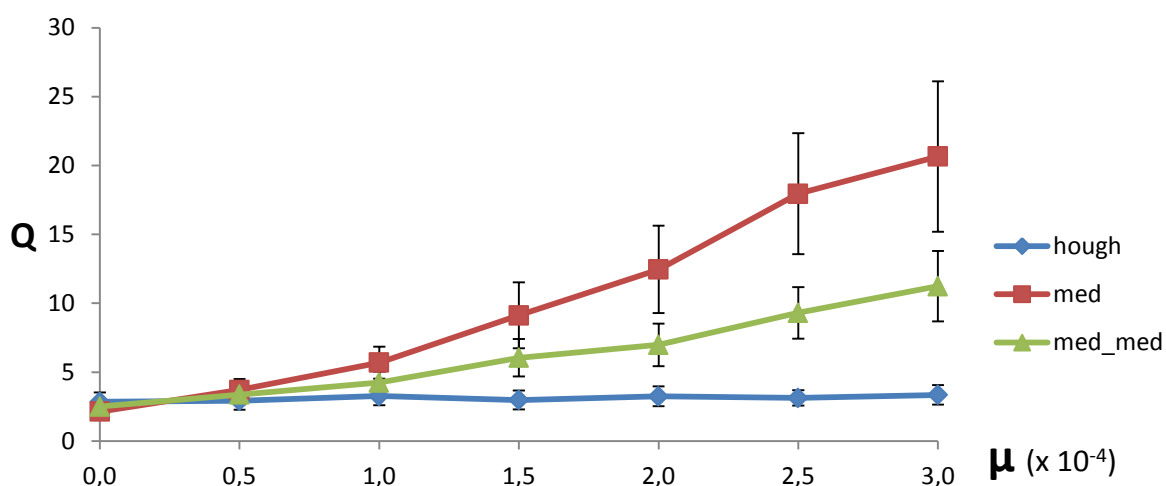


Рисунок 4 Зависимость средней ошибки методов от плотности выбросового шума

распределения. Помимо традиционного метода использовался модифицированный вариант, в котором сначала для каждой точки распределения выбиралась медиана наклонов всех прямых, проходящих через эту и какую-либо другую точку распределения, после чего из полученных значений выбиралась медиана («медиана медиан»). В обоих случаях для определения параметра сдвига прямой бралась медиана среди значений $\{y_i - tx_i\}$, где t – найденный описанными способами наклон (угловой коэффициент) прямой.

На рисунке 4 по оси абсцисс отложена плотность выбросового шума, а по оси ординат — средняя ошибка. В рамках каждой серии проводилось сто экспериментов. Вверх и вниз от точки на рисунке отложены двадцатипятипроцентные значения от среднеквадратичного отклонения ошибки. Видно, что в данном интервале плотности выбросового шума средняя ошибка метода БПХ является наименьшей и не растет, в то время как методы Тейла-Сена имеют пропорциональный рост ошибки. Также видно, что метод БПХ имеет наименьший разброс ошибки.

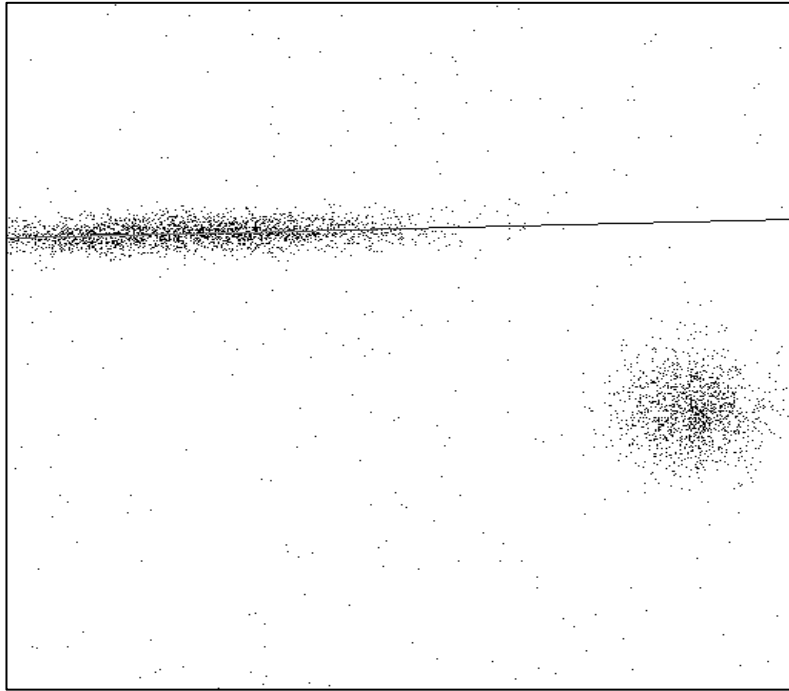


Рисунок 5 Пример смеси выбросового, аддитивного и "круглого" шумов

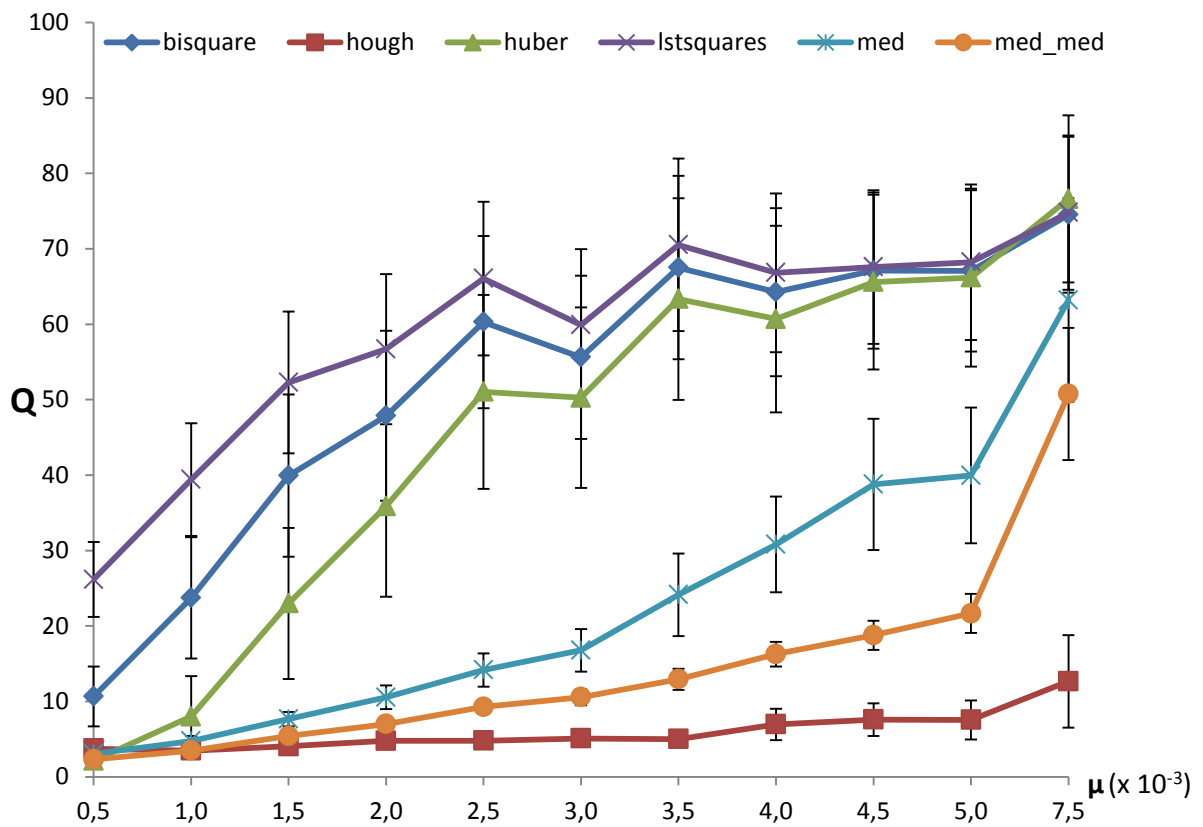


Рисунок 6 Зависимость средней ошибки методов от плотности «круглого» шума

8. Сравнение методов при наличии круглого шумового распределения

Отдельный интерес представляет случай, когда помимо нормально распределенного аддитивного и выбросового шумов присутствует еще и так называемый «круглый» шум, представляющий собой некоррелированный шум с круговым двумерным нормальным распределением. Подобного рода пример приведен на рисунке 5. Прямой линией на этом же рисунке показан ожидаемый результат работы алгоритма.

Было проведено исследование зависимости средней ошибки методов от плотности «круглого» шума при фиксированных параметрах выбросового и аддитивного шумов. Для каждого значения плотности «круглого» шума было сгенерировано сто распределений. Результаты приведены на рисунке 6. Видно, что срыв у МНК-методов происходит практически сразу, в отличие от медианных методов, в то время как у метода БПХ срыва не наблюдается.

9. Выводы

Решение задачи линейной регрессии с помощью комбинации гауссового сглаживания и преобразования Хафа является устойчивым к одновременно присутствующим аддитивному координатному и равномерному выбросовому шумам. Сравнение с некоторыми другими распространенными робастными методами показало, что предложенный алгоритм является более точным, когда плотность равномерного выбросового шума достаточно высока. Также продемонстрировано, что БПХ-метод дает лучший результат в случае, когда помимо аддитивного и равномерного шумов присутствует некоррелированный шум с круговым нормальным распределением.

Ссылки

- [1] D.P. Nikolaev, P.P. Nikolayev *Linear color segmentation and its implementation*, Computer Vision and Image Understanding. 2004, V.94 (Special issue on colour for image indexing and retrieval), pp. 115-139
- [2] L. Nolle. *Application of computational intelligence for target tracking*. In Proc. of 21st European Conference on Modelling and Simulation, 2007, pp. 289-293.
- [3] D.P. Nikolaev, S.M. Karpenko, I.P. Nikolaev, P.P. Nikolaev. *Hough transform: Underestimated Tool in Computer Vision Field*. In Proc. of 22nd European Conference on Modelling and Simulation, 2008, pp. 238-243.

- [4] P.J. Rousseeuw, A.M. Leroy. *Robust Regression and Outlier Detection*. Wiley series in probability and statistics, 2003.
- [5] Wilcox, R. Rand. *Theil-Sen Estimator, Introduction to Robust Estimation and Hypothesis Testing*, Academic Press, 2005, pp. 423–427