P-сплайны

1. Введение

P-сплайны – это замечательный инструмент для сглаживания. Вы можете использовать их для простого сглаживания, например, для вычисления тренда на диаграмме рассеяния или оценки плотности вероятности. Они также могут использоваться в более сложных приложениях, таких как модели с изменяющимися коэффициентами, таблицы смертности и пространственные модели.

Для использования P-сплайнов необходимо сделать несколько выборов: количество и степень B-сплайнов, их область определения, порядок штрафа и значение параметра сглаживания.

2. Иллюстрирование P-сплайнов

На рисунке 1 показана основная идея P-сплайнов, примененная к сглаживанию диаграммы рассеяния. Маленькие серые точки показывают смоделированные данные. Они соединены тонкими серыми линиями для ясности. Толстая синяя линия — это вычисленный тренд P-сплайна. Это сумма "гор" радужных цветов под ней. Это B-сплайны, масштабированные коэффициентами. Значения коэффициентов показаны большими цветными точками.

Легко увидеть, что для получения оптимальной подгонки (в смысле наименьших квадратов) тренда данных можно использовать линейную регрессию. Для каждого B-сплайна его значения во всех наблюдаемых точках x собираются в отдельный столбец матрицы, допустим, B. Минимизация $||y - B\alpha||^2$ даёт $\hat{\alpha}$ и $B\hat{\alpha}$ даёт лучшую подгонку. Как только $\hat{\alpha}$ доступно, мы можем вычислить подогнанную кривую с любым требуемым разрешением, как $\tilde{B}\hat{\alpha}$ заполнив столбцы $\tilde{B}$ значениями B сплайнов, вычисленными при этом разрешении. Заметим, что все B-сплайны имеют одинаковую ширину. В принципе, возможно иметь переменные ширины.

Описана регрессия на B-сплайнах; это может быть эффективным методом во многих случаях. Основной недостаток в том, что гладкость результата в основном определяется количеством B-сплайнов. Увеличение этого числа дает менее гладкий результат. Также при разреженных данных это может происходить.

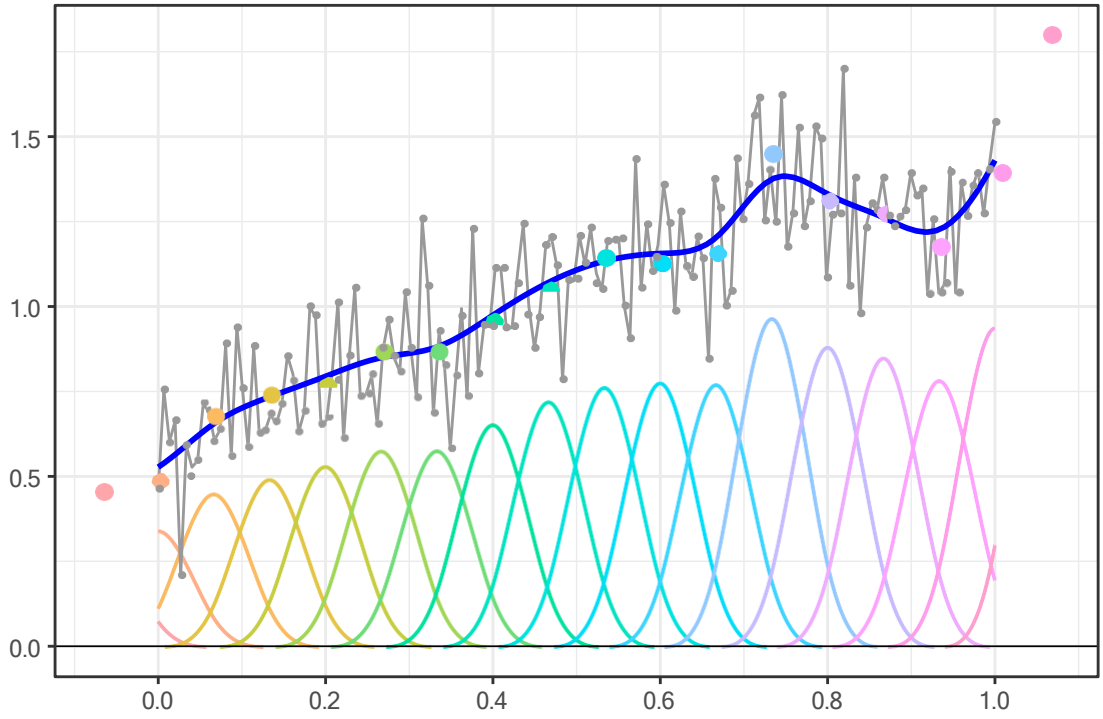

Рисунок 1

Рисунок 1: Основная идея P-сплайнов: сумма базисных функций B-сплайнов с постепенно изменяющимися высотами. Маленькие соединенные серые точки показывают смоделированные данные. Синяя кривая показывает подгонку P-сплайна, а большие точки показывают коэффициенты B-сплайнов (они имеют такие же цвета, как и сплайны).

Некоторые B-сплайны не имеют поддержки из-за отсутствия значений *x* "под" этими B-сплайнами. Один или несколько столбцов будут пустыми, и регрессия не будет работать. P-сплайны добавляют штраф для постоянного сглаживания и устранения проблем с отсутствующей поддержкой. Это простой штраф: ограничить различия между соседними элементами вектор-коэффициентов α. В простейшей форме штраф равен $\lambda \sum_j(\alpha_j-\alpha_{j-1})^2$.

Целевая функция: $||y - B\alpha||^2 + \lambda \sum_j (\alpha_j - \alpha_{j-1})^2$

Может быть записана компактно в виде: $||y - B\alpha||^2 + \lambda ||D\alpha||^2$

Параметр λ настраивает штраф: увеличение его значения даёт более гладкий результат. Здесь – это матрица, такая, что Dα формирует различия α.

Явное решение: $\hat{\alpha} = (B'B + \lambda D'D)^{-1} B'y$

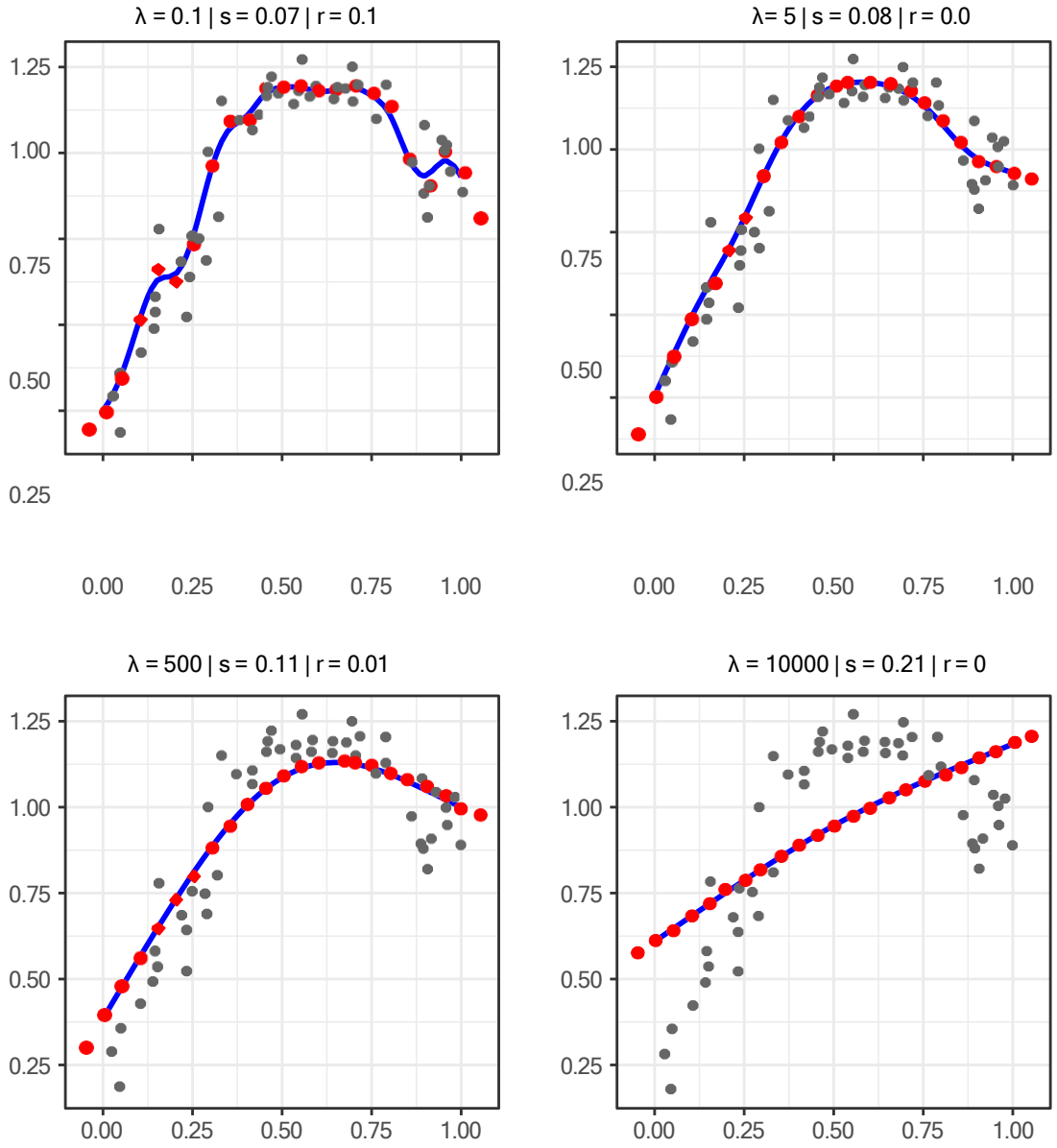

Рисунок 2

Рисунок 2: Иллюстрирует влияние этих штрафов, применяя тот же набор данных с теми же B-сплайнами, меняя λ. Иллюстрация подгонки P-сплайнов при изменении силы штрафа (параметр λ). Количество и ширина B-сплайнов не изменяются, но их коэффициенты (красные кружки) становятся более ограниченными при увеличении λ.

3. P-сплайны параметры

Чтобы использовать P-сплайны, необходимо выбрать значения пяти параметров

Область. Это область на оси, где определены B-сплайны. Часто минимальное и максимальное значения выбираются в качестве левых и правых границ (значение по умолчанию). В других случаях их можно округлить до удобных чисел. Не повредит выбрать область (намного) шире диапазона *x*. Экстраполяция происходит автоматически.
Область должна быть достаточно широкой, чтобы включать все наблюдаемые *x*). Если это не так, функция bbase (которая вычисляет базисную матрицу B-сплайнов) автоматически расширяет область до min(*x*) слева и max(*x*) справа. Если это произойдет, также будет выдано предупреждение.

Степень B-сплайнов. Значение по умолчанию — 3, что дает кубические B-сплайны, состоящие из плавно соединяющихся кубических полиномиальных сегментов. На практике редко нужна другая степень.

Количество B-сплайнов. Значение по умолчанию — 10, но больше — безопасный выбор. Для некоторых данных может потребоваться очень гибкая кривая, и 10 B-сплайнов может быть недостаточно. Невозможно иметь слишком много B-сплайнов, так как штраф убирает все особенности.

Порядок штрафа. Значение по умолчанию — 2. Это обычно хороший баланс между гладкостью подгонки и близостью к данным. В особых случаях рекомендуется штраф первого или третьего порядка.

Параметр сглаживания. Это λ, ключевой элемент P-сплайнов, и нет значения по умолчанию. Если ваши данные хорошо себя ведут, хорошее значение можно автоматически определить с помощью различных методов, но нет гарантии, что они всегда дадут значимый результат.

Всегда полезно изучить диапазон значений для λ и оценить результаты визуально. Такой диапазон должен быть большим и использовать линейную сетку значений для $\log_{10}(\lambda)$.

В большинстве случаев рассчитываются две базисные матрицы: одна для подгонки данных и другая для построения графика подогнанной кривой на сетке. Важно, чтобы область, степень и количество B-сплайнов во второй матрице совпадали с первой.

4. Преимущества P-сплайнов

P-сплайны объединяют (относительно много, равномерно расположенных) B-сплайнов с дискретным штрафом за шероховатость на их коэффициентах. Это дает им много практических и теоретических преимуществ:

-- B-сплайны имеют одинаковую форму и равномерно расположены; оптимальная установка узлов не является проблемой.

-- Благодаря штрафу количество B-сплайнов можно выбирать свободно. Невозможно иметь слишком много B-сплайнов.

-- B-сплайны любой степени можно вычислить быстро и легко. Во многих случаях линейные B-сплайны работают хорошо. Их очень легко вычислить. С множеством узлов они дают приятную кусочно-линейную подгонку.

-- Базисная матрица B-сплайнов является по своей природе разреженной. Наше программное обеспечение может вычислять очень большие B-сплайны в разреженной матрице намного быстрее. Штрафная матрица тоже разреженная. Используя программное обеспечение для работы с разреженными матрицами, наборы данных с миллионами наблюдений можно сглаживать за доли секунды.

-- Модель P-сплайнов параметрическая. Параметрами являются коэффициенты B-сплайнов. Это близко к локальной функции подгонки. Они имеют прямую и ясную интерпретацию. Это не относится к параметрическим моделям для обычных сплайнов.

-- Для P-сплайнов найденная подгонка получается за счет значений, которые представляют собой коэффициенты.

-- Штраф является ключевым элементом. Обычно он основан на разностях (высшего порядка) коэффициентов. Его порядок можно выбирать свободно, независимо от степени B-сплайнов. Более общие уравнения разностей могут использоваться в особых случаях, например, для периодических или циклических данных.

-- Дискретный штраф не является аппроксимацией к непрерывному. Популярная интегрированная вторая производная квадрата, т.е. та, которую мы знаем из сплайнов сглаживания или по работе O'Sullivan (1986), требует подгонки кривой, состоящей из полиномиальных частей третьей степени или выше; иначе штраф исчезает.

-- P-сплайны основаны на (штрафной) регрессии, поэтому ненормальные данные могут быть обработаны без труда, с адаптацией структуры обобщенной линейной модели.

-- Многочисленные расширения легко реализовать, такие как модели с аддитивными и переменными коэффициентами, квантильное и ожидаемое сглаживание, регрессия сигналов, составная ссылка, и другие.

-- P-сплайны можно интерпретировать и анализировать как смешанные модели. Параметры штрафа становятся отношениями дисперсий. Быстрые алгоритмы могут легко оценивать несколько параметров штрафа.

-- Байесовские P-сплайны легко реализуются с использованием марковских цепей или приближения Лапласа. Любая из структур автоматически вычисляет параметры настройки.

-- Эффективная размерность модели P-сплайнов четко определена и легко вычисляется. Это полезно для количественной оценки сложности модели, перекрестной проверки и вычисления AIC.

-- Тензорные произведения B-сплайнов и расширенные штрафы обобщают P-сплайны для многомерного сглаживания. Большие наборы данных могут быть обработаны напрямую. Данные на огромных сетках (1000 на 1000 ячеек и больше) не являются проблемой, поскольку алгоритмы массивов делают вычисления.

5. Литература

Eilers, P. H. C., Marx, B. D., and Durbán, M. 2015. Twenty years of P-splines. *SORT-Statistics and Operations Research Transactions*, **39**(2), 149–186.

Eilers, P.H.C, and Marx, B.D. 2021. *Practical Smoothing: The Joys of P-splines*. Cambridge University Press.

O'Sullivan, F. 1986. A statistical perspective on ill-posed inverse problems (with discussion).
Statistical Science, **1**, 505–527.

Ruppert, D. 2002. Selecting the number of knots for penalized splines.
Journal of Computa- tional and Graphical Statistics, **11**, 735–757.

Wand, M. P., and Ormerod, J. T. 2008. On semiparametric regression with O'Sullivan penalized splines. *Australian & New Zealand Journal of Statistics*, **50**, 179–198.