

Р-сплайны

1 Введение

Р-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. Этот документ имеет две цели: одна – дать (очень) краткое техническое введение; другая – объяснить, почему Р-сплайны настолько полезны. Мы также обсуждаем ряд заблуждений, с которыми мы сталкивались в литературе.

Для использования Р-сплайнов необходимо сделать несколько выборов: количество и степень В-сплайнов, их область определения, порядок штрафа и значение параметра сглаживания. Мы последовательно проработаем их и укажем разумные значения по умолчанию для большинства из них. Функции в пакете JOPS используют эти значения по умолчанию.

2 Иллюстрирование Р-сплайнов

На рисунке 1 показана основная идея Р-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 Р-сплайна. Это сумма "гор" радужных цветов под ней. Это В-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 В-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 подогапанную кривую с любым требуемым разрешением, как $\tilde{B}\hat{\alpha}$ заполнив столбцы \tilde{B} значениями B сплайнов,

вычисленными при этом разрешении.

Заметим, что все В-сплайны имеют одинаковую ширину. В принципе, возможно иметь переменные ширины.

Описана регрессия на В-сплайнах; это может быть эффективным методом во многих случаях. Основной недостаток в том, что гладкость результата в основном определяется количеством В-сплайнов.

Увеличение этого числа дает менее гладкий результат. Также при разреженных данных это может происходить.

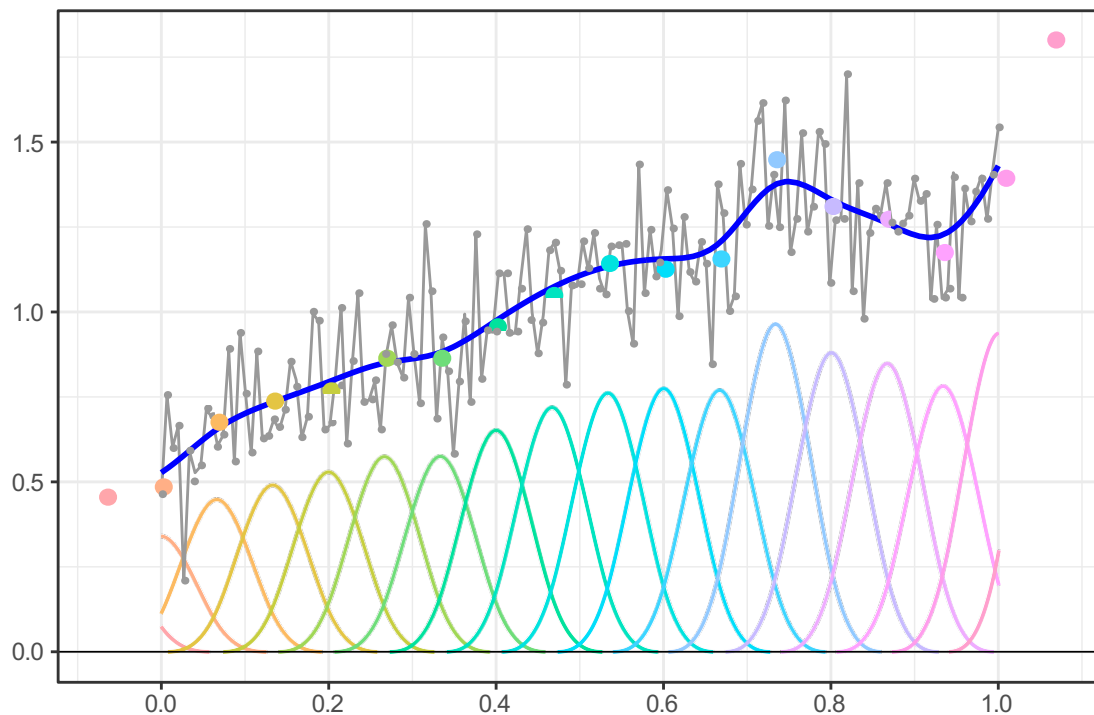


Рисунок 1

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

Некоторые В-сплайны не имеют поддержки из-за отсутствия значений x "под" этими В-сплайнами. Один или несколько столбцов будут пустыми, и регрессия не будет работать. Р-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 написана компактно в виде:

$$||y-B\alpha||^2 + \lambda ||D\alpha||^2$$

Параметр λ настраивает штраф: увеличение его значения дает более гладкий результат. Здесь — это матрица, такая, что $D\alpha$ формирует различия α . Явное решение: $\hat{\alpha} = (B'B + \lambda D'D)^{-1} B'y$.

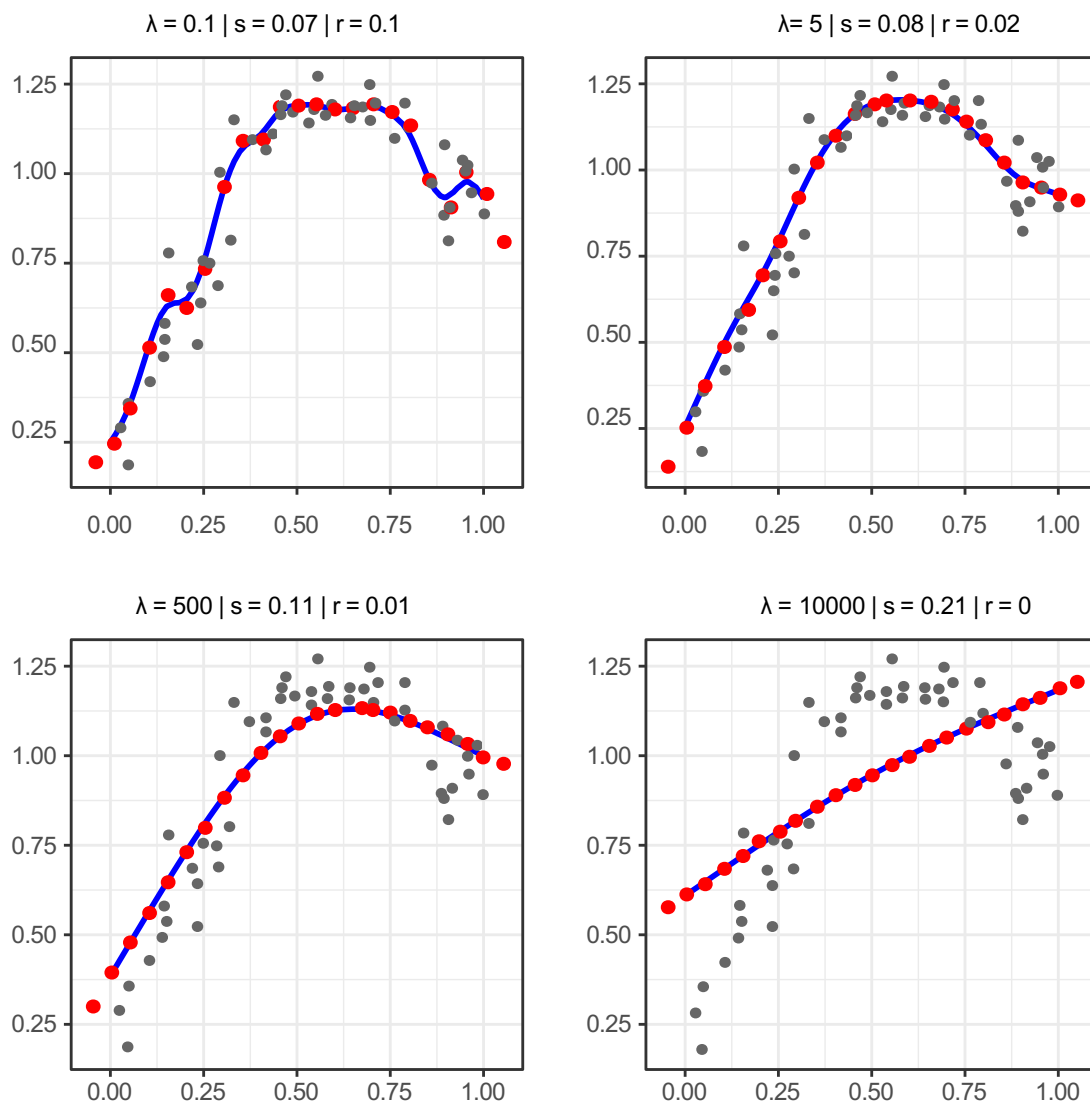


Рисунок 2

Рисунок 2: иллюстрирует влияние этих штрафов, применяя тот же

набор данных с теми же В-сплайнами, изменяя λ . Иллюстрация подгонки Р-сплайнов при изменении силы штрафа (параметр λ). Количество и ширина В-сплайнов не изменяются, но их коэффициенты (красные кружки) становятся более ограниченными при увеличении λ .

3 Р-сплайны параметры

Чтобы использовать Р-сплайны, необходимо выбрать значения пяти параметров

Область. Это область на оси, где определены В-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 В-сплайнов) автоматически расширяет область до $\min(x)$ слева и $\max(x)$ справа. Если это произойдет, также будет выдано предупреждение.

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

Количество В-сплайнов. Значение по умолчанию — 10, но больше — безопасный выбор. Для некоторых данных может потребоваться очень гибкая кривая, и 10 В-сплайнов может быть недостаточно. Невозможно иметь слишком много В-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 Р-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 В-сплайнов во второй матрице совпадали с первой.

3 Преимущества Р-сплайнов

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

- В-сплайны имеют одинаковую форму и равномерно расположены; оптимальная установка узлов не является проблемой.
- Благодаря штрафу количество В-сплайнов можно выбирать свободно. Невозможно иметь слишком много В-сплайнов.
- В-сплайны любой степени можно вычислить быстро и легко. Во многих случаях линейные В-сплайны работают хорошо. Их очень легко вычислить. С множеством узлов они дают приятную кусочно-линейную подгонку.
- Базисная матрица В-сплайнов является по своей природе разреженной. Наше программное обеспечение может вычислять очень большие В-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 Р-сплайнов параметрическая. Параметрами являются коэффициенты В-сплайнов. Это близко к локальной функции подгонки. Они имеют прямую и ясную интерпретацию. Это не относится к параметрическим моделям для обычных сплайнов.

Для Р-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 В-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.
- Р-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.
- Р-сплайны можно интерпретировать и анализировать как смешанные модели. Параметры штрафа становятся отношениями дисперсий. Быстрые алгоритмы могут легко оценивать несколько параметров штрафа.
- Байесовские Р-сплайны легко реализуются с использованием марковских цепей или приближения Лапласа. Любая из структур автоматически вычисляет параметры настройки.
- Эффективная размерность модели Р-сплайнов четко определена и легко вычисляется. Это полезно для количественной оценки сложности модели, перекрестной проверки и вычисления AIC.
- Тензорные произведения В-сплайнов и расширенные штрафы обобщают Р-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

Литература

- Eilers, P. H. C., Marx, B. D., and Durban, 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 Computational 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.