回帰曲線における knot の最適化

橋本 捺希（指導教員：吉田 裕亮）

1　スプライン回帰

スプライン回帰は, 機械学習手法の一つであり, 一次元データを小さな区間に分割し, 各区間ごとに滑らかな曲線をフィッティングすることができる [1]. この手法の利点は, データの特徴をより正確に捉えつつ, 曲線が滑らかで過剰に複雑になりにくい点にある. スプライン回帰では, knot（2 節）と呼ばれる区間の境界点を配置し, 各区間での曲線の滑らかさを保つために条件が設定される. これにより, データの変動を追随し, 曲線の形状を制御することが可能である.

knot の個数や位置の選択は, スプライン回帰の精度や柔軟性に大きく影響する. 本研究では, 様々な数や位置で knot を設定し, 最適なスプライン回帰について考察する.

2　knot

knot は, スプライン曲線などの幾何学的なオブジェクトにおいて, 曲線の滑らさと連続性を保つために使われる境界点である. knot はスプライン関数が切り替わる境界点であり, それぞれの knot において関数が異なる区間に適用されることで, 全体として滑らかな曲線が形成される.

knot の数が増えると, 曲線はデータにより適合しやすくなるが, 同時に過学習のリスクも高まる. 逆に, knot の数が少なすぎると, 曲線はデータの変動に対応しきれず, 適切なフィッティングが難しくなる. knot の役割は, その点での滑らかな推移を促進することで, 曲線が全体として自然な形状を持つように調整することである. knot の配置は, 関数がどの範囲でどれだけ急激に変化するかを制御し, データに最適に適合させるために調整する.

3　AIC

赤池情報量規準 (Akaike's Information Criterion) は, 一般に AIC と呼ばれ, 統計モデルの良さを評価するための指標である. AIC は, 式 (1) で与えられる.

$$\mathrm{AIC} = -2\ln L + 2k \tag{1}$$

L は最大尤度, k は自由パラメータの数である.

また, 各標本の誤差項が独立で確率分布が正規分布の場合, 式 (1) を式 (2) のように表すことができる.

$$\mathrm{AIC} = \sum_{i=0}^{n} \ln \sigma_i^2 + 2k \tag{2}$$

ここで n は標本サイズ, σ_i は各標本の標準誤差である.

本研究では, 各区間で生成した回帰曲線に対して式 (2) を用いて評価を行う. AIC の数値を比較したときに, 最小のものが良いモデルとされる. 値の大きさに関わらず, 2 つのモデルの AIC の差が 1 以上であれば優位な差が認められ, 1 未満であれば優位な差はないと判断される. AIC は, 式の簡易性や近似の正確性などのトータルでモデルを評価することができる.

4　数値実験概要

本研究では, いくつかの実データにおいて, 多項式の次数や, knot の数と位置を変化させ, スプライン回帰を行う. それにより生成されたモデルを, AIC を用いて評価し, 最良の knot を選択する手法を検討する. 最後に, それらの実験から, 最良の knot を自動で選択するためのアルゴリズムを検討する.

5　実データへの応用

2 つの実データに対して, 様々な knot でスプライン回帰を行い, モデルの評価と比較を行う.

5.1　給与データ

目的　「An Introduction to Statistical Learning」（当該書籍）の 7.4 節で扱われている年齢別給与データを使用して, スプライン回帰を行い, AIC を適用し, 当該書籍で示されている knot の取り方が最適であるかを検討する. 使用するデータは 18 歳から 80 歳の給与データで, 標本サイズ $n = 3000$ である.

予想　このデータは 1 つの凸グラフとなることが予測できるため, 少なくとも 2 次関数で近似を行うことができる. また, 60 歳以降に特徴的な変化があるため, その周辺に knot を設定すると良い結果が得られるのではないかと予想できる.

結果　本実験では, 当該書籍のモデルよりも AIC が小さいモデルを発見した. 当該書籍のモデルは, knot が 3 つ（$x = 25, 40, 60$）に設定された 3 次多項式のスプライン回帰であった. 当該書籍のモデルを図 1 に示す.

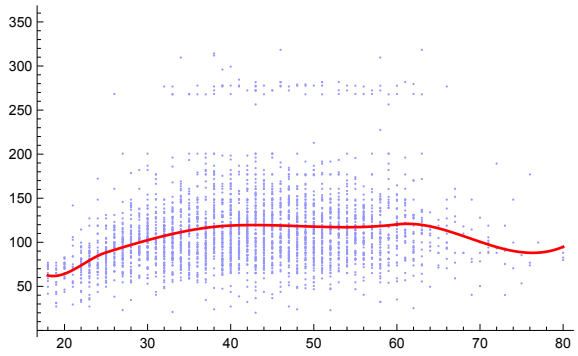

図 1: 当該書籍のモデル（knot3 つ）

一方, 本実験で見つけた AIC が最良のモデルは knot が 1 つで $x = 60$ に設定された 3 次多項式のスプライン回帰であった. AIC が最良のモデルを図 2 に示す.

図 1, 図 2 より, どちらのモデルも特徴的な変化を捉えたモデルが生成できた.

また, 2 つのモデルの knot と AIC をまとめたものを表 1 に示す.

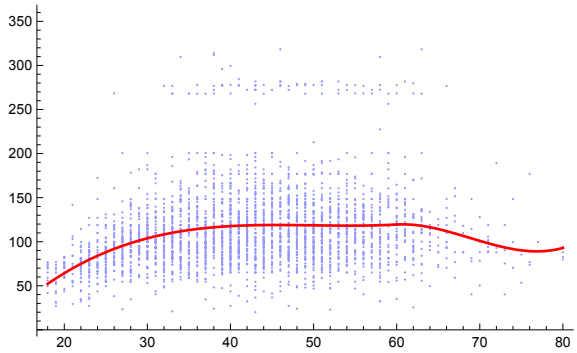

図 2: AIC が最良のモデル（knot1 つ）

表 1: 給与データの AIC

	knot 数	knot の位置	AIC
書籍	3	$x = 25, 40, 60$	22128.30
実験結果	1	$x = 60$	22122.75

表 1 より, AIC は当該書籍よりも実験結果の方が良い結果となり, AIC の差は 5.55 であった. AIC に 1 以上の差があるため, この差は有意であると言える. 以上の結果から, このデータにおける 3 次のスプライン回帰の knot は 1 つで十分であると言える.

5.2 最高気温データ

目的 気象庁のデータベースより 2 年分の東京の日最高気温データ [3] を使用して, スプライン回帰を行う. 日付を横軸（$-1 \leq x \leq 1$ の範囲にスケーリング）, 日最高気温を縦軸として, スプライン回帰の次数や knot の数と位置を変えながら, AIC を調べた. また AIC が最小となるのはどのときであるかを調べる.

予想 $x = 0.5$ の周辺に梅雨明け以降の日最高気温の特徴的な変化が確認できる. このデータは凹凸が 2 つ見られるため, スプライン回帰で特徴的な変化を捉えるためには, 3 次多項式以上, 複数の knots が必要であると予想した.

結果 多項式の次数や knot, AIC の結果をまとめた表を表 2 に示す.

表 2: 東京の最高気温データの AIC

次数	knot 数	knot の位置	AIC
3	1	$x = -0.1$	1807.58
	2	$x = 0.0, 0.8$	1775.97
	3	$x = -0.4, 0.2, 0.6$	1725.83
5	0	-	2164.75
7	0	-	1825.16
9	0	-	1829.16

表 2 より, AIC で最も良い結果は, 3 次多項式のスプライン回帰で knot が 3 つ, $x = -0.4, 0.2, 0.6$ のときであった. このときのグラフを図 3 に示す.

次に, 図 3 よりも $x = 0.5$ 周辺の特徴的な変化を捉えることができたモデルを図 5.2 に示す. 図 4 は, 3 次多項式のスプライン回帰で knot が 3 つ, $x = -0.5, 0.3, 0.6$ のときであった.

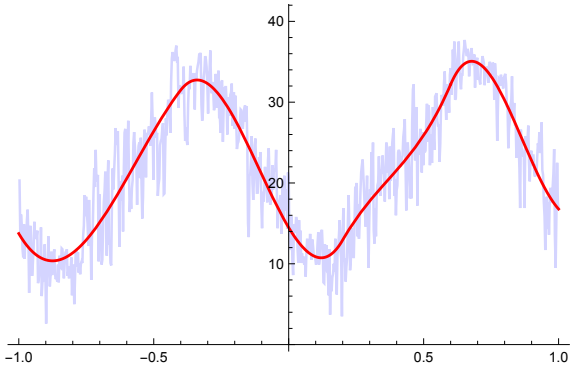

図 3: AIC が最小のスプライン回帰

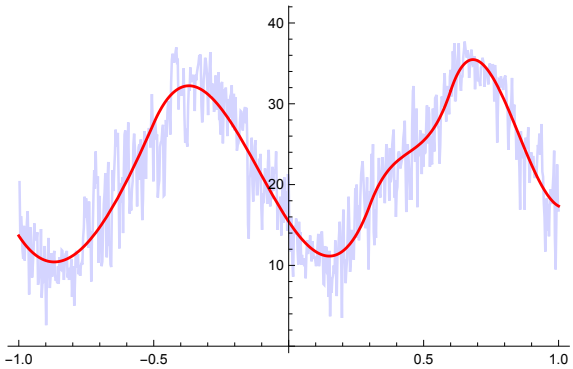

図 4: 特徴が捉えられるスプライン回帰

表 2 から分かるように, knot の多い方が AIC は良くなり, 特徴も捉えることができた. また, 多項式を 5 次, 7 次, 9 次に設定したときの, knot なしの回帰も行ったが, スプライン回帰よりも AIC の値が大きくなり, 特徴的な変化も捉える事ができなかったため, knot ありの方が良い回帰であると言える.

6 考察

多項式の次数によっても必要な knot の数が変わる. 一般に AIC が最小のモデルは, 過学習を起こさずにフィッテングすることができる. しかし細かな特徴を捉えたい場合, AIC ではその特徴がノイズや過学習として判断される可能性があるため, AIC の評価のみで判断するのは困難である. 以上のことから, AIC を考慮しつつ, 別の指標も加えてモデルを選択すれば, 欲しい特徴を捉えたモデルが生成できるのではないかと考えられる. 今後は, 自動で最適な knot の選択を行うアルゴリズムを検討したい.

参考文献

[1] G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning*, Springer, New York, NY (2017)

[2] “赤池情報量基準”. ウィキペディア (Wikipedia): フリー百科事典. 2023 . https://ja.wikipedia.org/wiki/赤池情報量基準/, (参照 2024-01-06)

[3] 国土交通省 気象庁. “過去の気象データ・ダウンロード”. 気象庁ホームページ. https://www.data.jma.go.jp/risk/obsdl/index.php, (参照 2023-11-29).