6　モデル選択と交差検証

これまで見てきたモデル $\hat{\boldsymbol{x}} = \arg\min \|\boldsymbol{y} - A\boldsymbol{x}\|^2 + \lambda\Phi(\boldsymbol{x})$ の形には常にパラメータ λ が入っていました。この λ は尤度関数に対する制約項の「重み」であり、ベイズ統計モデルでいうところの hyper-parameter です。このパラメータを調整することで欲しい結果を得ることができますが、自然科学への応用を考える場合は、むしろ解が一意に決まらないため問題視されることもあります。λ を変えると $\boldsymbol{x}$ の非ゼロ要素も変わり、結果としてモデルも変わります。したがって、モデルは λ に依存するため、データに対してどのモデルが（どの λ が）最適か選択する必要があります。本章ではそのような手法の 1 つである「交差検証（cross-validation: CV）」について主に説明し、最後に他の基準についても簡単に紹介します。

まず、過適合（over-fitting）の概念について説明します。図 1 左は単純な比例関係（点線）にある 2 つの量 x, y に対称な分布のノイズを y に加えたもの（白丸）です。比例の関係にあることがわかっている場合は $y = ax + b$ をあてはめて、最小二乗法でモデルパラメータ a と b を推定します（図 1 中央）。しかし、x と y の関係が未知の場合はどうでしょうか。例えば 5 次の多項式を当てはめると図 1 右のようになります。パラメータの数が増える分、モデルは y の値により近くなりますが、本来の一次式の関係からはむしろ外れたモデルになります。例えば、もう 1 つデータを観測して図 1 中央と右にあるような黒三角の点が得られたとしましょう。モデルとの差はむしろパラメータを増やした右図の方が大きくなってしまい、モデルの予測精度が落ちていることがわかります。これはモデルが過剰にデータに適合した結果、ノイズ成分まで説明しようとしたために起こった現象で、このような状態は「過適合」（over-fitting）と呼ばれます。

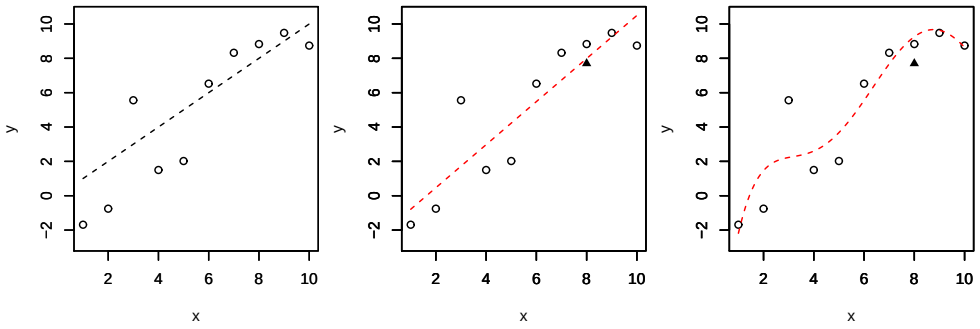

図 1: 左：点線で表される 1 次の関係にある量 x,y。y には正規分布のノイズが加えられている。中：一次式で線形回帰して推定されたモデル（赤線）。右：5 次の多項式で回帰して得られたモデル（赤線）。

この過適合を防ぐために有効な手法の 1 つが交差検証です。「K-分割交差検証」(K-fold cross-validation) ではデータを K 個のサブデータに分割します。この内、1 つのサブデータを除いた $K-1$ 個のデータ（訓練データ）からモデルパラメータを求めます。推定されたモデルの予測精度は除いておいた 1 つのサブデータ（テストデータ）で検証します。テストデータとモデルの残差の平均二乗誤差（mean square error; MSE）が予測精度の指標としてよく使われます。

K-分割交差検証の概念と例を図 2 で説明します。ここでは 2 次の関係 $y = -2 - 3/2x + 3/4x^2$

にある x,y を考えます。y は 1000 点用意します。実際には何次の多項式が適当なのかわからないためモデルは 7 次の項まで考慮します。このモデルは以下の線形の形で書くことができます。

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 & x_1^4 & x_1^5 & x_1^6 & x_1^7 \\ 1 & x_2 & x_2^2 & x_2^3 & x_2^4 & x_2^5 & x_2^6 & x_2^7 \\ & & \vdots & & & & & \\ 1 & x_N & x_N^2 & x_N^3 & x_N^4 & x_N^5 & x_N^6 & x_N^7 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ a_7 \end{pmatrix} + \mathcal{N}(0, \sigma^2) \qquad (1)$$

$\mathcal{N}$ は平均 0、分散 σ^2 の正規分布で、測定誤差を表します。この場合、最小二乗法 $\arg\min \|\boldsymbol{y} - M\boldsymbol{a}\|$ で解を得ることができますが、これだけだと過適合になる可能性があるため、$\boldsymbol{a}$ の 1 次ノルム最小化を使います。つまり、求めたい $\hat{\boldsymbol{a}}$ は

$$\hat{\boldsymbol{a}} = \arg\min \|\boldsymbol{y} - M\boldsymbol{a}\|_2^2 + \lambda \|\boldsymbol{a}\|_1 \qquad (2)$$

と書けます。ここで M は x に関する式 (1) の行列です。評価関数のこの形はこれまで見慣れたものです。

データ $\boldsymbol{y}$ はまず 10 分割され、1 つがテストデータ、残りが訓練データになります。この組み合わせは 10 通りあります。図 2 の手順とおり、まず訓練データに対してある λ のモデルを最適化します。そのモデルをテストデータに当てはめ、MSE を計算します。訓練データとテストデータは 10 通り試すことができるので、ある λ に対して MSE が 10 個計算でき、それらを加算平均して最終的な MSE とその標準誤差を計算します。図 2 の下の図はそのようにして計算された MSE を様々な λ ごとにプロットしたものです。計算は 1 章で紹介した `R` の `glmnet` パッケージにある交差検証用関数 `cv.glmnet` を使いました。グラフの上辺にある数字はモデルで使用された要素の数です。λ が非常に大きいときは 1 つの要素のみで最適化されているためモデルは単純すぎて、MSE が大きくなっていることがよくわかります。逆に λ が非常に小さいと過適合になり、やはり MSE は大きくなります。さて、MSE が最小値をとるのは図中の左の点線で示される位置で、要素を 4 つ使ったモデルです。しかし、標準誤差で示されているように不定性があるため、最適な λ がこの最小 MSE の時の λ だとは限りません。実際、真のモデルは 2 次式（要素 3 つ）ですが、最小 MSE では 3 次式（要素 4 つ）が使われており、正しく推定されていないことがわかります。よく使われる手法は「最小 MSE の標準誤差内にある最も大きな λ」を選択することです（one standard error rule）。これは不定性があるなかで、最も説明変数が少なくなるモデルを選んでいることに相当します。図 2 の場合は右側の点線の位置になり、この λ では要素が 3 つ使われており、正しい解が選択されています。

前章で紹介した Total Variation Minimization を使ったドップラートモグラフィーで交差検証を試すと、図 3 のようになります。前章とは違い、矮新星 TU Men のデータを使っています。図 3 を見ると、実際の問題でも、λ が大きいとモデルが単純になりすぎて MSE は大きくなり、λ が小さい領域では over-fitting になっていることがわかります。このケースでは $\lambda \sim 0.1$ で MSE 最小、”one standard error rule” では $\lambda \sim 0.5$ になり、後者が最適モデルと考えられます。図 4 は推定されたドップラーマップで、左図が単純過ぎるモデル ($\lambda = 5.0$)、中図が ”one standard error rule” で得られたもの ($\lambda = 0.5$)、右図が MSE 最小のモデル ($\lambda = 0.1$) です。左のマップが単純過ぎるのは一目でわかります。右のマップでは非対称で局所的な構造が強調さ

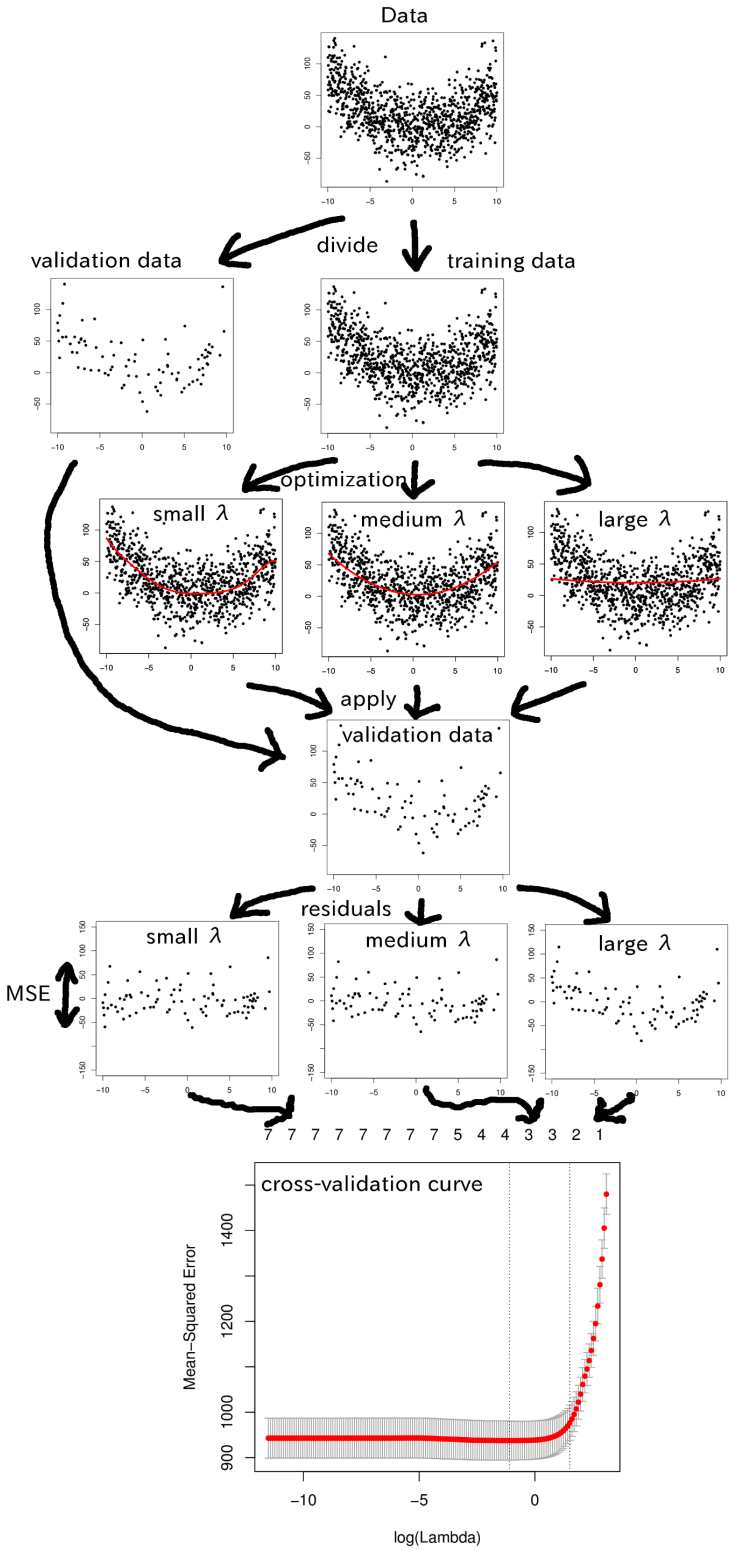

図 2: 交差検証の例。2 次の関係にある x, y のデータ。訓練データとテストデータに分け、訓練データに対して様々な λ でモデルを最適化し、そのモデルの予測精度をテストデータとの MSE で評価する。一番下の図は様々な λ に対する MSE の値。点線は MSE 最小をとる λ の値と、その標準誤差内に入る最大の λ の値。

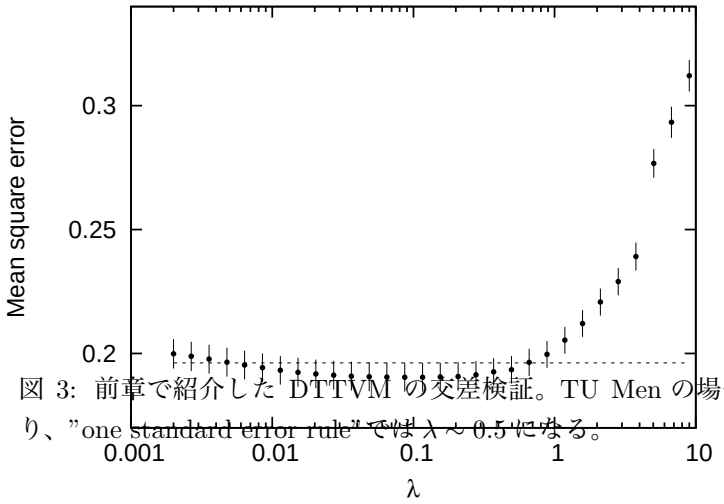

図 3: 前章で紹介した DTTVM の交差検証。TU Men の場合。$\lambda \sim 0.1$ で MSE 最小となり、”one standard error rule”では $\lambda \sim 0.5$ になる。

れて見えますが、中図で見える以上に細かい構造については、それらが真実の構造である確証はありません。データと交差検証からは「中図よりも細かい構造を加えても MSE に有意な差がない」ということが言えるわけです。

交差検証法はどのような問題にも使えるため便利です。一方で、最小 MSE には不定性があり、それは訓練用・テスト用のサブサンプルを生成する際に使用する乱数にも依存します。これらはデータ数が少ない場合により深刻な問題となり、試行ごとに最適な λ およびモデルが異なってしまうことがあります。このため、データ数が少ない場合は 1 つのデータだけをテスト用として、残りのデータを訓練用とする ”leave-one-out CV” が使われることもあります。また、”one standard-error rule”は慣習的なもので、閾値が ”one standard error”である必要はありません。hyper-parameter の推定やモデル選択については交差検証の他にも様々な指標を使うことができます。例えば階層ベイズ的に hyper-parameter の事後分布を推定したり、モデル選択理論における「情報量規準」を用いることもあります。「赤池情報量規準」（Akaike Information Criterion: AIC）は尤度 L とフリーパラメータの数 k を使って、$\mathrm{AIC} = -2\ln L + 2k$ と定義されます。AIC が最小になるモデルが、そのデータにとって予測誤差を最小にする最善のモデルとなります。一般に、AIC の方が交差検証よりも変数の多いモデルを選択する傾向があります。モデル選択については様々な研究が行われているので、詳しくは専門書をご覧ください。

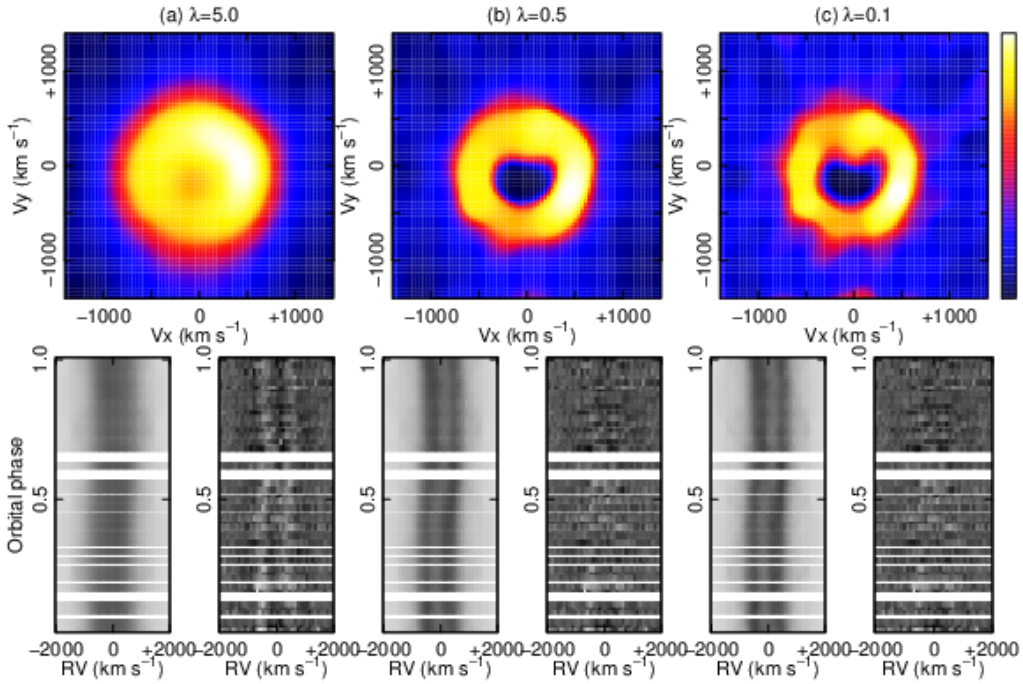

図 4: 矮新星 TU Men のデータを使って TVM によって得られたドップラーマップ（上段）とモデルスペクトル・観測との残差（下段）。3 組の結果は異なるパラメータ λ を用いて計算されたもの。