

Statistical Learning

Concepts + vocabulary to set up Modeling

Matthew McDonald

Statistical learning

Statistical learning: use data to learn a relationship between inputs (X) and an outcome (Y).

- Outcome / response / target: Y
- Inputs / features / predictors: $X = (X_1, X_2, \dots, X_p)$

Two reasons we build models

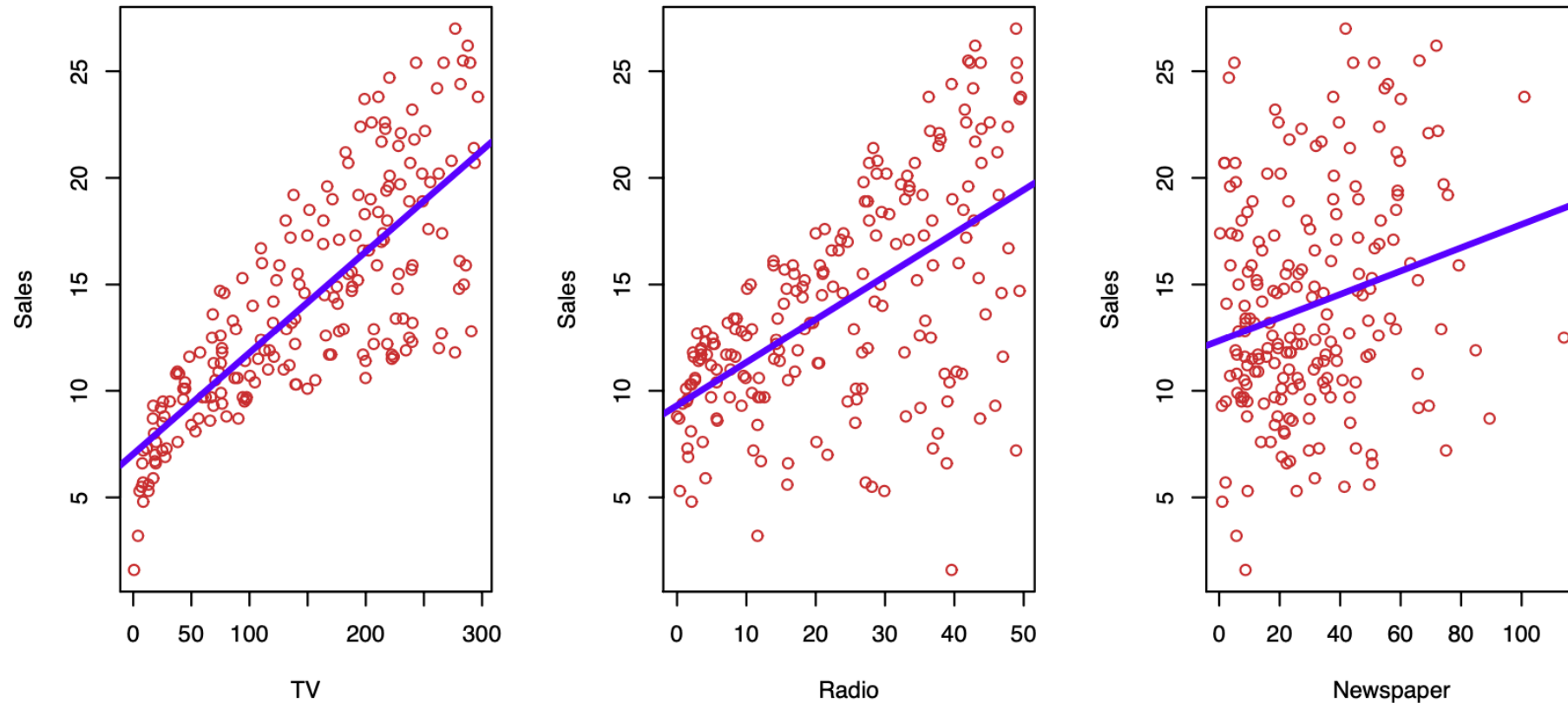
1) Prediction

- "What do I expect *next*?"
- Accuracy matters most

2) Inference (explanation / interpretation)

- "Which inputs matter, and how?"
- Clarity + assumptions matter

A helpful picture



- “Maybe we can do better than separate lines”
- Goal: learn $f(X)$ so $Y \approx f(X)$

The model-as-approximation view

We often write:

$$Y = f(X) + \varepsilon$$

- $f(\cdot)$: the *signal* / systematic component
- ε : noise, measurement error, missing variables (irreducible)

Supervised vs unsupervised

Supervised learning

- We observe **X** and **Y**
- Learn a rule to map $X \rightarrow Y$

Examples: - Regression (Y numeric) - Classification (Y category)

Unsupervised learning

- We observe **X** only
- Find structure: clusters, factors, embeddings, etc.

Regression vs classification

Regression

- Y is **quantitative**
- e.g., return, spread, volatility, loss given default

Classification

- Y is **qualitative**
- e.g., default / no-default, upgrade / downgrade, "risk-on" / "risk-off"

Parametric vs non-parametric (the big choice)

Parametric

- Assume a functional form
- Estimate a small number of parameters (e.g., linear regression)

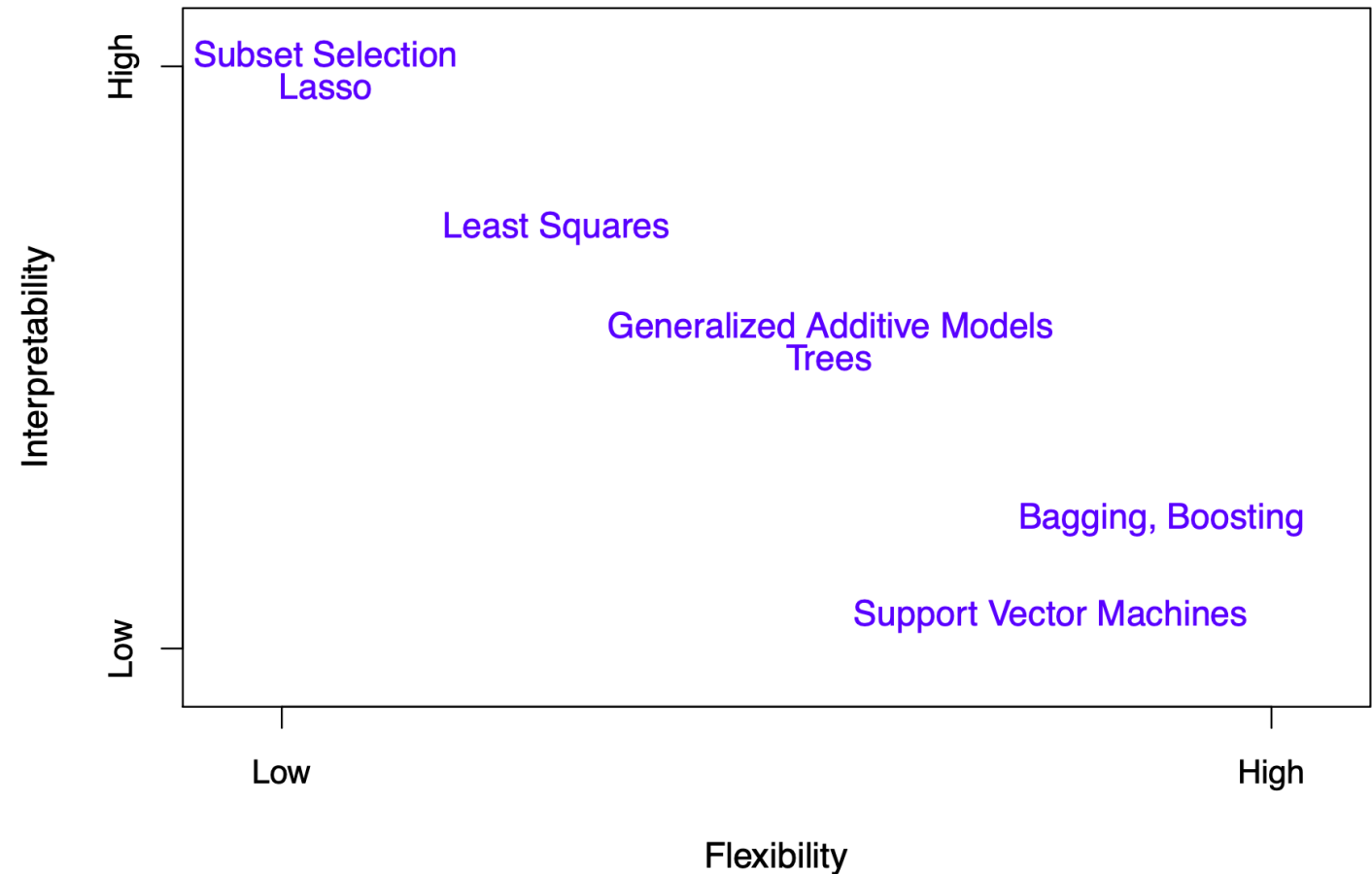
Non-parametric

- Fewer shape assumptions
- More flexible; can track complex patterns (but needs more data)

The real trade-off

Flexibility vs interpretability

- More flexible methods can fit more patterns
- But they become harder to explain
- And can overfit



A metric we'll use a lot: MSE

Mean Squared Error (MSE):

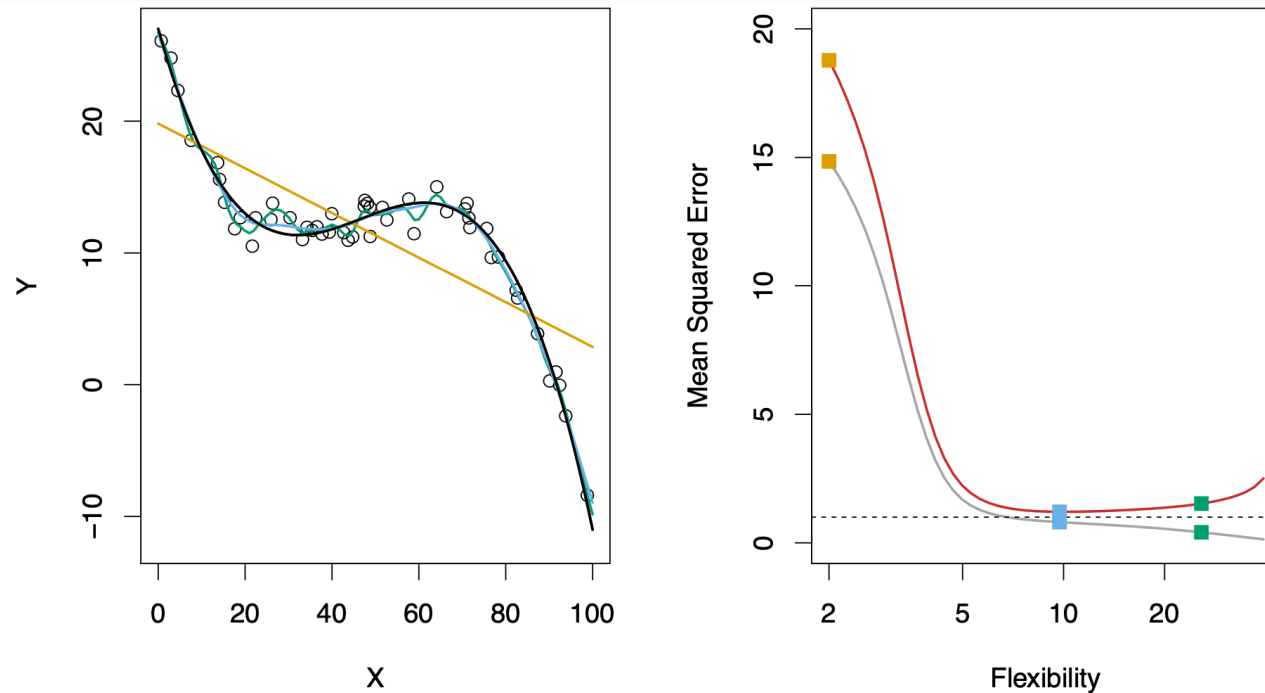
$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- Penalizes large errors more than small errors
- Convenient, common, and easy to compute
- Works naturally for regression
- **Units:** MSE is in **(units of Y)²**; RMSE is in the **same units as Y**

Overfitting

- A model can fit training data *too* well
- “Great in-sample” \neq “good out-of-sample”

Flexibility vs Performance



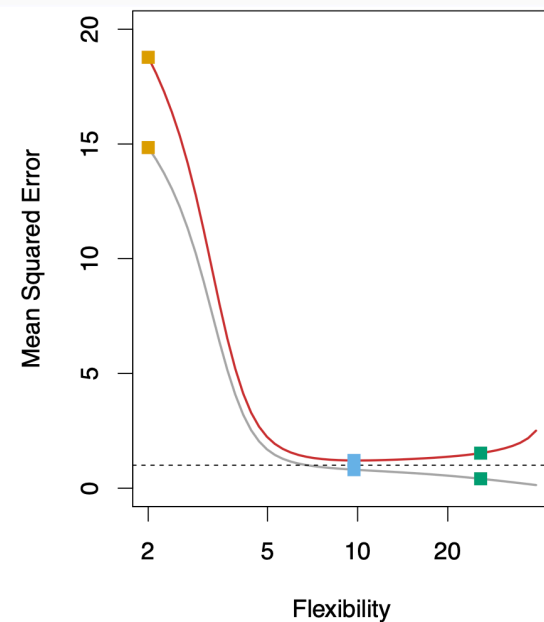
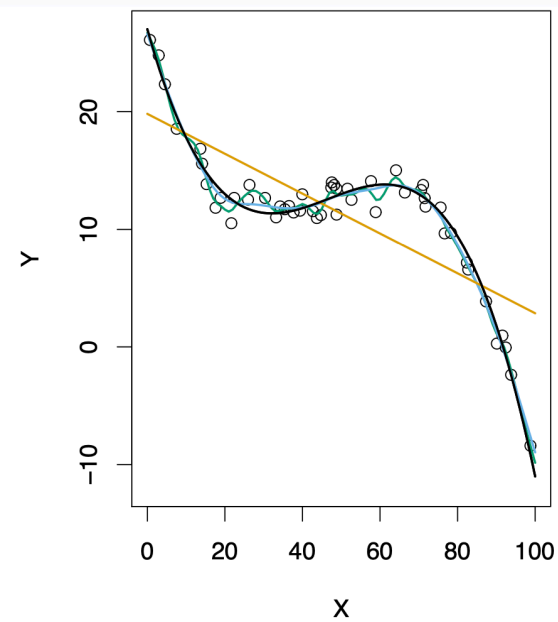
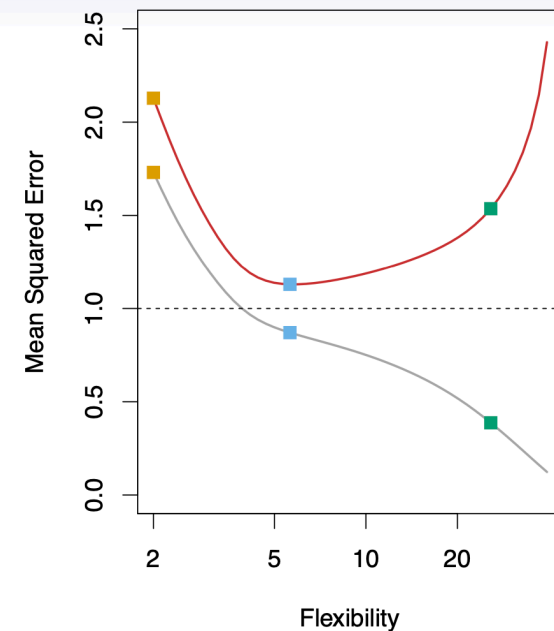
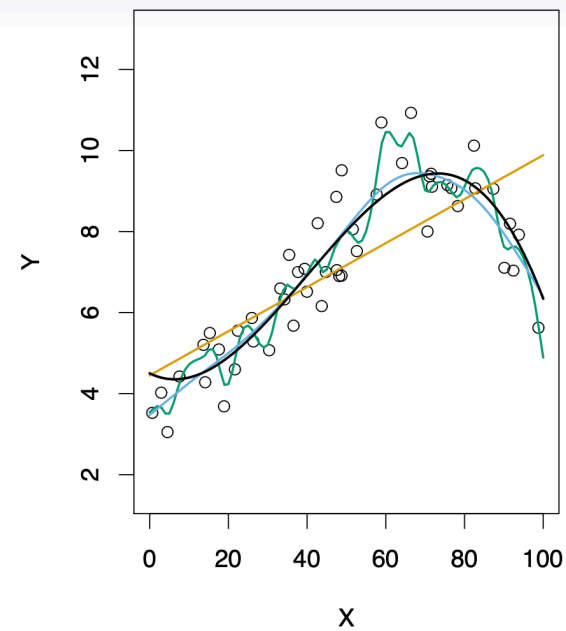
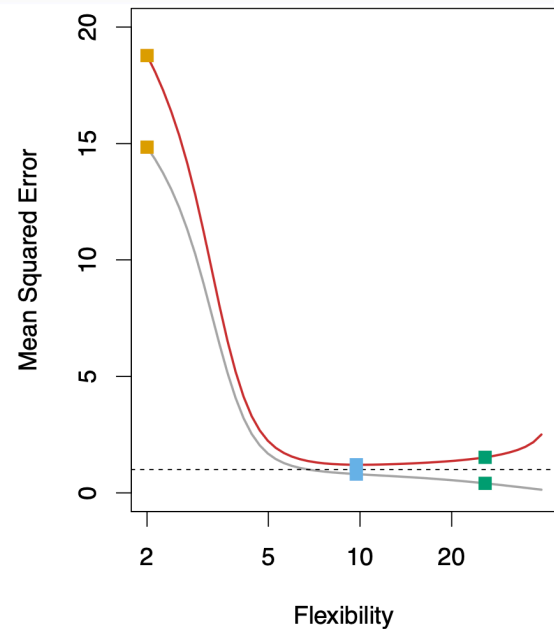
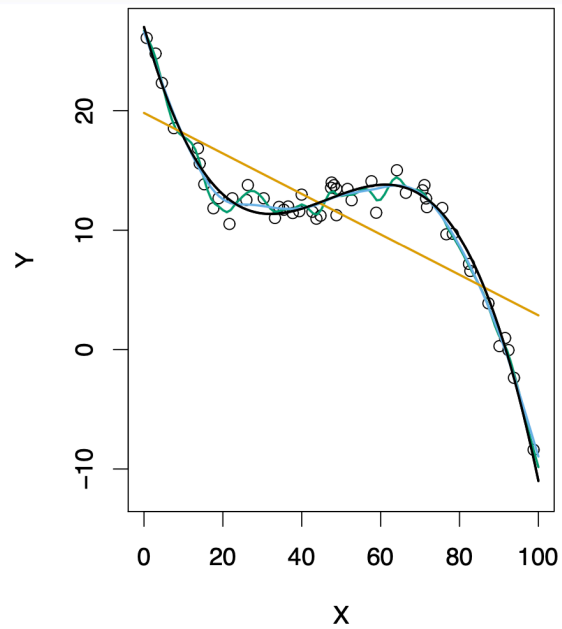
Left panel: data + fits

- points = observed data
- straight line = **rigid** (high bias)
- smoother curves = **more flexible**
- very wiggly = **overfit** (high variance)

Right panel: error vs flexibility

- gray = **training (in-sample)** error
- ↓
- red = **test (out-of-sample)** error is often **U-shaped**
 - choose flexibility near the **red minimum**

Three Examples



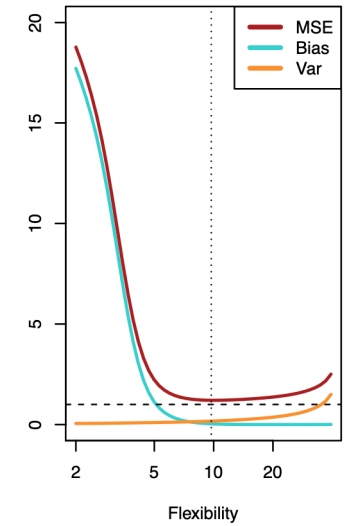
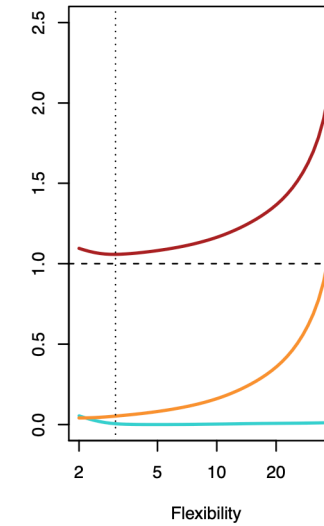
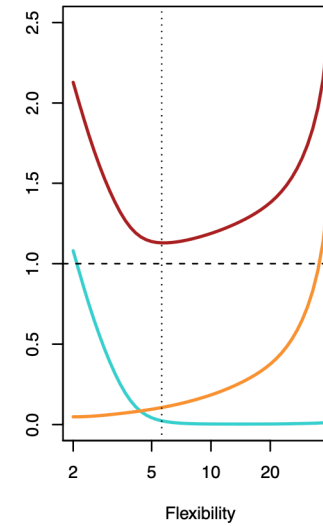
Bias–variance trade-off (intuition)

Test error reflects:

- **Bias**: too rigid / systematically wrong
- **Variance**: too wiggly / too sensitive to sample

As flexibility increases:

- bias typically ↓
- variance typically ↑



Choosing flexibility: the practical punchline

- There is no “best model” in the abstract
- The best model depends on:
 - the true relationship (unknown)
 - noise level
 - sample size
 - purpose (prediction vs inference)
- **We choose flexibility using out-of-sample thinking**

Classification: what changes?

- Output is a **label**
- We evaluate with metrics like:
 - misclassification rate / accuracy
 - later: precision/recall, ROC/AUC