- Problems with the approach we just saw:
  - We used all observations for fitting the model
  - We do not know if the model would generalize to other similar datasets?
  - If we have thousands of independent variables (X), do we include all variables in the model to forecast y?
- Before we proceed with forecasting techniques, it is very important to first understand the concepts of *model selection & model over-fitting*

Model Selection

- Given a set of candidate models, how do we select the most suitable model?
- One option, use Likelihood Ratio (LR) test, which is based on comparing the likelihood of two competing (nested) models:

$$\text{LR statistic} = 2\left(\log(L|\hat{\theta}_{MoreParas}) - \log(L|\hat{\theta}_{LessParas})\right)$$

Note that higher likelihood signifies better model fit.

- Problem with the LR test is that it does not take into account the complexity of the model
- Overly complex models result in what is called 'model overfitting'

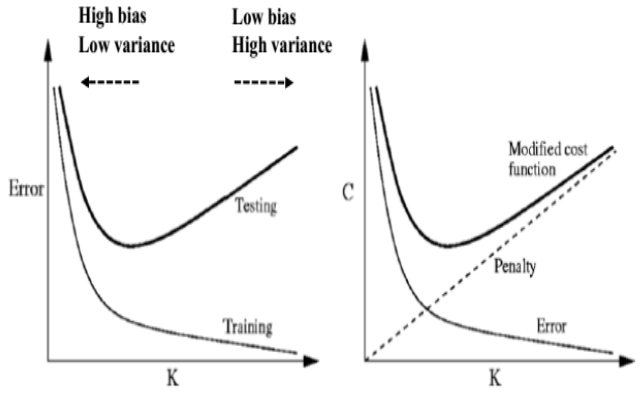
Model Overfitting

- Most real-life time series can be considered as the sum of a signal from a dynamical process (true underlying data generating process) plus some observational noise. $y_t = signal_t + noise_t$

- An excessively complex model, comprising too many parameters relative to the number of training observations, will tend to learn the noise along with the underlying signal structure, and hence, it will not be flexible enough to adapt to other similar datasets.

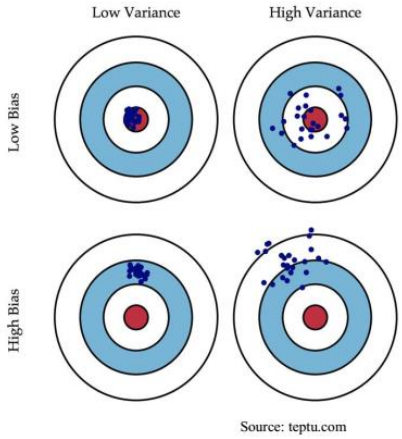
- This will result in **good in-sample fit** (low *bias*), but **poor generalization** (high *variance*).

Model Overfitting



- **Model overfitting** – a good fit on the in-sample data (or training data) does not translate into low prediction error on the out-of-sample data (or test data).
- **Solution** – trade-off between *bias* (goodness of fit) and *variance* (model's generalizability).

Graphical illustration of bias and variance



- **Bias** – quantifies the degree of similarity between the true observation (y) and the model response ($\hat{f}(X)$)
- **Variance** – quantifies the uncertainty in the model response
- Both depend on the model complexity

Bias variance decomposition

- Expressing the squared-error loss as a sum of bias and variance is known as the *bias-variance decomposition*

- For the true underlying data generating equation be given by:

$$y = f(X) + \varepsilon \text{, where } \varepsilon \sim IID(0,\sigma^2),$$

we perform a regression fit to obtain an approximate function $\hat{f}(X)$

- The expected (E) mean square error (MSE) can be computed as:

$$\begin{aligned} E[MSE(X)] &= E[(y - \hat{f}(X))^2] \\ &= E[(y - f(X) + f(X) - \hat{f}(X))^2] \\ &= \sigma^2 + [(E\hat{f}(X) - y)^2] + E[\hat{f}(X) - E\hat{f}(X))^2] \\ &= Irreducible\ error + Bias^2 + Variance \end{aligned}$$

Information criteria (IC)

- IC are commonly used to select a model from a set of competing models
- As opposed to LR test, IC selects a model by estimating the trade-off between bias and variance
- The rationale behind IC is to quantify the acceptable loss in predictive accuracy between different models for the gain in simplicity
- Given M competing models, a time series for length N, IC for a model comprising k number of parameters can be computed as:

$$IC = -\log L(\theta) + g(N,k)$$

where $L(\theta)$ is the likelihood function, θ is the parameter vector, and $g(N,k)$ is the penalty function

- Higher likelihood (or lower negative likelihood) signifies better fit, while an increase in k results in a higher penalty. Models with lower IC are preferable

Occam's Razor

- Occam's razor is a principle of parsimony attributed to the 14th-century English logician and Franciscan friar William of Ockham
- The principle states that a theory should rely on as few assumptions as possible, eliminating those that make no difference to the observable predictions of the theory
- Given multiple competing theories that are equally plausible, the principle of Occam's Razor suggests shaving of the unnecessary assumptions

Model parsimony

- While increasing the complexity of a model naturally gives more freedom to provide a better fit to the observations (resulting in low bias), a model with too many parameters will not distinguish between the underlying dynamics that we wish to extract and fluctuations due to factors such as measurement errors, non-stationarity and noise (resulting in high variance)

- We should aim to identify the simplest model that is compatible with the observations

- This provides motivation for seeking a parsimonious model (one with as few parameters as possible)

Akaike Information Criteria (AIC)

- Akaike (1974) proposed the AIC, which is a commonly used measure for model selection (14k citations, 73rd most cited paper)
- AIC rewards the model for goodness of fit and penalizes the number of free parameters employed in achieving that fit. AIC is computed as:

$$AIC = -2\ln L + 2k$$

where L is the maximized value of the likelihood for a model with k number of parameters

- For normally and independently distributed prediction errors, the AIC can be expressed as:

$$AIC = N\left\{\ln\left(2\pi\frac{RSS}{N}\right)+1\right\}+2k$$

where RSS is the residual sum of squares with N observations

- Lower AIC values are preferable
- AIC is useful in comparing competing models, but it does not tell anything about the suitability of the model for a given time series in the absolute sense

Bayesian Information Criteria (BIC)

- Schwarz (1978) proposed the Schwarz IC or BIC
- BIC is a measure of goodness of fit of a model:

$$BIC = -2\ln L + k\ln(N)$$

where L is the maximized value of the likelihood for the model with k number of parameters using a time series with N observations

- The BIC imposes a stronger penalty on the number of parameters, as compared to the AIC. Lower BIC values are preferable
- For normally and independently distributed prediction errors, the BIC can be expressed as:

$$BIC = N\left\{\ln\left(2\pi\frac{RSS}{N}\right)+1\right\}+k\ln(N)$$

LASSO

- Tibsharani (1996) proposed the least absolute shrinkage and selection operator (LASSO)
- LASSO employs a L_1 regularization scheme, and is commonly used for input variable selection
- Selecting only the most relevant subset of input variables improves the model's predictive accuracy and enhances interpretability
- Given a set of k input variables, the LASSO fits a linear model

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \ldots + \alpha_k x_{ik} + \varepsilon_t$$

whereby the model parameters are estimated using the following criterion

$$min_{\alpha_0,\alpha}\sum_{i=1}^{N}\left(y_i - \alpha_0 - \sum_{j=1}^{k}\alpha_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{k}|\alpha_j|$$

- The LASSO problem is equivalent to minimizing

$$\sum_{j=1}^{k}|\alpha_j| \le z$$

- As opposed to LASSO, which uses a L_1 penalty, Ridge regression uses a L_2 regularization scheme, whereas elastic nets use a combination of L_1 and L_2 regularization schemes

Model selection – Summary

- To ensure that good in-sample fit translates to accurate out-of-sample forecasts, it is imperative to try and avoid model overfitting, LR tests compare models based on goodness of in-sample fit, but ignore complexity
- More complex models (over parametrized models) may give good in-sample fit (low bias), but poor out-of-sample forecasts (high variance) due to potential overfitting (learning noise)
- To avoid overfitting, we seek a parsimonious model
- ICs try to achieve a trade-off between model bias and variance (for e.g., AIC, BIC)
- Regularization techniques have also been commonly used for variable selection as they enhance interpretability (for e.g., LASSO – L_1 norm, Ridge – L_2 norm, Elastic nets – L_1 and L_2 norm)

- Need for time series forecasting (making informed decisions, predicting extreme events)
- Different modelling strategies (linear vs nonlinear etc.)
- Regression analysis (modelling relationship between two variables)
- Importance of achieving a trade-off between Bias and Variance (model overfitting, bias variance decomposition, Occam's razor, principal of parsimony)
- Select a suitable model using IC (AIC/BIC) and regularization techniques (lasso, ridge regression, elastic nets)