Key formulas

- **Asymptotic distribution of ML estimator**
  - $\theta_{ML} \sim N(\theta_0, I(\theta_0)^{-1})$ as $n \to \infty$
  - where $I(\theta_0) = -\sum_{i=1}^{n} \frac{\partial^2 I(\theta_0; y_i)}{\partial\theta_0 \partial\theta_0'}$
- **KL Divergence**
  - $D_{KL} = \int_{-\infty}^{\infty} g(y) log \frac{g(y)}{p(y)} dy$
- **cross-entropy**
  - $\sum[-y_i log(p_i) - (1 - y_i) log(1 - p_i)]$
- **variance and bias**
  - $E[y - \hat{y}]^2 = E[\hat{y} - E[\hat{y}]]^2 + [E[\hat{y}] - E[y|x]]^2 + E[y - E[y|x]]^2$
  - $E[y - \hat{y}]^2 = Var[\hat{y}] + [E[\hat{y}] - E[y|x]]^2 + Var[y|x]$
  - $E[y - \hat{y}]^2 = Var[\hat{y}] + [bias(\hat{y})]^2 + Var[y|x]$
- **variance-bias trade-off**
  - more richly parametrized models (splines) and more flexible (KNN with small K) models, or smaller dataset (less bad periods), have:
    - **higher variance**
    - **lower bias**

Financial Series

- Low signal-to-noise ratio
  - even qualied humans can't use x explained most y
- Fat tails
- high persistence
  - can be thought of as a smaller number of independent observations
- Markets are dynamic
  - economy/institutions/ideas all change
  - long period data is not always good
  - central banks control interest rate more tightly (more persistant)
- substantial measure error
  - (GDP, ination, earnings, book value, debt quantity and quality)
  - Measurement errors tend to penalize complex models more severely

Maximum Log Likelihood

- good at: unbiased and efficient
- but rely on model assumption correct: errors be iid and normal (or student-t)
- gaussian errors: equal to OLS
- student t errors
  - $\sigma_t^2 = \sigma_g^2 \frac{v}{v-2}$
  - v small: great differece
  - large sample: OLS (gaussian errors) better point forecast (not full distribution)
  - small sample: OLS (gaussian errors) high variance and tail risk
- **Asymptotic distribution of ML estimator**
  - $\theta_{ML} \sim N(\theta_0, I(\theta_0)^{-1})$ as $n \to \infty$
  - where $I(\theta_0) = -\sum_{i=1}^{n} \frac{\partial^2 I(\theta_0; y_i)}{\partial\theta_0 \partial\theta_0'}$
- when misspecied models
  - minimizes the KL Divergence
    - $D_{KL} = \int_{-\infty}^{\infty} g(y) log \frac{g(y)}{p(y)} dy$
    - measure the information lost, still sensible
- modications of ML
  - Penalization
    - the penalization introduces some bias but reduces the variance
    - penalization disappear asymptotically
  - validation
    - a natural loss function: minus log-likelihood.

Loss function of classication problems

- why not error rate loss function
  - one category has a small average probability (default)
  - low signal-to-noise rate
- cross-entropy
  - loss function = $\sum[-y_i log(p_i) - (1-y_i) log(1-p_i)]$
- alternative
  - use a loss function close to the one that will be used when the model is deployed

Variance & Bias

- MSE into variance and bias
- $E[y-\hat{y}]^2 = E[\hat{y}-E[\hat{y}]]^2 + [E[\hat{y}] - E[y|x]]^2 + E[y-E[y|x]]^2$
- $E[y-\hat{y}]^2 = Var[\hat{y}] + [E[\hat{y}] - E[y|x]]^2 + Var[y|x]$
  - they are "variance", "bias" and "true error"
  - $E[\hat{y}]$ is point forecast of y given x
  - y is **really real y**, $E[y|x]$ is observed y
  - variance: 预测值的波动
  - bias: 模型误差
  - true error: 观测误差 or 真・随机扰动
- variance-bias trade-off
  - more richly parametrized models (splines) and more flexible (KNN with small K) models, or smaller dataset (less bad periods), have:
    - **higher variance**
    - **lower bias**

Linear model and KNN

- linear
  - OLS gives highest R2
  - linear projection has lowest MSE if model correct and error is NIID
    - $y_i^* = \alpha' x_i$
  - OLS is an asymptotically consistent estimator of the linear projection
- Comparison
  - OLS make strong assumptions about the form
  - KNN alternative and more flexible approach. (bias at the boundaries)
  - linear better:
    - if the linear is close to the true form.
    - high dimensions
  - **KNN better:**
    - **non-linearty, perhaps interation effects**
    - **high signal-to-noise ratio**
    - **large sample size**
    - **low dimension**
- use both
  - local regression

linear estimator

- estimator is linear if it can be expressed as a linear function of y

nonlinear estimator

- student t errors, min MAD estimators, logit model

Cross-Validation

- Leave-One-Out CV
  - unbiased estimates of the test error
  - higher variance and less bias (than K-Fold)
- one-standard-error rule
  - best model in test set with one standard error bound, select simplest within this bound
- Remarks on CV
  - nonparametric: makes no assumption about the models
  - general: can be applied to any model, with any loss function (MSE, likelihood or profit)
  - intuitive
- Cross validation on time series
  - problem with assumption: parameter stability
  - expanding window: training set becomes progressively larger

WLS and local regression

- WLS
  - when errors are very far from iid and we have good weights
- Local regression
  - capture additive nonlinearities and interaction effects (with interpretability)
    - Cross-products create leverage points (even gaussian)
  - similar to the "varying parameter model"
- *Gaussian kernel*
  - λ determines the amount of smoothing, replace k in KNN
  - as p (used in computing distance and weights) grows, all near the boundary, method breaks down
- *Exponential smoothing*
  - lower variance and higher bias
  - the distance is only relevant to compute the ordering

Splines

- Polynomial functions (are global)
  - very high variance particularly near the boundaries
- Dummy variables (are local)
  - have lower variance but have jumps
- Linear splines
  - In low signal-to-noise environments, linear splines may be more robust
- Polynomial splines
  - m = 3 is common (**cubic regression splines**)
- Natural cubic splines:
  - required to be linear at the boundary
  - good if K is large
- penalization for knots
  - Ridge if the function is smooth.
  - Lasso if the function changes rapidly in some parts and slowly in others

Generalized Additive Models (GAMs)

- Advantages:
  1. avoids curse of dimension
  2. high interpretability
- limitation
  1. restricted to be additive
  2. curse of dimension for interaction effects

interaction effects:

1. Use regression splines on standard interaction effects
   - GAM model on x1, x2, z=x1*x2
2. Interact basis expansions (tensor products)
   - Interact all the terms of the basis expansions
3. Local GAMs
   - local + splines
   - The local captures interaction effects
   - splines allows larger p and K
   - pick local variables (1-2 variables compute distance):
     - thought to drive the interaction effect
     - or use CV to pick

Ridge & Lasso

- ridge regression
  - $Min_{(\beta)}[(y - X\beta)'(y - X\beta) + \beta'\Lambda\beta]$
  - $\hat{\beta}^{ridge}(\lambda) = (X'X + \Lambda)^{-1}X'y$
  - Covariates should be standardized (also for PCR)
  - reducing over-fit
  - performs well when many covariates are highly correlated (than lasso)
- The Lasso
  - penalize term is absolute beta
  - set some coefficients to zero
- advantange and disadvantage
  - improve performance
    - large p
    - low signal-to-noise ratio
    - large "measurement" error
    - regression splines
  - perform poorly
    - covariates strongly positively correlated (bad forLasso)
    - only a small percentage actually matter

PCR

- Dimension reduction
  - reduce dimensiona while minimizing the loss of information
- Principal Components Regression
  - reduce variance while bias increases slightly
  - (eigen value)^2/N is the variance of factor
  - PCR do well when the first few principal components are sufficient
- PCR and Ridge
  - Ridge regression can be seen as a smooth version of principal components analysis
  - df of ridge: $\sum d_j^2/(d_j^2 + \lambda)$

Bootstrap and Bagging

- takes independent draws (with reimmission)
    - nonparametric
- when improves most
    - non-linear in parameters
    - different bootstrap samples have high variance
    - likelihood is multimodal
- **Block bootstrap**
    - In time series
        - data is not independent
        - highly persistent
        - not iid
    - Its a nonparametric bootstrap
        - the data is split into contiguous blocks of equal size M
        - draw blocks
        - y*{m+i} and y*{i} are nearly independent

subset selection

- cross-validated prediction error
    - AIC or BIC (small is good)
    - adjusted R2 (large is good)
- assumptions:
    1. Model form is correct
    2. The errors are iid
- requiring:
    1. parametric model
    2. p is "obvious"
    3. loss function is mse