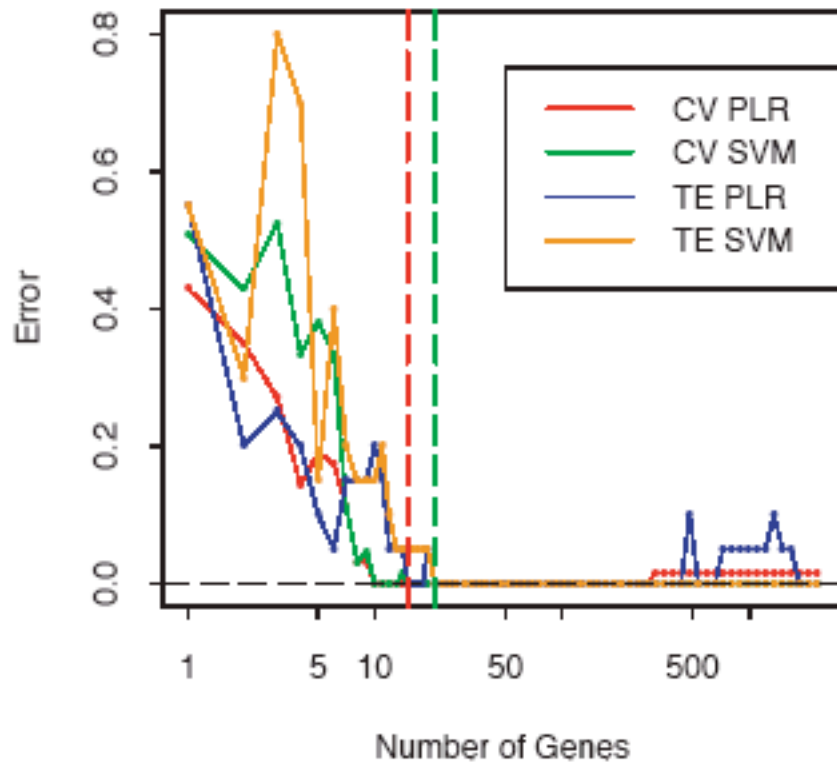


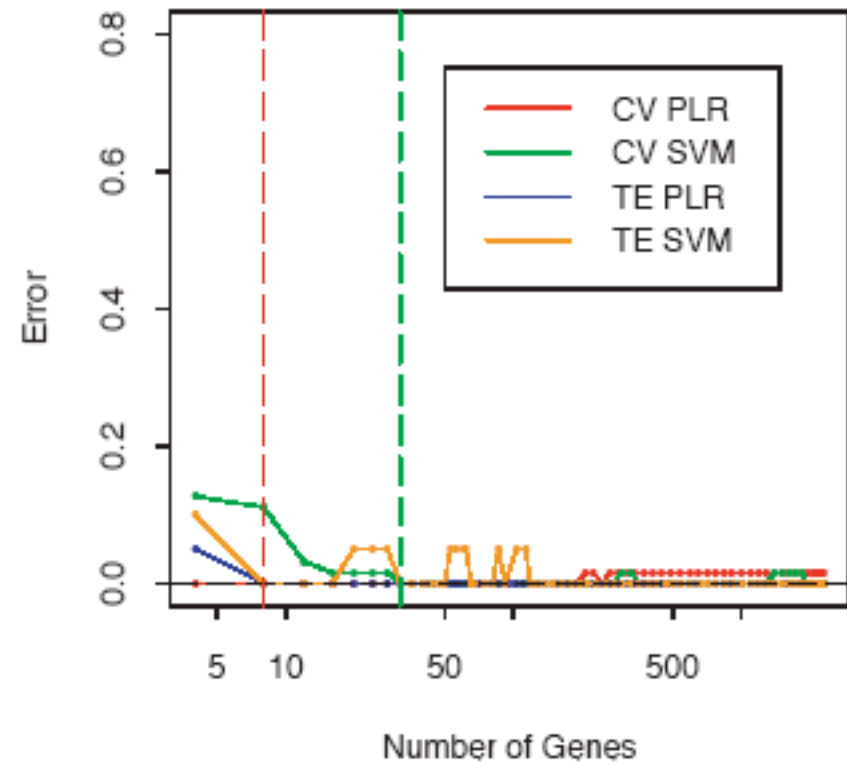
Model Selection

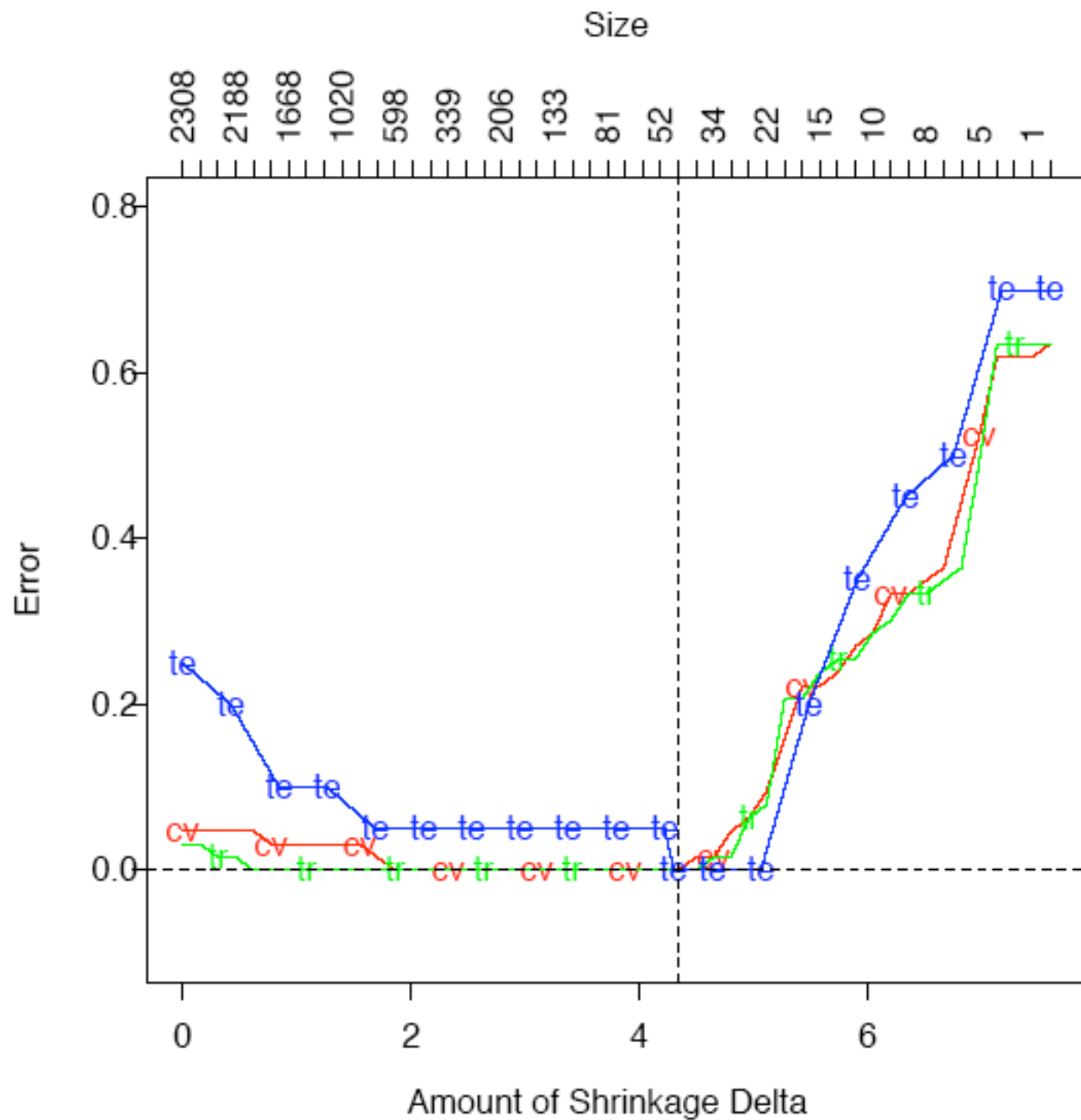
- Bias, Variance, Complexity
- Model Selection Criteria
- Cross-Validation

Use UR



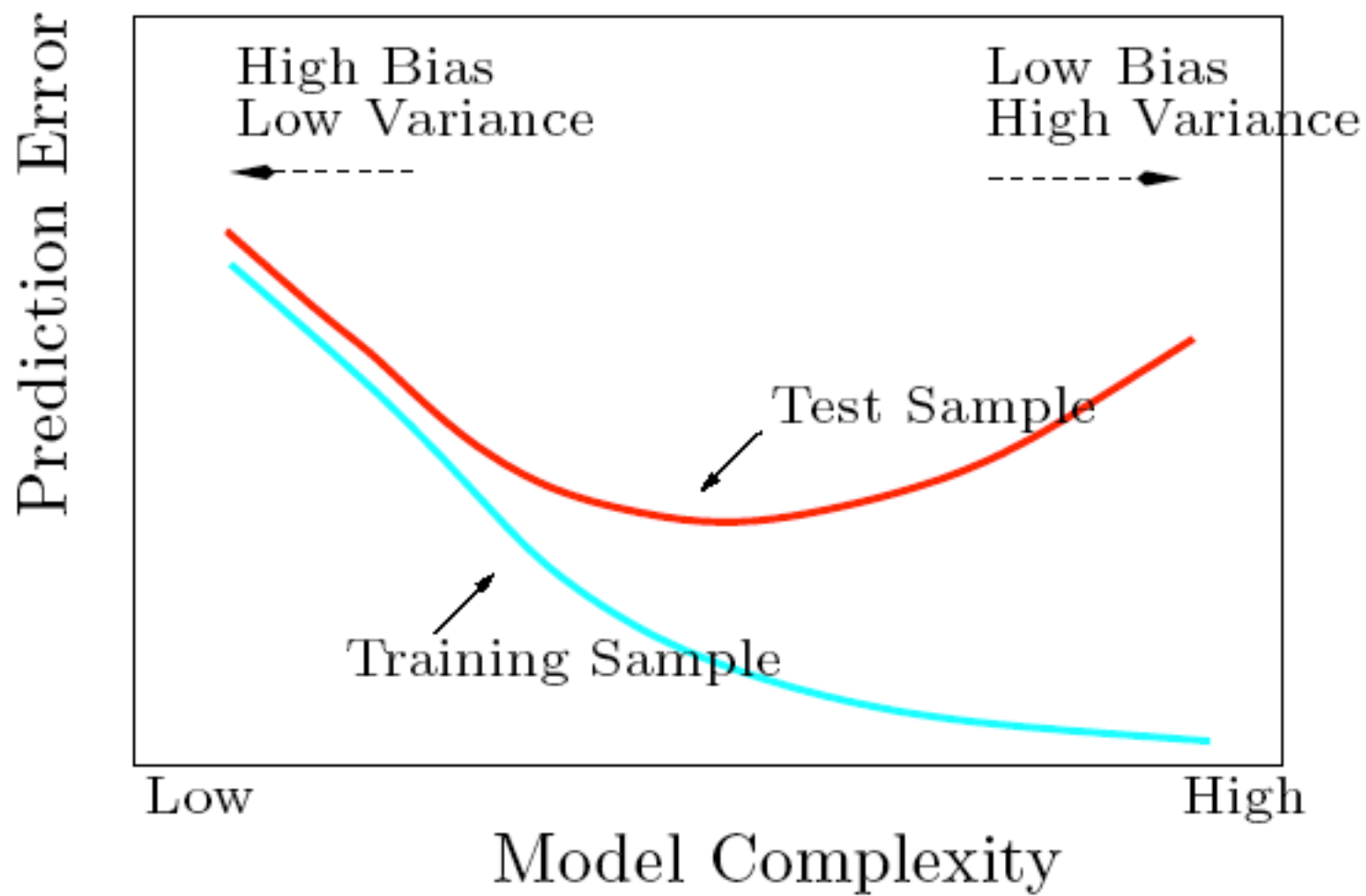
Use RFE





Selection & Assessment

- *Generalization* performance of a model pertains to its predictive ability on *independent* test data.
- Crucial for model choice and quality evaluation.
- These represent distinct goals:
 - *Selection*: determine performance of a series of competing models in order to pick best.
 - *Assessment*: having chosen a best model, estimate its prediction error on new data.
- Numerous criteria, strategies.



Bias - Variance Tradeoff

Outcome Y (assume continuous); input vector X ;
prediction model $\hat{f}(X)$.

$L(Y, \hat{f}(X))$: loss function for measuring errors
between Y and $\hat{f}(X)$. Common choices are:

$$L(Y, \hat{f}(X)) = \begin{cases} (Y - \hat{f}(X))^2 & \text{squared error} \\ |Y - \hat{f}(X)| & \text{absolute error} \end{cases}$$

Test or generalization error: expected prediction error over independent test sample

$\text{Err} = \mathbb{E}[L(Y, \hat{f}(X))]$ where X, Y drawn randomly from their joint distribution.

Training error: average loss over training sample:

$$\overline{\text{err}} = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}(x_i))$$

Typically, training error $<$ test error because same data is being used for fitting and error assessment.

Fitting methods usually adapt to training data so $\overline{\text{err}}$ overly *optimistic* estimate of Err.

Part of discrepancy due to where evaluation points occur. To assess optimism use *in-sample* error:

$$\text{Err}_{\text{in}} = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{Y^{\text{new}}} [L(Y_i^{\text{new}}, \hat{f}(x_i))]$$

Interest is in test or in-sample error of \hat{f}

\Rightarrow Optimal model minimizes these.

Assume $Y = f(X) + \epsilon$, $E(\epsilon) = 0$, $\text{Var}(\epsilon) = \sigma_\epsilon^2$.

Expected prediction error of fit $\hat{f}(X)$ at input point $X = x_0$ under squared error loss:

$$\begin{aligned}\text{Err}(x_0) &= E[(Y - \hat{f}(x_0))^2 | X = x_0] \\ &= \sigma_\epsilon^2 + [E\hat{f}(x_0) - f(x_0)]^2 + \\ &\quad E[\hat{f}(x_0) - E\hat{f}(x_0)]^2 \\ &= \sigma_\epsilon^2 + \text{Bias}^2(\hat{f}(x_0)) + \text{Var}(\hat{f}(x_0)) \\ &= \text{Irreducible Error} + \text{Bias}^2 + \text{Variance}.\end{aligned}$$

First term: variance of the outcome around its true mean $f(x_0)$; unavoidable.

Second term: squared bias – amount by which average of estimate $\hat{f}(x_0)$ differs from true mean.

Third term: variance – expected squared deviation of estimate around its mean.

Training Error Optimism

Training error typically less than true error.

Define the *optimism* as $op \equiv \text{Err}_{\text{in}} - \mathbb{E}(\overline{\text{err}})$.

For squared error and other loss functions have

$$op = \frac{2}{N} \sum_{i=1}^N \text{Cov}(\hat{y}_i, y_i)$$

\Rightarrow the amount by which $\overline{\text{err}}$ underestimates the true error depends on how strongly y_i affects its own prediction. The harder we fit the data, the greater $\text{Cov}(\hat{y}_i, y_i)$, thus increasing the optimism.

If \hat{y}_i is from a linear fit with p covariates

$$\text{Cov}(\hat{y}_i, y_i) = p\sigma_\epsilon^2$$

so
$$\text{Err}_{\text{in}} = \text{E}(\overline{\text{err}}) + 2 \cdot \frac{p}{N} \sigma_\epsilon^2$$

Prediction Error Estimation

General form of in-sample estimates is

$$\widehat{\text{Err}}_{\text{in}} = \overline{\text{err}} + \widehat{\text{op}}.$$

Applying to linear model with p parameters fit under squared error loss gives the C_p statistic:

$$C_p = \overline{\text{err}} + 2 \cdot \frac{p}{N} \hat{\sigma}_\epsilon^2.$$

Here $\hat{\sigma}_\epsilon^2$ is an estimate of the error variance obtained from a low-bias (large) model. Under this criterion we adjust the training error by a factor proportional to the number of covariates used.

Akaike Information Criterion is a generalization to situation where a log-likelihood loss function is used, e.g., binary, Poisson regression.

Criterion Selection Functions

Generic form for AIC is $AIC = -2 \cdot \text{loglik} + 2 \cdot p$

Bayes information criterion (BIC) is

$$BIC = -2 \cdot \text{loglik} + \log N \cdot p$$

For $N > e^2 \approx 7.4$, BIC penalty $>$ AIC penalty

\Rightarrow BIC favors simpler models.

Many variants; new feature – adaptive penalties.

When log-lik based on normal distⁿ we require an estimate for σ_{ϵ}^2 . Typically obtained as mean squared error of low-bias model \Rightarrow problematic for $p > n$ settings: *microarrays*.

Cross-validation does not require this.

Model Complexity

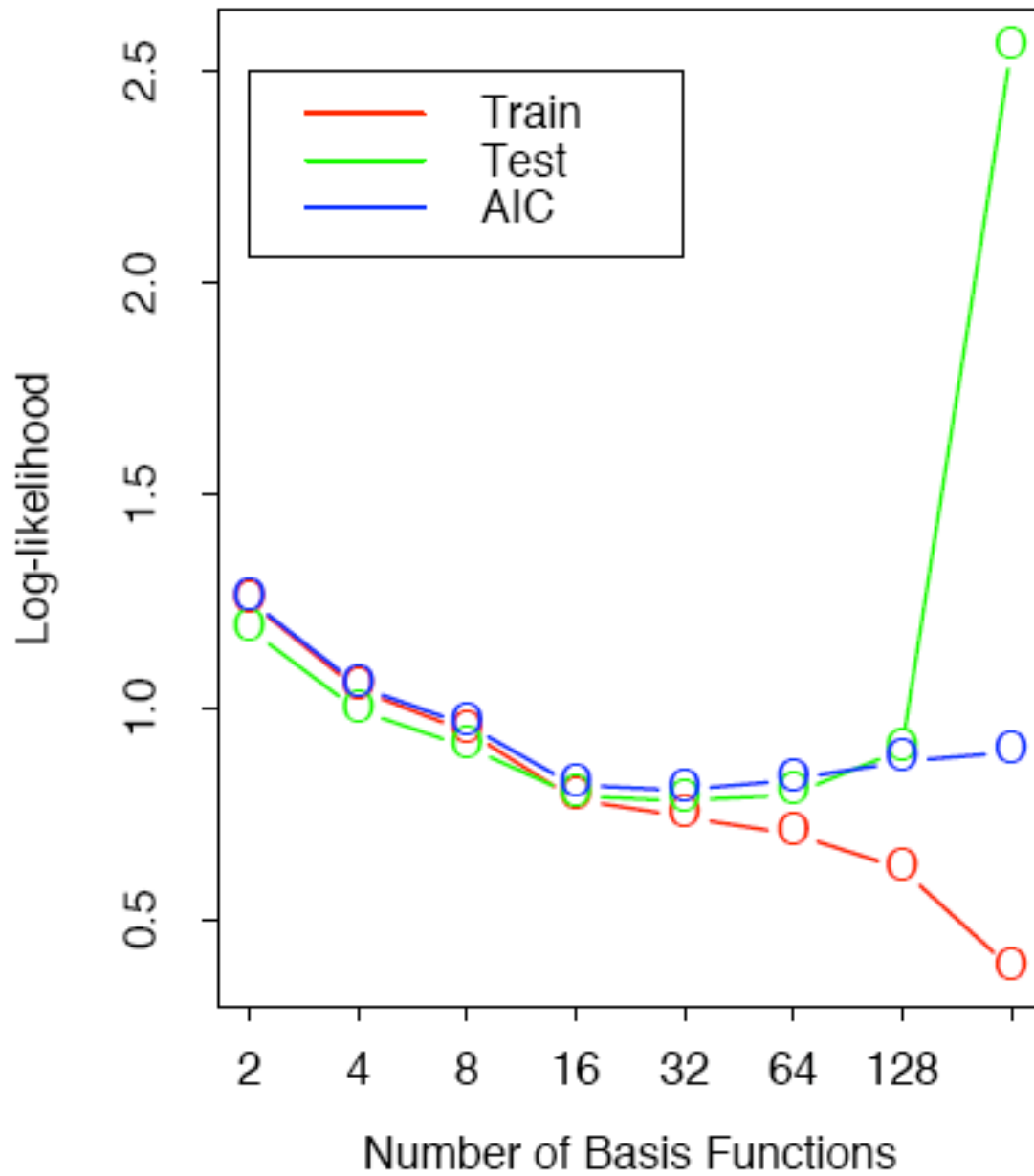
AIC, BIC and other model selection criteria include an optimism estimate (or penalty term) that involves the number of model parameters p .

If covariates are selected *adaptively* then no longer have $\text{Cov}(\hat{y}_i, y_i) = p\sigma_\epsilon^2$: if we *select* the best-fitting model with $q < p$ covariates, then $\text{Cov}(\hat{y}_i, y_i) > q\sigma_\epsilon^2$ and the *effective number of parameters* is $> q$.

Linear fitting methods: $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$ where \mathbf{S} depends only on x_i (not y_i). Includes linear regression and methods using quadratic penalties such as ridge regression and cubic smoothing splines. Define enp as $\text{trace}(\mathbf{S})$.

More complex fitting methods: enp via permutation.

Log-likelihood Loss



Cross-Validation

Simplest method for estimating prediction error.

Estimates *extra*-sample error $\text{Err} = \mathbb{E}[L(Y, \hat{f}(X))]$.

With enough data (large N) set aside portion as validation set. Use to assess model performance.

Not feasible with small $N \Rightarrow$ CV offers a finesse.

Randomly partition data into K equal-sized parts. For k th part, fit model to other $K - 1$ parts. Then calculate prediction error of resultant model when applied to k th part. Do this for $k = 1, \dots, K$ and combine the prediction error estimates.

Let $\kappa : \{1, \dots, N\} \mapsto 1, \dots, K$ map observations to their assigned partition. Let $\hat{f}^{-k}(x)$ denote fitted function with k th part removed.

Then CV prediction error estimate is

$$\text{CV} = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}^{-\kappa(i)}(x_i)).$$

Given a set of models $f(x, \alpha)$ indexed by tuning parameter α (SVMs, shrunken centroids) set

$$\text{CV}(\alpha) = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}^{-\kappa(i)}(x_i, \alpha)).$$

Find $\hat{\alpha}$ minimizing $\text{CV}(\alpha)$ and fit chosen model $f(x, \hat{\alpha})$ to all the data. CAUTION.

$K = N$: *leave-one-out* CV – approx unbiased for true prediction error but can be highly variable.

$K = 5$: lower variance but bias can be a problem.

Generally $K = 5$ or 10 recommended but clearly depends on $N \Rightarrow$ microarray applications??

