Variance-bias tradeoff

Miaoyan Wang

Department of Statistics
UW Madison

Purposes of Model Selection

- Recall a multiple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i, \quad \varepsilon_i \sim \text{iid } N(0, \sigma^2),$$

  can any of the $p-1$ explanatory variables be dropped to simplify the model?
- If the purpose is description/explanation/understanding, then
  - Parsimony is a key idea.
  - **Occam's razor**: All things being equal, the simplest solution tends to be the right one.
- If the purpose is prediction, then
  - Models are evaluated by predictive accuracy/power.

Bias-variance tradeoff

- Arguably, the goal of a regression analysis is to “build” a model that predicts well.
- One way to measure this performance is in the prediction mean squared error of the model

$$\begin{aligned}
\mathrm{MSE}_{pred}(\mathcal{M}) &= \mathbb{E}\left(Y_{\mathrm{new}} - \left(\hat{\beta}_0 + \sum_{j=1}^{p-1} \hat{\beta}_j X_{\mathrm{new},j}\right)\right)^2 \\
&= \mathrm{Var}(Y_{\mathrm{new}} - (\hat{\beta}_0 + \sum_{j=1}^{p-1} \hat{\beta}_j X_{\mathrm{new},j})) + \mathrm{Bias}(\hat{\beta})^2.
\end{aligned}$$

Derivation

$$
\begin{aligned}
\text{MSE}_{pred}(\mathcal{M}) &= \mathbb{E}(Y_{\text{new}} - \hat{Y})^2 \\
&= \text{Var}(Y_{\text{new}} - \hat{Y}) + \left[\mathbb{E}(Y_{\text{new}} - \hat{Y})\right]^2 \\
&= \text{Var}(Y_{\text{new}} - \hat{Y}) + \left(\mathbb{E}Y_{\text{new}} - \mathbb{E}\hat{Y}\right)^2
\end{aligned}
$$

Note that in the second line we used the property that $\mathbb{E}(Z^2) = \text{Var}Z + (\mathbb{E}Z)^2$ for random variable Z.

- $\hat{Y}$ comes from old data and Y_{new} comes from new data
- Earlier, we assume the new data and old data share the same model $\Rightarrow$ $\mathbb{E}Y_{\text{new}} = \mathbb{E}\hat{Y}$.
- In practice, the new data often comes from a different model compared to the old data $\Rightarrow \left(\mathbb{E}Y_{\text{new}} - \mathbb{E}\hat{Y}\right) = \text{Bias}(\hat{\beta})$.
- $\text{Bias}(\hat{\beta})$ denotes the total bias due to estimating $\beta_{\text{new},j}$ using $\hat{\beta}_{\text{old},j}$.
- E.g. Suppose $\mathbb{E}Y_{\text{new}} = \beta_{\text{new},0} + \sum_j \beta_{\text{new},j} X_j$, but we used $\hat{\beta}_{old,j}$ to estimate $\beta_{\text{new},j}$ which introduces bias.

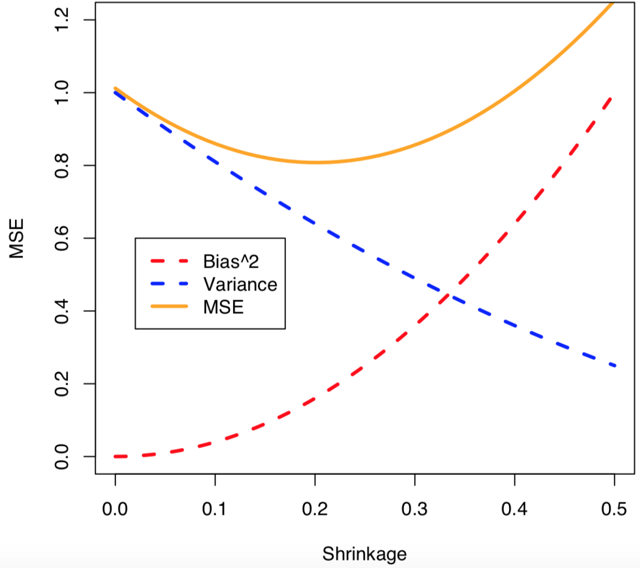

Figure credit: Prof. Jonathan Taylor from Stanford

Bias-variance tradeoff

- In choosing a model automatically, even if the "full" model is correct (unbiased), our prediction may be biased – a fact we have ignored so far.
- Inference (F, χ^2 tests, etc) is not quite exact for biased models.
- Sometimes, it is possible to find a model with lower MSE than an unbiased model! This is called the "bias-variance tradeoff."
- It is "generic" in statistics: almost always introducing some bias yields a decrease in MSE.

Shrinkage & Penalties

- Shrinkage can be thought of as "constrained" minimization.
- Minimize

$$\sum_{i=1}^{n} (Y_i - \mu)^2 \quad \text{subject to } \mu^2 \leq C$$

- Lagrange: equivalent to minimizing

$$\sum_{i=1}^{n} (Y_i - \mu)^2 + \lambda_C \mu^2$$

- Differentiating:

$$-2 \sum_{i=1}^{n} (Y_i - \hat{\mu}_C) + 2\lambda_C \hat{\mu}_C = 0$$

- Finally

$$\hat{\mu}_C = \frac{\sum_{i=1}^{n} Y_i}{n + \lambda_C} = K_C \bar{Y}, \quad K_C < 1.$$

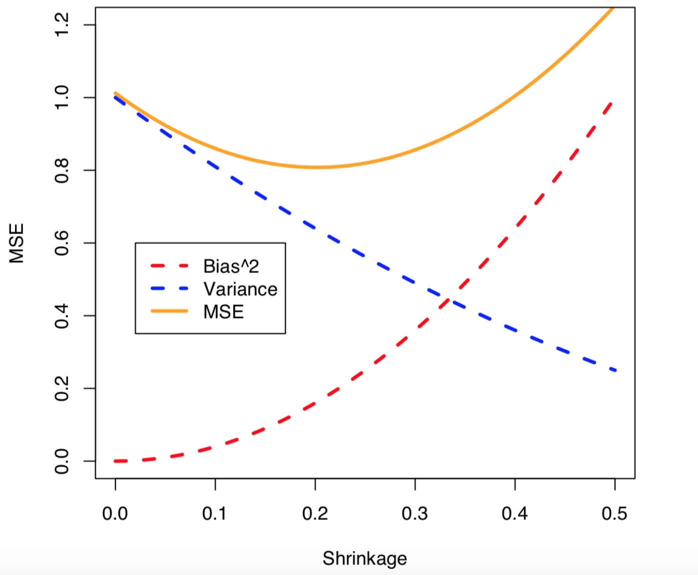

Figure credit: Prof. Jonathan Taylor from Stanford

Penalties & Priors

- Minimizing
$$\sum_{i=1}^{n}(Y_i-\mu)^2+\lambda\mu^2$$
is similar to computing "MLE" of μ if the likelihood was proportional to
$$\exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(Y_i-\mu)^2+\lambda\mu^2\right)\right)$$
- This is not a likelihood function, but it is a posterior density for μ if μ has a $N(0,\sigma^2/\lambda)$ prior.
- Hence, penalized estimation with this penalty is equivalent to using the MAP (Maximum A posteriori) estimator of μ with a Gaussian prior.

Biased regression: penalties

- Not all biased models are better — we need a way to find "good" biased model.
- Generalized one sample problem: penalize large values of β. This should lead to "multivariate" shrinkage of the vector β (next slide).
- Heuristically, "large β" is interpreted as "complex model". Goal is really to penalize "complex" models, i.e., Occam's razor.
- Equivalent Bayesian interpretation.
- If truth really is complex, this may not work! But, it will then be hard to build a good model anyways ... (statistical lore)

Ridge regression

- Assume that columns $(X_j)_{1\leq j\leq p-1}$ have zero mean, and length 1 (to distribute the penalty equally – not strictly necessary) and Y has zero mean, i.e. no intercept in the model.
- This is called the **standardized model.**

Ridge regression

A popular penalized regression technique:

$$\min_{\beta} \mathrm{SSE}_{\lambda}(\beta) = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p-1} X_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 .$$

- Corresponds (through Lagrange multiplier) to an L^2 constraint on β's.

Ridge regression

- Derivation gives
$$\hat{\beta}_\lambda = (X^t X + \lambda I)^{-1} X^t Y.$$
- This is identical to the previous $\hat{\mu}_C$ in matrix form.
- Essentially equivalent to putting a $N(0, CI)$ prior on the standardized coefficients.

Lasso regression

- Another popular penalized regression techniques.
- Use the standardized model

Lasso regression

$$\min_{\beta} \mathrm{SSE}_{\lambda}(\beta) = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p-1} X_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

- Corresponds (through Lagrange multiplier) to an L^1 constraint on β's.
- In theory works well when many β_j's are 0 and gives "sparse" solutions unlike ridge.
- Corresponds to a Laplace prior on standardized coefficients.
- R command: glmnet(...). Choose λ via cross-validation.