Cross-Validation with Confidence

Jing Lei

Department of Statistics, Carnegie Mellon University

UMN Statistics Seminar, Mar 30, 2017
### Overview

<table>
<thead>
<tr>
<th></th>
<th>Parameter est.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Overview

<table>
<thead>
<tr>
<th>Parameter est.</th>
<th>MLE, M-est., ...</th>
</tr>
</thead>
</table>

- Parameter estimation
- Maximum Likelihood Estimation (MLE)
- M-estimation (M-est.)
## Overview

<table>
<thead>
<tr>
<th>Parameter est.</th>
<th>MLE, M-est., ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point est.</td>
<td></td>
</tr>
</tbody>
</table>
### Overview

<table>
<thead>
<tr>
<th></th>
<th>Parameter est.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point est.</td>
<td>MLE, M-est., ...</td>
</tr>
<tr>
<td>Interval est.</td>
<td>Confidence interval</td>
</tr>
</tbody>
</table>
## Overview

<table>
<thead>
<tr>
<th>Parameter est.</th>
<th>Model selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point est.</td>
<td>MLE, M-est., ...</td>
</tr>
<tr>
<td>Interval est.</td>
<td>Confidence interval</td>
</tr>
</tbody>
</table>
### Overview

<table>
<thead>
<tr>
<th>Parameter est.</th>
<th>Model selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point est.</td>
<td>MLE, M-est., ...</td>
</tr>
<tr>
<td>Interval est.</td>
<td>Confidence interval</td>
</tr>
<tr>
<td></td>
<td>Cross-validation</td>
</tr>
</tbody>
</table>
### Overview

<table>
<thead>
<tr>
<th>Parameter est.</th>
<th>Model selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point est.</td>
<td>MLE, M-est., ...</td>
</tr>
<tr>
<td>Interval est.</td>
<td>Confidence interval</td>
</tr>
<tr>
<td></td>
<td>Cross-validation</td>
</tr>
<tr>
<td></td>
<td>CVC</td>
</tr>
</tbody>
</table>
Outline

• Background: cross-validation, overfitting, and uncertainty of model selection
• Cross-validation with confidence
  • A hypothesis testing framework
  • $p$-value calculation
  • Validity of the confidence set
• Model selection consistency for (low dim.) sparse linear models
• Examples
A regression setting

- Data: $D = \{(X_i, Y_i) : 1 \leq i \leq n\}$, i.i.d from joint distribution $P$ on $\mathbb{R}^p \times \mathbb{R}^1$
- $Y = f(X) + \varepsilon$, with $E(\varepsilon \mid X) = 0$
- Loss function: $\ell(\cdot, \cdot) : \mathbb{R}^2 \mapsto \mathbb{R}$
- Goal: find $\hat{f} \approx f$ so that
  \[ Q(\hat{f}) = \mathbb{E} [\ell(\hat{f}(X), Y) \mid \hat{f}] \]

is small.
A regression setting

- Data: \( D = \{(X_i, Y_i) : 1 \leq i \leq n\} \), i.i.d from joint distribution \( P \) on \( \mathbb{R}^p \times \mathbb{R}^1 \)
- \( Y = f(X) + \varepsilon \), with \( E(\varepsilon \mid X) = 0 \)
- Loss function: \( \ell(\cdot, \cdot) : \mathbb{R}^2 \to \mathbb{R} \)
- Goal: find \( \hat{f} \approx f \) so that

\[
Q(\hat{f}) \equiv \mathbb{E} [\ell(\hat{f}(X), Y) \mid \hat{f}]
\]

is small.

- The framework can be extended to unsupervised learning problems.
Model selection

- Candidate set: $\mathcal{M} = \{1, \ldots, M\}$. Each $m \in \mathcal{M}$ corresponds to a candidate model.
  1. $m$ can represent a competing theory about $P$ (e.g., $f$ is linear, $f$ is quadratic, variable $j$ is irrelevant, etc).
  2. $m$ can represent a particular value of a tuning parameter of a certain algorithm to calculate $\hat{f}$ (e.g., $\lambda$ in the lasso, number of steps in forward selection).
- Given $m$ and data $D$, there is an estimate $\hat{f}(D, m)$ of $f$.
- Model selection: find the best $m$
**Model selection**

- Candidate set: $\mathcal{M} = \{1, \ldots, M\}$. Each $m \in \mathcal{M}$ corresponds to a candidate model.
  1. $m$ can represent a competing theory about $P$ (e.g., $f$ is linear, $f$ is quadratic, variable $j$ is irrelevant, etc).
  2. $m$ can represent a particular value of a tuning parameter of a certain algorithm to calculate $\hat{f}$ (e.g., $\lambda$ in the lasso, number of steps in forward selection)
- Given $m$ and data $D$, there is an estimate $\hat{f}(D, m)$ of $f$.
- Model selection: find the best $m$
  1. such that it equals the true model
Model selection

• Candidate set: $\mathcal{M} = \{1, \ldots, M\}$. Each $m \in \mathcal{M}$ corresponds to a candidate model.
  
  1. $m$ can represent a competing theory about $P$ (e.g., $f$ is linear, $f$ is quadratic, variable $j$ is irrelevant, etc).
  2. $m$ can represent a particular value of a tuning parameter of a certain algorithm to calculate $\hat{f}$ (e.g., $\lambda$ in the lasso, number of steps in forward selection)

• Given $m$ and data $D$, there is an estimate $\hat{f}(D, m)$ of $f$.

• Model selection: find the best $m$
  
  1. such that it equals the true model
  2. such that it minimizes $Q(\hat{f})$ over all $m \in \mathcal{M}$ with high probability.
Cross-validation

- Sample split: Let $I_{\text{tr}}$ and $I_{\text{te}}$ be a partition of $\{1, ..., n\}$.
- Fitting: $\hat{f}_m = \hat{f}(D_{\text{tr}}, m)$, where $D_{\text{tr}} = \{(X_i, Y_i) : i \in I_{\text{tr}}\}$.
- Validation: $\hat{Q}(\hat{f}_m) = n_{\text{te}}^{-1} \sum_{i \in I_{\text{te}}} \ell(\hat{f}_m(X_i), Y_i)$.
- CV model selection: $\hat{m}_{\text{cv}} = \arg\min_{m \in M} \hat{Q}(\hat{f}_m)$.
- V-fold cross-validation:
  1. For $V \geq 2$, split the data into $V$ folds.
  2. Rotate over each fold as $I_{\text{tr}}$ to obtain $\hat{Q}^{(v)}(\hat{f}_m^{(v)})$
  3. $\hat{m} = \arg\min V^{-1} \sum_{v=1}^V \hat{Q}^{(v)}(\hat{f}_m^{(v)})$
  4. Popular choices of $V$: 10 and 5.
  5. $V = n$: leave-one-out cross-validation
Why can cross-validation be successful?

- To find the best model
  1. The fitting procedure \( \hat{f}(D,m) \) needs to be stable, so that the best model (almost) always gives the best fit.
  2. Conditional inference: Given \( (\hat{f}_m : m \in \mathcal{M}) \), cross-validation approximately minimizes \( Q(\hat{f}_m) \) over all \( m \).
Why can cross-validation be successful?

- To find the best model
  1. The fitting procedure $\hat{f}(D, m)$ needs to be stable, so that the best model (almost) always gives the best fit.
  2. Conditional inference: Given $(\hat{f}_m : m \in \mathcal{M})$, cross-validation approximately minimizes $Q(\hat{f}_m)$ over all $m$.

- The value of cross-validation is in conditional inference.
A simple negative example

• Model: \( Y = \mu + \epsilon \), where \( \epsilon \sim N(0, 1) \).

• \( \mathcal{M} = \{1, 2\}. \) \( m = 1: \mu = 0; m = 2: \mu \in \mathbb{R} \).

• Truth: \( \mu = 0 \)

• Consider a single split: \( \hat{f}_1 \equiv 0, \hat{f}_2 = \bar{\epsilon}_{tr} \).

• \( \hat{m}_{cv} = 1 \iff 0 < \hat{Q}(\hat{f}_2) - \hat{Q}(\hat{f}_1) = \bar{\epsilon}_{tr}^2 - 2\bar{\epsilon}_{tr}\bar{\epsilon}_{te}. \)

• If \( n_{tr}/n_{te} \approx 1 \), then \( \sqrt{n}\bar{\epsilon}_{tr} \) and \( \sqrt{n}\bar{\epsilon}_{te} \) are independent normal random variables with constant variances. So \( \mathbb{P}(\hat{m}_{cv} = 1) \) is bounded away from 1.
A simple negative example

- Model: \( Y = \mu + \varepsilon \), where \( \varepsilon \sim N(0, 1) \).

- \( \mathcal{M} = \{1, 2\} \). \( m = 1: \mu = 0; m = 2: \mu \in \mathbb{R} \).

- Truth: \( \mu = 0 \)

- Consider a single split: \( \hat{f}_1 \equiv 0, \hat{f}_2 = \bar{\varepsilon}_{\text{tr}} \).

- \( \hat{m}_{\text{cv}} = 1 \iff 0 < \hat{Q}(\hat{f}_2) - \hat{Q}(\hat{f}_1) = \bar{\varepsilon}_{\text{tr}}^2 - 2\bar{\varepsilon}_{\text{tr}}\bar{\varepsilon}_{\text{te}} \).

- If \( n_{\text{tr}}/n_{\text{te}} \approx 1 \), then \( \sqrt{n}\bar{\varepsilon}_{\text{tr}} \) and \( \sqrt{n}\bar{\varepsilon}_{\text{te}} \) are independent normal random variables with constant variances. So \( \mathbb{P}(\hat{m}_{\text{cv}} = 1) \) is bounded away from 1.

- (Shao 93, Zhang 93, Yang 07) \( \hat{m}_{\text{cv}} \) is inconsistent unless \( n_{\text{tr}} = o(n) \).
A simple negative example

- Model: \( Y = \mu + \varepsilon \), where \( \varepsilon \sim N(0, 1) \).
- \( \mathcal{M} = \{1, 2\} \). \( m = 1: \mu = 0; m = 2: \mu \in \mathbb{R} \).
- Truth: \( \mu = 0 \)
- Consider a single split: \( \hat{f}_1 \equiv 0, \hat{f}_2 = \bar{\varepsilon}_{tr} \).
- \( \hat{m}_{cv} = 1 \iff 0 < \hat{Q}(\hat{f}_2) - \hat{Q}(\hat{f}_1) = \bar{\varepsilon}_{tr}^2 - 2\bar{\varepsilon}_{tr}\bar{\varepsilon}_{te} \).
- If \( n_{tr}/n_{te} \approx 1 \), then \( \sqrt{n}\bar{\varepsilon}_{tr} \) and \( \sqrt{n}\bar{\varepsilon}_{te} \) are independent normal random variables with constant variances. So \( \mathbb{P}(\hat{m}_{cv} = 1) \) is bounded away from 1.
- (Shao 93, Zhang 93, Yang 07) \( \hat{m}_{cv} \) is inconsistent unless \( n_{tr} = o(n) \).
- \( V \)-fold does not help!
A closer look at the example

- Two potential sources of mistake: $\hat{f}(D_{tr}, m)$ is not stable, or CV does not work as expected.
A closer look at the example

- Two potential sources of mistake: \( \hat{f}(D_{tr}, m) \) is not stable, or CV does not work as expected.
- Cross-validation makes a mistake if
  \[ \hat{Q}(\hat{f}_2) - \hat{Q}(\hat{f}_1) = \bar{\varepsilon}_tr^2 - 2\bar{\varepsilon}_tr\bar{\varepsilon}_te < 0. \]

Here \( \bar{\varepsilon}_tr^2 \) is the signal, and \( 2\bar{\varepsilon}_tr\bar{\varepsilon}_te \) is the noise.
A closer look at the example

- Two potential sources of mistake: $\hat{f}(D_{tr}, m)$ is not stable, or CV does not work as expected.
- Cross-validation makes a mistake if $\hat{Q}(\hat{f}_2) - \hat{Q}(\hat{f}_1) = \bar{\epsilon}_{tr}^2 - 2\bar{\epsilon}_{tr}\bar{\epsilon}_{te} < 0$.
- But $Q(\hat{f}_2) - Q(\hat{f}_1) = \bar{\epsilon}_{tr}^2 > 0$. So the best model indeed gives the best fit. **The problem is in CV!**

Here $\bar{\epsilon}_{tr}^2$ is the signal, and $2\bar{\epsilon}_{tr}\bar{\epsilon}_{te}$ is the noise.
A closer look at the example

- Two potential sources of mistake: \( \hat{f}(D_{tr}, m) \) is not stable, or CV does not work as expected.

- Cross-validation makes a mistake if
  \[
  \hat{Q}(\hat{f}_2) - \hat{Q}(\hat{f}_1) = \bar{\varepsilon}_{tr}^2 - 2\bar{\varepsilon}_{tr}\bar{\varepsilon}_{te} < 0.
  \]

- But \( Q(\hat{f}_2) - Q(\hat{f}_1) = \bar{\varepsilon}_{tr}^2 > 0 \). So the best model indeed gives the best fit. The problem is in CV!

- Here \( \bar{\varepsilon}_{tr}^2 \) is the signal, and \( 2\bar{\varepsilon}_{tr}\bar{\varepsilon}_{te} \) is the noise.
A closer look at the example

- Two potential sources of mistake: \( \hat{f}(D_{tr}, m) \) is not stable, or CV does not work as expected.
- Cross-validation makes a mistake if
  \[ \hat{Q}(\hat{f}_2) - \hat{Q}(\hat{f}_1) = \bar{\varepsilon}^2_{tr} - 2\bar{\varepsilon}_{tr}\bar{\varepsilon}_{te} < 0. \]
- But \( Q(\hat{f}_2) - Q(\hat{f}_1) = \bar{\varepsilon}^2_{tr} > 0. \) So the best model indeed gives the best fit. The problem is in CV!
- Here \( \bar{\varepsilon}^2_{tr} \) is the signal, and \( 2\bar{\varepsilon}_{tr}\bar{\varepsilon}_{te} \) is the noise.

**Observation**

Cross-validation makes a mistake when it fails to take into account the uncertainty in the testing sample.
Cross-validation with confidence (CVC)

- We want to avoid making such a mistake as in the simple example.
- We want to use conventional split ratios, with V-fold implementation.
A fix for the simple example: hypothesis testing

- The fundamental question: When we see \( \hat{Q}(\hat{f}_2) < \hat{Q}(\hat{f}_1) \), do we feel confident to say \( Q(\hat{f}_2) < Q(\hat{f}_1) \)?
A fix for the simple example: hypothesis testing

• The fundamental question: When we see $\hat{Q}(\hat{f}_2) < \hat{Q}(\hat{f}_1)$, do we feel confident to say $Q(\hat{f}_2) < Q(\hat{f}_1)$?

• A standard solution uses hypothesis testing

$$H_0 : Q(\hat{f}_1) \leq Q(\hat{f}_2)$$

conditioning on $\hat{f}_1, \hat{f}_2$. 
A fix for the simple example: hypothesis testing

- The fundamental question: When we see $\hat{Q}(\hat{f}_2) < \hat{Q}(\hat{f}_1)$, do we feel confident to say $Q(\hat{f}_2) < Q(\hat{f}_1)$?
- A standard solution uses hypothesis testing
  
  $H_0: Q(\hat{f}_1) \leq Q(\hat{f}_2)$

  conditioning on $\hat{f}_1, \hat{f}_2$.

- Can do this using a paired sample $t$-test, say with type I error level $\alpha$. 
CVC for the simple example

- Recall that $H_0 : Q(\hat{f}_1) \leq Q(\hat{f}_2)$.

- When $H_0$ is not rejected, does it mean we shall just pick $m = 1$?

  No. Because if we consider $H'_0 : Q(\hat{f}_2) \leq Q(\hat{f}_1)$, $H'_0$ will not be rejected either (probability of rejecting $H'_0$ is bounded away from 0).

- Most likely, we do not reject $H_0$ or $H'_0$.

- We accept both fitted models $\hat{f}_1$ and $\hat{f}_2$, as they are very similar and the difference cannot be noticed from the data.
CVC for the simple example

- Recall that $H_0 : Q(\hat{f}_1) \leq Q(\hat{f}_2)$.
- When $H_0$ is not rejected, does it mean we shall just pick $m = 1$?
  - No. Because if we consider $H'_0 : Q(\hat{f}_2) \leq Q(\hat{f}_1)$. $H'_0$ will not be rejected either (probability of rejecting $H'_0$ is bounded away from 0.)
- Most likely, we do not reject $H_0$ or $H'_0$. 
CVC for the simple example

- Recall that $H_0 : Q(\hat{f}_1) \leq Q(\hat{f}_2)$.
- When $H_0$ is not rejected, does it mean we shall just pick $m = 1$?
- No. Because if we consider $H'_0 : Q(\hat{f}_2) \leq Q(\hat{f}_1)$. $H'_0$ will not be rejected either (probability of rejecting $H'_0$ is bounded away from 0.)
- Most likely, we do not reject $H_0$ or $H'_0$.
- We accept both fitted models $\hat{f}_1$ and $\hat{f}_2$, as they are very similar and the difference cannot be noticed from the data.
**Existing work**

- Ferrari and Yang (2014): F-tests, need a good variable screening procedure in high dimensions.
Existing work

• Ferrari and Yang (2014): F-tests, need a good variable screening procedure in high dimensions.
• Our approach: one step, with provable coverage and power under mild assumptions in high dimensions.
Existing work

- Ferrari and Yang (2014): F-tests, need a good variable screening procedure in high dimensions.
- Our approach: one step, with provable coverage and power under mild assumptions in high dimensions.
- Key technique: high-dimensional Gaussian comparison of sample means (Chernozhukov et al).
**CVC in general**

- Now suppose we have a set of candidate models $\mathcal{M} = \{1, ..., M\}$.
- Split the data into $D_{tr}$ and $D_{te}$, and use $D_{tr}$ to obtain $\hat{f}_m$ for each $m$.
- Recall that the model quality is $Q(\hat{f}) = \mathbb{E} \left[ \ell(\hat{f}(X), Y) \mid \hat{f} \right]$.
**CVC in general**

- Now suppose we have a set of candidate models $\mathcal{M} = \{1, ..., M\}$.
- Split the data into $D_{tr}$ and $D_{te}$, and use $D_{tr}$ to obtain $\hat{f}_m$ for each $m$.
- Recall that the model quality is $Q(\hat{f}) = \mathbb{E} \left[ \ell(\hat{f}(X), Y) | \hat{f} \right]$.
- For each $m$, test hypothesis (conditioning on $\hat{f}_1, ..., \hat{f}_M$)
  \[ H_{0,m} : \min_{j \neq m} Q(\hat{f}_j) \geq Q(\hat{f}_m). \]
- Let $\hat{p}_m$ be a valid $p$-value.
CVC in general

- Now suppose we have a set of candidate models $\mathcal{M} = \{1, \ldots, M\}$.
- Split the data into $D_{\text{tr}}$ and $D_{\text{te}}$, and use $D_{\text{tr}}$ to obtain $\hat{f}_m$ for each $m$.
- Recall that the model quality is $Q(\hat{f}) = \mathbb{E} \left[ \ell(\hat{f}(X), Y) \mid \hat{f} \right]$.
- For each $m$, test hypothesis (conditioning on $\hat{f}_1, \ldots, \hat{f}_M$)

$$H_{0,m} : \min_{j \neq m} Q(\hat{f}_j) \geq Q(\hat{f}_m).$$

- Let $\hat{p}_m$ be a valid $p$-value.
- $\mathcal{A}_{\text{cvc}} = \{m : \hat{p}_m > \alpha\}$ is our confidence set for the best fitted model: $\mathbb{P}(m^* \in \mathcal{A}_{\text{cvc}}) \geq 1 - \alpha$, where $m^* = \arg\min_m Q(\hat{f}_m)$.
Calculating $\hat{p}_m$

- Recall that $D_{tr}$ is the training data and $D_{te}$ is the testing data.
- The test and $p$-values are conditional on $D_{tr}$.
- Data: $n_{te} \times (M - 1)$ matrix ($I_{te}$ is the index set of $D_{te}$)
  \[
  \left[ \xi^{(i)}_{m,j} \right]_{i \in I_{te}, j \neq m}, \quad \text{where} \quad \xi^{(i)}_{m,j} = \ell(\hat{f}_m(X_i), Y_i) - \ell(\hat{f}_j(X_i), Y_i)
  \]
- Multivariate mean testing. $H_{0,m} : \mathbb{E}(\xi_{m,j}) \leq 0, \forall j \neq m$. 
Calculating $\hat{p}_m$

- Recall that $D_{\text{tr}}$ is the training data and $D_{\text{te}}$ is the testing data.
- The test and $p$-values are conditional on $D_{\text{tr}}$.
- Data: $n_{\text{te}} \times (M - 1)$ matrix ($I_{\text{te}}$ is the index set of $D_{\text{te}}$)
  \[
  \left[ \xi^{(i)}_{m,j} \right]_{i \in I_{\text{te}}, j \neq m}, \quad \text{where} \quad \xi^{(i)}_{m,j} = \ell(\hat{f}_m(X_i), Y_i) - \ell(\hat{f}_j(X_i), Y_i)
  \]
- Multivariate mean testing. $H_{0,m} : \mathbb{E}(\xi_{m,j}) \leq 0, \forall j \neq m$.
- Challenges
Calculating \( \hat{p}_m \)

- Recall that \( D_{tr} \) is the training data and \( D_{te} \) is the testing data.
- The test and \( p \)-values are conditional on \( D_{tr} \).
- Data: \( n_{te} \times (M - 1) \) matrix (\( I_{te} \) is the index set of \( D_{te} \))

\[
\begin{bmatrix}
\xi_{m,j}^{(i)} \\
\end{bmatrix}_{i \in I_{te}, j \neq m}, \text{ where } \xi_{m,j}^{(i)} = \ell(\hat{f}_m(X_i), Y_i) - \ell(\hat{f}_j(X_i), Y_i)
\]

- Multivariate mean testing. \( H_{0,m} : \mathbb{E}(\xi_{m,j}) \leq 0, \forall j \neq m. \)
- Challenges
  1. High dimensionality: \( M \) can be large.
Calculating $\hat{p}_m$

- Recall that $D_{tr}$ is the training data and $D_{te}$ is the testing data.
- The test and $p$-values are conditional on $D_{tr}$.
- Data: $n_{te} \times (M - 1)$ matrix ($I_{te}$ is the index set of $D_{te}$)

$$\left[\xi^{(i)}_{m,j}\right]_{i \in I_{te}, j \neq m}, \text{ where } \xi^{(i)}_{m,j} = \ell(\hat{f}_m(X_i), Y_i) - \ell(\hat{f}_j(X_i), Y_i)$$

- Multivariate mean testing. $H_{0,m}: \mathbb{E}(\xi_{m,j}) \leq 0, \forall j \neq m$.

- Challenges
  1. High dimensionality: $M$ can be large.
  2. Potentially high correlation between $\xi_{m,j}$ and $\xi_{m,j'}$. 
Calculating $\hat{p}_m$

- Recall that $D_{tr}$ is the training data and $D_{te}$ is the testing data.
- The test and $p$-values are conditional on $D_{tr}$.
- Data: $n_{te} \times (M - 1)$ matrix ($I_{te}$ is the index set of $D_{te}$)

$$
\begin{bmatrix}
\xi^{(i)}_{m,j}
\end{bmatrix}_{i \in I_{te}, j \neq m}, \text{ where } \xi^{(i)}_{m,j} = \ell(\hat{f}_m(X_i), Y_i) - \ell(\hat{f}_j(X_i), Y_i)
$$

- Multivariate mean testing. $H_{0,m}: \mathbb{E}(\xi_{m,j}) \leq 0, \forall j \neq m$.
- Challenges
  1. High dimensionality: $M$ can be large.
  2. Potentially high correlation between $\xi_{m,j}$ and $\xi_{m,j'}$.
  3. Vastly different scaling: $\text{Var}(\xi_{m,j})$ can be $O(1)$ or $O(n^{-1})$. 
Calculating \( \hat{p}_m \)

- \( H_{0,m} : \mathbb{E}(\xi_{m,j}) \leq 0, \forall j \neq m. \)

- Let \( \hat{\mu}_{m,j} \) and \( \hat{\sigma}_{m,j} \) be the sample mean and standard deviation of \((\xi_{m,j}^{(i)} : i \in I_{te})\).

- Naturally, one would reject \( H_{0,m} \) for large values of

\[
\max_{j \neq m} \frac{\hat{\mu}_{m,j}}{\hat{\sigma}_{m,j}}.
\]

- Approximate the null distribution using high dimensional Gaussian comparison.
Studentized Gaussian Multiplier Bootstrap

1. \( T_m = \max_{j \neq m} n_{\text{te}} \sqrt{\frac{\hat{\mu}_{m,j}}{\hat{\sigma}_{m,j}}} \)

2. Let \( B \) be the bootstrap sample size. For \( b = 1, \ldots, B \),
   
   2.1 Generate iid standard Gaussian \( \zeta_i, i \in I_{\text{te}} \).
   
   2.2 \( T^*_b = \max_{j \neq m} \frac{1}{\sqrt{n_{\text{te}}}} \sum_{i \in I_{\text{te}}} \frac{\xi_{m,i}^{(i)} - \hat{\mu}_{m,j}}{\hat{\sigma}_{m,j}} \zeta_i \)

3. \( \hat{p}_m = B^{-1} \sum_{b=1}^{B} 1(T^*_b > T_m) \).
Studentized Gaussian Multiplier Bootstrap

1. \[ T_m = \max_{j \neq m} \sqrt{n_{te}} \frac{\hat{\mu}_{m,j}}{\hat{\sigma}_{m,j}} \]

2. Let \( B \) be the bootstrap sample size. For \( b = 1, \ldots, B \),
   \[ T_b^* = \max_{j \neq m} \frac{1}{\sqrt{n_{te}}} \sum_{i \in I_{te}} \frac{\xi_{m,j}^{(i)} - \hat{\mu}_{m,j}}{\hat{\sigma}_{m,j}} \zeta_i \]

3. \[ \hat{p}_m = B^{-1} \sum_{b=1}^{B} 1(T_b^* > T_m). \]
   - The studentization takes care of the scaling difference.
Studentized Gaussian Multiplier Bootstrap

1. \( T_m = \max_{j \neq m} \sqrt{n_{te}} \frac{\hat{\mu}_{m,j}}{\hat{\sigma}_{m,j}} \)

2. Let \( B \) be the bootstrap sample size. For \( b = 1, \ldots, B \),
   2.1 Generate iid standard Gaussian \( \zeta_i, i \in I_{te} \).
   2.2 \( T^*_b = \max_{j \neq m} \frac{1}{\sqrt{n_{te}}} \sum_{i \in I_{te}} \frac{\xi^{(i)}_m - \hat{\mu}_{m,j}}{\hat{\sigma}_{m,j}} \zeta_i \)

3. \( \hat{\rho}_m = B^{-1} \sum_{b=1}^{B} 1(T^*_b > T_m) \).
   - The studentization takes care of the scaling difference.
   - The bootstrap Gaussian comparison takes care of the dimensionality and correlation.
Properties of CVC

- $\mathcal{A}_{cvc} = \{m : \hat{p}_m > \alpha\}$.
- Let $\hat{m}_{cv} = \arg\min_m \hat{Q}(\hat{f}_m)$. By construction $T_{\hat{m}_{cv}} \leq 0$.

**Proposition**

If $\alpha < 0.5$, then $\mathbb{P}(\hat{m}_{cv} \in \mathcal{A}_{cvc}) \to 1$ as $B \to \infty$.

- Proof: $\left[ \frac{1}{\sqrt{n_{te}}} \sum_{i \in I_{te}} \frac{\xi^{(i)}_{m,j} - \hat{\mu}_{m,j}}{\hat{\sigma}_{m,j}} \zeta_i \right]_{j \neq m}$ is a zero-mean Gaussian random vector. So the upper $\alpha$ quantile of its maximum must be positive.
- Can view $\hat{m}_{cv}$ as the “center” of the confidence set.
Coverage of $\mathcal{A}_{cvc}$

- Recall $\xi_{m,j} = \ell(\hat{f}_m(X), Y) - \ell(\hat{f}_j(X), Y)$, with independent $(X, Y)$.
- Let $\mu_{m,j} = \mathbb{E} \left[ \xi_{m,j} \mid \hat{f}_m, \hat{f}_m \right]$, $\sigma_{m,j}^2 = \text{Var} \left[ \xi_{m,j} \mid \hat{f}_m, \hat{f}_m \right]$.

**Theorem**

Assume that $(\xi_{m,j} - \mu_{m,j})/(A_n \sigma_{m,j})$ has sub-exponential tail for all $m \neq j$ and some $A_n \geq 1$ such that for some $c > 0$

$$A_n^6 \log^7 (M \lor n) = O(n^{1-c}).$$

1. If $\max_{j \neq m} \left( \frac{\mu_{m,j}}{\sigma_{m,j}} \right) = o \left( \sqrt{\frac{1}{n \log(M \lor n)}} \right)$, then $\mathbb{P}(m \in \mathcal{A}_{cvc}) \geq 1 - \alpha + o(1)$.

2. If $\max_{j \neq m} \left( \frac{\mu_{m,j}}{\sigma_{m,j}} \right) \geq CA_n \sqrt{\frac{\log(M \lor n)}{n}}$ for some constant $C$, and $\alpha \geq n^{-1}$, then $\mathbb{P}(m \in \mathcal{A}_{cvc}) = o(1)$. 
Coverage of $\mathcal{A}_{cvc}$

1. If $m^*$ is the best fitted model which minimizes $Q(\hat{f}_m)$ over all \( \{ \hat{f}_m : m \in \mathcal{M} \} \), then $\mu_{m^*, j}/\sigma_{m^*, j} \leq 0$ for all $j$. Thus $\mathbb{P}(m^* \in \mathcal{A}_{cvc}) \geq 1 - \alpha + o(1)$.

2. If $\mu_{m, j} = 0$ for all $j$, then $\mathbb{P}(m \in \mathcal{A}_{cvc}) = 1 - \alpha + o(1)$.

3. Part 2 of the theorem ensures that bad models are excluded with high probability.
Proof of coverage

• Let $Z(\Sigma) = \max N(0, \Sigma)$, and $z(1 - \alpha, \Sigma)$ its $1 - \alpha$ quantile.

• Let $\hat{\Gamma}$ and $\Gamma$ be sample and population correlation matrices of $(\xi^{(i)}_{m,j})_{i \in I_{te}, j \neq m}$. When $B \rightarrow \infty$,

$$
P(\hat{\rho}_m \leq \alpha) = P \left[ \max_j \sqrt{n_{te}} \frac{\hat{\mu}_{m,j}}{\hat{\sigma}_{m,j}} \geq z(1 - \alpha, \hat{\Gamma}) \right]
$$

• Tools (2, 3 are due to Chernozhukov et al.)
  1. Concentration: $\sqrt{n_{te}} \frac{\hat{\mu}_{m,j}}{\hat{\sigma}_{m,j}} \leq \sqrt{n_{te}} \frac{\mu_{m,j} - \mu_{m,j}}{\sigma_{m,j}} + o(1 / \sqrt{\log M})$
  2. Gaussian comparison: $\max_j \sqrt{n_{te}} \frac{\hat{\mu}_{m,j} - \mu_{m,j}}{\sigma_{m,j}} \overset{d}{\approx} Z(\Gamma) \overset{d}{\approx} Z(\hat{\Gamma})$
  3. Anti-concentration: $Z(\hat{\Gamma})$ and $Z(\Gamma)$ have densities $\lesssim \sqrt{\log M}$
V-fold CVC

- Split data into $V$ folds.

Rigorous justification is hard due to dependence between folds. But empirically much better.
V-fold CVC

- Split data into $V$ folds.
- Let $v_i$ be the fold that contains data point $i$. 

Rigorous justification is hard due to dependence between folds. But empirically much better.
V-fold CVC

- Split data into V folds.
- Let $v_i$ be the fold that contains data point $i$.
- Let $\hat{f}_{m,v}$ be the estimate using model $m$ and all data but fold $v$. 

$\xi(i)_m,j = \ell(\hat{f}_{m,v}(X_i),Y_i) - \ell(\hat{f}_{m,v}(X_i),Y_i)$, for all $1 \leq i \leq n$.

Calculate $T_m$ and $T^*_b$ correspondingly using the $n \times (M-1)$ cross-validated error difference matrix $\xi(i)_m,j$ $1 \leq i \leq n$, $j \neq m$. 

Rigorous justification is hard due to dependence between folds. But empirically much better.
**V-fold CVC**

- Split data into $V$ folds.
- Let $v_i$ be the fold that contains data point $i$.
- Let $\hat{f}_{m,v}$ be the estimate using model $m$ and all data but fold $v$.
- $\xi_{m,j}^{(i)} = \ell(\hat{f}_{m,v_i}(X_i), Y_i) - \ell(\hat{f}_{m,v_i}(X_i, Y_i))$, for all $1 \leq i \leq n$. 

Rigorous justification is hard due to dependence between folds. But empirically much better.
**V-fold CVC**

- Split data into $V$ folds.
- Let $v_i$ be the fold that contains data point $i$.
- Let $\hat{f}_{m,v}$ be the estimate using model $m$ and all data but fold $v$.
- $\xi^{(i)}_{m,j} = \ell(\hat{f}_{m,v_i}(X_i), Y_i) - \ell(\hat{f}_{m,v_i}(X_i, Y_i))$, for all $1 \leq i \leq n$.
- Calculate $T_m$ and $T_b^*$ correspondingly using the $n \times (M - 1)$ cross-validated error difference matrix $(\xi^{(i)}_{m,j})_{1 \leq i \leq n, j \neq m}$.
V-fold CVC

- Split data into $V$ folds.
- Let $v_i$ be the fold that contains data point $i$.
- Let $\hat{f}_{m,v}$ be the estimate using model $m$ and all data but fold $v$.
- $\xi_{m,j}^{(i)} = \ell(\hat{f}_{m,v_i}(X_i), Y_i) - \ell(\hat{f}_{m,v_i}(X_i, Y_i))$, for all $1 \leq i \leq n$.
- Calculate $T_m$ and $T^*_b$ correspondingly using the $n \times (M - 1)$ cross-validated error difference matrix $(\xi_{m,j}^{(i)})_{1 \leq i \leq n, j \neq m}$.
- Rigorous justification is hard due to dependence between folds. But empirically much better.
Example: the diabetes data (Efron et al 04)

- $n = 442$, with 10 covariates: age, sex, bmi, blood pressure, etc.
- Response is diabetes progression after one year.
- Including all quadratic terms, $p = 64$.
- 5-fold CVC with $\alpha = 0.05$, using Lasso with 50 values of $\lambda$.

Triangle: models in $\mathcal{A}_{cvc}$, solid triangle: $\hat{m}_{cv}$. 
Simulations: coverage of $A_{cvc}$

- $Y = X^T \beta + \epsilon$, $X \sim N(0, \Sigma)$, $\epsilon \sim N(0, 1)$, $n = 200$, $p = 200$
- $\Sigma = I_{200}$ (identity), or $\Sigma_{jk} = 0.5 + 0.5\delta_{jk}$ (correlated).
- $\beta = (1, 1, 1, 0, ..., 0)^T$ (simple), or $\beta = (1, 1, 1, 0.7, 0.5, 0.3, 0, ..., 0)^T$ (mixed).
- 5-fold CVC with $\alpha = 0.05$ using Lasso with 50 values of $\lambda$

| setting of $(\Sigma, \beta)$ | coverage | $|A_{cvc}|$ | cv is opt. |
|-----------------------------|----------|------------|------------|
| identity, simple            | .92 (.03)| 5.1 (.19)  | .27 (.04)  |
| identity, mixed             | .95 (.02)| 5.1 (.18)  | .37 (.05)  |
| correlated, simple          | .96 (.02)| 7.5 (.18)  | .18 (.04)  |
| correlated, mixed           | .93 (.03)| 7.4 (.23)  | .19 (.04)  |
Simulations: coverage of $\mathcal{A}_{cvc}$

- $Y = X^T \beta + \epsilon$, $X \sim N(0, \Sigma)$, $\epsilon \sim N(0, 1)$, $n = 200$, $p = 200$
- $\Sigma = I_{200}$ (identity), or $\Sigma_{jk} = 0.5 + 0.5 \delta_{jk}$ (correlated).
- $\beta = (1, 1, 1, 0, \ldots, 0)^T$ (simple), or $\beta = (1, 1, 1, 0.7, 0.5, 0.3, 0, \ldots, 0)^T$ (mixed).
- 5-fold CVC with $\alpha = 0.05$ using forward stepwise

| setting of $(\Sigma, \beta)$        | coverage | $|\mathcal{A}_{cvc}|$ | cv is opt. |
|-------------------------------------|----------|------------------------|------------|
| identity, simple                    | 1 (0)    | 3.7 (.29)              | .87 (.03)  |
| identity, mixed                     | .95 (.02)| 5.2 (.33)              | .58 (.05)  |
| correlated, simple                  | .97 (.02)| 4.1 (.31)              | .80 (.04)  |
| correlated, mixed                   | .93 (.03)| 6.3 (.36)              | .44 (.05)  |
How to use $\mathcal{A}_{\text{cvc}}$?

- We are often interested in picking one model, not a subset of models.
- $\mathcal{A}_{\text{cvc}}$ provides some flexibility of picking among a subset of highly competitive models.
How to use $\mathcal{A}_{cvc}$?

- We are often interested in picking one model, not a subset of models.
- $\mathcal{A}_{cvc}$ provides some flexibility of picking among a subset of highly competitive models.
  1. $\mathcal{A}_{cvc}$ may contain a model that includes a particularly interesting variable.
How to use $\mathcal{A}_{cvc}$?

- We are often interested in picking one model, not a subset of models.

- $\mathcal{A}_{cvc}$ provides some flexibility of picking among a subset of highly competitive models.
  1. $\mathcal{A}_{cvc}$ may contain a model that includes a particularly interesting variable.
  2. $\mathcal{A}_{cvc}$ can be used to answer questions like “Is fitting procedure A better than procedure B?”
How to use $\mathcal{A}_{cvc}$?

- We are often interested in picking one model, not a subset of models.
- $\mathcal{A}_{cvc}$ provides some flexibility of picking among a subset of highly competitive models.
  1. $\mathcal{A}_{cvc}$ may contain a model that includes a particularly interesting variable.
  2. $\mathcal{A}_{cvc}$ can be used to answers questions like “Is fitting procedure A better than procedure B?”
  3. We can also simply choose the most parsimonious model in $\mathcal{A}_{cvc}$. 
Now consider the linear regression problem:

\[ Y = X^T \beta + \epsilon. \]

Let \( J_m \) be the subset of variables selected using model \( m \)

\[ \hat{m}_{\text{cvc.min}} = \arg \min_{m \in \mathcal{A}_{\text{cvc}}} |J_m|. \]

\( \hat{m}_{\text{cvc.min}} \) is the simplest model that gives a similar predictive risk as \( \hat{m}_{\text{cv}} \).
A classical setting

- $Y = X^T \beta + \epsilon$, $X \in \mathbb{R}^p$, $\text{Var}(X) = \Sigma$ has full rank.
- $\epsilon$ has mean zero and variance $\sigma^2 < \infty$.
- Assume that $(p, \Sigma, \sigma^2)$ are fixed and $n \to \infty$.
- $\mathcal{M}$ contains the true model $m^*$, and at least one overfitting model.
- $n_{tr}/n_{te} \gtrsim 1$.
- Using squared loss, the true model and all overfitting models give $\sqrt{n}$-consistent estimates.
- Early results (Shao 93, Zhang 93, Yang 07) show that $\mathbb{P}(\hat{m}_{cv} \neq m^*)$ is bounded away from 0.
Consistency of $\hat{m}_{\text{cvc.min}}$

**Theorem**

Assume that $X$ and $\epsilon$ are independent and sub-Gaussian, and $\mathcal{A}_{\text{cvc}}$ is the output of CVC with $\alpha = o(1)$ and $\alpha \geq n^{-1}$, then

$$\lim_{n \to \infty} \mathbb{P}(\hat{m}_{\text{cvc.min}} = m^*) = 1.$$
Consistency of $\hat{m}_{cvc,\min}$

**Theorem**

Assume that $X$ and $\varepsilon$ are independent and sub-Gaussian, and $\mathcal{A}_{cvc}$ is the output of CVC with $\alpha = o(1)$ and $\alpha \geq n^{-1}$, then

$$\lim_{n \to \infty} \mathbb{P}(\hat{m}_{cvc,\min} = m^*) = 1.$$ 

- Sub-Gaussianity of $X$ and $\varepsilon$ implies that $(Y - X^T \beta)^2$ is sub-exponential.
Consistency of $\hat{m}_{cvc.\text{min}}$

**Theorem**

Assume that $X$ and $\varepsilon$ are independent and sub-Gaussian, and $A_{cvc}$ is the output of CVC with $\alpha = o(1)$ and $\alpha \geq n^{-1}$, then

$$\lim_{n \to \infty} \mathbb{P}(\hat{m}_{cvc.\text{min}} = m^*) = 1.$$ 

- Sub-Gaussianity of $X$ and $\varepsilon$ implies that $(Y - X^T \beta)^2$ is sub-exponential.
- Can allow $p$ to grow slowly as $n$ using union bound.
Example in low-dim. variable selection

- Synthetic data with $p = 5$, $n = 40$, as in [Shao 93].
- $Y = X^T \beta + \varepsilon$, $\beta = (2, 9, 0, 4, 8)^T$, $\varepsilon \sim N(0, 1)$.
- Generated additional rows for $n = 60, 80, 100, 120, 140, 160$.
- Candidates: $(1, 4, 5), (1, 2, 4, 5), (1, 3, 4, 5), (1, 2, 3, 4, 5)$
- Repeated 1000 times, using OLS with 5-fold CVC.
Simulations: variable selection with $\hat{m}_{\text{cvc.min}}$

- $Y = X^T \beta + \varepsilon$, $X \sim N(0, \Sigma)$, $\varepsilon \sim N(0, 1)$, $n = 200$, $p = 200$
- $\Sigma = I_{200}$ (identity), or $\Sigma_{jk} = 0.5 + 0.5\delta_{jk}$ (correlated).
- $\beta = (1, 1, 1, 0, \ldots, 0)^T$ (simple)
- 5-fold CVC with $\alpha = 0.05$ using forward stepwise

<table>
<thead>
<tr>
<th>setting of $(\Sigma, \beta)$</th>
<th>oracle</th>
<th>$\hat{m}_{\text{cvc.min}}$</th>
<th>$\hat{m}_\text{cv}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>identity, simple</td>
<td>1</td>
<td>1</td>
<td>.87</td>
</tr>
<tr>
<td>correlated, simple</td>
<td>1</td>
<td>.97</td>
<td>.80</td>
</tr>
</tbody>
</table>

Proportion of correct model selection over 100 independent data sets.

Oracle method: the number of steps that gives smallest prediction risk.
Simulations: variable selection with $\hat{m}_{\text{cvc.min}}$

- $Y = X^T \beta + \varepsilon$, $X \sim N(0, \Sigma)$, $\varepsilon \sim N(0, 1)$, $n = 200$, $p = 200$
- $\Sigma = I_{200}$ (identity), or $\Sigma_{jk} = 0.5 + 0.5 \delta_{jk}$ (correlated).
- $\beta = (1, 1, 1, 0, \ldots, 0)^T$ (simple)
- 5-fold CVC with $\alpha = 0.05$ using Lasso + Least Square

<table>
<thead>
<tr>
<th>setting of $(\Sigma, \beta)$</th>
<th>oracle</th>
<th>$\hat{m}_{\text{cvc.min}}$</th>
<th>$\hat{m}_{\text{cv}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>identity, simple</td>
<td>1</td>
<td>1</td>
<td>.88</td>
</tr>
<tr>
<td>correlated, simple</td>
<td>.87</td>
<td>.85</td>
<td>.71</td>
</tr>
</tbody>
</table>

Proportion of correct model selection over 100 independent data sets.

Oracle method: the $\lambda$ value that gives smallest prediction risk.
The diabetes data revisited

- Split $n = 442$ into 300 (estimation) and 142 (risk approximation).
- 5-fold CVC applied on the 300 sample points, with a final re-fit.
- The final estimate is evaluated using the 142 hold-out sample.
- Repeat 100 times, using Lasso with 50 values of $\lambda$. 

![Box plots of test error and model size for different cross-validation methods.](image)
Summary

• CVC: confidence sets for model selection

• $\hat{m}_{\text{cvc.min}}$ has similar risk as $\hat{m}_{\text{cv}}$ using a simpler model.

• Extensions
  • Validity of CVC in high dimensions and nonparametric settings.
  • Unsupervised problems
    1. Clustering
    2. Matrix decomposition (PCA, SVD, etc)
    3. Network models

• Other sample-splitting based inference methods.
Thanks!

Questions?

Paper:
“Cross-Validation with Confidence”, arxiv.org/1703.07904

Slides:
http://www.stat.cmu.edu/~jinglei/cvc_umn.pdf