Least Squares Model Averaging

Bruce E. Hansen

University of Wisconsin

January 2006

Revised: August 2006
Introduction

- This paper develops a model averaging estimator for linear regression.
- Model averaging is a generalization of model selection.
- We propose a Mallows’ criterion to select the weights.
- These weights are optimal in the sense that they asymptotically minimize the conditional squared error.
- A technical caveat is that the proof restricts attention to discrete weights, while continuous weights are desired.
- A substantive caveat is that we restrict attention to homoskedastic errors.
- (May be able to extend the method to heteroskedastic errors by using CV.)
- A companion paper applies the method to forecast combination.
Model Selection

- AIC: (Akaike, 1973)
- Mallows’ $C_p$: (Mallows, 1973)
- BIC: (Schwarz, 1978)
- Cross-Validation: (Stone, 1974)
- Generalized Cross Validation: (Craven and Wahba, 1979)
- Focused Information Criterion: (Claeskens and Hjort, 2003)
- GMM Selection: (Andrews and Lu, 2001)
- EL Selection: (Hong, Preston and Shum, 2003)
Model Averaging

- Bayesian Model Averaging:
  Draper (1995)
  Raftery, Madigan and Hoeting (1997)

- Exponential AIC weights:
  Buckland et. al. (1997)
  Burnham and Anderson (2002)

- Mixing estimator: Yang (2001)


Selection of terms in series expansion

- Andrews (1991a)
- Newey (1997)

Asymptotic Optimality of Model Selection Rules:

- Li (1987): Mallows, GCV and CV in homoskedastic regression
Model

\[ y_i = \mu_i + e_i \]
\[ \mu_i = \sum_{j=1}^{\infty} \theta_j x_{ji} \]
\[ E(e_i \mid x_i) = 0 \]
\[ E(e_i^2 \mid x_i) = \sigma^2. \]

The coefficients and regressors may be terms in a series expansion, or a “large” number of potential regressors.

The goal is to estimate \( \mu_i \) using a finite sample.

There are an infinite number of coefficients.

The regressors are ordered, at least by groups.
Approximating models

The $m^{th}$ approximating model uses the first $k_m$ regressors

$$y_i = \sum_{j=1}^{k_m} \theta_j x_{ji} + b_{mi} + e_i$$

Approximation error:

$$b_{mi} = \sum_{j=k_m+1}^{\infty} \theta_j x_{ji}.$$  

In matrix notation,

$$Y = X_m \Theta_m + b_m + e$$

Least-squares estimate of $\Theta_m$

$$\hat{\Theta}_m = (X_m'X_m)^{-1} X_m'Y$$

$$= \Theta_m + (X_m'X_m)^{-1} X_m' b_m + (X_m'X_m)^{-1} X_m' e$$
Model Selection

Let

\[
\hat{\sigma}^2(m) = \frac{1}{n} \left( Y - X_m \hat{\Theta}_m \right)' \left( Y - X_m \hat{\Theta}_m \right).
\]

The BIC, AIC, Mallows’, and CV selection rules picks \( m \) to minimize

\[
BIC(m) = \ln \hat{\sigma}^2(m) + \ln(n) \frac{k_m}{n}
\]

\[
AIC(m) = \ln \hat{\sigma}^2(m) + \frac{2k_m}{n}
\]

\[
C(m) = \hat{\sigma}^2(m) + \frac{2k_m \tilde{\sigma}^2}{n}
\]

\[
CV(m) = \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \hat{\mu}_{i,-i}(m) \right)^2
\]

where \( \hat{\mu}_{i,-i}(m) \) is the leave-one-out estimate of \( \mu_i \).

If \( m \) is small, the AIC, Mallows’ and CV rules are often quite similar.
Properties of Model Selection Rules

If the true model is finite, BIC is consistent; AIC/Mallows over-select.

If the true model is infinitie-dimensional, Li (Annals of Statistics, 1987) showed that Mallows and CV selection are asymptically optimal.

Mallows’ selected model

\[ \hat{m} = \arg \min_m C(m) \]

Average squared error:

\[ L_n(m) = \left( \mu - X_m \hat{\Theta}_m \right)' \left( \mu - X_m \hat{\Theta}_m \right) \]

Optimality: As \( n \to \infty \)

\[ \frac{L_n(\hat{m})}{\inf_m L_n(m)} \to_p 1 \]
Model Averaging

Let $W = (w_1, ..., w_M)'$ be a vector of non-negative weights.

It is an element of the unit simplex in $\mathbb{R}^M$

$$\mathcal{H}_n = \left\{ W \in [0, 1]^M : \sum_{m=1}^{M} w_m = 1 \right\}.$$ 

A model average estimator of $\Theta_M$ is

$$\hat{\Theta}(W) = \sum_{m=1}^{M} w_m \begin{pmatrix} \hat{\Theta}_m \\ 0 \end{pmatrix}.$$
The model average estimate of $\mu$ is

$$
\hat{\mu}(W) = X_M \hat{\Theta} \\
= \sum_{m=1}^{M} w_m X_m \hat{\Theta}_m \\
= \sum_{m=1}^{M} w_m X_m (X'_m X_m)^{-1} X'_m Y \\
= \sum_{m=1}^{M} w_m P_m Y \\
= P(W) Y
$$

where

$$
P(W) = \sum_{m=1}^{M} w_m P_m.
$$
Lemma 1

1. $\text{tr} (P(W)) = \sum_{m=1}^{M} w_m k_m \equiv k(W)$. 

2. $\text{tr} (P(W) P(W)) = \sum_{m=1}^{M} \sum_{l=1}^{M} w_m w_l \min (k_l, k_m) = W' \Gamma_M W$ where 

$$
\Gamma_M = \begin{bmatrix}
  k_1 & k_1 & k_1 & \cdots & k_1 \\
  k_1 & k_2 & k_1 & \cdots & k_2 \\
  k_1 & k_2 & k_2 & \cdots & k_3 \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  k_1 & k_2 & k_3 & \cdots & k_M \\
\end{bmatrix}.
$$

3. $\lambda_{\max} (P(W)) \leq 1$. 
Lemma 2. Let $L_n(W) = (\hat{\mu}(W) - \mu)'(\hat{\mu}(W) - \mu)$. Then

$$R_n(W) = E(L_n(W) \mid X) = W'(A_n + \sigma^2 \Gamma_M) W$$

where

$$\Gamma_M = \begin{bmatrix} k_1 & k_1 & k_1 & \cdots & k_1 \\ k_1 & k_2 & k_2 & \cdots & k_2 \\ k_1 & k_2 & k_3 & \cdots & k_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ k_1 & k_2 & k_3 & \cdots & k_M \end{bmatrix}, \quad A_n = \begin{bmatrix} a_1 & a_2 & a_3 & \cdots & a_M \\ a_2 & a_2 & a_3 & \cdots & a_M \\ a_3 & a_3 & a_3 & \cdots & a_M \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_M & a_M & a_M & \cdots & a_M \end{bmatrix},$$

and $a_m = b'_m (I - P_m) b_m$ are monotonically decreasing.

(Recall that $b_m$ are the approximation errors: $Y = X_m \Theta_m + b_m + e.$)
Lemma 2 shows that $R_n(W)$ is quadratic in $W$.

Contour sets are ellipses in $\mathbb{R}^M$.

The (infeasible) optimal weight vector minimizes $R_n(W)$ over $W \in \mathcal{H}_n$.

It necessarily puts non-zero weight on at least two models (unless $a_1 = a_m$).

To see this, suppose that $M = 2$ and WLOG $k_1 = 0$ and $a_2 = 0$. Then

$$R_n(W) = a_1 w_1^2 + k_2 w_2^2$$

which is uniquely minimized by

$$w_1 = \frac{\sigma^2 k_2}{a_1 + \sigma^2 k_2}$$
Weight Selection – existing approaches

How should the weight vector be selected in practice?

Bayesian Model Averaging (BMA). If $\pi_m$ are the prior probabilities then

$$w_m \propto \exp \left(-\frac{n}{2} BIC_m \right) \pi_m.$$  

If all models are given equal prior weight, then

$$w_m \propto \exp \left(-\frac{n}{2} BIC_m \right).$$

Smoothed AIC (SAIC) was introduced by Buckland et. al. (1997) and embraced by Burnham and Anderson (2002) and Hjort and Claeskens (2003)

$$w_m \propto \exp \left(-\frac{n}{2} AIC_m \right).$$

Hjort and Claeskens (2003) also discuss an Empirical Bayes Smoother.
**Mallows’ Criterion**

Define the residual vector

\[
\hat{e}(W) = Y - X_M \hat{\Theta}(W)
\]

and the average (or effective) number of parameters

\[
k(W) = \sum_{m=1}^{M} w_m k_m.
\]

The Mallow’s criterion for the model average estimator is

\[
C_n(W) = \hat{e}(W)' \hat{e}(W) + 2\sigma^2 k(W) = W' \bar{\hat{e}}' \bar{\hat{e}} W + 2\sigma^2 K' W
\]

where \( \bar{\hat{e}} = (\hat{e}_1, ..., \hat{e}_M) \) and \( K = (k_1, ..., k_M)' \).
The Mallows’ selected weight vector is

\[ \hat{W} = \arg \min_{W \in \mathcal{H}_n} C_n(W). \]

The solution is found numerically.

This is classic quadratic programming with boundary constraints, for which algorithms are widely available.
First Justification

The Mallows’ criterion is an unbiased estimate of the expected squared error.

**Lemma 3.** $EC_n(W) = EL_n(W) + n\sigma^2$

This is a classic known property of the Mallows’ criterion.
Second Justification

For some finite integer $N$.

Let the weights $w_m$ be restricted to $\{0, \frac{1}{N}, \frac{2}{N}, \cdots, 1\}$.

Let $\mathcal{H}_n(N)$ be the subset of $\mathcal{H}_n$ restricted to this set of weights.

Mallows’ weights restricted to discrete set $\mathcal{H}_n(N)$

$$\hat{W}_N = \arg\min_{W \in \mathcal{H}_n(N)} C_n(W)$$

These weights are optimal in the sense of asymptotic minimization of conditional squared error.
Theorem 1. As \( n \rightarrow \infty \), if

\[
\xi_n = \inf_{W \in \mathcal{H}_n} R_n(W) \rightarrow \infty
\]

almost surely, and

\[
E \left( |e_i|^{4(N+1)} | x_i \right) < \infty
\]

then

\[
\frac{L_n \left( \hat{W}_N \right)}{\inf_{W \in \mathcal{H}_n(N)} L_n(W)} \rightarrow_p 1.
\]

The proof is an application of Theorem 2.1 of Li (1987).

No restriction is placed on \( M \), the largest model.
Comment on the assumption
\[ \xi_n = \inf_{W \in \mathcal{H}_n} R_n(W) \to \infty \]
which is typical in the analysis of non-parametric models.

Recall
\[ R_n(W) = E(L_n(W) \mid X) \]
\[ L_n(W) = (\hat{\mu}(W) - \mu)'(\hat{\mu}(W) - \mu) \]
means that there is no correct finite dimensional model.

Otherwise, if there is a correct model \( m_0 \), then \( \inf_{W \in \mathcal{H}_n} R_n(W) = 0 \).

In this case, Mallows weights will not be optimal.
Technical Difficulty

• The proof of the theorem requires that the function $L_n(W)$ be uniformly well behaved as a function of the weight vector $W$.

• This requires a stochastic equicontinuity property over $W \in \mathcal{H}_n$.

• The difficulty is that the dimension of $\mathcal{H}_n$ grows with $M$, which is growing with $n$.

• This can be made less binding by selecting $N$ large.

• While $w_m$ is restricted to a set of discrete weights, there is no restriction on $M$, the largest model allowed.

• It would be desirable to allow for continuous weights $w_m$. 
Estimation of $\sigma^2$ for Mallows’ Criterion

The Mallows’ criterion depends on unknown $\sigma^2$.
It must be replaced with an estimate.
It is desirable for the estimate to have low bias.
One choice is

$$\hat{\sigma}_K^2 = \frac{1}{n-K} \left( Y - X_K\hat{\Theta}_K \right)' \left( Y - X_K\hat{\Theta}_K \right)$$

where $K$ corresponds to a “large” approximating model.

Theorem 2. If $K \to \infty$ and $K/n \to 0$ as $n \to \infty$ then $\hat{\sigma}_K^2 \to_p \sigma^2$. 
Simulation Experiment

\[ y_i = \sum_{j=1}^{\infty} \theta_j x_{ji} + e_i \]
\[ \theta_j = c \sqrt{2 \alpha} j^{-\alpha - 1/2} \]
\[ e_i \sim N(0, 1) \]

\( x_{1i} = 1 \). Remaining \( x_{ji} \) are iid \( N(0, 1) \).

The coefficient \( c \) is used to control \( R^2 = c^2/(1 + c^2) \)

\( n = 50, 150, 400, 1000 \)
\( \alpha = .5, 1.0 \) and 1.5

Larger \( \alpha \) implies that the coefficients \( \theta_j \) decline more quickly with \( j \).

\( M = 3n^{1/3} \). (maximum model order)
Methods

1. AIC model selection (AIC)

2. Mallows’ model selection (Mallows)

3. Smoothed AIC (S-AIC): \( w_m \propto \exp\left(-\frac{1}{2}AIC_m\right) \)

4. Smoothed BIC (S-BIC): \( w_m \propto \exp\left(-\frac{1}{2}BIC_m\right) \)

5. Mallows Model Averaging (MMA)

The estimators are compared by expected squared error (risk).

Averages across 100,000 simulations.

Risk normalized by dividing by infeasible best-fitting LS model \( m \)
Application

• U.S. Unemployment Rate Changes

• Monthly, 1995-2005 (120 Observations)

• AR(m) models, $0 \leq m \leq 12$

• AIC and Mallows select $\hat{m} = 3$

• $\hat{W}$ sets $\hat{w}_0 = .10, \hat{w}_1 = .17, \hat{w}_3 = .34, \hat{w}_9 = .39$
<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>AIC</th>
<th>MA</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta UR_{t-1}$</td>
<td>-.26</td>
<td>-.26</td>
<td>-.23</td>
</tr>
<tr>
<td>$\Delta UR_{t-2}$</td>
<td>.19</td>
<td>.21</td>
<td>.15</td>
</tr>
<tr>
<td>$\Delta UR_{t-3}$</td>
<td>.15</td>
<td>.19</td>
<td>.12</td>
</tr>
<tr>
<td>$\Delta UR_{t-4}$</td>
<td>-.01</td>
<td>-</td>
<td>-.01</td>
</tr>
<tr>
<td>$\Delta UR_{t-5}$</td>
<td>-.00</td>
<td>-</td>
<td>.01</td>
</tr>
<tr>
<td>$\Delta UR_{t-6}$</td>
<td>.16</td>
<td>-</td>
<td>.07</td>
</tr>
<tr>
<td>$\Delta UR_{t-7}$</td>
<td>.07</td>
<td>-</td>
<td>.03</td>
</tr>
<tr>
<td>$\Delta UR_{t-8}$</td>
<td>-.13</td>
<td>-</td>
<td>-.05</td>
</tr>
<tr>
<td>$\Delta UR_{t-9}$</td>
<td>.11</td>
<td>-</td>
<td>.06</td>
</tr>
<tr>
<td>$\Delta UR_{t-10}$</td>
<td>-.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\Delta UR_{t-11}$</td>
<td>.14</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\Delta UR_{t-12}$</td>
<td>.12</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Least Squares Forecast Averaging

Bruce E. Hansen

University of Wisconsin

March 2006
**Forecast Combination**

The goal is to construct a point forecast \( f_{n+1} \) of a target \( y_{n+1} \) given \( x_n \in \mathbb{R}^k \).

There are \( M \) models, each producing a forecast \( \hat{f}_{n+1}(m) \).

Model \( m \) has \( k_m \) free parameters.

A forecast combination takes the form

\[
\hat{f}_{n+1}(W) = \sum_{m=1}^{M} w_m \hat{f}_{n+1}(m)
\]

where \( W = (w_1, \ldots, w_m) \) are non-negative weights which satisfy

\[
\sum_{m=1}^{M} w_m = 1.
\]

Many methods have been proposed for the selection of weights.
Two Recent Successes:

(1) Equal Weights


\[ w_m = \frac{1}{M} \]

Excellent performance in empirical studies.

But inherently arbitrary.

Depends critically on the class of forecasts \( \hat{f}_{n+1}(m) \).

If a terrible model is included in the set, the forecast will suffer.

In this sense, “equal weights” is an incomplete forecast combination strategy.
(2) Bayesian Model Averaging (BMA)

Reference: Wright (2003ab).

BMA with uninformative priors approximately sets

$$w_m = \frac{\exp\left(-\frac{1}{2}BIC_m\right)}{\sum_{j=1}^{M} \exp\left(-\frac{1}{2}BIC_j\right)}.$$

where the BIC for model $m$ is

$$BIC_m = n \ln(\hat{\sigma}^2_m) + k_m \ln n.$$

Advantage: Excellent performance in empirical studies.

Deficiencies:

(a) Bayesian methods rely on priors, and thus are inherently arbitrary.

(b) BMA assumes that the truth is parametric. This paradigm is misspecified when models are approximations.
Mean-Squared Forecast Error

Define the forecast error $e_{n+1} = y_{n+1} - f_{n+1}$ and its variance $\sigma^2 = Ee_{n+1}^2$

The mean-squared forecast error is

$$MSE(W) = E \left(y_{n+1} - \hat{f}_{n+1}(W)\right)^2$$

$$= Ee_{n+1}^2 + E \left(\hat{f}_{n+1}(W) - f_{n+1}\right)^2$$

Thus minimizing the mean-square forecast error is the same as minimizing the estimated conditional mean.

I propose using the MMA weights for forecast combination among linear models.
Finite Sample Investigation – Static Model

Infinite-order iid regression

\[
y_i = \sum_{j=1}^{\infty} \theta_j x_{ji} + e_i
\]

\[
\theta_j = c \sqrt{2 \alpha} j^{-\alpha - 1/2}
\]

\[e_i \sim N(0, 1)\]

Remaining \(x_{1i} = 1\). Remaining \(x_{ji}\) are iid \(N(0, 1)\).

\[R^2 = c^2/(1 + c^2)\] is controlled by the parameter \(c\).

\(n = 50, 100, \text{ and } 200\).

\(\alpha \in \{.5, 1.0, 1.5, 2.0\}\).

Larger \(\alpha\) implies that the coefficients \(\theta_j\) decline more quickly with \(j\).

\(M = 3n^{1/3}\)
$n=50$, $\alpha=0.5$

$n=50$, $\alpha=1.0$

$n=50$, $\alpha=1.5$

$n=50$, $\alpha=2.0$
Finite Sample Investigation – Dynamic Model

Univariate time series $y_t$, generated by MA(m)

$$y_t = (1 + \psi L)^m e_t$$

$$e_t \sim N(0, 1)$$

$\psi \in [.1, .9]$  
$m \in \{1, 2, 3, 4\}$.  
Forecasting models: AR(p), $0 \leq p \leq M = 2n^{1/3}$

$n = 50, 100, \text{ and } 200$.  

Empirical Illustration

\[ y_t = \text{Unemployment Rate (Monthly)} \]

\[ x_t = \text{Interest Rate Spread (5 year Treasury - Fed Funds Rate)} \]

\textit{h-step-ahead Linear Forecasts}

\[ y_{t+h-1} = \alpha_0 + \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \beta_1 x_{t-1} + \cdots \beta_q x_{t-q} + e_t. \]

\[ p \in \{1, 2, 3, 4, 5, 6\} \]

\[ q \in \{0, 1, 2, 3, 4\} \]

30 Models

Sample Period: 1959:01 - 2006:02

Recursive Out-of-Sample Forecasts, Starting with 120 observations
Out-of-Sample Mean Squared Error
For Unemployment Rate Forecasts

<table>
<thead>
<tr>
<th></th>
<th>Equal</th>
<th>BIC Weights</th>
<th>MMA Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h = 1$</td>
<td>.88</td>
<td>.96</td>
<td>.89</td>
</tr>
<tr>
<td>$h = 3$</td>
<td>.85</td>
<td>.98</td>
<td>.90</td>
</tr>
<tr>
<td>$h = 6$</td>
<td>.87</td>
<td>.98</td>
<td>.90</td>
</tr>
<tr>
<td>$h = 12$</td>
<td>.85</td>
<td>.96</td>
<td>.91</td>
</tr>
</tbody>
</table>

Note: MSE is scaled relative to BIC selection
MMA Weights

6-step-ahead Forecast

Computed on Full Sample

<table>
<thead>
<tr>
<th>$q = 0$</th>
<th>$q = 1$</th>
<th>$q = 2$</th>
<th>$q = 3$</th>
<th>$q = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p = 1$</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>$p = 2$</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>$p = 3$</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>$p = 4$</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>$p = 5$</td>
<td>.02</td>
<td>.19</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>$p = 6$</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
</tbody>
</table>
### BIC Weights

**6-step-ahead Forecast**

Computed on Full Sample

<table>
<thead>
<tr>
<th></th>
<th>$q = 0$</th>
<th>$q = 1$</th>
<th>$q = 2$</th>
<th>$q = 3$</th>
<th>$q = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p = 1$</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>$p = 2$</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>$p = 3$</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.03</td>
</tr>
<tr>
<td>$p = 4$</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.01</td>
<td>.17</td>
</tr>
<tr>
<td>$p = 5$</td>
<td>.02</td>
<td>.04</td>
<td>.04</td>
<td>.08</td>
<td>.35</td>
</tr>
<tr>
<td>$p = 6$</td>
<td>.00</td>
<td>.03</td>
<td>.02</td>
<td>.04</td>
<td>.18</td>
</tr>
</tbody>
</table>
Summary

- Econometric models are approximations.
- Model selection should be guided by approximation accuracy.
- Squared error is a convenient criterion.
- Averaging estimates across models has the potential to improve estimation efficiency.
- Mallows Criterion may be used to select the model average weights.
- It has an asymptotic optimality property.
Cautionary Remarks

- The Mallows criterion assumes homoskedasticity.
- A cross-validation criterion may allows for heteroskedasticity (under investigation with Jeff Racine).
- Inference using model average estimates is difficult.
- There are no useful standard errors for the model average estimators.
- The estimates have non-standard distributions and are non-pivotal.
- It is unclear how to form confidence intervals from the estimates.