16 Model Averaging

16.1 Framework

Let g be a (non-parametric) object of interest, such as a conditional mean, variance, density, or distribution function. Let $\hat{g}_m$, $m = 1, ..., M$ be a discrete set of estimators. Most commonly, this set is the same as we might consider for the problem of model selection. In linear regression, typically $\hat{g}_m$ correspond to different sets of regressors. We will sometimes call the m'th estimator the m'th "model".

Let w_m be a set of weights for the m'th estimator. Let $\mathbf{w} = (w_1, ..., w_M)$ be the vector of weights. Typically we will require

$$\begin{aligned} 0 &\leq w_m \leq 1 \\ \sum_{m=1}^{M} w_m &= 1 \end{aligned}$$

The set of weights satisfying this condition is H_M, the unit simplex in $\mathbb{R}^M$.

An averaging estimator is

$$\hat{g}(\mathbf{w}) = \sum_{m=1}^{M} w_m \hat{g}_m$$

It is commonly called a "model average estimator".

Selection estimators are the special case where we impose the restriction $w_m \in \{0, 1\}$.

16.2 Model Weights

The most common method for weight specification is Bayesian Model Averaging (BMA). Assume that there are M potential models and one of the models is the true model. Specify prior probabilities that each of the potential models is the true model. For each model specify a prior over the parameters. Then the posterior distribution is the weighted average of the individual models, where the weights are Bayesian posterior probabilities that the given model is the true model, conditional on the data.

Given diffuse priors and equal model prior probabilities, the BMA weights are approximately

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

where

$$BIC_m = 2\mathcal{L}_m + k_m \log(n)$$

$\mathcal{L}_m$ is the negative log-likelihood, and k_m is the number of parameters in model m. BIC_m is the Bayesian information criterion for model m. It is similar to AIC, but with the "2" replaced by

log(n).

The BMA estimator has the nice interpretation as a Bayesian estimator. The downside is that it does not allow for misspecification. It is designed to search for the "true" model, not to select an estimator with low loss.

To remedy this situation, Burnham and Anderson have suggested replacing BIC with AIC, resulting in what has been called smoothed AIC (AIC) or weighted AIC (WAIC). The weights are

$$w_m = \frac{\exp\left(-\frac{1}{2}AIC_m\right)}{\sum_{j=1}^{M}\exp\left(-\frac{1}{2}AIC_j\right)}$$

where

$$AIC_m = 2\mathcal{L}_m + 2k_m$$

The suggestion goes back to Akaike, who suggested that these w_m may be interpreted as model probabilities. It is convenient and simple to implement. The idea can be applied quite broadly, in any context where AIC is defined.

In simulation studies, the SAIC estimator performs very well. (In particular, better than conventional AIC.) However, to date I have seen no formal justification for the procedure. It is unclear in what sense SAIC is producing a good approximation.

16.3 Linear Regression

In the case of linear regression, let X_m be regressor matrix for the m'th estimator. Then the list of all regressors. Then the m'th estimator is

$$\begin{aligned}\hat{\beta}_m &= \left(X_m'X_m\right)^{-1}X_m y \\ \hat{g}_m &= X_m\hat{\beta}_m \\ &= P_m y\end{aligned}$$

where

$$P_m = X_m\left(X_m'X_m\right)^{-1}X_m$$

The averaging estimator is

$$\begin{aligned}\hat{g}\left(\mathbf{w}\right) &= \sum_{m=1}^{M} w_m\hat{g}_m \\ &= \sum_{m=1}^{M} w_m P_m y \\ &= P\left(\mathbf{w}\right)y\end{aligned}$$

where

$$P(\mathbf{w}) = \sum_{m=1}^{M} w_m P_m$$

Let X be the matrix of all regressors. We can also write

$$\begin{aligned}
\hat{g}(\mathbf{w}) &= \sum_{m=1}^{M} w_m X_m \left(X_m' X_m\right)^{-1} X_m y \\
&= \sum_{m=1}^{M} w_m X_m \hat{\beta}_m \\
&= X \sum_{m=1}^{M} w_m \begin{pmatrix} \hat{\beta}_m \\ 0 \end{pmatrix} \\
&= X\hat{\beta}(\mathbf{w})
\end{aligned}$$

where

$$\hat{\beta}(\mathbf{w}) = \sum_{m=1}^{M} w_m \begin{pmatrix} \hat{\beta}_m \\ 0 \end{pmatrix}$$

is the average of the coefficient estimates. $\hat{\beta}(\mathbf{w})$ is the model average estimator for β. In linear regression, there is a direct correspondence between the average estimator for the conditional mean and the average estimator of the parameters, but this correspondence breaks down when the estimator is not linear in the parameters.

16.4 Mallows Weight Selection

As pointed out above, in the linear regression setting, $\hat{g}(\mathbf{w}) = P(\mathbf{w})\, y$ is a linear estimator, so falls in the class studied by Li (1987). His framework allows for estimators indexed by $\mathbf{w} \in H_M$

Under homoskedasticity, an optimal method for selection of $\mathbf{w}$ is the Mallows criterion. As we discussed before, for estimators $\hat{g}(\mathbf{w}) = P(\mathbf{w})\, y$, the Mallows criterion is

$$C(\mathbf{w}) = \hat{e}(\mathbf{w})' \hat{e}(\mathbf{w}) + 2\sigma^2 \operatorname{tr} P(\mathbf{w})$$

where

$$\hat{e}(\mathbf{w}) = y - \hat{g}(\mathbf{w})$$

is the residual.

In averaging linear regression

$$\begin{aligned}
\operatorname{tr} P(\mathbf{w}) &= \operatorname{tr} \sum_{m=1}^{M} w_m P_m \\
&= \sum_{m=1}^{M} w_m \operatorname{tr} P_m \\
&= \sum_{m=1}^{M} w_m k_m \\
&= \mathbf{w}'\mathbf{K}
\end{aligned}$$

where k_m is the number of coefficients in the m'th model, and $\mathbf{K} = (k_1, ..., k_M)'$. The penalty is twice $\mathbf{w}'\mathbf{K}$, the (weighted) average number of coefficients.

Also

$$\begin{aligned}
\hat{e}(\mathbf{w}) &= y - \hat{g}(\mathbf{w}) \\
&= \sum_{m=1}^{M} w_m (y - \hat{g}_m) \\
&= \sum_{m=1}^{M} w_m \hat{e}_m \\
&= \hat{\mathbf{e}}\mathbf{w}
\end{aligned}$$

where $\hat{e}_m$ is the $n \times 1$ residual vector from the m'th model, and $\hat{\mathbf{e}} = [\hat{e}_1, ..., \hat{e}_M]$ is the $n \times M$ matrix of residuals from all M models.

We can then write the criterion as

$$C(\mathbf{w}) = \mathbf{w}'\hat{\mathbf{e}}'\hat{\mathbf{e}}\mathbf{w} + 2\sigma^2\mathbf{w}'\mathbf{K}$$

This is quadratic in the vector $\mathbf{w}$.

The Mallows selected weight vector minimizes the criterion $C(\mathbf{w})$ over $\mathbf{w} \in H_M$, the unit simplex.

$$\hat{\mathbf{w}} = \underset{\mathbf{w} \in H_M}{\operatorname{argmin}} C(\mathbf{w})$$

This is a quadratic programming problem with inequality constraints, which is pre-programmed in Gauss and Matlab, so computation of $\hat{\mathbf{w}}$ is a simple command.

The Mallows selected estimator is then

$$\begin{aligned}
\hat{g} &= \hat{g}(\hat{\mathbf{w}}) \\
&= \sum_{m=1}^{M} \hat{w}_m \hat{g}_m
\end{aligned}$$

This is an

16.5 Weight Selection Optimality

As we discussed in the section on model selection, Li (1987) provided a set of sufficient conditions for the Mallows selected estimator to be optimal, in the sense that the squared error is asymptotically equivalent to the infeasible optimum. The key condition was

$$\sum_{\mathbf{w}} (nR(\mathbf{w}))^{-s} \to 0 \tag{1}$$

In Hansen (Econometrica, 2007), I show that this condition is satisfied if we restrict the set of weights to a discrete set.

Recall that H_M is the unit simplex in $\mathbb{R}^M$.

Now restrict $\mathbf{w} \in H_M^* \subset H_M$, where the weights in H_M^* are elements of $\{\frac{1}{N}, \frac{2}{N}, ..., 1\}$ for some integer N. In that paper, I show that Li's condition (1) over $\mathbf{w} \in H_M^*$ holds under the similar conditions as model selection, namely if the models are nested,

$$\xi_n = \inf_{\mathbf{w} \in H_M} nR(\mathbf{w}) \to \infty$$

and

$$E\left(e_i^{4(N+1)} \mid X_i\right) \leq \kappa < \infty.$$

Thus model averaging is asymptotically optimal, in the sense that

$$\frac{L(\hat{\mathbf{w}})}{\inf_{\mathbf{w} \in H_M^*} L(\mathbf{w})} \to_p 1$$

where, again

$$L(\mathbf{w}) = \frac{1}{n} \left(\hat{\mathbf{g}}\left(\mathbf{w}\right) - \mathbf{g}\right)' \left(\hat{\mathbf{g}}\left(\mathbf{w}\right) - \mathbf{g}\right)$$

The proof is similar to that for model selection in linear regression. The restriction of $\mathbf{w}$ to a discrete set is necessary to directly apply Li's theorem, as the summation requires discreteness.

The discreteness was relaxed in a paper by Wan, Zhang, and Zou (2008, Least Squares Model Combining by Mallows Criterion, working paper). Rather than proving (1), they provided a more basic derivation, although using stronger conditions. Recall that the proof requires showing uniform convergence results of the form

$$\sup_{\mathbf{w} \in H_M} \frac{|e' b(\mathbf{w})|}{nR(\mathbf{w})} \to_p 0$$

where

$$\begin{aligned} b(\mathbf{w}) &= \sum_{m=1}^{M} w_m b_m \\ b_m &= (I - P_m)\,\mathbf{g} \end{aligned}$$

Here is their proof: First,

$$\sup_{\mathbf{w}\in H_M} \frac{|e'b(\mathbf{w})|}{nR(\mathbf{w})} \leq \sum_{m=1}^{M} w_m \frac{|e'b_m|}{\xi_n} \leq \max_{1\leq m\leq M} \frac{|e'b_m|}{\xi_n}$$

Second, by Markov's and Whittle's inequalities

$$\begin{aligned} P\left(\max_{1\leq m\leq M} \frac{|e'b_m|}{\xi_n} > \delta\right) &\leq \sum_{m=1}^{M} P\left(\frac{|e'b_m|}{\xi_n} > \delta\right) \\ &\leq \sum_{m=1}^{M} \frac{E\,|e'b_m|^{2G}}{\delta^{2G}\xi_n^{2G}} \\ &\leq K \sum_{m=1}^{M} \frac{|b_m'b_m|^{G}}{\delta^{2G}\xi_n^{2G}} \\ &\leq K \sum_{m=1}^{M} \frac{\left(nR(\mathbf{w}_m^0)\right)^{G}}{\delta^{2G}\xi_n^{2G}} \end{aligned}$$

where $\mathbf{w}_m^0$ is the weight vector with a 1 in the m'th place and zeros elsewhere. Equivalently, $nR(\mathbf{w}_m^0)$ is the expected squared error from the m'th model. The final inequality uses the fact from the analysis for model selection that

$$nR(\mathbf{w}_m^0) = b_m'b_m + \sigma^2 k_m \leq b_m'b_m$$

Wan, Zhang, and Zou then assume

$$\frac{\sum_{m=1}^{M} \left(nR(\mathbf{w}_m^0)\right)^{G}}{\xi_n^{2G}} \to 0$$

This is stronger than the condition from my paper $\xi_n \to \infty$, as it requires that $\sum_{m=1}^{M} \left(nR(\mathbf{w}_m^0)\right)^{G}$ diverges slower than ξ_n^{2G}. They also do not directly assume that the models are nested.

16.6 Cross-Validation Selection

Hansen and Racine (Jacknife Model Averaging, working paper).

In this paper, we substitute CV for the Mallows criterion. As a result, we do not require homoskedasticity.

For the $m'th$ model, let $\tilde{e}_i^m$ denote the leave-one-out (LOO) residuals for the i'th observation, e.g.

$$\tilde{e}_i^m = y_i - X_i^{m\prime}\left(\mathbf{X}_{-i}^{m\prime}\mathbf{X}_{-i}^m\right)^{-1}\mathbf{X}_{-i}^{m\prime}\mathbf{y}_{-i}$$

and let $\tilde{e}_m$ denote the $n \times 1$ vector of the $\tilde{e}_i^m$. Then the LOO averaging residuals are

$$\tilde{e}_i(\mathbf{w}) = \sum_{m=1}^{M} w_m \tilde{e}_i^m$$

$$\tilde{e}(\mathbf{w}) = \sum_{m=1}^{M} w_m \tilde{e}^m = \tilde{\mathbf{e}}\mathbf{w}$$

where $\tilde{\mathbf{e}}$ is an $n \times M$ matrix whose mth column is $\tilde{e}_m$. Then the sum-of-squared LOO residuals is

$$CV(\mathbf{w}) = \tilde{e}(\mathbf{w})'\,\tilde{e}(\mathbf{w}) = \mathbf{w}'\tilde{e}'\tilde{e}\mathbf{w}$$

which is quadratic in $\mathbf{w}$.

The CV (or jacknife) selected weight vector $\hat{\mathbf{w}}$ minimizes the criterion $CV(\mathbf{w})$ over the unit simplex. As for Mallows selection, this is solved by quadratic programming.

The JMA estimator is then $\hat{g}(\hat{\mathbf{w}})$

In Hansen-Racine, we show that the CV estimator is asymptotically equivalent to the infeasible best weight vector, under the conditions

$$\begin{aligned}
0 &< \min_i E\left(e_i^2 \mid X_i\right) \leq \min_i E\left(e_i^2 \mid X_i\right) < \infty \\
E\left(e_i^{4(N+1)} \mid X_i\right) &\leq \kappa < \infty \\
\xi_n &= \inf_{\mathbf{w}\in H_M} nR(\mathbf{w}) \to \infty \\
\max_{1\leq m\leq M}\max_{1\leq i\leq n} X_i^{m\prime}\left(\mathbf{X}^{m\prime}\mathbf{X}^m\right)^{-1}\mathbf{X}^{m\prime}X_i^m &\to 0
\end{aligned}$$

16.7 Many Unsolved Issues

- Model averaging for other estimators: e.g. densities or conditional densities
- IV, GMM, EL, ET
- Standard errors?
- Inference