Model Averaging for Linear Regression

Erkki P. Liski

University of Tampere
Department of Mathematics, Statistics and Philosophy

Outline

- The Model
- Model selection
- Model average estimator (MAE)

Outline

- The Model
- Model selection
- Model average estimator (MAE)
- Why MAE?
- General structure of MAE

Outline

- The Model
- Model selection
- Model average estimator (MAE)
- Why MAE?
- General structure of MAE
- Selecting the model weights
- Finite Sample Performance

Homoscedastic linear regression

Variables The response y and the predictors $x_1, x_2, \ldots$

Homoscedastic linear regression

Variables The response y and the predictors $x_1, x_2, \ldots$

The Model

$$y = \mu + \varepsilon, \qquad \mu = \sum_{j=1}^{\infty} \beta_j x_j,$$

$$E(\varepsilon|\boldsymbol{x}) = 0, \qquad E(\varepsilon^2|\boldsymbol{x}) = \sigma^2,$$

$\beta_1, \beta_2, \ldots$ and σ^2 are unknown parameters, and $\boldsymbol{x} = (x_1, x_2, \ldots)$.

Homoscedastic linear regression

Variables The response y and the predictors $x_1, x_2, \ldots$

The Model

$$y = \mu + \varepsilon, \qquad \mu = \sum_{j=1}^{\infty} \beta_j x_j,$$

$$E(\varepsilon | \boldsymbol{x}) = 0, \qquad E(\varepsilon^2 | \boldsymbol{x}) = \sigma^2,$$

$\beta_1, \beta_2, \ldots$ and σ^2 are unknown parameters, and $\boldsymbol{x} = (x_1, x_2, \ldots)$.

Further

$E(\mu^2) < \infty$ and $\sum_{j=1}^{\infty} \beta_j x_j$ converges in mean-square.

Model Selection

Covariates K potential predictors $x_1, \ldots, x_K$ available.

Observe $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)$, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iK})$.

Model Selection

Covariates K potential predictors $x_1, \ldots, x_K$ available.

Observe $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)$, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iK})$.

Approximating Linear Model

$$y_i = \sum_{j=1}^{K} x_{ij}\beta_j + b_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n,$$

$$b_i = \sum_{j=K+1}^{\infty} \beta_j x_j \qquad \text{is the approximation error.}$$

Model Selection

Covariates K potential predictors $x_1, \ldots, x_K$ available.

Observe $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)$, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iK})$.

Approximating Linear Model

$$y_i = \sum_{j=1}^{K} x_{ij}\beta_j + b_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n,$$

$$b_i = \sum_{j=K+1}^{\infty} \beta_j x_j \qquad \text{is the approximation error.}$$

Multiple models are present.

Model m $\{x_i \mathbb{I}_{\{i \in m\}} | i = 1, 2, \ldots, K\} \subset \{1, 2, \ldots, K\}$.

A Class of Approximating Models $\mathcal{A}$

The $M \times K$ Incidence Matrix

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & 1 & \dots & 0 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & 1 & \dots & 0 & 1 \end{pmatrix} = \begin{pmatrix} \boldsymbol{a}_1^T \\ \vdots \\ \boldsymbol{a}_m^T \\ \vdots \\ \boldsymbol{a}_M^T \end{pmatrix}$$

for the models in $\mathcal{A}$. The 1's in row $\boldsymbol{a}_m$ display the predictors in the mth model.

A Class of Approximating Models $\mathcal{A}$

The $M \times K$ Incidence Matrix

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & 1 & \dots & 0 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & 1 & \dots & 0 & 1 \end{pmatrix} = \begin{pmatrix} \boldsymbol{a}_1^T \\ \vdots \\ \boldsymbol{a}_m^T \\ \vdots \\ \boldsymbol{a}_M^T \end{pmatrix}$$

for the models in $\mathcal{A}$. The 1's in row $\boldsymbol{a}_m$ display the predictors in the mth model.

The Regression Matrix of the Model m

$$\boldsymbol{X}_m = \boldsymbol{X}\,\mathrm{diag}(\boldsymbol{a}_m),$$

A Class of Approximating Models $\mathcal{A}$

The $M \times K$ Incidence Matrix

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & 1 & \dots & 0 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & 1 & \dots & 0 & 1 \end{pmatrix} = \begin{pmatrix} \boldsymbol{a}_1^T \\ \vdots \\ \boldsymbol{a}_m^T \\ \vdots \\ \boldsymbol{a}_M^T \end{pmatrix}$$

for the models in $\mathcal{A}$. The 1's in row $\boldsymbol{a}_m$ display the predictors in the mth model.

The Regression Matrix of the Model m

$$\boldsymbol{X}_m = \boldsymbol{X}\,\mathrm{diag}(\boldsymbol{a}_m),$$

$\boldsymbol{a}_m$ is the vector diagonal entries of $\mathrm{diag}(\boldsymbol{a}_m)$,
$\mathbf{X}$ denotes the $n \times K$ regression matrix.

Approximating Model m

takes the form

$$\boldsymbol{y} = \boldsymbol{X}_m \boldsymbol{\beta}_m + \boldsymbol{b}_m + \boldsymbol{\varepsilon}.$$

Approximating Model m

takes the form

$$\boldsymbol{y} = \boldsymbol{X}_m \boldsymbol{\beta}_m + \boldsymbol{b}_m + \boldsymbol{\varepsilon}.$$

The LSE of $\boldsymbol{\beta}_m$

$$\hat{\boldsymbol{\beta}}_m = (\boldsymbol{X}_m^T \boldsymbol{X}_m)^+ \boldsymbol{X}_m^T \boldsymbol{y}$$

and of $\boldsymbol{\mu}_m = \boldsymbol{X}_m \boldsymbol{\beta}_m$

$$\hat{\boldsymbol{\mu}}_m = \boldsymbol{H}_m \boldsymbol{y}$$

Approximating Model m

takes the form

$$\boldsymbol{y} = \boldsymbol{X}_m \boldsymbol{\beta}_m + \boldsymbol{b}_m + \boldsymbol{\varepsilon}.$$

The LSE of $\boldsymbol{\beta}_m$

$$\hat{\boldsymbol{\beta}}_m = (\boldsymbol{X}_m^T \boldsymbol{X}_m)^+ \boldsymbol{X}_m^T \boldsymbol{y}$$

and of $\boldsymbol{\mu}_m = \boldsymbol{X}_m \boldsymbol{\beta}_m$

$$\hat{\boldsymbol{\mu}}_m = \boldsymbol{H}_m \boldsymbol{y}$$

under $m \in M$, where

$$\boldsymbol{H}_m = \boldsymbol{X}_m (\boldsymbol{X}_m^T \boldsymbol{X}_m)^+ \boldsymbol{X}_m^T$$

is a projector.

Model Average Estimator (MAE)

- MAE is an alternative to model selection
- A model selection procedure can be unstable

Model Average Estimator (MAE)

- MAE is an alternative to model selection
- A model selection procedure can be unstable
- When is combining better than selection?
- How to measure the uncertainty in selection?

Model Average Estimator (MAE)

- MAE is an alternative to model selection
- A model selection procedure can be unstable
- When is combining better than selection?
- How to measure the uncertainty in selection?

Draper 1995 (JRSS B), Raftery, et. al. 1997 (JASA), Burnham & Anderson 2002 (Book), Hjort & Claeskens 2003 (JASA), Hansen 2007 (Econometrica)

Model Average Estimator (MAE)

- MAE is an alternative to model selection
- A model selection procedure can be unstable
- When is combining better than selection?
- How to measure the uncertainty in selection?

Draper 1995 (JRSS B), Raftery, et. al. 1997 (JASA), Burnham & Anderson 2002 (Book), Hjort & Claeskens 2003 (JASA), Hansen 2007 (Econometrica)

Model Average Estimator (MAE)

- MAE is an alternative to model selection
- A model selection procedure can be unstable
- When is combining better than selection?
- How to measure the uncertainty in selection?

Draper 1995 (JRSS B), Raftery, et. al. 1997 (JASA), Burnham & Anderson 2002 (Book), Hjort & Claeskens 2003 (JASA), Hansen 2007 (Econometrica)

MAE of $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$

$$\hat{\boldsymbol{\beta}}_w = \sum_{m=1}^{M} w_m \hat{\boldsymbol{\beta}}_m, \text{ weights } w_i \geq 0 \text{ with } \sum_{m=1}^{M} w_m = 1$$

$$\hat{\boldsymbol{\mu}}_w = \boldsymbol{H}_w \boldsymbol{y}, \quad \boldsymbol{H}_w = \sum_{m=1}^{M} w_m \boldsymbol{H}_m \text{ is the implied hat matrix.}$$

The Algebraic Structure of MAE

Define

$$\boldsymbol{A}\boldsymbol{A}^T = \begin{pmatrix} k_1 & k_{12} & \ldots & k_{1M} \\ k_{21} & k_2 & \ldots & k_{2M} \\ \vdots & \vdots & \ddots & \\ k_{M1} & k_{M2} & \ldots & k_M \end{pmatrix} = \boldsymbol{K}$$

The Algebraic Structure of MAE

Define

$$\boldsymbol{A}\boldsymbol{A}^T = \begin{pmatrix} k_1 & k_{12} & \dots & k_{1M} \\ k_{21} & k_2 & \dots & k_{2M} \\ \vdots & \vdots & \ddots & \\ k_{M1} & k_{M2} & \dots & k_M \end{pmatrix} = \boldsymbol{K}$$

Properties of $\boldsymbol{H}_w$, model weights $\boldsymbol{w}^T = (w_i, \dots, w_M)$.

(i) $\mathrm{tr}(\boldsymbol{H}_w) = \sum_{m=1}^{M} w_m k_m$.

(ii) $\mathrm{tr}(\boldsymbol{H}_w^2) = \boldsymbol{w}^T \boldsymbol{K} \boldsymbol{w}$.

(iii) $\lambda_M(\boldsymbol{H}_w) \leq 1$.

The Risk under Squared-Error Loss

Squared-Error Loss $L(\boldsymbol{w}) = \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_w\|^2.$

The Risk under Squared-Error Loss

Squared-Error Loss $L(\boldsymbol{w}) = \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_w\|^2$.

The Conditional Risk of $\hat{\boldsymbol{\mu}}_w$

$$
\begin{aligned}
R(\boldsymbol{w}) &= E(L(\boldsymbol{w})| \ \boldsymbol{x}_1 \ldots \boldsymbol{x}_n) \\
&= \|(\boldsymbol{I} - \boldsymbol{H}_w)\boldsymbol{\mu}\|^2 + \sigma^2 \boldsymbol{w}^T \boldsymbol{K} \boldsymbol{w} \\
&= \boldsymbol{w}^T (\boldsymbol{B} + \sigma^2 \boldsymbol{K}) \boldsymbol{w},
\end{aligned}
$$

The Risk under Squared-Error Loss

Squared-Error Loss $L(\boldsymbol{w}) = \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_w\|^2$.

The Conditional Risk of $\hat{\boldsymbol{\mu}}_w$

$$\begin{aligned} R(\boldsymbol{w}) &= E(L(\boldsymbol{w})|\ \boldsymbol{x}_1 \dots \boldsymbol{x}_n) \\ &= \|(\boldsymbol{I} - \boldsymbol{H}_w)\boldsymbol{\mu}\|^2 + \sigma^2 \boldsymbol{w}^T \boldsymbol{K} \boldsymbol{w} \\ &= \boldsymbol{w}^T (\boldsymbol{B} + \sigma^2 \boldsymbol{K}) \boldsymbol{w}, \end{aligned}$$

$$\boldsymbol{B} = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1M} \\ \vdots & \vdots & \ddots & \\ b_{M1} & b_{M2} & \dots & b_{MM} \end{pmatrix}, \text{ with } b_{mk} = \boldsymbol{b}_m^T (\boldsymbol{I} - \boldsymbol{H}_m)(\boldsymbol{I} - \boldsymbol{H}_k) \boldsymbol{b}_k.$$

The Risk under Squared-Error Loss

Squared-Error Loss $L(\boldsymbol{w}) = \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_w\|^2$.

The Conditional Risk of $\hat{\boldsymbol{\mu}}_w$

$$\begin{aligned} R(\boldsymbol{w}) &= E(L(\boldsymbol{w})|\, \boldsymbol{x}_1 \ldots \boldsymbol{x}_n) \\ &= \|(\boldsymbol{I} - \boldsymbol{H}_w)\boldsymbol{\mu}\|^2 + \sigma^2 \boldsymbol{w}^T \boldsymbol{K} \boldsymbol{w} \\ &= \boldsymbol{w}^T (\boldsymbol{B} + \sigma^2 \boldsymbol{K}) \boldsymbol{w}, \end{aligned}$$

$$\boldsymbol{B} = \begin{pmatrix} b_{11} & b_{12} & \ldots & b_{1M} \\ \vdots & \vdots & \ddots & \\ b_{M1} & b_{M2} & \ldots & b_{MM} \end{pmatrix}, \text{ with } b_{mk} = \boldsymbol{b}_m^T (\boldsymbol{I} - \boldsymbol{H}_m)(\boldsymbol{I} - \boldsymbol{H}_k) \boldsymbol{b}_k.$$

At least two non-zero w_i in the optimal $\boldsymbol{w}$.

Example: Suppose $M = 2$, $\boldsymbol{w}^T = (w, 1 - w)$. Then $w \in (0, 1)$ unless $b_{11} = b_{12}$ or $b_{22} = b_{12}$.

Selecting the Model Weights w_i

Mallows' Criterion (MMAE) for MAE (Hansen 2007)

$$C(\boldsymbol{w}) = \|(\boldsymbol{I} - \boldsymbol{H}_w)\boldsymbol{y}\|^2 + 2\sigma^2 k_w, \qquad k_w = \sum_{m=1}^{M} w_m k_m,$$

Selecting the Model Weights w_i

Mallows' Criterion (MMAE) for MAE (Hansen 2007)

$$C(\boldsymbol{w}) = \|(\boldsymbol{I} - \boldsymbol{H}_w)\boldsymbol{y}\|^2 + 2\sigma^2 k_w, \qquad k_w = \sum_{m=1}^{M} w_m k_m,$$

σ^2 is replaced with an estimate. Select $\hat{\boldsymbol{w}}$ such that

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmin}}\, C(\boldsymbol{w}).$$

Selecting the Model Weights w_i

Mallows' Criterion (MMAE) for MAE (Hansen 2007)

$$C(\boldsymbol{w}) = \|(\boldsymbol{I} - \boldsymbol{H}_w)\boldsymbol{y}\|^2 + 2\sigma^2 k_w, \qquad k_w = \sum_{m=1}^{M} w_m k_m,$$

σ^2 is replaced with an estimate. Select $\hat{\boldsymbol{w}}$ such that

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} C(\boldsymbol{w}).$$

Properties of $C(\boldsymbol{w})$: $E[C(\boldsymbol{w})] = E[L(\boldsymbol{w})] + n\sigma^2$ and

$$\frac{L(\hat{\boldsymbol{w}})}{\inf_{\boldsymbol{w}} L(\boldsymbol{w})} \longrightarrow_p 1 \qquad \text{as } n \to \infty.$$

Smoothed AIC **and** BIC **(**SAIC **&** SBIC**)**

$w_m = \exp(-\frac{1}{2}\mathrm{AIC}_m)/\sum_{i=1}^{M}\exp(-\frac{1}{2}\mathrm{AIC}_i)$

(*Bucland 1997, Burnham & Anderson 2002*),

$w_m = \exp(-\frac{1}{2}\mathrm{BIC}_m)/\sum_{i=1}^{M}\exp(-\frac{1}{2}\mathrm{BIC}_i)$,

Smoothed AIC **and** BIC **(***S*AIC **&** *S*BIC**)**

$$w_m = \exp(-\tfrac{1}{2}\mathrm{AIC}_m)/\textstyle\sum_{i=1}^{M}\exp(-\tfrac{1}{2}\mathrm{AIC}_i)$$

(*Bucland 1997, Burnham & Anderson 2002*),

$$w_m = \exp(-\tfrac{1}{2}\mathrm{BIC}_m)/\textstyle\sum_{i=1}^{M}\exp(-\tfrac{1}{2}\mathrm{BIC}_i),$$

the AIC and BIC criteria for model m are

$$\mathrm{AIC}_m = \ln\hat{\sigma}_m^2 + 2k_m \quad \text{and} \quad \mathrm{BIC}_m = \ln\hat{\sigma}_m^2 + k_m \ln n.$$

Smoothed AIC **and** BIC **(**SAIC **&** SBIC**)**

$$w_m = \exp(-\tfrac{1}{2}\mathrm{AIC}_m) / \sum_{i=1}^{M} \exp(-\tfrac{1}{2}\mathrm{AIC}_i)$$

(*Bucland 1997, Burnham & Anderson 2002*),

$$w_m = \exp(-\tfrac{1}{2}\mathrm{BIC}_m) / \sum_{i=1}^{M} \exp(-\tfrac{1}{2}\mathrm{BIC}_i),$$

the AIC and BIC criteria for model m are

$$\mathrm{AIC}_m = \ln \hat{\sigma}_m^2 + 2k_m \qquad \text{and} \qquad \mathrm{BIC}_m = \ln \hat{\sigma}_m^2 + k_m \ln n.$$

Smoothed MDL **(**SMDL**)**

$$w_m = \exp(-\mathrm{MDL}_m) / \sum_{i=1}^{M} \exp(-\mathrm{MDL}_i), \qquad \text{where}$$

Smoothed AIC **and** BIC **(**SAIC **&** SBIC**)**

$$w_m = \exp(-\tfrac{1}{2}\mathrm{AIC}_m)/\textstyle\sum_{i=1}^{M} \exp(-\tfrac{1}{2}\mathrm{AIC}_i)$$

(*Bucland 1997, Burnham & Anderson 2002*),

$$w_m = \exp(-\tfrac{1}{2}\mathrm{BIC}_m)/\textstyle\sum_{i=1}^{M} \exp(-\tfrac{1}{2}\mathrm{BIC}_i),$$

the AIC and BIC criteria for model m are

$$\mathrm{AIC}_m = \ln\hat{\sigma}_m^2 + 2k_m \qquad \text{and} \qquad \mathrm{BIC}_m = \ln\hat{\sigma}_m^2 + k_m \ln n.$$

Smoothed MDL **(**SMDL**)**

$$w_m = \exp(-\mathrm{MDL}_m)/\sum_{i=1}^{M} \exp(-\mathrm{MDL}_i), \qquad \text{where}$$

$$\mathrm{MDL}_m = n\ln\hat{s}_m^2 + k_m \ln F_m + \ln[k_m(n-k_m)], \quad F_m = \|\hat{\boldsymbol{\mu}}_m\|^2/k_m\hat{s}_m^2$$

(*Rissanen 2000 & 2007, Liski 2006*)

Finite Sample Performance

Simulation Model is the infinite order regression

$$y_i = \sum_{j=1}^{\infty} \beta_j x_{ji} + \varepsilon_i,$$

- $x_{ji} \sim \mathrm{N}(0,1)$ iid ($x_{1i} = 1$), $\varepsilon_i \sim \mathrm{N}(0,1)$ and $x_{ji} \perp\!\!\!\perp \varepsilon_i$.

Finite Sample Performance

Simulation Model is the infinite order regression

$$y_i = \sum_{j=1}^{\infty} \beta_j x_{ji} + \varepsilon_i,$$

- $x_{ji} \sim \mathrm{N}(0,1)$ iid ($x_{1i} = 1$), $\varepsilon_i \sim \mathrm{N}(0,1)$ and $x_{ji} \perp\!\!\!\perp \varepsilon_i$.
- $\beta_j = c\sqrt{2\alpha}\, j^{-\alpha-1/2}$ and the population $R^2 = \frac{c^2}{1+c^2}$.
- $50 \leq n \leq 1000$ and $M = 3n^{1/3}$.
- $0.5 \leq \alpha \leq 1.5$, for larger α the coefficients β_j decline more quicly.
- c is selected such that $0.1 \leq R^2 \leq 0.9$.

Finite Sample Performance

Simulation Model is the infinite order regression

$$y_i = \sum_{j=1}^{\infty} \beta_j x_{ji} + \varepsilon_i,$$

- $x_{ji} \sim \mathrm{N}(0,1)$ iid ($x_{1i} = 1$), $\varepsilon_i \sim \mathrm{N}(0,1)$ and $x_{ji} \perp\!\!\!\perp \varepsilon_i$.
- $\beta_j = c\sqrt{2\alpha}\, j^{-\alpha-1/2}$ and the population $R^2 = \frac{c^2}{1+c^2}$.
- $50 \leq n \leq 1000$ and $M = 3n^{1/3}$.
- $0.5 \leq \alpha \leq 1.5$, for larger α the coefficients β_j decline more quicly.
- c is selected such that $0.1 \leq R^2 \leq 0.9$.

Mean of predictive loss $\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_w\|^2$ over simulations.

Simulation Results

Comments

- AIC and Mallows' C similar, MMAE better than SAIC.

Simulation Results

Comments

- AIC and Mallows' C similar, MMAE better than SAIC.
- SAIC has lower risk than AIC

Simulation Results

Comments

- AIC and Mallows' C similar, MMAE better than SAIC.
- SAIC has lower risk than AIC
- MMAE better than SBIC in most cases.

Method	Performs well for
SBIC	n and R^2 small, α large
BIC	'small' models
AIC	'large´ models

Simulation Results

Comments

- AIC and Mallows' C similar, MMAE better than SAIC.
- SAIC has lower risk than AIC
- MMAE better than SBIC in most cases.

Method	Performs well for
SBIC	n and R^2 small, α large
BIC	'small' models
AIC	'large´ models

- SMDL better than MDL.

Simulation Results

Comments

- AIC and Mallows' C similar, MMAE better than SAIC.
- SAIC has lower risk than AIC
- MMAE better than SBIC in most cases.

Method	Performs well for
SBIC	n and R^2 small, α large
BIC	'small' models
AIC	'large´ models

- SMDL better than MDL.
- SMDL emulates the best performance criterion.

Simulation Results

Comments

- AIC and Mallows' C similar, MMAE better than SAIC.
- SAIC has lower risk than AIC
- MMAE better than SBIC in most cases.

Method	Performs well for
SBIC	n and R^2 small, α large
BIC	'small' models
AIC	'large´ models

- SMDL better than MDL.
- SMDL emulates the best performance criterion.
- SMDL has the best overall performance.

References

Buckland, S. T., Burnham, K. P. and Augustin, N. H. (1997), Model Selection: An Integral Part of Inference. *Biometrics*, 53, 603–618.

Burnham & Anderson (2002), *Model Selection and Multi-model Inference*, Springer

Draper, D. (1995), Assesment and Propagation of Model Uncertainty. *Journal of the Royal Statistical Society B*, 57, 45–70.

Hansen, B. E. (2007), Least Squares Model Averaging. *Econometrica*, Forthcoming.

Hjort, L. H. and Claeskens, G. (2003), Frequentist Model Average Estimators. *Journal of the American Statistical Association*, 98, 879–899.

- Liski, E. P. (2006), Normalized ML and the MDL Principle for Variable Selection in Linear Regression In: *Festschrift for Tarmo Pukkila on His 60th Birthday*, 159-172.
- Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997), Bayesian Model Averaging for Regression Models. *Journal of the American Statistical Association*, 92, 179–191.
- Rissanen, J. (2000). MDL Denoising. *IEEE Trans. Information Theory*, IT-46, pp. 2537–2543.
- Rissanen, J. (2007), *Information and Complexity in Statistical Modeling*, Springer