SEMIPARAMETRIC MODEL SELECTION WITH APPLICATIONS TO REGRESSION

Zhanlue Zhao *Huimin Chen* *X. Rong Li*

Department of Electrical Engineering
University of New Orleans, New Orleans, LA 70148
Email: {zzhao1, hchen2, xli}@uno.edu

ABSTRACT

In this paper we consider model selection problem using samples of small or moderate size where each model can have unknown parameter without a fully specified likelihood function. A semiparametric model selection criterion is proposed where the penalty-based model complexity term is used for the parameter with fully specified model structure and the kernel density estimation is used for the unknown noise distribution. A linear regression problem with various noise distributions is studied and the numerical results reveal that the semiparametric approach outperforms the penalty-based criteria and cross validation.

1. INTRODUCTION

In this paper we consider model selection using samples of a small or moderate size. When the likelihood function of each model has an explicit analytical form, the model selection problem is well studied and many approaches have been proposed. The model selection criterion usually introduces a penalty term to the best fitted likelihood function. Akaike information criterion (AIC) [1], Bayesian information criterion (BIC)[15], minimum description length (MDL) [13] and other criteria choose different penalty terms with various justifications based on certain distance measures between the selected model and the truth in the asymptotic regime. However, their performance with finite data records (and, in particular, short data records, which is the case of our interest) is unclear and the model selection performance is in general application dependent. Furthermore, when the likelihood function of each model is not fully specified, it is difficult, if not impossible, to derive a meaningful penalty based model selection criterion. In such cases, nonparametric approaches have been proposed among which the partition of data set into the training and testing subsets is often needed. Data oriented model selection such as cross validation [17], bootstrap [6], boosting [8], bagging [4] and their variants are applicable to models with likelihood functions unspecified. However, it has several disadvantages: (a) unreliable model validation when the data set is small, (b) self prone to overfitting if a large number of models is compared, (c) heavy computational burden due to various resampling routines. In contrast, penalty based approaches use the same data for training and validation but penalize models which are likely to overfit. Bayesian model selection [11], structural risk minimization [19], and PAC learning [10] fall largely into this category where the major issue is to design appropriate penalty term for a general model class.

When a model contains both unknown parameters with fully specified likelihood function and parameters without the likelihood function in a closed form, neither penalty based approaches nor data oriented approaches may be appropriate to the model selection problem. We propose a semiparametric model selection criterion where a kernel based approach is used to handle the nonparametric density estimation and the penalty term using either BIC or MDL is applied to the parametric likelihood function. We use a regression example with a linear observation model but unknown noise distribution to compare the model order selection accuracy. The results indicate a performance improvement using the semiparametric model selection criterion compared with cross validation and traditional penalty based criteria.

2. PARAMETRIC MODEL SELECTION

Consider a class of models $M_1, ..., M_K$ where model M_j assumes that the observation z is governed by a likelihood function $f_j(z|\theta_j)$, depending on the unknown parameter θ_j of dimension p_j $(j = 1, ..., K)$. Given z^n, a set of n independent observations, one needs to decide which model (associated with a particular value of the unknown parameter) characterizes the data best.

In a Bayesian setting, the prior distribution $p(\theta_j)$ of θ_j is assumed to be known. The parametric model associated with θ_j of dimension p_j is given by the likelihood function $f_j(z^n|\theta_j)$ and the likelihood function conditioned on model

Research Supported in part by ARO grant W911NF-04-1-0274, NASA/LEQSF grant (2001-4)-01 and UNO Office of Research Sponsored Program (ORSP) Investing in Research Excellence (IRE) 2005-06.

M_j is obtained as follows

$$f_j(z^n) = \int_{\Theta_j} f_j(z^n|\theta_j)p(\theta_j)d\theta_j$$

According to the maximum a posteriori (MAP) criterion [3], we choose model M_i if

$$i = \arg\max_j f_j(z^n)P(M_j) \quad (1)$$

where $P(M_j)$ is the prior probability of the model M_j associated with θ_j of dimension p_j. However, the unknown parameter θ_j is marginalized without using any observation.

In a non-Bayesian setting, one uses the data to estimate θ_j for model M_j. The model selection criterion is

$$i = \arg\max_j f_j(z^n|\hat{\theta}_j) \quad (2)$$

where the maximum likelihood estimate of θ_j is

$$\hat{\theta}_j = \arg\max_{\theta_j} f_j(z^n|\theta_j)$$

However, it has a tendency of fitting data with the most complicated model, which causes the overfitting problem. The penalty-based model selection criteria is devised to control the model complexity. It can be written in a general form as

$$i = \arg\min_j(-\log f_j(z^n|\hat{\theta}_j) + d_j(z^n)), \quad j = 1, ..., K \quad (3)$$

where $\hat{\theta}_j$ is the maximum likelihood estimate of θ_j and the penalty $d_j(z^n)$ represents the model complexity that varies for different criteria. For example, the Akaike information criterion (AIC) uses $d_j(z^n) = p_i$. Bayesian information criterion (BIC) uses $d_j(z^n) = p_i \log n/2$. The minimum description length (MDL) or stochastic information criterion (SIC) includes not only the penalty on the dimension of the unknown parameter but also the penalty associated with the Fisher information of the parameter, which has the asymptotic expansion given by [12, 13]

$$d_j(z^n) = \frac{p_j}{2}\log\left(\frac{n}{2\pi}\right) + \log\int |I(\theta_j)|^{1/2}d\theta_j \quad (4)$$

where $I(\theta_j)$ is the Fisher information matrix of a single observation.

3. SEMIPARAMETRIC MODEL SELECTION

3.1. Problem Formulation

In the above formulation, we assume that $f_j(z^n|\theta_j)$ has a known parametric form. However, in certain applications, a model may not have a fully specified likelihood function. In those cases, one needs to use data oriented model selection such as cross validation, bootstrap, boosting, and bagging to choose the best model in order to generalize well for unseen data. Error bounds can also be obtained for the worst case distribution using those data oriented techniques. An important issue is that certain prior knowledge of θ_j is ignored by those techniques especially when the parametric form of partial likelihood function is available which can not be inferred from the data.

In statistical modeling, $f_j(z^n|\theta_j)$ is nothing but a parametric model to characterize the given data z^n. In order to find the best model for a given data set, either a parametric or a nonparametric method can be used. In particular, given the samples of a small sample size, we know that the "best" model to characterize the data is its empirical distribution with δ bandwidth in the sense that it will always be supported by the data through hypothesis testing. For model selection, each model characterizes the data with a certain fitting error. To measure the difference carried by these errors is another recourse to take rather than to measure the difference by assuming a parametric model. Since the empirical distribution with δ bandwidth has no generalization capability for offset data, it is reasonable to assume that each observation carries its probability mass around its vicinity with a suitable bandwidth such that the data set can be meaningfully generalized. This motivates us to use the kernel density estimate to represent a small sample set instead of specifying a parametric model.

We assume that θ_j=$[\theta_{0j}^T\ \theta_{1j}^T]^T$ where $f_j(z^n|\theta_{0j})$ has a known parametric form if θ_{1j} is known. The parameter θ_{1j} does not have the likelihood function or distribution in a closed form. Note that both θ_{0j} and θ_{1j} need to be estimated from data as opposed to the marginalization using only the prior in the Bayesian approach. For example, θ_{1j} can be the estimation error with unknown distribution and $f_j(z^n|\theta_{0j})$ is fully specified for a known observation model. We call the model **semiparametric** to highlight the partial knowledge of the likelihood function. By applying the kernel density estimation to estimate $f_j(z^n|\theta_{0j})$ and using the penalty terms in the penalty-based parametric model selection criteria, the proposed semiparametric model selection criterion chooses model i if

$$i = \arg\min_j(-\log \hat{f}_j(z^n|\hat{\theta}_{0j}) + d_j(z^n)), \quad j = 1, ..., K \quad (5)$$

where $\hat{\theta}_{0j}$ is the maximum likelihood estimate of θ_{0j} assuming f_j is known. Under the Gaussian assumption, we use the least square estimate to efficiently compute $\hat{\theta}_{0j}$. $\hat{f}_j(\cdot)$ is the distribution of θ_{1j} obtained using the kernel density estimate to be explained next. The model complexity term $d_j(z^n)$ can be either BIC or MDL penalty. Alternatively, one can use data oriented approach to validate only the kernel density estimate $\hat{f}_j(\cdot)$.

3.2. KERNEL DENSITY ESTIMATION

We mainly adopt the results in [7] on kernel density estimation which has been extensively studied for decades. Given a data set of $\{X_i\}, i = 1, ..., N$, assuming each datum point carries the probability mass $1/N$ around its vicinity by a kernel function $K(\cdot)$, which is usually a nonnegative, symmetric, unimodal, smooth function. Then, the kernel density estimate is given by

$$\hat{f}_h(x) = \frac{1}{N}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{X_i - x}{h}\right) \tag{6}$$

where $\int_{-\infty}^{\infty} K(t)dt = 1$, $K(t) \geq 0, \forall\ t$, and h is a bandwidth parameter representing the window size.

The remaining two main tasks are to select the kernel function $K(\cdot)$ and the bandwidth h. Commonly used kernel functions include the Gaussian kernel

$$K(u) = (\sqrt{2\pi})^{-1}\exp(-u^2/2) \tag{7}$$

and the symmetric Beta family

$$K_r(u) = \frac{1}{\text{Beta}(1/2, \gamma+1)}(1-u^2)^{\gamma}I(|u| \leq 1) \tag{8}$$

The choices $\gamma = 0, 1, 2$ and 3 correspond to the uniform, the Epanechnikov, the biweight, and the triweight kernel functions, respectively. Note that different kernel functions have different support. Thus, even with the same bandwidth, different kernels use different amounts of information provided by the local data points around each observation. However, it is well-known both empirically and theoretically that the choice of kernel functions is not very important to the kernel density estimator. As long as they are symmetrical and unimodal, the resulting kernel density estimator performs nearly the same when the bandwidth h is optimally chosen in terms of minimizing the mean integrated square error (MISE) defined by

$$MISE = E\int_{-\infty}^{\infty}\{\hat{p}_h(x) - p(x)\}^2 dx \tag{9}$$

where $p(x)$ is a density and $\hat{p}$ is its estimator with the bandwidth h. The details can be found in [7]. Marron and Nolan also introduced the concept of canonical kernels to relate the equivalent amount of smoothing using two different kernels: Kernel K_2 using bandwidth h_2 performs nearly the same as kernel K_1 using the bandwidth

$$h_1 = \frac{\alpha(K_1)}{\alpha(K_2)}h_2 \tag{10}$$

where $\alpha(K) = \mu(K)^{-5/2}\,||K||_2^{2/5}$, $\mu(K) = \int_{-\infty}^{\infty} x^2K(x)dx$ is the variance of K and $||K||_2 = \int_{-\infty}^{\infty} K(x)^2dx$ is the L_2-norm.

The optimal bandwidth is selected by minimizing the MISE (9). When $p(x)$ is a Gaussian density with standard deviation σ, the optimal bandwidth is

$$h_{\text{opt},N} = (8\sqrt{\pi}/3)^{1/5}\alpha(K)\sigma N^{-1/5} \tag{11}$$

After replacing σ with the sample covariance and calculating the constant $\alpha(K)$ numerically, the optimal window width is

$$h = 1.06sn^{-1/5}, \tag{12}$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ is the sample variance and $\bar{X} = \frac{1}{n-1}\sum_{i=1}^{n} X_i$.

The above bandwidth selection criterion is only a simple rule of thumb and it often works well when the data are nearly Gaussian distributed. It may lead to oversmoothing when the underlying distribution is asymmetric or multimodal. In those cases, one has to tune the bandwidth with more sophisticated criteria such as plug-in bandwidth selectors [7].

4. APPLICATION TO LINEAR REGRESSION

We consider the following linear regression problem with model M_i being specified by

$$\mathbf{z} = \mathbf{H}_i\theta_i + \mathbf{v}_i \tag{13}$$

where $\mathbf{z}$ is the observation of an $n \times 1$ vector; $\mathbf{H}_i$ is a known $n \times p_i$ vector; θ_i is an unknown $p_i \times 1$ vector; $\mathbf{v}_i$ is the noise vector, the element of which is assumed to be independent, identically distributed with an unknown distribution. We want to find the best order p_i among different linear models. When the noise is Gaussian, the penalty based model selection criteria can be directly applied [5]. In this problem, we need to estimate $f_i(\mathbf{z}|\theta_i)$ in order to apply the penalty based criteria.

Since $f_i(\mathbf{z}|\theta_i) = f_{\mathbf{v}_i}(\mathbf{z} - \mathbf{H}_i\theta_i)$, we only need to estimate the density of $\mathbf{v}_i$ which is associated with model M_i. Specifically, for model M_i, $\mathbf{v}_i$ is an $n \times 1$ vector, the value of $f_i(\mathbf{z}(k)|\theta_i) = f_{\mathbf{v}_i}(\mathbf{v}_i(k))$, where $\mathbf{z}(k)$ is the kth element of $\mathbf{z}$ and $\mathbf{v}_i(k)$ is the kth element of $\mathbf{v}_i$. The kernel density estimate of f_i has the form

$$\begin{aligned} f_i(z(k)|\theta_i) &= f_{\mathbf{v}_i}(\mathbf{v}_i(k)) \\ &= \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{\mathbf{v}_i(k) - \mathbf{v}_i(j)}{h}\right) \end{aligned} \tag{14}$$

where $K(*)$ is a kernel function; h is the window width. Since θ_i is unknown, we have to use an estimated value $\hat{\theta}_i$ in the kernel density estimation. Here, we use the least squares estimate of θ_i, i.e,

$$\hat{\theta}_i = \left(\mathbf{H}_i^T\mathbf{H}_i\right)^{-1}\mathbf{H}_i^T\mathbf{z}_i$$

We use the Guassian kernel and the following bandwidth estimator proposed by Silverman ([16], pp. 45-47) to control possible over-smoothing

$$h = 0.9\hat{\sigma}/n^{-1/5} \tag{15}$$

where $\hat{\sigma} = \min(s, R/1.34)$, s is the sample standard deviation as in (12) and R is the interquartile range of the data. The constant 1.34 is derived from the fact that for a Gaussian distribution $N(z;\mu,\sigma^2)$, $P\{|z-\mu| < 1.34\sigma\} = 0.5$. The semiparametric model selection criterion is given by

$$\min_i(-\log \hat{f}_i(\mathbf{z}|\hat{\theta}_i) + d(\mathbf{z})), \quad i = 1, ..., K \tag{16}$$

We consider two penalty based criteria, namely, BIC and SIC, due to their popular usage in linear regression. When the noise is Gaussian, the BIC and SIC model selection criteria are [2]

$$\text{BIC}(p_i) = \frac{n}{2}\log R_i + \frac{p_i+1}{2}\log n$$

$$\text{SIC}(p_i) = \frac{n-p_i-2}{2}\log R_i + \frac{p_i+1}{2}\log n + \frac{1}{2}\log\left|\mathbf{H}_i^T\mathbf{H}_i\right|$$

where $R_i = \left|\left|\mathbf{z} - \mathbf{H}_i\hat{\theta}_i\right|\right|^2$. We compare the penalty based model selection criteria using BIC and SIC throughout the simulations with the proposed semiparametric criterion. A Gaussian kernel is used in the density estimation.

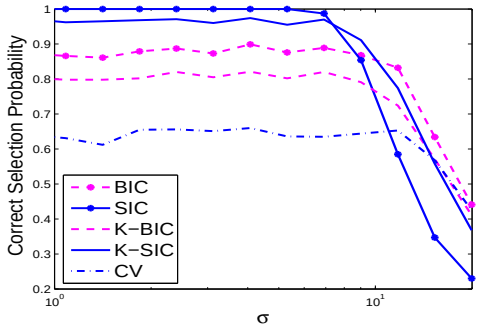
The signal is a polynomial (correct order=3)

$$s(t) = 0.4t + 0.1t^2; t = 0, 1, ..., n-1$$

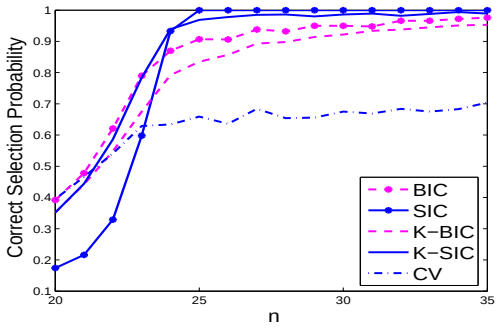
and the observation $y(t)$ is generated by the signal with additive noise, either Gaussian or non-Gaussian. The example has been used in [9] to evaluate the penalty based model selection criteria in terms of the probability of choosing the correct model order. Denote by BIC and SIC the penalty based model selection criteria assuming the noise is Gaussian. Denote by K-BIC and K-SIC the kernel based semiparametric model selection criteria. We also employ the leave-one-out cross validation method to select the correct model, which is denoted by CV. Leave-one-out cross validation runs N separate times which is the same as the data size. The linear regression models are trained on all data except for one point and a prediction is made using that point. The average error is computed and used to evaluate the model selection accuracy. We compare the probability of choosing the correct model order via 1000 Mont Carlo runs.

We consider the following three cases with different noise distributions. In Case 1, we investigate the difference among the performance of the penalty-based methods, the semiparametric methods and cross-validation method for Gaussian noise. In Cases 2 and 3, the performance differences for non-Gaussian noises are examined.

Case 1. Additive Gaussian noise is $N(w;0,\sigma^2)$. (a) σ increases from 0.1 to 20 and sample size is 30. (b) $\sigma = 8$ and sample size increases from 20 to 35.



(a) Probability of choosing correct model order vs. σ



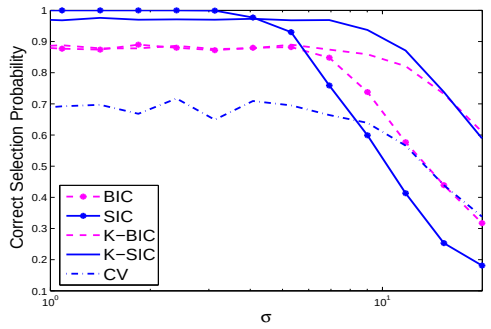
(b) Probability of choosing correct model order vs. sample size n

Fig. 1. Case 1

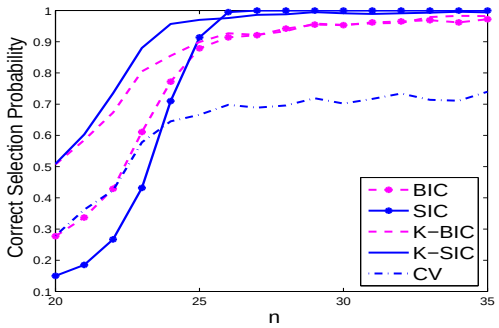
From Fig. 1(a) we can see that the penalty-based criteria perform better in the large SNR regime. This is not a surprise since the penalty terms are obtained assuming additive Gaussian noise. When the SNR decreases, the performance of the two approaches becomes similar. K-SIC performs even slightly better than SIC. SIC is better than BIC when the SNR is large, however, BIC is more robust as the SNR decreases. With 30 observations, the probability of choosing the correct model order by BIC, K-SIC and K-BIC can not further be improved as the SNR increases. As we can see from Fig. 1(b), BIC always performs better than K-BIC. When the sample size is larger than 25, SIC performs better than K-SIC. However, as sample size becomes smaller than 23, SIC has the worst performance. In all cases, CV has the worst performance among these methods.

Case 2. Additive non-Gaussian noise $v = w^2 \cdot \sigma$, $w \sim N(w;0,1)$. (a) σ increases from 0.1 to 20 and sample size is 30. (b) $\sigma = 8$ and sample size increases from 20 to 35.

As we can see from Fig. 2(a), the semiparametric approach has a larger region indicating better performance, and thus is more robust than penalty-based method when the SNR decreases, although the penalty-based parametric approach performs slightly better than semiparametric method

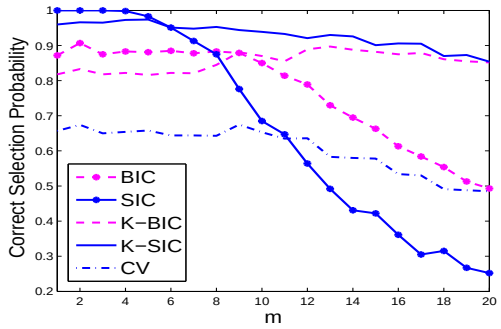


(a) Probability of choosing correct model order vs. σ

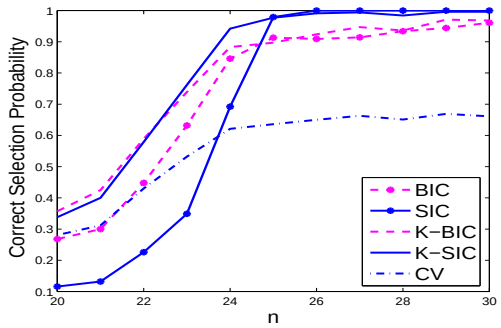


(b) Probability of choosing correct model order vs. sample size n

Fig. 2. Case 2



(a) Probability of choosing correct model order vs. σ



(b) Probability of choosing correct model order vs. sample size n

Fig. 3. Case 3

when the SNR increases. From Fig. 2(b), we can see that K-SIC and K-BIC have much better performance than SIC and BIC when sample size is smaller than 25. As sample size increases (as well as increasing SNR), these methods perform similarly, although SIC performs slightly better than K-SIC. Note that both approaches can choose the correct model order when sample size is large enough since the distance from the misspecified model to the truth has a minimum when the model order is 3. Again, CV is much worse than the others.

Case 3. Additive non-Gaussian noise $v \sim (0.3w_1 + 0.7w_2)$. (a) $w_1 \sim N(w; m, 5^2)$, $w_2 \sim N(w; -m, 5^2)$ and m increases from 1 to 20. (b) $w_1 \sim N(w; 10, 5^2)$, $w_2 \sim N(w; -10, 5^2)$ and sample size increases from 20 to 30.

As is clear from Fig. 3(a), the semiparametric approach has a larger region indicating better performance, and thus is more robust than penalty-based method as m increases. The penalty-based approach performs slightly better than semiparametric method when m is small. Moreover, the performance of the parametric method deteriorate very quickly with the increase of m. On the contrary, the semiparametric method remains good model selection accuracy. From Fig. 3(b), we can see that K-SIC and K-BIC have much better performance than SIC and BIC when the sample size is smaller than 25. When the sample size increases, these methods perform similarly and both approaches can choose the correct model order when the sample size is large enough. Again, CV is much worse than the others in both cases.

From the above simulation results, we can see that the semiparametric method outperforms the penalty-based method in the non-Gaussian noise cases but does not suffer too much even when the noise is Gaussian distributed. Even though the parametric method performs slightly better than the semiparametric method in some non-Gaussian situations, the latter performs more robust against heavy tail and multimodal noise distributions. For small sample size, SIC penalty has a better performance than BIC especially when the SNR is moderately large. The results are in line with the conclusion made in [5]. Cross validation as a powerful data oriented method, unfortunately, has inferior performance to the other methods. This concurs on the claim made by Rivals and Personnaz in [14], which says that the cross-validation has very poor performances for the selection of linear models as compared to the classic statistical tests. We have also examined the bandwidth of (12), and it does not work as well as the bandwidth of (15).

5. CONCLUSION

In the context of model selection, the parametric method needs to make a strong assumption that all models should have known analytical form of their likelihood functions.

The nonparametric method, in another extreme, does not assume any form of the likelihood function and one can only infer the distribution of the data with certain smoothness constraints. For models with a parametric likelihood function but unknown noise distribution, we have developed a semiparametric method for model selection that combines the strength of both the penalty-based method and the nonparametric kernel density estimation. In a model order selection problem for linear regression with an unknown noise distribution, the penalty-based kernel density estimation is quite robust in choosing the correct model order with either Gaussian or non-Gaussian noises. The proposed method performs close to that of the penalty-based method under Gaussian noise cases and much better under non-Gaussian noise cases as the SNR or sample size decreases. Although the semiparametric method does not always outperform the parametric method when the noise is non-Gaussian, the former performs more robust than the latter. This provides another course to take for model selection problem when the sample is of small or moderate size, which is most often preferable to the parametric and nonparametric model selection methods. In particular, the K-SIC criterion yields the best performance in most cases.

6. REFERENCES

[1] Akaike, H., "A New Look at the Statistical Model Identification", *IEEE Trans. on Automatic Control*, 19, pp. 716-723, 1974.

[2] Barron, A., Rissanen, J., and Yu B., "The Minimum Description Length Principle in Coding and Modeling", *IEEE Transactions on Information Theory*, 44(6), pp. 2743 - 2760, Oct. 1998.

[3] Bernardo, J.M., and Smith, A.F.M., *Bayesian Theory*, Wiley, New York, NY, 1994.

[4] Breiman, L., "Bagging Predictors", *Machine Learning*, 24, pp. 123-140, 1996.

[5] Chen, H., and Huang, S., "A Comparative Study on Model Selection and Multiple Model Fusion", *Int. Conf. on Information Fusion*, 2005.

[6] Efron, B., and Tibshirani, R.J., *An Introduction to the Bootstrap*, Chapman and Hall, New York, 1993.

[7] Fan, J. and Yao, Q., *Nonlinear Time Series: Nonparametric and Parametric Methods*, Springer-Verlag, New York, 2003.

[8] Freund, Y., "Boosting a Weaking Learning Algorithm by Majority", *Information and Computation*, 121(2), pp. 256-285, 1996.

[9] Kay, S., "Conditional Model Order Estimation", *IEEE Transactions on Signal Processing*, 49(9), pp. 1910-17, Sept. 2001.

[10] Kearns, M.J., and Vazirani, U.V., *An Introduction to Computational Learning Theory*, MIT Press, Cambridge, MA, 1994.

[11] MacKay, D., "Bayesian Interpolation", *Neural Computation*, 4, pp. 415-447, 1992.

[12] Rissanen, J., "Stochastic Complexity and the MDL Principle", *Econometric Reviews*, vol.6, pp. 85-102, 1987.

[13] Rissanen, J., "Fisher Information and Stochastic Complexity", *IEEE Trans. on Information Theory*, vol.42, pp. 40-47, Jan. 1996.

[14] Rivals, I. and Personnaz, L., "On Cross-Validation for Model Selection", in *Neural Computation*, 11(4), pp. 863-870, 1999.

[15] Schwartz, G., "Estimating the Dimension of a Model", *Annals of Statistics*, vol.6, pp. 461-464, 1978.

[16] Silverman, B.W., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York, 1986.

[17] Stone, M., "Cross-Validatory Choice and Assessment of Statistical Predictions", *Journal of the Royal Statistical Society*, B, 36, pp. 111-147, 1974.

[18] Towers, S., "Kernel Probability Density Estimation Methods", in *Proc. on Advanced Statistical Techniques in Particle Physics*, Durham, England, Mar. 2002.

[19] Vapnik, V.N., *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.