Messenger, R. C., and Mandell, L. M. (1972). A Modal Search Technique for Predictive Nominal Scale Multivariate Analysis. J. Am. Stat. Assoc. 67:768-772.

Morgan, J. N., and Messenger, R. C. (1973). THAID: A Sequential Analysis Program for the Analysis of Nominal Scale Dependent Variables. University of Michigan, Ann Arbor.

Morgan, J. N., and Sonquist, J. A. (1963). Problems in the Analysis of Survey Data and a Proposal. J. Am. Stat. Assoc. 58:415-434.

Perreault, W. D., and Barksdale, H. C., Jr. (1980). A Model-Free Approach for Analysis of Complex Contingency Data in Survey Research. J. Mark. Res. 17:503-515.

Scott, D. W., and Factor, L. E. (1981). Monte Carlo Study of Three Data-Based Nonparametric Probability Density Estimates. J. Am. Stat. Assoc. 76:9-15.

Shepard, R. N., and Carroll, J. D. (1966). Parametric Representation of Nonlinear Data Structures. In Multivariate Analysis, P. R. Krishnaiah (Ed.). Academic Press, New York.

Simonsen, R. H., and Anketell, D. L. (1966). Mechanization of the Curve Fitting Process: DATAN. Commun. ACM 9:299-304.

Sonquist, J. A., Baker, E. L., and Morgan, J. N. (1971). Searching for Structure. University of Michigan, Ann Arbor.

Tucker, L. R. (1958). Determination of Parameters of a Functional Relation by Factor Analysis. Psychometrika 23:19-23.

Turner, M. E., Monroe, R. J., and Henry, L. L., Jr. (1961). Generalized Asymptotic Regression and Non-linear Path Analysis. Biometrics 17: 120-143.

Wahba, G. (1977). A Survey of Some Smoothing Problems and the Method of Generalized Cross Validation for Solving Them. In Applications of Statistics, P. R. Krishnaiah (Ed.). North-Holland, Amsterdam.

4

Predicted Squared Error: A Criterion for Automatic Model Selection

ANDREW R. BARRON* / Adaptronics, Inc., Subsidiary of Flow General, Inc., McLean, Virginia

I. INTRODUCTION

Whereas conventional empirical modeling techniques require an assumed model structure, new procedures have been developed which generate the model structure as well as the model coefficients from a data base. These procedures include the GMDH and PNETTR algorithms for creating polynomial networks. Key to any automatic procedure for generating models is the criterion for ranking different model structures and selecting the best.

The objective of empirical modeling is to identify and train a model that will perform with low error on as yet unseen data. Experience has shown that this objective is met by selecting that model which minimizes an estimate of future performance that we call the predicted squared error (PSE). This criterion is incorporated in the PNETTR 4 algorithm developed by the author at Adaptronics, Inc. (see Chap. 2). This chapter presents a statistical analysis of PSE that explains why it is a good estimate of future performance.

First an intuitive understanding of PSE is helpful. PSE is the sum of two terms: the training squared error and overfit penalty. The training squared error (TSE) is given by the (empirical) average squared error of a model on n training observations. Let k be the number of coefficients in the model that are estimated so as to minimize TSE. The overfit penalty is given by $2\sigma_p^2(k/n)$, where σ_p^2 is a prior estimate of the true error variance that does not depend on the particular model being considered. Thus the predicted squared error is given by

*Present affiliation: Information Systems Laboratory, Stanford University, Stanford, California

$$PSE = TSE + 2\sigma_p^2 \frac{k}{n} \tag{1}$$

The PSE is used at all stages of network construction to rank and select the better model structures. The network that achieves the least PSE is the final product of network synthesis. A minimum will always be attained because TSE decreases with each additional coefficient but always remains nonnegative, whereas the overfit penalty linearly increases in the number of coefficients.

The TSE term favors models that perform well on the training data; however, by itself it can be a poor estimate of future performance. Factors that make TSE underestimate future error (this condition is often called "overfitting" the training data) include overly complicated model structure and many coefficients each adjusted to lower TSE. The overfit penalty term penalizes complex models. It will be shown that this penalty term accounts for the expected squared difference between the estimated model and the true model on future data and accounts for the bias of TSE below the true error variance. The presence of both TSE and penalty terms ensures that PSE favors simple models with low error.

Central to the derivation of PSE and also of independent interest is an understanding of the expected squared error on unseen data (data not yet available to train the model). Section II delves into this topic with some interesting and useful results. Section III discusses PSE as an estimate of this expected squared (future) error. Section IV relates PSE to hypothesis-testing procedures. Section V compares the PSE to other criteria for model selection: including those proposed by Akaike (1970, 1972), Mallows (1973), and Schwarz (1977).

II. EXPECTED PERFORMANCE ON FUTURE DATA

In general, data used to train a model will differ from data the model will encounter in the future. If future data are vastly different, we expect that the model will not perform as well. This is especially true if the model is required to extrapolate far beyond the range of the training data. On the other hand, if the training data are representative of future observations, we expect reasonable performance. This intuitive reasoning is substantiated by the results of this section. A simple expression is derived which provides an exact description of expected squared error when the model is linear in its coefficients and an approximate description when the model is nonlinear in its coefficients.

Consider for now models that are linear in their coefficients. For example, elements (the building blocks of polynomial networks) are linear in the coefficients even if quadratic or cubic terms in the inputs are included. Furthermore, polynomial networks are equivalent to models that are linear in the coefficients if they are composed of elements with nonlinear terms

in original input variables only (i.e., without nonlinear terms in intermediate inputs). Let $\underline{z}$ denote a row vector of transformed input variables that correspond to the terms in the model. For n training observations (input vectors $\underline{x}_i$ and dependent variables y_i, $i = 1, 2, \ldots, n$) consider the k by k symmetric matrix $\underline{R}_T$ composed of normalized (by 1/n) sums of cross-products of the transformed inputs. In matrix notation $\underline{R}_T = \underline{T}'\underline{T}/n$, where $\underline{T} = (\underline{z}_1, \underline{z}_2, \ldots, \underline{z}_n)'$ is the training data matrix consisting of the n transformed vectors. Now suppose that after training, the model will be applied to n_F "future" observations ($\underline{x}_{iF}$ and y_{iF}, $i = 1, 2, \ldots, n_F$; here y_{iF} need not be observable) with data matrix $\underline{F} = (\underline{z}_{1F}, \underline{z}_{2F}, \ldots, \underline{z}_{n_F F})'$ and $\underline{R}_F = \underline{F}'\underline{F}/n_F$. Note that $\underline{R}_T$ and $\underline{R}_F$ can be thought of as the "covariance" structure of the training and future data, respectively ($\underline{R}_T$ and $\underline{R}_F$ would be sample covariance matrices if the input data were regarded as random; however, here the input data are regarded as fixed). Now suppose that for some unknown value of the coefficient vector, the difference (errors) between the model output and the dependent variable are independent random variables with mean zero, and common variance σ^2 (no further assumptions regarding the error distribution are necessary; in particular, the errors need not be Gaussian). The expected squared error on the F data (of the model trained on the T data) is given by

$$\sigma^2 + \sigma^2 \frac{\text{trace}\,(\underline{R}_F \underline{R}_T^{-1})}{n} \tag{2}$$

This result is derived, using standard matrix manipulation, in the Appendix. [See also (Bibby and Toutenburg, 1977); the result above follows from their equations 1.5.5 and 1.5.13.]

The two terms in formula (2) correspond to two factors contributing to error on future data. The first term, σ^2, is the expected squared error of the ideal (but unknown) model. The second term is the expected squared difference (on the F data) between the trained model and the ideal model. This term shows that the expected performance depends on the degree of similarity between training and future data.

Although (2) is derived to motivate the PSE criterion for model selection, the result may be equally important for the areas of experimental design and model adaptation. Experimental design is concerned with locating training input data (in the space of possible input variable values), so that a model can be efficiently trained to perform well in the future. If we have some notion of how future input data will be dispersed, in particular, if we know (or can approximate) $\underline{R}_F$, then the training input data can be chosen*

*Cluster analysis (see Chap. 2) is an invaluable aid in understanding how input data are dispersed and in choosing the training data.

or generated to be representative of future data such that $\underline{R}_T \cong \underline{R}_F$. In this case the expected squared error (2) simplifies to

$$\sigma^2 + \sigma^2 \frac{k}{n} \tag{3}$$

It is tempting to try to reduce the future squared error by designing training data with $\underline{R}_T$ much "larger" than $\underline{R}_F$, so that trace $(\underline{R}_F \underline{R}_T^{-1}) << k$. Then we expect the future squared error to be less than (3). For fixed n, $\underline{R}_T$ larger means that the training data are more spread out. But unless we are confident that the model considered has the right structure, increasing $\underline{R}_T$ can cause a decrease in model accuracy due to interpolating between more distant training observations. Thus if we know that the model structure is correct, we should choose $\underline{R}_T$ such that trace $(\underline{R}_F \underline{R}_T^{-1}) << k$; however, if the structure is one of many possible (as with polynomial network training), it is better to choose $\underline{R}_T \cong \underline{R}_F$ such that trace $(\underline{R}_F \underline{R}_T^{-1}) \cong k$.

The result (2) can be very useful after a model has been trained (on data with "covariance" $\underline{R}_T$) and is being applied to new data for which the true values of the dependent variable y are unknown or unavailable. We wonder how accurate is the model's estimate of y. Result (2) indicates that the "covariance" $\underline{R}_F$ (of the new input data) should be monitored. If $\underline{R}_F$ is such that trace $(\underline{R}_F \underline{R}_T^{-1})$ is less than or comparable to k, the model should be satisfactory. However, if $\underline{R}_F$ is such that trace $(\underline{R}_F \underline{R}_T^{-1}) >> k$, the model is no longer suited for the data. In this case, we are attempting to extrapolate the model to data points consistently outside the range of training data. The model should be adapted or retrained.

Thus both before and after training models the skilled analyst or engineer works to ensure $\underline{R}_T \cong \underline{R}_F$ so that trace $(\underline{R}_F \underline{R}_T^{-1}) \cong k$. Then the expected squared error on future data is given by $\sigma^2 + \sigma^2(k/n)$. Note that models with large k (many estimated coefficients and hence high complexity) are not expected to perform well, unless there are enough training observations that k/n is adequately small.

The results given above have two shortcomings that must be addressed. One deficiency is the assumption that the model considered is of the correct form (i.e., for some unknown coefficient values the errors in the output of this model are independently distributed with zero mean and common variance). During the synthesis of polynomial networks, many different model structures are considered, not all of which can approximate the actual dependencies in the process. For "wrong" structures, there is an additional term in the expected squared (future) error, which is the average squared difference between the unknown correct model and the wrong model (where the coefficients of the wrong model are such that this average squared difference is minimum). When the correct structure is not known in advance, it is difficult to account for this term.

If a model selection procedure is successful in weeding out those model structures that cannot approximate the true relationships in the data, then results (2) and (3) accurately assess the expected squared (future) error for the remaining models. The PSE criterion is designed (in Sec. III) to be effective in both weeding out the clearly incorrect models and in estimating the future error for the remaining models.

The other deficiency is the assumption that the model considered is linear in its coefficients. This is not true for general polynomial networks. However, the results remain valid to the extent that polynomial networks can be approximated by some first-order Taylor expansion in the coefficients. Not all the coefficients of a general network are needed for this expansion. In particular, the coefficients that correspond to linear combinations of past element outputs can be regarded as fixed (since by varying the other coefficients we can obtain any linear combination of these element outputs). Similarly, not all of the coefficients corresponding to constant terms need to be counted or included in the expansion. Let the model function be denoted by $f(\underline{x}, \underline{\beta})$, where $\underline{x}$ is the vector of inputs and $\underline{\beta}$ is the column vector of k free coefficients, Let $\underline{z}_\beta$ be the row vector composed of the partial derivatives of the model $f(\underline{x}, \underline{\beta})$ with respect to the coefficients. A model is nonlinear in its coefficients whenever $\underline{z}_\beta$ depends on $\underline{\beta}$. Let $\hat{\underline{\beta}}$ be the vector of estimated coefficients. The first-order Taylor expansion is given by

$$f(\underline{x}, \underline{\beta}) \cong \underline{z}_{\hat{\beta}}(\underline{\beta} - \hat{\underline{\beta}}) + f(\underline{x}, \hat{\underline{\beta}}) \tag{4}$$

Note that this expansion may be highly nonlinear in the input variables $\underline{x}$ even though it is linear in the coefficients $\underline{\beta}$. The results summarized in formulas (2) and (3) will remain valid if the first-order expansion is an accurate approximation of $f(\underline{x}, \underline{\beta})$ for $\underline{\beta}$ equal to the ideal coefficient values.

The partial derivatives $\underline{z}_\beta$ correspond to $\underline{z}$ the vector of transformed inputs and these coincide when $\underline{z}_\beta$ does not depend on $\underline{\beta}$. For models that are nonlinear in coefficients, approximate $\underline{R}_T$ and $\underline{R}_F$ matrices can be constructed by evaluating (for each observation) these partial derivatives using the estimated coefficient values [e.g., $\underline{R}_T = \underline{T}'\underline{T}/n$, where $T = (\underline{z}_{1\hat{\beta}}, \underline{z}_{2\hat{\beta}}, \ldots, \underline{z}_{n\hat{\beta}})'$]. The partial derivatives can be computed analytically by applying the chain rule of calculus to the layers of polynomial elements (since a network is a composition of element functions). In practice, it is usually best to leave the derivatives in network form and reapply the chain rule whenever a value is desired.

Further research may resolve the issue of expected performance of models which are nonlinear in the coefficients. It may be possible to obtain a more general expression for the expected squared (future) error. No such results are known, but it is conjectured that the dominant terms are the same as in formulas (2) and (3). It is interesting to note that some of the

viewpoints (other than expected squared error) advocated for deriving model selection criteria use no linearity assumptions, yet result in criteria that are similar to PSE. These other viewpoints are mentioned in Sec. V.

III. ESTIMATING EXPECTED PERFORMANCE

This section demonstrates that PSE is a good estimate of the expected squared (future) error. When the training data are representative of future data and when the model considered has a structure that can approximate the "correct" model, then (from Sec. II) the expected squared error on future data is given by Eq. (3)

$$\sigma^2 + \sigma^2 \frac{k}{n}$$

To estimate (3) from the training data, it is natural to consider using an unbiased estimator of σ^2 which is given by

$$\hat{\sigma}_u^2 = \frac{n}{n-k} \text{TSE} = \frac{1}{n-k} \sum_{i=1}^{n} [y_i - f(\underline{x}_i, \hat{\underline{\beta}})]^2 \tag{5}$$

This yields the final prediction error FPE estimate of (3) proposed by Akaike [1970]:

$$\text{FPE} = \frac{n+k}{n-k} \text{TSE} \tag{6}$$

If, indeed, the model considered has the correct structure, then FPE is an unbiased estimate of $\sigma^2 + \sigma^2(k/n)$. Furthermore, if the errors are Gaussian, FPE has minimum variance among unbiased estimates. Akaike proposed FPE as a model selection criterion, and it has proved quite valuable in selecting subsets models from "complete" models in classical linear regression. This good performance is not surprising, because the classical linear regression setup restricts consideration to models that are no more complex than an assumed correct model. However, when training polynomial networks, many wrong (and typically complicated) models are considered that need to be rejected (the bonus, of course, is that we have more chance of finding an accurate and simple nonlinear model). Our experience is that FPE tends to favor some of the complicated overfit models. Why this is so will be evident from some of the analysis of this section.

The PSE estimator of $\sigma^2 + \sigma^2(k/n)$ is given by Eq. (1),

$$\text{PSE} = \text{TSE} + 2\sigma_p^2 \frac{k}{n}$$

where σ_p^2 is a prior estimate of σ^2 that does not depend on the model considered. A simple motivation for the PSE estimate is that the two sources of future error should be identified and estimated separately. The squared error of the ideal model can be estimated by the average squared error on the training data TSE. However, the expected value of TSE is $\sigma^2 - \sigma^2(k/n)$, where the subtracted term is the expected squared difference between the estimated model and ideal model on the training data. This expected squared difference on the training data plus the expected squared difference on future data [total of $2\sigma^2(k/n)$] is estimated using the penalty term [given by $2\sigma_p^2(k/n)$]. The fixed σ_p^2 in the penalty term is used because we do not want PSE to underestimate future squared error when the particular model considered is incorrect (e.g., an overly complex network with low TSE).

PSE is a biased estimator of (3): that is, the expected value of PSE exceeds (3) by the amount

$$\text{bias (PSE)} = \frac{2k}{n} (\sigma_p^2 - \sigma^2) \tag{7}$$

For simple models with few coefficients (small k relative to n), this bias is negligible. Only for overly large k (when typically we want to reject the model) is this bias significant. It is important that our prior σ_p^2 be at least as large as σ^2, so that these models will be rejected [some criteria correspond to using $\sigma_p^2 \cong \frac{1}{2}\sigma^2 \log n$; see Sec. V]. The bias of PSE is not bad; it is an important contribution to the ability of PSE to reject overly complicated (and usually wrong) models. In fact, the bias can help account for the additional term in the expected squared error when the model considered has the wrong structure. High TSE (error in the training set) usually rejects the overly simple models; the $2\sigma_p^2(k/n)$ penalty term (which includes the bias) is needed to reject the overly complex models.

In addition to rejecting wrong models, we want PSE to be an accurate estimate of the expected squared (future) error when the model has correct structure. A natural estimate of the accuracy of PSE is its mean-squared error: that is, the expected value of the squared difference between PSE and $\sigma^2 + \sigma^2(k/n)$. The mean-squared error is the sum of variance and squared bias of PSE.

$$\text{mse (PSE)} = \text{var (PSE)} + \text{bias}^2\text{(PSE)} \tag{8}$$

In order to compute the variance of PSE, we need an additional assumption on the distribution of the errors. If the errors were Gaussian, the sum of squared errors (SSE = nTSE) would be chi-squared on n - k degrees of freedom, which has variance $2(n-k)\sigma^4$. Therefore, as a benchmark for comparison, suppose that

$$\text{var (TSE)} = \frac{\text{var (nTSE)}}{n^2} = \frac{2(n-k)\sigma^4}{n^2} \tag{9}$$

Then the mean-squared error of PSE is given by

$$\text{mse (PSE)} = \frac{2(n-k)\sigma^4}{n^2} + 4\left(\frac{k}{n}\right)^2(\sigma_p^2 - \sigma^2)^2 \tag{10}$$

Note that the variance term is decreasing with increasing k. The decreasing variance and increasing bias implies that the probability that PSE underestimates $\sigma^2 + \sigma^2(k/n)$ decreases with increasing k. Similarly, the probability that PSE is less than $\sigma^2 + \sigma^2(k/n) - \epsilon$ (for any fixed threshold $\epsilon > 0$) decreases. (From the frequentist's point of view, the proportion of times PSE overfits the data decreases as we increase the number of coefficients considered.)

To see how accurate PSE is, we compare the mean-squared error of PSE to the mean-squared error of the unbiased estimator FPE. The mean-squared error of FPE has a variance term only:

$$\text{mse (FPE)} = \text{var}\left(\frac{n+k}{n-k}\,\text{TSE}\right) = \left(\frac{n+k}{n-k}\right)^2 \text{var (TSE)} = \frac{2(n+k)^2}{n^2(n-k)}\,\sigma^4 \tag{11}$$

Note that the variance of FPE is greater than the variance of PSE by a factor of $(n+k)^2/(n-k)^2$. The variance of FPE is increasing in k. Thus it is more likely that FPE is less than $\sigma^2 + \sigma^2(k/n) - \epsilon$ (for any threshold $\epsilon > 0$) as k increases. Loosely speaking, PSE has less probability of selecting an overfit model than does FPE.

From Eqs. (10) and (11), it can be shown that PSE has less mean-squared error than FPE if and only if

$$|\sigma_p^2 - \sigma^2| < \sigma^2\left[\frac{2n}{k(n-k)}\right]^{\frac{1}{2}} \tag{12}$$

Thus from the point of view of mean-squared error, PSE is superior whenever our prior σ_p^2 is reasonably close to the correct σ^2. For example, if $n = 32$ and $k = 8$, then PSE is better for $0.42\sigma^2 < \sigma_p^2 < 1.58\sigma^2$.

The mean-squared errors of PSE and FPE depend on the unknown error variance σ^2. We can compare the estimators further by computing weighted averages of the mean-squared error. Suppose that the weighted average of σ^2 is σ_p^2 and that the weighted average of $(\sigma^2 - \sigma_p^2)^2$ is $\gamma^2\sigma_p^4$. This analysis is equivalent to the Bayesian point of view that the parameter σ^2 has an a priori distribution with mean σ_p^2 and standard deviation $\gamma\sigma_p^2$. The Bayes risk is the average mean-squared error. To compute the risks for PSE and FPE, note that the average of σ^4 is the variance of σ^2 plus the squared mean of σ^2. Thus

$$\text{risk (FPE)} = \text{E(mse FPE)} = \frac{2(n+k)^2}{n^2(n-k)}\,\text{E}(\sigma^4) = \frac{2(n+k)^2}{n^2(n-k)}\,(\gamma^2+1)\sigma_p^4 \tag{13}$$

and similarly,

$$\text{risk (PSE)} = \left[\frac{2(n-k)}{n^2}(\gamma^2+1) + 4\left(\frac{k}{n}\right)^2\gamma^2\right]\sigma_p^4 \tag{14}$$

Under what conditions is the risk of PSE less than the risk of FPE? It is straightforward to show that risk (PSE) < risk (FPE) if and only if

$$\frac{\gamma^2}{\gamma^2+1} < \frac{2n}{k(n-k)} \tag{15}$$

For large n, the inequality can be simplified to risk (PSE) < risk (FPE) provided that the number of coefficients $k \leq 2(1+\gamma^2)/\gamma^2$. The right-hand side of inequality (15) is minimized for $k = n/2$ and then equals $8/n$. Thus the risk of PSE is less for all k when $n < 8(1+\gamma^2)/\gamma^2$. For example, suppose that it is vaguely known that σ^2 is about σ_p^2 with uncertainty (standard deviation) $\pm\frac{1}{2}\sigma_p^2$. Then PSE has less risk for all n if $k \leq 10$ and for all k if $n < 40$.

Clearly, PSE is the better estimator (in terms of mean-squared error) when the number k of estimated coefficients is small (e.g., $k \leq 10$). What can be said about the performance of PSE for larger k? The risk of PSE may be greater than the risk of FPE, but the percent difference remains small for $k^2 < n$:

$$\frac{\text{risk (PSE)} - \text{risk (FPE)}}{\text{risk (FPE)}} = \frac{k^2(n-k)}{(n+k)^2}\,\frac{2\gamma^2}{\gamma^2+1} - \frac{4kn}{(n+k)^2} \tag{16}$$

It should be remembered that these risks have been computed assuming a correctly specified model. Polynomial network training algorithms build the more complex models from the simpler models with smaller k. If incorrect decisions are made on the fewer coefficients, the large k model is incorrectly specified and the risk comparison is not valid. The importance of making correct decisions on small models, even if k will be large, suggests that PSE may be preferred over FPE. The definition of risk as average mean-squared error is misleading. It fails to account for the benefits of positive bias $\sigma_p^2 > \sigma^2$ for discouraging overfit and for helping to account for error due to incorrect model structure. Instead, mean-squared error treats both positive and negative bias as equally bad.

An objection to the PSE estimate is that it requires σ_p^2, a prior upper bound to the error variance σ^2 which might be hard to determine. Fortunately, there is a simple estimate which usually upper-bounds σ^2 and that does not depend on the model considered: specifically, the variation in the dependent variable y, given by

$$\sigma_0^2 = \frac{\Sigma(y_i - \bar{y})^2}{n} \quad \text{where} \quad \bar{y} = \frac{\Sigma y_i}{n} \tag{17}$$

The variation σ_0^2 will be greater than the TSE of every model considered (with the sole exception of the constant model). Thus if σ^2 is greater than σ_0^2, there is little hope of identifying a model. For reasonable data, σ^2 will be less than σ_0^2. If no prior value is provided, PNETTR 4 uses $\sigma_p^2 = \sigma_0^2/2$. [This choice corresponds to asserting that σ^2 is uniformly distributed or completely unknown in $0 < \sigma^2 < \sigma_0^2$ and hence that σ^2 has "mean" $\sigma_0^2/2$ and "standard deviation" $\pm(1/2\sqrt{3})\sigma_0^2$.] Since σ_0^2 depends indirectly on the random errors (of the true model), there is an additional contribution to the variance of PSE. However, this additional variance can be shown to be negligible (Barron, 1981).

IV. HYPOTHESIS TESTING

If the PSE criterion is viewed as a sequential hypothesis-testing procedure, we gain additional understanding of its behavior. Suppose that we have a nested sequence of linear hypotheses. For example, within a particular element of a polynomial network we may have up to eight terms and we wish to sequentially test whether to include individual terms (given that preceding terms have or have not been included). Let PSE(k) be the predicted squared error when a term (corresponding to the kth free coefficient in the entire network) is included and PSE(k - 1) when only preceding terms are included. The term is included if and only if

$$\mathrm{PSE}(k-1) - \mathrm{PSE}(k) > 0 \tag{18}$$

Multiplying by n and then adding $2\sigma_p^2$ to both sides, Eq. (18) can be rewritten as

$$\Delta\mathrm{SSE}(k) > 2\sigma_p^2 \tag{19}$$

where $\Delta\mathrm{SSE}(k) = \mathrm{SSE}(k-1) - \mathrm{SSE}(k)$ is the reduction of the sum of squared errors (SSE = nTSE) if the new term is included. The test (19) is recognized as the sequential chi-squared test assuming Gaussian errors. However, the test is robust in that regardless of the shape of the error distribution, the expected reduction in residual error is σ^2 (under the null hypothesis that the term should be excluded) or σ^2 plus the increase in variation of the model (under the hypothesis that the term should be included).

What would be the corresponding test if the final prediction error FPE were used as the criterion? The FPE criterion would include the term if and only if

$$\frac{n+k-1}{n-k+1}\mathrm{SSE}(k-1) - \frac{n+k}{n-k}\mathrm{SSE}(k) > 0 \tag{20}$$

Assuming that SSE(k) > 0 and multiplying (20) by the positive factor $(n-k)(n-k+1)/[(n+k-1)\,\mathrm{SSE}(k)]$ and then adding $2n/(n+k-1)$ to both sides, inequality (2) reduces to

$$(n-k)\frac{\Delta\mathrm{SSE}(k)}{\mathrm{SSE}(k)} > \frac{2n}{n+k-1} \tag{21}$$

This test is recognized as a sequential F-test assuming Gaussian errors. Note that the threshold on this test is automatically set (and is usually near 2). The F-test is not as robust as the chi-square-type test. The disadvantage of the F-test for the kth term is the sensitivity to incorrect decisions on the other k - 1 terms. If other terms have been included when they should have been omitted, then SSE(k) will be smaller—biasing the F-test high—so that this term has a greater chance of being included. Thus it is possible with the F-test to have a "snowballing" inclusion of terms and hence large overfit models.

This section has shown that in a restricted framework the PSE criterion can be viewed as a robust hypothesis-testing procedure. However, it is important to note that traditional hypothesis-testing procedures are not able to compare and rank models of entirely different structure as is essential in synthesizing polynomial networks. A criterion, such as PSE, that can assess the performance of a model irrespective of the other candidate models is necessary.

V. OTHER CRITERIA

A natural way to estimate future squared error is to withhold a subset of observations from the training data and to evaluate the (empirical) average squared error on this subset. If this evaluation subset is kept independent of the training process, and if the set is representative of the range of potential observations, it then provides a reasonable estimate of the performance of the model. When there are ample data for both training and evaluating, the practice above is strongly recommended. If the error on the evaluation set is comparable to the PSE, it gives the analyst additional confidence in the model selected. If the evaluation set error is much larger than PSE, it suggests that one or both of the subsets has not been designed to be representative of potential data.

Cross-validation, a common criterion for GMDH model selection, involves "withheld" data actively in the synthesis of the model. One subset of data is used to fit the coefficients of each model structure considered and a second subset is used to select the better structures. However, the selection subset does not provide an independent measure of the expected performance. Both the fitting and selection subsets are involved in training the model. If enough different model structures are considered, one can often be found that has low error on the fitting and selection sets, but will

not generalize well to new data. GMDH algorithms such as PNETTR 2 use fitting and selection sets but often need additional checks in network growth to avoid overfit. Additionally, careful attention to the partitioning of the observations into representative data groups (using a cluster algorithm—see Chap. 2) is required. If the number of observations is not large, the accuracy of the trained models is curtailed by the splitting into subsets. These difficulties led the author to develop PNETTR 3 and PNETTR 4 and the PSE criterion for model selection. The PSE criterion does not require data base partitioning. If desired, all the data may be used for training. Furthermore, PSE automatically restricts the network growth.

The PSE criterion resembles and was partially motivated by model selection criteria proposed by Mallows and Akaike. The criterion proposed by Mallows (1973) is to select that model which has minimum C_p, where he defines

$$C_p = \frac{SSE(k)}{\sigma_c^2} + 2k - n \tag{22}$$

If σ_c^2 is a prior estimate or upper bound to σ^2, then (setting $\sigma_p^2 = \sigma_c^2$) it is simple to show that minimizing C_p is equivalent to minimizing PSE. However, Mallows suggests using $\sigma_c^2 = SSE(p)/(n - p)$, where p is the order of a completely specified model and $k < p$. In that way, C_p is not a tool for creating models but rather a tool by which insignificant terms are removed from a known model. Even when a "completely specified" model is known, the C_p statistic may encourage overfit, since $SSE(p)/(n - p)$ is frequently less than σ^2. In the context of polynomial network synthesis, no correct structure is assumed known a priori (even in traditional linear regression, the assumption that the linear model in all inputs is "complete" seems doubtful). Using a fixed σ_p^2, which we believe to exceed σ^2, is preferred.

Akaike has proposed two criteria for model selection. The first (1970) is the final prediction error FPE criterion which has been analyzed in Secs. III and IV. The other is the Akaike information criterion AIC (1972). The AIC is based on a distribution assumed for the vector $\underline{y}$ of dependent variables and on the number of parameters adjusted to maximize the likelihood of $\underline{y}$.

$$AIC = -2 \log \ell(\underline{y}, \hat{\sigma}^2, \hat{\underline{\beta}}_k) + 2k \tag{23}$$

where $\ell(\underline{y}, \sigma^2, \underline{\beta})$ is the likelihood function which is maximized for $\sigma^2 = \hat{\sigma}^2$ and $\underline{\beta} = \hat{\underline{\beta}}_k$ (vector with k estimated coefficients). If the distribution is assumed to be independent Gaussian errors $e_i = y_i - f(\underline{x}_i, \underline{\beta})$, then minimizing AIC is equivalent to minimizing

$$\log (TSE) + \frac{2k}{n} \tag{24}$$

However, minimizing (24) has a serious flaw. The "minimum" is attained by having k sufficiently large that TSE = 0 (e.g., k = n linearly independent terms in a linear model). A more realistic criterion is obtained if σ^2 is assumed known. Then minimizing AIC is equivalent to minimizing

$$TSE + 2\sigma^2 \frac{k}{n} \tag{25}$$

Clearly, this is equivalent to the PSE criterion with σ^2 known ($\sigma_p^2 = \sigma^2$). The analysis of this chapter has shown that assuming σ^2 known is a stronger restriction than necessary. Possibly, a generalized AIC could be derived which incorporates vague knowledge of some parameters.

Akaike did not require linear models in his derivation of the AIC, but derived that asymptotically the log-likelihood is quadratic in the unknown parameters (i.e., it behaves like a Gaussian log-likelihood for a linear model with known error variance.

The AIC is one of several proposed criteria that depend explicitly on the assumed family of distribution. In principle such criteria are applicable to a wide range of problems. However, for a particular problem it is difficult to know what is the "true" family of distributions. PSE is a criterion that does not depend on the particular shape of distributions (e.g., Gaussian). Instead, PSE is derived from a specific "loss" function and is applicable whenever minimizing squared error on independent data is a realistic goal.

Schwarz (1977) proposed that if a parametric family of distribution is assumed, the model should be selected that is a posteriori most probable. He showed that if the log-likelihood function is of a common form (specifically, Koopman-Darmois, which includes the Gaussian), then for almost any prior distribution on the parameters, minimizing

$$-\log \ell(\underline{y}, \hat{\underline{\beta}}_k) + \frac{k}{2} \log n \tag{26}$$

is asymptotically equivalent to maximizing the a posteriori probability of the model. This procedure guarantees consistency (which means that asymptotically, i.e., as $n \to \infty$, the correct model will be selected). For models with independent Gaussian errors of known variance σ^2, minimizing (26) is equivalent to minimizing

$$TSE + \sigma^2 \frac{k}{n} \log_e n \tag{27}$$

This criterion corresponds to the PSE but with $\frac{1}{2}\sigma^2 \log_e n$ in the penalty term in place of σ_p^2. Thus for large n, Schwarz's criterion restricts model dimensionality (e.g., size of a polynomial network) more than does PSE (with σ_p^2 near σ^2). The quantity (27) is biased above the expected squared error on new data by a factor of

$$\frac{k}{n}\sigma^2(\log_e n - 2) \tag{28}$$

which remains negligible provided that the number of estimated coefficients k remains much less than $n/(\log_e n - 2)$.

Recently, a new philosophy for model selection has been proposed, first by Rissanen (1978, 1983) and then independently by this author (Barron, 1982). The goal proposed is to find that model which induces the shortest description for the data available. If a parametric family of distributions is assumed, then for each candidate model there is a description of the data that corresponds to a concatenation of a description of the model (including the estimated parameters) and a Shannon code for the data (given the parameters and input variables). Rissanen and Barron have each shown that minimizing

$$\frac{k}{2}\log n - \log \ell(\underline{y}, \hat{\underline{\beta}}_k) \tag{29}$$

is asymptotically equivalent to finding the shortest description. The first term amounts to using (1/2) log n bits for each of the coefficients and the second term corresponds to the length of the Shannon code. Note that this criterion is equivalent to (26), the criterion proposed by Schwarz. Moreover, the description length criterion (29) does not require the model to be linear in the coefficients. Furthermore, the notion of minimum description length permits improvements in (29) for finite n (Rissanen, 1983; Barron, 1982). If the Gaussian distribution is used to Shannon-code the data [i.e., $\underline{y} = (y_1, y_2, \ldots, y_n)'$ is described by describing the errors $\hat{e}_i = y_i - f(\underline{x}_i, \hat{\underline{\beta}}_k)$, i = 1, 2, ..., n, according to a zero-mean, covariance $\sigma^2 I$ Gaussian distribution], then minimizing (29) is equivalent to minimizing (27) given above.

Does the shortest description of data available now provide a good explanation of statistically similar data in the future? This may be a philosophical question. But the similarity of criteria based on minimum description length [such as (27)] and the predicted squared error (1) seems to be a first step toward a quantitative answer. The goals of good data description and good prediction are not incompatible; however, there are intriguing differences [e.g., σ_p^2 versus (1/2) σ^2 log n in the penalty term]. Whenever the primary objective of empirical modeling is to identify a model that will perform with low error on as yet unseen data, the predicted squared error criterion is strongly recommended.

APPENDIX

This appendix presents a derivation of the expected squared error of a model (trained on one set of data) when the model is applied to a new set of data. The result will be Eq. (2) discussed in the body of this chapter.

Let two sets of observation be denoted by

$$\{(\underline{x}_i, y_i),\ i = 1, 2, \ldots, n\} \quad \text{and} \quad \{(\underline{x}_{iF}, y_{iF}),\ i = 1, 2, \ldots, n_F\}$$

The first set is the training data, and the second can be thought out as future data. Suppose that $y_i = f(\underline{x}_i, \underline{\beta}) + e_i$, where f denotes a candidate model with k unknown coefficients represented by the column vector $\underline{\beta}$. Similarly, $y_{iF} = f(\underline{x}_{iF}, \underline{\beta}) + e_i$. Let $\hat{\underline{\beta}}$ be the coefficients estimated from the training data. We want to compute the expected squared error on the new data when using $\hat{\underline{\beta}}$.

$$E\left[\frac{1}{n_F}\sum_{i=1}^{n_F}[y_{iF} - f(\underline{x}_{iF}, \hat{\underline{\beta}})]^2\right] \tag{A-1}$$

Adding and subtracting the unknown coefficient values, (A-1) becomes

$$E\left[\frac{1}{n_F}\sum_{i=1}^{n_F}[y_{iF} - f(\underline{x}_{iF}, \underline{\beta}) + f(\underline{x}_{iF}, \underline{\beta}) - f(\underline{x}_{iF}, \hat{\underline{\beta}})]^2\right] \tag{A-2}$$

This expression can be expanded into three important terms:

$$E\left[\frac{1}{n_F}\sum_{i=1}^{n_F}[y_{iF} - f(\underline{x}_{iF}, \underline{\beta})]^2\right] + E\left[\frac{1}{n_F}\sum_{i=1}^{n_F}[f(\underline{x}_{iF}, \underline{\beta}) - f(\underline{x}_{iF}, \hat{\underline{\beta}})]^2\right] + 2E\left[\frac{1}{n_F}\sum_{i=1}^{n_F}[y_{iF} - f(\underline{x}_{iF}, \underline{\beta})][f(\underline{x}_{iF}, \underline{\beta}) - f(\underline{x}_{iF}, \hat{\underline{\beta}})]\right] \tag{A-3}$$

Substituting the error $e_i = y_{iF} - f(\underline{x}_{iF}, \underline{\beta})$, expression (A-3) simplifies to

$$E\left[\frac{1}{n_F}\sum_{i=1}^{n_F}e_{iF}^2\right] + E\left[\frac{1}{n_F}\sum_{i=1}^{n_F}[f(\underline{x}_{iF}, \underline{\beta}) - f(\underline{x}_{iF}, \hat{\underline{\beta}})]^2\right] + 2E\left[\frac{1}{n_F}\sum_{i=1}^{n_F}e_{iF}[f(\underline{x}_{iF}, \underline{\beta}) - f(\underline{x}_{iF}, \hat{\underline{\beta}})]\right] \tag{A-4}$$

The first term of (A-4) is the expected squared error of the ideal model on future data. Under the assumption of zero mean and common variance $[E(e_{iF}^2) = \sigma^2$ for each i] this term is just σ^2. Since the input vectors x_{iF} are regarded as fixed, the third term represents interaction between random

errors in training and future data. If we assume independence, this term is zero. (If the model is linear in the coefficients, then uncorrelated errors is sufficient for this term to be zero.) The Eq. (A-4) for expected squared error has now been reduced to

$$\sigma^2 + E\left[\frac{1}{n_F}\sum_{i=1}^{n_F}[f(\underline{x}_{iF},\underline{\beta}) - f(\underline{x}_{iF},\underline{\hat{\beta}})]^2\right] \tag{A-5}$$

Now assume that the model can be approximated as linear in the coefficients $f(\underline{x}_i,\underline{\beta}) = \underline{z}_i\underline{\beta}$, where $\underline{z}_i$ is a (row) vector of k possibly nonlinear transformations of the input variables. Similarly, $f(\underline{x}_{iF},\underline{\beta}) = \underline{z}_{iF}\underline{\beta}$. Define n by k training data matrix $\underline{T} = (\underline{z}_1,\underline{z}_2,\cdots,\underline{z}_n)'$ and n_F by k future data matrix $\underline{F} = (\underline{z}_{1F},\underline{z}_{2F},\cdots,\underline{z}_{n_FF})'$. Similarly, define column vectors for the dependent variables $\underline{y}$ and $\underline{y}_F$ and for the errors $\underline{e}$ and $\underline{e}_F$. Using the notation of matrix algebra, (A-5) becomes

$$\sigma^2 + \frac{1}{n_F}E[\|\underline{F}\underline{\beta} - \underline{F}\underline{\hat{\beta}}\|^2] \tag{A-6}$$

From traditional regression analysis, the coefficients that minimize the (empirical) average squared error on the training set are given by

$$\underline{\hat{\beta}} = (\underline{T}'\underline{T})^{-1}\underline{T}'\underline{y} \tag{A-7}$$

Also, since $\underline{y} = \underline{T}\underline{\beta} + \underline{e}$ we can write the difference $\underline{F}\underline{\beta} - \underline{F}\underline{\hat{\beta}}$ in terms of the data matrices and the error $\underline{e}$.

$$\underline{F}\underline{\beta} - \underline{F}\underline{\hat{\beta}} = \underline{F}\underline{\beta} - \underline{F}(\underline{T}'\underline{T})^{-1}\underline{T}'(\underline{T}\underline{\beta} + \underline{e}) = \underline{F}\underline{\beta} - \underline{F}\underline{\beta} - \underline{F}(\underline{T}'\underline{T})^{-1}\underline{T}'\underline{e} = -\underline{F}(\underline{T}'\underline{T})^{-1}\underline{T}'\underline{e} \tag{A-8}$$

Now substituting (A-8) into the expected squared error (A-6) yields

$$\sigma^2 + \frac{1}{n_F}E[\|\underline{F}(\underline{T}'\underline{T})^{-1}\underline{T}'\underline{e}\|^2] = \sigma^2 + \frac{1}{n_F}E[\underline{e}'\underline{T}(\underline{T}'\underline{T})^{-1}\underline{F}'\underline{F}(\underline{T}'\underline{T})^{-1}\underline{T}'\underline{e}] \tag{A-9}$$

The quantity in the brackets in (A-9) is a scalar. The trace of a scalar leaves the scalar untouched. Furthermore, within a trace operation, matrices commute. Thus (A-9) becomes

$$\sigma^2 + \frac{1}{n_F}E[\text{trace}\ (\underline{e}'\underline{T}(\underline{T}'\underline{T})^{-1}\underline{F}'\underline{F}(\underline{T}'\underline{T})^{-1}\underline{T}'\underline{e}]$$

$$= \sigma^2 + \frac{1}{n_F}E[\text{trace}\ (\underline{T}'\underline{e}\underline{e}'\underline{T}(\underline{T}'\underline{T})^{-1}\underline{F}'\underline{F}(\underline{T}'\underline{T})^{-1})] \tag{A-10}$$

Now if the errors are uncorrelated and have common variance σ^2, then $E(\underline{e}\,\underline{e}') = \sigma^2\underline{I}$. So the formula reduces to

$$\sigma^2 + \frac{1}{n_F}\sigma^2\ \text{trace}\ ((\underline{T}'\underline{T})(\underline{T}'\underline{T})^{-1}(\underline{F}'\underline{F})(\underline{T}'\underline{T})^{-1})$$

$$= \sigma^2 + \frac{1}{n_F}\sigma^2\ \text{trace}\ ((\underline{F}'\underline{F})(\underline{T}'\underline{T})^{-1}) \tag{A-11}$$

Defining $\underline{R}_T = \underline{T}'\underline{T}/n$ and $\underline{R}_F = \underline{F}'\underline{F}/n_F$, the notation is simplified. The result for the expected average squared error of the trained model when applied to new data is

$$\sigma^2 + \sigma^2\frac{\text{trace}\,(\underline{R}_F\underline{R}_T^{-1})}{n} \tag{A-12}$$

REFERENCES

Akaike, H. (1970). Statistical Predictor Identification. Ann. Inst. Stat. Math. 22:203-217.

Akaike, H. (1972). Information Theory and an Extension of the Maximum Likelihood Principle. In Proceedings of the Second International Symposium on Information Theory, B. N. Petrov and F. Csaki (Eds.), Akadémiai Kiadó, Budapest, pp. 267-281.

Barron, A. R. (1981). Properties of the Predicted Squared Error: A Criterion for Selecting Variables, Ranking Models, and Determining Order. Adaptronics, Inc., McLean, Va.

Barron, A. R. (1982). Complexity Approach to Estimating the Order of a Model. Electrical Engineering 378B Final Report, Information Systems Laboratory, Stanford University.

Bibby, J., and Toutenburg, H. (1977). Prediction and Improved Estimation in Linear Models. Wiley, New York.

Mallows, C. L. (1973). Some Comments on C_p. Technometrics 15:661-675.

Rissanen, J. (1978). Modeling by Shortest Data Description. Automatica 14: 465-471.

Rissanen, J. (1983). A Universal Prior for Integers and Estimation by Minimum Description Length. Ann. Stat. 11(2):416-431.

Schwarz, G. (1977). Estimating the Dimension of a Model. Ann. Stat. 6(2): 461-464.