Efficient Testing of Electronic Devices

Hans Engler

Department of Mathematics
Georgetown University
Washington, DC 20057

Michael Souders, Gerard N. Stenbakken

Electricity Division*
National Institute of Standards and Technology
Gaithersburg, MD 20899

Abstract

Testing electronic devices in order to assure the quality of individual products can be time-consuming and expensive. To speed up this process without compromising on reliability, testing strategies have been developed at NIST. They are based on the extraction of low-dimensional linear error models, using a priori knowledge of the structure of the device or data from an exhaustively tested training and validation set. Production testing then is performed only on a small subset of all possible test points which must be chosen optimally. This paper describes statistical and computational aspects of these procedures and reports on the theoretical background, heuristics, algorithms, and practical experiences.

1 The Need for Efficient Testing

Testing is a critical step for assuring the quality of electronic devices. For complicated devices, the cost of testing is quite significant and may exceed 20 % of the sale price. Efficient yet reliable testing strategies that assure the quality of every single device can therefore result in substantial savings.

For many device types, the number of test points is vastly larger than the number of parameters that are expected to determine the device behavior. For example, a 13-bit A/D converter has $2^{13} = 8192$ possible test points, but it can be described with only a few dozen parameters.

Efficient testing strategies try to identify error patterns that occur for a certain device type and build a mathematical model for it. For a given new device, the parameters for these patterns are then determined from measurements at a well-chosen reduced set of test points. The mathematical model is used to compute the device response at all test points. This note discusses an approach that has been developed at NIST (Koffman, Souders, Stenbakken, Lipe, Kinard 1996, Stenbakken & Souders 1991) with which a user can construct a model, choose a reduced test point set, assess its accuracy, and predict the behavior of a device under test. The method is now implemented in MATLAB as the *High dimensional Empirical Linear Prediction Toolbox* (HELP). It has a graphical user interface and allows the user to do device modeling, model diagnosis, and device prediction interactively and without detailed knowledge of the underlying statistical and computational procedures.

2 Linear Models

Consider a device whose behavior can be measured at m different test points. The actual behavior at each of these test points differs from the nominal one by a quantity that is here called the *device response*.

Let us denote the measured device response by the column vector y with m components. For a linear error model, one assumes that

$$y = Ax + \epsilon \tag{1}$$

where A is an $m \times p$ *model matrix*, specific to the device type and incorporating information that depends on the device design, its components, its production process etc. The $p \times 1$ vector x consists of parameters that are specific for an individual device. The column vector ϵ incorporates remainder terms and measurement errors. It is treated as a random vector with zero component means and standard deviations that are identical or similar in magnitude.

At the outset, neither the matrix A nor the number p of its columns are known. This note is mainly concerned with the situation where A has to be estimated from a training set of devices for which complete measurements are available. The result is called an *empirical*

[1] Electronics and Electrical Engineering Laboratory, Technology Administration, U.S. Department of Commerce. Official contribution of the National Institute of Standards and Technology; not subject to copyright in the United States.

model. Note that in the present form, A is not identifiable from observations of y, since it can be replaced with AB^{-1} and x can be replaced with Bx for any nonsingular matrix B. An estimate for A will be denoted by $\hat{A}$. In other approaches that are also implemented in HELP, linear models can be found by linearizing nonlinear circuit models or by including error patterns that are obtained by other engineering considerations, resulting in *physical* or *a priori* models. If several of these approaches are combined, the result is a *mixed* model.

3 Construction of Linear Models

To construct a linear empirical model, one assumes that an $m \times n$ matrix of training data Y is given, that is a matrix with columns y_i, $i = 1, \ldots, n$ that contain complete measurements of n devices in a training set. The goal is to extract a k-dimensional approximation of these responses, i.e. an $m \times k$ matrix $\hat{A}$ such that all columns of Y are nearly in its column space. The *model dimension* k may be smaller or larger than the unknown number p of parameters in equation (1). It must be determined along with $\hat{A}$.

To find a candidate for $\hat{A}$, one computes the *singular value decomposition* $Y = USV^T$. Here U is $m \times n$ with orthonormal columns, V is orthogonal, and $S = diag(s_1, s_2, \ldots, s_n)$ contains the *singular values* s_i that are non-negative and decreasing (Golub & Van Loan 1989). One then chooses the model matrix estimate as $\hat{A} = U_0$, consisting of the first k columns $u_1, u_2, \ldots, u_k$ of the left orthogonal factor U. It is known that U_0 has various optimal approximation properties among all rank-k matrices. In statistical terms, the columns of U are the principal components of the dispersion matrix $n^{-1}YY^T$ of the uncentered data matrix Y, and the s_i^2/n are the corresponding latent roots. For a theoretical analysis, let us make the following *normality assumptions*: The y_i are independent and have multivariate normal distributions $y_i \sim N_m(0, \Sigma)$ with covariance matrix $\Sigma = \tilde{U}\Lambda\tilde{U}^T$, the matrix $\tilde{U} = (\tilde{u}_1, \ldots, \tilde{u}_m)$ is orthogonal, $\Lambda = diag(\lambda_1, \ldots, \lambda_m)$ with $\lambda_1 \geq \lambda_2 \geq \ldots > 0$. Any data that come from a linear model as in (1) can be written in this form, if the measurement errors are independent and $N_1(0, \rho^2)$ - distributed and the device parameters come from a Gaussian distribution. If $rank(A) = p < m$, the λ_i will all be equal to ρ^2 for $i > p$. One expects to detect this if n is sufficiently large and choose k close to p. It is known that $u_i \to \tilde{u}_i$ and $s_i^2/n \to \lambda_i$ in probability, and the s_i^2/n are the maximum likelihood estimators of the λ_i (Muirhead 1982). The quantities with the smaller indices converge first, and the speed of convergence depends also on the separation of the λ_i. A Central Limit Theorem type argument supports these normality assumptions even if the distribution of the parameters is not Gaussian, but p is sufficiently large.

When the training data are fitted to the k-dimensional model U_0, the mean square residual per test point is

$$\tilde{\sigma}_k^2 = \frac{1}{mn} \sum_i \|y_i - U_0 U_0^T y_i\|^2 = \frac{1}{mn} \sum_{i>k} s_i^2 . \quad (2)$$

Of course, the mean square residual per test point for new devices will be larger. Asymptotic arguments and numerical evidence show that under normality assumptions it is approximately

$$\sigma_k^2 = \left(\frac{n}{n-k}\right)^2 \frac{m}{m-k} \tilde{\sigma}_k^2 . \quad (3)$$

If a validation set is available, i.e. complete measurement data from N additional devices $y_{n+1}, \ldots, y_{n+N}$, then an estimate for σ_k^2 can also be computed from their residuals.

The key problem now is the choice of the model dimension k. Formula (3) shows that this decision can be based on the values $s_1, s_2, \ldots$. There are a number of graphical procedures that are commonly used (Jackson 1991) and are implemented in HELP, including plots of the s_k (scree plot) or of their logarithms (LEV plot) and plots of σ_k against k. The dimension k is chosen by looking for an "elbow" in the scree plot or by making σ_k smaller than a predetermined quantity. For a selection of k that can be justified more rigorously, one sets

$$Q_k = \frac{\left(\sum_{i>k} s_i^2\right)^2}{\sum_{i>k} s_i^4} \quad (4)$$

for $k = 1, \ldots, n-1$. It is known that under normality assumptions $\frac{\sqrt{n-k}}{Q_k}$ becomes asymptotically normally distributed as $n \to \infty$ (Sugiura 1972). The number Q_k is never larger than $n-k$, and it is close to this number if the s_i are approximately the same for $i > k$. Thus Q_k can be viewed as an estimate of the dimension of the space that is spanned by the residual vectors $y_i - U_0 U_0^T y_i$: if Q_k is small, the residuals "look" as if they are concentrated on a low-dimensional set. The model dimension should then be increased to capture these few remaining error patterns. The recommendation is to pick the value of k for which Q_k is maximal or is reaching a plateau.

If there is actually a value k beyond which all λ_i are approximately the same, it can be reliably detected by *sphericity tests*. The test employed by HELP uses the test statistic

$$t_k = \frac{(n-k)^2(m-k)}{2} \left(\frac{1}{Q_k} - \frac{1}{n-k}\right) \quad (5)$$

that was proposed in (Sugiura 1972). One can show that if $\lambda_{k+1} = \ldots = \lambda_n$, then $t_k \to \chi^2_f$ in distribution as the number of test points m becomes large, with $f = \frac{(n-k-1)(n-k+2)}{2}$.

In practice, the recommendations from these approaches are often not consistent. Graphical methods tend to lead to unclear recommendations or to model dimensions that are obviously too low. Sphericity tests do not work well if the λ_i do not eventually become constant and then tend to give model dimensions that are too high. Using the quantity Q_k sometimes is a good compromise. The general recommendation is to choose k small (Jackson 1991). If several different values are proposed by HELP, a small choice for k around which several such values are clustered is usually safe, provided the user checks with formula (3) whether the root mean square residuals per test point will be acceptable.

After a model has been constructed, HELP allows the user to perform a number of additional diagnostic tests. These include figures of merit such as the R^2 - value for the training set fitted to its own model or for a validation set and scores from sign tests and rank-sum tests for the residuals for each test point.

4 Reducing the Test Point Set

One now has to choose a *reduced test point set* $J \subset \{1, \ldots, m\}$ such that the device behavior can be predicted reliably at all m test points from measurements at the test points in J. By the construction from the previous section, we can assume that the estimated model matrix $\hat{A} = U_0$ has orthonormal columns. Let U_{0J} denote the matrix that is obtained from U_0 by selecting the rows in J. Then the response of a new device can be predicted from measurements y_J^{new} at the test points in J by

$$\hat{y}^{new} = U_0 \left(U_{0J}^T U_{0J}\right)^{-1} U_{0J}^T y_J^{new} . \qquad (6)$$

The covariance matrix $C_J = U_0 \left(U_{0J}^T U_{0J}\right)^{-1} U_0^T$ describes the effect of random errors in y_J^{new} on the prediction if the errors have constant variances and are uncorrelated. Both assumptions are questionable, since measurement uncertainties are not exactly known. Also, at this stage of the modeling process, the errors contain remainder terms from fitting a model, and these are always correlated. Still, properties of C_J are useful for selecting J.

There are several possible optimality criteria for the choice of the reduced test point set J (Pukelsheim 1993). These include

a) maximizing $\det(U_{0J}^T U_{0J})$, i.e. minimizing the volume of prediction ellipsoids (D-optimality)

b) minimizing the trace of $(U_{0J}^T U_{0J})^{-1}$, i.e. minimizing an expected mean square of residuals (A-optimality)

c) minimizing $\max diag(C_J)$, i.e. minimizing diameters of prediction ellipsoids (G-criterion).

A famous theorem by Kiefer and Wolfowitz states that D-optimality and the G-criterion are equivalent for a continuous version of the design problem (Pukelsheim 1993).

To have an overdetermined problem and be able to perform diagnostics, a subset J with $|J| = t > k$ test points must be found. An exhaustive search for the best subset is obviously not feasible. In fact, the problem of finding a D-optimal selection is known to be NP-hard, and thus branch and bound methods provide no guarantee for a fast solution. Even approximate iterative methods such as DETMAX are still too expensive. In HELP, a "one-pass" method is used that builds up the reduced test point set J in a single sweep, consisting of two phases. In the *first phase*, a minimal subset of k test points is determined by applying QR-factorization with column pivoting to U_0^T (Golub & Van Loan 1989). In this phase, test points are added to J one at a time, corresponding to rows of U_0 that have maximal norms after having been orthogonalized with respect to all previously selected rows. This turns out to be the greedy algorithm for D-optimality. In the *second phase*, additional test points are selected by adding test points where the prediction variances are maximal, one at a time, until the desired size of J has been reached. The prediction variances must be recomputed after each selection. The second phase is the greedy algorithm for the G-criterion.

Numerical evidence shows that the reduced test point sets that are found with this procedure usually are not optimal (as is to be expected), but that they are better than most other choices and that their performance usually cannot be improved much. The reduced test point sets usually are also very good selections for other optimality criteria.

The distribution of the residual vectors $r = \hat{y}^{new} - y^{new}$ can be described in the asymptotic limit of large n if the data satisfy the normality assumptions of the previous section and if $k \geq p$ (Ding & Hwang 1996). A consequence of the theory in that paper is that approximately $r_i \sim N_1(0, \sigma_k^2 w_i)$, where σ_k^2 is as in (3), $w_i = 1 + \tilde{w}_i$ for $i \notin J$, $w_i = 1 - \tilde{w}_i$ for $i \in J$, with $(\tilde{w}_1, \ldots, \tilde{w}_m) = diag(C_J)$. In fact, the factors that determine r can be identified as measurement error e_m and truncation error e_t for a new device, measurement and truncation errors E_m and E_t for the training set, and an estimation error E_e for the model matrix U_0. Then there is a relation of the qualitative form $r \propto e_m + e_t + (E_m + E_t)(1 + E_e)x$, where x denotes the

parameters of the device under test, with a proportionality factor that can be computed from C_J. In particular, the device behavior itself contributes to a portion of the uncertainty: A device whose behavior is very close to nominal can be predicted well with almost any empirical model, regardless of the quality of the training data. Details can be found in (Ding & Hwang 1996).

Since $\sum_{i \in J} \tilde{w}_i = k$ and the $\tilde{w}_j$ tend to be smaller off J than on J, one also obtains that

$$E\left(\|\hat{y}^{new} - y^{new}\|^2\right) \approx \sigma_k^2 \left(m - 2k + \frac{mk}{|J|}\right) \quad (7)$$

This shows that increasing the model dimension k in order to improve the model's accuracy may have the opposite effect of increasing the mean square residuals, unless J is also increased. This happens because additional columns of U_0 eventually only model noise in the training set. Formula (7) can also serve as a guideline for selecting the model dimension when the size of J is fixed in advance due to cost constraints.

5 Detecting Nonmodel Errors

In order to detect whether a device under test is well described by the given linear model, one computes the residual sum of squares at the reduced test point set. Under normality assumptions and if k is not too small, the $\|\hat{y}_J^{new} - y_J^{new}\|^2/\sigma_k^2$ are nearly χ^2_{t-k} - distributed. A flag is raised by HELP if this quantity is larger than a suitable χ^2 quantile, and a *nonmodel factor*

$$\rho_{NM} = \max\{1, \|\hat{y}_J^{new} - y_J^{new}\|^2/((t-k)\sigma_k^2)\} \quad (8)$$

is computed. Of course one expects $\rho_{NM} > 1$ for about one out of two devices, even if they are well described by the model. If ρ_{NM} is consistently larger than 1 for subsequent production devices, there may be new error sources in the devices under test which are not decribed by the original model. This could be due to a change in production methods or in device components, or it could result from performing measurements for the device under test with less accuracy. However, large values of ρ_{NM} are also observed for devices from the same population as the training data merely if their response is large, due to the multiplicative error mechanism that was sketched in section 4. HELP uses ρ_{NM} to adjust the uncertainty bounds for the prediction.

The power of the χ^2 - based procedure for identifying nonmodel errors depends essentially on the number of degrees of freedom, $t - k$. The ratio k/t should also be sufficiently small, since typically $\tilde{w}_i \approx \frac{k}{t}$ or less. In simulations, it is observed that the procedure employed by HELP identifies devices with a nonmodel error component with a probability around 50% or larger, if the root mean square of this component is equal to the standard deviation of the measurement noise and if $t - k \geq 30$. If ρ_{NM} consistently exceeds a value of 2, this suggests that the model might have to be updated.

6 Prediction and Decision

Typically, the tolerance bounds that are permitted for an acceptable device are positive numbers B_i, $i = 1, \ldots, m$. A device is acceptable if its measured behavior at each test point is within the corresponding bounds, that is if

$$|y_i^{new}| \leq B_i\,, \quad i = 1, \ldots m\,. \quad (9)$$

On the basis of the predicted values $\hat{y}_i^{new}$, a device should be accepted if

$$|\hat{y}_i^{new}| \leq B_i - t_{\alpha,|J|}\sqrt{\rho_{NM}\sigma_k^2 w_i}\,, \quad i = 1, \ldots, m \quad (10)$$

where α is a bound for the desired type I error probability and the w_i are as in section 4. Thus the null hypothesis is that the device under test is unacceptable, i.e. "$|y_i^{new}| > B_i$ for some i".

HELP also offers options to compute individual and simultaneous confidence bounds for the measured behavior. These are

$$\hat{y}_i^{new} \pm t_{\alpha/2,|J|}\sqrt{\rho_{NM}\sigma_k^2 w_i} \quad (11)$$

and the Bonferroni versions

$$\hat{y}_i^{new} \pm t_{\alpha/2m,|J|}\sqrt{\rho_{NM}\sigma_k^2 w_i}\,. \quad (12)$$

In simulations, it is observed that the proportion of unacceptable devices that is identified correctly is larger than the advertised fraction $1-\alpha$, as is to be expected for composite null hypotheses. The nonmodel factor ρ_{NM} is observed to be essential for guaranteeing these detection probabilities even in the presence of moderate nonmodel effects. Similarly, coverage probabilities for the individual prediction intervals are observed to be close to $1-\alpha$ in simulations. However, simultaneous prediction intervals may be less reliable in the presence of nonmodel errors, that is, their coverage probability tends to be somewhat less than the advertised value $1 - \alpha$, due to the heavy reliance of the Bonferroni method on well-behaved tails.

Along with a large proportion of bad devices, a certain fraction of acceptable devices will also be rejected, resulting in a reduced yield of good devices. An increased nonmodel factor will be particularly harmful to the yield, so it has to be monitored carefully.

References

Ding, A. A., Hwang, J. T. G. (1996), Statistical intervals for high-dimensional empirical linear prediction, preprint, *Cornell Univ.*.

Golub, G. H., Van Loan, C. V. (1989), *Matrix Computations*, 2nd edition, Johns Hopkins University Press, Baltimore.

Jackson, J.E. (1991), *A User's Guide to Principal Components*, John Wiley & Sons, New York.

Koffman, A.D., Souders, T.M., Stenbakken, G.N., Lipe,T.E., Kinard, J.R. (1996), Empirical linear prediction applied to a NIST calibration service, *Proc. 1996 Natl. Conf. of Standards Laboratories (NCSL) Workshop and Symposium, Aug 25-29, 1996, Monterey, CA*, pp. 207-212.

Muirhead, R. (1982), *Aspects of Multivariate Statistical Theory*, John Wiley & Sons, New York.

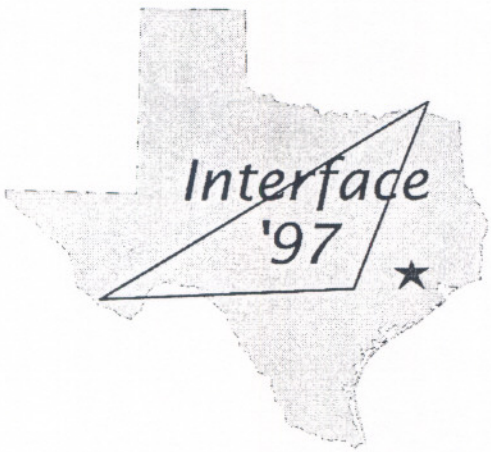
Pukelsheim, F. (1993), *Optimal Design of Experiments*, John Wiley & Sons, New York.

Stenbakken, G.N., Souders, T.M. (1991), Linear error modeling of analog and mixed-signal devices, *1991 Int. Test Conf. Proc., IEEE Computer Soc. Press*, pp. 573-581.

Sugiura, N. (1972), Locally best invariant test for sphericity and the limiting distributions, *Ann. Math. Stat.* vol. 43, pp. 1312-1316.

Computing Science and Statistics

Volume 29
Number 1



Mining and Modeling Massive Data Sets in Science, Engineering, and Business with a Subtheme in Environmental Statistics

Proceedings of the
29th Symposium on the Interface
Houston, TX, May 14-17, 1997

Editor

David W. Scott

INTERFACE
FOUNDATION
OF NORTH AMERICA

The papers and discussions in this Proceedings volume are reproduced exactly as received from the authors. These presentations are presumed to be essentially as given at the 29th Symposium on the Interface. The papers have not been reviewed and no claims are made by the editor or the publisher as to the originality or accuracy of their contents.

An example of the recommended citation of articles from this publication is:

> Friedman, J. H. (1998), "Data mining and statistics: What's the connection?" *Computing Science and Statistics*, 29(1), 3-9.

If more details are required, the editor and publisher may be added:

> Friedman, J. H. (1998), "Data mining and statistics: What's the connection?" *Computing Science and Statistics*, 29(1), (Scott, D.W., ed.), Interface Foundation of North America, Inc.: Fairfax Station, VA 22039-7460, 3-9.

Purchase of Previous Volumes

Volume 21-30 (1989-1997)	Interface Foundation of North America, Inc. c/o Patricia Joyce P. O. Box 7460 Fairfax Station, VA 22039-7460
Volume 22 (1990)	Springer-Verlag New York, Inc. 175 Fifth Avenue New York, NY 10010
Volume 18-21 (1986-1989)	American Statistical Association 1429 Duke Street Alexandria, VA 22314-3402

Interface Foundation of North America, Inc.
P. O. Box 7460
Fairfax Station, VA 22039-7460

ISBN 1-886658-04-8
PRINTED IN THE U.S.A. (1998)