

To appear in  
*Journal of the American Statistical Association*, June, 1999

## Prediction Intervals, Factor Analysis Models and HELP.

A. Adam Ding and J.T. Gene Hwang

### Abstract

We discuss a technique that provides prediction intervals based on a model called an empirical linear model. The technique, called HELP (which stands for *high-dimensional empirical linear prediction*), involves principal component analysis, factor analysis and model selection. In fact, a special case of the empirical model is the factor analysis model. A factor analysis model, however, does not generally aim at prediction. Therefore HELP can be viewed as a technique that provides prediction (and confidence) intervals based on a factor analysis model or a more generalized model possibly with unknown dimension to be estimated.

Although factor analysis models do not typically have justifiable theory due to nonidentifiability, we shall show that our intervals were justifiable asymptotically. An interval for a future response is called a *prediction interval*, whereas, an interval for the mean of the future response is called a *confidence interval*. These intervals were compared to the intervals of Hwang and Liu which were derived using standard asymptotic theory where the relevant covariance matrix has a fixed dimension. In contrast, our intervals are derived asymptotically with the dimension of the covariance matrix approaching infinity, a result much more difficult to obtain. However, the numerical results show that the intervals of this paper are much more satisfactory in many cases, including the motivating application.

The application that motivated us arises from the work of a group of electrical engineers led by Souders and Stenbakken at the National Institute of Standards and Technology (NIST). Their aim is to reduce the number of measurements of a high-dimensional variable of dimension,  $2^{13} = 8192$ , called the future “measurements”, by using the past measurements of similar electric components such as A/D converters. They claim that only 64 out of 8192 measurements need to be measured to predict the rest of unobserved measurements well. In this paper, we construct our intervals using only 64 measurements of the “future observations” and show that the intervals seem narrow enough to justify their claim.

---

A. Adam Ding is Assistant Professor, Department of Mathematics, Northeastern University, Boston, MA 02115 (E-mail: ding@neu.edu); and J.T. Gene Hwang, formerly Jiunn T. Hwang, is Professor, Department of Mathematics, Cornell University, Ithaca, NY 14853 (E-mail: hwang@math.cornell.edu). The authors wish to thank T.M. Souders and G.N. Stenbakken for their generosity in providing the data which are analyzed in Section 4.2. Moreover, Souders has been so kind in helping us to interpret his work (some with his co-authors) and to write Section 4.2.A. The valuable discussions of H.K. Liu are also very much appreciated.

# 1 Introduction.

In this paper, we shall address the general question of how to construct prediction intervals for a high-dimensional variable  $y$  based on observing a part of  $y$ . We also have measured exhaustively other similar variables  $y^i$  (called *training data*) satisfying the following model

$$y^i = \mu + \chi\beta^i + X\gamma^i + \epsilon^i, \quad (1)$$

where  $y^i$  and  $\epsilon^i$  are  $m$ -dimensional random vectors;  $\beta^i$  and  $\gamma^i$  are respectively column vectors with  $k$  and  $l$  components; and  $\chi$  and  $X$  are  $m \times k$  and  $m \times l$  matrices, where  $k$  and  $l$  are smaller than  $m$ . Except for the  $\epsilon^i$ 's, all the vectors or matrices on the right hand side of (1) are fixed and nonrandom. Further, it is assumed that  $y$ , an  $m$ -dimensional vector, satisfies a similar model

$$y = \mu + \chi\beta + X\gamma + \epsilon, \quad (2)$$

The components of  $\epsilon^i$  and  $\epsilon$  are assumed to be independently identically distributed (i.i.d.).

Model (1) looks similar to the usual multivariate model. However, what makes it interesting is that  $\chi$ , and its dimension  $k$  are unknown (whereas  $X$  is known). (Without loss of generality, we shall assume that the columns of  $\chi$  are linearly independent and hence the dimension of  $\chi$  is  $k$ .) Motivated by the problems to be discussed, we assume that we may choose to measure points that are called *test points*, which are a partial  $t$ -dimensional subvector  $y_1$  of  $y$ . In this paper, we shall use  $y_1$  together with the training data to construct statistical intervals for  $Ey$  or for  $y_2$ , the unobservable part of  $y$ . A statistical interval for  $Ey$  is called a *confidence interval* whereas a statistical interval for  $y_2$  is called a *prediction interval*.

In engineer's terminology, we shall call model (1) an *empirical linear model*. The name is chosen because the model has a design matrix  $\chi$  that has to be estimated empirically, i.e., estimated by data. The prediction procedure and the statistical intervals are therefore called the High-dimensional Empirical Linear Prediction (HELP). The technique HELP is an

integration of several different statistical procedures such as the choice of the partial vector of  $y$  to be measured, the estimation of  $\chi$  and  $k$ , and the statistical intervals.

It seems important to point out the relationship between the model considered here and those in the literature. In a special case where  $l = 0$ , and hence the term  $X\gamma^i$  in (1) is missing, the model, called a *purely empirical model*, reduces to the following model

$$y^i = \mu + \chi\beta^i + \epsilon^i, \quad i = 1, \dots, n. \quad (3)$$

which is known as a factor analysis model. Using the terminology of a factor analysis model,  $\chi$  is called the factor loading and  $\beta^i$ 's are called the common factors, both are unknown parameters. To relate our notation to Anderson (1984, Chapter 14),  $y$ ,  $\chi$ ,  $\beta$  and  $\epsilon$  are denoted respectively as  $X$ ,  $\Lambda$ ,  $f$  and  $U$ . In contrast to the purely empirical model (3), we shall call (1) as a *partially empirical model*.

As is well known, a factor analysis model is unidentifiable unless extra identifying assumptions are made. Much of the earlier theoretical work in factor analysis focuses on identifying parameters by proposing various constraints. See, for example, Thurstone (1947, p335), Reiersøl (1950), Koopmans and Reiersøl (1950), Anderson and Rubin (1956). The proposed identifications are typically different ways of simplifying the structure. Based on the various simplifications, results in factor analysis focus on estimation of the factor loadings. Interpretation of these estimators of factor loadings is tricky, since it depends on how to choose the identification condition or equivalently how to do rotation. Consequently, “factor analysis remains very subjective,” a quotation from Johnson and Wichern (1992) in the last paragraph on p442.

The aim of our study is markedly different from that of the factor analysis. Our aim is the prediction of  $Ey$  or  $y_2$  which are identifiable quantities and hence we do not need to identify the model or to make extra identifiability assumptions. Hence the prediction seems to be a “cure” of unidentifiability, as also was demonstrated in the recent example of Hwang and Ding (1997) on neural networks.

Although, as far as we know, the literature of factor analysis does not apply to our problem, the point estimators for  $\beta$  and  $\chi$  we use here depend on singular value decomposition or principal component analysis, techniques which have been applied to factor analysis models as well. See, for example, page 405 of Johnson and Wichern (1992).

Factor analysis models (or more generally empirical linear models) obviously have many applications. See Chapter 9 of Johnson and Wichern (1992). If the objective is to predict a future variable given some of its observable components, then the result of this paper is applicable.

To be concrete, we discuss some of the possible applications below. We begin with a generic example which motivates this work.

**Example 1.** After manufacturing a product, one often needs to measure exhaustively its characteristics in order to check whether it works properly. Several similar products are measured in this way and let  $y^i$  denote the  $i$ th discrepancy vectors, i.e., the difference between the measured characteristics and the targeted characteristics of the  $i$ th product. Hence each  $y^i$  is a  $m$ -dimensional column vector. For a new product, whose characteristics are represented by a column vector  $y$ , is it possible to measure only the  $t$ -dimensional partial subvector  $y_1$  of  $y$  and use it to predict well the other coordinates of  $y$ ? If we may do so with  $t$  much smaller than  $m$ , then the cost of testing is much reduced.  $\square$

A special case of Example 1 is the testing problem of electronic converters studied by a group of Engineers in the National Institute of Standards and Technology (NIST) led by Souders and Stenbakken. In their problem, the number of characteristics  $m$  is 8192 for a 13 bit converter. Measuring only  $t = 64$  characteristics of a converter and using  $n = 88$  exhaustively measured converters, they claim that they may predict the rest of  $(8192 - 64)$  characteristics well. Details will be discussed in Section 4.2.

**Example 2. Calibration of multi-range instruments.** Koffman and Stott at NIST applied HELP to calibrate multi-range instruments that are machines measuring voltages of

signals with wide variation in frequency, voltage, current, etc. To obtain exhaustive accurate NIST measurements of a multi-range instrument costs millions of dollars and may take several months. There are exhaustively measured data available from the past that correspond to other similar instruments. Using the earlier notations,  $y^i$ ,  $1 \leq i \leq n = 100$ , stands for the differences between the measurements of voltages of the  $i$ th multi-range instrument and its accurate NIST measurements of  $m (= 255)$  levels. By these training data, the dimension  $k$  of the subspace is estimated to be 20. With 50 chosen test points of the targeted instrument, they predict the rest 205 points. This saves lots of money and time.  $\square$

We now comment on the construction of statistical intervals. In this paper, we use asymptotic calculations by letting both  $m$  and  $n$  go to infinity. This relates to eigenvectors with large dimension  $m$ , a problem much more difficult than the standard asymptotic approach taken by Hwang and Liu in a Cornell University technical report where  $m$  remains fixed. The resulting intervals of this paper work much more satisfactory when  $m$  is larger than  $n$  as demonstrated numerically in Section 4.1.

The overall layout of the paper is as follows. In Section 2, we derive the confidence and prediction intervals assuming that  $k$  is known. These intervals have asymptotically valid coverage probabilities as  $m$  and  $n \rightarrow \infty$ . Section 3 deals with the case when  $k$  is unknown and is estimated. Numerical studies are reported in Section 4. Our studies confirm that the claim of Souders and Stenbakken that measuring 64 characteristic is sufficient to predict well the rest of the  $(8192 - 64)$  characteristics.

## 2 Empirical Linear Models and Statistical Intervals

In this section, we shall derive the statistical intervals for the empirical linear models (1) and (2). We shall first use (1) to estimate the space of the mean of the  $y$ 's, in particular the column space of  $\chi$ , a  $m \times k$  unknown matrix. Although the rank  $k$  is usually unknown, we shall assume that  $k$  is known in this section. The case of unknown  $k$  will be treated in

Section 3. In both cases we shall assume that  $k$  is fixed while  $m$  or  $n$  or both are becoming large. It is rather easy to deal with the part of the model corresponding to a known design matrix  $X$  called the *physical model*, i.e.,  $\mu + X\gamma^i$ . We project  $y^i$  onto the column space of  $X$  and then use its remaining residuals to estimate the column space of  $\chi$ . Precisely, we write the residual as

$$Y^* = M(y^1 - \bar{y}, \dots, y^n - \bar{y}).$$

where

$$M = I_m - X(X'X)^{-1}X'$$

$\bar{y}$  denotes  $\sum_{i=1}^n y^i/n$ , and  $I_m$  denotes the  $m \times m$  identity matrix.

We shall assume without loss of generality that

$$\chi'X = 0 \quad \text{and} \quad \bar{\beta} = \frac{1}{n} \sum_{i=1}^n \beta^i = 0 \quad (4)$$

Here and later, 0 denotes a zero matrix of appropriate size. To explain why we may assume the first equation of (4), we note that  $\chi\beta^i + X\gamma^i$  is a linear combination of the columns of  $\chi$  and  $X$ . Now let us replace  $\chi$  by a matrix  $\chi_*$  where each column of  $\chi_*$  is the component of the corresponding column of  $\chi$  orthogonal to the column space of  $X$ , e.g.  $\chi_* = M\chi$ , which obviously satisfies (4). We may then rewrite  $\chi\beta^i + X\gamma^i$  as a linear combination of columns of  $\chi_*$  and  $X$  (i.e.,  $\chi_*\beta^i + X\gamma_*^i$  where  $\gamma_*^i = \gamma^i + (X'X)^{-1}X'\chi\beta^i$ ). Consequently model (1) is unchanged after replacing  $\chi$  and  $\gamma^i$  by  $\chi_*$  and  $\gamma_*^i$ . For the second equation of (4), we may replace  $\beta^i$  by  $\beta^i - \bar{\beta}$  and  $\mu$  by  $\mu + \chi\bar{\beta}$  without changing the model (1).

Direct manipulation gives

$$Y^* = \chi B + M\mathcal{E}^* \quad (5)$$

where

$$\mathcal{E}^* = (\epsilon^1 - \bar{\epsilon}, \dots, \epsilon^n - \bar{\epsilon}), \quad \bar{\epsilon} = \frac{1}{n} \sum_{i=1}^n \epsilon^i,$$

and  $B = (\beta^1, \dots, \beta^n)$ .

We shall perform a *singular value decomposition* and write

$$\frac{\chi B}{\sqrt{mn}} = (S, S_0) \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V \\ V_0 \end{pmatrix} \quad (6)$$

Here the right-hand side is the product of three matrices; the first matrix (with size  $m \times m$ ) and the third matrix (with size  $n \times n$ ) are orthogonal matrices and the second matrix (with size  $m \times n$ ) is diagonal with decreasing nonnegative diagonal elements  $d_i$ , called the *singular values*. The number of positive singular values is the rank of  $\chi B$  which is  $k$ . In fact the diagonal matrix  $D$  of size  $k \times k$  has these positive singular values,  $d_1, d_2, \dots, d_k$ , as its diagonal elements. Note also that on the right hand side, we partitioned the first and the third matrices so that  $S$  and  $V$  are of size  $m \times k$  and  $k \times n$  respectively. Consequently, the right hand side equals SDV.

To estimate  $S$  and  $D$ , we apply a similar singular value decomposition to write

$$\frac{Y^*}{\sqrt{mn}} = (\widehat{S}, \widehat{S}_0) \begin{pmatrix} \widehat{D} & 0 \\ 0 & \widehat{D}_0 \end{pmatrix} \begin{pmatrix} \widehat{V} \\ \widehat{V}_0 \end{pmatrix} \quad (7)$$

Here the matrices on the right hand side are partitioned in the same way as that of (6), and hence the sizes of  $\widehat{S}$ ,  $\widehat{S}_0$ ,  $\widehat{D}$ ,  $\widehat{D}_0$ ,  $\widehat{V}$  and  $\widehat{V}_0$  are exactly the same as that of  $S$ ,  $S_0$ ,  $D$ ,  $D_0$ ,  $V$  and  $V_0$ . It seems reasonable to use  $\widehat{D}$  and  $\widehat{S}$  to estimate  $D$  and  $S$ , and  $\widehat{D}_0$  can be attributed to random error since it corresponds to a zero matrix in (6). In particular,  $\widehat{S}$  is a maximum likelihood estimation for  $S$  based on the normal assumption of  $\varepsilon$ . See, for example, Gabriel (1978). Theorem 2.1 below justifies that  $\widehat{D}$  and  $\widehat{S}$  are reasonable estimates of  $D$  and  $S$ , under some assumptions. The rationale of these assumptions is discussed now.

Note that  $\sum_{i=1}^k d_i^2$  equals the sum of all the squared elements of  $\frac{\chi B}{\sqrt{mn}}$ , or  $\frac{1}{mn} \sum_{i,j} a_{ij}^2$  where  $a_{ij}$  are the  $(i, j)$ th element of  $\chi B$ . It seems appropriate (by something similar to the law of large numbers) to expect for our application and many other applications that  $\frac{1}{mn} \sum a_{ij}^2$  (and hence  $d_i$ 's) converges to a finite number as  $m$  and  $n$  approach infinity. We shall therefore assume that

$$\lim d_i = \bar{d}_i < \infty \quad (8)$$

This also explains why we divide  $\chi B$  by  $\sqrt{mn}$  in (6): it is to simplify the expression (8). Note that  $d_i$  and hence  $\bar{d}_i$  is zero if  $i > k$ . We shall also assume that

$$\bar{d}_i\text{'s are distinct and positive if } i \leq k. \quad (9)$$

The proof the Theorem 2.1 here is based on a result by Yin, Bai and Krishnaiah (1988) and some delicate algebraic calculations. See Ding and Hwang (1997) or Ding (1996).

We shall use  $O_p(\cdot)$  and  $o_p(\cdot)$  to denote the probability big ‘‘O’’ and little ‘‘o’’. They may be scalars or matrices. For the case of matrices or vectors, we illustrate their definition using  $O_p(\sigma(\sqrt{\frac{1}{m}} + \sqrt{\frac{1}{n}}))$  in Theorem 2.1 as an example. This represents a matrix  $M$  whose largest singular value is of order  $\sigma(\sqrt{\frac{1}{m}} + \sqrt{\frac{1}{n}})$  in the sense of probability big ‘‘O’’. The matrix definition of probability little ‘‘o’’,  $o_p$ , can be similarly defined in terms of the largest singular value. In evaluating the order of  $M$ , it often is easier to evaluate the order of  $M'M$  or  $MM'$  and take the square root.

**Theorem 2.1** *Assume that the components of  $\epsilon^i$ 's are i.i.d. with zero mean and finite fourth moment. As  $m \rightarrow \infty$ ,  $n \rightarrow \infty$  and  $\frac{m}{n} \rightarrow c \in [0, \infty]$ ,*

$$|\hat{d}_i - d_i| \leq \sigma(\sqrt{\frac{1}{m}} + \sqrt{\frac{1}{n}})(1 + o_p(1)) \quad \text{for all } i, \quad (10)$$

where  $d_i = 0$  for  $i > k$ .

Additionally, if conditions (8) and (9) hold, then

$$\begin{cases} V\hat{V}' &= I_k + O_p(\sigma(\sqrt{\frac{1}{m}} + \sqrt{\frac{1}{n}})), \\ \hat{S}'S &= I_k + O_p(\sigma(\sqrt{\frac{1}{m}} + \sqrt{\frac{1}{n}})), \end{cases} \quad (11)$$

where  $I_k$  denotes a  $k \times k$  identity matrix.

Note that for two orthogonal matrices such as  $\hat{S}$  and  $S$ ,  $\hat{S}'S = I_k$  if and only if  $\hat{S} = S$ . Hence in this sense (11) implies the asymptotic consistency of  $\hat{S}$  as an estimator of  $S$ . We write it this way so that  $\hat{S}'$  and  $S$  multiply to be a finite dimensional matrix (size  $k \times k$ ), and hence it is easier to conceptualize. Similar comment applies to the matrices  $V$  and  $\hat{V}'$ .

Note that  $\chi$  has the same column space as  $S$  and hence the column space of  $\widehat{S}$  can be used to estimate the column space of  $\chi$ .

However, the main goal is to estimate  $y$  and  $Ey$  based on the observable subvector  $y_1$ , called the *test points*, of  $y$  corresponding to a new device. For notational convenience, we shall assume without loss of generality that  $y_1$  consists of the first  $t$  elements of  $y$ . Hence we write  $y' = (y'_1, y'_2)$ . In applications, the subvector  $y_1$  should be chosen according to some scheme, which will be discussed in Section 4.

Similar to  $y$  and  $y_1$ , we write

$$\widehat{\mu} = \begin{pmatrix} \widehat{\mu}_1 \\ \widehat{\mu}_2 \end{pmatrix} \quad \widehat{S} = \begin{pmatrix} \widehat{S}_1 \\ \widehat{S}_2 \end{pmatrix} \quad X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

where  $\widehat{\mu}_1$  consists of the first  $t$  elements of  $\widehat{\mu}$ , and  $\widehat{S}_1$  and  $X_1$  consist of the first  $t$  rows of  $\widehat{S}$  and  $X$ . To estimate  $y$  or  $Ey$ , note again that the column space of  $\chi$  is the same as  $S$ . Hence we may write (2) as

$$y = \mu + \sqrt{m}S\eta + X\gamma + \epsilon \quad (12)$$

Note that in the above expression, we have basically replaced  $\chi$  by  $\sqrt{m}S$ . To explain why we have a factor  $\sqrt{m}$ , we note that if the rows of  $\chi$  are i.i.d. random vectors, a situation that may occur in many applications including electronic converters, then  $\chi = O_p(\sqrt{m})$ . We also pointed out that  $S = O(1)$ . It therefore seems reasonable to multiply  $S$  by  $\sqrt{m}$  so that it is of the same order as  $\chi$ . In doing so, we may reasonably assume that  $\eta$  is a fixed constant just as  $\beta$  is.

For usual univariate regression, consistency results are established by making some assumptions about the asymptotic limit of the design matrix. Here we need similar assumptions. One such assumption was made in (8). The additional assumption of this type made here is that as  $t \rightarrow \infty$  and  $m \rightarrow \infty$ ,

$$\lambda_1/\lambda_k = O(1) \text{ and } \lambda_k^{-1} = O\left(\frac{m}{t}\right), \quad (13)$$

where  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k > 0$  denote the eigenvalues of  $(S_1X_1)'(S_1X_1)$ . Whether (13) is satisfied depends on the selection method of the test points. It is very difficult technically

to prove (13) for general selection methods. However, it is possible if the rows of  $\chi$  are i.i.d. and the test points are chosen independently of  $\chi$ . In such a case, it can be shown that

$$(S_1 X_1)'(S_1 X_1) = \frac{t}{m} I_{k+\ell} + O\left(\frac{\sqrt{t}}{m}\right),$$

which implies (13).

Although it is difficult to verify (13) for more sophisticated selection methods, the statistical intervals derived below using (13) have good coverage probabilities as asserted in the theorems below. Further the first assertion of (13) is really motivated by numerical evaluations relating to Table 3. In Table 3, for the purely empirical model and the method of maximizing  $\hat{\lambda}_k$  in choosing the test points,  $\hat{\lambda}_1 = .0342$  and  $\hat{\lambda}_k = .0032$ . Also,  $m = 8192$  and  $t = 64$ , indicating the plausibility of (13).

Now return to the estimation of  $Ey$  and  $y$ . If  $S$  were known, an intuitive estimator for  $Ey$  and  $y$  is

$$\hat{\mu} + (S, X)[(S_1, X_1)'(S_1, X_1)]^{-1}(S_1, X_1)'(y_1 - \hat{\mu}_1).$$

However, since  $S$  is unknown, we may substitute  $S$  and  $S_1$  by  $\hat{S}$  and  $\hat{S}_1$ , leading to the estimator

$$\hat{y} = \hat{\mu} + (\hat{S}, X)[(\hat{S}_1, X_1)'(\hat{S}_1, X_1)]^{-1}(\hat{S}_1, X_1)'(y_1 - \hat{\mu}_1). \quad (14)$$

**Theorem 2.2** *In addition to the assumptions in Theorem 2.1 and (13), we assume that  $m, n, t \rightarrow \infty$  in such a way that  $m/(m-t) = O(1)$ ,  $n = o(m^2)$ . Then*

$$\hat{y} - \mu^T = T + o_p\left(\sqrt{\frac{m}{n}} + \sqrt{\frac{m}{t}}\right), \quad (15)$$

where

$$T = A\eta + \bar{\epsilon} + (\hat{S}, X)[(\hat{S}_1, X_1)'(\hat{S}_1, X_1)]^{-1}(\hat{S}_1, X_1)'\epsilon_1, \quad (16)$$

where  $A = \frac{1}{\sqrt{n}}\mathcal{E}\hat{V}'[V\hat{V}']^{-1}D^{-1}$  is an  $m \times k$  matrix which is statistically independent of  $\epsilon$  and satisfies

$$A'A = \frac{m}{n}\sigma^2 D^{-2} + o\left(\frac{m}{n}\right).$$

Here  $\mathcal{E} = (\epsilon^1, \dots, \epsilon^n)$ .

Sketch of the proof. We begin with the identity

$$\frac{1}{\sqrt{mn}}Y^*\widehat{V}' = \widehat{S}\widehat{D}$$

which is established by post-multiplying (7) by  $\widehat{V}'$ . The left hand side equals to

$$SDV\widehat{V}' + \frac{1}{\sqrt{mn}}M\mathcal{E}^*\widehat{V}',$$

which leads to

$$S = (\widehat{S}\widehat{D} - \frac{1}{\sqrt{mn}}M\mathcal{E}^*\widehat{V}')(V\widehat{V}')^{-1}D^{-1}. \quad (17)$$

Substituting this  $S$  into  $\widehat{y} - \mu^T$ , we find that the leading term involving  $\widehat{S}\widehat{D}$  is canceled out. This leads to a smaller order term involving the second term on the right hand side of (17). After some careful evaluation of the order of the remaining terms, Theorem 2.2 can be established.

Therefore confidence sets and prediction sets can be constructed based on the distribution of the statistic  $T$ .

**Theorem 2.3** (*Spherical sets*) *Assume that  $t \leq n$ . Under the conditions of Theorem 2.2, and the additional assumption that all the components of  $\epsilon^i$ ,  $1 \leq i \leq n$  and  $\epsilon$  are i.i.d. normally distributed, then*

$$\frac{(\widehat{y} - \mu^T)'(\widehat{y} - \mu^T)}{\sigma^2} \leq c_\alpha^2 \quad (18)$$

*is a confidence set for  $\mu^T$  with an asymptotic coverage probability  $(1 - \alpha)$ . Similarly,*

$$\frac{(\widehat{y}_2 - y_2)'(\widehat{y}_2 - y_2)}{\sigma^2} \leq p_\alpha^2 \quad (19)$$

*is a prediction set for  $y$  with an asymptotic coverage probability  $(1 - \alpha)$ , where the  $c_\alpha^2$  and  $p_\alpha^2$  are respectively  $(1 - \alpha)$  quantiles of*

$$T_1 = \frac{m}{n}(\widehat{\eta}'\widehat{D}^{-2}\widehat{\eta} + 1) + e'[(\widehat{S}_1, X_1)'(\widehat{S}_1, X_1)]^{-1}e, \quad (20)$$

and

$$T_2 = \frac{m-t}{n}(\widehat{\eta}'\widehat{D}^{-2}\widehat{\eta} + 1) + e'([\widehat{S}_1, X_1)'(\widehat{S}_1, X_1)]^{-1} - I_{k+\ell}e + \chi_{m-t}^2; \quad (21)$$

Here  $\chi_{m-t}^2$ ,  $e$  and  $\widehat{S}_1$  are statistically independent, and  $e$  denotes a standard normal random vector of dimension  $k + \ell$ .

Sketch of the proof. From (15), we may show that  $(\widehat{y} - \mu^T)'(\widehat{y} - \mu^T)/\sigma^2$  equals

$$\eta' \left( \frac{A'A}{\sigma^2} \right) \eta + \frac{\bar{\epsilon}' \bar{\epsilon}}{\sigma \sigma} + \frac{\epsilon_1'}{\sigma} (\widehat{S}_1, X_1) [(\widehat{S}_1, X_1)'(\widehat{S}_1, X_1)]^{-2} (\widehat{S}_1, X_1) \frac{\epsilon_1}{\sigma} + o_p\left(\frac{m}{t}\right). \quad (22)$$

The first term is asymptotically equivalent to  $\frac{m}{n} \eta' D^{-2} \eta$  by (16) and can be estimated consistently by  $\frac{m}{n} \widehat{\eta}' \widehat{D}^{-2} \widehat{\eta}$ . The second term is asymptotically equivalent to  $\frac{m}{n}$  by the law of large numbers. And the third term has the same distribution as  $e' [(\widehat{S}_1, X_1)'(\widehat{S}_1, X_1)]^{-1} e$  by the independence of  $\widehat{S}$  and  $\epsilon_1$ .

Therefore, ignoring smaller order terms, (22) is distributed asymptotically as  $T_1$ . Hence (18) has asymptotically correct coverage  $1 - \alpha$ .

Remark 1: The cutoff points  $c_\alpha^2$  and  $p_\alpha^2$  can be obtained by simulations.

Remark 2: Up to now we assume that  $\sigma^2$  is known. In many applications  $\sigma^2$  is unknown. However, as in the example of Section 4.2, one often has repeated measurements of one or more converters and hence one can use these measurements to come up with an independent estimator  $\widehat{\sigma}^2$  of  $\sigma^2$ . In such a case  $\widehat{\sigma}^2$  approaches  $\sigma^2$  since the degrees of freedom increase linearly in  $m$ . Hence the above intervals with  $\sigma^2$  being replaced by  $\widehat{\sigma}^2$  will be asymptotically valid as well.

Remark 3: In practice, we often are interested in confidence and prediction intervals for the coordinates rather than the spherical sets (18) and (19). Theorem 2.3 readily induces the following simultaneous intervals for all the coordinates. For the  $j$ th coordinate of  $Ey$  or  $y$ , we obtain

$$\begin{aligned} \text{Confidence Intervals:} & \quad \widehat{y}_{(j)} \pm c_\alpha \sigma, & j = 1, \dots, m, \\ \text{Prediction Intervals:} & \quad \widehat{y}_{(j)} \pm p_\alpha \sigma, & j = t + 1, \dots, m. \end{aligned} \quad (23)$$

Although (23) is readily applicable to coordinates, they can be improved by a cube-shaped set.

Hence instead of a quadratic norm, we use the maximum norm  $\|\cdot\|_\infty$  where for any vector  $a = (a_1, \dots, a_m)'$ ,

$$\|a\|_\infty = \max |a_i|.$$

Similar to (18) we end up with a cube-shaped confidence set

$$\|\hat{y} - \mu^T\|_\infty \leq C_\alpha \sigma, \quad (24)$$

where  $C_\alpha$  is the  $(1 - \alpha)$  quantile of  $\|T_3\|_\infty$  and

$$T_3 = \frac{1}{\sqrt{n}}|\widehat{D}^{-1}\widehat{\eta}|e_1 + \frac{1}{\sqrt{n}}e_2 + (\widehat{S}, X)[(\widehat{S}_1, X_1)'(\widehat{S}_1, X_1)]^{-1}(\widehat{S}_1, X_1)'e_3, \quad (25)$$

where  $e_1$ ,  $e_2$  and  $e_3$  are statistically independent standard random vectors with dimensions  $m$ ,  $m$  and  $t$  respectively, and are independent of  $\widehat{S}$ . (Note that the term  $A\eta$  in (16) can be written as  $|D^{-1}\eta|e_1/\sqrt{n} + o_p(1/\sqrt{n})$ .)

The confidence set in (24) can also be written as confidence intervals:

$$\hat{y}_{(j)} \pm C_\alpha \sigma, \quad j = 1, \dots, m. \quad (26)$$

Similarly prediction intervals can be derived as

$$\hat{y}_{(j)} \pm P_\alpha \sigma, \quad j = t + 1, \dots, m, \quad (27)$$

with  $P_\alpha$  being the  $(1 - \alpha)$  quantile of  $\|T_4\|_\infty$ , and

$$T_4 = \frac{1}{\sqrt{n}}|\widehat{D}^{-1}\widehat{\eta}|e_1 + \frac{1}{\sqrt{n}}e_2 + (\widehat{S}_2, X_2)[(\widehat{S}_1, X_1)'(\widehat{S}_1, X_1)]^{-1}(\widehat{S}_1, X_1)'e_3 + e_4. \quad (28)$$

Here  $e_1$ ,  $e_2$ ,  $e_3$  and  $e_4$  are independent standard normal random vectors with dimensions  $m - t$ ,  $m - t$ ,  $t$  and  $m - t$  respectively. As before, all the random variables in the last sentence are statistically independent of  $\widehat{S}$ .

## 2.1 Purely empirical models

We make two remarks about the purely empirical linear model (3) (or the factor analysis model), which is a special case of the empirical linear model. All the theorems, in particular

Theorem 2.2 and 2.3, are all valid so long as  $X$  and  $X_1$  are removed in all the equations such as (14), (16), (20), (21), (25) and (28). Hence prediction and confidence intervals are obtained easily.

Secondly, we may easily derive Hwang and Liu's procedure below. They consider the asymptotics for the case  $n \rightarrow \infty$  and  $\frac{m}{n} \rightarrow 0$ . Hence  $\frac{t}{n} \leq \frac{m}{n} \rightarrow 0$ . Thus, the term  $A\eta$  in Theorem 2.2 is of order  $O_p(\sqrt{\frac{m}{n}}) = o(\sqrt{\frac{m}{t}})$ . Consequently, the dominating term is  $(\hat{S}, X)[(\hat{S}_1, X_1)'(\hat{S}_1, X_1)]^{-1}(\hat{S}_1, X_1)'\epsilon_1$ .

Assume that there is an independent estimator  $\hat{\sigma}^2$  for  $\sigma^2$  which has  $\chi_v^2/v$  distribution. Ignoring the smaller order term, we obtain Hwang and Liu's  $(1 - \alpha)$  confidence set

$$\frac{(\hat{y} - \mu^T)'(\hat{S}, X)(\hat{S}_1, X_1)'(\hat{S}_1, X_1)(\hat{S}, X)'(\hat{y} - \mu^T)}{(k + \ell)\hat{\sigma}^2} \leq F_{1-\alpha}, \quad (29)$$

where  $F_{1-\alpha}$  is the  $1 - \alpha$  quantile of an F-distribution with  $k + \ell$  and  $v$  degrees of freedom.

Since both  $\hat{y}$  and  $(\hat{S}, X)[(\hat{S}_1, X_1)'(\hat{S}_1, X_1)]^{-1}(\hat{S}_1, X_1)'\epsilon_1$  are on the  $k + \ell$ -dimensional subspace  $\{(\hat{S}, X)\underline{z} : \underline{z} \in R^{k+\ell}\}$ , (29) is a  $k + \ell$ -dimensional ellipsoid within the above subspace. By Scheffé-type argument, the ellipsoid induces the simultaneous intervals

$$|\hat{y}_{(j)} - \mu_{(j)}^T|^2 \leq B_j^2 \quad (30)$$

where  $\hat{y}_{(j)}$  and  $\mu_{(j)}^T$  are respectively the  $j$ th coordinates of  $\hat{y}$  and  $\mu^T$ , and

$$B_j^2 = k(\hat{S}_{(j)}, X_{(j)})'[(\hat{S}_1, X_1)'(\hat{S}_1, X_1)]^{-1}(\hat{S}_{(j)}, X_{(j)})\hat{\sigma}^2 F_{1-\alpha}, \quad (31)$$

where  $(\hat{S}_{(j)}, X_{(j)})'$  denotes the  $j$ th row of  $(\hat{S}, X)$ .

### 3 Estimation of $k$ .

In the previous sections,  $k$ , the dimension of the column space of  $\chi$ , was assumed to be known. Since  $\chi$  is unknown,  $k$  is often unknown as well. Therefore it is important to discuss how to estimate  $k$ , which is discussed in this section.

To do so, we shall assume that we have a statistically independent estimate of the variance  $\sigma^2$ , so that we can replace  $\sigma$  by  $\hat{\sigma}$  in all the asymptotic results derived below. This information can be obtained by repeatedly measuring one or several devices and is available in Souders and Stenbakken's experiments (see Section 4).

It is easy to derive a consistent method of estimating dimension  $k$  from Theorems 2.1:

$$\hat{k} = \text{the largest } i \text{ such that } \hat{d}_i > (1 + \delta)\sigma\left(\sqrt{\frac{1}{m}} + \sqrt{\frac{1}{n}}\right), \quad (32)$$

where  $\delta$  is an arbitrary positive number.

**Corollary 3.1** *Under conditions for Theorem 2.1,  $P(\hat{k} = k) \rightarrow 1$*

Due to this Corollary, we may replace  $k$  by  $\hat{k}$  in the previous statistical intervals. Asymptotic property remains valid with this replacement. As stated now the proof does not apply to the case  $\delta = 0$ , which need more technically involved upper bounds on the eigenvalues. However, the choice of  $\delta = 0$  seems to work well in simulation. This is what is used in Table 2 of Section 4.

## 4 Numerical Studies

To study the coverage probability of the proposed intervals, we did two numerical studies. The first was a Monte Carlo simulation; the second used the converter example of Souders and Stenbakken.

### 4.1 Simulation

For the first study, samples were generated from the purely empirical model (3), where the standard deviation  $\sigma$  of elements of  $\varepsilon^i$  is 0.1 and  $k = 3$ . Both  $\sigma$  and  $k$  are assumed to be known. Different procedures are applied respectively on the sample and their coverage and lengths are recorded. Then the process was repeated by generating 1000 such samples. Table 1 reports the coverage probabilities and average lengths for three procedures: the

intervals (30) proposed by Hwang and Liu, our proposed spherical confidence sets (23) and the cube-shaped sets (26).

In Table 1, we use a fixed  $\eta = (3.833, -2.452, 0.3385)$ . The coordinates of  $\eta$  are generated from three independent normal random variables with variance  $3^2$ ,  $2^2$  and  $1^2$  respectively. We also studied two other  $\eta$  generated similarly,  $(6.155, 2.167, 0.0826)$  and  $(-3.960, -3.049, -0.45924)$ . The coverage probabilities and the average half-lengths are similar and are not reported here.

Table 1: Coverage probabilities and average half lengths (given in parentheses) for Hwang & Liu's confidence intervals, the proposed spherical and cube-shaped confidence region at nominal level 95%.

k=3; $\sigma=0.1$ ; nominal 95%; $\eta = (3.833, -2.452, 0.3385)$					
			Hwang & Liu	Spherical	Cube-shaped
n=10	t=10	m=30	6.1%(.36)	88.7%(.72)	95.3%(.38)
		m=100	0.4%(.43)	87.0%(1.37)	95.1%(.46)
		m=500	0.0%(.51)	85.7%(3.12)	95.4%(.54)
	t=40	m=100	0.0%(.16)	85.6%(.86)	94.1%(.28)
		m=500	0.0%(.19)	83.9%(1.92)	95.4%(.32)
		t=100	m=500	0.0%(.11)	79.0%(1.72)
n=40	t=10	m=30	50.0%(.37)	95.8%(.63)	94.9%(.29)
		m=100	15.5%(.44)	92.9%(1.21)	94.6%(.36)
		m=500	0.6%(.51)	92.6%(2.72)	95.0%(.42)
	t=40	m=100	0.0%(.16)	94.4%(.55)	95.1%(.16)
		m=500	0.0%(.19)	93.4%(1.25)	95.4%(.19)
		t=100	m=500	0.0%(.11)	95.0%(.93)
n=100	t=10	m=30	78.3%(.36)	94.7%(.61)	95.0%(.28)
		m=100	50.8%(.43)	94.9%(1.17)	94.9%(.34)
		m=500	11.8%(.51)	93.6%(2.67)	96.0%(.40)
	t=40	m=100	3.3%(.16)	94.6%(.50)	95.1%(.14)
		m=500	0.0%(.19)	93.2%(1.13)	95.9%(.17)
		t=100	m=500	0.0%(.12)	94.0%(.76)
n=500	t=10	m=30	92.1%(.35)	93.8%(.60)	94.9%(.27)

To address more values of  $\eta$ , we let  $\eta$  be 1000 realizations of a 3-dimensional independent normal random vector with the diagonal covariance matrix  $diag(3^2, 2^2, 1^2)$ . Based on the 1000 runs, the resultant coverage probabilities and average half lengths are reported in Table 2.

From the discussion in Section 2.1, we know that Hwang & Liu's procedures are asymp-

Table 2: Coverage probabilities and average half lengths (given in parentheses) for Hwang & Liu's confidence intervals, the proposed spherical and cube-shaped confidence region at nominal level 95%.

k=3; $\sigma=0.1$ ; nominal 95%; $\eta$ randomly generated.					
			Hwang & Liu	Spherical	Cube-shaped
n=10	t=10	m=30	18.6%(.37)	83.6%(.71)	89.3%(.36)
		m=100	8.3%(.44)	81.9%(1.35)	89.6%(.43)
		m=500	1.7%(.52)	81.7%(3.05)	88.5%(.51)
	t=40	m=100	1.2%(.16)	82.1%(.94)	84.6%(.27)
		m=500	0.1%(.19)	81.8%(2.10)	82.7%(.32)
		m=500	0.0%(.11)	82.8%(2.02)	83.1%(.30)
n=40	t=10	m=30	58.2%(.36)	94.1%(.63)	94.1%(.29)
		m=100	31.9%(.44)	93.2%(1.22)	94.4%(.36)
		m=500	11.7%(.50)	94.7%(2.71)	94.4%(.41)
	t=40	m=100	6.6%(.16)	91.8%(.58)	94.3%(.17)
		m=500	1.6%(.19)	92.6%(1.30)	92.9%(.20)
		m=500	.4%(.11)	92.2%(1.08)	92.8%(.16)
n=100	t=10	m=30	78.3%(.36)	95.2%(.61)	94.7%(.28)
		m=100	55.6%(.44)	94.6%(1.18)	96.6%(.34)
		m=500	27.8%(.51)	94.0%(2.67)	94.5%(.39)
	t=40	m=100	18.7%(.16)	93.6%(.50)	94.8%(.14)
		m=500	5.5%(.19)	94.8%(1.14)	95.1%(.17)
		m=500	1.2%(.12)	93.8%(.80)	95.0%(.12)
n=500	t=10	m=30	93.2%(.36)	95.8%(.60)	94.7%(.27)

totically equivalent to the proposed procedures when  $n > m$ .

Hence the procedure of Hwang & Liu should have reasonably good coverage probability. The results for the case  $m = 30, n = 500$  and  $t = 10$  in Tables 1 and 2 confirm the theory.

But for the case where  $m \gg n$ , Hwang & Liu's confidence intervals are no longer valid. This is reflected by the extremely low coverage (near zero) in the cases where  $m = 500$ . However, Hwang and Liu's prediction intervals probably have good coverage probabilities, which are not studied here. Also, their procedure has the advantage of being simpler. In Tables 1 and 2, we consider only the confidence region.

Tables 1 and 2 also report coverage probabilities for the spherical confidence region (18) of Theorem 2.3. The coverage probabilities are much closer to the nominal level .95, especially for  $n \geq 40$ . The average half lengths (in the parentheses) reported here average over the radii  $c_\alpha \sigma$  of the balls, each radius being the half length of the one-dimensional interval (23). The determination of each  $c_\alpha$  is based on a simulation of 1000 repetitions.

Tables 1 and 2 show that the cube-shaped confidence sets perform amazingly well in both coverage probabilities and half lengths, having coverage probabilities close to the nominal level 95% and half lengths much smaller than the spherical sets. Further, even when the coverage probability of Hwang and Liu's intervals are reasonable, the cube-shaped intervals have smaller average length. See the last row of Tables 1 and 2.

## 4.2 Application to Electronic Converters.

### A. Scientific Background

Electronic A/D or D/A converters are very important electronic devices for modern life. The A/D converter is an electronic device that converts the analog signals (usually voltages) into digital signals (binary outputs). Conversely D/A converters convert digital signals into analog signals. These converters serve as important bridges between the physical world, in which measurements are typically continuous and hence in analog signals, and the computer world, in which typically digital signals are processed. For example, a CD player

uses a converter. For more details about how electronic converters work, see Boylestad and Nashelsky (1996, p748-751). Part of the description is reported in Filliben and Li (1997, page 286) also.

To insure that a converter, say an A/D converter, works properly, however, one needs to make sure that an input analog signal within a certain voltage will be converted into the intended digital output. These voltage ranges are called *transition levels*. In other words, one needs to make sure that the transition levels are roughly correct.

The Engineers at NIST studied the statistical prediction problem described in Example 1 of Section 1 by assuming the partially empirical linear models (1) and (2) or the purely empirical linear model (3). Here  $y^i$  denotes the column vector of error transition levels, i.e., the differences between the measured transition levels and the targeted transition levels. Furthermore,  $\chi$  and  $X$  can be interpreted as the first order partial derivatives of the response variable with respect to the physical characteristics (denoted as  $\beta$  and  $\gamma$ ) of the dominating resistors, capacitors and inductors of the converter. Note that most of these partial derivatives are unknown which justify the assumptions that  $\chi$  is unknown.

The NIST engineers mainly focused on developing a point estimator for  $y$  or  $Ey$ . This involves estimating  $\chi$  in (1) or  $S$  in (12). In doing so, Stenbakken and Souders (1987) first proposed QR decomposition and then later recommended the more efficient singular-value decomposition (which results in a maximum likelihood estimator) as being used in this paper. Their other major effort is to come up with an algorithm to choose test points. See the next section B as well as Stenbakken and Souders (1987).

They apply their point estimator to a data set to be analyzed below and concluded that they only need to measure 64 (out of 8192) test points in order to predict well. They conclude this by setting aside some measurements and use them to study the prediction error. See Souders and Stenbakken (1991). This conclusion is justified by the confidence intervals in this paper.

While the point estimation works well, there is a desperate need for studying the uncer-

tainty. This is why the work of Hwang and Liu, which constructs confidence and prediction intervals, was welcomed. Their procedure was described in (30) and was studied in Section 4.1. In a slightly different problem, Stenbakken (1996) modified the intervals of Hwang and Liu and constructed intervals for the new converter when the model of the new converter may have been changed perhaps due to changes in the manufacturing process.

Another somewhat different approach was taken by Filliben and Li (1997), which applies the *Principal Hessian Direction* technique to analyze the structure of a converter. Although their approach does not apply to a problem involving more than one converter, their technique can serve as a useful tool to search for a plausible physical model (corresponding to  $X$  in (1)).

## B. Application of our intervals

The data that Souders and Stenbakken generously provided us are mainly the error transition levels of 88 13-bit converters. To study how well our intervals work, we use the common delete-one cross-validation technique to estimate the coverage probability and the expected half lengths of the prediction intervals. More specifically, we select one converter, consider it as the “new” one and use the data of the remaining 87 converters together with  $t = 64$  observations of the “new” converter to construct a prediction intervals for the coordinates of the “new” converter. Doing this for each of the converters, we have a total of 88 experiments; the coverage probability can then be estimated by counting the proportion of times that the intervals contain the values corresponding to the “new” converters. One could also calculate the average length of the 88 intervals and use it to estimate the expected length. These results are reported in Table 3.

Both the partially empirical models (1) and (2) and the purely empirical model (3) are studied. In the partially empirical model, the known design matrix  $X$  is taken to be the design matrix corresponding to the  $2^{13}$  factorial design (see p377 of Box, Hunter and Hunter (1978)). See also Filliben and Li (1997) for a careful discussion about the relationship between the factorial design and converter data. Hence the size of  $X$  is  $8192 \times 14$ . We use

Table 3: Estimated coverage probabilities and average half lengths (in parentheses) for simultaneous prediction intervals in the converter data.

Test points selection method	model	
	purely empirical	partially empirical
first 64	93.2%(1.75)	
1, 1+96, 1+96+97, ...	94.3%(.324)	97.8%(.412)
maximize $\hat{\lambda}_k$	94.3%(.241)	96.6%(.130)
minimize length of (30)	97.7%(.129)	96.6%(.128)
Souders & Stenbakken	97.7%(.128)	97.7%(.129)

(32) with  $\delta = 0$  to estimate  $k$ . For the purely empirical model,  $\hat{k} = 19$  and for the empirical model  $\hat{k} = 9$ . In addition to the data described above, Souders and Stenbakken also provided us with four repeated measurements of an extra converter. For each level, we may calculate the sample variance. The average of all these 8192 sample variances, denoted as  $\hat{\sigma}^2$ , is an unbiased estimate for  $\sigma^2$ . This calculation gives  $\hat{\sigma} = .0219$ .

Several methods for selecting the 64 test points are used. Our intervals appear to have correct coverage regardless of which method is used to select the test points. But there are clear differences in the lengths of the resultant intervals. The first row provides cross-validation estimated coverage probabilities and the average lengths of prediction intervals based on the first 64 test points. (There is no entry in the partially empirical column for the first 64 test points since the matrix  $[(\hat{S}_1, X_1)'(\hat{S}_1, X_1)]$  is singular in this case.) The intervals are quite long so that they may be useless in practice. Note that the transition error is scaled in such a way so that one unit equals the voltage difference between two neighboring targeted transition levels. Consequently, in many situations, it is required that the error transition level should be less than 0.5. Prediction intervals with average half length 1.75 seem too wide.

The second row contains the results using test points that are chosen in a deterministic way. The test points are more evenly spread out for the entire range and the gap between

indices is increased by one each time to avoid becoming periodic.

As we can see in (28), roughly speaking, the length of the intervals increases as  $[(\widehat{S}_1, X_1)'(\widehat{S}_1, X_1)]^{-1}$  increases. Therefore, it is reasonable to choose the  $t$  test points that maximize the matrix  $[(\widehat{S}_1, X_1)'(\widehat{S}_1, X_1)]$  in some sense. One method is to use a *random maximization algorithm* to maximize  $\widehat{\lambda}_k$ , the smallest singular value of  $[(\widehat{S}_1, X_1)'(\widehat{S}_1, X_1)]$ . That is, randomly choose  $B$  different sets, each containing 64 test points. Compute the  $\widehat{\lambda}_k$  for each set of test points, and choose the set with largest  $\widehat{\lambda}_k$ . We chose test points based on this algorithm with  $B = 10,000$  runs. The results of using such chosen test points are reported in the third row.

Similarly we can also use a *random minimization algorithm* to minimize

$$\max_j (\widehat{S}_{(j)}, X_{(j)})' [(\widehat{S}_1, X_1)'(\widehat{S}_1, X_1)]^{-1} (\widehat{S}_{(j)}, X_{(j)}), \quad (33)$$

where  $(\widehat{S}_{(j)}, X_{(j)})$  is the  $j$ th row of  $(\widehat{S}, X)$ . (This incidentally also minimizes the maximum length of (30)). The results of using such chosen test points are reported in the fourth row. It seems to work as well as Stenbakken and Souders' method to be discussed below.

Stenbakken and Souders (1987) also proposed a method to choose test points. Their method uses a QR decomposition and chooses the initial  $k$  points by a greedy algorithm that maximizes the determinant of  $[(\widehat{S}_1, X_1)'(\widehat{S}_1, X_1)]$ . Then, the remaining  $t - k$  points are chosen by a greedy algorithm that minimizes the quantity (33). In the last row, we report the results using their test points.

The half lengths of the intervals in the last two rows are very similar. This seems to indicate that the random maximization (minimization) algorithm works as well as Stenbakken and Souders' approach. Note that it is much simpler to program the random maximization algorithm.

Also, for all the selection methods, our prediction intervals have good coverage probabilities. The selection method only affects the length of the intervals. It agrees with the intuition that a better selection method would end up with more informative test points, hence producing shorter intervals. Judging from these intervals and using the selection meth-

ods corresponding to the last two rows, the lengths (about .13) seem small enough and hence 64 test points predict sufficiently well the 8192 test points. The saving is amazing: HELP reduces the measurements of a new converter to less than one percent. Our results confirm the claim of Souders and Stenbakken (1991) that 64 test points are sufficient.

One surprising discovery demonstrated in Table 3, however, is that the purely empirical model works as well as or sometimes even better than the partially empirical models in most cases. (If a purely empirical model is used, the test points selection techniques are the same as what was described above for the partially empirical model, except  $X_1$  should be removed.) This may be due to the fact that HELP is working: it automatically picks up good or even better column space to substitute for  $X$  when  $X$  is not available.

Finally the Figures 1-3 report the result of applying HELP to the prediction of the first converter using the test points selected by Souders and Stenbakken's algorithm and using the purely empirical model. Figure 1 gives the raw data  $y^1$ , the estimated value of  $y^1$  and the residual which is substantially reduced in magnitude. Figure 2 plots various intervals and Figure 3 is its enlargement. We also plotted three corresponding pictures for the partially empirical model which turn out to be similar.

It is interesting to observe that in Figure 2 there is a big difference in length between Hwang and Liu's prediction interval and their confidence interval: the ratio of average lengths is 22.2. However, the cube-shaped prediction interval is similar in length to its confidence interval with a ratio of 1.6. As a result, the cube-shaped prediction interval is still quite informative. Both spherical confidence intervals and prediction intervals are too wide.

## 5 Conclusion

In this paper, we propose a new way of studying empirical linear models, namely by using asymptotic calculations involving large dimensional matrices. Statistical intervals constructed using this approach are demonstrated, by using simulation and by applying to real

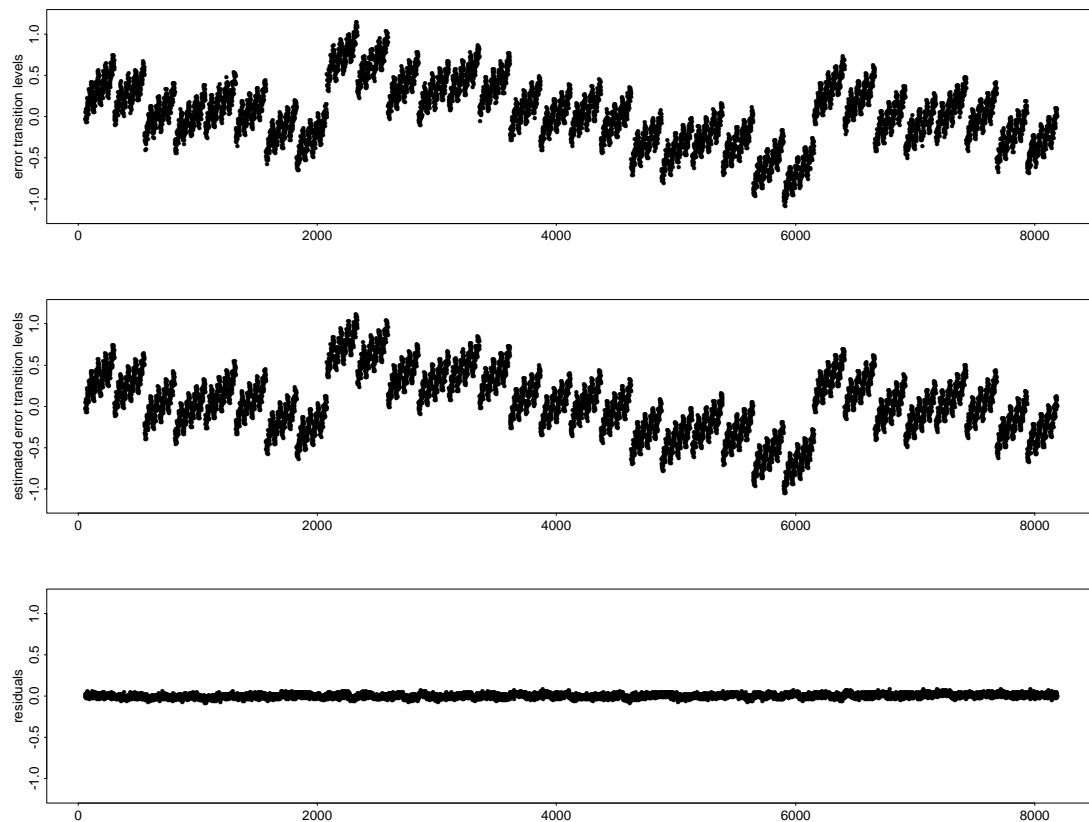


Figure 1: Results on the first converter using purely empirical model: the top, middle and bottom pictures plot respectively, the raw data, the predicted value and the residuals.

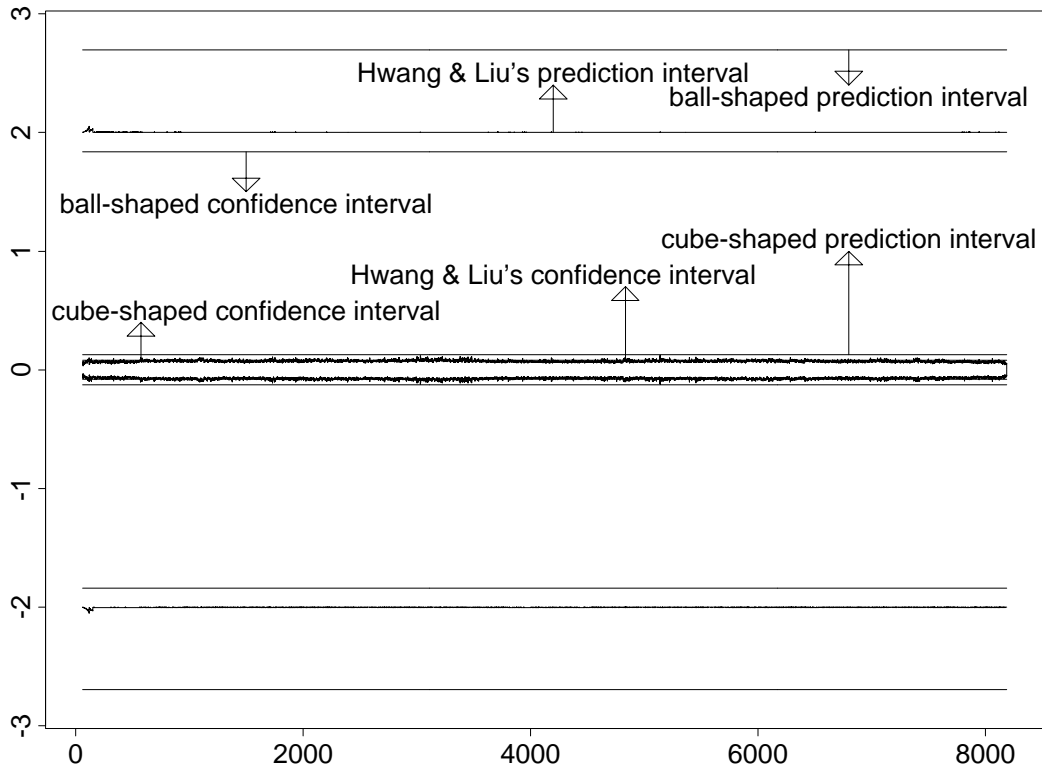


Figure 2: 95% Confidence intervals and prediction intervals (centered at the origin) on the first converter using purely empirical model. The newly proposed intervals (cube-shaped and spherical) are straight lines, while the Hwang & Liu's intervals appear as ragged curves. See Figure 3 for an enlarged picture.

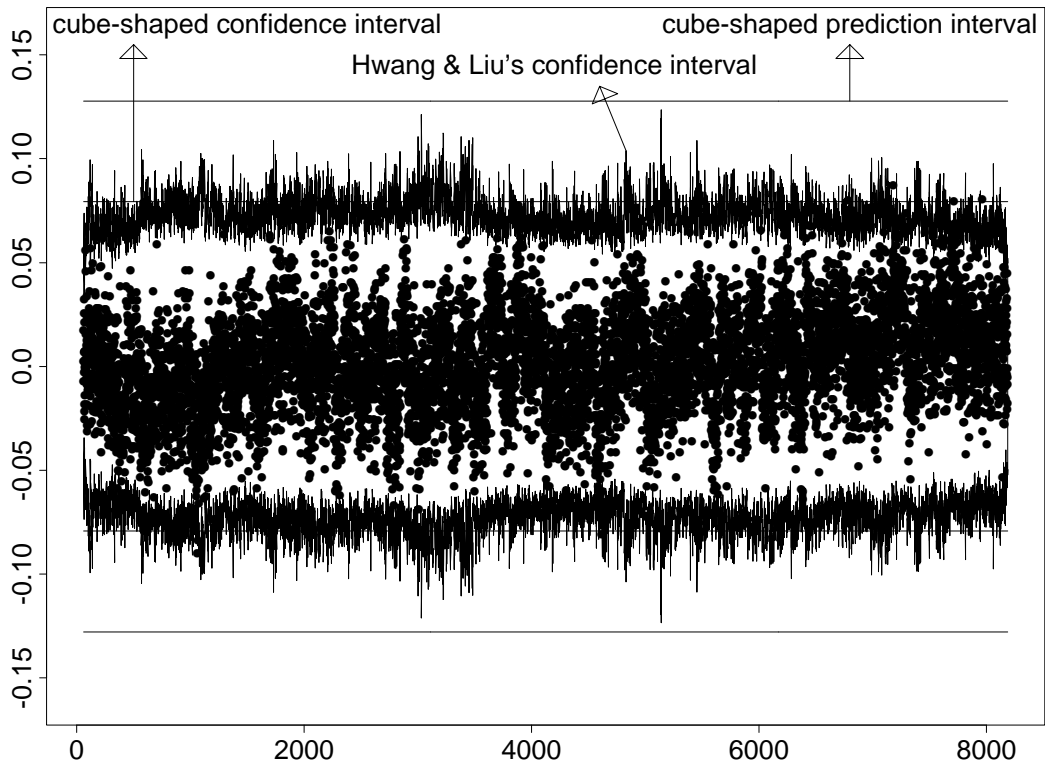


Figure 3: Residuals and 95% intervals (centered at the origin) on the first converter using the purely empirical model. Hwang & Liu's confidence intervals are represented by ragged curves and the cube-shaped intervals are represented by straight lines. Residuals appear as the dots in between these intervals.

data. Simulation studies indicate that this method substantially improves upon the existing intervals of Hwang and Liu.

To predict a high-dimensional variable, HELP may require measuring only a small subset of the variable. The saving in measurements may be quite amazing. In the case of converters, one may reduce the measurements to less than one percent.

## References

- [1] Anderson, T. W. (1984), *An Introduction to Multivariate Statistical Analysis, 2nd ed.*, Wiley, New York.
- [2] Anderson, T. W. and Rubin, Herman (1956), *Statistical Inference in Factor Analysis*, Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability (Jerzy Neyman, ed.) 5, 111-150. University of California Press, Berkeley, California.
- [3] Box, G.E.P., Hunter, W.G. and Hunter, J.S. (1978), *Statistics for Experimenters*, Wiley, New York.
- [4] Boylestad, R. and Nashelsky, L. (1996), *Electronic Devices and Circuit Theory, 6th ed.*, Prentice Hall, London.
- [5] Ding, A.A. (1996), *Prediction Intervals and Confidence Intervals for Neural Networks and HELP*, Ph.D. thesis, Cornell University, Ithaca, New York.
- [6] Filliben, J.J. and Li, K-C. (1997), *A Systematic Approach to the Analysis of Complex Interaction Patterns in Two-level Factorial Designs*, Technometrics 39, 3, 286-297.
- [7] Gabriel, R. R. (1978), *Least Squares Approximation of Matrices by Additive and Multiplicative Models*, JRSS B 40, 2, 186-196.
- [8] Hwang, J.T.G. and Ding, A.A. (1997), *Prediction Intervals for Artificial Neural Networks*, Journal of American Statistical Association, 92, 748-756.

- [9] Johnson, R.A. and Wichern, D.W. (1992), *Applied Multivariate Statistical Analysis*, Prentice Hall, New Jersey.
- [10] Koopmans, T.C. and Reiersøl, O. (1950), *The Identification of Structural Characteristics*, Annals of Mathematical Statistics 21, 165-181.
- [11] Reiersøl, O. (1950), *On the Identification of Parameters in Thurstone's Multiple Factor Analysis*, Psychometrika 15, 121-149.
- [12] Stenbakken, G. N. (1996), *Effects of Nonmodel Errors on Model-based Testing*, IEEE Transaction on Instrumentation and Measurement 45, 2, 384-388.
- [13] Stenbakken, G. N. and Souders, T. M. (1987), *Test Point Selection and Testability Measures via QR Factorization of Linear Models*, IEEE Trans. Instrum. Meas. Vol. IM-36, No. 2, 406-410, June.
- [14] Souders, T. M. and Stenbakken, G. N. (1991), *Cutting the High Cost of Testing*, IEEE Spectrum, March 1991, p 48-51.
- [15] Thurstone, L.L. (1947), *Multiple Factor Analysis*, University of Chicago, Chicago, Illinois.
- [16] Yin, Y. Q., Bai, Z. D. and Krishnaiah, P. R. (1988), *On the Limit of Largest Eigenvalue of the Large Dimensional Sample Covariance Matrix*, Probability Theory and Related Fields 78, 509-521.