Goodness-of-fit and confidence intervals of approximate models

Lourens J. Waldorp*[1,2], Raoul P.P.P. Grasman[1], and Hilde M. Huizenga[1]

1 University of Amsterdam, Dept. of Psychology, Amsterdam, the Netherlands
2 University Maastricht, Dept. of Cognitive Neuroscience, Maastricht, the Netherlands
*email: waldorp@psy.uva.nl

keywords model fit, approximation error, sandwich parameter covariance, robust covariance

Abstract

If the model for the data are strictly speaking incorrect, then how can one test whether the model fits? Standard goodness-of-fit (GOF) tests rely on strictly correct or incorrect models. But in practice the correct model is not assumed to be available. It would still be of interest to determine how good or how bad the approximation is. But how can this be achieved? If it is determined that a model is a good approximation and hence a good explanation of the data, how can reliable confidence intervals be constructed? In this paper an attempt is made to answer the above questions. Several GOF tests and methods of constructing confidence intervals are evaluated both in a simulation and with real data from the internet based daily news memory test.

1 Introduction

One of the challenges in psychology is explaining the data obtained in experimental research. Such an explanation is often given in terms of a model. It appears quite difficult to determine whether a model, and hence an explanation, is adequate. One of the causes of this is that no model is the truth (White, 1981; Golden, 1995). Consequently, testing whether it is the truth seems pointless. At least four questions then arise concerning model fit or goodness-of-fit (GOF), which require answers. First, given that a model is incorrect, that is, the model is not true, how good is the approximation, and how do you test this? Second, a common assumption is that the data are uncorrelated. This is often, however, not the case. In standard analyses these correlations are ignored. So how can these correlations be incorporated adequately in the analyses? Third, if the model is incorrect, what distribution should be used to test the hypothesis that a model fits the data? And fourth, to interpret the model, that is, to interpret the parameters of the model, confidence intervals of parameters are used. If approximations to the truth are used, how can we get reliable confidence intervals?

To see why and in what way these questions are important, an example is first given to illustrate the problems involved. Consider the example of a regression analysis of a memory study on forgetting of learned nonsense syllables (described in Reisberg, 2001, p. 204). In this study, participants were asked to recall a list of nonsense syllables at several retention intervals. Typically, an increase in retention interval showed a decrease in the mean number of recalled syllables, as can be seen in Fig. 1. A regression analysis for this example involves a model which predicts the mean of the participants for each of the retention intervals. At first glance, an exponential model seems well suited (dashed line in Fig. 1). Suppose that we used a linear function (dotted line in Fig. 1) to approximate the true underlying process. The linear function is obviously incorrect, but how bad is it and how can that be

tested? In some cases a linear approximation could be reasonable. For example, if only a subset of the retention intervals were to be analyzed, in which the curve is (nearly) linear, then the linear approximation could be considered good.

The parameters α and β of the linear function $\alpha + \beta x$, where x contains the six retention intervals, can be estimated by, for example, least squares. Since the same participants were measured repeatedly, a correlation is expected between the retention intervals. This has to be taken into account in determining whether the model fits the data. But the question is how. Standard GOF tests do not account for the correlations in the data. The interpretation of the model, our explanation of the data, depends in part on the confidence we have in the estimates of the parameters, that is, it depends on the confidence intervals. A small confidence interval of β means that the slope of the line can vary to a small extent. And so the slope is probably significantly different from zero. If this is true, then the linear function might be a good explanation of the data. But either correlations in the data or the assumption of a true model can yield too small confidence intervals (Waldorp, 2005). This could lead to the incorrect conclusion that a model is a good explanation for the data.

Testing whether the linear function fits, means testing the null hypothesis that there is a relatively small difference between the data and the prediction from the linear function. Such an hypothesis test often requires that the reference distribution, that is, the distribution of the statistic, is correct. Quite often the reference distribution is chi-square. Unless the data are normally distributed, it is often the case that the GOF test is chi-square distribution is inappropriate. So when only a relatively small number of observations is available, using the chi-square distribution as a reference distribution can lead to rejecting too often the null hypothesis that the model fits. It is therefore difficult to determine whether the linear approximation is adequate.

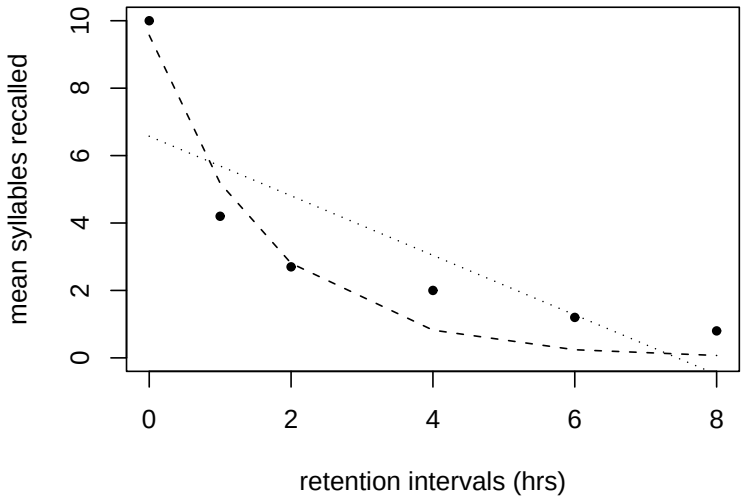

Fig. 1. Forgetting data described in Reisberg (2001, p. 204) (dots), the estimated exponential function (dashed line), and the estimated linear function (dotted line), both estimated by ordinary LS.

A possible solution to the three first problems is a test statistic known as Hotelling's t-square (T^2). Hotelling's test is usually associated with testing the hypothesis that the population means are equivalent in multivariate analysis of variance (MANOVA) (Johnson and Wichern, 2004). But basically it is tested whether the difference between the data and the prediction is zero, which is the same as other GOF tests. If in MANOVA the hypothesis were to be changed to the statement that the prediction of the model for the mean of the data is correct, then Hotelling's test would correspond to a GOF test as required. The advantages of using Hotelling's test as opposed to most other GOF tests are that (1) it can be used for models that are not true, (2) the correlations in the data are incorporated, and (3) it is an exact statistic, that is to say, the assumed distribution is exactly correct if the data are normally distributed. If the data are not normally distributed, then a sufficient number of observations can still lead to correct results, as with other GOF tests.

Of course in practice we need to estimate the parameters of the model, and so the assumption of MANOVA that the hypothesis (prediction from the model) is fixed, is not satisfied. In fact, considering that the estimated parameters are random variables, the model with estimated parameters is also a random variable. Consequently, the distribution of such a modified Hotelling's test is unknown. By using a projection based on the first-order derivative of the (incorrect) model of the data, the distribution of this statistic can be determined. Although this statistic will no longer be exact, it is still expected to outperform both the traditional and original Hotelling's test.

A related problem is that in GOF it is assumed that the model is either correct or incorrect whereas in model specification it is often assumed that the model is only an approximation, and so always incorrect. A GOF test will therefore nearly always indicate that an approximate model does not fit. Following Browne and Cudeck (1993) a measure of approximate fit is used to quantify the amount of misfit of the model.

To determine reliable confidence intervals for approximate (incorrect) models, traditional methods are inappropriate (Golden, 1995). It is shown that either bootstrap or sandwich estimates should be used when the model is approximate or when the assumption on the noise correlations is incorrect (White, 1980, 1982; Hastie et al., 2001; Kauermann and Carroll, 2001; Golden, 1995).

The paper is organized as follows. After a brief definition of approximate models, GOF testing is discussed. Next four different estimators of standard errors for confidence intervals are presented. In a numerical example three GOF tests and four methods of determining standard errors are compared. Finally, the results of the GOF and standard errors are applied to real data from the daily news memory test (Meeter et al., 2004).

2 Theory

2.1 Misspecified or approximate models

Consider n independent observations $Y_1, \ldots, Y_n$ of a vector with p variables (retention intervals in the example) $Y_j = (Y_{1j}, \ldots, Y_{pj})'$. The observations are identically distributed according to the probability distribution function $F_0 = F_{\theta_0}(y)$, parameterized by the q vector θ_0, the true value. The mean and variance of each Y_j are denoted by μ and Σ, respecively. If Y is normally distributed, for example, then the true vector θ_0 contains the mean μ and the unique elements of the variance matrix Σ. A model is a set S_θ of distributions F_θ indexed by the parameter vector θ. The model S_θ does not always contain the true distribution F_0. Such a model is said to be misspecified or approximate. It is generally assumed that a model is an approximation (see White, 1982; Golden, 1995; Zucchini, 2000, for a similar definition of approximate models).

To illustrate the principle of the above definition of approximations, consider the regression analysis of the data in the example. Lets assume that the true model F_0 is a normal distribution, and that for each observation $Y_j = \mu_{\theta_0}(x) + e_j$, where $\mu_{\theta_0}(x) = \alpha \exp(-\beta x)$ and e_j is distributed as $N(0, \Sigma)$. So the mean of F_0 is the function $\mu_{\theta_0}(x)$ with parameters $\theta_0 = (\alpha, \beta)'$, and we assume Σ known. Suppose we try to approximate the exponential function for the mean by a linear function, $f_\theta(x) = \theta_1 1_p + \theta_2 x$, where 1_p is a p vector of ones. This means that the model S_θ contains all normal distributions with Σ and means obtainable by the linear function. Then the correct mean of F_0 cannot be obtained using $f_\theta(x)$. Hence, the model S_θ does not contain F_0, and is therefore misspecified.

Define an objective function Q_θ which is to be minimized. This could, for instance, be the negative log-likelihood function, or the least squares function (LS). Since regression analysis is the main focus of the paper, the LS function is adopted. The LS function is defined by $Q_\theta(y) = (y-\mu)'\Sigma^{-1}(y-\mu)$. This is in fact the generalized LS (GLS) function, since the (known) covariance matrix Σ is used to account for correlated noise e. In the example, the retention intervals are correlated, and so Σ can downweight correlated noise. In practice, Σ is not known and has to be estimated. An unbiased, nonparametric estimate of Σ is (Muirhead, 1982)

$$S = \frac{1}{n-1} \sum_{j=1}^{n} (Y_j - \bar{Y})(Y_j - \bar{Y})'. \tag{1}$$

The nonparametric estimate S is convenient since usually the interest is not in the noise structure and no structured model is used in this estimate. If the matrix Σ is simply taken as being proportional to the identity matrix, that is $\Sigma = \sigma^2 I_p$, then Q_θ is called the ordinary LS (OLS) function.

2.2 Goodness of fit

We are looking for a GOF test that can distinguish between a correct and incorrect model with high accuracy. But if the model is incorrect, then an estimate of how good or bad the approximation is, should be available.

To use a GOF test, the null hypothesis should be available, together with reasonable distributional assumptions about the data. In the forgetting example, the null hypothesis could be $H_0 : \mu = f_\theta$, the mean of the distribution of the data is structured according to an exponential function $f_{\theta_0} = \alpha \exp(-\beta x)$. Alternatively, the null hypothesis could be that $f_\theta = \theta_1 1_p + \theta_2 x$.

A distributional assumption could be that the data Y_j are $N(\mu, \Sigma)$, that is, the data are multivariate normally distributed with mean μ and covariance Σ. Then if a specified (fixed) θ is available and if Σ is known, a traditional statistic is given by

$$C_\theta = n(\bar{Y} - \mu_\theta)' \Sigma^{-1} (\bar{Y} - \mu_\theta) \tag{2}$$

which is χ^2_p distributed under H_0 that the mean μ_θ is correct (Muirhead, 1982), where χ^2_p refers to the chi-square distribution with p degrees of freedom. This test is referred to as the Gaussian Residual (GR) test. The test rejects the null hypothesis for values $C_\theta > \chi^2_p(\alpha)$, where $\chi^2_p(\alpha)$ is the upper quantile of the χ^2_p distribution at significance level α. If no specified θ is available and Σ is unknown, then the estimates $\mu_{\hat{\theta}}$ and S can be used. If $\hat{\theta}$ is normally distributed, then $C_{\hat{\theta}} = n(\bar{Y} - \mu_{\hat{\theta}})' S^{-1} (\bar{Y} - \mu_{\hat{\theta}})$ converges to a χ^2_{p-q} distribution as n grows large (Van der Vaart, 1998). The convergence of $C_{\hat{\theta}}$ in distribution to χ^2_{p-q} is based on the convergence in probability of $\hat{\mu}$ and S to the constants μ_0 and Σ, respectively (Ferguson, 1996). But the convergence can be 'slow', that is, the chi-square distribution is appropriate only if the number of observations n (subjects in the retention example) is very large. The rate of convergence is even slower when the ratio of the number of variables (retention intervals) to the number of observations p/n is large. Consequently, the GR test will reject H_0 too often when there are not enough observations to justify the limiting chi-square distribution.

Alternatively, Hotelling's test could be used if the Y_j are $N(\mu, \Sigma)$ or if n is large. This test is known from MANOVA, where it is tested whether the mean from the data is equivalent to that of the assumed population. Hotelling's test can also be used to test whether a hypothesized function for the mean of the data fits. The two main advantages are that (1) estimating the correlations in the data are accounted for, and (2) it is an exact statistic, that is to say, the assumed distribution is exactly correct if the data are normally distributed. This make Hotelling's test well suited to use in cases where a small number of observations is available. Hotelling's test is defined as (Muirhead, 1982, p. 98)

$$T^2_\theta = n(\bar{Y} - \mu_\theta)' S^{-1} (\bar{Y} - \mu_\theta) \tag{3}$$

and is under H_0 that the mean μ_θ is correct distributed as $kF_{p,n-p}$, with $k = (n-1)p/(n-p)$, and $F_{p,n-p}$ refers to the central F-distribution with p and $n-p$ degrees of freedom. Although this test is more appropriate for small samples than the GR test and does not assume Σ known, it does assume that some θ is known and fixed. Often only an estimate $\hat{\theta}$ is available in regression analysis. If the estimate is included in Hotelling's T^2 then the distribution of this statistic is unknown. This is because the additional random variable $\mu_{\hat{\theta}}$ is included so that the variance of the residual $\bar{Y} - \mu_{\hat{\theta}}$ is no longer simply Σ; it now depends on both $\bar{Y}$ and $\mu_{\hat{\theta}}$. If the distribution of the residual were known, then a modified $T^2_{\hat{\theta}}$ could be used. Such a test would have the advantage of an exact distributional result and would allow using the estimates $\hat{\theta}$ and S for the test. Unfortunately, in general no such exact test exists. An approximate version does exist, however, as is shown next.

The approximate test $T^2_{\hat{\theta}}$ utilizes the estimates $\hat{\theta}$ and S, so hardly any prior knowledge is required. The modified version $T^2_{\hat{\theta}}$ approximates the distribution of $\bar{Y} - \mu_{\hat{\theta}}$ by a projection (see the Appendix), such that only the degrees of freedom of the numerator p of the original Hotelling's T^2_θ have to be adjusted to $p-q$. The projection uses the first-order derivatives of the possibly misspecified model, such that a tangent plane is as close as possible to the true model in the least squares sense (see White, 1980, for approximation in least squares sense). If the model is misspecified in terms of the mean, then this can be defined as $h = \sqrt{n}(\mu_0 - f_\theta)$, where f_θ is a possibly misspecified function for the mean. This result is stated as a theorem. A proof of the theorem is given in the Appendix.

Theorem 1. *Assume that Y_j are multivariate normally distributed with mean $\mu_0 = \mu(\theta_0)$ and covariance Σ, and that the function f_θ is continuous and has continuous first-order partial derivatives and converges in probability to $f_* = f_{\theta_*}$ as $n \to \infty$. Then*

$$T^2_{\hat{\theta}} = n(\bar{Y} - f_{\hat{\theta}})' S^{-1} (\bar{Y} - f_{\hat{\theta}}) \sim K F_{p-q,n-p}(\delta_h) \tag{4}$$

with $K = (n-1)(p-q)/(n-p)$ and $F_{p-q,n-p}(\delta_h)$ the noncentral F-distribution with noncentrality parameter $\delta_h = n(\mu_0 - f_*)'\Sigma^{-1}(\mu_0 - f_*)$. The distribution of $T^2_{\hat{\theta}}$ when the null-hypothis is true ($\delta_h = 0$) is the central F-distribution $F_{p-q,n-p}$.

According to the theorem we know the distribution of the modified Hotelling's test for a function that does not have to be correct in the strict sense, but the function must be sufficiently smooth. We need to establish that the modified Hotelling's test statistic satisfies the basic requirement of consistency. Informally, a test that is able to distinguish between the null and alternative hypotheses exactly for large n, is said to be asymptotically consistent (Van der Vaart, 1998). Such a test satisfies two conditions. The first is that as $n \to \infty$ the test rejects the null hypothesis when it is true (Type I error) at most with probability α, with α typically 0.05. A test that satisfies this is called asymptotically of level α. The second condition is that the test rejects a model when it is not true (power) with probability 1 as $n \to \infty$. It is shown that the approximate Hotelling's test is asymptotically consistent.

The power function is defined as the probability that $T^2_{\hat{\theta}}$ is larger than a critical value for which the null hypothesis is rejected (Van der Vaart, 1998). It is convenient to write the power function in terms of the reparameterization $h = \sqrt{n}(\theta - \theta_*)$, where θ_* could be the true value θ_0. Let $||\tilde{Z}||^2$ be distributed as $KF_{p-q,n-p}$, then the power function is

$$P_\theta(T^2_{\hat{\theta}}/(n-1) > c_\alpha) \to 1 - P_h(||\tilde{Z} + \sqrt{\delta_h}||^2 \leq c_\alpha)$$

as $n \to \infty$, where $c_\alpha = (p-q)/(n-p)F_{p-q,n-p}(\alpha)$ denotes the critical value based on the central F-distribution ($\delta_h = 0$) and the significance level α. It can be seen that for $||h|| > 0$, when the alternative hypothesis H_1 is true, the power tends to 1. This also means that the Type II error (not rejecting H_0 when it is false) tends to zero. If the null hypothesis H_0 is true, on the other hand, then $||h|| \to 0$ and $\theta_* = \theta_0$. Then $P_h(||\tilde{Z}||^2 \leq c_\alpha) \to 1$ as $\alpha \to 0$, and so the test is asymptotically a level α test. These two results together yield the test asymptotically consistent.

The noncentrality parameter is a (Mahalanobis) distance between the true and approximate function with respect to the inverse covariance matrix Σ. It can be used to determine an overall measure of approximation without assuming explicitly that the proposed model is completely correct. Such additional information can be used to determine whether the approximate model is a good enough approximation. This measure has the advantage that "close" and "distant" between the true and approximate model are defined. For this to work, a good estimate of the noncentrality parameter is required.

Browne and Cudeck (1993) give such an estimate and define a measure for which close and distant are defined, the so called root mean square error of approximation (RMSEA). Assume that the conditions of theorem 1 are satisfied. Then the expectation of $T^2_{\hat{\theta}}$ is equal to $\frac{p-q}{n} + \delta_h$, since the covariance matrix S is independent of the projected residual (see the Appendix). This leads to an estimate of the noncentrality parameter

$$\hat{\delta}_h = \max\left\{T^2_{\hat{\theta}} - \frac{p-q}{n}, 0\right\}. \tag{5}$$

With the estimate $\hat{\delta}_h$, the RMSEA can be computed by $\hat{\epsilon} = \sqrt{\hat{\delta}_h/(p-q)}$. The RMSEA can be interpreted as a discrepancy measure between the true and approximate model per degree of freedom (since $T^2_{\hat{\theta}}$ is χ^2_{p-q} asymptotically). The RMSEA is scale-free and a model fits perfectly if $\hat{\epsilon}$, with the true δ_h, is zero. A close fit for an approximate model is if $\hat{\epsilon}$ is no larger than 0.1 (Browne and Cudeck, 1993). It has also been suggested that the RMSEA be used for testing approximate models (Browne and Cudeck, 1993), but this requires estimates of upper and lower bounds of $\hat{\epsilon}$ which can be difficult to determine.

2.3 Confidence intervals

An indication of the accuracy of parameter estimates of θ can be obtained by computing confidence intervals. Confidence intervals are often based on the asymptotic normal distribution. For instance, a simple 95% confidence interval for the single parameter a based on the estimate $\hat{a}$, which is assumed $N(a, \sigma^2/n)$, is defined as (e.g., Seber and Wild, 1989, Ch. 5)

$$\hat{a} \pm t_{p-q}(\alpha/2)\hat{\sigma}/\sqrt{n}, \tag{6}$$

where $t_{p-q}(\alpha/2)$ is the quantile of the Student t-distribution at $\alpha/2$ with $p-q$ degrees of freedom, and $\hat{\sigma}/\sqrt{n}$ is an estimate of the standard error (se). It can be seen that an accurate confidence interval depends critically on the estimate of the standard error.

There are several ways to compute the standard error. If either the model or the noise characteristics are incorrect, the standard method to estimate the standard error could be inaccurate (see e.g., White, 1980, 1982; Kauermann and Carroll, 2001; Golden, 1995). In this section four methods to compute the standard error estimates are compared: (1) the inverse of the second-order derivative (Hessian) matrix, (2) the Hessian with a different noise variance estimate, (3) the sandwich estimator, and (4) the nonparametric bootstrap estimate. The computation of the standard errors are discussed in terms of the linear function $f_\theta(x) = \theta_1 1_p + \theta_2 x$.

Let X be a $p \times q$ fixed matrix, and θ a q vector, then the linear function can be written as $f_\theta(x) = X\theta$. In the approximate function of the example of forgetting data $X = (1_p, x)$ and $\theta = (\theta_1, \theta_2)$. The LS function for this approximation is

$$Q_\theta(Y) = \frac{1}{n}\sum_{j=1}^{n}(Y_j - X\theta)'\Sigma^{-1}(Y_j - X\theta). \tag{7}$$

The estimate $\hat{\theta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\bar{Y}$ is obtained by minimizing $Q_\theta(Y)$ with respect to θ. The Hessian method is to compute the second-order derivative of Q_θ (e.g., Seber and Wild, 1989). This is $J = 2(X'\Sigma^{-1}X)$. The standard errors are the square root of the diagonal elements of

$$2J^{-1} = (X'\Sigma^{-1}X)^{-1} = \begin{pmatrix} 1_p'\Sigma^{-1}1_p & 1_p'\Sigma^{-1}x \\ x'\Sigma^{-1}1_p & x'\Sigma^{-1}x \end{pmatrix}^{-1} \tag{8}$$

This assumes that the linear function is true (see the Appendix on the sandwich estimator). If $\Sigma = \sigma^2 I_p$, then the familiar covariance matrix $2J^{-1} = \sigma^2(X'X)^{-1}$ is obtained. The variance estimate σ^2 can be estimated in two ways: from the model $\hat{\sigma}^2 = Q_{\hat{\theta}}(Y)/(p-q)$, with Q the LS function, or based on the mean of the observations $s^2 = \sum_{i,j}(Y_{ij} - \bar{Y}_i)^2/p(n-1)$. If the approximate function is used, it can be

expected that $2s^2J^{-1}$ will give better results than $2\hat{\sigma}^2J^{-1}$ because s^2 is not affected by bias in the model as is $\hat{\sigma}^2$. Often, the expectation of J is used instead of J itself. In linear models, however, the expected Hessian and the Hessian are equivalent.

The sandwich estimator does not assume that the linear function is true. It is derived from computing the covariance of $\hat{h} = \sqrt{n}(\hat{\theta} - \theta_*)$ which for the linear approximating function can be rewritten as $(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(\mu_0 - X\theta_* + e_j)$ (see e.g., Van der Vaart, 1998; White, 1980). This has mean zero and covariance (see the Appendix)

$$J^{-1}IJ^{-1} = J^{-1}4\begin{pmatrix} 1_p'\Sigma^{-1}\Sigma_*\Sigma^{-1}1_p & 1_p'\Sigma^{-1}\Sigma_*\Sigma^{-1}x \\ x'\Sigma^{-1}\Sigma_*\Sigma^{-1}1_p & x'\Sigma^{-1}\Sigma_*\Sigma^{-1}x \end{pmatrix} J^{-1}, \tag{9}$$

where $\Sigma_* = D + \Sigma = (\mu_0 - X'\theta)(\mu_0 - X'\theta)' + \Sigma$. It can be seen that if the linear function were true, then $D = 0$ and $J^{-1}IJ^{-1} = 2J^{-1}$, which is seen to be equivalent to the Hessian method. (See the Appendix for a more elaborate discussion of the sandwich estimator.)

Finally, the nonparametric bootstrap estimates are computed by taking a random sample of size n with replacement, B times from the data $Y_1, \ldots, Y_n$ and estimating θ each time. Then the standard error is obtained by taking the square root of the diagonal of (Davison and Hinkley, 1997)

$$\frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i - \bar{\theta})(\hat{\theta}_i - \bar{\theta})', \tag{10}$$

where $\hat{\theta}_i$ is the *i*th estimate and $\bar{\theta}$ is the average of B estimates $\hat{\theta}$.

In estimating the standard errors the covariance matrix Σ of e_j also presents a problem. The nonparametric and unbiased estimate S is available but is subject to sampling error. If the ratio of the number of observations n to the number of variables (retention intervals in the example) p is unfavorable, then it is to be expected that the standard errors will be affected by the poor estimate (Waldorp et al., 2001). Therefore, it is expected that for low n the standard error estimates will be less accurate than for large n.

3 Numerical example

Simulations are presented to evaluate GOF tests and methods of constructing confidence intervals. In the present simulations we used the set-up of the example of the forgetting curve introduced in the introduction to show the small sample behavior of the three GOF tests: GR (C_θ), original Hotelling's (T^2_θ), and modified Hotelling's ($T^2_{\hat{\theta}}$). We assume that the true function is exponential. The parameters for the true

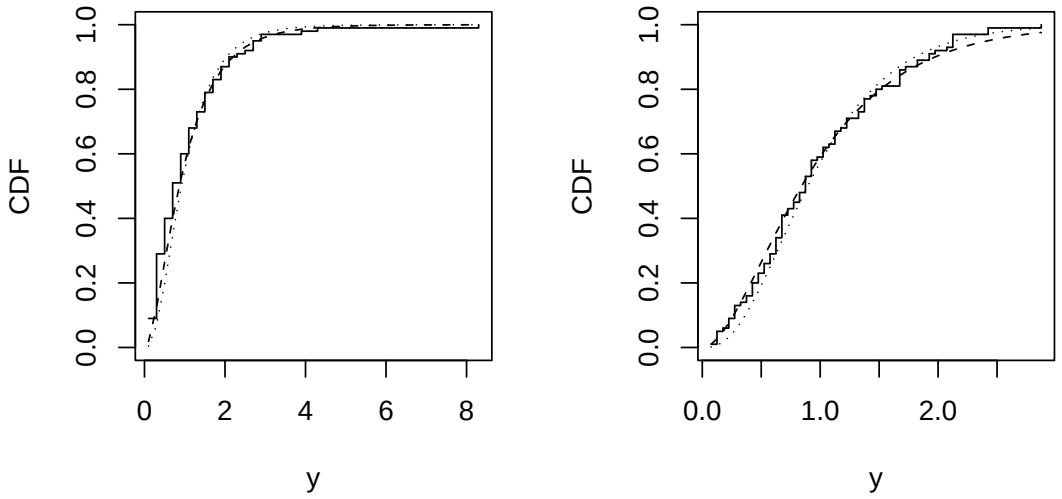

Fig. 2. CDF of the central F-distribution of $T^2_{\hat{\theta}}$ for the simulated exponential forgetting data estimated by the exponential function (see text for details). Left: The empirical CDF (solid line), the theoretical central CDF of the modified Hotelling's test with $F(4, 24)$ (dashed line), and the theoretical CDF of the original Hotelling's test $F(6, 24)$ (dotted line). Right: the same only with CDF $F(4, 196)$ and $F(6, 196)$.

exponential model are $\theta_0 = (9.569, 0.613)$, which were estimated by OLS from the data of Fig 1. The approximate function is linear $f_\theta(x) = \theta_1 1_p + \theta_2 x$. First the simulations of the GOF tests will be presented followed by a simulation study on the confidence intervals.

To mimic the temporal dependency in a time series such as the forgetting data, an AR(1) process was assumed to govern the temporal correlation. This gives rise to a Toeplitz matrix which has elements $(\Sigma)_{ij} = \gamma^{|i-j|}$, with $|\gamma| < 1$ (Chatfield, 1989). Since in the example the retention intervals were in units of hours, it seems warranted to assume a strong correlation, which is taken to be $\gamma = 0.8$. In each run $N(0, \psi^2\Sigma)$ noise is generated for $n = 30, 50, 100,$ and 200, with ψ^2 a scaling parameter such that 10% of the maximum of μ_{θ_0} was used as the noise variance for the averaged data $\bar{Y}$. The absolute correlations generated with this matrix were between 0.35 and 0.80 with an average of 0.51. The matrix S is estimated and is used in the LS function $Q_\theta(Y, S)$, from which the three statistics can be computed: the approximate version $T^2_{\hat{\theta}}$ with $p - q$ and $n - p$ degrees of freedom (df) from theorem 1, the original version T^2_θ with p and $n - p$ degrees of freedom, and $C_{\hat{\theta}}$ with p degrees of freedom. In the example the number of retention intervals is $p = 6$ and the number of parameters for both models is $q = 2$.

It is first shown that the distribution of theorem 1 is valid. Fig. 2 shows that the theoretical cumulative distribution function (CDF) of $T^2_{\hat{\theta}}$ approximates the empirical distribution function well. This is true for both $n = 30$ and 200. It is difficult to see, however, whether the CDF with 4 (modified) or with 6 (original) df is more appropriate. The quantile-quantile (QQ) plots in Fig. 3 show that the theoretical CDF with 4 df (left panel) corresponds best to the empirical CDF in the larger quantiles.

The consequence of inaccurate "tail behavior" is that the nominal significance level does not correspond to the true level, and so the null hypothesis is rejected too often. In Fig. 4 it can be seen that the GR test does not correspond to the theoretical χ^2_4 distribution even when $n = 200$ (right panel).

The tail probabilities of the tests give information on how often the null hypothesis is rejected when it is true. This should always be at most α, which in this example is set to 0.05. From the left panel of Fig. 5 it can be seen that the approximate test with 4 df is a level α test but that if either 6 df or the GR test is used, it rejects the null hypothesis too often when few observations are available. All three tests are asymptotically level α tests. To inspect the power of the tests the linear function is used as null hypothesis. The tests should reject this hypothesis as often as possible, that is the power should be close to 1. From the right panel of Fig. 5 it can be seen that the power for each of the three test statistics is close to 1.

Next the result of theorem 1 is tested when the model is misspecified, that is, the incorrect linear model is tested. Fig. 6 shows that, if the model is approximate, then the CDF is only good if the number of observations is large enough ($n = 200$). As there is a difference in slope between the empirical and theoretical CDF with 4 df the noncentrality parameter is not singly responsible for the misfit. This is confirmed by the accuracy of the noncentrality parameter: for $n = 30$ the relative difference $(\hat{\delta}_h - \delta_h)/\hat{\delta}_h$ is 0.036, and for $n = 200$ this is 0.048. This also means that the estimate of the RMSEA is quite accurate. These results suggest that the projection in the approximate Hotelling's test causes the misfit of the expected and sample distribution.

The four methods to estimate the standard errors are compared both with the approximate linear function and with the true exponential function. The methods are Hessian using the model for the variance $2\hat{\sigma}^2 J^{-1}$, Hessian using the means for the variance $2s^2 J^{-1}$, sandwich estimator $J^{-1} I J^{-1}$, and the bootstrap estimate $\text{se}(\theta_B)$

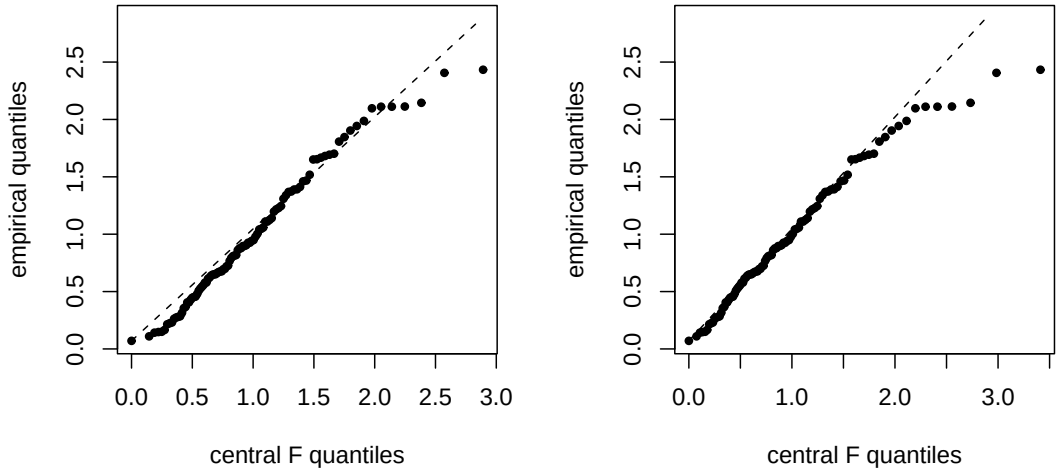

Fig. 3. Q-Q plots of the empirical CDF (solid dots) and the theoretical CDF (dashed line) with modified $F(4, 194)$ (left) and original $F(6, 194)$ (right) Hotelling's test.

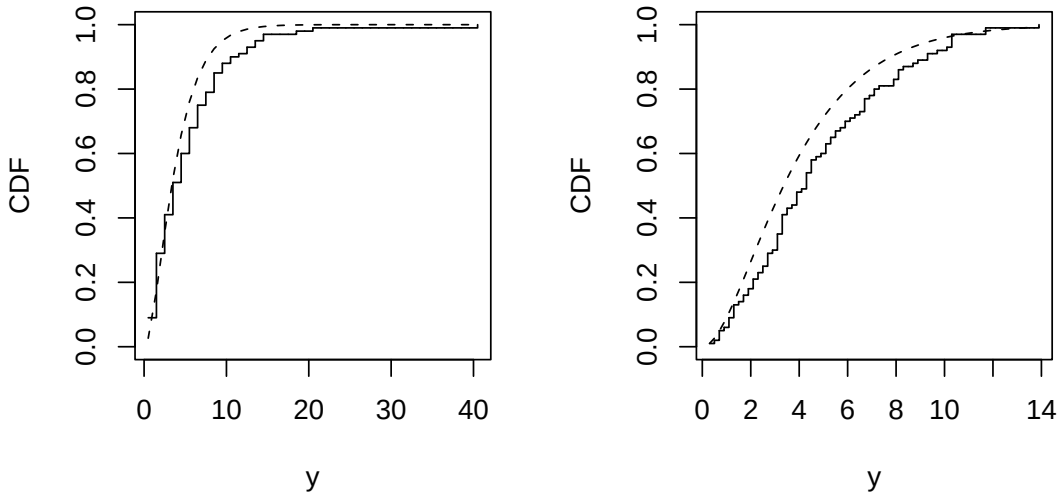

Fig. 4. The empirical CDF (solid line) based on $n = 30$ (left) and 200 (right) and the theoretical CDF (dashed line) of the χ^2_4 distribution.

based on $B = 100$. The same parameters for the exponential function and noise properties are used as mentioned above.

To compare these methods both when the true and the approximate function are used, three noise conditions are created: (1) e_j is $N(0, \sigma^2 I_p)$ and is estimated with $S = s^2 I_p$ or $\hat{\sigma}^2 I_p$, (2) e_j is $N(0, \Sigma)$ and is estimated with S, and (3) e_j is $N(0, \Sigma)$ but is estimated assuming $S = s^2 I_p$ or $\hat{\sigma}^2 I_p$. These conditions can make clear when each of the different methods to compute se is most accurate. To evaluate the methods the ratio of the estimated se and the "true" se is computed, which should equal 1 if the method is good. The "true" standard errors are computed from the standard deviation of the estimates from 100 simulations.

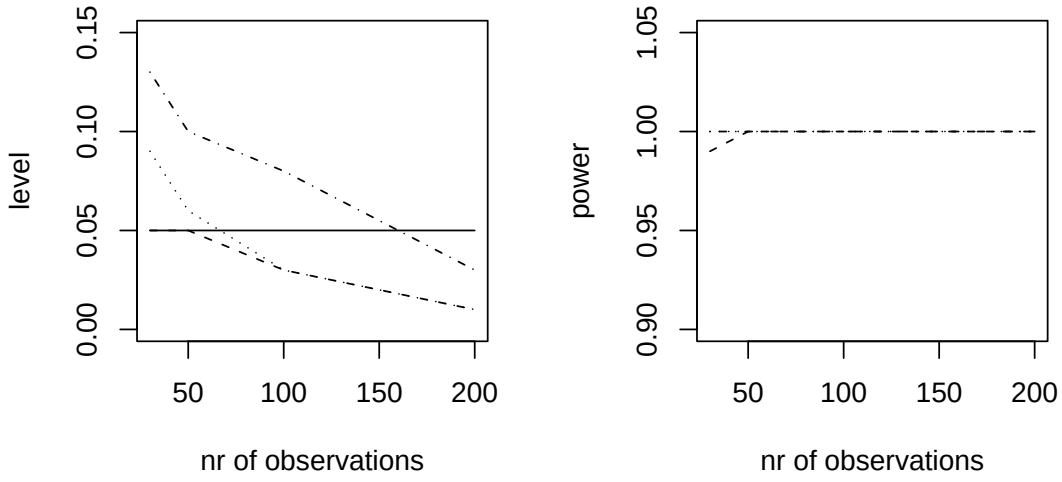

Fig. 5. Left: The level of the tests modified $T^2_{\hat{\theta}}$ (dashed line), original T^2_{θ} (dotted line), and GR $C_{\hat{\theta}}$ (dahsed-dotted line). The solid line is at 0.05, the nominal level of the test. Right: The power of the three tests.

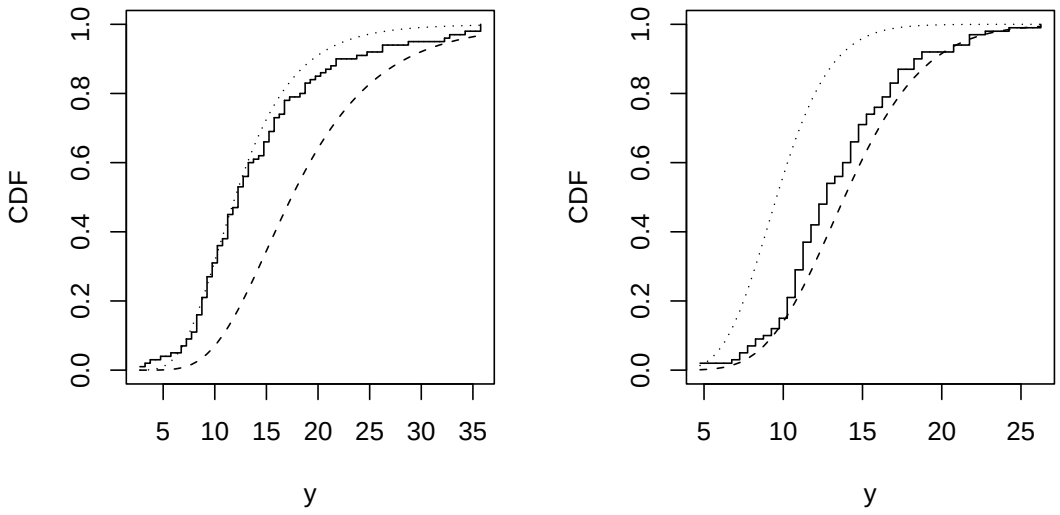

Fig. 6. The empirical and theoretical CDF of the F distribution based on $n = 30$ (left) and 200 (right) when the approximate (linear) function is used. The dotted line represents the theoretical CDF of the original Hotelling's test and the dashed line that of the approximate Hotelling's test.

In Fig. 7 the ratio of the estimated se to the true se shows that there is not much difference between the methods except when $e \sim N(0, \Sigma), S = s^2 I$. This is because the estimate $\hat{\sigma}^2$ depends on the assumption that the noise is uncorrelated, which is not the case.

When the approximate (linear) function is used it is seen in Fig. 8 that the biased estimate $\hat{\sigma}^2$ results in overestimation of the se with $2\hat{\sigma}^2 J^{-1}$. When the noise is correlated the other methods tend to underestimate the se in a similar way. If the noise structure is assumed incorrectly, that is $e \sim N(0, \Sigma), S = s^2 I$, then the sandwich and bootstrap estimator are accurate, but the others are not.

In the situation where nothing is known about the appropriateness of the model or of the noise, then both the sandwich and bootstrap se estimate appear most accurate.

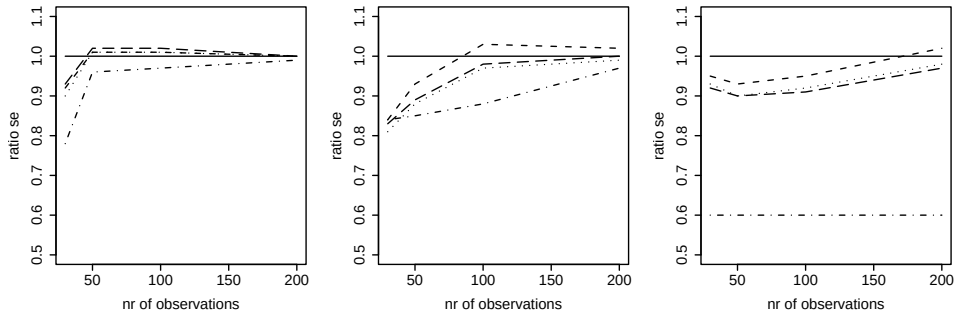

Fig. 7. The ratio of estimated and true se for the true (exponential) function for the different types of estimators: Hessian using the model $2\hat{\sigma}^2 J^{-1}$ dashed-dotted line, Hessian using the means $2s^2 J^{-1}$ long-dashed line, sandwich $J^{-1} I J^{-1}$ dotted line, bootstrap se(θ_B) short–dashed line. Left: white noise, middle: colored noise, right: colored noise but estimated as if white.

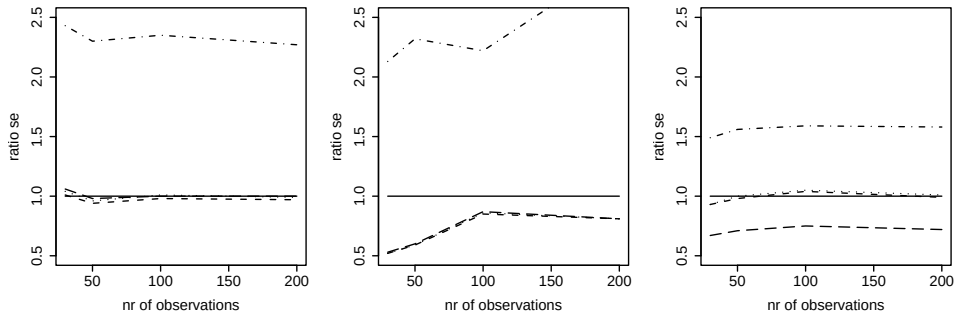

Fig. 8. Same as in Fig. 7 for the approximate (linear) function.

Even though the Hessian method $\hat{\sigma}^2 J^{-1}$ can be slightly more optimal in the situation when both the model and noise structure are known, the difference between methods is small compared to the inaccuracy when the model and noise structure could be incorrect.

4 Application to Daily News Memory Test

In the Daily News Memory Test (DNMT) participants were asked through internet to fill out a questionnaire about news events (Meeter et al., 2004). The questions considered in the present analysis were 4-alternative forced choice questions, in all about 30 to 40 filled out by 4239 Dutch participants. The analysis contained 60 out of the 365 days of retention intervals, since after 60 days the number of observations for each day dropped considerably and could therefore not be used in the present analysis. Corresponding to the analysis of Meeter et al. (2004) retention intervals were grouped in three consecutive days (bins), resulting in $p = 20$ intervals. Absolute correlations between the retention intervals ranged from 0.04 to 0.53 with an average of 0.23. For more details about the test and the participants see Meeter et al. (2004). The data are shown in the right panel of Fig. 9.

The GOF test requires normally distributed data whereas recall data is typically zero-one, Bernoulli distributed and so the sum is distributed as Binomial $\sum_{j=1}^{n} X_{ij}$, where for retention interval i, X_{ij} equals either 1 or 0. A normal approximation to the Binomial was used by creating bins at each of the retention intervals $i = 1, \ldots, p$ with length, $s_i = [n_i/n]$, where n_i is the number of recalled items at (bin) interval i, n is the number of repetitions set equal for all intervals, and $[x]$ denotes the integer part of x. The averages $\frac{1}{s_i} \sum_{j \in I_{i,k}} X_{ij}$, with $I_{i,k}$ the kth bin with s_i observations, are then assumed to be distributed as normal. For $n = 25$ the number of recalled items in the bins for each of the retention intervals s_i varied between 36 and 81, with an average of 73. An example of the quality of the approximation can be seen in the left panel of Fig. 9. The estimated normal curve (dashed line) shows that the approximation of the normal distribution is acceptable.

A model for the data used in Meeter et al. (2004) and presented in Chessa and

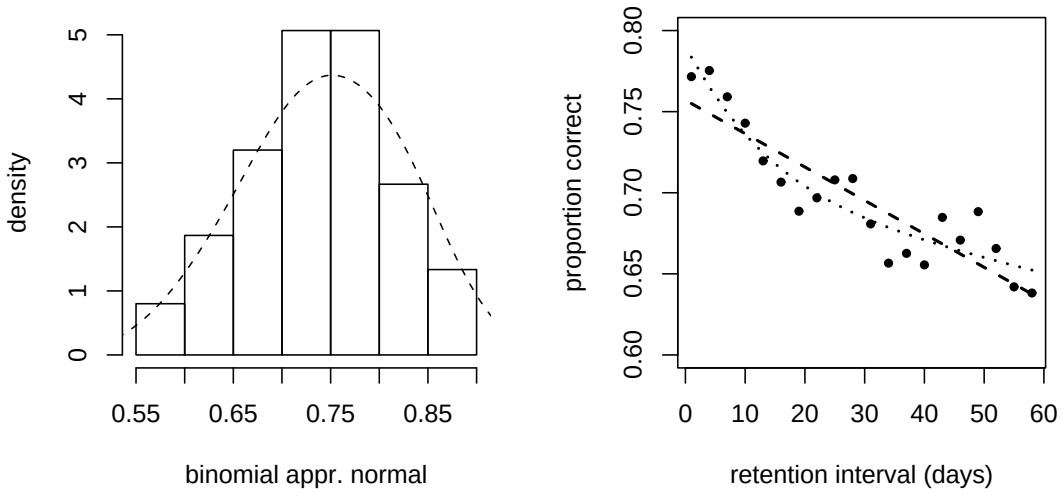

Fig. 9. Left: Example of the Binomial approximating the normal distribution. An estimate of the normal curve is shown in dashed line. Right: Data from the DNMT for 60 days (with bins of 3 consecutive days) (dots), and two fitted models, the MCM (dotted line) and the linear model (dashed line).

Table 1
Model fits for the memory chain model (MCM) and the linear model

	MCM	linear
$\hat{\theta}$	$(1.56, 0.09, -0.06, 0.002)$	$(0.76, -0.002)$
$2\hat{\sigma}^2 J^{-1}$	$(0.97, 0.42, 0.33, 0.02)$	$(0.09, 0.003)$
$J^{-1}IJ^{-1}$	$(0.04, 0.01, 0.01, 0.0003)$	$(0.003, 0.0008)$
GOF	$F(16, 5) = 0.0013, p = 1$	$F(18, 5) = 0.002, p = 1$
RMSEA	$\hat{\epsilon} = 0$	$\hat{\epsilon} = 0$

Murre (2002) is the memory chain model (MCM). This model assumes that memory strength can be modeled by a number of points in memory. These points could be either copies of the memory or some aspects associated with the memory. A memory is retrieved if one or more of these points can be obtained. Forgetting is modeled by the disappearance of the points. The MCM function is

$$f_\theta(x) = 1 - \exp\left[-\theta_1\left(\exp(-\theta_2 x) + \frac{\theta_3}{\theta_4 - \theta_2}[\exp(-\theta_4 x) - \exp(-\theta_2 x)]\right)\right].$$

The MCM model fits well to the data as can be seen in Fig. 9, right panel, dotted line. The parameters for the MCM model and the linear model, their standard errors, and their GOF and RMSEA indices are given in Table 1.

As can be seen, the standard errors of the parameters are larger when the standard Hessian method is used compared the sandwich method. This corresponds to the simulation results in the previous section. The GOF indicates that both the MCM and linear model fit well. No distinction between models can be made from

this index, which seems to correspond to the good fits for both models shown in Fig. 9. Furthermore, the RMSEA indicates that both models are good approximations. Therefore, it is concluded in this case that the linear model is sufficient in explaining the data.

5 Discussion

A model can be misspecified in terms of the mean and (co)variance. What was required was a way to determine goodness-of-fit (GOF) and reliable confidence intervals for interpretation of the model. To this end three tests were evaluated and four different methods of constructing confidence intervals were compared. The three tests were: the Gaussian Residual test, Hotelling's test, and a modified Hotelling's test that takes into account estimated parameters. Hotelling's modified test was seen to work well for possibly misspecified models, provided enough observations were available and the function for the mean is continuous. If the model does not fit, the root mean error of approximation (RMSEA) provides a way to quantify how good or bad the approximation is. The RMSEA is based on the noncentrality parameter of Hotelling's test and was seen to work well. So, if the model is misspecified, it can be determined to what to what extent it can be used to explain the data.

Hotelling's (modified) test also naturally takes into account the correlations in the data. This circumvents an often violated assumption that there is no dependency in the data. It was seen in the simulations and in the example of the daily news memory test that Hotelling's test accounts for this dependency well. Additionally, it was shown that for misspecified models reliable confidence intervals can be constructed by using either the bootstrap or sandwich estimate for standard errors of the parameters of the model. The standard Hessian method tends to overestimate the standard errors when the model is incorrect, which leads to confidence intervals that are too wide. Significance testing will then result in accepting the null hypothesis too often. With either the sandwich or bootstrap method, significance testing can be performed at the nominal level (usually 0.05).

The RMSEA can also be used to compare the amount of approximation of other possibly misspecified models. This was shown in the example of the daily news memory test. Of course, eventually, a model selection scheme should be used to distinguish between different models. The methods of Hjort and Claeskens (2003) and Claeskens and Hjort (2003) are good examples of model selection and model averaging where the sampling error of the selection process is taken into account. Their results also do not assume that the true model is known or that is among the set of possible models. Initially, however, the tool from the GOF test can serve as a rudimentary selection method.

A drawback of Hotelling's approach in general is that if not all means for $j =$

$1, \ldots, n$ are the same, then the T^2_θ (and $T^2_{\hat{\theta}}$) cannot distinguish between the weak hypothesis that the average of the (different) means is zero, or the hypothesis that all means are zero (Jensen and Ramirez, 1991). If it is reasonable to assume that the means are approximately equal then tests such as Hotelling's are appropriate.

Appendix

Proof of theorem 1. By using a linear approximation of $\hat{h}$, it is shown that $T^2_{\hat{\theta}}$ approximately has an $KF_{p-q,n-p}$ distribution. Let $\dot{f}_\theta$ denote the $p \times q$ matrix with first-order partial derivatives of f_θ with respect to the q parameters in θ. If $f_\theta(x)$ is sufficiently smooth with continuous first-order derivative, then a linear approximation by a Taylor-exapansion can be used as $f_{\hat{\theta}} = f_{\theta_*} + \dot{f}_{\theta_*}(\hat{\theta} - \theta_*) + o_p(n^{-1/2})$. This gives the approximation of the residual $\bar{Y} - f_{\hat{\theta}}$ solved for the difference $\hat{\theta} - \theta_*$ (see Th. 5.23, Ch. 5 Van der Vaart, 1998, for regularity conditions)

$$\hat{h} = \sqrt{n}(\hat{\theta} - \theta_*) = J_*^{-1} \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \dot{f}_*' \Sigma^{-1} (Y_j - f_{\hat{\theta}}) + o_p(1)$$

where $J_* = \dot{f}_*' \Sigma^{-1} \dot{f}_* + O_p(1)$ if $f_\theta(x)$ is regular and if the residual has a (biased) normal distribution. (see proof Th. 3.3 Appendix of White, 1981, for more restrictive but more readable conditions). With the two approximations, it can be seen that the residual $\bar{Y} - f_{\hat{\theta}}$ is distributed as $(I_p - P_*)\bar{Y} + o_p(1)$, with projection matrix $P_* = \dot{f}_*(\dot{f}_*' \Sigma^{-1} \dot{f}_*)^{-1} \dot{f}_*' \Sigma^{-1}$, since

$$\begin{aligned} \bar{Y} - f_{\hat{\theta}} &= \bar{Y} - f_{\theta_*} + \dot{f}_{\theta_*}(\hat{\theta} - \theta_*) + o_p(n^{-1/2}) \\ &= (I_p - P_*)(\bar{Y} - f_{\theta_*}) + o_p(n^{-1/2}) = (I_p - P_*)\bar{Y} + o_p(n^{-1/2}). \end{aligned}$$

The variance of this is $(I_p - P_*)\Sigma(I_p - P_*)' = \Sigma - \Omega$ with rank $m - q$. The matrix $\Omega = \dot{f}_* J_*^{-1} \dot{f}_*'$ is the variance matrix of $f_{\hat{\theta}}(x)$. Using the Moore-Penrose inverse for $\Sigma - \Omega$ in Hotelling's with S, also in $\hat{\Omega}$, T^2_a now becomes

$$\begin{aligned} T^2_{\hat{\theta}}/(n-1) &= n(\bar{Y} - f_{\hat{\theta}})'[(n-1)(S - \hat{\Omega})]^+(\bar{Y} - f_{\hat{\theta}}) \\ &\stackrel{d}{=} n(\bar{Y} - f_*)'(I_p - P_*)'[(n-1)S]^{-1}(I_p - P_*)(\bar{Y} - f_*). \end{aligned}$$

The distribution of $T_{\hat{\theta}}$ is noncentral F if $(I_p - P_*)Y_j$ is $N((I_p - P_*)\mu, \Sigma)$ and $(I_p - P_*)\bar{Y}$ and S are independent (Muirhead, 1982, p. 98). Normality is true by assumption. Independence can be shown by writing $\bar{Y} = \frac{1}{n} Y' 1_n$ and $S = \frac{1}{n} Y'(I_n - \frac{1}{n} 1_n 1_n')Y$, where $Y' = (Y_1, \ldots, Y_n)$, and realizing that $1_p'(I_n - \frac{1}{n} 1_n 1_n') = 0$. The degrees of freedom of the noncentral F-distribution are $p - q$ and $n - p$. This can be seen by considering an orthogonal and scaling transformation O and $\Sigma^{-1/2}$, respectively, such that $O(I_p - \Sigma^{-1/2} P_* \Sigma^{1/2})\bar{Y}$, which has the same distribution as $(||(I_p - P_*)Z||, 0, \ldots, 0)' = \tilde{Z}$, where Z is standard normal. Then by Theorem

1.4.2 (Muirhead, 1982, p. 26) $\tilde{Z}'\tilde{Z} = Z'(I_p - P_*)Z$ is $\chi^2_{p-q}(\delta_h)$. The noncentrality parameter δ_h follows from $(I_p - P_*)(\mu_0 - f_*) = \mu_0 - P_*\mu_0$, where $P_*\mu_0$ is the projection of μ_0 onto the tangent plane of f_θ at θ_*. The variance V of $\tilde{Z}$ yields for the first element in $\tilde{Z}$, $v_{11} - v'_{12}V_{22}^{-1}v_{21}$, which is χ^2_{n-p} (Muirhead, 1982). The ratio of these independent chi-square random variables gives the result.

Sandwich estimator. It is shown how an estimate of the standard error can be obtained for a local parameter $\hat{h} = \sqrt{n}(\hat{\theta} - \theta_*)$. Let X be a $p \times q$ fixed matrix, and θ a q vector, then the linear function can be written as $f_\theta(x) = X\theta$. In the approximate function of the example of forgetting data $X = (1_p, x)$ and $\theta = (\theta_1, \theta_2)$, where 1_p is a p vector of ones. The LS function for this approximation is

$$Q_\theta(Y) = \frac{1}{n}\sum_{j=1}^{n}(Y_j - X\theta)'\Sigma^{-1}(Y_j - X\theta).$$

Then the estimate $\hat{\theta}$ obtained by minimizing the LS function is $(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\bar{Y}$. The parameter $\hat{h}$ can be rewritten as

$$(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(\mu_0 - X\hat{\theta} + e_j).$$

To see that its mean is zero at θ_*, consider the prediction mean squared error (PMSE). The PMSE is in this case

$$E\{Q_\theta(Y)\} = (\mu_0 - X\theta)'\Sigma^{-1}(\mu_0 - X\theta) + p,$$

with first-order derivative $-2X'\Sigma^{-1}(\mu_0 - X\theta)$. This derivative is proportional to the mean $E\{\hat{h}\}$ above. Since θ_* minimizes the PMSE, its first-order derivative is zero at θ_*. Then, since $-2X'\Sigma^{-1}(\mu_0 - X\theta_*) = 0$, it follows that $E\{\hat{h}\} = 0$ as well if $\hat{\theta}$ converges in probability to θ_*. The covariance of $\hat{h}$ can now be obtained from the rewritten form and using the fact that its mean is zero; the covariance is then

$$E\{\hat{h}\hat{h}'\} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}(D + \Sigma)\Sigma^{-1}X(X'\Sigma^{-1}X)^{-1},$$

where $D = (\mu_0 - X'\theta)(\mu_0 - X'\theta)'$. This is the so-called sandwich matrix. It can be seen that if the mean function is correct, then $D = 0$ and the variance matrix reverts to the familiar result $(X'\Sigma^{-1}X)^{-1}$.

Acknowledgements

The authors would like to express their thanks to M. Meeter, J.M.J. Murre, and S.M.J. Janssen for the use of their data, and extensive discussions about the analyses. We would also like to thank the reviewers and the editor for their invaluable comments.

References

Browne, M., Cudeck, R., 1993. Alternative ways of assessing model fit. In: Bollen, K., Long, J. (Eds.), Testing structural equation models. Sage Publicaations, pp. 136–162.

Chatfield, C., 1989. The analysis of time series: An introduction. Bury St Edmunds: Chapman and Hall.

Chessa, T., Murre, J., 2002. A model of learning and forgetting I: The forgetting curve. Tech. Rep. 02-01, University of Amsterdam, NeuroMod.

Claeskens, G., Hjort, N., 2003. The focused information criterion. Journal of the American Statistical Association 98 (464), 900–916.

Davison, A., Hinkley, D., 1997. Bootstrap methods and their application. New York: Cambridge University Press.

Ferguson, T., 1996. A course in large sample theory. Bury st Edmunds: Chapman and Hall.

Golden, R., 1995. Making correct statistical inferences using the wrong probability model. Journal of Mathematical Psychology 39, 3–20.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer-Verlag, New York.

Hjort, N., Claeskens, G., 2003. Frequentist model average estimators. Journal of the American Statistical Association 98 (464), 879–899.

Jensen, D., Ramirez, D., 1991. Misspecified T^2 tests I. location and scale. Communications in Statistics: Theory and Methods 20 (1), 249–259.

Johnson, R., Wichern, D., 2004. Applied Multivariate Statistical Analysis, 5th Ed. New York: Prentice Hall.

Kauermann, G., Carroll, R., 2001. A note on the efficiency of sandwich covariance extimation. Journal of the American Statistical Association 96 (456), 1387–1396.

Meeter, M., Murre, J., Janssen, S., 2004. Remembering the news: A forgetting study with 14,000 participants. Memory and Cognition, in press .

Muirhead, R., 1982. Aspects of multivariate statistical theory. New York: John Wiley & Sons.

Reisberg, D., 2001. Cognition: Exploring the science of the mind. New York: Norton & Company.

Seber, G., Wild, C., 1989. Nonlinear regression. Toronto: John Wiley and Sons.

Van der Vaart, A., 1998. Asymptotic Statistics. New York: Cambridge University Press.

Waldorp, L., 2005. The wald test and cramer-rao bound for misspecified models in electromagnetic source analysis. IEEE Transactions on Signal Processing 53 (9), 3427.

Waldorp, L., Huizenga, H., Dolan, C., Molenaar, P., 2001. Estimated generalized least squares electromagnetic source analysis based on a parametric noise covariance model. IEEE Transactions on Biomedical Engineering 48, 737–741.

White, H., 1980. Using least squares to approximate unknown regression functions. International Economic Review 21 (1), 149–170.

White, H., 1981. Consequences and detection of misspecified and nonlinear regres-

sion models. Journal of the American Statistical Association 76 (374), 419–433.
White, H., 1982. Maximum likelihood estimation of misspecified models. Econometrica .
Zucchini, W., 2000. An introduction to model selection. Journal of mathematical psychology 44, 41–61.