Flight Dynamics Support Services II
FDSS-II-16-0049
CONTRACT # NNG14VC09C

FDSS-II
Task Order 16
TECHNICAL MEMORANDUM

Covariance Realism Evaluation Approaches

Issue Date: 10 JUL 2015

Prepared by:
M.D. Hejduk

Submitted by:
Ryan Frigm
Flight Safety Functional Lead

Omitron, Inc.
7051-A Muirkirk Meadows
Beltsville, MD 20705

Preface

This document describes assessing the realism of a covariance produced as part of the state estimates for spacecraft.

If you have any questions, please contact the following:

Ryan Frigm
FDSS-II Task Order 16 Functional Lead
GSFC Building 28
Greenbelt, Maryland 20771

Table of Contents

Purpose

The purpose of this white paper is to assist owner/operators in assessing the realism of a covariance produced as part of the state estimates for their spacecraft. Because it is a practical guide rather than theoretical treatment, this paper does not address broader questions of covariance formation or describe in detail the different methods for covariance propagation. What it does attempt to do is to enumerate the data types needed for covariance realism investigations, introduce the appropriate test statistics, and give some guidance for the evaluation and interpretation of such test statistics. Most of the information here derives from the author's experience with this problem over a period of many years, as there are few published studies on the subject and no practical guides. The particular focus is on determining the realism of covariance generation for actively-maintained spacecraft in the presence of precision position data (e.g., on-board GPS or telemetry tracking information); this is a more straightforward subset of the general covariance realism problem, which must consider unmaintained objects that lack this position information. Covariance realism assessment consists of three parts: collection/calculation of position error data, calculation of covariance realism test statistics, and proper assessment of those test statistics. After some introductory discussion, each of the three parts of this process will be addressed in detail.

General Remarks

It is important to remember that a state estimate is an estimate of a *mean* state (*i.e.* mean position and mean velocity), and a covariance is a stochastic characterization of the expected errors about that mean state. The governing assumption is that these errors in the individual components conform to a Gaussian distribution; this is how a covariance matrix alone is able to represent the error distribution without requiring higher-order tensors. The activity of covariance realism is to determine how well actual state errors conform to the Gaussian distribution that the covariance specifies.

This statement engenders a number of remarks. First, the present treatment will confine itself to the assessment of the realism of the position portion of the covariance. This is not necessarily the preferred position to take, as there are reasonable arguments for evaluating the covariance in element space (usually equinoctial elements) and considering all six elemental representations in any evaluation.[1] However, since the principal purpose of covariance generation at present is for conjunction assessment (CA) applications, and only the position portion of the covariance is used in most CA processes, it is acceptable in this case to limit oneself to a Cartesian representation and focus only on the position portion of the covariance. Furthermore, a previous study has demonstrated that potential non-Gaussian covariance behavior brought about by working in a Cartesian rather than curvilinear framework does not affect the outcome of higher-probability CA events, those with probabilities of collision (Pc) greater than 1E-04.[2] For the problem at hand, limiting the analysis to the position portion of the covariance is reasonable.

Second, because a covariance represents an error distribution, its adequacy as a representation can be evaluated only by examining a set of error data and determining whether these data conform to the distribution specified by the covariance. In actuality, this is usually not possible in so straightforward a way because a particular covariance is propagated to a particular point in time and relevant only at that moment, and at that moment there is only one state estimate and thus only one error point. It is not possible to determine in any definitive way whether a single point conforms to a distribution; one can determine a percentile level for that point (*e.g.*, for a normal distribution with mean 0 and standard deviation 1, a value of 3 or greater should occur only 0.13% of the time) and discern that such a point will not occur frequently in such a distribution, but it will occur occasionally, and the situation in the above example may well be one such instance. Typically the way this problem is addressed is to generate a test statistic that compares each error point to its associated covariance, to determine what the distribution of such test statistics should be, and to evaluate the conformity of the set of test statistics to the expected distribution. More will be said about this in the following sections that directly address the calculation of test statistics and evaluation of their statistical properties.

[1] J. Woodburn and S. Tanygin, “Coordinate Effects on the use of Orbit Error Uncertainty,” International Symposium on Space Flight Dynamics 24 (Baltimore, MD), May 2014.

[2] R. Ghrist and D. Plakalovic, Impact of Non-Gaussian Error Volumes on Conjunction Assessment Risk Analysis,” AAS/AIAA Astrodynamics Conference, Minneapolis, MN, August 2012.

Third, a covariance is always a covariance propagated to a particular time. If it is not propagated at all, it is an epoch covariance and is purported to reflect the error of the fit[3]; if it is propagated, it is intended to be a reflection of the state estimate error of that propagated state. Many different covariance propagation methods are available, each of which achieves a different level of fidelity; so one source of covariance irrealism can be the propagation method itself. It is thus important to group realism evaluations by propagation state; for example, all the calculated test statistics for covariances propagated from 1.5 to 2.5 days could be collected into a single pool and evaluated as a group, and the evaluation could be said to be applicable to covariances propagated about two days. It should also be pointed out that covariance irrealism, however it may be evaluated, is unlikely to correlate directly with propagation time in the sense that if the covariance is made realistic for a particular propagation state, it cannot be expected to be realistic for all propagation states. When pursuing covariance realism remediation, one will therefore need to decide whether some sort of omnibus improvement over all common propagation states is desired or whether it is preferable to focus on a particular propagation state. Many aspects of the JSpOC OD process are tuned to optimize performance at 3 days of propagation; this would not be a bad propagation state to choose for O/O covariance realism optimization, although one could also choose something a bit shorter if that were to align more closely with maneuver commitment timelines.

Fourth, in the context of a covariance error calculation, the individual state error values will need to be presumed to constitute independent samples, free of correlation between them. Such an assumption is common in most statistical processes, and it is not surprising to see it arise here. While true sampling independence is unlikely to be achieved, there are certain measures that can be taken to promote it. For example, if one has set up a covariance realism assessment scenario by comparing a predicted ephemeris (with covariance) to a definitive ephemeris, it would be best to take only one point of comparison for each propagation state bin that one wishes to evaluate. In evaluating the two-day propagation state, there would be a temptation to take all of the ephemeris points from, say one hour before the two-day propagation point to one hour after that; if the ephemeris were generated with one-minute time-steps, then this would produce 120 points instead of just one, which on the surface would seem to give a nice sample set for a statistical evaluation. However, because the propagated and definitive ephemeris points are highly correlated, in fact very little additional information is brought to the evaluation by including the entire dataset. Additionally, all of the statistical evaluation techniques that will be deployed to form a realism evaluation assume sample independence; in such a case as the one described above, they will miscarry and produce what is likely to be an overly optimistic result. In the present scenario, it is best to limit oneself to a single comparison point per propagation state from this pair of ephemerides and seek out a group of such ephemeris pairs rather than use closely-grouped data points from them to try to broaden the sample set.

[3] It is questionable whether the *a priori* covariance emerging from the batch process does indeed do this, as it is produced by a calculation that involves only the amount of tracking, times of the tracks, and *a priori* error variances for each observable; it does not give a statistical summary of the actual correction residuals. For this reason, some commentators have argued for an alternative formulation of the covariance that is essentially a description of residual error.

The procedure recommended above eliminates a particular strain of correlation, but it leads naturally to the consideration of a second type: that between successive ephemerides. Definitive ephemerides are generally formed by piecing together sections of ephemeris from essentially the middle of the fit-spans of moving-window batch ODs, and filters process data sequentially but with a forgetting-rate matrix that reduces the influence of older data as a function of data age; but in each case a fair amount of ephemeris time needs to pass before the generated points are statistically independent of a given predecessor set. The independence condition could be fulfilled by forcing the generation of products from entirely different datasets, but this is not a practical procedure. In past projects, trying to reduce the data overlap to less than 50%, meaning that subsequent ephemerides to be used in covariance realism evaluations need to be spaced so that they share less than 50% of the generating data with the previous ephemeris, has been considered acceptable in practice, although a lower figure would be desirable. One need not become overly fastidious about this particular issue, but it is important to take whatever practical steps one can to reduce the influence of sample data correlation on the realism assessment.

Part I: Error Computation

State errors are computed by comparing a predicted state estimate to some sort of truth source. The predicted data typically come from a predicted ephemeris, with a predicted covariance associated with each ephemeris point. Truth data can come from a variety of sources: on-board GPS precision position data, precision telemetry tracking data, or a precision ephemeris constructed from either of these. Ideally, both the predicted and truth data should align temporally (*i.e.,* their time points should be exactly the same), and they should both possess covariance data at each time point; in actuality, it is rare that both of these conditions are met, and often neither of them are.

Covariance data are often not available for the truth data source. Precision tracking or GPS data can sometimes include covariance information as an output from a tracking or combinatorial filter, and when precision ephemerides are constructed, sometimes covariances can be synthesized from the abutment differences observed between the joined “pieces” of ephemeris; but it is frequently the case that even when this is theoretically possible the construction software failed to include this as a feature. If no covariance data are available for the truth source, one must then determine whether the errors in the truth data are so much smaller than the errors in the predicted ephemeris under evaluation that the errors in the truth data can safely be neglected. It is a matter of opinion when this is the case, but most commentators would probably agree that an order of magnitude difference would be an acceptable decrement; and in some cases even smaller ratios could be tolerated. These differences, however, must be considered at the individual component level. If the error in one component tends to dominate the entire arrangement (as is often the case with the in-track component due to drag mis-modeling), then the overall error of each point in the truth ephemeris can be much smaller than the overall error in each point in the predicted ephemeris, yet component errors for the cross-track and radial components could be of a similar magnitude in both truth and prediction. In such a case, the omnibus test statistic to be described in the next section will be distorted because it will treat the normalized errors in each component essentially equally.

If the time-points between truth and predicted data do not align, then some sort of intermediate-value-determination scheme must be used for one of the two datasets. The highest-fidelity approach would be to use an actual numerical propagator to produce aligning values; but if the points for the source to be adjusted are spaced sufficiently closely, then a number of different interpolation approaches produce results quite adequate for a covariance realism assessment. The author has found that a fifth-order Hermite interpolator, using a J2 model for each point’s acceleration terms, has performed quite well in past covariance realism assessment efforts; source code for this interpolation scheme is available from the author if desired. If both data sources (truth and prediction) are interpolatable but only one of the two data sources possesses accompanying covariance data (presumably the predicted source), then it is usually preferable to interpolate the data source that lacks the covariance, as this eliminates the question of how to handle covariance interpolation. If both sources possess covariance, then one would in general interpolate the source with the closer point spacing. If the time spacing is quite close (*e.g.,* one minute), then it is probably not strictly necessary to interpolate the covariance at all;

acceptable performance should be realized by choosing the covariance at either side of the interpolation interval. However, if for completeness or additional accuracy one wishes to interpolate the covariance, Alfano outlines certain techniques and provides test results.[4]

Once position and covariance data for the predicted and truth datasets are aligned at the time-point of interest, calculation of the error is straightforward: it is simply the (subtracted) difference between states, so long as they both be rendered in the same coordinate system. Typically an inertial coordinate system is chosen for this, such as ECI. The subtraction convention does not actually matter given the manner in which the test statistic is computed, but for consistency one can follow the "observed – expected," or truth – prediction, paradigm. The two covariances for each comparison point are simply added together, presuming, of course, that they are in the same coordinate system. Thus for each point of comparison one will have a position difference (with three components) and a corresponding combined covariance.

While one can conduct an entire covariance realism evaluation in the ECI reference frame, it is often more meaningful to move it into the RIC (sometimes called the UVW) frame—a reference frame that is centered on the object itself. If only one of the evaluation data sources has an associated covariance, then it probably makes sense to make that source the center for the RIC coordinate frame. If a covariance is provided for both, then it does not really matter which is selected to ground the RIC frame. The use of the RIC frame is helpful in moving from a finding of irrealism to remediation suggestions.

[4] S.Alfano, "Orbital Covariance Interpolation," 14th AAS/AIAA Space Flight Mechanics Conference, Maui, HI, Feb. 2004.

Part II: Test Statistic Computation

As stated previously, a position covariance, which is the portion of the matrix to be tested for proper error representation, describes a three-dimensional distribution of position errors about the object's nominal estimated state; and the test procedure is to calculate a set of these state errors and determine whether their distribution matches that described the position covariance matrix. To understand the particular test procedure, it is best to consider the problem first in one dimension, perhaps the in-track component of the state estimate error. Given a series of state estimates for a given trajectory and an accompanying truth trajectory, one could calculate a set of in-track error values, here given the designation ε, as the differences between the estimated states and the actual true positions. According to the assumptions previously discussed about error distributions, this group of error values should conform to a Gaussian distribution. As such, one can proceed to make this a "standardized" normal distribution, as is taught in most introductory statistics classes, by subtracting the sample mean and dividing by the sample standard deviation:

$$\frac{\varepsilon-\mu}{\sigma} \quad . \tag{1}$$

This should transform the distribution into a Gaussian distribution with a mean of 0 and a standard deviation of 1, a so-called "z-variable." Since it is presumed from the beginning that the mean of this error distribution is 0, the subtraction as indicated in the numerator of Eq. 1 is unnecessary, simplifying the expression to:

$$\frac{\varepsilon}{\sigma} \quad . \tag{2}$$

It will be recalled that the sum of the squares of *n* standardized Gaussian variables constitutes a chi-squared distribution of *n* degrees of freedom. As such, the square of Eq. 2 should constitute a one-degree-of-freedom chi-squared distribution. This particular approach of testing for normality—evaluating the square of the sum of one or more z-variables—is a convenient approach for the present problem, as all three state components can be evaluated as part of one calculation (ε_u represents the vector of state errors in the radial direction, ε_v the in-track direction, and ε_w the cross-track direction):

$$\frac{\varepsilon_u^2}{\sigma_u^2}+\frac{\varepsilon_v^2}{\sigma_v^2}+\frac{\varepsilon_w^2}{\sigma_w^2}=\chi^2_{3\,dof} \quad . \tag{3}$$

One could calculate the standard deviation of the set of errors in each component and use this value to standardize the variable, but it is the covariance matrix that is providing, for each sample, the expected standard deviation of the distribution; and since the intention here is to test whether this covariance-supplied statistical information is correct, the test statistic should use the variances from the covariance matrix rather than a variance calculated from the actual sample of state estimate errors. For the moment, it is helpful to presume that the errors align themselves such that there is no correlation among the three error components (for any given example it is always possible to find a coordinate alignment where this is true, so the presumption here is not far-fetched; it is merely allowing that that particular coordinate alignment happens to be the *UVW* coordinate frame). In such a situation, the covariance matrix would lack any off-diagonal terms and thus look like the following:

$$C = \begin{bmatrix} \sigma_u^2 & 0 & 0 \\ 0 & \sigma_v^2 & 0 \\ 0 & 0 & \sigma_w^2 \end{bmatrix}, \tag{4}$$

and its inverse would be straightforward:

$$C^{-1} = \begin{bmatrix} 1/\sigma_u^2 & 0 & 0 \\ 0 & 1/\sigma_v^2 & 0 \\ 0 & 0 & 1/\sigma_w^2 \end{bmatrix}. \tag{5}$$

If the errors for a single state are formulated as:

$$\varepsilon = \begin{bmatrix} \varepsilon_u & \varepsilon_v & \varepsilon_w \end{bmatrix}, \tag{6}$$

then the pre-and post-multiplication of the covariance matrix inverse by the vector of errors (shown in Eq. 3-6) will produced the expected chi-squared result:

$$\varepsilon C^{-1} \varepsilon^T = \begin{bmatrix} \varepsilon_u & \varepsilon_v & \varepsilon_w \end{bmatrix} \begin{bmatrix} 1/\sigma_u^2 & 0 & 0 \\ 0 & 1/\sigma_v^2 & 0 \\ 0 & 0 & 1/\sigma_w^2 \end{bmatrix} \begin{bmatrix} \varepsilon_u \\ \varepsilon_v \\ \varepsilon_w \end{bmatrix} = \frac{\varepsilon_u^2}{\sigma_u^2} + \frac{\varepsilon_v^2}{\sigma_v^2} + \frac{\varepsilon_w^2}{\sigma_w^2} = \chi^2_{3dof} \tag{7}$$

What is appealing about this formulation is that, as the covariance becomes more complex and takes on correlation terms, the calculation procedure need not change: the matrix inverse will formulate these terms so as properly to apportion the variances among the *U*, *V*, and *W* directions, and the chi-squared variable can still be computed with the $\varepsilon C^{-1} \varepsilon^T$ formulary. For such a situation in two-dimensions (chosen here for illustrative purposes because the expression is less complex), in which the error quantities are ε_x and ε_y and the correlation coefficient is ρ, the entire equation, with correlation terms included, assumes the form:

$$\varepsilon C^{-1} \varepsilon^T = \frac{1}{(1-\rho)^2} \left(\frac{\varepsilon_x^2}{\sigma_x^2} + \frac{\varepsilon_y^2}{\sigma_y^2} - \frac{2\rho \varepsilon_x \varepsilon_y}{\sigma_x \sigma_y} \right) \tag{8}$$

One can observe that if the correlation coefficient is zero, the equation reduces to the form shown in Eq. 7 above. As the correlation coefficient moves from zero to a more substantial value, the test statistic encounters a trade-off is encountered between the overall inflating effect of the (1-ρ) multiplier and the subtracted correlation term.

Part III: Test Statistic Evaluation

It is very well that a test statistic can be derived that, if the covariance is realistic, should conform to a known statistical distribution; but of course there needs to be some method for testing this dataset to determine if in fact it does conform to the expected distribution. Such a desire leads the investigation to the statistical sub-discipline of “goodness of fit” (GOF).

Every student of college statistics learns about analysis of variance (ANOVA), the particular procedure for determining whether two groups of data can essentially be considered the same or different. More precisely, it is a procedure for determining whether the experimental distribution, produced by the research hypothesis, can be considered to come from the parent distribution represented by the null hypothesis; and the operative statistic arising from the analysis is the *p*-value: the likelihood that the research hypothesis is a sample drawn from the null hypothesis’s parent distribution. If this value becomes small, such as only a few percent, it means that there are only, say, two or three chances in one hundred that the differences between the two samples (null and research) can be explained by sampling error alone, which in this case would be likely to lead to the rejection of the null hypothesis and the embrace of the research hypothesis. This procedure is a specific example of statistical hypothesis testing.

A similar procedure can be applied to evaluate GOF, namely, to evaluate how well a sample distribution corresponds to a hypothesized parent distribution. In this case, the general approach is the reverse of the typical ANOVA situation: it is to posit for the null hypothesis that the sample distribution does indeed conform to the hypothesized parent distribution, with a low *p*-value result counseling the rejection of this hypothesis. This approach does favor the association of the sample and the hypothesized distribution, which is why it is often called “weak-hypothesis testing”; but that is not an unreasonable method: what is being sought is not necessarily the “true” parent distribution but rather an indication of whether it is reasonable to propose the hypothesized distribution as the parent distribution. Such a view is appropriate to the present purpose, namely whether sets of the test statistic described previously can be reasonably ascribed to a 3-DoF chi-squared parent distribution.

There are several different mainstream techniques for goodness-of-fit weak-hypothesis testing: moment-based approaches, chi-squared techniques (not in any way linked to the fact that the present application will be testing for conformity to a chi-squared distribution), regression approaches, and empirical distribution function (EDF) methods. The easiest and most direct of these is simply a test of the first moment of the distribution, which, if normalized by the degrees of freedom of the distribution, should be unity (or, in the present case, should take on an unnormalized value of 3). The square root of this mean, the so-called Mahalanobis distance, is a good estimate of a single-value scale factor describing the covariances departure from reality, and as such it is a convenient way to compare the results from different covariance correction techniques, as well as estimate a single-value scale factor that could potentially be used to remediate an irrealism situation.

While this test is easy to apply and expeditious for generating comparative results, in comparison to other GOF tests it lacks power. Indeed, many different distributions could have the same mean yet be substantially different in the overall behavior or in the tails: matching of the mean is a necessary but not sufficient condition for matching a distribution. To evaluate the match between entire distributions, the EDF methodology is generally considered to be both the most powerful and most fungible to different applications. For this reason, it does not make sense to apply formal GOF tests to first-moment results (*i.e.*, matching of the mean); these should be used merely as a methodology to compare performance for the corrected versus uncorrected cases and intra-correction-methodology performance.

The general EDF approach is to calculate and tabulate the differences between the CDF of the sample distribution and that of the hypothesized distribution, to calculate a GOF statistic from these differences, and to consult a published table of *p*-values for the particular GOF statistic to determine a significance level. Specifically, there are two GOF statistics in use with EDF techniques: supremum statistics, which draw inferences from the greatest deviation between the empirical and idealized CDF (the Kolmogorov-Smirnov statistics are perhaps the best known of these); and quadratic statistics, which involve a summation of a function of the squares of these deviations (the Cramér – von Mises and Anderson-Darling statistics are the most commonly used). It is claimed that the quadratic statistics are the more powerful approach, especially for samples in which outliers are suspected; so it is this set of GOF statistics that were employed for the present analysis. The basic formulation for both the Cramér – von Mises and Anderson-Darling approaches is of the form:

$$Q = n\int_{-\infty}^{\infty}\left[F_n(x) - F(x)\right]^2 \psi(x)dx \ ; \tag{9}$$

the two differ only in the weighting function ψ that is applied. The Cramér – von Mises statistic is the simpler:

$$\psi(x) = 1 \tag{10}$$

setting ψ to unity; the Anderson-Darling is the more complex, prescribing a function that weights data in the tails of the distribution more heavily than those nearer the center:

$$\psi(x) = \left\{F(x)\left[1 - F(x)\right]\right\}^{-1} \tag{11}$$

Given the likelihood that the evaluated datasets will contain outliers, it is appropriate to choose the somewhat more permissive Cramér – von Mises statistic for this investigation.

It is a straightforward exercise to calculate the statistic in Eq. 9, discretized for the actual individual points in the CDF for each trajectory (that is, changing the integral into a summation). This calculates the Cramér – von Mises statistic, from this point on called the “Q-statistic,” as suggested in Eq. 11. The step after this is, for each Q-statistic result, to consult a published table of *p*-values (determined by Monte Carlo studies) for this test to determine the *p*-value associated with each Q-statistic.[5] The usual procedure is to set a *p*-value threshold (*e.g.*, 5%, 2%, 1%) and then to determine whether the sample distribution produces a *p*-value greater than this threshold (counseling the retention of the null hypothesis: sample distribution conforms to hypothesized distribution) or less than this threshold (counseling rejection of the null hypothesis: sample distribution cannot be said to derive from the hypothesized distribution as a parent). One can also interpolate to determine the precise *p*-value for each test situation. MATLAB source code that performs this test and evaluates the results is available from the author.

Finally, it should be noted that, even though there are normalization provisions within the EDF formulation, the results to some degree do depend on the size of the sample. In a way this is considered already by accommodation within the *p*-value tables for sample size; but because the quadratic-sum nature of the test statistic, the procedure can still be overwhelmed by large sample sizes. One approach to mitigating this is to pick a standard sample size—perhaps somewhere in the neighborhood of 50 samples—and calculate the test statistic in a resampled manner, producing a CDF of the *p*-values attained for each sample. As an example, suppose that for a particular evaluation there are 100 error values with associated covariance and therefore 100 test statistic ($\varepsilon C^{-1}\varepsilon^T$) results. When running the GOF test, one might choose 1000 random, 50-point samples from this set of 100 values and test each, producing a CDF chart of the 1000 *p*-values obtained from the resampling investigation.

What is an acceptable level for a *p*-value result—one that would indicate that the error distribution matches that of the covariance? In GOF practice, rarely is a significance level greater than 5% required; and levels of 2% or even 1% are often accepted. It probably can be said that values much less than 1% cannot allow the conclusion that there is any real conformity to the hypothesized distribution. At the same time, it should be added that the calculation is rather sensitive to outliers and that this should be kept in mind when interpreting results.

If results from different remediation approaches match the hypothesized distribution closely enough, then comparison of different *p*-value levels can serve as an indication of the relative performance of these different approaches. However, if the performance is such that the hypothesized distribution is not approached all that closely by any of the correction mechanisms, then a situation can be encountered in which the results fall off of the published tables of *p*-values; and it becomes extremely difficult to compare the results of the different methods conclusively. In such a case, it may be possible to draw some broad comparative results from looking at the Q-values rather than the associated *p*-values; but in all likelihood it will be necessary to revert to simply a comparative results set, such as first-moment tests.

[5] R.B. D'Agostino and M.A. Stephens, *Goodness-of-Fit Techniques*, New York: Marcel Dekker, Inc., 1986.

Additional Consideration: Data Outliers

As mentioned previously, GOF test results are sensitive to outliers; this is true whether one interprets results visually or uses a formal technique. The deleterious impact on test results is mitigated somewhat through the use of the resampling approach outlined above, by which the contribution of the outliers to each individual test is lessened; but the fact remains that bad data do enter the OD process, and the failure to take some account of this can produce situations where the covariance realism assessment problem becomes intractable.

The entire covariance realism assessment process is grounded on the notion that individual component errors are normally distributed, and this situation allows for certain techniques to identify outliers. The usual "x-sigma" filter is a naïve and unscientific method for outlier identification and is especially difficult to justify for the Gaussian distribution, where there is some developed theory for this problem. A superior approach is the Grubbs outlier test, which provides a formal statistical test for outliers but works only in situations with a single outlier and cannot be applied recursively.[6] In the case of multiple outliers, the procedure of Rosner is applicable but must be applied with an *a priori* guess of the potential outliers.[7] That is, one must first inspect the data to assemble a proposed set of outliers and then can test this group as outliers for a particular significance level. Because only a single component can be assessed at a time, it is probably best to design a tool that can examine a particular error point's behavior in all three components (compared to the rest of the main distribution) in order to determine which set of points might constitute an outlier set and test that set. Several such tries may be needed before a set of points can be identified as outliers to a given significance level. Copies of the journal articles outlining the procedure and source code that instantiates part of this is available from the author.

[6] F.E. Grubbs, "Procedures for detecting Outlying Observations in Samples," *Technometrics* 11 (1969), 1-21.

[7] B. Rosner, "On the Detection of Many Outliers," *Technometrics* 17 (1975), 221-227 and "Percentage Points for the RST Many Outlier Procedure," *Technometrics* 19 (1977), 307-312.