Please cite this paper as: J. T. Horwood, J. M. Aristoff, N. Singh, A. B. Poore, and M. D. Hejduk, "Beyond covariance realism: a new metric for uncertainty realism," in Proceedings of the SPIE, Signal and Data Processing of Small Targets, Vol. 9092, Baltimore, MD, May 2014

Beyond Covariance Realism: A New Metric for Uncertainty Realism

Joshua T. Horwood[a], Jeffrey M. Aristoff[a], Navraj Singh[a], Aubrey B. Poore[a], and Matthew D. Hejduk[b]

[a]Numerica Corporation, 4850 Hahns Peak Drive, Suite 200, Loveland CO, 80538

[b]Astrorum Consulting, 10006 Willow Bend Drive, Woodway TX, 76712

ABSTRACT

In the space surveillance tracking domain, it is often necessary to assess not only the covariance consistency or covariance realism of an object's state estimate, but also the realism (proper characterization) of its full estimated probability density function. In other words, there is a need for "uncertainty realism." We propose a new metric (applicable to any tracking domain) that generalizes the covariance realism metric based on the Mahalanobis distance to one that tests uncertainty realism. We then review various goodness-of-fit and distribution matching tests that exploit the uncertainty realism metric and describe how these tests can be applied to assess uncertainty realism in off-line simulations with multiple Monte-Carlo trials or on-line with real data when truth is available.

Keywords: metrics, covariance consistency, covariance realism, uncertainty realism, Mahalanobis distance, non-Gaussian, space surveillance, space situational awareness

1. INTRODUCTION

A point estimator is one that uses sample data to calculate a single value or statistic. Two desirable properties of a point estimator are that it be *unbiased* and *consistent*. The estimator is unbiased if the difference between the expected value and the true value of the parameter being estimated is zero. It is consistent if the expected value of the estimator converges in probability to the true value of the parameter being estimated as the sample size increases. Under Gaussian assumptions, *covariance realism** is the proper characterization of the covariance (statistical uncertainty) in the state of a system. Covariance realism requires that the estimate of the mean be the true mean (i.e., the estimate is unbiased) and the covariance possesses the right size, shape, and orientation (i.e., consistency). Relaxing any Gaussian assumptions, *uncertainty realism* is the proper characterization of the probability density function (statistical uncertainty) of the state of a system. For non-Gaussian probability density functions (PDFs), higher-order cumulants (beyond a state and covariance) are needed to properly characterize the errors. Covariance realism is a necessary but not sufficient condition for achieving uncertainty realism. In the space surveillance tracking domain, the PDF of a space object's orbital state sometimes can become non-Gaussian, especially in applications requiring non-linear uncertainty propagation such as uncorrelated track (UCT) resolution from sparse data with long time gaps. Thus, in such applications, there is a need to represent the uncertainty using a general (non-Gaussian) PDF and to assess more general uncertainty realism.

Metrics for evaluating covariance realism are defined and evaluated by Drummond, Ogle, and Waugh[1] and the references contained therein. Of particular interest is what Drummond *et al.* call the track covariance consistency metric (which is called the covariance realism metric throughout this article). Specifically, the metric assesses, for a particular object, the mean normalized chi-squared statistic of the track (orbit) assigned to that object. This covariance realism metric is defined using the Mahalanobis distance[2] and has several properties that make it well-suited for assessing covariance realism:

Further author information: (Send correspondence to J.T.H.)
J.T.H.: E-mail: joshua.horwood@numerica.us, Telephone: 1 970 461 2000

*The air and missile defense communities commonly use the term *covariance consistency*[1] rather than covariance realism.

1. It is compatible with any choice of coordinate system used to represent uncertainty. Hence, the metric need not be restricted to Cartesian representations of uncertainty.[†]

2. It takes into account each component of the orbital state and each component of the covariance matrix used to represent the (Gaussian) uncertainty.

3. It is dimensionless and non-intrusive (i.e., it does not require any additional knowledge about how the state and covariance are generated).

4. It is computationally tractable and statistically rigorous.

5. It can be used both in off-line simulations with multiple Monte-Carlo trials (including a single run of a simulation) and also on-line with real data assuming a reference or truth trajectory is available.

In the context of space surveillance tracking, the covariance realism (or Mahalanobis distance) metric has been used by several groups.[5–8] A main contribution of this paper is a proposed definition of a new tracking metric (applicable to any tracking domain) that generalizes the covariance realism metric based on the Mahalanobis distance to one that tests the proper characterization of the state PDF (i.e., uncertainty realism). Under weak assumptions, this new *uncertainty realism metric* is also a chi-squared random variable. As such, analogous tests for covariance realism (both in off-line simulations with multiple Monte-Carlo trials and online with real data) can be extended to use the uncertainty realism metric, some of which are reviewed in this paper, such as Pearson's chi-squared goodness-of-fit test[9] and the Cramér-von Mises goodness-of-fit test.[10–12] We acknowledge that the uncertainty realism metric and the corresponding tests for uncertainty realism proposed in this paper only address *aleatoric* uncertainties (which fit naturally in a probabilistic framework) and not *epistemic* uncertainties (which are generally treated by non-probabilistic methods).

The plan of this paper is as follows. In Section 2, we provide the background on the proposed metrics for covariance and uncertainty realism and their use in Monte-Carlo simulations. In particular, we review the classical Mahalanobis distance and its use for testing covariance realism. We then discuss how the Mahalanobis distance can be generalized to non-Gaussian PDFs, thereby defining a metric for uncertainty realism. Tests that use the uncertainty realism metrics in off-line simulations are subsequently proposed, including an "averaged version" of these metrics followed by our recommended (and more powerful) tests that use distribution matching techniques based on Pearson's chi-squared goodness-of-fit test and the Cramér-von Mises criterion. Section 2 also reviews some of the other distribution matching tests as well as tests for normality and why we have chosen not to pursue them. In Section 3, we illustrate an example application of the recommended metrics and statistical tests described in the previous section to assess uncertainty realism in the problem of uncertainty propagation in space surveillance. A more detailed study is carried out in the companion paper.[3] Based on this example and the results in the companion paper, we provide recommendations and make conclusions in Section 4.

2. BACKGROUND

2.1 Mahalanobis Distance

The track (orbit) covariance realism metric evaluates the consistency of the uncertainty corresponding to the orbit state estimate against some known truth state. It is assumed that the orbital state uncertainty is Gaussian so that it can be represented by a single state and covariance matrix. Recall that the random vector $\boldsymbol{x} \in \mathbb{R}^n$ is jointly distributed as a Gaussian distribution if and only if its joint PDF has the form

$$p_x(\boldsymbol{x}; \boldsymbol{\mu}, \mathbf{P}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \mathbf{P}) \equiv \frac{1}{\sqrt{\det(2\pi\mathbf{P})}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\mathbf{P}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]. \quad (1)$$

[†]Even though the companion paper[3] recommends the use of orbital element coordinates for representing and propagating orbital state uncertainty in space surveillance, in some applications, a Cartesian representation of the uncertainty is required. In such instances, Aristoff *et al.*[4] demonstrate how uncertainty in orbital elements can be faithfully (and efficiently) transformed to Cartesian space using Gaussian mixtures.

In this definition, $\boldsymbol{\mu} \in \mathbb{R}^n$ denotes the mean (which also coincides with the mode) and $\mathbf{P}$ is an $n \times n$ symmetric positive-definite matrix called the covariance.

Under these Gaussian assumptions, the covariance realism metric is defined using the Mahalanobis distance. Let $\boldsymbol{x}$ be a given (Gaussian) orbital state estimate at a certain time t and let $\mathbf{P}$ be its corresponding estimated covariance. Further denote $\boldsymbol{x}_{truth}$ as the truth state of the target at time t. The *Mahalanobis distance* between the estimated orbit state and truth target state is defined as

$$\mathcal{M}(\boldsymbol{x}; \boldsymbol{x}_{truth}, \mathbf{P}) = (\boldsymbol{x} - \boldsymbol{x}_{truth})^T \mathbf{P}^{-1} (\boldsymbol{x} - \boldsymbol{x}_{truth}). \tag{2}$$

The expected value of $\mathcal{M}$ is n, where n is the dimension of the state vector $\boldsymbol{x}$. Moreover, it follows that $\mathcal{M} \sim \chi^2(n)$; i.e., $\mathcal{M}$ is chi-squared distributed with n degrees of freedom.

Given a significance level α (typically 0.01 or 0.001), one can derive a two-sided $100(1-\alpha)\%$ confidence interval for the distribution $\chi^2(n)/n$ given by $[\chi^2(n;\alpha/2)/n, \chi^2(n;1-\alpha/2)/n]$, where $\chi^2(n;\beta)$ is the $100\beta\%$ quantile of the distribution $\chi^2(n)$. One rejects the null hypothesis that the covariance $\mathbf{P}$ centered at the estimate $\boldsymbol{x}$ is realistic given the truth $\boldsymbol{x}_{truth}$ if $\mathcal{M}(\boldsymbol{x}; \boldsymbol{x}_{truth}, \mathbf{P})/n$ falls outside this two-sided confidence interval. For example, if $\alpha = 0.001$ and $n = 6$, the two-sided 99.9% confidence interval for this test is $[0.0499, 4.017]$.

We remark that (2) can be modified to accommodate the case in which the truth state $\boldsymbol{x}_{truth}$ is imprecise or "fuzzy." If the uncertainty in $\boldsymbol{x}_{truth}$ is provided as a covariance matrix $\mathbf{P}_{truth}$ then, in place of (2), one can use

$$\mathcal{M}(\boldsymbol{x}; \boldsymbol{x}_{truth}, \mathbf{P}_{truth}, \mathbf{P}) = (\boldsymbol{x} - \boldsymbol{x}_{truth})^T (\mathbf{P} + \mathbf{P}_{truth})^{-1} (\boldsymbol{x} - \boldsymbol{x}_{truth}). \tag{3}$$

2.1.1 Example

The covariance realism metric based on the Mahalanobis distance (2) or (3) has many applications both at the sensor- and system-level. One example of its use for validating covariance realism during uncertainty propagation is the following. First, we make the following basic assumptions: (i) the orbital state uncertainty is identically Gaussian and is realistic[‡] at some initial epoch t_0, and (ii) the truth state $\boldsymbol{x}_{truth}$ is available at some future time $t > t_0$. With these assumptions, we propagate the Gaussian at epoch t_0 to time t, approximate the propagated uncertainty by a Gaussian with state (mean) $\boldsymbol{x}$ and covariance $\mathbf{P}$, and evaluate the metric (2). The application of this metric is only valid if the propagated uncertainty is represented by a Gaussian distribution. A generalization of the covariance realism metric to an uncertainty realism metric that relaxes the assumptions that the initial or propagated uncertainty be Gaussian is described in the Subsection 2.2.

2.2 Generalization of the Mahalanobis Distance

Let $p_x(\boldsymbol{x}; \boldsymbol{\Theta})$ denote a general PDF in the n-dimensional orbit state $\boldsymbol{x}$ with some parameter set $\boldsymbol{\Theta}$. The proposed generalization of the Mahalanobis distance which serves as a metric for *uncertainty realism* is defined by

$$\mathcal{U}(\boldsymbol{x}; \boldsymbol{\Theta}) = -2 \ln \left[\frac{p_x(\boldsymbol{x}; \boldsymbol{\Theta})}{p_x(\hat{\boldsymbol{x}}; \boldsymbol{\Theta})} \right], \tag{4}$$

where $\hat{\boldsymbol{x}}$ is the mode of $\boldsymbol{x}$:

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} \; p_x(\boldsymbol{x}; \boldsymbol{\Theta}).$$

We remark that for a Gaussian PDF, as defined in (1), the parameter set $\boldsymbol{\Theta}$ encapsulates the mean $\boldsymbol{\mu}$ and covariance $\mathbf{P}$. Further, in such a case, the uncertainty realism metric (4) reduces to

$$\mathcal{U}(\boldsymbol{x}; \boldsymbol{\Theta}) = \mathcal{U}(\boldsymbol{x}; \boldsymbol{\mu}, \mathbf{P}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \mathbf{P}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}),$$

which is precisely the Mahalanobis distance (2) defined earlier. For a Gauss von Mises distribution,[3,13,14] it can be shown that $\mathcal{U}(\boldsymbol{x}; \boldsymbol{\Theta})$ is also chi-squared distributed with n degrees of freedom (under certain weak assumptions).

[‡]In this example, we will not be concerned about how this initial Gaussian covariance is generated and if it is indeed realistic. If such a covariance is generated from a batch process (e.g., non-linear least squares, differential correction) using multiple sensor observations, the covariance realism would be highly dependent on the realism of the sensor measurement errors.

Thus, in analogy to the statistical test for covariance realism described at the end of Subsection 2.1, the metric (4) provides a test for uncertainty realism.

The uncertainty realism metric (4) is also chi-squared distributed under the following scenario. Suppose again that $p_x(\boldsymbol{x};\boldsymbol{\Theta}) = \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\mathbf{P})$, and let $\boldsymbol{x} = \boldsymbol{\Phi}(\boldsymbol{y})$ where $\boldsymbol{\Phi}$ is a volume-preserving transformation[§] (i.e., the determinant of its Jacobian is unity). By the change of variables theorem, it follows that $p_y(\boldsymbol{y};\boldsymbol{\Theta}) = \mathcal{N}(\boldsymbol{\Phi}(y);\boldsymbol{\mu},\mathbf{P})$ and, from the definition of the uncertainty realism metric (4), $\mathcal{U}(\boldsymbol{y};\boldsymbol{\Theta}) = \mathcal{M}(\boldsymbol{\Phi}(y);\boldsymbol{\mu},\mathbf{P})$. Thus, quite remarkably, the uncertainty realism metric for any (possibly non-linear) volume-preserving transformation of a Gaussian random vector also possesses the chi-squared property.

The property presented above is significant with regards to uncertainty propagation of a space object's orbital state under two-body dynamics. Since the dynamics are dominated by conservative forces (i.e., gravity), any initial Gaussian distribution that is propagated under said dynamics will have the form $\det(\partial\boldsymbol{\Phi}/\partial\boldsymbol{y})\,\mathcal{N}(\boldsymbol{\Phi}(y);\boldsymbol{\mu},\mathbf{P})$, where $\det(\partial\boldsymbol{\Phi}/\partial\boldsymbol{y}) \approx 1$. This is a consequence of Liouville's theorem[15] in Hamiltonian mechanics. Thus, the uncertainty realism metric (4) applied to this distribution will be (approximately) chi-squared distributed. In practice, we approximate this propagated distribution by some parametric family of distributions such as a multivariate Gaussian. If the propagated distribution cannot be properly characterized by a single Gaussian, we can approximate it using, for example, a Gauss von Mises distribution[3,13,14] or a Gaussian mixture.[16–20] For the latter, if the Gaussian mixture well approximates[¶] the actual distribution $\det(\partial\boldsymbol{\Phi}/\partial\boldsymbol{y})\,\mathcal{N}(\boldsymbol{\Phi}(y);\boldsymbol{\mu},\mathbf{P})$, then one can apply the definition (4) directly to the Gaussian mixture knowing that the resulting test statistic is (approximately) chi-squared distributed.

2.2.1 Statistical Interpretation

The statistical interpretation of the Mahalanobis distance metric (2) and its generalization, given by (4), can be understood by considering the general setting of a multivariate random vector $\boldsymbol{x}$ with support on a differentiable manifold $\mathfrak{M}$. We express its PDF in the form $p_x(\boldsymbol{x};\boldsymbol{\Theta}) = \mathrm{e}^{-f(\boldsymbol{x};\boldsymbol{\Theta})/2} \Leftrightarrow f(\boldsymbol{x};\boldsymbol{\Theta}) = -2\ln p_x(\boldsymbol{x};\boldsymbol{\Theta})$. Suppose now that a point $\boldsymbol{x}_* \in \mathfrak{M}$ is given and one wishes to test the null hypothesis H_0 that $\boldsymbol{x}_*$ is not a statistically significant realization of the random vector $\boldsymbol{x}$ (i.e., $\boldsymbol{x}_*$ is a representative draw from $\boldsymbol{x}$). The p-value for a one-sided test is

$$p = \Pr[\boldsymbol{x} \in \Omega_*] = \int_{\Omega_*} \mathrm{e}^{-f(\boldsymbol{x};\boldsymbol{\Theta})/2}\,\mathrm{d}\boldsymbol{x}, \tag{5}$$

where $\Omega_* = \{\boldsymbol{x} \mid f(\boldsymbol{x};\boldsymbol{\Theta}) > f(\boldsymbol{x}_*;\boldsymbol{\Theta}) \equiv C\}$. Smaller p-values imply that the realization $\boldsymbol{x}_*$ lies farther out on the tails of the PDF (see Figure 1(a)). The null hypothesis H_0 is rejected at the significance level α (typically 0.01 or 0.001) if $p < \alpha$. Figure 1(b) shows the setup for the analogous two-sided hypothesis test. For a given significance level α, one determines the contours C_L and C_U such that

$$\int_{\Omega_L} \mathrm{e}^{-f(\boldsymbol{x};\boldsymbol{\Theta})/2}\,\mathrm{d}\boldsymbol{x} = \int_{\Omega_U} \mathrm{e}^{-f(\boldsymbol{x};\boldsymbol{\Theta})/2}\,\mathrm{d}\boldsymbol{x} = \frac{1}{2}\alpha,$$

where $\Omega_L = \{\boldsymbol{x} | f(\boldsymbol{x};\boldsymbol{\Theta}) < C_L\}$ and $\Omega_U = \{\boldsymbol{x} | f(\boldsymbol{x};\boldsymbol{\Theta}) > C_U\}$. (Note that the yellow shaded region in Figure 1(b) has probability α.) A two-sided test with significance level α rejects the null hypothesis H_0 if $f(\boldsymbol{x}_*;\boldsymbol{\Theta}) < C_L$ or $f(\boldsymbol{x}_*;\boldsymbol{\Theta}) > C_U$.

2.2.2 Examples

Suppose $p_x(\boldsymbol{x};\boldsymbol{\Theta}) = \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\mathbf{P})$ (i.e., $\boldsymbol{x}$ is Gaussian) and a realization $\boldsymbol{x}_* \in \mathbb{R}^n$ is given, then

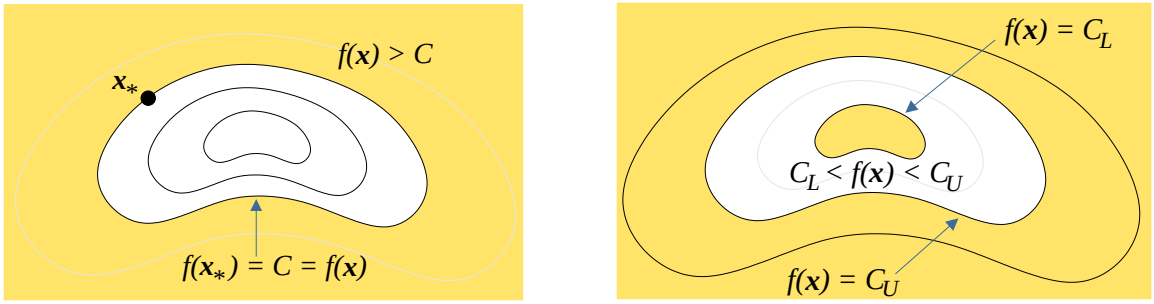
$$f(\boldsymbol{x};\boldsymbol{\Theta}) = -2\ln\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\mathbf{P}) = \mathcal{M}(\boldsymbol{x};\boldsymbol{\mu},\mathbf{P}) + \ln\det(2\pi\mathbf{P}),$$

where $\mathcal{M}$ is the Mahalanobis distance (2). The integration region Ω_* in (5) is

$$\Omega_* = \{\boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x};\boldsymbol{\Theta}) > f(\boldsymbol{x}_*;\boldsymbol{\Theta})\} = \{\boldsymbol{x} \in \mathbb{R}^n \mid \mathcal{M}(\boldsymbol{x};\boldsymbol{\mu},\mathbf{P}) > \mathcal{M}(\boldsymbol{x}_*;\boldsymbol{\mu},\mathbf{P})\}.$$

[§]One example of a volume-preserving transformation is the solution flow of a *conservative* dynamical system that propagates some initial state x_0 at time t_0 to a state $\boldsymbol{x}$ at time t.

[¶]In principle, any PDF can be approximated by a Gaussian mixture to within any desired accuracy (in the L^1 sense) due to a result of Alspach and Sorenson.[21]



(a) One-sided test (b) Two-sided test

Figure 1. Setup for the statistical significance tests. The yellow shaded regions are those regions in which the null hypothesis is rejected. Dependence on the parameter set $\boldsymbol{\Theta}$ in $f(\boldsymbol{x};\boldsymbol{\Theta})$ is omitted in the figure.

Substituting this information into (5) yields

$$p = \Pr[\boldsymbol{x} \in \Omega_*] = \Pr[\mathcal{M}(\boldsymbol{x};\boldsymbol{\mu},\mathbf{P}) > \mathcal{M}(\boldsymbol{x}_*;\boldsymbol{\mu},\mathbf{P})] = \Pr[\chi^2(n) > \mathcal{M}(\boldsymbol{x}_*;\boldsymbol{\mu},\mathbf{P})].$$

Thus, the resulting p-value can be computed by evaluating the complementary cumulative distribution function (i.e., tail distribution) of $\chi^2(n)$ at the Mahalanobis distance (2) evaluated at the realization $\boldsymbol{x}_*$.

The uncertainty realism metric (4) can also be applied in analogy to the example discussed in Subsection 2.1 on uncertainty propagation. We make the analogous basic assumptions that (i) the orbital state uncertainty is represented by some general PDF (that need not be Gaussian) and is realistic at some initial epoch t_0, and (ii) the truth state $\boldsymbol{x}_{truth}$ is available at some future time $t > t_0$. With these assumptions, we propagate the initial PDF at epoch t_0 to time t yielding the PDF $p_x(\boldsymbol{x};\boldsymbol{\Theta})$, and evaluate the uncertainty realism metric $\mathcal{U}(\boldsymbol{x}_{truth};\boldsymbol{\Theta})$ from the definition (4). Based on some confidence interval of the underlying chi-squared distribution, we might reject the null hypothesis that the propagated PDF "captures the truth"; i.e., $\boldsymbol{x}_{truth}$ is a representative draw from $p_x(\boldsymbol{x};\boldsymbol{\Theta})$.

2.2.3 Important Remarks

It is important to understand that covariance realism does not always imply uncertainty realism. Said in another way, if one has properly characterized the first two cumulants (i.e., mean and covariance) of the state PDF, then it does not necessarily imply that those two cumulants alone are sufficient for characterizing the full PDF. Consequently, while our proposed tests for uncertainty realism can detect a breakdown in uncertainty realism, such a breakdown does not say anything about the realism of the covariance[‖]. This is a moot point; one needs to strive for uncertainty realism. Henceforth, we will use the terms "uncertainty realism metric" and "uncertainty realism test" exclusively.

2.3 Averaged Uncertainty Realism Metric

As an uncertainty realism metric, one can consider the value of $\mathcal{U}$ (or $\mathcal{M}$ in the Gaussian case), averaged over all orbits or at each time instance (averaged over one or many Monte-Carlo trials, or with real data), together with upper and lower bounds for a particular confidence interval. Specifically, let $\mathcal{U}^{(i)}$ be the uncertainty realism metric (4) computed in the i-th Monte-Carlo trial. Let k be the total number of independent trials. Then,

$$\bar{\mathcal{U}} \equiv \frac{1}{nk}\sum_{i=1}^{k}\mathcal{U}^{(i)} \sim \frac{1}{nk}\chi^2(nk). \tag{6}$$

One can test the null hypothesis that the (normalized) average of the samples $\mathcal{U}^{(i)}$, given by $\bar{\mathcal{U}}$, is consistent with the mean of the distribution $\chi^2(nk)/(nk)$. For example, if $n = 6$ and $k = 100$, a two-sided 99.9% confidence

[‖] If one is only interested in testing covariance realism without regard to uncertainty realism, some options are provided by Vallado and Seago.[22]

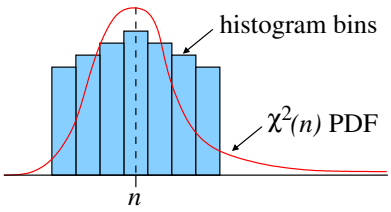


Figure 2. Depiction of a distribution matching test. Shown is a histogram of Monte-Carlo trials with the expected $\chi^2(n)$ matching distribution. In this example, the sample mean (e.g., as would be computed from the uncertainty realism test described in Subsection 2.3) is consistent with the expected value of the matching distribution; the sample variance (and higher-order moments) would not be consistent. Thus, a distribution matching test would correctly reject the null hypothesis that the samples come from a $\chi^2(n)$ distribution; a first-order moment (mean) matching test would not.

interval for the distribution (6) is $[0.8209, 1.2010]$. Note that this is a necessary (but not sufficient) test of uncertainty realism, and that $\mathbb{E}[\bar{\mathcal{U}}] = 1$. Moreover, $\mathrm{Var}[\bar{\mathcal{U}}] \to 0$, as $k \to \infty$ (i.e., the confidence interval becomes infinitesimally small). Confidence intervals for other significance levels and other common values of n and k are provided in Table 1 of the appendix.

As motivated in Section 2.4, a stronger test for uncertainty realism (in an off-line setting with multiple Monte-Carlo trials) would be to consider the distribution of the Monte-Carlo samples $\mathcal{U}^{(i)}$ and perform a distribution matching or goodness-of-fit test.[9] This test would indirectly consider the consistency of the sample with both the mean and higher-order cumulants of the matching chi-squared distribution. The example below highlights some of the differences between the averaged uncertainty realism metric and metrics based on distribution matching and motivates the need for the latter more powerful tests.

2.3.1 Example

An off-line simulation with multiple Monte-Carlo trials that applies the metric (6) for assessing uncertainty realism during uncertainty propagation is the following. First, we make the basic assumption that the orbital state uncertainty is identically Gaussian and is realistic at some initial epoch t_0. With this assumption, we perform the following operations:

1. Propagate the Gaussian at epoch t_0 to time t and approximate the propagated uncertainty by a Gaussian with state (mean) $\boldsymbol{\mu}$ and covariance $\mathbf{P}$.
2. Sample random particle states[**] $\boldsymbol{x}^{(i)}(t_0)$, $i = 1, \ldots, k$, from the initial Gaussian distribution at epoch t_0.
3. Propagate each particle state $\boldsymbol{x}^{(i)}(t_0)$ from time t_0 to time t yielding a propagated particle state $\boldsymbol{x}^{(i)}(t)$.
4. For $i = 1, \ldots, k$, compute $\mathcal{U}^{(i)} = \mathcal{M}\big(\boldsymbol{x}^{(i)}(t); \boldsymbol{\mu}, \mathbf{P}\big)$.
5. Compute the averaged uncertainty realism metric $\bar{\mathcal{U}}$ from (6) in conjunction with the metrics $\mathcal{U}^{(i)}$ computed in the previous step[††].
6. Perform the hypothesis test described in the text following Equation (6).

Figure 2 illustrates the differences between the conclusions one might make when applying the averaged uncertainty realism metric described here and a distribution matching test. If one were to plot a histogram of the Monte-Carlo samples $\mathcal{U}^{(i)}$, it could look like the one shown in the figure. In this example, computing

[**] A random draw $\boldsymbol{x}$ from a multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance $\mathbf{P}$ can be obtained as follows. Let $\boldsymbol{z}$ be a vector where each component is an independent random draw from the standardized Gaussian (i.e., the Gaussian with mean 0 and variance 1); this functionality is provided in most programming languages and scientific and statistical software. Then, the required $\boldsymbol{x}$ is the vector $\boldsymbol{\mu} + \mathbf{A}\boldsymbol{z}$, where $\mathbf{A}$ is the lower-triangular Cholesky factor[23] of the covariance $\mathbf{P}$ such that $\mathbf{P} = \mathbf{A}\mathbf{A}^T$.

[††] Alternatively or in addition to Steps 5 and 6, one can compute the normalized Pearson test statistic or the Cramér-von Mises test statistic from the $\mathcal{U}^{(i)}$, and perform the corresponding hypothesis tests described in Subsection 2.4.

the average of the $\mathcal{U}^{(i)}$ would yield a value of n, in agreement with the expectation of the $\chi^2(n)$ distribution. Consequently, one would be tempted to assert that one has uncertainty realism since the test statistic (i.e., $\bar{\mathcal{U}}$) would lie in the center of any confidence interval. Indeed, consistency of the sample mean with the expected value of the matching distribution is only a necessary condition for uncertainty realism. Clearly, as the figure shows, the variance (as well as the higher-order moments) of the $\mathcal{U}^{(i)}$ do not match those of the target chi-squared distribution; the histogram is a poor fit. Therefore, a distribution matching test, such as one based on Pearson's test or the Cramér-von Mises test described in the next subsection, is a better test for uncertainty realism in this case.

2.4 Distribution Matching Tests

It is helpful to keep in mind that the purpose of the multivariate PDF is to represent the statistical distribution of the actual state errors. Durable testing of the adequacy of this PDF, therefore, must be an exercise in distribution matching. Said another way, it must be a determination of whether the statistical distribution of a set of state errors, usually calculated by the comparison of state estimates to an externally-determined precision reference orbit, matches to a reasonable degree the distribution represented by the hypothesized multivariate PDF. Testing approaches that do less than this, such as distribution-matching of only a single component of the state error or the testing only for a matching of distribution mean values (as in the averaged uncertainty realism metric described in Subsection 2.3), render results that can be of some use; but they will not allow a definitive assessment of the realism of the uncertainty. It is thus important to discuss fully-formed methods of uncertainty realism and the particular virtues that they possess so that the advantages and drawbacks of more abbreviated methods can be thrown into relief.

Most multivariate distribution matching approaches proceed by calculating test statistics (e.g., a Mahalanobis distance) that can then be evaluated for conformity to a canonical univariate distribution (e.g., a χ^2 distribution) and thus employ standard goodness-of-fit (GOF) techniques. GOF approaches comprise both specialty tests for a specific distribution type (a number of these, for example, exist for the Gaussian distribution) and more general tests that are capable of evaluating conformity to a number of different distributions. Because covariance realism evaluation usually requires the evaluation of conformity to both the Gaussian and the chi-squared distribution, and because uncertainty realism evaluation can involve testing for conformity to a variety of different distributions (including potentially distributions that have no analytic representation), one should focus on these more fungible techniques, principal among them the traditional chi-squared distribution matching test and the family of empirical distribution function (EDF) tests. An example of the former, Pearson's chi-squared GOF test, is discussed in Subsection 2.4.1, while examples of the latter, including the Kolmogorov-Smirnov, Cramér-von Mises, and Anderson-Darling tests, are described in Subsection 2.4.2. Other possible GOF and EDF tests and our reasons for not using them to assess covariance and uncertainty realism in space surveillance applications are briefly discussed in Subsection 2.5.

2.4.1 Pearson's Chi-Squared Goodness-of-Fit Test

Pearson's chi-squared distribution matching test can be considered the most ecumenical in that it can be deployed as a test of conformity to any distribution and can be used for discrete as well as continuous distributions. Further, as a staple of most introductory statistics courses, it has the additional advantage of familiarity to most scientists and engineers. When testing a random sample to see if it can be considered to belong to a hypothesized parent distribution, the basic procedure is to (i) divide up the possible range of values of the random sample into a set of m discrete cells; (ii) determine the number of values that actually fall into each cell and the number that should have fallen into each under the assumed parent distribution; and (iii) compute the chi-squared test statistic (defined below) and compare it to a critical value from the χ^2 distribution. It what follows, we describe these steps in more detail.

Specifically, let $x^{(i)}$, $i = 1, \ldots, k$, denote the observed sample trials. We wish to test the null hypothesis that these observations belong to a particular matching distribution. Suppose the observations $x^{(i)}$ are grouped into m cells or bins where o_j is the number of observations contained in the j-th bin. The *normalized Pearson test statistic* is defined by

$$P_\chi = \frac{1}{m-1-p} \sum_{j=1}^{m} \frac{(o_j - e_j)^2}{e_j}, \tag{7}$$

where e_j is the expected number of observations contained in j-th bin as determined from the definition of the j-th bin and the properties of the matching distribution. The normalized Pearson test statistic P_χ in (7) asymptotically approaches the distribution $\chi^2(m-1-p)/(m-1-p)$ where the integer p is the number of co-variates used in fitting the matching distribution. For example, if the matching distribution is fixed (e.g., to a $\chi^2(6)$ distribution), then $p = 0$. The approximation of P_χ by a chi-squared distribution breaks down if the expected frequencies e_j are too low. One way to ensure that the e_j are sufficiently large (so that the asymptotic property of the Pearson statistic holds) is to choose m "equiprobable" bins so that $e_j = k/m$ for all j. In other words, the bins are of variable width all having the same expected probabilities. One prescription for choosing the number of bins is

$$m = \max(5, \min(100, 0.01k)).$$

We now specialize the framework of this goodness-of-fit test to the case when the observed samples $x^{(i)}$ are $\mathcal{U}^{(i)}$ corresponding to the uncertainty realism metric (4) of the i-th trial. (Recall that Subsection 2.3.1 provides an example of how the $x^{(i)}$ are generated in an uncertainty propagation scenario.) In both cases, the target matching distribution is $\chi^2(n)$, where n is the dimension of the orbital state space. For sake of example, we take $n = 6$. We define the m bins so that the expected number of observations e_j is k/m for all j. Let $b^{(i)} \in \{1, \ldots, m\}$ denote the bin number in which each observation $x^{(i)}$ belongs. By virtue of the choice of the expected frequencies e_j, these bin numbers can be determined from the cumulative distribution function (CDF) of the matching distribution. For a $\chi^2(6)$ matching distribution, its CDF enjoys a particularly simple form:

$$F(x) = \begin{cases} 1 - \frac{1}{8}\mathrm{e}^{-x/2}(x^2 + 4x + 8), & x > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

It follows that

$$b^{(i)} = \lceil mF(x^{(i)}) \rceil,$$

for $i = 1, \ldots, k$. The number of observations contained in each bin can be readily determined as follows: (i) initialize $o_j = 0$, for $j = 1, \ldots, m$, and (ii) for $i = 1, \ldots, k$, do $o_{b^{(i)}} \leftarrow o_{b^{(i)}} + 1$. Finally, we compute the normalized Pearson test statistic:

$$P_\chi = \frac{1}{m-1} \sum_{j=1}^{m} \frac{(o_j - e_j)^2}{e_j} = \frac{1}{m-1} \sum_{j=1}^{m} \frac{(o_j - k/m)^2}{k/m}.$$

Given a significance level α (typically 0.01 or 0.001), we can derive a one-sided $100(1-\alpha)\%$ confidence interval for the distribution $\chi^2(m-1)/(m-1)$ given by $[0, \chi^2(m-1; 1-\alpha)/(m-1)]$, where $\chi^2(m-1; \beta)$ is the $100\beta\%$ quantile of the distribution $\chi^2(m-1)$. One rejects the null hypothesis that the observed samples $x^{(i)}$ belong to a $\chi^2(6)$ distribution if the normalized Pearson test statistic P_χ computed above falls outside this confidence interval.

2.4.2 Empirical Distribution Function (EDF) Tests

The simplicity and flexibility of the Pearson GOF test make it quite appealing, but it unfortunately harbors considerable disadvantages as well. The first is a relative lack of power when used to test against continuous distributions. Discretizing a continuous distribution by dividing it up into individual cells eliminates some of the distribution's information; techniques that can preserve continuity will generally render more powerful results. The second is an inherent arbitrariness introduced by the number of cells selected. Certain studies have shown that the test is more powerful when performed with cells sized so as to be equiprobable (as detailed in Subsection 2.4.1), and cell quantities optimized for the use of the test against a hypothesized Gaussian distribution can be recommended.[24] However, there is no set of such recommendations for the general case, with both the test outcomes and the resulting statistical inference affected by different bin quantities and sizes. Given these problems, it is typically best to select other GOF approaches unless one is performing tests against a hypothesized distribution that is natively discrete.

The leading alternative candidate for this type of GOF testing is the family of tests based on the *empirical distribution function (EDF)* of the sample data and the hypothesized distribution. The basic idea here is to

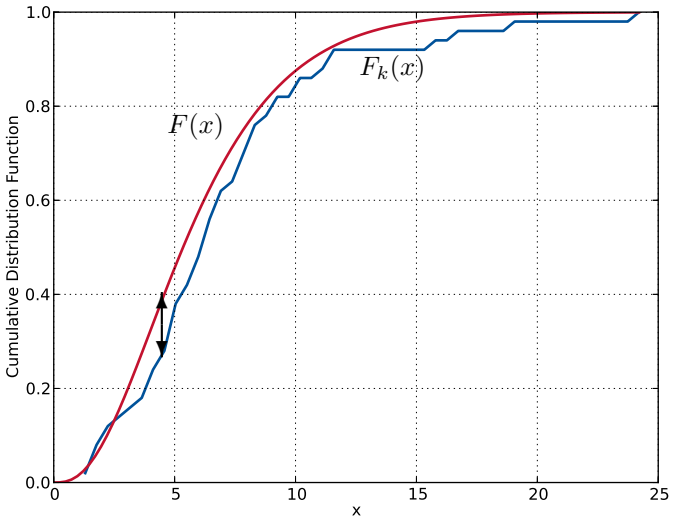


Figure 3. Depiction of the Kolmogorov-Smirnov test. The test statistic, up to a normalization in the sample size k, is the largest deviation between the hypothesized CDF $F(x)$ and the empirical CDF $F_k(x)$, as indicated by the black arrow.

(i) calculate the differences between a cumulative distribution function (CDF) for the hypothesized distribution and an empirical CDF (ECDF) for the sample distribution; (ii) encapsulate these differences in terms of a GOF test statistic; and (iii) determine the likelihood, by the use of published p-value tables, that the sample actually could have the hypothesized distribution as a parent. Constructing the CDF for the hypothesized distribution and the ECDF for the sample distribution is straightforward enough; plotting them on the same graph can reveal the differences visually. At this point, there are two different approaches to determining the overall amount of deviation. The supremum approach is to catalog the largest deviation between the hypothesized and empirical result and examine the p-value associated with this large a deviation; this approach is the basis of the *Kolmogorov-Smirnov* test.

In Figure 3, the hypothesized distribution's CDF $F(x)$ is given in red and the ECDF $F_k(x)$ for the sample is shown in blue. The x-axis is the actual function value, and the y-axis is the cumulative probability. The largest deviation between the hypothesized and empirical CDF is indicated by the arrow. In applying the Kolmogorov-Smirnov test, one computes this deviation and adjusts it for the sample size k:

$$D_k = \sqrt{k} \sup_x |F_k(x) - F(x)|.$$

One then compares the Kolmogorov-Smirnov test statistic D_k to a table of p-values in order to determine the likelihood that the sample set could have originated from the hypothesized distribution. The formulation of the hypothesis test here is somewhat unusual in that the null hypothesis is that the two distributions (hypothesized and sample) are in fact the same, and this hypothesis is rejected for low p-values. This approach is called "weak-hypothesis testing" and is occasionally criticized for being too permissive. Note that it must be remembered that the purpose here is not to identify the true underlying distribution, only to determine whether the hypothesized distribution is a reasonable candidate for a parent distribution of the sample. P-values from 0.1–1% are typically used as the rejection threshold in GOF testing. It is the authors' experience that at this significance level, mismatched distributions are clearly rejected, even with small sample sizes.

The second strain of EDF GOF test approaches are the *quadratic statistics*. These approaches sum up all of the deviations and use this sum as the test statistic. The canonical summation equation is the following:

$$Q_k = k \int_{-\infty}^{\infty} \left[F_k(x) - F(x)\right]^2 \psi(F(x))\, \mathrm{d}F(x). \tag{9}$$

The weighting factor $\psi(F(x))$ is typically either set to unity to produce the *Cramér-von Mises statistic* or to a function that will give more weight to the tails, such as $\psi(F(x)) = 1/[F(x)(1-F(x))]$, to produce the *Anderson-Darling statistic*. Testing proceeds in the same way as that described for the Kolgomorov-Smirnov test. The test statistic Q_k is calculated, and tables of p-values are consulted to determine the significance level for the test statistic indicated. A good source for these tables is the monograph of D'Agostino and Stephens[12] on this subject. Tables exist for all of the major distributions (e.g., normal, gamma, chi-squared, von Mises), both as fully-specified distributions ("Case 0") and as distributions in which distribution parameters are represented by estimators ("Cases 1-3").

With regards to assessing uncertainty realism using an EDF test in the context of the example of Subsection 2.3.1, we recommend using a GOF test based on a quadratic statistic rather than the Kolmogorov-Smirnov test. The latter is generally considered less powerful than quadratic tests because it considers only one value (i.e., the largest deviation). Of the two quadratic tests reviewed here, we recommend the Cramér-von Mises test over the Anderson-Darling test because the latter, though usually considered more powerful, is more fragile due to sensitivity of the test on the tails. Specializing (9) to $\psi(F(x)) = 1$, the resulting Cramér-von Mises test statistic is[‡‡]

$$Q_k = \frac{1}{12k} + \sum_{i=1}^{k} \left[\frac{2i-1}{2k} - F(x^{(i)}) \right]^2, \tag{10}$$

where the $x^{(i)}$, $i = 1, \ldots, k$, are the observed samples in increasing order. Table 2 in the appendix provides one-sided confidence intervals for the Cramér-von Mises test statistic (10) for common significance levels and sample sizes k.

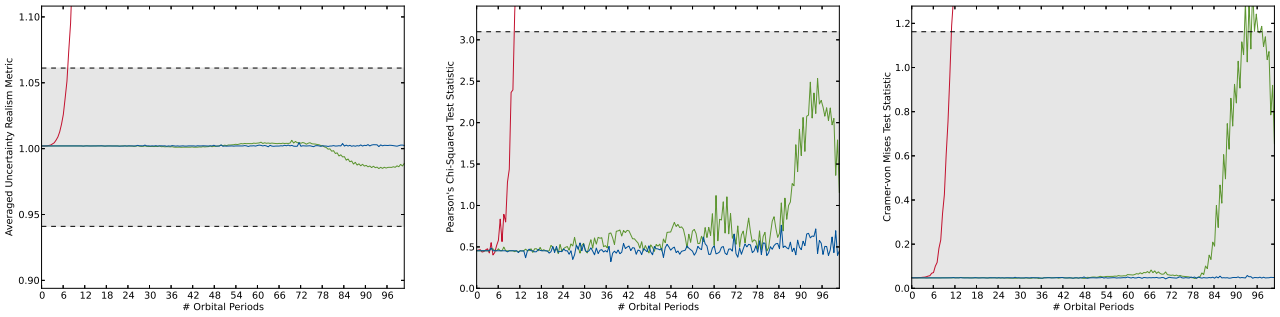
2.5 Other Distribution Matching Tests

There are a number of additional GOF testing approaches that are common in modern engineering practice that were not pursued here, for reasons of both suitability and convenience. The Akaike Information Criterion[26] is an entropy-based approach that has achieved currency for model adequacy assessment. While in some ways it is more powerful than traditional hypothesis testing, its chief drawback is that it can serve only as a comparative test to evaluate the relative performance of two or more models; it cannot give an absolute evaluation, in a p-value sense, of the conformity of the data to any single hypothesized distribution. Regression/correlation GOF tests, such as the well-known Shapiro-Wilk test,[27] and moment-based tests, such as the combination third-fourth moment test,[28] are very much legitimate GOF tests that often can confer substantial power. However, the ability to deploy them requires certain *a priori* products: for regression tests, it is a set of data weighting coefficients appropriate to the hypothesized distribution; and for moment-based tests, it is the null distributions of those moments for the hypothesized distribution. The present authors were not able to locate published sources for these *a priori* products for the chi-squared distribution, which is one of the principal hypothesized distributions to be tested for the present application. While it may be possible to establish these *a priori* products through private Monte-Carlo studies, this was seen as unnecessary labor when other GOF tests – equally powerful – already exist with the deployment products necessary for the testing of all of the hypothesized distributions presently considered. For this reason, the present collection of GOF test was limited to the traditional Pearson test and the mainstream EDF tests (Kolmogorov-Smirnov, Cramér-von Mises, and Anderson-Darling).

3. EXAMPLES

We now provide an example that tests uncertainty realism in the context of the uncertainty propagation scenario described in Section 2.3.1 and examines the power and effectiveness of the averaged uncertainty realism test, the Pearson goodness-of-fit (GOF) test, and the Cramér-von Mises test. With additional details in the companion paper,[3] the high-level setup for these tests is as follows. The initial orbital state at epoch describes a high accuracy low Earth orbit (LEO) object; its uncertainty is taken to be Gaussian in the osculating equinoctial orbital element coordinate system. The LEO object state and covariance are propagated using the prediction step of the unscented Kalman filter[29] (UKF); individual sigma points in the UKF are propagated using an implicit Runge-Kutta-based method[30,31] in conjunction with a 32×32 gravity model and lunar-solar perturbations. A

[‡‡]A detailed explanation of the derivation of (10) from (9) can be found in the monograph of Darling.[25]



(a) Averaged uncertainty realism test (b) Pearson goodness-of-fit test (c) Cramér-von Mises test

Figure 4. Application of the (a) averaged uncertainty realism test, (b) Pearson goodness-of-fit test, and (c) Cramér-von Mises test to an uncertainty propagation scenario in space surveillance. The coordinate system in which the state and covariance are represented is Cartesian position-velocity coordinates **(red)**, osculating equinoctial orbital elements **(green)**, or J_2 equinoctial orbital elements **(blue)**. The 99.9% confidence region is shown in the shaded area. The less-powerful tests in (a) and (b) suggest that uncertainty realism does not break down in osculating element space **(green)** over the entire duration of propagation, whereas the more-powerful Cramér-von Mises test (c) indicates a breakdown after about 90 orbital periods of propagation.

total of 1000 Monte-Carlo trials are used; 10 (equiprobable) bins are defined when applying the Pearson test. Three variations of the scenario are considered which differ only by the choice of coordinate system used to the represent the state and covariance. These coordinate systems are (i) osculating equinoctial orbital elements, (ii) Cartesian position-velocity coordinates, and (iii) J_2 equinoctial orbital elements. For scenarios that use the latter two coordinate systems, the initial state and covariance (defined in the osculating elements as described above) are converted to (and approximated by) a Gaussian in the target coordinate system using the unscented transform*.

Figure 4 presents the results of the uncertainty realism tests outlined above. All tests show that uncertainty realism breaks down the fastest when the uncertainty is represented in Cartesian coordinates (see the red curves). This is not surprising since the non-linearity in the dynamics is the greatest in these coordinates, and non-linearity implies that the initial Gaussian density quickly becomes non-Gaussian. In contrast, the osculating equinoctial orbital elements absorb the most dominant term in the non-linear dynamics (i.e., the $1/r^2$ term in the gravity) while, in addition, the J_2 equinoctial elements absorb the J_2 perturbation in the gravity.[3] As such, representing uncertainty in one of these orbital element systems mitigates the departure from "Gaussianity," "extends the life" of the UKF, and preserves uncertainty realism longer. The results inferred from the averaged uncertainty realism metric in Figure 4(a) are deceptive for the osculating equinoctial case (green curve) and suggest that uncertainty realism is maintained over the entire duration of propagation (since the test statistic fails to pierce the 99.9% confidence interval). The application of the Pearson GOF test in Figure 4(b), a more powerful test for uncertainty realism, suggests that a breakdown is imminent (in the osculating element case), though the Pearson test statistic is still within the 99.9% confidence interval. On the other hand, the Cramér-von Mises test statistic, plotted in Figure 4(c), confirms a potential breakdown in uncertainty realism after about 90 orbital periods. By using the J_2 variant of the equinoctial elements to represent uncertainty (blue curve), the Cramér-von Mises test statistic is nearly constant over the entire duration of propagation which strongly suggests that the uncertainty remains Gaussian in such coordinates and its uncertainty (and covariance) realism does not degrade. We also remark that the Cramér-von Mises test tends to produce a smoother (less noisy) test statistic when comparing panels (b) and (c) of Figure 4 due in part to the test's power, robustness, and the fact that it is "bin agnostic." Clearly, this example shows that first-moment tests such as the averaged uncertainty realism test are not recommended since they have very little determinative power. Pearson's test statistic is an improvement, but it is not to be preferred over the more fungible and powerful Cramér-von Mises test.

*The initial state uncertainty is well-approximated by a single Gaussian in all three coordinate systems, as evidenced in Figure 4, since the averaged uncertainty realism metric, Pearson test statistic, and Cramér-von Mises test statistic are (approximately) equal at the initial epoch and well within their respective 99.9% confidence intervals.

In the companion paper, we perform a more thorough study of the test described above using a broader range of initial conditions. Additional methods for non-linear filtering are also evaluated, namely, the prediction steps of the extended Kalman and Gauss von Mises filters,[3,13,14] as are different coordinate systems for representing uncertainty.

4. RECOMMENDATIONS AND CONCLUSIONS

This paper has defined a series of metrics and statistical tests to assess covariance realism (and more general uncertainty realism) during propagation of space object state uncertainty and to enable quantitative comparisons between different approaches. Chief among the metrics is a new metric that generalizes the traditional Mahalanobis distance used in covariance realism tests to non-Gaussian distributions and possesses an analogous chi-squared property of the resulting test statistic.

In scenarios involving real data with truth on a single object (i.e., a high accuracy orbit) or in off-line simulations with only a few Monte-Carlo trials (between 1 and 10) in which there are insufficient number of samples to perform a more powerful goodness-of-fit (GOF) or distribution matching test, we recommend using the averaged uncertainty realism metric. The use of such a metric provides a necessary condition for uncertainty realism based on the chi-squared property of the computed test statistic.

In scenarios involving real data with truth on multiple objects or in off-line simulations amenable to a large number of Monte-Carlo trials (at least 10), the use of first-moment tests (such as the one based on the averaged uncertainty realism metric) are not recommended due to their limited determinative power. In fact, the application of such tests can lead to deceptive results as demonstrated in Section 3. Instead, the Cramér-von Mises test is recommended over the Pearson GOF test or the related Anderson-Darling test because the former tends to be more robust and less sensitive to observations in the tails of the distribution and the number of samples used in the test.

ACKNOWLEDGMENTS

This work was funded, in part, by a grant and a Phase II STTR from the Air Force Office of Scientific Research (FA9550-11-1-0248, FA9550-12-C-0034).

REFERENCES

[1] Drummond, O. E., Ogle, T. L., and Waugh, S., "Metrics for evaluating track covariance consistency," in [*SPIE Proceedings: Signal and Data Processing of Small Targets 2007*], **6699** (2007).

[2] Mahalanobis, P. C., "On the generalised distance in statistics," *Proceedings of the National Institute of Sciences of India* **2**(1), 49–55 (1936).

[3] Horwood, J. T., Aristoff, J. M., Singh, N., and Poore, A. B., "A comparative study of new non-linear uncertainty propagation methods for space surveillance," in [*SPIE Proceedings: Signal and Data Processing of Small Targets 2014*], **9092** (2014).

[4] Aristoff, J. M., Horwood, J. T., Singh, N., and Poore, A. B., "Non-linear uncertainty propagation in orbital elements and transformation to Cartesian space without loss of realism," in [*Proceedings of the 2014 AAS/AIAA Astrodynamics Specialist Conference*], (August 2014).

[5] DeMars, K. J., Jah, M. K., and Schumacher, Jr., P. W., "The use of short-arc angle and angle rate data for deep-space initial orbit determination and track association," in [*Proceedings of the Eighth US/Russian Space Surveillance Workshop*], (October 2009).

[6] Hill, K., Sabol, C., and Alfriend, K. T., "Comparison of covariance-based track association approaches with simulated radar data," in [*Proceedings of the AAS Kyle T. Alfriend Astrodynamics Symposium*], (May 2010). Paper AAS-10-318.

[7] Sabol, C., Sukut, T., Hill, K., Alfriend, K. T., Wright, B., Li, Y., and Schumacher, P., "Linearized orbit covariance generation and propagation analysis via simple Monte Carlo simulations," in [*Proceedings of the 20th AAS/AIAA Space Flight Mechanics Meeting*], (February 2010). Paper AAS-10-134.

[8] Aristoff, J. M., Horwood, J. T., Singh, N., and Poore, A. B., "Multiple hypothesis tracking (MHT) for space surveillance: theoretical framework," in [*Proceedings of the 2013 AAS/AIAA Astrodynamics Specialist Conference*], (August 2013).

[9] Snedecor, G. W. and Cochran, W. G., [*Statistical Methods*], Iowa State University Press, 8-th ed. (1989).

[10] Durbin, J. and Knott, M., "Components of Cramér-von Mises statistics I," *Journal of the Royal Statistical Society Series B* **34**(2), 290–307 (1972).

[11] Stephens, M. A., "EDF statistics for goodness of fit and some comparisons," *Journal of the American Statistical Association* **69**(347), 730–737 (1974).

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

[13] Horwood, J. T. and Poore, A. B., "Orbital state uncertainty realism," in [*Proceedings of the Advanced Maui Optical and Space Surveillance Technologies Conference*], (September 2012).

[14] Horwood, J. T. and Poore, A. B., "Gauss von Mises distribution for improved uncertainty realism in space situational awareness," *SIAM Journal of Uncertainty Quantification* (2014). (To Appear).

[15] Goldstein, H., [*Classical Mechanics*], Addison-Wesley, Reading, MA (1965).

[16] DeMars, K. J., Jah, M. K., Cheng, Y., and Bishop, R. H., "Methods for splitting Gaussian distributions and applications within the AEGIS filter," in [*Proceedings of the 22nd AAS/AIAA Space Flight Mechanics Meeting*], (February 2012). Paper AAS-12-261.

[17] Horwood, J. T., Aragon, N. D., and Poore, A. B., "Gaussian sum filters for space surveillance: theory and simulations," *Journal of Guidance, Control, and Dynamics* **34**(6), 1839–1851 (2011).

[18] Horwood, J. T. and Poore, A. B., "Adaptive Gaussian sum filters for space surveillance," *IEEE Transactions on Automatic Control* **56**(8), 1777–1790 (2011).

[19] Ito, K. and Xiong, K., "Gaussian filters for nonlinear filtering problems," *IEEE Transactions on Automatic Control* **45**(5), 910–927 (2000).

[20] Terejanu, G., Singla, P., Singh, T., and Scott, P. D., "Uncertainty propagation for nonlinear dynamic systems using Gaussian mixture models," *Journal of Guidance, Control and Dynamics* **31**(6), 1623–1633 (2008).

[21] Alspach, D. and Sorenson, H., "Nonlinear Bayesian estimation using Gaussian sum approximations," *IEEE Transactions on Automatic Control* **17**(4), 439–448 (1972).

[22] Vallado, D. A. and Seago, J. H., "Covariance realism," in [*Proceedings of the 2009 AAS/AIAA Astrodynamics Specialist Conference*], (August 2009). Paper AAS 09-304.

[23] Golub, G. and Van Loan, C., [*Matrix Computations*], John Hopkins University Press, Baltimore, MD (1996).

[24] Mann, H. B. and Wald, A., "On the choice of the number of class intervals in the application of the chi-square test," *Annals of Mathematical Statistics* **13**(3), 306–317 (1942).

[25] Darling, D. A., "The Kolmogorov-Smirnov, Cramér-von Mises tests," *Annals of Mathematical Statistics* **28**(4), 823–838 (1957).

[26] Burnham, K. P. and Anderson, D. R., [*Model Selection and Multimodel Inference*], Springer, New York (2002).

[27] Shapiro, S. S. and Wilk, M. B., "An analysis of variance test for normality (complete samples)," *Biometrika* **52**, 591–611 (1965).

[28] D'Agostino, R. and Pearson, E. S., "Tests for departure from normality: empirical results for the distributions of b_2 and $(b_1)^{1/2}$," *Biometrika* **60**(3), 613–622 (1973).

[29] Julier, S. J., Uhlmann, J. K., and Durant-Whyte, H. F., "A new method for the nonlinear transformation of means and covariances in filters and estimators," *IEEE Transactions on Automatic Control* **55**, 477–482 (2000).

[30] Aristoff, J. M., Horwood, J. T., and Poore, A. B., "Orbit and uncertainty propagation: a comparison of Gauss-Legendre-, Dormand-Prince-, and Chebyshev-Picard-based approaches," *Celestial Mechanics and Dynamical Astronomy* **118**, 13–28 (2014).

[31] Aristoff, J. M., Horwood, J. T., and Poore, A. B., "Implicit Runge-Kutta-based methods for fast, precise, and scalable uncertainty propagation," (under review) (2014).

[32] Csörgö, S. and Faraway, J. J., "The exact and asymptotic distributions of the Cramér-von Mises statistics," *Journal of the Royal Statistical Society Series B* **58**(1), 221–234 (1996).

APPENDIX

For applications of the averaged uncertainty realism test described in Section 2.3, Table 1 provides two-sided confidence intervals for the corresponding test statistic $\bar{\mathcal{U}}$ given by Equation (6). For applications of the Cramér-von Mises goodness-of-fit test described in Section 2.4.2, Table 2 provides one-sided confidence intervals for the test statistic Q_k given by Equation (10). The confidence intervals in Table 2 were extracted from the manuscript of Csörgö and Faraway.[32]

Table 1. Two-sided confidence intervals for the test statistic $\bar{\mathcal{U}} \sim \chi^2(nk)/(nk)$, used in the averaged uncertainty realism test in Section 2.3, for common significance levels, state space dimensions n, and sample sizes k.

	Two-Sided Confidence Interval for $\bar{\mathcal{U}} \sim \chi^2(nk)/(nk)$			
(n, k)	90%	95%	99%	99.9%
(6, 100)	[0.906967, 1.096823]	[0.890031, 1.116282]	[0.857548, 1.154969]	[0.820868, 1.200960]
(7, 100)	[0.913733, 1.089515]	[0.897967, 1.107444]	[0.867685, 1.143045]	[0.833416, 1.185295]
(8, 100)	[0.919203, 1.083639]	[0.904391, 1.100344]	[0.875906, 1.133483]	[0.843616, 1.172757]
(6, 1000)	[0.970160, 1.030219]	[0.964533, 1.036099]	[0.953598, 1.047654]	[0.941013, 1.061171]
(7, 1000)	[0.972360, 1.027965]	[0.967142, 1.033399]	[0.956997, 1.044076]	[0.945314, 1.056558]
(8, 1000)	[0.974135, 1.026149]	[0.969248, 1.031226]	[0.959742, 1.041197]	[0.948790, 1.052848]
(6, 10000)	[0.990522, 1.009515]	[0.988716, 1.011347]	[0.985191, 1.014934]	[0.981111, 1.019107]
(7, 10000)	[0.991224, 1.008808]	[0.989551, 1.010503]	[0.986285, 1.013822]	[0.982505, 1.017682]
(8, 10000)	[0.991790, 1.008238]	[0.990224, 1.009823]	[0.987168, 1.012926]	[0.983629, 1.016535]

Table 2. One-sided confidence intervals for the Cramér-von Mises test statistic (10) for common significance levels and sample sizes k.

	One-Sided Confidence Interval for the Cramér-von Mises test statistic			
k	90%	95%	99%	99.9%
10	[0.008333, 0.34510]	[0.008333, 0.45441]	[0.008333, 0.71531]	[0.008333, 1.07428]
20	[0.004167, 0.34617]	[0.004167, 0.45778]	[0.004167, 0.72895]	[0.004167, 1.11898]
50	[0.001667, 0.34682]	[0.001667, 0.45986]	[0.001667, 0.73728]	[0.001667, 1.14507]
200	[0.000417, 0.34715]	[0.000417, 0.46091]	[0.000417, 0.74149]	[0.000417, 1.15783]
1000	[0.000083, 0.34724]	[0.000083, 0.46119]	[0.000083, 0.74262]	[0.000083, 1.16120]
∞	[0, 0.34730]	[0, 0.46136]	[0, 0.74346]	[0, 1.16204]