Nguyen, Theanh, Chan, Tommy, & Thambiratnam, David
(2014)
Controlled Monte Carlo data generation for statistical damage identification employing Mahalanobis squared distance.
Structural Health Monitoring, *13*(4), pp. 461-472.

https://doi.org/10.1177/1475921714521270

Controlled Monte Carlo data generation for statistical damage identification employing Mahalanobis squared distance

Theanh Nguyen, Tommy HT Chan, David P Thambiratnam

Abstract

The use of Mahalanobis squared distance (MSD) based novelty detection in statistical damage identification has become increasingly popular in recent years. The merit of the MSD-based method is that it is simple and requires low computational effort to enable the use of a higher-dimensional damage sensitive feature which is generally more sensitive to structural changes. MSD-based damage identification is also believed to be one of the most suitable methods for modern sensing systems such as wireless sensors. Although possessing such advantages, this method is rather strict with the input requirement as it assumes the training data to be multivariate normal which is not always available particularly at an early monitoring stage. As a consequence it may result in an ill-conditioned training model with erroneous novelty detection and damage identification outcomes. To date, there appears to be no study on how to systematically cope with such practical issues especially in the context of a statistical damage identification problem. To address this need, this paper proposes a controlled data

School of Civil Engineering and Built Environment, Queensland University of Technology, Australia

Corresponding author:
Theanh Nguyen, School of Civil Engineering & Built Environment, Queensland University of Technology, 2 George St, Brisbane, GPO Box 2434, QLD 4001, Australia.
Email: theanh.nguyen@qut.edu.au

generation scheme which is based upon the Monte Carlo simulation methodology with the addition of several controlling and evaluation tools to assess the condition of output data. By evaluating the convergence of the data condition indices, the proposed scheme is able to determine the optimal setups for the data generation process and subsequently avoid unnecessarily excessive data. The efficacy of this scheme is demonstrated via applications to a benchmark structure data in the field.

Keyword

Statistical damage identification, Mahalanobis squared distance (MSD), novelty detection, multivariate normal (multinormal), data generation, Monte Carlo, data condition assessment

Introduction

It is well-known that environmental and operational variations (EOVs) can prevent genuine structural damage in real civil structures from being identified since their effects can be larger than those from the genuine structural damage.[1, 2] One of the most popular approaches to deal with this, especially when measures of EOVs are not fully available, is based on statistical pattern recognition. In this case, machine learning algorithms are oftentimes used to learn the underlying trend induced by EOVs and create a robust damage index which can be considered to be invariant under the EOV presence. Amongst different methods in this approach, Mahalanobis squared distance (MSD) based damage identification is believed to be one of the best in unsupervised learning mode i.e. only using data from undamaged structures.[3, 4] In this regard, one will simply turn MSD-based (multivariate) outlier analysis into a novelty detection method and attempt to identify a potentially damaged observation as an outlier.[5, 6] Well-known for its simplicity and computational efficiency, MSD-based method has good potential to be cooperated on embedded modern sensing systems such as wireless sensors [4, 7]. However, the proper use of the standard MSD for the novelty detection purpose theoretically requires the training data needs to be multivariate normal (short as multinormal) or also known as multi-Gaussian.[5, 8] Due to the unavailability of complete multinormal data in many practical applications, one can obtain an approximation by increasing the observation-to-variable ratio.[9, 10] In practical structural monitoring, however, this is not always experimentally available particularly at an early monitoring stage. To systematically cope with such an adverse situation, this paper present a controlled data generation scheme which is based upon the Monte Carlo simulation methodology cooperated with several controlling and evaluation tools to assess the output data condition. By evaluating the convergence of the data condition indices, the proposed data generation scheme is able to determine the optimal simulation input parameters that need to be used and subsequently avoid improper simulation setups or unnecessarily excessive data. The efficacy of this scheme is demonstrated via applications to benchmark experimental data in the field. The layout of this paper is as follows. The next section provides descriptions of MSD-based damage identification and the controlled Monte Carlo data generation scheme. The benchmarks and their dataset used in this study are then briefly described. In the two last sections, detailed analyses and discussions are first provided before the key findings are summarised in the conclusion.

Damage identification and data generation methods

MSD-based damage identification

There are two main types of data used in statistical damage identification process. In general, the primary (or raw) data acquired by sensors is not directly used but is transformed into a damage sensitive feature (DSF) which then become input data for the statistical training model. This secondary data is oftentimes in a much lower dimension compared to the primary one so as to alleviate the computational effort and to extract the most meaningful structural information. Typical examples for this can be found in the case of common DSFs such as modal parameters and auto-regressive (AR) vectors.[7, 11-16]

Suppose that a training dataset consists of p (i.e. DSF dimension) variables and n observations. If its shape approximates a multinormal distribution, this dataset can be represented by the sample mean vector ($\bar{x}$) and the sample covariance matrix (S). In this case, these two parameters are often referred as "sufficient statistics". By using the standard MSD technique as a multivariate outlier analysis [6], each feature vector (x_i) for either the training or testing purposes will be converted into a damage index in terms of distance measure (d_i) as follows

$$d_i = (x_i - \bar{x})^T \mathrm{S}^{-1}(x_i - \bar{x}) \tag{1}$$

In damage identification context, the mean and covariance should be formulated as an exclusive measure, or in other words, consisting of no potential outlier from the testing phase.[6] After computing all training distances, the assumption of a multinormal distribution again allows the estimation of the threshold from the basis of chi-square distribution for the training distances.[5] It is because under such an assumption, one can specify a statistical threshold for the distances based on a distribution quantile or equivalently a confidence level.[5, 8] There might be a trade-off in choosing the confidence level: using very high level of confidence level might not be able to detect a lightly damaged case that is known as one class of Type II errors but the least critical. However, such confidence level can assist in avoiding as many as possible false-positive indication of damage (i.e. Type I errors).[5]

In the testing phase, whenever a new DSF comes, its corresponding distance can be used to compare against the threshold to determine whether it corresponds to a normal or damaged state. In this sense, the anticipation is that the more severe a damaged state is, the more significant the difference between its actual distance and the threshold becomes. This has been observed in prior studies in this area.[4, 6, 14]

As seen earlier, even though the MSD-based damage identification possesses a simple computational structure, the success of this method depends on whether its assumption of data distribution (i.e. multinormal) can be adequately satisfied. Since complete multinormal data is seldom available in practice, the overall remedy, stemming from the central limit theorem (CLT) and the law of large numbers (LLN), is to increase the observation size (n) relative to number of variables (p).[9, 10] One simple and inexpensive approach to realise this remedy in the context of measured data shortage is using the controlled Monte Carlo data generation scheme.

Controlled Monte Carlo data generation

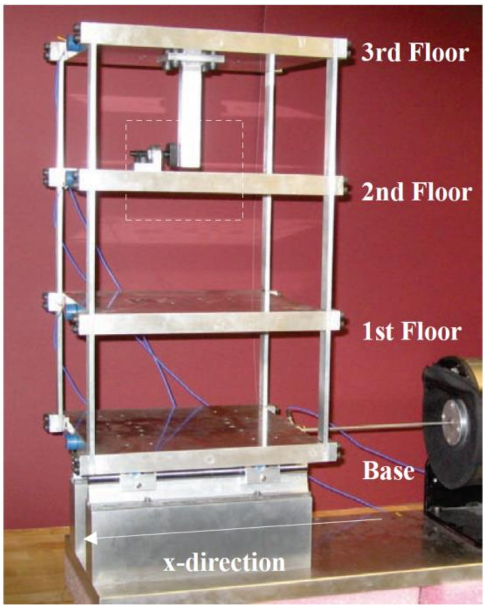
As previously mentioned, the controlled data generation scheme developed in this paper originates from the Monte Carlo simulation methodology. In a broad sense, a Monte Carlo method today refers to any simulation method that involves the use of random numbers and was termed by Neumann and Ulam in the 1940's.[17, 18] Being easy and inexpensive, this approach is particularly applicable for evaluation of highly multidimensional and complex problems.[19] To conduct a Monte Carlo simulation, one just needs to define a model that represents the population or phenomenon of interest and a criterion to generate random numbers for the model. The latter commonly involves the use of a user-selected probability distribution. Once completed, the data generated from the model can then be used as though they were actual observations.[18] In the damage identification context, Monte Carlo data generation has also seen its applicability since the DSFs are often in high dimension. However, prior studies in the field have mainly applied the Monte Carlo simulation methodology in an ad hoc manner. The conventional trend in such studies was to generate large number of observations from the data seed of a single or few DSF(s) by applying certain amount of random Gaussian noise onto each copy.[6, 15] Even though the noise was constructed from a Gaussian distribution, its magnitude and the sample quantity were generally set in a rather uncontrolled manner. Another general suggestion from prior research is that using lower levels of noise allows more lightly damaged cases to be detected.[20] However, a possible problem for applying a too low level of noise in data generation is that subsequently generated observations might not be sufficiently random with respect to initial observations to improve the data condition (and this issue will be examined in the application section). Obviously, a more systematic data generation scheme is in need particularly when considering real structural monitoring circumstances with a certain number of observations initially available to form the seed. Such a type of seed apparently reflects more accurately the training conditions of structures but also requires a more thorough data generation scheme to be cooperated.

To cater to this need, the present paper proposes an enhanced data generation scheme termed as controlled Monte Carlo data generation (CMCDG). This is realised by adding into the conventional scheme two controlling tools that are in fact two data condition assessment methods and a robust probability-based evaluation procedure to assist these methods. Of the two condition assessment methods, the first one is based on evaluating the condition of the generated data through the condition of its sample covariance matrix which is represented by a well-known and robust index, i.e. (2-norm) condition number (COND) in linear algebra.[21, 22] On the other hand, the second method is based on one of the most popular graphical tools for evaluating multinormality of data i.e. the quantile-quantile (Q-Q) plot of a beta distribution or, in certain cases, a chi-square distribution.[9, 10, 23] In this study, the beta Q-Q plot is employed since it is generally more accurate than the chi-square counterpartner.[10] To evaluate multinormality of a dataset, the actual plot of data is compared with the theoretical one and a significant discrepancy in the plot would indicate that the data no longer belongs to a multinormal distribution. Since the number of datasets generated by CMCDG for statistical evaluations is large, the root-mean-square error (RMSE), one of the most commonly-used discrepancy measures, between the theoretical and actual Q-Q plots will be used as another condition index. The mathematical expression of this measure will be included in the application section. The rationale of employing these two methods to evaluate CMCDG process is as follows. First, under the regulation of CLT and LLN, the sample covariance matrix (S) converges in probability to the actual population covariance matrix (Σ) as number of random observations (n) increases.[9] It is therefore sensible to anticipate that, as *n* increases, COND (S) also converges in probability to COND (Σ). Similarity can be seen for the second method. As *n* increases, the Q-Q plot is expected to converge in probability to the theoretical line and its RMSE is therefore anticipated to converge in probability to zero.

Inherent in the way that the two data condition assessment methods is implemented in CMCDG is a robust probability-based evaluation procedure with two robust measures i.e. the median and inter-quartile range (IQR) [24] to examine the central tendency and dispersion of COND and beta Q-Q RMSE. By tracking the convergence of these measures, CMCDG is able to determine the optimal noise level and possibly minimum number of data replications that need to be set in the simulation process. Details of CMCDG and its controlling and evaluation components are illustrated in the application section.

Description of the benchmark structure and data

The benchmark dataset used in this study is from Los Alamos national laboratory (LANL), USA and has been intensively used in recent statistical damage identification studies.[4, 7] This data was collected by four accelerometers from a benchmark building model (Figure 1) with varied practical conditions (Table 1) including stiffness deviation due to temperature change and mass difference (e.g. caused by traffic). Nonlinear damage was generated by contacting a suspended column with a bumper mounted on the floor below to simulating fatigue crack that can open and close under loading conditions, or loose connections in structures. Different levels of damage were created by adjusting the gap between the column and the bumper. In total, there were 9 undamaged states and 8 damaged states each of which consists of a number of tests performed to take into account excitation variability. In this study, the largest dataset available for public use with 50 tests for each state is used.[25] According to the test description [7], state 14 can be considered as the most severe one since it corresponds to the smallest gap case which induces the highest impact of contact. State 10 is the least severe damaged scenario whereas state 11, 12 and 13 can represent mid-level damage scenarios. Other states (i.e. 15, 16 and 17) are the variant states of either state 10 or 13 with mass added effect.



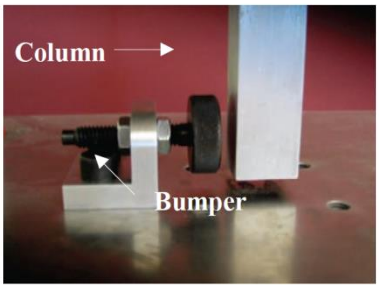


Figure 1. The test structure (left) and damage simulation mechanism (right) at LANL.[7]

Table 1. Data labels of the structural state conditions (adapted from LANL).[7]

Label	Feature	Description
State 1	Undamaged	Baseline condition
State 2	Undamaged	Mass = 1.2 kg added at the base
State 3	Undamaged	Mass = 1.2 kg added on the 1st floor
State 4	Undamaged	State 4-9: 87.5% stiffness reduction at various positions to simulate temperature impact (see [7] for details)
State 5	Undamaged	
State 6	Undamaged	
State 7	Undamaged	
State 8	Undamaged	
State 9	Undamaged	
State 10	Damaged	Gap = 0.20 mm
State 11	Damaged	Gap = 0.15 mm
State 12	Damaged	Gap = 0.13 mm
State 13	Damaged	Gap = 0.10 mm
State 14	Damaged	Gap = 0.05 mm
State 15	Damaged	Gap = 0.20 mm & mass = 1.2 kg added at the base
State 16	Damaged	Gap = 0.20 mm & mass = 1.2 kg added on the 1st floor
State 17	Damaged	Gap = 0.10 mm & mass = 1.2 kg added on the 1st floor

Analyses and discussions

The data used in this study is from the second floor sensor which is close to the damage location to guarantee the sensitivity of the method when classifying different-level damage cases. The testing data, established by taking 20 first tests in each of 9 undamaged states and all tests of damaged structure, therefore has 580 (i.e. 20×9+50×8) observations. With 30 remaining tests in each undamaged state for the training purpose, differently sized learning data can be formed by varying number of training tests (i.e. from as low as 1 up to 30) taken in each learning state. This is to illustrate the impact of the observation size reflected through the two data condition assessment methods (as previously mentioned) by means of pure experimental data. For the sake of simplicity, this number of tests per learning state will be referred below as "state observation size". DSF used in this investigation is the autoregressive (AR) vector which has also been used in recent studies using this dataset.[4, 7] Each raw data time series is first standardized to zero mean and unit variance before being transformed into an AR vector with a user-selected model order. Even though there are a number of order estimation techniques [4, 7], in this study, the heuristic technique, based on directly observing RMSE of the AR model, is adopted. The basis for this adoption is that it reflects the actual impact of order change on prediction capacity of AR model which, in the opinion of the present authors, is the most crucial. For the sake of completeness, the following part will present a brief description of AR model and the order estimation method based on RMSE.

The AR (p) model, for a regularly sampled time series process Y with n observations can be described by the following formulae

$$y_i = \hat{y}_i + \varepsilon_i \tag{2}$$

$$\hat{y}_i = \sum_{j=1}^{p} \phi_j y_{i-j} \tag{3}$$

where y_i, $\hat{y}_i$ and ε_i are the measured signal the predicted signal and the residual error, respectively at the discrete time index i while ϕ_j is the jth AR variable which can be estimated by one of a number of techniques such as Burge, least squares and Yule-Walker.[26] RMSE of the time series predicted by an AR (p) model with respect to the measured signal is therefore as follows

$$\mathrm{RMSE}(p) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{4}$$

To find an appropriate model order, RMSE is plotted as a function of the model order which in turn can be estimated by minimizing the RMSE value. Figure 2 shows the average RMSE of AR models with the orders ranging from 1 to 40 for each of the 9 undamaged states. One can see that, RMSE becomes significantly steady for all 9 states from the order of 10 which suggests that one should choose the order at least from this value. In the following sections, this suggested starting order (i.e. 10) and one rather high (i.e. 30), along with one medium (i.e. 15) at some points when necessary will be used in the succeeding sections.

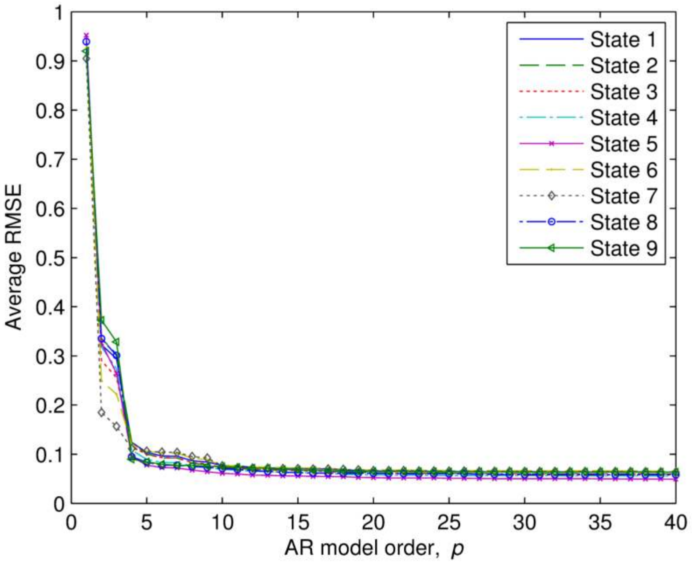


Figure 2. RMSE of AR models of increasing order for each undamaged state.

MSD-based damage identification performance on pure experimental data

At the model order *p*, one DSF (i.e. AR vector) for each observation in either training or testing data is computed by least squares technique. This leads to 270 by *p* training data and 580 by *p* testing data. The threshold distance which is used to differentiate between the undamaged and damaged states is established based on the highest confidence level (i.e. 100%). This can avoid as many as possible the Type I error which, in the opinion of the present authors, is more crucial than the ability of detecting lightly damaged cases which one might achieve by using a lower confidence level. Using this confidence level, the MSD training model is able to correctly detect almost all damage cases – only 1 out of 400 Type II error tests is occasionally found across the lower-dimensional DSF (i.e. AR10 and AR15). The high-dimensional DSF (AR30) herein has seen no Type II error indicating that it is slightly more sensitive to damage than AR10 and AR15. Overall, the results have confirmed that the previously selected confidence level is appropriate.

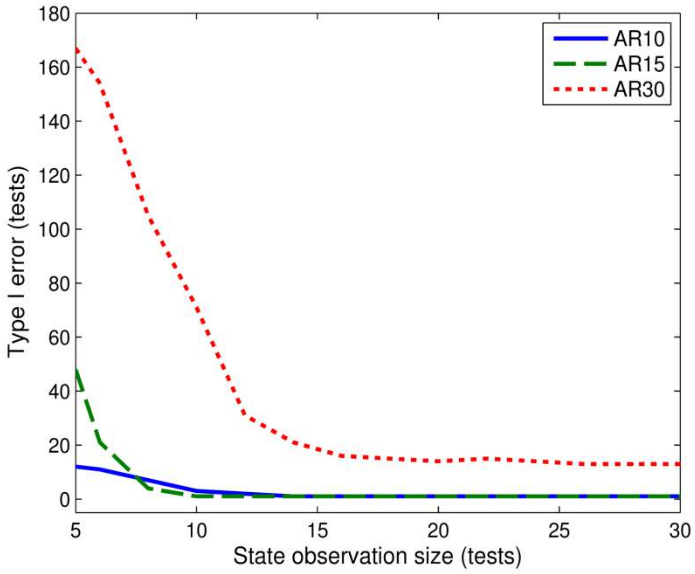


Figure 3. Type I error of increasing observation size.

In spite of using the highest confidence level, the result of Type I errors significantly differs from that of the Type II errors especially for the smaller range of observations. Figure 3 plots the number of false positive errors (out of total 180 tests) against the state observation size in the range between 5 to the maximum (i.e. 30) as previously described. It can be seen that, the Type I error becomes significant for most of the DSF dimensions when less than a quarter of the maximum training data is available and is generally higher for higher dimensions. This is most likely due to the fact that, with higher number of variables, higher-order AR models require more observations to be as sufficiently trained as lower-order models. It is worth noting that this problem is well known as "curse of dimensionality" [5] and the use of CMCDG herein should be seen to mitigate this problem.

Performance of two condition assessment methods on pure experimental data

In Figure 4, the condition number of the MSD model is plotted against the state observation size across three DSF dimensions in normal linear scale as well as logarithmic (log) scale to facilitate the comparison at different ranges.

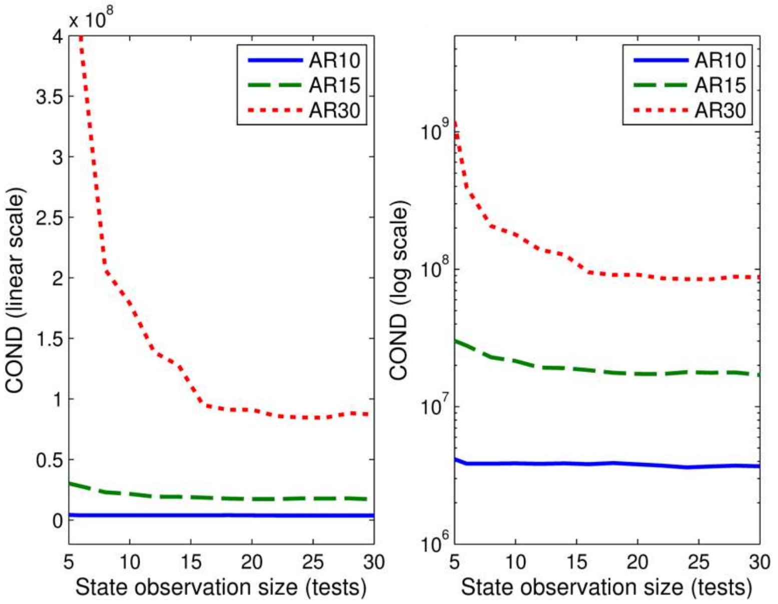


Figure 4. COND in linear (left) and log (right) scales.

From this figure, one could see that the condition number tends to converge after certain number of observations which is larger for higher DSF dimensions. Overall, it can be seen that the convergence trend of this condition number is in fairly good agreement with the performance result presented in Figure 3.

To construct a beta Q-Q plot, the training distance in formula (1) first needs to be scaled by a factor related to the sample size (n) as follows

$$d_i^* = \frac{n\,d_i}{(n-1)^2} \tag{5}$$

If the training data is multinormal, this scaled distance would follow a beta distribution. The scaled distance is then ranked in ascending order and plotted with the corresponding beta quantiles.[10] For illustration purpose, Figure 5 shows the beta Q-Q plots of AR10 (at the state observation of 5 and 13 tests) and AR30 (at the state observation of 8 and 16 tests). These two (one small and one medium) datasets are selected to represent two (one unstable and one improved) conditions of the data, respectively.

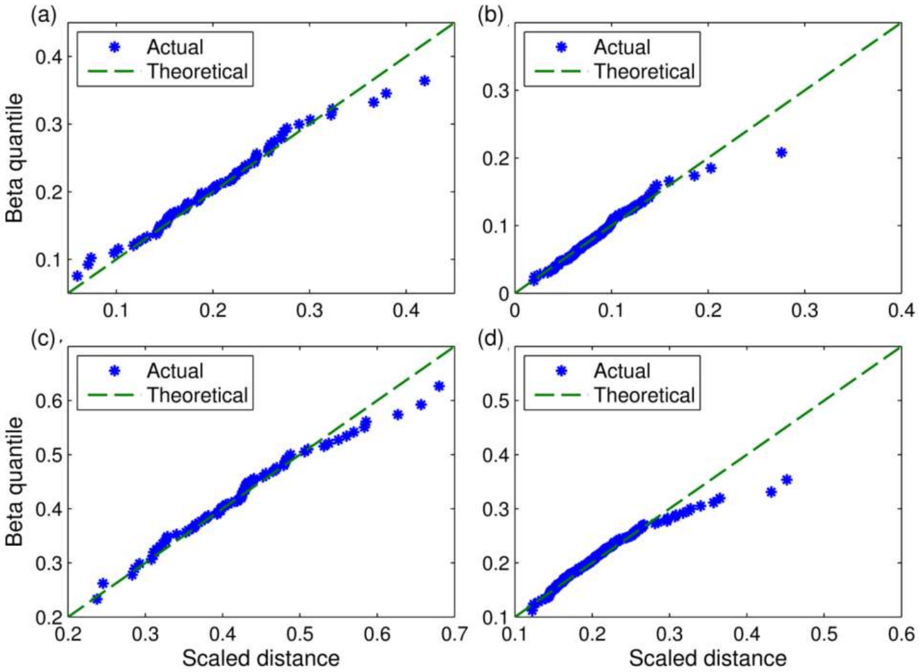


Figure 5. Beta Q-Q plot of (a) AR10-05 tests, (b) AR10-13 tests, (c) AR30-08 tests and (d) AR30-16 tests.

From Figure 5, one can see that increasing number of observations generally improves the agreement between the actual and theoretical plots for most of the data points. This reveals that it is feasible to use a good-of-fitness measure between the two plots as another data condition index (besides COND) to evaluate a huge number of datasets generated from CMCDG process. As previously mentioned, the measure adopted is RMSE which is one of the most commonly-used measures for this type of purpose.

Performance of CMCDG on premature data

Previous results have shown that the condition of the experimental data will require certain numbers of observations to reach a stable point. Before that, data can be considered as premature and will therefore need a compensation solution such as from CMCDG to improve its condition. In this section, CMCDG will be applied on two premature training datasets each of which is for each DSF type, i.e. at the state observation size of 5 tests (for AR10) and 8 tests (for AR30) as preliminarily checked by COND and beta Q-Q plot as shown in Figure 4 and 5. With such limited observations, the main problem for these two premature datasets is the Type I errors as previously discussed and presented in Figure 3. Out of a total of 180 tests, the original Type I errors of these two (AR10 and AR30) training datasets are 13 and 105 tests (or 7.2% and 58.3% in terms of the error rate), respectively. Under the CMCDG scheme, each premature dataset is first employed as the seed to generate a (user-specified) number of additional datasets of the same size as the seed (by means of random noise) and all the datasets are then tiled one after another to obtain the final data. The random noise herein is generated based on its optimal level in root-mean-square (RMS) sense with respect to the largest deviation of the training DSFs. In this study, the optimal level of noise is determined by the convergence basis of median and IQR of COND. As an example, Figure 6(a) and (b) shows the probability distribution of COND values at different noise levels (from 0.05 to 5%) when running 10,000 simulations to evaluate the case of using CMCDG generating 19 additional data replications. Note that the presented noise levels on Figure 6 are unequally distributed to accommodate different ranges of noise. From Figure 6(a) and (b), one can clearly see that the median and IQR of COND are very large if very low level of noise is employed such as at 0.05 or 0.1%. This is because when noise levels that are too low are applied to the data generation process, subsequently generated observations will have inadequate randomness with respect to the initial observations in the seed as previously discussed. In this case, the covariance matrix becomes more computationally unstable (reflected by larger and more widely variable COND values [27]) than those formulated by later ranges of noise levels.

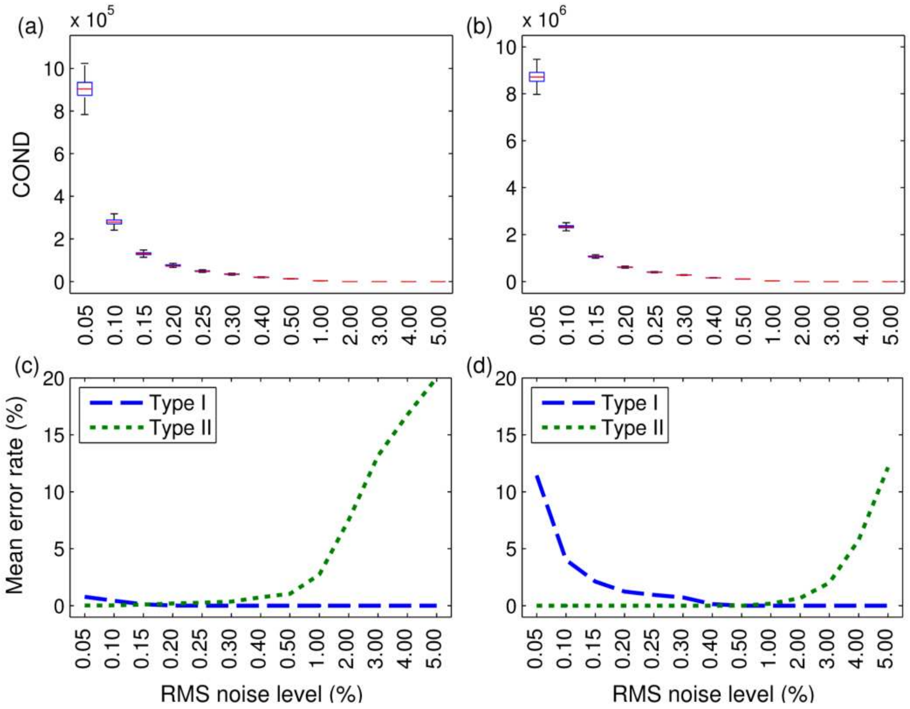


Figure 6. COND and mean error rate of increasing noise level: AR10 (left) and AR30 (right).

However, when the noise level increases, COND rapidly decreases in both median and IQR values. This results in unnoticeable difference in these values from the noise level of around 0.4% onward even though the noise increment later is set at 1%. For correlation purposes, the corresponding mean Type I and Type II errors are also shown in Figure 6(c) and (d) in relative sense with respect to a total of 180 Type I and 400 Type II tests. One can first see that the Type I error result is generally in good agreement with the convergence trend of COND. Note that higher Type I error rate for

AR30 (in comparison with AR10) at low noise levels should not be seen as abnormal since the initial rate of the premature AR30 data is 58.3% (while that of AR10 is only 7.2%) as previously mentioned. On the other hand, Figure 6(c) and (d) appear to show certain impact for the Type II error at high noise levels. However, checking the details across multiple noise levels from 0.5% (for AR10) or 1% (for AR30) up to 5% has revealed that all the Type II errors for both AR10 and AR30 merely belong to the most lightly damaged states (i.e. state 10 and its two variants, state 15 and 16 as illustrated in Table 1). Detecting such a damage state may be desirable but not always in the highest priority of damage identification as previously discussed in the regard to choosing the confidence level. Nevertheless, using a higher-dimensional DSF (such as AR30 that has lower Type II error rate) and/or a correct noise level (close to such an optimal level as 0.4% herein) will enhance the damage identification outcome. This also reaffirms the need to determine of an optimal noise level such as being considered in the CMCDG scheme herein since this can lead to a more satisfactory solution.

To find a possibly minimum number of data replications to be used in CMCDG, the same approach used to produce Figure 6 will be implemented with a minor swap. The noise level is fixed (at 0.3% for AR10 and 0.5% for AR30) while number of data replications is varied. Figure 7 shows the probability distribution of COND and beta Q-Q RMSE along with the mean rate of the Type I error. Again, one can see that both COND and RMSE tend to rapidly converge in both median and IQR values after a certain number of data replications. The figure also shows that the convergence trends of these two indices are in excellent agreement with each other and with that of the Type I error. On the other side, the Type II error results can be retained as more or less the same as those from pure experimental data as previously presented. Once again, there is no single error for AR30 while AR10 only fails to detect one or two most lightly damaged cases out of total 400 tests. This is probably mostly due to the nature of lower-dimensional DSF such as AR10 which is less sensitive to damage than high-dimensional DSF like AR30 as previously remarked. This also highlights the feasibility of CMCDG in assisting the use of the high-dimensional DSF that may result in higher capability of detecting structural damage.

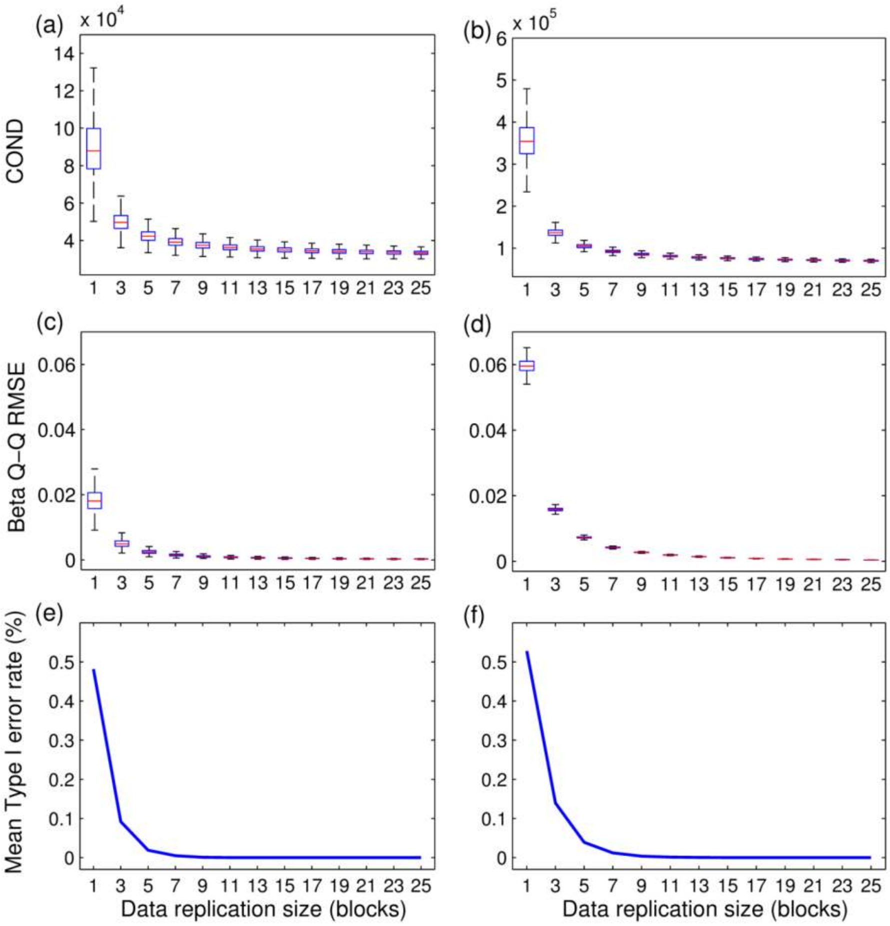


Figure 7. COND, Q-Q RMSE and Type I error rate of increasing replication size: AR10 (left) and AR30 (right).[*]

Based on the convergence of these two condition indices, one can adopt 15 as a possibly minimum number of additional data replications that need to be generated in CMCDG for both DSF types of this demonstration example. At this replication size, both post-

[*] As they are (nearly) zero, Type II error rates have been omitted for a better display of Type I errors

CMCDG datasets (of both DSF types) face no single Type I error across 180 total tests. Compared to aforementioned initial error rates (7.2% and 58.3%) of original datasets, this obviously reflects excellent improvements for the Type I testing performance for both DSF types in general and for high-dimensional DSF (AR30) in particular.

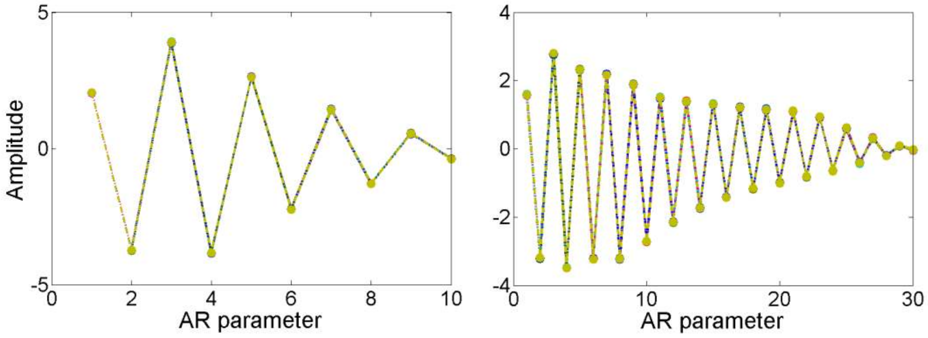


Figure 8. Overlay of one typical seed observation and its 15 variants: AR10 (left) and AR30 (right).

In Figure 8 for each DSF type, one typical seed (initial) observation and its 15 variants generated by CMCDG are overlaid together and one can see that they are almost identical. This means that the noise addition process in CMCDG does not induce significant variations on the amplitude of the observation. Instead, the efficacy of CMCDG is mainly from the generation of multiple additional random observations to provide a sufficiently large random dataset as directed by CLT and LLN. Finally, to illustrate detailed effectiveness of CMCDG on the training data multinormality, the beta Q-Q plots of two typical datasets generated by CMCDG using aforementioned selected noise levels (0.3% and 0.5%) and replication sizes (15 blocks for both DSF types) are shown in Figure 9. Compared to those of original (pre-CMCDG) datasets as in Figure 5(a) and 5(c), there are inarguable improvements in terms of the agreement between actual and theoretical lines of the post-CMCDG datasets of both DSF types. This once again confirms the effectiveness of the CMCDG scheme.

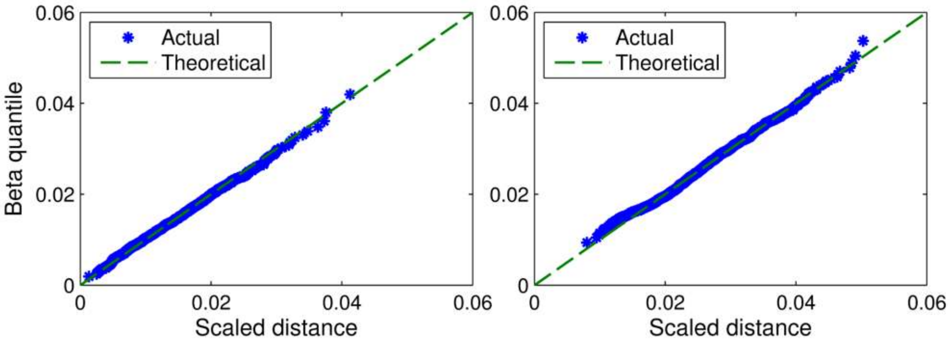


Figure 9. Post-CMCDG beta Q-Q plots: AR10 (left) and AR30 (right).

From results presented, it has become apparent that one can conquer the data shortage by employing CMCDG without having to suffer from data burden that is more likely to be confronted during the application of the uncontrolled data generation approach.

Conclusions

This paper has proposed an enhanced data generation scheme named CMCDG which can be used to compensate for the shortage of data such as at an early monitoring stage. Targeting a more systematic approach, CMCDG is constructed by adding into the conventional data generation approach two condition assessment methods cooperated with a robust probability-based evaluation procedure. Stemming from a computationally efficient method in linear algebra, COND has been shown to be a simple but useful condition index. This indicator can be used for not only assessing the data condition but also statistically evaluating the effect of random disturbance at different levels such as random noise. Based on the latter usage, the optimal noise level and the possibly minimum number of data replications to be used in CMCDG can be derived so that the generated data can be used for reliable damage identification while being kept reasonable in size. As a different approach, the second assessment method can first act as a convenient tool for graphically examining the status of any single dataset. To use in CMCDG besides COND to work with huge number of simulated datasets, the previous graphical evaluation method is transformed into a single condition indicator which is actually one of the most common good-of-fitness measures, RMSE, to track the discrepancy between actual data and theoretical data. The rationale of utilising the convergence basis of all of data condition indices for determining optimal input for CMCDG has been proved under the regulation of two well-known theorems i.e. CLT and LLN. These two theorems have also been found to be the theoretical bases not only for the CMCDG scheme developed herein but also for the traditional data generation approach. The implementation and application of CMCDG to a benchmark data have shown that CMCDG and its added components can compensate well for the data shortage, improve computational stability and therefore the reliability of MSD-based damage identification. This has also highlighted an important role of CMCDG in assisting the high-dimensional DSF such as AR30 that is likely to have higher sensitivity toward a lightly damaged case. Finally, as been shown to be able to improve multinormality of data, CMCDG can be seen as a promising scheme not only for novelty detection based damage identification but also for statistically-based structural analysis in a broader field.

Acknowledgments

The first author gratefully appreciates the financial support for his research from Vietnam Government and Queensland University of Technology (QUT). Useful discussions with Prof. Keith Worden of University of Sheffield (UK), generous data sharing made by LANL Engineering Institute (USA) in general and from Dr. Charles Farrar in particular are also acknowledged with thanks.

References

1. Farrar CR, Sohn H and Worden K. Data Normalization: A Key for Structural Health Monitoring. In: *Proceedings of the 3rd International Workshop on Structural Health Monitoring*, Stanford, CA, USA 2001 September 17-19, pp. 1229-38.

2. Sohn H, Farrar C, Hemez F, Shunk D, Stinemates D and Nadler B. A review of structural health monitoring literature: 1996-2001. Report: LA-13976-MS, Los Alamos National Laboratory, USA, 2003

3. Manson G, Worden K and Allman D. Experimental validation of a structural health monitoring methodology. Part II. Novelty detection on a Gnat aircraft. *J Sound Vib* 2003; 259(2): 345-63.

4. Figueiredo E, Park G, Farrar CR, Worden K and Figueiras J. Machine learning algorithms for damage detection under operational and environmental variability. *Struct Health Monit* 2011; 10(6): 559-72.

5. Farrar CR and Worden K. *Structural health monitoring: a machine learning perspective*. Chichester, West Sussex: Wiley, 2013.

6. Worden K, Manson G and Fieller NRJ. Damage detection using outlier analysis. *J Sound Vib* 2000; 229(3): 647-67.

7. Figueiredo E, Park G, Figueiras J, Farrar C and Worden K. Structural health monitoring algorithm comparisons using standard data sets. Report: LA-14393-MS, Los Alamos National Laboratory, USA, 2009

8. Filzmoser P, Garrett RG and Reimann C. Multivariate outlier detection in exploration geochemistry. *Comput Geosci* 2005; 31(5): 579-87.

9. Johnson RA and Wichern DW. *Applied multivariate statistical analysis*. 5th ed. Upper Saddle River, NJ: Prentice Hall, 2002.

10. Rencher AC. *Methods of multivariate analysis*. 2nd ed. New York: John Wiley & Sons, 2002.

11. Farrar CR, Doebling SW and Nix DA. Vibration-based structural damage identification. *Philos Trans R Soc A* 2001; 359(1778): 131-49.

12. Worden K and Manson G. The application of machine learning to structural health monitoring. *Philos Trans R Soc A* 2007; 365(1851): 515-37.

13. Nguyen T, Chan THT and Thambiratnam DP. Effects of wireless sensor network uncertainties on output-only modal-based damage identification. *Aust J Struct Eng* 2014; 15(1): 15-25; http://dx.doi.org/10.7158/S12-041.2014.15.1

14. Gul M and Catbas FN. Statistical pattern recognition for structural health monitoring using time series modeling: theory and experimental verifications. *Mech Syst Sig Process* 2009; 23(7): 2192-204.

15. Worden K, Sohn H and Farrar CR. Novelty detection in a changing environment regression and interpolation approaches. *J Sound Vib* 2002; 258(4): 741-61.

16. Sohn H, Farrar C, Hunter N and Worden K. Applying the LANL statistical pattern recognition paradigm for structural health monitoring to data from a surface-effect fast patrol boat. Report: LA-13761-MS, Los Alamos National Laboratory, USA, 2001

17. Martinez WL and Martinez AR. *Computational statistics handbook with MATLAB*. Boca Raton: CRC Press, 2002.

18. Wikipedia. Monte Carlo method, http://en.wikipedia.org/wiki/Monte_Carlo_method (2002, accessed August, 2012).

19. Dunn WL and Shultis JK. *Exploring Monte Carlo Methods*. Burlington: Elsevier, 2011.

20. Worden K, Farrar CR, Manson G and Gyuhae P. The fundamental axioms of structural health monitoring. *Proc R Soc A* 2007; 463(2082): 1639-64.

21. Golub GH and Van Loan CF. *Matrix computations*. 3rd ed. Baltimore: Johns Hopkins University Press, 1996.

22. Strang G. *Linear algebra and its applications*. 4th ed. Belmont, CA: Thomson Brooks/Cole, 2006.

23. Sharma S. *Applied multivariate techniques*. New York: John Wiley & Sons, 1995.

24. Martinez WL and Martinez AR. *Exploratory data analysis with MATLAB*. Boca Raton, Florida: Chapman & Hall/CRC, 2005.

25. Los Alamos National Laboratory. SHMTools and mFUSE, http://institute.lanl.gov/ei/software-and-data/SHMTools/ (2010, accessed August, 2012)

26. Ljung L. *System Identification Toolbox™ 7 User's Guide*. Natick, MA: MathWorks, 2011

27. Statistics Toolbox Development Team. *Statistics Toolbox^{TM} 7 User's Guide*. Natick, MA: MathWorks, 2011