MSET PERFORMANCE OPTIMIZATION THROUGH REGULARIZATION

J. WESLEY HINES and ALEXANDER USYNIN
Nuclear Engineering Department
The University of Tennessee
Knoxville, TN 37996-2300
E-mail : jhines2@utk.edu

Received October 6, 2004
Accepted for Publication February 12, 2005

The Multivariate State Estimation Technique (MSET) is being used in Nuclear Power Plants for sensor and equipment condition monitoring. This paper presents the use of regularization methods for optimizing MSET's predictive performance. The techniques are applied to a simulated data set and a data set obtained from a nuclear power plant currently implementing empirical, on-line, equipment condition monitoring techniques. The results show that regularization greatly enhances the predictive performance. Additionally, the selection of prototype vectors is investigated and a local modeling method is presented that can be applied when computational speed is desired.

KEYWORDS : regularization, inferential MSET, predictive performance

1. INTRODUCTION

The optimization of predictive performance of empirical modeling applied to condition monitoring is of primary importance for early fault detection. One source of uncertainty that is of particular interest comes from the collinearity of the predictor variables, or predictor vectors, which causes an ill-posed problem. Hadamard [1] defined a well-posed problem as a problem that satisfies the three following conditions:

- The solution for the problem exists.
- The solution is unique.
- The solution is stable or smooth under small perturbations of the data; i.e. small perturbations in the data should produce small perturbations in the solution.

The input collinearity, which is found in data sets with large mutual correlations, makes the solution non-unique, causing the problem to be ill-posed and requiring special considerations to ensure consistent, reliable results.

Previous work shows that kernel-based techniques, such as kernel regression, local linear regression, and local polynomial regression, can be properly regularized through the selection of the optimal kernel width [2]. The research presented herein investigates the effect of kernel width selection and ridge regularization of the MSET memory matrix as two methods to optimize inferential MSET modeling predictive performance.

Early work has focused on the use of regularization methods for producing reliable, consistent, low noise empirical predictions for sensor calibration verification using a variety of empirical models [3,4]. It applied regularization methods such as ridge regression [5], complexity regularization [6, 7], local regularization [8, 9] to models such as non-linear partial least squares [10], neural networks [11] and linear regression. A survey of these techniques applied to equipment monitoring is available [12].

This prior research provides a solid background in regularization techniques, but did not apply them to the MSET model, which is currently being used for on-line monitoring at close to a dozen nuclear power plants. This research presents the results of the application of regularization methods to MSET, with the objective of minimizing predictive error. Additionally, the selection method and number of prototype vectors used in the memory matrix is investigated.

2. INFERENTIAL MSET REGULARIZATION

Let us consider the following nonlinear model:

$$y = f(x) + \varepsilon \tag{1}$$

where X is a $n \times m$ data matrix of n observations of m predictor variables;

y is a *n x 1* vector of the response variable;
$f(X)$ is a unknown nonlinear function representing the true relationship between the response and predictor variables;
ε is a *n x 1* vector of the random errors distributed according to the normal distribution ($\varepsilon \sim N(0, \sigma^2 I)$).

The objective is to estimate the unknown function $f(X)$ from the given matrix of observations (X, y) and the given assumption that the random errors ε come from a normal distribution.

The inferential MSET estimator [13] is one approach to solve such a problem. The MSET algorithm was originally developed by Jack Mott [14] and later refined and applied to nuclear power plant surveillance by researchers at Argonne National Laboratory [15]. Being a "lazy" learning technique [16, 17], the inferential MSET estimator represents the solution to the problem as a weighted sum of given past observations. This model belongs to a class of techniques called non-parametric techniques. Equation 2 represents the inferential MSET formula in the matrix form [2].

$$\hat{Y}_{MSET} = \frac{1^T (X_{tr}^T \otimes X_{new})(X_{tr}^T \otimes X_{tr})^{-1}}{\left[1^T (X_{tr}^T \otimes X_{tr})^{-1} (X_{tr}^T \otimes X_{new}) 1\right]} Y_{tr} \qquad (2)$$

where X^T_{tr} is the matrix of training data, which is commonly called the memory matrix, X_{new} is the query data vector, 1is a column of ones, and Y_{tr} is the training response data. The symbol $\otimes$ stands for a non-linear similarity operator termed a kernel function

$$K_h(x, x').$$

A kernel function for $x \in R^d$ has the following properties [18]:

1. K(x, x') takes on a maximum value where x=x'.
2. |K(x, x')| decreases with |x-x'|.
3. K(x, x') is a general function of 2d variables

The most common choice of kernel is the Gaussian operator,

$$K_h(x, x_i) = \frac{1}{\sqrt{2\pi}h} e^{\frac{(x-x_i)^2}{2h^2}} \qquad (3)$$

where x_i is the data point around which the kernel is placed, x is the data point being compared to x_i, and h is the smoothing parameter or bandwidth. SmartSignal (SS) Corporation licensed MSET from Argonne National Laboratory and subsequently extended and modified the basic MSET technology in developing their commercial Equipment Condition Monitoring (SmartSignal eCM™) software [19] which uses a proprietary kernel. Because the SS similarity operator is proprietary, this paper will employ the common Gaussian kernel. Many researchers believe that the kernel choice does not have a major impact on the algorithm performance [18].

The kernel bandwidth (h) is a user selectable parameter in the MSET model and an inappropriate choice would be a source of potential underperformance; therefore, it should be optimized. In terms of kernel bandwidth, the behavior of the MSET estimator is similar to the behavior of the well-known kernel regression estimator (Equation 4). A narrow kernel bandwidth overfits the solution by depending on only a few noisy training observations, while a large kernel bandwidth over-smooths the predictions by depending on too many.

$$\hat{Y}_{KR} = \frac{1^T (X_{tr}^T \otimes X_{new})}{\left[1^T (X_{tr}^T \otimes X_{new}) 1\right]} Y_{tr} \qquad (4)$$

The experiments performed in this research reveal that the inferential MSET estimator is less sensitive to the kernel bandwidth than the standard kernel regression estimator. This is because the matrix of training data: $D = (X'_{tr} \otimes X_{tr})$, performs as a matrix of bandwidth correction factors. The D matrix makes the Gaussian kernel approximate a high-order kernel so that the smoothing effects of the modified kernel become less aggressive. In other words, applying the memory matrix as in Equation 2 changes the shape of the kernel and changes the effective kernel bandwidth. The optimization of the kernel bandwidth parameter will be discussed in more detail in following sections.

As can be noted from Equation 2, if the memory matrix is near singular, its inversion leads to an unstable solution. In other words, inversion of an ill-conditioned memory matrix presents a potential additional problem to obtaining a stable solution. To address this issue, ridge regularization of the memory matrix is applied to the inferential MSET estimator.

2.1 Ridge Regularization of the Memory Matrix

One of the possible reasons for getting a high-variance prediction may be the ill conditionality of the memory matrix. When the memory matrix of observations, X, has very similar prototype vectors, it becomes ill-conditioned and the elements of the inverted memory matrix become very large. When this happens, the MSET estimator becomes an amplifier. If the given observations are contaminated with random noise, obtaining a stable prediction becomes difficult, since the dominant component of the predicted output is the amplified noise. The ill-conditioned memory matrix may cause the obtained prediction to be excessively noisy as shown in Figure 1. This figure presents an example using simulated data in which the noise in the input variables get amplified and the MSET prediction is extremely noisy.

Figure 2 shows how the ill-conditioned nature of the memory matrix impacts the inferential MSET estimator

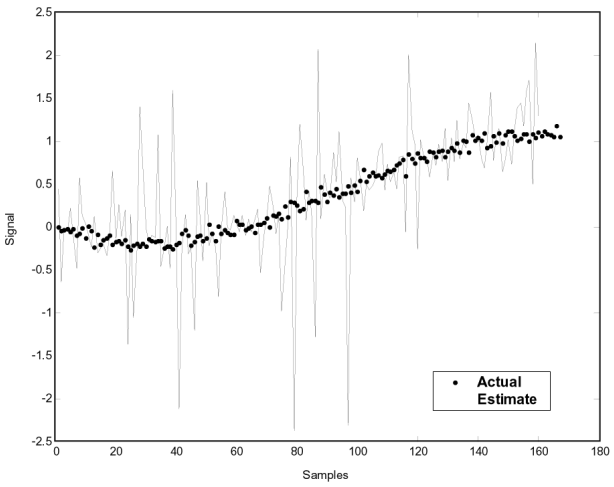

Fig. 1. An Ill-conditioned Memory Matrix Leads to a High Variance Response.

predictive performance. Here, the figure of merit is the leave-one-out cross-validation (LOO CV) error (the upper plot). The lower plot represents the growth of the condition number of the memory matrix as the kernel width is increased.

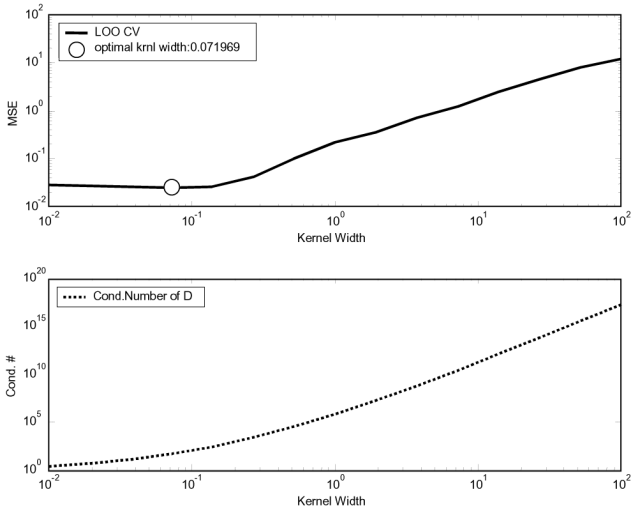

Fig. 2. LOO CV Error Versus the Kernel Bandwidth Parameter; Condition Number of the Memory Matrix Versus the Kernel Bandwidth Parameter.

In this example, increasing the kernel bandwidth increases the memory matrix condition number. When the condition numbers grows beyond $10^2 - 10^3$ (100-1000), the prediction degrades, as shown by an increase in predictive error. This is a common problem and is well discussed in terms of linear regression for which the most common solution is ridge regression [5]. Decreasing the condition number of the memory matrix greatly improves the prediction stability. One possible approach to improving the conditionality of the memory matrix is to apply ridge regularization to the matrix when it is inverted. The ridge regularized MSET estimator takes the following form:

$$\hat{Y}_{MSET} = \frac{1^T (X_{tr}^T \otimes X_{new})(X_{tr}^T \otimes X_{tr} + \lambda I)^{-1}}{\left[1^T (X_{tr}^T \otimes X_{tr} + \lambda I)^{-1} (X_{tr}^T \otimes X_{new}) 1\right]} Y_{tr}, \qquad (2)$$

where λ is the ridge regularization parameter ($\lambda > 0$).

In general, finding the optimal value of λ is a one-dimensional optimization problem, which can be solved using a non-linear method such as conjugate gradient descent. Our experiments show that setting λ equal to one is a good default value for most models when the data is standardized to have a variance of one, the kernel width is set equal to one, and the entire data set is used as the memory matrix.

2.2 MSET Gaussian Kernel Width Optimization

When using the Gaussian kernel similarity operator, one can vary the bandwidth to obtain stable and smoothed predictions. In the simplest case, there is one common bandwidth parameter *h*, which is used for each input variable. In this case, the bandwidth matrix is represented as a diagonal matrix with only one single value of *h*. To account for the different input variable's contributions, the bandwidth matrix may be diagonal with the different bandwidth values for each input. In such a case, the optimization problem becomes *p*-dimensional, where *p* is the number of input variables. The most complicated case is when the bandwidth matrix is a full matrix. The optimization problem becomes p^2-dimensional and is very computationally burdensome even if *p* is relatively small.

This research considers the *p*-dimensional problem of optimizing the Gaussian kernel. That is, the diagonal bandwidth matrix consists of different values of *h*.

2.2.1 Experimental Methodology

The ridge regularization method was applied to the data set obtained from a nuclear power plant. The data set is comprised of 136 critical process variables observed each hour during an 11-month-period. Including the entire data set, which contains 7955 observations, as the memory matrix would create a significant computational burden; therefore, a portion of the data was selected to represent the entire data set. A one-month period of observations was selected as the training data set and the subsequent one-month period was selected to be the validation set. The data sets were downsampled so that only 2-hour observations were retained in the training and validation data sets. Thus, 451 training vectors and 471 validation observations were used for the numerical experiments.

The response variable (predicted variable) was an arbitrarily chosen variable. The selection of the best

predictor variables for the selected response variable was performed using the LASSO regression based method [20]. The following is an outline of the experimental methodology.

1. A given data set is divided into 2 parts: training and validation. The validation set is used to measure the predictive performance by means of the mean prediction error.
2. The training data is standardized so that each input has a mean of zero and variance of one. The same standardization parameters (training set mean and variance) are used to standardize the validation set.
3. Several prototype vectors are selected from the training set so that they properly cover the training space.
4. The cost function minimum is obtained using a conjugate gradient descent method for the two following cases.
   a). Both the kernel width vector and the ridge parameter are optimized.
   b). Only the kernel width vector is optimized. The ridge parameter is selected to be constant and equal to one.

In non-linear optimization, the minimum point is not guaranteed to be the global minimum, but will be an improvement to the starting point of a kernel width of 1 and regularization parameter of 1. This starting point is a good, and common, choice because the data is standardized to have a variance of one.

When using the Gaussian kernel, the estimator hat matrix can be obtained. This allows the use of a complexity based cost function. Mallow's CL [21] is chosen to be the cost function because it takes complexity into consideration and provides a more robust measure of future predictive performance.

$$f(h,\lambda) = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n} + \frac{2\sigma}{n} trace(H), \qquad (6)$$

where Y is a vector of the training samples;
$\hat{Y}$ is a vector of the estimates;
n is the number of training samples;
H is the hat matrix of the MSET estimator, which depends on the bandwidth parameters h;
σ^2 is the variance of random noise in the response variable.

The random noise variance cannot be obtained precisely, but it can be estimated using the following:

$$\hat{\sigma}^2 = \frac{2}{3(n-2)}\sum_{i=2}^{n-1}\left(Y_i - \frac{1}{2}\left(Y_{i-1} - Y_{i+1}\right)\right)^2. \text{ [22]} \qquad (7)$$

2.2.2 Experimental Results

The experiments are conducted using both the nuclear data set and a simulated data set consisting of partially correlated data with both linear and non-linear associations. The simulated data, which are corrupted by various levels of Gaussian noise, is used as a more controllable data set so that more conclusive results can be reported.

One of the important choices made by MSET users is the number of data observations selected for use in the memory matrix. Usual statistical theory states that more data means more information and results in better predictive performance. However, currently, reduced data memory matrices are employed to produce low variance predictions. To explore the effects of the number of prototype vectors in the memory matrix, a reduced data set is compared to the full training data set. Table 1 summarizes the results obtained for the Gaussian kernel-based MSET estimator.

The predictions made using the larger training prototypes set resulted in better predictive performance when compared to that obtained using the reduced number of prototype vectors. Since adding additional

Table 1. The Mean Prediction Errors Provided by the Gaussian Kernel Based MSET Estimator with the Optimized Bandwidth Parameters Using the Entire Training Vector Data Set and a Reduced Prototype Vector Set Equal to 4 Time the Number of Predictors.

		Gaussian Kernel-based MSET				
		Constant Ridge = 1		Ridge Parameter Is Optimized		
		MSE	Cond. #	MSE	Cond. #	Optimal Ridge
Nuclear Data	Reduced Memory Matrix Model 28 prototypes	0.0090	8.20	0.0084	42.5	0.138
	Entire Dataset Based Model 451 prototypes	0.0083	57.0	0.0045	4.66	9.74
Simulated Data	Reduced Memory Matrix Model 24 prototypes	0.3740	19	0.3340	1	2.07×10^4
	Entire Dataset Based Model 584 prototypes	0.313	474	0.312	1.1	1011

vectors can make the matrix more ill-conditioned, this improvement takes place only if the memory matrix is regularized using the ridge method. The use of very large data sets without regularization produces high variance predictions due to ill-conditioning caused by similar vectors. It can be concluded that ridge regularization produces a better mean prediction error than the mean prediction error obtained by reducing the number of training prototypes down to four times the number of predictors.

Since the optimization problem is non-linear and of *P*-dimension, the optimization procedure may fall into a local minima. This problem was shown to exist by Buckner [9]. One possible solution to the problem of local minima may be the use of several starting points for the conjugate gradient descent optimization. However, this will consume additional computational time and cannot guarantee the avoidance of local minima.

3. MEMORY MATRIX PROTOTYPE VARIABLE SELECTION

In this section, we will discuss a technique that makes use of a reduced size memory matrix, termed "local MSET". Kernel regression is termed a local technique because only prototype vectors similar to the query point are used in the locally weighted regression equation. However, the standard implementation of the MSET algorithm allows all of the prototype vectors to have an effect on the prediction, so it is not a true local technique. If one wishes MSET to be a true local technique, then the memory matrix should only include prototype vectors similar, or near, the query point.

One additional non-parametric modeling consideration is that of boundary effects. A prediction obtained with a kernel-based technique may be unduly biased near the input space boundary. When queried near the boundary of the training data range, kernel methods becomes less accurate since fewer samples can be averaged at the boundary. The weighted average value tends to be biased towards the center of the training region. The boundary effect impacts all kernel regression estimators including the MSET estimator.

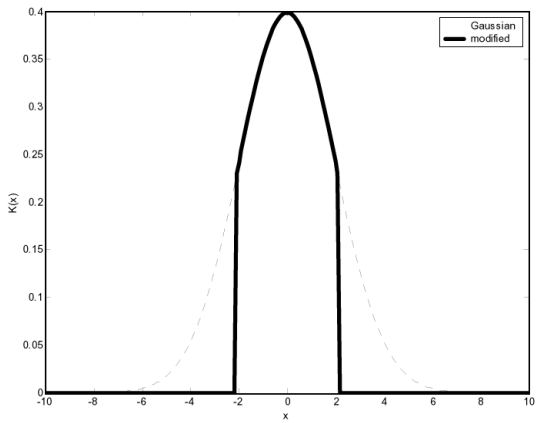

Fig. 3. The Gaussian Kernel Modified to Account for the Local Set of the Prototype Vectors.

One possible solution to reduce the biasing, is to use a relatively small set of prototype vectors to obtain a prediction at a query point. There are several methods to select the similar prototype vectors. In the most common, which is similar to the k-nearest neighbor classifier, the prediction is constructed by considering some constant number (k) of the prototype vectors that are most similar to the point of interest. Another method is to choose a region around the query point. In this case, a kernel regression estimator is constructed by selecting a crisp set of local prototypes as shown in Figure 3.

In the case of the MSET estimator, the idea of a local prototype matrix may seem to be contrary to the currently used method of selecting prototype vectors covering the entire training sample range. The currently used technique selects prototype vectors that are the most different from each other and bound the entire range of data. The local memory matrix selects prototype vectors from the entire training set that lie close to the query point, and therefore are most similar to each other. Because of the similarity of the prototype vectors, the local memory matrix is commonly ill-conditioned and ridge regularization must be used to produce robust, low-variance predictions.

3.1 Methodology

To determine how well the local method performs in comparison to the standard "global" prototype selection, prediction performance is quantified using the following two selection methods:

1. The "global" method selects the set of prototype vectors from the entire range of data; the selected set of prototypes remains the same for each query point.
2. The local method selects a number of prototype vectors nearest to the query point. A number of 15-20 vectors is a good empirical number that provides a plausible average value.

To provide equal conditions for both test methods, we use the same number of prototypes for both the methods. In the global experiment, the number of prototype vectors is set equal to four times the number of model predictors. Therefore, the number of selected vectors used to build the local memory matrix is also set equal to four times the number of model predictors. To calculate the prediction for each query point, one needs to select a new set of local prototype vectors.

3.2 Experimental Results

Figure 4 presents the results of a comparison of the

local memory matrix and standard global memory matrix methods. Ridge regularization was applied to both the "local" and "global" memory matrices. As can be seen, if almost no regularization is used (ridge parameter = 10^{-4}), the local memory matrix has a smaller prediction error and thus outperforms the global technique. In some cases, such as that shown in the simulated data model of Figure 4, a properly regularized global technique will perform better than a local method. However, a properly regularized local model will usually perform best. These results are consistent with results using data sets from a variety of equipment condition monitoring applications including data sets from automotives, airlines, and fossil power plants [23].

In Section 2, it was stated that optimized ridge regularization of a global memory matrix results in better

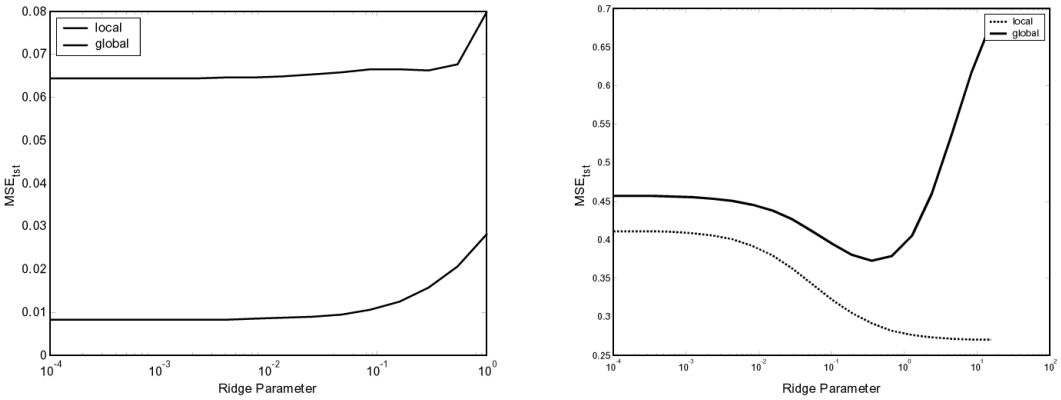

Fig. 4. A Comparison Between the Local Method and a Reduced "global" Memory Matrix. The Left Plot is for the Nuclear Data model. The Right Plot is for the Simulated Data Model.

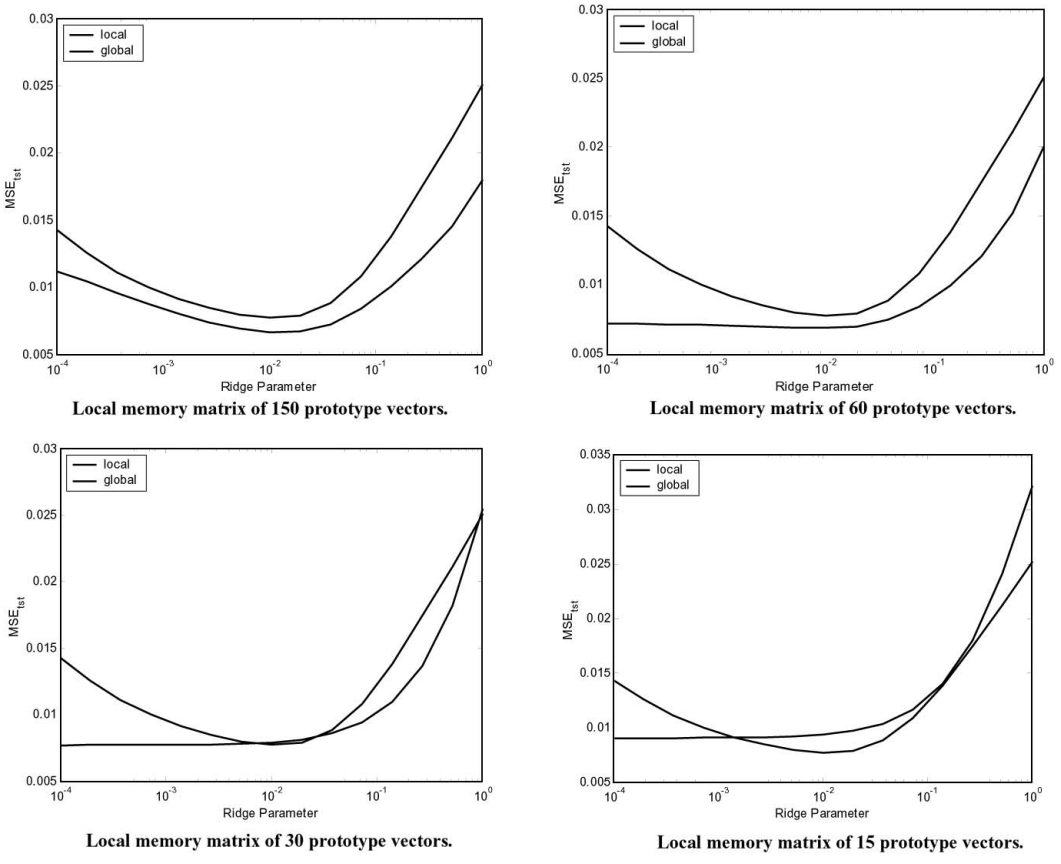

Fig. 5. A Comparison of Various Local Memory Matrix Methods with the Global Memory Matrix Method that Uses the Entire Training Set.

prediction performance than the common method of reducing the number of prototype vectors. Figure 5 presents a comparison between the local memory matrix method and the global memory matrix method for different sized prototype vector sets. The global memory matrix is composed of 199 prototype vectors. As can be seen, the local method performs better than the global technique, except for the last case when the local method is limited to 15 vectors and the global method is optimally regularized.

In Figure 5 we also see that when the number of local prototype vectors is small, regularization is not as important. In all cases without regularization, the local method performs better because vectors dissimilar to the query point are not allowed to adversely influence the prediction.

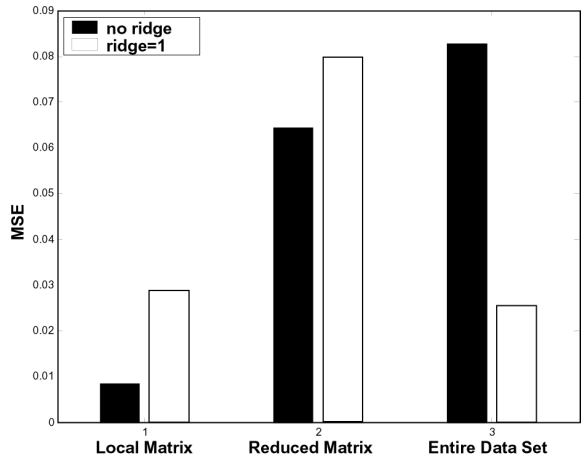

Fig. 6. Mean Squared Errors Provided by the "Local" and "Global" Memory Matrix Based MSET Estimators

Figure 6 summarizes the results of a comparison of the MSET estimators employing all three memory matrix methods. The three types are the "local" method, the "global" method using a reduced matrix, and the "global" method using all available training data. As can be seen, the "local" memory matrix outperforms the global techniques. The memory matrix using the entire range of data without any regularization performs the worst. However, if the full matrix is properly regularized, it produces a predictive error comparable to the local memory matrix based method.

These results make sense. From information theory we know that the more information, the better. So it would be expected that the entire set of observations provides a good prediction. However, the similar vectors cause the matrix to be ill-conditioned and thus cause poor performance; regularization can reduce those effects. A reduced memory matrix is not ill-conditioned because the vectors are not similar and thus does not require regularization. The local method does not contain as much information as the matrix with the entire set of observations, but it does contain the ones that are the most similar and thus the ones that contain the predictive information. Therefore, the local method contains the important information and may, or may not, require regularization depending on the specific data set. In the example above, regularization was not needed.

4. CONCLUSIONS

MSET regularization can be performed through the use of two different methods:

1. The use of the ridge regularization of the memory matrix.
2. The optimization of the vector of kernel bandwidth parameters.

The elements of the inverted memory matrix can be treated as correction factors for the kernel shape. Therefore, the proper ridge regularization of the memory matrix impacts the MSET predictive performance as much as the use of optimal kernel bandwidth parameters. To find the optimal ridge parameter one needs to solve a 1-dimensional minimization problem, which is more usable than the p-dimensional problem to be solved in the case of kernel bandwidth parameter optimization. Additionally, one can obtain a plausible prediction through the use of a default ridge parameter value equal to 1 when the data is standardized, the kernel width is set equal to one, and the entire data set is used as the memory matrix. These conclusions came from months of experimentation on several diverse data sets in an attempt to develop a methodology that is easy to integrate and generalize into an automated monitoring system.

A second conclusion is that increasing the number of training patterns hinders predictive performance unless regularization is used. This may be stated differently. If the number of prototype vectors is reduced, regularization may not be necessary. However, predictive performance can be optimized when more prototype vectors are used in conjunction with optimal regularization techniques. In most cases, the optimal technique uses a local memory matrix and regularization is not necessary. This method is less computationally intensive than using a large global matrix with regularization, which would be the second best method.

Regularization methods have been proven to improve MSET predictive performance when applied in an optimal manner. The results are general and applicable to a wide range of applications including nuclear power plant sensors and equipment condition monitoring.

REFERENCES

[1] Hadamard, J., Lectures on Cauchy's Problem in Linear Partial Differential Equations, Yale University Press, New Haven, 1923.

[2] Gribok, A.V., J.W. Hines, A. Urmanov, and R.E. Uhrig,

"Use of Kernel Based Techniques for Sensor Validation in Nuclear Power Plants", *Statistical Data Mining and Knowledge Discovery*, pp. 217-231, Chapman and Hall/CRC Press, 2004.

[3] Hines, J.W., A.V. Gribok, I. Attieh, and R.E. Uhrig, "Regularization Methods for Inferential Sensing in Nuclear Power Plants", *Fuzzy Systems and Soft Computing in Nuclear Engineering*, Ed. Da Ruan, Springer, 1999.

[4] Gribok, A.V., J.W. Hines, A. Urmanov, and R.E. Uhrig, "Regularization of Ill-Posed Surveillance and Diagnostic Measurements", *Power Plant Surveillance and Diagnostics*, Eds. Da Ruan and P. Fantoni, Springer, 2002.

[5] Hoerl, A.E., and R.W. Kennard, "Ridge regression: biased estimation for nonorthogonal problems", *Technometrics*, 12, pp.55-67, 1970.

[6] Urmanov, A.M., A.V. Gribok, J.W. Hines, and R.E. Uhrig, "Complexity-penalized model selection for feedwater inferential measurements in nuclear power plants", ANS International Topical Meeting on Nuclear Plant Instrumentation, Controls, and Human-Machine Interface Technologies (NPIC&HMIT 2000), Washington, DC, 2000.

[7] Urmanov, A., A. Gribok, J.W. Hines, and Brandon Rasmussen, "Application of Information Complexity in Principal Component Regression Modeling of the Venturi Meter Drift", published in the proceedings of the *Maintenance and Reliability Conference* (MARCON 2000), Knoxville, TN, May 6-9, 2001.

[8] Hines, J.W., A. Gribok, A. Urmanov and M. Buckner, "Selection of Multiple Regularization Parameters in Local Ridge Regression Using Evolutionary Algorithms and Prediction Risk Optimization", *Inverse Problems in Engineering*, Vol. 11, No. 3, pp. 215-227, 2003.

[9] Buckner, M., "Learning From Data with Localized Regression and Differential Evolution", Ph.D. Dissertation, The University of Tennessee, Nuclear Engineering Department, 2003.

[10] Rasmussen, B., J.W. Hines, and R.E. Uhrig , "Nonlinear Partial Least Squares Modeling for Instrument Surveillance and Calibration Verification", by published in the proceedings of the Maintenance and Reliability Conference (MARCON 2000), Knoxville, TN, May 7-10, 2000.

[11] Hines, J.W., A.V. Gribok, I. Attieh, and R.E. Uhrig, "Neural Network Regularization Techniques for a Sensor Validation System", American Nuclear Society Annual Meeting, San Diego, California, June 4-8, 2000.

[12] Hines, J.W, A. Gribok, A. Urmanov and R.E. Uhrig, "Heuristic, Systematic, and Informational Regularization for Process Monitoring", *International Journal of Intelligent Systems* on Intelligent Systems for Process Monitoring, Wiley Publishers, Vol. 17, No. 8, pp. 723-750, 2002.

[13] Singer, R.M., K.C. Gross, J.P. Herzog, R.W. King, and S.W. Wegerich, "Model-Based Nuclear Power Plant Monitoring and Fault Detection: Theoretical Foundations," Proc. 9th Intl. Conf. on Intelligent Systems Applications to Power Systems, Seoul, Korea, 1996.

[14] Mott, J, Young, and R.W. King, "Pattern Recognition Software for Plant Surveillance", US DOE Report, 1987.

[15] Gross, K. C., R. M. Singer, S. W. Wegerich, J. P. Herzog, R. Van Alstine, and F. K. Bockhorst, "Application of a Model-based Fault Detection System to Nuclear Plant Signals," Proc. 9th Intl. Conf. on Intelligent Systems Applications to Power Systems, Seoul, Korea, 1997.

[16] Aha, D.W., Editorial. *Artificial Intelligence Review*, 11(1-5), 1-6, Special Issue on Lazy Learning, 1997.

[17] Atkeson, C.G., Moore, A.W., Schaal, S., "Locally Weighted Learning", *Artificial Intelligence Review*, 11(1-5), 11-73, 1997.

[18] Cherkassky, V., and F. Mulier, *Learning From Data*, John Wiley & Sons, 1998.

[19] Wegerich, S, R. Singer, J. Herzog, and A. Wilks, "Challenges Facing Equipment Condition Monitoring Systems", MARCON 2001, Gatlinburg TN, May 6-9, 2001.

[20] Tibshirani, R., "Regression shrinkage and selection via the lasso", *Journal of the Royal Statistical Society* B 58, 267-288, 1996.

[21] Mallows, C.L., "Some comments on CP", Technometrics, 15, No. 4, pp. 661-675, 1973.

[22] Muller, Hans-Georg "Nonparametric regression analysis of longitudinal data", Springer-Verlag, New York, p.94, 1988.

[23] Hines, J.W., and A Usynin, "Regularization Methods for MSET and other Kernel Techniques", Research Report for SmartSignal Inc., September 2003.