Using Principal Components in a Proportional Hazards Model with Applications in Condition-Based Maintenance
Author(s): D. Lin, D. Banjevic and A. K. S. Jardine
Source: *The Journal of the Operational Research Society*, Vol. 57, No. 8 (Aug., 2006), pp. 910-919

Published by: Palgrave Macmillan Journals on behalf of the Operational Research Society
Stable URL: http://www.jstor.org/stable/4102404
Accessed: 23-12-2015 08:49 UTC

REFERENCES
Linked references are available on JSTOR for this article:
http://www.jstor.org/stable/4102404?seq=1&cid=pdf-reference#references_tab_contents

You may need to log in to JSTOR to access the linked references.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Operational Research Society and Palgrave Macmillan Journals are collaborating with JSTOR to digitize, preserve and extend access to *The Journal of the Operational Research Society*.

http://www.jstor.org

This content downloaded from 138.73.1.36 on Wed, 23 Dec 2015 08:49:05 UTC
All use subject to JSTOR Terms and Conditions

Journal of the Operational Research Society (2006) 57, 910–919 © 2006 Operational Research Society Ltd. All rights reserved. 0160-5682/06 $30.00

www.palgrave-journals.com/jors

Using principal components in a proportional hazards model with applications in condition-based maintenance

D Lin, D Banjevic and AKS Jardine*

University of Toronto, Toronto, Ontario, Canada

This paper proposes the application of a principal components proportional hazards regression model in condition-based maintenance (CBM) optimization. The Cox proportional hazards model with time-dependent covariates is considered. Principal component analysis (PCA) can be applied to covariates (measurements) to reduce the number of variables included in the model, as well as to eliminate possible collinearity between the covariates. The main issues and problems in using the proposed methodology are discussed. PCA is applied to a simulated CBM data set and two real data sets obtained from industry: oil analysis data and vibration data. Reasonable results are obtained.

Journal of the Operational Research Society (2006) **57,** 910–919. doi:10.1057/palgrave.jors.2602058
Published online 31 August 2005

Keywords: condition-based maintenance; optimal maintenance policy; principal component analysis; proportional hazards model; hazard regression

Introduction

Condition-based maintenance (CBM) is an area that has been attracting more and more attention in industry. Owing to higher system complexity and higher demand of system reliability/availability, traditional time-based preventive maintenance has become inefficient in many cases. Recognizing this, many companies have shifted their maintenance programs to a smarter way of doing maintenance—to CBM. CBM is a maintenance strategy based on collected condition data that are related to the system health or status. This maintenance strategy results in a major repair/replacement, for example, when there is objective evidence of an impending failure. A CBM program, if correctly and effectively implemented, can save significant maintenance costs while reducing the number of functional and catastrophic failures and maintaining the required system reliability/availability. For more description and discussion of CBM, see the books by Williams *et al* (1994) and Moubray (1997).

The traditional way of doing CBM is:

(1) Collect off-line and on-line data;
(2) Set certain warning limits on the condition variables and/or features extracted from sensor signals, either based on some simple statistical analyses, such as descriptive statistics and trend analysis on the off-line data, or based on some conservative engineering-based experience;
(3) Establish a maintenance policy based on the warning limits and the current condition data.

A more advanced approach to CBM in step (2) is to build a statistical model for time to failure using data from step (1), such as the proportional hazard model (PHM), and to calculate an optimal maintenance policy based on the PHM and the requirements on system reliability/availability and/or cost considerations. Then, this policy is applied in the third step. Using PHM has an advantage over using warning limits on condition variables/features, since PHM relates the system working age, the covariates (condition variables and/or features), and the maintenance history to the risk of system failure. In addition, by estimating a PHM, one can assess how much relevance a covariate has to the risk of system failure.

Cox's PHM (Cox, 1972a, b) and its extensions have been widely applied to survival analysis in biology, medical science and engineering. A PHM with time-dependent covariates is a reasonable model to describe the failure mechanism in the CBM setting. Based on this model, an optimal maintenance decision policy that minimizes maintenance costs can be calculated (Makis and Jardine, 1992). The theoretical development has been implemented in the CBM optimization software EXAKT (Banjevic *et al*, 2001), for which a number of applications in different industrial areas such as electrical utility industry (Chevalier *et al*, 2004), food process industry (Jardine *et al*, 1999), mining industry

*Correspondence: AKS Jardine, Department of Mechanical and Industrial Engineering, University of Toronto, 5 King's College Road, Toronto, Ontario, M5S 3G8 Canada.
E-mail: jardine@mie.utoronto.ca

This content downloaded from 138.73.1.36 on Wed, 23 Dec 2015 08:49:05 UTC
All use subject to JSTOR Terms and Conditions

(Jardine *et al*, 2001), nuclear industry (Jardine *et al*, 2003), defence sector (Jeffris *et al*, 2004; Lin *et al*, 2004), construction sector (Monnot *et al*, 2004), and petrochemical industry (Vlok *et al*, 2002) have been reported. It appears that some practical issues have to be addressed to implement the CBM optimization methodology more effectively. We address two of these issues in this paper. One may be termed the 'curse of dimensionality', that is, the number of possible system states grows exponentially with the number of covariates. The other issue is the collinearity problem, that is, some linear combinations of covariates may be highly correlated. The former may lead to unaffordable computational costs. The latter may lead to ineffective estimation of the models and thus to unreliable results.

A usual method to solve the problem of dimensionality is to reduce the number of covariates. This solution is often referred to as dimension reduction. A common way to eliminate the collinearity is by transformation of covariates. There are many approaches to dimension reduction and/or collinearity elimination. Among them, the principal component analysis (PCA) is an approach that can be used to reduce dimension and eliminate collinearity in multiple linear regression (Jolliffe, 1986; Chatterjee *et al*, 2000). However, there has been little attempt to apply principal components (PCs) in proportional hazards regression, that is, to include the PCs, not the original covariates, in the hazard regression model. An approach in combining PCA and proportional hazards regression is to use PCA to reveal important factors (covariates) and then include those important factors, but not the PCs, in the proportional hazards regression (Danielyan *et al*, 1986). Another approach in a medical application focuses on dimension reduction by PCA and selecting a number of PCs that contributed the most to variance for the proportional hazards regression (Li and Li, 2004, http://repositories.cdlib.org/cbmb/surv2/, accessed 18 May 2005). Yet there has been no reported attempt to apply PCs and proportional hazards regression in the area of CBM except the MSc thesis work (Gao, 2003). The approach is to apply dynamic PCA on both original covariates and the covariates with shifted time lag. However, only the first few PCs are included in the model based solely on their variance. This way of selecting the PCs may not be appropriate as it may lead to exclusion of significant PCs which have small variance. This problem will be discussed later.

In this paper, we apply PCA and proportional hazards regression to CBM. The remainder of this paper is organized as follows. First, the CBM model and the optimal CBM policy are introduced. Then the proportional hazards model with time-dependent covariates is described and PCA is briefly introduced. Finally, the proportional hazards regression with PCs is described, and then used to analyse the simulated data and two real CBM data sets, followed by the conclusion of the paper.

CBM model and the optimal CBM policy

To find the optimal policy for a CBM program, following the work (Banjevic *et al*, 2001), we need to build the following three models:

(1) Failure model describing the effects of working age and condition variables (covariates) on the hazard of system failure;
(2) Covariate behaviour model describing the evolution of covariates over time;
(3) Cost model evaluating the cost of the CBM program.

The failure model used in this paper is the Weibull proportional hazards model with time-dependent covariates. The model is described in the next section. A finite state space Markov process model is used as the covariate behaviour model. The state space of covariates is discretized and the covariate readings are used to estimate the transition probability matrix. The control-limit replacement policy, that is, to replace/repair when the hazard exceeds a predetermined level (Banjevic *et al*, 2001), is assumed. The expected replacement cost per unit time is then calculated as a function of the control-limit level (cost function). Finally, the optimal control-limit level is found by minimizing the cost function.

Building the CBM model, as described above, may result in several model candidates. A few criteria to select the best model can be used. They include the model fit, the comparisons of the MTTF's (mean time to failure), MTBR's (mean time between replacements), and the cost performance evaluation. These criteria will be elaborated upon later in the paper.

Proportional hazards model with time-dependent covariates

A proportional hazards model with time-dependent covariates is defined by the hazard function

$$h(t, \mathbf{Z}(t)) = h_0(t)\exp(\gamma'\mathbf{Z}(t))$$

where $h_0(t)$ is the baseline hazard function, $\mathbf{Z}(t)$ is the vector of covariates at time t, and γ is the vector of coefficients. The baseline hazard function $h_0(t)$ can be either specified (parametric PHM) or unspecified (semi-parametric PHM). In semi-parametric PHM, the parameter γ is estimated by maximizing the partial likelihood that does not depend on $h_0(t)$. The baseline hazard function $h_0(t)$ can then be estimated using a non-parametric estimator of either the baseline cumulative hazard, or the baseline survival function (Therneau and Grambsch, 2000). In parametric PHM, the baseline hazard function $h_0(t)$ is specified as a parametric function of certain form, such as Weibull, lognormal, etc. The model parameters can be then estimated by maximizing the full likelihood. The proportional hazards model with

This content downloaded from 138.73.1.36 on Wed, 23 Dec 2015 08:49:05 UTC
All use subject to JSTOR Terms and Conditions

Weibull baseline hazard (Weibull PHM) is the most popular in reliability and maintenance, and has the hazard function

$$h(t, \mathbf{Z}(t)) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} \exp(\boldsymbol{\gamma}'\mathbf{Z}(t))$$

where β and η are the shape and scale parameters, respectively.

Assume that the sample data consists of n possibly right-censored histories denoted by $(T_i, \delta_i, (\mathbf{Z}_i(t), 0 \leqslant t \leqslant T_i))$, where δ_i is the censoring indicator, and T_i is the observed failure time if $\delta_i = 1$, or the censored time if $\delta_i = 0$. $(\mathbf{Z}_i(t), 0 \leqslant t \leqslant T_i)$ are the covariate readings for observed history i from the beginning to time T_i, $i = 1, 2, \ldots, n$. Then the likelihood function is

$$L(\beta, \eta, \boldsymbol{\gamma}) = \prod_{i=1}^{n} h(T_i, \mathbf{Z}_i(T_i))^{\delta_i} \exp(-H(T_i, \mathbf{Z}_i))$$

where $H(\cdot)$ denotes the cumulative hazard function, that is,

$$H(T_i, \mathbf{Z}_i) = \int_0^{T_i} h(u, \mathbf{Z}_i(u))\mathrm{d}u$$

By maximizing the likelihood function, the model parameters, β, η, $\boldsymbol{\gamma}$ can be estimated. Note that for the calculation of the likelihood function, the complete covariate realization $(\mathbf{Z}_i(t), 0 \leqslant t \leqslant T_i)$ should be known, at least in principle. In practice, however, it is not common or feasible to record covariate readings continuously. Instead, covariate readings are recorded at discrete times $0 \leqslant t_{i1} < t_{i2} < \cdots < t_{im_i} \leqslant T_i$. A simple approach for incorporating this historical covariate data into the likelihood function is to assume that the covariate function $\mathbf{Z}_i(t)$ is a stepwise constant function with jumps only at discrete times $t_{i1}, t_{i2}, \ldots, t_{im_i}$.

Principal component analysis

Denote by $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$ a vector of p random variables. The idea of PCA is to find a set of p uncorrelated linear combinations of the p random variables such that the kth linear combination has the kth largest variance among all possible (normalized to one) linear combinations, $k = 1, 2, \ldots, p$. We then expect that most variations of the p random variables can be described by the first $q(<p)$ linear combinations and hence that the number of original variables can be reduced without losing too much information. The kth linear combination is called the kth principal component.

If $\mathbf{x}$ has a known covariance matrix $\boldsymbol{\Sigma}$, the PCs can be calculated as $\boldsymbol{\alpha}_k'\mathbf{x}$, $k = 1, 2, \ldots, p$, where $\boldsymbol{\alpha}_k$ is the eigenvector of $\boldsymbol{\Sigma}$ corresponding to the kth largest eigenvalue. If $\boldsymbol{\Sigma}$ is unknown and sample data of size m from the population of $\mathbf{x}$ is available, the sample covariance matrix $\mathbf{S} = \mathbf{X}'\mathbf{X}/(m-1)$ is used instead of $\boldsymbol{\Sigma}$ for the calculation of PCs, where $\mathbf{X}$ is the centered observation matrix of dimension $m \times p$. A more efficient way of obtaining PCs from sample data is via singular value decomposition (SVD) of $\mathbf{X}/\sqrt{m-1}$, that is, from $\mathbf{X}/\sqrt{m-1} = \mathbf{UDV}'$, where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices, $\mathbf{D}$ is a diagonal matrix of the same dimension as $\mathbf{S}$ with non-negative diagonal elements in non-increasing order. In this case, $\mathbf{V}$ is the matrix of principal component coefficients and the squares of the diagonal elements of $\mathbf{D}$ are the eigenvalues of the sample covariance matrix $\mathbf{S}$.

The above derivation of PCs is based on the covariance matrix. A main drawback of this approach is that the PCs are sensitive to the magnitudes of the elements of $\mathbf{x}$. If there is a large difference in the variances of the elements of $\mathbf{x}$, the elements of $\mathbf{x}$ with the largest variances will tend to dominate the first few PCs. In this situation, PCA based on the correlation matrix rather than the covariance matrix is more appropriate. PCA based on the correlation matrix can be performed by standardizing the elements of $\mathbf{x}$. For a more detailed discussion on PCA, see the books by Jolliffe (1986) and Jackson (1991). This issue will be discussed further in later sections of this paper.

Principal components proportional hazards regression

Similar to principal components linear regression (PCLR), the idea of 'principal components proportional hazards regression' (PCPHR) is to use a set of PCs instead of the original variables as covariates (regressors) in the PHM. The objective of applying PCPHR is to eliminate possible collinearity among covariates for more accurate parameter estimation and to reduce the number of covariates included in the proportional hazards regression model.

A main issue in PCPHR is the selection of PCs as covariates. One way of selecting PCs is to select the first few PCs that account for the main part of variation in the original variables, or to delete those PCs with a small variance. In this way the number of covariates in the model can be reduced when the original variables are highly correlated. However, some PCs that account for a small amount of variation in the original variables may be important for predicting risk of system failure and may thus be significant covariates in the model. For a note on the similar problem in PCLR see the paper by Jolliffe (1982). Therefore, the procedure of PC selection based solely on variance may lead to the exclusion of significant PCs.

Another way of selecting PCs can be based on the P-values of the PCs in the model. This approach can prevent excluding significant PCs. However, PCs with very small variance may lead to inaccurate parameter estimation. A compromise is to exclude PCs with very small variance and at the same time exclude nonsignificant PCs based on the P-values. This approach is adopted in this paper.

In PCA, samples from the population of $\mathbf{x}$ are assumed to be independent. However, since the covariates in our model

This content downloaded from 138.73.1.36 on Wed, 23 Dec 2015 08:49:05 UTC
All use subject to JSTOR Terms and Conditions

are time dependent, samples taken at different times for the same unit may be dependent. However, the independency assumption is important only for the inference; it does not affect the numerical calculation. Since the objective of PCA in this paper is mainly descriptive, that is, to find a convenient transformation of the data, rather than inferential, the violation of the independence assumption can be ignored. Another problem is that we implicitly assume that the correlation structure of original covariates is preserved in time, so that the PCs have the same covariance structure over time, which may not be the case. In industrial applications, the covariates may have a low correlation during the 'normal' life of a unit (close to random noise), and show an increase with a much different correlation structure before an oncoming failure or inappropriate operation. One feasible way to get around this problem is to include in PCA only the condition data collected in the normal operating state. Alternatively, the dynamic PCA (Gao, 2003) could be applied to this situation, but it will not be discussed here.

Using the PCs instead of the original variables as covariates to build the PHM means applying a linear transformation to the original covariate data and building the PHM based on the transformed data. The likelihood function $L(\beta, \eta, \gamma)$ based on transformed data, when all PCs are included in the model, is simply obtained by replacing $(\mathbf{Z}_i(t), 0 \leqslant t \leqslant T_i)$ by its linear transformation. In that case, the maximum likelihood estimates of the parameters in the model based on the original data and on the transformed data should have the same relationship as the original and transformed data, if rounding errors are ignored. However, after starting the covariate selections in these two models, the likelihoods may not be directly comparable (because they will likely depend on different sets of the original variables) and then no direct relationship between the two estimates of parameters may exist. Nevertheless, PHM based on PC provides us with an alternative way of building a PHM and it can always be compared with the PHM based on the original covariates through the test model fit and/or the deviance change test. The method using PCs is powerful especially when there are high correlations between covariates and the number of covariates is greater than the sample size (number of histories, not number of measurements).

Analysis of a simulated data set

A PHM model with three uncorrelated covariates, Cov1, Cov2 and Cov3, was used to simulate a data set of 70 histories, of which 42 ended by failure, and 28 ended by suspension, with random censoring. Cov1 and Cov2 were used in the PHM as significant covariates, and Cov3 was not used as a significant covariate. Also, all three covariates were generated independently, given the value of the age variable.

The simulated data set is used to test the basic concept of PCPHR discussed previously. We will first apply PCPHR to the original data set with covariates *Cov*1, *Cov*2 and *Cov*3 to test the ability of PCA to reduce the number of variables in the model. Even if the number of original covariates is only three, and they have low pairwise correlations, which can be seen by the scatter plots in Figure 1, PCPHR will still provide a simpler model in this case. Then, we will test the ability of PCA to reduce the effect of collinearity on the model.

First we used the original covariates Cov1, Cov2 and Cov3 to build the model. As expected, Cov1 and Cov2 appeared as significant. This model is denoted by COV12. We then used the three PCs, created from standardized Cov1, Cov2 and Cov3, as covariates to build the PHM. It appeared that only the first PC is a significant covariate, and so it is included in the final model, denoted by PC1_COV123. We also applied PCA to standardized Cov1 and Cov2, the real significant covariates, and used the two PCs as covariates. Similarly, only the first PC is significant and included in the final PHM, which is denoted by PC1_COV12. The results are summarized in Table 1. For the purpose of comparison, we included the 'simple' Weibull model (denoted by SW) that does not include covariates, as well as the original model with all three covariates (denoted by COV123). Note that the PHM PC1_COV123 implicitly includes the nonsignificant covariate Cov3. However, the weight for Cov3 is comparatively small. Parameter estimates are actually very close for the four PH models.

We used Kolmogorov–Smirnov goodness-of-fit test (Koziol, 1980) to check the model fit for each model in Table 1. The results are summarized in Table 2. We see that all models fit the data very well and that the model fit using PCs is slightly better than the one including the original covariates. This is possibly due to the fact that PCs are uncorrelated (orthogonal).

We further developed the optimal replacement policy for each PHM candidate in Table 1. The optimal replacement policy is defined as one that minimizes the average total cost per unit working age of replacements (preventive and reactive maintenance) as discussed earlier. We used the costs of failure replacement and preventive replacement of $5000 and $1000, respectively (cost ratio 5:1). The inspection interval was set to 250 h (as it was used in simulation, on average). We compared the MTTF, MTBR and the cost analysis results for the five optimal replacement policies associated with the five models. The results are summarized in Table 3. The 'Expected cost per hour' is the theoretical average cost per hour corresponding to the optimal policy. The 'Average cost per hour when the optimal policy is applied' is the actual average cost that would have been obtained had the optimal decision policy been in force while operating the system. Since the SW model fits the data well, we can use the MTTF calculated from it as a reference. A good PHM candidate should not give an estimate of MTTF

This content downloaded from 138.73.1.36 on Wed, 23 Dec 2015 08:49:05 UTC
All use subject to JSTOR Terms and Conditions

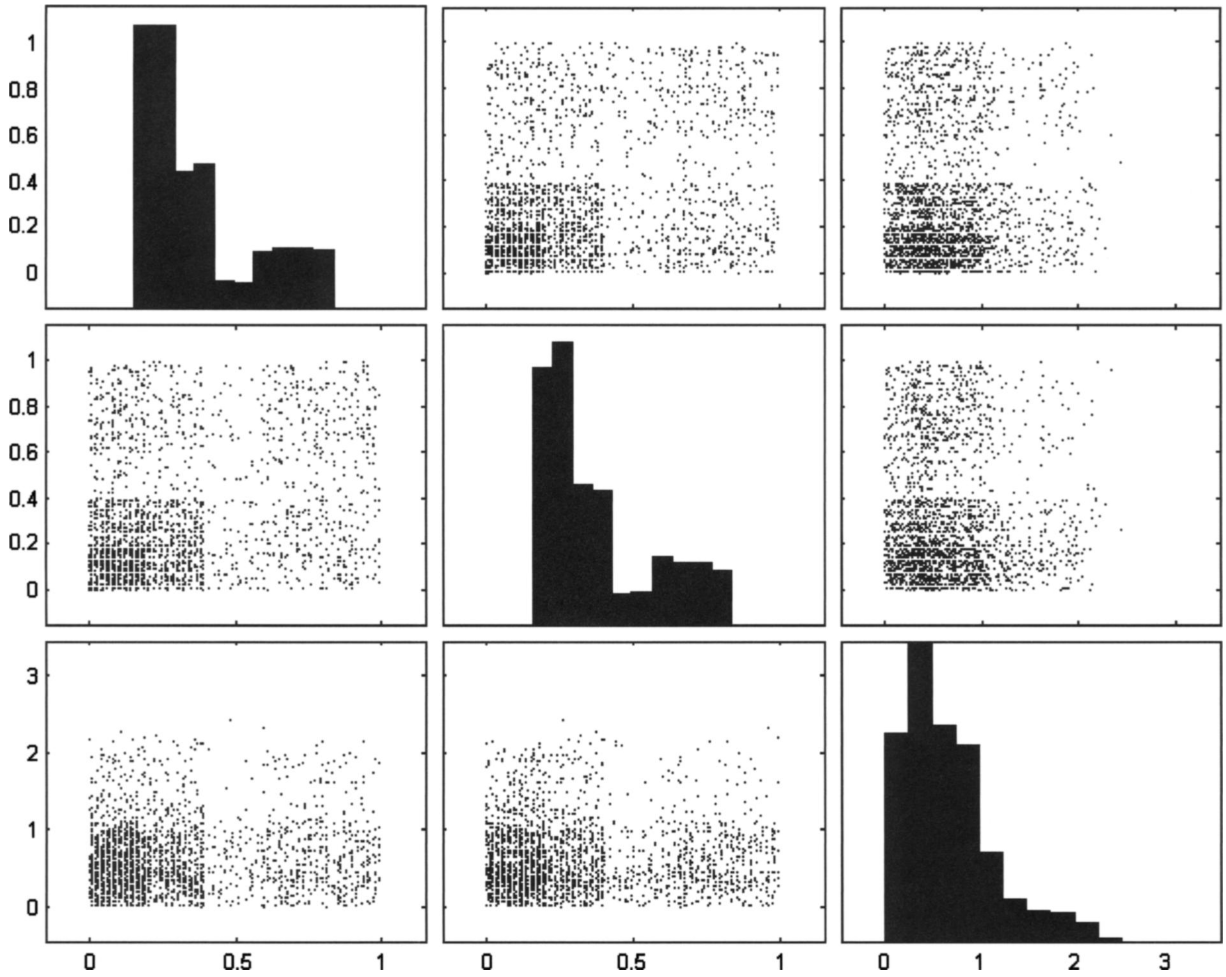

Figure 1 Scatter plot matrix of Cov1, Cov2 and Cov3.

Table 1 PHM candidates

Model	*Hazard function*
SW	$h(t) = \frac{3.9178}{11603}\left(\frac{t}{11603}\right)^{2.9178}$
COV12	$h(t) = \frac{3.6040}{17339}\left(\frac{t}{17339}\right)^{2.6040} \exp(1.1374\,\text{Cov1} + 1.6902\,\text{Cov2})$
COV123	$h(t) = \frac{3.6066}{17718}\left(\frac{t}{17718}\right)^{2.6066} \exp(1.1149\,\text{Cov1} + 1.6970\,\text{Cov2} + 0.1287\,\text{Cov3})$
PC1_COV123	$h(t) = \frac{3.6263}{13408}\left(\frac{t}{13408}\right)^{2.6263} \exp(0.5406\,\text{PC1\_COV123})$ $= \frac{3.6263}{17561}\left(\frac{t}{17561}\right)^{2.6263} \exp(1.3735\,\text{Cov1} + 1.3109\,\text{Cov2} + 0.2131\,\text{Cov3})$
PC1_COV12	$h(t) = \frac{3.6138}{13463}\left(\frac{t}{13463}\right)^{2.6138} \exp(0.5494\,\text{PC1\_COV12})$ $= \frac{3.6263}{17126}\left(\frac{t}{17126}\right)^{2.6138} \exp(1.3900\,\text{Cov1} + 1.31845\,\text{Cov2})$

SW, Simple Weibull; PCI_COV123 is the first PC obtained from standardized Cov1, Cov2 and Cov3; PC1_COV12 is the first PC obtained from standardized Cov1 and Cov2. The covariates in the hazard functions are functions of *t*, which is suppressed in the formulas.

This content downloaded from 138.73.1.36 on Wed, 23 Dec 2015 08:49:05 UTC
All use subject to JSTOR Terms and Conditions

Table 2 Kolmogrov–Smirnov goodness-of-fit test results

Model	*Statistics*	*P-value*
SW	0.0831	0.7008
COV12	0.0857	0.6642
COV123	0.0793	0.7541
PC1_COV123	0.0675	0.8971
PC1_COV12	0.0764	0.7927

Table 3 Model comparisons in MTTF, MTBR and costs for the simulated data

Model	*MTTF (h)*	*MTBR (h)*	*Expected cost per hour ($)*	*Average cost per hour when the optimal policy is applied ($)*
SW	10 504	6370	0.325	0.361
COV12	10 930	5963	0.222	0.231
COV123	11 119	6361	0.209	0.218
PC1_COV123	11 009	6045	0.218	0.227
PC1_COV12	11 041	6081	0.217	0.226

significantly different from one given by the SW model. We observe that the results for the four PHM candidates are close to each other. Their MTTFs are slightly greater than the one for the SW model. Their MTBRs are slightly smaller than the one for the SW model. However, their expected costs per hour are significantly lower than the one for the SW model. The costs per hour when the policy is applied are close to expected, particularly for the four PHMs.

We may conclude from the above comparisons that there is no significant difference among the four PH models in Table 1. Any of them can be used as a reasonable model to develop the optimal replacement policy. Of course, using the either of the last two, PC1–COV123 and PC1–COV12, would be simpler since each of them includes only one final covariate from the transformed data.

We want to consider more closely the problem of using the original or standardized covariates when applying the PCA. As mentioned earlier, the elements of **x** with largest variances will dominate the PCs, regardless of their influence on the hazard. This may result in a PHM missing some significant original covariate, or a model in which none of the PCs can be excluded. We applied PCA to original, non-standardized covariates Cov1, Cov2 and Cov3, and then used a similar procedure as before to build final PHMs. The results obtained are similar to the ones reported in Table 1. This may be explained by the same magnitude of standard deviations of the original covariates (the first two have almost the same value, and the third one is twice as large). The situation will be quite different with real data in the next section.

We also investigated the effect of collinearity on model selection and inclusion of PCs with very low variation by using the following example. Two additional covariates are included in the original simulated data:

$$\text{Cov4} = 50\text{Cov1} + 45\text{Cov2}, \quad \text{Cov5} = 5\text{Cov2} + 2\text{Cov3}$$

We applied backward selection (exclude the least significant covariate one at a time), starting with five covariates Cov1, Cov2, ..., Cov5, to obtain the final model. The final model includes only one covariate, Cov2, which is significant, but excludes the other significant covariate, Cov1. Even obtaining this model is a matter of 'chance', because the *P*-values of covariates excluded from the model as nonsignificant are very close to the *P*-values of non-excluded covariates, and all are close to one. The final model also depends on the order in which the covariates are included at the beginning of the selection, that is, the numerical procedure is sensitive to round-off errors. This shows that collinearity among covariates may result in a completely incorrect model. Obviously, this is an extreme case, because the covariates are linearly dependent. We also tried to add some random noise to Cov4 and Cov5. At lower noise levels (with a magnitude up to 10^{-3} of the covariate mean) the estimation procedure converged very slowly. A reasonable final model and much faster convergence were obtained when the noise level was increased to a magnitude of about 10^{-2} of the covariate mean. This was expected since the added random noise lowered the level of collinearity among the covariates. To test the PCPHR approach, we first applied PCA to the data with five covariates (standardized, as before). Theoretically, two of the PCs should be equal to zero and contribute nothing to the total variation. Their actual contribution, after calculation, was close to 0% in the total variation, but not exactly zero, due to round-off errors. It indicates that we should exclude these two PCs from the initial PHM. Starting with the first three PCs and applying backward selection, the final model ended up with only the first PC as a significant covariate. Theoretically, we cannot start with all five PCs in the model, because the likelihood would not depend at all on the 'gamma' parameter of two covariates that are equal to zero. In practice, the last two PCs were close to zero, but not exactly equal to zero (they behave as a random noise with very low variation). So the calculation can start, but produces unreliable and incorrect results, similar to those discussed above for the collinearity problem with five original covariates.

Obviously, in practical problems we do not expect to have exact linear dependence between original variables, but the collinearity problem does appear frequently, particularly in cases with a large number of variables. From this example, we may see that collinearity among covariates can easily lead to exclusion of significant covariates and thus to an incorrect model. We also see that PCA can help to eliminate collinearity among covariates, by excluding PCs with distinctively low variations.

This content downloaded from 138.73.1.36 on Wed, 23 Dec 2015 08:49:05 UTC
All use subject to JSTOR Terms and Conditions

Analysis of real data sets

In this section, we will apply the PCPHR to two real data sets: oil analysis data obtained from a mining company, and vibration analysis data obtained from a pulp and paper company.

Oil analysis data set from a mining company

The data set consists of event and diagnostic oil analysis data for transmissions on haul trucks. Event data include the information when a particular transmission (new or renewed) was installed on a truck, when transmission oil was replaced (event denoted here by OC), when the transmission failed, or was suspended due to preventive replacement. Owing to vendor specification, a transmission should be preventively replaced after 12 000 h of operation. Diagnostic data include spectrometric oil analysis of 20 metals obtained from regular inspections. Inspections are performed roughly every 600 h. Oil changes are performed roughly every 1200 h. There are altogether 51 histories: 20 failures, 16 suspensions (histories ended by preventive maintenance) and 15 temporary suspensions (transmissions still operating at the moment of data collection). The raw dataset contained some errors and missing OC events. After corrections, the dataset includes 342 OC records and 704 inspection records.

We first examined the inspection data by looking at various scatter plots and correlation coefficients. All covariates have from low to moderate correlations, with absolute values of correlation coefficients ranging from close to zero to less than 0.8 (see Figure 2 for some of the scatter plots). So, we do not expect to have PCs contributing a very low percentage of variation. Actually, all PCs were found to contribute more than 1% of the variation, except the last one, which is very close to 1%. By including all PCs in the model at the beginning and applying backward selection, we obtained the final model that includes seven PCs. We still wanted to simplify this model for further implementation in

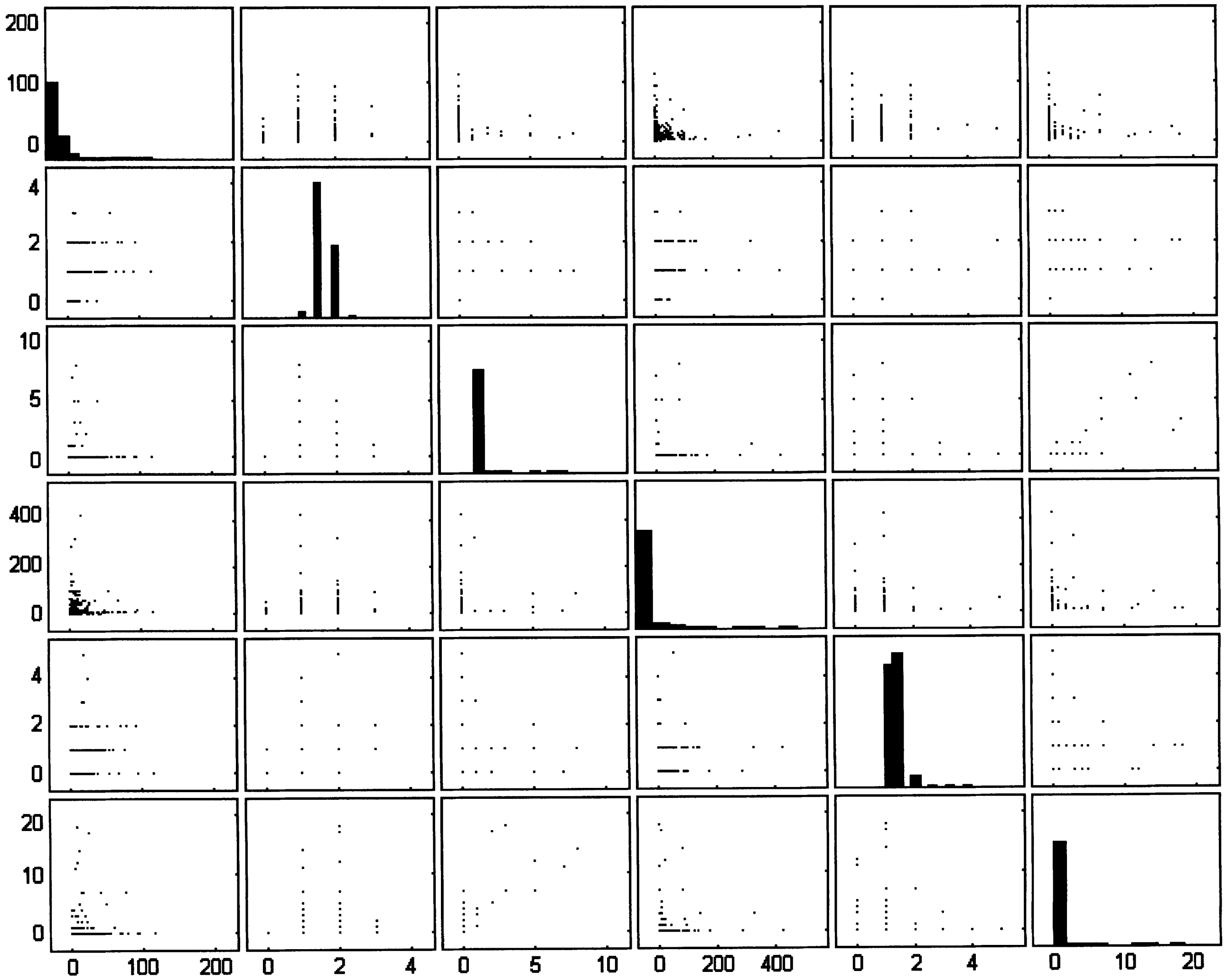

Figure 2 Scatter-plot matrix of the six significant covariates.

This content downloaded from 138.73.1.36 on Wed, 23 Dec 2015 08:49:05 UTC
All use subject to JSTOR Terms and Conditions

decision analysis. We tried to eliminate some original covariates that are not significant. We included all the 20 original metal readings in the model and applied backward selection again. After a few steps the model included 11 covariates: sodium (Na), potassium (K), iron (Fe), aluminium (Al), titanium (Ti), phosphorus (P), zinc (Zn), calcium (Ca), magnesium (Mg), molybdenum (Mo) and vanadium (V). We applied PCA to these 11 covariates and used the PCs to build the model again and obtained a model including four PCs.

We again considered the 11 original covariates. After more detailed analysis and backward selection, we reduced it to six significant covariates: Fe, Al, Ti, Mg, Mo and V. The scatter plot matrix of these six significant covariates is presented in Figure 2. Applying PCA to these six covariates (standardized), and using the PCs in the model, the final model included three PCs: PC2, PC3 and PC6 (model PC_236). The component loadings of these three PCs are summarized in Table 4. We tried to apply PCA to the six original (non-standardized) covariates and found that the final model included all six PCs. In this case, PCA on non-standardized covariates did not help to reduce dimension. This is likely due to the fact that the covariates have very different magnitudes and each PC is dominated mainly by one covariate, as discussed earlier in the introduction to PCA.

From Table 4, we see that the main contributors to PC6 are Ti and V, and their contribution is almost equal to the difference between these two metals (note that the covariates are standardized). It was also noted that PC6 has a much smaller contribution to the total variation, 3.25%, compared to more than 11% for other PCs. To test the importance of the PC with low variation in the model, we excluded PC6 and built the model with two covariates, PC2 and PC3 (model PC_23). For comparison, we also built the SW model. All three models (SW, PC_23 and PC_236) have satisfactory goodness-of-fit.

We developed the optimal replacement policies for these three models. The failure replacement cost and preventive replacement cost were estimated by the maintenance personnel as $5220 and $1560, respectively. The MTTF, MTBR and the costs for the optimal replacement policies associated with these three models are summarized in Table 5. Notice that the expected cost for the model PC_236 ($0.234) is significantly smaller than the average cost obtained when this policy is applied ($0.373). If we use the MTTF for SW model as a reference, we may conclude that the model PC_236 significantly overestimates the MTTF and thus significantly underestimates the expected cost. The MTTF for the model PC_23 is slightly smaller than the one for the SW model. The expected cost and the average cost obtained when the policy is applied are very close for the other two models SW and PC_23. These costs are significantly lower for the model PC_23 than for the SW model. This analysis shows that the PHM with covariates PC2 and PC3, and the corresponding optimal replacement policy, performs better than the other two models for the transmissions in this example.

Vibration analysis data set from a pulp and paper company

This data set contains event records and vibration measurements taken from water pumps every month. The pumps essentially work 24 h per day, 7 days per week. Working age was not provided in the data, so it was calculated in days since the beginning of the history. Vibration signals were taken at seven different locations. For each vibration signal, the overall amplitude and the amplitude for six different frequency bands were recorded. So, altogether there are 49 vibration variables recorded. The data set consists of 15 histories ended by failures and 18 histories ended by temporary suspensions. The inspection data includes 850 records of vibration measurements.

Table 4 Component loadings of the significant PCs

PC	*Fe*	*Al*	*Ti*	*Mg*	*Mo*	*V*
PC2	0.4140	0.5665	−0.2196	−0.0669	0.6579	−0.1486
PC3	−0.4171	0.3703	−0.0620	0.8238	−0.0111	−0.0792
PC6	−0.0464	−0.0522	−0.7022	0.0152	0.0012	0.7084

Table 5 Model comparison in MTTF, MTBR and costs for the oil analysis data

Model	*MTTF (h)*	*MTBR (h)*	*Expected cost per hour ($)*	*Average cost per hour when the optimal policy is applied ($)*
SW	14820	10415	0.430	0.446
PC_23	13764	10496	0.375	0.376
PC_236	18443	11285	0.234	0.373

This content downloaded from 138.73.1.36 on Wed, 23 Dec 2015 08:49:05 UTC
All use subject to JSTOR Terms and Conditions

Table 6 Model comparisons in MTTF, MTBR and costs for the vibration data

Model	*MTBF (days)*	*MTBR (days)*	*Expected cost per day ($)*	*Average cost per day when the optimal policy is applied ($)*
SW	1059	922	162.7	177.2
PC5	1268	920	122.5	154.5

The problem of this data set is that the number of covariates (49) is greater than the sample size (33). So, one cannot include all the covariates and use the backward selection to obtain the final PHM in this case. A solution to this problem is to reasonably reduce the number of covariates included in the model, or to try forward selection. One way to reduce the number of covariates is to apply PCA.

We did preliminary correlation analysis of the variables by viewing scatter plots and calculating correlation coefficients. Some of the variables are highly correlated, with correlation coefficients greater than 0.9. So we can expect to have some PCs whose percent contribution to variation is very low, and then the problem is to decide whether to include them in the model, as already discussed. However, in order to reduce the dimensionality, we decided to exclude them from the model, and, thus, to test the methodology of applying PCA in model building. We transformed the 49 original covariates (standardized) into 49 PCs. We applied backward selection to the first 17 PCs that each contribute at least 1% of the total variation. The final model included only one covariate, the fifth principal component (model PC_5). The SW model was also estimated. We found that both the SW model and PC_5 model fit the data.

We then calculated the optimal replacement policy for the SW model and for the PC_5 model using an estimate of $55 000 for the preventive replacement cost and $175 000 for the failure replacement cost. The results for MTTF, MTBR and the costs of the policies are summarized in Table 6. We observe that the MTTF for the model PC_5 is somewhat larger than for the model SW. Both the expected and the average cost using the model PC_5 seems significantly lower than the one using the SW model. From these observations, we accept that the model PC5 is a reasonably good model. We have to admit that calculating any kind of confidence intervals for the costs, which might be useful for comparison, would be extremely difficult. So we did not use this method in the paper.

Conclusion

This paper investigates the application of the principal components (PCs) proportional hazards regression to condition-based maintenance. We proposed to apply PCA to the original covariates in condition monitoring data and to use the PCs instead of the original covariates to build the PHM. The method may help in elimination of collinearity among original covariates and reduction of the number of covariates included in the PHM. We have demonstrated, by applying the methodology to a simulated data set, and two real data sets (oil analysis data set and the vibration analysis data set), that the PCs proportional hazards regression is able to produce a simpler model than one that would be obtained directly from the original data, and a reasonable replacement policy associated with the model. In particular, in a situation where the number of original covariates is greater than the sample size and some of the original covariates are highly correlated, this methodology is a very useful approach to the problem and obtains a reasonable solution.

Acknowledgements—We thank MMO (Materials and Manufacturing Ontario) of Canada for their financial support of this work and the industrial companies from North America for providing the data for this study.

References

Banjevic D, Jardine AKS, Makis V and Ennis M (2001). A control-limit policy and software for condition-based maintenance optimization. *INFOR* **39**: 32–50.

Chatterjee S, Hadi AS and Price B (2000). *Regression Analysis by Example*, 3rd edn. Wiley: New York.

Chevalier R, Garnero MA, Jardine AKS, Banjevic D and Montgomery N (2004). Optimizing CM data from EDF main rotating equipment using proportional hazard model. *Surveillance 5 Conference*. France, 11–13 October, 2004.

Cox DR (1972a). Regression models and life-tables (with discussion). *J Roy Stat Soc B* **34**: 187–220.

Cox DR (1972b). The statistical analysis of dependencies in point processes. In: Lewis PAW (ed). *Stochastic Point Processes*. Wiley: New York, pp 55–66.

Danielyan SA, Zharinov GM and Osipova TT (1986). Application of the principal components method and the proportional hazards regression model to analysis of survival data. *Biometrical J* **28**: 73–79.

Gao Y (2003). *Application of DPCA to oil data PH model building and comparison of optimal CBM policies*. MSc thesis, University of Toronto: Canada.

Jackson JE (1991). *A User's Guide to Principal Components*. Wiley: New York.

Jardine AKS, Banjevic D, Khan K, Wiseman M and Lin D (2003). An optimized policy for the interpretation of inspection data from a CBM program at a nuclear reactor station. *COMADEM 2003*. Vaxjo, Sweden, 27–29 August, 2003.

This content downloaded from 138.73.1.36 on Wed, 23 Dec 2015 08:49:05 UTC
All use subject to JSTOR Terms and Conditions

Jardine AKS, Banjevic D, Wiseman M and Buck S (2001). Optimizing a mine haul truck wheel motors' condition monitoring program: use of proportional hazards modeling. *J Qual Maint Eng* **7**: 286–301.

Jardine AKS, Joseph T and Banjevic D (1999). Optimizing condition-based maintenance decisions for equipment subject to vibration monitoring. *J Qual Maint Eng* **5**: 192–202.

Jeffris T, Banjevic D, Jardine AKS and Montgomery N (2004). Oil analysis of marine diesel engines: optimizing condition-based maintenance decisions. *COMADEM 2004*. Robinson College, Cambridge, England, 23–25 August, 2004.

Jolliffe IT (1982). A note on the use of principal components in regression. *Appl Stat* **31**: 300–303.

Jolliffe IT (1986). *Principal Component Analysis*. Springer-Verlag: New York.

Koziol JA (1980). Goodness-of-fit tests for randomly censored data. *Biometrika* **67**: 693–696.

Li L and Li H (2004). *Dimension reduction methods for microarrays with application to censored survival data.* Posted at the Scholarship Repository, University of California.

Lin D, Wiseman M, Banjevic D and Jardine AKS (2004). An approach to signal processing and condition-based maintenance for gearboxes subject to tooth failure. *Mech Systems and Signal Process* **18**: 993–1007.

Makis V and Jardine AKS (1992). Optimal replacement in the proportional hazards model. *INFOR* **30**: 172–183.

Monnot M *et al* (2004). Smartly interpreting oil analysis results from the hydraulic system of backhoes. *STLE Confrrence*. Toronto, Canada, 17–19 May, 2004.

Moubray J (1997). *Reliability-Centred Maintenance*, 2nd edn. ButterworthHeinemann: Oxford, Boston.

Therneau TM and Grambsch PM (2000). *Modeling Survival Data. Extending the Cox Model.* Springer: New York.

Vlok PJ, Coetzee JL, Banjevic D and Jardine AKS (2002). Optimal component replacement decisions using vibration monitoring and the proportional hazards model. *J Opl Res Soc* **53**: 193–202.

Williams JH, Davies A and Drake PR (1994). *Condition-Based Maintenance and Machine Diagnostics.* Chapman & Hall: London.

Received November 2004;
accepted June 2005 after one revision

This content downloaded from 138.73.1.36 on Wed, 23 Dec 2015 08:49:05 UTC
All use subject to JSTOR Terms and Conditions