Multi-task Learning for Joint Prediction of the Remaining Useful Life and Failure Type of Train Wheelsets

by
Weixin Wang
May 20, 2015

A thesis submitted to the
Faculty of the Graduate School of the
State University of New York at Buffalo
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Industrial and Systems Engineering
University at Buffalo, SUNY, USA
ACKNOWLEDGEMENT

I would like to give my sincere gratitude to my advisor, Dr. Qing He, for your invaluable guidance and help, without which I would not have been able to accomplish this.

Also a special thank you to my parents. Their constant support and encouragement is always my strength.
# TABLE OF CONTENTS

LIST OF FIGURES ........................................................................................................................................... 5
LIST OF TABLES ............................................................................................................................................... 6
ABSTRACT ....................................................................................................................................................... 7

CHAPTER 1 INTRODUCTION .......................................................................................................................... 8
  1.1 Motivation .................................................................................................................................................. 8
  1.2 State-of-practice in railcar monitoring and maintenance ........................................................................ 9
  1.3 Problem Statement and Research Objectives ....................................................................................... 12
  1.4 Thesis Organizations ............................................................................................................................ 13

CHAPTER 2 LITERATURE REVIEW .............................................................................................................. 15
  2.1 Railway predictive maintenance ........................................................................................................... 15
    2.1.1 Condition-based monitoring (CBM) ............................................................................................. 15
    2.1.2 Predictive maintenance ................................................................................................................ 15
    2.1.3 Train wheel failure prediction ..................................................................................................... 17
  2.2 Methodology in Failure Prediction ...................................................................................................... 19
    2.2.1 RUL ............................................................................................................................................... 19
    2.2.2 Multi-task multi-modal learning ................................................................................................. 21

CHAPTER 3 DATA DESCRIPTION and PREPARATION ............................................................................... 23
  3.1 Data presentation .................................................................................................................................. 23
    3.1.1 Maintenance Data ......................................................................................................................... 23
    3.2.3 WILD Data .................................................................................................................................... 26
    3.1.4 Data extraction and cleaning ........................................................................................................ 28
  3.2 Data analysis under different tasks ..................................................................................................... 28
    3.2.1 RUL analysis ............................................................................................................................... 28
    3.2.2 Competing risks analysis for failure type .................................................................................... 29

CHAPTER 4 METHODOLOGY and ALGORITHM ......................................................................................... 32
  4.1 Multi-task feature selection (MTFS) ..................................................................................................... 33
  4.2 Algorithm .............................................................................................................................................. 36
    4.2.1 Gradient descent algorithm ......................................................................................................... 36
    4.2.2 Objective function in matrix format ........................................................................................... 37
    4.2.3 Proof of convexity ....................................................................................................................... 37
CHAPTER 5 EXPERIMENT RESULTS ................................................................. 39
  5.1 MTFS Result .................................................................................. 39
    5.1.1 MTFS experiment procedure ..................................................... 39
  5.2 SVM Result .................................................................................... 43
    5.2.1 Classification with SVM .............................................................. 44
    5.2.2 Regression with SVR ................................................................. 45
    5.2.3 Cross-validation results ............................................................. 46
CHAPTER 6 CONCLUSIONS and FUTURE WORK ........................................... 48
  6.1 Conclusions .................................................................................. 48
  6.2 Future work ................................................................................... 49
References ............................................................................................ 50
LIST OF FIGURES

FIGURE 1 DERAILMENT TRAIN
FIGURE 2 WHEEL CAUSED DERAILMENT .......... 9
FIGURE 3 ROLLING CONTACT FATIGUE
FIGURE 4 OUT OF ROUNDNESS .................... 10
FIGURE 5 RE-PROFILE WHEELSETS
FIGURE 6 WHEEL LATHE ........................... 12
FIGURE 7 PREVENTIVE MAINTENANCE APPROACH .................................................. 16
FIGURE 8 KEY STEPS OF CBM .......................................................... 16
FIGURE 9 STATISTICAL DATA-DRIVEN APPROACHES FOR RUL ESTIMATION .......................... 20
FIGURE 10 AXLE AND TRUCK ARRANGEMENT IN AN EQUIPMENT ........................................ 24
FIGURE 11 BOXCAR
FIGURE 12 TANK ........................................ 25
FIGURE 13 FLAT
FIGURE 14 GONDOLA .................................. 25
FIGURE 15 COMPETING RISK FOR DIFFERENT CAR KIND
WHEEL SIZE .................................................. 29
FIGURE 17 COMPETING RISK FOR CAR TYPE AND WHEEL SIZE ............................................. 31
FIGURE 18 PROCEDURE OF METHODOLOGY .................................................. 32
FIGURE 19 DIAGRAM FOR L2/L1 REGULARIZATION .................................................. 35
FIGURE 20 GRADIENT DESCENT ALGORITHM .................................................. 37
FIGURE 21 PROCEDURE OF MTFS .................................................. 39
FIGURE 22 OBJECTIVE VALUE CONVERGE CURVE WHEN $\lambda = 1$
FIGURE 23 OBJECTIVE VALUE CONVERGE CURVE WHEN $\lambda = 10$ ................................................. 40
FIGURE 24 RELATION BETWEEN SELECTED FEATURE NUMBER AND $\lambda$ .............................................. 41
FIGURE 25 ROC CURVE .................................................. 45
FIGURE 26 APE FOR CAR KIND&WHEEL SIZE
FIGURE 27 APE FOR REPAIR TYPE ............. 46
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>Maintenance Data Presentation</td>
<td>23</td>
</tr>
<tr>
<td>Table 2</td>
<td>Mileage Data</td>
<td>26</td>
</tr>
<tr>
<td>Table 3</td>
<td>Wild Data and Data Connections</td>
<td>27</td>
</tr>
<tr>
<td>Table 4</td>
<td>Statistical Summary of Rule</td>
<td>28</td>
</tr>
<tr>
<td>Table 5</td>
<td>Part of Experiment Results</td>
<td>41</td>
</tr>
<tr>
<td>Table 6</td>
<td>Selected Features and Descriptions</td>
<td>42</td>
</tr>
<tr>
<td>Table 7</td>
<td>Confusion Matrix Result for One Trial</td>
<td>44</td>
</tr>
<tr>
<td>Table 8</td>
<td>Comparing the Cross-Validation Result Between Single Task Learning and Multi-Task Learning</td>
<td>47</td>
</tr>
</tbody>
</table>
ABSTRACT

The failures of train wheels account for half of all train derailments. Both remaining useful life (RUL) and failure types of wheels are critical for wheel maintenance. RUL prediction is a regression task, whereas failure type is a classification task. Most existing approaches to multi-task learning usually deal with homogeneous tasks, such as purely regression tasks, or entirely classification tasks, thus cannot utilize the intrinsic useful correlation information among different variables. In this paper, we propose a general methodology to jointly predict two tasks by using a common input space to achieve more desirable results. We formulate this problem as a convex optimization, a combination of linear regressions and logistic regressions, and model the joint sparsity as L2/L1 norm of the model parameters in order to couple feature selection across tasks. In our experiments, we use real-world data from a Class I railroad in North America. Competing risk analysis is also applied to failure time data, in order to compare failure probability of different failure types for different car kinds and wheel sizes. To validate our method, we perform support vector regression to predict RUL and support vector machine to predict failure type. And we show that our method outperforms the single task learning method.

Keywords: Wheelsets; Multi-task learning; Support vector machine (SVM); Remaining useful life (RUL); Failure type
CHAPTER 1 INTRODUCTION

1.1 Motivation

Wheelsets are the top rolling stock maintenance item in North America. Transportation Technology Center, Inc (TTCI) gave an estimation that wheelsets replacement costs $828 million annually, while the number is still increasing (Cummings 2012). Huge money is spent on wheelsets since it is the most maintenance intensive part of the railway system and is the most vulnerable if maintenance is neglected. For example, the worst rail disaster (see fig 1) in the history of Germany, occurred on 3 June 1998, caused 101 death and around 100 people injured at Eschede. In the later investigation, the incident was caused by a steel band coming loose from a wheel (see fig 2), which seemed hardly foreseeable before accident (Brumsen 2011).

Yet a real question remains: Is the derailment really unpredictable? From a maintenance report generated by the train's on-board computer, that two months prior to the Eschede disaster, conductors and other train staff filed eight separate complaints about the noises and vibrations generated from the bogie with the defective wheel; the company did not replace the wheel. Deutsche Bahn (Railway Company) claimed that its inspections were proper at the time and that the engineers could not have predicted the wheel fracture (RAILWAY TECHNICAL WEB PAGES). Why these inspection results could not convince the company to change the wheel? Or simply because it was too expensive to stop the high-speed train and replace a new wheel? We cannot tell after almost twenty years. But we can learn that no matter how much we emphasize on the importance of maintenance, it is not superfluous. Since it not only impacts system performances, costs and quality of service, but
also protects human lives.

1.2 State-of-practice in railcar monitoring and maintenance

Rolling stock maintenance can be programmed in one of the three ways: by time, by mileage or by condition-based monitoring (CBM). Of these three methods, condition monitoring is the most recent. Our research is based on CBM, and we will introduce it in detail in section 2.1.

Traditionally, maintenance was carried out on the time basis, usually related to safety items like braking and wheel condition. Many rail companies later adopted mileage-based system, although this is more difficult to operate as you have to keep records of all vehicle mileages and this is time consuming unless you have a modern train control and data gathering system. Modern trains should be able to run for some weeks without a maintenance inspection. And there are often special rules for high-speed trains and for heavy freight.

When we aim to prevent from derailment, the worst failure, we actually expect to rectify defects in every level, from rolling contact fatigue (fig 3) to out-of-roundness of wheel (fig 4). In real-world operations, there are many causes a rail wheel can leave its ideal shape then begin to generate problems, such as high impact loads and potential accidents (RAILWAY
The roundness of wheel is very critical to rail safety, especially for high-speed rail. Take the Eschede disaster as example again: The wheel was alleged to have had an eccentricity (the difference between major and minor axes of the ellipse) of 1.1mm, against a limit of 0.6mm.

To avoid a disaster like this, rail companies have to regularly inspect wheelsets. Current railcar inspection practice requires a car inspector to walk or ride a vehicle along the entire length of a train, visually inspecting the mechanical components on each car. Several tools of the trade include a hammer to “sound out” cracked or broken wheels, a piece of chalk or other marking devices to identify cars with possible defects, and a lantern for inspecting trains at night (Schlake 2010).

To improve the effectiveness of railcar condition monitoring and take advantage of CBM, the Association of American Railroads (AAR) began to promote the development and implementation of Automated Condition Monitoring Technologies (ACMT) and Technology Driven Train Inspection (TDTI) programs. ACMT consists of durable wayside systems located adjacent to or directly beneath the railroad tracks. Various sensing mechanisms have been designed to collect data from passing trains in the form of force, temperature, audio, or...
visual-based measurements. The ultimate goal of TDTI is the implementation of a network of inspection sites where each component of the train will be inspected automatically in real time and critical health information for each car will be documented and made available to railroad mechanical department personnel (Schlake 2010).

CBM is achieved by checking the operation of the equipment and only changing something if it shows signs of wear beyond preset limits. The checking is often done using on-board monitoring and storing the data gathered in a computer for downloading at the maintenance facility. Of course, it is a recent development made available by the introduction of information technology on trains. Such systems are now becoming so sophisticated that it is possible to have failure predictions of some items of equipment. A combination of on-board data gathering and depot maintenance systems have been developed into complete maintenance management systems on lines where modern rolling stock has been introduced (RAILWAY TECHNICAL WEB PAGES).

Reprofiling wheels is a slow and expensive process. Removing the wheels requires the train to be lifted and this is an expensive operation and very time-consuming. During maintenance, they remove wheels from a train by an equipment called wheel-drop. Wheel lathe (Fig 5-6) is another type of widely used facility to profile defected wheels. These are normally designed so that the wheels can be reprofiled while still on the train. Modern wheel lathes can also reprofile a wheelset that has been removed from the train. Some modern wheels lathes are designed to turn both wheelsets on a bogie at the same time.

The design and maintenance of train wheelsets have been improved considerably over recent years, reducing the periods between visits for reprofiling. Even so, there are still
persistent cases of railways running into unforeseen or unusual wheel wear problems and the wheel/rail interface still needs a lot more research before it is fully understood (RAILWAY TECHNICAL WEB PAGES).

![Figure 5 Re-profile wheelsets](image)

![Figure 6 Wheel lathe](image)

**Figure 5** Re-profile wheelsets  **Figure 6** Wheel lathe

1.3 problem Statement and Research Objectives

Although wheelset maintenance is expensive, it will be more expensive to deal with In-service-failure (ISF) such as derailments if maintenance has been neglected. Moreover, some researchers proved that re-profiling wheelset after certain mileage would nearly double the wheel service life thus minimizing total life cycle costs (Braghin, Lewis et al. 2006). Therefore, it is critical to know the remaining time before the occurrence of a failure given the current and historical wheel conditions. We define a random variable:

\[ Z(t) = T - t \mid T > t \]

where \( T \) denotes the random variable of time-to-failure, and suppose the wheel has survived until time \( t \). In the literature of reliability engineering, \( T \) is also known as remaining useful life (RUL).

In addition to RUL, we can also predict the failure type, which will typically result in two different repair actions: replacement with a new wheelset (N), or wheelset turning (T) failure.
type can benefit preventive maintenance, reduce maintenance cost, and decrease derailment risk. To better define the problem, we add a class attribute $C_i$ to each wheelset $i$ using following equation:

$$C_i(t) = \begin{cases} 0 & \text{when } t = T \\ 1 & \text{when } t = \text{N} \end{cases}$$

where $T$ and $t$ represent RUL and given time, respectively. 0 denotes that the wheelset only needs to be turned or reprofiled (T) and 1 the wheelset has to be replaced (N).

The objective of this thesis is to develop a prognostic machine learning model to concurrently predict RUL and failure type of wheelsets in a freight car with readings from wayside detectors. Competing risk models will be employed to analyze the failure probability under different failure types and different car kinds and wheel size. Multi-task learning method will be applied for common feature selection, followed by prediction with Support Vector Regression (SVR) and Support Vector Machine (SVM). The model can assist railway companies in making decisions about rail vehicle operation and maintenance, by providing further information on the cause of wheel defects and degradation.

This paper only considers wheelset failure prediction. Existing maintenance practice always replaces or re-profile both wheels in one wheelsets once a failure detected, since a wheel degenerates rapidly once the other wheel is defected.

1.4 Thesis Organizations

The remaining of the paper is organized as follows. Chapter 2 briefly reviews the predictive maintenance, railway failure prediction and multi-task learning. Chapter 3 discusses the data collection and initial analysis. Chapter 4 focuses on feature selection with multi-task learning. A new methodology combining least square loss and negative maximum likelihood
of logistic regression is proposed and solved by gradient decent method. Following methodology, chapter 5 gives experiment result based on data given on chapter 2. Support vector classification and support vector regression are both involved in prediction. A conclusion of proposed methodology and feature work is given in chapter.
CHAPTER 2 LITERATURE REVIEW

2.1 Railway predictive maintenance

2.1.1 Condition-based monitoring (CBM)

CBM aims to record the current (real-time) condition of a system (Li and Goodall 2004). CBM has been applied and studied for a long time. The previous studies generally fall into two categories: measurement analysis and model-based approach.

Measurement analysis traditionally relied on signal processing and knowledge-based techniques, e.g. statistical limit value checking, Power-spectral-density (PSD) analysis, correlation analysis, etc. (Sunder, Kolbasseff et al. 2001). Palo stated that condition monitoring of railway vehicles is mainly performed using wheel impact load detectors (WILD) and truck performance detectors (TPD). In addition, based on the analysis of such measurements, it is found that differences for various wheel positions determined the robustness of the measurement system (Palo, Schunnesson et al. 2012).

Model-based approach is well-developed recently. Li et al. derived Rao–Blackwellized particle filter (RBPF) based method, applied to parameter estimation for CBM, and tested it by railway data (Li, Goodall et al. 2007). Li proposed a Kalman filter-based innovation method to detect and isolated faults in railway vehicle suspension system based on the derived vehicle dynamic model (Li and Goodall 2004).

2.1.2 Predictive maintenance

The earliest maintenance technique is basically breakdown maintenance (Jardine, Lin et al. 2006). Maintenance can be categorized into corrective maintenance and preventive maintenance, shown in Figure 7. With experience and increasing knowledge of technical
systems, maintenance activities have evolved towards a more preventive approach based on time intervals (Lagnebäck 2007). The preventive maintenance approach can be divided into two main groups, predetermined maintenance and condition-based maintenance (CBM) (Niu and Pecht 2009).

Since our research is based on CBM, we only give a brief review about it. Jardine summarized three key steps of CBM as follows:

He also pointed out that diagnostics and prognostics are two important aspects in a CBM program (Jardine, Lin et al. 2006). An insightful study in 20 years ago had revealed that implement of CBM would revealed a potential saving of 1.5 billion £ per annum to railway
field and suggested that a modest 5% increase in machine availability could secure 30% profitability improvement (Fararooy and Allan 1995).

The most related research, published by Marcus, elaborated possibility to apply CBM system to railway. He introduced main layers in the on-rail-vehicle CBM system but also pointed out three major difficulties, which is false alarms, maintenance planning and running-in priority (Marcus 2002).

2.1.3 Train wheel failure prediction

As we discussed in the first chapter, wheel failures cost significantly to railway operations. Even worse, it may contribute irrevocable loss for both companies and human beings. To avoid those consequences, there is a pressing need to predict wheel failure accurately.

2.1.3.1 Identification of failure events

The first question to be answered is to how to define a failure event. Yang declared a failure event when a wheel vertical impact measurement exceeds 140 kips or a computed value (noted WildCal) greater than 170 kips (Yang and Létourneau 2005). Li and He used bad order data directly, consisting of order dates and types. Once a bad order is issued, the car is scheduled to be checked or repaired in the workshop (Li and He). Hajibabai et al. categorized an instance of train stop as failure; otherwise, with no train stop nor repair record, it was categorized as a non-failed wheel (Hajibabaia, Saat et al. 2012). Stratman stated that Current AAR’s wheel impact load limit for wayside detectors, which is 90,000 pounds, is no longer reliable, since many catastrophic wheel failures occurred below the 90,000 pound (Stratman, Liu et al. 2007). Therefore, it would be unreliable if we only consider failure identification with pre-determined thresholds, which varies from case to
case, depending on many factors such as detector system, experience and research perspective.

2.3.1.2 Failure prediction

Failure prediction is the next step of failure identification. In failure prediction, there are two major methods: (1) develop criteria & performance and defined indices and pertinent thresholds for wheel failure, and (2) statistical learning and data mining based methods (Li and He).

Instead of using present thresholds, Stratman proposed two efficient criteria, based on structural health monitoring trends using WILD detectors, for removal of wheels with a high likelihood of failure, which can remove 15.8% about-to-fail wheels (Stratman, Liu et al. 2007). However, these criteria both need at least three days consecutive increasing impact readings, which subject to the location of detectors and the route of the railcar.

Yang proposed a Multiple Classifier System (MCS) capable of predicting 97% of wheel failures while maintaining a reasonable false alert rate (8%) (Yang and Létourneau 2005). The model only gave the probability of failure but cannot tell the occurrence time of the failure.

Li and He developed a Random Forests based methodology to assess the current health and predict RUL of both trucks (bogies) and wheels of a railcar. They also compared the efficiency of three types of detectors (Li and He). But they didn’t consider the failure types and the following repair actions. Actually the costs of reprofiling a wheel and replacing a wheel are totally different.

Hajibabaia et al. developed a logistic regression model to classify wheel failures with classification accuracy of 90% and 10% false alarm rate (Hajibabaia, Saat et al. 2012).
However, they only take into account wheels within 30 days of train stop as “bad wheels”, which means the model cannot predict the failure times which are greater than 30 days.

A.EKBERG et al. defined three indices to quantify fatigue impact. They also set corresponding pertinent thresholds. If one or more inequalities are fulfilled, fatigue is predicted to occur. They noticed that all the conservative approximation occurred at the same lateral position. A more precise estimation can be made by studying the lateral spread of fatigue impact from a given load (Ekberg, Kabo et al. 2002).

Liu developed a multiaxial high-cycle fatigue initiation life prediction model, based on the critical plane approach, for railroad wheels. Both the initiation crack plane orientation and fatigue initiation life can be predicted based on the proposed model (Liu, Stratman et al. 2006). But interactive effects of involved parameters, such as wheel diameter, vertical loads and hardness of material, were not considered while the wheel contact problem is highly nonlinear.

2.2 Methodology in Failure Prediction

In this section, we will review the previous methodologies related to our prediction model, which includes RUL and multi-task learning.

2.2.1 RUL

Remaining useful life (RUL), also called remaining service life, residual life or remnant life, is nowadays in fashion, both in theory and applications (Banjevic 2009). The concept of the RUL is upon individual explanations, regarding the word “useful” (Si, Wang et al. 2011). RUL refers to the time left before observing a failure given the current machine age and condition, and the past operation profile (Ahmadzadeh and Lundberg 2014). Prognostics is also called the
prediction of an asset’s lifetime as its objective is to predict the RUL (Jardine, Lin et al. 2006). Mean residual life (MRL) was a popular concept once a time. But due to the large variation of RUL proved by Banjevic, it is may not be very useful to estimate RUL (Banjevic 2009). Some studies published a complete and detailed review about statistical data-driven approaches for RUL estimation models, which are classified into two broad types of models, that is, the RUL estimation models based on the directly observed state processes, and those cannot be observed directly (Si, Wang et al. 2011). There are several sub-branches of each classification given as follow.

![Figure 9 Statistical Data-driven Approaches for RUL Estimation [REF]](image)

In this paper, the main objective is to predict RUL of wheelsets under condition-based monitoring. RUL estimation is one of the key factors in condition-based maintenance (Wang 2007). RUL estimation can be used to provide decision support for maintenance actions (Jardine, Lin et al. 2006). Definition of failure is crucial to the interpretation of RUL (Cheng and Pecht 2009). However, it also varies from systems and railroads. In this paper, we assume that the definition of the failure is known to the railroad and we will present some cases of rail industry to show different understandings of failure in the next
section.

2.2.2 Multi-task multi-modal learning

Different from the conventional single-task feature selection, the multi-task feature selection simultaneously selects a common feature subset relevant to all tasks (Zhang, Shen et al. 2012). The very early work about multi task learning is Caruana in 1997. They gave a framework of neural networks with one or more hidden layers, which are trained for each task and they all share the same hidden weights (Caruana 1997).

Typical approaches to information transfer among tasks include (Xue, Liao et al. 2007): sharing hidden nodes in neural networks (Caruana 1997); placing a common prior in hierarchical Bayesian (Ando and Zhang 2005); sharing parameters of Gaussian processes (Lawrence and Platt 2004) and structured regularization in kernel methods (Evgeniou et al., 2005), among others (Xue, Liao et al. 2007).

Our algorithm shares some similarities with recent work in Yang (Yang, Kim et al. 2009) where they deal with heterogeneous tasks including both continuous and discrete outputs from a common set of input variables. Two main differences are that their formulation uses L infinite regularization and that, in our formulation, L2/L1 norm is applied. And second order derivative method like interior-point method is adopted, however, we choose gradient method because of number of tasks and data volume.

The idea of taking L2/L1 norm is after carefully consideration of previous work. Schmidt illustrated detail derivation process of adding L2 norm and an extra parameter \( \lambda \) (Schmidt 2005). However, the current trend of replacing L2 by L1 norm is also proposed and so is the reason. Furthermore, a combination of L1 and L2 norm seems to be a better approach to
combine tasks and ensure that common features will be selected across the (Evgeniou and Pontil 2007). It’s well-know that using L1 norm leads to sparse solutions, which is perfect for us. Since we want to some components of learned vector to be zero then we can select those feature which matter. Note that this L1/L2 regularization scheme reduces to the L1 regularization in the single-task case, and can thus be seen an extension of it where instead of summing the absolute values of coefficients associated to features we sum the Euclidean norms of coefficient blocks. The L2-norm is just used here as a measure of magnitude and one could also use Lp-norms(Obozinski, Taskar et al. 2006).

We also inspired by implement multi-task learning in medical area. Zhang and Shen generated a methodology can effectively estimate the MMSE and ADAS-Cog scores and the classification label(Zhang, Shen et al. 2012). It made the first investigation on jointly predicting multiple regression and classification variables from the baseline multi-modal data. This work is based on Zhang’s previous work of classification ADS(Zhang, Wang et al. 2011). Even it only includes multi task trough pure classification, it still gives us some clue about multimodal prediction. In addition, a paper working on HIV derive a solution that produces resampling weights which match the pool of all examples to the target distribution of any given task. It’s motivated by the problem of predicting the outcome of a therapy attempt for a patient who carries an HIV virus with a set of observed genetic properties. Multi-task learning enables them to make predictions even for drug combinations with few or no training examples and substantially improves the overall prediction accuracy(Bickel, Bogojeska et al. 2008).
We collected our data from one of Class I railroads in North America. The entire datasets include: maintenance data, bad order data, mileage data and WILD data. Independent presentations will be provided to summary of data fields, data cleaning and processing of every dataset. Also, a statistical summary of RUL is included and competing risks analysis is applied to each failure type.

### 3.1 Data presentation

#### 3.1.1 Maintenance Data

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQP_ID</td>
<td>Faulty equipment identification number</td>
</tr>
<tr>
<td>BAD_CD</td>
<td>Bad order code for the reason of maintenance</td>
</tr>
<tr>
<td>APL_CD</td>
<td>Applied job code</td>
</tr>
<tr>
<td>BAD_DT</td>
<td>Date of receiving bad order</td>
</tr>
<tr>
<td>RPR_DT</td>
<td>Date of wheelset repair/maintenance</td>
</tr>
<tr>
<td>RPR_LOC</td>
<td>Repaired axle position on the car</td>
</tr>
<tr>
<td>SIZE</td>
<td>Wheel size</td>
</tr>
<tr>
<td>CAR_KIND</td>
<td>Car type.</td>
</tr>
<tr>
<td>FIR_PRP</td>
<td>Last repair type. N=New wheelsets, T=Turned wheelsets</td>
</tr>
</tbody>
</table>

Maintenance data was collected from railcar workshops. Once a failure is diagnosed by a bad order, the faulty equipment will be scheduled to visit workshop depending on the severity level of failure. There are dozens of reasons that may contribute a bad order, such like thin brakes and warm bearings. In this paper, only wheel bad orders are considered. For wheel repair, technicians usually replace or turn the defected wheel in pairs, depending on damage intensity. The corresponding feature FIR_PRP contains two categories: New (N) and Turn (T). New, simply means a new wheeset replacement, whereas turn involves different repair
behavior such as: re-profile a wheel or polish a wheel. Repair axle number is indicated by RPR_LOC and illustrated as Figure 10. CAR_KIND uses the first letter of different car type. G represents gondola. H represents hopper. M represents miscellaneous, which includes tank, boxcar and so on. Different car types have different designs in terms of loading and shipping purposes. Fig 11-14 present the major car types.

Figure 10 Axle and truck arrangement in an equipment
3.1.2 Mileage Data
Mileage data was derived from revenue mile data. The recording mileage is daily basis, which needs further processing and aggregation.

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQP_ID</td>
<td>Equipment identification number</td>
</tr>
<tr>
<td>DATE</td>
<td>Recording date</td>
</tr>
<tr>
<td>MILES</td>
<td>Recording mileage</td>
</tr>
</tbody>
</table>

3.2.3 WILD Data

WILD is built into the track to detect defective wheels, weighting each wheel on the train several times when the wheel passes by a detector in a certain distance (Lechowicz and Hunt 1999). WILD systems are quite expensive and therefore only installed at a few strategic locations on the rail network (Yang and Létourneau 2005).

WILD uses strain-gauge-based technologies to measure the performance of a railcar in a dynamic mode (Li, Parikh et al. 2014). The strain gauges quantify the force applied to the rail through a mathematical relationship between the applied load and the strain caused to the rail web or rail foot (Stratman, Liu et al. 2007). As indicated in (Li and He), WILD is the most important detector to predict wheel RUL.

Each WILD site collects 16 vertical force values and 16 lateral force values per rail; the strain gages were spaced for maximum exposure to all wheel sizes. To reduce data storage space, the 16 vertical and 16 lateral force values for each rail were used to calculate a vertical average force, vertical peak force, lateral average force, and lateral peak force measurement for each wheel (Stratman, Liu et al. 2007).

WILD generates different levels of data include: train data, equipment data, truck data and wheel data. For each level, some important features are shown in Table 3. So does the connection of each level's data.
Train Data
Each train has at least one locomotive and several equipment. Those equipment may belong to different companies.

Equip Data
Each equipment has at least two trucks. Or more than 20 trucks in rare cases.

Truck Data
Each truck has two axles. Those trucks could be of different types.

Wheel Data
Each axle has two wheels, left and right.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DET_ID</td>
<td>Unique identifier for an e-detector</td>
</tr>
<tr>
<td>CAR_CNT</td>
<td>Count of rail cars in a train.</td>
</tr>
<tr>
<td>MAX_PK_KP</td>
<td>Max peak wheel load reading in kips</td>
</tr>
<tr>
<td>MAX_KP_RA</td>
<td>Max ratio between average &amp; peak kips</td>
</tr>
<tr>
<td>LOCO_ID</td>
<td>Locomotive name</td>
</tr>
<tr>
<td>MAX_HUNT</td>
<td>The max train car truck hunting index</td>
</tr>
<tr>
<td>TRN_KIND</td>
<td>Kind of train</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DET_ID</td>
<td>Unique identifier for an elec detector</td>
</tr>
<tr>
<td>EQP_INIT</td>
<td>Equipment initial</td>
</tr>
<tr>
<td>AXLE_CNT</td>
<td>If axel count for a car was correct</td>
</tr>
<tr>
<td>TRCK_CNT</td>
<td>Truck amount in one equipment</td>
</tr>
<tr>
<td>LST_MAINT</td>
<td>User id that changed this row</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DET_ID</td>
<td>Unique identifier for an e-detector</td>
</tr>
<tr>
<td>TRCK_SEQ</td>
<td>The sequence of a truck on a car</td>
</tr>
<tr>
<td>WGT_TONS</td>
<td>The weight of a truck recorded in tons</td>
</tr>
<tr>
<td>HUNT_IDX</td>
<td>The truck hunting index for each truck</td>
</tr>
<tr>
<td>RMĐTNS_TS</td>
<td>Indicates timestamp</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DET_ID</td>
<td>Unique identifier for an e-detector</td>
</tr>
<tr>
<td>AXLE_NBR</td>
<td>Identify an axle on specific piece equip</td>
</tr>
<tr>
<td>AXLE_SIDE</td>
<td>R=right, l=left</td>
</tr>
<tr>
<td>AVG_KIPS</td>
<td>Average load reading in kips for a wheel</td>
</tr>
<tr>
<td>PEAK_KIPS</td>
<td>Peak load reading kips for wheel</td>
</tr>
<tr>
<td>LAT_KIPS</td>
<td>Ave lateral load reading kips for a wheel</td>
</tr>
<tr>
<td>LAT_PEAK</td>
<td>Peak lateral load reading kips for a wheel</td>
</tr>
</tbody>
</table>
3.1.4 Data extraction and cleaning

More than 50 GB raw data was collected in this paper. Data ranges from Jan 2010 to Mar 2012. Only equipment with more than one repair will remain in the final table since the last repair type (failure type) will be used to predict the following repair type (failure type). Data records with RUL less than 60 days were excluded from this study. When RUL is very short, it is more reasonable to apply failure detection other than RUL prediction (Li and He).

Moreover, to address relatively near failures, data records with RUL greater than 180 days were removed from the final dataset as well. Furthermore, to study how axle side (left or right) influences on RUL and failure type, data with only one side records were also omitted from the raw data. Although it has more than 10 types of car in the original dataset, some car types only account for a very small proportion. We merged those car types to one called miscellaneous type, denote as M. This results in only four types: gondola, hopper, hopper without cover and miscellaneous. After data extraction and cleaning, the final dataset consists of 2459 observations and 110 features (M= 2459, N=61). The entire dataset is fed into MTFS.

3.2 Data analysis under different tasks

3.2.1 RUL analysis

After data cleaning, RUL was narrowed to a range from 60 days to 180 days. The statistical summary of RUL in each sub category is presented in Table 4. As one can see, RUL will be very similar for either New or Turn, as the last repair type. And RUL for each car kind is very close to each other. Further, 36 inch wheels survive longer than 33 inch.

Table 4 Statistical summary of RUL
### Attributes Category

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Min</th>
<th>1st Qu</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIR_RPR.N</td>
<td>61</td>
<td>87</td>
<td>118</td>
<td>118.5</td>
<td>152</td>
<td>179</td>
</tr>
<tr>
<td>FIR_RPR.T</td>
<td>61</td>
<td>93</td>
<td>128</td>
<td>123.9</td>
<td>155</td>
<td>179</td>
</tr>
<tr>
<td>SIZE.33</td>
<td>61</td>
<td>78</td>
<td>102</td>
<td>112</td>
<td>141</td>
<td>177</td>
</tr>
<tr>
<td>SIZE.36</td>
<td>61</td>
<td>88</td>
<td>121</td>
<td>120</td>
<td>153</td>
<td>179</td>
</tr>
<tr>
<td>CAR_KIND.H</td>
<td>61</td>
<td>96.5</td>
<td>126</td>
<td>121.9</td>
<td>148.5</td>
<td>178</td>
</tr>
<tr>
<td>CAR_KIND.L</td>
<td>61</td>
<td>86.25</td>
<td>118</td>
<td>119.7</td>
<td>153</td>
<td>179</td>
</tr>
<tr>
<td>CAR_KIND.M</td>
<td>61</td>
<td>89</td>
<td>120</td>
<td>120.2</td>
<td>153</td>
<td>179</td>
</tr>
<tr>
<td>CAR_KIND.G</td>
<td>61</td>
<td>88</td>
<td>119</td>
<td>119.5</td>
<td>153</td>
<td>179</td>
</tr>
</tbody>
</table>

#### 3.2.2 Competing risks analysis for failure time and failure types

Standard survival data measure the time span from some time origin until the occurrence of one type of event. If several types of events occur, a model describing progression to each of these competing risks is needed (Putter, Fiocco et al. 2007). In this paper, two failure risks are addressed including New and Turn, which involve different maintenance costs. This paper applied the competing risk model to analyze failure time which will be observed by the differences between the first repair time and the next bad order time (Scrucca, Santucci et al. 2007).

**Figure 15** Competing risk for different car kind  
**Figure 16** Competing risk for different wheel size

G, H, L and M represent four different car kinds, where 33 and 36 are wheel sizes. N and T
indicate the last repair type. As one can see from Figure 15-16, no matter what car kind or wheel size it is, failure type of “New” occurs much more frequently than type “Turn”. For every car type (Figure 15), the probability of falling to “New” failure is very close to each other. So is their “Turn” risk. But “New” risk is apparently higher than “Turn”. Surprisingly, it is found that in the most of time, wheel failure cannot be fixed without replacing a new wheelset. Such expenditure spent on wheelset replacement could be expected to decrease if more preventive maintenance is applied. Furthermore, when we consider wheel size (Figure 16), category of “33N” generates the highest failure probability among all categories, which indicates that existing maintenance needs to be improved when applied on 33 inches wheel.

We further investigate the probability of failing under categories classified by both car type and wheel size. In Fig 17, all solid lines represent type “New” failure and dash lines represent type “Turn” failure. It shows a similar pattern as the previous two figures that probability of “New” failure is higher than the “Turn” failure. One exception is that category G with 36 inches wheel, represented by red line, have high failure probability in both failure types. Same case happened to miscellaneous car, its failure of “Turn” becomes particularly high, even higher than its “New” risk after 20th month. One possible reason is that “M” is a category mixed with several different car types which only takes a small portion in the whole dataset.
Figure 17 Competing risk for car type and wheel size
CHAPTER 4 METHODOLOGY and ALGORITHM

Our methodology consists of two major parts: Multi-task feature selection (MTFS) and Multi-modal support vector machine (SVM) and support vector regression (SVR). In this paper, we have two tasks, including regression task of RUL and classification task of corresponding failure type. MFTS aims to select common feature for both tasks: RUL prediction and failure type classification. Specifically, we assume that the related tasks share a common relevant feature subset but with a varying amount of influence on each task, and thus adopt a multi-task feature selection method to obtain a common feature subset for different tasks simultaneously. The underlying assumption is that prediction of RUL will provide useful information for classification of failure types and vice versa. Based on jointly selected features; we implement SVM/SVR to conduct the prediction and classification for these tasks, respectively. Figure 18 illustrates the flow chart of the proposed methodology.
4.1 Multi-task feature selection (MTFS)

Different from the conventional single-task feature selection, multi-task feature selection simultaneously selects a common feature subset relevant to all tasks. MTFS is especially important for diagnosis of failure time and type, since they are both essentially determined by the same underlying attributes, i.e., car kind, wheel load, wheel size, etc (Zhang, Shen et al. 2012).

The learning algorithm simultaneously learns all the tasks through two alternating steps. The first step consists of independently learning the parameters of the regression/classification tasks. The second step consists of learning, in an un-supervised way, a low-dimensional representation for these task parameters, which we show to be equivalent tolerance common features across the tasks. The number of common features learned is controlled by the regularization parameters (Evgeniou and Pontil 2007).

Regression task

RUL, denoted as $y_i^r$ in this paper, is defined as number of days between current measurement date and the following bad order date. Obviously, $y_i^r$ is a continuous variable that we can model it as linear regression problem where:

$$y_i^r = a_0 + \sum_{j=1}^{m} x_{ij}a_j + \epsilon, i = 1, ..., n$$

Suppose we have $n$ observations, while each observation has $m$ features. $x_{ij}$ represents $i^{th}$ observation’s $j^{th}$ feature. And $a_j$ represents coefficient for $j^{th}$ feature, while $a_0$ is the intercept and $\epsilon$ the residual.
Accordingly, least square is adopted as the loss function of regression task, which is formulated as follows:

\[ L_r = \sum_{i=1}^{n} (y_i^r - a_0 - \sum_{j=1}^{m} x_{ij} a_j)^2 \]

Classification task

Wheelset repair can be either replacement or reprofiling. We assume that wheelset goes back to original brand new state after either repair activity is applied. If the wheelset is badly damaged so that maintenance technicians have to replace the whole wheelset, we denote its corresponding failure type for replacement as 1, otherwise denote it as 0. Let \( y_c \), a binary variable, represents the failure type. We use the logistic regression to model it (Czepiel 2002).

\[ y_i^c = \frac{1}{1 + \exp \left( -b_0 - \sum_{j=1}^{m} x_{ij} b_j \right)} , i = 1, ... , n \]

Classification task shares the same inputs with regression task. Identically, \( b_j \) represents classification coefficient for \( j^{th} \) feature and \( b_0 \) the intercept. To estimate parameters \( b_j \), one can minimize the negative log-likelihood (Yang, Kim et al. 2009) given as below:

\[ L_c = - \sum_{i=1}^{n} \left\{ y_i^c \left( b_0 + \sum_{j=1}^{m} x_{ij} b_j \right) + \log \left[ 1 + \exp \left( b_0 + \sum_{j=1}^{m} x_{ij} b_j \right) \right] \right\} \]

Regularization

The key point of MTFS is to first apply L2-norm for both tasks, which forces weights corresponding to the \( j^{th} \) across multiple tasks to be grouped together and tends to be selected jointly as a group (Zhang, Shen et al. 2012). Furthermore, we also applied L1 regularization to obtain a sparse solution of MTFS results, in which the weights of groups of
features are forced to be zero. Figure 19 illustrates the details of regularization processes.

We do not penalize $a_0$ and $b_0$ since we do not want the model to depend on the mean of the $y$ vector (Schmidt 2005).

![Diagram for L2/L1 regularization](image)

Let $L_n$ be the regularization penalty, shown as follows:

$$L_n = \sum_{j=1}^{n} (a_j^2 + b_j^2)$$

**Objective function**

Based on previously defined three components, the objective function can be easily conducted as follows:

$$\text{Min: } L = \theta L_r + L_c + \lambda L_n$$

Subject to: $L_r = \sum_{i=1}^{n} (y_i^r + a_0 + \sum_{j=1}^{m} x_{ij} a_j)^2$

$$L_c = - \sum_{i=1}^{n} \left\{ y_i^c \left( b_0 + \sum_{j=1}^{m} x_{ij} b_j \right) + \log \left[ 1 + \exp \left( b_0 + \sum_{j=1}^{m} x_{ij} b_j \right) \right] \right\}$$

$$L_n = \sum_{j=1}^{n} (a_j^2 + b_j^2)$$
where λ is a regularization parameter which determines the sparsity level and it controls the number of selected features. i.e the “group sparsity”. Because of the characteristic of ‘group sparsity’, the solution of MTFS results in a weight matrix whose elements in some rows (groups) are all zeros. For feature selection, we just keep those features with non-zero weights. At present, there are many algorithms developed to solve MTFS. We will adopt gradient descent algorithm to solve the above optimization formulation. Note that at preprocessing stage, we perform a common feature normalization step, i.e., subtracting the mean and then dividing the standard deviation (of all training subjects) for each feature value.

**4.2 Algorithm**

**4.2.1 Gradient descent algorithm**

Several different methods can be used to solve the proposed objective function, such as gradient descent, steepest descent, Newton’s method and interior method. Second-order derivatives methods have fast convergence to approach a global minimum of convex objective functions, but they also involve computing Hessian matrix and its inverse matrix, which most likely would be infeasible in the high-dimensional setting. Considering this and we only have two tasks, we adopt gradient descent method(Boyd and Vandenberghe 2004). This method iteratively updates each element of the parameter vector once at a time, using a closed-form update equation given all of the other elements. Besides, the loss functions of linear regression and logistic regression have different forms. The gradient descent method optimizes their original loss function without any transformation so that it is more intuitive to see how the two heterogeneous tasks affect each other.
Given an initial guess the weight matrix $w = w_{\text{initial}}$

Repeat

Step 1 Compute gradient $g(a_j, b_j)$;

Step 2 Update $w$ by $w_{\text{new}} = w - \gamma^* g(a_j, b_j)$;

Step 3 Checking stop criterion: quit if mean $\{\text{abs} (w_{\text{new}} - w) / \text{abs} (w)\} < \text{tolerance}$;

Step 4 Update $w$ by $w = w_{\text{new}}$.

Figure 20 Gradient descent algorithm

$\gamma$ is the learning rate. It controls how big the step is when updating the parameters. If $\gamma$ is very large, then it corresponds to very aggressive gradient procedure. If $\gamma$ is very small, then it corresponds to small steps.

4.2.2 Objective function in matrix format

$$ L_r = (Y_r - X \cdot A)' \cdot (Y_r - X \cdot A) $$

$$ L_c = -Y_c' \cdot (X \cdot B) + \sum_{i=1}^{m} \log (1 + \exp (X \cdot B)) $$

$$ L_n = A' \cdot A + B' \cdot B $$

*Input data $X$: mxn matrix

*Output for regression (RUL) $Y_r$: nx1 vector

*Output for classification (failure type) $Y_c$: nx1 vector

*Loss function for regression $L_r$: nx1 vector

* Loss function for classification $L_c$: nx1 vector

*Coefficient vector for regression $A$: nx1 vector

* Coefficient vector for classification $B$: nx1 vector

4.2.3 Proof of convexity
Before applying gradient decent method, we need to prove that the objective function is convex. If it’s not a convex function, there is no optimal solution that is guaranteed. Since each coefficient is a decision variable, it will be more convenient to solve the problem with its matrix format (Bazaraa, Sherali et al. 2013).

The convexity of a function can be proved by its Hessian matrix. If the Hessian matrix is positive semi-definite on the interior of the convex set, it is a convex function (Ruszczyński 2006). Hessian matrix is a square matrix of second-order partial derivatives of a scalar-valued function. We denote objective function’s Hessian matrix as H. H is a 2nx2n square matrix, consisting of 4 partitions: P1 to P4, shown as below.

\[
H = \begin{bmatrix}
P_1 & P_2 \\
P_3 & P_4
\end{bmatrix}
\]

\[
P_1 = \frac{\partial^2 (L)}{\partial^2 (A)} = 2 \cdot X' \cdot X + 2I
\]

\[
P_2 = P_3 = \frac{\partial^2 (L)}{\partial (A) \partial (B)} = 2I
\]

\[
P_4 = \frac{\partial^2 (L)}{\partial^2 (B)} = X' \cdot (1 - \pi) \cdot \pi' \cdot X + 2I
\]

Since P2 and P3 are nxn identity matrix, which can be simplified to 0. Thus H can be transferred to block diagonal matrix D.

\[
D = \begin{bmatrix}
P_1 & 0 \\
0 & P_4
\end{bmatrix}
\]

For D, it’s straightforward to see that for every non-zero column vector z of n real numbers, 

\[
z^T D^* z \text{ is greater or equal to 0. Therefore, } H \text{ is a convex function.}
\]
CHAPTER 5 EXPERIMENT RESULTS

This chapter consists of two major parts: Multi-task feature selection (MTFS) and support vector regression (SVR)/support vector machine (SVM), according to the flow chart (fig 4.1) described in Chapter 4. Part 1 (MTFS) includes selection procedure and a set of selected features based on the original feature pool presented in chapter 3. Part 2 (SVR/SVM) is also divided into two parts: regression for RUL prediction and classification for failure types. Several evaluation indicators, such as MAPE, precision, and recall, are used to measure the efficiency of the proposed model.

5. 1 MTFS Result

After data extraction and cleaning in Chapter 3, the final dataset consists of 2459 observations and 110 features (M= 2459, N=61). The entire dataset is fed into MTFS.

5.1.1 MTFS experiment procedure

We have demonstrated that our algorithm is able to find global optimum in Chapter 4. It guarantees that for a fixed dataset, our result is combination of most significant features. The main idea of MTFS is to select significant features through adjusting value of parameters λ. Different λ results different subset of original feature pool.

![Figure 21 Procedure of MTFS](image)

The above figure shows details of the procedure of MTFS. In the first step, the value of λ and θ vary from 0.001 to 1000. When λ increases, it reduces the value of coefficients towards 0
gradually. And because of the fixed stopping criteria, the iteration could stop when it reaches the tolerance level. Figure 5.2 (a) and (b) show how λ affect objective value and iteration numbers. As one can see, when λ is larger, the objective value tends to converge faster.

For different tasks, it is reasonable to set different selection criteria since the loss functions of linear regression and logistic regression have different forms. For example, features are considered to be significant if the corresponding coefficients are greater than 0.01 in regression task. However, the key point of MTFS is to obtain common features. Within the feature subset, which is demonstrated as significant through regression task, we need to remove those features that are non-significant in classification tasks. Figure 5.3 displays how different λ affects the number of selected features, which decreases with an increasing λ. However, in the range [0, 0.6] for the ratio λ/λ_{max}, the number of selected features goes up and down, which means some noise exists. In the range [0.6, 1], when λ is very big, all coefficients are forced to 0 thus no feature is selected.
In step 3, we select a solution with a proper number of features. Over-fitting generally occur if a model contains too many features. In this case, the model describes random error or noise instead of the underlying relationship. Or the model could become under-fitting when it cannot capture the underlying trend of the data if too few features are included.

Hundreds of experiments are conducted to search a suitable $\lambda$ and the corresponding number of selected features. Some results are presented in Table 5. Compared to $Lr$ (least square) $Lc$ varies dramatically. The reason is that maximum likelihood is very sensitive since it includes exponential function. The selected feature number decreases as Fig 24 displays.

### Table 5 Part of experiment results

<table>
<thead>
<tr>
<th>Trial #</th>
<th>$\theta$</th>
<th>$\lambda/\lambda_{max}$</th>
<th>L</th>
<th>$L_c$</th>
<th>$\lambda\cdot\ln$</th>
<th>$\theta\cdot L_r$</th>
<th>Iteration#</th>
<th>Selected Feature #</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>0.01</td>
<td>434.8994</td>
<td>184.9594</td>
<td>67.56367</td>
<td>182.3664</td>
<td>8553</td>
<td>36</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>0.12</td>
<td>1132.455</td>
<td>196.0116</td>
<td>754.0382</td>
<td>182.4052</td>
<td>498</td>
<td>31</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>0.28</td>
<td>1253.554</td>
<td>425.6094</td>
<td>644.5609</td>
<td>183.3835</td>
<td>237</td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>0.42</td>
<td>1371.195</td>
<td>547.0092</td>
<td>640.1895</td>
<td>183.9963</td>
<td>213</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>0.41</td>
<td>1359.133</td>
<td>544.0439</td>
<td>631.1071</td>
<td>183.9816</td>
<td>174</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>0.47</td>
<td>1402.299</td>
<td>584.0818</td>
<td>634.038</td>
<td>184.179</td>
<td>131</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>100</td>
<td>0.53</td>
<td>1451.84</td>
<td>609.2766</td>
<td>658.2805</td>
<td>184.2824</td>
<td>125</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>100</td>
<td>0.68</td>
<td>1512.169</td>
<td>699.1607</td>
<td>628.3714</td>
<td>184.6374</td>
<td>115</td>
<td>2</td>
</tr>
</tbody>
</table>
The highlighted trial (#4) is selected as the final solution of MTFS since 18 features is a proper subset based on original 110 features Table 6 displays all the selected feature.

<table>
<thead>
<tr>
<th>LABELS</th>
<th>EXPLANATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHL_A_KIPS.R</td>
<td>Average load reading taken in kips for an individual car wheel</td>
</tr>
<tr>
<td>N_WHL_A_KIPS.R</td>
<td>Normalized WHL_A_KIPS for right wheel</td>
</tr>
<tr>
<td>TTRCK_WGT_TONS</td>
<td>Weight of a train car truck recorded in tons</td>
</tr>
<tr>
<td>EDR_EQP_SPD</td>
<td>Speed of an equipment at the instance an e-detector read</td>
</tr>
<tr>
<td>CAR_CNT</td>
<td>Count of rail cars in a train.</td>
</tr>
<tr>
<td>EDR_HMDTY_PCT</td>
<td>Humidity percentage</td>
</tr>
<tr>
<td>EDR_WIND_DIR</td>
<td>Wind direction</td>
</tr>
<tr>
<td>VNDR_TRN_TYP.F</td>
<td>A general type of train as assigned by a vendor. F= Freight</td>
</tr>
<tr>
<td>MAX_PK_KP</td>
<td>Max peak wheel load reading in kips</td>
</tr>
<tr>
<td>TTRCK_MAX_TONS</td>
<td>Maximum weight of a train car truck recorded</td>
</tr>
<tr>
<td>VNDR_LD_CD.L</td>
<td>Load status for a car or train. 0=Empty 1=loaded</td>
</tr>
<tr>
<td>VNDR_LD_D.M</td>
<td>Empty status for a car or train 0=loaded 1=Empty</td>
</tr>
<tr>
<td>MAX_HUNT</td>
<td>The max train car truck hunting index</td>
</tr>
<tr>
<td>TRN_KIND.C</td>
<td>Kind of train. C=COAL</td>
</tr>
<tr>
<td>FIR_APPLD_JCD</td>
<td>Last repair applied job code</td>
</tr>
<tr>
<td>CAR_KIND.G</td>
<td>Car kind. G=GONDOLA</td>
</tr>
<tr>
<td>SIZEE.36</td>
<td>Wheel size. 36=36 inches</td>
</tr>
<tr>
<td>FIR_REP</td>
<td>Last repair type</td>
</tr>
</tbody>
</table>

To verify the solution, we further examine the physical meaning of each selected feature.

* WHL_A_KIPS: Kips represents the downward or lateral force of a wheel when it passes over WILD, recorded in thousands of pounds. High average kips usually indicate potential wheel defects.

* TTRCK_WGT_TONS: Railcar truck (or bogie) is the complete assembly of parts including wheels, axles, bearings, and all associated connecting components. The function of a truck is
to provide support, mobility, and guidance to a railroad car. Continuous heavy truck load will cause high wheel load and eventually cause problems on wheelsets.

* VNDR_TRN_TYP: Passenger cars and freight cars have different load and usage pattern. So does the failure patterns.

* MAX_PK_KP: The maximum peak car wheel load is also a useful indicator for bad wheels.

* MAX_HUNT: The truck hunting index is the difference between kips (downward or lateral force) readings of wheels common to a truck. Truck hunting is usually caused by a swaying motion, which can be violent, damaging track and wheels and potentially causing derailment at a critical speed. Wheel defects will cause truck defects as well as truck hunting.

* TRN_KIND: It indicates different function types of train transportation, such as coal, chemistry product or military. Different types of train will generate different failure types.

* CAR_KIND: Car kind is related to train kind. As we described in chapter 3, gondola is usually used to carry coal. These two features jointly determine the characteristics of railcar.

* SIZEE.36: 36 inch wheel is the most widely used wheel in our dataset.

* FIR_REP: The last repair behavior has direct influence on both RUL and the following failure type.

5.2 SVM Result

To get reliable result, we use k-fold cross-validation to testify our prediction results based on jointly selected features. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. 5-fold cross-validation is used in this paper, which means 80% data is used as training data, 20% as test data.
5.2.1 Classification with SVM

In this paper, SVM model is trained to distinguish between “New” (N) and “Turn” (T). A confusion matrix is used to summarize the classification results. Confusion matrix is a specific table layout that allows visualization of the performance of an algorithm. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. Test data consists of 491 observations. The resulting confusion matrix is shown in Table 7. Larger numbers in both up-left box and bottom-right box indicate better classification results. For example in table 7, one experiment result shows that, 319 actual New and 58 actual Turn are correctly classified. But some error do exist, like 46 New is incorrectly classified as Turn.

<table>
<thead>
<tr>
<th>Actual</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>319</td>
</tr>
<tr>
<td>T</td>
<td>68</td>
</tr>
</tbody>
</table>

Table 7 Confusion Matrix Result for one trial

To visualize the performance of SVM classification result, receiver operating characteristic (ROC) curve is also introduced as Figure 25. The curve is created by plotting the true positive rate against the false positive rate at various threshold settings. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. This ROC curve is above the 45-degree diagonal but not very far from it, which means there is still room for improvement.
5.2.2 Regression with SVR

For the regression task, we evaluate its performance through absolute percentage error (APE) and mean absolute percentage error (MAPE) of predicted RUL. Since car kind of gondola and wheel size with 36 inches dominate in the dataset, we divide the dataset into 4 subsets, which is gondola & 36, gondola & non-36 inches, 36 inches & non-gondola and 36 inches & gondola. In our dataset, there is no gondola car with 33 inches wheel size involved. From the perspective of car kind and wheel size, we can conclude that it’s relatively more robust to predict category in non-gondola & non-36 inches since the range of its APE is smaller, compared to other categories (see Fig 26). In addition, there is no evident difference between APE of type “New” and type “Turn” (Fig 27). It is found that the SVR model works equally well for both failure types.
* G= gondola   NG= Non-gondola
* 36= wheel size with 36 inches   N36= wheel size beyond 36 inches

\[ APE(i) = \left( \frac{|Predict(i) - Actual(i)|}{Actual(i)} \right) \]

\[ MAPE = \text{mean}(APE(i)) \]

*Predict(i) = \( i \)th predicted RUL
*Actual(i) = \( i \)th actual RUL

5.2.3 Cross-validation results

Cross validation and repeated test are both applied in this thesis. For regression task, we take the mean of APE (MAPE) to measure the accuracy of the regression. For classification task, we calculate precision and recall based on confusion matrix to evaluate classification result.

<table>
<thead>
<tr>
<th>Actual</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Yes</td>
<td>TP</td>
</tr>
<tr>
<td>No</td>
<td>FN</td>
</tr>
</tbody>
</table>

*True positive (TP)
*True negative (TN)
*False positive (FP)
* False negative (FN)

Define

\[
\text{Precision} = \frac{TP}{TP + FP} \\
\text{Recall} = \frac{TP}{TP + FN}
\]

Table 8 Comparing the cross-validation result between single task learning and multi-task learning

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Single Task Feature Selection</th>
<th>MTFS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Method</td>
<td>SVR</td>
</tr>
<tr>
<td>Value</td>
<td>MAPE</td>
<td>Precision</td>
</tr>
<tr>
<td></td>
<td>0.280</td>
<td>0.800</td>
</tr>
</tbody>
</table>

As one can see, the results from MTFS outperform the results from single task, which is conducted by L1 regularization. The MAPE is reduced by 3%, with some minor improvement in precision and recall. The results from cross-validation validate that multi-task learning works well in train wheel failure prediction which jointly considers failure time and failure type.
CHAPTER 6 CONCLUSIONS and FUTURE WORK

6.1 Conclusions

In this thesis, we propose a multi-task learning method to jointly select common features for prediction of wheelsets Remaining Useful Life (RUL) and failure types. Such method combines linear regression loss, logistic regression loss and L2/L1 regularization for multi-task feature selection (MTFS). In our experiments, wheel measurement data from WILD is incorporated with and bad order data and repair data to a comprehensive table, which is further divided to training dataset and test dataset to perform cross-validation. We demonstrate that using L2/L1 regularizations not only selects features but also leads to “group sparsity”, which identifies the input variables that are commonly relevant to multiple tasks.

Besides, those jointly selected features have shown the consistence between each other. For example, TRAIN_TYPE. C, indicating whether the train transports coal, is selected together with CAR_KIND.G, representing if a railcar is a gondola, designed for ore product. The selection of these two features implies the success of MTFS.

Competing risk analysis revealed that wheelsets tend to fall into severe failure which requires a replacement (“New”). Most of time, mechanics have to replace the defected wheelsets with a new wheelsets, no matter what car kind or wheel size is. And it’s always replaced by pairs. Only a small portion of wheel failure is rectified by turning the wheels or reprofiling the wheels (“Turn”). This finding indicates that existing operations in wheelset maintenance is not running in a very efficient manner.
The regularization parameter $\lambda$ greatly influences the process of MTFS and the total iteration numbers. When $\lambda$ increases, both two sets of coefficient decrease thus less feature will be selected. At the same time, the program will stop earlier and the objective function will reach the optimal earlier. On the other hand, different $\lambda$ will contribute several candidate subsets of original features. We need to decide which subset is a proper solution to our problem. In this study, 18 features is selected and considered as inputs for later prediction part.

The prediction consists of two components, SVM for failure type classification and SVR for RUL prediction. In failure type classification, most of failures are classified correctly. Type I and II errors do exist, but only take a small portion in test data. ROC curve proves same conclusion. As for RUL prediction, the prediction results show that the MAPE of multi-task learning reduces single-task learning by 3%. It is also found that the wheelsets with car kind gondola and 36 inch wheel are relatively easier to be predicted by comparing across different categories.

In summary, our experimental results show that our proposed multi-task learning method can effectively predict failure time and failure type of wheelsets concurrently.

6.2 Future work

One future research direction is to apply more descents method to reduce computing time. We can also extend to second-order derivatives method to prepare for larger-scale data and more tasks. In addition, some other machine learning methods could be developed to increase prediction accuracy for RUL.
References


Li, Z. and Q. He "Prediction of Railcar Remaining Useful Life by Multiple Data Source Fusion."


Marcus, B. (2002). "Condition Based Maintenance on Rail Vehicles–Possibilities for a more effective maintenance strategy."


Schlake, B. (2010). Impact of automated condition monitoring technologies on railroad safety and efficiency, University of Illinois at Urbana-Champaign.


