工业化的兴起导致了对能源的大量需求，化石能源的枯竭以及环境污染促进了新能源的崛起[2,3]。然而，大多数可再生能源，如太阳能和风能，本质上是间歇性的，依靠自然现象来发电，此时能源存储便成为新的需求 [4-10]。可充电锂离子电池作为一种储能技术，由于拥有高密度、大容量、长寿命等优点，已经得到了广泛应用[11,12]。但广泛使用的同时也带来了续航、动力不足、电池爆炸等问题，如图1所示。如果能在电池老化之前对电池寿命进行预测的话，在避免上述问题的同时还可以加快电池开发周期、验证新的工艺、对电池进行二次回收[16-18]，将为电池的制造、使用和优化带来重大机遇[19-21]。

目前的电池寿命估计所采用的模型主要可以分为等效电路模型(ecm)[22-23]、电化学模型[24-26]和数据驱动模型[27-32]三种。电化学模型和等效电路模型的准确性和鲁棒性有限，因此这两种模型并不是一个很好的可行解决方案。相反，数据驱动模型不需要了解电池内部的复杂化学反应，没有复杂的建立电路的过程等优势被研究者广泛应用。同时随着近几年的研究展开，发现了电池数据集中带有噪声是不可避免的，这主要源于充放电过程中的环境干扰，如温度变化、湿度波动的影响。此外实验条件并不能完全模拟现实，因此，越来越多的研究开始关注带有噪声的电池寿命预测。

线性参数估计问题出现在信号处理等广泛的科学学科中[33-34]。如[35]和[36]所示，在所有感兴趣的变量都具有参数线性关系且所有测量值都受到噪声污染的情况下，总最小二乘法是参数估计的最佳选择。

但实际情况中电池厂商提供的电池信息数据集来源不同，故由于温度、人为干扰以及传感器等造成的误差大不相同，此时便不能简单的假设数据集的噪声服从同一个分布，直接使用TLS/OLS并不能很好的建立电池寿命预测模型，故本文在建立线性模型计算电池的寿命时进行改进，对带有不同噪声分布的电池样本进行加权之后，使用TLS/OLS进行预测，经循环迭代能够准确的计算出噪声分布的标准差的同时建立适应不同噪声分布的预测模型对电池寿命进行预测，预测结果显示我们的方法比传统的TLS/OLS方法更好。

本文使用三个基于商用锂离子电池的公共数据集，三个数据集分别称为“数据集1”、“数据集2”和“数据集3”，分别由41个、43个和40个样本组成，虽然这些数据集总共提取了20个特征，但我们进一步根据领域专业知识手动选择3个重要特征的子集，具体含义见table1，我们将三个数据集按照9:1划分之后合并作为训练集和测试集，每次实验随机打乱样本顺序，每次运行均独立随机生成训练和测试数据集。为每种方法报告1000个RMSE值的中位数，以便误差度量不会因随机波动而产生强烈偏差

Data-driven prediction of battery cycle life before capacity degradation

如图3所示，我们设置了四种不同的噪声比例。图a,b,c,d是噪声水平增大的实验结果：（1）随着噪声增大，TLS和OLS效果明显变差，而改进的算法受噪声水平的影响不明显，具有较强的稳定性。（2）结合EM思想改进的算法（TLS\_EM、OLS\_EM）比传统算法（TLS、OLS）效果更佳,说明了改进的算法更能适应带有噪声的电池数据集。（3）TLS\_EM效果优于OLS\_EM，（TLS效果也优于OLS），在所有测量值都收到噪声污染的情况下，TLS比LS有更大的优势。

图e,f,g,h是训练集比例增大的实验结果：（1）随着训练集占比增大，四种方法效果更好，有了更多的训练数据，模型预测能力提升。（2）不论训练集比例大小，改进的算法优于传统算法，说明了融入EM思想的算法有效性 。（3） 在绝大部分情况下（训练集占比大于25%）TLS\_EM效果优于OLS\_EM，说明了TLS\_EM比OLS\_EM适用性更强。

图4展示了算法经过循环迭代噪声收敛过程。（算法通过TLS/OLS拟合样本数据得到模型系数和，根据模型系数和对样本数据进行预测得到新一轮的电池寿命预测值，将预测值和真实值对比得到新一轮的误差，由此求出标准差对样本数据加权后通过TLS/OLS求出下一轮的模型系数和）

本文在建立线性模型计算电池的寿命时进行改进，对带有不同噪声分布的电池样本进行加权之后，使用TLS/OLS进行预测，经循环迭代能够在准确的计算出噪声分布的标准差的同时建立适应不同噪声分布的预测模型，进而对电池寿命进行预测。预测结果显示我们的方法有着更好的效果。

The rise of industrialization led to a large demand for energy, and the depletion of fossil energy and environmental pollution promoted the rise of new energy [2,3]. However, most renewable energy sources, such as solar and wind, are intermittent in nature and rely on natural phenomena to generate electricity, at which point energy storage becomes a new demand [4-10]. As an energy storage technology, rechargeable lithium-ion batteries have been widely used due to their advantages such as high density, large capacity and long life [11,12]. However, widespread use also brings problems such as battery life, lack of power, and battery explosion, as shown in Figure 1. If the battery life can be predicted before the battery aging, the above problems can be avoided, and the battery development cycle can be accelerated, new processes can be validated, and the battery can be recycled [16-18], which will bring major opportunities for the manufacturing, use and optimization of the battery [19-21].

At present, the models used for battery life estimation can be mainly divided into equivalent circuit model (ecm)[22-23], electrochemical model [24-26] and data-driven model [27-32]. The accuracy and robustness of the electrochemical model and the equivalent circuit model are limited, so these two models are not a good viable solution. In contrast, data-driven models do not need to understand the complex chemical reactions inside the battery, there is no complex process to build the circuit and other advantages are widely used by researchers. At the same time, with the development of research in recent years, it is found that the noise in the battery data set is inevitable, which is mainly due to the environmental interference during the charging and discharging process, such as temperature change and humidity fluctuation. In addition, experimental conditions do not fully simulate reality, so more and more research has begun to focus on the prediction of battery life with noise.

Linear parameter estimation problems arise in a wide range of scientific disciplines such as signal processing [33-34]. As shown in [35] and [36], total least squares is the best choice for parameter estimation when all variables of interest have parametric linear relationships and all measurements are noise polluted.

However, in the actual situation, the battery information data sets provided by battery manufacturers come from different sources, so the errors caused by temperature, human interference and sensors are very different. At this time, it is impossible to simply assume that the noise of the data sets obey the same distribution, and the direct use of TLS/OLS cannot establish a good battery life prediction model. Therefore, this paper made improvements in the establishment of a linear model to calculate battery life. After weighted battery samples with different noise distributions, TLS/OLS was used for prediction. The standard deviation of noise distribution could be accurately calculated through cyclic iteration, and a prediction model adapted to different noise distributions was established to predict battery life. The prediction results show that our method is better than the traditional TLS/OLS method.

When dataset samples obey noise of different distributions,, we can know that our goal is to minimize the error when the samples obey different noises：

,

To solve this problem, we give each sample a different weight:

,

Assuming that the noise obeys a Gaussian distribution with zero mean and different variances: , then obeys a Gaussian distribution . The likelihood function is:

,

Maximizing the likelihood function is equivalent to minimizing the objective function , and we can get by comparison.

EM algorithm can solve the problem with hidden variables well. In this paper, we assume that the standard deviation of noise obeys the distribution as hidden variables, and improve the traditional OLS/TLS algorithm. The specific steps are as follows:

1. Initialize model coefficients w,b

2. Predict the battery life according to the model coefficient, and calculate the error between the predicted value and the real value to update the noise standard deviation of the three data sets.

1 , ()

3. The samples are weighted according to the formula, and a new round of model coefficients are obtained by using TLS/OLS.

4. Repeat steps 2 and 3 until convergence.

In this paper, three public datasets based on commercial lithium-ion batteries, called "Dataset 1", "Dataset 2" and "Dataset 3", are composed of 41, 43 and 40 samples, respectively. Although these datasets extract a total of 20 features, we further manually select subsets of 3 important features based on domain expertise. The specific meaning is shown in table1. We divide the three data sets according to 9:1 and combine them as training sets and test sets. Each experiment randomly scramps the sample order, and each run independently randomly generates training and test data sets. A median of 1,000 RMSE values is reported for each method so that the error measure is not strongly biased by random fluctuations.

Table 1: Features for battery lifetime modeling

| Feature Name | Description |
| --- | --- |
| *x*1 | Variance of the difference in the discharge capacity curves as a function of voltage between the 10-th and 100-th cycles |
| *x*2 | Slope of the capacity fade curve fitted by a linear function |
| *x*3 | Discharge capacity of the 2-nd cycle |

As shown in Figure 3, we set four different noise ratios. Figure a,b,c, and d show the experimental results with the increase of noise level: 1) With the increase of noise, the effects of TLS and OLS become significantly worse, while the improved algorithm is not significantly affected by noise level and has strong stability. 2) The improved algorithms TLS\_EM and OLS\_EM combined with EM idea have better effects than the traditional algorithms TLS and OLS, indicating that the improved algorithm is more suitable for the battery data set with noise. 3) The effect of TLS\_EM is better than that of OLS\_EM, and the effect of TLS is also better than that of OLS). In the case that all the measured values receive noise pollution, TLS has greater advantages than LS.

Figures e ,f ,g and h are the experimental results of increasing the proportion of training sets: 1) With the increase of the proportion of training sets, the four methods have better effects, and with more training data, the prediction ability of the model is improved. 2) Regardless of the proportion of the training set, the improved algorithm is superior to the traditional algorithm, which shows the effectiveness of the algorithm with EM idea. 3) In most cases, the training set accounts for more than 25%)TLS\_EM is better than OLS\_EM, which shows that TLS\_EM is more applicable than OLS\_EM . Figure 4 shows the iterative noise convergence process of the algorithm.

In this paper, the linear model was established to calculate the battery life. After the battery samples with different noise distributions were weighted, TLS/OLS was used for prediction. After cyclic iteration, the standard deviation of noise distribution could be accurately calculated while the prediction model adapted to different noise distributions could be established to predict the battery life. The results show that our method has better results.