TLS和LS是传统的回归预测方法，本文考虑了数据集来自三种不同的批次，不同批次的电池噪声分布不同，在此情况下，使用了em算法和传统方法结合通过多次迭代使模型适应噪声干扰，能够对电池寿命进行准确预测。

引言

引言：锂电池作为新能源的一种，在化石能源迅速消耗的今天作为其替代，不仅能解决能源消耗殆尽，还能保护环境。锂电池因为其体积小、续航时间长、 成本低、使用寿命时间长等原因，被广泛使用在汽车、航空航天等领域。

然而锂电池的广泛使用也带来了一些新的问题，锂电池的不当使用造成电池加速老化，存在较大的安全隐患。随时电池的多次循环使用其容量减少和内阻增加导致性能恶化，容易出现内短路、热失控等问题，导致设备故障甚至带来灾难性事故。锂电池寿命预测是电池管理系统中的一个重要组成部分，它可以通过预测电池寿命来确定何时需要更换电池或对电池进行维护。虽然经过多年的优化改造，但对于锂电池寿命的预测仍然存在不足，是当今研究的热点。

电池寿命预测可以帮助用户提前了解电池的寿命状况，避免电池寿命到期后出现电池失效或存在安全隐患的情况。能提高电池系统的可靠性和安全性；同时通过对电池寿命进行预测，可以提前发现电池故障，及时对电池进行维护或更换，从而降低电池系统的维护成本；通过对电池寿命进行预测，可以优化电池的充电和放电策略，从而提高电池系统的续航里程和电池寿命；锂电池寿命预测技术对于新能源汽车的发展至关重要，它可以提前预测电池寿命，让用户更加放心地使用新能源汽车，同时也可以降低新能源汽车的维护成本，提高其使用寿命。

关于锂电池寿命预测的研究已经取得了很多进展，但现有的锂电池寿命的研究体系主要集中在两个方面，基于经验的方法和基于性能的方法。

基于经验的方法：循环周期数法、安时法和加权安时法、面向事件的老化累积方法。但基于经验的方法对数据的依赖性太强，以至于换一个类型的电池，经验模型效果不好，鲁棒性差。

基于性能的方法：基于模型的方法，基于数据驱动的方法，基于融合模型的方法。基于模型的方法大多建立在电池电化学机理的研究基础上，通过分析电池的退化现象或构建等效 电路而建立数学模型，常用的模型有基于 Wiener 过 程的模型［6，7］，Thevenin 模型［8，9］，自回归模型［10］，单粒子 (single particle, SP) 模型， 伪二维 (pseudo-two-dimensions, P2D) 模型等。其中基于数据驱动的方法不需要了解锂电池内部的物理化学反应，仅仅根据电池的历史循环数据就能很好的预测锂电池的寿命，成为当下研究热点。举例，支持向量机［14］Deng Y W，Ying H J，E J Q，et al． Feature parameter extraction and intelligent estimation of the State-ofHealth of lithium-ion batteries［J］． Energy，2019，176: 91 － 102．，神经网络［15 ～ 17］［15］ 吴忠强，尚梦瑶，申丹丹，等． 基于 BSA-ＲELM 的纯 电动汽 车 锂 离 子 电 池 SOC 估 计［J］． 计 量 学 报和相关向量机［18，19］( relevance vector machine，ＲVM)［18］ Song Y C，Liu D T，Hou Y D，et al． Satellite lithiumion battery remaining useful life estimation with an iterative updated ＲVM fused with the KF algorithm［J］． Chinese Journal of Aeronautics，2018，31( 1) : 31 － 40．［19］ Yang W A，Xiao M H，Zhou W，et al． A hybrid prognostic approach for remaining useful life prediction of Lithium-Ion batteries ［J］． Shock and Vibration， 2016，2016: 1 － 15等都用于预测锂电池的寿命。其中关于使用神经网络的方法主要是集中于直接使用神经网络进行预测还有就是使用神经网络的时序模型，不需要电池的其他信息，只需要知道电池寿命的历史放电容量就能对电池有一个比较准确的预测。文献［37］考虑了数据少时 存在预测精度低问题，对长短期记忆神经网络进行了改进，将该方法与 BP 和传统长短期记忆神经网络进行了对比分析，发现改进的长短期记忆神经网络预测结果误差最小; ［37］ 李瑞津，刘斌，张学敏，等． 基于改进 LSTM 的变电站铅酸电池寿 命 预 测［J］． 电 池，2020，50 ( 6 ) : 560-564．基于融合的方法：主要分为模型与数据驱动的方法和多种数据驱动方法的结合，包括主要涉及电池充放电截止 电压和标称电压，利用径向基神经网络修正无迹卡尔曼的状态估计，从而预测电池寿命的研究。［51］梅枭央． 基于融合算法的锂离子电池剩余使用寿 命预测［D］． 武汉: 华中科技大学，2019．在基于多模型融合和无迹粒 子滤波的电池寿命预测［52］作者使用电压、电流、温度 和 EIS 数据等，考虑了不同种电池衰减情况对电池寿命进行预测。［52］ 王帅． 数据驱动的锂离子电池剩余寿命预测方法 研究［D］． 哈尔滨: 哈尔滨工业大学，2017．

本文考虑了数据集来自三种不同的批次，不同批次的电池噪声分布不同，对于原始的TLS和LS，我们假设其误差是服从零均值标准差相同的高斯分布的，然而现实情况下，误差可能服从不同的高斯分布。在此情况下，仅使用TLS或者LS对电池寿命预测会有较大的误差，本文将EM算法和原始方法结合，将每个带噪声的样本进行加权计算最优函数，将每一个批次的电池误差的标准差作为隐变量，通过多次迭代使模型适应噪声干扰，能够对电池寿命进行准确预测。

数据集介绍

数据集介绍：由124个在快速充电条件下循环失效的商用锂离子电池组成。这些锂离子磷酸盐（LFP）/石墨电池由A123系统公司（APR18650M1A）制造，在设置为48°C的强制对流温度室中的30通道Arbin LBT恒电位仪上的水平圆柱形夹具中循环。 电池的标称容量为1.1 Ah，标称电压为3.3 V。这项工作的目标是优化锂离子电池的快速充电。因此，此数据集中的所有单元格都使用一步或两步快速充电策略收费。此策略的格式为“C1（Q1）-C2”，其中 C1 和 C2 分别是第一和第二个恒流步骤，Q1 是电流切换的充电状态 （SOC， %）。第二个电流步骤以80%SOC结束，之后电池以1C CC-CV充电。上限和下限截止电位分别为3.6 V和2.0 V，符合制造商的规格。这些截止电位对于所有电流步骤都是固定的，包括快速充电;经过一些循环后，电池可能会在快速充电期间达到上限截止电位，从而导致显着的恒压充电。所有细胞在4C下放电。本文将三个批次的电池全部提取，并按照9:1划分训练测试集。所有电池均采用一步或两步充电策略进行循环。充电时间从~8到13.3分钟（0-80%SOC）不等。每个策略通常测试两个细胞，除了3.6C（80%）。在充电期间和放电后分别达到1%SOC后放置1分钟和80秒休息。我们循环到标称容量的 80% （0.88 Ah）。在每次测试开始时执行初始 C/10 周期。恒压阶跃的截止电流在充电和放电时均为C/50。红外测试的脉冲宽度为 30 ms。

TLS，LS，EM算法以及创新

TLS算法：

考虑一组特征,表示特征总数，代表转置矩阵向量，我们的目标是学习一个从特征向量到电池寿命的映射：

是第个基函数，式中为第个基函数，为第M个模型系数，为零均值高斯分布后的建模误差，为基函数总数。

给定一组样本，模型系数可以通过最小化总平方误差来确定[15]:

其中表示向量的L2范数，且：

上述方法在文献中被称为OLS回归，其目的是寻找未知模型系数的最大似然解[15]。

在实际应用中，每个基函数可能是有噪声的，因为特征向量的测量存在测量误差。在这种情况下，Eq.(1)应改写为:

式中表示与第个基函数相关的测量误差。本文假设服从零均值高斯分布。

为了推导出(6)中未知模型系数的最大似然解，我们进一步做了两个假设。首先，和通过适当缩放和可以归一化为标准高斯分布。

第二，和在统计上是独立的。

有了这两个假设，很容易看出观察到一个样本的可能性等于:

=

其中和分别表示和的第个样本。进一步，通过假设数据集中的所有样本在统计上是独立的，观察到这个样本的可能性等于:

因此，通过求解以下优化问题，可以求出最大似然解w:

其中表示矩阵的Frobenius范数

注意，最小化(9)中的成本函数相当于最大化(8)中的似然。这种方法在文献[13]中被称为TLS回归。

(9)中的优化问题因其二次等式约束而非凸[18]。因此，很难找到既高效(即低计算成本)又鲁棒(即保证全局收敛)的最优解。在本节中，我们将原始的非凸问题转换为易于解决的等效版本。

首先，我们注意到，如果向量w已知，Eq.(9)可简化为以下凸优化问题:

其中和是未知问题。Eq.(12)最小化一个受一组线性等式约束的凸二次代价函数。利用KKT条件可解析导出其最优解[18]。为此，我们计算式(12)的拉格朗日量:

其中λ是包含K个拉格朗日乘子的向量。通过求L(ε， E， λ)对ε和E的偏导数并令它们为零，我们有:

ε和E的最优解必须满足式(14)-式(15)和线性等式约束:

求解(14)-(16)线性方程组得到:

, (17)

(18)

将式(17)-式(18)代入式(9)，则TLS回归的代价函数可表示为:

(19)

注意，(19)中的成本函数在ε和E上已经最小化，现在它是w的非凸函数。

为了有效地最小化式(19)中的代价函数，我们构造如下优化问题[18]:

(20)

等于

(20)

其中，

(22)

(23)

且是的第(M+1)个元素。研究(21)，我们注意到它的代价函数与的大小无关。

因此，在不存在约束的情况下，求解以下优化问题，可以得到相同的最优成本函数值:

(24)

Eq.(24)是众所周知的瑞利商最小化问题[18]。它的最优解是b的最小特征值对应的特征向量。一旦已知，我们可以将乘以它的(M+1)-元素M+1来确定(21)的最优解θ:

(25)

因此，取θ的前M个元素即可得到未知模型系数的最优解w: 261 2T M枭龙w L。

(26)

算法1总结了TLS回归的主要步骤:

算法1:Total-Least-Squares (TLS) Regression从一组基函数，一组样本。

2. 计算矩阵G在(3)和向量y在(5)。

3.根据式(22)构造矩阵B。

4. 计算B的最小特征值对应的特征向量。

5. 通过将归一化计算向量在(25)中。

6. 计算包含最优模型系数的向量w在(26)。

EM：

给定数据集，假设样本间相互独立，我们想要拟合模型到数据的参数。根据分布得到似然函数：，也可以最大化似然函数以求解，但当存在隐变量时，似然函数表示为：

对于某个样本i，我们使用表示样本i隐含变量z的分布概率，则

引入，并利用Jensen不等式：

似然函数变换为：

不等式取等号：当为常数时，不等式可以取等号，对等式变换并两边和:

E-Step：求条件期望，本质是求出来自哪种类别的概率，为简化将log中的去掉，

关于：

两边同时乘以且求和即可得到期望：

可以发现：；

M-Step: 求，更新参数

创新思想：

对于样本 , , 假设满足关系式 , 则就可以使用最小二乘法最小化误差计算：

然而在真实情况下，每个数据样本的质量并不一致，数据样本中可能存在异常。我们希望为不同的数据样本分配不同的重要性，这可以通过在回归损失中分配权重来实现：

我们假设，使得每个数据样本的误差产生具有明显误差的高斯分布。那么数据样本遵循分布。我们希望优化最大似然函数：

最大化的对数等价于最小化，因此， 由于分布的参数不能预先知道，我们需要以EM方式优化似然性，同时考虑作为隐藏变量。 给定模型系数的初始组(可以使用OLS计算)，我们可以估计每个数据样本的方差，并为下一次迭代更新权重。

实验介绍：

**TLS/LS：**( TLS: ，噪声服从同一分布

**噪声不同分布加权：**

**权值求解：**假设，则，

即可得权重

对于LS加权原有的数据集X,Y需要与相乘，再使用LS/TLS求解

EM算法：

1）初始化w,b

2）利用w,b对电池寿命预测，记录预测结果和真实值之差,求出三个电池批次误差的标准差，并根据分类批次对不同批次电池数据进行加权作为新加权数据

3）利用加权数据更新模型系数（直接使用TLS/LS求解）

4）重复步骤2和3直至收敛得到W\_em,b\_em,在测试集上面得到em\_rmse

数据集的划分：训练：测试=9:1，三批次划分再合并，关于噪声：噪声水平0-19\*0.05左右，原始rmse：tls/ls直接求解得到模型系数W,b，在测试集上得到rmse。

实验结果

在本节中，我们使用了三个基于商用锂离子电池的公共数据集[1]来验证我们提出的TLS方法在电池寿命预测中的有效性。这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[1]，但我们进一步根据领域专业知识手动选择5个重要特征的子集，表示为x = [x1 x2 cheese cheese x5] T。在本例中，由于每个数据集中可用的样本数量很少，为了避免过拟合，我们只使用5个特征进行回归建模。这些选定特征的物理意义总结在表1中。为了减少建模任务的非线性，我们对电池寿命y和第一特征x1取对数[1]。通过这些非线性变换，我们采用以下线性模型模板来预测log(y): 27 log logyw w x w x w x w x w 1 2 2 3 3 4 4 5 5 6x1。(27)

为了测试和比较的目的，已经实施了两种回归建模方法:(i) TLS和(ii) OLS。为了评估建模精度，我们将每个数据集划分为(i)一个训练数据集来确定未知的模型系数，(ii)一个测试数据集来评估基于均方根误差(RMSE)的建模精度[19]。我们随机抽取30%的样本组成测试数据集，其余70%的样本组成训练数据集。此外，我们将训练数据集的大小从70%的样本改变为20%的样本，以便在只有极少量的训练样本时研究TLS和OLS的鲁棒性。为了提高数值稳定性，我们将预测结果log(y)和所有特征{log(x1)， x2, x3, x4, x5}归一化，使它们在训练数据集上具有零均值和单位方差。上述实验重复200次，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200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。

结论