arXiv:2301.03005v1 [stat.AP] 8 Jan 2023

动态在线预测模型及其在汽车理赔频率数据中的应用

Jiakun Jiang * Zhengxiao Li † Liang Yang ‡

摘要

理赔频率预测建模是非寿险定价和风险管理中的一项重要任务，需要根据被保险人群、监管法规和技术的变化频繁更新。现有方法要么采用临时性做法，例如参数模型校准，要么较少以预测为目的。本文提出一种动态泊松状态空间模型（DPSS），可在获得新的索赔信息时持续更新参数。DPSS模型同时支持时变系数和时不变系数。为考虑时变系数随时间变化的平滑趋势，本文使用平滑样条对时变系数进行建模。平滑参数通过最大似然客观选取。模型利用按预设时间区间累积的批量数据进行更新，从而更好地近似底层的泊松密度函数。所提方法还可扩展至零膨胀泊松分布和负二项分布的假设。在模拟中，我们表明，与现有方法相比，新模型的预测准确度显著更高。我们将该方法应用于中国一个为期六年的真实汽车保险理赔数据集，并通过与文献中的竞争模型结果进行比较，展示了该方法的优越性。

关键词： 状态空间模型；动态预测；时变；索赔频率数据

文章历史： 2023年1月10日

*北京师范大学统计与数据科学中心，中国珠海。

†对外经济贸易大学保险与经济学院，中国北京。电子邮箱：li zhengxiao@uibe.edu.cn。

‡通讯作者：西南财经大学金融学院，中国成都。电子邮箱：yangliang@swufe.edu.cn。

1 引言

预测未来索赔数量是精算师在保险行业中的主要工作之一，尤其是在保险费率厘定过程中。一个表现良好的预测模型应能长期保持准确性。精算科学中现有的大多数预测方法都采用静态预测模型，例如广义线性模型及其扩展（Klein 等，2015年，Lambert，1992年，Lee，2021年，Yip 和 Yau，2005）。许多保险案例表明，静态预测模型的表现往往容易随时间推移而恶化。这通常是由于被保险人群体随时间发生变化、理赔行为不规律、立法变化或技术进步所致（Chang 和 Shi，2020年，Dong 和 Chan，2013，Henckaerts 和 Antonio，2022，Zhu 等，2011年）。针对这一问题，人们提出了一些方法，例如模型重新拟合和重新校准。然而，这些方法通常是临时采用的。本文提出一种动态预测模型，可随时间推移，依次更新参数估计值以及对未来索赔数量的预测。该模型非常适用于数据生成过程本身不断变化的大规模数据流。此外，该模型可在线实现，使模型参数能够随着新索赔数据的到来而快速更新。

基于多年保险索赔数据预测未来理赔频率，最常用的方法是信度理论，该方法通过随机效应对时间依赖结构进行建模（Denuit 和 Lu，2021年，Lee 等人，2020年）。近年来，若干包含动态随机效应的信度模型已被应用于纵向情境，以进行经验费率厘定；在这些模型中，投保人的风险由个体未观测异质性过程决定（Ahn 等人，2022，Lu，2018，Pinquet，2020，Taylor，2018）。然而，保险动态费率厘定领域仍面临一些挑战。首先，上述预测模型必须依赖纵向数据，这构成了极大的限制：被保险人群体每年都在变化，因此保险公司难以收集并整理个人保单的纵向数据。这些方法会因排除提前退保和未续保的保单而导致大量索赔信息流失。其次，在动态随机效应模型中预测未来保险费时，保险公司将面临计算负担，因为贝叶斯信度保费通常难以处理（Li 等人，2021）。第三，尽管现有方法允许随机效应随时间变化（Pinquet，2020），但它们仍属于静态模型，即假定回归参数在整个数据期间保持固定；当风险因素的影响随时间变化时，这会导致模型参数估计难以准确追踪参数——

估计值（Chang 和 Shi，2020年）。此外，这些模型在捕捉响应变量与协变量随时间变化的复杂非线性关系方面过于受限。这促使我们提出一种相对简单的预测模型结构，将索赔数量表示为若干重要费率因子的函数，同时允许模型参数随时间变化，并捕捉复杂的时间非线性效应。由此，我们开展了对变系数模型的研究（Fan 和 Zhang，2008，Hastie 和 Tibshirani，1993年）。具体而言，在这一情境下，变系数模型易于解释，因此我们可以通过报告拟合得到的个体回归参数，在统计模型与保险费率厘定机制之间建立标准接口。然而，统计学中的变系数模型主要关注估计和统计推断，而非未来预测，因为核估计量和基于样条的估计量都不适用于预测。

状态空间模型是用于建模动态系统的强大工具，既适用于高斯响应，也适用于非高斯响应；其应用领域包括生物学（Bhattacharjee 等人，2018年）、医学（Jiang 等人，2021年）、流行病学（O'Neill，2002）和保险（Ahn 等人，2022，Fung 等人，2017，Lai 和 Sun，2012年）。在这种方法中，通过一组潜在参数随时间对动态系统进行建模，这些参数决定观测数据分布的变化。层次模型结构既可以进行统计推断，也可以在考虑观测到的历史轨迹的基础上预测未来结果变量。对于高斯分布，参数估计存在闭式解。对于非高斯分布，例如我们示例中的理赔次数数据，其解通常依赖复杂的数值积分。为便于计算，通常使用正态近似，例如拉普拉斯近似（Lewis 和 Raftery，1997年a）。

在保险应用中，系统特征会随时间缓慢变化，也可能受到保险监管法规变更的影响。一种常用方法是将关注参数建模为随时间变化的平滑函数，例如平滑样条。Wahba (1978) 和 Wahba(1983) 表明，高斯结果的平滑样条估计等价于高斯随机过程的后验估计。Gu (1992) 表明，借助拉普拉斯近似，这一结论也适用于非高斯结果。Wecker 和 Ansley (1983) 展示了如何使用状态空间模型表述拟合平滑样条。最近，Jiang 等人 (2021) 讨论了基于状态空间模型的临床决策制定中的动态二分类问题。然而，这种方法不适用于保险费率厘定，因为理赔次数数据通常表现出

过度离散且零值过多的特征。受基于多年理赔数据对汽车保险理赔频率进行建模的启发，我们提出一种动态泊松状态空间模型（DPSS），其中时变系数由状态方程建模。泊松分布这一基础分布假设可以扩展为负二项分布和零膨胀泊松分布，以捕捉计数数据中的零膨胀和过度离散特征。与传统泊松模型类似，DPSS模型将索赔数量视为响应变量，并采用对数连接函数将响应变量与协变量联系起来。不同于传统泊松模型，DPSS模型中的时变系数采用平滑样条建模，类似于Wecker 和 Ansley(1983)。所提模型中的平滑参数通过训练数据的最大似然法估计。理赔频率的向前 K 步预测基于一种类似于卡尔曼滤波器的算法，因此当有新的索赔数据时，可以动态更新平滑参数、模型系数以及预测结果。所提方法的优势在于，当有新的索赔数据时，只需针对新增观测值进行更新；而在标准状态空间模型中，每当有新的观测值时都要更新模型。由于计数型结果的滤波步骤通常涉及拉普拉斯近似，逐次更新一个观测值会导致较大的累积近似误差。此外，在实际应用中，逐条数据更新模型通常既不可行，也非理想做法。因此，我们提出在预先设定的固定时间区间更新模型，每次在一批理赔数据可用时进行更新。

尽管精算科学文献中已有一支研究方向探讨状态空间模型在死亡率预测（Chang and Shi, 2020年）和非寿险定价（Ahn 等人，2022）中的应用，但很少有研究讨论如何将状态空间模型嵌入变系数模型，以进行未来预测。据我们所知，这是首项在理赔频率建模框架下研究动态预测模型的研究。通过采用时变系数模型并运用状态空间建模工具，我们的 DPSS 模型显著填补了动态理赔频率建模与预测领域的空白。所提方法具有较高的计算效率，因此可用于在线监测。还应指出，尽管所提方法旨在预测理赔频率，但只要假设响应变量服从其他连续重尾分布（如伽马分布）或半连续分布（如复合泊松分布或 Tweedie 分布），便可轻松将其扩展至理赔严重度和纯保费的预测（Jørgensen

和 Smyth，1994）。

为展示所提方法的应用，我们分析了一个新的中国汽车保险数据集。该数据集涵盖2010年至2015年约440,000份保单的理赔历史和自报风险特征。我们的目标是从一个参数静态的费率厘定模型出发，通过引入时变系数，建立一种动态机制来调整费率因子的基准价格。强有力的实证证据表明，该模型优于静态预测模型。此外，我们还展示了 DPSS 模型估计得出的时变效应的实证证据，包括保险年限、汽车用途、汽车座位数、车辆产地、保险保障、性别、保单类型 和发动机排量 [1]。大多数费率因子的估计回归系数均呈现出2012年之后的下降趋势。我们认为，这些动态模式可能是由中国汽车监管政策的变化所致：监管机构收紧了保险公司设定保险费的权限，旨在推动汽车保险费率标准统一。因此，识别并理解时变效应及其成因，对理论研究和精算费率厘定实践至关重要。

本文的结构如下。在第 2 节中，我们首先概述部分变系数模型和状态空间模型，然后介绍一类新的动态泊松状态空间模型。第 3 节讨论估计和预测程序。第4节通过模拟研究，展示 DPSS 模型相较于保险费率厘定文献中几种成熟方法的优势。为说明其实际应用，我们在第5节对 DPSS 模型进行分析，该模型拟合于一个涵盖近六年时间段的中国汽车保险数据集。第 6 节给出一些结论，并探讨未来可能的扩展方向。

2 方法论

2.1 部分变系数模型

经典的广义线性模型无法捕捉协变量的时变效应。更复杂的模型是变系数模型（Hastie 和 Tibshirani，1993年）。它们允许回归系数随另一个变量（例如时间）以平滑的方式变化。设（y_i、$\boldsymbol{x}_i$、t_i）为每个受试者在观测时间 $i = 1$、$\cdots$、n 所观测到的数据 i，其中 t_i，$0 \leq t_1 < t_2 < \cdots < t_n \leq 1$、$\boldsymbol{x}_i = (x_{i1}$、$\cdots$、$x_{iq})^T$ 表示一个 q维协变量向量，而 y_i 是

[1] 这些协变量的描述可见表 4

保险索赔次数（索赔数量）。我们针对 y_i 提出如下模型，其中 λ_i 是强度（预期索赔频率），即:

$$y_i \sim \text{Poisson}(\lambda_i),$$

with

$$\begin{aligned}\log(\lambda_i) &= \beta_0(t_i) + \sum_{j=1}^{q_1} \beta_j(t_i) x_{ij} + \sum_{k=1}^{q_2} \alpha_k x_{i,q_1+k} \\ &= \boldsymbol{x}_{i1}^T \boldsymbol{\beta}(t_i) + \boldsymbol{x}_{i2}^T \boldsymbol{\alpha}, \end{aligned} \tag{2.1}$$

其中，$\boldsymbol{x}_{i1} = (1, x_{i1}, \cdots, x_{iq_1})^T$和 $\boldsymbol{x}_{i2} = (x_{i,q_1+1}, \cdots, x_{i,q})^T$分别是具有时变系数$\boldsymbol{\beta}(t) = (\beta_0(t)\ \beta_1(t)\ \cdots, \beta_{q_1}(t))^T$和时不变系数的协变量 $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_{q_2})^T$，且 $q_1 + q_2 = q$。自 Hastie 和 Tibshirani（1993）提出变系数回归模型以来，统计学文献中对其进行了大量研究；全面综述可参见 Fan 和 Zhang（2008）及 Park 等人（2015）。然而，现有文献大多关注估计和统计推断问题，而非预测。

注1. 在汽车保险定价的背景下，我们通常假设索赔数量 Y_i服从均值为 $\mathbb{E}(Y_i) = w_i\lambda_i$的泊松分布，其中 w_i 是风险暴露量，而 $\lambda_i = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})$。于是有 $\log \mathbb{E}(Y_i) = \log(w_i) + \boldsymbol{x}_i^T\boldsymbol{\beta}$，其中 $\log(w_i)$ 可视为系数固定为 1 的协变量。在（2.1）中，我们将 $\log(w_i)$ 视为具有未知系数的协变量。因此，传统的泊松回归模型可视为（2.1）的一个特例。

注释2。尽管泊松分布是对计数数据建模的自然选择，但人们通常会发现，保险索赔数据中的零值数量超过了泊松模型所预测的零值数量。为刻画索赔计数分布中零值过多的现象，通常使用 Lambert(1992年)提出的零膨胀泊松（ZIP）模型拟合汽车保险索赔数据（Yip 和 Yau，2005）。为处理过度离散问题（此时，泊松分布固有的期望与方差相等这一假设必须改为方差大于期望），负二项分布通过引入一个决定分布尺度的第二个参数，为扩展泊松分布提供了一个便捷的框架，相关例子见Hilbe (2011年）。利用广为人知的负二项（NB）和零膨胀泊松（ZIP）分布，我们进一步将动态泊松状态空间模型（DPSS）扩展为动态负二项状态

空间模型（DNBSS）和动态零膨胀泊松状态空间模型（DZIPSS）。模型设定和估计程序的详细内容见附录A和B 。

参数可以通过最大化以下惩罚对数似然函数来估计：

$$L_c\left(\boldsymbol{\alpha}, \boldsymbol{\beta}(t); \boldsymbol{y}\right) = \sum_{i=1}^{n} \left\{ y_i \left[\boldsymbol{x}_{i1}^T \boldsymbol{\beta}\left(t\right) + \boldsymbol{x}_{i2}^T \boldsymbol{\alpha} \right] - \exp\left[\boldsymbol{x}_{i1}^T \boldsymbol{\beta}\left(t_i\right) + \boldsymbol{x}_{i2}^T \boldsymbol{\alpha} \right] - \log\left(y_i!\right) \right\} - \frac{1}{2} \sum_{j=0}^{q_1} \tau_j \int \left[\beta''_j\left(t\right) \right]^2 dt, \tag{2.2}$$

ˆ 其中 τ_j 是平滑参数。众所周知，使 $\beta_j(t)$ 关于 $j = 0$、1，$\cdots$，q_1取得最小值的函数是三次平滑样条。对于来自指数族的数据，Gu (1992年建立了平滑样条模型与贝叶斯模型之间的联系，据此可通过高斯随机过程的后验估计得到平滑样条的估计。遵循 Gu 的贝叶斯方法，$\beta_j(t)$可建模为

$$\beta_j\left(t\right) = B_{1j} + B_{2j} t + b_j^{1/2} \int_0^t W_j\left(s\right) ds, \quad j = 0, \cdots, q, \tag{2.3}$$

其中，B_{1j}和 B_{2j}服从弥散先验 $[B_{1j}, B_{2j}]^T \sim N(0, \tau I)$，其中 $\tau \to \infty$；$W_j(s)$为维纳过程，b_j 为平滑参数。在弥散先验下，当 $b_j = 1/\tau_j$ 时，近似后验均值就是平滑样条估计量 $\beta_j(t)$；若通过拉普拉斯方法近似后验分布，则这一结果成立（Gu，1992年）。Wahba（1983）进一步指出，随机模型（2.3）可以用函数 $\beta_j(t_i)$及其一阶导数 $\beta'_j(t_i)$构成的状态向量表示为：

$$\begin{bmatrix} \beta_j(t_i) \\ \beta'_j(t_i) \end{bmatrix} = T_i \begin{bmatrix} \beta_j(t_{i-1}) \\ \beta'_j(t_{i-1}) \end{bmatrix} + U_j(t_i, t_{i-1}),$$
$$T_i = \begin{bmatrix} 1 & t_i - t_{i-1} \\ 0 & 1 \end{bmatrix}, \tag{2.4}$$

其中

$$U_j(t_i, t_{i-1}) \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tau_j^{-1} Q_i \right), \; Q_i = \begin{bmatrix} (t_i - t_{i-1})^3/3 & (t_i - t_{i-1})^2/2 \\ (t_i - t_{i-1})^2/2 & (t_i - t_{i-1}) \end{bmatrix}. \tag{2.5}$$

参数 τ_j 是平滑参数，用于控制平滑度与偏差之间的权衡。当 $\tau_j^{-1}=0$， $\beta_j(t)$ 被限制为一条截距为 $\beta_j(0)$、斜率为 $\beta_j'(0)$ 的直线。因此，模型 (2.1) 可用状态空间表示为

$$\begin{aligned}\log(\lambda_i) &= \boldsymbol{z}_i^T\boldsymbol{\gamma}(t_i), \quad i=1,\cdots,n\\ \boldsymbol{\gamma}(t_i) &= \boldsymbol{T}_i\boldsymbol{\gamma}(t_i)+\boldsymbol{\eta}(t_i,t_{i-1}), \quad \boldsymbol{\eta}(t_i,t_{i-1})\sim N(\mathbf{0},\boldsymbol{Q}_i),\end{aligned} \tag{2.6}$$

其中

$$\begin{aligned}\boldsymbol{z}_i &= (1,0,x_{i1},0,x_{i2},\cdots,x_{iq_1},0,x_{i,q_1+1},\cdots,x_{iq})^T,\\ \boldsymbol{\gamma}(t) &= \left(\beta_0(t),\beta'_0(t),\beta_1(t),\beta'_1(t),\cdots,\beta_{q_1}(t),\beta'_{q_1}(t),\alpha_1,\cdots,\alpha_{q_2}\right)^T,\\ \boldsymbol{\eta}(t_i,t_{i-1}) &= \left(U_0^T(t_i,t_{i-1}),\cdots,U_{q_1}^T(t_i,t_{i-1})\right)^T,\\ \boldsymbol{T} &= \operatorname{diag}\left(\overbrace{T,\cdots,T}^{q_1+1},\overbrace{1,\cdots,1}^{q_2}\right),\\ \boldsymbol{Q} &= \operatorname{diag}\left(\overbrace{\lambda_0^{-1}Q,\cdots,\lambda_{q_1}^{-1}Q}^{q_1+1},\overbrace{0,\cdots,0}^{q_2}\right).\end{aligned}$$

这里，我们将模型（2.6）称为动态泊松状态空间（DPSS）预测模型。变动系数与常系数的模型设定差异体现在转移矩阵 $\boldsymbol{T}_i$ 和方差矩阵 $\boldsymbol{Q}_i$中。 $\boldsymbol{T}_i$ 中与 $\boldsymbol{\alpha}$相对应的单位转移矩阵以及 $\boldsymbol{Q}_i$ 中的零方差意味着 $\boldsymbol{\alpha}$ 是常系数。在状态空间模型中，每当有新的观测值可用时，系数估计值都会随之更新。对于非高斯状态空间模型，序贯更新通常采用数值积分或正态近似。对计数分布进行数值近似会产生较大的近似误差，而且误差可能随时间累积。此外，在实际应用中，逐个样本更新模型往往既不可行，也不可取。因此，我们建议在固定时间区间内利用批量数据更新估计值，这有助于获得更准确的近似。

2.2 使用批量数据进行在线监测的模型设定

我们首先将时间域 $[0,1]$ 划分为 S 个等间隔区间 $[(s-1)/S,\ s/S]$ 其中 $s=1,\ \cdots,\ S$。令 $\boldsymbol{y}_s=\{y_{s_m},m=1,2,\cdots,n_s\}$ 为满足 $t_{s_m}\in[(s-1)/S,\ s/S]$的观测值 y_{s_m} 的集合，并且

~对应于 $\boldsymbol{x}_s = \{x_{s_m}, m = 1, 2, \cdots, n_s\}$。令 $t = (2s-1)/2S$ 为第 s 个区间的中点。我们使用批量数据 $(\boldsymbol{x}_s,\ \boldsymbol{y}_s)$ 更新模型。具体而言，我们在中点 t 处评估区间 $\sim[(s-1)/S,\ s/S]$ 内的数据。因此，批量数据的状态空间模型可表示为

$$\begin{aligned}
\log\left(\lambda_{s_m}\right) &= \boldsymbol{z}_{s_m}^T \boldsymbol{\gamma}\left(\tilde{t}_s\right), \quad s = 1, \cdots, S,\ m = 1, 2, \cdots, n_s \\
\boldsymbol{\gamma}\left(\tilde{t}_s\right) &= \tilde{\boldsymbol{T}} \boldsymbol{\gamma}\left(\tilde{t}_{s-1}\right) + \boldsymbol{\eta}\left(\tilde{t}_s, \tilde{t}_{s-1}\right), \quad \boldsymbol{\eta}\left(\tilde{t}_s, \tilde{t}_{s-1}\right) \sim N\left(\mathbf{0}, \tilde{\boldsymbol{Q}}\right),
\end{aligned} \tag{2.7}$$

其中

$$\begin{aligned}
\boldsymbol{z}_{s_m} &= (1, 0, x_{s_m 1}, 0, x_{s_m 2}, \cdots, x_{s_m q_1}, 0, x_{s_m, q_1+1}, x_{s_m, q_1+2} \cdots, x_{s_m, q})^T, \\
\boldsymbol{\gamma}(t) &= \left(\beta_0(t), \beta'_0(t), \beta_1(t), \beta'_1(t), \cdots, \beta_{q_1}(t), \beta'_{q_1}(t), \alpha_1, \cdots, \alpha_{q_2}\right)^T, \\
\tilde{\boldsymbol{T}} &= \operatorname{diag}\left(\overbrace{\tilde{T}, \cdots, \tilde{T}}^{q_1+1}, \overbrace{1, \cdots, 1}^{q_2}\right), \tilde{T} = \begin{bmatrix} 1 & 1/S \\ 0 & 1 \end{bmatrix}, \\
\tilde{\boldsymbol{Q}} &= \operatorname{diag}\left(\overbrace{\lambda_0^{-1}\tilde{Q}, \cdots, \lambda_{q_1}^{-1}\tilde{Q}}^{q_1+1}, \overbrace{0, \cdots, 0}^{q_2}\right), \tilde{Q} = \begin{bmatrix} (1/S)^3/3 & (1/S)^2/2 \\ (1/S)^2/2 & 1/S \end{bmatrix}.
\end{aligned}$$

3 模型估计与预测

3.1 估计与预测

由于状态空间模型具有递归性质，对于给定的平滑参数，可以使用由每个时间点的预测步骤和滤波步骤组成的卡尔曼滤波算法，高效地估计模型系数。还可以使用相同的算法计算并最大化似然，以估计平滑参数。在获得平滑参数的估计值后，可以通过运行卡尔曼滤波器的预测步骤而不运行滤波步骤，进行 K 步超前预测。当新数据可用时，可以从原始估计值开始，在新的观测值上运行该算法，从而高效地更新系数。我们首先概述该算法的整体流程，其中包括以下步骤：

（A）在训练数据集上拟合模型，

（a）根据似然选择平滑参数。

(b) 使用卡尔曼滤波算法估计系数。

(B) 卡尔曼滤波器的前向 K 步预测。

(C) 当有新数据到来时：

(a) 根据所有可用数据的似然更新平滑参数。

(b) 执行滤波步骤，以估计当前时间点的系数。

(D) 重复执行 (B)-(C)。

注释 3. 在步骤 (C) 中，每当有新数据可用时，都可以更新平滑参数。但是，由于平滑参数控制的是函数的平滑度，而平滑度是函数的一项全局特征，通常不会频繁变化。因此，我们建议每隔一段时间更新一次平滑参数。

接下来，我们将介绍卡尔曼滤波器、平滑参数的最大似然估计以及动态预测步骤。

3.2 估计

本节介绍如何使用卡尔曼滤波算法拟合状态空间模型。设$\boldsymbol{Y}^s=(\boldsymbol{y}_1,\ \cdots,\ \boldsymbol{y}_s)$和 $\boldsymbol{X}^s=(\boldsymbol{x}_1,\ \cdots,\ \boldsymbol{x}_s)$分别为截至时间 $t=s/S$的结果变量和协变量的历史观测值集合。给定平滑参数 $\lambda_j,\ \ j=0,\ 1,\ \cdots,\ q_1$后，该算法依次重复预测步骤和滤波步骤来估计系数。为简化记号，$\widetilde{}$表示 $\boldsymbol{\gamma}_s\equiv\boldsymbol{\gamma}(t_s)$。模型拟合的一般算法按以下步骤进行：

(i) 从 $s=0$开始，取弥散先验 $p(\boldsymbol{\gamma}_0|\boldsymbol{Y}^0)\sim N(\mathbf{0},\ c\mathbf{I})$作为 $\boldsymbol{\gamma}_1$的初始先验分布，其中 c 是一个较大的常数。

(ii) 预测步骤：计算

$$\mathbb{E}(\boldsymbol{\gamma}_s|\boldsymbol{Y}^{s-1},\boldsymbol{X}^{s-1})=\widetilde{\mathbf{T}}\mathbb{E}(\boldsymbol{\gamma}_{s-1}|\boldsymbol{Y}^{s-1},\boldsymbol{X}^{s-1}),$$
$$\mathrm{Var}(\boldsymbol{\gamma}_s|\boldsymbol{Y}^{s-1},\boldsymbol{X}^{s-1})=\widetilde{\mathbf{T}}\mathrm{Var}(\boldsymbol{\gamma}_{s-1}|\boldsymbol{Y}^{s-1},\boldsymbol{X}^{s-1})\widetilde{\mathbf{T}}'+\widetilde{\mathbf{Q}}.$$

(iii) 滤波步骤：通过正态近似计算后验均值 $\mathbb{E}(\boldsymbol{\gamma}_s|\boldsymbol{Y}^{s-1},\ \boldsymbol{X}^{s-1},\ \boldsymbol{y}_s,\ \boldsymbol{x}_s)$和方差Var($\boldsymbol{\gamma}_s|\boldsymbol{Y}^{s-1},\ \boldsymbol{X}^{s-1},\ \boldsymbol{y}_s,\ \boldsymbol{x}_s$)。

(iv) 依次重复步骤 ii) 和 iii)，其中 $s=1,\ 2,\ \cdots,\ S$。

在滤波步骤中，系数 $\boldsymbol{\gamma}_s$ 的后验分布可写为

$$\begin{aligned} p\left(\boldsymbol{\gamma}_s|\boldsymbol{Y}^{s-1},\boldsymbol{X}^{s-1},\boldsymbol{y}_s,\boldsymbol{x}_s\right) &\propto p\left(\boldsymbol{y}_s|\boldsymbol{\gamma}_s,\boldsymbol{x}_s\right)p(\boldsymbol{\gamma}_s|\boldsymbol{Y}^{s-1},\boldsymbol{X}^{s-1}) \\ &= \prod_{i=s_1}^{s_{n_s}} p\left(y_i|\boldsymbol{\gamma}_s,x_i\right)\cdot p\left(\boldsymbol{\gamma}_s|\boldsymbol{Y}^{s-1},\boldsymbol{X}^{s-1}\right). \end{aligned} \tag{3.1}$$

这里 $p(\boldsymbol{\gamma}_s|\boldsymbol{Y}^{s-1},\ \boldsymbol{X}^{s-1})$ 服从均值和方差分别为 $\mathbb{E}(\boldsymbol{\gamma}_s|\boldsymbol{Y}^{s-1},\ \boldsymbol{X}^{s-1})$ 和 $\mathrm{Var}(\boldsymbol{\gamma}_s|\boldsymbol{Y}^{s-1},\ \boldsymbol{X}^{s-1})$ 的正态分布。然而，由于 $p\left(\boldsymbol{y}_s|\boldsymbol{\gamma}_s,\ \boldsymbol{x}_s\right)$ 是泊松分布的乘积，后验分布（3.1）没有闭式表达式。我们用正态分布近似后验分布，其中近似正态分布的均值为后验分布的众数。

记为

$$\varphi\left(\boldsymbol{\gamma}_s\right)=\sum_{i=s_1}^{s_{n_s}}\log p\left(y_i|\boldsymbol{\gamma}_s,x_i\right)+\log p\left(\boldsymbol{\gamma}_s|\boldsymbol{Y}^{s-1},\boldsymbol{X}^{s-1}\right). \tag{3.2}$$

可使用牛顿-拉夫森方法最大化 $\varphi(\boldsymbol{\gamma}_s)$，以求得其众数。初始值取为 $\boldsymbol{\gamma}_s^{(0)}=\mathbb{E}(\boldsymbol{\gamma}_s|\boldsymbol{Y}^{s-1},\ \boldsymbol{X}^{s-1})$。具体过程是迭代以下方程直至收敛，第 k 次迭代为

$$\boldsymbol{\gamma}_s^{(k)}=\boldsymbol{\gamma}_s^{(k-1)}-\left\{D^2\varphi\left(\boldsymbol{\gamma}_s^{(k-1)}\right)\right\}^{-1}D\varphi\left(\boldsymbol{\gamma}_s^{(k-1)}\right),$$

其中 $D\varphi$ 和 $D^2\varphi$ 分别为一阶和二阶导数算子。设 $\boldsymbol{\gamma}_s^*$ 为收敛值，则后验均值和方差为

$$\begin{aligned} &\mathbb{E}(\boldsymbol{\gamma}_s|\boldsymbol{Y}^{s-1},\boldsymbol{X}^{s-1},\boldsymbol{y}_s,\boldsymbol{x}_s)=\boldsymbol{\gamma}_s^*, \\ &\mathrm{Var}(\boldsymbol{\gamma}_s|\boldsymbol{Y}^{s-1},\boldsymbol{X}^{s-1},\boldsymbol{y}_s,\boldsymbol{x}_s)=\{-D^2\ell(\boldsymbol{\gamma}_s^*)\}^{-1}. \end{aligned} \tag{3.3}$$

3.3 平滑参数的选择

在该模型中，平滑参数 λ_j 分别控制 $\beta_j(t)$ $j=0$、1、$\cdots$、q_1 的平滑度。λ_j 的选择对模型拟合和预测起着关键作用。我们提出通过最大化以下对数似然来估计平滑参数，

$$(\widehat{\lambda}_0,\cdots,\widehat{\lambda}_{q_1})=\arg\max_{\lambda_0,\cdots,\lambda_{q_1}}\sum_{s=1}^{S}\sum_{i=s_1}^{s_{n_s}}\log p(y_i|\boldsymbol{Y}^{s-1},\boldsymbol{X}^{s-1}),$$

其中

$$p(y_i|\boldsymbol{Y}^{s-1},\boldsymbol{X}^{s-1}) = \int p(y_i|\boldsymbol{\gamma}_s,\boldsymbol{Y}^{s-1},\boldsymbol{X}^{s-1})p(\boldsymbol{\gamma}_s|\boldsymbol{Y}^{s-1},\boldsymbol{X}^{s-1})d\boldsymbol{\gamma}_s,$$

且无法以闭式形式求得。一种常用方法是拉普拉斯近似。Lewis 和 Raftery (1997b)认为这种近似相当准确。拉普拉斯近似得出

$$p(y_i|\boldsymbol{Y}^{s-1},\boldsymbol{X}^{s-1}) \approx (2\pi)^{(q+q_1+2)/2}\left|\{-D^2\ell_i(\boldsymbol{\gamma}_s^*)\}^{-1}\right|^{1/2} p(y_i|\boldsymbol{\gamma}_s^*,\boldsymbol{Y}^{s-1},\boldsymbol{X}^{s-1})p(\boldsymbol{\gamma}_s^*|\boldsymbol{Y}^{s-1},\boldsymbol{X}^{s-1})$$

其中 $\ell_i(\boldsymbol{\gamma}) = \log\, p(y_i|\boldsymbol{\gamma},\boldsymbol{Y}^{s-1},\boldsymbol{X}^{s-1}) + \log\, p\left(\boldsymbol{\gamma}|\boldsymbol{Y}^{s-1},\boldsymbol{X}^{s-1}\right)$。尽管每当有新数据可用时都可以更新平滑参数，但在实践中我们通常不会频繁更新它们，因为曲线的平滑度往往不会随时间发生太大变化。

3.4 K 步超前预测

~时间点 t_S 的 K 步超前预测模型系数如下

$$\begin{aligned}
&\mathbb{E}(\boldsymbol{\gamma}_{S+K}|\boldsymbol{Y}^S,\boldsymbol{X}^S) = \widetilde{\mathbf{T}}^K E(\boldsymbol{\gamma}_S|\boldsymbol{Y}^S,\boldsymbol{X}^S),\\
&\mathrm{Var}(\boldsymbol{\gamma}_{S+K}|\boldsymbol{Y}^S,\boldsymbol{X}^S) = \widetilde{\mathbf{T}}^K\mathrm{Var}(\boldsymbol{\gamma}_S|\boldsymbol{Y}^S,\boldsymbol{X}^S)\widetilde{\mathbf{T}}'^K + \widetilde{\mathbf{T}}^{K-1}\widetilde{\mathbf{Q}}\widetilde{\mathbf{T}}'^{K-1} + \cdots + \widetilde{\mathbf{T}}\widetilde{\mathbf{Q}}\widetilde{\mathbf{T}}' + \widetilde{\mathbf{Q}}.
\end{aligned}$$

随后可以使用这些系数和预测期望，计算给定协变量下的预测强度。

4 模拟

本节中，我们使用模拟数据评估所提模型的预测性能。我们将所提方法与所有系数均为常数的广义线性模型进行比较（分别称为泊松、负二项和零膨胀泊松方法，对应泊松、负二项和零膨胀泊松分布）。

我们使用以下模型模拟计数型结果变量，

$$P(Y_i = k) = \frac{\lambda_i^k e^{-\lambda_i}}{k!},\ \log(\lambda_i) = \beta_0(t_i) + \beta_1(t_i)X_{i1} + \alpha X_{i2},$$

其中， t_i 由均匀分布 $U[0,1]$ 生成， X_{i1} 和 X_{i2} 则分别独立地由均匀分布 $U[0,1]$生成。 $\beta_0(t)$是由 $\beta_0(t) = t-2$，生成的时变截距，而 $\beta_1(t)$ 是 X_{i1} 的时变系数，由 $\beta_1(t) = 0.2\log(t) + 0.5$ 生成；$\alpha = 0.25$ 是常系数。随时间变化的平均强度（预期索赔频率），即$\frac{1}{n}\sum_{i=1}^{n}\lambda_i$，约为0.24。我们在样本量为 $n = 10^5$的情况下评估模型性能。前75%的数据用于拟合模型，剩余25%的数据用于测试模型性能。

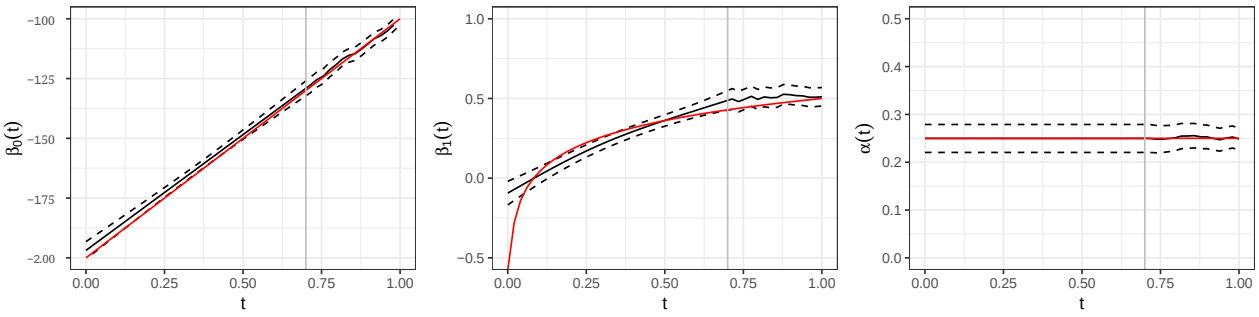


图1：使用所提方法得到的截距（左图）、 X_1（中图）和 X_2（右图）系数及其95%置信带。样本量为 $n = 10^5$。红线表示各参数的真实值。系数图在 $t = 0.75$处分界，分别对应训练数据中的估计系数和验证数据中的一步超前预测。

对于我们提出的 DPSS 模型，我们首先将数据划分为若干批次。在该数据集中，我们将时间划分为 50 个等间隔的时间区间， $[(s-1)/50, s/50], s = 1, 2, \cdots, 50$。每个时间区间平均包含 2000 名受试者，对应的总样本量为 10^5。我们采用正态分布 $N(0, 100 \cdot I_5)$ 作为模型系数的初始先验分布，其中 I_5 为 5 维单位矩阵。我们使用第 3.3 节提出的准则，基于前 75% 的数据为 $\beta_0(t)$ 和 $\beta_1(t)$ 选择平滑参数，分别记为 λ_0 和 λ_1 。在剩余数据上进行预测时， λ_0 和 λ_1 均固定不变。

我们比较了所提方法得到的各个系数的估计函数曲线及其对应的95%置信带。图1 展示了截距（左图）、 X_1（中图）和 X_2（右图）的估计系数及其对应的95%置信带。系数图在 $t = 0.75$ 处由垂直虚线分隔，分别对应训练数据和验证数据。图中展示了训练集中的估计系数（$t \leq 0.75$）以及验证集中的一步超前预测系数（$t > 0.75$）。通过所提方法得到的估计系数曲线

表 1：模拟数据集中的测试数据的泊松偏差。

模型	泊松	NB	ZIP	DPSS
偏差	0.9354	0.9354	0.9178	0.8557

接近真实曲线（红线）。此外，95%置信带覆盖了训练数据和验证数据中的真实值。

我们使用两个不同的指标来评估预测索赔结果的准确性。第一个是泊松偏差，用于衡量拟合模型 ¡¯的预测与观测结果的接近程度。其定义为

$$D(y_i, \hat{y}_i) = 2\sum_{i=1}^{n} \left[y_i \log(y_i/\hat{y}_i) - (y_i - \hat{y}_i) \right], \tag{4.1}$$

其中，ˆy_i 表示基于估计模型参数对观测 i 的预测均值。泊松偏差越小，表示模型拟合越好。第二个指标是预测数据与观测数据之间的差异，其定义为

$$D(\hat{\lambda}_i, y_i) = \sum_{i=1}^{n_{test}} I(y_i = k) - \sum_{i=1}^{n_{test}} \frac{\hat{\lambda}_i^k e^{-\hat{\lambda}_i}}{k!},\ k = 0, 1, 2, \cdots,$$

ˆ 其中 n_{test} 为测试数据的样本量， λ_i 为相应的预测索赔频率。表 1和表 2 显示，与静态模型（泊松、负二项和零膨胀泊松）相比，我们所提方法的预测性能显著更优，因为静态模型无法捕捉随时间变化的效应，因此表现不佳，这也在意料之中。

表 2：模拟数据集测试数据中预测理赔频率与观测理赔频率的差异。

计数（k）	泊松	NB	ZIP	DPSS
0	3015	3055	3055	24
1	-2139	-2209	-2213	-58
2	-750	-727	-720	9
3	-112	-105	-107	21
4	-14	-13	-13	3
6	-1	-1	-1	-1

5 实证分析

我们利用一个来自中国的新型汽车保险数据集检验 DPSS模型的性能。该数据集涵盖 2010年至2015年收集的数据，包括约 440,000份保险合同的自报风险特征和索赔历史（含索赔数量）。索赔数据集来自三种类型的汽车保险，包括碰撞险、第三者责任险和交强险。表 3 展示了索赔数量随时间的分布。我们观察到，大多数保单持有人没有索赔（82.10%），一部分有一到两次索赔（16.4%），其余保单持有人的索赔次数超过三次。未发生索赔的概率逐年上升，尤其是在2014年至2015年间。我们认为，中国监管机构自2007年至2015年对汽车保险行业实施了全面的第二轮改革，随后又于2015年至2020年实施了第三轮改革，改革涉及定价、费用结构和产品保障范围等方面，对索赔数量的分布产生了重大影响。尤其在2014至2015年前后，中国通过扩大保险保障范围并采用统一的费率厘定机制，启动了汽车保险改革，以解决中国汽车保险市场“高保费、低赔付”的问题。监管政策的变化促使我们在预测样本期之后的理赔频率时，将时间趋势纳入统计模型。

该数据集中个人保单风险特征的信息包括：保险年限，性别，年龄，车龄，汽车价值，保单类型，汽车用途，发动机排量，汽车座位数 和保险保障。这些变量的汇总统计量见表 4。图2 展示了汽车保险数据集中的协变量在各年份的分布情况。大多数保单持有人（70.90%）在整个保单年度内都面临提出索赔的风险；其余保单的保险年限介于0和1之间，这表明这些保单的承保日期可能晚于年初，或者合同在年末前退保。几乎所有保单持有人（89.12%）的年龄在25至60岁之间，这意味着保险组合中的年轻和年长保单持有人较少，且超过半数的保单持有人（75.96%）为女性。大多数保单（78.59%）是续保保单或从其他保险公司转入的保单。超过半数的车辆（76.20%）排量小于1.6升；半数车辆（50.73%）为国产车；79.05%的车辆载客座位数少于6个，66.57%的车辆为轿车。超过半数的保单（60.60%）未涵盖以下三种保险：碰撞险、第三者责任险和强制交通

保险。

Ta表 3：汽车保险数据集中各年份索赔数量的分布 .

计数	按年份划分的百分比						
	2010	2011	2012	2013	2014	2015	总体
0	77.20%	77.90%	78.70%	78.70%	82.10%	94.70%	82.10%
1	15.00%	14.80%	14.20%	14.50%	13.60%	4.70%	12.60%
2	5.10%	5.00%	5.00%	5.00%	3.50%	0.50%	3.80%
3	1.70%	1.60%	1.50%	1.40%	0.70%	0.10%	1.10%
4+	1.00%	0.70%	0.50%	0.40%	0.20%	0.00%	0.40%
obs.	39,103	51,604	75,298	91,418	106,076	79,012	442,511

表4：汽车保险数据集中风险特征的描述。

变量	**Type**	**描述**
expo	连续	按有效保单期限计量的风险暴露量：0-1
age	连续	保单持有人的年龄：18-99
车龄	连续	车辆年龄：0-13
汽车价值	连续	车辆购买价格的对数变换：9.8（最小值）-16（最大值）
保单类型	分类变量	保单类型（两个类别）：新车的新承保保单（新车），来自其他保险公司的续保保单或转入保单（续保或转保）
汽车用途	分类变量	车辆类别（两个类别）：轿车 和非轿车
车辆产地	分类变量	两个类别：国产 或 非国产 车辆
发动机排量	分类变量	车辆排量（两个类别）：$> 1.6L$ 或 $\leq 1.6L$
汽车座位数	分类变量	车辆座位数（两个类别）：≥ 6 座或 < 6 座
保险保障	分类变量	车辆保险保障范围（两个类别）：承保碰撞险、第三者责任险和交强险（全部保障），或不承保 _ 三种保险（其他保障） _
性别	分类变量	男性或女性

我们首先拟合一个DPSS模型，以描述索赔数量随时间变化的情况，并允许截距和上述所有预测变量的系数随时间变化。具体而言，按年份将数据集分为六个批次。前五个批次（2010—2014年的数据，记为 $\mathcal{L}(Y_i(k)\ \boldsymbol{x}_i(k))_{0\leq i\leq n},\ k$ =2010，.....，2014）用作模型拟合和平滑参数选择的训练数据；第六个批次（2015年的数据，记为 $\mathcal{L}(Y_i(k)\ \boldsymbol{x}_i(k))_{0\leq i\leq n},\ k = 2015$）用作

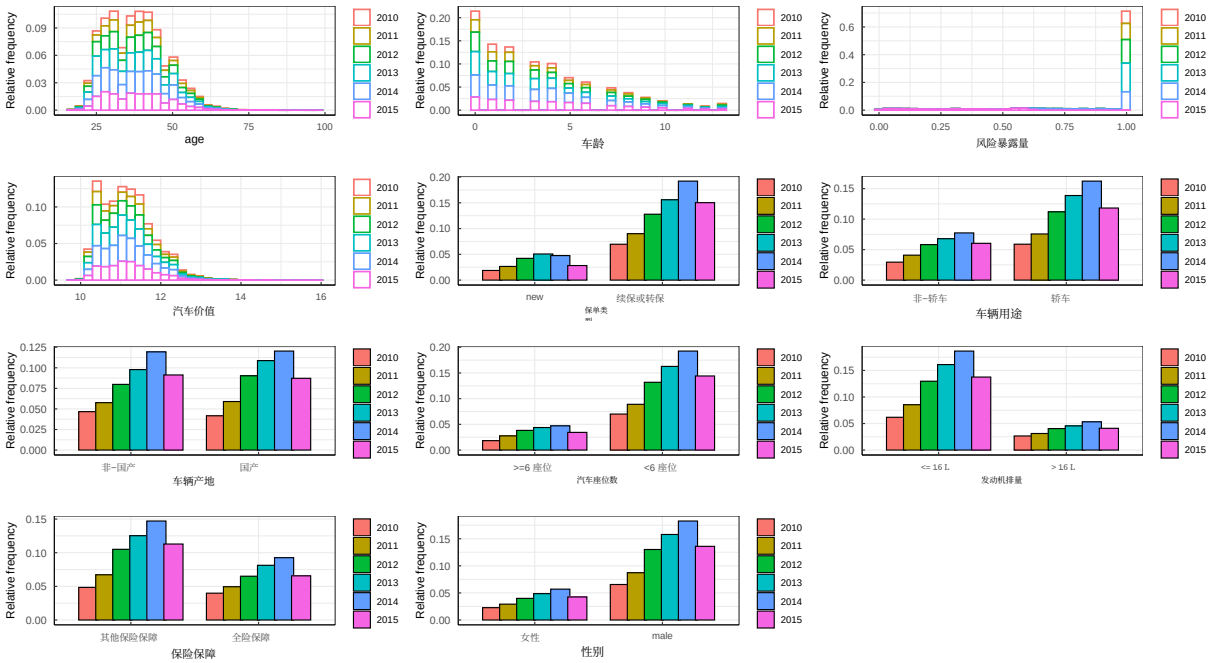


图2：协变量相对频率保险年限、性别、年龄、车龄、汽车价值、保单类型、汽车用途，车辆产地、发动机排量、汽车座位数 和保险保障随年份变化的情况。

用于样本外分析的测试数据。初步分析表明，只有年龄、车龄和汽车价值的系数随时间保持不变，而其他系数则随时间变化。现在，我们考虑以下 DPSS：

$$\text{对数 } \lambda_i(t_i) = \beta_0(t_i) + \beta_1(t_i) \text{ 对数 保险年限 } + \beta_2(t_i)\text{车辆用途}_i + \beta_3(t_i)\text{汽车座位数}_i + \beta_4(t_i)\text{保单类型}_i + \beta_5(t_i)\text{保险保障}_i + \beta_6(t_i)\text{性别}_i + \beta_7(t_i)\text{发动机排量}_i + \beta_8(t_i)\text{车辆产地}_i + \alpha_1\text{年龄}_i + \alpha_2\text{车龄}_i + \alpha_3\text{汽车价值}_i.$$

请注意，分类变量（性别、保单类型、汽车用途、车辆产地、发动机排量、汽车座位数 和保险保障）采用二元编码。为展示所提出模型的优越性，我们还考虑了五种竞争回归模型：泊松回归模型（泊松）、负二项回归模型（NB）、零膨胀泊松回归模型（ZIP），以及动态负二项状态空间模型（DNBSS）和动态零膨胀泊松状态空间模型（DZIPSS）。在静态模型中，将年份 t_i 作为连续协变量引入均值参数，以捕捉历年的时间趋势。泊松分量的均值和额外零值的概率

均假设为 ZIP 模型中协变量的函数。 el.

图3(a)-(i)展示了 DPSS 模型中截距和八个协变量的时变系数估计值及其 95%置信带，这些结果对应训练数据中的 2010—2014年和测试数据中的 2015年。从图 3(c)-(e)可以看出，轿车的理赔频率高于非轿车；座位数少于六个的车辆比座位数多于六个的车辆索赔更频繁；国产车辆的索赔数量多于非国产车辆。轿车与非轿车之间的理赔频率差异从 2010年到 2011年略有下降，随后从 2011年到 2013年上升，并在 2013年后下降。图 3(d)和(e)也呈现出类似的时间趋势。从图 保险保障变量的时间图（图 3(f)）可以看出，高保险保障保单与低保险保障保单之间的理赔频率差异在 2012年之前呈上升趋势，之后则呈下降趋势。估计系数为正，表明高保险保障保单的理赔频率相对较高。这可能有两个原因：(1) 驾驶习惯不佳的驾驶员更倾向于购买保险保障较高的保险；(2) 保险保障较高的驾驶员开车时较为随意，从而导致更多保险索赔。图3(g)显示，女性的索赔数量多于男性，但两者的理赔频率差异随时间推移而缩小，甚至在 2014年后趋近于零。图 3(i)中可以观察到 保单类型的时间趋势呈下降态势，这表明续保或转入保单的理赔频率高于新保单。另一个有趣的发现是，图 3(h)中 发动机排量的估计系数符号随年份增加由负转正。总体而言，我们可以得出结论，大多数费率因子的时间趋势在 2011—2012年左右发生变化，且其对理赔频率的影响随时间推移而减弱。时间图表明，随着年份增加，费率因子在费率分类中的作用逐渐减小，这与第二轮保险费率改革的时间进程相符。该轮改革于 2007—2015年实施，旨在降低保险费率并提高保险保障。本轮改革最重要的法规包括中国保险行业协会于 2012年发布的《机动车商业保险示范条款》[2]，以及中国银行保险监督管理委员会于 2015年发布的《关于深化商业车险条款费率管理制度改革的意见》[3]。监管部门在汽车保险费率制定过程中建议采用统一的机动车保险条款，以避免保险公司通过降低保险费来吸引客户、引发恶性竞争。这些不断变化的法规削弱了保险公司根据风险分类制定保险费的权利，这体现在大多数费率因子的回归系数估计值呈下降趋势。

[2]http://www.iachina.cn/art/2012年/3/14/art_22_9744.html[3]
http://www.gov.cn/gongbao/content/2015年/content_2868890.htm

时期内，为避免保险公司通过降低保险费吸引客户而引发恶性竞争，汽车保险费率制定采用了统一条款。这些法规的变化削弱了保险公司依据风险分类制定保险费的权利，这反映在大多数费率因子的回归系数估计值呈下降趋势。

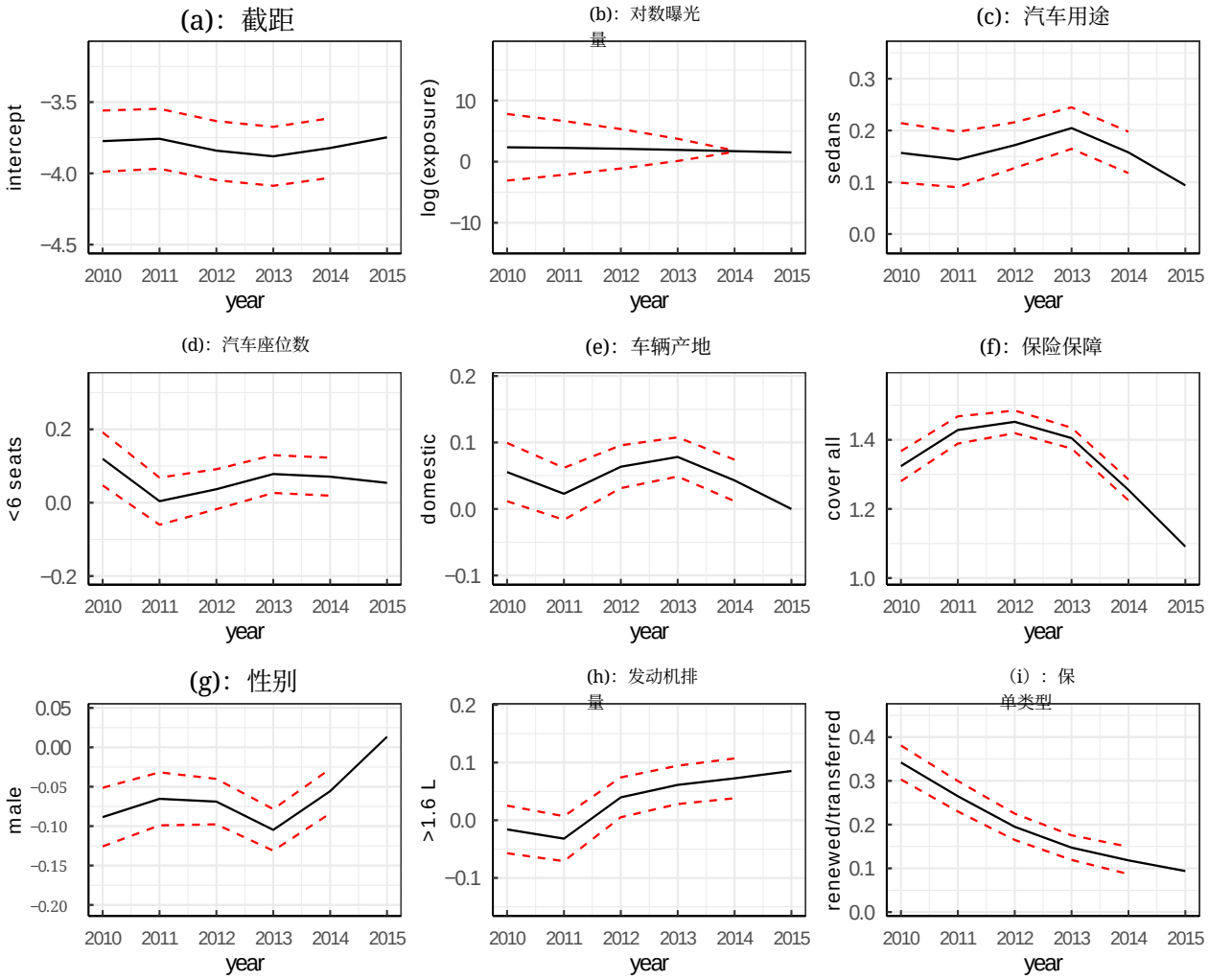


图3：使用所提方法得到的截距（左）和八个协变量的系数及95%置信带图。系数图在 $t = 0.75$处分隔，分别对应训练数据中的估计系数和测试数据中的一步超前预测。女性在性别变量中、新保单在保单类型中、非轿车在车辆用途中、非国产在车辆产地中、 $\leq$ 1.6L在发动机排量中、 ≥ 6 座在汽车座位数中，以及其他保障在保险保障中被用作参考水平。

为了比较 泊松、负二项、零膨胀泊松、DPSS、DNBSS 和 DZIPSS 模型的表现方面，我们通过在测试数据上应用各模型来预测 ct

理赔频率。表5报告了测试数据中预测索赔与实际观测索赔之间的差异。可以看出，我们提出的动态预测模型在预测无索赔情况下的理赔频率方面表现要好得多。图4展示了测试数据中整个保险组合的预测总索赔额。静态预测模型（泊松、负二项、零膨胀泊松）高估了总索赔额，而 DPSS 和 DNBSS 低估了真实值。DZIPSS 的预测表现最佳，因为其偏差最低。动态模型的预测结果比静态模型更接近观测到的总索赔额。我们还使用式（4.1）中定义的泊松偏差、文献（Lee，2021年）讨论的提升度，以及Frees等人（2011年）提出的基尼指数，评估预测索赔结果的准确性。提升度和基尼指数是辅助指标，无须假设底层分布即可评估模型。提升度是最高十分位预测的平均响应与最低十分位（双向提升度）或总体（单向提升度）平均响应之比。提升度可以视为衡量模型区分观测值能力的指标。提升度越高，说明模型越能将极端值与平均值区分开来。另一个类似的辅助指标是基尼指数，它用于衡量模型的区分能力，尤其适用于零值占比高且特征高度右偏的索赔数据。基尼指数越高，观测值之间的区分度越大。换言之，基尼指数越高，区分优质风险与劣质风险的能力越强。从表6可以看出，动态模型（DPSS、DNBSS 和 ZIPSS）的偏差小于静态模型（泊松、负二项和零膨胀泊松），这表明引入费率因子的时间趋势可以改善模型拟合表现。DPSS 和泊松模型的单向提升度分别为 3.522 和 3.550，这表明根据单向提升度，DPSS 模型划分出的风险最高的 10% 贡献了约 35.22% 的总索赔额，而静态泊松模型划分出的风险最高的 10% 贡献了约 35.50%。这意味着保险公司可以通过核保剔除风险最高的 10%，而根据 DPSS 的结果，剩余 90% 的保单仅占原始损失的 64.78%。根据 DPSS 模型，保险公司平均可降低收费 28.02%（1-64.78/90）；而根据静态泊松模型，平均可降低 28.33%（1-64.50/90）。动态模型的单向和双向提升度较低，这表明动态模型区分优质风险与劣质风险的能力并未更强。这些结果符合中国第二轮保险费率改革的目标。该轮改革于2007—2015年实施，旨在弱化费率

表5：真实数据集测试数据中索赔预测值与观测值之间的差异。

计数（k）	静态模型			动态模型		
	泊松	NB	ZIP	DPSS	DNBSS	DZIPSS
0	-1353	-1058	-1032	526	685	-222
1	1342	828	838	-309	-555	268
2	27	172	172	-181	-113	-25
3	-12	44	21	-31	-15	-16
4	-3	11	1	-5	-2	-4
5	-1	2	0	-1	-1	-1
6	0	1	0	0	0	0

表 6：真实数据集测试数据的泊松偏差、提升度和基尼指数。

模型	静态模型			动态模型		
	泊松	NB	ZIP	DPSS	DNBSS	DZIPSS
泊松偏差	0.300	0.300	0.299	0.299	0.300	0.296
双向提升	1.735	1.711	1.707	1.608	1.581	1.701
单向提升	3.550	3.530	3.537	3.522	3.507	3.576
基尼指数	0.934	0.934	0.934	0.934	0.934	0.934

因素。表 6 显示，就基尼指数而言，这六个模型的风险区分能力没有显著差异。

从图 5中，我们还可以利用提升图来评估候选模型的表现（Lee，2021年）。为计算这一指标，我们对预测值进行排序，并将观测值分为10个十分位组。图5 展示了各十分位组的平均响应与平均预测值。若各点与45度线对齐，则表明模型具有较高的预测性能。从图 5中可以看出，动态模型在预测高频风险类别方面表现出色，因为预测值与响应几乎相同。图 6 展示了双重提升图，可直接比较DPSS模型与泊松模型的预测性能。关于双重提升图的详细讨论，见Lee （2020年）的研究。我们可以看到，蓝线（损失比率）与红线（指示费率变化）之间存在显著的正相关关系，相关系数为0.963，这表明DPSS模型能够解释很大一部分泊松模型未能捕捉到的残差。

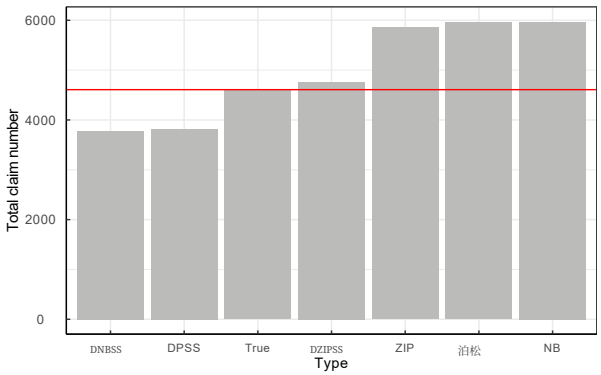


图4：测试数据中索赔总数的预测值。索赔总数的真实观测值为4,068，以红线标出。

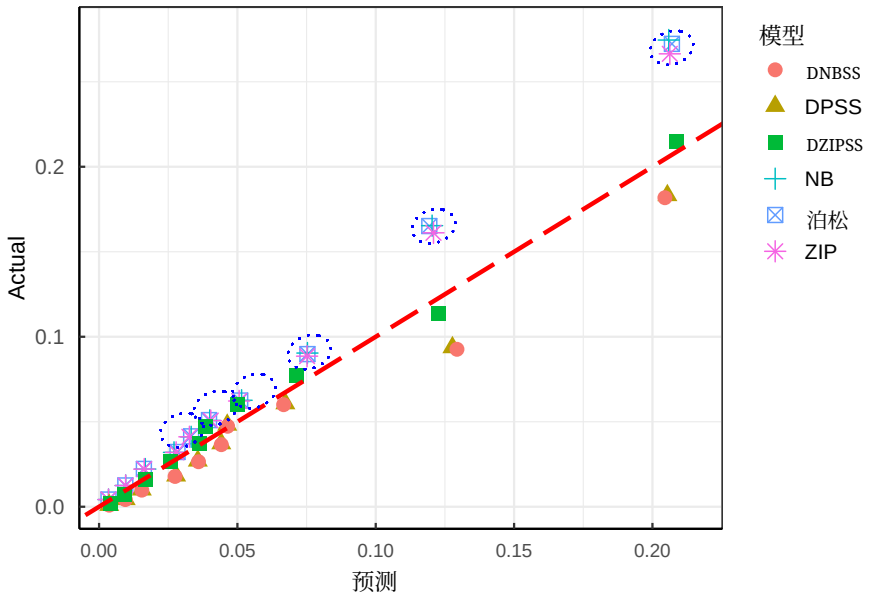


图5： 竞争模型的提升图。静态模型以深蓝色虚线圈出 .

6 结论

我们提出并研究了动态泊松状态空间模型及其扩展，不仅能够处理零值占比过高、极不平衡的索赔数据，还能通过状态空间建模框架随时间更新理赔频率预测。与现有精算科学文献中经常讨论的状态空间模型所采用的贝叶斯估计方法不同（Ahn 等人，2022），我们提出的模型将参数表示为随时间变化的平滑样条，并通过最大似然法进行估计。理赔频率预测基于

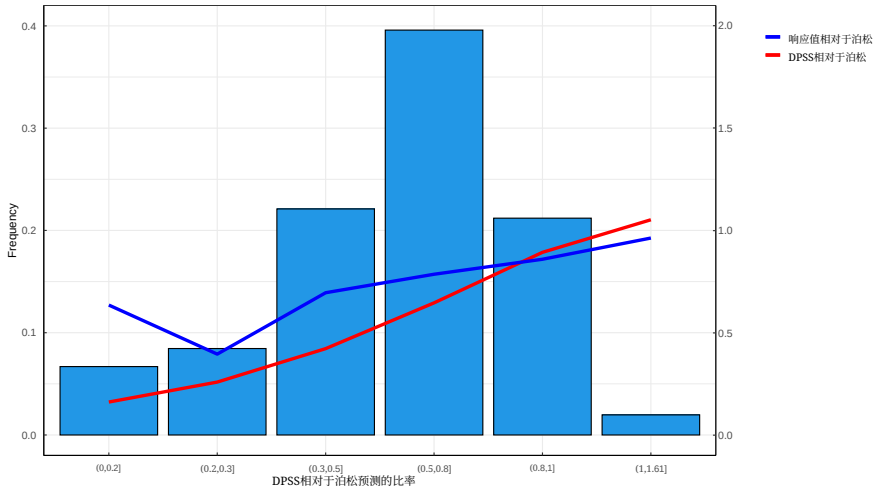


图6：DPSS模型相对于泊松模型的双重提升图。观测值按照DPSS预测与泊松预测的比率排序，并按其所属区间分组（比率为 $0.6 \in (0.5, 0.8]$）。对于每个组，实际理赔次数与泊松预测总量之比以蓝线表示，DPSS预测总量与泊松预测总量之比以红线表示。

基于卡尔曼滤波器类算法，并且在获得新的索赔信息时可以动态更新。所提方法克服了传统变系数模型的局限性；传统模型主要关注估计和统计推断，而非未来预测。该方法还专为在线监测而设计，因为所提算法比贝叶斯估计方法具有更高的计算效率。在一个真实世界的保险理赔数据集中，我们在泊松、负二项和零膨胀泊松假设下，将六年的理赔次数数据建模为若干协变量的函数，并每年更新一次模型。实证结果表明，DPSS中大多数费率因子的估计系数动态模式，可以用中国2010年至2015年的汽车保险费率改革过程来解释。若干实用指标显示，所提方法在预测理赔频率方面具有优势。

值得一提的是，我们在本研究中提出的 DPSS 属于单变量计数模型；当汽车理赔的频数包含多种类型时，可将其扩展为多变量计数模型，例如第三方人身伤害索赔数量（Y_1）、自身车辆损失索赔数量（Y_2）和第三方财产损失索赔数量（Y_3）。本研究的另一项扩展方向，是在动态状态空间框架下开展空间分析，以便探究协变量对保险索赔数据更复杂的影响。此外，

基于状态空间方程的动态建模方法，还可在经验费率厘定的背景下，针对纵向索赔数据作进一步探讨。

补充材料

代码：动态泊松状态空间模型及模拟研究的 R 源代码。压缩包中包含 readme 文件（dpss.zip）。

致谢

Zhengxiao Li 感谢国家自然科学基金（项目编号：72271056）、对外经济贸易大学优秀青年学者项目（项目编号：20YQ16）以及对外经济贸易大学“中央高校基本科研业务费”（项目编号：CXTD13-02）提供的经费支持。Jiakun Jiang 感谢国家自然科学基金（项目编号：12101054）和全国统计科学研究计划（项目编号：2022LY034）提供的经费支持。

附录

动态零膨胀泊松状态空间模型（DZPSS）

假设响应变量 y_i 服从均值为 $e_i\lambda_i$ 、零膨胀概率为 ϕ_i的零膨胀泊松分布，即

$$Y_i \sim \begin{cases} 0, & \text{with probability} \quad \phi_i, \\ \text{Poisson}\left(\lambda_i\right), & \text{with probability } 1-\phi_i, \end{cases} \tag{.1}$$

在（.1）中， $0 \leq \phi_i < 1$，因此可以对数据中的额外零值进行显式建模。在 0 处增加的点质量使零膨胀泊松模型能够产生比泊松分布预测更多的取值为 0 的响应，从而适应零膨胀现象。

影　　响可以建模为具有时变和时不变系数的协变量的影响

$$\begin{aligned}\log(\lambda_i) &= \boldsymbol{B}_{i,q_1}^T\boldsymbol{\beta}(t_i) + \boldsymbol{B}_{i,q_2}^T\boldsymbol{\varphi} = \boldsymbol{B}_i^T\boldsymbol{\gamma}(t_i),\\ \operatorname{logit}(\phi_i) &= \log\left(\frac{\phi_i}{1-\phi_i}\right) = \boldsymbol{G}_{i,q_1}^T\boldsymbol{\rho}(t_i) + \boldsymbol{G}_{i,q_2}^T\boldsymbol{\alpha} = \boldsymbol{G}_i^T\boldsymbol{\gamma}(t_i),\end{aligned} \tag{.2}$$

其中，$\boldsymbol{B}_{i,q_1} = (1, B_{i,1}, \cdots, B_{i,q_1})^T$和 $\boldsymbol{G}_{i,q_1} = (1, G_{i,1}, \cdots, G_{i,q_1})^T$是系数随时间变化的协变量 $\boldsymbol{\beta}(t) = (\beta_0(t) \cdots, \beta_{q_1}(t))^T$和 $\boldsymbol{\rho}(t) = (\rho_0(t), \rho_{q_1}(t))^T$。$\boldsymbol{B}_{i,q_2} = (1, B_{i,1}, \cdots, B_{i,q_2})^T$和$\boldsymbol{G}_{i,q_2} = (1, G_{i,1}, \cdots, G_{i,q_2})^T$是系数时不变的协变量 $\boldsymbol{\varphi} = (\varphi_1, \cdots, \varphi_{q_2})^T$和$\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_{q_2})^T$。 $\boldsymbol{B}_i = \left(\boldsymbol{B}_{i,q_1}^T, \boldsymbol{B}_{i,q_2}^T, 0, \cdots, 0\right)^T$和 $\boldsymbol{G}_i = \left(0, \cdots, 0, G_{i,q_1}^T, G_{i,q_2}^T\right)^T$是具有系数向量 $\boldsymbol{\gamma}(t) = \left(\boldsymbol{\beta}(t)^T, \boldsymbol{\varphi}^T, \boldsymbol{\rho}(t)^T, \boldsymbol{\alpha}^T\right)^T$的 $2(q_1 + q_2 + 1)$ 维协变量向量。

参数可以通过最大化以下惩罚对数似然函数来估计：

$$\begin{aligned}L_c(\boldsymbol{\gamma}; \boldsymbol{y}) = &\sum_{y_i=0}\log\left\{e^{\boldsymbol{G}_i^T\boldsymbol{\gamma}(t_i)} + \exp\left[-e^{\boldsymbol{B}_i^T\boldsymbol{\gamma}(t_i)}\right]\right\} + \sum_{y_i>0}\left[y_i\boldsymbol{B}_i^T\boldsymbol{\gamma}(t_i) - e^{\boldsymbol{B}_i^T\boldsymbol{\gamma}(t_i)} - \log(y_i!)\right]\\ &- \sum_{i=1}^{n}\log\left[1 + e^{\boldsymbol{G}_i^T\boldsymbol{\gamma}(t_i)}\right] - \frac{1}{2}\sum\nolimits_{j=0}^{q_1}\tau_{1j}\int\left\{\boldsymbol{\beta}_j''(t)\right\}^2 dt - \frac{1}{2}\sum\nolimits_{j=0}^{q_1}\tau_{2j}\int\left\{\boldsymbol{\rho}_j''(t)\right\}^2 dt,\end{aligned}$$

与（2.1）中提出的 DPSS 类似，DZIPSS（.2）也可以表示为状态空间表示形式：

$$\begin{aligned}\log(\lambda_i) &= \boldsymbol{U}_i^T\boldsymbol{\zeta}(t_i),\\ \operatorname{logit}(\phi_i) &= \boldsymbol{V}_i^T\boldsymbol{\zeta}(t_i),\\ \boldsymbol{\zeta}(t_i) &= \boldsymbol{T}_i\boldsymbol{\zeta}(t_{i-1}) + \boldsymbol{\eta}(t_i, t_{i-1}), \quad \boldsymbol{\eta}(t_i, t_{i-1}) \sim N(\mathbf{0}, \boldsymbol{Q}_i),\end{aligned} \tag{.3}$$

其中

$$\boldsymbol{U}_i = (1, 0, B_{i1}, 0, B_{i2}, \cdots, B_{iq_1}, 0, B_{iq_1+1}, B_{iq_1+2}, \cdots, B_{iq_1+q_2}, 0, 0, \cdots, 0)^T,$$
$$\boldsymbol{V}_i = (0, 0, \cdots, 0, 1, 0, G_{i1}, 0, G_{i2}, \cdots, G_{iq_1}, 0, G_{iq_1+1}, G_{iq_1+2}, \cdots, G_{iq_1+q_2})^T,$$
$$\boldsymbol{\zeta}(t) = \left(\beta_0(t_i), \beta_0{}'(t_i), \cdots, \beta_{q_1}(t_i), \beta_{q_1}{}'(t_i), \boldsymbol{\varphi}^T, \rho_0(t_i), \rho_0{}'(t_i), \cdots, \rho_{q_2}(t_i), \rho_{q_2}{}'(t_i), \boldsymbol{\alpha}^T\right)^T,$$
$$\boldsymbol{T}_i = \mathrm{diag}(\overbrace{T_i, \cdots, T_i}^{q_1+1}, \overbrace{1, \cdots, 1}^{q_2}, \overbrace{T_i, \cdots, T_i}^{q_1+1}, \overbrace{1, \cdots, 1}^{q_2}),$$
$$\boldsymbol{Q}_\mathrm{i} = \mathrm{diag}(\overbrace{\tau_{10}^{-1}Q_i, \cdots, \tau_{1q_1}^{-1}Q_i}^{q_1+1}, \overbrace{0, \cdots, 0}^{q_2}, \overbrace{\tau_{20}^{-1}Q_i, \cdots, \tau_{2q_1}^{-1}Q_i}^{q_1+1}, \overbrace{0, \cdots, 0}^{q_2}).$$

估计与预测算法与模型（2.6）类似。

B 动态负二项状态空间模型（DNBPSS）

负二项模型可以由一个针对离散变量 Y分布的两阶段模型导出。我们假设存在一个未观测随机变量 Θ ，其服从伽马分布$\Theta \sim \mathsf{Gamma}(\alpha,\ \alpha)$，均值为 1，方差为 $1/\alpha$。然后，模型假定在给定 Θ的条件下， Y 服从均值为 $\mu\Theta$的泊松分布。于是， Y 的边缘分布为均值为 μ 、方差为 $\mu + 1/\alpha\mu^2$的负二项分布。在动态负二项状态空间模型中，系数随时间变化和时不变的协变量效应可以引入负二项分布的均值 μ 中，即

$$\begin{aligned} \log(\mu_i) &= \boldsymbol{z}_i^T \boldsymbol{\gamma}(t_i), \quad i = 1, \cdots, n \\ \boldsymbol{\gamma}(t_i) &= \boldsymbol{T}_i \boldsymbol{\gamma}(t_i) + \boldsymbol{\eta}(t_i, t_{i-1}), \quad \boldsymbol{\eta}(t_i, t_{i-1}) \sim N(\mathbf{0}, \boldsymbol{Q}_i), \end{aligned} \tag{.4}$$

其中

$$\begin{aligned} \boldsymbol{z}_i &= (1, 0, x_{i1}, 0, x_{i2}, \cdots, x_{iq_1}, 0, x_{i,q_1+1}, \cdots, x_{iq})^T, \\ \boldsymbol{\gamma}(t) &= \left(\beta_0(t), \beta'_0(t), \beta_1(t), \beta'_1(t), \cdots, \beta_{q_1}(t), \beta'_{q_1}(t), \alpha_1, \cdots, \alpha_{q_2}\right)^T, \\ \boldsymbol{\eta}(t_i, t_{i-1}) &= \left(U_0^T(t_i, t_{i-1}), \cdots, U_{q_1}^T(t_i, t_{i-1})\right)^T, \\ \boldsymbol{T} &= \mathrm{diag}\left(\overbrace{T, \cdots, T}^{q_1+1}, \overbrace{1, \cdots, 1}^{q_2}\right), \boldsymbol{Q} = \mathrm{diag}\left(\overbrace{\lambda_0^{-1}Q, \cdots, \lambda_{q_1}^{-1}Q}^{q_1+1}, \overbrace{0, \cdots, 0}^{q_2}\right). \end{aligned}$$

参考文献

Jae Youn Ahn、Himchan Jeong 和 Yang Lu。集体风险模型的简易贝叶斯状态空间方法。《斯堪的纳维亚精算期刊》，第 1–21 页，2022 年。

Atanu Bhattacharjee、Gajendra K Vishwakarma 和 Abin Thomas。基因表达数据分析中的贝叶斯状态空间建模：生物标志物预测的应用。《数学生物科学》，305：96–101，2018年。

Le Chang 和 Yanlin Shi。死亡率的动态建模与一致预测：一种时变系数时空自回归方法。《斯堪的纳维亚精算期刊》，2020年（9）：843–863，2020年。

Michel Denuit 和 Yang Lu。应用于非寿险的 Wishart-伽马随机效应模型。《风险与保险杂志》，88（2）：443–481，2021年。

A.X.D. Dong 和 J.S.K. Chan。使用带广义贝塔分布的动态模型进行理赔准备金评估的贝叶斯分析。《保险：数学与经济学》，53（2）：355–365，2013年。

Jianqing Fan 和 WenyangZhang。变系数模型的统计方法。《统计学及其交叉领域》，1（1）：179–195，2008年。

Edward W Frees、Glenn Meyers 和 A David Cummings。使用基尼指数概述保险评分。《美国统计学会会刊》，106(495):1085–1098，2011年。

Man Chung Fung、Gareth W. Peters 和 Pavel V. Shevchenko。基于状态空间框架进行死亡率建模的统一方法：特征描述、识别、估计与预测。《精算科学年刊》，11(2):343–389，2017年。

Chong Gu。惩罚似然回归：贝叶斯分析。《中华统计学志》，第255–264页，1992年。

Trevor Hastie 和 Robert Tibshirani。变系数模型。《英国皇家统计学会会刊：B辑（方法论）》，55(4):757–779，1993年。

Roel Henckaerts 和 Katrien Antonio。利用远程信息处理收集的驾驶行为数据动态更新汽车保险价格的附加价值。《保险：数学与经济学》，105:79–95，2022年。

Joseph M Hilbe。《负二项回归》。剑桥大学出版社，2011年。

Jiakun Jiang、Wei Yang、Erin M Schnellinger、Stephen E Kimmel 和 Wensheng Guo。用于临床决策的动态逻辑状态空间预测模型。《生物统计学》，2021年。

Bent Jørgensen 和 Gorden Smyth。将 Tweedie复合泊松模型拟合到保险索赔数据。《斯堪的纳维亚精算期刊》，1994：69–93，1994年1月。

Nadja Klein、Thomas Kneib 和 Stefan Lang。针对零膨胀和过度离散计数数据的位置、尺度和形状的贝叶斯广义加性模型。美国统计学会会刊，110(509)：405–419，2015年。

Tze Leung Lai 和 Kevin Haoyu Sun。演化可信度理论：一种广义线性混合建模方法。《北美精算期刊》，16(2)：273–284，2012年。

Diane Lambert。零膨胀泊松回归及其在制造缺陷中的应用。《技术计量学》，34(1)：1–14，1992年。

Simon CK Lee. 精算定价中负二项回归的 Delta提升实现。《风险》，8(1)，2020年。

Simon CK Lee. 通过结合提升法的零膨胀泊松回归处理不平衡保险数据。《ASTIN公报：国际精算学会期刊》，51(1):27–55，2021年。

Woojoo Lee、Jeonghwan Kim 和 Jae Youn Ahn. 用于经验费率厘定的泊松随机效应模型：局限性与替代方案。保险：数学与经济学，91: 26–36，2020年。

Steven M. Lewis 和 Adrian E. Raftery. 使用拉普拉斯—梅特罗波利斯估计量，通过后验模拟估计贝叶斯因子。《美国统计学会会刊》，92(438):648–655，1997年a。

Steven M Lewis 和 Adrian E Raftery. 使用拉普拉斯—梅特罗波利斯估计量，通过后验模拟估计贝叶斯因子。《美国统计学会会刊》，92(438):648–655，1997b。

Hong Li、Yang Lu 和 Wenjun Zhu。《动态贝叶斯费率厘定：一种马尔可夫链近似法》。《北美精算期刊》，25(2)：186–205，2021年。

Yang Lu。《保险中的动态脆弱性计数过程：估计、定价与预测的统一框架》。《风险与保险杂志》，85(4)：1083–1102，2018年。

Philip D O'Neill。使用马尔可夫链蒙特卡罗方法对随机流行病模型进行贝叶斯推断的教程式介绍。《数学生物科学》，180(1-2):103–114，2002年。

Byeong U Park、Enno Mammen、Young K Lee 和 Eun Ryung Lee。变系数回归模型：综述与新进展。《国际统计评论》，83(1):36–64，2015年。

Jean Pinquet。具有动态随机效应和逐期非负可信度的泊松模型。《ASTIN公报：国际精算学会期刊》，50(2):585–618，2020年。

Greg Taylor。演化层级可信度。《ASTIN公报：国际精算学会期刊》，48(1): 339–374，2018年。

Grace Wahba。非正规先验、样条平滑以及回归中防范模型误差的问题。《英国皇家统计学会会刊：B辑（方法论）》，40(3):364–372，1978年。

Grace Wahba。交叉验证平滑样条的贝叶斯"置信区间"。英国皇家统计学会会刊：B辑（方法论），45(1):133–150，1983年。

William E Wecker 和 Craig F Ansley。非线性回归与样条平滑的信号提取方法。美国统计学会会刊，78(381):81–89，1983年。

Karen CH Yip 和 Kelvin KW Yau。含额外零值的一般保险索赔频率数据建模。保险：数学与经济学，36(2):153–163，2005年。

Ying Zhu、Barry K Goodwin 和 Sujit K Ghosh。技术变革下的产量风险建模：动态产量分布与美国农作物保险计划。农业与资源经济学杂志，第192–210页，2011年。