Advances in Applied Mathematics 应用数学进展, 2023, 12(3), 1299-1305
Published Online March 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.123132

Hans汉斯

增量索赔准备金的高斯过程回归模型

邓旭生，卢志义*

天津商业大学理学院，天津

收稿日期：2023年2月27日；录用日期：2023年3月22日；发布日期：2023年3月29日

摘　要

在保险实践中，索赔管理、业务实践、货币政策和立法变化经常发生，这将会对在同一日历年发生的索赔产生相同影响。对来自同一日历年的多个索赔之间的相关关系进行建模有可能进一步提高预测精度。本文尝试利用高斯过程回归(GPR)和复合核方法对经过对数转换后的增量索赔数据进行建模，从而引入同一日历年间的相关关系，提升预测精度。我们对来自NAIC数据库的三条业务线进行了实证分析，比较并展示了我们模型的性能，为今后的研究提供了新的思路。

关键词

索赔准备金，高斯过程回归，核函数，日历年影响

Gaussian Process Regression Model for Incremental Claim Reserves

Xusheng Deng, Zhiyi Lu*

School of Science, Tianjin University of Commerce, Tianjin

Received: Feb. 27th, 2023; accepted: Mar. 22nd, 2023; published: Mar. 29th, 2023

Abstract

In insurance practice, claims management, business practices, monetary policy and legislative changes occur frequently, which will have the same impact on claims occurring in the same calendar year. Modelling correlations between multiple claims from the same calendar year has the potential to further improve forecasting accuracy. In this paper, Gaussian process regression (GPR) and composite kernel method are used to model the log-transformed incremental claim data, so as to introduce the correlation between the same calendar year and improve the prediction accuracy.

*通讯作者。

文章引用: 邓旭生, 卢志义. 增量索赔准备金的高斯过程回归模型[J]. 应用数学进展, 2023, 12(3): 1299-1305.
DOI: 10.12677/aam.2023.123132

We conducted empirical analysis on three lines of business from NAIC database, compared and demonstrated the performance of our models, which provided a new idea for future research.

Keywords

Claims Reserving, Gaussian Process Regression, Kernel Function, Calendar-Year Effects

1. 引言

在非寿险保险实践中，索赔准备金的估计是一项重要任务：即利用已付索赔的历史来预测未来索赔的出现模式。一个主要的挑战是量化未来索赔出现的各自的不确定性或分布。这对于确定一家保险公司所需的准备金、风险资本以及由此产生的偿付能力至关重要。在过去的几十年间，人们提出了多种用于预测索赔准备金的模型。早期的工作倾向于建立在链接比方法和寻找某些类型的回归之间的相似之处[1]。自英国精算师将广义线性模型引入到非寿险精算领域后，引发了许多关于这一模型的研究，例如 Taylor 和 McGuire [2]、毛泽春和吕立新[3]、孟生旺[4]、Peters 等[5]。Antonio 和 Beirant 则进行了广义线性混合模型在该领域的应用研究[6]，Shi 等人遵循贝叶斯方法，使用伽马分布来对索赔过程进行建模[7]，Peters 等人则引入了唯一链接函数。而 Ajne 首先提出索赔准备金之间存在相关性，在这方面的研究有多元链梯法[8]，多元加法损失准备金方法，包括 Hess 等人[9]和 Merz 和 Wüthrich [10]。最近，Merz 和 Wüthrich 将链梯法和加性损失预留方法结合到一个集成框架中，以考虑多个流量三角形之间的异质性[11]。Zhang 和 Dukic 添加了关联函数来解释多个产品线之间的相关性[12]，而 Shi 和 Hartman 则使用贝叶斯层次模型来解释这些相关性[13]。

最近，Lopes 等人提出使用支持向量机和高斯过程回归模型来估计已发生但未报告的索赔准备金，首次将高斯过程回归引入该领域[14]。Lally 和 Hartman 用带输入翘曲的高斯过程回归和几个常用的协方差函数来估计索赔[15]。

本文运用高斯过程回归以及加性组合核函数对经过对数转换后的增量索赔数据进行了建模，引进了同一日历年间索赔数据之间的相关性。在第二节我们简要介绍了高斯过程回归的原理和一些常用的核函数，在第三节描述了我们提出的模型，在第四节我们将模型应用到了三个来自 NAIC 的实际数据集上并与其他模型进行了对比，最后我们探讨了高斯过程回归在索赔准备金预测问题上其他可能的发展方向。

2. 高斯过程回归

2.1. 基本原理

高斯过程回归是以高斯过程为先验，直接在函数空间进行贝叶斯推断的一种贝叶斯非参数模型。高斯过程是多维高斯分布的推广，它完全由均值函数和协方差函数确定：

$$f(x) \sim GP\left(m(x), K(x, x')\right) \tag{1}$$

其中 $m(x)$ 为均值函数， $K(x, x')$ 为协方差函数(核函数)。

考虑回归问题 $y = f(x) + \varepsilon$，假设噪声 $\varepsilon \sim N\left(0, \sigma_\varepsilon^2\right)$。可得到观测值 $\boldsymbol{y}$ 与在测试点 $\boldsymbol{x}_* = \left(x_{n+1}, \cdots, x_{n+m}\right)^{\mathrm{T}}$

处预测值 $\boldsymbol{f}_*$ 的联合分布

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{f}_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\boldsymbol{x}) \\ m(\boldsymbol{x}_*) \end{bmatrix}, \begin{bmatrix} K(\boldsymbol{x},\boldsymbol{x}) + \sigma_n^2 \boldsymbol{I} & K(\boldsymbol{x},\boldsymbol{x}_*) \\ K(\boldsymbol{x}_*,\boldsymbol{x}) & K(\boldsymbol{x}_*,\boldsymbol{x}_*) \end{bmatrix} \right) \tag{2}$$

其中均值函数通常设为 0，即 $m(\boldsymbol{x}) = m(\boldsymbol{x}_*) = 0$，$K(\boldsymbol{x},\boldsymbol{x})$ 为 $n \times n$ 阶对称半正定协方差矩阵；$K(\boldsymbol{x}_*,\boldsymbol{x}) = K(\boldsymbol{x},\boldsymbol{x}_*)^{\mathrm{T}}$ 为 $m \times n$ 阶矩阵，有 $\left[K(\boldsymbol{x}_*,\boldsymbol{x})\right]_{ij} = K(x_{n+i}, x_j)$；$K(\boldsymbol{x}_*,\boldsymbol{x}_*)$ 为 $m \times m$ 阶矩阵，且 $\left[K(\boldsymbol{x}_*,\boldsymbol{x}_*)\right]_{ij} = K(x_{n+i}, x_{n+j})$。由此可以得到 $\boldsymbol{f}_*$ 的后验分布

$$\boldsymbol{f}_* \mid \boldsymbol{x}, \boldsymbol{x}_*, \boldsymbol{y} \sim \mathcal{N}\left(\overline{\boldsymbol{f}_*}, \operatorname{cov}(\boldsymbol{f}_*)\right) \tag{3}$$

其中

$$\overline{\boldsymbol{f}_*} = \mathbb{E}\left[\boldsymbol{f}_* \mid \boldsymbol{x}, \boldsymbol{x}_*, \boldsymbol{y}\right] = m(\boldsymbol{x}_*) + K(\boldsymbol{x}_*,\boldsymbol{x})\left[K(\boldsymbol{x},\boldsymbol{x}) + \sigma_\varepsilon^2 \boldsymbol{I}\right]^{-1}\left(\boldsymbol{y} - m(\boldsymbol{x})\right) \tag{4}$$

$$\operatorname{cov}(\boldsymbol{f}_*) = K(\boldsymbol{x}_*,\boldsymbol{x}_*) - K(\boldsymbol{x}_*,\boldsymbol{x})\left[K(\boldsymbol{x},\boldsymbol{x}) + \sigma_n^2 \boldsymbol{I}\right]^{-1} K(\boldsymbol{x},\boldsymbol{x}_*). \tag{5}$$

如此便可得到 $\boldsymbol{f}_*$ 的均值和方差。

高斯过程回归可以通过选择不同的核函数来捕捉不同的统计特征，其中最常用的是平方指数核(SE)：

$$K_{SE} = \eta^2 \exp\left(-\frac{\left|x_i - x_j\right|^2}{2l^2}\right) \tag{6}$$

其中 η^2 为信号方差，l 为长度尺度，$\theta = \left\{\eta^2, l, \sigma_n^2\right\}$ 即为模型的超参数。一般情况下，通过二型极大似然法来估计超参数，在得到最优参数后即完成了模型训练，便可利用式(4) (5)对任何测试点进行预测。但上述方法存在收敛于局部最大值的风险，因此本文使用了由 Neal [16]和 Flaxman 等人[17]提出的分层贝叶斯方法。该方法要求我们为协方差函数的超参数 θ 定义一个联合概率分布：

$$\theta \sim p(\theta) \tag{7}$$

$$f \mid X, \theta \sim N_n\left(M(X), K_\theta(X,X)\right) \tag{8}$$

其中 $K_\theta(X,X)$ 是输入数据和超参数 θ 的函数。

2.2. 核函数

核函数以其在支持向量机中的应用而广为人知，在高斯过程的应用中，核函数的作用则是定义数据间的相似度，以构造相似矩阵。

其中最常用的核函数便是平方指数核函数，它的形式为：

$$k_{SE}(x,x') = \eta^2 \exp\left(-\frac{\left|x - x'\right|^2}{2l^2}\right) \tag{9}$$

它只有两个参数：长度尺度 l，用于确定函数中“摆动”的长度；输出方差 σ^2 确定函数与其平均值的平均距离。每个内核前面都有这个参数，起到比例因子的作用。

平方指数核函数已成为高斯过程和支持向量机应用中的默认核函数。这是因为它有一些不错的属性。例如，这个核函数是无限可微的，这意味着具有这个核函数的高斯过程具有所有阶的均方导数，从中抽取的曲线都将非常平滑。

另外，在本文中我们还将使用指数核函数，它与平方指数核函数的区别仅在于不对数据点之间的欧

氏距离进行平方：

$$k_{EXP}\left(x,x'\right)=\eta^2\exp\left(-\frac{\left|x-x'\right|}{2l^2}\right) \tag{10}$$

当我们需要的数据结构不能由任何已知的简单核函数表示时，就需要构建具有所需属性的“定制”核函数。两种最常用的构建组合核函数的方法是加法和乘法。其内在原理为两个核函数的和或者积一定是一个核函数，Rasmussen 和 Williams 给出了具体证明[18]。

$$k_a+k_b=k_a\left(x,x'\right)+k_b\left(x,x'\right) \tag{11}$$

$$k_a\times k_b=k_a\left(x,x'\right)\times k_b\left(x,x'\right) \tag{12}$$

特别的，当在对多维函数建模时，对每个子维度上定义的核函数求和可以得到跨维度的可加性结构。而这种结构允许我们做出远离训练数据的预测。将每个子维度上定义的核函数相乘则更具灵活性，能在近距离内拟合数据。

3. 索赔准备金预测模型

3.1. 建模对象

索赔准备金预测问题中常用的建模对象是累积索赔 *CC*，它以事故年 *AY* 和进展年 *DL* 为索引。表 1 展示了一个经典的聚合索赔损失三角形。*x* 轴单位是进展年，*y* 轴单位是事故年。有许多变量可以从中推导出来，例如增量索赔和增量损失率，它们也都可以用于建立索赔准备金估计模型。

Table 1. Cumulative loss upper triangle
表 1. 累积赔款流量三角形

进展年(*DL*) / 事故年(*AY*)	0	1	2	…	*j*	…	*J*
0	$CC_{0,0}$	$CC_{0,1}$	$CC_{0,2}$	...	$CC_{0,j}$	...	$CC_{0,J}$
1	$CC_{1,0}$	$CC_{1,1}$	$CC_{1,2}$	...	$CC_{1,j}$	...	
2	$CC_{2,0}$	$CC_{2,1}$	$CC_{2,2}$	...	$CC_{2,j}$		
...	...	...	...	...			
i	$CC_{i,0}$	$CC_{i,1}$	$CC_{i,2}$				
...	...	...					
I	$CC_{I,0}$						

Lally 和 Hartman 直接对累积索赔 *CC* 建模，并且将累积索赔数据可视化为三维曲面，将预测问题视为外推任务。Ludkovski 则开发了以增量损失率为建模对象的高斯过程回归模型[19]。但所有这些模型都没有考虑具有同一日历年的索赔数据间的相关关系。

每年新增的观测值在损失三角形中建立一条新的对角线.同一日历年的观测结果受到由政治因素、通货膨胀等因素带来的同一水平影响。因而，这些同一日历年的观测结果应该具有更高的相关关系。

对增量损失建模相较对累积损失建模的最大优点就是可以捕捉日历年(*CY*)带来的相关关系。另外，可以通过对数转换或直接要求增量损失大于零来确保最终预测结果单调递增。我们可以得到一个由三个输入(*AY*, *DL*, *CY*)和一个输出 $I_{(AY,DL,CY)}$。

因此在本文中，我们选择增量损失 I 作为模型对象。增量损失和日历年可由下式得到：

$$I_{i,j}=\begin{cases}CC_{i,0} & j=0\\ CC_{i,j}-CC_{i,j-1} & j=1,\cdots,J\end{cases} \tag{13}$$

$$CY=AY+DL \tag{14}$$

3.2. 变量转换

Lally 和 Hartman 通过实现输入扭曲强制 $CC_{(AY,DL)}$ 随进展年单调递增，但这导致了预测值在事故年维度上过于平滑的问题。本文通过对增量损失进行对数变换后进行预测，再通过指数变换将预测值转换回原始维数来确保增量损失大于零。虽然这种转换总是产生非常高的预测方差。但在以增量损失为模型对象的情况下，对数指数变换仍是目前较为合理的方法。

3.3. 我们的模型

正如第二节中我们强调的，指定合适的均值函数和核函数对高斯过程回归模型至关重要。

均值函数我们选择了零均值函数，这既是常规选择，也符合下三角形中增量数据随日历年不断增加逐渐趋近于零的数据结构。

我们使用以下核函数来引入历年效应。

平方指数(SE)加性核函数：

$$\begin{aligned}K\left(X-X'\right)&=K_{SE}\left(AY,AY'\right)+K_{SE}\left(DL,DL'\right)+K_{SE}\left(CY,CY'\right)\\&=\eta_1^2\exp\left(-\frac{\left|AY-AY'\right|^2}{2l_1^2}\right)+\eta_2^2\exp\left(-\frac{\left|DL-DL'\right|^2}{2l_2^2}\right)+\eta_3^2\exp\left(-\frac{\left|CY-CY'\right|^2}{2l_3^2}\right)\end{aligned} \tag{15}$$

指数(EXP)加性核函数：

$$\begin{aligned}K\left(X-X'\right)&=K_{EXP}\left(AY,AY'\right)+K_{EXP}\left(DL,DL'\right)+K_{EXP}\left(CY,CY'\right)\\&=\eta_1^2\exp\left(-\frac{\left|AY-AY'\right|}{2l_1^2}\right)+\eta_2^2\exp\left(-\frac{\left|DL-DL'\right|}{2l_2^2}\right)+\eta_3^2\exp\left(-\frac{\left|CY-CY'\right|}{2l_3^2}\right)\end{aligned} \tag{16}$$

正如第二节中所说的，加性核允许我们做出远离训练数据的预测。通过这种方式，大大提高了模型的外推能力。

超参数上的先验分布称为超先验。下面详细介绍了与我们提出的每个 GP 模型相关联的超先验模型。两个内核中涉及的超参数都服从以下先验分布：

$$\eta_{1,2,3}\sim H_n\left(\sigma=5\right) \tag{17}$$

$$l_{1,2,3}\sim Gamma\left(4,4\right) \tag{18}$$

其中 $H_n\left(\sigma=5\right)$ 为标准差为 5 的半正态分布，$Gamma\left(4,4\right)$ 为形状参数和逆尺度参数都为 4 的伽马分布。

4. 实证结果

在本节中，我们将我们的模型应用于从 NAIC 获得的数据集，包括医疗事故(Medical Malpractice，公司代码 = 669)、私人客车(PP auto，公司代码 = 1538)和工人赔偿(Worker's Comp，公司代码 = 1767) [20]。模型的结果将以均值和预测区间呈现，以便与其他方法进行比较。

损失储量模型的主要任务是建立储量预测的最佳估计值。为了测量模型的准确性，我们在以上三个

数据集的上三角形上训练我们的模型。因为每个数据集都是完整的，所以我们可以轻松的得到最终的观察值。我们预测了每个数据集的未偿债权负债，并将结果与观察到的未偿债权负债和其他方法的结果进行比较。表 2 列出了上述每个案例研究的损失准备金预测平均值和 95%最高密度区间(HDI)，其中我们将最接近观察值的三个模型的预测值进行了加粗，并且为最窄的预测区间加上了下划线。

Table 2. The observed and the predicted values of each model
表 2. 观察值与各个模型的预测值

	Medical Malpractice	PP Auto	Worker's Comp
观察值	164633	37397	307810
链梯法	**131996**	21181	184832
Guszcza 2008 Weibull [21]	199058	**42724**	236555
Guszcza 2008 Loglogistic [21]	266415	51219	**308420**
Lally, Hartman	(118722, **164003**, 209008)	(6492, 22034, 35388)	(137802, 268228, 389561)
SE 加性核函数	(101833, 119740, 135513)	(41848, **45452**, 50547)	(276823, **285126**, 295875)
指数加性核函数	(143100, **157289**, 168731)	(36606, **39371**, 42125)	(302204, **307204**, 312483)

加性核的结构限制了预测结果的可能结果数，因此运用了加性核函数模型的索赔准备金估计的预测区间较小，减少了预测的不确定性。另外，我们通过选择增量索赔作为模型对象来考虑相同日历年的相关性，并且通过对数变换保证了增量损失不为负，避免了输入翘曲引起的过度平滑问题。预测结果也证明了，引入日历年数据相关性后，核函数的选择依然会在很大程度上决定预测结果的准确性。

5. 总结与展望

本研究的主要目的是探讨高斯过程回归在索赔准备金预测问题中的进一步应用。我们的方法在一定程度上提升了预测精度，并且提供了一个新的研究方向。还有其他几个方向有待进一步研究，我们现在简要讨论一下。

正如我们之前提到的，将预测转换为原始尺度的对数指数转换总是会产生非常高的预测方差。我们认为直接以对数高斯过程为先验的完全对数高斯过程回归可以在很大程度上解决这一问题，这也是我们未来研究的重点。

另一个可能的发展方向是多重三角形。通过借鉴其他损失三角形的信息，有可能克服由单个损失三角形引起的模型不确定性问题，或者识别出多个损失三角形之间的共同趋势。

参考文献

[1] Ajne, B. (1994) Additivity of Chain-Ladder Projections. *ASTIN Bulletin*: *The Journal of the IAA*, **24**, 311-318. https://doi.org/10.2143/AST.24.2.2005072

[2] Taylor, G. and McGuire, G. (2004) Loss Reserving with GLMs: A Case Study. Partachi 2010 Statistica IMO, 327-391.

[3] 毛泽春, 吕立新. 用双广义线性模型预测非寿险未决赔款准备金[J]. 统计研究, 2005(8): 52-55.

[4] 孟生旺. 非寿险准备金评估的广义线性模型[J]. 统计与信息论坛, 2009, 24(6): 3-7.

[5] Peters, G.W., Shevchenko, P.V. and Wüthrich, M.V. (2009) Model Uncertainty in Claims Reserving within Tweedie's Compound Poisson Models. *ASTIN Bulletin*, **39**, 1-33. https://doi.org/10.2143/AST.39.1.2038054

[6] Antonio, K. and Beirlant, J. (2008) Issues in Claims Reserving and Credibility: A Semiparametric Approach with

Mixed Models. *Journal of Risk and Insurance*, **75**, 643-676. https://doi.org/10.1111/j.1539-6975.2008.00278.x

[7] Shi, P., Basu, S. and Meyers, G.G. (2012) A Bayesian Log-Normal Model for Multivariate Loss Reserving. *North American Actuarial Journal*, **16**, 29-51. https://doi.org/10.1080/10920277.2012.10590631

[8] Zhang, Y. (2010) A General Multivariate Chain Ladder Model. *Insurance*: *Mathematics and Economics*, **46**, 588-599. https://doi.org/10.1016/j.insmatheco.2010.03.002

[9] Hess, K.T., Schmidt, K.D. and Zocher, M. (2006) Multivariate Loss Prediction in the Multivariate Additive Model. *Insurance*: *Mathematics and Economics*, **39**, 185-191. https://doi.org/10.1016/j.insmatheco.2006.02.004

[10] Merz, M. and Wüthrich, M.V. (2008) Prediction Error of the Multivariate Chain Ladder Reserving Method. *North American Actuarial Journal*, **12**, 175-197. https://doi.org/10.1080/10920277.2008.10597509

[11] Merz, M. and Wüthrich, M.V. (2009) Prediction Error of the Multivariate Additive Loss Reserving Method for Dependent Lines of Business. *Variance*, **3**, 131-151.

[12] Zhang, Y. and Dukic, V. (2013) Predicting Multivariate Insurance Loss Payments under the Bayesian Copula Framework. *Journal of Risk and Insurance*, **80**, 891-919. https://doi.org/10.1111/j.1539-6975.2012.01480.x

[13] Shi, P. and Hartman, B.M. (2016) Credibility in Loss Reserving. *North American Actuarial Journal*, **20**, 114-132. https://doi.org/10.1080/10920277.2015.1109456

[14] Lopes, H., Barcellos, J., Kubrusly, J. and Fernandes, C. (2012) A Non-Parametric Method for Incurred but Not Reported Claim Reserve Estimation. *International Journal for Uncertainty Quantification*, **2**, 39-51. https://doi.org/10.1615/Int.J.UncertaintyQuantification.v2.i1.40

[15] Lally, N. and Hartman, B. (2018) Estimating Loss Reserves Using Hierarchical Bayesian Gaussian Process Regression With Input Warping. *Insurance*: *Mathematics and Economics*, **82**, 124-140. https://doi.org/10.1016/j.insmatheco.2018.06.008

[16] Neal, R.M. (1997) Monte Carlo Implementation of Gaussian Process Models for Bayesian Regression and Classification. ArXiv Preprint Physics/9701026.

[17] Flaxman, S., Gelman, A., Neill, D., *et al*. (2015) Fast Hierarchical Gaussian Processes. Manuscript in Preparation.

[18] Rasmussen, C.E. and Williams, C.K.I. (2006) Gaussian Processes for Machine Learning. MIT Press, Cambridge. https://doi.org/10.7551/mitpress/3206.001.0001

[19] Ludkovski, M. and Zail, H. (2022) Gaussian Process Models for Incremental Loss Ratios. *Variance*, **15**.

[20] Meyers, G. (2011) National Association of Insurance Commissioners Schedule P Data. https://www.casact.org/publications-research/research/research-resources/loss-reserving-data-pulled-naic-schedule-p

[21] Guszcza, J. (2008) Hierarchical Growth Curve Models for Loss Reserving. *Casualty Actuarial Society Forum*, **3**, 146-173.