2022.3.21

固定效应项正则

第 i 条序列、第 t 时刻、对分位数 τ 的固定效应模型组合结果 $Q_{it,\tau}$ 如下所示：

$$Q_{it,\tau} = \alpha_{i,\tau} + \boldsymbol{f_{it,\tau}^T}\boldsymbol{\beta_\tau}$$

其中 $\boldsymbol{f_{it,\tau}^T} = (f_{it1,\tau}, f_{it2,\tau}, \ldots, f_{itk,\tau})$ 是 k 个模型的 τ 分位数预测结果，$\boldsymbol{\beta_\tau}$ 是模型组合的系数，$\alpha_{i,\tau}$ 是固定效应截距项。对该模型在 `rgpd` 的系数估计，是最小化下式：

$$\min_{\alpha,\beta} \sum_{j=1}^{q}\sum_{t=1}^{T}\sum_{i=1}^{N} w_j \rho_{\tau_j}(y_{it} - \alpha_{i,\tau_j} - \boldsymbol{f_{it,\tau_j}^T}\boldsymbol{\beta_{\tau_j}}) + \lambda \sum_{i=1}^{N} |\alpha_{i,\tau_j}|$$

这里 $\rho_{\tau_j}()$ 是 Pinball Loss 损失函数，λ 是固定效应的正则项系数。w_j 是同时对多个分位数进行优化时，对不同分位数施加的权重。因为实验中每次回归只优化一个分位数，故 w_j 及对应的优化分位数个数 q 均为1。这里探讨是否引入正则项对组合的影响。

这里对比在上次的5个模型组合条件下，对固定效应项施加正则（默认 Lasso 的系数是1）以及不加正则项的结果，比较组合模型的预测结果以及系数变化。首先是组合结果 Pinball loss 对比：

分位数	0.01	0.025	0.165	0.25	0.5	0.75	0.835	0.975	0.99
固定效应有惩罚	0.0146	0.0363	0.2168	0.3093	0.4943	0.4873	0.4165	0.1340	0.0710
固定效应无惩罚	0.0217	0.0411	0.2197	0.3104	0.4951	0.4927	0.4263	0.1634	0.1234
LSTM 元学习	0.0146	0.0364	0.2207	0.3142	0.5058	0.5018	0.4354	0.1518	0.0716

可以看到，没有惩罚项，则预测效果要变差很多，两端的预测差距则更为明显，甚至两端的预测结果不如不含截距项的组合。而系数对比则进一步体现了有无惩罚项的差距，有惩罚项和无惩罚项的预测结果回归系数如下：

分位数	0.01	0.025	0.165	0.25	0.5	0.75	0.835	0.975	0.99
gam	6.6E-14	2.4E-14	1.7E-12	6.0E-02	7.6E-02	2.2E-12	3.9E-02	8.5E-02	1.1E-01
vz	1.0E+00	3.3E-01	4.9E-13	5.9E-15	1.0E-14	2.2E-13	2.0E-02	1.3E-01	2.1E-01
wss	-2.0E-12	1.7E-01	7.0E-13	-1.4E-19	-1.3E-14	-7.0E-14	9.3E-05	3.3E-02	5.9E-02
poisson_damped	1.7E-13	8.4E-14	9.5E-13	2.1E-14	7.6E-02	6.2E-12	1.8E-01	1.9E-01	2.3E-01
poisson_undamped	1.4E-12	4.5E-01	6.7E-01	7.2E-01	7.7E-01	1.0E+00	7.6E-01	5.8E-01	4.1E-01
系数和	1.0000	0.9451	0.6667	0.7800	0.9226	1.0000	1.0006	1.0198	1.0176

分位数	0.01	0.025	0.165	0.25	0.5	0.75	0.835	0.975	0.99
gam	1.5E-12	4.4E-15	7.4E-13	2.5E-13	2.6E-14	3.6E-14	1.3E-13	4.1E-13	7.6E-12
vz	-3.0E-01	-5.3E-01	-1.1E-12	-1.1E-13	-4.3E-15	-9.5E-15	-4.7E-14	-1.4E-13	-1.3E-12
wss	7.0E-01	2.7E-01	-2.5E-12	-1.2E-13	-1.0E-14	-1.0E-14	-4.6E-14	-3.1E-13	-4.8E-12
poisson_damped	-1.2E-11	-2.1E-14	-5.3E-13	1.5E-13	-1.3E-15	1.0E-14	7.4E-14	1.5E-12	2.7E-11
poisson_undamped	-1.2E-11	-1.7E-14	-7.2E-13	4.3E-14	-1.8E-15	1.0E-15	5.4E-14	7.9E-13	1.5E-11
系数和	0.3986	-0.2693	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

可以看到，有惩罚项的回归系数求和基本接近1，并且可以选出合适的模型、排除不合适的模型；而无惩罚项的模型系数大多数为0，说明发生了过拟合，只靠截距项进行组合预测，组合的解释力下降严重。因此，如果对每个样本都设置固定效应，应该加入惩罚项。（联想到上次汇报时 LSTM 模型加入截距后发生的过拟合，故无论怎样引入截距，都应考虑惩罚项问题）。

惩罚项应该还有调整的空间，打算在模型池确定之后再加以调整。

截距项

这一节在5模型组合的条件下，探讨不引入固定效应但保留截距，以及完全不存在截距的回归组合效果。首先比较有截距、无截距但权重无限制、无截距且权重设置为1（由无约束的系数归一化得到；尝试过通过同时减去某一项构造约束分位数回归，但是结果过于奇怪，原因是包含较多的0，不易回归）的 Pinball loss 结果：

分位数	0.01	0.025	0.165	0.25	0.5	0.75	0.835	0.975	0.99
有截距	0.0146	0.0372	0.2230	0.3184	0.5491	0.5070	0.4354	0.1380	0.0717
无截距，无约束	0.0146	0.0372	0.2230	0.3184	0.5491	0.5070	0.4354	0.1380	0.0717
无截距，有约束	0.0146	0.0364	0.2232	0.3196	0.5101	0.5070	0.4372	0.1421	0.0747
固定效应	0.0146	0.0363	0.2168	0.3093	0.4943	0.4873	0.4165	0.1340	0.0710
LSTM 元学习	0.0146	0.0364	0.2207	0.3142	0.5058	0.5018	0.4354	0.1518	0.0716

在取消固定效应后，基本结果都会变差，说明固定效应是 Global 组合必需的部分。去固定效应后，有截距和无截距的结果一样，从下文系数的对比中可以解释，二者系数几乎是一样的。加入归一化约束后，两个结果互有高低，没有完全体现出无约束损失小于有约束的理论优势（因为也不是受约束回归）。此外，对于普通分位数回归与 LSTM 元学习组合的结果对比，发现普通回归基本无法战胜基于特征的元学习，说明：基于 LSTM 的元学习方法在加权组合中应是一较好方法，因其利用了特征的信息，并且使用更复杂的模型求解权重。

为进一步观察，展示有/无截距的系数结果，展示如下：

有截距的回归结果：

分位数	0.01	0.025	0.165	0.25	0.5	0.75	0.835	0.975	0.99
截距	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0324	0.0482
gam	0.0000	0.0000	0.0153	0.0969	0.0582	0.0267	0.0932	0.1340	0.1404
vz	0.6914	0.8095	0.0000	0.0000	-0.0591	0.0000	0.0843	0.3936	0.4031
wss	0.2320	0.1190	0.3571	0.1372	0.1192	0.0000	0.0374	0.0000	0.0635
poisson_damped	0.0000	0.0000	0.0000	0.0000	0.1270	0.3200	0.2998	0.2702	0.2773
poisson_undamped	0.0000	0.0000	0.5000	0.6197	0.3773	0.6533	0.5227	0.3037	0.2246

无截距的回归结果：

分位数	0.01	0.025	0.165	0.25	0.5	0.75	0.835	0.975	0.99
gam	0.0000	0.0000	0.0153	0.0969	0.0582	0.0267	0.0932	0.1270	0.1366
vz	0.6914	0.8095	0.0000	0.0000	-0.0591	0.0000	0.0843	0.3757	0.3960
wss	0.2320	0.1190	0.3571	0.1372	0.1192	0.0000	0.0374	0.0227	0.0743
poisson_damped	0.0000	0.0000	0.0000	0.0000	0.1270	0.3200	0.2998	0.2781	0.2829
poisson_undamped	0.0000	0.0000	0.5000	0.6197	0.3773	0.6533	0.5227	0.3008	0.2245

两表对比，在无固定效应的回归中，有无截距的差距并不大，可能是因为各模型分位数预测结果已与真实数据分位数较接近的缘故。综上，在 Global 组合机制中，以固定效应表示的截距起到了重要作用，需要重视。

对于基于回归的模型组合，有以下三种形式：

- $Q_{it,\tau} = \boldsymbol{f}_{\boldsymbol{it,\tau}}^{\boldsymbol{T}}\boldsymbol{\beta_\tau}$ （线性加权平均）
- $Q_{it,\tau} = \alpha_\tau + \boldsymbol{f}_{\boldsymbol{it,\tau}}^{\boldsymbol{T}}\boldsymbol{\beta_\tau}$ （有截距的线性组合，但不使用固定效应）
- $Q_{it,\tau} = \alpha_{i,\tau} + \boldsymbol{f}_{\boldsymbol{it,\tau}}^{\boldsymbol{T}}\boldsymbol{\beta_\tau}$ （固定效应回归组合）

如果 Q 是（均值）点预测，则三个组合可看成不同的线性回归，其（样本内）残差平方和由上到下依次减小；如果 Q 是分位数预测，基于回归的直觉，应该也是类似的结论，但由于分位数回归的参数不解析，能否给出一个理论上的证明？

模型池更新——引入参数分布模型

模型池中加入 Snyder 等（2012）的模型：数据分布设置为泊松、负二项、hurdle-shifted 泊松分布，而与分布相关的均值有以下三种变化模式：不变、平稳变化、非平稳变化。分布与均值变化的公式如下所示：

Table 1
Count distributions used in the empirical study.

Distribution	Mass function	Parameter restrictions	Mean (μ)
Poisson	$\frac{\lambda^y}{y!}\exp(-\lambda)$	$\lambda > 0$	λ
Negative binomial	$\frac{\Gamma(a+y)}{\Gamma(a)y!}\left(\frac{b}{1+b}\right)^a\left(\frac{1}{1+b}\right)^y$	$a > 0, b > 0$	$\frac{a}{b}$
Hurdle shifted Poisson	$\begin{cases} q & y = 0 \\ p\lambda^{y-1}\exp(-\lambda)/(y-1)! & y = 1, 2, \ldots \end{cases}$	$p \geq 0,\ q > 0, \lambda > 0, p + q = 1$	$p(\lambda + 1)$

Table 2
Recurrence relationships for the mean.

Relationship	Recurrence relationship	Restrictions
Static	$\mu_t = \mu_{t-1}$	
Damped dynamic	$\mu_t = (1 - \phi - \alpha)\mu + \phi\mu_{t-1} + \alpha y_{t-1}$	$\mu > 0, \phi > 0, \alpha > 0$ $\phi + \alpha < 1$
Undamped dynamic	$\mu_t = \delta\mu_{t-1} + \alpha y_{t-1}$	$\delta > 0, \alpha > 0$ $\delta + \alpha = 1$

参数估计方法可使用极大似然估计，且对于均值不变的模型，泊松与 hurdle-shifted 泊松分布的 MLE 是可以用均值来解析求解的；负二项分布均值不时变模型的两个参数使用矩估计得到（为了计算速度）。由于 hurdle-shifted 泊松分布在求似然迭代的过程中会遇到数值计算问题（负数求对数），以及其与负二项分布作用基本相同（解决过度分散问题），故在参数分布模型的引入中只考虑泊松与负二项。各模型的预测误差结果及组合结果展示如下：

分位数	**0.01**	**0.025**	**0.165**	**0.25**	**0.5**	**0.75**	**0.835**	**0.975**	**0.99**
quantGAM	0.0195	0.0455	0.2459	0.3494	0.5638	0.5604	0.4787	0.1601	0.0901
VZ	0.0146	0.0364	0.2296	0.3320	0.5374	0.5342	0.4566	0.1446	0.0742
WSS	0.0146	0.0365	0.2336	0.3427	0.5740	0.5672	0.4793	0.1519	0.0781
poisson_static	0.0405	0.0704	0.2718	0.3675	0.5535	0.5544	0.4901	0.2237	0.1570
poisson_damped	0.0237	0.0491	0.2394	0.3325	0.5135	0.5114	0.4454	0.1766	0.1141
poisson_undamped	0.0244	0.0496	0.2371	0.3280	0.5073	0.5100	0.4498	0.2020	0.1438
nb_static	0.0147	0.0367	0.2349	0.3463	0.6026	0.6370	0.5432	0.1534	0.0779
nb_damped	0.0154	0.0376	0.2285	0.3303	0.5393	0.5341	0.4617	0.1638	0.0888
nb_undamped	0.0147	0.0364	0.2247	0.3234	0.5166	0.5198	0.4578	0.1830	0.1164
固定效应回归组合	**0.0146**	**0.0362**	**0.2166**	**0.3087**	**0.4940**	**0.4866**	**0.4162**	**0.1341**	**0.0707**
简单平均	0.0182	0.0408	0.2277	0.3250	0.5236	0.5207	0.4440	0.1399	0.0731
倒数损失平均	0.0167	0.0390	0.2269	0.3255	0.5265	0.5225	0.4449	0.1392	0.0723
面板回归组合	0.0146	**0.0363**	**0.2168**	**0.3093**	**0.4943**	**0.4873**	**0.4165**	**0.1340**	**0.0710**

（倒数损失平均，指以各基准方法 Pinball loss 倒数作权值的加权平均；固定效应组合固定效应惩罚项系数为1）

加入了7种基准模型后，发现：

- 负二项比泊松模型预测效果好；
- 均值是平稳过程变化与非平稳过程变化的结果互有优劣，可能非平稳过程稍好；但是时变参数结果一般好于非时变参数。
- 即使池内有一些效果不佳的模型，固定效应回归组合也能找到较好的组合；也能从下面的系数中看出：往往只有几个模型的系数有显著正值，说明回归组合同时有选择的作用。
- **共线性**！

分位数	**0.01**	**0.025**	**0.165**	**0.25**	**0.5**	**0.75**	**0.835**	**0.975**	**0.99**
gam	0.00	0.00	0.00	0.01	0.05	0.05	0.05	0.08	0.10
vz	-0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.12
wss	-0.25	-0.24	-0.09	-0.01	0.00	0.00	0.00	0.08	0.08
poisson_damped	0.00	0.00	0.00	0.00	0.02	0.10	0.13	0.16	0.20
poisson_undamped	0.00	0.20	0.47	0.57	0.66	0.55	0.46	0.30	0.16
poisson_static	0.00	0.00	0.00	0.00	0.13	0.05	0.07	0.01	0.00
nb_static	0.46	0.43	0.48	0.32	-0.08	-0.01	0.00	-0.09	-0.05
nb_damped	0.00	0.02	0.00	0.00	0.00	0.00	0.02	0.04	0.04
nb_undamped	1.15	0.80	0.21	0.12	0.11	0.26	0.27	0.27	0.24
系数和	1.33	1.21	1.08	1.01	0.90	0.99	1.00	0.91	0.90

特征不时变的影响

以之前的5模型组合为例，使用不时变的特征进行网络训练，观察预测损失的变化。二者差别不大，但多数分位数的结果，使用时变特征的结果会更好；但是，时变特征中只有近7期/28期均值、标准差有明显时变性，因此需要再仔细寻找时变特征，进行元模型训练。

分位数	0.01	0.025	0.165	0.25	0.5	0.75	0.835	0.975	0.99
LSTM 元学习	0.0146	0.0364	0.2207	0.3142	0.5058	0.5018	0.4354	0.1518	0.0716
LSTM 元学习-不时变特征	0.0146	0.0364	0.2212	0.3175	0.5073	0.5064	0.4397	0.1400	0.0717

之后计划

- 池内基准方法：打算再加入两个机器学习方法（比如 LSTM/LightGBM 等），一个针对分布，一个针对分位数，以接近M5的方法；
- 进一步考虑时变模式及基于特征的问题，看看最近的预测组合是怎么做时变的（点、分位数、密度都可以参考）；
- 考虑在回归组合中加入特征作为协变量（解释性、时变性、进一步修正）；进一步研究固定效应项的作用（希望找到一些理论结果），以及如何在其它 Global 方法中引入截距项。