Заметки по многошаговому прогнозированию

Зехов Матвей

22 октября 2019 г.

Содержание

1 Основные стратегии многошагового прогнозирования

1.1 Модель

Общая идея всех подходов: Больше параметров -> точнее модель, более гибкая, меньше сдвиг, больше дисперсия. И наоборот.

Если нам важно более направление, чем значение прогноза, то следует выбирать модель с меньшим смещением, так как большое смещение искривляет направление. А если у нас короткие ряды, склонные к переобучению, то - с маленькой дисперсией.

Нам необходимо спрогнозировать временной ряд из T наблдений на H шагов вперёд. Будем предполагать, что данные пришли к нам из модели (возможно нелинейной) следующего вида:

$$y_t = f\left(x_{t-1}\right) + \varepsilon_t \text{ with } x_t = \left[y_t, \ldots, y_{t-d+1}\right]'$$

ε_t - iid, среднее ноль, дисперсия σ^2, $\kappa = \mathbb{E}\left(\varepsilon^4\right) > 0$.

Процесс специфицируется функцией f, размерностью эмбеддингов d и ошибкой ε_t Цель – оценить условное среднее $\mu_{t+h|t} = \mathrm{E}\left(y_{t+h}|x_t\right)$, и мы будем пытаться использовать различные стратегии аппроксимации $\mu_{t+h|t}$.

При горизонте прогнозирования один имеем просто $\mu_{t+1|t} = f\left(\boldsymbol{x}_t\right)$. Если f линейная, можно записать более общую формулу.

$$\mu_{t+h|t} = \begin{cases} f\left(\left[f^{(h-1)}\left(\boldsymbol{x}_t\right), \ldots, f^{(h-d)}\left(\boldsymbol{x}_t\right)\right]'\right), & \text{if } h > 0 \\ x_t' w_h, & \text{if } 1 - d \leq h \leq 0 \end{cases}$$

где w_h имеет единицу на $j = 1 - h$ позиции и нули в иных. Если функция f линейна, то условное среднее можно посчитать рекурсивно. В иных случаях при $h > 1$ и нелинейной f нет простой формы подсчёта $\mu_{t+h|t}$

Каждая стратегия включает оценку одной или более моделей, которые не обязательно той же формы, что и f и могут иметь иную размерность эмбеддингов. Для одношаговых прогнозов мы будем оценивать модель $y_t = m\left(\boldsymbol{x}_{t-1}; \boldsymbol{\theta}\right) + e_t$ где $\boldsymbol{x}_t = \left[y_t, \ldots, y_{t-p+1}\right]'$. Мы будем оценивать форму m, параметры θ и размерность p. Если форма m совпадает с формой f, то будем писать $m \asymp f$. В идеале мы бы хотели, чтобы $p = d, m \asymp f$ и θ были близки к истинным параметрам, но мы не будем придерживаться этих предположений из-за возможных ошибок спецификации.

Обозначим $\hat{m}^{(h)}\left(\boldsymbol{x}_t\right)$ прогнозы конкретной стратегии на горизонте h и обозначим $m^{(h)} = \mathbb{E}\left[\hat{m}^{(h)}\left(\boldsymbol{x}_t\right)|\boldsymbol{x}_t\right]$

Тогда MSE прогноза на горизонте h можно записать следующей формулой:

$$\mathrm{MSE}_h = \mathbb{E}\left[\left(y_{t+h} - \hat{m}^{(h)}\left(\boldsymbol{x}_t\right)\right)^2\right] =$$

$$\underbrace{\mathbb{E}\left[\left(y_{t+h} - \mu_{t+h|t}\right)^2\right]}_{\text{Noise}} + \underbrace{\left(\mu_{t+h|t} - m^{(h)}\left(\boldsymbol{x}_t\right)\right)^2}_{\text{Bias } \mathrm{b}_h^2} + \underbrace{\mathbb{E}\left[\left(\hat{m}^{(h)}\left(\boldsymbol{x}_t\right) - m^{(h)}\left(\boldsymbol{x}_t\right)\right)^2\right]}_{\text{Variance}}$$

Декомпозиция получена следующим образом:

$$\begin{aligned}
&\mathrm{E}\left[(y - \hat{f})^2\right] \\
&= \mathrm{E}\left[(f + \varepsilon - \hat{f})^2\right] \\
&= \mathrm{E}\left[(f + \varepsilon - \hat{f} + \mathrm{E}[\hat{f}] - \mathrm{E}[\hat{f}])^2\right] \\
&= \mathrm{E}\left[(f - \mathrm{E}[\hat{f}])^2\right] + \mathrm{E}\left[\varepsilon^2\right] + \mathrm{E}\left[(\mathrm{E}[\hat{f}] - \hat{f})^2\right] + 2\mathrm{E}[(f - \mathrm{E}[\hat{f}])\varepsilon] + 2\mathrm{E}[\varepsilon(\mathrm{E}[\hat{f}] - \hat{f})\Big] \\
&+ 2\mathrm{E}[(\mathrm{E}[\hat{f}] - \hat{f})(f - \mathrm{E}[\hat{f}])] \\
&= (f - \mathrm{E}[\hat{f}])^2 + \mathrm{E}\left[\varepsilon^2\right] + \mathrm{E}\left[(\mathrm{E}[\hat{f}] - \hat{f})^2\right] + 2(f - \mathrm{E}[\hat{f}])\mathrm{E}[\varepsilon] + 2\mathrm{E}[\hat{f}] - \hat{f}\Big] + 2\mathrm{E}[\mathrm{E}[\hat{f}] - \hat{f}](f - \mathrm{E}[\hat{f}]) \\
&= (f - \mathrm{E}[\hat{f}])^2 + \mathrm{E}\left[\varepsilon^2\right] + \mathrm{E}\left[(\mathrm{E}[\hat{f}] - \hat{f})^2\right] \\
&= (f - \mathrm{E}[\hat{f}])^2 + \mathrm{Var}[y] + \mathrm{Var}[\hat{f}] \\
&= \mathrm{Bias}[\hat{f}]^2 + \mathrm{Var}[y] + \mathrm{Var}[\hat{f}] \\
&= \mathrm{Bias}[\hat{f}]^2 + \sigma^2 + \mathrm{Var}[\hat{f}]
\end{aligned}$$

Более того, сдвиг можно разложить ещё на две составляющие:

$$B_h = \mathbb{E}_{x_t}\left[\left(\mu_{t+h|t} - \mathbb{E}_{Y_T}\left[m\left(x_t; \hat{\theta}_{Y_T}; h\right)\right]\right)^2\right]$$
$$= \mathbb{E}_{x_t}[\underbrace{\left(\mu_{t+h|t} - m\left(x_t; \theta^*; h\right)\right)}_{A} \quad + \underbrace{m\left(x_t; \theta^*; h\right) - \mathbb{E}_{Y_T}\left[m\left(x_t; \hat{\theta}_{Y_T}; h\right)\right]\Big)^2}_{\text{B}}]$$

В этой части сознательно использованы несколько другие, более точные обозначения, которые позволят разобраться полнее. Интуитивно понять аналогию несложно. Часть А отображает различие между условным средним семейства моделей, которое мы рассматриваем. Например, если первое - нелинейное, а мы оцениваем линейную модель. Тогда правая часть А будет наилучшей из достижимых оценкой линейных параметров. Следует обратить внимание, что подставлены именно оптимальные параметры. Тогда А обозначает ограничения выбранной нами модели относительно истинной зависимости.

Часть В отображает ошибку вследствие ограничинности временных рядов до T наблюдений. Она выражает как конечность выборки влияет на сдвиг. Это можно понять по тому, что в левой части В подставлены оптимальные параметры, а в правой - оценённые по конечному ряду.

Следовательно, даже если мы подберём оптимальную модель, то мы не сможем избавиться от части В вследствие окнечной выборки.

Важно заметить, что компонента шума не зависит ни от рассматриваемой модели, ни от сратегии построения прогноза. Она зависит только от процесса генерации данных.

При росте числа наблюдений дисперсия будет сходиться к нулю, и далее она исключается из рассмотрения. Далее разложим левую часть (Noise) для горизонта h = 2. Аналогично для любого другого.

$$y_{t+2} = f\left(y_{t+1}, \ldots, y_{t-d+2}\right) + \varepsilon_{t+2} = f\left(f\left(x_t\right) + \varepsilon_{t+1}, \ldots, y_{t-d+2}\right) + \varepsilon_{t+2}$$

Используя разложение Тейлора до второго порядка (как я понял, относительно переменной ε_t), получим:

$$y_{t+2} \approx f\left(f\left(\boldsymbol{x}_t\right), \ldots, y_{t-d+2}\right) + \varepsilon_{t+1} f_{x_1} + \frac{1}{2}\left(\varepsilon_{t+1}\right)^2 f_{x_1 x_1} + \varepsilon_{t+2},$$

где f_{x_1} – первая производная f по её первому аргументу, а $f_{x_1 x_1}$ – вторая производная дважды по первому аргументу

Тогда шум можно раскрыть по следующей формуле (правая часть это математическое ожидание левой во второй строке):

$$\mathbb{E}\left[\left(y_{t+2} - \mu_{t+2|t}\right)^2\right]$$
$$\approx \mathbb{E}\left[\left(f\left(f\left(\boldsymbol{x}_t\right), \ldots, y_{t-d+2}\right) + \varepsilon_{t+1} f_{x_1} + \tfrac{1}{2}\varepsilon_{t+1}^2 f_{x_1 x_1} + \varepsilon_{t+2} - f\left(f\left(\boldsymbol{x}_t\right), \ldots, y_{t-d+2}\right) - \tfrac{1}{2}\sigma^2 f_{x_1 x_1}\right)^2\right]$$
$$= \sigma^2\left(1 + f_{x_1}^2\right) + \tfrac{1}{4}\left(\kappa - \sigma^4\right) f_{x_1 x_1}^2$$

Тогда MSE на горизонте $h = 2$ будет равно:

$$\text{MSE}_2 \approx \sigma^2\left(1 + f_{x_1}^2\right) + \frac{1}{4}\left(\kappa - \sigma^4\right) f_{x_1 x_1}^2 + \left(\mu_{t+h|t} - m^{(h)}\left(x_t\right)\right)^2$$

1.2 Рекурсивная стратегия

При рекурсивной стратегии оценивается модель

$$y_t = m\left(\boldsymbol{x}_{t-1}; \boldsymbol{\theta}\right) + e_t, \qquad \text{where } \boldsymbol{x}_t = \left[y_t, \ldots, y_{t-p+1}\right]'$$

где $\mathbb{E}[e_t] = 0$, целевая функция: $\mathbb{E}\left[\left(y_{t+1} - m\left(\boldsymbol{x}_t; \boldsymbol{\theta}\right)\right)^2 | \boldsymbol{x}_t\right]$, а параметры оцениваются следующим образом:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \sum_t \left(y_t - m\left(\boldsymbol{x}_{t-1}; \boldsymbol{\theta}\right)\right)^2$$

Прогнозы вычисляются рекурсивно относительно предыдущих:

$$\hat{m}^{(h)}\left(x_t\right) = \begin{cases} \hat{m}\left(\left[\hat{m}^{(h-1)}\left(x_t\right), \ldots, \hat{m}^{(h-p)}\left(x_t\right)\right]'\right), & \text{if } h > 0 \\ x_t' w_h, & \text{if } 1 - p \leq h \leq 0 \end{cases}$$

где $\hat{m}(x) = m(x; \hat{\theta})$. Иногда это называют итеративным многошаговым прогнозированием.

Элемент смещения $\delta\left(z_t; \boldsymbol{\theta}\right)$ возникает из-за:

- ☼ Недостатка гибкости рассматриваемой модели.
- ☼ Потенциального пропуска регрессоров во входных данных
- ☼ Неадекватного алгоритма оценки параметров θ.

Вариативность прогноза описывается элементом $\eta\left(z_t; \boldsymbol{\theta}\right) \varepsilon_\eta$. Она возникает вследствие следующих причин:

- ☼ Ограниченности длины ряда при оценке θ
- ☼ Включение в z потенциально избыточных или бессмысленных переменных.
- ☼ Избыточная гибкость модели и переобучение

Предсказания рекурсивной стратегии:

$$m_h\left(x_t; \hat{\theta}\right) = \underbrace{\underbrace{\underbrace{f\left(\boldsymbol{x}_t\right)}_{\mu_{t+h|t}} + \delta\left(z_t; \boldsymbol{\theta}\right)}_{m(z_t;\theta)} + \eta\left(z_t; \boldsymbol{\theta}\right) \varepsilon_\eta}_{m\left(z_t;\hat{\theta}\right)}$$

Выбор квадратичного функционала для $\hat{\theta}$ обеспечивает $m^{(1)}\left(x_t\right) = \mu_{t+1|t}$, и, как следствие, $m \asymp f$ и $p = d$. Следовательно, одношаговый прогноз будет несмещённым. Однако это не сохраняется для порядка 2 и более:

$$\begin{aligned} b_2 = \mu_{t+2|t} - m^{(2)}\left(\boldsymbol{x}_t\right) &\approx f\left(f\left(\boldsymbol{x}_t\right), \ldots, y_{t-d+2}\right) + \frac{1}{2}\sigma^2 f_{x_1 x_1} - m\left(m\left(\boldsymbol{x}_t\right), \ldots, y_{t-p+2}\right) \\ &\approx \left[f\left(f\left(\boldsymbol{x}_t\right), \ldots, y_{t-d+2}\right) - m\left(m\left(\boldsymbol{x}_t\right), \ldots, y_{t-p+2}\right)\right] + \frac{1}{2}\sigma^2 f_{x_1 x_1} \end{aligned}$$

Можно аналогично раскрыть и форму:

$$\begin{aligned} & m\left(x_t; \hat{\theta}; 2\right) = m\left(m\left(z_t; \hat{\theta}\right), \ldots, y_{t-p+2}; \hat{\theta}\right) \approx f\left(f\left(x_t\right), \ldots, y_{t-p+2}\right) \\ & + \delta\left(f\left(x_t\right), \ldots, y_{t-p+2}; \theta\right) + \eta\left(f\left(x_t\right), \ldots, y_{t-p+2}; \theta\right) \varepsilon_{\eta 2} \\ & + \delta\left(z_t; \theta\right) m_{z_1} + \frac{1}{2}\left[\delta\left(z_t; \theta\right)\right]^2 m_{z_1 z_1} + \eta\left(z_t; \theta\right) \varepsilon_{\eta_1} m_{z_1} + \frac{1}{2}\left[\eta\left(z_t; \theta\right) \varepsilon_{\eta_1}\right]^2 m_{z_1 z_1} \end{aligned}$$

Следовательно, даже при идеальных условиях $m \asymp f$ и $p = d$, смещение будет отсутствовать только при линейности истинной функции f.

Для этой стратегии MSE будет равна:

$$\begin{aligned}&\mathrm{MSE}_2^{\text{recursive}}\\&\approx \sigma^2\left(1+f_{x_1}^2\right)+\tfrac{1}{4}\left(\kappa-\sigma^4\right)f_{x_1x_1}^2+\left(\left[f\left(f\left(x_t\right),\ldots,y_{t-d+2}\right)-m\left(m\left(x_t\right),\ldots,y_{t-p+2}\right)\right]+\tfrac{1}{2}\sigma^2 f_{x_1x_1}\right)^2\end{aligned}$$

При выполнении условий $m \asymp f$ и $p = d$ формула упрощается до

$$\mathrm{MSE}_2^{\text{recursive}} \approx \sigma^2\left(1+f_{x_1}^2\right)+\frac{1}{4}\kappa f_{x_1x_1}^2$$

Когда модель неправильно специфицирована или параметры оценены неверно, ашибка существенно больше.

Есть различные вариации:

С оценкой своего параметра для каждого горизонта:

$$\hat{\boldsymbol{\theta}}_h = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}} \sum_t \left[y_t - m^{(h)}\left(\boldsymbol{x}_{t-h};\boldsymbol{\theta}\right)\right]^2$$

Такую стратегию назовём RECMULTI

Минимум по первым H горизонтам:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}} \sum_{h=1}^{H}\sum_t \left[y_t - m^{(h)}\left(\boldsymbol{x}_{t-h};\boldsymbol{\theta}\right)\right]^2$$

Достоинства: мало вычислений, только одна модель, низкая дисперсия, может быть лучше прямого метода, если модель правильно специфицирована

Недостатки: Смещение, только одна модель

В случае AR(p) гауссовского-процесса генерации данных и AR(k) оцениваемой модели, то при $k \geq p$ можно показать, что оценки МНК асимптотически эквивалентны ML и все оценки с индексом $i > p = 0$. В общем, параметры прогноза, которые зависят от параметров оценённой модели по инвариантности оценок ML можно тоже оценить.

$\hat{\mathbf{a}}_n(h,k) = \hat{A}_n^{h-1}(k)\hat{\mathbf{a}}_n(1,k)$

is asymptotically equivalent to the MLE of

$\mathbf{a}(h,k) = (a_1(h,k),\ldots,a_k(h,k))' = A^{h-1}(k)\mathbf{a}(1,k)$

where $k \geq p, A^0(k) = I_k$, and

$A(k) = \left(\mathbf{a}(1,k)|\frac{I_{k-1}}{\mathbf{0}'_{k-1}}\right)$

1.3 Прямая стратегия

Прямая стратегия: давайте оценивать для каждого горизонта свою модель.

$$y_t = m_h\left(y_{t-h},\ldots,y_{t-h-p_h};\boldsymbol{\theta}_h\right)+e_{t,h}$$

Тогда параметры для каждой модели будем оценивать МНК по соответствующим пространствам параметров и $\hat{m}^{(h)}\left(\boldsymbol{x}_t\right) = m_h\left(\boldsymbol{x}_t;\hat{\boldsymbol{\theta}}_h\right)$

$$\hat{\boldsymbol{\theta}}_h = \underset{\boldsymbol{\theta}_h\in\boldsymbol{\Theta}_h}{\operatorname{argmin}} \sum_t \left[y_t - m_h\left(\boldsymbol{x}_{t-h};\boldsymbol{\theta}_h\right)\right]^2$$

Такой подход обычно называют "direct multi-step". Для оценки h моделей требуется больше времени. Теперь мы оцениваем не единую модель, а несколько независимых моделей

Из-за квадратичных отклонений и $m^{(h)}(x_t) = m_h(x_t)$ (истинная модель для горизонта h это h-ая модель)б bias модели при горизонте 2 равен:

$$b_2 = \mu_{t+2|t} - m^{(2)}(\boldsymbol{x}_t) \approx f(f(\boldsymbol{x}_t), \ldots, y_{t-d+2}) + \frac{1}{2}\sigma^2 f_{x_1x_1} - m_2(y_t, \ldots, y_{t-p_2+1})$$

Тогда эта стратегия буде вести к несмещённым оценкам когда $m_2(y_t, \ldots, y_{t-p+1}) \asymp f(f(x_t), \ldots, y_{t-d+2}) + \frac{1}{2}\sigma^2 f_{x_1x_1}$. Эти условия будут выполняться в стучае, если m_2 будет достаточно гибкой моделью.

Можно также дополнительно раскрыть формулу прогноза следующим образом:

$$m\left(\boldsymbol{x}_t; \hat{\boldsymbol{\theta}}_2; 2\right) = \underbrace{\underbrace{\mu_{t+2|t} + \delta(\boldsymbol{r}_t; \boldsymbol{\theta}_2)}_{m_2(\boldsymbol{r}_t; \boldsymbol{\theta}_2)} + \eta(\boldsymbol{r}_t; \boldsymbol{\theta}_2)\varepsilon_\eta}_{m_2\left(\boldsymbol{r}_t; \hat{\boldsymbol{\theta}}_2\right)}$$

Следует обратить внимание, что в отличие от рекурсивной стратегии, в прямой сразу может возникнуть условное среднее без рекурсивного вызова функции от x_t

Во всех иных стратегиях прогноз будет выглядеть следующим образом:

$$g(\boldsymbol{x}_t; \hat{\boldsymbol{\gamma}}; 2) = \underbrace{\underbrace{\mu_{t+2|t} + \delta(\boldsymbol{r}_t; \boldsymbol{\gamma})}_{m_2(\boldsymbol{r}_t; \boldsymbol{\gamma})} + \eta(\boldsymbol{r}_t; \boldsymbol{\gamma})\varepsilon_\eta}_{m_2(\boldsymbol{r}_t; \hat{\gamma})}$$

где гамма, это набор параметров, различный для каждой стратегии. Все стратегии за исключение рекурсивной используют h-шаговую ошибку в качестве целевой функции.

Вспоминая формулу

$$\mathrm{MSE}_2 \approx \sigma^2\left(1 + f_{x_1}^2\right) + \frac{1}{4}\left(\kappa - \sigma^4\right) f_{x_1x_1}^2 + \left(\mu_{t+h|t} - m^{(h)}(x_t)\right)^2$$

можем получить аналогичную для прямого случая

$$\mathrm{MSE}_2^{\mathrm{direct}} \approx \sigma^2\left(1 + f_{x_1}^2\right) + \frac{1}{4}\left(\kappa - \sigma^4\right) f_{x_1x_1}^2 + \left[f(f(x_t), \ldots, y_{t-d+2}) + \frac{1}{2}\sigma^2 f_{x_1x_1} - m_2(y_t, \ldots, y_{t-p_2+1})\right]^2$$

Если истинная стратегия несмещённая, то уравнение упростится до

$$\mathrm{MSE}_2^{\mathrm{direct}} \approx \sigma^2\left(1 + f_{x_1}^2\right) + \frac{1}{4}\left(\kappa - \sigma^4\right) f_{x_1x_1}^2$$

В отличие от рекурсивного случае здесь m_2 отражает не всю истинную модель и может быть достаточно гибкой чтобы подстроиться. Следовательно, несмещённый случай возможен. не только в линейном случае.

Получается, в идеальных условиях, когда $m \asymp f$ and $p = d$ для рекурсивной стратегии и прямая стратегия несмещённая, можно посчитать разницу ошибок:

$$\mathrm{MSE}_2^{\mathrm{recursive}} - \mathrm{MSE}_2^{\mathrm{direct}} \approx \frac{1}{4}\sigma^4 f_{x_1x_1}^2$$

Были работы, которые показали, что даже при линейности всех истинных зависимостей и стационарности ряда рекусивная модель хуже, чем прямая.

У прямого метода помимо вычислительной сложности есть ещё проблема построения слишком разных моделей на разных горизонтах. Следовательно прогнозы строятся на независимых моделях с различными формами и условной информцией. Это как раз и увеличивает вариацию прогноза, но зато смещение низкое на нелинейных данных.

Может работать лучше рекурсивной стратегии, если она модель неверно специфицирована, так как ошибки спецификации обычно создают сдвиг. Ещё один недостаток: автокорреляция ошибок.

2 Комбинации прямой и рекурсивной стратегий

2.1 DirRec

Эта смешанная стратегия использует различные наборы параметровθ_h для каждого горизона h, как в прямой стратегии, но включает предыдущие прогнозы вместе с другими входнымиз начениями:

$$\hat{\boldsymbol{\theta}}_h = \underset{\boldsymbol{\theta}_h \in \boldsymbol{\Theta}_h}{\operatorname{argmin}} \sum_t \left[y_t - \left[m_h \left(\hat{m}_{h-1}, \ldots, \hat{m}_1, \boldsymbol{r}_{t-h}; \boldsymbol{\theta}_h \right) \right] \right]^2$$

где $\hat{m}_h$ – это сокращение для $m_h \left(\boldsymbol{r}_{t-h}; \hat{\boldsymbol{\theta}}_h \right)$

Тогда прогнозы для каждого горизонта из соответствующей модели получаются следующим образом: $\hat{\mu}_{T+h|T} = m_h \left(\hat{m}_{h-1}, \ldots, \hat{m}_1, \boldsymbol{r}_T; \hat{\boldsymbol{\theta}}_h \right)$

2.2 MSRV

Ещё одна комбинация прямого и рекурсивного подходов. Если мы представим горизонты прогнозирования следующим образом: $H = L \times R$, то эта стратегия первые L моделей оценит как прямые, а потом использует их R раз для получения H прогнозов:

$$\hat{\mu}_{T+h|T} = \begin{cases} m_l \left(\boldsymbol{r}_T; \hat{\boldsymbol{\theta}}_l \right) & \text{if } h \leq L \\ m_l \left(\hat{m}_{l-1}, \ldots, \hat{m}_1, \boldsymbol{r}'_T; \hat{\boldsymbol{\theta}}_l \right) & \text{if } h > L \end{cases}$$

где $l = (h-1) \,\%L + 1, \hat{m}_l$ это сокращение для . $m_l \left(\boldsymbol{r}_T; \hat{\boldsymbol{\theta}}_l \right)$, и $\boldsymbol{r}'_T \subset \boldsymbol{r}_T$

Преимущество: время вычислений. Однако, из-за рекурсии точность также страдает.

2.3 Стратегия RECTIFY

Основная идея - взять плюсы двух первых. Она начинает с рекурсивных прогнозов и улучшает их так, что они становятся несмещёнными и имеют меньшую ошибку.

Начнём с рекурсивных прогнозов на линейной модели. Они будут смещёнными даже когда процесс правильно специфицирован. Потом скорректируем их моделированием прогнозных ошибок через прямую стратегию и получим несмещённые прогнозы.

Достоинство подхода в том, что он соединяет все модели прямого подхода через единую базовую модель, снизая разрегулировку, порождённую независимыми моделями. Грубо говоря, прямые модели станут более похожими друг на друга.

Обозначим базовую линейную модель как $y_t = z \left(x_{t-1}; \boldsymbol{\theta} \right) + e_t$. Из неё получим рекурсивные прогнозы. После этого мы изменим прогозы базовой модели, применяя модели прямого прогнозирования к ошибкам рекурсиного прогноза. То есть мы будем фитить следующие модели:

$$y_t - z^{(h)} \left(y_{t-h}, \ldots, y_{t-h-p}; \hat{\boldsymbol{\theta}} \right) = r_h \left(y_{t-h}, \ldots, y_{t-h-p_h}; \gamma_h \right) + e_{t,h}, \forall h = 1 \ldots H$$

Все параметры оценены через OLS. Прогнозы получены через суммирование базовой и дополнительной модели: $\hat{m}^{(h)} \left(\boldsymbol{x}_t \right) = \hat{z}^{(h)} \left(\boldsymbol{x}_t \right) + \hat{r}_h \left(\boldsymbol{x}_t \right)$

Пусть $m^{(h)} \left(\boldsymbol{x}_t \right) = \mathbb{E} \left[\hat{m}^{(h)} \left(\boldsymbol{x}_t \right) | \boldsymbol{x}_t \right]$. Тогда смещение можно записать как

$$
\begin{aligned}
b_2 &= \mu_{t+2|t} - m^{(2)}\left(\boldsymbol{x}_t\right) \\
&\approx f\left(f\left(\boldsymbol{x}_t\right), \ldots, y_{t-d+2}\right) + \frac{1}{2}\sigma^2 f_{x_1 x_1} - \left[z\left(z\left(\boldsymbol{x}_t\right), \ldots, y_{t-p+2}\right) + r_2\left(y_t, \ldots, y_{t-p_2+1}\right)\right] \\
&= \left[f\left(f\left(\boldsymbol{x}_t\right), \ldots, y_{t-d+2}\right) - z\left(z\left(\boldsymbol{x}_t\right), \ldots, y_{t-p+2}\right) + \frac{1}{2}\sigma^2 f_{x_1 x_1}\right] - r_2\left(y_t, \ldots, y_{t-p_2+1}\right)
\end{aligned}
$$

Следовательно, стратегия будет несмещённой, когда $r_2\left(y_t, \ldots, y_{t-p_2+1}\right) \asymp f\left(f\left(x_t\right), \ldots, y_{t-d+2}\right) - z\left(z\left(x_t\right), \ldots, y_{t-p+2}\right) + \frac{1}{2}\sigma^2 f_{x_1 x_1}$, то есть когда дополнительная модель достаточно гибкая, чтобы оценить условное среднее ошибок базовой модели.

Смещение базовой модели скорректировано дополнительной моделью. Следовательно, отпадает необходимость условия $z \asymp f$. В случае статьи рассматривалась модель линейной авторигрессии с отбором лага по AIC. Конечно, модель будет смещённой для нелинейных моделей, но она поможет моделировать большую часть сигнала f и будет давать относительно маленькую дисперсию из - за своей линейности. Дополнительная модель должна быть достаточно гибкой и в статье рассматриваются оценки KNN для неё.

Можно как и ранее вывести MSE для этой стратегии.

$$
\begin{aligned}
\mathrm{MSE}_2^{\text{rectify}} \approx & \sigma^2\left(1 + f_{x_1}^2\right) + \frac{1}{4}\left(\kappa - \sigma^4\right) f_{x_1 x_1}^2 \\
& + \left[\left[f\left(f\left(x_t\right), \ldots, y_{t-d+2}\right) - z\left(z\left(x_t\right), \ldots, y_{t-p+2}\right) + \frac{1}{2}\sigma^2 f_{x_1 x_1}\right] - r_2\left(y_t, \ldots, y_{t-p_2+1}\right)\right]^2
\end{aligned}
$$

Когда стратегия несмещённая, она имеет такую же MSE, как и прямая стратегия.

$$
\mathrm{MSE}_2^{\text{recursive}} - \mathrm{MSE}_2^{\text{rectify}} = \mathrm{MSE}_2^{\text{recursive}} - \mathrm{MSE}_2^{\text{direct}} \approx \frac{1}{4}\sigma^4 f_{x_1 x_1}^2
$$

Однако, надо не забывать учитывать дисперсию. Хотя у обеих стратегий они стремятся к нулю, у модели rectify меньшая дисперсия на конечных выборках.

2.4 Оценка суммы сдвига и дисперсии

Для горизонта $h = 2$ можно записать следующую формулу:

$$
B_2\left(\boldsymbol{x}_t\right) + V_2\left(\boldsymbol{x}_t\right) \approx \left(\mu_{t+2|t} - m\left(z_t; \boldsymbol{\theta}; 2\right)\right)^2 + \mathbb{E}_{\boldsymbol{Y}_T}\left[\left(m\left(z_t; \hat{\boldsymbol{\theta}}; 2\right) - m\left(z_t; \boldsymbol{\theta}; 2\right)\right)^2 | \boldsymbol{x}_t\right]
$$

2.5 Регрессионные методы и отбор моделей

В статье - линейная авторегрессия и KNN.

Линейная модель отбирает порядок p по AIC.

KNN - нелинейная и непараметрическая модель. Использовалось взвешенное среднее. k - количество взвешенно усреднённых соседей. Чем больше соседей - тем больше сглаживание, меньше дисперсия, больше смещение и наоборот

2.6 Оценка смещения и дисперсии

$$
\begin{aligned}
\mathrm{MSE}_h &= \frac{1}{LR}\sum_{i=1}^{L}\sum_{j=1}^{R}\left(y_j(h) - \hat{m}_{D^i}^{(h)}\left(\boldsymbol{x}_j\right)\right)^2 \\
&= \text{ Noise }_h + \mathrm{Bias}_h^2 + \text{ Variance }_h
\end{aligned}
$$

Модели генерации данных:

☼ AR(6) - процесс, $\varepsilon_t \sim \text{NID}(0,1)$

$$y_t = 1.32y_{t-1} - 0.52y_{t-2} - 0.16y_{t-3} + 0.18y_{t-4} - 0.26y_{t-5} + 0.19y_{t-6} + \varepsilon_t$$

☼ STAR, $\varepsilon_t \sim \text{NID}\left(0, \sigma^2\right)$, $\sigma^2 = [0.05^2, 0.1^2]$

$$y_t = 0.3y_{t-1} + 0.6y_{t-2} + \left(0.1 - 0.9y_{t-1} + 0.8y_{t-2}\right)\left[1 + e^{(-10y_{t-1})}\right]^{-1} + \varepsilon_t$$

3 Стратегии с множественными горизонтами

Все стратегии, описанные выше, можно отнести к стратегиям с одним горизонтам, то есть стратегиям, когда модель рассматривает каждый горизонт отдельно. Другой тип моделей будет рассматривать несколько горизонтов за один приём. Целевая функция для оценки параметров будет учитывать одновременно ошибки нескольких горизонтов и, соответственно, оценивать один набор параметров.

Идея - учесть общие характеристики моделей нескольких различных горизонтов вследствие автокорреляции.

Учёт параметров в такой присоединённой манере позволит:

☼ Учесть взаимозависимости между различными горизонтами прогнозирования и улучшить обобщающую способность модели.

☼ Избежать разрегулировки вследствие использования слишком разных моделей на каждом горизонте

☼ Это помогает при маленьком размере семпла, используя дополнительные семплы из связанных задач ??

Для получения прогноза мультигоризонтной модели, сравнимого с одногоризонтной, нужно чтобы на горизонте h

3.1 Прямые стратегии

Dirjoint и Recjoint дают больший вес поздним горизонтам, так как усредняют ошибки. Ошибки дальних горизонтах обычно сильно выше. СЛедовательно, возникает ещё смещение на ближних горизонтах. Dirjoint всё же лучше по смещению.

3.1.1 Стратегия DIRJOINT

Эта стратегия передаёт параметры между всеми горизонтами и оценивает параметры по средней ошибке на всех горизонтах.

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}} \sum_t \frac{1}{H} \sum_{h=1}^{H} \left[y_t - m\left(\boldsymbol{r}_{t-h}; \boldsymbol{\theta}\right)\right]^2$$

Эта стратегия применялась в контексте нейронынх сетей. Очень легко понять на примере простого многослойного пресептрона, в котором входной слой имеет размерность p, а выходной - размерность H. Прогнав все наблюдения на нём и минимизировав ошибку как раз и получим формулу выше. Если есть аналитическое решение в разумный срок, то ок, а иначе - прошагаем градиентным спуском.

3.1.2 Стратегия SJOINT

Эта модель также предполагает разделение моделей на L различных многогоризонтных моделей, в которых параметры оцениваются следующим образом:

$$\hat{\boldsymbol{\theta}}_l = \underset{\theta_l \in \boldsymbol{\Theta}_l}{\operatorname{argmin}} \sum_t \frac{1}{R} \sum_{h=(l-1)R+1}^{l \times R} \left[y_t - m_l \left(\boldsymbol{r}_{t-h}; \boldsymbol{\theta}_l \right) \right]^2$$

Горизонты группируются через параметр l, и для каждого из них будет свой набор параметров $\boldsymbol{\theta}_l$

3.1.3 Стратегия DIRJOINTL

Эта модель усредняет ошибки на текущем горизонте, i предыдущих горизонтах и j последующих горизонтах. Аналогично, можно использовать в нейросетях.

$$\hat{\boldsymbol{\theta}}_h = \underset{\boldsymbol{\theta}_h \in \boldsymbol{\Theta}_h}{\operatorname{argmin}} \sum_t \frac{1}{i+j+1} \sum_{h'=h-i}^{h+j} \left[y_t - m_h \left(\boldsymbol{r}_{t-h'}; \boldsymbol{\theta} \right) \right]^2$$

Как по ощущениям, все эти модели пытаются так или иначе сохранить несмещённость, но уменьшить дисперсию, уйдя от количества параметров. Например, DIRJOINT уже начинает сильно напоминать рекурсивную стратегию. Разве что усреднение по горизонтам может дать чуть большую устойчивость. Да и то слегка сомнительно.

3.1.4 Стратегия RECJOINT

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmin}} \sum_t \frac{1}{H} \sum_{h=1}^{H} \left[y_t - m^{(h)} \left(z_{t-h}; \boldsymbol{\theta} \right) \right]^2$$

Различие с обычной рекурсивной стратегией в том, что здесь при оптимизации параметров учитывается влияние всех горизонтов. Эта штука очень похожа на BPTT из LSTM.

3.1.5 Стратегия RECJOINTL

Аналогично можем усреднять не все горизонты, а плюс-минус вперёдд и назад.

$$\hat{\boldsymbol{\theta}}_h = \underset{\boldsymbol{\theta}_h \in \boldsymbol{\Theta}_h}{\operatorname{argmin}} \sum_t \frac{1}{i+j+1} \sum_{h'=h-i}^{h+i} \left[y_t - m^{(h)} \left(\boldsymbol{r}_{t-h'}; \boldsymbol{\theta} \right) \right]^2$$

Количество горизонтов, включённых в какую-либо оптимизируемую функцию далее будем обозначать как $L_h \subseteq \{1 \dots, H\}$. В случае, например, DIRHOINTL: $L_h = \{h - i, \dots, h, \dots, h + i\}$

3.2 Рекурсивные стратегии

3.2.1 Стратегия RECJOINT

Логичная идея: а давайте также усредним по всем горизонтам, но уже рекурсивную стратегию в рамках модели, прогнозирующей рекурсивно.

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmin}} \sum_t \frac{1}{H} \sum_{h=1}^{H} \left[y_t - m^{(h)} \left(z_{t-h}; \boldsymbol{\theta} \right) \right]^2$$

Аналогично, можно усреднять не все горизонты, а несколько вперёд и назад:

$$\hat{\boldsymbol{\theta}}_h = \underset{\boldsymbol{\theta}_h \in \boldsymbol{\Theta}_h}{\operatorname{argmin}} \sum_t \frac{1}{i+j+1} \sum_{h'=h-i}^{h+i} \left[y_t - m^{(h)} \left(\boldsymbol{r}_{t-h'}; \boldsymbol{\theta} \right) \right]^2$$

4 Эксперименты и симуляции

4.1 Параметры симуляций и результаты.

На длинных рядах прямая стратегия превосходит рекурсивную и имеет малый сдвиг.

Хорошая практика – обрезать первые три тысячи сгенерированных наблюдений для стабилизации временного ряда.

На малеьких горизонтах прогнозирования рекурсивная и прямая стратегия очень похожи.

Прямые прогнозы постепенно сходятся к средним прогнозам с ростом длины ряда и горизонта прогнозирования.

Нелинейность ряда и нелинейность модели прогнозирования ухудшает ухудшают сдвиг в рекурсивной стратегии. Рассматривая KNN-KNN можно заметить, что сильно уменьшается сдвиг, но сильно увеличиваеся дисперсия во всех моделях. Следовательно, не лучшая идея брать нелинейную модель в качестве базовой. В данном случае LINARP-KNN сработала лучше. Особенно заметно на коротком горизонте. В сравнении с прямой стратегией этот вариант сильно лучше прямой стратегии на коротких горизонтах и примерно сопоставим на длинных. Также она сильно лучше на коротких рядах и приближается к прямой стратегии с ростом длины ряда.

В принципе, использование rectify всегда может быть оправданным чтобы избежать выбора между рекурсивной и прямой стратегиями.