贝叶斯线性回归在期货交易中的应用

华泰期货研究所 量化策略组
陈维嘉
量化研究员
☎ 0755-23991517
✉ chenweijia@htgwf.com
从业资格号：T236848
投资咨询号：TZ012046

贝叶斯方法简介

贝叶斯方法提供了一种通过计算假设概率来预测未来概率的方法，这种方法是基于假设的先验概率、给定假设下观察到不同数据的概率以及观察到的数据本身而得出的。其方法为，将关于未知参数的先验信息与样本信息综合，再根据贝叶斯公式，得出后验信息，然后根据后验信息去推断未知参数的方法，再由此推断出未来待预测变量的概率分布。

贝叶斯线性回归模型与经典的线性回归模型有很大区别，后者把回归系数看作是固定的未知参数，而前者则把回归系数看作是一个未知的概率分布，然后根据可获得的样本对这些未知分布进行推断。在计算待预测变量的分布时需要根据回归系数的分布在给定自变量的情况下进行采样，从而得到待预测变量的分布。因此通常训练贝叶斯模型和利用该类模型做预测的计算量都会比常规线性回归要大。

本报告尝试利用沪铜期货的基本面周频数据对沪铜期货下一周的收益率分布进行预测，研究结果标明利用贝叶斯模型对期货收益率预测有一定效果。根据该模型的预测结果对 2016 年 3 月至 2017 年 6 月的纯样本外数据进行回测获得年化收益 17.8%，波动率 22%，夏普率 0.8。

贝叶斯方法背景介绍：

比较常用的线性回归其实是频率主义学派的观念：他们认为回归参数是客观存在的，即使是未知的，但都是固定值，不会改变。而贝叶斯学派认为，待定参数和概率一样是一个人对于一件事的信念强度，是主观的。频率学派认为进行一定数量的重复实验后，如果出现某个现象的次数与总次数趋于某个值，那么这个比值就会倾向于固定。最简单的例子就是抛硬币了，在理想情况下，我们知道抛硬币正面朝上的概率会趋向于1/2。但贝叶斯提出了一种截然不同的观念，他们认为概率不应该这么简单地计算，而需要加入先验概率的考虑。先验概率也就是说，我们先设定一个假设，然后我们通过一定的实验来修正这个假设，这就是后验。应用在回归分析上也是类似的道理，我们先对线性回归系数的分布作出一个假设，随后根据训练集的样本去修正这些分布。

这个修正过程通常有两类方法，一种是根据系数的先验分布生成随机数采样模拟，计算此样本观测到的待预测变量的概率，再据此调整先验分布，这就是马尔科夫链蒙特卡诺（MCMC）方法。另外一种则是把待预测变量的概率密度函数参数化表示，把随机采样问题转化为优化问题，即最小化待预测变量的概率密度函数与其真实概率密度之间的距离，这就是变分推断（Variational Inference）。有关研究认为这两种方法的精度比较接近，但变分推断有可能低估预测变量的方差。变分推断与马尔科夫链蒙特卡诺相比最大的优势在于计算量被大大减少了，所以这种方法比较适用于较为复杂的模型。

理论上贝叶斯模型与频率模型相比具有较大优势，首先这类模型更加符合人们对事物的认知过程。人们在认知过程中总会有一个预期，这个就是先验分布。如果是一个完全陌生的事物，那么这个预期就相当于是一个均匀分布，即各种可能性均相等。人们总是通过实践（即观测样本）来积累经验修正认知，这就相当于贝叶斯模型中的推断，即不断修正先验分布。

另外与频率模型相比，贝叶斯模型能把不确定性也考虑进来。我们如果使用贝叶斯模型来对资产未来收益率进行预测的话是可以得到一个概率分布的，这里包含的信息要比单纯的线性回归得出的一个预期收益要多，这个分布在资产配置上可以发挥一定作用，虽然这个分布的实际意义有待进一步研究。由于在预测时模型考虑了回归参数的整个概率分布，所以得出的结果通常对样本中的异常值并不敏感。

最后，也是最重要的一点就是有研究认为模型的贝叶斯化可以看作是正则化的一种有效手段。经典线性模型为了防止过度拟合通常要在目标函数中加入惩罚项，又或者对更大的模型，比如神经网络，需要压制神经元活性(dropout)。而模型的贝叶斯化则是防止过度拟合的一种天然手段。

交易模型简介:

首先考虑多个因子的时间序列，当中包含N个可观测因子和T个时间段。这里使用的可观测因子包括成交量，商品库存和商品上下游产物等信息，因子个数范围一般在 20-30 之间，使用资产价格或周频回报率。但在输入 VAR 模型前会通过主成分分析(PCA)的方法减少至 10 个以下。T为周频数据，时间长度为 1 年-4 年不等，根据预测效果进行调节。

利用贝叶斯模型可对这些期货收益率进行预测

$$\mathrm{r_t} = \mathrm{A_1 f_{t-1}} + \mathrm{A_2 f_{t-2}} + \mathrm{A_3 f_{t-3}} + b \ldots + N(0,\sigma^2) \quad (1)$$

其中$\mathrm{f_{t-1}, f_{t-2}, f_{t-3}}$,..为当前周，上一周和上两周…的特征因子。$\mathrm{A_1, A_2, A_3}$，…和$b$为贝叶斯模型的系数和截距，其先验分布均为标准正态分布。同时假设下周回报率为$\mathrm{r_t}$的预测误差服从正态分布$N(0,\sigma^2)$，其方差σ^2服从标准对数正态分布。这里的参数$\boldsymbol{\theta} = \{\mathrm{A_1, A_2, A_3}, \ldots b, \sigma^2\}$的后验分布都是通过历史数据进行马尔科夫链蒙特卡诺或者变分推断。在推断出后验分布后就可以计算$\mathrm{r_t}$的后验分布了，也就是给定了当前因子$\mathbf{f_{t-1}} = \{\mathrm{f_{t-1}, f_{t-2}, f_{t-3}}, \ldots\}$下的条件概率分布$P(\mathrm{r_t}|\mathbf{f_{t-1}}, \boldsymbol{\theta})$。

由于使用的市场因子数据有 20-30 个，数量较多，直接导入上述贝叶斯模型进行计算运算量较为庞大，所以这里使用了 PCA 的方法以较少变量个数。对N个市场因子计算其日回报率(利率、汇率等因素直接使用)$\mathrm{X_1}, X_2, X_3, \cdots X_N$做线性变换得到$N$个主成分:

$$C_j = a_{j1}X_1 + a_{j2}X_2 + \cdots + a_{jN}X_N, \mathrm{j=1,2,\cdots,N} \quad (2)$$

这N个主成分$C_1, C_2, C_3, \cdots C_N$有以下特征:

(1) 第一主成分C_1的方差最大，第二成分C_2的方差其次，从第一主成分到第二主成分方差依次递减。

(2) 各个主成分间不相关，即对任意主成分C_i和C_j的相关系数为 0。

(3) 变换前N个随机变量$X_1, X_2, X_3, \cdots X_N$的总方差与$N$个主成分$C_1, C_2, C_3, \cdots C_N$的总方差相等。

模型训练与选择

这里使用滚动时间窗口的剩余一步交叉验证法来训练和选择模型超参数。在上述模型中，可以选择的超参数包括 PCA 后的保留因子个数，回溯周数和训练窗口长度 3 个，这三个超参数可以根据模型的样本外预测效果来进行选择。这里以沪铜为例，我们搜集了沪铜自 2010 年 1 月起至 2017 年 6 月份的数据，如果训练窗口长度为 3 年，那么我们就用 2010 年 1 月至 2012 年 12 月约 150 个周频样本对贝叶斯模型进行训练，然后预测下一周沪铜主力合约的收益率，作为一个样本外的预测数据。接着把训练窗口往下移动一周，再对模型进行重新训练和预测。这样如果训练窗口为 3 年的话，就有约 300 个样本外数据，其中前 75%的数据作为验证集用于筛选模型超参数，后 25%的数据作为测试集，检验模型的预测效果。这里虽然只有 3 个超参数，但是可能的组合却上百，我们在这里预先设定超参数的可选范围，从而得到约 70 组超参数，对每个参数组合计算验证集中所有样本的后验概率

$$\sum_{t}^{T} \log P(\mathrm{r_t}|\mathbf{f_{t-1}}, \boldsymbol{\theta_{t-1}}) \tag{3}$$

其中 T 为验证集中的样本个数，$\boldsymbol{\theta_{t-1}}$为每次推断的模型系数分布。目前使用 32 线程，2.6Ghz 的双 CPU 进行超参数筛选大约需要 10 小时。

模型评价

在利用上述方法筛选出模型以后，我们也有必要了解贝叶斯模型对样本内，也就是训练集数据的解释程度。其中一个
练集数据的解释程度。其中一个方法是利用公式

(3)考察训练集的后验概率均值，这个在经典线性回归里就相当于拟合度 R。但在贝叶斯方法里，我们也需要考察模型对数据分布的解释程度，一种常用的方法是后验预测检验（Posterior Predictive Check, PPC）。当模型训练完成后，可以考察训练集因变量y，即期货主力合约收益率的某一统计量$\mathrm{T}(y)$，如均值，最大值和最小值等。然后对训练集内每一个样本计算其后验预测分布y_{rep}，得到训练集里每个样本的$\mathrm{T}(y_{rep})$，从而得到$\mathrm{T}(y_{rep})$的分布。对比$\mathrm{T}(y)$和$\mathrm{T}(y_{rep})$的分布就能获得该模型对样本内数据的解释程度。

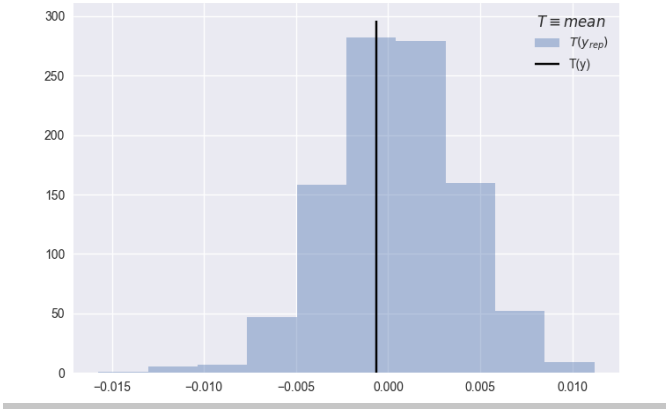
我们使用从 2014 年 1 月初至 2016 年 12 月底约 150 周的沪铜主力合约数据训练模型，首先考察后验预测均值的分布，如图 1：所示，这段时期，沪铜主力合约的收益率均值$\mathrm{T}(y)$在 0 附近，用黑线表示。后验均值分布$\mathrm{T}(y_{rep})$较为均匀地分布在$\mathrm{T}(y)$左右两侧，因此直观上看该贝叶斯模型能对均值这个统计量能做出较好的解释。后验预测检验通常会使用p值来衡量模型对某个统计量的解释程度，其定义如下

$$p = P\{\mathrm{T}(y_{rep}) \geq \mathrm{T}(y)\} \tag{4}$$

这实际上就是计算训练集里$\mathrm{T}(y_{rep}) \geq \mathrm{T}(y)$的数量后除以训练集的样本数量，一般$p$值越接近 0.5，表示对该统计量解释度越好。图 1：中的p值为 0.51，表明该贝叶斯模型在均值上解释较好。

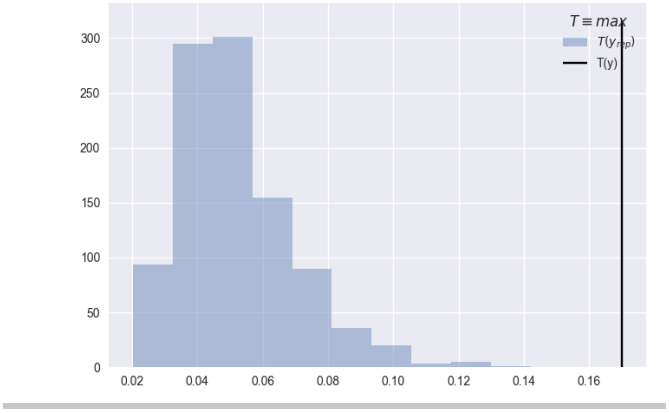
我们也可以利用后验预测检验来考察该贝叶斯模型对沪铜主力合约尾部风险的解释程度，统计量为最大值和最小值的情况分别在图 2：和图 3：中给出。由图可见，黑线距离尾部较大，也就是说在估计最大值和最小值这两个统计量上，该贝叶斯模型有较大的偏差。同时这两个统计量的p值分别为 0.0 和 0.99，所以该贝叶斯模型在极端风险的估计上会存在较大偏差。一种比较简单直接的解决方案是在公式(1)中引入一个服从泊松分布的随机项。

图 1：　后验预测检验—均值



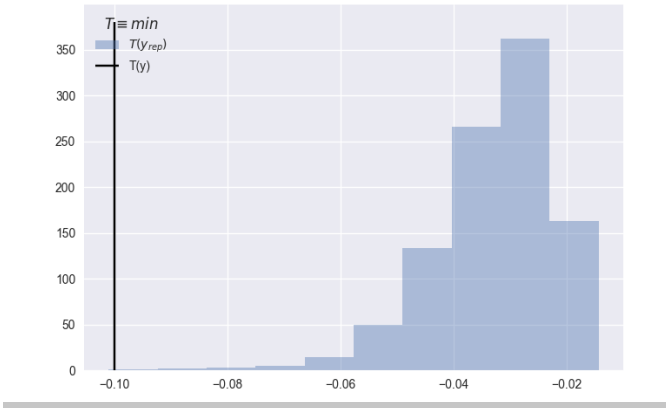
数据来源：Wind 华泰期货研究所

图 2：后验预测检验—最大值



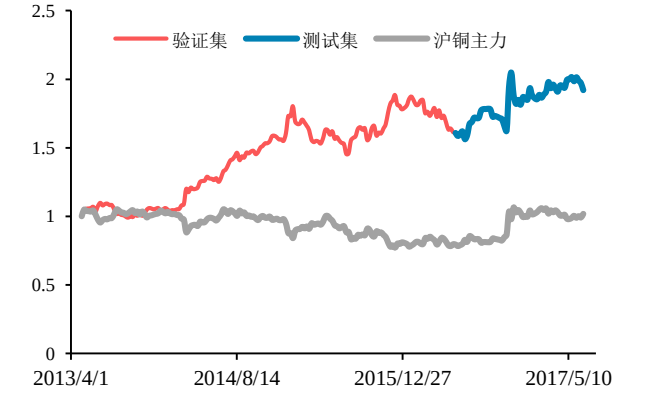
数据来源：Wind 华泰期货研究所

图 3：　后验预测检验—最小值



数据来源：Wind 华泰期货研究所

图 4：沪铜交易回测



数据来源：Wind 华泰期货研究所

模型的交易回测

我们根据筛选出的最优超参数组合来选择模型进行交易，模型先预测出沪铜主力合约下一周的收益率分布再根据此计算出沪铜主力合约上涨和下跌的概率，如果上涨概率大于0.5就做多，下跌概率大于0.5就做空。每周最后一个交易日进行预测和调仓，假设以收盘价交易。虽然在2013年4月至2017年6月期间沪铜主力合约的变化不大，但是无论验证集和测试集在这段时间里都能取得较好的年化收益，分别为14.8%和16.7%。夏普率分别为0.9和0.7。虽然利用这个模型交易沪铜的夏普率不高，但是把类似方法应用到其他商品期货上就有可能获得一个夏普率较高的商品组合策略。

结果讨论

本报告首先对贝叶斯模型进行了基本介绍，然后以沪铜期货为例，尝试基于基本面因子，利用贝叶斯线性回归模型对其主力合约的周度收益进行模型训练，与经典线性回归模型相比，贝叶斯模型需要更长的训练和筛选时间，而且对极端风险也没法准确解释，但是该模型在实际交易中能取得一定效果。

免责声明

公司总部

地址：广州市越秀区先烈中路65号东山广场东楼5、11、12层

电话：400-6280-888

网址：www.htgwf.com