

基于机器学习的 PPI 预测模型——量化知宏观系列之一

报告日期: 2022-03-08

分析师: 何宁

执业证书号: S0010521100001

电话: 15900918091

邮箱: hening@hazq.com

相关报告

- 1.《华安证券_宏观研究_宏观点评_CPI-PPI 触底了吗? 价格传导如何》2021-11-10
- 2.《华安证券_宏观研究_宏观专题_CPI-PPI 见底回升, 如何配置资产?——通胀系列之一》2021-12-09
- 3.《华安证券_宏观研究_宏观点评_五大主线打开 2022 经济密码——中央经济工作会议点评》2021-12-12
- 4.《华安证券_宏观研究_宏观点评_是终点也是起点——12 月 PMI 点评》2021-12-31
- 5.《华安证券_宏观研究_宏观点评_数说价格传导之谜——通胀系列之二》2021-01-06
- 6.《华安证券_宏观研究_宏观点评_平稳收官——2021 年通胀点评》2022-01-12
- 7.《华安证券_宏观研究_宏观点评_价格传导加强, 困境反转来了?》2022-02-16

主要观点:

价格和通胀是宏观分析研究的重要组成部分, 本文通过梳理常见的 PPI 预测方法, 分析其缺陷和难点所在, 尝试基于机器学习的方法解构 PPI 并建模预测。

●一、传统 PPI 预测有哪些方法, 哪些痛点?

PPI 预测主要有三类: 领先指标预测法、高频指标降频法和工业品价格合成法。

1) 领先指标预测法, 通过挖掘符合经济运行规律的具有领先性的指标, 确定最优领先期后进行回归预测, 例如 CRB 工业原料指数领先 PPI 2-3 个月, 由于相关性和领先期不稳定, 往往精度较低。

2) 对高频指标做降频处理, 简洁有效, 生产资料价格指数与当期 PPI 相关性高达 0.98。由于自变量和因变量的具体构成关系不明确, 导致无法调参和优化模型。

3) 利用工业品价格合成 PPI, 逻辑关系明晰, 可信度和可调试性较强, 理论上拟合效果在三种预测类型中最优。此前的相应研究存在两大难点: 1) 生产资料价格种类繁多, 传统线性 OLS 模型共线性问题严重, 模型方差较大、容易过拟合, 需要合适的变量筛选以确定 PPI 核心驱动因素; 2) 统计局对 PPI 分项权重披露信息较少, 如何理解统计过程中的权重以及调权方式尤为重要。

●二、如何理解 PPI 权重和权重调整?

PPI 行业分项权重为分行业工业销售产值的比重, 一般用工业行业的营收占比作为替代。统计局在计算当期 PPI 同比是基于去年同期的行业权重。

1) 根据统计局解释, “CPI 中每一种商品和服务的权重是根据这一类商品和服务的消费支出占整个篮子里的总支出比重作为权重的”。2019 年非洲猪瘟快速扩散, 全国能繁母猪与生猪存栏双降, 猪肉价格于 19 年 11 月和 20 年 2 月达到两个高位, 而猪肉同比权重在 20 年 11 月攀至顶点的 4.6%, 从历史曲线来看恰好滞后猪肉价格约 12 个月。PPI 和 CPI 指数的编制与统计保持一致, 均采用链式拉式公式, 合理推断统计局在计算 PPI 同比时采用去年同期 (T-12 期) 分行业工业销售产值的比重, 在计算 PPI 环比时采用上期 (T-1 期) 权重。

2) 根据 T-12 期行业营收占比加权计算的 PPI 最接近实际值。我们将 39 个行业分项 PPI 同比按各期营收占比进行加权, 其中 T-12、T-11 和 T-10 期与实际 PPI 同比的相关性均高于 99.85%, 但 T-12 期平均偏差最小、效果最好。

●三、基于机器学习的 PPI 预测模型

1) 数据预处理: 2016 月 1 月至 2021 年 12 月为数据始末时间, 删除时间序列过短的普通硅酸盐水泥、天然橡胶。取生产资料价格月均值、计算同比得到 48 个特征变量。

2) 基于 LASSO 筛选变量解决共线性问题。LASSO 在 OLS 回归的系数上增加一个惩罚项, 求得模型的稀疏化解 (部分系数为零)。分割数据集后, LASSO 回归从 48 个自变量中提取出 9 个关键变量, 剔除不显著的因子, 保留了无缝钢管、柴油、汽油、石蜡、无烟煤 5 个变量, 模型解释力度 R^2 为 97.1%。

3) 针对权重调整问题, 我们将变量滞后 12 期以拟合当期生产资料销售产值波动对下年同比权重的影响。优化后, 模型解释力度上升至 99.0%。这种替代方法的潜在假设是工业品价格的波动能代表工业品销售产值的变动, 但供给侧改革期间, 上游原材料价升量跌, 因而 2017 年初 PPI 预测和 PPI 实际值存在一定偏离。

●四、模型预测 2 月 PPI 同比为 9.1%

1) 3 月 4 日统计局发布了 2 月下旬的流通领域生产资料价格, 2 月煤炭、石油、化肥涨幅居前, 系俄乌危机下大宗商品迎来新一轮上行, 全球通胀压力加剧。

2) 2 月无缝钢管、柴油、汽油、石蜡、无烟煤、柴油滞后项、无烟煤滞后项分别为 11.0%、51.3%、41.1%、43.5%、93.6%、-10.3%、-9.5%, 基于 Lasso 回归的机器学习模型预测 2 月 PPI 为 9.1%。

●风险提示

俄乌冲突带来干扰, 模型可能存在失效风险。

正文目录

1 传统 PPI 预测方法及其痛点	4
2 如何理解 PPI 权重及其调整	5
3 基于机器学习的 PPI 预测方法	7
3.1 数据预处理	8
3.2 LASSO 特征筛选处理共线性问题	9
3.3 引入滞后项拟合商品权重调整	10
4 预测 2022 年 2 月 PPI 同比 9.1%	10

图表目录

图表 1 生产资料价格指数相关性最高，适合预测即期 PPI 同比	4
图表 2 PPI 预测方法梳理	4
图表 3 2021 年工业行业权重	5
图表 4 猪肉价格领先猪肉权重 12 个月	6
图表 5 修正后，猪肉权重计算逻辑更明晰	6
图表 6 以 T-12 期营收占比为权重，平均偏差 0.18%	6
图表 7 以 T-1 期营收占比为权重，平均偏差 0.29%	6
图表 8 按各期营收占比为权重，测算 PPI 同比	7
图表 9 流通领域重要生产资料价格和 PPI 对比	7
图表 10 流通领域生产资料价格相关性矩阵	8
图表 11 LASSO 回归筛选关键特征	9
图表 12 基于 LASSO 筛选特征、构建模型	10
图表 13 增加滞后项，模型解释力度上升至 99.0%	10
图表 14 2022 年 2 月流通领域生产资料价格涨跌幅	11
图表 15 基于机器学习的 PPI 预测	12

价格和通胀是宏观分析研究的重要组成部分，本文通过梳理常见的 PPI 预测方法，分析其痛点与难点所在，尝试基于机器学习的方法解构 PPI 并建模预测。

1 传统 PPI 预测方法及其痛点

PPI 预测常见的类型有三种：**领先指标预测法**、**高频指标降频法**和**工业品价格合成法**。

领先指标预测法，通过挖掘符合经济运行规律的具有领先性的指标，确定最优领先期后进行回归预测，例如 CRB 工业原料指数领先 PPI 2-3 个月，由于相关性和领先期不稳定，往往精度较低。

对高频指标做降频处理往往是一种简洁有效的办法，例如生产资料价格指数在统计品类和编制方法上和 PPI 较为相近，即期相关性高达 0.98。由于自变量和因变量的具体构成关系不明确，导致无法调参和优化模型。

利用工业品价格合成 PPI，逻辑关系明晰，可信度和可调试性较强，理论上拟合效果在三种预测类型中最优。此前的相关研究存在两大难点：1) 生产资料价格间种类繁多，传统线性模型直接拟合共线性问题严重，模型方差较大、容易过拟合，需要合适的变量筛选以确定 PPI 核心驱动因素；2) 统计局对 PPI 分项权重披露信息较少，如何理解统计过程中的权重以及调权方式尤为重要。因此，本文尝试基于机器学习的方法，利用统计局发布的 50 种流通领域生产资料价格构建 PPI 预测模型。

图表 1 生产资料价格指数相关性最高，适合预测即期 PPI 同比

相关性	南华工业品指数：同比	南华综合指数：同比	CRB 现货综合：同比	CRB 现货工业原料：同比	大宗商品价格指数：同比	生产资料价格指数：同比
即期	0.6309	0.6602	0.7737	0.6936	0.8677	0.9834
领先 1 期	0.7236	0.7575	0.8360	0.7723	0.9175	0.9682
领先 2 期	0.7597	0.8029	0.8622	0.8148	0.9166	0.9160
领先 3 期	0.7552	0.8100	0.8612	0.8307	0.8792	0.8639

资料来源：WIND，华安证券研究所测算

图表 2 PPI 预测方法梳理

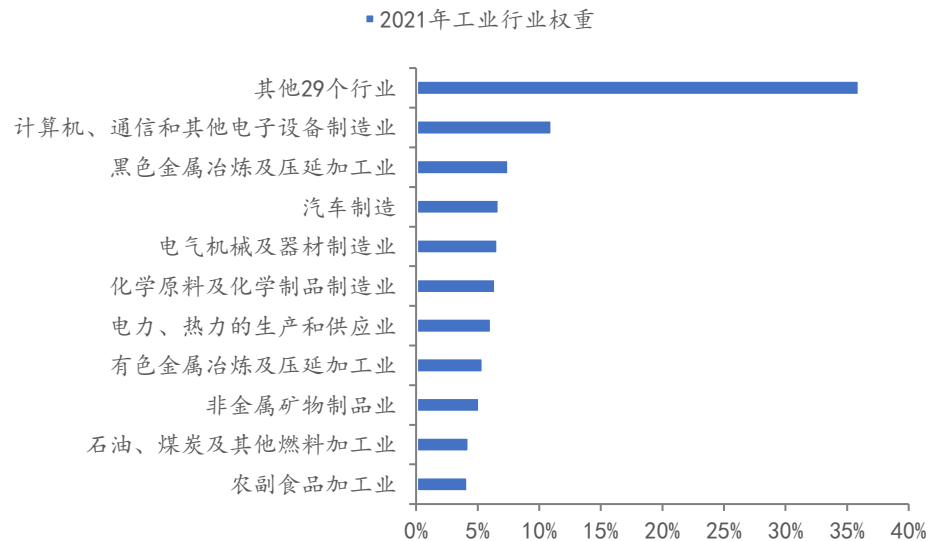
	领先指标预测法	高频指标降频法	工业品价格合成法
底层数据	南华工业品指数、CRB 现货价格指数、大宗商品价格指数等	生产资料价格指数	1. 流通领域重要生产资料价格 2. 关键行业商品价格
计量方法	确定指标的最优领先期，滞后相应期数后进行线性回归	线性回归	明晰工业品价格和 PPI 构成的关系，确定权重和调权方式，构造模型
预测期限	视领先期而定，可即期、可远期	即期	即期
精度	☆☆	☆☆☆☆	☆☆☆☆☆
优点	方法简便、具有一定领先期	方法简便、精度高	精度高、可信度高
缺点	领先期不稳定、相关性不稳定、精度低	构成关系不明确、难以调参	价格间共线性问题严重、需要合适的方法确定权重

资料来源：国家统计局，华安证券研究所整理

2 如何理解 PPI 权重及其调整

PPI 根据二分法分为生产资料和生活资料，按行业划分有 39 个工业行业，根据统计局解释，PPI 分项的权重为分行业工业销售产值的比重，一般用工业行业的营收占比作为替代。2021 年，计算机通信和其他电子设备制造业、黑色金属冶炼及压延加工业、汽车制造业、电气机械及器材制造业、化学原料及化学制品制造业分列前五，占比为 11.1%、7.6%、6.8%、6.7%和 6.5%。

图表 3 2021 年工业行业权重



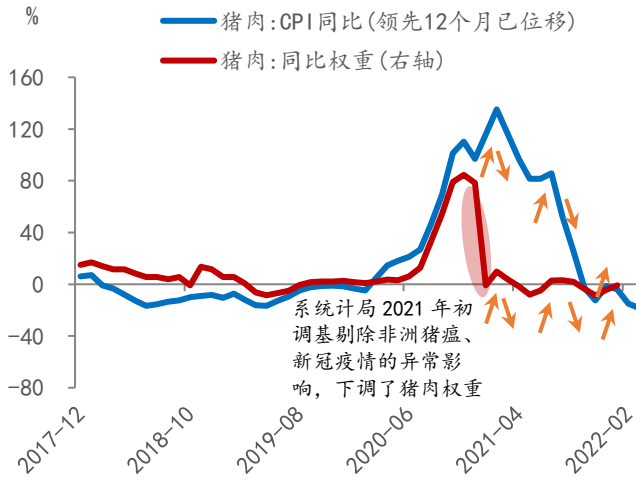
资料来源：WIND，华安证券研究所

一个容易被忽略的事实在于，统计局在计算当期 PPI 同比是基于去年同期的行业权重。若使用当期或者近期的权重进行预测，则会导致较大的偏差。

(1) 统计局解释支持该判断。根据统计局毛盛勇司长 2019 年前三季度国民经济运行情况发言，“CPI 中每一种商品和服务的权重是根据这一类商品和服务的消费支出占整个篮子里的总支出比重作为权重的。比如说，9 月的猪肉环比权重就是 8 月猪肉的消费支出在整个消费支出中的比重作为它的权重；9 月猪肉同比涨幅，用上年 9 月猪肉消费支出占整个消费支出的比重作为权重。这样每个月基期是固定的、篮子是固定的，但是权重是有变化的。”

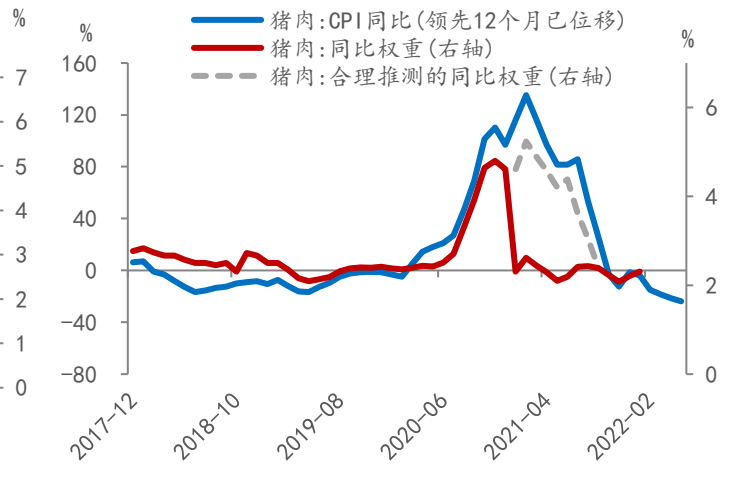
不妨观察猪肉价格和猪肉权重的关系，2019 年非洲猪瘟快速扩散，全国能繁母猪、生猪存栏双降，猪肉价格于 19 年 11 月和 20 年 2 月达到两个高位，而猪肉同比权重在 20 年 11 月攀至顶点的 4.6%，从历史曲线来看恰好滞后猪肉价格约 12 个月。21 年 2 月猪肉权重理应达到高位，但实际权重却较低，系统计局 2021 年初调基剔除非洲猪瘟、新冠疫情的异常影响，下调猪肉权重所致（除统计局已发布解释外，最有力的证明在于 2021 年 1 月猪肉权重非常完美的等于 2020 年 12 月的一半）。PPI 和 CPI 指数的编制与统计保持一致，均采用链式拉式公式，合理推断统计局在计算 PPI 同比时采用去年同期（T-12 期）分行业工业销售产值的比重，在计算 PPI 环比时采用上期（T-1 期）权重。

图表 4 猪肉价格领先猪肉权重 12 个月



资料来源: WIND, 华安证券研究所

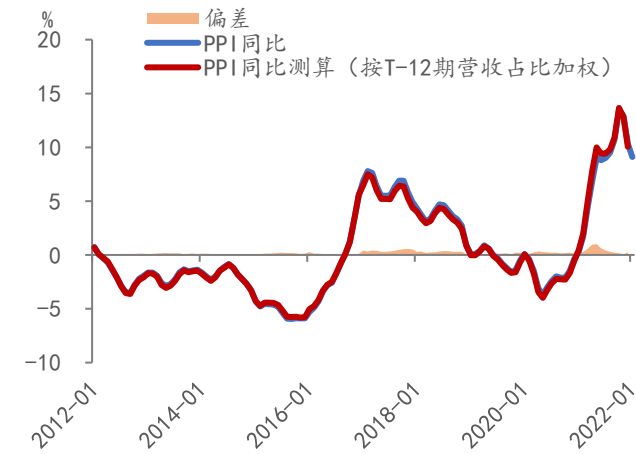
图表 5 修正后, 猪肉权重计算逻辑更明晰



资料来源: WIND, 华安证券研究所测算

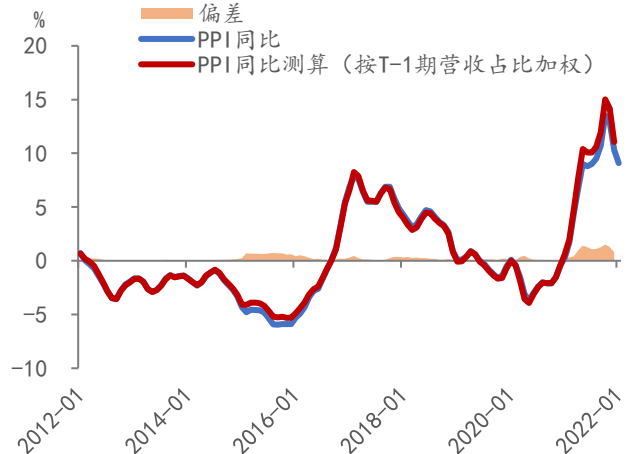
(2) 根据 T-12 期行业营收占比计算的 PPI 最接近实际值。统计局每月公布 PPI 同比和 39 个工业行业的分项 PPI, 我们将行业分项 PPI 按各期营收占比加权得到 PPI 拟合值, 拟合效果和 Tn 线性关系显著, 其中 T-12、T-11 和 T-10 期与实际 PPI 同比的相关性均高于 99.85%, 但 T-12 期平均偏差¹最小、效果最好。

图表 6 以 T-12 期营收占比为权重, 平均偏差 0.18%



资料来源: WIND, 华安证券研究所测算

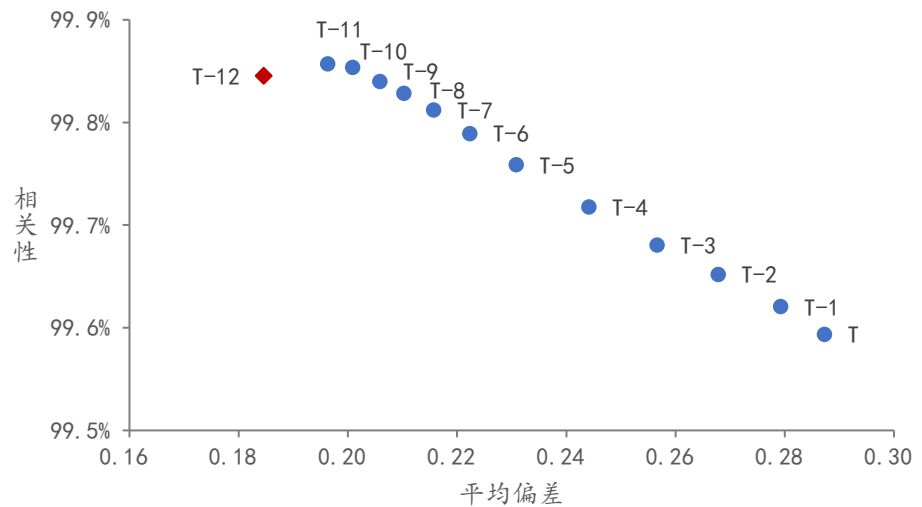
图表 7 以 T-1 期营收占比为权重, 平均偏差 0.29%



资料来源: WIND, 华安证券研究所测算

¹ 平均偏差计算方式为, 取每个月 PPI 拟合值与 PPI 实际值的绝对偏差, 求其均值。

图表 8 按各期营收占比为权重，测算 PPI 同比



资料来源：WIND，华安证券研究所测算

3 基于机器学习的 PPI 预测方法

如前述，我们理解了 PPI 指数在编制过程中的权重调整方式，但 PPI 行业分项不能用于预测，因而我们使用流通领域重要生产资料价格（下称“重要生产资料价格”）来替代，使用该数据的合理性在于：（1）重要生产资料价格高频，分别于每月 4/14/24 日公布上一旬数据，在时间上领先于统计局发布的 PPI；（2）商务部发布的生产资料价格指数与 PPI 相关性高，则统计局发布的流通领域重要生产资料价格理论上拟合效果更好。

图表 9 流通领域重要生产资料价格和 PPI 对比

	流通领域重要生产资料价格	PPI
指标含义	重要生产资料经营企业的批发和销售价格，与出厂价格不同，生产资料市场价格既包含出厂价格，也包含有经营企业的流通费用、利润和税费等。	工业生产者出厂价格指数反映工业企业产品第一次出售时的出厂价格的变化趋势和变动幅度。
统计范围	9 大类、50 种工业产品价格。	40 个工业行业大类，201 个工业行业中类，581 个工业行业小类的工业产品，1300 多个基本分类的工业产品价格。
计算方法	同一种规格品的各旬价格根据旬初第一个工作日和旬末最后一个工作日的均价平均计算。	企业上报的报表包括报告期单价和上月单价，产品报告期单价为报告月 5 日、20 日两次所采单价的简单平均值。
发布时间	每月 4 日、14 日、24 日发布上一旬数据，节假日顺延。	每月 9 日或 13 日。

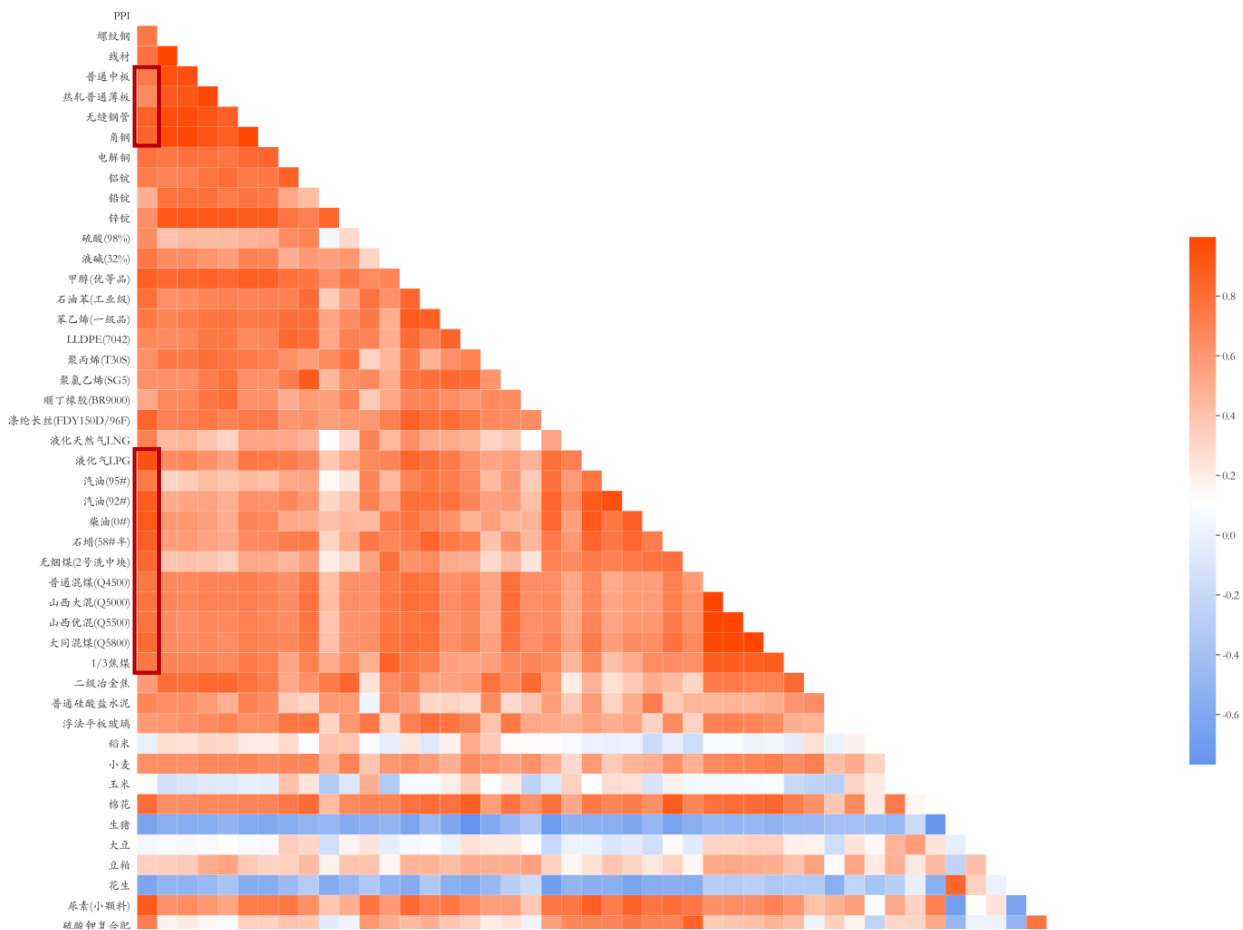
资料来源：WIND，华安证券研究所

3.1 数据预处理

考虑到基期更换的影响，我们选择 2016 月 1 月至 2021 年 12 月作为数据始末时间。在数据预处理上，剔除时间序列过短的普通硅酸盐水泥、天然橡胶。由于重要生产资料价格旬度更新，因此我们取当月三个旬度的重要生产资料价格均值、计算当月同比作为特征变量（共 48 个）。

从相关性矩阵来看，钢、油、煤和 PPI 高度相关，农产品价格相关性较弱、部分呈负相关。

图表 10 流通领域生产资料价格相关性矩阵



预览已结束，完整报告链接和二维码如下：

https://www.yunbaogao.cn/report/index/report?reportId=1_39291

