In [1]:
```
import  numpy as np
import matplotlib.pyplot as plt
import pandas as pd

plt.rcParams['font.sans-serif'] = ['Songti SC']
plt.rcParams['axes.unicode_minus'] = False


df = pd.read_excel('平安银行 (1).xlsx')

print("原始数据预览：")
print(df.head())
```

```
原始数据预览：
          日期     收盘价      涨跌幅      主力净流入  Unnamed: 4      超大单净流入  Unn
amed: 6  \
0        NaT     NaN      NaN         净额         净占比          净额         净占比
1 2024-10-08   12.88   0.0549     -3.59亿     -0.0473   -9006.75万     -0.0118
2 2024-09-30   12.21   0.0692   2198.93万      0.0034      1.02亿      0.0158
3 2024-09-27   11.42   0.0242   -461.72万     -0.0011   2276.75万      0.0052
4 2024-09-26   11.15   0.0619      2.88亿      0.0863      2.80亿      0.0837

   Unnamed: 7
0         成交额
1  7602663936
2  6467438592
3  4337985024
4  3341759488
```

In [2]:
```
print("缺失值情况：")
print(df.isnull().sum())

df.columns = df.columns.str.strip()
date_col = [c for c in df.columns if '日期' in c or 'date' in c.lower()]
date_col = date_col[0]

df[date_col] = pd.to_datetime(df[date_col])
df = df.sort_values(date_col)
```

```
缺失值情况：
日期             1
收盘价            1
涨跌幅            1
主力净流入           3
Unnamed: 4     0
超大单净流入          4
Unnamed: 6     3
Unnamed: 7     0
dtype: int64
```

In [3]:
```
for col in df.columns:
    if col == date_col:
        continue

    if df[col].dtype in ['float64', 'int64']:
        # 资金相关字段
        if '净流入' in col:
            df[col] = df[col].fillna(0)  # 缺失视为无流动
        # 占比类字段
```

```
        elif '占比' in col or '%' in col:
            df[col] = df[col].interpolate(method='linear')  # 线性插值
        # 其他连续数值字段
        else:
            df[col] = df[col].interpolate(method='linear')
    else:
        # 非数值字段：前向填充
        df[col] = df[col].ffill()

print("局部缺失补齐后缺失统计：")
print(df.isnull().sum())
```

```
局部缺失补齐后缺失统计：
日期             1
收盘价            0
涨跌幅            0
主力净流入          0
Unnamed: 4     0
超大单净流入         0
Unnamed: 6     0
Unnamed: 7     0
dtype: int64
```

In [4]:
```
# 假设数据有 '日期' 列
df['日期'] = pd.to_datetime(df['日期'])
# 按日期排序
df = df.sort_values('日期')
# 去除重复日期（取均值）
df = df.groupby(date_col, as_index=False).mean(numeric_only=True)

# 检查交易日缺失
all_days = pd.date_range(start=df['日期'].min(), end=df['日期'].max(), freq='B')
missing_days = all_days.difference(df['日期'])
print(f"\n缺失交易日数量: {len(missing_days)}")
print("缺失日期：", missing_days[:10])
```

```
缺失交易日数量: 16
缺失日期： DatetimeIndex(['2024-05-01', '2024-05-02', '2024-05-03', '2024-06-10',
               '2024-06-20', '2024-07-05', '2024-08-05', '2024-09-03',
               '2024-09-13', '2024-09-16'],
              dtype='datetime64[ns]', freq=None)
```

In [5]:
```
# 交易日补全
df = df.set_index(date_col).reindex(all_days)
df.index.name = date_col

# 数值型字段用前后值插补（保持趋势平滑）
for col in df.select_dtypes(include=[np.number]).columns:
    df[col] = df[col].interpolate(method='linear')

# 分类型字段（如股票代码）用前向填充
for col in df.select_dtypes(exclude=[np.number]).columns:
    df[col] = df[col].ffill()

print("最终数据条数：", len(df))
```

```
最终数据条数： 132
```