Uncovering sparsity and heterogeneity in firm-level return predictability using machine learning

Theodoros Evgeniou
Ahmed Guerciouer
Rodolfo Prieto
INSEAD

Contributions
For the problem of firm-level month-ahead return prediction, and interpreting characteristic importance,
• We find statistical evidence (using the bootstrap) that heterogeneity matters for predictability.
• By incorporating heterogeneity in predictive models, we improve their out-of-sample performance.

Specifications of predictive heterogeneity
To define groupings of firms, we consider two alternatives:
1. Firm industry memberships – based on SIC codes.
2. Inferring (possibly) latent group memberships from observable characteristics – by applying k-means clustering to characteristic means.

Incorporating heterogeneity in linear predictive models
• Index firms by i, and let ε_t be a high-dimensional (M×1)-vector of a firm’s characteristics.
• We apply ML regularization techniques to classical pooled linear models with common coefficients:
  \[ y_{i,t} = \gamma_0 + \gamma_1 x_{i,t} + \gamma_2 z_{i,t} + \cdots + \gamma_M y_{i,M} + \varepsilon_t \]
  \[ = y'_i' \beta + \varepsilon_t \] (1)
• Furthermore, we incorporate heterogeneity in predictive relationships in by-group models, given a mapping from a firm i to its (unique) group j, by employing group-specific coefficients:
  \[ y_{i,t} = \gamma_{0,j} + \gamma_{1,j} x_{i,t} + \gamma_{2,j} z_{i,t} + \cdots + \gamma_{M,j} y_{i,M} \]
  \[ = y'_i' \beta_j + \varepsilon_t \] (2)
• We also combine the two stages to specify composite two-stage models, that take the form
  \[ y_{i,t} = \gamma_{0,j} + \gamma_{1,j} x_{i,t} + \cdots + \gamma_{M,j} y_{i,M} + \varepsilon_t \] (3)

1. estimate a pooled model on the entire cross-section of returns, then
2. estimate by-group model on the residuals of the first-stage pooled model.
3. Need to tune multiple regularization parameters (e.g. lasso λ) for by-group and two-stage models.

Motivations for predictive heterogeneity
• Equilibrium asset pricing models with multiple state variables, such as Marris, Santos, and Varousek (2004) and Koijen and Yogo (2010), imply heterogeneity in firm-specific predictive relationships.
• Patton and Wheeler (2019) find evidence for risk premia deviations that are specific to groupings of firms (rather than the whole cross-section) in a modified conditional CAPM.

Data & evaluation
• 109 predictive characteristics: 101 are firm-specific (Green, Hand, and Zhang 2017) and 8 are market-level (Welch and Goyal 2007).
• Time period: 1990-2015 (inclusive).
• Our out-of-sample evaluation uses the same R² metric and takes the same expanding window approach as Gu, Kelly, and Xi (2020).

Overall predictability, measured by out-of-sample R² (%)
• Clustering firms into groups in predictive relationships, then incorporating this specification of heterogeneity when estimating a regularized linear model achieves an out-of-sample R² = 1.05%.
• For context, Gu, Kelly, and Xi (2020) achieved an out-of-sample R² = 0.40% on their sample using a deep neural network.

Table 1: Using industry memberships to define heterogeneity in predictive relationships.

<table>
<thead>
<tr>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

Uncovering sparsity & heterogeneity in characteristic importance
• In the overall predictability results (above), heterogeneous lasso-based linear models performed well.
• Selection by the lasso is a measure of variable importance. These lasso-selected predictive variables vary between clusters of firms, and are a sparse subset of the 109 total variables employed.
• In contrast to Gu, Kelly, and Xi (2020), the important predictive variables are mostly a subset of low-frequency cash and profitability-related coefficients (cfhp, cahs, and the market-level D/P ratio (dp/dp500), rather than higher-frequency price-based predictors.

Table 5: Frequency of selection (%) across slices of our database, according to the by-cluster lasso model.

References


