A Few Words on Variable Selection

Prof. Sharon McNicholas

STATS 780/CSE 780

Introduction

- We have seen some approaches to dimension reduction, e.g., PCA, factor analysis, the PGMM family, etc.
- McNicholas (2016) refers to dimension reduction as implicit or explicit.
- We have also seen another, and explicit, form of dimension reduction known as variable selection.
- We carried out variable selection when we built (logistic) regression models, e.g., using the `step()` function, the AICs, or the adjusted R^2 values.

Variable Selection

- Consider clustering and classification problems.

- Dimension reduction is important because the presence of variables that are not helpful in discriminating groups can have a deleterious effect on clustering, or classification, performance.

- We have seen the PGMM family, which can be considered an implicit form of dimension reduction.

- There are also variable selection approaches available, which give an explicit dimension reduction.

VSCC

- The VSCC technique (Andrews and McNicholas, 2014) finds a subset of variables that simultaneously minimizes the within-group variance and maximizes the between-group variance.

- Note that the within-group variance for variable j can be written

$$\mathcal{W}_j = \frac{\sum_{g=1}^{G} \sum_{i=1}^{n} z_{ig}(x_{ij} - \mu_{gj})^2}{n}.$$

- The variance within variable j that is not accounted for by $\mathcal{W}_j$, i.e., $\sigma_j^2 - \mathcal{W}_j$, provides an indication of the variance between groups.

- If the data are standardized to have equal variance across variables, then any variable minimizing the within-group variance also maximizes the leftover variance.

VSCC contd.

- If V represents the space of currently selected variables, then variable j is selected if
$$|\rho_{jr}| < 1 - \mathcal{W}_j^m$$
for all $r \in V$, where $m \in \{1, \ldots, 5\}$ is fixed.

- Further details on the VSCC approach are given by Andrews and McNicholas (2014) and McNicholas (2016,Chapter 4).

- The VSCC approach can be used for clustering or semi-supervised classification, and is supported by the `vscc` package (Andrews and McNicholas, 2013) for R.

clustvarsel and selvarclust

- Raftery and Dean (2006) propose a variable selection method that utilizes a greedy search of the model space; their approach is based on Bayes factors.

- Given data $\mathbf{x}$, the Bayes factor B_{12} for model $\mathcal{M}_1$ versus model $\mathcal{M}_2$ is
$$B_{12} = \frac{p(\mathbf{x} \mid \mathcal{M}_1)}{p(\mathbf{x} \mid \mathcal{M}_2)}.$$
where
$$p(\mathbf{x} \mid \mathcal{M}_k) = \int p(\mathbf{x} \mid \boldsymbol{\theta}_k, \mathcal{M}_k) p(\boldsymbol{\theta}_k \mid \mathcal{M}_k) d\boldsymbol{\theta}_k,$$
$\boldsymbol{\theta}_k$ is the vector of parameters for model $\mathcal{M}_k$, and $p(\boldsymbol{\theta}_k \mid \mathcal{M}_k)$ is the prior distribution of $\mathcal{M}_k$ (Kass and Raftery, 1995).

clustvarsel and selvarclust cont.

- Like VSCC, the approach of Raftery and Dean (2006) simultaneously selects a variable subset, the number of components, and the model, i.e., the GPCM covariance structure.

- The approach of Raftery and Dean (2006) is implemented within the `clustvarsel` package Dean et al. (2015) for R.

- A related approach, `selvarclust`, is described by Maugis et al. (2009a,b).

Comments

- Because the number of free model parameters for some of the GPCM models is quadratic in data dimensionality, `clustvarsel`, `selvarclust`, and `vscc` are largely ineffective in high dimensions.

- In addition to approaches like PGMM, there are other very effective approaches for implicit dimension reduction, e.g., GMMDR (Scrucca, 2010) and HD-GMM (Bouveyron et al., 2007).

- Two good starting places for further reading are Bouveyron and Brunet-Saumard (2014) and McNicholas (2016, Ch. 4).

References

- Andrews, J. L. and P. D. McNicholas (2013). vscc: Variable Selection for Clustering and Classification. R package version 0.2.
- Andrews, J. L. and P. D. McNicholas (2014). 'Variable selection for clustering and classification'. *Journal of Classification* **31**(2), 136–153.
- Bouveyron, C. and C. Brunet-Saumard (2014). 'Model-based clustering of high-dimensional data: A review'. *Computational Statistics and Data Analysis* **71**, 52–78.
- Bouveyron, C., S. Girard, and C. Schmid (2007). 'High-dimensional data clustering'. *Computational Statistics and Data Analysis* **52**(1), 502–519.
- Dean, N., Raftery, A. E., and Scrucca, L. (2015). clustvarsel: Variable Selection for Model-Based Clustering. R package version 2.2.
- Kass, R. E. and A. E. Raftery (1995). 'Bayes factors'. *Journal of the American Statistical Association* **90**(430), 773–795.
- McNicholas (2016). *Mixture Model-Based Classification*. Boca Raton: Chapman & Hall/CRC Press.
- Raftery, A. E. and N. Dean (2006). 'Variable selection for model-based clustering'. *Journal of the American Statistical Association* **101**(473), 168–178.
- Scrucca, L. (2010). 'Dimension reduction for model-based clustering'. *Statistics and Computing* **20**(4), 471–484.