How to learn from a lot: Empirical Bayes in high-dimensional prediction settings

Mark van de Wiel

Empirical Bayes (EB) is widely acknowledged as a useful technique to borrow information across variables of the same type. In the broadest sense it is a collection of techniques which estimate the prior of a parameter from the data, where the prior may also be implicitly defined by a regularization parameter in a frequentist setting. We focus on application of EB to high-dimensional prediction and classification settings, with an emphasis on genomics applications. EB is very useful in high-dimensional settings, because it a) avoids cross-validation of multiple tuning parameters; b) allows use of external data to improve predictive performance and variable selection; and c) does generally not 'over-shrink' due to the multitude of variables. However, depending on the prediction framework (e.g. penalized regression, discriminant analysis, random forest, etc.), development of EB-based predictors or classifiers may be cumbersome. Therefore, we provide an overview of methods that can be used to apply EB in high-dimensional prediction settings. These methods are based on: i) Laplace and ii) Gibbs-sampling-based approximation of the marginal likelihood; on iii) Moments of the parameters; iv) Univariate effect-size estimates; and v) Bagging multiple predictors. Rather than considering details, we focus on the basic philosophies behind each of the methods. We discuss and illustrate several prediction frameworks to which EB applies. Here, we pay special attention to application of EB to frameworks that allow multiple tuning parameters to either a) bridge sparse and dense situations, such as the elastic net or b) enable incorporating information on a priori defined groups of predictor variables. We illustrate the methods on clinical prediction problems based on modern cancer genomics data sets, including (mi)RNAseq and methylation ones. For that purpose, we show that EB estimation of different penalty parameters for groups of variables may enhance p