13

Bayesian Models for Integrative Genomics

FRANCESCO C. STINGO AND MARINA VANNUCCI

13.1 Introduction

The practical utility of variable selection is well recognized, and this topic has been the focus of much research. Bayesian methods for variable selection have several appealing features. They address the selection and prediction problems in a unified manner, allow rich modeling via the implementation of Markov Chain Monte Carlo (MCMC) stochastic search strategies and incorporate optimal model averaging prediction strategies; they extend quite naturally to multivariate responses and many linear and nonlinear settings; they can handle the "small n–large p" setting (i.e., situations in which the number of measured covariates is much larger than the sample size); and they allow past and collateral information to be easily accommodated into the model through the priors.

The flexibility of the variable selection approach, in particular the fact that it can handle the"large p–small n" paradigm, has made Bayesian methods particularly relevant for the analysis of genomic studies, in which high-throughput technologies allow thousands of variables to be measured on individual samples. In this chapter we discuss recent contributions from our group on methods for integrative genomics. First, we focus on methods that integrate external biological information into the analysis of gene expression data. We consider a linear model that predicts a phenotype based on predictors synthesizing the activity of genes belonging to same pathways and encode into the prior model information on gene–gene networks, as retrieved from available databases. Variable selection is achieved via latent indicators that induce mixture priors with a spike at zero on the regression coefficients of the model. We then extend the variable selection techniques and the use of network priors to mixture models for gene expression data, where we treat both unsupervised (i.e., clustering) and supervised settings for pattern recognition. Unlike the case of the linear