Bayesian Statistics
Reading List

Compiled by Edoardo Costantini

2020-02-03

Books

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

> Use this resource as main reference every time you are in doubt.

Hoff, P. D. (2009). *A first course in bayesian statistical methods* (Vol. 580). Springer.

> A good starting point for your journey into Bayesian Statistics. The book strikes a good balance between rigorous presentation of the material and easy-to-grasp examples. Furthermore, most of the chapters are accompanied by R code that reduces to a minimum the abstraction level of the concepts presented.

General knowledge

Gelman, A., & Raghunathan, T. E. (2001). [conditionally specified distributions: An introduction]: Comment. *Statistical Science*, *16*(3), 268–269. Retrieved from http://www.jstor.org/stable/2676690

> The authors discuss the use of conditional distributions not to approximate joint models but for the purpose of multiple imputation. This is simply a short compendium that facilitate the understanding of SRMI/FCS Multiple Imputation for someone coming from a more traditional Bayesian background.

Gelman, A., & Speed, T. (1993). Characterizing a joint probability distribution by conditionals. *Journal of the Royal Statistical Society: Series B (Methodological)*, *55*(1), 185–188.

> A clarification on the conditions under which a set of conditional (and marginal) distributions uniquely specifies a joint distribution.

Regularized Regression

Griffin, J., & Brown, P. (2005). Alternative prior distributions for variable selection with very many more variables than observations. *University of Kent Technical Report*.

> The authors present scale mixture of normal prior distributions that can be used to perform variable selection in Bayesian regression analysis, laying the groundwork for the development of actual bayesian counterparts to the frequentist Lasso penalty for regression coefficient estimates.
>
> This paper describes clearly how the mixture of normal distribution priors proposed by Griffin and Brown (2011), to achieve some form of lasso regularisation in a bayesian framework, can be considered as generalisation of the one proposed by Park and Casella (2008).

Griffin, J. E., & Brown, P. J. (2011). Bayesian hyper-lassos with non-convex penalization. *Australian & New Zealand Journal of Statistics*, *53*(4), 423–442.

> With this study, the authors propose priors for Bayesian Lasso that are generalisations of Park and Casella's (2008) one

Kyung, M., Gill, J., Ghosh, M., Casella, G., et al. (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, *5*(2), 369–411.

> A comprehensive description of the Bayesian lasso and its frequentists counterparts.

Li, L., & Yao, W. (2018). Fully bayesian logistic regression with hyper-lasso priors for high-dimensional feature selection. *Journal of Statistical Computation and Simulation*, *88*(14), 2827–2851.

> The author implements a multinomial bayesian regression with lasso priors by proposing a new t-prior and adapting the priors proposed by Griffin and Brown (2011) for variable selection. The paper was developed along with an R-package (HTLR) that can be used to perform analyses and predictions.

Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686.

> The authors propose a conditional Laplace prior for the regression coefficients of a bayesian regression model that produce posterior mode estimates that have the same interpretation as Lasso estimates.

Tree modelling

Chipman, H. A., George, E. I., & McCulloch, R. E. (1998). Bayesian cart model search. *Journal of the American Statistical Association*, *93*(443), 935–948.

> The authors propose a Bayesian approach for finding of classification and regression trees that outperforms, mainly in terms of predictive performance, trees grown according to more traditional routines. This approach has two components:

- a prior specification for the set of possible CART models, which entails the specification of a prior for the tree space and one for the terminal nodes of each tree.
  - the specification of a prior probability that terminal node η is split (p_{SPLIT}), and the probability of assigning splitting rule ρ to η if it is split (p_{RULE}). The prior assigned to p_{SPLIT}, depends on the number of splits above η in a way that favours trees that have terminal nodes that do not vary much in depth (i.e. number of splits above). The prior for p_{RULE} is a uniform specification: all the variables have uniform probability of being chosen as splitting variables, and all the observed values on these variables have the same probability of being chosen as splitting values.
  - priors for terminal nodes parameters (e.g. vector of terminal node means) are usually specified assuming independence of terminal node parameters across terminal nodes. For regression trees, one could use normal priors centred around some shared mean value with a shared (or not) error variance; for classification trees Dirichlet distributions (multivariate generalisation of a beta prior) are recommended.
- a stochastic search of a promising CART (guided by the prior) through a Metropolis-Hastings search algorithm. After deciding on an initial tree, the proposal tree is obtained through one of four strategies (randomly selected): grow, prune, change, swap (see paper for details) and an acceptance ratio is computed based on the current and new tree draw. Such procedure simulates a Markov Chain sequence of trees that converges to the tree posterior distribution of interest $P(T|Y, X)$.

Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266–298.

The authors propose and ensemble method for approximating the expected value of some continuous and dichotomous Y, conditioned on a p-dimensional X matrix of predictors, by a sum of m regression trees. This Bayesian Additive Regression Tree method uses a sum-of-trees to approximate $f(x) = E(Y|X)$ and it does so by imposing a regularising prior that weakens each of the m trees making up the sum-of-trees. By doing so, each tree part of the sum is explaining a small but unique portion of f, and in this sense BART is not the same as averaging single trees fitted to approximate f, as boosting, bagging and random forests do. A BART model is made up of two fundamental parts:

- the sum-of-trees model $Y = \Sigma_{j=1}^{m} g(x; T_j, M_j) + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$ where $g(\ldots)$ is the function that assigns one of the terminal node parameters in M_j of tree T_j to all observation in x
- a (set of) regularising prior(s) for all of the parameters in the sum-of-trees model: all of the T_j and M_j, and σ. These priors keep the effect of each individual tree small. Interestingly: the T_j (independent) prior is operationalised exactly as for for the Bayesian CART in Chipman 1998 (see

equation 7 of both papers) in terms of a the probability that a split is non terminal ($p_{SPLIT} = \alpha(1 + d)^{\beta}$) and uniform probabilites of selecting a variable as a splitting one and an observed value as a splitting value; the prior for the components of M_j, μ_{ij} are normal priors with shared values for the mean and variance hyperparameters, which ends up being a specific version of the standard conjugate normal prior specified in equation 9 of Chipman 1998: finally the prior for $\sigma **2$ is also the same the one specified for the Bayesian CART, just expressed with the alternative inverse-chi-square parametrisation.

- choice of m - BART needs to a fixed number of m trees to be defined. This number is usually required to be high for prediction tasks, but when small (say 5 to 10) makes BART an interesting variable selection routine.

Compared to Chipman 1998, BART produces a posterior probability of a collection of m different trees conditioned on a dataset (Y,X) $p((T_1, M_1), ..., T_j, M_j), ..., T_m, M_m), \sigma^2|Y, X)$, instead of the posterior probability of one single tree $P(T|Y, X)$. However, the draws of each (T_j, M_j) can be done by drawing T_j with the Metropolis–Hastings algorithm defined in Chipman 1998 using a modified version of the data (y minus the fit from the sum of trees excluding the current tree), and the draws of M_j can be done from normal distributions. By iterating K times, meaning repeating the draw of all of the m trees and σ^2 K times, we obtain a sequnce of f^* draws that approximates the posterior disitrbution of $p(f|y)$, whatever $f(.)$ is.