The Journal of Open Source Software

VBLinLogit: Variational Bayesian linear and logistic regression

Jan Drugowitsch[1]

1 Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA

DOI: 10.21105/joss.01359

Software

- Review
- Repository
- Archive

Submitted: 10 March 2019
Published: 27 June 2019

Summary

Linear and logistic regression are essential workhorses of statistical analysis, whose Bayesian treatment has received much recent attention (Bishop, 2006; Gelman et al., 2013; Hastie, Tibishirani, & Friedman, 2011; Murphy, 2012). Using Bayesian statistics for linear and logistic regression allows specifying prior beliefs over certain model parameters, which makes it particularly useful for small and/or high-dimensional datasets. Bayesian regression furthermore provides an estimate of the uncertainty about estimated regression coefficients, as well as uncertainty about predictions arising from the regression. Both are again particularly important for small and/or high-dimensional datasets, as the use of such data might result in highly uncertain predictions, and allow the user to be explicit about this uncertainty.

`VBLinLogit` is a MATLAB/Octave library that provides a variational Bayesian implementation of Bayesian models for both linear and logistic regression. It uses variational Bayesian inference (Beal, 2003; Bishop, 2006; Murphy, 2012) as a method for approximating Bayesian computations, as these computations would otherwise be intractable for the used regression models. It is significantly faster than Markov Chain Monte Carlo (MCMC) methods (Gilks, Richardson, & Spiegelhalter, 1995), another form of approximate Bayesian inference, which makes it applicable to high-dimensional problems for which standard MCMC might be too slow.

A specific regression variant implemented by this library is automatic relevance determination (ARD), which uses a model that automatically determines which data dimensions are relevant for the regression, discarding the others (Wipf & Nagarajan, 2008). It does so without a separate "validation set", as would be required by alternative methods, like the Lasso (Tibshirani, 1996). Therefore, it can be used when the small size of the dataset makes the use of such a separate "validation set" prohibitive.

The scripts encompassing the library were written to be light-weight and thus do not depend on external libraries. They include variants with and without ARD. Their use is deliberately kept simple. The core arguments to the regression scripts are a matrix of predictors, as well as a vector of response variable. Additional parameters specifying the prior and hyperprior parameters are usually optional. Details about the specifics of the used models, the variational Bayes derivation, and detailed use of the scripts included in the library can be found in Drugowitsch (2013).

Additional details, novelty, and relation to other approaches

Use of the library is particularly beneficial if the data is sparse. Sparsity can occur either if few training examples are available, or if the dimensionality of the input (i.e., the number of dependent variables) is large. Data sparsity becomes particularly challenging if the input dimensionality exceeds the number of training examples. In this case, the regression is underdetermined: multiple solutions exist that fit the training set equally well. However, only some of them yield good predictions on a separate test set.

A common approach to handle underdetermination is to make additional assumptions about potential solutions. Specifically, it is commonly assumed that the regression weights (i.e., the regression coefficients) that form the solution and describe how the output (i.e., the independent variable) varies with the inputs, take small values. This is known as *regularization*. Regularization isn't only beneficial if the regression is underdetermined, but also if the data is noisy. The different methods discussed here differ in how exactly they introduce regularization by making different a-priori assumptions about the regression weights.

The Bayesian methods provided in this library implement two different sets of assumptions. They either assume that all regression coefficients are equally small and tunes how small they are overall (that's the variant without ARD), or they adjust "smallness" of each regression coefficient individually (that's the variant with ARD). The latter is particularly beneficial if the data includes some spurious input dimensions that don't determine the outputs. In this case, the ARD variant might be able to set the associated regression weights to zero, effectively ignoring these input dimensions. As mentioned further above, another benefit of Bayesian regularization is that it doesn't need a separate "validation set" to tune its parameters. How well this actually works depends on how close the data matches the assumptions underlying the different methods. Thus, different regularization approaches might work better or worse on different datasets. There currently exists no single best method that works best for all datasets.

More specifically, the models underlying variational Bayesian linear and logistic regression implemented in this library are Bayesian hierarchical models with priors on the regression coefficient, as well as hyper-priors on the prior parameters. For the ARD variants, the hyper-priors are assigned to each of the regressors separately, which supports pruning eventually irrelevant coefficients (Wipf & Nagarajan, 2008). This happens without the need for a separate validation set, unlike comparable sparsity-inducing methods like the Lasso (Tibshirani, 1996). Bishop (2006) describes ARD only in the context of type-II maximum likelihood (MacKay, 1992; Neal, 1996; Tipping, 2001), in which case the (hyper-)parameters are tuned by maximizing the marginal likelihood (or model evidence). The library instead provides implementations for the full Bayesian treatment, that finds the ARD hyper-posteriors by variational Bayesian inference.

Since release `R2017a`, MATLAB also provides some functions for Bayesian linear regression, but none for Bayesian logistic regression. For linear regression, it provides variants with and without variable selection. Neither of the variants without variable selection use or infer hyper-priors. Therefore, they do not support inferring prior parameters in hierarchical models, unlike this library. The variants with variable selection either use a straight-foward Bayesian formulation of Lasso, or Stochastic Search Variable Selection (SSVS) (George & McCulloch, 1993). Both are different from ARD, and the advantages of either method remain to be clarified.

Some of the scripts provided by this library have been included in the `pmtk3` software accompanying Murphy (2012). Part of the library has furthermore been used for various scientific publications across multiple domains (e.g., Kanitscheider, Coen-Cagli, Kohn, & Pouget, 2015; Oh et al., 2016; Ruxanda, Zahfir, & Muraru, 2018; Wang & Rehder, 2017).

References

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference* (PhD thesis). Gatsby Computational Neuroscience Unit, University College London.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer-Verlag.

Drugowitsch, J. (2013). Variational Bayesian inference for linear and logistic regression. arXiv:1310.5438 [stat.ML].

Gelman, A., Carlin, J. B., Sern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC texts in statistical science (3rd ed.). Chapman; Hall.

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881–889. doi:10.1080/01621459.1993.10476353

Gilks, W. R., Richardson, S., & Spiegelhalter, D. (Eds.). (1995). *Markov chain monte carlo in practice*. CRC interdisiplinary statistics. Chapman; Hall. doi:10.1201/b14835

Hastie, T., Tibishirani, R., & Friedman, J. (2011). *The elements of statistical learning*. Springer series in statistics (2nd ed.). Springer-Verlag. doi:10.1007/978-0-387-84858-7

Kanitscheider, I., Coen-Cagli, R., Kohn, A., & Pouget, A. (2015). Measuring Fisher information accurately in correlated neural populations. *PLoS Computational Biology*, *11*(6), e1004218. doi:10.1371/journal.pcbi.1004218

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation*, *4*(3), 415–447. doi:10.1162/neco.1992.4.3.415

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. Adaptive computation and machine learning series. The MIT Press.

Neal, R. M. (1996). *Bayesian learning for neural networks*. Springer-Verlag. doi:10.1007/978-1-4612-0745-0

Oh, H., Beck, J. M., Zhu, P., Sommer, M. A., Ferrari, S., & Egner, T. (2016). Satisficing in split-second decision making is characterized by strategic cue discounting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(12), 1937–1956. doi:10.1037/xlm0000284

Ruxanda, G., Zahfir, C., & Muraru, A. (2018). Predicting financial distress from Romanian companies. *Technological and Economic Development of Economy*, *24*(6), 2318–2337. doi:10.3846/tede.2018.6736

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society B*, *58*(1), 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine ensembles. *Journal of Machine Learning Research*, *1*, 211–244. doi:10.1109/icdm.2012.58

Wang, S., & Rehder, B. (2017). Multi-attribute decision-making is best characterized by an attribute-wide reinforcement learning model. *bioRxiv*. doi:10.1101/234732

Wipf, D. P., & Nagarajan, S. S. (2008). A new view of automatic relevance determination. In J. C. Platt, D. Koller, Y. Singer, & S. T. Roweis (Eds.), *Advances in neural information processing systems 20* (pp. 1625–1632). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/3372-a-new-view-of-automatic-relevance-determination.pdf