Supplementary Material for "Sparse models for predicting psychosocial impairments in patients with PTSD: an empirical Bayes approach"

Heiner Stuke[1,2], Kathlen Priebe[1], Veith A. Weilnhammer[1,2], Hannes Stuke[3], and Nikola Schoofs[1]

[1]Department of Psychiatry and Neurosciences

Charité Universitätsmedizin Berlin

[2]Berlin Institute of Health

[3]CertaintyLab Ltd Berlin

Author Note

Correspondence concerning this article should be addressed to Heiner Stuke, Department of Psychiatry and Neurosciences, Charité Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany, heiner.stuke@charite.de

Supplementary Material for "Sparse models for predicting psychosocial impairments in patients with PTSD: an empirical Bayes approach"

Data analysis

Linear Regression and regularization

The data $\mathcal{D}$ analysed in the present study consists of the n = 192 × p = 12 predictor matrix $\mathcal{X}$ (with the variables described above) and the 192 × 1 outcome vector y (with the degree of psychosocial impairments). According to a linear regression model, the outcome is modeled as a sum of predictor variables, each weighted by the 12 × 1 β coefficients (plus Gaussian-distributed noise), yielding the log likelihood function:

$$\mathcal{L} = logp(\mathcal{D}|\beta) = \sum_n log(\mathcal{N}(y_n|X_n \cdot \beta, \epsilon)) \tag{1}$$

where $\mathcal{N}$ denotes the Gaussian distribution. Maximizing this function finds the values of the coefficients β, for which the likelihood of the data under the model is maximal. However, as outlined in the introduction, the parameter values with a maximum likelihood of the data can be overfitted and thus generalize poorly to new data that are not part of the training process. To avoid overly high coefficients and the related overfitting, an additional penalty term for the β values can be added (regularization). The most frequently used penalty terms are L2 regularization, where the objective function becomes:

$$\mathcal{L} = \sum_n log(\mathcal{N}(y_n|X_n \cdot \beta, \epsilon)) - \lambda \sum_p \beta_p^2 \tag{2}$$

and L1 regularization with the objective function:

$$\mathcal{L} = \sum_n log(\mathcal{N}(y_n|X_n \cdot \beta, \epsilon)) - \lambda \sum_p |\beta_p| \tag{3}$$

Linear regression with L2 regularization is also referred to as Ridge Regression (Hoerl & Kennard, 1970), regression with L1 regression as Lasso Regression (Tibshirani, 1996). It becomes clear that both objective functions of regularized regression represent a compromise between the model fit for the current data (the first term with the likelihood of

the data) and the complexity of the model. The model complexity is taken as the sum of the squares of the coefficients in the ridge regression and as the sum of the absolute values of the coefficients in the lasso regression. This difference leads to specific effects of ridge versus lasso regularization with respect to the coefficient values: Ridge regression particularly punishes high coefficient values (due to the square, the punishment for example of a coefficient value of 3 compared to 2 is much higher than for 1 compared to 0). Thus, ridge regression effects low parameter values, which reduces overfitting and increases the generalizability of the results (Bishop, 2006; Fox, 2015). With lasso regularization, the punishment is independent from the coefficient magnitude (for example, the punishment of a coefficient value of 3 compared to 2 is the same as for 1 compared to 0). Hence, lasso regularization more strongly encourages to set coefficient values entirely to 0, i.e., to remove a predictor from the model. It is thus used for predictor selection (Bishop, 2006). In both cases, the λ parameter, which determines the regularization strength (interpretable as a weighing between model fit and complexity), is a hyperparameter, which has to be set manually or with cross-validation (where it is set to the value with the lowest prediction error in the test sets).

Bayesian Regression

In Bayesian regression, the posterior probability of the regression weights β is a function of their prior probability and the likelihood of the data given the parameters:

$$p(\beta|\mathcal{D}) = \frac{p(D|\beta)p(\beta)}{\int p(D|\beta)p(\beta)d\beta} \tag{4}$$

Through specific choices of the prior distribution, Bayesian pendants of Ridge Regression and Lasso Regression can be obtained.

Bayesian Ridge

As a simple case, one can assume a spherical Gaussian shrinkage (i.e., zero-centered) prior for β:

$$p(\beta) = \mathcal{N}(\beta|0, \sigma) \tag{5}$$

where $\mathcal{N}$ denotes the Gaussian distribution and σ is a scalar matrix. Since the log prior probability of β decreases quadratically with deviations from zero, this type of prior corresponds to regression with L2-regularization. As explained above, this type of regularization particularly reduces high coefficient values and the related overfitting. For the bayesian ridge, the posterior distribution over β is a multivariate Gaussian

$$p(\beta|\mathcal{D}) = \mathcal{N}(\beta|m, S) \tag{6}$$

where explicit formula for the mean vector m and the covariance matrix S are available (Bishop, 2006), chapter 3.1. The variance of the prior σ inversely determines the regularization strength (this can be understood intuitively in such a way that a prior that peaks sharply at zero implies a low prior probability for high parameter values). The prior variance can be set manually or its most likely value can be calculated from the data by maximizing the marginal likelihood (Empirical Bayes). Using point estimates for ϵ and σ, the marginal likelihood of the data (i.e., the likelihood of the data over the entire prior) is given as:

$$p(\mathcal{D}) = \int \mathcal{N}(y|X \cdot \beta, \epsilon) * \mathcal{N}(\beta|0, \sigma) d\beta \tag{7}$$

The inference on this model entails maximizing the marginal likelihood given by Equation 7 with respect to σ and ϵ and computing the posterior distribution over β based on the maximum marginal likelihood estimates for σ and ϵ. In this paper, we used an implementation of this procedure provided in the BayesianRidge class of Scikit-learn for Python (Pedregosa et al., 2011). Mean values and credible intervals were obtained by drawing 10000 samples from this posterior and then computing mean and 95% credible intervals (i.e., the interval around the mean in which the parameter value lies with 95% probability according to the posterior marginal distribution for each parameter). Finally, we computed the probability with which the parameter value lies in a prespecified interval of "no or negligible" impact. In the classical interpretation of effect sizes for correlation analysis (Cohen, 2013), a correlation coefficient of 0.1 is considered a "small effect".

Following this interpretation, we defined the interval of "no or negligible" impact of a predictor as a coefficient value between -0.1 and 0.1 (i.e., not even a small effect). Integrating the marginal distribution of the posterior over this interval allowed us to directly test the probability that the coefficient value of a certain predictor lies in this prespecified interval. This intregration can be done approximately by evaluating the samples drawn from the posterior (i.e., calculating the percentage of samples lying in this interval of no or negligible effect). Please note that due to the explicit parametric form of the multivariate Gaussian posterior, the credible intervals and the probability density mass in the interval of no or negligible impact can also be calculated explicitly (as shown in the publicly available source code).

Bayesian Lasso

In the Bayesian Lasso, a spherical Laplace distribution instead of a Gaussian distribution is used as a prior:

$$p(\beta) = \phi(\beta|0, \sigma) \tag{8}$$

where ϕ denotes the Laplace distribution and σ is a scalar matrix. Since the log prior probability of β decreases linearly with deviations from zero, this type of prior corresponds to regression with L1-regularization (Lasso Regression, Bishop (2006), chapter 3.1). As explained above, this type of regularization more strongly enforces coefficient values of 0 and thus effects predictor selection. For the Bayesian Lasso, the posterior given by Equation 4 does not have a parametric form with explicit formulas for the parameters (Park & Casella, 2008). For this reason, either sampling procedures such as Markov Chain Monte Carlo or approximate inference have to be used. Moreover, even having one global maximum, the Bayesian Lasso posterior is not symmetric (Figure 1), which is why mode (maximum a posteriori value) and mean of the distribution are not identical. Crucially, if samples are drawn from the posterior distribution and their mean value is computed, it will therefore be shifted away from 0, even if the distribution's maximum is at 0. This prevents the effect hoped for by the LASSO, namely setting of parameters to 0. This also applies to

global approximations of the posterior with a symmetrical distribution, which have their maximum close to the mean value and not at the maximum of the posterior (Figure 1).

In the present study, inference on the Bayesian Lasso model was performed firstly using stochastic variational inference (SVI, Hoffman et al. (2013) and Wingate and Weber (2013)) implemented in Pyro for Python (Bingham et al., 2019) and secondly using a local approximation (Laplace approximation). In the former case (SVI), we used a multivariate normal distribution as an approximate posterior distribution q over β and delta distributions (point estimates) for σ and ϵ:

$$q(\beta|\mathcal{D}) = \mathcal{N}(\beta|m_q, S_q) \tag{9}$$

Here, the objective function to be maximized with respect to the variance of the prior σ, the variance of the likelihood ϵ as well as the mean value m_q and covariance matrix S_q of the approximate posterior is the evidence lower bound on the marginal likelihood:

$$\mathcal{L} = ELBO = \int q(\beta) \log(p(\mathcal{D}|\beta)p(\beta))d\beta - \int q(\beta) \log q(\beta)d\beta \tag{10}$$

The evidence lower bound is a lower bound of the marginal likelihood of the data under the model. Maximizing it minimizes the dissimilarity between the approximate posterior q and the true posterior (Beal, 2003). Therefore, conceptually, the maximization process can be thought of as finding the multivariate Gaussian distribution which resembles the true posterior distribution as close as possible. We refer to standard textbooks (Bishop, 2006) and specific reviews (Blei et al., 2017) for details on variational methods for Bayesian inference.

Secondly, a local approximation to the posterior at the maximum aposteriori estimate was obtained using the Laplace approximation. Here, a maximum a posteriori (MAP) estimate of β was firstly obtained and then a local Gaussian approximation was computed by relating the covariance matrix of the posterior to the second derivative of the likelihood function at the MAP value (Evans & Swartz, 2000). Again, mean values and credible intervals were obtained by drawing 10000 samples from the approximate posteriors

and then computing mean and 95% credible intervals.

References

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference* (Doctoral dissertation). UCL (University College London).

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., & Goodman, N. D. (2019). Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research*, *20*(1), 973–978.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, *112*(518), 859–877.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Evans, M., & Swartz, T. (2000). *Approximating integrals via Monte Carlo and deterministic methods* (Vol. 20). OUP Oxford.

Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage Publications.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, *14*(5).

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Wingate, D., & Weber, T. (2013). Automated variational inference in probabilistic programming. *arXiv preprint arXiv:1301.1299*.

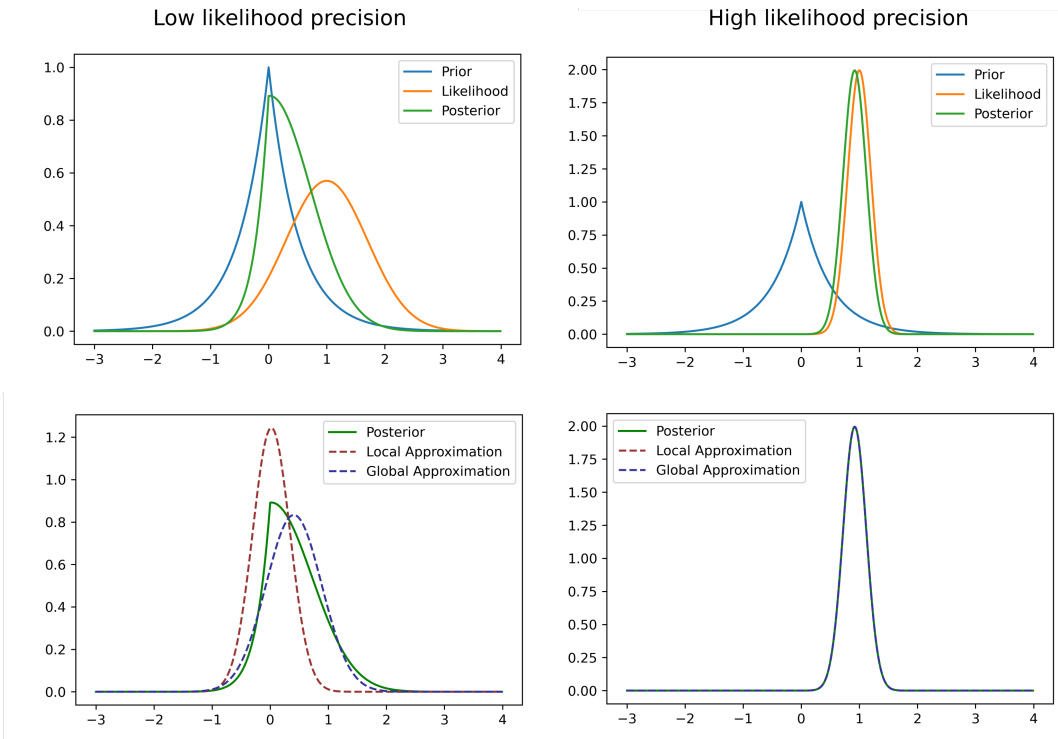

Figure 1

One-dimensional example of local and global approximations to the Bayesian Lasso posterior for low likelihood precision (uncertain β estimates, left column) and high likelihood precision (certain β estimates, right column). The above row shows the prior distribution (a zero-centered Laplace distribution), the Gaussian likelihood distribution and the resulting posterior distribution. It can be seen that in the case of high likelihood precision, the posterior distribution is largely determined by the prior and thus the maximum a posteriori estimate is strongly shrunken to 0. Conversely, in the case of high likelihood precision, the posterior closely resembles the likelihood distribution.

The lower row shows the local approximation to the posterior based on the local Laplace approximation and the global approximation based on variational Bayes. It can be seen that the mean value of the local approximation corresponds to the mode of the posterior distribution, whilst the mean value of the global approximation is close to the mean value of the posterior. As a sidenote please note that maximizing the ELBO does not lead to an approximate posterior whose mean necessarily matches the mean of the true posterior, because it minimizes the KL divergence between q and p, whilst moment matching between approximate and true posterior corresponds the minimizing the KL divergence between p and q. We refer to (Bishop, 2006), chapter 10, for readers with further interest in this topic.