Second exam (A2)

Class: Bayesian Statistics
Instructor: Luiz Max Carvalho
TA: Isaque Pim

26 June 2024

- You have 10 (ten) days to complete the exam;
- You must hand in a single PDF file typeset in your favourite flavour of TeX;
- Enunciate and prove every non-elementary result;
- For computational experiments – in case there are any –, give detailed descriptions of the implementation, running time, etc.
- The exam is worth $\min\{\text{your score}, 100\}$ marks.

1. The well-calibrated Bayesian

Dawid (1982) approaches the thorny issue of "calibration" (*reliability*) in a Bayesian context and how *coherence*, a concept so treasured by subjectivist Bayesians, fares in the quest for calibration.

a) (20 marks) Give a general account of the problem at hand and discuss why it is important to discuss a Bayesian approach to calibration. Be thorough.

b) (20 marks) Discuss the problem of forecasting "seemingly unrelated" quantities: what can *post hoc* analysis of the forecasts inform us about an individuals calibration? How can one factor in the sequential nature of many forecasting tasks? Consult the related literature to give a rich description of the problem.

c) (20 marks) Define what an admissible selection procedure is. Restate and prove the main theorem in Section 3 of Dawid (1982). What is the main takeaway of this result? Comment on the relationship between ξ_i and Y_i for the proof to go through and how one could interpret this relation.

d) (10 marks) Explain why in the sequential forecasting setting, the issue of poor calibration is much more serious. From a technical standpoint, why does ξ_i need to be measurable w.r.t $\mathfrak{B}_i$?

e) (10 marks) Discuss the implication of the theorem above to Bayesian coherence – **Hint:** use the mathematical definitions in Section 6. In particular, reason through what it means to assume that $\mathfrak{B}_0 \subseteq \mathfrak{B}_1 \subseteq \ldots$. Further, reason about what it means to specify a distribution over $\bigcup_{i=0}^{\infty} \mathfrak{B}_i$ *ahead of time*.

f) (20 marks) Argue that re-calibration of forecasts is incoherent. Show how requiring *strict* coherence is equivalent to Cromwell's law – you are strongly encouraged to read and discuss Lindley (1980). Analyse the conflict between coherence and calibration within the subjectivist framework; argue that as soon as the forecaster is seen as a "calibratable" machine, the tension goes away.

g) * (10 bonus marks) Show how abandoning coherence can be seen as a particular incarnation of the "All models are wrong, some are useful" philosophy.

2. A note on Metropolis-Hastings kernels for general state spaces

The Metropolis-Hastings algorithm is a cornerstone of applied Bayesian Statistics, allowing researchers to relatively effortlessly set up samplers to produce approximate samples from high-dimensional distributions on complicated spaces. There is, however, a lot of theoretical nuance is involved in guaranteeing the correctness of the algorithm. Tierney (1998) gives a general account and collects many interesting results that underpin the theoretical guarantees we are able to give about MH samplers.

a) (20 marks) Give a general account of the Metropolis-Hastings algorithm, its history and general construction. Pay particular attention to the mathematical nuances of moving to general state spaces.

   Hint: Define the main ingredients involved (acceptance probability, transition kernels, reversibility, detailed balance) and give examples. It would be convenient if you fixed a simple yet non-trivial statistical example as the motivation for the discussion.

b) (10 marks) Consider Proposition 1: why is it interesting that μ and μ^T are mutually singular on R^c? What is the interpretation of $r(x, y)$?

c) (10 marks) Now, let us analyse Theorem 2. Contrast conditions (i) and (ii) in terms of their feasibility – how easy they are to verify and implement. Show that the common Metropolis-Hastings acceptance satisfies these conditions by construction. Moreover, show that[1] for $x, y \in R$

$$\alpha_{\mathrm{B}}(x, y) = \frac{\pi(dy)\mu_R(dy, dx)}{\pi(dx)\mu_R(dx, dy) + \pi(dy)\mu_R(dy, dx)},$$

   and $\alpha_{\mathrm{B}}(x, y) = 0$ otherwise satisfies both conditions. This is called Barker's acceptance probability (Barker, 1965).

d) (20 marks) Consult Green (1995) in order to properly discuss the model selection setting; Show how the results discussed so far can be applied in this situation. Discuss what is needed from the dimension-changing kernels.

e) (20 marks) Define Peskun ordering and give examples. Discuss the implication that $P_1 \succeq P_2$ to the ordering of their asymptotic variances, carefully proving any results that are needed along the way – **Hint:** Prove and discuss Lemma 3 and Theorem 4. Then, discuss why the ordering of asymptotic variances gives no guarantees on the ordering of empirical variances for all chain sizes (numbers of iterations) N.

f) (20 marks) Compare the mixture-of-maximal-kernels versus the maximal-mixture-of-kernels approaches in Proposition 5. Build an empirical example to illustrate the relative merits of each strategy, paying attention to the cost of computing the transition kernel densities – one must consider situations where those are cheap or expensive to compute.

[1] Note that we are following the paper here and μ_R is a σ-finite measure on $(E \times E, \mathcal{E} \otimes \mathcal{E})$ restricted to the symmetric set R as defined in the paper.

g) * (30 bonus marks) Recall that π is the target distribution. Take $f \in L^2(\pi)$ and define $\mu_\pi := \mathbb{E}_\pi[f(X)]$ and $\sigma_\pi^2 := \mathbb{E}_\pi[\{f(X) - \mu_\pi\}^2]$. Suppose there holds a central limit theorem of the form

$$\frac{\sum_{t=1}^N f(X_t) - \mu_\pi}{\sqrt{N}} \Rightarrow \text{Normal}(0, \sigma_{\text{MH}}^2),$$

as $N \to \infty$ for $(X_t)_{t=0}^N$ sampled according to an MH with the usual acceptance probability. Show that

$$\sigma_{\text{MH}}^2 \leq \sigma_{\text{B}}^2 \leq 2\sigma_{\text{MH}}^2 + \sigma_\pi^2,$$

where σ_{B}^2 is the asymptotic variance of a chain constructed using Barker's acceptance probability.

3. The Bayesian LASSO

The least absolute shrinkage and selection operator (LASSO, Tibshirani (1996)), along with other sparsity-inducing approaches, is a fundamental tool in applied research where a handful of important effects swim in a sea of irrelevant ones. Park and Casella (2008) give an account of the Bayesian approach to the LASSO, discussing the main aspects of prior specification, implementation and inference, including for the regularisation parameter.

a) (20 marks) Give a general account of the problem at hand and discuss why a Bayesian solution would be desirable. Be thorough.

b) (10 marks) Justify using a Laplace prior instead of, for instance, a normal prior. Go beyond conjugacy considerations, but do not overlook them. Explain the choice for a conditional Laplace prior from both a mathematical/computational point of view and a scientific perspective.

c) (10 marks) Discuss the hierarchy suggested by Park and Casella (2008). Show how this hierarchy could be modified in order to facilitate MAP estimation. Carefully discuss each choice of priors and reflect on how one would go about including information about, say, the measurement noise parameter, σ^2.

d) (10 marks) Discuss why marginalise μ in the computation and when that is and is not desirable. Show how to recover inferences about μ in the marginalised case.

e) (20 marks) The classical LASSO is known for its shrinkage properties. Now, let us computationally investigate the shrinkage properties of the Bayesian LASSO. We will use the artificial scenarios 1, 2, 3, and 4 from Section 7.2 of the seminal LASSO Tibshirani (1996) paper as our basis for exploration. Implement the Bayesian LASSO using a tool of your choice, such as Stan (or by implementing the Gibbs sampler from scratch). Clearly specify your priors, assess model convergence, and report relevant metrics to ensure convergence and mixing. Quantification of shrinkage should be performed by verifying whether 0 is included in a significant credible interval (using a 95% credible interval). This involves measuring correct inclusions (cases where the true parameter value was zero and was correctly not included in the credible interval) and false inclusions. Visualise and interpret the results.

f) (20 marks) Discuss and contrast the two approaches presented in the paper for "choosing" the LASSO parameter, λ. How should one manage prior specification to maintain conjugacy? Analyse what is required of a prior on λ – or transformations thereof – in order to achieve computational tractability. Consider the justification for the hyperparameters of the prior on λ given by Park and Casella (2008): does that convince you? Using the diabetes data analysed in the paper, re-run their experiment with different hyperparameter choices and report the results in terms of both estimation and computational efficiency. Be thorough in your description of the sampler(s) employed.

g) (10 marks) In the "Huberised LASSO", explain the proposed choice of prior for $\boldsymbol{D}_\sigma$.

h) * (20 bonus marks) Prove or disprove the following conjecture:

Conjecture 1 (Multimodality of the LASSO with unconditional prior) *Consider a setting (model and notation) as in Park and Casella (2008). If one specifies the joint prior over the unknown parameters as*

$$\pi_\lambda(\boldsymbol{\beta}, \sigma^2) = \pi_V(\sigma^2) \prod_{j=1}^{p} \frac{\lambda}{2} \exp\left(-\lambda|\beta_j|\right),$$

then, for any choice of $\lambda \geq 0$ and some choice of π_V, the posterior has more than two modes when $p > 1$.

4. A weakly informative default prior distribution for logistic and other regression models

Logistic regression is an ubiquitous tool in applied statistical modelling. In practice, issues of separation and sparsity crop up constantly. A Bayesian approach follows naturally as a way of making models identifiable and tractable by smoothing over ridges in the likelihood caused by separation and providing shrinkage when many predictors are considered. As always, there is the issue of specifying the priors in a way that is both sensible and general, in the sense that it can be used as a default prior in applied research. Gelman et al. (2008) discusses a tentative solution within the Student-t family, evaluating it on a large corpus of logistic regression examples.

a) (20 marks) Give a general account of the problem at hand and discuss why a Bayesian solution would be desirable, focusing on the regularisation properties of the prior in the presence of separation and sparsity. Be thorough.

b) (10 marks) Explain why and how to scale predictors for binary regression, drawing directly from Raftery (1996) to expand the explanation given in Section 2.1 of Gelman et al. (2008). Further, explain how a default prior for all coefficients would penalise interactions in such a model and discuss the implications in terms of sparsity.

c) (10 marks) Explain the approximate EM proposed by the authors, paying particular attention to the exploitation of the hierarchical structure inherent in their prior choice.

d) (20 marks) As the results in Sections 4.1 and 4.2 indicate, the comparative advantage of the proposed class of priors relative to, say, standard normal priors. An important consideration when choosing priors is the implications of a given choice to certain *observables* in the model. Using the framework laid out by Gabry et al. (2019) and references therein, evaluate the Cauchy, Laplace and Normal priors discussed by Gelman et al. (2008) using **prior predictive checks** (PPC). In particular, reason through which kinds of statistics (observable or otherwise) would be helpful in evaluating the priors for this particular class of models. Consider at least two link functions commonly employed with binary regression.

e) (40 marks) The authors conduct a comprehensive comparison of Bayesian estimators using their proposed priors and frequentist methods using multiple datasets from the UCI Machine Learning Repository. They employ Bayesian point estimations and cross-validation to assess predictive performance. Today we have many tools that facilitate a fully Bayesian analysis of the proposed priors. We will now compare the predictive performance of the priors computationally. For this analysis, select five datasets from the UCI repository. Using Stan[2] fit the proposed models, assess convergence and report relevant metrics. Compare the predictive performance of

[2]Feel free to code the probability model on your own or to use ready-to-use packages such as *rstanarm*.

each choice of prior using LOO-CV [3]. Do the results agree with the point estimate analysis? Discuss.

f) ** (50 bonus marks) An issue with heavy tails is that some estimators (e.g. the Bayes estimator under quadratic loss) might not exist. Let us investigate the existence of posterior means under the recommended Cauchy prior. Consider the logistic model

$$\text{logit}(\theta_i) = \boldsymbol{x}_i^T \boldsymbol{\beta},$$

where $\boldsymbol{x}$ is a an $n \times P$ matrix of predictors and $\boldsymbol{\beta}$ is a P-dimensional vector of coefficients.

Now, let us consider the definition of separation:

Definition 1 (Complete and solitary separation) *Let $A_0 = \{i : y_i = 0\}$ and $A_1 = \{i : y_i = 1\} = \{1, \ldots, n\} \setminus A_0$. We say **complete separation** occurs if there exists a vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_P$ such that for all $i = 1, \ldots, n$ it holds that*

$$\boldsymbol{x}_i^T \boldsymbol{\alpha} > 0 if i \in A_1 \text{ and } \quad \boldsymbol{x}_i^T \boldsymbol{\alpha} < 0 \quad \text{otherwise.}$$

We say $\boldsymbol{x}_j$ is a solitary separator iff there exists $\boldsymbol{\alpha}$ such that

$$\alpha_j \neq 0, \alpha_k = 0 \quad \forall k \neq j.$$

Now, prove the following claim:

***Claim 1 (Existence of posterior means)** Suppose the joint prior density for the coefficients (including the intercept) is given by*

$$\pi_B(\boldsymbol{\beta}) = \prod_{j=1}^{P} \pi_j(\beta_j) = \prod_{j=1}^{P} \left(\pi\sigma_j(1 + \beta_j^2/\sigma_j^2)\right)^{-1},$$

*where $\sigma_j > 0$ for all j are the prior scales. Then $\mathbb{E}[|\beta_j| \mid \boldsymbol{y}, \boldsymbol{x}] < \infty$ if and only if $\boldsymbol{x}_j$ is **not** a solitary separator.*

In light of the claim, comment on the role of centring the covariates for the existence of the posterior mean of each coefficient.

[3] LOO-CV - Leave-one-out cross-validation. You may use the R package *loo* for this task `https://cran.r-project.org/web/packages/loo/index.html`

5. Spike-and-slab meets LASSO: A review of the spike-and-slab

Sparsity-inducing priors are a staple of modern applied Statistics. Sparsity can be achieved in two main ways: spike-and-slab priors (George and McCulloch, 1993) and continuous shrinkage priors, such as the horseshoe (Carvalho et al., 2010). Bai et al. (2021) provide an analysis of the connection between these two apparently separate worlds by bridging the LASSO (Tibshirani, 1996) and the spike-and-slab.

a) (20 marks) Give a general account of the problem of sparsity in high-dimensional regression and discuss why a Bayesian solution would be desirable. Describe succinctly the SSVS and continuous shrinkage approaches (**Hint:** consult the references above) and argue for the desirability of the spike-and-slab LASSO.

b) (10 marks) Formulate the problem of high-dimensional sparse regression as a penalised likelihood problem and discuss the form of the penalisation term. Discuss choice between penalisations in terms of their *separability*.

c) (10 marks) Describe the spike-and-slab LASSO, paying particular attention to the prior specification aspect of the modelling endeavour. What is the role of the prior on the mixing parameter? How should one structure the "slab" component? Are there any theoretical reasons that underpin certain choices (e.g. matching the regular LASSO *via* MAP)?

d) (20 marks) Explain the adaptive nature of the spike-and-slab penalty and argue about the importance of having such a mechanism. How reasonable is the simplifying assumption that $\theta_j \approx \hat{\theta} = E[\theta \mid \boldsymbol{\beta}]$? Discuss the role of λ_0 and λ_1 in the specification of the penalty and contrast their interpretation with the computational subtleties described in Section 4.2.

e) (40 marks) [4] Prediction problems in which the number of features p is much larger than the number of observations N, often written as $p \gg N$, have emerged in many practical problems recently. For example, gene expression arrays typically have 50 to 100 samples and 5,000 to 20,000 variables. Let's explore the performance of SS-LASSO by comparing it with similar methods *via* a computational experiment. Using R, generate samples from the same linear model scenario in section 6.1 of Bai et al. (2021). Compare the traditional LASSO (you can use the *glmnet* package for this task), a Spike-and-Slab with conjugate priors (you can use *spikeSlabGAM* for this task), SCAD and MCP (use package *ncvreg*) and the studied method SS-LASSO. Report discovery rates and other relevant metrics you find necessary. Additionally, compare the estimates of *spikeSlabGAM* and SS-LASSO. Does SS-LASSO overshrink parameter estimates compared to *spikeSlabGAM*?

f) * (30 bonus marks) Show that under a separable penalty as in the SS-LASSO, the predictive density can be written as a linear mixture of the

[4] SS-LASSO - Spike-and-Slab LASSO; SCAD - Smoothly Clipped Absolute Deviation; MCP - Minimax Concave Penalty

predictive densities under the spike and slab priors respectively. Exhibit the form of this mixture, including the functional form of the weights.

Bibliography

Bai, R., Ročková, V., and George, E. I. (2021). Spike-and-slab meets LASSO: A review of the spike-and-slab LASSO. *Handbook of Bayesian variable selection*, pages 81–108.

Barker, A. A. (1965). Monte Carlo calculations of the radial distribution functions for a proton-electron plasma. *Australian Journal of Physics*, 18(2):119–134.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Dawid, A. P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. (2019). Visualization in bayesian workflow. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 182(2):389–402.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360 – 1383.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732.

Lindley, D. V. (1980). *The Bayesian approach to statistics*. University of California (Berkeley). Operations Research Center.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the american statistical association*, 103(482):681–686.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, 83(2):251–266.

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

Tierney, L. (1998). A note on Metropolis-Hastings kernels for general state spaces. *Annals of Applied Probability*, pages 1–9.