# Penalized Complexity Priors

Valerie Han, Sebastian McCrimmon, Benjamin Jacobs

September 28, 2023

# Background & Motivation

1. Overview of existing priors
2. "Base model" framework
3. Desired qualities for a prior

# Objective and weakly informative priors

- Objective prior
  - Priors that try to be as non-informative as possible
  - Ex: Jeffreys' prior, reference prior
- Weakly informative prior
  - In between objective and expert priors
  - Ex: Should not give prior mass at 20 ft when estimating the height of an adult human

# Ad hoc, risk adverse, and computationally convenient prior specification

- Ad hoc approach
  - Prior that has been previously used for a similar problem
  - Often inappropriate for the new application
- Compuationally convenient prior
  - Ex: conjugate prior
- Risk averse prior
  - Ex: independent priors on each hyperparameter
  - May have unintentional consequences like not having enough shrinkage/sparsity to limit the flexibility of the model and avoid over-fitting

# “Base model”

- Consider model component with density $\pi(x|\xi)$ controlled by flexibility parameter $\xi$. Then, the base model is the “simplest” model in the class. For notational convenience, let $\xi = 0$ for the base model.
- Interpret $\pi(x|\xi)$ as a flexible extension of the base model where increasing values of $\xi$ correspond to increasing deviations from the base model
- Ex: Gaussian random effects
  - Let $x|\xi \sim N(0, \xi I)$.
  - Base model has $\xi = 0$, i.e., no random effects
- Prior $\pi(\xi)$ on $\xi$ should place sufficient mass on the base model (otherwise, the prior “forces overfitting” / “overfits”)
  - Ex: spike-and-slab priors (computationally unpleasant)

# Desired qualities for a prior

1. Prior should not be noninformative.
    - Want more stability of inference.
    - If model is overparametrized (too flexible), posterior will be over-fitted.
2. Prior should be aware of the model structure.
    - If a subset of the parameters control a single aspect of the model, the prior for these should be set jointly.
    - Try to use parametrization of model where each parameter only controls one aspect of the model.
3. Prior specification should be explicit about what needs to be changed when applying it to a similar but different problem.

# Desired qualities for a prior

4. Prior should limit the flexibility of an over-parametrized model
   - Shrinkage properties
5. Should set a joint prior for non-identifiable parameters.
6. Prior should (at least locally) be indifferent to the parametrization used.
7. Prior should be computationally feasible.
8. Prior should perform well; estimators produced using the prior should have appropriate theoretical properties.

# Section 3: Defining the PC prior

- Principle 1: Occam's razor
  - The simpler model should be preferred until there is enough evidence to support a more complicated model.
- Principle 2: Measure of complexity
  - The measure of complexity is based off the Kullback-Leibler divergence, which is defined as

$$KLD(\pi(\mathbf{x}|\xi)||\pi(\mathbf{x}|\xi = 0)) = \int \pi(\mathbf{x}|\xi) \log(\frac{\pi(\mathbf{x}|\xi)}{\pi(\mathbf{x}|\xi = 0)})d\mathbf{x} \tag{1}$$

  - We define the distance between two models with densities $f$ and $g$ as $d(f||g) = \sqrt{2KLD(f||g)}$.

# The constant rate penalization principle

- Principle 3: Constant rate penalisation
  - There is a prior placed on the deviation from the base model, which is measured using $d$. The prior satisfies the memory-less property, meaning

    $$\frac{\pi_d(d+\delta)}{\pi_d(d)} = r^{\delta} \tag{2}$$

    where $d, \delta \geq 0$ and $0 < r < 1$.
  - The mode of the prior is at $d = 0$ (the base model).
  - The constant rate penalization assumption implies that $\pi(d) = \lambda \exp(-\lambda d)$.
  - After applying the change of variables formula, the prior on the $\xi$ is defined as

    $$\pi(\xi) = \lambda \exp(-\lambda d(\xi)) \left| \frac{\partial d(\xi)}{\partial \xi} \right| \tag{3}$$

# How do you choose $\lambda$?

- Principle 4: User-defined scaling
  - The analyst can select $\lambda$ by controlling the prior mass in the tail. More formally, the analyst can choose the value of $\lambda$ that satisfies the following equation

$$Prob(Q(\xi) > U) = \alpha \tag{4}$$

  - $Q(\xi)$ is an interpretable transformation of the flexibility parameter.
  - $U$ is a user-defined upper bound that specifies a "tail-event" and $\alpha$ is the weight we put on this event.

# The PC prior for the flexibility parameter of a Gaussian random effect

- We'll derive the PC prior for the flexibility parameter of a p-dimensional Gaussian random effect $\mathbf{x}$, where $\mathbf{x} \sim MVN(0, \xi\mathbf{I})$.
- The prior is

$$\pi(\xi) = \frac{\lambda}{2\sqrt{\xi}} \exp(-\lambda\sqrt{\xi}) \tag{5}$$

# What should $\lambda$ be for this example?

- Let $Q(\xi) = \sqrt{\xi}$.
- In order for $Prob(\sqrt{\xi} > U) = \alpha$, $\lambda = \frac{-\log(\alpha)}{U}$.
- After integrating out $\xi$, the marginal standard derivation of **x** will be about $0.31U$ when $\alpha = 0.01$. Thus, choosing $(U = 0.968, \alpha = 0.01)$ will give $Stdev(\mathbf{x}) \approx 0.3$.
- More intuitive than choosing hyperparameters of a given prior.

# Section 4: Properties of PC Priors

1. Behavior and asymptotics near the base model
2. Hypothesis testing
3. Sparsity Priors

# Is there a relationship to the Jeffreys' prior?

If $\lambda\xi \approx 0$, then the following holds:

$$\pi(\xi) = I(\xi)^{1/2} exp\{-\lambda \int_0^\xi \sqrt{I(s)} ds\} + \text{higher order terms}$$

which implies that near the base model, the prior is a 'tilted' version of the Jeffreys' prior.

# Large Sample behavior under the base model

1. Asymptotic results like the Bernstein-von Mises theorem require that the true parameter value lie not on the edge of the parameter space.
2. but when the base model is true, then $\xi = 0$ which is often at the very edge of its parameter space.
3. The authors cite (Self and Liang, 1987) that if the prior density of the base model is finite, the large sample behaviour of the posterior is identical to that of the maximum likelihood estimator.
4. So the exponential distribution on $\xi$ ensures this happens (by putting finite density at $\xi = 0$).

# Hypothesis testing

1. The authors do not recommend using PC priors for Bayesian hypothesis testing.
2. THEOREM 2. Under the conditions of Bochkina and Green (2014), the Bayes factor for the test $H_0$: $\xi = 0$ against $H_1 : \xi > 0$, is consistent when the prior for $\xi$ does not overfit. That is, $B_{01} \to \infty$ under $H_0$ and $B_{01} \to 0$ under $H_1$, where $B_{01}$ denotes the Bayes factor for candidate model $M_0$ against candidate model $M_1$.
3. HOWEVER, the rates of convergence are not the same under the null as under the alternative, suggesting suboptimal finite-sample properties.
4. The authors allude to some potential solutions to this and provide sources, but don't go into detail.

# Sparsity Priors/applications to variable selection

Example:

$$y_i \sim \pi(y_i|\beta),\ \beta \sim N(0, D),\ D_{i,i}^{-1} \sim^{iid} \pi(\tau)$$

1. the base model is a sparse vector $\beta$.
2. The authors claim that a correct application of their PC-prior principles would first put an exponential distribution on the number of nonzero components, and then put i.i.d. PC priors on each of the selected components conditional on their inclusion.
3. the other idea is to just put a PC prior on $\tau$, but this doesn't work well
4. the problem is essentially that putting iid PC priors on the coefficients would put most probability on the event that they are all small, and what we really would want is most of the probability on the event that a few are large, and the rest are small.

# Extensions

1. Disease mapping with Besag-York-Mollié models
2. Multivariate (probit) models
3. Hierarchical models