Bayesian Inference under Small Sample Sizes Using General Noninformative Priors

Daniil Dorin

MIPT, 2024

October 1, 2024

Motivation

Main idea

- Sample sizes in many engineering fields are often small due to time, economic, and physical constraints. Traditional statistical methods may not be suitable under such conditions.
- Bayesian inference is a preferred probabilistic approach in these situations, allowing the integration of prior knowledge with observed data. However, the choice of prior distributions is crucial, especially with limited data, as it can significantly influence the inference results.
- This paper introduces a general uninformative prior for Bayesian inference in problems characterized by small sample sizes, aiming to reduce subjectivity and provide objective inference.

Definitions

Given:

- Likelihood: $p(\mathcal{D}|\mathbf{w})$, $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$.
- Prior distribution: $p(\mathbf{w}|\mathbf{h})$, probability of the parameters $\mathbf{w}$ given the prior parameters $\mathbf{h}$.

Posterior distribution $p(\mathbf{w}|\mathcal{D}, \mathbf{h})$:

$$p(\mathbf{w}|\mathcal{D}, \mathbf{h}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w}|\mathbf{h})}{\int p(\mathcal{D}|\mathbf{w})p(\mathbf{w}|\mathbf{h})d\mathbf{w}}.$$

Predictive Performance:

$$p(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}, \mathbf{y}_{\text{train}}, \mathbf{X}_{\text{train}}) = \int p(\mathbf{y}_{\text{test}}|\mathbf{w}, \mathbf{X}_{\text{test}})p(\mathbf{w}|\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}})d\mathbf{w}$$

Definitions

Uninformative prior

If the contribution of the prior is negligible with respect to that provided by the data, then we say that the prior is uninformative.

- Jeffreys prior (invariant under reparameterization):

$$p(w) \propto \sqrt{\det I(w)},$$

  where $I(w)$ is the Fisher information matrix.

Normal-Inverse-Gamma distribution (NIG)

Suppose

$$\theta \mid \sigma^2, \mu, \Sigma \sim \mathrm{N}(\mu, \sigma^2/\Sigma), \quad \sigma^2 \mid \alpha, \beta \sim \Gamma^{-1}(\alpha, \beta).$$

Then (θ, σ^2) has a normal-inverse-gamma distribution, denoted as

$$(\theta, \sigma^2) \sim \text{N-}\Gamma^{-1}(\alpha, \beta, \Sigma, \mu).$$

Proposed method

Assumption

Assume the normal likelihood of the data.

Idea

Uninformative priors, including the Jeffreys, Asymptotically Locally Invariant (ALI), for (θ, σ^2) are mostly in the form of $1/\sigma^q$, $q \in 1, 2, \ldots$ The uniform prior can be seen as a special case of $1/\sigma^q$ as $q = 0$.

It is shown as follows that these uninformative priors can be obtained as certain limiting states of NIG conjugates of (θ, σ^2), $\dim \theta = k$. For example, the NIG distribution with parameters $(\alpha, \beta, \Sigma, \mu)$ can reduce to the Jeffreys prior:

$$p(\theta, \sigma^2) \to \frac{1}{\sigma^2}, \quad \text{when} \quad \begin{cases} \alpha \to -k/2 \\ \beta \to 0^+ \\ \Sigma^{-1} \to \mathbf{0} \\ |\mu| < \infty \end{cases}$$

Proposed method

Using the conjugacy property of the normal likelihood and the NIG prior, the following estimates are derived:

Table 1. $1/\sigma^q$-type of prior as limiting state cases of the NIG distribution and its NIG posteriors.

Prior (θ, σ^2)	**NIG**$(\alpha, \beta, \theta, \sigma^2) \to$ **Prior**	**Posterior** $\to$ **NIG**$(\alpha^*, \beta^*, \theta, \sigma^2)$
flat	$\alpha = -k/2 - 1$ $\beta \to 0$ $\Sigma^{-1} \to 0$	$\alpha^* = \alpha + n/2 = (n-k-2)/2$ $\beta^* = \beta + \mathrm{SSE}/2 = \mathrm{SSE}/2$ $\Sigma^* = (\Sigma^{-1} + \mathbf{x}^T\mathbf{x})^{-1} = (\mathbf{x}^T\mathbf{x})^{-1}$
$\frac{1}{\sigma}$	$\alpha = -k/2 - 1/2$ $\beta \to 0$ $\Sigma^{-1} \to 0$	$\alpha^* = \alpha + n/2 = (n-k-1)/2$ $\beta^* = \beta + \mathrm{SSE}/2 = \mathrm{SSE}/2$ $\Sigma^* = (\Sigma^{-1} + \mathbf{x}^T\mathbf{x})^{-1} = (\mathbf{x}^T\mathbf{x})^{-1}$
$\frac{1}{\sigma^2}$	$\alpha = -k/2$ $\beta \to 0$ $\Sigma^{-1} \to 0$	$\alpha^* = \alpha + n/2 = (n-k)/2$ $\beta^* = \beta + \mathrm{SSE}/2 = \mathrm{SSE}/2$ $\Sigma^* = (\Sigma^{-1} + \mathbf{x}^T\mathbf{x})^{-1} = (\mathbf{x}^T\mathbf{x})^{-1}$
$\frac{1}{\sigma^3}$	$\alpha = -k/2 + 1/2$ $\beta \to 0$ $\Sigma^{-1} \to 0$	$\alpha^* = \alpha + n/2 = (n-k+1)/2$ $\beta^* = \beta + \mathrm{SSE}/2 = \mathrm{SSE}/2$ $\Sigma^* = (\Sigma^{-1} + \mathbf{x}^T\mathbf{x})^{-1} = (\mathbf{x}^T\mathbf{x})^{-1}$
$\frac{1}{\sigma^4}$	$\alpha = -k/2 + 1$ $\beta \to 0$ $\Sigma^{-1} \to 0$	$\alpha^* = \alpha + n/2 = (n-k+2)/2$ $\beta^* = \beta + \mathrm{SSE}/2 = \mathrm{SSE}/2$ $\Sigma^* = (\Sigma^{-1} + \mathbf{x}^T\mathbf{x})^{-1} = (\mathbf{x}^T\mathbf{x})^{-1}$
$\frac{1}{\sigma^5}$	$\alpha = -k/2 + 3/2$ $\beta \to 0$ $\Sigma^{-1} \to 0$	$\alpha^* = \alpha + n/2 = (n-k+3)/2$ $\beta^* = \beta + \mathrm{SSE}/2 = \mathrm{SSE}/2$ $\Sigma^* = (\Sigma^{-1} + \mathbf{x}^T\mathbf{x})^{-1} = (\mathbf{x}^T\mathbf{x})^{-1}$

Proposed method

Bayesian prediction

In this case, the Bayesian prediction can be calculated analytically:

$$p(\mathbf{y}_{\text{test}} \mid \mathbf{y}_{\text{train}}) = \mathrm{MVT}_{n-k} \begin{pmatrix} \mathbf{X}_{\text{test}}(\mathbf{X}_{\text{train}}^T \mathbf{X}_{\text{train}})^{-1}(\mathbf{X}_{\text{train}}^T \mathbf{y}_{\text{train}}), \\ \frac{\mathrm{SSE}}{n-k}\left(\mathbf{I} + \mathbf{X}_{\text{test}}(\mathbf{X}_{\text{train}}^T \mathbf{X}_{\text{train}})^{-1}\mathbf{X}_{\text{test}}^T\right) \end{pmatrix},$$

where MVT is multivariate t–distribution. Sum of squared errors is defined as $\mathrm{SSE} = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T \boldsymbol{\theta})^2$.

Advantages of treating the $1/\sigma^q$-type of uninformative prior

- The Bayesian posteriors of the model prediction and parameters all have analytical forms, allowing for efficient evaluations without resorting to the MCMC techniques.

Experiments

Data: The dataset contains data on low-cycle fatigue of materials obtained as a result of tests according to the ASTM E739-10 standard. These data are used to assess the reliability of materials under low-cycle fatigue.

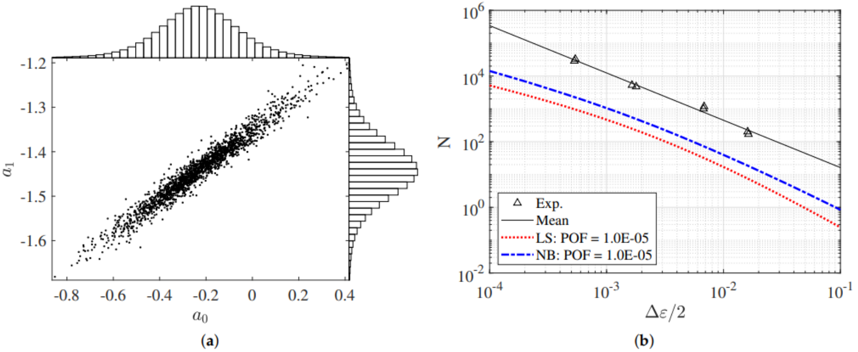


Figure 1. (**a**) MCMC samples drawn from the noninformative Bayesian posterior with Jeffreys' prior ($q = 2$) and (**b**) the fatigue life results at POF $= 10^{-5}$ obtained with the regular least squares (LS) estimator and noninformative Bayesian (NB).

Experiments

Data: Aeroengine Turbine Disk Lifing problem. The number of tested piece is usually less than five due to the cost and time constraints.

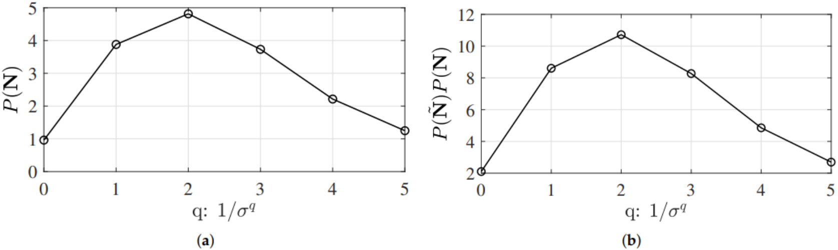
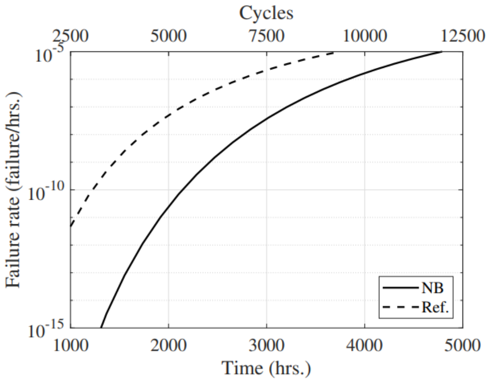


Figure 2. **(a)** The fitting performance $P(N)$ evaluated with different $1/\sigma^q$ priors and **(b)** the prediction performance $P(\tilde{N})P(N)$ evaluated with different $1/\sigma^q$ priors. One data point $\tilde{N}$ is considered for the prediction.

For the $1/\sigma^q$-type of uninformative prior, the classical Jeffreys prior $1/\sigma^2$ yielded an optimal fitting and prediction performance in terms of the likelihood or Bayes factors.

Experiments

Results of uninformative Bayesian (NB) comparable with those obtained using the existing method (Ref.) incorporating the empirical evidence and assumptions.



Literature

1. **Main article:** He J. et al. Bayesian inference under small sample sizes using general noninformative priors // Mathematics.