Bayesian inference for quantiles of a log-normal variable in presence of censored observations

Enrico Fabrizi[1], **Aldo Gardini**[2] and Carlo Trivisano[2]
aldo.gardini2@unibo.it

[1]DISES, Università Cattolica del S. Cuore
[2]Dipartimento di Scienze Statistiche 'P. Fortunati', Università di Bologna

Workshop "Extremes", Hannover
December, 2^{nd} 2021

Outline

1. Introduction
   - Bayesian inference under log-normality assumption
   - Target of the presentation
2. Inferential setting
   - Methodological results
   - Proposed prior specification
3. Results from simulations
4. An example on real data
5. Conclusions and possible extensions

Log-Normality assumption

Let us consider a random sample $(Y_1, \ldots, Y_n)$ for which we assume log-normality:

$$Y_i|\xi, \sigma^2 \overset{indep}{\sim} \log \mathcal{N}\left(\xi, \sigma^2\right), \quad \forall i.$$

Beside estimating the parameters ξ and σ^2, we might be interested on making inference on:

- **Moments:** $M_q = \mathbb{E}\left[Y^q|\xi, \sigma^2\right] = \exp\left\{q\xi + \frac{q^2\sigma^2}{2}\right\}$.
- **Quantiles:** $Q_p = \exp\left\{\xi + \Phi^{-1}(p)\sigma\right\}$. Where $Q_p : \mathbb{P}\left[Y \leq Q_p|\xi, \sigma^2\right] = p$.
- **Predictions:** posterior predictive distribution $\tilde{Y}|\mathbf{y}$.

Warnings in Bayesian inference

- Within the Bayesian framework, prior distributions must be specified for ξ and σ^2.
- To make inference on M_q and Q_p a reasonable prior specification is crucial because of the **exponential transformation**.
- Specific focus on σ^2.

Popular **proper** priors such as:

$$\sigma^2 \sim \text{IG}(a, b), \quad \sigma \sim \text{Half-}t$$

lead to:

$$\mathbb{E}\left[M_q^r|\mathbf{y}\right] = \infty, \ \mathbb{E}\left[Q_p^r|\mathbf{y}\right] = \infty \text{ and } \mathbb{E}\left[\tilde{Y}^r|\mathbf{y}\right] = \infty;$$

i.e. infinite posterior moments.

Some remarks on posterior moments

Some facts about the non-existence of posterior moments:

- Not possible to evaluate posterior mean (i.e. the Bayes estimator under quadratic loss) and posterior variance.
- Quantiles of the posterior distribution remains well-defined.
- The issue is masked when the sample size increases and numerical methods (MCMC algorithms) are used: **possibly misleading results**.
- Issues related to the tail heaviness of the posterior distribution.
- Equivalent issues also when ξ is a generic linear predictor (e.g. $\mathbf{x}_i^T \beta$).

Overview on the topic

Issues tackled in the literature:

- Log-normal mean (Fabrizi and Trivisano, 2012).
- Quantiles (Gardini et al., 2020).
- Conditional means under log-normal mixed model (Gardini et al., 2021).

Focus of the presentation: estimation of quantiles in presence of censored data (multiple right/left censoring).

Practical framework

The inferential problem is of interest in fields such as environmental monitoring and exposure assessment, where:

- Concentration data are used:
  - Usually skewed.
  - Left censoring frequent due to instrumental detection limits.
- Quantiles represents important threshold to check.
- Only small samples are often available.

The sample: some notation

Let us consider the situation in which the n-dimensional sample is partitioned into:

- n_o completely observed units: $Y_1, \ldots Y_{n_o}$.
- n_c censored observations, possibly at different levels. Distinguishing:
  - n_l left censored units: $Y_1^<, \ldots, Y_{n_l}^<$.
  - n_r right censored units: $Y_1^>, \ldots, Y_{n_r}^>$.

The model

For each observation we assume:

$$Y_i|\xi,\sigma^2 \overset{indep}{\sim} \log\mathcal{N}\left(\xi,\sigma^2\right), \quad \forall i.$$

Denoting with y_i the observed values we can write the likelihood as:

$$p(\mathbf{y}|\xi,\sigma^2) = \prod_{i=1}^{n_o}\left[\frac{1}{\sigma y_i\sqrt{2\pi}}\exp\left\{-\frac{(\log y_i-\xi)^2}{2\sigma^2}\right\}\right]\times$$
$$\times\prod_{j=1}^{n_l}\mathbb{P}\left[Y_j\leq y_j^{<}|\xi,\sigma^2\right]$$
$$\times\prod_{k=1}^{n_r}\mathbb{P}\left[Y_k\geq y_k^{>}|\xi,\sigma^2\right]$$

Prior setting

1. Vague prior for the mean in the log-scale $\xi \sim \mathcal{N}(0, V_0^2)$.
2. To specify the prior on σ^2 the following result needs to be considered.

Teorem

$\mathbb{E}\left[Q_p^r|\mathbf{y}\right] < +\infty$ if:

- the prior of σ^2 includes an exponential term $\exp\{-t\sigma^2\}$;
- t must fulfil:

$$t > \frac{r^2}{2n}.$$

Prior for σ^2: the GIG distribution

The flexibility of the three-parameters **GIG distribution** is appealing. If $W \sim GIG(\lambda, \delta, \gamma)$, then the density is:

$$p(w) = \frac{w^{\lambda-1}}{2K_\lambda(\delta\gamma)} \exp\left\{-\left(\frac{\delta^2}{2w} + \frac{\gamma^2 w}{2}\right)\right\}, \quad w \in \mathbb{R}^+.$$

Some interesting features:

- When $\gamma \to 0$ and $\lambda < 0$ the inverse gamma distribution arises.
- When $\delta \to 0$ and $\lambda > 0$ the gamma distribution arises.
- Conjugate prior for the variance in the normal model.

Hyperparameters specification

The existence condition for the r-th moment translates into the condition:

$$\gamma > \frac{r}{\sqrt{n}}.$$

How to fix γ?

- Delicate parameter: it rules the right tail. Higher values of γ leads to lighter tails.
- If r is the order of the moment of interest, we propose to fix:

$$\gamma_0 = \frac{r+1}{\sqrt{n}},$$

to guarantee the stability of the results.

Hyperparameters specification: weakly informative setting

Starting from the fact that the posterior of σ^2 remains a GIG distribution, we can consider an approximation of its posterior expectation:

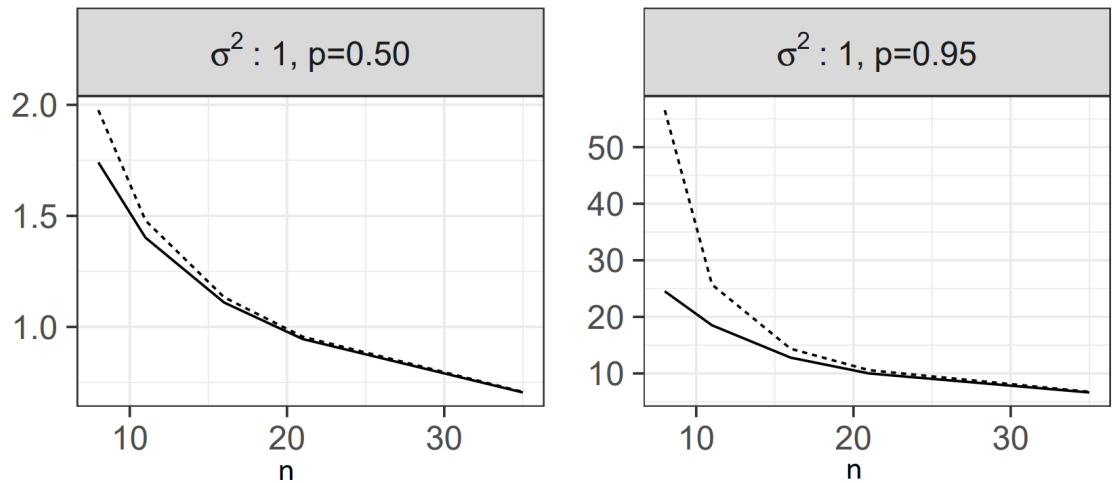
$$\mathbb{E}\left[\sigma^2|\mathbf{y}\right] \cong \frac{\lambda + \sqrt{(\lambda - n/2)^2 + (nV^2 + \delta^2)\,\gamma^2}}{\gamma^2}$$
$$\cong \frac{(nV^2 + \delta^2)}{-2\lambda + n - 1}.$$

The prior distribution $\sigma^2 \sim GIG(\lambda = 0, \delta = 0.01, \gamma = \gamma_0)$ leads to:

$$\mathbb{E}\left[\sigma^2|\mathbf{y}\right] \cong V^2.$$

Simulation study: synthesis of results

Aims of the simulation: compare the **average width** of the credible intervals under GIG priors (——) to a ML-based method (- - - -).



The coverage levels are similar, reaching the nominal one.

Simulation study: synthesis of results

Aspects to further investigate:

- Behaviour under different proportions of censored values.
- Comparison with other priors for σ^2.

Real data example: groundwater monitoring

Data: Total Organic Carbon (TOC) in a background monitoring well (from Gibbons et al., 2009). 3 values over $n = 10$ below the detection limit:

$$5,\ 7,\ <1,\ 3,\ <1,\ 4,\ 6,\ 5,\ <1,\ 6.$$

Goal of the analysis: estimating quantiles in the right tail (e.g. $p = 0.9$) to evaluate water contamination.

Aim of the example: show the central role of the prior distribution on σ^2 in making inference on Q_p with small samples. Compared:

- Frequentist procedures.
- Bayesian procedure with $\sigma^2 \sim GIG(0, 0.01, \gamma_0)$.
- Bayesian procedure with $\sigma^2 \sim IG(0.1, 0.1)$.

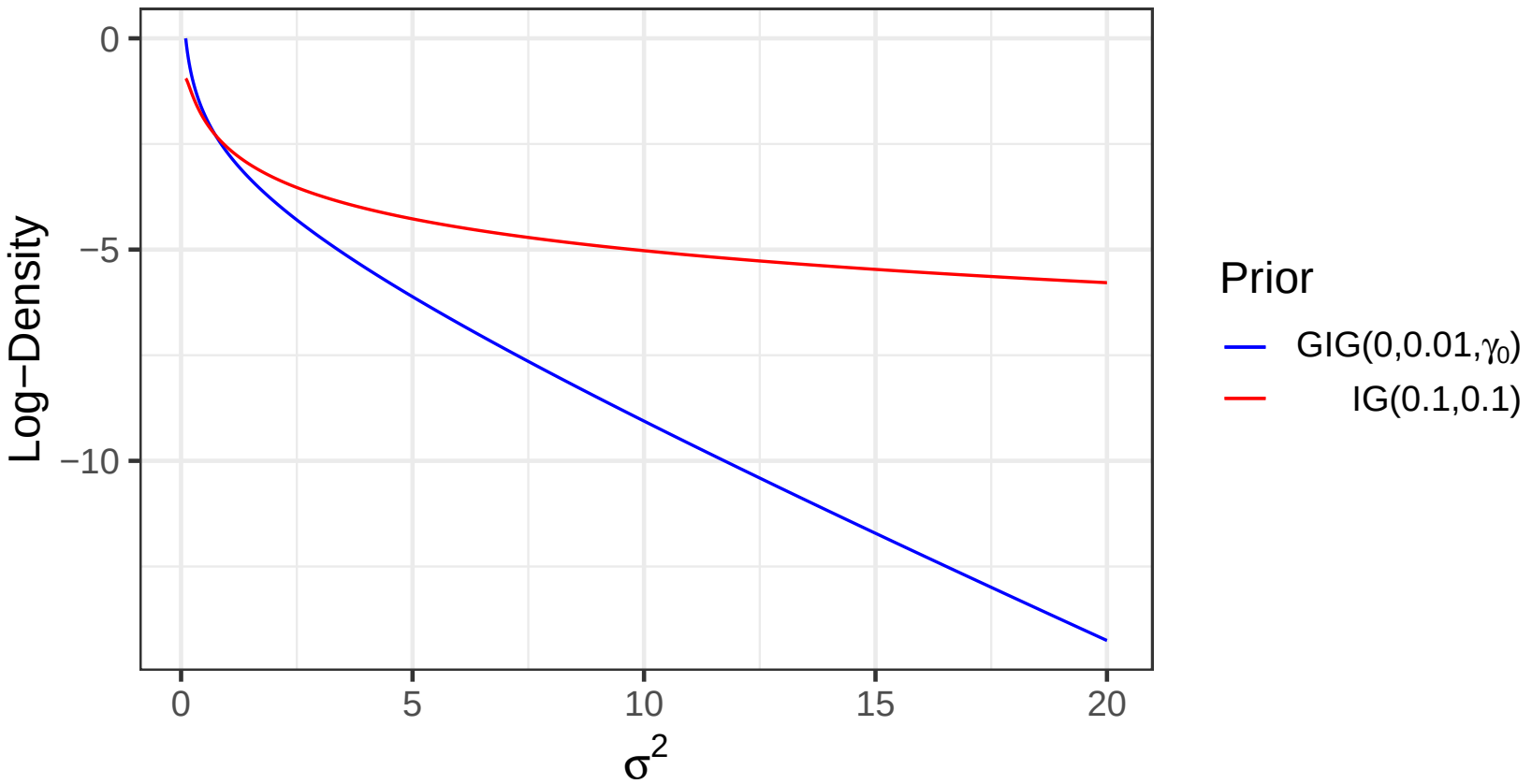
Real data example: practical aspects

- **Hyperparameter specification.** We need to determine γ_0: fixing $r = 2$ and having $n = 10$:

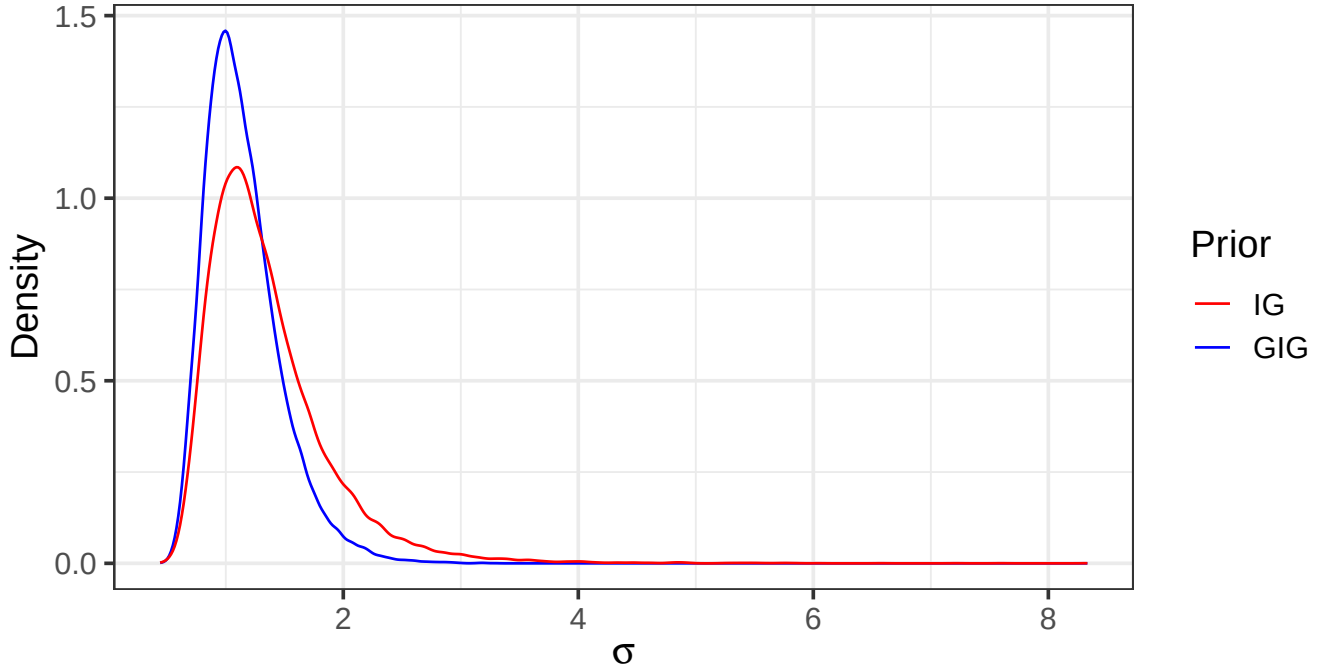
$$\gamma_0 = 0.95.$$

- **Obtaining the posterior of Q_p.** With censored data no closed form results. MCMC samples from the posterior of ξ and σ^2 required. We propose to use `Stan`:
  - Probabilistic language easy to use with `R` interface.
  - Flexibility in the likelihood declaration.
  - Possible to add non-standard distributions (GIG non included by default).

Real data example: prior of σ^2



Real data example: inference on σ



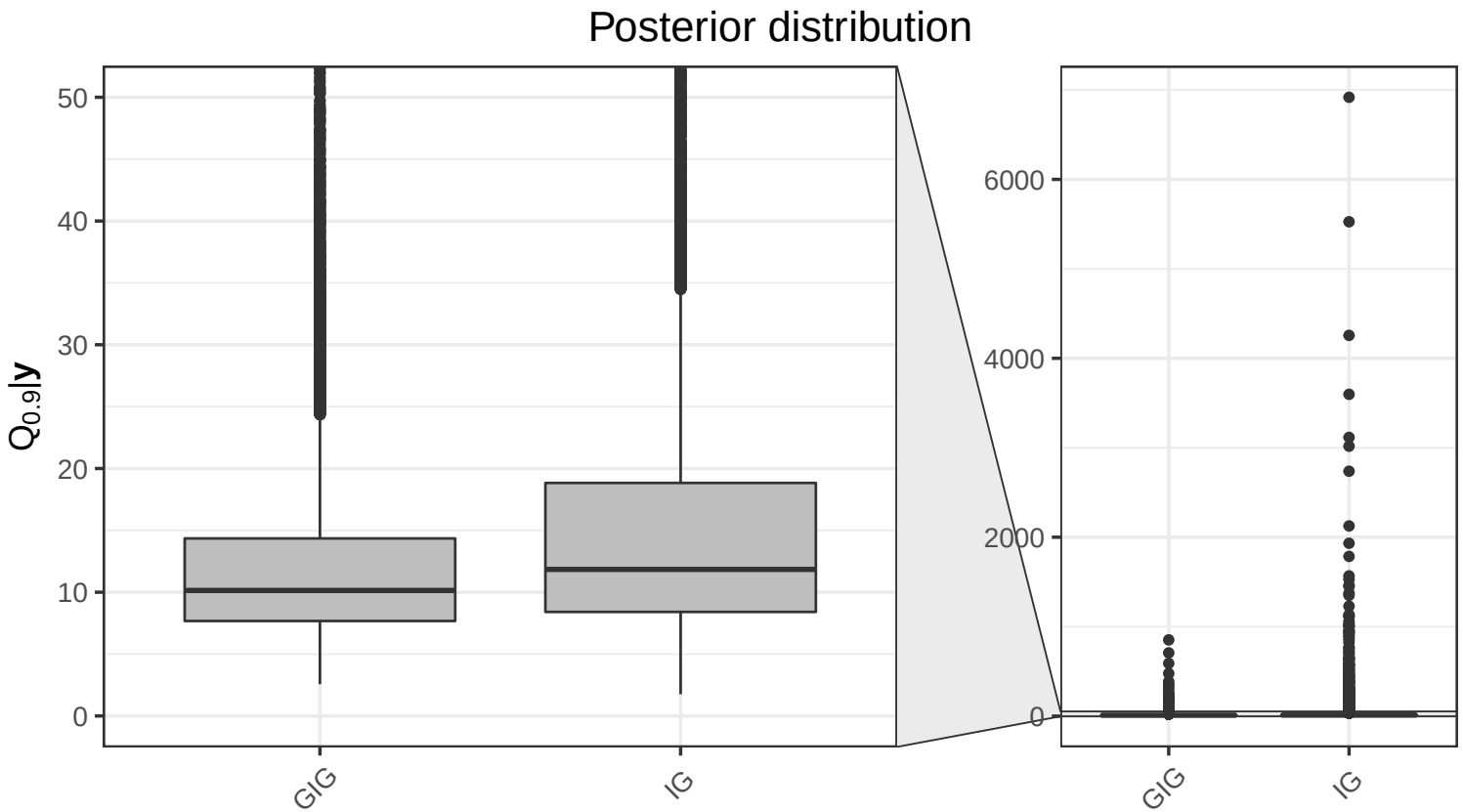
Bayesian estimates

Prior	Mean	Median	S.D.
IG	2.10	1.54	2.11
GIG	1.42	1.20	0.87

Frequentist estimates

Method	Estimate
ML	1.06

Real data example: posterior of Q_p



Real data example: posterior of Q_p II

Bayesian estimates

Prior	Mean	Median	S.D.	95% C.I.
IG	21.36 [∞]	11.84	85.87 [∞]	[5.13;78.23]
GIG	12.86	10.14	13.37	[4.93;35.63]

Frequentist estimates

Method	Estimate	95% C.I.
ML	10.04	[4.96;42.62]

Concluding remarks and possible developments

- Inference under the log-normality assumption should be carried out carefully:
  - Possible strong impact of the priors on the posterior results.
  - Issues with the existence of posterior moments.
- **Focus on quantile estimation with censored data:**
  - Extensive simulation study to investigate the impact of the proportion of censored data.
  - Real data problems (also with underlying more complex models).

Concluding remarks and possible developments

- **Extend the functions in the `BayesLN` package to deal with censored data.**
- **Other interesting inferential problems:** spatial log-normal process affected by the same issues.
  - Used to produce both predictions and estimates of quantiles.
  - Presence of correlation parameters: investigate their possible involvement in the existence conditions.

References

Fabrizi, E. and Trivisano, C. (2012). Bayesian estimation of log-normal means with finite quadratic expected loss. *Bayesian Analysis*, 7(4):975–996.

Gardini, A., Trivisano, C., and Fabrizi, E. (2020). Bayesian inference for quantiles of the log-normal distribution. *Biometrical Journal*, 62(8):1997–2012.

Gardini, A., Trivisano, C., and Fabrizi, E. (2021). Bayesian analysis of anova and mixed models on the log-transformed response variable. *Psychometrika*.

Gibbons, R. D., Bhaumik, D. K., and Aryal, S. (2009). *Statistical methods for groundwater monitoring*, volume 59. John Wiley & Sons.