Bayesian model selection method and its application in modeling Covid-19 pandemic development

Truong-Vinh Hoang

MATH4UQ, RWTH Aachen
July 23, 2020

- Presenting Bayesian approach for model selection
  reference: Chapter 5, Ando, Tomohiro. *Bayesian model selection and statistical modeling*. CRC Press, 2010.
- Apply for modeling the Covid-19 evolution

Outline

General framework

Posterior, marginal likelihood

Setting: Select a model from a set of candidate models: $M_1, \ldots, M_r$

- Model M_k is characterized by probability density $f_k(\boldsymbol{x}|\boldsymbol{\theta_k})$ where $\boldsymbol{\theta}_k$ is a p_k-dimensional vector of unknown parameter
- $\pi_k(\boldsymbol{\theta}_k)$ is the prior distribution for parameter vector $\boldsymbol{\theta}_k$
- Posterior probability of the model M_k given data set $\boldsymbol{X}_n = \{x_1, \ldots, x_n\}$ of i.i.d samples

$$P\left(M_k|\boldsymbol{X}_n\right) = \frac{P\left(M_k\right)\int f_k\left(\boldsymbol{X}_n|\boldsymbol{\theta}_k\right)\pi_k\left(\boldsymbol{\theta}_k\right)d\boldsymbol{\theta}_k}{\sum_{\alpha=1}^{r} P\left(M_\alpha\right)\int f_\alpha\left(\boldsymbol{X}_n|\boldsymbol{\theta}_\alpha\right)\pi_\alpha\left(\boldsymbol{\theta}_\alpha\right)d\boldsymbol{\theta}_\alpha}$$

- $P\left(M_k\right)$ is the prior: e.g.
  - Uniform
  - Poisson prior: $P(M_k) \propto \lambda^{p_k}\exp(-\lambda), \quad \lambda > 0 \rightarrow$ higher probability on simpler models
  - $P(M_k) \propto \prod_{j=1}^{p} \pi_j^{\gamma_{k,j}}(1-\pi_j)^{1-\gamma_{k,j}}$, where π_j are the prior probability that the $j-$th predictor is included in the model
- $\int f_k\left(\boldsymbol{X}_n|\boldsymbol{\theta}_k\right)\pi_k\left(\boldsymbol{\theta}_k\right)d\boldsymbol{\theta}_k$ is the **marginal likelihood**.

General framework

Bayes factor

Bayes factor (BF): comparing models, for testing hypotheses.

$$\text{BF } (M_k, M_j) \equiv \frac{P(\boldsymbol{X}_n|M_k)}{P(\boldsymbol{X}_n|M_j)}$$

Measures the evidence for M_k vs. M_j based on the data information [posterior odds = Bayes factor $\times$ the prior odds]

$$\frac{P(M_k|\boldsymbol{X}_n)}{P(M_j|\boldsymbol{X}_n)} = \frac{P(\boldsymbol{X}_n|M_k)}{P(\boldsymbol{X}_n|M_j)} \times \frac{P(M_k)}{P(M_j)}$$

Bayes factor can be reduced to classical likelihood ratio

1. Introduction

2. Theoretical background
   - General framework
   - Evaluation of marginal likelihood
     - Analytical method: Exact calculation of marginal likelihood
     - Laplace's method for computing the marginal likelihood
     - Bayesian information criterion (BIC)
   - Expected predictive likelihood approach

3. Examples with Covid19 data

Methods to evaluate marginal likelihood $f(\boldsymbol{X}_n|\boldsymbol{\theta}_k, M_k)$

- Analytical method
  - E.g. Binomial model with conjugate prior
- Laplace's approximation
  - Bayesian information criterion (BIC)
  - Bayesian information criterion vs. Akaike's Information criterion
  - Generalized BIC
- Numerical method

Exact calculation of marginal likelihood

Binomial model with conjugate prior

- Observations: n independent samples
  $\boldsymbol{X}_n = \{x_1, x_2, \ldots, x_n\} \sim \text{Bin}(n, p)$ with unknown p,
- Conjugate prior: beta distribution with parameter α, β for p

Exact calculation of marginal likelihood

Binomial model with conjugate prior

- Observations: n independent samples
  $\boldsymbol{X}_n = \{x_1, x_2, \ldots, x_n\} \sim \text{Bin}(n, p)$ with unknown p,
- Conjugate prior: beta distribution with parameter α, β for p
- Let $y_n = \sum_{i=1}^{n} x_i$, the marginal likelihood:

$$
\begin{aligned}
& f\left(\boldsymbol{X}_n | \, \alpha, \beta\right) = \\
& \int_0^1 \left[\begin{pmatrix} n \\ y_n \end{pmatrix} p^{y_n} (1-p)^{n-y_n} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1} \right] dp \\
= & \begin{pmatrix} n \\ y_n \end{pmatrix} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma\left(y_n+\alpha\right)\Gamma\left(n+\beta-y_n\right)}{\Gamma(n+\alpha+\beta)}
\end{aligned}
$$

Exact calculation of marginal likelihood

Binomial model with conjugate prior

- Observations: n independent samples $\boldsymbol{X}_n = \{x_1, x_2, \ldots, x_n\} \sim \text{Bin}(n, p)$ with unknown p,
- Conjugate prior: beta distribution with parameter α, β for p
- Let $y_n = \sum_{i=1}^{n} x_i$, the marginal likelihood:

$$f(\boldsymbol{X}_n | \alpha, \beta) = \int_0^1 \left[\binom{n}{y_n} p^{y_n}(1-p)^{n-y_n} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \right] dp$$

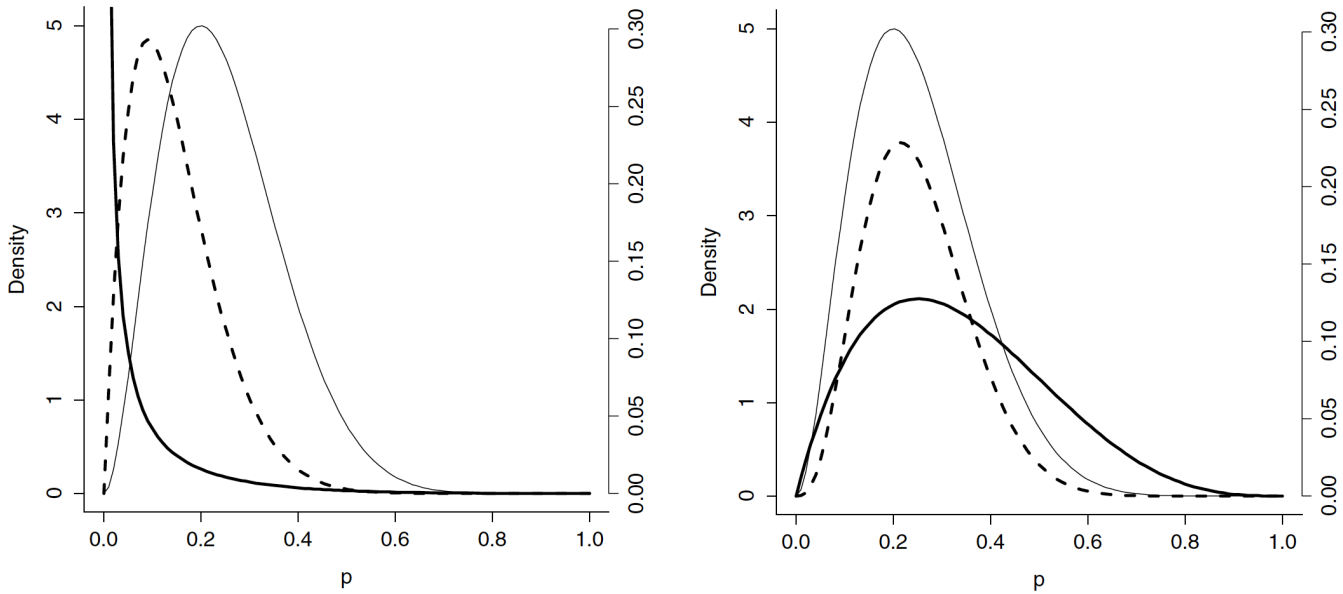
$$= \binom{n}{y_n} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y_n+\alpha)\Gamma(n+\beta-y_n)}{\Gamma(n+\alpha+\beta)}$$

- Example: Observation: $n = 10$, $y_n = 2$
- Model selection: $M_1 : p \sim \text{Beta}(0.1, 4)$; $M_2 : p \sim \text{Beta}(2, 4)$; $M_3 : p \sim \text{Beta}(4, 4)$; $M_4 : p \sim \text{Beta}(8, 4)$

Exact calculation of marginal likelihood

Binomial model with conjugate prior

Models: $M_1 : p \sim \text{Beta}(0.1, 4)$; $M_2 : p \sim \text{Beta}(2, 4)$; $M_3 : p \sim \text{Beta}(4, 4)$; $M_4 : p \sim \text{Beta}(8, 4)$



left M_1, right M_2

likelihood function (thin line), prior (thick line), posterior (dash line)

$f(\boldsymbol{X}_{10}|M_1) = 0.0277$, $f(\boldsymbol{X}_{10}|M_2) = 0.1648$, $\rightarrow$ M_2 has the best BF

$f(\boldsymbol{X}_{10}|M_3) = 0.0848$, $f(\boldsymbol{X}_{10}|M_4) = 0.0168$.

Laplace's method

Asymptotic evaluation of the marginal likelihood

Assumption: the posterior density $\pi(\boldsymbol{\theta}|\boldsymbol{X}_n)$ is sufficiently well-behaved (highly peaked at the posterior mode $\widehat{\boldsymbol{\theta}}_n$). Let

$$s_n(\boldsymbol{\theta}) = \log\left\{f\left(\boldsymbol{X}_n|\boldsymbol{\theta}\right)\pi(\boldsymbol{\theta})\right\}$$

Taylor series expansion of s_n yields (first order derivative at $\widehat{\boldsymbol{\theta}}_n = 0$):

Laplace's method

Asymptotic evaluation of the marginal likelihood

Assumption: the posterior density $\pi(\boldsymbol{\theta}|\boldsymbol{X}_n)$ is sufficiently well-behaved (highly peaked at the posterior mode $\widehat{\boldsymbol{\theta}}_n$). Let

$$s_n(\boldsymbol{\theta}) = \log\left\{f\left(\boldsymbol{X}_n|\boldsymbol{\theta}\right)\pi(\boldsymbol{\theta})\right\}$$

Taylor series expansion of s_n yields (first order derivative at $\widehat{\boldsymbol{\theta}}_n = 0$):

$$s_n(\boldsymbol{\theta}) \approx s_n\left(\hat{\boldsymbol{\theta}}_n\right) - \frac{n}{2}\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n\right)^T S_n\left(\widehat{\boldsymbol{\theta}}_n\right)\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n\right)$$

where $S_n\left(\widehat{\boldsymbol{\theta}}_n\right) = -\frac{1}{n}\frac{\partial^2 \log\{f(\boldsymbol{X}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n}$

$$\exp\left\{s_n(\boldsymbol{\theta})\right\} \approx \exp\left\{s_n\left(\widehat{\boldsymbol{\theta}}_n\right)\right\} \times \exp\left\{-\frac{n}{2}\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n\right)^T S_n\left(\widehat{\boldsymbol{\theta}}_n\right)\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n\right)\right\}$$

$$\propto \mathcal{N}\left(\widehat{\boldsymbol{\theta}}_n, n^{-1}S_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right)$$

Laplace's method

Asymptotic evaluation of the marginal likelihood

Assumption: the posterior density $\pi(\boldsymbol{\theta}|\boldsymbol{X}_n)$ is sufficiently well-behaved (highly peaked at the posterior mode $\widehat{\boldsymbol{\theta}}_n$). Let

$$s_n(\boldsymbol{\theta}) = \log\left\{f\left(\boldsymbol{X}_n|\boldsymbol{\theta}\right)\pi(\boldsymbol{\theta})\right\}$$

Taylor series expansion of s_n yields (first order derivative at $\widehat{\boldsymbol{\theta}}_n = 0$):

$$s_n(\boldsymbol{\theta}) \approx s_n\left(\hat{\boldsymbol{\theta}}_n\right) - \frac{n}{2}\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n\right)^T S_n\left(\widehat{\boldsymbol{\theta}}_n\right)\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n\right)$$

where $S_n\left(\widehat{\boldsymbol{\theta}}_n\right) = -\frac{1}{n}\frac{\partial^2 \log\{f(\boldsymbol{X}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n}$

$$\exp\left\{s_n(\boldsymbol{\theta})\right\} \approx \exp\left\{s_n\left(\widehat{\boldsymbol{\theta}}_n\right)\right\} \times \exp\left\{-\frac{n}{2}\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n\right)^T S_n\left(\widehat{\boldsymbol{\theta}}_n\right)\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n\right)\right\}$$

$$\propto \mathcal{N}\left(\widehat{\boldsymbol{\theta}}_n, n^{-1}S_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right)$$

Laplace's method for computing the marginal likelihood

Marginal likelihood

- Marginal likelihood

$$P(\boldsymbol{X}_n|M) = \int \exp\{s_n(\boldsymbol{\theta})\}\, d\boldsymbol{\theta}$$

$$\approx \exp\left\{s_n\left(\widehat{\boldsymbol{\theta}}_n\right)\right\} \times \int \exp\left\{-\frac{n}{2}\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_n\right)^T S_n\left(\widehat{\boldsymbol{\theta}}_n\right)\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_n\right)\right\} d\boldsymbol{\theta}$$

$$= f\left(\boldsymbol{X}_n|\widehat{\boldsymbol{\theta}}_n\right)\pi\left(\widehat{\boldsymbol{\theta}}_n\right) \times \frac{(2\pi)^{\frac{p}{2}}}{n^{\frac{p}{2}}\left|S_n\left(\widehat{\boldsymbol{\theta}}_n\right)\right|^{\frac{1}{2}}}$$

where p is the dimension of vector $\boldsymbol{\theta}$

- $\mathrm{BF}(M_k, M_j) \approx \frac{f_k\left(\boldsymbol{X}_n|\widehat{\boldsymbol{\theta}}_{kn}\right)}{f_j\left(\boldsymbol{X}_n|\widehat{\boldsymbol{\theta}}_{jn}\right)} \times \frac{\pi_k\left(\widehat{\boldsymbol{\theta}}_{kn}\right)}{\pi_j\left(\widehat{\boldsymbol{\theta}}_{jn}\right)} \times \frac{\left|S_{jn}\left(\widehat{\boldsymbol{\theta}}_{jn}\right)\right|^{\frac{1}{2}}}{\left|S_{kn}\left(\widehat{\boldsymbol{\theta}}_{kn}\right)\right|^{\frac{1}{2}}} \times \left(\frac{2\pi}{n}\right)^{\frac{p_k-p_j}{2}}$$

Laplace's method for computing the marginal likelihood

Marginal likelihood

- Marginal likelihood

$$P(\boldsymbol{X}_n|M) = \int \exp\{s_n(\boldsymbol{\theta})\}\, d\boldsymbol{\theta}$$

$$\approx \exp\left\{s_n\left(\widehat{\boldsymbol{\theta}}_n\right)\right\} \times \int \exp\left\{-\frac{n}{2}\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_n\right)^T S_n\left(\widehat{\boldsymbol{\theta}}_n\right)\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_n\right)\right\} d\boldsymbol{\theta}$$

$$= f\left(\boldsymbol{X}_n|\widehat{\boldsymbol{\theta}}_n\right)\pi\left(\widehat{\boldsymbol{\theta}}_n\right) \times \frac{(2\pi)^{\frac{p}{2}}}{n^{\frac{p}{2}}\left|S_n\left(\widehat{\boldsymbol{\theta}}_n\right)\right|^{\frac{1}{2}}}$$

where p is the dimension of vector $\boldsymbol{\theta}$

- $\mathrm{BF}(M_k, M_j) \approx \frac{f_k\left(\boldsymbol{X}_n|\widehat{\boldsymbol{\theta}}_{kn}\right)}{f_j\left(\boldsymbol{X}_n|\widehat{\boldsymbol{\theta}}_{jn}\right)} \times \frac{\pi_k\left(\widehat{\boldsymbol{\theta}}_{kn}\right)}{\pi_j\left(\widehat{\boldsymbol{\theta}}_{jn}\right)} \times \frac{\left|S_{jn}\left(\widehat{\boldsymbol{\theta}}_{jn}\right)\right|^{\frac{1}{2}}}{\left|S_{kn}\left(\widehat{\boldsymbol{\theta}}_{kn}\right)\right|^{\frac{1}{2}}} \times \left(\frac{2\pi}{n}\right)^{\frac{p_k-p_j}{2}}$$
- Deterministic approaches: first two terms
- Third term: prefer model with less sensitivity
- Last term: complexity of the model (p_k)

Laplace's method for computing the marginal likelihood

Bayesian information criterion

We will consider two cases:

- $\log \pi(\boldsymbol{\theta}) = O_p(1)$, $n \gg 1 \rightarrow$ prior information is ignored
- $\log \pi(\boldsymbol{\theta}) = O_p(n)$

O_p: stochastic boundedness:
if $X_n = O_p(a_n)$ For any $\epsilon > 0$, there exists a finite $m > 0$ and a finite $N > 0$ such that, $P(|X_n/a_n| > m) < \varepsilon$, $\forall\ n > N$.

Bayesian information criterion

Assumption: $\log \pi(\boldsymbol{\theta}) = O_p(1)$, $n \gg 1 \rightarrow$ prior information is ignored
The marginal likelihood is approximated as:

$$P\left(\boldsymbol{X}_n | M\right) \approx f\left(\boldsymbol{X}_n | \widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}\right) \pi\left(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}\right) \times \frac{(2\pi)^{\frac{p}{2}}}{n^{\frac{p}{2}} \left| J_n\left(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}\right) \right|^{\frac{1}{2}}}$$

J_n is minus the Hessian matrix of function $\frac{1}{n} \log f(\boldsymbol{X}_n | \widehat{\boldsymbol{\theta}})$ at $\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}$

Bayesian information criterion

Assumption: $\log \pi(\boldsymbol{\theta}) = O_p(1)$, $n \gg 1 \rightarrow$ prior information is ignored
The marginal likelihood is approximated as:

$$P\left(\boldsymbol{X}_n|M\right) \approx f\left(\boldsymbol{X}_n|\widehat{\boldsymbol{\theta}}_{\text{MLE}}\right) \pi\left(\widehat{\boldsymbol{\theta}}_{\text{MLE}}\right) \times \frac{(2\pi)^{\frac{p}{2}}}{n^{\frac{p}{2}} \left|J_n\left(\widehat{\boldsymbol{\theta}}_{\text{MLE}}\right)\right|^{\frac{1}{2}}}$$

J_n is minus the Hessian matrix of function $\frac{1}{n} \log f(\boldsymbol{X}_n|\widehat{\boldsymbol{\theta}})$ at $\widehat{\boldsymbol{\theta}}_{\text{MLE}}$
Derive $-2\log(P\left(\boldsymbol{X}_n|M\right)$, ignore constant terms when $n \gg 1 \rightarrow$ Schwarz's BIC

$$\text{BIC} = -2 \log f\left(\boldsymbol{X}_n|\widehat{\boldsymbol{\theta}}_{\text{MLE}}\right) + p \log n$$

BIC gives a rough approximation to log(BF)

$$\log(\text{BF}(M_k, M_j)) = \log\left[P\left(\boldsymbol{X}_n|M_k\right)\right] - \left[P\left(\boldsymbol{X}_n|M_j\right)\right] \approx (\text{BIC}_j - \text{BIC}_k)/2$$

Bayesian information criterion

Assumption: $\log \pi(\boldsymbol{\theta}) = O_p(1)$, $n \gg 1 \rightarrow$ prior information is ignored
The marginal likelihood is approximated as:

$$P\left(\boldsymbol{X}_n|M\right) \approx f\left(\boldsymbol{X}_n|\widehat{\boldsymbol{\theta}}_{\text{MLE}}\right)\pi\left(\widehat{\boldsymbol{\theta}}_{\text{MLE}}\right) \times \frac{(2\pi)^{\frac{p}{2}}}{n^{\frac{p}{2}}\left|J_n\left(\widehat{\boldsymbol{\theta}}_{\text{MLE}}\right)\right|^{\frac{1}{2}}}$$

J_n is minus the Hessian matrix of function $\frac{1}{n}\log f(\boldsymbol{X}_n|\widehat{\boldsymbol{\theta}})$ at $\widehat{\boldsymbol{\theta}}_{\text{MLE}}$
Derive $-2\log(P\left(\boldsymbol{X}_n|M\right)$, ignore constant terms when $n \gg 1 \rightarrow$ Schwarz's BIC

$$\text{BIC} = -2\log f\left(\boldsymbol{X}_n|\widehat{\boldsymbol{\theta}}_{\text{MLE}}\right) + p\log n$$

BIC gives a rough approximation to log(BF)

$$\log(\text{BF}(M_k, M_j)) = \log\left[P\left(\boldsymbol{X}_n|M_k\right)\right] - \left[P\left(\boldsymbol{X}_n|M_j\right)\right] \approx (\text{BIC}_j - \text{BIC}_k)/2$$

$$\lim_{n\to\infty} \frac{(\text{BIC}_j - \text{BIC}_k)/2 - \log(\text{BF}(M_k, M_j)}{\log(\text{BF}(M_k, M_j)} = 0$$

BIC does not involve J_n

BIC vs Akaike information criterion (AIC)

Model selection by minimizing ICs: $IC_k = -2 \log f_k(\boldsymbol{X}_n | \widehat{\theta}_{\text{MLE } k}) + c_{n,p_k}$

$c_{n,p}$: penalty to encourage the selection of a parsimonious model (fewer parameter dimension)

- BIC: $c_{n,p} = p \log(n)$, AIC: $c_{n,p} = 2p$

[1] discussed later in Expected predicitve likelihood approach

BIC vs Akaike information criterion (AIC)

Model selection by minimizing ICs: $IC_k = -2 \log f_k(\boldsymbol{X}_n | \widehat{\boldsymbol{\theta}}_{\text{MLE } k}) + c_{n,p_k}$

$c_{n,p}$: penalty to encourage the selection of a parsimonious model (fewer parameter dimension)

- BIC: $c_{n,p} = p \log(n)$, AIC: $c_{n,p} = 2p$

The best model $f_0(z|\boldsymbol{\theta}_0)$ that has the lowest Kullback-Leibler divergence [1] from true model $G(z)$, or equivalently maximum expected log-likelihood:

$$\int \log f_0(z|\boldsymbol{\theta}_0) dG(z) = \max_k \left\{ \sup_{\theta_k} \int \log f_k(z|\boldsymbol{\theta}_k) dG(z) \right\}$$

[1] discussed later in Expected predicitve likelihood approach

BIC vs Akaike information criterion (AIC)

Model selection by minimizing ICs: $IC_k = -2 \log f_k(\boldsymbol{X}_n | \widehat{\boldsymbol{\theta}}_{\text{MLE } k}) + c_{n,p_k}$

$c_{n,p}$: penalty to encourage the selection of a parsimonious model (fewer parameter dimension)

- BIC: $c_{n,p} = p \log(n)$, AIC: $c_{n,p} = 2p$

The best model $f_0(z|\boldsymbol{\theta}_0)$ that has the lowest Kullback-Leibler divergence [1] from true model $G(z)$, or equivalently maximum expected log-likelihood:

$$\int \log f_0(z|\boldsymbol{\theta}_0) dG(z) = \max_k \left\{ \sup_{\theta_k} \int \log f_k(z|\boldsymbol{\theta}_k) dG(z) \right\}$$

Consistency in picking model $\log f_0(z|\boldsymbol{\theta}_0)$ requires for all non-optimal f_k:

$$\liminf_{n\to\infty} \left[\frac{1}{n} \int \log f_0(\boldsymbol{X}_n|\boldsymbol{\theta}_0) dG(\boldsymbol{X}_n) - \frac{1}{n} \int \log f_k(\boldsymbol{X}_n|\boldsymbol{\theta}_{k0}) dG(\boldsymbol{X}_n) \right] > 0$$

[1] discussed later in Expected predicitve likelihood approach

BIC vs Akaike information criterion (AIC)

Model selection by minimizing ICs: $IC_k = -2\log f_k(\boldsymbol{X}_n|\widehat{\theta}_{\text{MLE } k}) + c_{n,p_k}$
$c_{n,p}$: penalty to encourage the selection of a parsimonious model (fewer parameter dimension)

- BIC: $c_{n,p} = p\log(n)$, AIC: $c_{n,p} = 2p$

The best model $f_0(z|\boldsymbol{\theta}_0)$ that has the lowest Kullback-Leibler divergence [1] from true model $G(z)$, or equivalently maximum expected log-likelihood:

$$\int \log f_0(z|\boldsymbol{\theta}_0)dG(z) = \max_k \left\{ \sup_{\theta_k} \int \log f_k(z|\boldsymbol{\theta}_k)dG(z) \right\}$$

Consistency in picking model $\log f_0(z|\boldsymbol{\theta}_0)$ requires for all non-optimal f_k:

$$\lim\inf_{n\to\infty} \left[\frac{1}{n}\int \log f_0(\boldsymbol{X}_n|\boldsymbol{\theta}_0)dG(\boldsymbol{X}_n) - \frac{1}{n}\int \log f_k(\boldsymbol{X}_n|\boldsymbol{\theta}_{k0})dG(\boldsymbol{X}_n) \right] > 0$$

if $f_0(z|\boldsymbol{\theta}_0)$ is unique, the consistency of selection holds;
if $f_0(z|\boldsymbol{\theta}_0)$ is not-unique (e.g f_k, f_j), only BIC satisfies the consistency in picking the simplest model: $\log\left(\frac{f_k(\boldsymbol{X}_n|\boldsymbol{\theta}_{k0})}{f_j(\boldsymbol{X}_n|\boldsymbol{\theta}_{j0})}\right) = O_p(1)$, and $c_{n,p} \to \infty$

[1] discussed later in Expected predicitve likelihood approach

Generalized Bayesian information criterion

Consider $\log \pi(\boldsymbol{\theta}) = O_p(n)$. The marginal likelihood is approximated as:

$$P\left(\boldsymbol{X}_n|M\right) \approx f\left(\boldsymbol{X}_n|\widehat{\boldsymbol{\theta}}_n\right)\pi\left(\widehat{\boldsymbol{\theta}}_n\right) \times \frac{(2\pi)^{\frac{p}{2}}}{n^{\frac{p}{2}}\left|S_n\left(\widehat{\boldsymbol{\theta}}_n\right)\right|^{\frac{1}{2}}}$$

$S_n(\widehat{\boldsymbol{\theta}})$ is minus the Hessian matrix of function $\frac{1}{n}\log f(\boldsymbol{X}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ at the mode of posterior distribution $\widehat{\boldsymbol{\theta}}_n$
Generalized Bayesian information criterion:

$$\text{GBIC} = -2\log f\left(\boldsymbol{X}_n|\widehat{\boldsymbol{\theta}}\right) - 2\log\pi(\widehat{\boldsymbol{\theta}}) + p\log n + \log|S_n(\widehat{\boldsymbol{\theta}})| - p\log(2\pi)$$

Choosing model with the largest posterior $\equiv$ the model that minimizes the criterion.
Difficulty: require to compute the Hessian matrix

1. Introduction

2. Theoretical background
   - General framework
   - Evaluation of marginal likelihood
     - Analytical method: Exact calculation of marginal likelihood
     - Laplace's method for computing the marginal likelihood
     - Bayesian information criterion (BIC)
   - Expected predictive likelihood approach

3. Examples with Covid19 data

Expected predictive likelihood approach

Motivation: the main weakness of BF is its sensitivity to the prior
Let g be the true density of observations, $\boldsymbol{Z}_n = \{z_1, \ldots, z_n\}$ is a set of i.i.d. unseen future observations
$f(\boldsymbol{Z}_n|\boldsymbol{X}_n, M)$: Bayesian predictive distribution

$$f(\boldsymbol{Z}_n|\boldsymbol{X}_n, M) = \int f(\boldsymbol{Z}_n|\boldsymbol{X}_n, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{X}_n)d\boldsymbol{\theta}$$

Predictive ability of a given model M using Kullback-Leibler divergence

$$\begin{aligned} KL(g||f) &= \int \left[\log \frac{g(\boldsymbol{Z}_n)}{f(\boldsymbol{Z}_n|\boldsymbol{X}_n, M)}\right] g(\boldsymbol{Z}_n)d\boldsymbol{Z}_n \\ &= \int \log g(\boldsymbol{Z}_n) g(\boldsymbol{Z}_n)d\boldsymbol{Z}_n - \int \log\left(f(\boldsymbol{Z}_n|\boldsymbol{X}_n, M)\right) g(\boldsymbol{Z}_n)d\boldsymbol{Z}_n \end{aligned}$$

Expected predictive likelihood approach

Motivation: the main weakness of BF is its sensitivity to the prior
Let g be the true density of observations, $\boldsymbol{Z}_n = \{z_1, \ldots, z_n\}$ is a set of i.i.d. unseen future observations
$f(\boldsymbol{Z}_n|\boldsymbol{X}_n, M)$: Bayesian predictive distribution

$$f(\boldsymbol{Z}_n|\boldsymbol{X}_n, M) = \int f(\boldsymbol{Z}_n|\boldsymbol{X}_n, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{X}_n)d\boldsymbol{\theta}$$

Predictive ability of a given model M using Kullback-Leibler divergence

$$\begin{aligned} KL(g||f) &= \int \left[\log \frac{g(\boldsymbol{Z}_n)}{f(\boldsymbol{Z}_n|\boldsymbol{X}_n, M)}\right] g(\boldsymbol{Z}_n)d\boldsymbol{Z}_n \\ &= \int \log g(\boldsymbol{Z}_n)g(\boldsymbol{Z}_n)d\boldsymbol{Z}_n - \int \log\left(f(\boldsymbol{Z}_n|\boldsymbol{X}_n, M)\right) g(\boldsymbol{Z}_n)d\boldsymbol{Z}_n \end{aligned}$$

Expected log-predictive likelihood $\rightarrow$ a general approach for evaluating the goodness of fit (validation step)

$$\eta(M) = \int \log f(\boldsymbol{Z}_n|\boldsymbol{X}_n, M)g(\boldsymbol{Z}_n)d\boldsymbol{Z}_n$$

1. Introduction

2. Theoretical background
   - General framework
   - Evaluation of marginal likelihood
     - Analytical method: Exact calculation of marginal likelihood
     - Laplace's method for computing the marginal likelihood
     - Bayesian information criterion (BIC)
   - Expected predictive likelihood approach

3. Examples with Covid19 data

Problem setting

We will compare SIR models with different setting in modeling of Covid-19 pandemic evolution in Germany, Italy, Uruguay, and Saudi Arabia.
Model selection is performed in three steps

- Introduction of SIR models
- Parameter identification using Bayesian approach
- Evaluation of model marginal likelihood and Bayes factor

Modeling of Covid-19 pandemic evolution

SIR models

- **SIR model**

$$\begin{cases} \dfrac{dS}{dt} = -\beta \dfrac{IS}{N} \\ \dfrac{dI}{dt} = \beta \dfrac{IS}{N} - \gamma I \\ \dfrac{dR}{dt} = \gamma I \end{cases}$$

- S, I, R and N refer to the susceptible, infected, recovered and total populations, respectively
- $\boldsymbol{\theta} = \{\beta, \gamma, S_0, I_0\}$ are parameter to be identified

Modeling of Covid-19 pandemic evolution

SIR models

- **SIR model**

$$\begin{cases} \dfrac{dS}{dt} = -\beta \dfrac{IS}{N} \\ \dfrac{dI}{dt} = \beta \dfrac{IS}{N} - \gamma I \\ \dfrac{dR}{dt} = \gamma I \end{cases}$$

- S, I, R and N refer to the susceptible, infected, recovered and total populations, respectively
- $\boldsymbol{\theta} = \{\beta, \gamma, S_0, I_0\}$ are parameter to be identified

We consider two models $P(M_1) = P(M_2) = 0.5$:

- M_1: β is varying with time while γ is constant ($p_1 = 7$)
- M_2: both β and γ are time-dependent ($p_2 = 12$)

β in M_1, M_2, and γ in M_2 are represented as piecewise linear functions with a time interval of 20 days (except for Uruguay with 10 days interval).

Bayesian approach for parameter identification and model selection

- Let π_k be the prior densities of the parameters $\boldsymbol{\theta}_k$ of model M_k
- $\boldsymbol{X}$: observations of new infected and recovered cases during time intervals of **14** days
  - reduce the sensitivity due to data errors
- Using Bayesian approach, the posterior of the parameters $\boldsymbol{\theta}_i$ is obtained as

$$\pi_i(\boldsymbol{\theta}_i|\boldsymbol{X}) \propto f(\boldsymbol{X}|\boldsymbol{\theta}_i, M_i)\pi_i(\boldsymbol{\theta}_i)$$

  where f is the likelihood function
- Computational method: MH-MCMC
- Marginal predictive likelihood of model M (estimated using MCMC samples)

$$f(\boldsymbol{Z}|\boldsymbol{X}, M) = \int f(\boldsymbol{Z}|\boldsymbol{\theta}_i, M_i)\pi_i(\boldsymbol{\theta}_i|\boldsymbol{X})\mathrm{d}\boldsymbol{\theta}_i$$

- $\boldsymbol{Z}$: observations of new infected and recovered cases during time intervals of **20** days

Bayesian parameter identification

Likelihood function

Construct likelihood function is a difficult task

Motivation: the number of infected cases $>$ that confirmed

- using the conventional Gaussian likelihood function $\rightarrow$ Probability of (predicted cases are smaller than confirmed case) ~ 0.5
- Seroprevalence studies of anti-SARS-CoV-2 IgG antibodies suggest 200% in Spain and 1100% in Switzerland (estimate from small population samples)
- **Assumption**: number of infections/recovers is up to 120% of that confirmed (belief based likelihood)
- (not ideal assumption)

Bayesian parameter identification

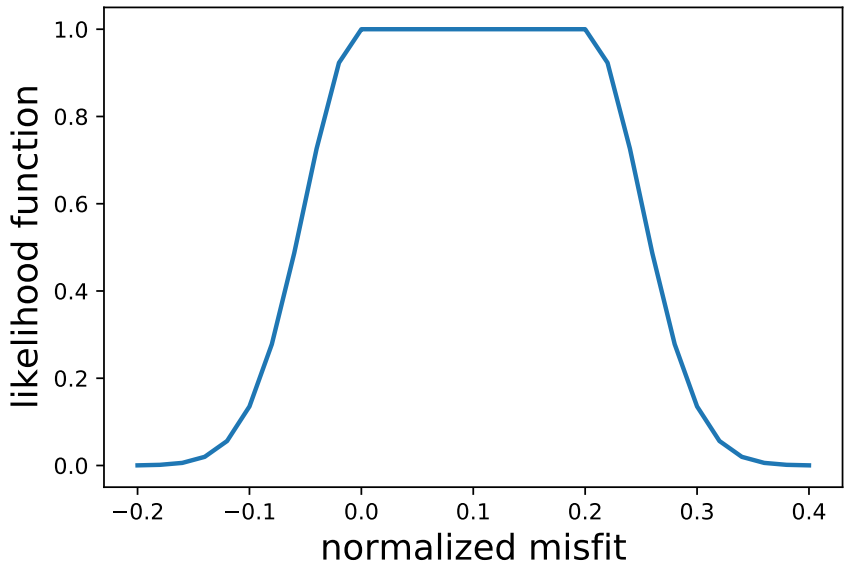
Likelihood function

Assumption: number of infections/recovers is up to 120% of that confirmed

Let $\overline{\text{err}}$ be the normalized misfit of the model prediction: $\overline{\text{err}} = \dfrac{x_i - y_i(\boldsymbol{\theta}_k)}{x_i}$

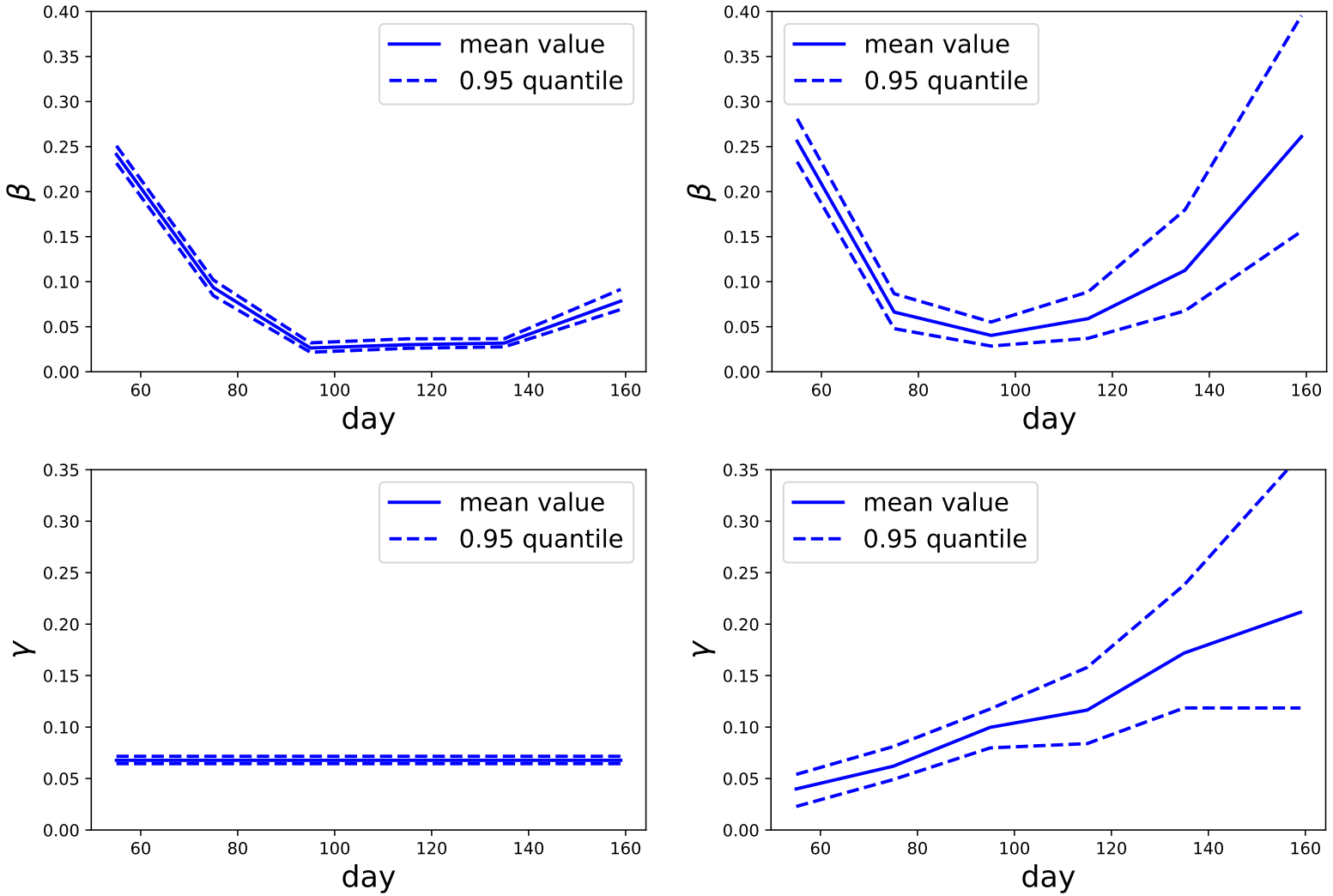
where x_i is an observation, and y_i is the corresponding model prediction

$$f(x_i|\boldsymbol{\theta}_k) = \begin{cases} 1 & \text{if } 0 \leq \overline{\text{err}} \leq 0.2 \\ \mathcal{N}(0, 0.05^2) * \sqrt{2\pi}0.05 & \text{if } \overline{\text{err}} < 0 \\ \mathcal{N}(0.2, 0.05^2) * \sqrt{2\pi}0.05 & \text{if } \overline{\text{err}} > 0.2 \end{cases}$$



Results: Model 1 (left) Model 2 (right)

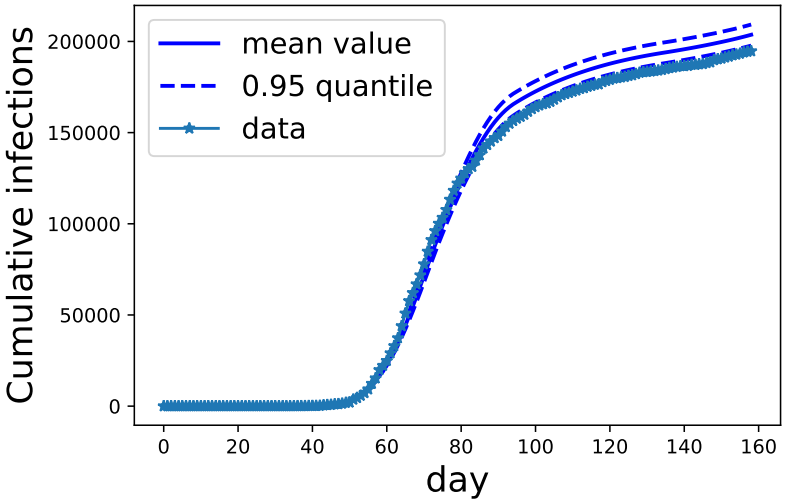
Germany (day 60: 22.03.2020, day 160: 28.06.2020)

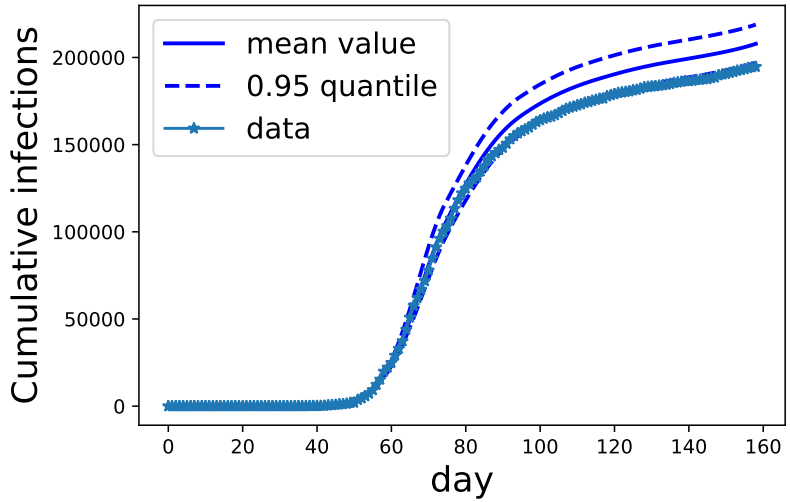


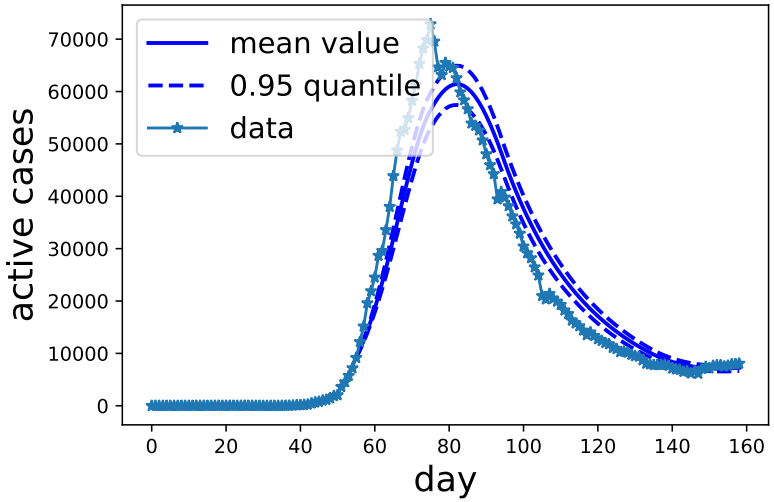
MH-MCMC with 1 million samples

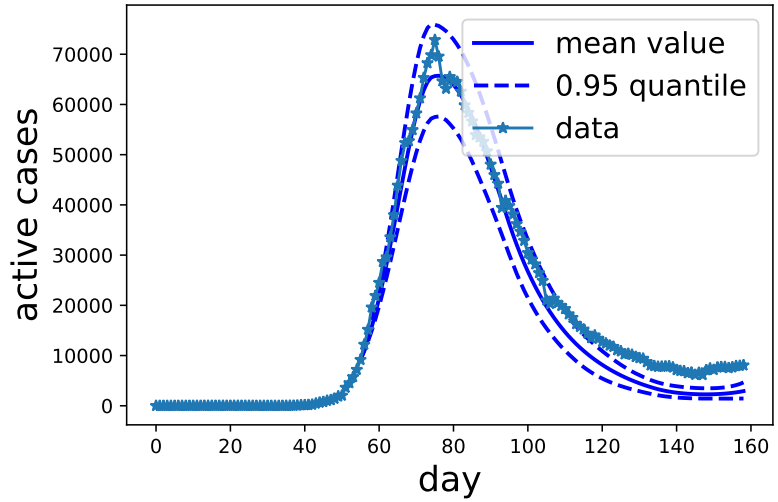
Results: Model 1 (left) Model 2 (right)

Germany, day 60: 22.03.2020, day 159: 28.06.2020







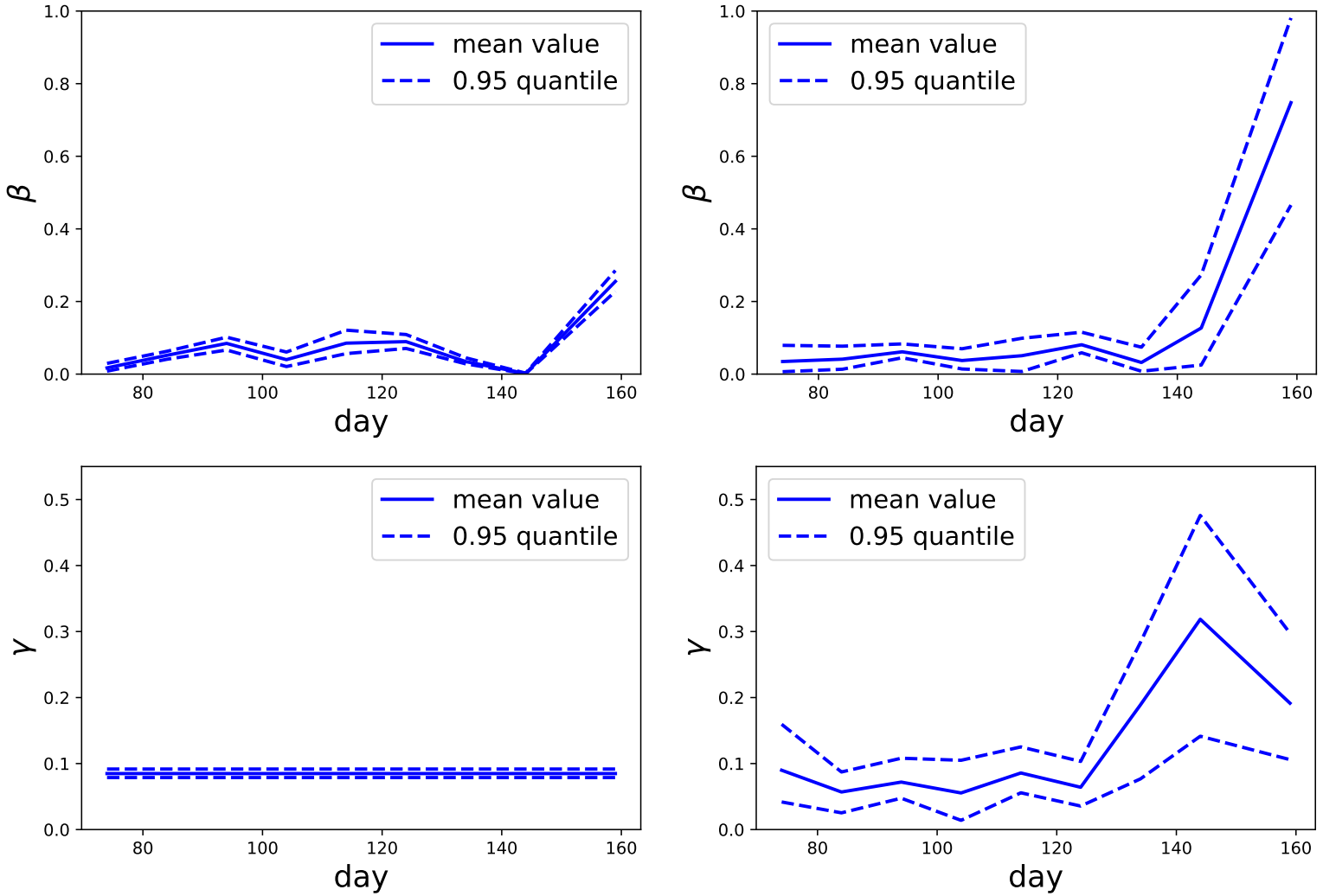


$$f(\boldsymbol{Z}|\boldsymbol{X}, M_1) = 1.e-8 \qquad f(\boldsymbol{Z}|\boldsymbol{X}, M_2) = 0.19$$

(Current setting: $f = 1$ indicate best models)

Results: Model 1 (left) Model 2 (right)

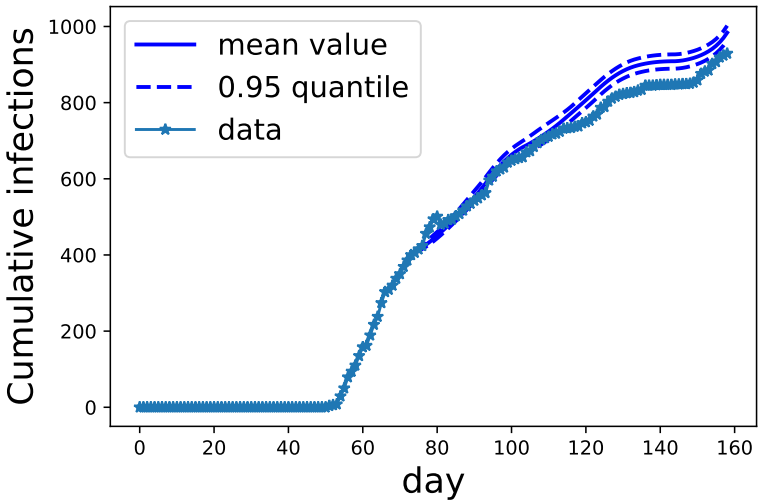
Uruguay, day 70: 01.04.2020, day 159: 28.06.2020

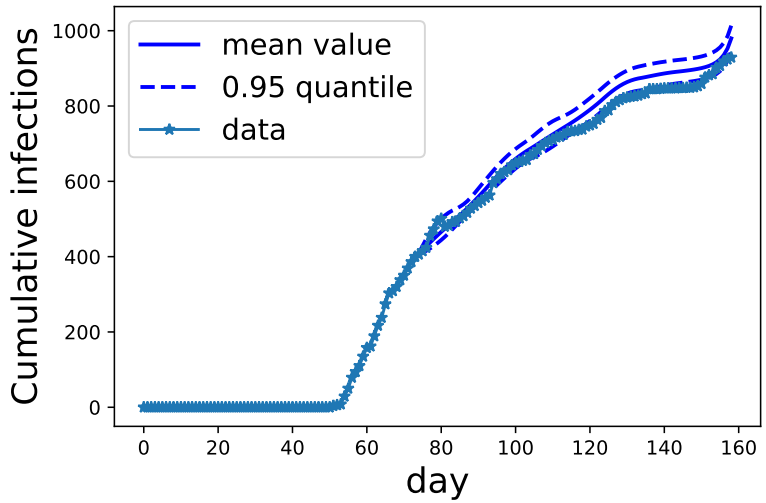


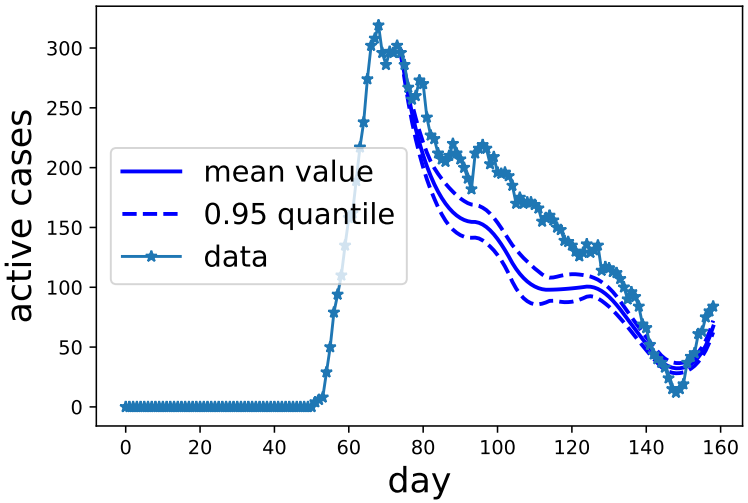
β and γ in Model 2 are represented as piecewise linear functions with time interval 10 days.

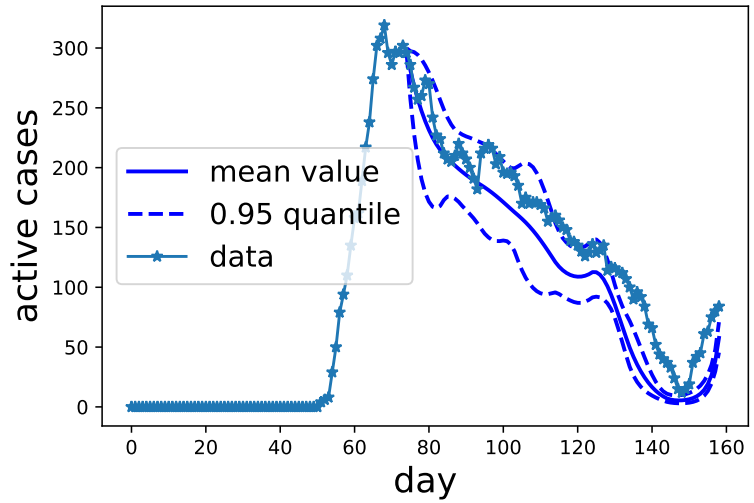
Results: Model 1 (left) Model 2 (right)

Uruguay, day 70: 01.04.2020, day 159: 28.06.2020





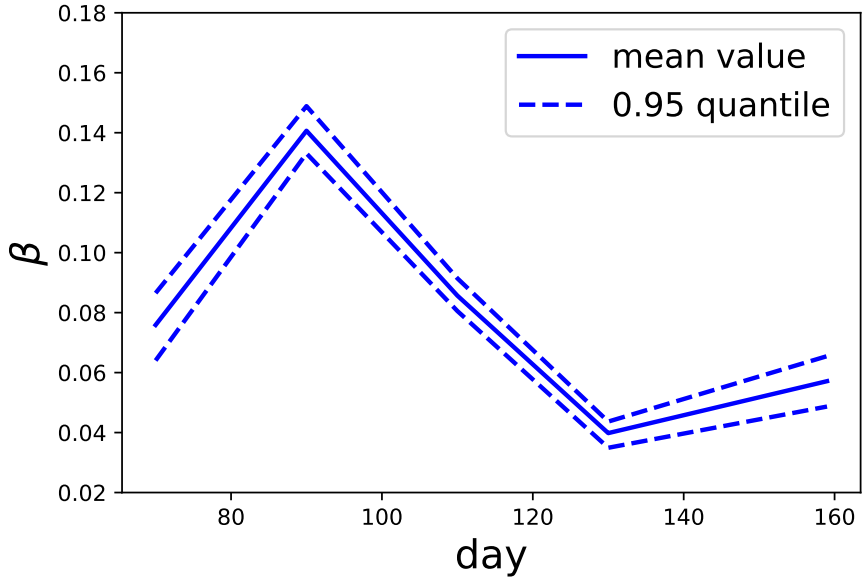


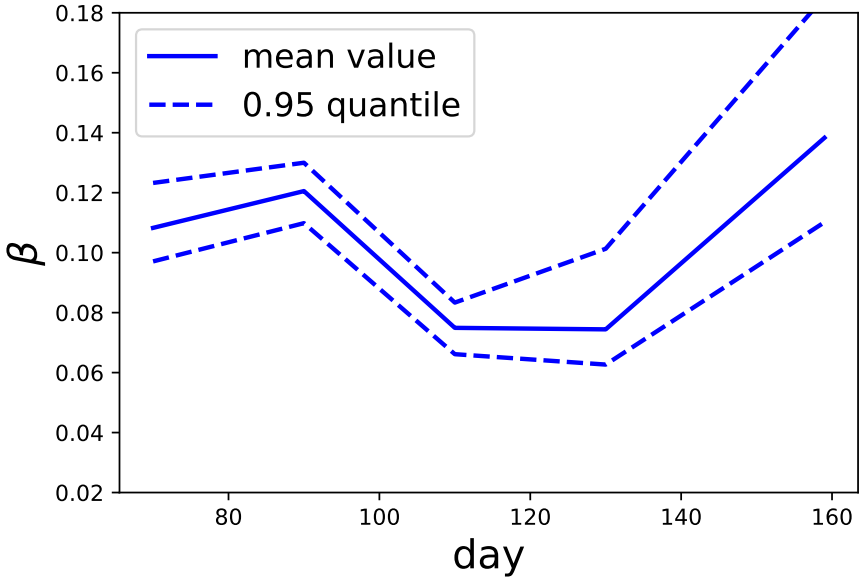


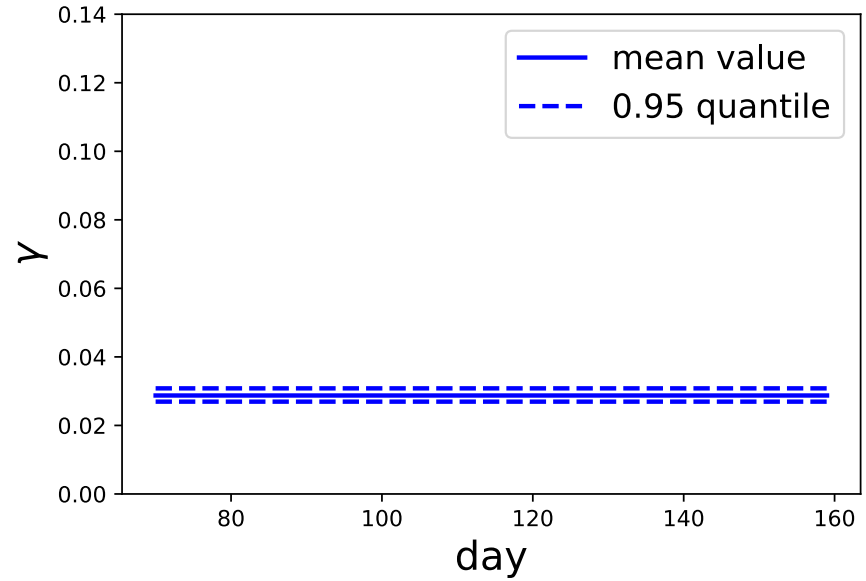
$$f(\boldsymbol{Z}|\boldsymbol{X}, M_1) = 3e-11 \qquad f(\boldsymbol{Z}|\boldsymbol{X}, M_2) = 0.02$$

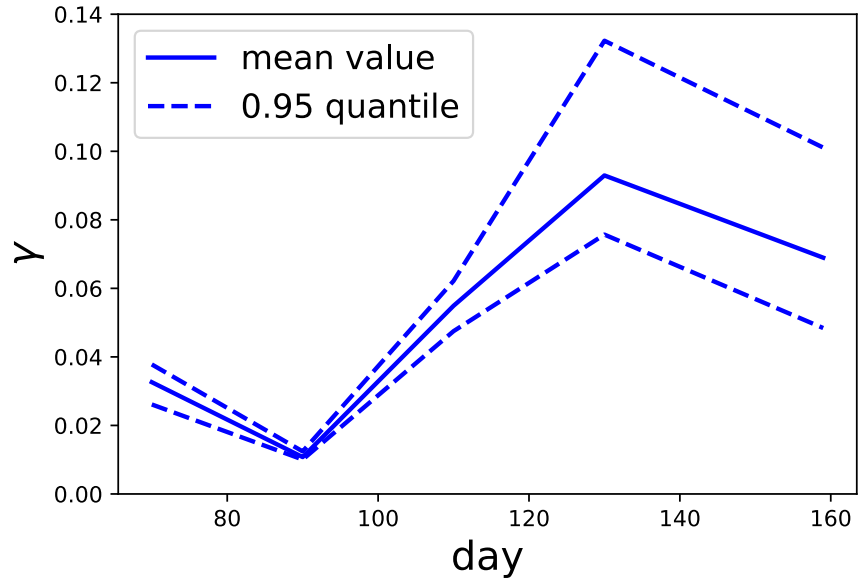
Results: Model 1 (left) Model 2 (right)

Saudi Arabia, day 70: 01.04.2020, day 159: 28.06.2020



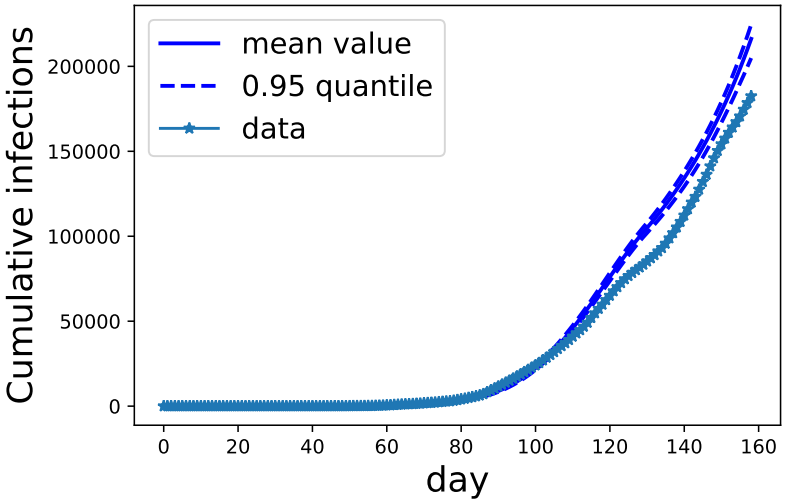


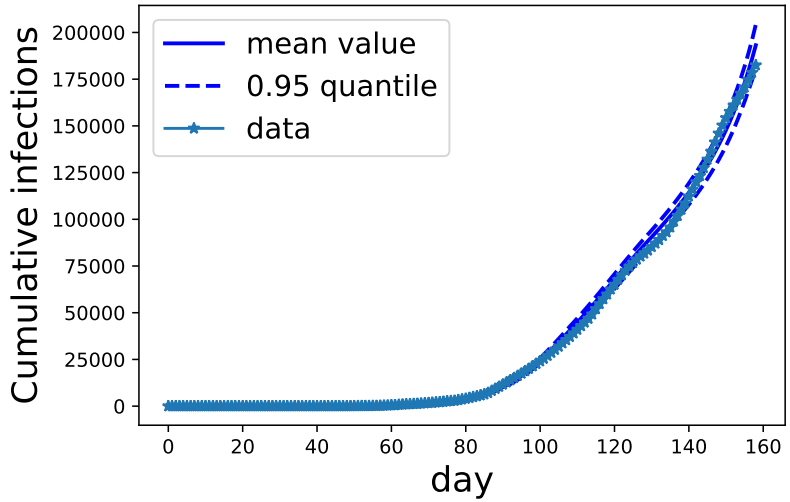


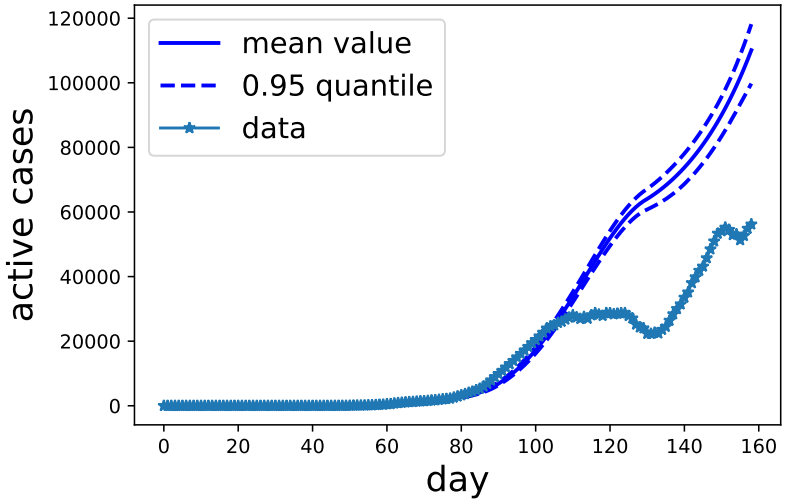


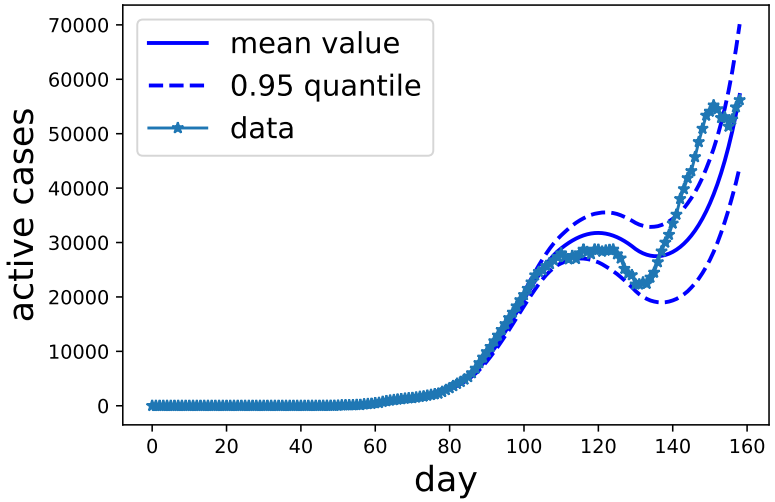
Results: Model 1 (left) Model 2 (right)

Saudi Arabia, day 70: 01.04.2020, day 159: 28.06.2020





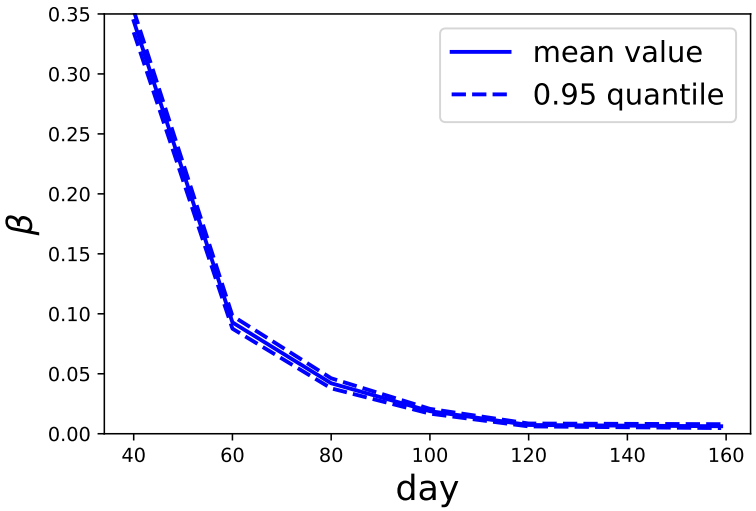


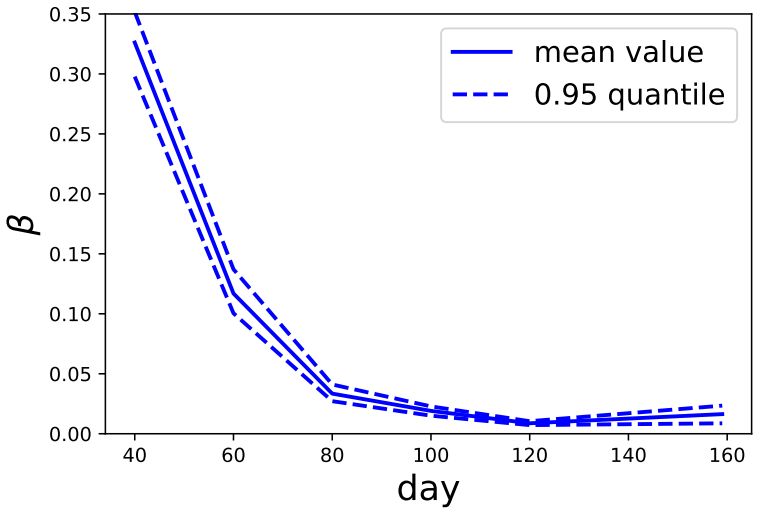


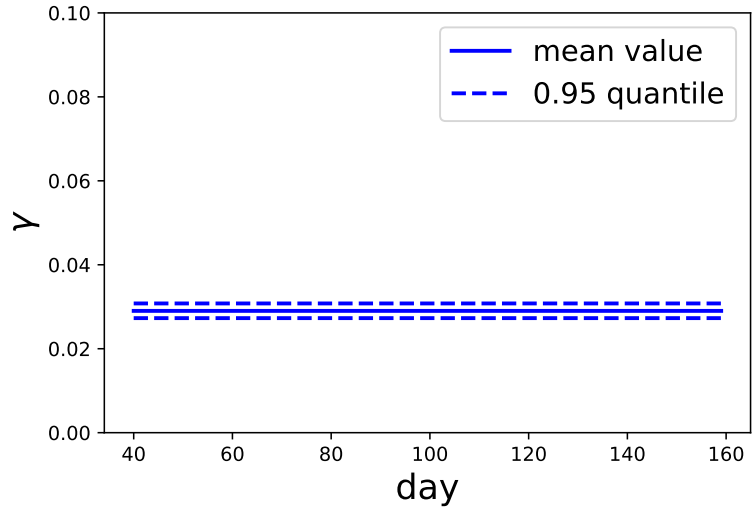
$$f(\boldsymbol{Z}|\boldsymbol{X}, M_1) = 3e-4 \qquad f(\boldsymbol{Z}|\boldsymbol{X}, M_2) = 0.05$$

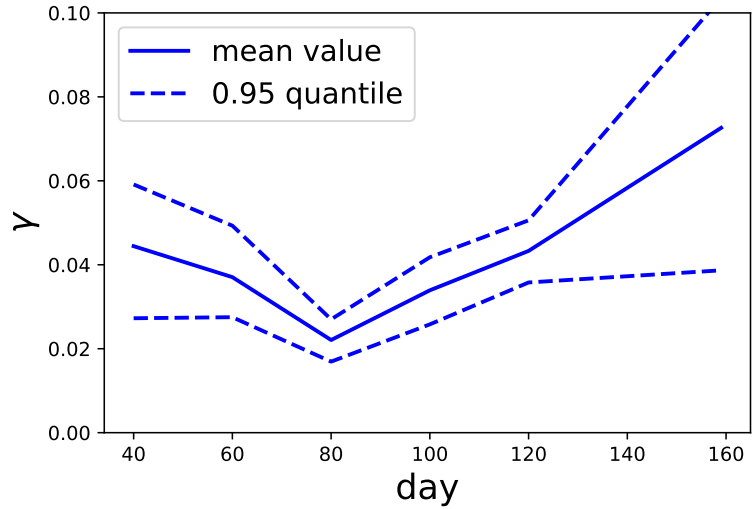
Results: Model 1 (left) Model 2 (right)

Italy, day 40: 02.03.2020, day 159: 28.06.2020



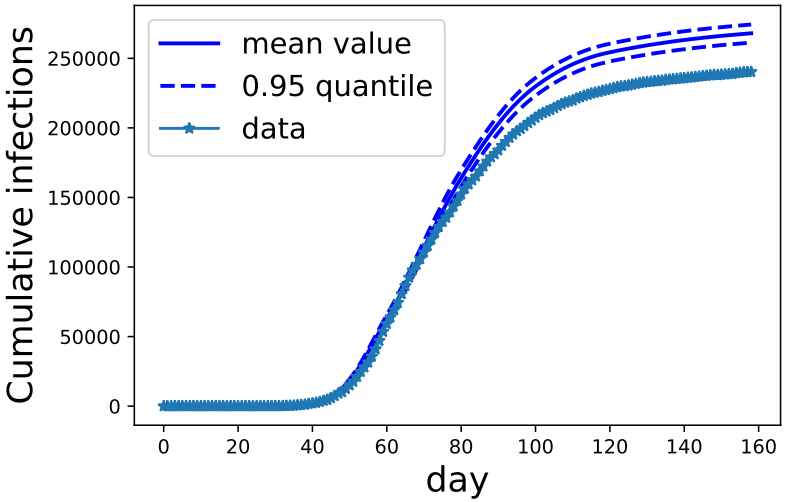


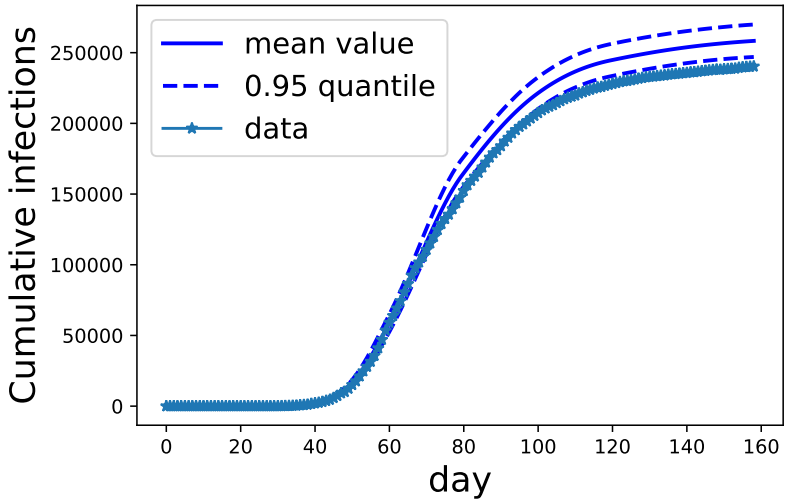


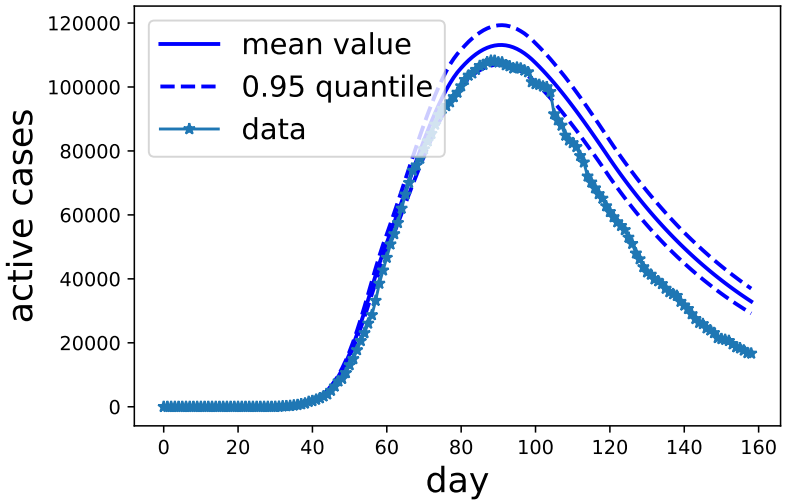


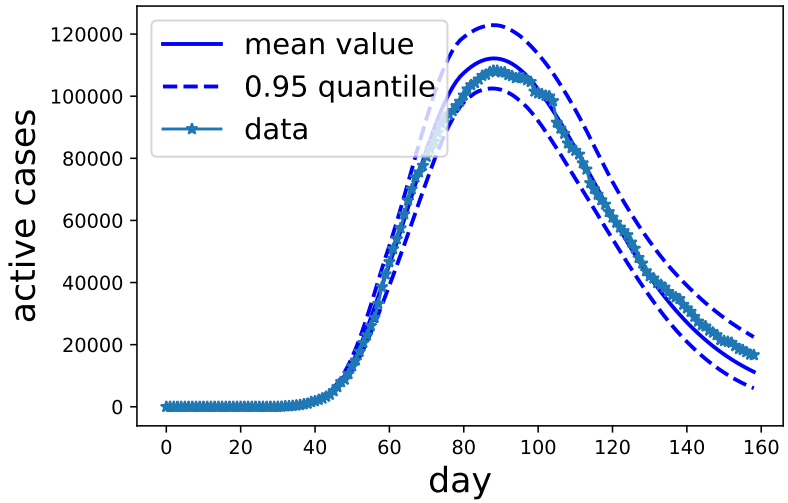
Results: Model 1 (left) Model 2 (right)

Italy, day 40: 02.03.2020, day 159: 28.06.2020









$$f(\boldsymbol{Z}|\boldsymbol{X}, M_1) = 0.07 \qquad f(\boldsymbol{Z}|\boldsymbol{X}, M_2) = 0.45$$

Summary

	$f(\boldsymbol{Z}\|\boldsymbol{X}, M_1)$	$f(\boldsymbol{Z}\|\boldsymbol{X}, M_2)$
Germany	1e-8	0.19
Uruguay	3e-11	0.02
Saudi Arabia	3e-4	0.05
Italy	0.07	0.45

There is strong evidence supporting Model 2

Conclusion

- An overview of Bayesian approach for model selection
- Different information criterions (BIC, GBIC, AIC)
- Predictive Bayesian distribution
- Apply for modeling the evolution of Covid-19 pandemic

Open problems:

- How to construct likelihood function
- Compare with SEIR, (E: exposed)
- Efficient method for identification