1 Introduction and Set-Up

In this paper we summarize the methods, for which we compared the predictive performance. The data set used is California Housing Prices Nugent 2017. The chosen likelihood is Gaussian linear regression, so that the observations y_i, can be represented as

$$y_i = x_i^\top w + \epsilon_i,$$

where x_i is the vector of features for the given data point i, w is the vector of parameters, and ϵ_i is Gaussian noise, such that $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Thus, the distribution of targets is $y_i \sim \mathcal{N}(x_i^\top w, \sigma^2)$. Therefore, in this set up, the parameters we need to learn are w and σ^2. The choice of prior varies for the cases we will describe below, and we will specify it later in the text.

First, the data was split into training and testing $4:1$. Then, pre-processing of the data was completed – the features were standardized, using StandardScaler from sklearn. Then, an intercept column (a column of ones) was added to the feature matrix X, so that the intercept is included as the first element of the coefficient vector w.

Additionally, the target values y_i were log transformed. This is an intuitive step to take – y_i represent the prices of houses, therefore, they cannot be negative.

The experiments were conducted in the following setting: we tested each of the methods on three bootstraps of the original training set, in order to investigate reproducibility of results for each of the bootstraps. The methods tested include GVI Knoblauch et al. 2022 with β, γ-divergence loss functions, and NLL, PCUQ Shen et al. 2025 with MMD and CRPS as scoring rule, and BayesBag Huggins and Miller 2024.

2 Experiments

2.1 GVI: $\beta-$divergence loss

For Gaussian linear regression and β-divergence loss, it is possible to solve the GVI objective both in closed form and using black box inference. We investigate both cases.

The objective optimized by GVI methodology can be written as

$$q^*(w,\sigma^2) = \arg\min_{q\in Q} \mathbb{E}_{q(w,\sigma^2)}\Big[\sum_{i=1}^{n} \ell(y_i, x_i, w, \sigma^2)\Big] + D(q(w,\sigma^2)||\pi(w,\sigma^2)), \tag{1}$$

where ℓ is the chosen loss function, D is the chosen divergence, $\pi(w,\sigma^2)$ is the prior, and Q is the variational family, within which we approximate the posterior. In both closed form case and black box inference case, D was chosen to be KLD. In this subsection we consider loss to be $\beta-$divergence loss, which for the likelihood being Gaussian linear regression can be formally expressed as

$$\begin{aligned}\ell_\beta(y_i, x_i, w, \sigma^2) = \frac{-1}{\beta-1}\mathcal{N}(y_i|x_i^\top w, \sigma^2)^{\beta-1} + \frac{1}{\beta}\int \mathcal{N}(z|x_i^\top w, \sigma^2)^\beta dz = \\ \frac{-1}{\beta-1}(2\pi\sigma^2)^{-(\beta-1)/2}\exp\{\frac{-(\beta-1)(y_i - x_i^\top w)^2}{2\sigma^2}\} + \\ \frac{1}{\beta}\int\Big[\Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^\beta \exp\frac{-\beta(z-x_i^\top w)^2}{2\sigma^2}\Big]dz = \\ \frac{-1}{\beta-1}(2\pi\sigma^2)^{-(\beta-1)/2}\exp\{\frac{-(\beta-1)(y_i - x_i^\top w)^2}{2\sigma^2}\} + \\ \beta^{-3/2}(2\pi\sigma^2)^{-(\beta-1)/2}.\end{aligned} \tag{2}$$

Now let us describe how we proceed in each of the considered cases – closed form objective and black-box inference.

2.1.1 Closed form objective

Prior and Variational Family Set-Up As the authors point out in Knoblauch et al. 2022, it is possible to compute (1) in the closed form, when the prior and variational family are chosen to be conjugate to the likelihood. Therefore, in the case of likelihood being Gaussian linear regression, the conjugate prior is Normal-Inverse-Gamma, so that

$$\pi(\sigma^2) \sim \text{InverseGamma}(a_0, b_0),$$

and

$$\pi(w\ |\sigma^2) \sim \mathcal{N}(\mu_0, \sigma^2 V_0).$$

Therefore, the joint prior $\pi(w, \sigma^2)$ is Normal-Inverse-Gamma(μ_0, V_0, a_0, b_0). In our test setting, we set the prior hyperparameters as follows: $\mu_0[0]$ – the prior mean of the intercept, is set to be equal to the empirical mean of the training dataset. The prior means of features coefficients are set to 0, as the data is standardized. The prior covariance matrix V_0 of the parameter vector w, is set to be: $V_0[0] = 2.0$, and $V_0[1:n] = 1.0$. The hyperparameters of the Inverse-Gamma prior on σ^2, are set as $a_0 = 2.1$, which ensures a finite variance for σ^2, and the parameter b_0 is initialized using the sample variance of the training dataset. Let $\widehat{\mathrm{Var}}(y)$ denote the empirical variance of y in the training set; then we set

$$b_0 = (a_0 - 1)\,\widehat{\mathrm{Var}}(y),$$

so that the prior mean of σ^2 matches the empirical variance of the observed data.

Next, we discuss the chosen variational family. In the closed-form case, we choose a variational family that is conjugate to the likelihood, matching the form of the prior. For Gaussian linear regression, the conjugate variational family for approximating the posterior over w and σ^2 is the Normal–Inverse-Gamma, where

$$\begin{aligned} q(\sigma^2) &\sim \text{Inverse-Gamma}(a, b), \\ q(w \mid \sigma^2) &\sim \mathcal{N}(\mu, \sigma^2 V), \text{ and} \\ q(w, \sigma^2) = q(\sigma^2) q(w \mid \sigma^2) &\sim \text{Normal-Inverse-Gamma}(\mu, V, a, b). \end{aligned} \tag{3}$$

Let p be equal to the number of features $+1$ (as we added an intercept). We parameterize the Normal–Inverse-Gamma variational family as $\mu \in \mathbb{R}^{\text{l}}$, $V = LL^\top$, where $LL^\top$ is a Cholesky factorization of a matrix V. Such parametrization of V ensures it is symmetric positive definite. The parameters of Inverse-Gamma a and b are just restricted to be positive. We when we optimize the posterior in the given variational family, we initialize the variational parameters at the prior hyperparameters.

KL-Divergence As denoted in eq. (1), KL-divergence between the prior and the approximated posterior is the part of the objective. Therefore, it needs to be derived. In

the derivation process we use the result of Llera and Beckmann 2016. The derivation is as follows,

$$\begin{aligned}
\mathrm{KL}\Big[q(w,\sigma^2)\;||\;\pi(w,\sigma^2)\Big] &= \mathrm{KL}\Big[q(\sigma^2)\;||\;\pi(\sigma^2)\Big] + \mathbb{E}_{q(\sigma^2)}\Big(\mathrm{KL}\Big[q(w\mid\sigma^2)\;||\;\pi(w\mid\sigma^2)\Big]\Big)\\
&= (a-a_0)\psi(a) + b_0\Big(\frac{a}{b}\Big) - a + \log\frac{b^{a_0+1}\Gamma(a_0)}{b\,b_0^{a_0}\Gamma(a)}\\
&+ \frac{1}{2}\mathbb{E}_{q(\sigma^2)}\Big[\log\frac{|V_0|}{|V|} - p + (\mu-\mu_0)^\top\frac{1}{\sigma^2}(\mu-\mu_0) + \mathrm{tr}(V_0^{-1}V)\Big]\\
&= (a-a_0)\psi(a) + b_0\Big(\frac{a}{b}\Big) - a + \log\frac{b^{a_0+1}\Gamma(a_0)}{b\,b_0^{a_0}\Gamma(a)}\\
&+ \frac{1}{2}\Big[\log\frac{|V_0|}{|V|} - p + (\mu-\mu_0)^\top\frac{a}{b}(\mu-\mu_0) + \mathrm{tr}(V_0^{-1}V)\Big]. \qquad (4)
\end{aligned}$$

Computing the closet form objective To compute the objective we need to optimize in the closed form, we need to compute

$$\mathbb{E}_{q(w,\sigma^2)}\Big[\sum_{i=1}^{n}\ell_\beta(y_i,x_i,w,\sigma^2)\Big] = \sum_{i=1}^{n}\mathbb{E}_{q(w,\sigma^2)}\Big[\ell_\beta(y_i,x_i,w,\sigma^2)\Big].$$

Let us compute the expectation of the loss function for a single data point,

$$\begin{aligned}
&\mathbb{E}_{q(w,\sigma^2)}\Big[\ell_\beta(y_i,x_i,w,\sigma^2)\Big] = \mathbb{E}_{q(\sigma^2)}\Big[\mathbb{E}_{q(w|\sigma^2)}\Big(\ell_\beta(y_i,x_i,w,\sigma^2)\Big)\Big]\\
&= \mathbb{E}_{q(\sigma^2)}\Bigg[\mathbb{E}_{q(w|\sigma^2)}\Bigg[\underbrace{\frac{-1}{\beta-1}(2\pi\sigma^2)^{-(\beta-1)/2}\exp\Big\{\frac{-(\beta-1)(y_i-x_i^\top w)^2}{2\sigma^2}\Big\}}_{\text{Term 1}}\Bigg]\Bigg]\\
&+ \mathbb{E}_{q(\sigma^2)}\Bigg[\mathbb{E}_{q(w|\sigma^2)}\Bigg[\underbrace{\beta^{-3/2}\,(2\pi\sigma^2)^{-(\beta-1)/2}}_{\text{Term 2}}\Bigg]\Bigg].
\end{aligned}$$

Let us fist compute the nested expectation of Term 2, which is

$$\mathbb{E}_{q(\sigma^2)}\Big[\mathbb{E}_{q(w|\sigma^2)}\Big[\beta^{-3/2}\,(2\pi\sigma^2)^{-(\beta-1)/2}\Big]\Big] = \beta^{-3/2}(2\pi)^{-(\beta-1)/2}\frac{\Gamma(a+\frac{\beta-1}{2})}{\Gamma(a)}b^{-(\beta-1)/2} = (*),$$

using the formula of the moment of Inverse-Gamma.

Now, let us compute the nested expectation of Term 1, which is

$$
\begin{aligned}
&\mathbb{E}_{q(\sigma^2)}\left[\mathbb{E}_{q(w|\sigma^2)}\left[\frac{-1}{\beta-1}(2\pi\sigma^2)^{-(\beta-1)/2}\exp\left\{\frac{-(\beta-1)(y_i-x_i^\top w)^2}{2\sigma^2}\right\}\right]\right] \\
&= \mathbb{E}_{q(\sigma^2)}\left[\frac{-1}{\beta-1}(2\pi\sigma^2)^{-(\beta-1)/2}\,\mathbb{E}_{q(w|\sigma^2)}\left(\exp\left\{\frac{-(\beta-1)(y_i-x_i^\top w)^2}{2\sigma^2}\right\}\right)\right] \\
&= \frac{-1}{\beta-1}(2\pi)^{-(\beta-1)/2}\,\mathbb{E}_{q(\sigma^2)}\left[(\sigma^2)^{-(\beta-1)/2}\,\mathbb{E}_{q(w|\sigma^2)}\left(\exp\left\{\frac{-(\beta-1)(y_i-x_i^\top w)^2}{2\sigma^2}\right\}\right)\right] \\
&= \frac{-1}{\beta-1}(2\pi)^{-(\beta-1)/2}\,\mathbb{E}_{q(\sigma^2)}\left[(\sigma^2)^{-(\beta-1)/2}\,\frac{\exp\left\{-\frac{(\beta-1)(y_i-x_i^\top\mu)^2}{2\sigma^2\left(1+(\beta-1)\,x_i^\top V x_i\right)}\right\}}{\sqrt{1+(\beta-1)\,x_i^\top V x_i}}\right] \\
&= \frac{-1}{\beta-1}(2\pi)^{-(\beta-1)/2}\,\frac{1}{\sqrt{1+(\beta-1)\,x_i^\top V x_i}}\,\mathbb{E}_{q(\sigma^2)}\left[(\sigma^2)^{-(\beta-1)/2}\exp\left\{-\frac{(\beta-1)(y_i-x_i^\top\mu)^2}{2\sigma^2\left(1+(\beta-1)\,x_i^\top V x_i\right)}\right\}\right] \\
&= \frac{-1}{\beta-1}(2\pi)^{-(\beta-1)/2}\,\frac{1}{\sqrt{1+(\beta-1)\,x_i^\top V x_i}}\,\frac{b^a}{\Gamma(a)}\,\Gamma\left(a+\frac{\beta-1}{2}\right)\left(b+\frac{(\beta-1)(y_i-x_i^\top\mu)^2}{2\left(1+(\beta-1)\,x_i^\top V x_i\right)}\right)^{-\left(a+\frac{\beta-1}{2}\right)} \\
&= (**).
\end{aligned}
$$

Therefore, the objective optimized can be written as,

$$
\begin{aligned}
q^*(w,\sigma^2) = \underset{q\in Q\sim\text{Normal-Inverse-Gamma}(\mu,V,a,b)}{\arg\min} \sum_{i=1}^{n} \Bigg[& \frac{-1}{\beta-1}(2\pi)^{-(\beta-1)/2}\,\frac{1}{\sqrt{1+(\beta-1)\,x_i^\top V x_i}}\,\frac{b^a}{\Gamma(a)} \\
& \Gamma\left(a+\frac{\beta-1}{2}\right)\left(b+\frac{(\beta-1)(y_i-x_i^\top\mu)^2}{2\left(1+(\beta-1)\,x_i^\top V x_i\right)}\right)^{-\left(a+\frac{\beta-1}{2}\right)} \\
& + \beta^{-3/2}(2\pi)^{-(\beta-1)/2}\frac{\Gamma(a+\frac{\beta-1}{2})}{\Gamma(a)}b^{-(\beta-1)/2}\Bigg] \\
& + (a-a_0)\psi(a) + b_0\left(\frac{a}{b}\right) - a + \log\frac{b^{a_0+1}\Gamma(a_0)}{b\,b_0^{a_0}\Gamma(a)} \\
& + \frac{1}{2}\Big[\log\frac{|V_0|}{|V|} - p + (\mu-\mu_0)^\top\frac{a}{b}(\mu-\mu_0) + \operatorname{tr}(V_0^{-1}V)\Big]. \qquad (5)
\end{aligned}
$$

Training For training the model we used Adam (following the authors of Knoblauch et al. 2022), with learning rate = 1e-2, number of steps = 500, batch size = 64. These hyperparameters were set after experimenting on a multiple of possible values of these

hyperparameters.

2.1.2 Black-Box Inference

We also try the other variational family and prior, which are not conjugate to the likelihood, in order to compare the closed form inference with the black box inference for β-divergence loss. Let us denote that the form of the loss function itself stays the same.

Prior and Variational Family Set-Up We choose prior to be Gaussian on the vector of features w and Inverse-Gamma on σ^2, so that

$$\pi(w) \sim \mathcal{N}(\mu_0, V_0),$$

and

$$\pi(\sigma^2) \sim \text{InverseGamma}(a_0, b_0).$$

Let us notice that in this set-up, σ^2 and w are apriori independent. The prior hyperparameters are chosen the same as in the previous section.

The independence is kept in variational family as well. We choose the Normal-Log-Normal variational family, where

$$q(w) \sim \mathcal{N}(\mu, V), q(\sigma^2) \sim \text{LogNormal}(m, s^2).$$

As we choose a mean-field variational family and assume independence of w and σ^2, the joint variational posterior factorizes as

$$q(w, \sigma^2) = q(w)\, q(\sigma^2).$$

We parametrize $\mu \in \mathbb{R}^p$, $m \in \mathbb{R}$, and take V to be diagonal with positive variances. Therefore, we represent V through a positive vector containing its diagonal entries. The standard deviation of $\log \sigma^2$ is parametrized to be positive.

KL-Divergence As in the previous case, KL-divergence between the approximated posterior $q(w, \sigma^2)$ and prior $\pi(w, \sigma^2)$ is a part of the objective. Let us derive it,

$$\begin{aligned}
\mathrm{KL}\big[q(w,\sigma^2) \,\|\, \pi(w,\sigma^2)\big] &= \mathrm{KL}[q(w) \,\|\, \pi(w)] + \mathrm{KL}\big[q(\sigma^2) \,\|\, \pi(\sigma^2)\big] \\
&= \frac{1}{2}\Big[\operatorname{tr}(V_0^{-1}V) + (\mu-\mu_0)^\top V_0^{-1}(\mu-\mu_0) - p + \log\frac{|V_0|}{|V|}\Big] \\
&+ \Big(a_0 m - \log s - \tfrac{1}{2}\big(1+\log(2\pi)\big) - a_0\log b_0 + \log\Gamma(a_0) + b_0\, e^{-m+\frac{1}{2}s^2}\Big).
\end{aligned} \tag{6}$$

Therefore, the objective we are optimizing is

$$\begin{aligned}
q^*(w,\sigma^2) = \operatorname*{arg\,min}_{q\in Q\sim\text{Normal-Log-Normal}(\mu,V,m,s)} &\sum_{i=1}^{n} \mathbb{E}_{q(w,\sigma^2)}\Big[\frac{-1}{\beta-1}(2\pi\sigma^2)^{-(\beta-1)/2}\exp\{\frac{-(\beta-1)(y_i - x_i^\top w)^2}{2\sigma^2}\} \\
&+ \beta^{-3/2}(2\pi\sigma^2)^{-(\beta-1)/2}\Big] \\
&+ \frac{1}{2}\Big[\operatorname{tr}(V_0^{-1}V) + (\mu-\mu_0)^\top V_0^{-1}(\mu-\mu_0) - p + \log\frac{|V_0|}{|V|}\Big] \\
&+ \Big(a_0 m - \log s - \tfrac{1}{2}\big(1+\log(2\pi)\big) - a_0\log b_0 + \log\Gamma(a_0) + b_0\, e^{-m+\frac{1}{2}s^2}\Big).
\end{aligned}$$

The expectation of the loss function cannot be computed in the closed form for the non-conjugate variational family. Therefore, we need to estimate it. For that, we sample the random noise $\epsilon \sim \mathcal{N}(0, I)$, and set

$$w = \mu + V^{1/2}\epsilon,$$

and random noise $\eta \sim \mathcal{N}(0, 1)$

$$\sigma^2 = \exp\big(m + s\,\eta\big),$$

which allows us to construct reparameterized samples from the variational distribution $q(w, \sigma^2) = q(w)\, q(\sigma^2)$. Using these samples, we form a Monte Carlo approximation of the intractable expectation and optimize the variational objective with noisy gradients.

Training For training the model we used Adam (following the authors of Knoblauch et al. 2022), with learning rate = 1e-3, number of steps = 1000, batch size = 64. These hyperparameters were set after experimenting on a multiple of possible values of these hyperparameters.

2.2 GVI: NLL

For NLL we also compare the results for two variational families, as in the case of β-divergence loss. The key difference from β-divergence loss is that for NLL, even for the case of Normal-Log-Normal mean field family, the objective we optimize has a closed form. We skip the specifications of priors, variational families, and KL divergence, as they stay unchanged.

Let us specify the NLL function itself, for the case of gaussian linear regression. does not depend on the choice of prior or variational family,

$$\mathrm{NLL}(y_i, x_i, w, \sigma^2) = \frac{n}{2}\log(2\pi) + \frac{n}{2}\log(\sigma^2) + \frac{(y_i - x_i^\top w)}{2\sigma^2}. \tag{7}$$

Now let us describe how we proceed in each of twp cases.

Normal-Inverse-Gamma Variational Family: w and σ^2 are not independent The objective we optimize is of form (1). Therefore, we need to compute the expectation of the loss function (7) in closed form. We derive it as,

$$\begin{aligned}\mathbb{E}_{q(w,\sigma^2)}\left[\sum_{i=1}^{n}\mathrm{NLL}(y_i, x_i, w, \sigma^2)\right] &= \frac{n}{2}\log(2\pi) + \frac{n}{2}\mathbb{E}_{q(\sigma^2)}\Big[\log(\sigma^2)\Big] + \sum_{i=1}^{n}\mathbb{E}_{q(\sigma^2)}\left[\mathbb{E}_{q(w|\sigma^2)}\Big[\frac{(y_i - x_i^\top w)^2}{2\sigma^2}\Big]\right] \\ &= \frac{1}{2}\left[n\log(2\pi) + n\big(\log b - \psi(a)\big) + \frac{a}{b}\sum_{i=1}^{n}(y_i - x_i^\top \mu)^2 + \sum_{i=1}^{n} x_i^\top V x_i\right].\end{aligned} \tag{8}$$

In this case the optimized objective is the sum of eq. (8) and eq. (4).

Training For training the model we used Adam, with learning rate = 1e-2, number of steps = 500, batch size = 64.

Normal-Log-Normal Mean-Field Variational family: w and σ^2 are independent
The objective we optimize is of form (1). We need to compute the expectation of the loss function (7) in closed form. We derive it as,

$$\begin{aligned}\mathbb{E}_{q(w,\sigma^2)}\left[\sum_{i=1}^{n}\mathrm{NLL}(y_i,x_i,w,\sigma^2)\right] &= \frac{n}{2}\log(2\pi)+\frac{n}{2}\,\mathbb{E}_{q(\sigma^2)}[\log(\sigma^2)]+\frac{1}{2}\,\mathbb{E}_{q(\sigma^2)}\left[\frac{1}{\sigma^2}\right]\sum_{i=1}^{n}\mathbb{E}_{q(w)}\left[(y_i-x_i^\top w)^2\right] \\ &= \frac{n}{2}\log(2\pi)+\frac{n}{2}m+\frac{1}{2}\exp\left(-m+\frac{s^2}{2}\right)\sum_{i=1}^{n}\left[(y_i-x_i^\top\mu)^2+x_i^\top V x_i\right].\end{aligned} \tag{9}$$

Therefore, in this case the optimized objective is the sum of eq. (9) and eq. (6).

Training For training the model we used Adam, with learning rate = 1e-3, number of steps = 1000, batch size = 64.

2.3 GVI: γ-divergence loss

In case of γ-divergence (and the corresponding loss), we were able to find several parameterizations of it. One is presented in Jewson et al. 2024, and is of the form

$$\begin{aligned}\ell^\gamma(y_i,x_i,w,\sigma^2) = \frac{-1}{\gamma-1}\left[\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{\gamma-1}\exp\left\{\frac{-(\gamma-1)(y_i-x_i^\top w)^2}{2\sigma^2}\right\}\right] \\ \frac{1}{\gamma}\frac{1}{\left[\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{\gamma}\int\exp\left\{\frac{-\gamma(z-x_i^\top w)^2}{2\sigma^2}\right\}dz\right]}^{\frac{\gamma-1}{\gamma}}.\end{aligned} \tag{10}$$

It can be noticed that this form is a monotonic transformation of the reparameterized version of the original γ-divergence loss, introduced in Hideitsu Fujisawa and Eguchi 2008. Jewson et al. 2024 perform this transformation to obtain a form of γ-divergence loss, which can guaranty a closed form objective in case of conjugate variational family. However, in the gaussian linear regression setting, although the objective can be written in closed form, its form imposes constraints that prevent the variational parameters from being optimized properly. In particular, when we take expectation with respect to $q(w,\sigma^2)$, where $q(w,\sigma^2)$ is Normal-Inverse-Gamma, one of the terms of the resulting

objective is $\sqrt{1+(\gamma-1)x_i^\top V x_i}$. Clearly, for this term to exist, $1+(\gamma-1)x_i^\top V x_i$ should be non-negative, which is equivalent to $x_i^\top V x_i \leq \frac{1}{1-\gamma}$. Thus, this puts a constraint on V, a variational parameter which should be properly optimized with no artificial constraints, for robust results. Conducting experiments with the loss function of form (10) has confirmed the stated above – the results were numerically meaningful only when constraints were put on V, but in this case, as this parameter was not properly optimized, the performance was significantly worse than for any other considered method. Therefore, we do not consider the version of γ-divergence loss, stated above. For the further experiments we adopted another version of γ-divergence loss, introduced by Kawashima and Hironori Fujisawa 2016. It is important to note that this form is an empirical approximation to the full $\gamma-$cross-entropy. In the case of likelihood being a gaussian linear regression, the loss is of the form

$$
\begin{aligned}
L^\gamma(y, X, w, \sigma^2) &= -\frac{1}{\gamma}\log\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i - x_i^\top w)^2}{2\sigma^2}\right)\right]^\gamma\right\} \\
&+ \frac{1}{1+\gamma}\log\left\{\frac{1}{n}\sum_{i=1}^{n}\int\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y - x_i^\top w)^2}{2\sigma^2}\right)\right]^{1+\gamma} dy\right\} \\
&= \frac{\log n}{\gamma} + \frac{1}{2}\log(2\pi\sigma^2) + \frac{-1}{\gamma}\log\Big(\sum_{i=1}^{n}\exp\Big\{\frac{-\gamma(y_i - x_i^\top w)^2}{2\sigma^2}\Big\}\Big) \\
&+ \frac{-\gamma\log(2\pi\sigma^2) - \log(\gamma+1)}{2(1+\gamma)}.
\end{aligned} \tag{11}
$$

Same as for β-divergence loss and NLL, we set experiments for $\gamma-$divergence loss of the form (11) both with the mean-field Normal-Log-Normal variational family, and conjugate Normal-Inverse-Gamma variational family.

Normal-Inverse-Gamma Variational Family: w and σ^2 are not independent In this case the objective optimized cannot be computed in closed form. Therefore, in the training process we need to sample from the variational family, and evaluate the expectation of (11) empirically, adding the KL-divergence term of the form (4). To draw samples we use a two-step sampling. As sampling from Inverse-Gamma directly is impossible, we first sample the precision $\tau = 1/\sigma^2$ from the Gamma distribution, using

rsample() in PyTorch. Then we invert back, to get σ^2. Second, conditional on σ^2, we sample w. We draw $\epsilon \sim \mathcal{N}(0, 1)$, and compute $w = \mu + \sigma L^\top$.

Training For training we use Adam with 1000 steps, learning rate of 1e-3, and batch batch size of 64.

Normal-Log-Normal Mean-Field Variational Family: w and σ^2 are independent To optimize the objective, which is not available in closed form, we have to sample from the variational family to empirically estimate the expectation of the loss from (11). We sample in the same way as in 2.1.2. The full objective consists of the estimated expectation, and KL-divergence of the form (6).

Training For training we use Adam with 1000 steps, learning rate of 1e-3, and batch batch size of 64.

2.4 PCUQ: MMD

2.5 PCUQ: CRPS

2.6 BayesBag

3 Evaluation

We evaluate the performance of each method using several metrics. The first step is to compute the predictive distribution, and to evaluate how close the predictions are to the true target values. We do this evaluation using (1) Simple visualization of residuals for 100 data points, (2) NLPD score, (3) CRPS score. Additionally, we plot the posterior distributions and predictive means for 100 points, across all 3 bootstraps, to evaluate reproducibility. [we will need to introduce a criterion to evaluate reproducibility and posteriors and predictives formally] Posterior predictive distribution can be formally

expressed as,

$$p(y^* \mid x^*, X, y) = \iint \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y^* - x^{*\top} w)^2}{2\sigma^2}\right) \left[\frac{1}{(2\pi\sigma^2)^{p/2}|V|^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(w-\mu)^\top V^{-1}(w-\mu)\right)\right] \frac{b^a}{\Gamma(a)}(\sigma^2)^{-(a+1)} \exp\left(-\frac{b}{\sigma^2}\right)\, dw\, d\sigma^2. \tag{12}$$

It is important to the denote that (12) can be computed in closed form in case of conjugate likelihood and variational family. For the case of Normal-Inverse-Gamma variational family and gaussian linear regression likelihood, the resulted predictive posterior of the form (12) is a Student-t-distribution$(x^{*\top}\mu^*, \frac{b^*}{a^*}(1 + x^{*\top} V x^*), 2a^*)$, where μ^*, V^*, a^*, b^* are the variational parameters of the optimal distribution q^*. In the case of mean-field variational family, the closed form predictive distribution does not exist. Therefore, sampling is necessary to obtain the empirical approximation to it. Additionally, we use NLPD and CRPS scoring rules to evaluate the predictive distribution. NLPD is the negative log of predictive distribution, defined in (12). CRPS is a metric which evaluates both accuracy to the true test values, and closeness of predictives for the same data point to one another. Let M be the number of samples for each test data point, then the empirical CRPS for an observed value y and predictive samples $\{s_1, \ldots, s_M\}$ is defined as

$$\widehat{\mathrm{CRPS}}(y;\, s_{1:M}) = \frac{1}{M}\sum_{m=1}^{M}|s_m - y| \;-\; \frac{1}{2M^2}\sum_{m=1}^{M}\sum_{k=1}^{M}|s_m - s_k|.$$

The first term measures the average distance between the predictive samples and the true value, and the second term accounts for all pairwise distances between samples.

4 Results

The results are presented in the tables and the plots below (at the section Figures). We used $\beta = 1.3$ and $\gamma = 0.7$. CRPS and NLPD scores are as follows (means for 3 bootstraps):

Table 1: Averaged Scores (Normal–Inverse–Gamma Variational Family)

Method	NLPD	CRPS
NLL	0.9105	0.4050
BETA	0.8853	0.3822
GAMMA	1.1662	0.4460

Table 2: Averaged Scores (Mean-Field Variational Family)

Method	NLPD	CRPS
BETA	0.8820	0.3822
GAMMA	8301.7497	1.1593
NLL	36756.5458	1.2452

5 Next steps

Implement the gaps in the sections above. Conduct experiments in order to evaluate stability – instead of testing on different bootstraps, test with changing the prior or the model slightly.

References

Fujisawa, Hideitsu and Shinto Eguchi (2008). "Statistical inference based on the γ-divergence". In: *Journal of Japan Statistical Society* 38.1, pp. 3–31. DOI: 10.1016/S0047-259X(08)00045-6.

Huggins, Jonathan H. and Jeffrey W. Miller (2024). "Reproducible parameter inference using bagged posteriors". In: *Electronic Journal of Statistics* 18.1, pp. 1549–1585. DOI: 10.1214/24-EJS2237. URL: https://doi.org/10.1214/24-EJS2237.

Jewson, Jack, Jim Q. Smith, and Chris Holmes (2024). "On the Stability of General Bayesian Inference". In: *Bayesian Analysis* TBA.TBA, pp. 1–31. DOI: 10.1214/24-BA1502. URL: https://doi.org/10.1214/24-BA1502.

Kawashima, Takayuki and Hironori Fujisawa (2016). "Robust and Sparse Regression via γ-divergence". In: *arXiv preprint*. eprint: arXiv:1604.06637.

Knoblauch, Jeremias, Jack Jewson, and Theodoros Damoulas (2022). "An Optimization-centric View on Bayes' Rule: Reviewing and Generalizing Variational Inference". In: *Journal of Machine Learning Research* 23. Ed. by Frank Wood, pp. 1–109. URL: https://jmlr.org/papers/volume23/19-1047/19-1047.pdf.

Llera, Alberto and Christian F. Beckmann (2016). "Estimating an Inverse Gamma Distribution". In: *arXiv preprint arXiv:1605.01019*. version v2, July 7, 2016. URL: https://arxiv.org/abs/1605.01019.

Nugent, Cam (2017). *California Housing Prices*. https://www.kaggle.com/datasets/camnugent/california-housing-prices. Accessed: 2025-11-29.

Shen, Zhen et al. (2025). "Prediction-centric uncertainty quantification via MMD". In: *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*. Vol. 258. Proceedings of Machine Learning Research. PMLR. URL: https://arxiv.org/abs/2410.11637.

6 Figures

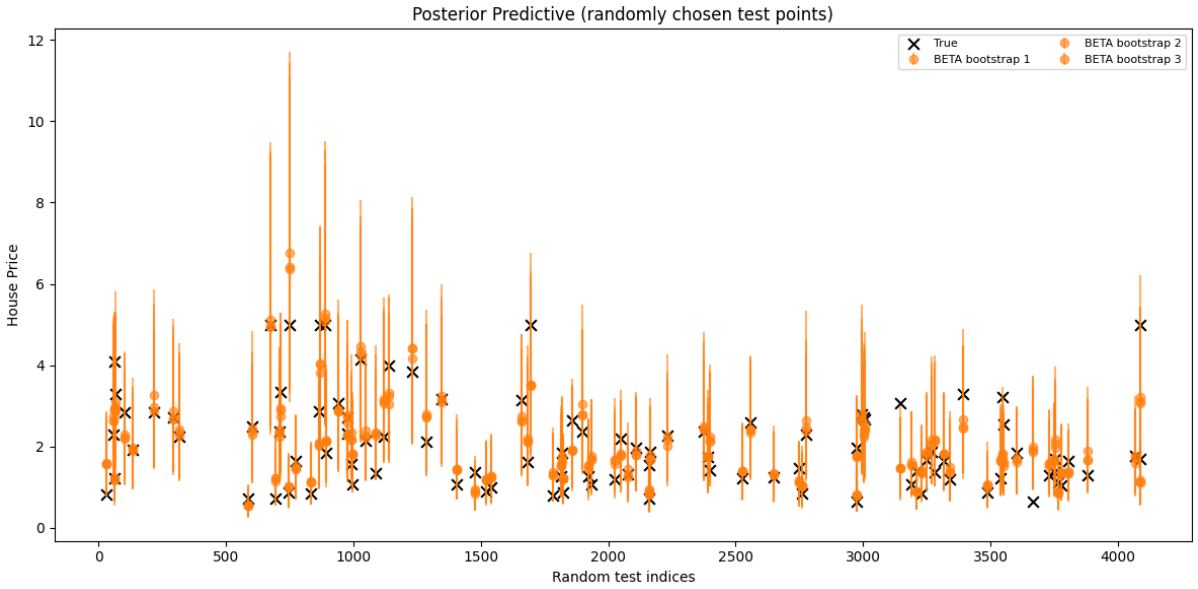

Figure 1: beta_conj_plot.png

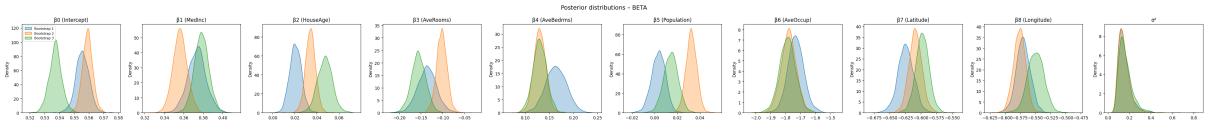

Figure 2: beta_conj_posterior.png

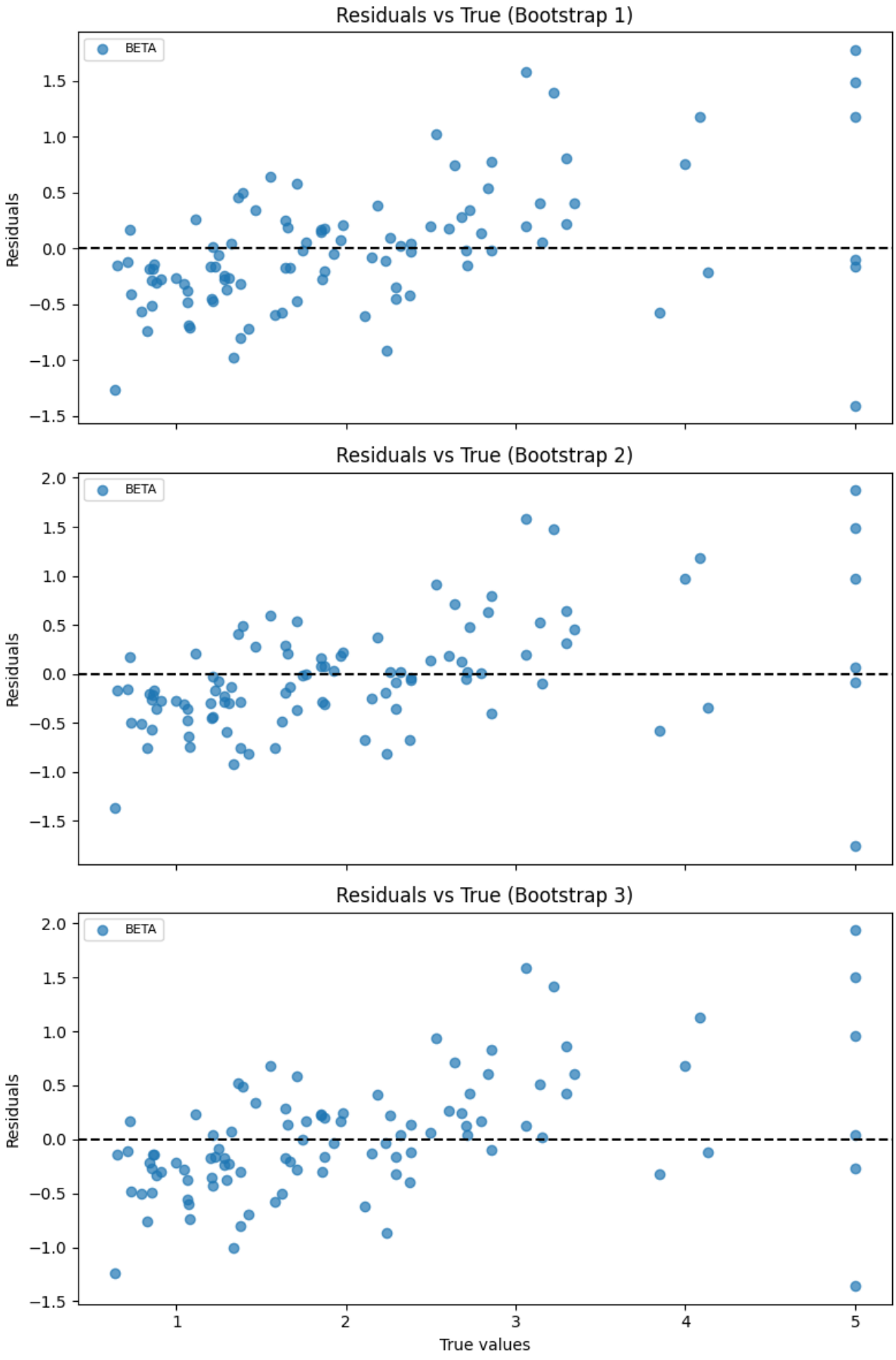

Figure 3: beta_conj_res.png

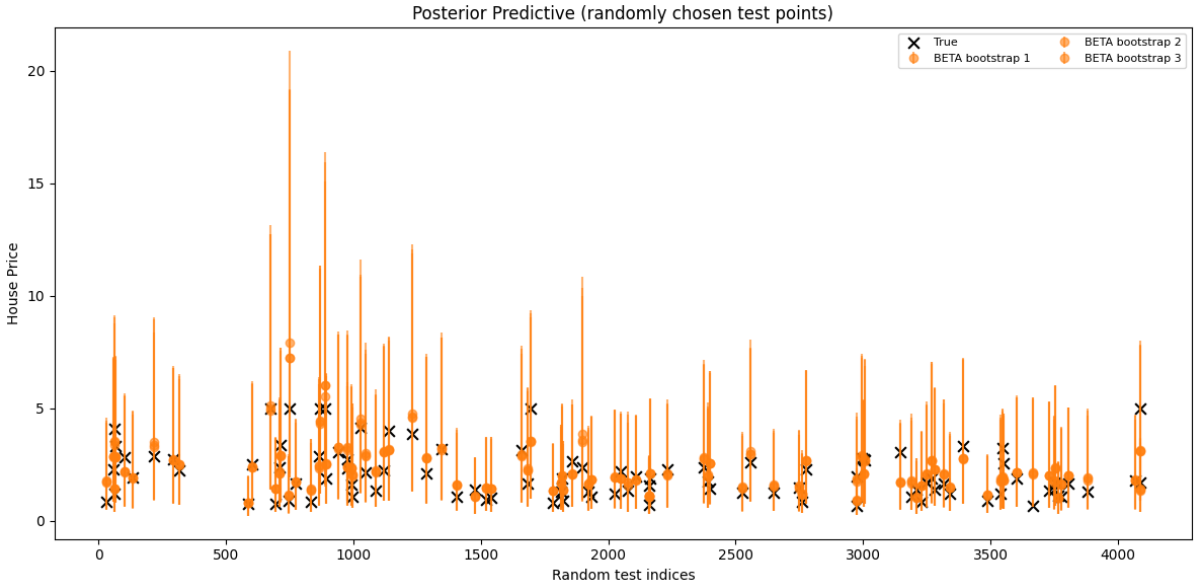

Figure 4: beta_mean_field_plot.png

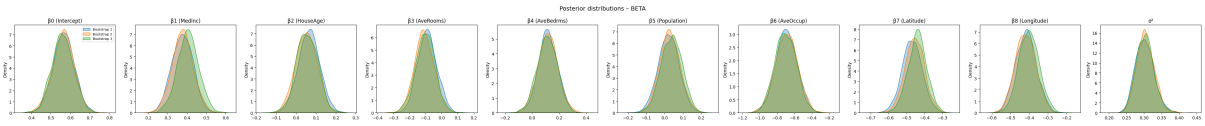

Figure 5: beta_mean_field_posteriors.png

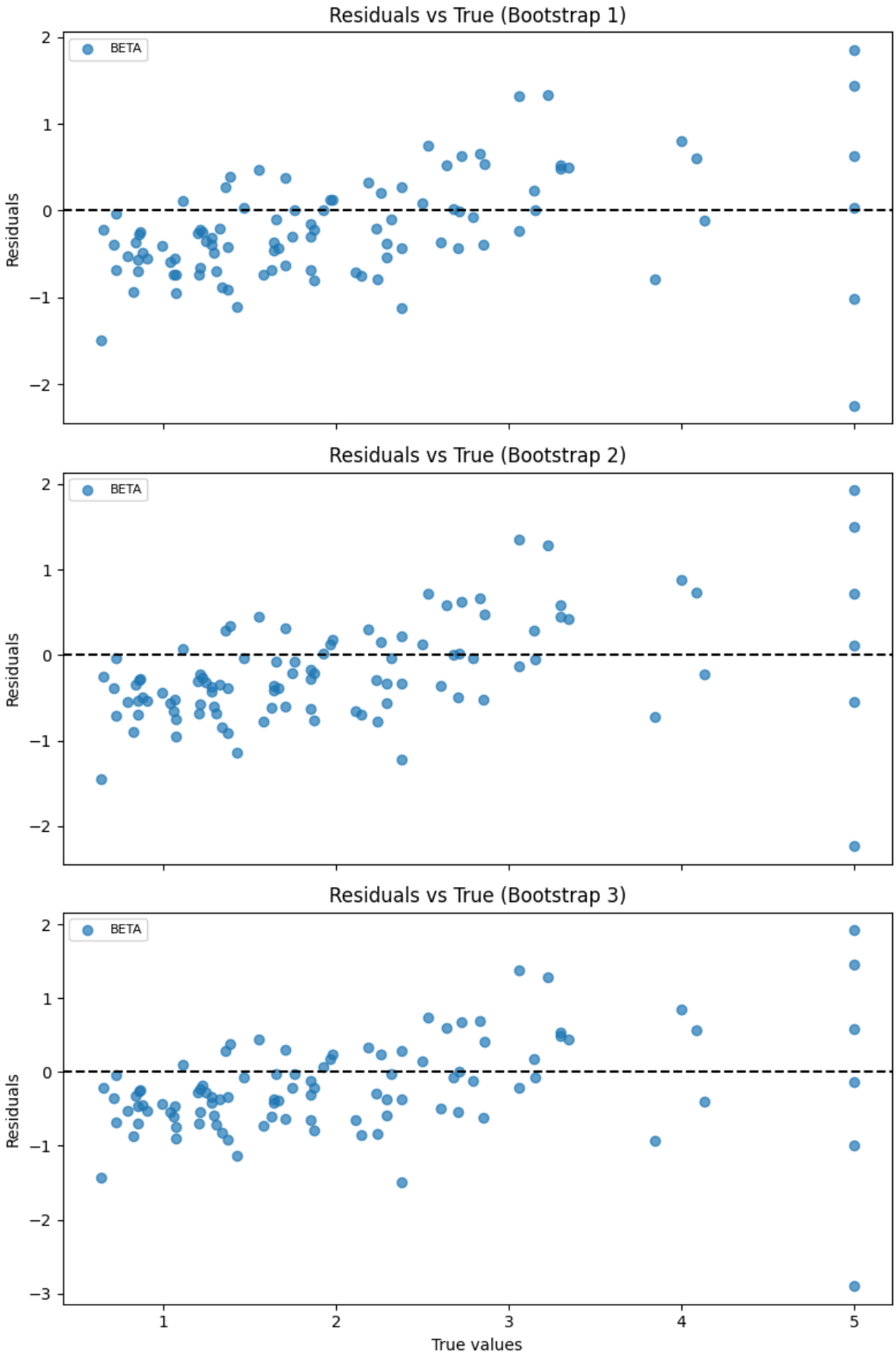

Figure 6: beta_mean_field_residuals.png

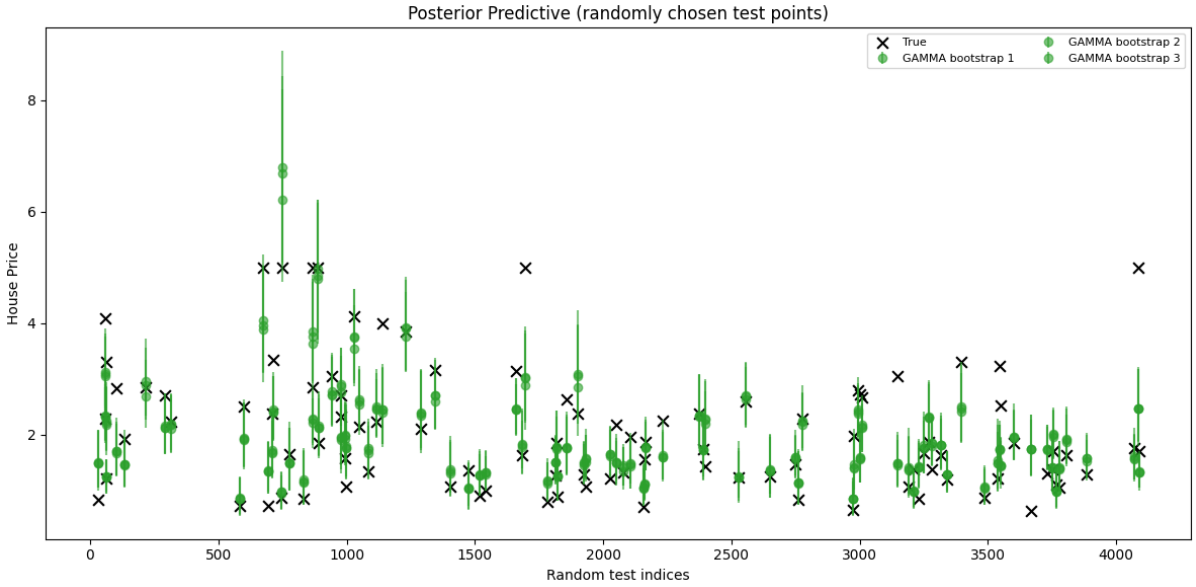

Figure 7: gamma_conj_plot.png

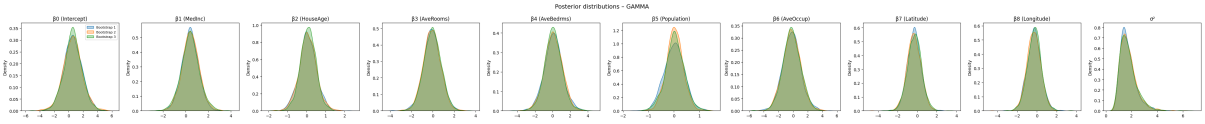

Figure 8: gamma_conj_posteriors.png

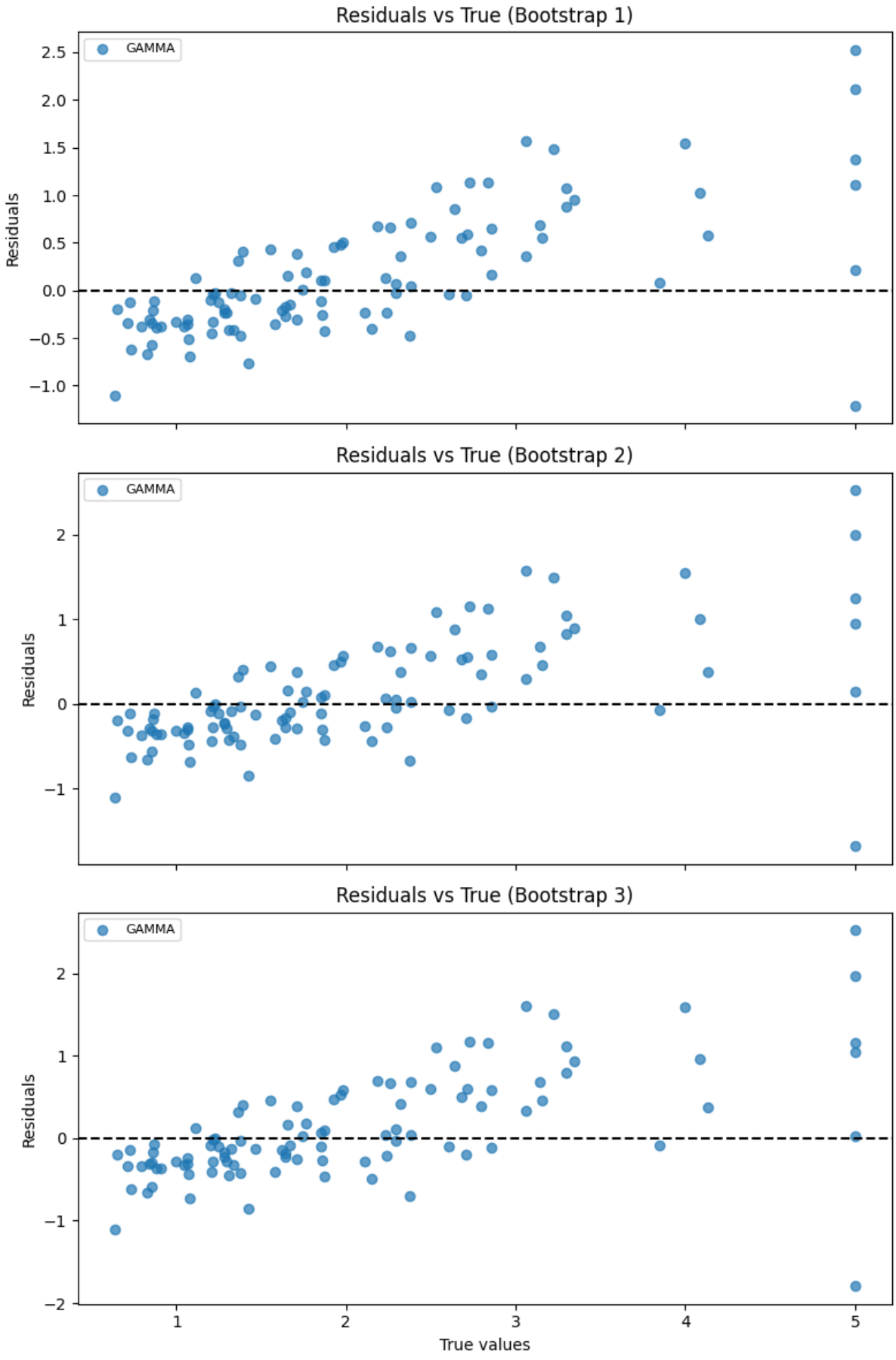

Figure 9: gamma_conj_residuals.png

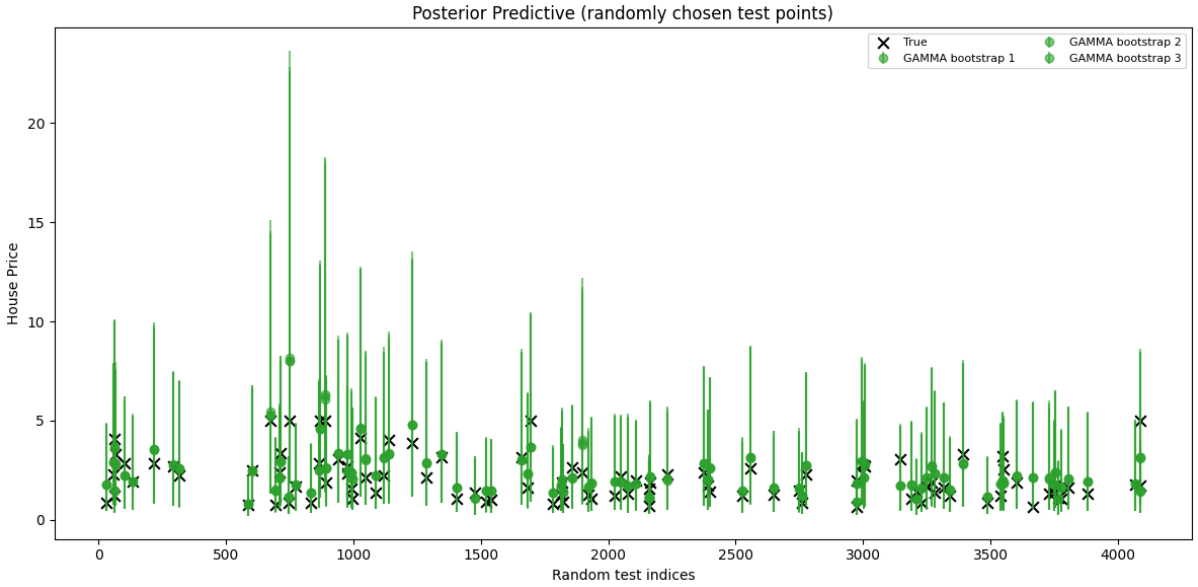

Figure 10: gamma_mean_field_plot.png

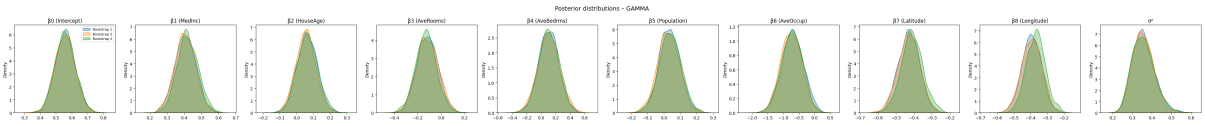

Figure 11: gamma_mean_field_posteriors.png

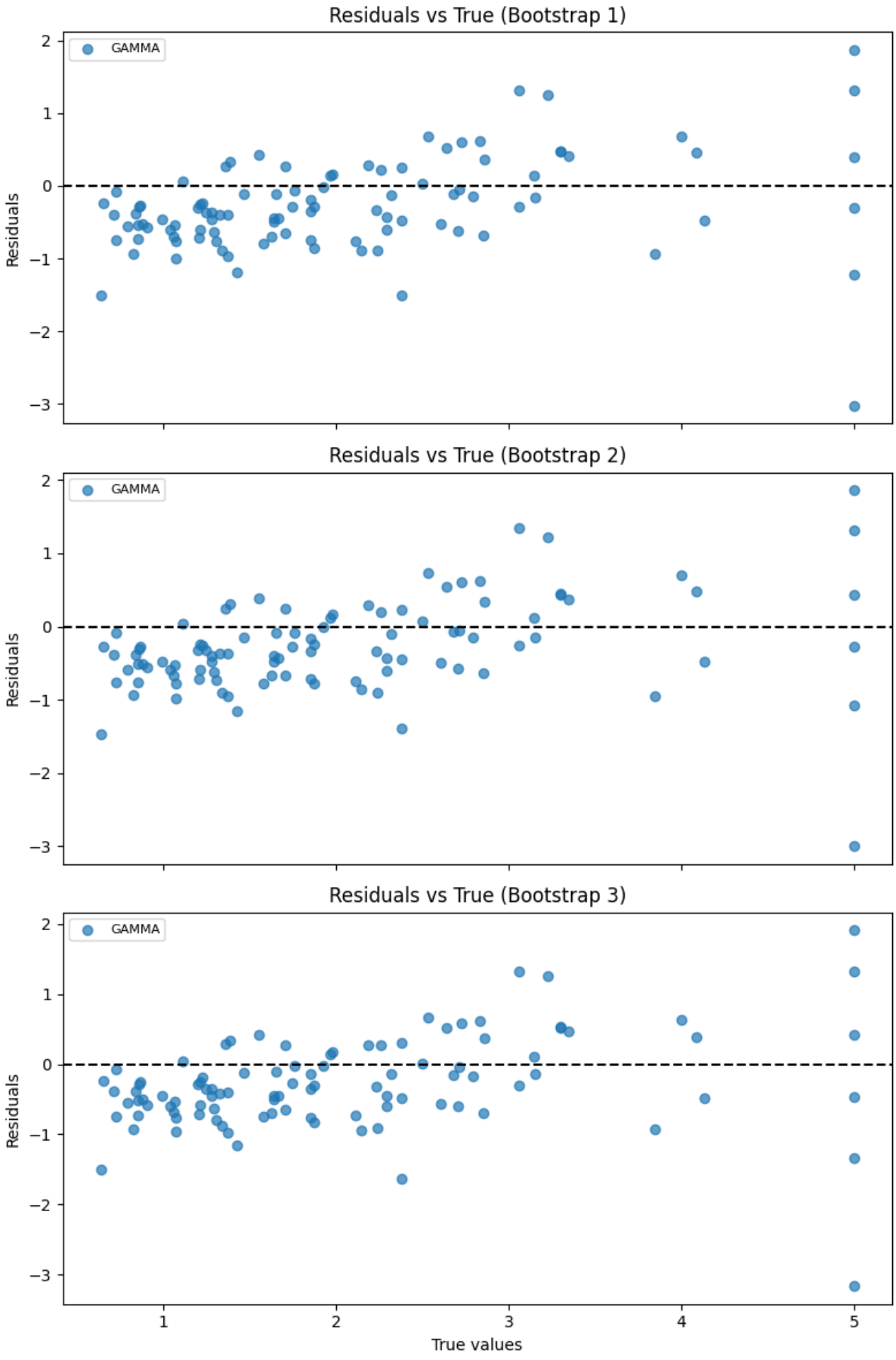

Figure 12: gamma_mean_field_residuals.png

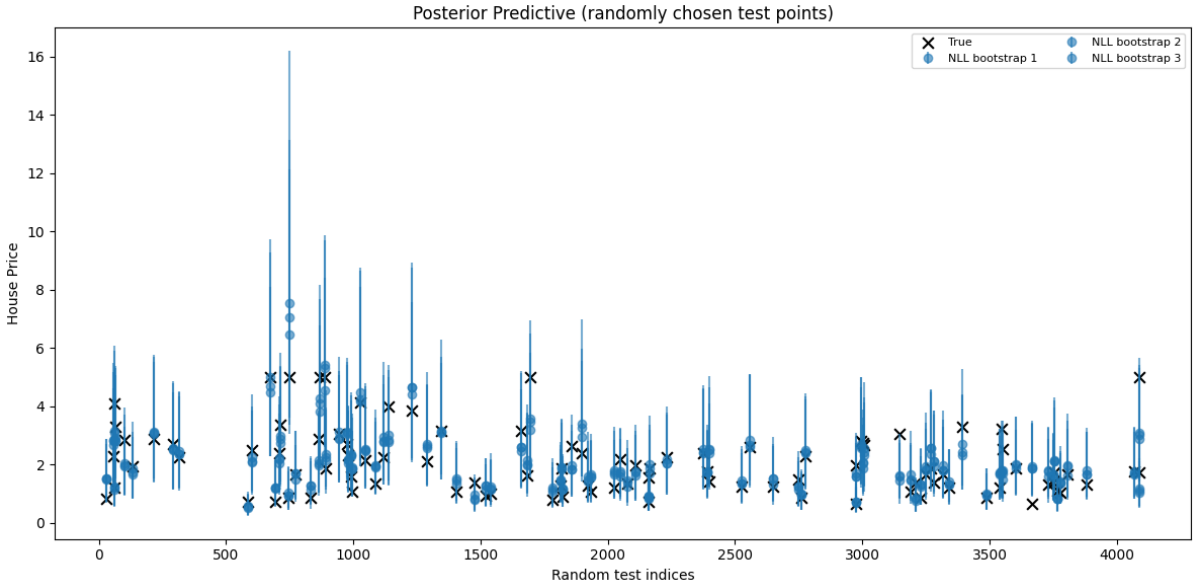

Figure 13: nll_conj_plot.png

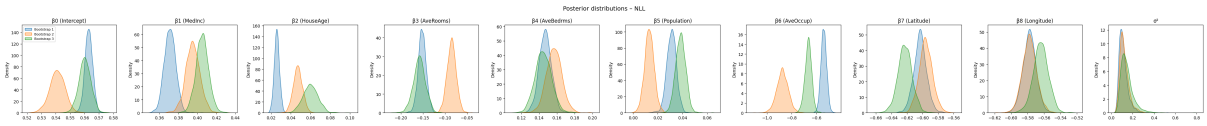

Figure 14: nll_conj_posterior.png

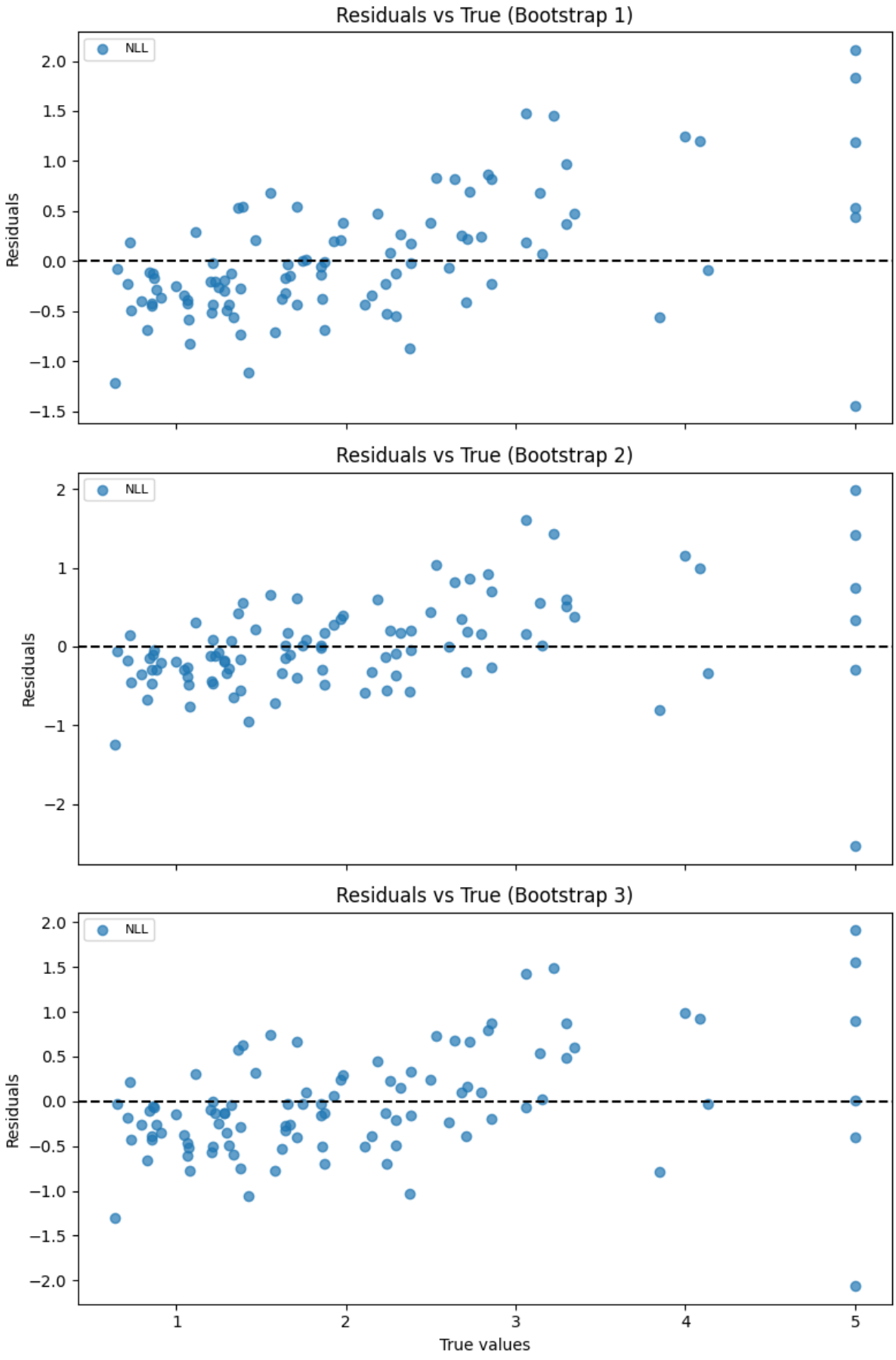

Figure 15: nll_conj_residuals.png

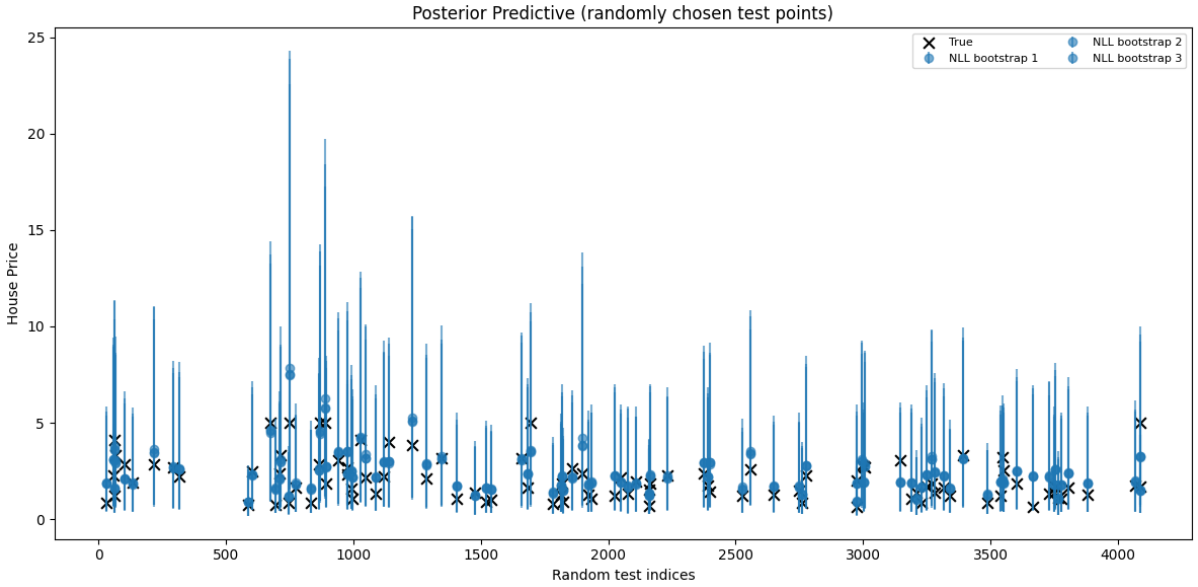

Figure 16: NLL_mean_field_plot.png

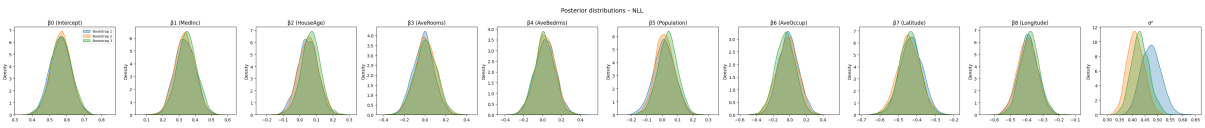

Figure 17: NLL_mean_field_posteriors.png

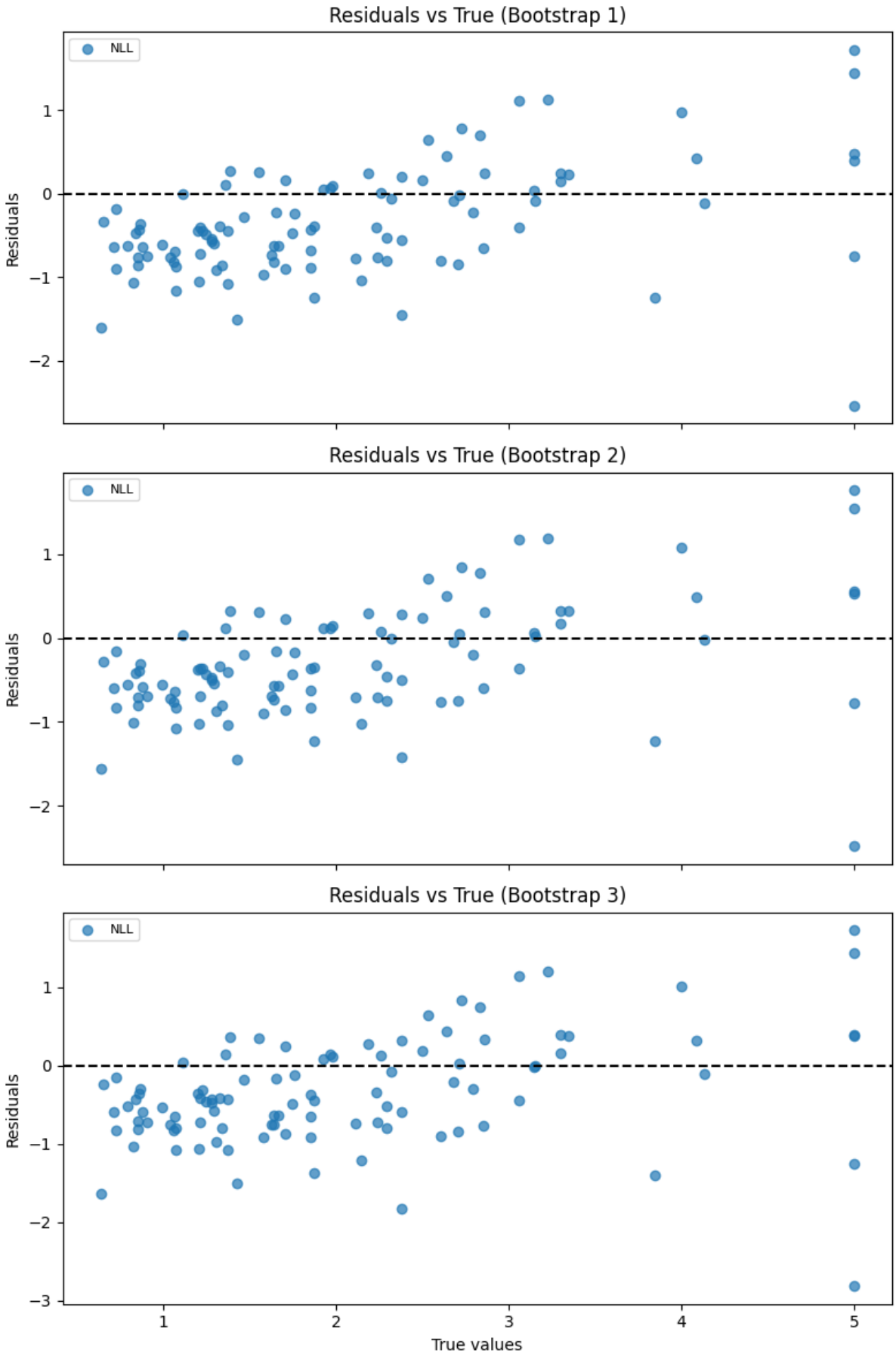

Figure 18: NLL_mean_field_residuals.png