Noname manuscript No.
(will be inserted by the editor)

Criterion Constrained Bayesian Hierarchical Models

Qingying Zong · **Jonathan R. Bradley**

Received: date / Accepted: date

Abstract The goal of this article is to improve the predictive performance of a Bayesian hierarchical statistical model by incorporating a criterion typically used for model selection. In this article, we view the problem of prediction of a latent real-valued mean as a model selection problem, where the candidate models are from an uncountable infinite set (i.e., the parameter space of the mean represents the candidate set of models). Specifically, we select a subset of our Bayesian hierarchical statistical model's parameter space with high predictive performance (as measured by a criterion). Explicitly, we truncate the joint support of the data and the parameter space of a given Bayesian hierarchical model to only include small values of the covariance penalized error (CPE) criterion. The CPE is a general expression that contains several information criteria as special cases. Simulation results show that as long as the truncated set does not have near-zero probability, we tend to obtain a lower squared error than Bayesian model averaging. Additional theoretical results are provided as the foundation for these observations. We apply our approach to a dataset consisting of American Community Survey (ACS) period estimates to illustrate that this perspective can lead to improvements in a single model.

Keywords Bayesian hierarchical model · Markov chain Monte Carlo · Posterior predictive p-value · Information theory · Gaussian Processes

Qingying Zong
Department of Statistics, Florida State University,
117 N. Woodward Ave, Tallahassee, Fl 32306,
E-mail: qz16b@fsu.edu

Jonathan R. Bradley
Department of Statistics, Florida State University,
117 N. Woodward Ave, Tallahassee, Fl 32306,
E-mail: bradley@stat.fsu.edu

1 Introduction

Statistical model selection using a criterion often involves selecting a single model (see, for example, Akaike 1973). In this article, we make use of selection criteria to improve the predictive performance of a given statistical model instead of using it to select a model from B candidates. In particular, the selection criterion is used to select a subset of the parameter space. The selected set is chosen so that the values in the set have high predictive performance as measured by a criterion.

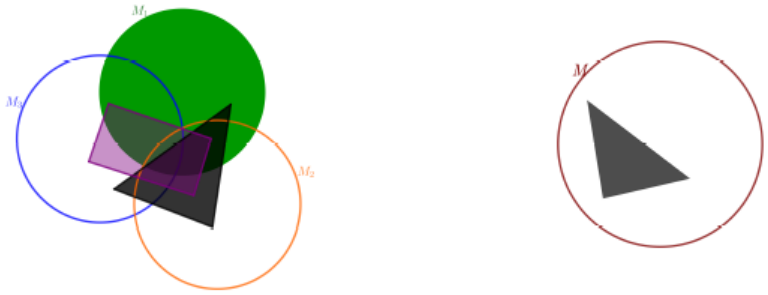

Fig. 1: M_1, M_2, and, M_3 are abstract representations of the parameter space of three candidate statistical models. Their union is the "expanded parameter space," the black shaded triangle region represents values associated with high predictive performance, and the purple shaded rectangle provides an abstract representation of the posterior distribution based on the sparsity sparsity-inducing priors. The right panel is the setting, where we select a subset of the parameter space with high predictive performance (according to a criterion) of a single model.

This perspective to truncate the support to values with high predictive performance (according to a criterion) is similar to the use of sparsity-inducing priors. The difference with our approach is that we are selecting a region we believe to have high predictive performance through the use of selection criteria, where sparsity-inducing priors use Bernoulli (e.g., see Ishwaran and Rao, 2005, for the spike and slab prior) or "near Bernoulli" priors (e.g., see Carvalho et al., 2009, for the horseshoe prior) to effectively select a subset of an expanded parameter space. We provide Figure 1 to show how the procedure for prediction is motivated by methods for model selection. In the left panel of Figure 1, we give a representation of traditional model selection (green circle), sparsity-inducing priors (purple rectangle), and our approach (black triangle). In the right panel of Figure 1, we give the setting we are primarily interested in, which is taking a given statistical model, say M_1, and improving the predictive performance by constraining the parameter space to a high predictive performance region.

In this article, our goal is to improve the predictive performance of a Bayesian hierarchical model using selection criteria (i.e., the strategy described in Figure 1). Specifically, we propose defining the joint distribution of the data and parameters to

be proportional to,

$$\text{Likelihood} \times \text{prior} \times I(\text{criterion} < \kappa), \tag{1}$$

where $I(\cdot)$ is the indicator function and $\kappa > 0$ is prespecified. This incorporates a criterion directly into the support of the model, and we will show (theoretically and empirically) that one can obtain gains in predictive performance using specifications of (1). This is a novel strategy to incorporate a selection criterion into a Bayesian hierarchical model.

The use of selection criteria in Bayesian hierarchical models has a rich history. There are methods that average models based on selection criteria (e.g., Burnham and Anderson, 2003; Chen and Huang, 2012, among others). Similarly, Wasserman (2000) estimates a quantity under each candidate model and then averages the estimates with respect to how probable each model is, which bears some similarity to the Bayesian model averaging (BMA, Hoeting et al., 1999) that defines weights according to the posterior probability of each candidate model. One criticism of model averaging approaches is that they include poor-performing models in their averages, where sparsity-inducing prior similar to ours can remove these models.

An important issue that is avoided by (1) is that the use of selection criteria has sampling variability that is not incorporated directly into the selected model. For example, consider the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). The AIC selects the model that minimizes an approximated Kullback-Leibler divergence to the true data generating process (see discussion in Acquah 2010). The BIC is designed to approximate a Bayes factor (see discussion in Acquah 2010) and therefore is often used when there are random effects. The values of these criteria are functions of the dataset itself, and hence, have sampling variability. Thus, as new data are generated, the "best" model may change. This is true for a majority of the selection criteria used in the literature. For example, Vaida and Blanchard (2005)'s conditional AIC penalizes the in-sample error (a function of data) using the effective degrees of freedom (Hodges and Sargent, 2001) and has sampling variability. Similarly, Huang and Chen (2007) select spatial models where the penalty is based on the generalized degrees of freedom (Ye, 1998), which again has sampling variability.

The approach in (1) accounts for the sampling variability through the support of the model, which is a general strategy developed by Yekutieli (2012). Yekutieli (2012) addresses sampling variability for Bayesian models by truncating the support of the data to always produce the same selected covariates in a model selection problem. Truncating the support of the data in this manner removes the sampling variability of the selected parameters, by removing this variability in the data generating mechanism. Our approach to truncation uses perspective to remove the variability of the criterion. Criteria such as AIC, Stein's unbiased risk estimate (SURE), and Mallows' Cp can be interpreted as a type of covariance penalized error (CPE) (Efron, 2004; Efron and Hastie, 2016; Tibshirani and Rosset, 2018; Holbrook et al., 2020). As such, we use this general expression when adopting Yekutieli (2012)'s truncation approach to incorporate a criterion into the Bayesian hierarchical model to improve prediction. Thus, the CPE is not treated as a plug-in estimator and instead is used to constrain

the support of a Bayesian hierarchical model. Our method truncates the data and parameter space based on a selection criterion, which incorporates the criterion directly into the model in a principled way (i.e., through the support of the statistical model). Consequently, we refer to our model as the truncated CPE model. In this article, we choose the CPE, however, our constrained Bayesian perspective is flexible enough to incorporate several other criteria.

The truncated CPE model can be considered as a type of Calibrated Bayes (CB) (Box, 1980; Rubin, 1984; Little, 2006, 2011). CB uses Bayesian methods for inference and uses frequentist methods for model development and assessment (Little, 2006, 2011). That is, the Bayesian model is "calibrated" to have a frequentist property. Typically, in CB, one selects models that produce posterior credibility intervals with (approximately) their nominal frequentist coverage under repeated sampling (Little, 2012). Our method, however, calibrates a Bayesian hierarchical model to have small values of CPE, which is another frequentist property.

We provide a result that shows *every* proper Bayesian model can be expressed as a type of truncated CPE model. In particular, one can augment a Bayesian model with a uniformly distributed random variable (in a manner similar to Damien et al., 1999) so that the posterior distribution can be expressed as a truncated CPE model. In our method, we explicitly make the truncation tighter, which can lead to better predictive performance. That is, we analytically show that the truncated CPE model leads to better predictions in terms of squared error than the corresponding untruncated Bayesian hierarchical model taken to be a BMA model for normal data. Recently Torkashvand et al. (2016) has shown similar positive results empirically for a different constrained Bayesian model in a specific functional/process modeling setting. That is, we can improve upon the predictive performance (as measured by a criterion) of a single model as long as the truncating event is admissible. The size of the truncating event also has important practical implications. In particular, we can compare models through acceptance rates when implementing a Gibbs sampler, where we reject when the CPE is "too large". That is, one model may reject more parameter values than another because its parameter space implies large values of CPE.

The remainder of this paper is organized as follows. In Section 2, we review several approaches for model selection that motivate our proposed model whose goal is prediction. In Section 3, we introduce the truncated CPE model and provide theoretical support. In particular, we show that every proper Bayesian model can be interpreted as a truncated CPE model and show that our specifications can lead to higher predictive performance than a given Bayesian hierarchical model for normal data in terms of squared error under reasonable conditions. We illustrate this through a simulation study in Section 4. In Section 5, we analyze American Commutation Survey (ACS) period estimates over census tracts in central Missouri. We apply our approach to space-time change of support (Bradley et al., 2015). This example demonstrates a case where only a single candidate Bayesian model is available, and that one can obtain better out-of-sample performances using the proposed truncated CPE model. We end with a discussion in Section 6. For ease of exposition, proofs are provided in the Appendix.

2 Motivation: Review of Model Selection Approaches

As discussed in the Introduction, our inferential goal is prediction and not model selection. However, we are highly motivated by existing approaches used in model selection. Thus, in Section 2, we provide reviews of several model selection procedures that are relevant to our proposed model. In Section 3, we introduce our truncated CPE model.

2.1 A Review of Covariance Penalized Errors

Denote the observed data with $Y_1, Y_2, \cdots, Y_n$ and the n-dimensional observed data vector with $\mathbf{y} \equiv (Y_1, Y_2, ..., Y_n)'$. We assume the following additive model

$$Y_i = \mu_i + \varepsilon_i; \;\; i = 1, 2, \ldots, n, \tag{2}$$

where $\mu_i \in \mathbb{R}$ is unknown; the ε_i's are mean zero, variance $\sigma_i^2 > 0$, and are independent of ε_j and μ_k for $j \neq i$, and $k = 1, ..., n$. We refer to (2) as the "additive model". For simplicity, we will use a shorthand for set, such as $\{\sigma_i^2\}$ represents $\{\sigma_1^2, \ldots, \sigma_n^2\}$. Now, suppose there are B candidate models to predict $\boldsymbol{\mu}$. These models all result in different predictors for $\boldsymbol{\mu}$, which we denote with $\hat{\boldsymbol{\mu}}_b : \mathbb{R}^n \to \mathbb{R}^n$; $b = 1, 2, ..., B$. For example, $\hat{\boldsymbol{\mu}}_b \equiv (\hat{\mu}_{1,b}, ..., \hat{\mu}_{n,b})'$ may be the posterior mean of $\boldsymbol{\mu}$ using model b.

Let $(\mu_i - \hat{\mu}_{i,b})^2$ be the true prediction error measured by squared error discrepancy for component i. The true prediction error is not observed because $\{\mu_i\}$ is not observed. In practice, one can more easily compute the in-sample error for component i, $(Y_i - \hat{\mu}_{i,b})^2$. Efron (1983, 1986, 2004) derives an important expression of the prediction error,

$$E[(\mu_i - \hat{\mu}_{i,b})^2] + \sigma_i^2 = E[(Y_i - \hat{\mu}_{i,b})^2 + 2cov(\hat{\mu}_{i,b}, Y_i)]; \;\; i = 1, \ldots, n, \; b = 1, \ldots, B, \tag{3}$$

where the expectation is taken with respect to $\mathbf{y}|\boldsymbol{\mu}, \{\sigma_i^2\}$. Equation (3) leads to the following criterion referred to as the CPE,

$$CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}_b) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_{i,b})^2 + 2 \sum_{i=1}^{n} cov(\hat{\mu}_{i,b}, Y_i); \;\; b = 1, \ldots, B, \tag{4}$$

which is unbiased for

$$\sum_{i=1}^{n} E(\mu_i - \hat{\mu}_{i,b})^2 + \sum_{i=1}^{n} \sigma_i^2.$$

These fundamental results show that, on average, the apparent in-sample error needs to be corrected by a penalty (i.e., a covariance, hence the name CPE) to be an unbiased estimation for the overall prediction error (Efron and Hastie, 2016). This CPE criterion is well-known to be a general expression of several criteria introduced in the literature. For example, AIC, Mallow's C_p (Mallows, 1973), and Stein's unbiased risk estimator (Stein, 1981) are all special cases of the CPE (see Efron, 2004; Efron and Hastie, 2016; Tibshirani and Rosset, 2018; Holbrook et al., 2020 for discussions).

This criterion, while very useful, has a limitation that we focus on in this article. Namely, the CPE is a statistic (more formally a method of moments estimate

of $\sum_{i=1}^{n} E(\mu_i - \hat{\mu}_{i,b})^2 + \sum_{i=1}^{n} \sigma_i^2$), and hence has sampling variability. This sampling variability can have an effect on the chosen models. Consider the following simulated example to illustrate the issue of sampling variability in selection criteria:

- Simulate 1000 replicates with $n = 200$.
- Consider a multiple regression model with $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$, each a 200-dimensional vector, where the elements are chosen independently from a standard normal distribution.
- Let the 200×4 matrix $\mathbf{X}_b = [\mathbf{1}_{200}, \mathbf{x}_1\delta_1, \mathbf{x}_2\delta_2, \mathbf{x}_3\delta_3] = (\mathbf{x}'_{1b}, \ldots, \mathbf{x}'_{200b})'$, where δ_i is either zero or one, $\mathbf{1}_{200}$ is a 200-dimensional vector of ones, define $\boldsymbol{\mu} = \mathbf{X}_b\boldsymbol{\beta}$, where the value of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$ is arbitrarily chosen to be $(2, 1, 1, 0)'$.
- For a given 200-dimensional data vector $\mathbf{y}$ with $\sigma_i^2 = \sigma^2$, we consider implementing the following models for $\boldsymbol{\mu}$:

$$\begin{aligned}
I(b=1) &= I(\delta_1 = \delta_2 = \delta_3 = 0) \\
I(b=2) &= I(\delta_1 = 1, \delta_2 = \delta_3 = 0) \\
I(b=3) &= I(\delta_1 = \delta_3 = 0, \delta_2 = 1) \\
I(b=4) &= I(\delta_1 = \delta_2 = 0, \delta_3 = 1) \\
I(b=5) &= I(\delta_1 = \delta_2 = 1, \delta_3 = 0) \\
I(b=6) &= I(\delta_1 = \delta_3 = 1, \delta_2 = 0) \\
I(b=7) &= I(\delta_2 = \delta_3 = 1, \delta_1 = 0) \\
I(b=8) &= I(\delta_1 = \delta_2 = \delta_3 = 1),
\end{aligned} \tag{5}$$

where $I(\cdot)$ is an indicator function. Then let $\hat{\boldsymbol{\mu}}_b$ be the ordinary least squares estimator with eight different choices of covariates based on (5). The Mallow's C_p is given by

$$C_p(\mathbf{y}, \hat{\boldsymbol{\mu}}_b) = \sum_{i=1}^{200} (Y_i - \hat{\mu}_{i,b})^2 + 2\sigma^2 p(b); \;\; b = 1, \ldots, 8,$$

where $p(b)$ is the number of non-zero regression coefficients identified in the model b. Note that for

$$\hat{\boldsymbol{\mu}}_b = \mathbf{X}_b(\mathbf{X}'_b\mathbf{X}_b)^{-1}\mathbf{X}'_b\mathbf{y},$$

we have the covariance in Equation (4) is given by

$$\sum_{i=1}^{200} cov(\mathbf{x}'_{ib}(\mathbf{X}'_b\mathbf{X}_b)^{-1}\mathbf{X}'_b\mathbf{y}, Y_i) = trace(\mathbf{X}_b(\mathbf{X}'_b\mathbf{X}_b)^{-1}\mathbf{X}'_b)\sigma^2 = p(b)\sigma^2,$$

which shows that Mallow's C_p is a special case of the CPE when selecting covariates using the ordinary least squares (e.g., see Efron, 2004; Efron and Hastie, 2016; Tibshirani and Rosset, 2018; Holbrook et al., 2020, among others). Then denote the selected model with

$$\hat{b} = arg \min_{b=1,\ldots,8} C_p(b).$$

From Table 1, we present the proportion of times $\hat{b} = b$ by σ over 1000 independent replicates of the vector $\mathbf{y}$. For each σ, 83% of the time we roughly select the correct

Table 1: The proportion of times $\hat{b} = b$ by σ over 1000 independent replicates of the vector $\mathbf{y}$.

	b							
	1	2	3	4	5	6	7	8
$\sigma = 0.5$	0	0	0	0	83.5%	0	0	16.5%
$\sigma = 1$	0	0	0	0	85.2%	0	0	14.8%
$\sigma = 2$	0	0	0	0	83.0%	0	0	17.0%
$\sigma = 3.5$	0	0.4%	0.4%	0	82.3%	0	0	16.9%

value of $b = 5$, but we consistently (over σ) select the incorrect full model around 17% of time. This is consistent with the literature, where several (but not all) selection criteria tend to select more complicated models (Rao and Wu, 1989; Maraun and Widmann, 2018). This also demonstrates the weakness of the selection criteria discussed in the Introduction. That is, high sampling variability in C_p can lead to choose incorrect models.

2.2 A Review of Bayesian Model Averaging (BMA)

Bayesian model averaging addresses model uncertainty (as demonstrated in Table 1) by directly modeling b with a prior distribution. Let $\pi(b)$ be the prior mass for model b such that $\sum_{j=1}^{B} \pi(b = j) = 1$. Under BMA, inference on the quantity of interest (here is $\boldsymbol{\mu}$), can be obtained through the probability density function (pdf) of $\boldsymbol{\mu} \mid \mathbf{y}$. This can be computed with

$$\pi(\boldsymbol{\mu} \mid \mathbf{y}) = \sum_{j=1}^{B} \pi(\boldsymbol{\mu}|\mathbf{y}, b = j)\pi(b = j|\mathbf{y}), \tag{6}$$

which is a weighted average of the distribution of $\boldsymbol{\mu}$ given each model and data, and the weights are posterior probabilities of the model. The choice of prior specifications for the candidate models have an important impact in practice. Let's revisit the small simulation example in Section 2.1, where notice $b = 1, 2, 3, 4, 6, 7$ in Table 1 are nearly never selected using Mallow's C_p. This leads us to consider the case where $\pi(b = j) = 1/8$ (for all j) and the case

$$\pi(b = j) = \begin{cases} 1/2 & \text{j = 5,8} \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Consider the case $\sigma = 3.5$. Figure 2 contains the histogram of $\sum_{i=1}^{n}(\mu_i - \hat{\mu}_{i,v})^2 - \sum_{i=1}^{n}(\mu_i - \hat{\mu}_{i,w})^2$, where $\hat{\mu}_{i,v}$ is the posterior mean using $\pi(b = j) = 1/8$, and $\hat{\mu}_{i,w}$ is the posterior mean using Equation (7). The majority of values in Figure 2 are consistently positive, which suggests better predictions when using $\pi(b = j)$ in (7). Here, we can see that the choice of model priors has a clear impact by informally using Table 1 (or the CPE) to reduce the parameter space (of b). The improvements, by using (7), are not surprising. Poor-performing values in the parameter space are averaged in BMA when $\pi(b = j) = 1/8$, but are not averaged when using $\pi(b = j)$ in

(7). The prior distribution $\pi(b)$ in (7) informally incorporates CPE, which is formed via Table 1, but does not account for the sampling variability of CPE. Thus, our goal is to formally incorporate CPE, by accounting for the variability of CPE.

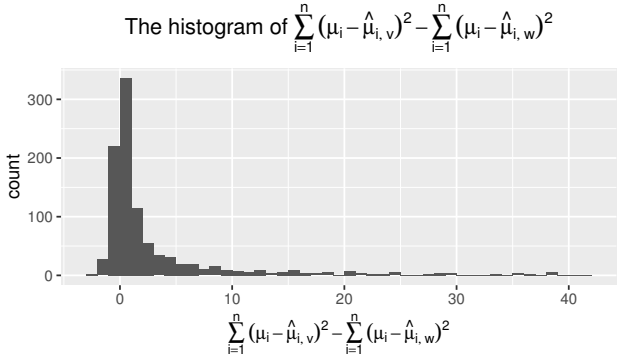

Fig. 2: The histogram of $\sum_{i=1}^{n}(\mu_i-\hat{\mu}_{i,v})^2-\sum_{i=1}^{n}(\mu_i-\hat{\mu}_{i,w})^2$ (i.e. the difference in squared error) by $\sigma=3.5$ over 1000 independent replicates of the vector $\mathbf{y}$.

3 Methodology: Improving the Predictive Performance of a Given Bayesian Model Using CPE

In Section 3, we describe how we use the existing perspectives to improve the predictions of a given Bayesian hierarchical model using selection criteria (i.e., CPE). In Section 3.1, we state our model. Then in Section 3.2, we show that every proper Bayesian hierarchical model can be written as the same form as our proposed truncated CPE model. Then in Section 3.3, we show that we can improve upon a given Bayesian hierarchical model under reasonable conditions for Gaussian data. Finally, in Section 3.4, we give an interpretation of the constrained posterior distribution assuming the unconstrained Bayesian hierarchical model.

3.1 The Proposed Model: The Truncated CPE Model

The statistical model we use for inference is defined as the product of the following conditional and marginal probability density functions:

$$\pi(\mathbf{y},\boldsymbol{\mu},\boldsymbol{\theta}_b,b|\kappa)\propto f(\mathbf{y}|\boldsymbol{\mu},\{\sigma_i^2\})\pi(\boldsymbol{\mu}|\boldsymbol{\theta}_b,b)\pi(\boldsymbol{\theta}_b|b)\pi(b)I\{CPE<\kappa\};\ \ b=1,\ldots,B, \tag{8}$$

where $I(\cdot)$ is the indicator function, $f(\mathbf{y}|\boldsymbol{\mu},\{\sigma_i^2\})$ is the data model (if is referred to as the likelihood when written as a function of $\boldsymbol{\mu}$ and $\{\sigma_i^2\}$), where the i-th component of $\mathbf{y}=(Y_1,\ldots,Y_n)'$ have known variance σ_i^2, $\boldsymbol{\theta}_b$ is the generic real-valued parameter vector, $\pi(\boldsymbol{\mu}|\boldsymbol{\theta}_b,b)$ is the process model, $\pi(\boldsymbol{\theta}_b|b)$ is the prior for $\boldsymbol{\theta}_b$, $\pi(b)$ is the prior probability of the model b, the value of $\kappa>0$ is a pre-specified real value and is

crucial for our model (see Section 3 and 4 for more discussion). In general $\pi(\boldsymbol{\mu}|\boldsymbol{\theta}_b, b)$ is not restricted to be a Gaussian linear model as is used for illustration in Section 2.1. The model in Equation (8) allows for many special cases. For example, in our application $B = 1$, and we show that (8) can lead to improvements in a single model.

We introduce $\hat{\boldsymbol{\mu}}$ into our notation for $CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b))$ when using CPE to estimate the overall prediction error, where $\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b)$ is a generic predictor of $\boldsymbol{\mu}$ (we give our specification of $\hat{\boldsymbol{\mu}}$ in (11)). We also introduce the possible functional dependence on $\boldsymbol{\theta}_b$ and b into our notation for $\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b)$. This strategy is inspired by Yekutieli (2012)'s method incorporating an estimator into the support of the model. His selection-adjusted Bayes inference method involves truncating the support of the data. Our method differs because it involves truncating the support based on CPE. The $CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b))$ is directly incorporated into the model through $I\{CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b)) < \kappa\}$, and hence, does not have unaccounted for variability in Equation (8). Specifically, we mean that the joint posterior distribution of our model is given by

$$f(\boldsymbol{\mu}, \boldsymbol{\theta}_b, b|\mathbf{y}, \kappa) \tag{9}$$
$$= \frac{f(\mathbf{y}|\boldsymbol{\mu}, \{\sigma_i^2\})\pi(\boldsymbol{\mu}|\boldsymbol{\theta}_b, b)\pi(\boldsymbol{\theta}_b|b)\pi(b)I\{CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b)) < \kappa\}}{\sum_{q=1}^{B} \int\int_{CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_q, q)) < \kappa} f(\mathbf{y}|\boldsymbol{\mu}, \{\sigma_i^2\})\pi(\boldsymbol{\mu}|\boldsymbol{\theta}_q, q)\pi(\boldsymbol{\theta}_q|q)\pi(q)I\{CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_q, q)) < \kappa\} d\boldsymbol{\mu} d\boldsymbol{\theta}_q},$$

for $b = 1, \ldots, B$, which does not treat $CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b))$ as a plug-in estimator (causing unaccounted for variability) but rather uses CPE to constrain the support of a Bayesian hierarchical model. Equation (9) is well defined provided that κ is not specified so small that the integral is equal to zero. More empirically motivated discussions on the choice of κ are given by Section 4 and 5.

To our knowledge, Yekutieli (2012) is the first to avoid the issue of unchecked sampling variability by incorporating an estimator into the support of the model in a Bayesian context (effectively changes the data generating mechanism). Yekutieli (2012) uses this approach to adjust the posterior distribution for selecting a model, which is different from our goal of using CPE to improve prediction in a Bayesian hierarchical model. This general approach is not unique to our setting and has been done in several different areas, including approximate methods for the likelihood conditional on selection (Panigrahi et al., 2016), false discovery rates/multiple testing, false coverage rates, and empirical Bayesian analysis (Benjamini, 2010; Benjamini et al., 2009; Catelan et al., 2010; Zhao and Hwang, 2012; Bradley and Zong, 2021, among others).

The joint posterior distribution in Equation (9) explicitly shows how we combine BMA and classical model selection criteria to improve prediction of the BMA. Specifically, a prior is placed on the model b, and the parameter space of this model is constrained to a "good predictive set" by using $I\{CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b)) < \kappa\}$. That is, for example, $I\{CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b)) < \kappa\}$ subsets the three models represented by circles to the black region in Figure 1.

3.2 Comparison to Untruncated Bayesian Hierarchical Models

A key point that motivates the truncated CPE model in (8) is that *every* proper Bayesian hierarchical model can be interpreted as a type of truncated CPE model. Specifically, one can augment any proper Bayesian hierarchical model, using the technique introduced in Damien et al. (1999), so that the CPE is bounded above. We formally state this result below in Theorem 1.

Theorem 1 *Suppose* $\pi(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\theta}_b,\boldsymbol{\theta}_D,b) = f(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\theta}_D)h(\mathbf{y},\boldsymbol{\theta}_b,b)$*,* $\pi(b) > 0$*,* $\sum_b \pi(b) = 1$*,* $\pi(\boldsymbol{\mu}|\boldsymbol{\theta}_b,b)$*,* $\pi(\boldsymbol{\theta}_D)$*, and* $\pi(\boldsymbol{\theta}_b|b)$ *are proper densities,* $h(\mathbf{y},\boldsymbol{\theta}_b,b)$ *is a non-negative real-valued function such that* $0 < \int f(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\theta}_D)h(\mathbf{y},\boldsymbol{\theta}_b,b)d\mathbf{y} < \infty$*,* $f(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\theta}_D)$ *is a proper model with mean* $\boldsymbol{\mu}$*, and* $\boldsymbol{\theta}_D$ *is a generic finite dimensional real-valued parameter vector. Then, for u uniformly distributed on* $(0,1)$ *and* $r = \frac{CPE(\mathbf{y},\hat{\boldsymbol{\mu}})}{2log(f(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\theta}_D))} + 1$*, we have that the posterior distribution*

$$\pi(\boldsymbol{\mu},\boldsymbol{\theta}_D,\boldsymbol{\theta}_b,b|\mathbf{y}) = \frac{1}{\pi(\mathbf{y})}\int_0^1 \pi(\boldsymbol{\mu},\boldsymbol{\theta}_D,\boldsymbol{\theta}_b,b,\mathbf{y}|u)\pi(u)du,$$

where $\pi(\mathbf{y})$ *is the density for the marginal distribution of the data,*

$$\begin{aligned}&\pi(\mathbf{y},\boldsymbol{\theta}_D,\boldsymbol{\mu},\boldsymbol{\theta}_b,b|u) \qquad (10)\\ =&f(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\theta}_D)^r\pi(\boldsymbol{\theta}_D)\pi(\boldsymbol{\mu}|\boldsymbol{\theta}_b,b)\pi(\boldsymbol{\theta}_b|b)\pi(b)I\{CPE(\mathbf{y},\hat{\boldsymbol{\mu}}) < \kappa^*\}h(\mathbf{y},\boldsymbol{\theta}_b,b),\end{aligned}$$

and $\kappa^* = -2log(u)$.

Proof : See Appendix B.

When $h(\mathbf{y},\boldsymbol{\theta}_b,b) \equiv 1$ then Equation (10) in Theorem 1 shows that any generic Bayesian hierarchical model is a truncated CPE model, where the CPE is truncated above by κ^* in (10). That is, Equation (10) with $h(\mathbf{y},\boldsymbol{\theta}_b,b) \equiv 1$ is directly analogous to the truncated CPE model in (8).

Setting $h(\mathbf{y},\boldsymbol{\theta}_b,b) = I\{CPE(\mathbf{y},\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b,b)) < \kappa\}$ in Theorem 1 implies that the proposed truncated CPE model in (8) truncates the CPE above by $min(\kappa^*,\kappa)$, since the product

$$I\{CPE(\mathbf{y},\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b,b)) < \kappa^*\}I\{CPE(\mathbf{y},\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b,b)) < \kappa\} = I\{CPE(\mathbf{y},\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b,b)) < min(\kappa^*,\kappa)\}.$$

Thus, one can interpret our truncated CPE model in (8) as a minor modification to any proper Bayesian hierarchical model, where one replaces the implicit bound on the CPE (i.e., κ^*) with $min(\kappa^*,\kappa)$. Changing κ^* to $min(\kappa^*,\kappa)$ has two important consequences. First, changing κ^* (or $h(\mathbf{y},\boldsymbol{\theta}_b,b) \equiv 1$) to $min(\kappa^*,\kappa)$ (or $h(\mathbf{y},\boldsymbol{\theta}_b,b) = I(CPE(\mathbf{y},\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b,b)) < \kappa)$) changes the data model from a normal distribution, for example, to a type of truncated normal distribution. However, as shown in Theorem 1 the implied posterior for either choice of data model (truncated or untruncated) stays the same (i.e., Equation (10) and (8) are analogous). Furthermore, changing the distribution of the data is reasonable in our criterion-based setting, as we are allowing for the possibility of model misspecification. Second, changing κ^* to $min(\kappa^*,\kappa)$ can lead to smaller squared prediction error, which we discuss in detail in the subsequent Section 3.3.

3.3 Squared Prediction Error Properties

Constraining a Bayesian hierarchical model based on the CPE implicitly constrains the unobserved $\sum_{i=1}^{n}(\mu_i-\hat{\mu}_i)^2$. To investigate this, consider the setting where the predictor $\hat{\boldsymbol{\mu}}$ is specified to be the Best Linear Unbiased Prediction (BLUP) (Ravishanker and Dey, 2020),

$$\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b,b)=\mathbf{m}(\boldsymbol{\theta}_b,b)+\boldsymbol{\Sigma}(\boldsymbol{\theta}_b,b)\boldsymbol{\Sigma}_Y^{-1}(\boldsymbol{\theta}_b,b)\{\mathbf{y}-\mathbf{m}(\boldsymbol{\theta}_b,b)\} \tag{11}$$

where $\mathbf{m}(\boldsymbol{\theta}_b,b)$ is the mean of the process model $\pi(\boldsymbol{\mu}|\boldsymbol{\theta}_b,b)$, $\boldsymbol{\Sigma}(\boldsymbol{\theta}_b,b)$ is the process model's covariance, and $\boldsymbol{\Sigma}_Y(\boldsymbol{\theta}_b,b)$ is the covariance of $\mathbf{y}$ from $f(\mathbf{y}|\{\sigma_i^2\})$, which equals to $\int f(\mathbf{y}|\boldsymbol{\mu},\{\sigma_i^2\})\pi(\boldsymbol{\mu}|\boldsymbol{\theta}_b,b)d\boldsymbol{\mu}$. Notice $\pi(\boldsymbol{\mu}|\boldsymbol{\theta}_b,b)$ is not necessarily restricted to be a linear model as is used for illustration in Section 2.1. The i-th component of $\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b,b)$ is denoted with $\hat{\boldsymbol{\mu}}_i(\boldsymbol{\theta}_b,b)$. The CPE for this specification of $\hat{\boldsymbol{\mu}}$ is computed using $\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b,b)$ as follows (Efron, 2004):

$$CPE(\mathbf{y},\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b,b))=\{\mathbf{y}-\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b,b)\}'\{\mathbf{y}-\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b,b)\}+2trace\{\boldsymbol{\Sigma}\boldsymbol{\Sigma}_Y^{-1}\},$$

where the penalty term is referred to as the effective degrees of freedom (Hodges, 2013). Then the following result shows that κ can be chosen in a manner that leads to smaller squared prediction error.

Theorem 2 *Assume* $\mathbf{y}|\boldsymbol{\mu}^{(t)},\{\sigma_i^2\}\sim N(\boldsymbol{\mu}^{(t)},\mathbf{D})$, *where* $\mathbf{D}=diag(\sigma_1^2,\ldots,\sigma_n^2)$, *the true unobserved mean* $\boldsymbol{\mu}^{(t)}=(\mu_1^{(t)},\ldots,\mu_n^{(t)})'$. *Let* $\hat{\boldsymbol{\mu}}_{tc}(\kappa)=(\hat{\mu}_{1,tc}(\kappa),...,\hat{\mu}_{n,tc}(\kappa))'$ *be the element-wise posterior median of* $\hat{\boldsymbol{\mu}}$ *using the model in (9). That is, let* $\hat{\mu}_{i,tc}(\kappa)$ *be the value such that*

$$0.5=\int_{-\infty}^{\hat{\mu}_{i,tc}}\pi(\hat{\mu}_i|\mathbf{y},\kappa)d\hat{\mu}_i, \tag{12}$$

where $\pi(\hat{\mu}_i|\mathbf{y},\kappa)=\sum_{b=1}^{B}\int_R\pi(\boldsymbol{\theta}_b,b|\mathbf{y},\kappa)d\boldsymbol{\theta}_b,\pi(\boldsymbol{\theta}_b,b|\mathbf{y},\kappa)=\int f(\boldsymbol{\mu},\boldsymbol{\theta}_b,b|\mathbf{y},\kappa)d\boldsymbol{\mu}$, *and* $R=\{(\boldsymbol{\theta}_b,b):\hat{\mu}_i=\hat{\mu}_i(\boldsymbol{\theta}_b,b)\}$. *Then,*

$$E\Big\{\sum_{i=1}^{n}[\mu_i^{(t)}-\hat{\mu}_{i,tc}(\kappa)]^2\Big\}<E\Big\{\sum_{i=1}^{n}(\mu_i^{(t)}-\hat{\mu}_{i,m})^2\Big\}, \tag{13}$$

where $\hat{\mu}_{i,m}$ *is a generic real-valued predictor of* $\mu_i^{(t)}$, *the expectation is taken with respect to* $\mathbf{y}|\boldsymbol{\mu}^{(t)},\{\sigma_i^2\}$, *and* $\kappa=E\{\sum_{i=1}^{n}(\mu_i^{(t)}-\hat{\mu}_{i,m})^2\}+\sum_{i=1}^{n}\sigma_i^2$. *We assume this choice of* κ *produces a model in (9) that is proper.*

Proof : See Appendix B.

Theorem 2 shows that $\hat{\boldsymbol{\mu}}_{tc}(\kappa)$ improves the unobserved squared prediction error of any predictor $\{\hat{\mu}_{i,m}\}$ given the conditions in Theorem 2. For example, $\hat{\boldsymbol{\mu}}$ can arise from a standard Bayesian mixed effects model, and $\{\hat{\mu}_{i,m}\}$ may be the predictor from a Bayesian hierarchical model (possibly with smaller squared error than $\hat{\boldsymbol{\mu}}$). Hence, our final predictor $\hat{\boldsymbol{\mu}}_{tc}(\kappa)$ modifies the posterior distribution for an arguably simple predictor $\hat{\boldsymbol{\mu}}$ (through truncation) to have smaller squared error than a possibly more complicated $\{\hat{\mu}_{i,m}\}$. Theorem 2 is extremely general because no assumptions are placed on the true unobserved $\boldsymbol{\mu}^{(t)}$ to obtain improvements in squared error (i.e.,

(13)). That is, a semi-parametric assumption (i.e., $\mathbf{y}|\boldsymbol{\mu}^{(t)},\{\sigma_i^2\} \sim N(\boldsymbol{\mu}^{(t)},\mathbf{D})$), is all that is needed for (13) to hold. We say the assumption $\mathbf{y}|\boldsymbol{\mu}^{(t)},\{\sigma_i^2\} \sim N(\boldsymbol{\mu}^{(t)},\mathbf{D})$, is semi-parametric because no parametric assumptions are placed on $\boldsymbol{\mu}^{(t)}$ in Theorem 2. This perspective follows the same strategy/philosophy of the semi-parametric conditional AIC literature (e.g., Vaida and Blanchard, 2005, for an early reference). The conditional Gaussian assumption is both standard and reasonable in Bayesian statistics and will likely continue to be a reasonable assumption since data are often recorded as averages and the central limit theorem can be applied. The current landscape of complex Bayesian hierarchical models including multivariate spatial data, multiscale spatial data (e.g., see our Section 5 for example), gradients, Dirichlet process mixture model, data fusion/assimilation, the analysis of high dimensional data, and other areas have been developed under a similar Gaussian assumption (e.g., see Gelfand and Schliep, 2016, for a thorough review). Finally, unlike other Bayesian strategies using CPE (i.e., see the plug-in strategies in Section 2.1), our approach is unaffected by sampling variability of CPE. That is, (13) holds even through CPE has sampling variability.

Also, Theorem 2 shows that, on average, our model is restricted to a good predictive performing set, where "good predictive performance" is defined as $CPE(\mathbf{y},\hat{\boldsymbol{\mu}}) < E\{\sum_{i=1}^n(\mu_i^{(t)}-\hat{\mu}_{i,m})^2\}+\sum_{i=1}^n\sigma_i^2$. This result cannot be directly used in practice since for the term, $E\{\sum_{i=1}^n(\mu_i^{(t)}-\hat{\mu}_{i,m})^2\}+\sum_{i=1}^n\sigma_i^2$, the expectation is taken with respect to $f(\mathbf{y}|\boldsymbol{\mu}^{(t)},\{\sigma_i^2\})$ and the true model for $\boldsymbol{\mu}^{(t)}$ is assumed unknown. By choosing κ, we implicitly choose $\{\hat{\mu}_{i,m}\}$ in our framework. However, Theorem 2 does suggest a choice of κ exists that can lead to good predictive results. As a result, in practice several values of κ are considered, where small values would imply a better prediction error. However, one should keep in mind the admissibility of the set $\{CPE[\mathbf{y},\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b,b)] < \kappa\}$ when choosing small values of κ (e.g., $\kappa = 0$ would be inadmissible). Of course, the value of κ can not be chosen to be so small that one concentrates on a set with zero posterior mass. More explicitly, $\kappa > \inf_{(\mathbf{y},\theta_b,b)\in\delta}\{CPE(\mathbf{y},\theta_b,b)\}$, where δ is the joint support of the data and all parameters of the unconstrained model. As a simple degenerate example, consider $Y \mid p \sim Bernoulli(p), \hat{p} = (0.5+Y)/(0.5+2), \pi(p=0.6)=0.5$, and $\pi(p=0.2)=1/2$. Then $CPE(Y=1,\hat{p})=1.88$, and $CPE(Y=0,\hat{p})=3.79$. If $\kappa < 1.88$, our approach will fail in this degenerate example. One should also be concerned when κ is too large. In the BMA setting, the median model is known to be optimal from a predictive standpoint (Barbieri and Berger, 2004; Farcomeni, 2010), and hence, if κ is large enough that the CPE of the median model is contained in the support then there would be no need to enforce a constraint.

Equation (13) in Theorem 2 allows one the flexibility to consider several assumptions on $\pi(\boldsymbol{\mu}|\boldsymbol{\theta}_b,b)$. For example, one can consider a random process/function representation. A random process/function is different from a random vector, albeit clearly related. A random process/function is different from a random vector, albeit clearly related. A random process $\mu(\mathbf{s})$ defines a random vector $(\mu(\mathbf{s}_1),\ldots,\mu(\mathbf{s}_n))'$ for any given collection of locations $\mathbf{s}_j \in \mathbb{R}^d, j = 1,\ldots,n$. To state that our approach can be interpreted from a random process/function perspective one needs to define how the locations arise (i.e., the spatial domain and intensity function). This is a standard

consideration in Kriging and optimal prediction, for example, see Huang and Chen (2007, cf. Theorem 4), where they assume locations are observed at random in an open subset of the real numbers when using a CPE for model selection in spatial statistics. See Appendix C for more details.

3.4 An Interpretation of the Constrained Posterior Distribution Assuming the Unconstrained Bayesian Hierarchical Model

The constrained Bayesian hierarchical model is non-traditional, and thus, it would be useful to give an interpretation assuming the traditional (unconstrained) Bayesian model is true. That is, suppose the data generating mechanism is the following,

$$\mathbf{y} \mid \boldsymbol{\mu}, \{\sigma_i^2\} \sim N(\boldsymbol{\mu}, \mathbf{D}), \tag{14}$$

and suppose we assume the following prior distribution

$$\pi(\boldsymbol{\mu}, \boldsymbol{\theta}_b, b) = \pi(\boldsymbol{\mu} \mid \boldsymbol{\theta}_b, b)\pi(\boldsymbol{\theta}_b \mid b)\pi(b), \tag{15}$$

similar to Equation (8), where we now remove the truncation. Suppose we "choose to observe" the event $I\{CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b)) < \kappa\} = 1$ in addition to $\mathbf{y}$. The event is an unconventional observation since it can not be observed in practice. That is, we can not simultaneously observe $\mathbf{y}, \boldsymbol{\theta}_b$ and b to ensure $I\{CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b)) < \kappa\} = 1$. However, the event $I\{CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b)) < \kappa\} = 1$ can be treated as an observation in a Bayesian analysis in practice. That is, in a Bayesian analysis one can use Bayes rule to derive the distribution of parameters given the data. In the same manner, the unconstrained model in (14) and(15) can be used to derive the conditional distribution $f(\boldsymbol{\mu}, \boldsymbol{\theta}_b, b \mid \mathbf{y}, I\{CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b)) < \kappa\} = 1)$, which treats both $\mathbf{y}$ and the event $I\{CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b)) < \kappa\} = 1$ as observed since both are given. That is, when assuming the unconstrained Bayesian hierarchical model in (14) and (15), we obtain the following expression

$$\begin{aligned} &f(\boldsymbol{\mu}, \boldsymbol{\theta}_b, b \mid \mathbf{y}, I\{CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b)) < \kappa\} = 1) \\ &= \frac{f(\mathbf{y}|\boldsymbol{\mu}, \{\sigma_i^2\})\pi(\boldsymbol{\mu}|\boldsymbol{\theta}_b, b)\pi(\boldsymbol{\theta}_b|b)\pi(b) I\{CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b)) < \kappa\}}{\sum_{q=1}^{B} \int\int f(\mathbf{y}|\boldsymbol{\mu}, \{\sigma_i^2\})\pi(\boldsymbol{\mu}|\boldsymbol{\theta}_q, q)\pi(\boldsymbol{\theta}_q|q)\pi(q) I\{CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_q, q)) < \kappa\} d\boldsymbol{\mu} d\boldsymbol{\theta}_q}. \end{aligned} \tag{16}$$

Since $f(\boldsymbol{\mu}, \boldsymbol{\theta}_b, b \mid \mathbf{y}, I\{CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b)) < \kappa\} = 1) = f(\boldsymbol{\mu}, \boldsymbol{\theta}_b, b, \mathbf{y}, I\{CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b)) < \kappa\} = 1) / f(\mathbf{y}, I\{CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b)) < \kappa\} = 1)$, $f(\boldsymbol{\mu}, \boldsymbol{\theta}_b, b, \mathbf{y}, I\{CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b)) < \kappa\} = 1) = f(I\{CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b)) < \kappa\} = 1 \mid \mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\theta}_b, b) f(\mathbf{y} \mid \boldsymbol{\mu}, \{\sigma_i^2\})\pi(\boldsymbol{\mu}, \boldsymbol{\theta}_b, b)$, and the data model $f(I\{CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b)) < \kappa\} = 1 \mid \mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\theta}_b, b) = I\{CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b)) < \kappa\}$ assuming the traditional unconstrained Bayesian model in (14) and (15).

Notice the right-hand-side of (16) is the same as (9). Theorem 2 shows that if we "choose to observe" this additional unconventional datum $I\{CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b)) < \kappa\} = 1$ (i.e., use the unconstrained model in (14) and (15) to produce (16)) we can improve predictions. Since Theorem 2 is specific to prediction, the scientific utility of using $f(\boldsymbol{\mu}, \boldsymbol{\theta}_b, b \mid \mathbf{y}, I\{CPE(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b)) < \kappa\} = 1)$ instead of $f(\boldsymbol{\mu}, \boldsymbol{\theta}_b, b \mid \mathbf{y})$ is restricted to prediction.

4 Simulation Study: Improving the Predictive Performance of a BMA

In this section, we perform an "empirical simulation study." By this, we mean the data generating mechanism is calibrated towards the dataset. This strategy is done in an effort to produce a realistic simulated dataset that differs from the model we fit. This aids in producing realistic simulated data and assessing departures from model assumptions. Thus, we generate data from the following statistical model:

$$\mathbf{y} \sim N(\mathbf{L}, \sigma^2 \mathbf{I}_n), \tag{17}$$

where $\mathbf{I}_n$ is an $n \times n$ identify matrix and $n = 112$.

Let $\mathbf{L} = (L_1, L_2, ..., L_n)'$ be an n-dimensional dataset (`www.biostat.umn.edu/~brad/data2.html`,) consisting of the log thickness of radioactive materials at each of $n = 112$ sites contained within the Radioactive Waste Management Complex region associated with the Idaho National Engineering and Environmental Laboratory. We use covariates A-B Elevation, and Surf Elevation. The value of σ^2 is chosen in a way that controls the signal to noise ratio (SNR). Specifically, we choose *SNR* and solve for σ^2 in $SNR = \frac{1}{111}\sum_{i=1}^{112}(L_i - \bar{L})^2/\sigma^2$, where $\bar{L} = \frac{1}{112}\sum_{i=1}^{112} L_i$. We give our choices for *SNR* when describing our analysis of variance (ANOVA) later in this section.

The model we fit to the simulated data is product of the following distributions:

$$\begin{aligned}
&\text{Data Model: } \mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\theta}_b, b \sim N(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n) I\{CPE[\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b)] < \kappa\} \\
&\text{Process Model: } \boldsymbol{\mu}|\boldsymbol{\theta}_b, b, \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{C}(\phi(b), \tau^2)) \\
&\text{Parameter Model 1: } \boldsymbol{\beta} \sim N(\mathbf{0}_p, 10\mathbf{I}_p) \\
&\text{Parameter Model 2: } \tau^2 \sim IG(1, 0.01) \\
&\text{Parameter Model 3: } \pi(b) = \frac{1}{6}; \qquad b = 1, \ldots, 6
\end{aligned} \tag{18}$$

where $\phi(1) = 10, \phi(2) = 15, \ldots, \phi(6) = 35$, $IG(\cdot)$ is the inverse gamma distribution, $\mathbf{X}$ is a $n \times p$ matrix, $p = 3$ since we take the intercept and the aforementioned 2 covariates into consideration, and $\boldsymbol{\beta}$ are the associated coefficients. The (i, j)-th element of $n \times n$ matrix $\mathbf{C}(\phi(b), \tau^2)$ is specified as $\tau^2 exp(-\phi(b)||\mathbf{s}_i - \mathbf{s}_j||)$, $\sigma^2 > 0$ is assumed as a known value, $||\mathbf{s}_i - \mathbf{s}_j||$ is the Euclidean distance between the i-th and j-th location, $\mathbf{0}_n$ is a n-dimensional zero vector, and $\boldsymbol{\theta}_b \equiv (\boldsymbol{\beta}', \tau^2)'$. In Appendix D, we derive the full-conditional distributions associated with this model. In terms of implementation, each iteration of this Gibbs sampler is not any more complex than standard Gibbs sampling. However, what changes is that the number of iterations increases as our truncation on CPE leads one to an additional rejection step.

We consider two crucial factors that influence $\hat{\boldsymbol{\mu}}$ of them and specify their levels for an analysis of variance (ANOVA) as follows: *SNR* with 3 levels, $SNR = 3, 5, 10$; the values for κ are set equal to the d-th quantile of the set $\{CPE[\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b^{[1]}, b^{[1]})], \ldots, CPE[\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b^{[G]}, b^{[G]})]\}$ for levels $d = 0.1, 0.5, 0.9$, where $\boldsymbol{\theta}_b^{[i]}, b^{[i]}$ are the i-th Markov Chain Monte Carlo (MCMC) replicate for $\boldsymbol{\theta}_b$ and b respectively, and the CPE from the untruncated model is $CPE[\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b)]$; G is the length of the MCMC. We simulate 100 independent replicates of the data vector $\mathbf{y}$ and implement our model as

Table 2: Two-way ANOVA table. The degrees of freedom (DF), sum of squares error (Sum Sq), mean squared error (Mean Sq), F statistics, and P-value are listed. We include two main effects, *SNR* and d, and the interaction between *SNR* and d.

	DF	Sum Sq	Mean Sq	F value	Pr(>F)
SNR	2	1.4417	0.72085	44.1791	$< 2.2 \times 10^{-16}$
d	2	1.8038	0.90192	55.2765	$< 2.2 \times 10^{-16}$
SNR:d	4	0.3221	0.08052	4.9348	0.0006138
Residuals	891	14.5380	0.01632		

well as BMA, both of which are computed using a Gibbs sampler (see Appendix D). We evaluate the models using the sum of squared residuals, and we define as "Response" in our ANOVA, whose form is $\sum_{i=1}^{n}(\mu_i - \hat{\mu}_{i,tc})^2 - \sum_{i=1}^{n}(\mu_i - \hat{\mu}_{i,m})^2$. Notice that this Response can be estimated using the CPE. That is, $CPE[\mathbf{y}, \hat{\boldsymbol{\mu}}_{tc}(\boldsymbol{\theta}_b, b)] - CPE[\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b, b)]$ is unbiased for our "Response." We use an MCMC with the length of $12,000$ and a burn-in of $2,000$ and use trace plots to assess convergence visually for a single replicate of the simulated $\mathbf{y}$.

We analyze the effect of the aforementioned factors SNR and Quantile d on the Response by using an ANOVA with 100 independent replicates of the vector $\mathbf{y}$ per factor level combination. From Table 2, we can see that the main effects and the interaction between them are highly significant. To visualize the main effects and the interaction, we use boxplots and an interaction plot. In Figure 3, we see that as SNR increases, the boxplot for the Response shows less variability, but is centered below zero. Negative values suggest that the truncated model surpasses BMA in terms of squared errors. As Quantile d increases, the boxplot for the Response is less negative for $d = 0.1$ and 0.9 than it is when $d = 0.5$. From Figure 4, it can be seen that the interaction is due to the fact that the slope of the line for $d = 0.5$ is much steeper than the lines for $d = 0.1$ and 0.9. Also, the behavior when $d = 0.9$ is very similar to that of $d = 0.1$. When $SNR = 3$, BMA does not outperform our method when $d = 0.1$ and 0.9 and does considerably worse when $d = 0.5$ in practice. These results conform to intuition. When d approaches 1, there should be no difference between the truncated model and BMA. Following our discussion after Theorem 2, small values of κ may imply inadmissibility, which violates the condition of our theorem.

Based on above results, the values of Response (i.e., $\sum_{i=1}^{n}(\mu_i - \hat{\mu}_{i,tc})^2 < \sum_{i=1}^{n}(\mu_i - \hat{\mu}_{i,m})^2$) for $d = 0.5$ are uniformly less than zero. Thus, we suggest using $d = 0.5$ in practice. When d is 0.1 or 0.9, the values of Response are less than zero, but still less preferable when it comes to sums of squared error as when $d = 0.5$. In practice, one might use an information criterion to choose κ. Therefore, our method does as appear to improve the prediction accuracy with respect to the sum of squared residuals.

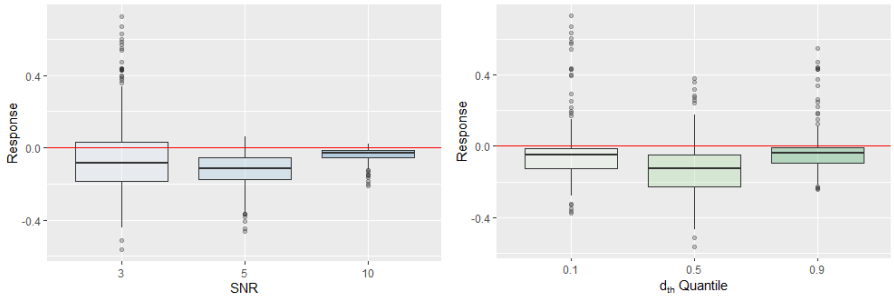

Fig. 3: Main effect plots of *SNR* (left panel) and d-th Quantile (right panel). The horizontal red solid line in each panel stands for Response equal to zero. Response that are negative indicates the truncated model outperforms BMA when it comes to squared errors.

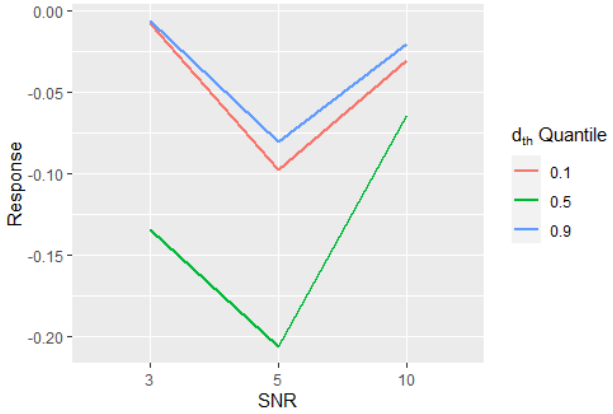

Fig. 4: Plot of average of the Response by *SNR* and d. Response that are negative implies the truncated model outperforms BMA regarding squared errors. The blue solid line indicates $d = 0.9$, the green solid line indicates $d = 0.5$, and the orange solid line indicates $d = 0.1$.

5 Real Data Analysis

The American Community Survey (ACS) is an ongoing survey conducted by the U.S. Census Bureau annually and published on the website (`https://www.census.gov/programs-surveys/acs`). The purpose of ACS is to provide up-to-date estimates that are related to society and economy for a variety of geographies to the U.S. public. The U.S. Census Bureau launched the ACS in 2005. Since then, the public-use ACS estimates are released yearly on the basis of 1-year, 3-year, or 5-year periods.

However, 3-year estimates, which were available for areas of greater than 20,000 population, were terminated in 2013. The 1-year estimates are accessible for areas of at least 65,000 people, while no population restriction is put for the 5-year estimates.

Motivated by an application of the ACS data, Bradley et al. (2015) proposed the Spatio-Temporal Change of Support (STCOS) methodology. This novel methodology was developed based on the fact that one may be interested in getting estimates on spatial and/or temporal domains, which differ from the observed domains. The model results in a mixed effects model, where the coefficients of the random effects are structured to account for the multiple space/time scales. This is an example where the Gaussian mixed effects model is used to analyze the data but there is no completing method in the literature (i.e. $B = 1$). Thus, this application provides a good example of how our methodology can be used to obtain gains in prediction even though $B = 1$. To illustrate our approach, we adopt the STCOS analysis of income data from Raim et al. (2021). This dataset consists of all released 1-year, 3-year, and 5-year period ACS estimates of median household income over various geographies, such as counties and census block groups, within Missouri. The ACS estimates consist of point estimates, margins of errors (MOE), and variance estimates. For this application, we adapt our methodology to the STCOS model applied to ACS median household income data recorded over the 2017 5-year period at the block-group level to predict median household income in four neighborhoods of Boone County, Missouri.

The truncated STCOS as a Bayesian hierarchical model can be written as

$$
\begin{aligned}
&\text{Data Model:} \quad Y_i \mid \mu_i, \boldsymbol{\theta}, \sigma^2 \sim \mathrm{N}\left(\mu_i, \sigma^2\right) I\{\sum_i CPE_i[Y_i, \hat{\mu}_i(\boldsymbol{\theta})] < \kappa\} \\
&\text{Process Models:} \quad \mu_i \mid \boldsymbol{\beta}, \boldsymbol{\eta}, \sigma_\xi^2 \sim \mathrm{N}\left(\boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{\psi}_i'\boldsymbol{\eta}, \sigma_\xi^2\right), \quad \boldsymbol{\eta} \mid \sigma_K^2 \sim \mathrm{N}\left(\mathbf{0}, \sigma_K^2\mathbf{K}\right) \\
&\text{Parameter Model 1:} \quad \boldsymbol{\beta} \mid \sigma_\mu^2 \sim \mathrm{N}\left(\mathbf{0}, \sigma_\mu^2\boldsymbol{I}\right) \\
&\text{Parameter Model 2:} \quad \sigma_\mu^2 \sim \mathrm{IG}\left(a_\mu = 1, b_\mu = 2\right) \\
&\text{Parameter Model 3:} \quad \sigma_K^2 \sim \mathrm{IG}\left(a_K = 1, b_K = 2\right) \\
&\text{Parameter Model 4:} \quad \sigma_\xi^2 \sim \mathrm{IG}\left(a_\xi = 1, b_\xi = 2\right), \qquad (19)
\end{aligned}
$$

where $\boldsymbol{X}_i = \left(\frac{|A_i \bigcap B_1|}{|A_i|}, \ldots, \frac{|A_i \bigcap B_{n_B}|}{|A_i|}\right)'$, $\boldsymbol{\psi}_i = (\psi_1(A_i, \ell_i, t_i), \ldots, \psi_r(A_i, \ell_i, t_i))'$, $B_1, \ldots, B_{n_B}$ are fine-scale grid points over the spatial domain, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_\xi^2, \sigma_K^2)'$, $|A|$ is denoted as the total surface area for areal unit A, A_i is the areal unit associate with the i-th observation, ℓ_i is the period associated the i-th observation, t_i is the time point associate with the i-th observation, $I(\cdot)$ is the indicator function, the matrix $\mathbf{K}$ is a structure covariance matrix based on a random walk (details about this structure can be found in the paper of Raim et al. 2021 for this example with 421 observations and 5 bisquare basis functions) and is multiplied with a free parameter σ_K^2 to fully define the covariance of the random coefficient $\boldsymbol{\eta}$, $a_\mu = a_K = a_\xi = 1$, and $b_\mu = b_K = b_\xi = 2$. We have dropped b in our notion for $\hat{\mu}_i(\boldsymbol{\theta}_b, b)$ because $B = 1$, where B is the number of can-

didate models. Set $\boldsymbol{\psi}_j(A,\ell,t) = \frac{1}{\ell|A|}\sum_{j=t-\ell+1}^{t}\int_A g_j(\boldsymbol{s},j)d\boldsymbol{s}$, where $g_j(\boldsymbol{s},j)$ represents a collection of spatio-temporal bisquare basis functions.

The STCOS model is a highly structured Bayesian mixed effect model for Gaussian data, where the random effect coefficients are different spatio-temporal scales, and the covariates are the percentage of overlapping regions between the data's spatial support and a fine-scale grid. The purpose of this application is to show that our methodology can benefit prediction accuracy for a given Bayesian hierarchical model.

Fig. 5: WAIC from the truncated model with κ set equal to the d-th quantile of CPE from the untruncated model. Smaller value of WAIC suggest better out-of-sample predictive accuracy.

To assess out-of-sample performance, we use the Watanabe–Akaike information criterion (WAIC). In Figure 5, WAIC of the truncated model versus κ set equal to different d-th quantile of CPE from the model in (19) without any truncation. The sequence of d is chosen between 0.1 and 1 by 0.01. WAIC decreases as d increases, then increases as d increases, and is fairly constant between 0.5 to 1. When κ is chosen to be the 0.17 quantile, WAIC reaches the smallest value. Therefore, we set κ to be the 0.17 quantile of CPE from (19) without any truncation for inference. When comparing Table 3 with Table 4, we see that predictions are fairly similar, but the measures of variability, in general, are larger for the untruncated model. Thus, this comparison, along with the WAIC values in Figure 5, suggests that we may be outperforming the "untruncated CPE model."

Table 3: Untruncated model-based estimates of 2017 median household income in four neighbors of Boone County: Central, East, North, and Paris

Region	Posterior Mean	Posterior Standard Deviation
Central	27047.33	1895.125
East	43765.68	2453.249
North	43483.82	2854.626
Paris63Corridor	19563.84	3910.908

6 Discussion

We propose a new approach that improves the prediction of a Bayesian hierarchical model using a criterion for model selection. Our new approach uses covariance penalized error (CPE) (Efron, 2004; Efron and Hastie, 2016; Tibshirani and Rosset, 2018; Holbrook et al., 2020) as a model selection criterion, and selects a subset of parameter space based on the values of CPE. Explicitly, the subset is formed by truncating the joint support of the data and the parameter space to only include small values of CPE. We show in Theorem 2 that our choice of truncation can lead to improvements in squared error. We provide additional motivation for this truncated CPE model by showing that every Bayesian model for normal data can be interpreted as a type of truncated CPE model in Theorem 1. We provide additional developments in terms of the implication of the data generating mechanism.

The simulation study shows that when we truncate half the MCMC replicates, after a burin-in, we consistently obtain smaller squared prediction error than the original Bayesian model over three different signal-to-noise specifications. The results also show that if you truncate too much or too little, we see little to no improvement on the basis of the squared errors. Hence, the selection of κ appears to be an important choice, and in our real data example, we suggest using WAIC. One limitation of our method is that it is not clear how much of a decrease is expected in the prediction error that can be achieved when using our method. But, empirically speaking, there are clear improvements (roughly 23 percents decrease) to the out-of-sample error according to the WAIC in the real data study, where there are noticeable changes to the estimate of the variability of the predictions. The real data study of ACS period estimates demonstrates that prediction accuracy improvements can be achieved when applying our methodology to a single model. Moreover, the added computational effort to implement our method is minimal, since one merely needs to add an accept-reject step to a standard MCMC.

Section 5 demonstrates the wide-applicability of our method, where we are able to improve predictions of a modern multi-scale spatio-temporal model, namely, STCOS from (Bradley et al., 2015). However, our improved predictions come at a cost to computation, which limits the type of models that one might choose to constrain. We require one to compute the BLUP predictor in (11) for every iteration of the Gibbs sampler for the unconstrained STCOS, and samples are then rejected according to our constraint on the CPE. Consequently, the BLUP predictor needs to be computationally feasible for our constrained modeling approach to be computationally feasible, which is not always true for every multi-level spatio-temporal model. However, this

is true for STCOS, where r ($< n$) is chosen to be small, which makes the BLUP is computationally feasible. Thus, there is a natural computational limitation on our method in big data settings, where our constrained model requires the BLUP to be computationally feasible.

Table 4: Truncated model-based estimates of 2017 median household income in four neighbors of Boone County: Central, East, North, and Paris

Region	Posterior Mean	Posterior Standard Deviation
Central	27005.93	1719.283
East	43688.09	2442.975
North	43243.18	2753.046
Paris63Corridor	19686.20	3941.126

The term κ is an unknown parameter, and its specification can lead to either improvements or no changes. There are other approaches to estimate κ. For example, best subset selection is a traditional method that sets $\hat{\kappa}$ equal to the arg-min of an information criterion for parameters that take discrete values (Lee et al., 2018). On the other hand, a natural extension of our method is to place a prior distribution on κ, as our use of the WAIC to estimate κ has unchecked variability not accounted for in the model. The theoretical results in this article may provide some guidance. For example, κ^* in Theorem 1 follows a chi-square distribution, and the original Bayesian model is a re-scaled (to the power r) truncated CPE model with a chi-square prior placed on the upper bound. Thus, priors on κ that imply a stochastic ordering relative to a chi-square distribution is an interesting topic of future research.

A non-Bayesian version of our method can be applied by optimizing a constrained loss function. For example, one can maximize the likelihood subject to the constraint that the CPE is smaller than κ. Similar constrained optimization has been done in the past, however not in the context of out-out-sample criteria (e.g.,support vector regression, Smola and Schölkopf 2004, constrains on the in-sample absolute error). Our truncated CPE model can be generalized beyond normal data and squared error, using the CPE based on the q-class of error measures as described in Efron (1986, 2004). In fact, our Theorem 1 applies to this situation. We provided empirical results in Appendix E that suggest improved predictions using the truncated CPE model for dichotomous data. However, our motivating Theorem 2 does not apply to the non-squared error q-class. In non-Gaussian setting, the CPE requires a bootstrap/plug-in estimate that bias the CPE, and Theorem 2 requires the CPE to be unbiased for the true squared error as is the case for our conditional Guardian setting. We treat this as a topic of future research work.

Acknowledgments

Jonathan R. Bradley's research was partially supported by the U.S. National Science Foundation (NSF) under NSF grant SES-1853099 and the National Institutes of

Table 5: The frequency of an ordering of the sums of squared prediction error (i.e., $\sum_i(\mu_i - \hat{\mu}_i)^2$) over 1000 independent replicates of the vector $\mathbf{y}$. For example, "TCPE <OW <BMA" represents the case where the sums of squared prediction error for our proposed truncated CPE (TCPE) is smaller than that of Occam's Window (OW), which is smaller than that of BMA. The value of κ in the tuncated CPE model was chosen to be the 20-th quantile of the CPE for the BMA model.

	Frequency
TCPE < OW < BMA	414
TCPE < BMA < OW	116
BMA< TCPE < OW	32
BMA< OW < TCPE	51
OW< BMA < TCPE	175
OW< TCPE < BMA	212

Health (NIH) under grant 1R03AG070669-01. Qingying Zong's research is partially supported by NSF grant SES-1853099.

Appendix A A Review of Occam's Window

The intuition on modifying the support of a BMA is similar to that of Occam's Window (Madigan and Raftery, 1994; Onorante and Raftery, 2016). Occam's Window proceeds as a typical forward or backward stepwise algorithm with the added rules that a model is eliminated if its performance is greatly less than the best model among candidate models, and if a model is rejected then so are all of its sub-models. Consequently, at each step of the procedure, added constraints are imposed to narrow the search through competing models. In Table 5, we see the Occam's Window procedure tends to outperform traditional BMA in terms of sums of squared prediction error. This conforms to intuition as Occam's window has a more informative stepwise search through the space of potential covariates. In Section 3, we use this general strategy that Occam's Window uses in each step of a stepwise algorithm (i.e., constraining the space based on criteria), and apply it specifically to the problem of improving the prediction of a given Bayesian hierarchical model.

Appendix B Technical Results

Proof of Thoerem 1

By definition $\pi(\boldsymbol{\mu}, \boldsymbol{\theta}_D, \boldsymbol{\theta}_b, b|\mathbf{y}) = \frac{1}{\pi(\mathbf{y})} f(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\theta}_D)\pi(\boldsymbol{\theta}_D)\pi(\boldsymbol{\mu}|\boldsymbol{\theta}_b, b)\pi(\boldsymbol{\theta}_b|b)\pi(b)h(\mathbf{y}, \boldsymbol{\theta}_b, b)$, and when writing $f(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\theta}_D) = f(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\theta}_D)^r f(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\theta}_D)^{1-r}$, we have $\pi(\boldsymbol{\mu}, \boldsymbol{\theta}_D, \boldsymbol{\theta}_b, b|\mathbf{y}) = \frac{1}{\pi(\mathbf{y})} f(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\theta}_D)^r f(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\theta}_D)^{1-r}\pi(\boldsymbol{\theta}_D)\pi(\boldsymbol{\mu}|\boldsymbol{\theta}_b, b)\pi(\boldsymbol{\theta}_b|b)\pi(b)h(\mathbf{y}, \boldsymbol{\theta}_b, b)$. Introduc-

ing u in a similar manner to Damien et al. (1999) we have,

$$\pi(\boldsymbol{\mu},\boldsymbol{\theta}_D,\boldsymbol{\theta}_b,b|\mathbf{y})$$
$$=\frac{1}{\pi(\mathbf{y})}\int_0^1 f(\mathbf{y}\mid\boldsymbol{\mu},\boldsymbol{\theta}_D)^r\pi(\boldsymbol{\theta}_D)\pi(\boldsymbol{\mu}|\boldsymbol{\theta}_b,b)\pi(\boldsymbol{\theta}_b|b)\pi(b)I\{u<f(\mathbf{y}\mid\boldsymbol{\mu},\boldsymbol{\theta}_D)^{1-r}\}h(\mathbf{y},\boldsymbol{\theta}_b,b)du.$$

Within the expression of the indicator take the log and multiply by -2 to obtain.

$$\pi(\boldsymbol{\mu},\boldsymbol{\theta}_D,\boldsymbol{\theta}_b,b|\mathbf{y})$$
$$=\frac{1}{\pi(\mathbf{y})}\int f(\mathbf{y}\mid\boldsymbol{\mu},\boldsymbol{\theta}_D)^r\pi(\boldsymbol{\theta}_D)\pi(\boldsymbol{\mu}|\boldsymbol{\theta}_b,b)\pi(\boldsymbol{\theta}_b|b)\pi(b)I\{-2(1-r)logf(\mathbf{y}\mid\boldsymbol{\mu},\boldsymbol{\theta}_D)$$
$$<-2log(u)\}h(\mathbf{y},\boldsymbol{\theta}_b,b)du.$$

Then, upon substituting the expression $r=\frac{CPE(\mathbf{y},\hat{\boldsymbol{\mu}})}{2logf(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\theta}_D)}+1$, we obtain the result.

Proof of Thoerem 2

By construction $CPE(\mathbf{y},\hat{\boldsymbol{\mu}}_{tc}(\kappa))<\kappa$, we have $CPE(\mathbf{y},\hat{\boldsymbol{\mu}}_{tc}(\kappa))<E\{\sum_{i=1}^n(\mu_i^{(t)}-\hat{\mu}_{i,m})^2\}+\sum_{i=1}^n\sigma_i^2$ for $\kappa=E\{\sum_{i=1}^n(\mu_i^{(t)}-\hat{\mu}_{i,m})^2\}+\sum_{i=1}^n\sigma_i^2$. By Stein's lemma (Stein, 1981), upon taking the expected value, we obtain the result.

Appendix C Corollary 1

Theorem 2 can be extended from a multivariate-vector to a random process. To do this, we introduce notation that treats Y, μ and ε as processes: $Y(\mathbf{s})=\mu(\mathbf{s})+\varepsilon(\mathbf{s})$, where $Y(\mathbf{s})$ is the observation at location $\mathbf{s}\in D\subset\mathbb{R}^{\mathrm{d}}$, $\varepsilon(\mathbf{s})$ is normally distributed with mean zero, constant variance $\sigma^2>0$, and $\varepsilon(\mathbf{s}_i)$ is independent of $\varepsilon(\mathbf{s}_j)$ for $i\neq j$ and $\mathbf{s}_i,\mathbf{s}_j\in D$. Let $\mathbf{y}=(Y(\mathbf{s}_1),...,Y(\mathbf{s}_n))'$, $\boldsymbol{\mu}=(\mu(\mathbf{s}_1),...,\mu(\mathbf{s}_n))'$, where $\mathbf{s}_1,...,\mathbf{s}_n$ are known locations associated with the observed data. Then, the model in (8) stays the same. For $\mathbf{s}_0\in D$, define the Kriging Predictor (Cressie, 1993) as $\hat{\mu}(\mathbf{s}_0)=\mu(\mathbf{s}_0)+cov\big(\mu(\mathbf{s}_0),\mathbf{y}\big)\boldsymbol{\Sigma}_Y^{-1}\{\mathbf{y}-\boldsymbol{\mu}\}$.

Corollary 1 *Let D be a spatial domain and $f(\cdot):D\to\mathbb{R}$ be an intensity function. Suppose we observe normal data $Y(\boldsymbol{s}_i)$ with true real-valued mean $\mu^{(t)}(\mathbf{s}_i)$ and variance $\sigma^2>0$ for $i=1,\ldots,n$. The notation $\hat{\mu}_{tc}(\mathbf{s})$ represents the posterior median of the Kriging predictor $\hat{\mu}(\mathbf{s})$ and $\hat{\mu}_m(\mathbf{s})$ be a generic real-value predictor of $\mu^{(t)}(\mathbf{s})$. Let $\boldsymbol{s}_1,\ldots,\boldsymbol{s}_n$ be independent and identically distributed according to $f(\boldsymbol{s})$. Then, as $n\to\infty$,*

$$\int_D E\{\mu^{(t)}(\mathbf{s})-\hat{\mu}_{tc}(\mathbf{s},\kappa)\}^2f(\mathbf{s})d\mathbf{s}<\int_D E\{\mu^{(t)}(\mathbf{s})-\hat{\mu}_m(\mathbf{s})\}^2f(\mathbf{s})d\mathbf{s},\qquad(20)$$

where $\kappa=\sum_{i=1}^n E\{\mu^{(t)}(\mathbf{s}_i)-\hat{\mu}_m(\mathbf{s}_i)\}^2+n\sigma^2$. We are assuming that this choice of κ leads to a proper model in (9).

Proof : $\frac{1}{n}\sum_{i=1}^n E\{\mu^{(t)}(\mathbf{s}_i)-\hat{\mu}_{tc}(\mathbf{s}_i,\kappa)\}^2<\frac{1}{n}\sum_{i=1}^n E\{\mu^{(t)}(\mathbf{s}_i)-\hat{\mu}_m(\mathbf{s}_i)\}^2$, which follows from Theorem 2 that. Then apply the law large numbers (Billingsley, 2013) as n approaches infinity to obtain the result.

Appendix D Derivation of full-conditional distributions for Gibbs Sampling

Let $\mathbf{C}(\phi(b),\tau^2)=\tau^2\mathbf{H}(\phi(b))$, and let $\boldsymbol{\mu}=\mathbf{X}\boldsymbol{\beta}+\mathbf{w}$, where $\mathbf{w}|\tau^2,\phi(b)\sim N(\mathbf{0},\mathbf{C}(\phi(b),\tau^2))$. In our Gibbs sampler, we update $\boldsymbol{\beta}$, and $\mathbf{w}$, which implicitly updates $\boldsymbol{\mu}$. We provide the derivations of the full-conditional distributions associated with the model in (18) with a list as follows.

– Full-conditional distribution for $\mathbf{w}$:

$$
\begin{aligned}
&f(\mathbf{w}|\cdot)\\
\propto& exp\left\{-\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{w})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{w})}{2\sigma^2}-\frac{\mathbf{w}'\mathbf{H}(\phi(b))^{-1}\mathbf{w}}{2\tau^2}\right\}I\{CPE[\mathbf{y},\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b,b)]<\kappa\}\\
\propto& exp\left\{\frac{-\mathbf{w}'\mathbf{w}}{2\sigma^2}+\frac{2\mathbf{w}'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2}-\frac{\mathbf{w}'\mathbf{H}(\phi(b))^{-1}\mathbf{w}}{2\tau^2}\right\}I\{CPE[\mathbf{y},\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b,b)]<\kappa\}\\
\propto& exp\left[\frac{-\mathbf{w}'\left\{\frac{1}{\sigma^2}\mathbf{I}_n+\frac{1}{\tau^2}\mathbf{H}(\phi(b))^{-1}\right\}\mathbf{w}}{2}+\frac{2\mathbf{w}'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right]I\{CPE[\mathbf{y},\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b,b)]<\kappa\}\\
\propto& exp\left\{\frac{-\mathbf{w}'\boldsymbol{\Sigma}_w^{-1}\mathbf{w}}{2}+\frac{2\mathbf{w}'\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_w(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right\}I\{CPE[\mathbf{y},\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b,b)]<\kappa\}\\
\propto& exp\left(\frac{-\mathbf{w}'\boldsymbol{\Sigma}_w^{-1}\mathbf{w}}{2}+\frac{2\mathbf{w}'\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\mu}_w}{2}\right)I\{CPE[\mathbf{y},\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b,b)]<\kappa\}\\
\propto& N(\boldsymbol{\mu}_w,\boldsymbol{\Sigma}_w)I\{CPE[\mathbf{y},\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b,b)]<\kappa\},
\end{aligned}
$$

where $\boldsymbol{\mu}_w=\frac{\boldsymbol{\Sigma}_w(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^2}$, $\boldsymbol{\Sigma}_w^{-1}=\frac{1}{\sigma^2}\mathbf{I}_n+\frac{1}{\tau^2}\mathbf{H}(\phi(b))^{-1}$.

– Full-conditional distribution for $\boldsymbol{\beta}$:

$$
\begin{aligned}
f(\boldsymbol{\beta}|\cdot)&\propto exp\left\{\frac{-(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{w})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{w})}{2\sigma^2}-\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{20}\right\}I\{CPE[\mathbf{y},\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b,b)]<\kappa\}\\
&\propto exp\left\{\frac{-\boldsymbol{\beta}'\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta}}{2}+\frac{2\boldsymbol{\beta}'\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\Sigma}_\beta\mathbf{X}'(\mathbf{y}-\mathbf{w})}{2\sigma^2}\right\}I\{CPE[\mathbf{y},\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b,b)]<\kappa\}\\
&\propto N(\boldsymbol{\mu}_\beta,\boldsymbol{\Sigma}_\beta)I\{CPE[\mathbf{y},\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b,b)]<\kappa\},
\end{aligned}
$$

where $\boldsymbol{\mu}_\beta=\frac{\boldsymbol{\Sigma}_\beta\mathbf{X}'(\mathbf{y}-\mathbf{w})}{\sigma^2}$, $\boldsymbol{\Sigma}_\beta^{-1}=\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}+\frac{1}{10}\mathbf{I}_p$.

– Full-conditional distribution for τ^2:

$$f(\tau^2|\cdot) \propto \frac{I\{CPE[\mathbf{y},\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b,b)]<\kappa\}}{|\tau^2\mathbf{H}(\phi(b))|^{1/2}} exp\left[\frac{-\mathbf{w}'\left\{\frac{1}{\tau^2}\mathbf{H}(\phi(b))^{-1}\right\}\mathbf{w}}{2}-\frac{0.01}{\tau^2}\right]\left(\frac{1}{\tau^2}\right)^{1/2}$$
$$\propto exp\left[\frac{-\frac{1}{2}\mathbf{w}'\mathbf{H}(\phi(b))^{-1}\mathbf{w}-0.01}{\tau^2}\right]\left(\frac{1}{\tau^2}\right)^{n/2+2} I\{CPE[\mathbf{y},\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b,b)]<\kappa\}$$
$$\propto IG(0.5n+1, 0.01+0.5\mathbf{w}'\mathbf{H}(\phi(b))^{-1}\mathbf{w}) I\{CPE[\mathbf{y},\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}_b,b)]<\kappa\},$$

where $|\mathbf{H}(\phi(b))|$ is the determinant of $\mathbf{H}(\phi(b))$.

– Full-conditional distribution for b:

$$f(b|\cdot) = \frac{exp\left[\frac{-\mathbf{w}'\mathbf{H}(\phi(b))^{-1}\mathbf{w}}{2\tau^2}\right]|\mathbf{H}(\phi(b))|^{-1/2}}{\sum_{q=1}^{6} exp\left[\frac{-\mathbf{w}'\mathbf{H}(\phi(q))^{-1}\mathbf{w}}{2\tau^2}\right]|\mathbf{H}(\phi(q))|^{-1/2}};\ \ b=1,\ldots,6.$$

Appendix E A Small Preliminary Simulation Study for Bernoulli Observations

We conduct a simulation study for Bernoulli observations and illustrate that the truncated CPE model can be applied to non-Gaussian settings. We generate n independent observations Y_i from Bernoulli($\mu_i \equiv 0.25$) and use the following Bayesian hierarchical model to fit the simulated binary data vector $\mathbf{y} = (Y_1,\ldots,Y_n)'$:

$$\begin{aligned}&\text{Data Model: } \mathbf{y}|\boldsymbol{\mu},a \sim \text{Ber}(\boldsymbol{\mu}) I\{CPE[\mathbf{y},\hat{\boldsymbol{\mu}}(a)]<\kappa\}\\ &\text{Parameter Model: } \boldsymbol{\mu} \sim \text{Beta}(a,1)\\ &\text{Hyper Parameter Model: } a \sim \text{Gamma}(\text{shape}=2,\text{rate}=0.1),\end{aligned} \tag{21}$$

where $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_n)'$; Ber$(\cdot)$, Beta$(\cdot,\cdot)$, and Gamma$(\cdot,\cdot)$ are shorthand for Bernoulli, beta, and gamma distribution, respectively. From Efron (2004), the formula for CPE in this setting is

$$CPE[\mathbf{y},\hat{\boldsymbol{\mu}}] = \sum_{i=1}^{n}\{-2Y_i log(\hat{\mu}_i) - 2(1-Y_i)log(1-\hat{\mu}_i) + 2cov\{log[\hat{\mu}_i/(1-\hat{\mu}_i)], Y_i\}.$$

We let $\hat{\mu}_i(a) = \frac{a+\sum_{i=1}^{n} Y_i}{a+1+n}$, which is the posterior predictive mean for a given value of a.

We consider three settings, where κ are set to be d-th Quantile (d = 0.1, 0.5, and 0.9) of the set $\{CPE[\mathbf{y},\hat{\boldsymbol{\mu}}(a^{[1]})],\ldots,\hat{\boldsymbol{\mu}}(a^{[2,000]})]\}$. The i-th Markov Chain Monte Carlo (MCMC) replicate for a is $a^{[i]}$ and $CPE[\mathbf{y},\hat{\boldsymbol{\mu}}(a)]$ is the CPE calculated from the untruncated model. We apply our model and the untruncated model by using a Gibbs sampler with length of 2,000 and burin-in of 1,000. The true prediction error in this setting is,

$$\sum_{i=1}^{n}\Big[-2log(1-\hat{\mu}_i) - 2\mu_i log(\frac{\hat{\mu}_i}{1-\hat{\mu}_i})\Big],$$

which can be estimated by $CPE[\mathbf{y}, \hat{\boldsymbol{\mu}}]$ as described by Efron (2004).

To analyze the effect of Quantile d, we simulate 100 independent replicates of the data vector $\mathbf{y}$. We plot

$$\sum_{i=1}^{n}[-2log(1-\hat{\mu}_{i,tc}) - 2\mu_i log(\frac{\hat{\mu}_{i,tc}}{1-\hat{\mu}_{i,tc}})] - \sum_{i=1}^{n}[-2log(1-\hat{\mu}_i) - 2\mu_i log(\frac{\hat{\mu}_i}{1-\hat{\mu}_i})].$$

Here negative values suggest that the truncated CPE model has better predictive performance than the untruncated model. In Figure 6, as Quantile d decreases, the boxplot for the variable BerResponse shows more variability but is more centered negative. In this simulation, we show that when we use binomial deviance as the error measure for Bernoulli data and select κ appropriately, our method can improve the prediction in a simple setting.

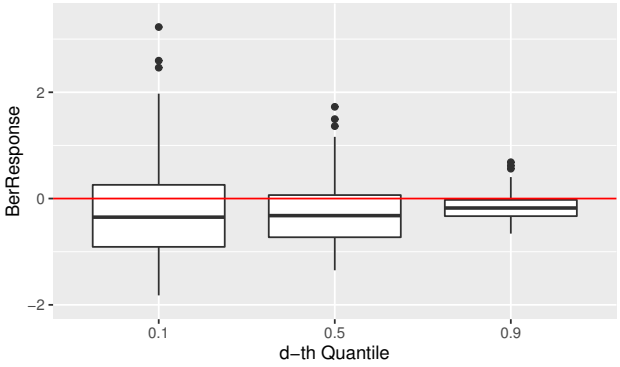

Fig. 6: The boxplot of BerResponse for d-th Quantile. The horizontal red solid line stands for BerResponse equal to zero. BerResponse that are negative indicates the truncated model outperforms the untruncated model in terms of binomial deviance.

References

Acquah, H. D.-G. (2010). "Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship." *Journal of Development and Agricultural Economics*, 2, 1, 001–006.

Akaike, H. (1973). "Maximum likelihood identification of Gaussian autoregressive moving average models." *Biometrika*, 60, 2, 255–265.

Barbieri, M. M. and Berger, J. O. (2004). "Optimal predictive model selection." *The Annals of Statistics*, 32, 3, 870–897.

Benjamini, Y. (2010). "Discovering the false discovery rate." *Journal of the Royal Statistical Society: series B (statistical methodology)*, 72, 4, 405–416.

Benjamini, Y., Heller, R., and Yekutieli, D. (2009). "Selective inference in complex research." *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367, 1906, 4255–4271.

Billingsley, P. (2013). *Convergence of probability measures*. John Wiley & Sons.

Box, G. E. (1980). "Sampling and Bayes' inference in scientific modelling and robustness." *Journal of the Royal Statistical Society: Series A (General)*, 143, 4, 383–404.

Bradley, J. R., Wikle, C. K., and Holan, S. H. (2015). "Spatio-temporal change of support with application to American Community Survey multi-year period estimates." *Stat*, 4, 1, 255–270.

Bradley, J. R. and Zong, Q. (2021). "Empirical Bayesian analysis through the lens of a particular class of constrained Bayesian hierarchical models." *Stat*, 10, 1, e403.

Burnham, K. P. and Anderson, D. R. (2003). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). "Handling sparsity via the horseshoe." In *Artificial Intelligence and Statistics*, 73–80.

Catelan, D., Lagazio, C., and Biggeri, A. (2010). "A hierarchical Bayesian approach to multiple testing in disease mapping." *Biometrical Journal*, 52, 6, 784–797.

Chen, C.-S. and Huang, H.-C. (2012). "Geostatistical model averaging based on conditional information criteria." *Environmental and ecological statistics*, 19, 1, 23–35.

Cressie, N. (1993). "Spatial statistics." *New York*.

Damien, P., Wakefield, J., and Walker, S. (1999). "Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61, 2, 331–344.

Efron, B. (1983). "Estimating the error rate of a prediction rule: improvement on cross-validation." *Journal of the American statistical association*, 78, 382, 316–331.

— (1986). "How biased is the apparent error rate of a prediction rule?" *Journal of the American statistical Association*, 81, 394, 461–470.

— (2004). "The estimation of prediction error: covariance penalties and cross-validation." *Journal of the American Statistical Association*, 99, 467, 619–632.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference, Student Edition: Algorithms, Evidence, and Data Science*. Institute of Mathematical Statistics Monographs. Cambridge University Press.

Farcomeni, A. (2010). "Bayesian constrained variable selection." *Statistica Sinica*, 1043–1062.

Gelfand, A. E. and Schliep, E. M. (2016). "Spatial statistics and Gaussian processes: A beautiful marriage." *Spatial Statistics*, 18, 86–104.

Hodges, J. S. (2013). *Richly parameterized linear models: additive, time series, and spatial models using random effects*. CRC Press.

Hodges, J. S. and Sargent, D. J. (2001). "Counting degrees of freedom in hierarchical and other richly-parameterised models." *Biometrika*, 88, 2, 367–379.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). "Bayesian model averaging: a tutorial." *Statistical science*, 382–401.

Holbrook, A., Lumley, T., and Gillen, D. (2020). "Estimating prediction error for complex samples." *Canadian Journal of Statistics*, 48, 2, 204–221.

Huang, H.-C. and Chen, C.-S. (2007). "Optimal geostatistical model selection." *Journal of the American Statistical Association*, 102, 479, 1009–1024.

Ishwaran, H. and Rao, J. S. (2005). "Spike and slab variable selection: frequentist and Bayesian strategies." *Annals of statistics*, 33, 2, 730–773.

Lee, Y., Nelder, J. A., and Pawitan, Y. (2018). *Generalized linear models with random effects: unified analysis via H-likelihood*, vol. 153. CRC Press.

Little, R. (2011). "Calibrated Bayes, for statistics in general, and missing data in particular." *Statistical Science*, 26, 2, 162–174.

Little, R. J. (2006). "Calibrated Bayes: a Bayes/frequentist roadmap." *The American Statistician*, 60, 3, 213–223.

— (2012). "Calibrated Bayes, an alternative inferential paradigm for official statistics." *Journal of official statistics*, 28, 3, 309.

Madigan, D. and Raftery, A. E. (1994). "Model selection and accounting for model uncertainty in graphical models using Occam's window." *Journal of the American Statistical Association*, 89, 428, 1535–1546.

Mallows, C. L. (1973). "Some comments on C p." *Technometrics*, 15, 4, 661–675.

Maraun, D. and Widmann, M. (2018). *Statistical downscaling and bias correction for climate research*. Cambridge University Press.

Onorante, L. and Raftery, A. E. (2016). "Dynamic model averaging in large model spaces using dynamic Occam's window." *European Economic Review*, 81, 2–14.

Panigrahi, S., Taylor, J., and Weinstein, A. (2016). "Integrative methods for post-selection inference under convex constraints." *arXiv preprint arXiv:1605.08824*.

Raim, A. M., Holan, S. H., Bradley, J. R., and Wikle, C. K. (2021). "Spatio-temporal change of support modeling with R." *Computational Statistics*, 36, 1, 749–780.

Rao, R. and Wu, Y. (1989). "A strongly consistent procedure for model selection in a regression problem." *Biometrika*, 76, 2, 369–374.

Ravishanker, N. and Dey, D. K. (2020). *A first course in linear model theory*. CRC Press.

Rubin, D. B. (1984). "Bayesianly justifiable and relevant frequency calculations for the applied statistician." *The Annals of Statistics*, 1151–1172.

Smola, A. J. and Schölkopf, B. (2004). "A tutorial on support vector regression." *Statistics and computing*, 14, 3, 199–222.

Stein, C. M. (1981). "Estimation of the mean of a multivariate normal distribution." *The annals of Statistics*, 1135–1151.

Tibshirani, R. J. and Rosset, S. (2018). "Excess optimism: How biased is the apparent error of an estimator tuned by SURE?" *Journal of the American Statistical Association*.

Torkashvand, E., Jozani, M. J., and Torabi, M. (2016). "Constrained Bayes estimation in small area models with functional measurement error." *Test*, 25, 4, 710–730.

Vaida, F. and Blanchard, S. (2005). "Conditional Akaike information for mixed-effects models." *Biometrika*, 92, 2, 351–370.

Wasserman, L. (2000). "Bayesian model selection and model averaging." *Journal of mathematical psychology*, 44, 1, 92–107.

Ye, J. (1998). "On measuring and correcting the effects of data mining and model selection." *Journal of the American Statistical Association*, 93, 441, 120–131.

Yekutieli, D. (2012). "Adjusted Bayesian inference for selected parameters." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74, 3, 515–541.

Zhao, Z. and Hwang, J. T. G. (2012). "Empirical Bayes false coverage rate controlling confidence intervals." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74, 5, 871–891.