Exact Bayesian Geostatistics Using Predictive Stacking

BY LU ZHANG

Division of Biostatistics, Department of Population and Public Health Sciences, University of Southern California, USA

lzhang63@usc.edu

WENPIN TANG

Department of Industrial Engineering and Operations Research, Columbia University, USA

wt2319@columbia.edu

AND SUDIPTO BANERJEE

Department of Biostatistics, University of California Los Angeles, Los Angeles, USA

sudipto@ucla.edu

SUMMARY

We develop Bayesian predictive stacking for geostatistical models. Our approach builds an augmented Bayesian linear regression framework that subsumes the realisations of the spatial random field and delivers exact analytically tractable posterior inference conditional upon certain spatial process parameters. We subsequently combine such inference by stacking these individual models across the range of values of the hyper-parameters. We devise stacking of means and posterior densities in a manner that is computationally efficient without the need of iterative algorithms such as Markov chain Monte Carlo (MCMC) and can exploit the benefits of parallel computations. We offer novel theoretical insights into the resulting inference within an infill asymptotic paradigm and through empirical results showing that stacked inference is comparable to full sampling-based Bayesian inference at a significantly lower computational cost.

.

Some key words: Bayesian inference; conjugate spatial models; Gaussian processes; Geostatistics; stacking.

1. INTRODUCTION

Geostatistics (Cressie, 1993; Chilés & Delfiner, 1999; Zimmerman & Stein, 2010; Banerjee, 2019) refers to the study of a spatially distributed variable of interest, which in theory is defined at every point over a bounded study region of interest. Customary geostatistical modelling proceeds from a latent stochastic process over space that specifies the probability law for the point-referenced measurements on the variable as a partial realisation of the process over a finite set of locations. Inference is sought for the underlying spatial process posited to be generating the data, which is subsequently used for spatial predictions over the domain ("kriging" Stein, 1999) to better understand the scientific phenomenon under study. The spatial process, is often assumed to be stationary and empirically estimated from measurements at sampled locations using the "variogram" and characterised by parameters representing the sill, the nugget, the range and, possibly, the smoothness of the process. We collectively refer to these as process parameters.

Formal likelihood-based inference for this process is, however, thwarted by the absence of classical consistent estimators of the process parameters in a customarily preferred infill asymptotic paradigm (see, e.g., Stein, 1999; Zhang, 2004; Zhang & Zimmerman, 2005; Kaufman & Shaby, 2013; Tang et al., 2021). Bayesian inference for geostatistical processes (Handcock & Stein, 1993; Berger et al., 2001; Banerjee et al., 2014; Li et al., 2023), while not relying upon asymptotic inference, is also not entirely straightforward. Specifically, irrespective of how many spatial locations yield measurements, the likelihood does not mitigate the effect of the prior distributions on the inference. This is undesirable since proper prior elicitation for spatial process parameters is challenging. Objective priors for spatial process models have also been pursued, but interpreting such information in practice and their implications in scientific contexts are not uncontroversial. The related question of how effectively (or poorly) the realised data can identify these process parameters (in an exact sense from finite samples) has also been the subject of commentary (see, e.g., Hodges, 2013; Bose et al., 2018; De Oliveira & Han, 2022).

The aforementioned, rather substantial, literature has focused extensively on inference for geostatistical parameters and their sensitivity to modelling assumptions from diverse perspectives. It is, therefore, not unreasonable to pursue methods that will yield robust inference for the spatial process and for spatial predictions of the outcome at arbitrary points ("kriging") while circumventing inference on the weakly identified parameters. Instead of seeking families of prior distributions for such parameters, recent efforts at computationally efficient algorithms for geostatistical models have proposed multi-fold cross-validation methods (Finley et al., 2019) to fix the values of weakly identified parameters. However, the metrics for ascertaining optimal values of such parameters are somewhat arbitrary and may not offer robust inference. Instead, our current contribution develops Bayesian predictive stacking of geostatistical models.

Stacking is a model averaging procedure for generating predictions (Wolpert, 1992; Breiman, 1996a; Clyde & Iversen, 2013). Stacking methods and algorithms in diverse data analytic applications are rapidly evolving and a comprehensive review is beyond the scope of this manuscript. Significant developments of stacking methodology in Bayesian analysis have been achieved in recent years (Le & Clarke, 2017; Yao et al., 2018, 2020, 2021), but, to the best of our knowledge, developments in the context of spatial data analysis are lacking. Stacking can be regarded as an alternative to Bayesian model averaging (Madigan et al., 1996; Hoeting et al., 1999). Assume that there are G candidate models $M = \{M_1, \ldots, M_G\}$. Following Bernardo & Smith (1994), Bayesian model comparison customarily considers three settings: (i) M-closed where a true data generating model exists and is included in M; (ii) M-complete where a true model exists but is not included in M; and (iii) M-open where we do not assume the existence of a true data generating model. While Bayesian model averaging is often the preferred solution in the first setting, predictive stacking has advantages in the M-complete and M-open settings. Given the complex nature of spatial dependence, the M-closed assumption is rarely tenable in practical geostatistics and we prefer stacking to model averaging.

Much of the theoretical properties of conventional stacking rely upon properties of exchangeable models that are not enjoyed by geostatistical models. Inferential behaviour of posterior distributions is, therefore, studied within an infill asymptotic paradigm. We absorb spatial process realisations into a convenient augmented Bayesian linear regression framework (Section 2). We offer some novel theoretical insights into the consistency of the posterior and posterior predictive distributions. Each model $M_g \in M$ is indexed according to fixed values of certain spatial covariance parameters so that the exact posterior distribution is analytically tractable. Section 3 develops geostatistical predictive stacking by combining the exact inference across the G models in M. We develop (i) stacking of means, which combines posterior predictive means, $\mathrm{E}_g(y(s_0) \mid y_{obs}, M_g)$, from each of the G models; and (ii) stacking of posterior predictive den-

sities (Yao et al., 2018), which combines posterior predictive densities $p(\tilde{y}(s_0) \mid y_{obs}, M_g)$ for $g = 1, 2, \ldots, G$, where $\tilde{y}(s_0)$ is the random variable denoting model predictions of the outcome at an arbitrary point s_0 and y_{obs} denotes the observed data on the outcome. We obtain exact Bayesian inference by exploiting exact distribution theory and avoiding iterative algorithms such as Markov chain Monte Carlo (MCMC). These methods are evaluated theoretically as well as empirically through simulation experiments (Section 4) demonstrating that stacked inference is comparable to full Bayesian inference using MCMC at significantly less computational expense. An illustrative data analysis is presented in Section 5 to further corroborate results seen in the simulations. We conclude with some discussions and pointers to future work in Section 6.

2. Bayesian hierarchical spatial process models

2.1. Conjugate Bayesian spatial model

We explore a regression model for a spatially indexed outcome $y(s)$ at a location s in a bounded region $D \subset \mathbb{R}^d$,

$$y(s) = x(s)^\top \beta + z(s) + \varepsilon(s), \tag{2.1}$$

where $x(s)$ is a $p \times 1$ vector of spatially referenced predictors, β is a $p \times 1$ vector of slopes measuring the trend, $z(s) \sim \text{GP}(0, \sigma^2 R(\cdot\,;\phi))$ is a zero-centred spatial Gaussian process on $\mathbb{R}^d$ with spatial correlation function $R(\cdot\,;\cdot)$ depending on spatial range parameter ϕ, and σ^2 is a scale (spatial variance) parameter. The white noise process $\varepsilon(s) \sim N(0, \tau^2)$ with variance τ^2 captures measurement error.

Let $\mathcal{S} = \{s_1, \ldots, s_n\}$ be a set of n spatial locations, each $s_i \in D$, yielding measurements $y = (y(s_1), \ldots, y(s_n))^\top$ with known values of predictors at these locations collected in the $n \times p$ matrix $X = (x(s_1), \ldots, x(s_n))^\top$. We also define the finite-dimensional realisation of the spatial process as $z = (z(s_1), \ldots, z(s_n))^\top$, and let $R(\phi) = (R(s_i, s_j))_{1 \le i,j \le n}$ be the $n \times n$ spatial correlation matrix constructed from the correlation function. A customary Bayesian hierarchical model is constructed as

$$\begin{aligned} y \mid z, \beta, \tau^2 &\sim N(X\beta + z, \tau^2 I_n); \quad z \mid \sigma^2, \phi \sim N(0, \sigma^2 R(\phi)) \\ \beta \mid \sigma^2 &\sim N(\mu_\beta, \sigma^2 V_\beta); \quad \sigma^2 \sim \text{IG}(a_\sigma, b_\sigma); \end{aligned} \tag{2.2}$$

where we fix the spatial correlation parameters ϕ and the noise-to-spatial variance ratio $\alpha^2 := \frac{\tau^2}{\sigma^2}$, and μ_β, V_β, a_σ, and b_σ are fixed hyper-parameters specifying the prior distributions for β and σ^2. This ensures closed-form conjugate marginal posterior and posterior predictive distributions (Finley et al., 2019; Banerjee, 2020). In order to harness familiar results from conjugate Bayesian linear regression, we cast the spatial model in (2.2) into an augmented linear system:

$$\underbrace{\begin{pmatrix} y \\ 0 \\ 0 \end{pmatrix}}_{y} = \underbrace{\begin{pmatrix} X & I_n \\ I_p & 0 \\ 0 & I_n \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \beta \\ z \end{pmatrix}}_{\theta} + \underbrace{\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}}_{\omega}; \tag{2.3}$$

where $\omega \sim N(0, \sigma^2 V_y)$, $V = \begin{pmatrix} \alpha^2 I_n & 0 & 0 \\ 0 & V & 0 \\ 0 & 0 & R(\phi) \end{pmatrix}$ and $p(\beta, \sigma^2) \propto \text{IG}(\sigma^2 \mid a_\sigma, b_\sigma)$.

LEMMA 1. *The posterior distribution of* $(\ ;\ ^2)$ *from (2.2) is*

$$p(\ ;\ ^2 j y) = \underbrace{IG(\ ^2 j a\ ;b\ )}_{p(\ ^2 j y)} \quad \underbrace{N(\ j \hat{\ };\ ^2 M\ )}_{p(\ j\ ^2;y)}; \tag{2.4}$$

where $a\ = a\ + n=2$, $b\ = b\ + \frac{1}{2}(y\ \ X\ \hat{\ })^\top V\ ^{1}(y\ \ X\ \hat{\ })$, $M\ ^{1} = X^\top V\ ^{1}X$ *and* $\hat{\ } = M\ X^\top V\ ^{1}y$. *The posterior distribution* $p(\ j y)$ *is a multivariate Student's t with degrees of freedom* $2a\ $, *location* $\hat{\ }$ *and scale matrix* $(b\ =a\ )M\ $.

Proof. The proof follows from a straightforward adaptation of familiar results from the Normal-Gamma family of distributions (see (Murphy, 2015) and Section 3.1 in (Banerjee, 2020)) to the setting in (2.3).

Furthermore, let $\tilde{\ } = \{s_1;:::;s_m\}$ be a set of m unknown points in D, $\tilde{z}$ and $\tilde{y}$ be the $m\ \ 1$ vectors with elements $z(s_i)$ and $y(s_i)$ for $i = 1;2;:::;m$. Let $\tilde{X} = (x(s_1);:::;x(s_m))^\top$ be the $m\ \ p$ matrix that carries the values of predictors at $\tilde{\ }$ and let $J\ (\tilde{\ }) = (R\ (s;s^{'}))_{fs2\ ;s^{'}2\tilde{\ }g}$ Then, spatial predictive inference follows from the posterior distribution

$$p(\tilde{z};\tilde{y} j y) = \int p(\tilde{y} j \tilde{z};\ ;\ ^2)p(\tilde{z} j z;\ ^2)p(\ ;\ ^2 j y)\, d\ \ d\ ^2; \tag{2.5}$$

which is again a multivariate t distribution with degrees of freedom $2a\ $, location $\tilde{\ }$ and scale matrix $(b\ =a\ )\tilde{M}$ where

$$\tilde{\ } = W\hat{\ };\ \tilde{M} = W M\ W^\top + M_2\ ;\ M_1 = I_m \quad J^\top(\tilde{\ })R\ ^{1}(\ )J\ (\tilde{\ });$$
$$W = \begin{bmatrix} 0 & J^\top(\tilde{\ })R\ ^{1}(\ ) \\ \tilde{X} & J^\top(\tilde{\ })R\ ^{1}(\ ) \end{bmatrix};\ M_2\ ^{1} = \begin{bmatrix} \frac{1}{2}I_m + M_1\ ^{1} & \frac{1}{2}I_m \\ \frac{1}{2}I_m & \frac{1}{2}I_m \end{bmatrix}:$$

Specifically, the predictive distributions $p(z(s_0) j y)$ and $p(y(s_0) j y)$ are also available in analytic form as non-central t distributions for any single point $s_0 \in D$. Bayesian inference can proceed from exact posterior samples obtained from (2.4) as follows. We first draw values of $^2 \sim IG(a\ ;b\ )$ followed by a single draw of $\sim N(\hat{\ };\ ^2 M\ )$ for each drawn value of ^{2}. This yields samples $\{\ ;\ ^2\}$ from (2.4). Predictive inference for the latent process $z(s_0)$ and the outcome $y(s_0)$ is obtained by sampling from (2.5) by drawing a value of $\tilde{z} \sim N(\ _z(\ );\ ^2 V_{zj})$ with $\ _z(\ ) := J^\top(\tilde{\ })R\ (\ )\ ^{1}z$ and $V_{zj} := R\ (\tilde{\ }) \quad J^\top(\tilde{\ })R\ (\ )\ ^{1}J\ (\tilde{\ })$ for each value of $\{\ ;\ ^2\}$ drawn above (see Section 3.4 in Banerjee (2020)), then drawing a value of $\tilde{y} \sim N(\tilde{X}\ + \tilde{z};\ ^2\ ^2 I_m)$ for each drawn value of (extracted from), ^{2} and $\tilde{z}$.

This tractability is only possible if the range decay and the noise-to-spatial variance ratio ^{2} are fixed. While the data can inform about these parameters, they are inconsistently estimable (Zhang, 2004) often resulting in poorer convergence. Alternate approaches using K-fold cross-validation have been explored with limited success (Finley et al., 2019). Therefore, our approach will conduct the exact inference using (2.4) and (2.5) and stack the inference over the different fixed values of $\{\ ;\ ^2\}$. Next, we investigate the concentration of the posteriors and the effect of incorrectly specified hyperparameters on the posterior inference.

2.2. Posterior inference for spatial process models

Let L be the Cholesky decomposition of V such that $V = L L^\top$, and L a non-singular square matrix such that $R(\ )^{-1} = L^\top L$. The linear model (2.3) can be rewritten as

$$\underbrace{\begin{pmatrix} y \\ L^{-1}y \end{pmatrix}}_{y_y} = \underbrace{\begin{pmatrix} X & I_n \\ L^{-1}X & 0 \end{pmatrix}}_{X_y} \underbrace{\begin{pmatrix} \beta \\ z \end{pmatrix}}_{\gamma} + \begin{pmatrix} \epsilon_{y;1} \\ \epsilon_{y;2} \end{pmatrix}; \quad \epsilon_{y;3} \tag{2.6}$$

where $\epsilon_y \sim N(0, \sigma^2 I_{2n+p})$. To explore posterior concentrations within an in-fill paradigm, where we assume a fixed size of our spatial domain D and an increasing number of spatial locations inside D, we will use the concept of equivalence of probability measures.

Assumption 1 *(Equivalence).* Let P_0 be the probability distribution of the process $y(s)$ defined by the model (2.1) with true parameter values $\{\beta_0, \sigma_0^2, \tau_0^2, \phi_0\}$. For each $\epsilon > 0$, there is $\delta^2 > 0$ such that the probability distribution of the process $y(s)$ defined by the model (2.1) with parameter values $\{\beta_0, \delta^2, \tau_0^2, \phi_0\}$ is equivalent to P_0.

Note that the above Assumption holds when the latent process $z(s)$ follows a Matérn model in dimension $d \in \{1, 2, 3\}$, which we provide more details in Section 2.3. In this case, the probability distribution of the process $y(s)$ defined by the model (2.1) with parameters $\{\beta_0, \frac{\sigma_0^2}{\tau_0^2}, \tau_0^2, \phi_0\}$, with ν a smoothness parameter of the Matérn model, is equivalent to P_0 (see e.g. Tang et al. (2021, Section 2.1)). The following theorem explores the posterior (in)consistency of parameter inference.

THEOREM 1 (POSTERIOR INFERENCE (IN)CONSISTENCY). *Let* P_0 *be the probability measure of the model* (2.1) *with parameter values* $\{\beta_0, \sigma_0^2, \tau_0^2, \phi_0\}$*, and let Assumption 1 hold. Let* $H = X_y(X_y^\top X_y)^{-1}X_y^\top$ *be the* $(2n+p) \times (2n+p)$ *orthogonal projector onto the column space of* X_y *and let* $H = \begin{pmatrix} H_{11} & H_{12} \\ H_{12}^\top & H_{22} \end{pmatrix}$ *be a* 2×2 *partition of* H *so that* H_{22} *is the lower right* $n \times n$ *block formed by rows and columns indexed from* $n+p+1$ *to* $2n+p$*. Assume that* $\mathrm{Tr}(H_{22})/n \to \infty$ *as* $n \to \infty$*. Then under* P_0*,*

$$\lim_{n \to \infty} p(\sigma^2 \mid y(\ _n)) = \delta_{\sigma^2}; \tag{2.7}$$

where $\ _n = \{s_1, s_2, \ldots, s_n\}$ *is a sequence of sets of* n *spatial locations in the fixed domain* D*,* $y(\ _n) = (y(s_1), y(s_2), \ldots, y(s_n))^\top$*,* $\sigma^2 := \frac{\sigma_0^2}{\tau_0^2} + \delta^2(1-\nu)$ *and* δ_{σ^2} *denotes the Dirac measure at* σ^2*.*

Proof. See Section A in the Appendix.

Theorem 1 suggests that the posterior distribution of the scale parameter σ^2 in the spatial process does not necessarily concentrate to the true generating value. We also provide conditions for the posterior consistency of ν in the following corollary.

COROLLARY 1. *Let* P_0 *be the probability distribution of the model* (2.1) *with parameter values* $\{\beta_0, \sigma_0^2, \tau_0^2, \phi_0\}$ *under Assumption 1. Define* $U = (X_y^\top X_y)^{-1} = \begin{pmatrix} U_{11} & U_{12} \\ U_{12}^\top & U_{22} \end{pmatrix}$*, where* U_{11} *is* $p \times p$*,* $B = U \begin{pmatrix} X^\top X & X^\top \\ X & I_n \end{pmatrix} U = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$*, where* B_{11} *is* $p \times p$*,* $C = U_{11}V^{-1}U_{11}$*, and* $D = U_{12}R(\ _n)^{-1}U_{12}^\top$ *and assume the following additional conditions:*

$$\mathrm{Tr}(U_{11}),\ \mathrm{Tr}(B_{11}),\ \mathrm{Tr}(C),\ \mathrm{Tr}(D) \to 0 \quad \text{as } n \to \infty: \tag{2.8}$$

Then, $\lim_{n \to \infty} p(\quad | y(\quad_n)) = \quad_0$ *under* P_0*, where* $\quad_n$ *and* $y(\quad_n)$ *are defined as in Theorem 1.*

Proof. See Section B in the Appendix.

Figure 1 summarises some numerical experiments to empirically explore $\mathrm{Tr}(H_{22})/n \to$ and (2.8) for some typical examples. The study domain D is $[0, 1]^2$, and locations in are chosen uniformly on D. We generate data using the Matérn covariogram for $z(s)$ (see (2.12) in Section 2.3). We consider two types of predictors $x(s)$. For the first type, titled "with intercept", $x(s)$ consists of an constant 1 for intercept and a predictor generated by a standard normal. For the second type, labelled "without intercept", $x(s)$ is composed of two predictors sampled from a standard normal. We explore the trends of the target quantities with different hyper-parameter values in the covariogram of $z(s)$ and different types of $x(s)$ as sample size increases. Figure 1 reveals clearly that $\mathrm{Tr}(H_{22})/n$ increases as sample size increases for all examples. Since $\mathrm{Tr}(H_{22})/n$ is bounded above by 1, the condition $\mathrm{Tr}(H_{22})/n \to$ for some constant is likely to hold for all examples. On the other hand, condition (2.8) is observed to be more sensitive to the choice of $x(s)$. It does not seem to hold for the cases when contains an intercept. Examples on the one-dimensional $[0, 1]$ exhibit similar behaviour. Moreover, we prove in Theorem 3 that in the special case where there is no trend term $x(s)^\top$, and the latent process $z(s)$ follows a Matérn model in dimension $d \in \{1, 2, 3\}$, the posterior distribution of ^{2} converges to the degenerate distribution at $^2 = ^2$. Hence, in general, this result will be inconsistent unless the noise-to-spatial variance ratio ^{2} is fixed to be $^2_0 = ^2_0 / ^2_0$.

Next, we consider Bayesian posterior predictive inference at a new location $s_0 \in D$. Let $Z_n(s_0)$ be a random variable distributed as $p(z(s_0) | y(\quad_n))$ and $Y_n(s_0)$ be distributed as $p(y(s_0) | y(\quad_n))$. We study the prediction error $E_0(Z_n(s_0) \quad z(s_0))^2$ for the latent process, and $E_0(Y_n(s_0) \quad y(s_0))^2$ for the response variable. Let X_y and y_y be as in (2.6) and U is as in Corollary 1. The following theorem quantifies these posterior prediction errors.

THEOREM 2 (POSTERIOR PREDICTIVE CONSISTENCY). *Consider* $s_0 \in D$*. For any given* > 0*, denote* $\mathrm{Cov}(z, z(s_0) | \ ^2)$ *and* $R(\quad_n)$ *by* $^2 J_{,n}$ *and* $R_{,n}$*, respectively. Let* $Z_n(s_0)$ *have the density* $p(z(s_0) | y(\quad_n))$*, and* $Y_n(s_0)$ *have the density* $p(y(s_0) | y(\quad_n))$*, and let*

$$F_n = J^\top_{,n} R^{-1}_{,n} (U_n X_y^\top y_y)_{[p+1:p+n]} \quad \text{and}$$
$$G_n = 1 \quad J^\top_{,n} R^{-1}_{,n} (R_{,n} + U_{n[p+1:p+n; p+1:p+n]}) R^{-1}_{,n} J_{,n};$$

where we use the suffix n *to highlight the dependency of* $\quad_n$*, and we remove the suffix* n *of* X_y *and* y_y *when there is no confusion. Under the assumptions in Theorem 1, we have*

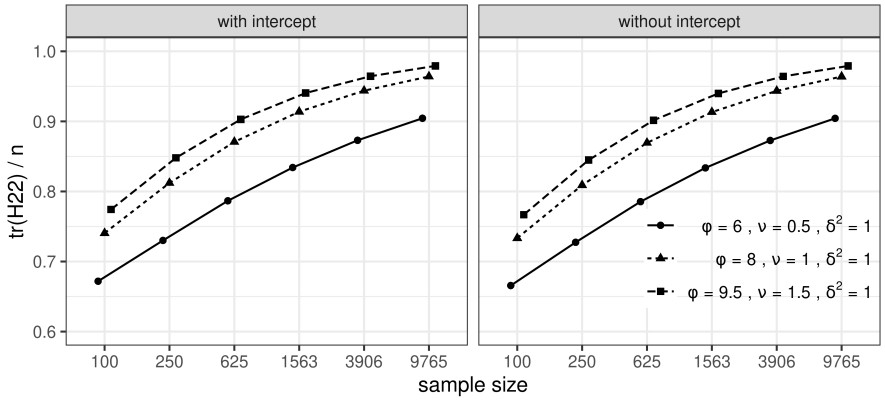
$$E_0(Z_n(s_0) \quad z(s_0))^2 = E_{1,n} + E_{2,n} + o(1); \tag{2.9}$$

where E_0 *on the left side supports* $z(s_0)$*,* $y(\quad_n)$ *and the external randomisation of* $p(z(s_0) | y(\quad_n))$*, and* $E_{1,n} = E_0(z(s_0) \quad F_n)^2$ *and* $E_{2,n} = \ ^2 G_n$*. Moreover, if* $E_{1,n}, E_{2,n} \to 0$ *as* $n \to \infty$*, then posterior inference for the latent process is consistent in the sense that* $E_0(Z_n(s_0) \quad z(s_0))^2 \to 0$ *as* $n \to \infty$*, while posterior predictive inference for* $y(s_0)$ *satisfies*

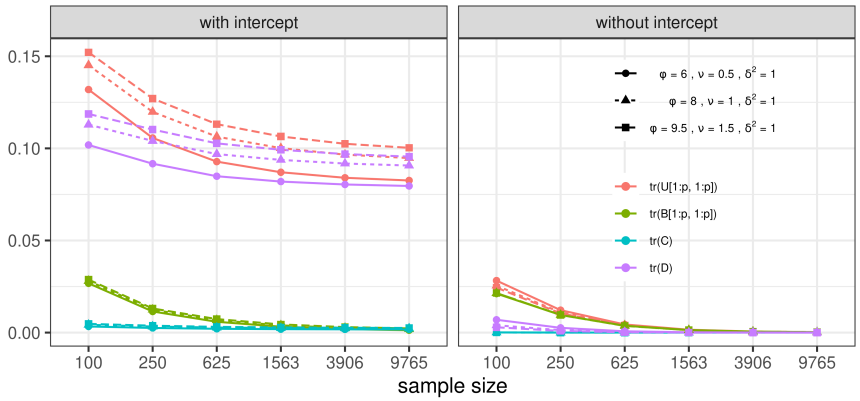
$$E_0(Y_n(s_0) \quad y(s_0))^2 \to \ ^2_0 + \ ^2 \ ^2 \quad \text{as } n \to \infty: \tag{2.10}$$

Proof. See Section C in the Appendix.

In the decomposition (2.9), the term $E_{1,n}$ arises in the deviation from the posterior mean, while the term $E_{2,n}$ is from the posterior uncertainty. The conditions $E_{1,n}, E_{2,n} \to 0$ is analytically intractable in the general case. If, however, is we detrend the outcome so that $x(s)^\top = 0$, then $E_{1,n}$ and $E_{2,n}$ are simplified to inherit a rich structure under the Matérn covariance model for the latent process $z(s)$. In this case, we provide evidence to support the posterior predictive consistency for the latent process in the next subsection.



(a)



(b)

Fig. 1: The top row (a) provides plots of $\mathrm{Tr}(H_{22})/n$ when $x(s)$ consists of an intercept only (left), and with additional covariates (right) for different parameter values. The bottom row (b) plots the quantities $\mathrm{Tr}(U_{11})$, $\mathrm{Tr}(B_{11})$, $\mathrm{Tr}(C)$, and $\mathrm{Tr}(D)$ with an intercept only (left) and with covariates (right)

2.3. *Conjugate Bayesian model without trend*

We now consider a zero-centred spatial process for $y(s)$ at a location $s \in D$ modelled as

$$y(s) = z(s) + \varepsilon(s). \tag{2.11}$$

Let $D \subset \mathbb{R}^d$ be a compact domain with $d \in \{1, 2, 3\}$. We also assume that the correlation function $R(\cdot;\cdot)$ is specified by the isotropic Matérn covariogram

$$R(s; s') := \frac{(\|s - s'\|)}{\Gamma(\nu)2^{\nu-1}} K_\nu(\|s - s'\|); \tag{2.12}$$

where $\nu > 0$ is a smoothness parameter, $\Gamma(\cdot)$ is the Gamma function, and $K_\nu(\cdot)$ is the modified Bessel function of the second kind of order ν (Abramowitz & Stegun, 1965, Section 10). It is

known that the spectral density of the (isotropic) Matérn model without nugget is

$$f_{\text{Matérn}}(u) = C\frac{2\ 2}{(\ ^2 + u^2)^{\ +d=2}} \quad \text{for some } C > 0: \tag{2.13}$$

Fix the smoothness parameter $\ > 0$, call the process (2.11) the Matérn model with parameter values $f\ ^2;\ ;\ ^2g$. Since there is no $x^\top$ term, the conjugate Bayesian model (2.2) simplifies to

$$y j\ ^2 \quad N(0;\ ^2(R(\ ) + \ ^2 I_n)); \quad ^2 \quad \text{IG}(a\ ; b\ ); \tag{2.14}$$

and the corresponding augmented linear regression becomes

$$\underset{y}{\begin{pmatrix} y \\ 0 \end{pmatrix}} = \underset{x}{\begin{pmatrix} I_n \\ I_n \end{pmatrix}} z + \underset{\ ^2}{\begin{pmatrix} 1 \\ 2 \end{pmatrix}}; \tag{2.15}$$

where $\ N(0;\ ^2 V_y\ )$ and $V_y = \begin{pmatrix} ^2 I_n & 0 \\ 0 & R(\ ) \end{pmatrix}$.

The following theorem shows the posterior inconsistency of the scale $\ ^2$ under the conjugate model (2.14).

THEOREM 3 (POSTERIOR INFERENCE FOR THE MATÉRN MODEL). *Assume that the location set* $\ = fs_1;:::;s_ng$ *satisfies*

$$\max_{s2D}\ \min_{1\ i\ n} js\ \ s_ij\ \ n^{\ \frac{1}{d}}: \tag{2.16}$$

Let P_0 *be the probability distribution of the Matérn model* (2.11) *with the true parameter values* $(\ ^2_0;\ _0;\ ^2_0)$. *Under* P_0,

$$\lim_{n\ !\ 1} p(\ ^2 j y(\ _n)) = \ _{\frac{2}{0}=\ ^2}: \tag{2.17}$$

Consequently, $\lim_{n\ !\ 1} p(\ ^2 j y(\ _n)) = \ ^2_0$.

Proof of Theorem 3. See Section D in the Appendix.

Note that the posterior inference of $\ ^2$ is independent of the range decay $\ $ chosen in the conjugate Bayesian model (2.14). The scale $\ ^2$ is posterior inconsistent unless the noise-to-spatial variance ratio $\ ^2 = \ ^2_0=\ ^2_0$, while the nugget $\ ^2$ is posterior consistent.

Posterior prediction for $z(\ )$ *and* $y(\ )$: Recall the decomposition (2.9) for the posterior prediction error for the latent process $z(s)$. The following proposition simplifies the two sources of error $E_{1;n}$ and $E_{2;n}$, when the outcome $y(s)$ follows a Matérn model in the presence of a nugget.

THEOREM 4 (POSTERIOR PREDICTIVE CONSISTENCY FOR THE MATÉRN MODEL). *Let* $s_0\ 2\ D$. *Then, we have the decomposition* (2.9), *where* $E_{1;n}$ *is the prediction error of the best linear predictor for a Matérn model with parameters* $f\ ^{02};\ ;\ ^{02}g$ *satisfying* $\ ^2 = \ ^{02}_{02}$, *and*

$$E_{2;n} := \ ^{\frac{2}{0}}_{\frac{2}{2}}\ 1 + J^\top_{;n}\ (I_n + \ ^2 R(\ _n))^{\ 1}\ \ I_n\ \ R(\ _n)^{\ 1} J_{\ ;n}: \tag{2.18}$$

Moreover, if $E_{1;n}; E_{2;n}\ !\ 0$ *as* $n\ !\ 1$, *then the latent process* $z(s)$ *is posterior predictive consistent in the sense that* $\mathsf{E}_0(Z_n(s_0)\ \ z(s_0))^2\ !\ 0$ *as* $n\ !\ 1$ *and, hence,* $\mathsf{E}_0(Y_n(s_0)\ \ y(s_0))^2\ !\ 2\ ^2_0$ *as* $n\ !\ 1$.

Proof. See Section E in the Appendix.

Theorem 4 reveals that the posterior mean of the conjugate Bayesian model (2.14) is identified as the best linear predictor of *any* Matérn model with parameters $f\ ^{02};\ ;\ ^{02}g$ provided $^{02}=\ ^{02} = \ ^2$. This observation connects the Bayesian modelling to a frequentist approach in

that the deviation error $E_{1;n}$ is viewed as the prediction error of the best linear predictor of a Matérn model in the presence of a nugget. Next we provide evidence to support the posterior predictive consistency for the latent process in the sense that $E_{1;n}; E_{2;n} \to 0$ as $n \to \infty$.

The deviation error $E_{1;n}$: It is expected that the prediction error of the best linear predictor of a Matérn model in the presence of a nugget tends to 0 as long as the fill distance condition (2.16) holds. However, it is hard to prove this statement in the general case. To provide some ideas, we consider a particular one-dimensional example.

We assume without loss of generality that $D = [-1; 1]$, and $= \{i/n; -n \le i \le n\}$ so the fill distance condition (2.16) is satisfied. Also let $_1 = \{i/n; -\infty < i < \infty\}$ be the infinite evenly spaced grid. Define

$$\hat{b} := \mathrm{E}(z(0) \mid y(s); s \in \setminus \{0\}) \quad (\text{resp. } \hat{b}_1 := \mathrm{E}(z(0) \mid y(s); s \in {}_1 \setminus \{0\}));$$

be the best linear predictor of $z(0)$ based on the observations $y(s)$ on $\setminus\{0\}$ (resp. ${}_1 \setminus \{0\}$). Note that $\hat{b}$ (resp. $\hat{b}_1$) may well be computed using misspecified parameter values $\{\ ; {}^2\}$. Let

$$e := \mathrm{E}_0(z(0) - \hat{b})^2 \quad (\text{resp. } e_1 := \mathrm{E}_0(z(0) - \hat{b}_1)^2); \tag{2.19}$$

be the prediction error of the best linear predictor based on the observations on the finite grid $\setminus\{0\}$ (resp. the infinite grid ${}_1 \setminus \{0\}$). We make the following assumption which relates e to e_1 as $n \to \infty$.

Assumption 2 *(Infinite approximation).* The difference $e - e_1 \to 0$ as $n \to \infty$.

Basically, this assumption suggests that as the grid becomes finer and finer, observations outside any fixed bounded interval have little impact on the prediction of $z(0)$. However, it is not easy to prove such a statement which is conjectured in (Stein, 1999, page 97). The plan is to prove $e_1 \to 0$, and by Assumption 2 it implies that $e \to 0$ as $n \to \infty$. This program is recorded in the following proposition.

PROPOSITION 1 (PREDICTION ERROR FOR THE MATÉRN MODEL). *Let* $D = [-1; 1]$ *and* $= \{i/n; -n \le i \le n\}$. *Under Assumption 2, we have*

$$e \to 0 \quad \text{as } n \to \infty; \tag{2.20}$$

where e *is the prediction error of the best linear predictor defined as in* (2.19).

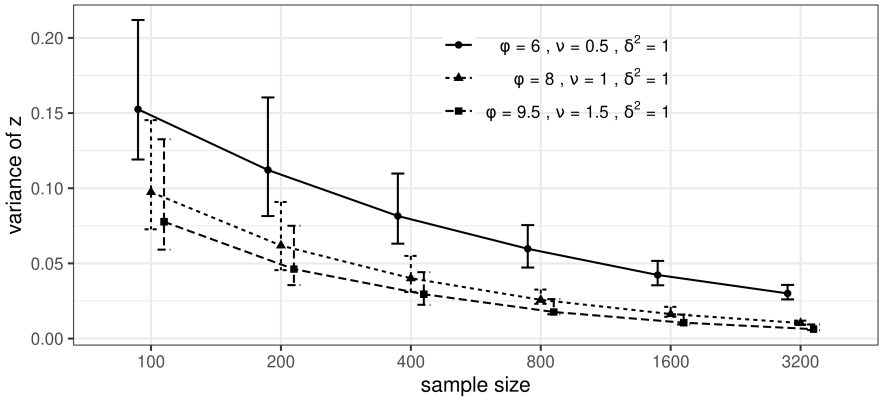
Proof. See Section F in the Appendix.

The posterior variance $E_{2;n}$: The error term $E_{2;n}$ defined by (2.18) is also analytically intractable. Here, we provide a numerical study to investigate the behaviour of $E_{2;n}$ when $n \to \infty$ in the general case. We first generate the location set by uniformly sampling n locations in $[0; 1]$ or $[0; 1]^2$, then we compute $E_{2;n}$ for every location in with ${}^2 = 1; \ {}^2 = 1$ and different values of and . We expand the location set sequentially to sets with larger sample sizes by adding locations that are uniformly sampled in the study domain. For each expanded set , we recompute the $E_{2;n}$ for all locations in . Figure 2 plots the median, the 2.5th and 97.5th percentiles of $E_{2;n}$ for different sample sizes. The values of $E_{2;n}$ for points in a fixed domain, shown in Figure 2, decrease rapidly as the sample size increases, although the rate diminishes when d increases from 1 to 2. This suggests that the decreasing rate is related to dimension.

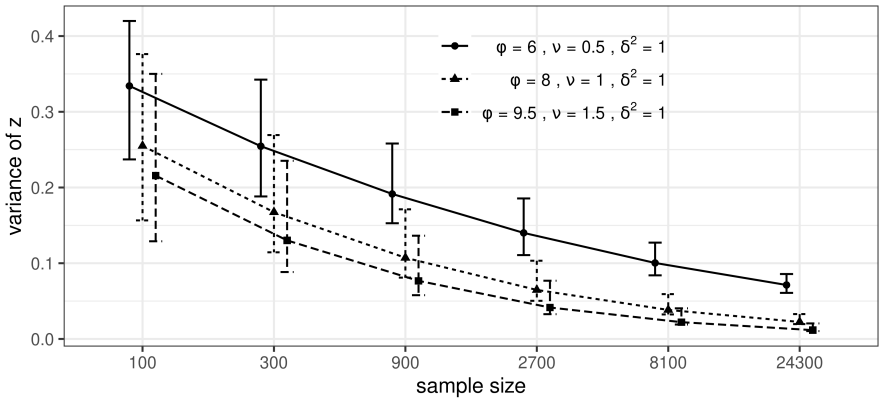
3. STACKING ALGORITHMS FOR SPATIAL MODELS

3.1. Stacking of means

R The Bayes predictor for $y(s_0)$ under model M_g, for each $g = 1; \ldots; G$, is $\mathrm{E}_g(y(s_0) \mid y) = \int y(s_0) p(y(s_0) \mid y; M_g) dy(s_0)$, where $\mathrm{E}_g(y(s_0) \mid y)$ is the expectation of the posterior predictive



(a)



(b)

Fig. 2: The median of $E_{2;n}$ for locations uniformly sampled on $[0; 1]$ (a) and $[0; 1]^2$ (b). The error bars indicate the 97.5th and 2.5th percentiles. The sample size n ranges from 100 to 3,200 and from 100 to 24,300 for the experiments on $[0; 1]$ and $[0; 1]^2$, respectively.

mean computed from M_g. Stacking methods will combine the G Bayes predictors as a weighted average

$$\sum_{g=1}^{G} w_g \mathrm{E}_g(y(s_0) \mid y); \tag{3.1}$$

where $fw_1; : : : ; w_G g$ are the weights for combination. Define the leave-one-out (LOO) Bayes predictor for $y(s_i)$ under model M_g as

$$\hat{y}_{g(-i)}(s_i) = \mathrm{E}_g(y(s_i) \mid y_{-i}; M_g) = \int\int y(s_i)\, p(y(s_i) \mid \cdot; M_g)\, p(\cdot \mid y_{-i}; M_g)\, d\cdot\, dy(s_i)$$

where y_{-i} is the data without the ith observation. Stacking determines the optimal weights as

$$\arg\min_w \sum_{i=1}^{n} \Big(y(s_i) - \sum_{g=1}^{G} w_g \hat{y}_{g(-i)}(s_i) \Big)^2 : \tag{3.2}$$

The following proposition studies the posterior prediction error using the stacking predictor, where we only require that the weights be bounded.

PROPOSITION 2. *Let $s_0 \in D$, and $w := (w_1, \ldots, w_G)$ be the stacking weights defined by* (3.2) *such that $|w_g|$ is bounded for each $1 \leq g \leq G$. Let the assumptions in Theorem 2 or Theorem 4 hold for each model M_g. We have*

$$\mathrm{E}_0\left(y(s_0) - \sum_{g=1}^{G} w_g \mathrm{E}_g(y(s_0) \mid y)\right)^2 \to \tau_0^2 \quad \text{as } n \to \infty. \tag{3.3}$$

Proof. See Section G in the Appendix.

The following theorem shows that the stacking predictor asymptotically minimises the posterior prediction error.

THEOREM 5. *Let $s_0 \in D$, and $w := (w_1, \ldots, w_G)$ be the stacking weights defined by* (3.2) *such that $|w_g|$ is bounded for each $1 \leq g \leq G$. Let*

$$E_{1;g;i} := \mathrm{E}_0(z(s_i) - \hat{y}_g(s_i))^2; \quad 1 \leq g \leq G \text{ and } 1 \leq i \leq n;$$

be the deviation error for the latent process $z(s)$ by leaving the i^{th} observation out under the model M_g. Assume that for each $1 \leq g \leq G$,

$$\frac{1}{n}\sum_{i=1}^{n} E_{1;g;i} \to 0 \quad \text{as } n \to \infty. \tag{3.4}$$

Also, if the assumptions in Theorems 2 or 4 hold for each model M_g, then as $n \to \infty$ we obtain

$$\mathrm{E}_0\left(y(s_0) - \sum_{g=1}^{G} w_g \mathrm{E}_g(y(s_0) \mid y)\right)^2 \mathrm{E}_0\left(\frac{1}{n}\sum_{i=1}^{n} \left(y(s_i) - \sum_{g=1}^{G} w_g \hat{y}_g(s_i)\right)^2\right) \to 0. \tag{3.5}$$

Proof. See Section H in the Appendix.

Equation (3.4) implies that for each model M_g the average deviation error for the latent process goes to 0 as the sampling resolution becomes finer. This condition is consistent with the fact that in the one-dimensional grid we typically have $E_{1;g;i} \asymp n^{-\min(2\nu_0, \frac{2}{2\nu+1})}$ (see Proposition 1); hence $\frac{1}{n}\sum_{i=1}^{n} E_{1;g;i} \to 0 \to 0$ as $n \to \infty$.

Theorem 5 holds for candidate models with a misspecified smoothness parameter ν in the Matérn kernel for $d = 1$. Note that Clyde & Iversen (2013, p.487) proves a similar result, while the established theoretical results about stacking assume exchangeability which is not available in geostatistical models. The innovation in our theoretical results emerge from studying the behaviour of the posterior and predictive distributions within the infill asymptotic paradigm. The proof of Theorem 5 can be extended to the case where the LOO Bayes prediction $\hat{y}_g(s_i)$ is replaced by a much cheaper prediction based on K-fold cross-validation.

3.2. Stacking of predictive densities

Following the generalised Bayesian stacking framework established in Yao et al. (2018), we devise a second stacking algorithm for spatial analysis, which we refer to as *stacking of predictive densities*. This algorithm finds the distribution in the convex hull $\mathcal{C} = \{\sum_{g=1}^{G} w_g \, p(\cdot \mid M_g) : \sum_g w_g = 1; w_g \geq 0\}$ that is optimal according to some proper scoring functions. Here $p(\cdot \mid M_g)$ refers to the distribution of interest under model M_g. Let $S_1^G = \{w \in [0,1]^G : \sum_{g=1}^{G} w_g = 1\}$

and $p_t(\ \ jy)$ be the true posterior predictive distribution. Using the logarithmic score (corresponding to the KL divergence), we seek w so that

$$\max_{w2S_1^G} \frac{1}{n} \sum_{i=1} \log \left(\sum_{k=1} w_g p(y(s_i) jy_{\ i}; M_g) \right) : \tag{3.6}$$

The optimal distribution $\sum_{g=1}^{G} w_g p(y(s_i) jy_{\ i}; M_g)$ provides a "likelihood" of observing $y(s_i)$ on location s_i given other data. Therefore, $\prod_{i=1}^{n} \sum_{g=1}^{G} w_g p(y(s_i) jy_{\ i}; M_g))$ serves as a pseudo-likelihood that measures the performance of prediction based on the weighted average of the LOO predictors for all observed locations. Also, it provides an approximation of the expected log point-wise predictive density (ELPD)

$$\sum_{i=1} \int p_t(y(s_i)) \log \left(\sum_{g=1} w_g p(y(s_i) jy_{\ i}; M_g) \right) dy(s_i) \tag{3.7}$$

It is worth pointing out that the weights for stacking of means are not necessarily positive, while those for stacking of predictive densities must be non-negative. Relaxing this restriction imparts greater flexibility in prediction and improves prediction accuracy for stacking of means. On the other hand, we can no longer interpret negative weights as a reflection of our prior beliefs regarding . In our subsequent experiments, we restrict the stacking weights for both algorithms to be non-negative so that the diagnostic metric MLPD (the average log pointwise predictive density, introduced in Section 4) is valid for both algorithms.

3.3. Predictive stacking in finite sample spatial analysis

A critical step in solving stacking weights is the computation of the Bayes predictor and predictive density. Computing the exact LOO Bayes predictor and predictive densities for all observed locations $fs_1; : : : ; s_n g$ requires refitting a model n times. For a Gaussian latent variable model with the number of parameters larger than the sample size n, there are limited choices for approximating LOO predictors accurately without the onerous computation (see, e.g., Vehtari et al., 2016). Instead of the LOO cross-validation, we prefer using the much cheaper K-fold cross-validation to generate the predictions. Using K-fold cross-validation instead of LOO in stacking is first implemented in Breiman (1996b), where the simulation evinces that 10-fold cross-validation can provide more efficient predictors than LOO cross-validation. Here, we use $K = 10$ in the implementations in Section 4.

We offer the pseudo-code algorithms for stacking of means and stacking of predictive densities using the conjugate Bayesian spatial regression model in Section I. In the algorithms, we first partition the data into K-folds based on locations. Letting $X = [x(s_1) : \quad : x(s_n)]^{\mathsf{T}}$ be the design matrix, we use $X[k], y[k], \ _n[k]$ to denote the predictors, response and observed locations from k-th fold, respectively, and $X[\ k], y[\ k] \ _n[\ k]$ to denote the respective data not in k-th fold. The values of the prefixed hyper-parameters $f \ ; \ ; \ ^2 g$ of the conjugate Bayesian spatial regression model are picked from the grid G_{all}, which is expanded over the grids of candidate values as the cartesian product $G \quad G \quad G^2$ for $f \ ; \ ; \ ^2 g$. For stacking of means, we compute the expected latent process $z^{(k)}_{(\ ; \ ; \ ^2)}$ on locations in fold k for each $f \ ; \ ; \ ^2 g$ from G_{all} and then generate the corresponding expectation $y^{(k)}_{(\ ; \ ; \ ^2)}$ to obtain the stacking weights.

For stacking of predictive densities, we compute the log point-wise predictive density, $(lp_{(\ ; \ ; \ ^2)}(s))$, of $y(s)$ at location s for locations in each fold for all candidate models to find the stacking weights. The closed form of the log point-wise predictive density is derived in Section J

of the Appendix. We also devise a Monte Carlo algorithm for stacking of predictive densities, which is expounded in Section K of the Appendix. Solving the weights for stacking of means is a quadratic programming problem. We use the *solve.QP()* function offered by the *quadprog* package in the R statistical computing environment and the solver *Mosek* to solve the weights in our R and Julia implementations, respectively. The stacking weights for stacking of predictive densities are calculated by *Mosek* in both our R and Julia implementations.

4. SIMULATION

4.1. Simulation settings

We present two simulations to examine the predictive performance of the proposed stacking algorithms. The data sets for experiments are generated by model (2.1) on locations sampled uniformly over a unit square $[0; 1]^2$, where the correlation function R is specified by the Matérn covariogram (2.12). The sample size n of the simulated data sets ranges from 200 to 900, and we randomly pick $n_h = 100$ observations for checking predictive performance. The vector $x(s)$ consists of an intercept and a single predictor generated from a standard normal distribution. The true value of the parameters for generating the data in the first simulation are $= (1; 2)^\top$, $= 7$, $^2 = 1$, $^2 = 1$ and $= 1$. In the second simulation, we alter the hyper-parameters in the Matérn covariogram to $= 20$, $^2 = 1$, $^2 = 0.3$ and $= 0.5$.

We consider customary values of smoothness in spatial analysis, $2\ G = \{0.5; 1; 1.5; 1.75\}$. The candidate values for are selected so that the "effective spatial range", which refers to the distance where spatial correlation drops below 0.05, covers 0.1 and 0.6 times $\sqrt{2}$ (the maximum inter-site distance within a unit square) for all candidate values of . Here we set $G = \{3; 14; 25; 36\}$. Finally, we specify $G_{^2} = \{0.1; 0.5; 1; 2\}$ as the candidate grid for ^{2}. We assign an IG(a ; b) prior with $a = b = 2$ for ^{2}. The prior of follows a zero centred Gaussian with the covariance matrix equal to a diagonal matrix whose diagonal elements are 4, i.e., N(; V) where $= \mathbf{0}$ and $V = 4\ I$. For each simulated data set, we use stacking of means and stacking of predictive densities to obtain the expected outcome $\hat{y}(s)$ based on the held out observed locations. The predictive accuracy is evaluated by the mean squared prediction error over a set of n_h hold-out locations in set S_h (MSPE $= \sum_{s2S_h}((\hat{y}(s) \quad y(s))^2)/n_h$). We also compute the posterior expected values of the latent process $\hat{z}(s)$ for $z(s)$ on all of the n sampled locations in S and evaluate the mean squared error for $z(s)$ (MSEZ $= \sum_{s2S}(\hat{z}(s) \quad z(s))^2/n$). To further evaluate the distribution of predicted values, we compute the mean log point-wise predictive density for the n_h held out locations (MLPD $= \sum_{s2S_h}\{\log(\sum_{g=1}^{G} w_g p(y(s) \mid y; M_g))\}/n_h$).

Apart from stacking, we also implemented a fully Bayesian model with priors on the hyper-parameters using Markov chain Monte Carlo (MCMC) sampling for comparison. In addition, we carried out exact Bayesian inference using the conjugate model in Section 2.1 with hyper-parameters fixed at the exact value (denoted as M_0). We use the same priors for ^{2} and as those in stacking implementations. For the rest of the priors needed in full MCMC sampling, we assign uniform priors U(3; 36) for and U(0.25; 2) for , and an IG(2; 2) prior for ^{2}. Sampling is fitted through the *spLM* function in the *spBayes* package in R. The diagnostic metrics are computed based on 1,000 posterior samples retained after convergence was diagnosed over a burn-in period of 10,000 initial iterations. The algorithm for recovering the expected z and the log point-wise predictive density based on the output of `spLM` is presented in Section L of the Appendix. We monitor all diagnostic metrics for prediction for all competing algorithms. To measure uncertainty of the diagnostic metrics, we generate 60 data sets for each sample size in each simulation, fit each data set with the four competing methods and record the diagnostic metrics of each model fitting.

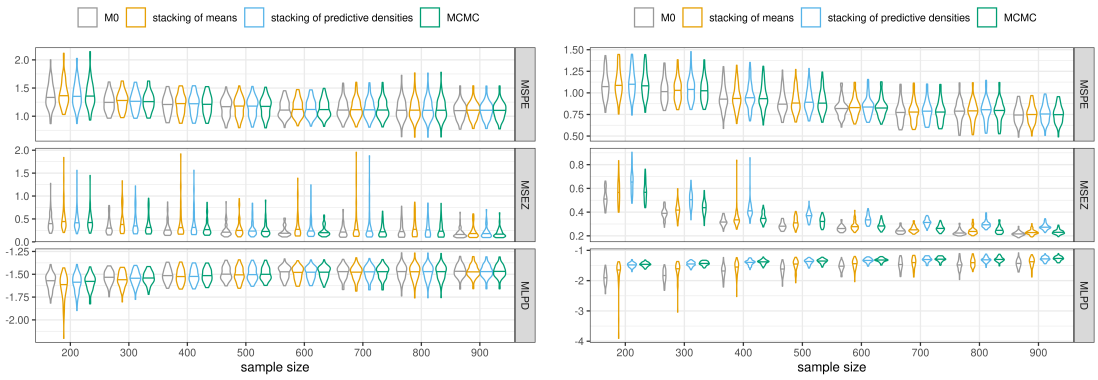


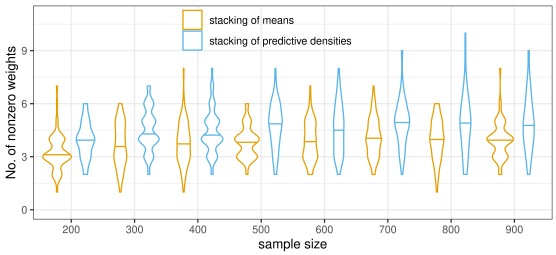
Fig. 3: *Distributions of the diagnostic metrics for prediction performance for the first simulation (a) and the second simulation (b). Each distribution is depicted through a violin plot. The horizontal line in each violin plot indicates the median.*

4.2. Predictive performances

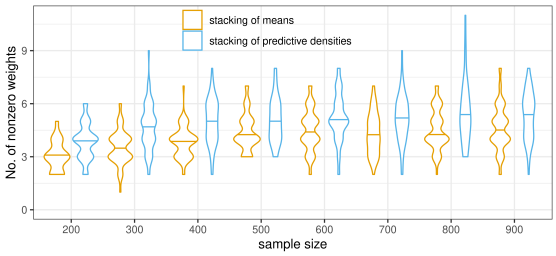
Interestingly, the candidate algorithms exhibit different behaviours in the two simulation studies. Figure 3 summarises the comparison of the predictive performance. In the first simulation there seem to be no pronounced distinctions between the prediction performance for all competing models. In the second simulation, however, we observe that stacking of means outperforms stacking of predictive densities on having better estimates of the latent process on both the observed and unobserved locations (based on MSEZ), while stacking of predictive densities outperforms stacking of means in terms of the log point-wise predictive density (based on MLPD). These results are expected since we optimise the prediction error in the stacking of means, and we maximise the log predictive densities in the stacking of predictive densities.

Treating the fully Bayesian model with priors on all hyperparameters (fitted using MCMC) as a benchmark, we find that stacking of predictive densities is very competitive in terms of MLPD. The performance of latent process estimation for the full Bayesian model falls between stacking of means and stacking of predictive densities. The prediction accuracy for the response on unobserved locations for all competing algorithms are very close. Based on the medians of the MSPEs for all fittings, stacking of means slightly outperforms the full Bayesian model, and both are slightly better than stacking of predictive densities, The conjugate Bayesian model M_0 provides the best point estimates for both response and latent processes based on MSPE and MSEZ, while it performs worse in terms of MLPD. These results seem to indicate that, when choosing stacking algorithms, stacking of means is preferred for point estimation and stacking of predictive densities is preferable for interval estimation.

In Figure 4, we compare the counts of the non-zero weights in stacking and find that stacking of means tends to produce a slightly smaller number of non-zero weights than stacking of predictive densities. The number of non-zero weights is small for both stacking algorithms, and this phenomenon is also observed and discussed in Breiman (1996b). On average, there are around 3.7 and 4.5 out of 64 weights that are greater than 0.001 in the simulation studies for stacking of means and stacking of predictive densities, respectively. And the number is relatively consistent when the sample sizes increase. These results suggest that stacking methodology is memory efficient, which is beneficial for large-scale spatial data analysis.



(a)



(b)

Fig. 4: Distributions of the counts of nonzero weights in the first (a) and the second (b) simulation. The distribution of the counts are described through violin plots whose horizontal lines indicate the medians.

In Section M.1, we pick one run for each simulation study and illustrate the interpolated maps of the predicted response on held out locations and expected latent processes over all locations generated by different fitting algorithms. The predicted response resembles the de-noised response ($x(s)^\top \beta + z(s)$, not $y(s)$), and the estimated latent process shares a similar pattern with the raw data. The latent process estimated by stacking of predictive densities is observed to be slightly smoother than those estimated by other algorithms, while the predictions of the response on unobserved locations fitted by different fitting algorithms are almost indistinguishable.

4.3. Running time comparisons

We provide both R and Julia code for the two proposed stacking algorithms. The code for the simulation studies are available from the GitHub repository `https://github.com/LuZhangstat/spatial_stacking`. Comparisons in predictive performances presented above are conducted in R. For the running time comparisons reported here, the stacking algorithms are implemented using Julia-1.6.6 and the MCMC sampling algorithms are run in R-4.2.0. We report the time for obtaining weights for stacking, and we consider the sampling time for $f\ ;\ ;\ ^2;\ ^2g$ using MCMC (no sampling of $f\ ;zg$ and no predictions). The timing comparisons are based upon experiments on a Windows 10 pro platform, with 64 GB of RAM and a 1-Intel Core i7-7700K CPU @ 4.20GHz processor with 4 cores each and 2 threads per core—totalling 8 possible threads for parallel computing. Figure 5 summarises the running time for the three competing algorithms. On average, the stacking of means is 526 times faster than MCMC in the simulation study. Stacking of predictive densities is slightly slower than stacking of means but is still around 430 times faster than MCMC sampling. These experiments clearly establish that predictive stacking algorithms are efficient alternatives to the full MCMC sampling algorithms for estimating latent spatial processes and predicting geostatistical outcomes.

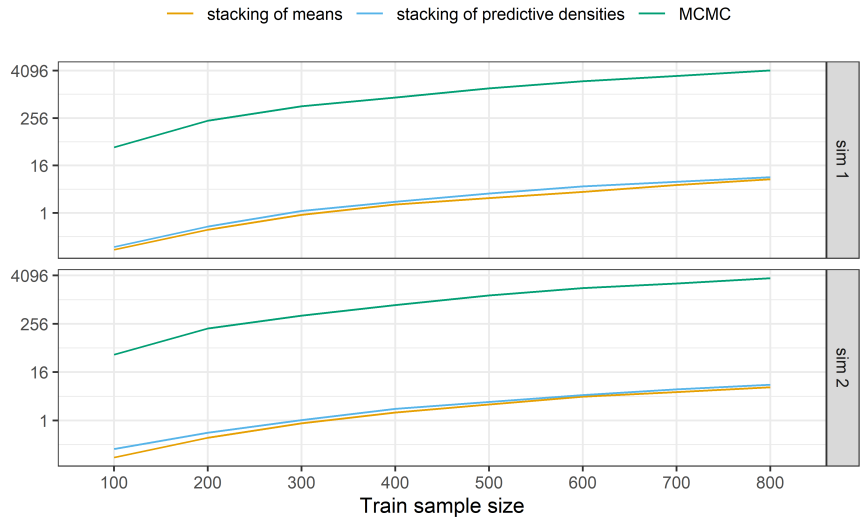


Fig. 5: Running time comparison for stacking and MCMC sampling

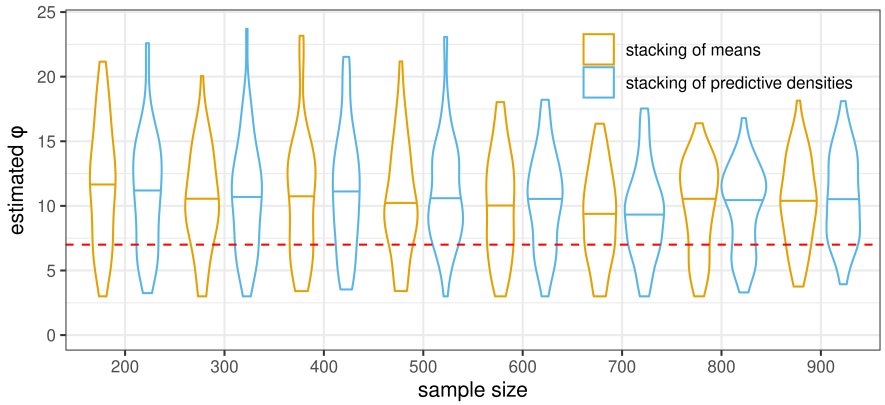
4.4. *Inference of prefixed hyper-parameters*

One limitation of stacking, compared to full Bayesian inference (e.g., using MCMC sampling algorithms), is that it does not provide interval estimates for the prefixed hyper-parameters. Moreover, our experiments show that stacking cannot provide a reliable point estimate of the hyper-parameters. If we treat the grid of the candidate values for the hyper-parameters in our stacking algorithms as a discrete uniform prior, then, intuitively, we should be able to achieve point estimates for those fixed hyper-parameters by a weighted average based on stacking weights. In Figure 6, we compare the point estimates of based on stacking for the simulation studies. It is clear that stacking of means yields unstable estimates. Stacking of predictive densities has a smaller variance, but the bias can be large. Meanwhile, since is not identifiable, we observe that the posterior interval estimates for inferred from MCMC algorithms are wide, showing that the inference for is relatively unstable for all candidate algorithms in this simulation study. The comparisons of the other two hyper-parameters are provided in the Section M.2.

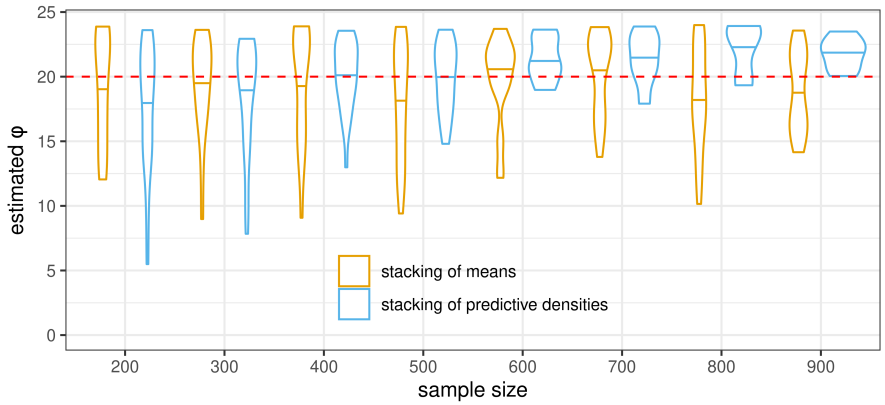
5. WESTERN EXPERIMENTAL FOREST INVENTORY DATA ANALYSIS

We illustrate our proposed stacking methodology using the Western Experimental Forest (WEF) inventory data from a long-term ecological research site in western Oregon. This data set contains coordinates, species, diameter at breast height (DBH) and other measurements for trees in the experimental forest. The raw data is provided in the R package *spBayes*, and the code is available from the GitHub repository `https://github.com/LuZhangstat/spatial_stacking`. In this analysis, we focus on the prediction of DBH for living trees using the location and species information. There are 1,954 records in all after cleaning up some of the data. We randomly hold out 500 points for prediction and use the remaining observations to train the model and draw posterior inference.

We first fit a Bayesian linear regression model (BLM) using Hamiltonian Monte Carlo algorithms implemented in the Stan Bayesian modelling environment through the R interface *rstanarm*. We run the model fitting with the default settings, where the default priors are described in



(a)



(b)

Fig. 6: Distributions of the estimated in the first (a) and the second (b) simulation. The distribution of the counts are described through violin plots whose horizontal lines indicate the medians. The red dashed horizontal line indicates the actual value of

Gabry & Goodrich (2020). Then, we use the same prior for in our stacking implementations. For the prior of ^{2} in stacking implementations, we set up the hyper-parameters in the inverse-Gamma to have mean at the square of the mean of the default prior for in BLM. We use the same candidate values for and ^{2} as our simulation studies in Section 4. The largest and smallest candidate values for are the maximum and minimum values so that the effective spatial range covers 0.1 and 0.6 times the maximum inter-site distance for Matérn covariogram with all candidate smoothness values. We pick four candidate values for in this data analysis. Finally, we evaluate predictive performance using the root mean squared prediction error (square root of the MSPE defined in Section 4.1) and the average log point-wise predictive density (MLPD defined in Section 4.1) for all held out locations. Table 1 summarises the result. Stacking delivers an RMSPE that is lower by around 9% and yields better MLPD in this data analysis.

Table 1: Summary Statistics of Western experimental forest inventory data analysis

	Bayesian linear regression	stacking of means	stacking of predictive densities
RMSPE	22.86	20.70	20.79
MLPD	-4.55	-4.45	-4.44

6. CONCLUSION AND FUTURE WORK

We develop geostatistical inference using Bayesian stacking. We offer theoretical insights for inferential behaviour of posterior distributions in fixed-domain or infill settings and explore the performance of our methods through simulations and analysis of a forestry data set. The empirical results reveal that our devised stacking methods deliver predictions comparable to full Bayesian inference obtained using MCMC samples, but at significantly lower costs. Our proposed stacking algorithm can be implemented in parallel and made storage efficient and, hence, can become a powerful alternative to full Bayesian inference using MCMC samples. Future directions can build upon our current framework to extend Bayesian stacking for multivariate geostatistics using conjugate matrix-variate normal-Wishart families (Zhang et al., 2021) and conjugate exponential families for non-Gaussian data (Bradley et al., 2020). Another extension is will be in the direction of stacked Bayesian inference for high-dimensional geostatistics (for example, building on the conjugate framework in Banerjee, 2020).

REFERENCES

ABRAMOWITZ, M. & STEGUN, A. (1965). *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. Dover.

BANERJEE, S. (2019). Geostatistics for environmental processes. In *Handbook of Environmental and Ecological Statistics*, A. E. Gelfand, M. Fuentes, J. A. Hoeting & R. L. Smith, eds. CRC press, Boca Raton, FL, pp. 81–96.

BANERJEE, S. (2020). Modeling massive spatial datasets using a conjugate bayesian linear modeling framework. *Spatial Statistics* **37**, 100417.

BANERJEE, S., CARLIN, B. P. & GELFAND, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press, Boca Raton, FL.

BERGER, J. O., OLIVEIRA, V. D. & SANSÓ, B. (2001). Objective bayesian analysis of spatially correlated data. *Journal of the American Statistical Association* **96**, 1361–1374.

BERNARDO, J. & SMITH, A. (1994). *Bayesian Theory*. Chichester, UK: John Wiley and Sons.

BOSE, M., HODGES, J. S. & BANERJEE, S. (2018). Toward a diagnostic toolkit for linear models with gaussian-process distributed random effects. *Biometrics* **74**, 863–873.

BRADLEY, J. R., HOLAN, S. H. & WIKLE, C. K. (2020). Bayesian hierarchical models with conjugate full-conditional distributions for dependent data from the natural exponential family. *Journal of the American Statistical Association* **115**, 2037–2052.

BREIMAN, L. (1996a). Stacked regressions. *Machine Learning* **24**, 49–64.

BREIMAN, L. (1996b). Stacked regressions. *Machine learning* **24**, 49–64.

CHILÉS, J. & DELFINER, P. (1999). *Geostatistics: Modeling Spatial Uncertainty*. John Wiley: New York.

CLYDE, M. & IVERSEN, E. S. (2013). Bayesian model averaging in the m-open framework. *Bayesian theory and applications* **14**, 483–498.

CRESSIE, N. (1993). *Statistics for Spatial Data*. Wiley-Interscience, New York, revised ed.

DE OLIVEIRA, V. & HAN, Z. (2022). On information about covariance parameters in gaussian matern random fields. *Journal of Agricultural, Biological and Environmental Statistics* **27**, 690–712.

FINLEY, A. O., DATTA, A., COOK, B. C., MORTON, D. C., ANDERSEN, H. E. & BANERJEE, S. (2019). Efficient algorithms for bayesian nearest neighbor gaussian processes. *Journal of Computational and Graphical Statistics* **28**, 401–414.

GABRY, J. & GOODRICH, B. (2020). Prior distributions for rstanarm models. `http://mc-stan.org/rstanarm/articles/priors.html`.

HANDCOCK, M. S. & STEIN, M. L. (1993). A bayesian analysis of kriging. *Technometrics* **35**, 403–410.

HODGES, J. S. (2013). Richly parameterized linear models: Additive, time series, and spatial models using random effects .

HOETING, J. A., MADIGAN, D., RAFTERY, A. E. & VOLINSKY, C. T. (1999). Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors. *Statistical Science* **14**, 382 – 417.

KAUFMAN, C. G. & SHABY, B. A. (2013). The role of the range parameter for estimation and prediction in geostatistics. *Biometrika* **100**, 473–484.

LE, T. & CLARKE, B. (2017). A bayes interpretation of stacking for m-complete and m-open settings. *Bayesian Analysis* **12**, 807–829.

LI, C., SUN, S. & ZHU, Y. (2023). Fixed-domain posterior contraction rates for spatial gaussian process model with nugget. *Journal of the American Statistical Association* **0**, 1–21.

MADIGAN, D., RAFTERY, A. E., VOLINSKY, C. & HOETING, J. (1996). Bayesian model averaging. In *Proceedings of the AAAI Workshop on Integrating Multiple Learned Models, Portland, OR*.

MURPHY, K. P. (2015). Conjugate bayesian analysis of the gaussian distribution, 2007. *URL https://www. cs. ubc. ca/˜ murphyk/Papers/bayesGauss.pdf* .

RASMUSSEN, C. E. & WILLIAMS, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press, Cambridge, MA.

STEIN, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag, New York.

TANG, W., ZHANG, L. & BANERJEE, S. (2021). On identifiability and consistency of the nugget in Gaussian spatial process models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83**, 1044–1070.

VEHTARI, A., MONONEN, T., TOLVANEN, V., SIVULA, T. & WINTHER, O. (2016). Bayesian leave-one-out cross-validation approximations for gaussian latent variable models. *The Journal of Machine Learning Research* **17**, 3581–3618.

WOLPERT, D. H. (1992). Stacked generalization. *Neural networks* **5**, 241–259.

YAO, Y., PIRεs, G., VEHTARI, A. & GELMAN, A. (2021). Bayesian hierarchical stacking: Some models are (somewhere) useful. *Bayesian Analysis* **1**, 1–29.

YAO, Y., VEHTARI, A. & GELMAN, A. (2020). Stacking for non-mixing bayesian computations: The curse and blessing of multimodal posteriors. *arXiv preprint arXiv:2006.12335* .

YAO, Y., VEHTARI, A., SIMPSON, D. & GELMAN, A. (2018). Using stacking to average bayesian predictive distributions (with discussion). *Bayesian Analysis* **13**, 917–1007.

ZHANG, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association* **99**, 250–261.

ZHANG, H. & ZIMMERMAN, D. L. (2005). Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika* **92**, 921–936.

ZHANG, L., BANERJEE, S. & FINLEY, A. O. (2021). High-dimensional multivariate geostatistics: A Bayesian matrix-normal approach. *Environmetrics* **32**, e2675.

ZIMMERMAN, D. & STEIN, M. (2010). Classical geostatistical methods. In *Handbook of spatial statistics*, A. E. Gelfand, P. Diggle, P. Guttorp & M. Fuentes, eds. CRC press, Boca Raton, FL, pp. 29–44.

A. PROOF OF THEOREM 1

Proof. For ease of presentation, we denote $y(\ _n)$ by y_n in the proofs. By Assumption 1, there is a probability distribution $P^0 \quad P_0$, which corresponds to the model with parameters $(\ _0; \ ^{02}; \ ; \ ^2_0)$ for some $\ ^{02}$. Under P^0, there exists a $\ _0$ such that

$$y_y \quad X_y\ _0 j\ ^2 \quad N\ (0; V^0)\ ; \tag{A.1}$$

where $V^0 = \begin{pmatrix} 0 & & & 1 \\ \ ^2_0 I_n & 0 & 0 & \\ 0 & \ ^2 I_p & 0 & \\ 0 & 0 & \ ^{02} I_n & \end{pmatrix}$. Consider the sequence $\ _n \quad p(\ ^2 j y_n)$. Under P^0, $\ _n$ $IG(a\ _{;n}; b\ _{;n})$, where $a\ _{;n} = a\ + n=2$ and

$$\begin{aligned} b\ _{;n} &= b\ + \frac{1}{2}(y_y \quad X_y\ _0)^\top (I_{2n+p} \quad H)(y_y \quad X_y\ _0) \\ &= b\ + \frac{1}{2}[Q(y_y \quad X_y\ _0)]^\top \begin{pmatrix} 0 & 0 \\ 0 & I_n \end{pmatrix} [Q(y_y \quad X_y\ _0)]; \end{aligned} \tag{A.2}$$

The expectation under P^0 for $b_{;n}$ is

$$\begin{aligned} E^0(b_{;n}) &= b + \frac{1}{2}E^0(y_y^\top (I_{2n+p} - H)y_y) \\ &= b + \frac{1}{2}{}_0^\top X_y^\top (I_{2n+p} - H)X_y{}_0 + \mathrm{Tr}((I_{2n+p} - H)V^0) \\ &= b + \frac{1}{2}\mathrm{Tr}((I_{2n+p} - H)V^0) \quad (\text{since } (I_{2n+p} - H)X_y = O) \\ &= b + \frac{1}{2}\mathrm{Tr}\begin{pmatrix} Q_{21}^\top Q_{21} & O \\ O & {}^2I_p \end{pmatrix} + \frac{{}^02}{2}\mathrm{Tr}(Q_{22}^\top Q_{22}); \end{aligned} \tag{A.3}$$

where $Q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix}$ is an orthogonal matrix such that $H = Q^\top \begin{pmatrix} I_{n+p} & 0 \\ 0 & 0 \end{pmatrix} Q$ with Q_{22} being the lower right $n \times n$ block of Q in a 2×2 partition. Using $Q_{21}Q_{21}^\top = I_n - Q_{22}Q_{22}^\top$ and some further simplification we obtain

$$E^0(b_{;n}) = b + \frac{n}{2}\left(\frac{{}^2_0}{2}(1 - \mathrm{Tr}(Q_{22}^\top Q_{22}/n)) + {}^{02}\mathrm{Tr}(Q_{22}^\top Q_{22}/n)\right) + \frac{p}{2}\underbrace{{}^2\frac{{}^2_0}{2}}_{o(n)}. \tag{A.4}$$

Since $E^0(\ _n) = E^0(b_{;n})/(a_{;n} - 1)$ and $\mathrm{Tr}(Q_{22}^\top Q_{22}) = n - \mathrm{Tr}(H_{22})$, where $\mathrm{Tr}(H_{22})/n \to 0$ as $n \to \infty$, we obtain $\lim_{n\to\infty} E^0(\ _n) = {}^2$. The variance of $\ _n$ under P^0 is given by

$$\begin{aligned} V^0(\ _n) &= E^0[V(\ _n|y_n)] + V^0[E(\ _n|y_n)] \\ &= E^0\left[\frac{b_{;n}^2}{(a_{;n} - 1)^2(a_{;n} - 2)}\right] + V^0\left[\frac{b_{;n}}{a_{;n} - 1}\right]. \end{aligned} \tag{A.5}$$

Further note that

$$\begin{aligned} V^0(b_{;n}) &= \frac{3}{4}\mathrm{Tr}\left[\frac{{}^2_0}{2}(I_n - Q_{22}^\top Q_{22}) + {}^{02}Q_{22}^\top Q_{22}\right]^2 + o(n) \\ &\le C\left[n + \mathrm{Tr}(Q_{22}^\top Q_{22}) + \mathrm{Tr}\left[(Q_{22}^\top Q_{22})^2\right]\right]; \end{aligned} \tag{A.6}$$

for some $C > 0$ independent of n. Since $Q_{21}Q_{21}^\top + Q_{22}Q_{22}^\top = I_n$, we have $\mathrm{Tr}(Q_{22}^\top Q_{22})$ and $\mathrm{Tr}\left[(Q_{22}^\top Q_{22})^2\right]$ are bounded from above by n. Hence,

$$\frac{V^0(b_{;n})}{n^2} \to 0 \quad \text{and} \quad \frac{E^0(b_{;n}^2)}{n^{2+}} \to 0 \text{ for any } > 0. \tag{A.7}$$

Combining (A.5) and (A.7) yields $\lim_{n\to\infty} V^0(\ _n) = 0$. By Chebyshev's inequality, $\ _n$ converges in probability under P^0, and hence under P_0, to ^{2}.

B. PROOF OF COROLLARY 1

Proof. Again we denote $y(\ _n)$ by y_n. The conditional posterior distribution $p(\ |{}^2; y_n)$ can be derived from Lemma 1 as $N(\ |(M\ m)_{[1:p]}, {}^2(M\ )_{[1:p;1:p]})$, where $M\ ^{-1} = V\ ^{-1} + X^\top R(\ _n) + {}^2I_n\ ^{-1}X$ and $m\ = V\ ^{-1}\ + X^\top {}^2R(\ _n) + I_n\ ^{-1}y_n$. Let $\ _n\ p(\ |y_n)$. Some straightforward algebra yields

$$E^0\|\ _n\ {}_0\|^2 = \frac{{}^2_0}{4}\mathrm{Tr}(B_{11}) + E^0(\ _n)\mathrm{Tr}(C) + {}^{02}\mathrm{Tr}(D) + E^0(\ _n)\mathrm{Tr}(U_{11}). \tag{B.1}$$

Since $E^0(\ _n) \to {}^2 < 1$ from Theorem 1, the proof follows from (2.8) and (B.1).

Derivation of (B.1)

$$
\begin{aligned}
&E^0\|k_{n,\theta} - {}_0k\|^2 = E^0[E^0\{f(\theta_n,\theta) - {}_0f(\theta)\}^\top\{f(\theta_n,\theta) - {}_0f(\theta)\}\,|\,y_n] \\
&= E^0[E^0\{f(\theta_n,\theta) - E^0(f(\theta_n,\theta)|y_n) + E^0(f(\theta_n,\theta)|y_n) - {}_0f(\theta)\}^\top\{f(\theta_n,\theta) - E^0(f(\theta_n,\theta)|y_n) + E^0(f(\theta_n,\theta)|y_n) - {}_0f(\theta)\}\,|\,y_n] \\
&= E^0[E^0\{f(\theta_n,\theta) - E^0(f(\theta_n,\theta)|y_n)\}^\top\{f(\theta_n,\theta) - E^0(f(\theta_n,\theta)|y_n)\}\,|\,y_n] \\
&\quad + 2E^0\Big[E^0\Big\{(E^0(f(\theta_n,\theta)|y_n) - {}_0f(\theta))^\top\underbrace{\{f(\theta_n,\theta) - E^0(f(\theta_n,\theta)|y_n)\}}_{=0}\Big\}\,|\,y_n\Big] \\
&\quad + E^0[E^0\{f(E^0(f(\theta_n,\theta)|y_n) - {}_0f(\theta))^\top(E^0(f(\theta_n,\theta)|y_n) - {}_0f(\theta))\,|\,y_n] \\
&= E^0[\mathrm{Tr}\{f V^0(f(\theta_n,\theta)|y_n)\}] + E^0[(E^0(f(\theta_n,\theta)|y_n) - {}_0f(\theta))^\top(E^0(f(\theta_n,\theta)|y_n) - {}_0f(\theta))]
\end{aligned}
\tag{B.2}
$$

Let $\theta_n \sim p(\theta|y_n)$. By the definition of θ_n, we have

$$
\begin{aligned}
V^0(f(\theta_n,\theta)|y_n) &= V^0[E^0(f(\theta,\theta^*)|y_n)|y_n] + E^0(V^0[f(\theta,\theta^*)|y_n]|y_n) \\
&= \underbrace{V^0[(M\,m)_{[1:p]}|y_n]}_{=0} + E^0(\sigma^2(M)_{[1:p],[1:p]}|y_n) \\
&= E^0(\sigma^2|y_n)(M)_{[1:p],[1:p]} = E^0(\sigma^2|y_n)U_{11}; \\
E^0(f(\theta_n,\theta)|y_n) &= E^0\{E^0(f(\theta,\theta^*)|y_n)|y_n\} = (M\,m)_{[1:p]}.
\end{aligned}
\tag{B.3}
$$

With $M = (X_y^\top X_y)^{-1} = U$, $m = X_y^\top y_y$, under P^0,

$$
\begin{aligned}
(M\,m)_{[1:p]}|\sigma^2 &\sim N(\beta_0, [M X_y^\top V^0 X_y M]_{[1:p],[1:p]} \\
&\sim N\Big(\beta_0, U\begin{pmatrix} \frac{\phi_0^2}{\sigma^4}X^\top X + \sigma^2 V^{-1} & \frac{\phi_0^2}{\sigma^4}X^\top \\ \frac{\phi_0^2}{\sigma^4}X & \frac{\phi_0^2}{\sigma^4}I_n + \sigma_0^2 R^{-1}(\theta_n) \end{pmatrix}_{[1:p],[1:p]} U\Big)
\end{aligned}
\tag{B.4}
$$

Therefore, $E^0[E^0(f(\theta_n,\theta)|y_n)] = {}_0f(\theta)$, and

$$
\begin{aligned}
&E^0[(E^0(f(\theta_n,\theta)|y_n) - {}_0f(\theta))^\top(E^0(f(\theta_n,\theta)|y_n) - {}_0f(\theta))] \\
&= \mathrm{Tr}\{f V^0[E^0(f(\theta_n,\theta)|y_n)]\} \\
&= \mathrm{Tr}\{f V^0[(M\,m)_{[1:p]}]\} \\
&= \mathrm{Tr}\{f \underbrace{V^0[E^0(\underbrace{(M\,m)_{[1:p]}|\sigma^2}_{=0})]}_{=0} + E^0\{V^0[(M\,m)_{[1:p]}|\sigma^2]\}\}
\end{aligned}
\tag{B.5}
$$

We separate the variance $V^0[(M\,m)_{[1:p]}|\sigma^2]$ as presented in (B.4) into the following three parts

$$
\underbrace{\frac{\phi_0^2}{\sigma^4}U\begin{pmatrix} X^\top X & X^\top \\ X & I_n \end{pmatrix}U_{[1:p],[1:p]}}_{B_{11}} + \sigma^2\underbrace{U\begin{pmatrix} V^{-1} & 0 \\ 0 & 0 \end{pmatrix}U_{[1:p],[1:p]}}_{C} + \sigma_0^2\underbrace{U\begin{pmatrix} 0 & 0 \\ 0 & R^{-1}(\theta_n) \end{pmatrix}U_{[1:p],[1:p]}}_{D}.
$$

Then

$$
E^0[(E^0(f(\theta_n,\theta)|y_n) - {}_0f(\theta))^\top(E^0(f(\theta_n,\theta)|y_n) - {}_0f(\theta))] = \frac{\phi_0^2}{\sigma^4}\mathrm{Tr}(B_{11}) + E^0(\sigma^2)\,\mathrm{Tr}(C) + \sigma_0^2\,\mathrm{Tr}(D).
\tag{B.6}
$$

Together with (B.2), (B.3), we can obtain (B.1).

C. PROOF OF THEOREM 2

Proof. For ease of presentation, we denote $y(\ _n)$ by y_n. Observe that

$$\begin{aligned}\mathsf{E}_0(Z_n(s_0) \quad z(s_0))^2 &= \mathsf{E}_0\, fZ_n(s_0) \quad \mathsf{E}(z(s_0)jy_n) + \mathsf{E}(z(s_0)jy_n) \quad z(s_0)g^2 \\ &= \mathsf{E}_0\, fz(s_0) \quad \mathsf{E}(z(s_0)jy_n)g^2 + \mathsf{E}_0 f\mathsf{V}(z(s_0)jy_n)g;\end{aligned} \tag{C.1}$$

where the second equality follows from the fact that $z(s_0) \quad \mathsf{E}(z(s_0)jy_n)$ is independent of y_n. Note that

$$p(z(s_0)jy_n) = \int p(z(s_0)jy_n;\ ^2;\ )\ \ p(\ jy_n;\ ^2)\ \ p(\ ^2jy_n)\,d\ ^2d\ : \tag{C.2}$$

By standard Gaussian conditioning (see e.g. (Rasmussen & Williams, 2006, Section 2.2)),

$$p(z(s_0)jy_n;\ ^2;\ ) = N(J^{\mathsf{T}}_{;n}R^{\ 1}_{;n}z;\ ^2(1 \quad J^{\mathsf{T}}_{;n}R^{\ 1}_{;n}J_{\ ;n})): \tag{C.3}$$

By (C.2), (C.3) and Lemma 1, the posterior predictive mean is

$$\mathsf{E}(z(s_0)jy_n) = F_n: \tag{C.4}$$

Further by the law of total variance and Theorem 1, we get

$$\begin{aligned}&\mathsf{V}(z(s_0)jy_n) = \mathsf{E}f\mathsf{V}(z(s_0)jy_n;\ ^2;\ )g + \mathsf{V}f\mathsf{E}(z(s_0)jy_n;\ ^2;\ )g \\ &\ !\quad ^2(1 \quad J^{\mathsf{T}}_{;n}R^{\ 1}_{;n}J_{\ ;n}) + \ ^2J^{\mathsf{T}}_{;n}R^{\ 1}_{;n}U_{n[p+1:p+n;\,p+1:p+n]}R^{\ 1}_{;n}J_{\ ;n} = \ ^2G_n:\end{aligned} \tag{C.5}$$

Combining (C.1), (C.4) and (C.5) yields the decomposition (2.9), and hence the posterior predictive consistency for $z(s_0)$ holds if $E_{1;n}; E_{2;n}\ !\ 0$ as $n\ !\ 1$.

Further, let $\ _n$ have the density $p(\ jy_n)$. Under conditions (2.8), by Corollary 1 we have $\mathsf{E}_0 jj\ _n$ $_0jj^2\ !\ 0$ as $n\ !\ 1$. As a result,

$$\mathsf{E}_0(Y_n(s_0) \quad y(s_0))^2 = \mathsf{E}_0(x(s_0)^{\mathsf{T}}(\ _n \quad _0))^2 + E_{1;n} + E_{2;n} + \ ^2_0 + \ ^2\mathsf{E}_0(\ ^2jy_n);$$

which clearly converges to $\ ^2_0 + \ ^2\ ^2$ as $n\ !\ 1$.

D. PROOF OF THEOREM 3

The proof of this theorem breaks into several lemmas. Recall the definition of b from Lemma 1. The lemma below provides a simple expression for $b_{\ ;n}$, which is specific to the conjugate model (2.14).

LEMMA 2. *We have* $b_{\ ;n} = b + \frac{1}{2}y(\ _n)^{\mathsf{T}}(\ ^2I_n + R(\ _n))y(\ _n)$.

Proof. Note that

$$M^{\ 1} = x^{\mathsf{T}}V_y^{\ 1}x = \ ^2I_n + R(\ ) \quad \text{and} \quad m = x^{\mathsf{T}}V_y^{\ 1}y = \ ^2y: \tag{D.1}$$

By the Woodbury matrix identity,

$$\begin{aligned}m^{\mathsf{T}}M\ m &= \ ^2y^{\mathsf{T}}(I_n + \ ^2R^{\ 1}(\ ))^{\ 1}y \\ &= \ ^2y^{\mathsf{T}}y \quad y^{\mathsf{T}}(\ ^2I_n + R(\ ))^{\ 1}y = y^{\mathsf{T}}V_y^{\ 1}y \quad y^{\mathsf{T}}(\ ^2I_n + R(\ ))^{\ 1}y;\end{aligned}$$

which yields the desired result.

The next lemma studies the asymptotic behaviour of b in the case where the range decay $= \ _0$, that is is fixed at the true value.

LEMMA 3. *Let* $= \ _0$, *and assume that* $\max_{s2D}\min_{1\ \ i\ \ n} js\ \ s_ij \ \ n^{\ \frac{1}{d}}$. *Then*

$$\frac{b_{\ ;n} \quad b}{n} \ \ !\ \ \frac{\ ^2_0}{2\ ^2}; \quad \mathsf{P}_0\text{-almost surely:} \tag{D.2}$$

Proof. Let Q_n be the orthogonal matrix such that $Q_n R_0(\phi_n) Q_n^\top = \begin{pmatrix} \lambda_1^{(n)} & & \\ & \ddots & \\ & & \lambda_n^{(n)} \end{pmatrix}$, where $\lambda_i^{(n)}$ is the i-th largest eigenvalue of matrix $R_0(\phi_n)$. Thus, under P_0,

$$Q_n y(\phi_n) \sim N\left(0, \begin{pmatrix} \sigma_0^2 \lambda_1^{(n)} + \tau_0^2 & & \\ & \ddots & \\ & & \sigma_0^2 \lambda_n^{(n)} + \tau_0^2 \end{pmatrix}\right).$$

By Lemma 2, we get

$$2(b_{\cdot;n} - b) = \sum_{i=1}^{n} \frac{\sigma_0^2 \lambda_i^{(n)} + \tau_0^2}{\lambda_i^{(n)} + \tau^2} u_i^2, \tag{D.3}$$

where $u_i \overset{i.i.d.}{\sim} N(0,1)$ for $i = 1, \ldots, n$. By Tang et al. (2021, Corollary 2), there exists $C > 0$ independent of n such that $\lambda_i^{(n)} \geq Cni^{\frac{2}{d}-1}$ for all $1 \leq i \leq n$. This implies that

$$\sum_{i=1}^{n} \frac{\sigma_0^2 \lambda_i^{(n)} + \tau_0^2}{\lambda_i^{(n)} + \tau^2} \leq \frac{n\tau_0^2}{2} \quad \text{as } n \to \infty. \tag{D.4}$$

By the law of large numbers, (D.2) follows from (D.3) and (D.4).

Proof of Theorem 3. Let $\theta^{02} := \sigma_0^2 \tau_0^2 / \tau^2$, and let P^0 be the probability distribution of the Matérn model with parameters $(\theta^{02}, \phi, \tau_0^2)$. By Tang et al. (2021, Theorem 1), P^0 is equivalent to P. Further by Lemma 3,

$$\frac{b_{\cdot;n} - b}{n} \to \frac{\tau_0^2}{2\tau^2}, \quad P^0\text{-almost surely},$$

which also holds P_0-almost surely. Now by Lemma 1,

$$E_0(\sigma^2 | y(\phi_n)) = \frac{b_{\cdot;n}}{a_{\cdot;n}} \to \frac{\tau_0^2}{2} \quad \text{and} \quad V_0(\sigma^2 | y(\phi_n)) = \frac{b_{\cdot;n}^2}{(a_{\cdot;n}-1)^2(a_{\cdot;n}-2)} \to \frac{1}{n}, \tag{D.5}$$

which yields (2.17) by Chebyshev's inequality. This implies the posterior consistency of $\sigma^2 = \sigma^2 \tau^2$.

E. Proof of Theorem 4

Proof. By (D.1), we have

$$p(z | y, \sigma^2) = N((I_n + \sigma^2 R(\phi)^{-1})^{-1} y, \sigma^2 (\sigma^2 I_n + R(\phi)^{-1})^{-1}).$$

Combining with (C.4), we get the posterior predictive mean

$$\begin{aligned} E(z(s_0) | y_n) &= J_{\cdot;n}^\top R(\phi_n)^{-1} E(z | y_n) \\ &= J_{\cdot;n}^\top R(\phi_n)^{-1} (I_n + \sigma^2 R(\phi_n)^{-1})^{-1} y_n = J_{\cdot;n}^\top (\sigma^2 I_n + R(\phi_n))^{-1} y_n, \end{aligned}$$

which is the best linear predictor corresponding to a Matérn model with parameter values $\{\theta^{02}, \phi, \tau^{02}\}$ satisfying $\sigma^2 = \theta^{02}/\tau^{02}$. Further by Theorem 3, the formula (C.5) reduces to

$$V(z(s_0) | y_n) \to \frac{\tau_0^2}{2}(1 - J_{\cdot;n}^\top R(\phi_n)^{-1} J_{\cdot;n}) + \sigma_0^2 J_{\cdot;n}^\top (\sigma^2 I_n + R(\phi_n))^{-1} R(\phi_n)^{-1} J_{\cdot;n},$$

which gives (2.18). The rest of the theorem easily follows.

F. PROOF OF PROPOSITION 1

Proof. To simplify the notation, we denote by $:= \frac{1}{n}$ the inter-spacing of the grid. Let f_0 (resp. f) be the spectral density of the Matérn model with true parameter values $f \overset{2}{0}; \ _0; \ \overset{2}{0}g$ and the smoothness parameter $_0$ (resp. possibly misspecified parameter values $f \ ^2; \ ; \ ^2g$ the smoothness parameter). Define

$$\not{f}_0 \ (u) = \ ^1 \sum_{k= \ 1} f_0 \ \frac{u + 2 \ k}{} \quad \text{and} \quad \not{f} \ (u) = \ ^1 \sum_{k= \ 1} f \ \frac{u + 2 \ k}{} \ :$$

By (Stein, 1999, Chapter 3, (13)), the prediction error of $y(0)$ based on $y(s), s \ 2 \ _1 \ n \ f0g$ is

$$4 \ ^2 \int \frac{\not{f}_0 \ (u)}{\not{f} \ (u)^2} du \int \not{f} \ (u) \ ^1 du \ ^2;$$

and hence the prediction error of $z(0)$ based on $y(s), s \ 2 \ _1 \ n \ f0g$ is

$$e_1 \ = \frac{4 \ ^2 \int \frac{\widetilde{f}_0 \ (u)}{\widetilde{f} \ (u)^2} du}{\int \not{f} \ (u) \ ^1 du \ ^2} \ \overset{2}{0}:$$

Recall from (2.13) the spectral density of the Matérn model without nugget. We write

$$\not{f}_0 \ (u) = \ \overset{2}{0} g_0 \ (u) + \frac{\overset{2}{0}}{2} \quad \text{and} \quad \not{f} \ (u) = \ ^2g \ \ (u) + \frac{^2}{2};$$

so that

$$e_1 \ = \frac{4 \ ^2 \int \frac{\overset{2}{0} g_0 \ (u) + \frac{1}{2} \ \overset{2}{0}}{(\ ^2g \ \ (u) + \frac{1}{2} \ ^2)^2} du}{\int (\ ^2g \ \ (u) + \frac{1}{2} \ ^2) \ ^1 du \ ^2} \ \overset{2}{0} \tag{F.1}$$

Note that $g \ \ (u) \ \ cu \ ^2 \ ^1$ for some $c > 0$. It is known that (see e.g. (Tang et al., 2021, Section 2.3))

$$\int ^2g \ \ (u) + \frac{1}{2} \ ^2 \ ^1 du \ \ \frac{4 \ ^2}{^2 + C \ \frac{^2}{2 \ ^2 + 1}} \quad \text{for some } C > 0: \tag{F.2}$$

Furthermore, $g \ \ (u) \ \ ^2$ for u large and $g \ \ (u) \ \ ^1$ for u small. We prove that $e_1 \ ! \ 0$ as $n \ ! \ 1$. Further by Assumption 2, we get $e \ ! \ 0$ as $n \ ! \ 1$. Therefore,

$$\int \frac{\overset{2}{0} g_0 \ (u) + \frac{1}{2} \ \overset{2}{0}}{(\ ^2g \ \ (u) + \frac{1}{2} \ ^2)^2} du = \frac{4 \ ^2 \ \overset{2}{0}}{4} + O(\ ^{\min(2 \ _0; 2 \ ; 1)}): \tag{F.3}$$

Combining (F.1), (F.2) and (F.3) yields

$$e_1 \ = \ \overset{2}{0} + O(\ ^{\min(2 \ _0; 2 \ ; 1)}) \ \ 1 + \frac{C \ \frac{^2}{2 \ ^2 + 1}}{2} \ ! \ \overset{2}{0} \ \ ^{\min(2 \ _0; \frac{^2}{2 \ ^2 + 1})}:$$

Thus, we have $e_1 \ ! \ 0$ as $n \ ! \ 1$, and by Assumption 2 we get (2.20).

G. Proof of Proposition 2

Proof. For ease of presentation, we give the proof in the setting of Theorem 4. Note that

$$\mathrm{E}_0\left[y(s_0) - \sum_{g=1}^{G} w_g \mathrm{E}_g(y(s_0)\,|\,y)\right]^2 = \tau_0^2 + \mathrm{E}_0\left[\sum_{g=1}^{G} w_g\,(z(s_0) - \mathrm{E}_g(z(s_0)\,|\,y))\right]^2$$
$$\tau_0^2 + \sum_{g=1}^{G} w_g^2 \sum_{g=1}^{G} \mathrm{E}_0\,(z(s_0) - \mathrm{E}_g(z(s_0)\,|\,y))^2; \qquad \text{(G.1)}$$

where we apply the Cauchy-Schwarz inequality in (G.1). For each $1 \le g \le G$, the term $\mathrm{E}_0\,(z(s_0) - \mathrm{E}_g(z(s_0)\,|\,y))^2$ corresponds to the deviation error $E_{1;n}$ for the model M_g, which goes to 0 as $n \to \infty$. Since $|w_g|$ is bounded for each g, the bound (3.3) follows readily from (G.1).

H. Proof of Theorem 5

Proof. By Proposition 2, it suffices to prove that

$$\mathrm{E}_0\left(\frac{1}{n}\sum_{i=1}^{n} y(s_i) - \sum_{g=1}^{G} w_g \hat{y}_g(s_i)\right)^2 \to \tau_0^2 \quad \text{as } n \to \infty. \qquad \text{(H.1)}$$

For ease of presentation, we prove (H.1) in the setting of Theorem 4. Note that

$$\mathrm{E}_0\left(\frac{1}{n}\sum_{i=1}^{n} y(s_i) - \sum_{g=1}^{G} w_g \hat{y}_g(s_i)\right)^2 = \tau_0^2 + \frac{1}{n}\underbrace{\mathrm{E}_0\left[\sum_{i=1}^{n}\sum_{g=1}^{G} w_g\,(z(s_i) - \hat{y}_g(s_i))\right]^2}_{B}$$
$$\tau_0^2 + \frac{1}{n}\sum_{g=1}^{G} w_g^2 \sum_{i=1}^{n} E_{1;g;i}. \qquad \text{(H.2)}$$

By the boundedness of $|w_g|$ and the condition (3.4), the limit (H.1) follows easily from (H.2).
Taking a closer look at (H.2), we can see that B summarises the average squared prediction errors for $z(s_i); i = 1; \ldots; n$, in a LOO cross-validation. Hence, the stacking weights obtained from (3.2) minimises the average squared prediction errors for the latent process over the observed locations.

I. PSEUDO-CODES FOR STACKING ALGORITHMS

Algorithm 1. Stacking of means.

Data: $X, y, \ _n$: Design matrix, outcome and location set
$, V , a , b$: Prior parameters
$G \ ; G \ ; G \ _2$: Grids of $\ ; \ ; \ ^2$
Result: G_{all}: Grid spanned by $G \ ; G \ ; G \ _2$
w: Stacking weights
Compute $X^{(k)}_{\text{prod}} = X^\intercal[\ k]X[\ k]$, $X^{(k)}_y = X^\intercal[\ k]y[\ k]$ and record the number of observations n_k in fold k for $k = 1;:::;K$, where $X[\ k]\ y[\ k]$ denotes the predictors and response for observations not in fold k;
foreach $f \ ; \ g$ *in grid expanded by grid of* $\ $ *and* $\ $ **do**
  for $k = 1$ **to** K **do**
    Calculate $R_{(\ ^1 k; \ k)} = fR \ ^1(s; s^0; \ \ ; \ )g_{s;s^0 2 \ _n[\ k]}$ $O((n \ \ nk)^3)$;
    Store $R_{(k; \ k)} = fR(s; s^0; \ \ ; \ )g_{s2 \ _n[k];s^0 2 \ _n[\ k]}$ $O(n \ \ nk)$;
    for ^{2} *in the grid of* ^{2} **do**
      Compute the Cholesky decomposition L of
      $M \ ^1 = L \ L^\intercal = \begin{matrix} ^2X^{(k)}_{\text{prod}} + V \ ^1 & ^2X^\intercal[\ k] \\ ^2X[\ k] & R_{(\#\ ^1 k; \ k)} + \ ^2I_{n \ \ nk} \end{matrix}$ $O(n^3)$;
      Compute $u = m \ = \begin{matrix} V \ ^1 \ + \ ^2X^{(k)}_y \\ ^2y[\ k] \end{matrix}$ $O(n \ \ p)$;
      Update $u = L \ ^1 u$;
      Update $u = L \ ^\intercal u$, where $u = (\ ^\intercal; \ _z{}^\intercal)^\intercal$ $O(n^2)$;
      Compute the expected latent process on locations in fold k
      $z^{(k)}_{(\ ; \ ; \ ^2)} = R_{(k; \ k)} \ R_{(\ ^1 k; \ k)} \ \ _z$;
      Compute expected outcome on locations in fold k $y^{(k)}_{(\ ; \ ; \ ^2)} = X[k] \ \ + z^{(k)}_{(\ ; \ ; \ ^2)}$;
Compute Stacking weights w based on $fy^{(k)}_{(\ ; \ ; \ ^2)} g^{k=1;:::;K}_{(\ ; \ ; \ ^2) 2 G_{all}}$;

Algorithm 2. Stacking of predictive densities

Data: $X, y, \quad _n$: Design matrix, outcome and location set
$, V \ , a \ , b$: Prior parameters
$G \ ; G \ ; G \ _2$: Grids of $\ ; \ ; \ ^2$
J: number of samples for log point-wise predictive density estimation

Result: G_{all}: Grid spanned by $G \ ; G \ ; G \ _2$
w: Stacking weights

Compute $X^{(k)}_{\text{prod}} = X^\top[\ \ k]X[\ \ k], X^{(k)}_y = X^\top[\ \ k]y[\ \ k], ky[\ \ k]k^2 = y^\top[\ \ k]y[\ \ k]$ and record the number of observations n_k in fold k for $k = 1; : : : ; K$, where $X[\ \ k]\, y[\ \ k]$ denotes the predictors and response for observations not in fold k;

foreach $f \ ; \ g$ *in grid expanded by grid of* $\ \ and$ **do**

  for $k = 1$ **to** K **do**

    Calculate $R_{(\ ^1 k; \ k)} = fR \ ^1(s; s^0; \ \ ; \ )g_{s;s^0 2 \ _n[\ \ k]}$ $O((n \ \ nk)^3)$;

    Store $R_{(k; \ k)} = fR(s; s^0; \ \ ; \ )g_{s2S[k];s^0 2 \ _n[\ \ k]}$ $O(n \ \ nk)$;

    for ^{2} *in the grid of* ^{2} **do**

      Compute the Cholesky decomposition L of
      $M \ ^1 = L \ L^\top = {}^2X^{(k)}_{\text{prod}} + V \ ^1 \quad {}^2X^\top[\ \ k] \ \# \ " \ {}^2X[\ \ k] \quad R_{(\ ^1 k; \ k)} + \ {}^2I_{n \ \ nk}$ $O(n^3)$;

      Compute $u = m \ = V \ ^1 \ + \ {}^2X^{(k)}_y \ {}^2y[\ \ k]$ $O(n \ \ p)$;

      Update $u = L \ ^1 u$;

      Compute $b \ = b \ + 0{:}5(\ \ ^2ky[\ \ k]k^2 + \ ^\top V \ ^1 \quad u^\top u)$ and $a \ = a \ + 0{:}5(N \ \ n_k)$;

      Update $u = L \ ^\top u$, where $u = (\ \ ^\top; \ _z{}^\top)^\top$ $O(n^2)$;

      Generate the posterior expected outcome on locations in fold k $\hat{y}^k_{(\ ; \ ; \ ^2)} = X[k] \quad + R_{(k; \ k)} \ R_{(\ ^1 k; \ k)} \quad z$

      Compute $lpc = \ \ 0{:}5 \log(2 \ ) + \log \ (a \ + 1{=}2) \ \ \log \ (a \ ) + a \ \log b$;

      foreach $s \ 2 \ S[k]$ **do**

        Generate $h_s = \ x(s) \ R_{(k; \ k)}(s) \ \ R^{(inv)}$ where $R_{(k; \ k)}(s)$ is the row with elements $fR(s; s^0; \ \ ; \ )g_{s^0 2 \ _n[\ \ k]}$ in $R_{(k; \ k)}$

        Compute $V_s = kL \ ^1 h_s^\top k^2 + \ ^2$.

        Compute the log point-wise predictive density of $y(s)$ at location s, by
        $lp_{(\ ; \ ; \ ^2)}(s) =$
        $lpc \quad 0{:}5 \log(V_s) \quad (a \ + 1{=}2) \log fb \ + (y(s) \quad \hat{y}^k_{(\ ; \ ; \ ^2)}(s))^2{=}(2V_s)g$ (J.1);

Compute Stacking weights w based on $flp_{(\ ; \ ; \ ^2)}(s)g^{s2 \ \ n}_{(\ ; \ ; \ ^2)2G_{all}}$;

J. Derive the closed form of point-wise predictive density

In this subsection, we derive the posterior predictive density of the outcome $y(s_0)$ on location s_0. We follow the notations in Section 2. First, we know that $y(s_0)j \ \ ; \ ^2; y;$ follows a Gaussian with mean $x(s_0) \ + U(\ )R \ ^1(\ _n)z$ and variance $\ ^2 \ ^2$. Since the conditional posterior distribution $j \ ^2; y;$ follows N$(M \ m \ ; \ ^2M \ )$, the conditional posterior distribution $y(s_0)j \ ^2; y;$ still follows a Gaussian N$(\ _{s_0}; \ ^2V_{s_0})$ where

$$_{s_0} = \underbrace{x(s_0) \ U(\ )R \ ^1(\ _n)}_{h_{s_0}} M \ m \ ; V_{s_0} = h_{s_0} M \ h_{s_0}^\top + \ ^2 :$$

Next, through equation (2.4)

$$\begin{aligned}
p(y(s_0)\,|\,y;\ ) &= \int p(y(s_0)\,|\ ^2;y;\ )p(\ ^2\,|\ ;y)d\ ^2 \\
&= \int \mathrm{N}(y(s_0)\,|\ _{s_0};\ ^2V_{s_0})\mathrm{IG}(\ ^2\,|\,a\ ;b\ )d\ ^2 \\
&= \int \frac{1}{(2\ V_{s_0}\ ^2)^{1=2}}\exp\ \frac{(y(s_0)\ \ _{s_0})^2}{2\ ^2V_{s_0}}\ \ \frac{b^a}{(a\ )}\ ^{2(\ a\ \ 1)}\exp\ \ \frac{b}{2}\ \ d\ ^2 \\
&= \frac{b^a}{(2\ V_{s_0})^{1=2}\ (a\ )}\int\ ^{2(\ a\ \ 1=2\ \ 1)}\exp\ \ \frac{1}{2}\ \ b\ +\ \frac{(y(s_0)\ \ _{s_0})^2}{2V_{s_0}}\ \ \ d\ ^2 \\
&= \frac{(a\ +1=2)b^a}{(2\ V_{s_0})^{1=2}\ (a\ )}\ \ b\ +\ \frac{(y(s_0)\ \ _{s_0})^2}{2V_{s_0}}\ \ ^{(a\ +1=2)}
\end{aligned}$$

The log point-wise predictive density is

$$\begin{aligned}
lp(y(s_0)\,|\,y;\ ) = \ & 0{:}5\log(2\ V_{s_0}) + a\ \ \log b\ \ \ (a\ +1=2)\log\ \ b\ +\ \frac{(y(s_0)\ \ _{s_0})^2}{2V_{s_0}} \\
& +\log\ \ (a\ +1=2)\ \ \ \log\ \ (a\ )
\end{aligned} \tag{J.1}$$

K. STACKING OF PREDICTIVE DENSITIES ALGORITHM (MONTE CARLO VERSION)

We present a Monte Carlo algorithm to estimate the log of point-wise predictive density for outcome in fold k given observations not in fold k. For each k, we generate J posterior samples of ^{2} and $(\ ^\top;z^\top)^\top = $, (i.e., $f\ ^{2(j)};\ ^{(j)}g$ for $j = 1;:::;J$), using data not in fold k. Then we calculate the corresponding expected outcome for location s in fold k, $\hat{y}^{k;j}_{(\ ;\ ;\ ^2)}(s)$ for $j = 1;:::;J$. Next, we compute the predictive density of $y(s)$ conditional on the prediction $\hat{y}^{k;j}_{(\ ;\ ;\ ^2)}(s)$ and the nugget (variance of the noise process, which equals the product of ^{2} and the j-th posterior sample $^{2(j)}$) for each j. The conditional predictive distribution of $y(s)$ follows

$$p(y(s)\,|\ ^{2(j)};\ ^{(j)}) = \mathrm{N}(y(s)\,|\,y^{k;j}_{(\ ;\ ;\ ^2)};\ ^2\ ^{2(j)})\ : \tag{K.1}$$

Finally, the log point-wise predictive density (LPD) of $y(s)$ at location s is estimated by

$$\begin{aligned}
lp_{(\ ;\ ;\ ^2)}(s) = \log & \int p(y(s)\,|\ ^2;\ )p(\ ^2;\ \ |y[\ \ k])d\ ^2d \\
& \log\left\{\frac{1}{J}\sum_{j=1}^{J} p_{(\ ;\ ;\ ^2)}(y(s)\,|\ ^{2(j)};\ ^{(j)})\right\};
\end{aligned} \tag{K.2}$$

and we can compute the stacking weights based on the estimated LPDs. The following is the Monte Carlo version of the stacking of predictive densities algorithm.

Algorithm 3. Stacking of predictive densities (Monte Carlo Version)

Data: $X, y, \quad_n$: Design matrix, outcome and location set
$, V , a , b$: Prior parameters
$G ; G ; G_2$: Grids of $; ; ^2$
J: number of samples for log point-wise predictive density estimation

Result: G_{all}: Grid spanned by $G ; G ; G_2$
w: Stacking weights

Compute $X^{(k)}_{\text{prod}} = X^\top[\ k]X[\ k], X^{(k)}_y = X^\top[\ k]y[\ k], ky[\ k]k^2 = y^\top[\ k]y[\ k]$ and record the number of observations n_k in fold k for $k = 1;:::;K$, where $X[\ k]\ y[\ k]$ denotes the predictors and response for observations not in fold k;

foreach $f \ ; \ g$ *in grid expanded by grid of* $\quad$ *and* $\quad$ **do**

  for $k = 1$ **to** K **do**

    Calculate $R_{(\ ^1{}_{k;\ k)}} = fR\ ^1(s; s^0; \ ; \ )g_{s;s^0 2\ _n[\ k]} \qquad O((n \quad nk)^3)$;

    Store $R_{(k;\ k)} = fR(s; s^0; \ ; \ )g_{s2S[k];s^0 2\ _n[\ k]} \qquad O(n \ \ nk)$;

    for ^{2} *in the grid of* ^{2} **do**

      Compute the Cholesky decomposition L of

$$M\ ^1 = L\ L^\top = \begin{matrix} {}^2X^{(k)}_{\text{prod}} + V\ ^1 & {}^2X^\top[\ k] \\ {}^2X[\ k] & R_{(\ ^1{}_{k;\ k)}} + {}^2I_{n\ \ nk} \end{matrix} \qquad O(n^3)\ ;$$

      Compute $u = m = \begin{matrix} V\ ^1 + {}^2X^{(k)}_y \\ {}^2y[\ k] \end{matrix} \qquad O(n \ \ p)$;

      Update $u = L\ ^1u$;

      Compute $b = b + 0{:}5(\ {}^2ky[\ k]k^2 + \ ^\top V\ ^1 \quad u^\top u)$ and $a = a + 0{:}5(N \quad n_k)$;

      Generate $^{2(1)};:::;\ ^{2(J)} \quad$ Inverse-Gamma$(a\ ; b\ )$;

      Generate $^{(j)} \quad \mathrm{N}(M\ m\ ;\ ^{2(j)}M\ )$ by taking $v^{(j)} \quad \mathrm{N}(0;\ ^{2(j)}I_{n\ \ nk+p})$ and computing $^{(j)} = L\ ^\top(v^{(j)} + u)$ for $j = 1;:::;J$;

      Generate the posterior samples of the expected outcome on locations in fold k $\hat{y}^{(k;j)}_{(\ ;\ ;\ ^2)} = X[k]\ ^{(j)} + R_{(k;\ k)}\ R_{(\ ^1{}_{k;\ k)}}\ ^{(j)}_z, j = 1;:::;J$;

      foreach $s \ 2 \ S[k]$ **do**

        Compute the posterior samples of the log-density of observation $y(s)$ at location $s, p_{(\ ;\ ;\ ^2)}(y(s) \mid \ ^{2(j)};\ ^{(j)})$ by taking the density of $\mathrm{N}(\hat{y}^{(k;j)}_{(\ ;\ ;\ ^2)}(s);\ ^2\ ^{2(j)})$ at $y(s)$ for $j = 1;:::;J$ (K.1);

        Compute the expected log point-wise predictive density of $y(s)$ at location s, by $lp_{(\ ;\ ;\ ^2)}(s) = \log\ \frac{1}{J}\ \sum_{j=1}^{J} p_{(\ ;\ ;\ ^2)}(y(s) \mid \ ^{2(j)};\ ^{(j)})$ (K.2);

Compute Stacking weights w based on $f lp_{(\ ;\ ;\ ^2)}(s) g^{s2}_{(\ ;\ ;\ ^2)2G_{all}}$;

L. RECOVER EXPECTED z AND LOG POINT-WISE PREDICTIVE DENSITY FOR MCMC SAMPLING

The package *spBayes* doesn't record the posterior samples of the latent process $z(s)$. In this subsection, we illustrate how to recover the expected $z(s)$ for both observed and unobserved location and compute MLPD for the simulation studies based on the outputs returned by *spLM*. To achieve our goal, we need to recover the posterior samples of $z(s)$ on all locations given the recorded MCMC samples of parameters $; ; ^2; ^2$ and $$. Denote $z(s)$ on observed and unobserved locations as z_o and z_u, respectively, and

denote $z(s)$ on all locations as z . Based on (2.2),

$$
\begin{aligned}
p(z_o, z_u \mid y, \beta, \sigma^2, \tau^2, \phi) &\propto N(y \mid X\beta + I_n \otimes z_o, \tau^2 I) \; N(z_o, z_u \mid 0, \sigma^2 R(\phi)) \\
&\propto \exp\left\{ -\frac{1}{2} \left[\sigma^{-2} (I_n \otimes z_o - (y - X\beta))^\top (I_n \otimes z_o - (y - X\beta)) + z^\top \sigma^{-2} R^{-1}(\phi) z \right] \right\} \\
&\propto \exp\left\{ -\frac{1}{2} \left[z^\top \sigma^{-2} R^{-1}(\phi) + \begin{pmatrix} \sigma^{-2} I_n & 0 \\ 0 & 0 \end{pmatrix} \right] z - z^\top \begin{pmatrix} (y - X\beta)/\sigma^2 \\ 0 \end{pmatrix} + (y - X\beta)^\top / \sigma^2 \otimes z \right\} \\
&\propto N(z \mid M_z^{-1} m_z, M_z^{-1}),
\end{aligned}
$$

where combines the observed and unobserved location sets and

$$
M_z = \sigma^{-2} R^{-1}(\phi) + \begin{pmatrix} \sigma^{-2} I_n & 0 \\ 0 & 0 \end{pmatrix}, \quad m_z = \begin{pmatrix} (y - X\beta)/\sigma^2 \\ 0 \end{pmatrix}.
$$

Let $\{\beta^{(j)}, \sigma^{2(j)}, \tau^{2(j)}, \phi^{(j)}, z^{(j)}\}$ for $j = 1, \ldots, J$ denote the recorded MCMC samples. We generate posterior samples for z using the above full conditional posterior distribution for each iteration j and then compute the average as the expected z . We further compute the LPD of $y(s)$ any held out location s by

$$
\begin{aligned}
lp(s) &= \log \int p(y(s) \mid \beta, z(s), \sigma^2) p(\beta, z(s), \sigma^2 \mid y) \, d\beta \, dz(s) d\sigma^2 \\
&= \log \left\{ \frac{1}{J} \sum_{j=1}^{J} p(y(s) \mid \beta^{(j)}, z(s)^{(j)}, \sigma^{2(j)}) \right\} \\
&= \log \left\{ \frac{1}{J} \sum_{j=1}^{J} N(y(s) \mid x(s)^\top \beta^{(j)} + z(s)^{(j)}, \sigma^{2(j)}) \right\}
\end{aligned}
$$

M. FIGURES FOR SIMULATION STUDIES

M.1. Interpolated maps for the simulation studies

M.2. Distributions of the estimated and τ^2

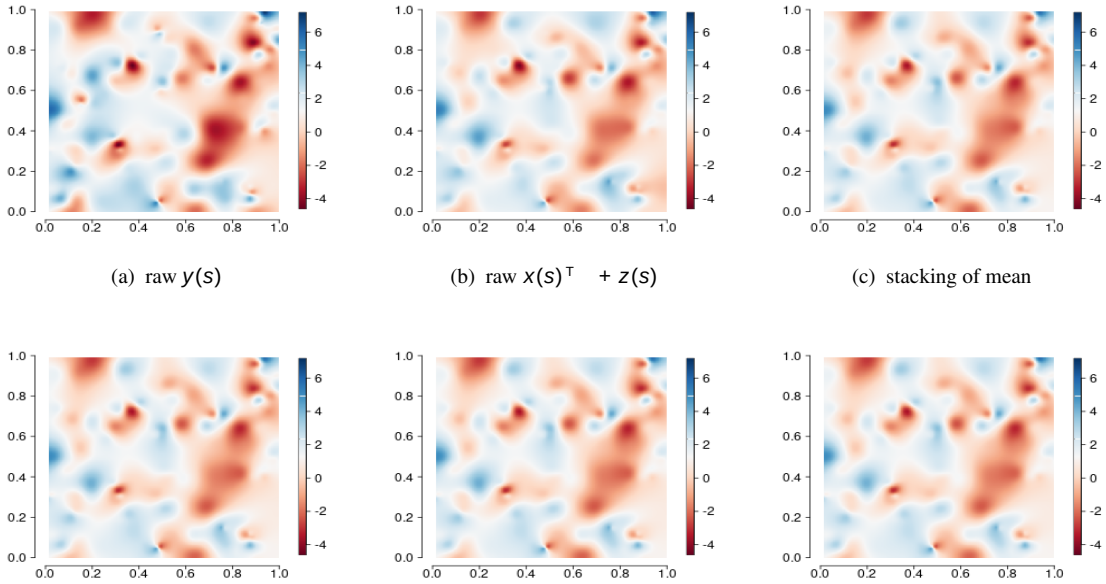
N. COMPUTE THE STACKING WEIGHTS FOR STACKING OF MEANS (IN R CODE)

Let us format the expected outcome $\{y^{(k)}_{(\beta, z, \sigma^2)}\}^{k=1,\ldots,K}_{(\beta, z, \sigma^2) \in G_{all}}$ computed in Algorithm 1 by an $N \times G$ matrix $\hat{Y}$. Each column of $\hat{Y}$ stores the expected outcome $\{y^{(k)}_{(\beta, z, \sigma^2)}\}^{k=1,\ldots,K}$ for each candidate model, and it shares the same order of observed locations as the outcome y. Let $w = (w_1, w_2, \ldots, w_G)^\top$ be the stacking weights, we need to find the weights that satisfy

$$
\operatorname*{argmin}_w \{(y - \hat{Y}w)^\top (y - \hat{Y}w)\},
$$

under the constrain $\sum_g^G w_g = 1$. Now we modify it into a quadratic programming (QP) problem.

$$
\begin{aligned}
&(y - \hat{Y}w)^\top (y - \hat{Y}w) \\
&= f(y - \sum_g^G w_g \hat{Y}_G) - \sum_g^{G-1} w_g (\hat{Y}_g - \hat{Y}_G) g^\top f(y - \sum_g^G w_g \hat{Y}_G) - \sum_g^{G-1} w_g (\hat{Y}_g - \hat{Y}_G) g \\
&= (\tilde{y} - \tilde{Y}w)^\top (\tilde{y} - \tilde{Y}w)
\end{aligned}
$$

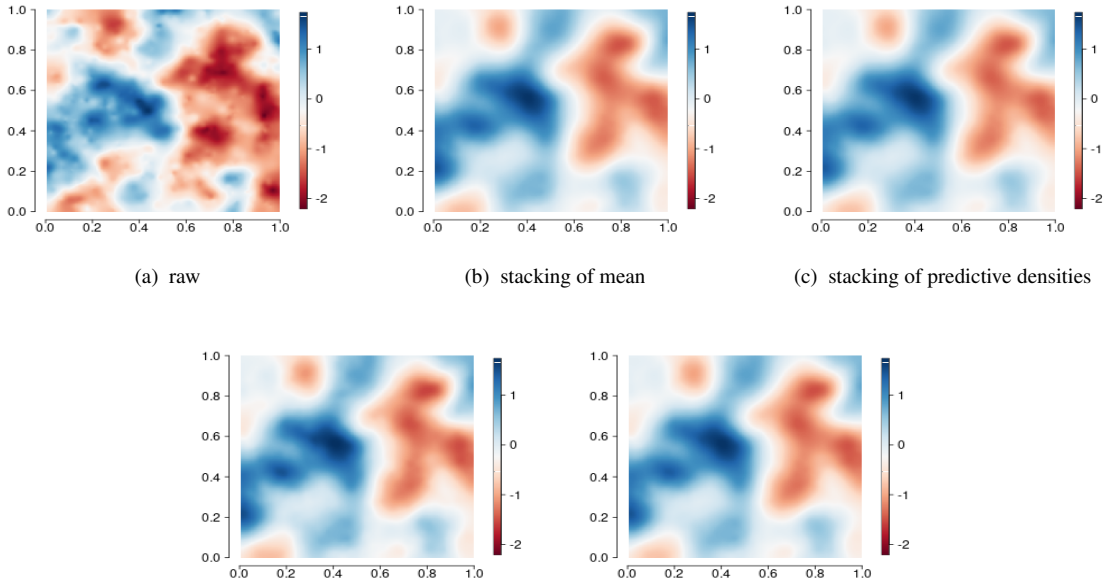


(a) raw $y(s)$ (b) raw $x(s)^\top \beta + z(s)$ (c) stacking of mean

(d) stacking of predictive densities (e) M0 (f) MCMC

Fig. 7: Interpolated maps of (a) the response $y(s)$, (b) the denoised response $x(s)^\top \beta + z(s)$ and (c-f) the expected $y(s)$ on the $n_h = 100$ held out locations generated by all competing algorithms for one data set in the first simulation.

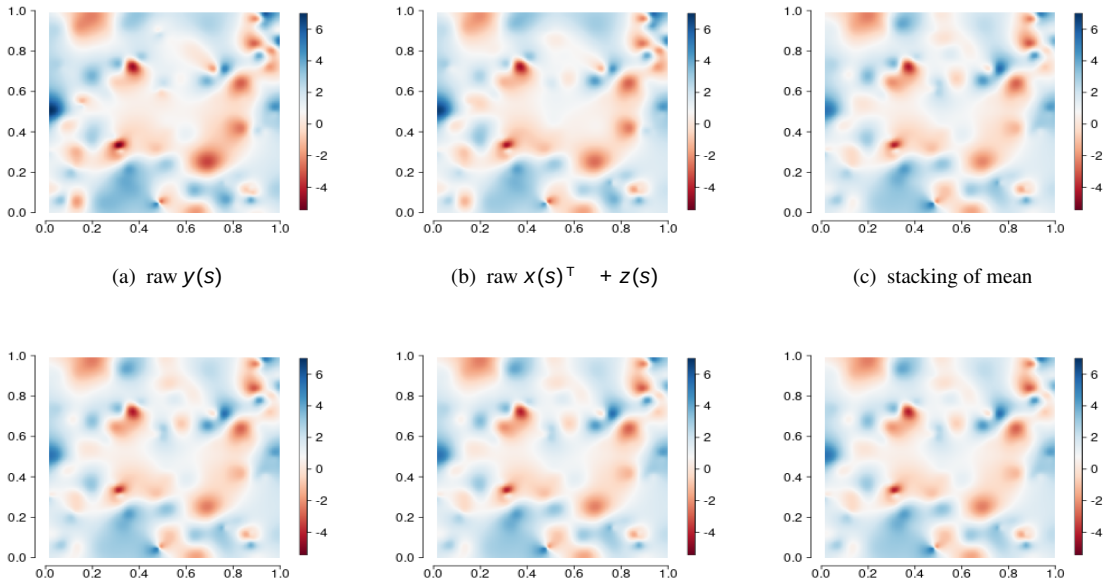
where $\tilde{y} = y - \hat{Y}_G$, $\tilde{Y} = [(\hat{Y}_1 - \hat{Y}_G) : \cdots : (\hat{Y}_{G-1} - \hat{Y}_G)]$, and $w = (w_1, \ldots, w_{G-1})^\top$. And the QP problem has constrains $\sum_{g=1}^{G-1} w_g \leq 1$ and $w_g \geq 0$ for $g = 1, \ldots, G-1$.



(a) raw (b) stacking of mean (c) stacking of predictive densities

(d) M0 (e) MCMC

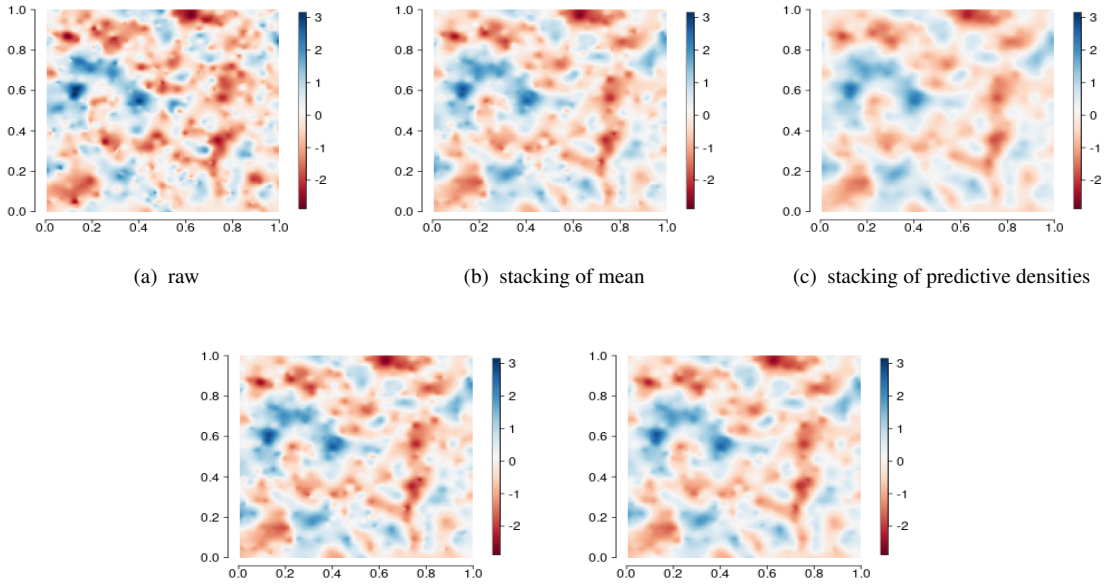
Fig. 8: Interpolated maps of (a) the latent process $z(s)$ and (b-g) the expected $z(s)$ on all $n = 900$ sampled locations generated by all competing algorithms for one data set in the first simulation. The $n = 900$ locations include both observed and unobserved locations



(a) raw $y(s)$ (b) raw $x(s)^\top \quad + z(s)$ (c) stacking of mean

(d) stacking of predictive densities (e) M0 (f) MCMC

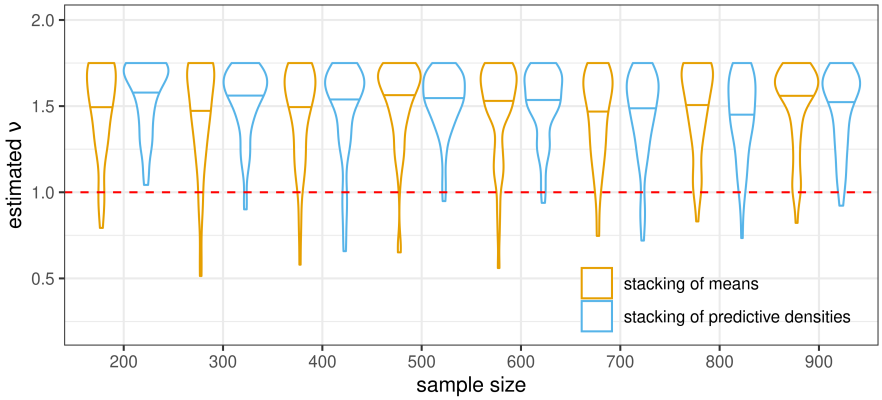
Fig. 9: Interpolated maps of (a) the response $y(s)$, (b) the denoised response $x(s) \quad + z(s)$ and (c-f) the expected $y(s)$ on the $n_h = 100$ held out locations generated by all competing algorithms for one data set in the second simulation.



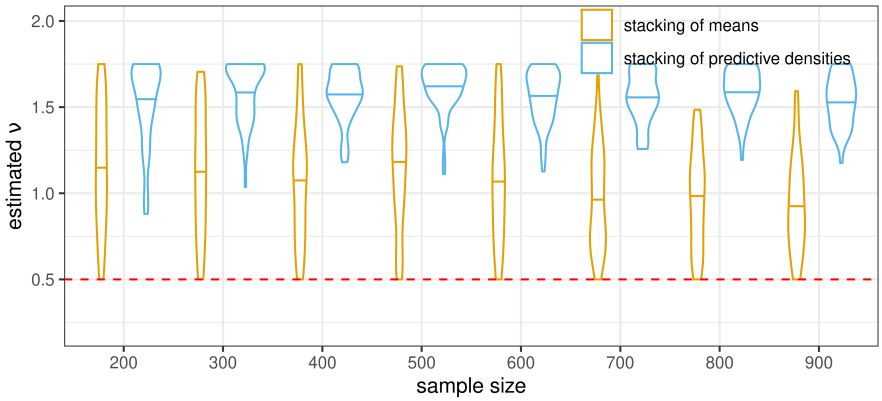
(a) raw (b) stacking of mean (c) stacking of predictive densities

(d) M0 (e) MCMC

Fig. 10: Interpolated maps of (a) the latent process $z(s)$ and (b-g) the expected $z(s)$ on all $n = 900$ sampled locations generated by all competing algorithms for one data set in the second simulation. The $n = 900$ locations include both observed and unobserved locations



(a)



(b)

Fig. 11: Distributions of the estimated in the first (a) and the second (b) simulation. The distribution of the counts are described through violin plots whose horizontal lines indicate the medians. The red dashed horizontal line indicates the actual value of