Vecchia-Laplace approximations of generalized Gaussian processes for big non-Gaussian spatial data

Paper by Daniel Zilber and Matthias Katzfuss

April 21 2020

Presentation by Peter Gao

Overview

Generalized Gaussian processes

First, a tangent:

Does the term Gaussian process encompass Gaussian random field?

Yes: Zilber and Katzfuss; Gelfand (2016), Sampson and Guttorp (1992), Fuentes (2005), Stein (1999)…

No: Wikipedia, Rozanov (1982), Lindgren (2012)…

Generalized Gaussian processes

How should we model dependent non-Gaussian data?

Spatial generalized linear mixed models or generalized GPs:

- Latent Gaussian process
- non-Gaussian likelihood from exponential family

Generalized Gaussian processes

In practice:

Working with GGPs may be expensive (cost grows cubically with data size), so use methods like:

- MCMC
- Expectation propagation
- Variational methods
- Laplace approximations

Gaussian process approximations

Ways to decrease computational cost:

- Low rank approximations
- Enforcing sparsity in covariance/precision matrices
- Vecchia approximations

Generalized Gaussian process approximations

Extend to non-Gaussian data by combining

- Low-rank GP
- Approximation of non-Gaussian likelihood

For example, INLA-SPDE approach:

- Sparse-precision approximation of a GP with Matérn covariance
- Laplace approximation for marginal posteriors conditioning on non-Gaussian observations

…. cost still $\mathcal{O}(n^{3/2})$ or even $\mathcal{O}(n^2)$ in higher dimensions!

Vecchia-Laplace approximations

Can we achieve an approximation with linearly scaling cost?

The authors propose to combine

- Vecchia approximation for latent GP
- Laplace approximation for non-Gaussian likelihood

Review

Generalized Gaussian processes

$$z_i \mid \mathbf{y} \sim_{ind.} g_i(z_i \mid y_i)$$
$$y(\cdot) \sim GP(\mu, K)$$

$\mathbf{z}$ conditionally independent observations at locations in $\mathscr{D} \subset \mathbb{R}^d$

$\mathbf{y}$ latent GP

Generalized Gaussian processes

$$p(\mathbf{y} \mid \mathbf{z}) = \frac{\mathcal{N}_n(\mathbf{y} \mid \boldsymbol{\mu}, \mathbf{K}) \prod_{i=1}^{n} g_i(z_i \mid y_i)}{p(\mathbf{z})}$$

Want to estimate the posterior of $\mathbf{y}$…

Laplace approximation

$$p(\mathbf{y} \mid \mathbf{z}) = \frac{\mathcal{N}_n(\mathbf{y} \mid \boldsymbol{\mu}, \mathbf{K}) \prod_{i=1}^{n} g_i(z_i \mid y_i)}{p(\mathbf{z})}$$

May be difficult to deal with $p(\mathbf{z})$… estimate by assuming posterior is Gaussian with mean and precision equal to the mode and negative Hessian at the mode of $\log p(\mathbf{y} \mid \mathbf{z})$

… requires optimization to find mode of $\log p(\mathbf{y} \mid \mathbf{z})$!

Laplace approximation

Newton-Raphson optimization turns out to be equivalent to computing posterior mean of $\mathbf{y}$ with Gaussian pseudo-data.

That is, although our data is non-Gaussian, we can find the true posterior mean by equating our observations with Gaussian pseudo-data.

!

Pseudo-data

distribution	likelihood $g(z\|y)$	pseudo-data t_y	pseudo-variance $d(y)$
Gaussian	$\mathcal{N}(y, \tau^2)$	z	τ^2
Bernoulli	$\mathcal{B}(\text{logit}(y))$	$y + \frac{(1+e^y)^2}{e^y}(z - \frac{e^y}{1+e^y})$	$(1+e^{-y})(1+e^y)$
Poisson	$\mathcal{P}(e^y)$	$y + e^{-y}(z - e^y)$	e^{-y}
Gamma	$\mathcal{G}(a, ae^{-y})$	$y + (1 - z^{-1}e^y)$	aze^{-y}

Table 1: Examples of popular likelihoods, together with the Gaussian pseudo-data and pseudo-variances implied by the Laplace approximation. The non-canonical logarithmic link function is used for the Gamma likelihood to ensure that the second parameter, ae^{-y}, is positive.

Pseudo-data

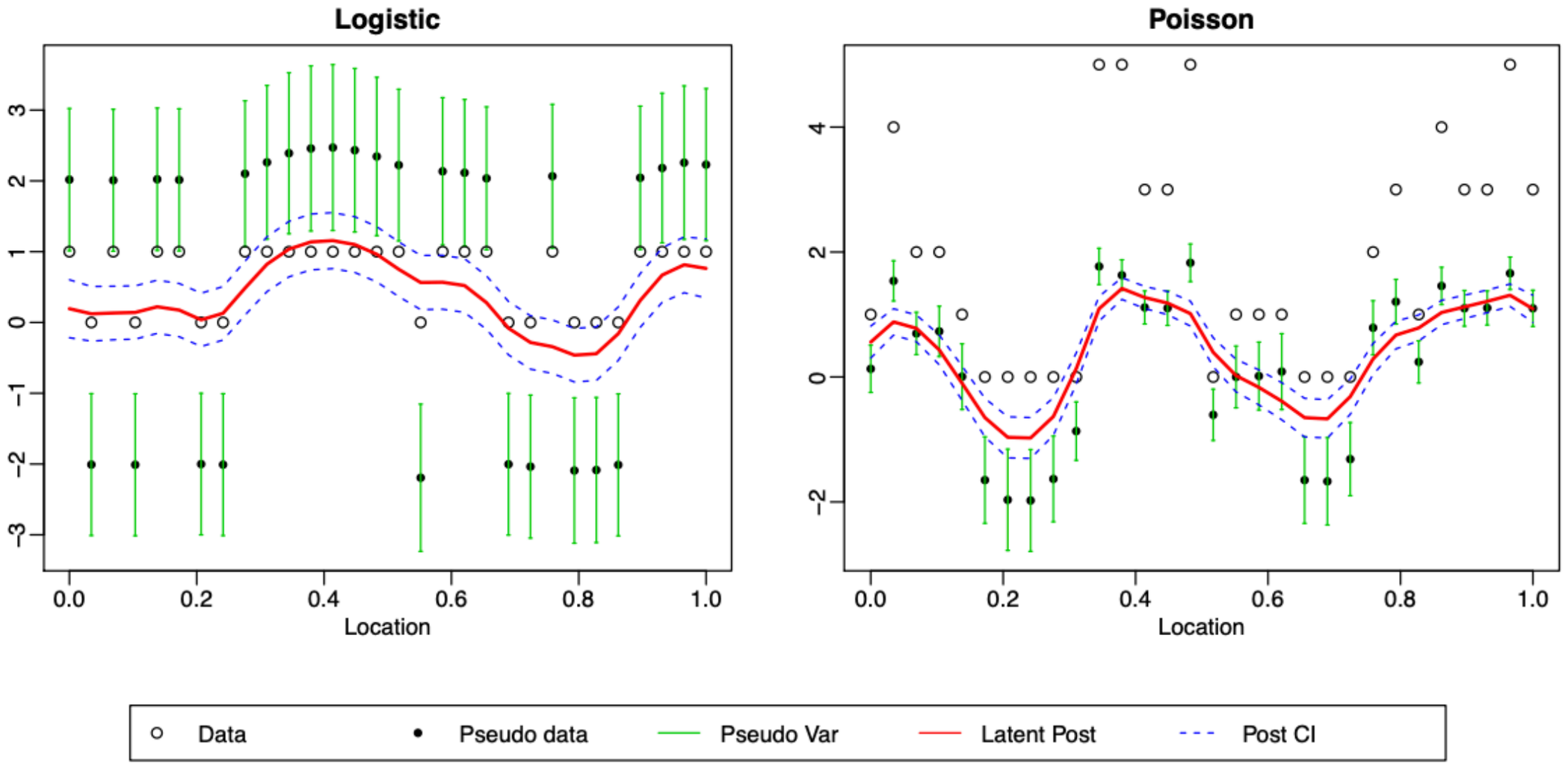


Figure 1: Pseudo-data $\mathbf{t}_{\boldsymbol{\alpha}}$ plus or minus half the standard deviation of the pseudo-noise for simulated data $\mathbf{z}$ in one spatial dimension, along with the latent posterior mode $\boldsymbol{\alpha}$ plus or minus half the posterior standard deviation. Note that the data are on a different scale than the pseudo-data due to the link function.

Vecchia approximation

$\mathbf{y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{K})$ vector of GP realizations

$\mathbf{t} \mid \mathbf{y} \sim \mathcal{N}_n(\mathbf{y}, \mathbf{D})$ pseudo data (diagonal covariance)

Then, let $\mathbf{x} = \mathbf{y} \cup \mathbf{t}$ and apply the approximation

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i \mid \mathbf{x}_{i:i-1}) \approx \prod_{i=1}^{2n} p(x_i \mid \mathbf{x}_{c(i)})$$

for some conditioning set $c(i)$. Still need to choose this set wisely!

Conditioning sets

Interweaved (IW) ordering:

$$\mathbf{x} = (y_1, t_1, \ldots, y_n, t_n)^T$$

$$\widehat{p}_{IW}(\mathbf{x}) = \prod_{i=1}^{n} p(t_i \mid y_i) p(y_i \mid \mathbf{y}_{q_y(i)}, \mathbf{t}_{q_t(i)})$$

Conditioning sets

If $x_j = t_i$, we only condition on y_i, because $\mathbf{D}$ is diagonal and therefore t_i is conditionally independent of all other variables in $\mathbf{y}$ and $\mathbf{t}$ given y_i. If $x_j = y_i$, we condition on $\mathbf{y}_{q_y(i)}$ and $\mathbf{t}_{q_t(i)}$, where $q(i) = q_y(i) \cup q_t(i)$ is the conditioning index vector consisting of the indices of the nearest m locations previous to i in the ordering. For splitting $q(i)$ into $q_y(i)$ and $q_t(i)$, we attempt to maximize $q_y(i)$ while ensuring linear complexity (Katzfuss and Guinness, 2019). Specifically, for $i = 1, \ldots, n$, we set $q_y(i) = (k_i) \cup (q_y(k_i) \cap q(i))$, where $k_i \in q(i)$ is the index whose latent-conditioning set has the most overlap with $q(i)$: $k_i = \arg\max_{j \in q(i)} |q_y(j) \cap q(i)|$, choosing the closest k_i in space to $\mathbf{s}_i$ in case of a tie. In one-dimensional space with coordinate ordering, this results in $q_y(i) = q(i) = (\max(1, i-m), \ldots, i-1)$ and $q_z(i) = \emptyset$. In higher-dimensional space, we may not be able to condition entirely on $\mathbf{y}$, so the remaining conditioning indices are assigned to $q_t(i) = q(i) \setminus q_y(i)$. These conditioning rules guarantee that $\mathbf{U}$ and $\mathbf{V}$ are both highly sparse with at most m nonzero off-diagonal elements per column. Katzfuss and Guinness (2019) showed that these matrices, and the resulting posterior mean and precision matrix, can be obtained in $\mathcal{O}(nm^3)$ time.

Conditioning sets

Interweaved (IW) ordering:

$$\mathbf{x} = (y_1, t_1, \ldots, y_n, t_n)^T$$

$$\widehat{p}_{IW}(\mathbf{x}) = \prod_{i=1}^{n} p(t_i \mid y_i) p(y_i \mid \mathbf{y}_{q_y(i)}, \mathbf{t}_{q_t(i)})$$

Response First (RF) ordering

$$\mathbf{x} = (t_1, \ldots, t_n, y_1, \ldots, y_n)^T$$

$$\widehat{p}_{RF}(\mathbf{x}) = \prod_{i=1}^{n} p(t_i) p(y_i \mid \mathbf{y}_{q_y(i)}, \mathbf{t}_{q_t(i)})$$

Vecchia-Laplace approximation

$\mathbf{y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{K})$ vector of GP realizations

$\mathbf{t} \mid \mathbf{y} \sim \mathcal{N}_n(\mathbf{y}, \mathbf{D})$ pseudo data based on observations $\mathbf{z}$

Apply Vecchia approximation to joint distribution of $\mathbf{x} = \mathbf{y} \cup \mathbf{t}$ and then compute posterior mean of $\mathbf{y}$ given $\mathbf{t}$… yielding mode of posterior of $\mathbf{y}$

Next, apply Laplace approximation

Parameter inference

Given unknown model parameters $\boldsymbol{\theta}$, need to perform inference based on integrated likelihood:

$$\mathscr{L}(\boldsymbol{\theta}) = p(\mathbf{z} \mid \boldsymbol{\theta}) = \int p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}) p(\mathbf{y} \mid \boldsymbol{\theta}) d\mathbf{y}$$

Again, use Laplace approximation of integrated likelihood:

$$\mathscr{L}(\boldsymbol{\theta}) \approx p(\mathbf{t}) \times \frac{p(\mathbf{z} \mid \mathbf{y})}{p(\mathbf{t} \mid \mathbf{y})}$$

evaluated at the posterior mode!

Parameter Estimation

Approximation of integrated likelihood is equivalent to approximation of posterior $p(\boldsymbol{\theta} \mid \mathbf{z})$ with flat priors.

Still essentially an ML-based approach

Prediction

Predictions at unobserved locations can be made by simply appending the corresponding random variables to get $\tilde{x} = \mathbf{t} \cup \mathbf{y} \cup \mathbf{y}^*$

Approximation properties

How can we assess the quality of the approximation?

Two sources of errors: Vecchia and Laplace approximations.

Authors use simulation to show accuracy.

Questions:

- Does the choice of likelihood affect quality of Laplace approximation?
- How to choose size of conditioning set for Vecchia approximation?

Simulations

Simulate mean zero GP with Matérn covariance on a unit square grid, and then conditionally generate observation data.

Compare with Laplace approximation w/o Vecchia and Hamiltonian Monte Carlo

RRMSE

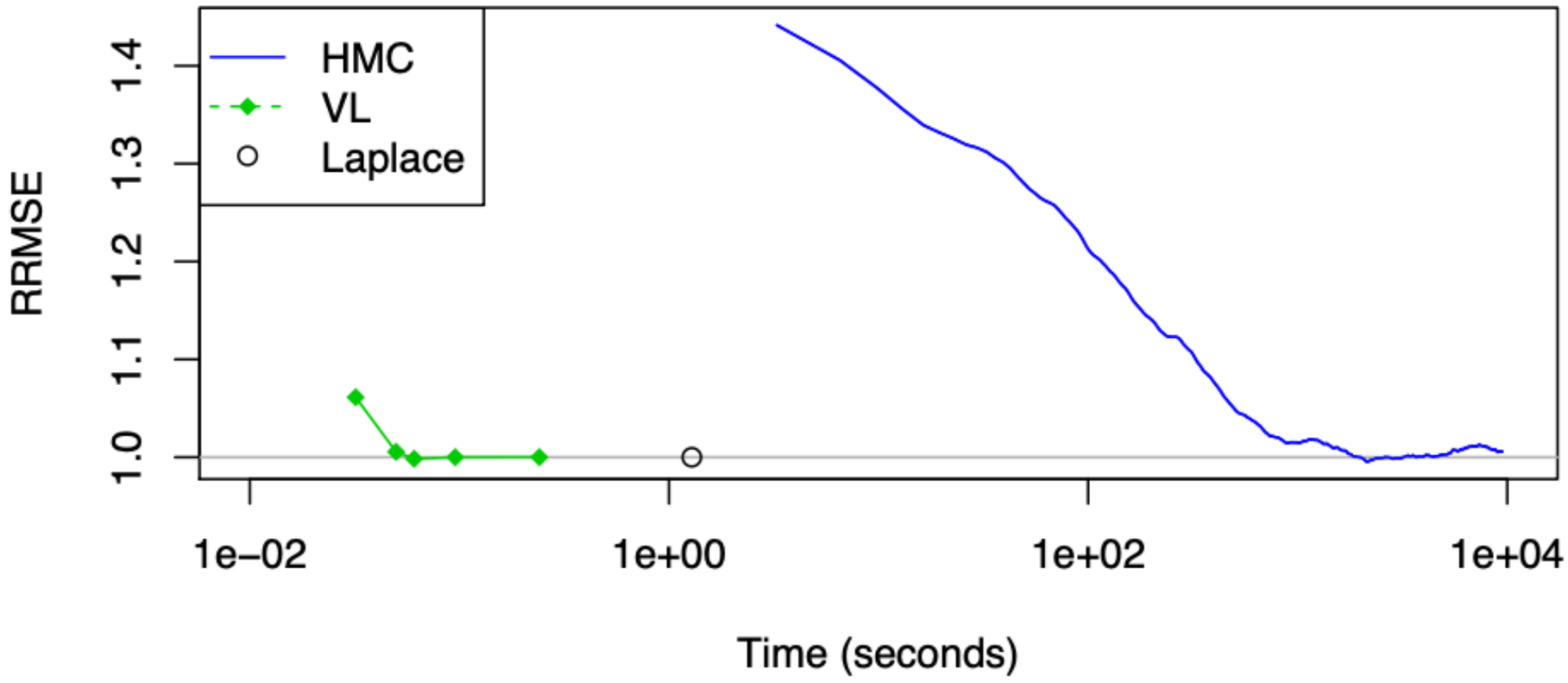


Figure 2: RRMSE versus time (on a log scale) for Bernoulli data of size $n = 625$. Laplace is run once until convergence. For VL, we considered $m \in \{1, 5, 10, 20, 40\}$. The number of HMC iterations varies from 5,100 to 300,000 in increments of 100, with the first 5,000 considered burn-in.

Comparing accuracy

VL-IW:

Vecchia-Laplace approximation based on interweaved ordering for finding conditioning sets

LowRank

Modified predictive process approximation (equivalent to Vecchia-Laplace except conditioning based on maxmin ordered latent variables)

Laplace:

Laplace approximation without Vecchia (computationally expensive)

Run time

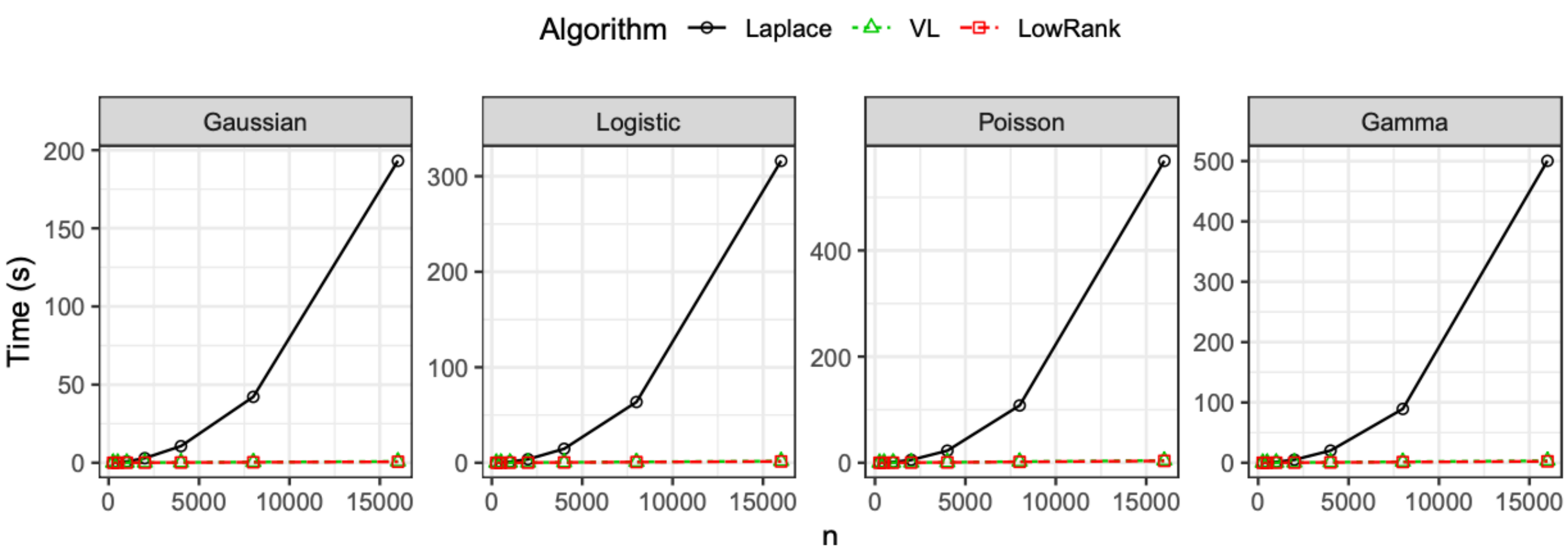
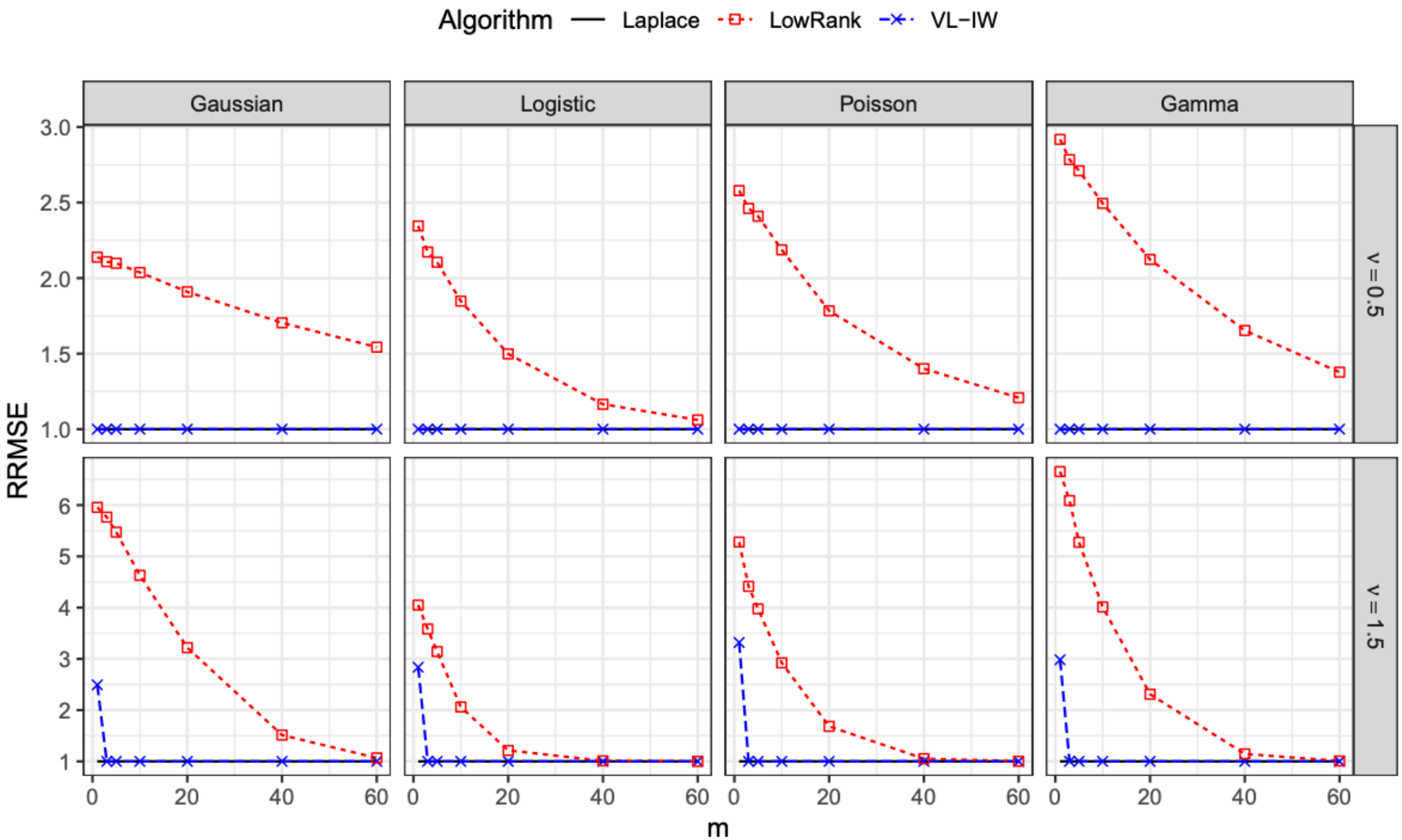


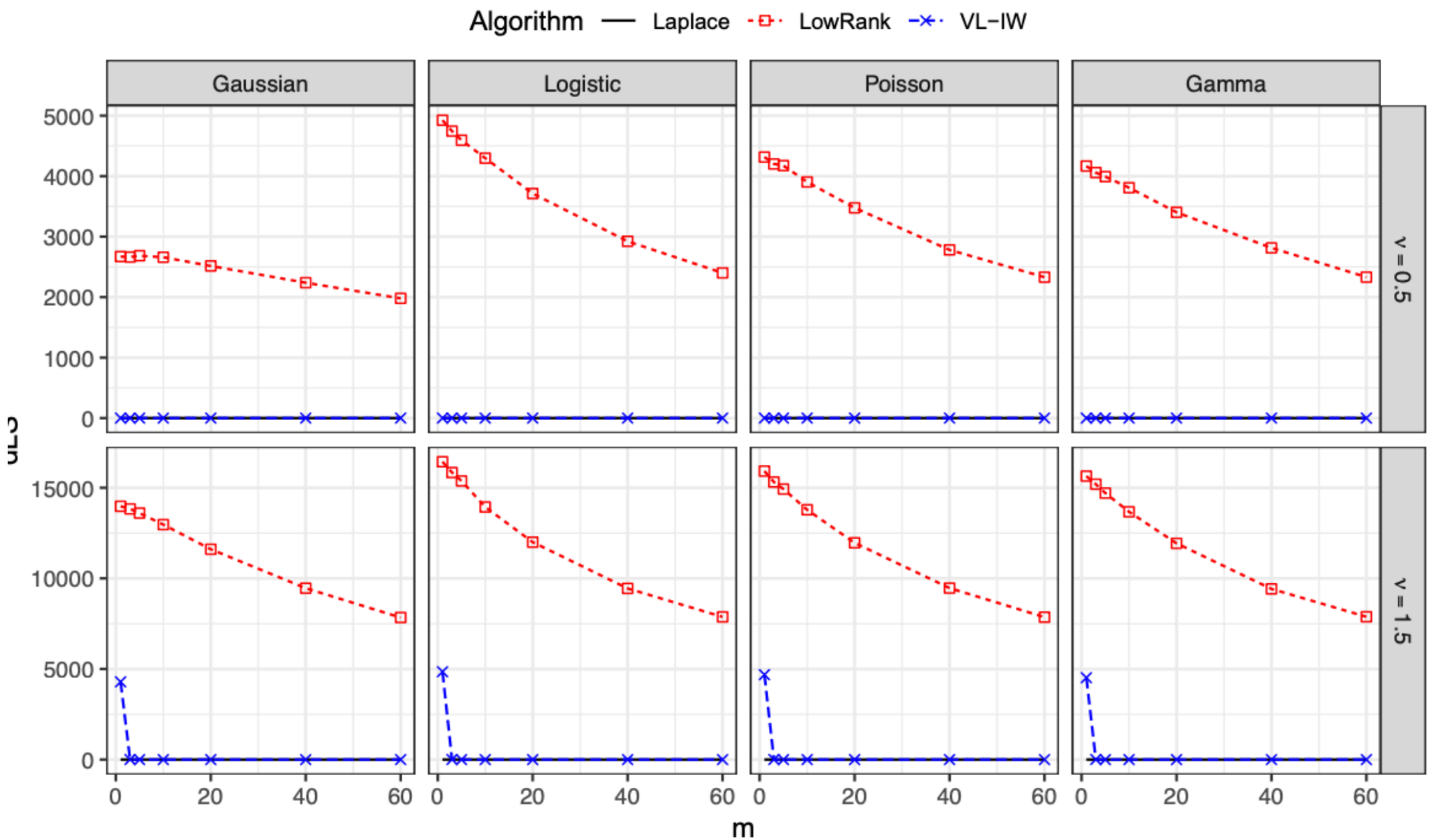
Figure 3: For sample size n between 250 and 16,000, computing time for the Laplace approximation based on Newton-Raphson, compared to VL and LowRank using Algorithm 1 with $m = 10$

Accuracy in 1D



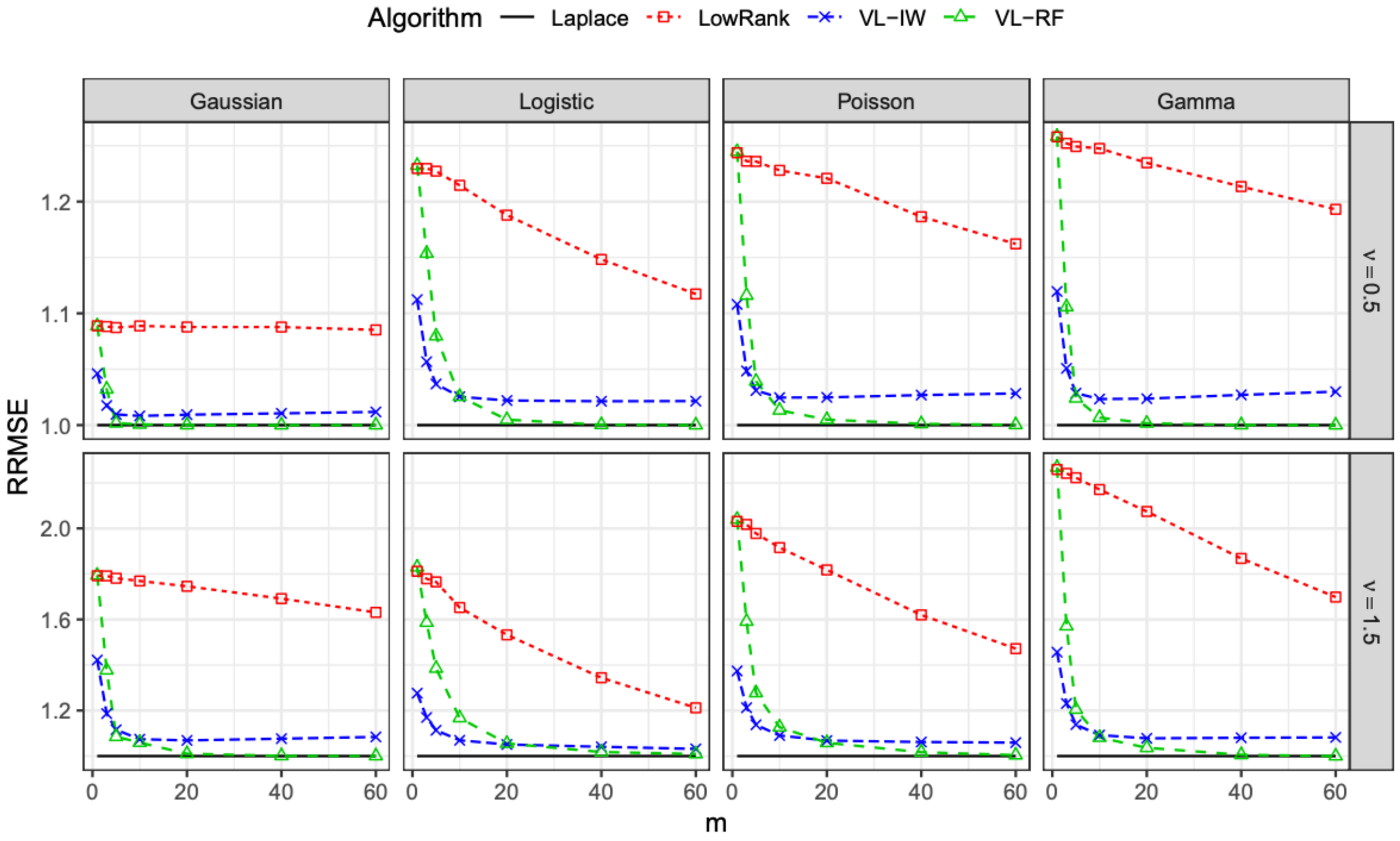
(a) RMSE (relative to Laplace)

Accuracy in 1D



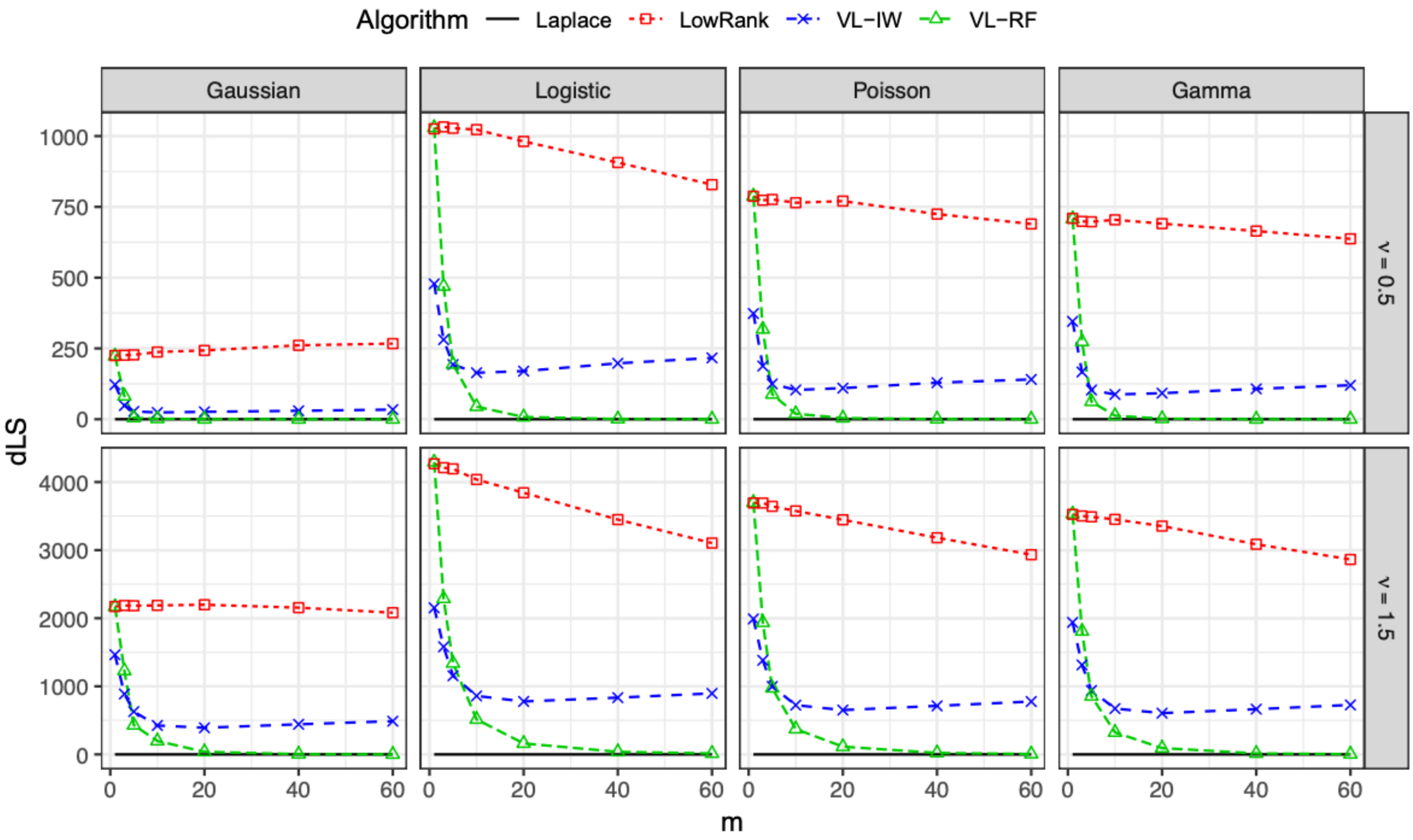
(b) Difference in log score (relative to Laplace)

Accuracy in 2D



(a) RMSE (relative to Laplace)

Accuracy in 2D



(b) Difference in log score (relative to Laplace)

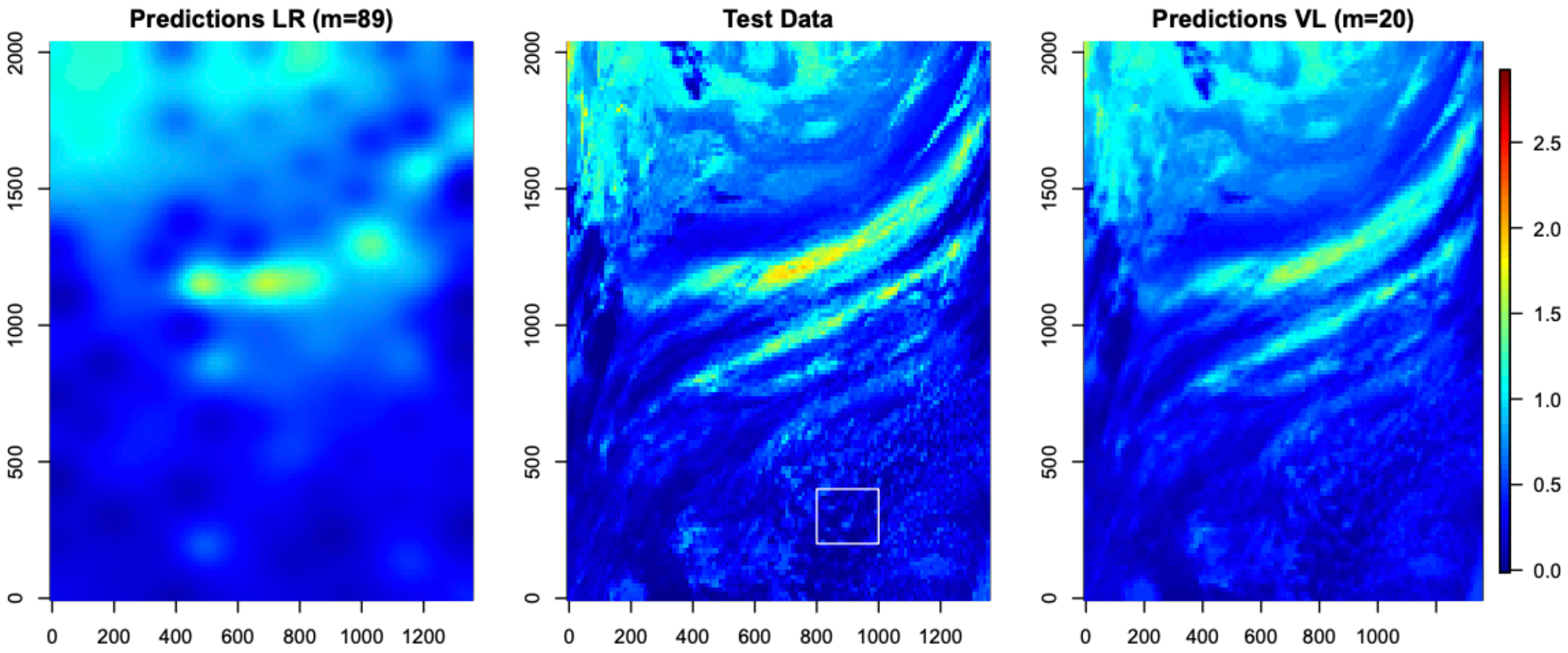
Applications

Authors apply their method to water vapor data (continuous and positive).

$$z(\mathbf{s}_i) \mid y(\mathbf{s}_i) \sim_{ind} \mathcal{G}(a, ae^{-y(\mathbf{s}_i)})$$

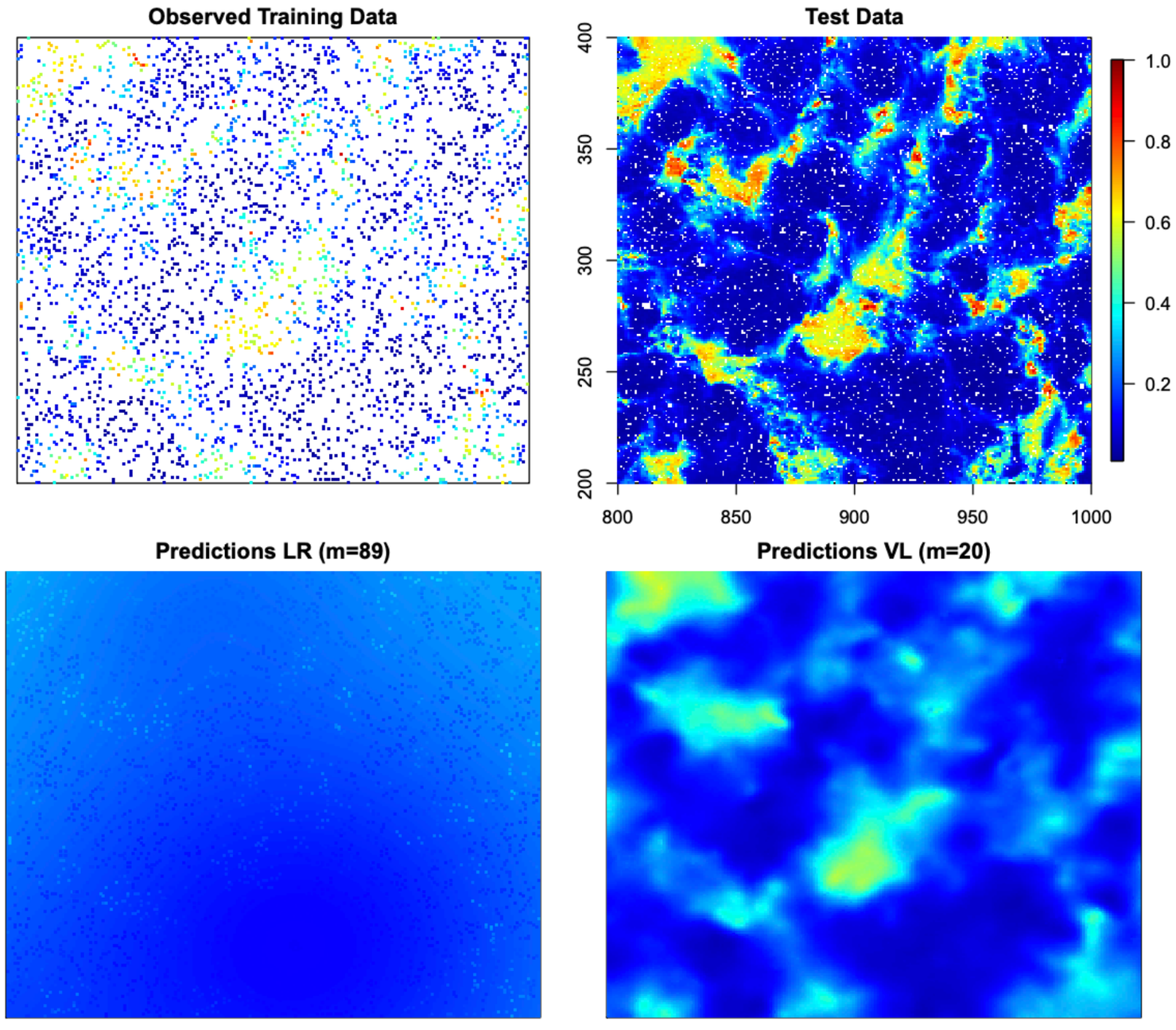
On 1354 × 2030 = 2,746,820 grid of 1km pixels. For analysis, the authors used 500000 data points.

Applications



(a) Entire spatial domain

Applications



(b) Zooming into the white square shown in Panel (a)

Discussion

Key idea is to combine Vecchia approximation for latent GP and Laplace approximation for posterior marginals of latent variables given non-Gaussian data.

Vecchia approximation relies heavily on the ordering of the model variables and the conditioning set. These choices can dramatically affect runtime and accuracy.

Code available in the R package GPVecchia.

Discussion

Use integrated nested Laplace approximation (INLA) to improve marginal posteriors

Other ideas/questions?