COMPARING PSEUDO-INPUT SELECTION METHODS IN SPARSE GAUSSIAN PROCESS REGRESSION WITH APPLICATIONS TO HEART RATE DATA

HENGRUI LUO AND GIOVANNI NATTINO

1. Introduction(2-3 paragraphs)

Gaussian process (GP) models are widely used in a variety of fields, such as engineering, economy and medicine (Rasmussen and Williams, 2006). They are a family of flexible models whose mean and covariance function are fully specified. They have been successfully applied in data mining and forecasting. However, GPs suffer from a substantial limitation, which is the high computational complexity of fitting the model when the sample size is large. Among the first attempts to address this limitation, Snelson and Ghahramani (2006) proposed a Bayesian model that reduces the information conveyed by the observed data to a smaller set of global variables, also known as inducing variables or pseudo-inputs. The idea of the authors was to identify the location of the pseudo-inputs by maximizing the marginal likelihood. Following this line, Titsias (2009a) introduced the sparse variational approximation framework, which is based on the minimization of the Kullabach-Leibler (KL) divergence between the true and the approximate posterior, simpler posterior based on the pseudo-inputs. The author derived a lower bound of the KL divergence that can be optimized with efficient algorithms. Simulation analyses and applications to real data have demonstrated that these approaches perform relatively well (Hensman et al., 2013).

In this report, we investigate the role of the choice of the pseudo-inputs in sparse GP modeling. We consider five approaches that tackle GP fitting with pseudo-inputs: deterministic location of the pseudo-input locations on an equispaced grid; simple random sampling and preferential sampling of the pseudo-inputs among the observed data points; maximization of the marginal likelihood (Snelson and Ghahramani, 2006); minimization of the Kullback-Leibler (KL) divergence between posterior and approximate posterior (Damianou and Lawrence, 2013). For each method, we briefly describe the theoretical background and we compare the performances of these approaches in a simulation study.

We finally apply sparse GP to real data. We use GP regression to model heart rate (HR) recordings. We consider the HR stream of a single subject under different level of stress (at rest and running at four different speeds). These data have been recorded in a study of HR kinetics by Zakynthinaki (2015) and are publicly available. By applying pseudo-input GP methods to this HR dataset, our goal is two-fold. First, we investigate the capability of pseudo-input GP regression to describe the underlying HR signal, comparing its performance with standard GP regression on the full dataset. Second, we heuristically evaluate whether the optimal locations of pseudo inputs might inform about characteristic patterns of the analyzed signal.

2. Background(1-2 pages)

Sparse GP regression based on pseudo-inputs has been introduced by the seminal paper by Snelson and Ghahramani (2006). The idea of the authors is described in a Bayesian framework. They considered a GP model with independent noise structure, with the form:

$$\boldsymbol{y} = \mathbf{f}(\boldsymbol{x}) + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N\left(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n\right), \tag{2.1}$$

where $\mathbf{f}(\boldsymbol{x})$ is a GP defined on the locations $\boldsymbol{x}$, $\boldsymbol{\varepsilon}$ is a vector of independent noise on the observations and $\boldsymbol{y}$ is the vector of observed responses. All of the vectors are n-dimensional. In particular, it can be shown that the computations to express the posterior and predictive posterior distributions require the inversion of a $n \times n$ covariance matrix. When n is large, this step dictates the high computational cost of GP fitting.

Snelson and Ghahramani (2006) proposed to replace the GP prior on the mean function $\mathbf{f}(\boldsymbol{x})$ by an approximating prior with masses on m pseudo-inputs $\bar{\boldsymbol{x}}$, with $m << n$. This is a data reduction operation: points in a high dimensional space $\boldsymbol{x} \in \mathcal{X}$— used to fit the exact model—are replaced with their corresponding representation $\bar{\boldsymbol{x}}$ with respect to a lower dimensional space defined by a subset of the basis of $\mathcal{X}$. Friedman et al. (2001) noted the analogy between basis truncation and this type of data reduction, which has been denoted as a “subset of regressors” method (Quiñonero-Candela and Rasmussen, 2005).

As a data reduction method, some information is naturally loss in the process. In the approach described by Snelson and Ghahramani (2006), the main loss is due to the approximation of the covariance matrix of the marginal likelihood with the Nyström method. Denote with K_{nn} the covariance matrix of the GP process $\mathbf{f}(\boldsymbol{x})$. The exact likelihood

$$p(\boldsymbol{y}\,|\boldsymbol{x}\,) = N\left(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n + K_{nn}\right) \tag{2.2}$$

is approximated by

$$p(\boldsymbol{y}\,|\boldsymbol{x}, \bar{\boldsymbol{x}}\,) = N\left(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n + diag\left[K_{nn} - K_{nm}K_{mm}^{-1}K_{nm}^T\right] + K_{nm}K_{mm}^{-1}K_{nm}^T\right), \tag{2.3}$$

where K_{nm} is the $n \times m$ matrix with the covariances between the GP at the n original inputs (i.e., $\mathbf{f}(\boldsymbol{x})$) and the GP at the m pseudo-inputs (i.e., $\mathbf{f}(\bar{\boldsymbol{x}})$). The matrix K_{mm} is the covariance matrix of the GP at the m pseudo-inputs. The approximate likelihood can be derived assuming a GP model on the pseudo input with form $p(\mathbf{f}(\bar{\boldsymbol{x}})\,|\boldsymbol{x}\,) = N\left(\mathbf{0}_n, K_{mm}\right)$. The full dataset is approximated by a regression model on the pseudo-inputs, whose likelihood $p(\boldsymbol{y}\,|\boldsymbol{x}, \bar{\boldsymbol{x}}, \mathbf{f}(\bar{\boldsymbol{x}})\,)$ depends on the value of the GP at the pseudo-input locations. The marginal likelihood in equation (2.3) is computed by marginalizing over these unobserved locations, i.e., $p(\boldsymbol{y}\,|\boldsymbol{x}, \bar{\boldsymbol{x}}\,) = \int p(\boldsymbol{y}\,|\boldsymbol{x}, \bar{\boldsymbol{x}}, \mathbf{f}(\bar{\boldsymbol{x}})\,) p(\mathbf{f}(\bar{\boldsymbol{x}})\,|\boldsymbol{x}\,) d\mathbf{f}(\bar{\boldsymbol{x}})$.

The choice of the pseudo-input locations $\bar{\boldsymbol{x}}$ is essential to approximate the likelihood based on the complete dataset. The idea is that good pseudo-input locations can balance the loss of information in the approximation with the computational efficiency gained in the data reduction. Even confining the selection of pseudo-inputs to the observed locations of the GP, greedy algorithms exploring optimal combinations of m pseudo-inputs among the n observations are computationally unsustainable. Indeed, the complexity of these type of algorithm is $\mathcal{O}\left(\binom{m}{n}\right)$. Therefore, the pseudo-inputs must be chosen using other criteria.

A naive choice is to consider the pseudo inputs as fixed, known locations and to use the marginal likelihood of the approximated model to carry out inference on the other parameters, i.e., σ^2 and the parameter(s) of the covariance kernel. For example, the locations can be chosen as a random subsample of the observed data points or with space-filling designs. In these cases, estimators for σ^2 and for the covariance kernel parameters can be interpreted as partial maximum likelihood estimators (PMLE), since the marginal likelihood is only maximized with respect to a subset of the parameters.

Despite its simplicity, naive approaches based on arbitrary choices of the locations might be suboptimal in specific applications. To overcome this limitation, Snelson and Ghahramani (2006) suggested to jointly maximize the marginal likelihood with respect to pseudo-input locations and parameters of the model.

Damianou and Lawrence (2013) and Titsias (2009a) criticized this framework because of the impossibility to evaluate the reliability of the approximation. Furthermore, the m-dimensional augmentation of the parameter space exposes the approach to the risk of over-fitting (Titsias, 2009a). To address these issues, Titsias (2009a) described a variational formulation of the problem. The author introduced an approximating model and proposed to estimate its parameters by minimizing the KL divergence between the full-data posterior and the posterior of the approximate model. In this way, such a formulation directly targets the identification of an approximate model that is as close as possible to the original one—according to the KL divergence criterion. The computational costs of fitting the approximate model can be further reduced if the minimization is based on a lower bound of the KL divergence, which has a more convenient expression than the KL divergence itself (Titsias, 2009a). In particular, Hensman et al. (2013) describe a very efficient stochastic gradient descent approach to conduct this optimization. Natural connections of this methodology to optimal designs in the full dataset have been described by da Silva Ferreira et al. (2015).

We apply sparse GP regression to model HR data. GP are well-suited to model streams of vital signs, such as HR recordings, because they can account for random noise and "knowledge of the generative physiology" (Colopy et al., 2016). Current research focuses on identifying patterns of HR streams of critical patients, to detect signals of deterioration (Almeida and Nabney, 2016). In other contexts, the HR can be used to evaluate the level of physical preparation and design training/rehabilitation activities (Zakynthinaki, 2015). Notably, HR streams often consist in a massive amount of data points. This is because clinical researchers are interested in detecting patterns over long time intervals, and the measurements are taken with high frequency. Moreover, recording the HR in beats/minute is very easy and can be performed with cheap wearable devices. In this context, the large amount of available data makes standard GP regression computationally expensive. We compare the performance of pseudo-input selection methods in sparse GP regression on HR data.

3. Methodology(5-7 pages)

We consider different methods to choose pseudo-inputs in sparse GP regression. Our real-data application is one-dimensional, since we model the HR data over time. Therefore, we focus on one-dimensional GP.

We evaluate five different criteria to select the pseudo-input locations. Three of them are naive selections, that is, they are not optimal with respect to a measure or criterion: 1) pseudo-inputs locations over an equispaced grid; 2) simple random sampling and 3) preferential sampling among the locations of the data points. We also considered two methods identifying optimal pseudo-input locations with respect to some criterion: 4) maximization of the marginal likelihood of the approximate model; 5) minimization of the KL divergence between true and approximate posterior distributions. We briefly describe the five methods for pseudo-input selections in section 3.1.

Classically, the number m of pseudo-inputs is fixed *a priori* to an arbitrary value. To the best of our knowledge, the literature does not offer guidelines for the choice of m in sparse GP regression. This is an important specification. If m is set to a value that is too large, the computational gain might be insufficient for very large datasets. On the other hand, if m is too small, the resulting approximate GP might fail to capture important features of the full-data process. Titsias (2009a) discusses how the value of m might be treated as a free parameter. In this case, m could be identified within the optimization process that is used to select the pseudo-input locations. However, in the practical computations, the author fixes the number of pseudo-inputs to arbitrarily chosen values.

We designed a simple simulation study to compare the considered approaches. First of all, we evaluated the capability of the five pseudo-input GP regression methods to describe a draw from a GP using different values of m. Secondly, we fixed m and we evaluated the reliability of the methods to describe a collection of GP draws. The setup is described in detail in section 3.2.

The methods that performed well in the simulation analysis were subsequently applied to our real data. We carefully describe our application to the HR data in section 3.3.

3.1. **Compared Methods.**

3.1.1. *Equispaced Grid.* The pseudo-input locations are considered as the points of a m-dimensional equispaced grid in the design space. This choice can be interpreted as a simple and deterministic method to identify a space-filling design in one-dimensional spaces. Notably, grid designs are outperformed by more sophisticated approaches in higher dimensional spaces. For example, fixed the number of points, Latin Hypercube Designs combined with space-filling criteria provide locations characterized by better coverage of the space and representation of the lower-dimensional projections.

3.1.2. *Simple Random Sampling (SRS) of Observed Locations.* The pseudo-input locations are considered as a m-dimensional simple random sample (SRS) of the n locations observed in the dataset. The rationale of such a naive criterion is to approximate the model with pseudo-inputs locations that are representative of the large number of observed locations. Sampling the observed locations with equal probability is an easy way to achieve a design with higher representation of the regions where there are several observation. As the grid design, such a naive choice of pseudo-input locations has minimal computational costs.

3.1.3. *Preferential Sampling (PS) of Observed Locations.* The main idea of the PS method is to prefer pseudo-input locations whereas the variability of the response is large. A simple approach to obtain this type of pseudo-inputs is to sample m locations with probabilities proportional to $\Delta y/\Delta x$. The motivation of this choice is that regions where the changing rate is larger are likely to contain more information, which should be taken into account to fit the GP. The density of pseudo-input locations selected with PS will be higher in regions with larger changing rates. For clarity, we provide a brief pseudo code that describes the PS pseudo-input selection.

```
n = length(y)
ySort, xSort = sort y and x by increasing value of x
yDiff = abs(ySort[2:n] - ySort[1:(n-1)])
xDiff = abs(xSort[2:n] - xSort[1:(n-1)])
p = (yDiff/xDiff) / sum(yDiff/xDiff)
sample m elements from x[2:n] with probabilities p
```

Algorithm 1: Pseudo code for PS method. The algorithm samples m pseudo-inputs among the n observed locations $\boldsymbol{x}$. The vector $\boldsymbol{y}$ contains the observed values of the response.

3.1.4. *Maximization of the Marginal Likelihood (ML).* The pseudo-input locations are determined by jointly maximizing the marginal likelihood in equation (2.3) with respect to the pseudo-input locations and the parameters of the model (Snelson and Ghahramani, 2006). By maximizing the likelihood function, the approach targets the "most likely" locations according to the observations. In the implementation, we use the closed form expression of marginal likelihood for pseudo-inputs given in Snelson and Ghahramani (2006).

The augmentation of the parameter space by a m-dimensional vector of parameters (i.e., the pseudo-input locations) provides a clear computational advantage over the fit of the GP on the full dataset. The maximization of the likelihood in equation (2.3) only requires the inversion of a $m \times m$ matrix. The computational cost is therefore reduced from $\mathcal{O}(n^3)$ to $\mathcal{O}(nm^2)$. The gain is substantial if $m << n$.

However, this apparent advantage has an important drawback, which is the increased complexity of the optimization problem. In particular, the parameter space where the optimization takes place is increased by m dimensions. Even in scenarios where n is only moderately large and $m << n$, a reasonable number of pseudo inputs cannot be smaller than 5-10 units, which is a non-trivial increase in dimensionality. Moreover, the function that we optimize is characterized by several peaks in many scenarios. This can be easily shown by noting that, if $\boldsymbol{x}$ corresponds to the global optimum of the marginal likelihood, any permutation of its component are globabl optima of the marginal likelihood as well. This complicates the optimization procedure.

Because of the high complexity of the optimization problem, we found that the optima identified by (non-stochastic) gradient descent algorithms were sensitive to the starting points in the parameter space. We therefore opted to stable annealing optimization, using the implementation in the `GenSA` package that has been described by Xiang et al. (2013). Simulated annealing algorithms are appropriate for global optimization in high-dimensional spaces, in particular when the identification of an approximate global optima is preferred to a very precise identification of a local optimum (Xiang et al., 2013). In our results, this method turned out to be numerically stable. The downsize of this approach is the higher computational cost compared to gradient-based optimization algorithms.

3.1.5. *Minimizing the Lower Bound of KL Divergence between Posterior and Approximate Posterior.* The idea of the variational learning of the pseudo-inputs is to directly approximate the true posterior distribution $p\left(\mathbf{f}(\boldsymbol{x}), \mathbf{f}(\bar{\boldsymbol{x}}) \mid \boldsymbol{y}\right) = p\left(\mathbf{f}(\boldsymbol{x}) \mid \mathbf{f}(\bar{\boldsymbol{x}}), \boldsymbol{y}\right) p\left(\mathbf{f}(\bar{\boldsymbol{x}}) \mid \boldsymbol{y}\right)$, where $p\left(\mathbf{f}(\boldsymbol{x}) \mid \mathbf{f}(\bar{\boldsymbol{x}}), \boldsymbol{y}\right)$ is the posterior distribution given the pseudo-inputs locations $\bar{\boldsymbol{x}}$ and the corresponding values of the GP $\mathbf{f}(\bar{\boldsymbol{x}})$. Titsias (2009a) proposed to use the approximate posterior $q\left(\mathbf{f}(\boldsymbol{x}), \mathbf{f}(\bar{\boldsymbol{x}})\right) = p\left(\mathbf{f}(\boldsymbol{x}) \mid \mathbf{f}(\bar{\boldsymbol{x}}), \boldsymbol{y}\right) \phi\left(\mathbf{f}(\bar{\boldsymbol{x}})\right))$, by appropriately selecting the function ϕ. The author proposed to select ϕ^* minimizing the KL divergence of the two distributions, that is:

$$\phi^* = \arg\min_{\phi} \ \mathrm{KL}\left(q(\mathbf{f}(\boldsymbol{x}), \mathbf{f}(\bar{\boldsymbol{x}}) \| p\left(\mathbf{f}(\boldsymbol{x}), \mathbf{f}(\bar{\boldsymbol{x}}) \mid \boldsymbol{y}\right)\right) \tag{3.1}$$

Titsias (2009a) and Titsias (2009b) provide a lower bound $F_V\left(\bar{\boldsymbol{x}}, \phi\right)$ of the KL divergence, whose global minimum coincides with the global minimum of the KL divergence. The author suggests the optimization of the lower bound, because its optimization is computationally more tractable. The optimization reduces to:

$$\phi^* = \arg\min_{\phi} F_V\left(\bar{\boldsymbol{x}}, \phi\right), \quad \text{with} \ \ F_V\left(\bar{\boldsymbol{x}}, \phi\right) = \int p\left(\mathbf{f}(\boldsymbol{x}) \mid \mathbf{f}(\bar{\boldsymbol{x}})\right) \log \frac{p(\boldsymbol{y} \mid \mathbf{f}(\boldsymbol{x})) p\left(\mathbf{f}(\bar{\boldsymbol{x}})\right)}{\phi\left(\mathbf{f}(\bar{\boldsymbol{x}})\right)} d\mathbf{f}(\boldsymbol{x}) d\mathbf{f}(\bar{\boldsymbol{x}}) \tag{3.2}$$

The authors provide a closed form expression for ϕ^*. In particular, the optimization of $F_V\left(\bar{\boldsymbol{x}}, \phi^*\right)$ with respect to $\bar{\boldsymbol{x}}$ provide the optimal location of the pseudo-inputs which correspond to the approximating posterior that is the "closest" to the true posterior in terms of the KL divergence between two probability distributions.

Although the optimization of the lower bound is less computationally demanding than the one of the KL divergence, the challenges discussed for the ML method in section 3.1.4 apply to this methodology as well. The target function is defined in an high-dimensional space and is characterized by many local optima. Similarly to the case of ML optimization, we used the stable annealing optimization described by Xiang et al. (2013).

3.2. **Simulation Analysis.** We considered two families of simulations to study different properties of the sparse GP regression methods.

3.2.1. *Effect of* m. The first set of simulations compare the performances of the sparse GP regression methods across several number of pseudo-inputs m. We considered two draws of one-dimensional GP with independent noise, with the form described in equation (2.1). Each draw was formed of $n = 100$ observations $(\boldsymbol{y}, \boldsymbol{x})$. The locations $\boldsymbol{x}$ were sampled from a uniform distribution on the $(0, 1)$ interval. The observations were generated with the Gaussian covariance kernel $K(x_1, x_2) = \sigma_K^2 \rho^{(x_1 - x_2)^2}$.

The first draw was sampled from a family of moderately smooth GPs, by setting $\rho = .1$. The parameter was set to a much smaller value ($\rho = 10^{-9}$) when sampling the second draw. A graphical representation of the two sets of observations is reported in the Appendix A.1. We will refer to these two realizations as "smooth" and "wiggly" draws. The variance parameter σ_K was set equal to 1 in both the cases. The noise on the observations was generated with independent normal samples from a normal distribution with $\sigma = 0.05$ (see equation (2.1)).

Each pseudo-input GP method was applied to the two realizations using several values of m. We considered sparse GP regression with $m = 5, 7, 10, 12, 15, 20, 30, 40$ and 50 pseudo-inputs. For each method and for each value of m, we evaluated the reliability of the fitted GP using the mean squared prediction error on the observed data, defined as $1/n \left\| \widehat{y}(\boldsymbol{x}) - \boldsymbol{y} \right\|^2$, where $\widehat{y}(\boldsymbol{x})$ is the posterior predictive mean. In addition, we visually evaluated the reliability of the fitted GP by plotting the posterior predictive mean (and the 95% pointwise prediction interval) on the scatterplots of the observations.

3.2.2. *Performances across GP Realizations.* A second set of simulations investigated the reliability of sparse GP regression over several GP draws. Using a setup similar to the first family of simulations, we drew 50 realizations from a family of smooth GP ($\rho = .1$) and 50 realizations of a wiggly GP($\rho = 10^{-9}$). Each realization consisted in $n = 100$ observations. As before, we used a Gaussian covariance kernel, the locations $\boldsymbol{x}$ were sampled uniformly on $(0, 1)$ interval and the covariance and noise parameters were set to $\sigma_K = 1$ and $\sigma = 0.05$. We applied all of the methods described in section 3.1. We fixed the number of pseudo-inputs to a value that showed good performances in the first simulation analysis, that is, $m = 10$.

For each method and for each GP draw, we evaluated the prediction accuracy of the pseudo-input GP with the mean squared prediction error on the training dataset. Again, we plotted the posterior predictive mean function with the appropriate 95% prediction interval, and we visually evaluated the capability of the predictive function to describe the observations.

3.3. **HR Analysis.** We used the pseudo-input GP regression to model HR data. The data were recorded within a clinical study investigating the HR kinetic (Zakynthinaki, 2015). A single 33-year old runner was asked to run on a treadmill at four constant speeds: $v = 13.4, 14.4, 15.7$ and 17.0km/h. The HR in beats/minute was recorded for about 6 minutes from the beginning of the exercise. The four sets of HR recordings are labeled as Exercise 1 to 4 respectively. After each exercise, the HR of the subjects was measured for about 8 minutes from the end of the exercise. The data consist of eight streams of HR recordings (four exercise and four recovery recordings). We discarded the portion of the data corresponding to transient phases, to avoid the complication of fitting an appropriate trend in our model and to make sure that our data respected stationarity assumptions. We therefore discarded the first 150 seconds of the exercise recordings and the first 100 seconds of the recovery recordings. A graphical representation of the eight HR streams is reported in the Appendix A.2.

For each of the eight streams, we applied the pseudo-inputs GP regression approaches standing out for reliability in the simulation study. We considered the grid, ML and KL pseudo-input GP regression methods.

Each HR stream counted about 500 recordings (mean: 495, min-max:444-559). In each fit, we considered a fixed number of $m = 40$ pseudo-inputs. We visually compared the posterior predictive mean with the fit of the GP based on the full dataset (implemented in the "km" function of the "DiceKriging" package). The 95% prediction interval of the full-data and of the pseudo-input GP were reported on the plot as well. We quantified the performances of the three pseudo-input GP methods with the mean square prediction error.

Finally, we visually compared the optimal locations of the pseudo-inputs in ML and KL methods across the eight streams of HR recordings. Our goal was to evaluate whether the identified pseudo-input locations provided any information about the level of stress of the subject. Being the result of a selection of locations conveying "most" of the information of the signal (in likelihood and KL divergence sense), the rationale of our heuristical evaluation is that the intensity of the effort might lead to different concentration of the pseudo-inputs.

4. Results(3-4 pages)

The R code that we used for the simulation study and for the application to the HR data is reported in the Appendix A.4. We describe the results of our analysis in the following sections.

4.1. **Simulation Study.**

4.1.1. *Effect of m.* The results of the first set of simulations are reported in Figure 4.1. For each considered method, we report the mean squared prediction error of the predictive posterior mean function. The left and right panels report the results for the smooth and wiggly GP realizations, respectively. The graphical representations of the fitted GPs for each method are reported in Figures A.1 and A.2 in the Appendix for $m = 5, 10$ and 50.

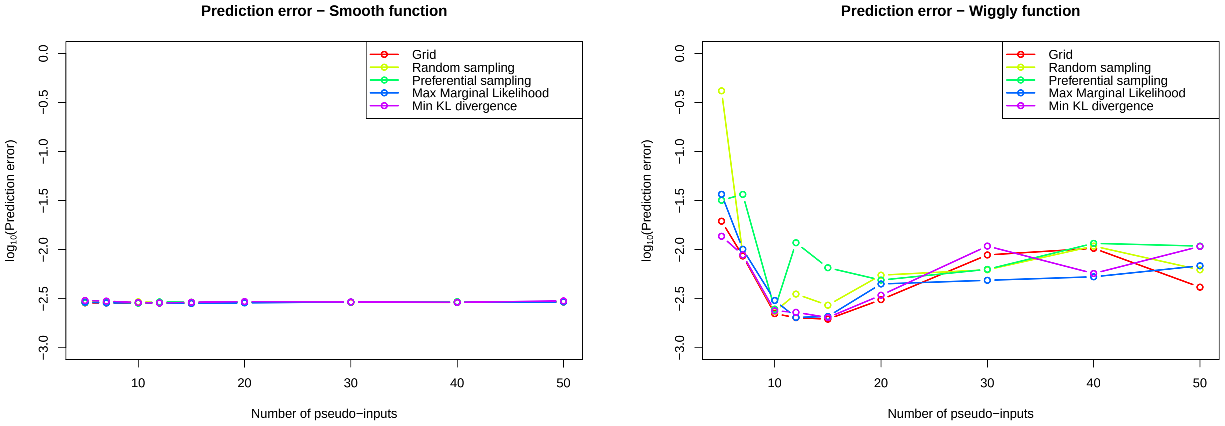


Figure 4.1. Mean squared prediction error of the predictive posterior mean function for the smooth (left panel) and wiggly (right panel) GP realizations. See Figures A.1 and A.2 for a graphical representations of the fitted GPs for $m = 5, 10$ and 50 (representations of other values of m are omitted for space-related constraints).

Notably, the number of pseudo-inputs does not affect the prediction error in the smooth realization. The mean squared prediction error is small and constant across all the considered values of m. Interestingly, even a very small number of pseudo-inputs was sufficient to appropriately reconstruct the GP. To visually appreciate how well the sparse GP regression methods describe the observations, we report the predictive posterior distributions for $m = 5, 10$ and 50 in Figure A.1 of the Appendix.

Conversely, the number of pseudo-inputs has a clear impact on the prediction error when the wiggliness increases. In all of the methods, the prediction error is maximum for very small numbers of pseudo-inputs ($m = 5$). In our simulation, the error drops quickly as m increases, reaching the smallest values for m around 10/15. Interestingly, higher values of m do not appear to be characterized by smaller prediction errors. Overall, sparse GP methods based on sampling of the pseudo inputs (SRS and PS) show larger "noise" in the prediction error, which was expected given the underlying randomness of the selection process.

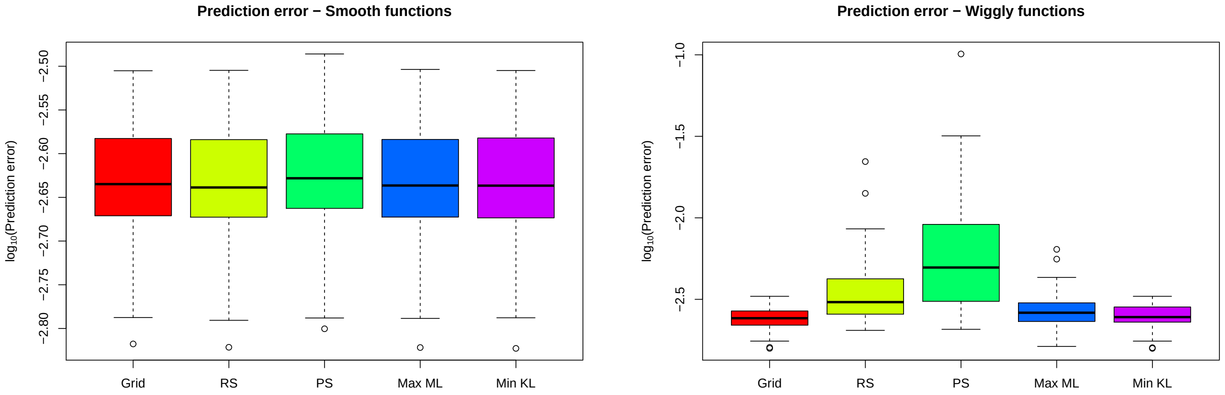


FIGURE 4.2. Distribution of mean squared prediction error of the predictive posterior mean function over 50 realizations for the smooth (left panel) and wiggly (right panel) GP families.

4.1.2. *Performances across GP Realizations.* Figure 4.2 reports the result of the second family of simulations. We use boxplots to summarize the distributions of the mean squared prediction errors across simulations. The errors in samples from smooth and wiggly GP are reported in the left and right panels, respectively. As discussed in section 3.2, we fixed the number of pseudo-inputs to $m = 10$ in all of the simulations.

The left panel confirms what suggested by the first set of simulations. The performances of the different approaches across smooth GP realization are extremely similar, in terms of prediction error. Very simple naive methods based on arbitrary locations (grid) or random samples of the observed locations (SRS) are able to describe smooth processes as precisely as sophisticated approaches that search for "optimal" locations in the high-dimensional space of pseudo-input locations (ML and KL methods).

When the GP are drawn from a distribution of wiggly processes, however, the results are very different. Overall, the prediction error for the grid, ML and the KL method is clearly smaller than the one associated to the random sampling methods (SRS and PS). Interestingly, the naive approach based on deterministic, space-filling locations (grid) performed very well.

4.2. **HR Analysis.** The results of the simulation analysis suggest that 1) a number of pseudo-inputs that is around 10% provided a good approximation of the underlying GP and 2) the naive grid method and the ML and KL approaches outperformed the other considered methods. For these reasons, we used the grid, ML and KL approaches to fit a GP on the eight HR stream data and we fixed the number of pseudo-inputs to $m = 40$ in all of the fits.

Figure 4.3 provides the predictive means for the three methods in two of the eight sets of HR recordings, specifically for the exercise with speed $v = 13.4$km/h (Exercise 1) and for the recovery session after this exercise. GP fits for the other exercises and recovery sessions are reported in Figure A.3 and A.4 of the Appendix. Each panel also provides the prediction based on the standard full-data GP regression.

All of the methods capture the overall trends of the HR recordings in each scenario. However, sparse GP regression approaches fail to precisely describe smaller fluctuations and spikes. The estimate of the noise in the data was smaller in the grid-based method than the other two approaches, in particular in the fit of the HR records during the physical activity.

We report the considered pseudo-input locations as green points at the bottom of each panel. In the grid method, such locations are arbitrarily fixed to the m-th dimensional grid spanning the time interval. For the ML and KL methods, whose pseudo-input locations are the result of a numerical optimization, we also report the starting locations fed to the numerical algorithm as blue points at the top of the panel. Histograms of the pseudo-input locations over time are reported in Figures A.5 and A.6 in the Appendix. We visually analyzed these distributions, investigating whether patterns in the selected pseudo-input locations might provide information about the level of stress of the subjects. We did not find any clear difference in these distributions.

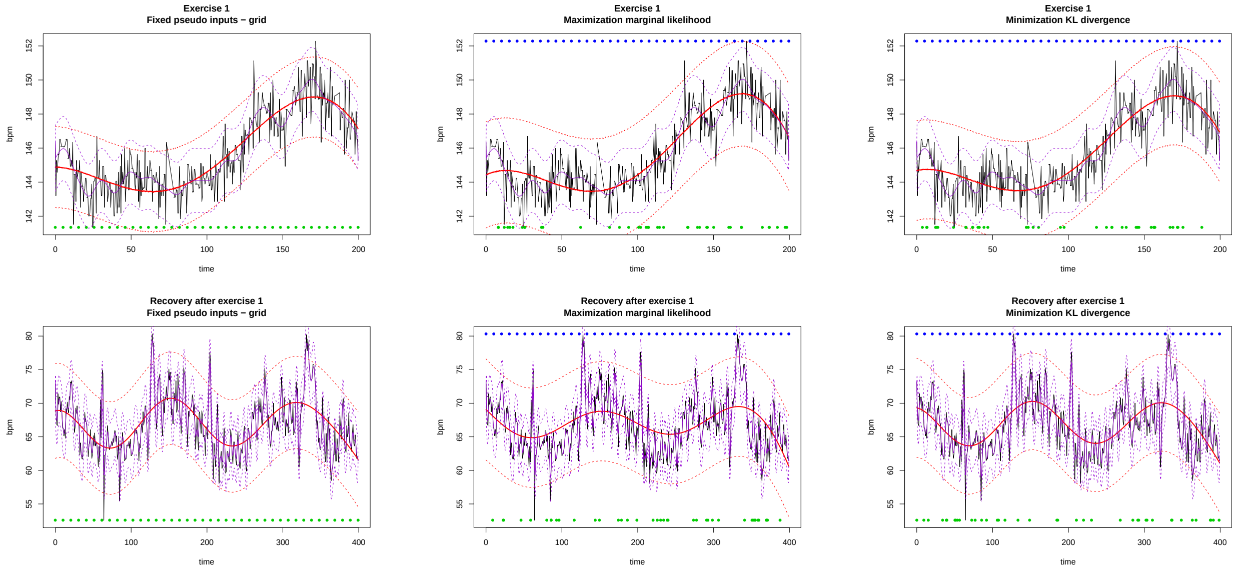


Figure 4.3. Posterior predictive means (red solid line) and 95% prediction intervals (red dashed lines) of the pseudo-input GP based on grid (left panels), ML (central panels) and KL (right panels) methods. In each panel, we also report the prediction based on standard full-data GP regression (purple solid lines) and the 95% prediction intervals (purple dashed lines). The considered pseudo-input locations (for all the methods) are reported as green points at the bottom of the panel. Starting locations for the numerical algorithm (for the ML and KL methods) are reported as blue points at the top of the panel.

5. Discussion(1-2 pages)

We explored the performances of different methods to select pseudo-input locations in sparse GP regression. With a simulation study, we investigated how such performances depend on the number of pseudo-inputs and on the degree of smoothness of the underlying data generating process.

Our results suggest that naive criteria for pseudo-input selection perform as well as complex approaches targeting "optimal" choices. In particular, distributing the locations on a uniform grid over the design space appears to be an approach at least as reliable as the identification of the pseudo-input locations by maximization of the marginal likelihood or minimization of the KL divergence between posterior and approximate posterior distributions.

This result was observed also on our HR data analysis. Predictive values and prediction intervals based on arbitrary selection of the pseudo-inputs performed just as well as more complicated approaches optimizing over the pseudo-input locations.

Interestingly, the prediction error did not appear to monotonically decrease with the increase of the number of pseudo-inputs in our simulations. On the contrary, for a sample from a very wiggly family of GPs, the performances of sparse GP methods worsened when the number of pseudo-inputs was very large (30-50% of the observations). This phenomenon might appear as contradictory at a first glance, since one might expect that smaller reductions in the dimensionality of the data should correspond to higher level reconstruction of the underlying signal. However, there is a trade-off between the data reduction and the complexity of the optimization problem to solve. The larger the number of pseudo-inputs, the smaller the degree of data reduction, the higher the number of parameters that are considered in the model.

Such a trade-off also manifests in terms of computational costs, in particular in the methods that treat the locations of the pseudo-inputs as parameters to consider in the optimization (ML and KL methods). The larger is the number of pseudo-inputs, the higher is the dimensionality of the parameter space where the optimization takes place.

Our analysis has important limitations. First of all, our simulation study is very limited in terms of scenarios explored and number of simulations. We considered only two degrees of smoothness of the GP and only simulations based on the Gaussian covariance kernel. We also considered only the Gaussian covariance kernel in the application to the real data. Previous research has adopted more complex kernels defined as a weighted sum of Matérn and Gaussian kernels (Colopy et al., 2016). As a preliminary analysis, we tried to used a simple kernel for our model. Further research should investigate whether the predictive capability of the explored approaches can be improved by using different covariance kernels.

Another limitation of our work is that we only focused on the in-sample prediction error. We evaluated the reliability of sparse GP methods on the same sample that we used to train the GP. Further analysis should evaluate the reliability of these approaches in out-of-sample observations. Such evaluation should use simple split-sample techniques or algorithms that are more computationally intensive, such as "leave-one-out" or cross-validation procedures.

We did not address the problem of possible non-stationarity in the data. Under non-stationarity conditions, the performance of out-of-sample predictions might be very poor even if the sparse GP method perfectly reconstructs the underlying signal, that is, even if the approximate posterior is extremely close to the true predictive posterior based on the full data. For example, Paciorek (2003) discussed the potential problems of non-stationary data. Unfortunately, a well-established procedure detecting non-stationarity of underlying signals is still not available. This is an important issue to keep in mind when using sparsification procedures, whose reliability stricly depends on the assumption of stationarity. Examples of models addressing the non-stationarity nature of the data are provided by Atkinson and LLoyd (2007), Toal and Keane (2012) and da Silva Ferreira et al. (2015).

We heuristically explored whether the locations of the pseudo-inputs selected by ML and KL methods provided any information about the level of physical activity of the subject. To the best of our knowledge, the interpretation of the pseudo-input locations has not be fully explored in the literature. If the locations of the pseudo-inputs in HR modeling could provide information about the stress of the subject, such location might inform about the deterioration of the patient and help physicians in early detection of clinical emergencies. We did not spot obvious patterns in our results. Further studies should explore this research question.

Finally, we only focused on modeling of one-dimensional processes, since our project was motivated by the analysis of HR streams. Although in many research fields the processes to model are function of multidimensional inputs, a similar choice was taken in the study by Hensman et al. (2013), where the authors only considered two-dimensional GP. To give an example in our specific field of application, the value of other vital signs (such as the blood pressure or the oxygen saturation) might be used to explain HR recorded values. Importantly, the performances of the considered methods might be radically different in higher dimensional spaces. For example, the grid-based pseudo-input design performed very well in our simulations but has clear drawbacks in multi-dimensional spaces. In facts, the number of points in a grid grows exponentially with the number of dimensions and the grid points have a very poor representation in lower-dimensional projections. Space-filling methods, such as Latin Hypercube Designs, should be explored in multi-dimensional GP modeling.

References

Vânia G Almeida and Ian T Nabney. Early warnings of heart rate deterioration. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, pages 940–943. IEEE, 2016.

Peter M Atkinson and Christopher D LLoyd. Non-stationary variogram models for geostatistical sampling optimisation: An empirical investigation using elevation data. *Computers & Geosciences*, 33(10):1285–1300, 2007.

Glen Wright Colopy, Marco AF Pimentel, Stephen J Roberts, and David A Clifton. Bayesian gaussian processes for identifying the deteriorating patient. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, pages 5311–5314. IEEE, 2016.

Gustavo da Silva Ferreira, Dani Gamerman, et al. Optimal design in geostatistics under preferential sampling. *Bayesian Analysis*, 10(3):711–735, 2015.

Andreas Damianou and Neil Lawrence. Deep gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215, 2013.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013.

Christopher Joseph Paciorek. *Nonstationary Gaussian processes for regression and spatial modelling.* PhD thesis, PhD thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania, 2003.

Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6(Dec):1939–1959, 2005.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, 2006.

Edward Snelson and Zoubin Ghahramani. Sparse gaussian processes using pseudo-inputs. In *Advances in neural information processing systems*, pages 1257–1264, 2006.

Michalis K Titsias. Variational learning of inducing variables in sparse gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, pages 567–574, 2009a.

Michalis K Titsias. Variational model selection for sparse gaussian process regression. Technical report, Technical report, School of Computer Science, University of Manchester, 2009b.

David John James Toal and Andy J Keane. Non-stationary kriging for design optimization. *Engineering Optimization*, 44(6):741–765, 2012.

Yang Xiang, Sylvain Gubian, Brian Suomela, and Julia Hoeng. Generalized simulated annealing for global optimization: The gensa package. *R Journal*, 5(1), 2013.

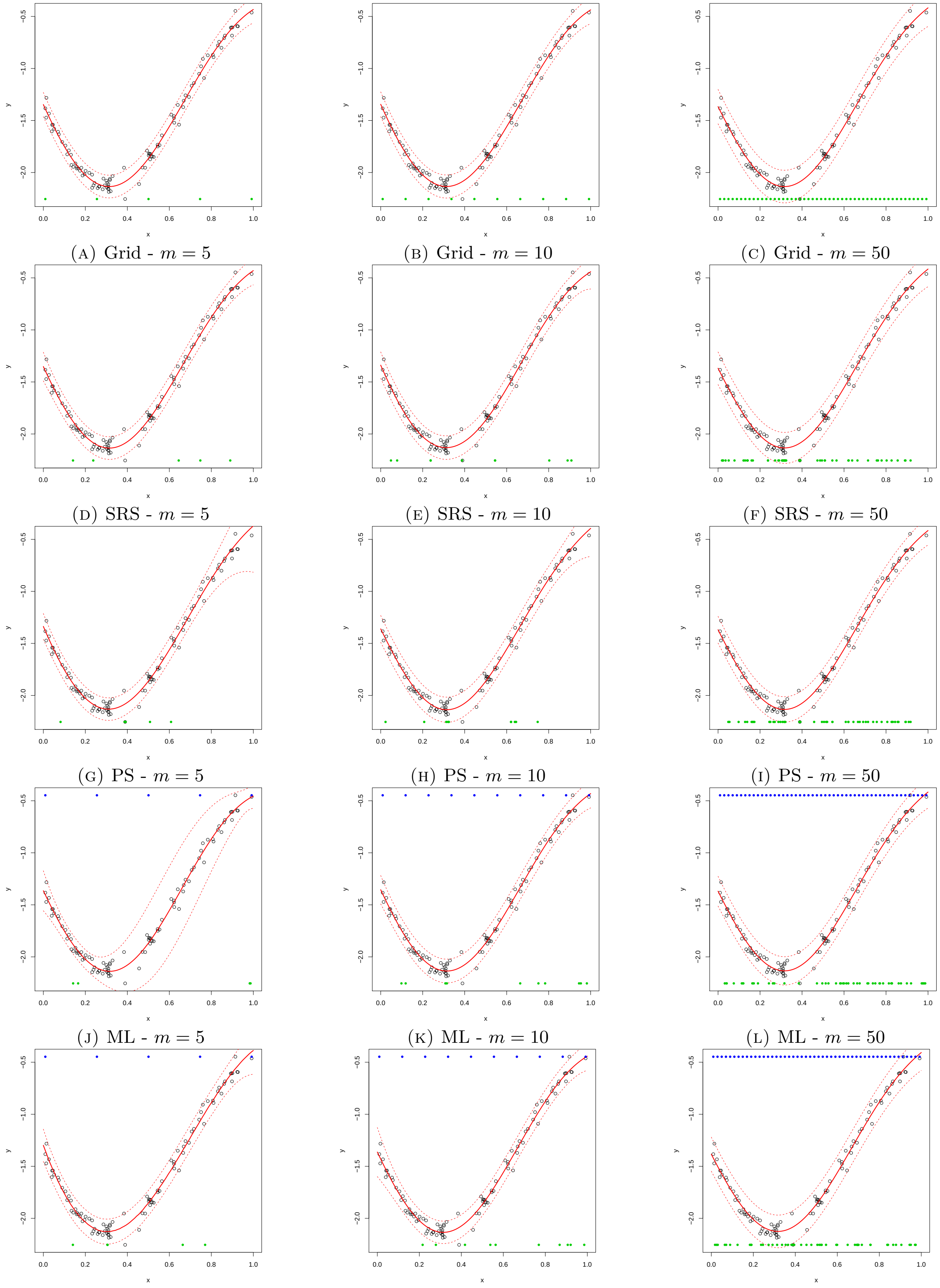
Maria S Zakynthinaki. Modelling heart rate kinetics. *PloS one*, 10(4):e0118263, 2015.

Appendix A.

A.1. **Graphical Representation of Simulations for Effect of m.** Figures A.1 and A.2 provide a graphical representations of the fitted GPs for the smooth and wiggly GP realizations, respectively. We only provide the graphical representation of the cases m =5, 10 and 50. The fitted models for other values of m are omitted for space-related constraints. In each panel, we report the considered pseudo-input locations as green points at the bottom of the panel. For the ML and KL methods, whose pseudo-input locations are the result of a numerical optimization, we report the starting locations for the numerical algorithm as blue points at the top of the panel.

A.2. **Graphical Representation of HR data.** Figure A.3 and A.4 report the posterior predictive means (and 95% prediction intervals) of the pseudo-input GP fit on the HR data with ML and KL approaches, respectively. In all the cases, we fixed the number of pseudo-inputs to $m = 40$. In each panel we also report the predictions based on the full data GP (and the 95% prediction intervals). Notably, the predictions capture the overall fluctuations of the HR recordings in each scenario.

Figure A.5 and A.6 provide a graphical representation of the distribution of the pseudo-input locations in each scenario.



(A) Grid - $m = 5$

(B) Grid - $m = 10$

(C) Grid - $m = 50$

(D) SRS - $m = 5$

(E) SRS - $m = 10$

(F) SRS - $m = 50$

(G) PS - $m = 5$

(H) PS - $m = 10$

(I) PS - $m = 50$

(J) ML - $m = 5$

(K) ML - $m = 10$

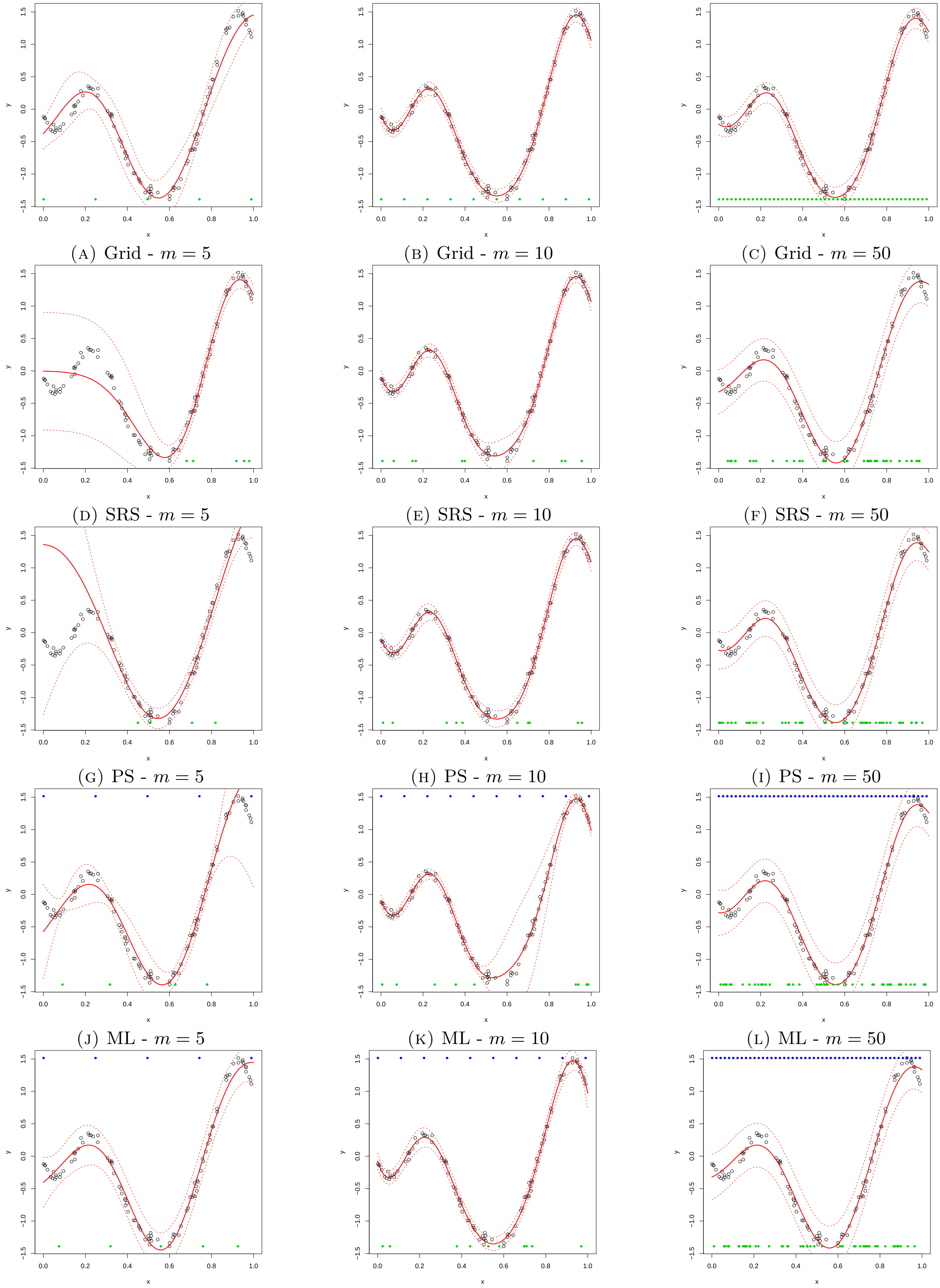
(L) ML - $m = 50$

(M) KL - $m = 5$

(N) KL - $m = 10$

(O) KL - $m = 10$

FIGURE A.1. Posterior predictive mean (solid line) and 95% prediction intervals (dashed lines) for the five considered approaches on the smooth GP realization. The considered pseudo-input locations (for all the methods) are reported as green points at the bottom of the panel. Starting locations for the numerical algorithm (for the ML and KL methods) are reported as blue points at the top of the panel.



(A) Grid - $m = 5$

(B) Grid - $m = 10$

(C) Grid - $m = 50$

(D) SRS - $m = 5$

(E) SRS - $m = 10$

(F) SRS - $m = 50$

(G) PS - $m = 5$

(H) PS - $m = 10$

(I) PS - $m = 50$

(J) ML - $m = 5$

(K) ML - $m = 10$

(L) ML - $m = 50$

(M) KL - $m = 5$

(N) KL - $m = 10$

(O) KL - $m = 10$

FIGURE A.2. Posterior predictive mean (solid line) and 95% prediction intervals (dashed lines) for the five considered approaches on the wiggly GP realization. The considered pseudo-input locations (for all the methods) are reported as green points at the bottom of the panel. Starting locations for the numerical algorithm (for the ML and KL methods) are reported as blue points at the top of the panel.

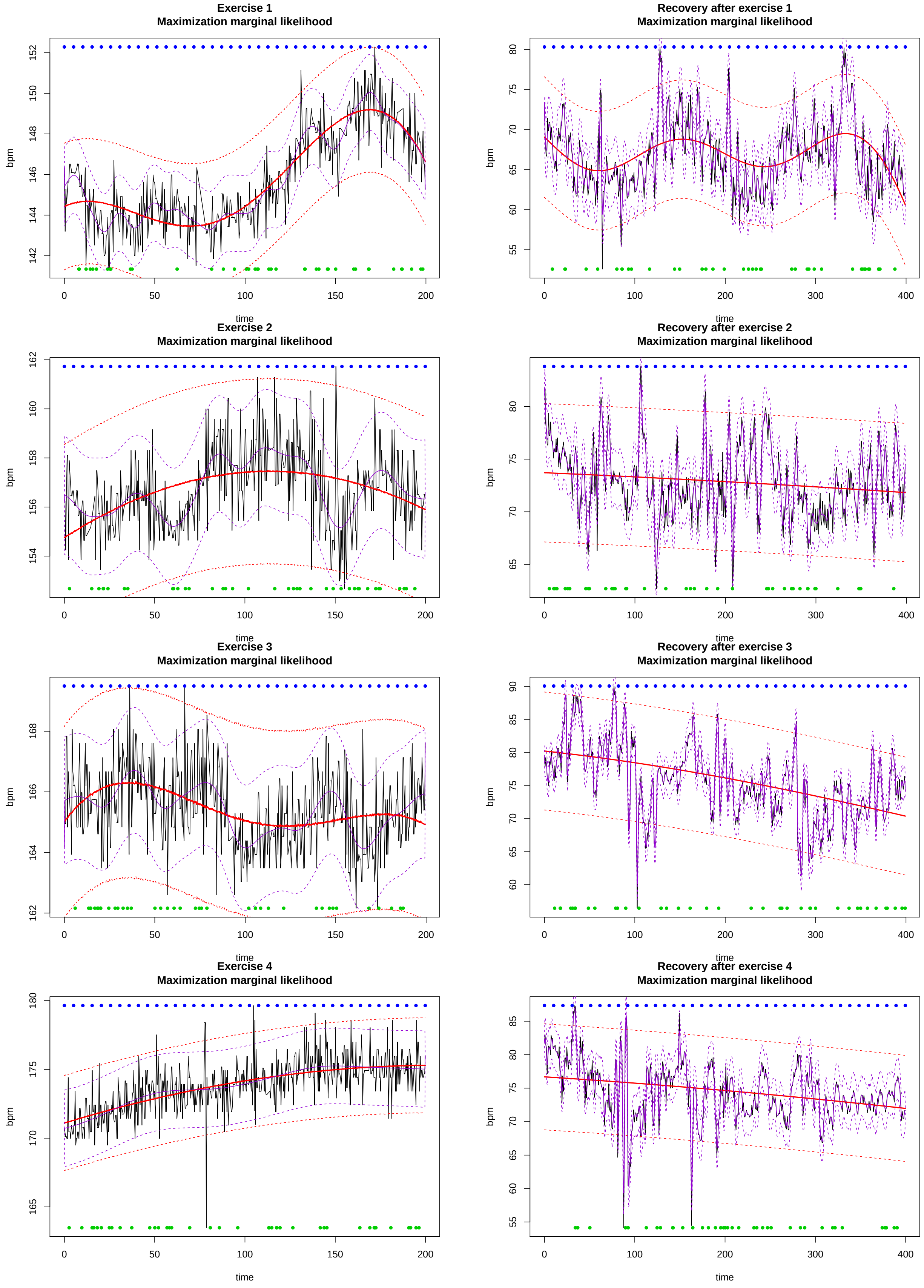


FIGURE A.3. Posterior predictive means (solid line) and 95% prediction intervals (dashed lines) of the pseudo-input GP fit based on the maximization of the marginal likelihood (red lines) and on the standard full-data GP regression (purple lines).

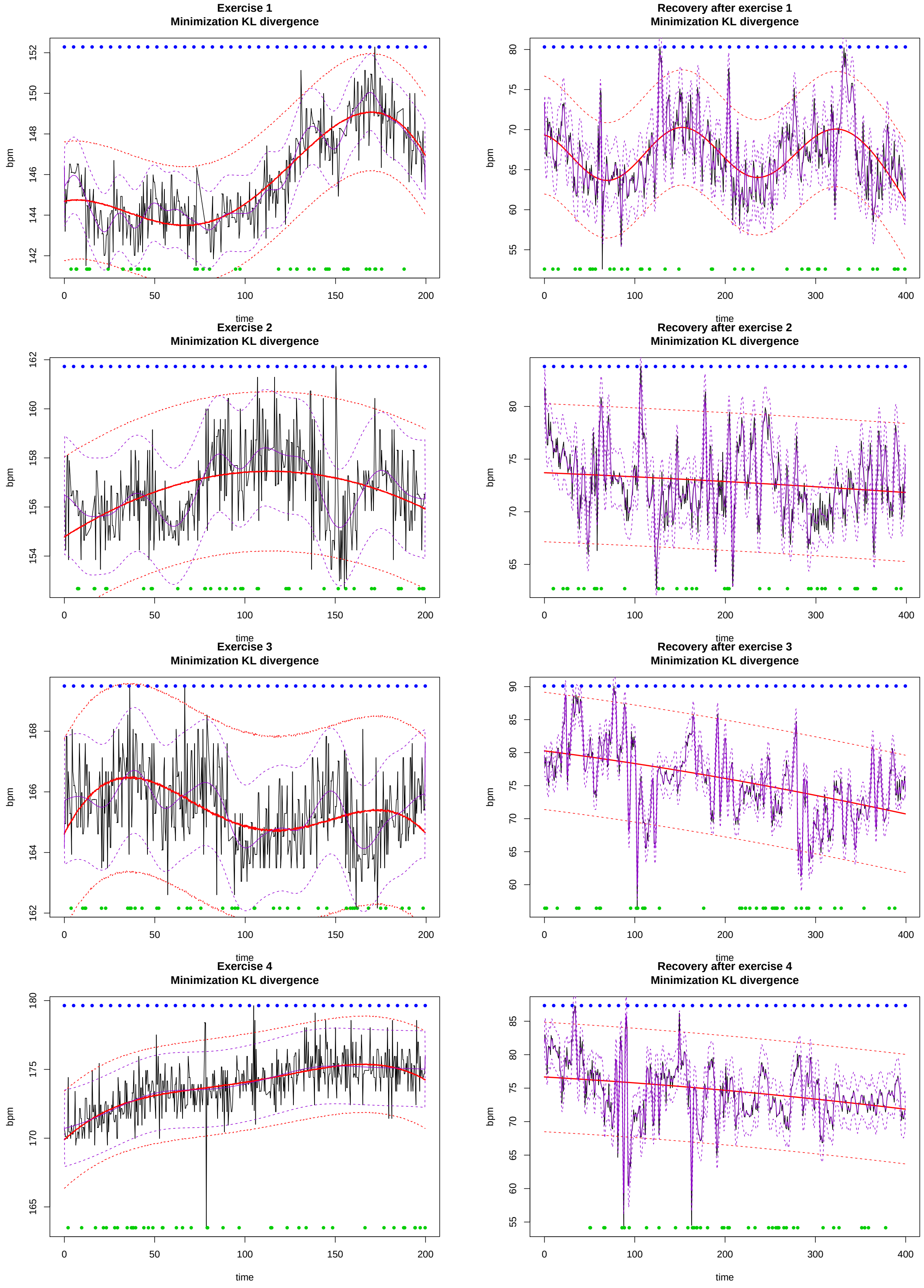


FIGURE A.4. Posterior predictive means (solid line) and 95% prediction intervals (dashed lines) of the pseudo-input GP fit based on the minimization of the KL divergence (red lines) and on the standard full-data GP regression (purple lines).

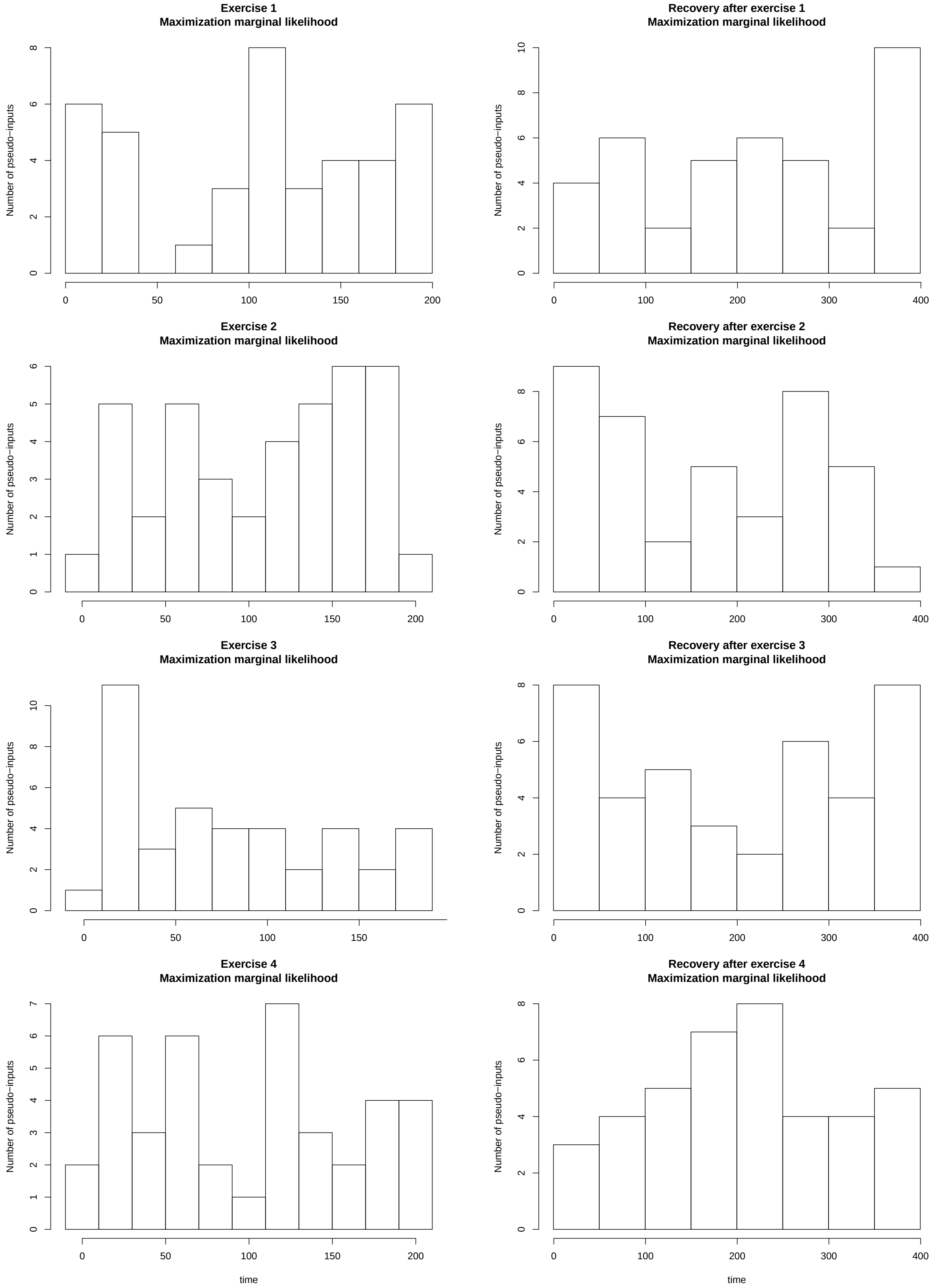
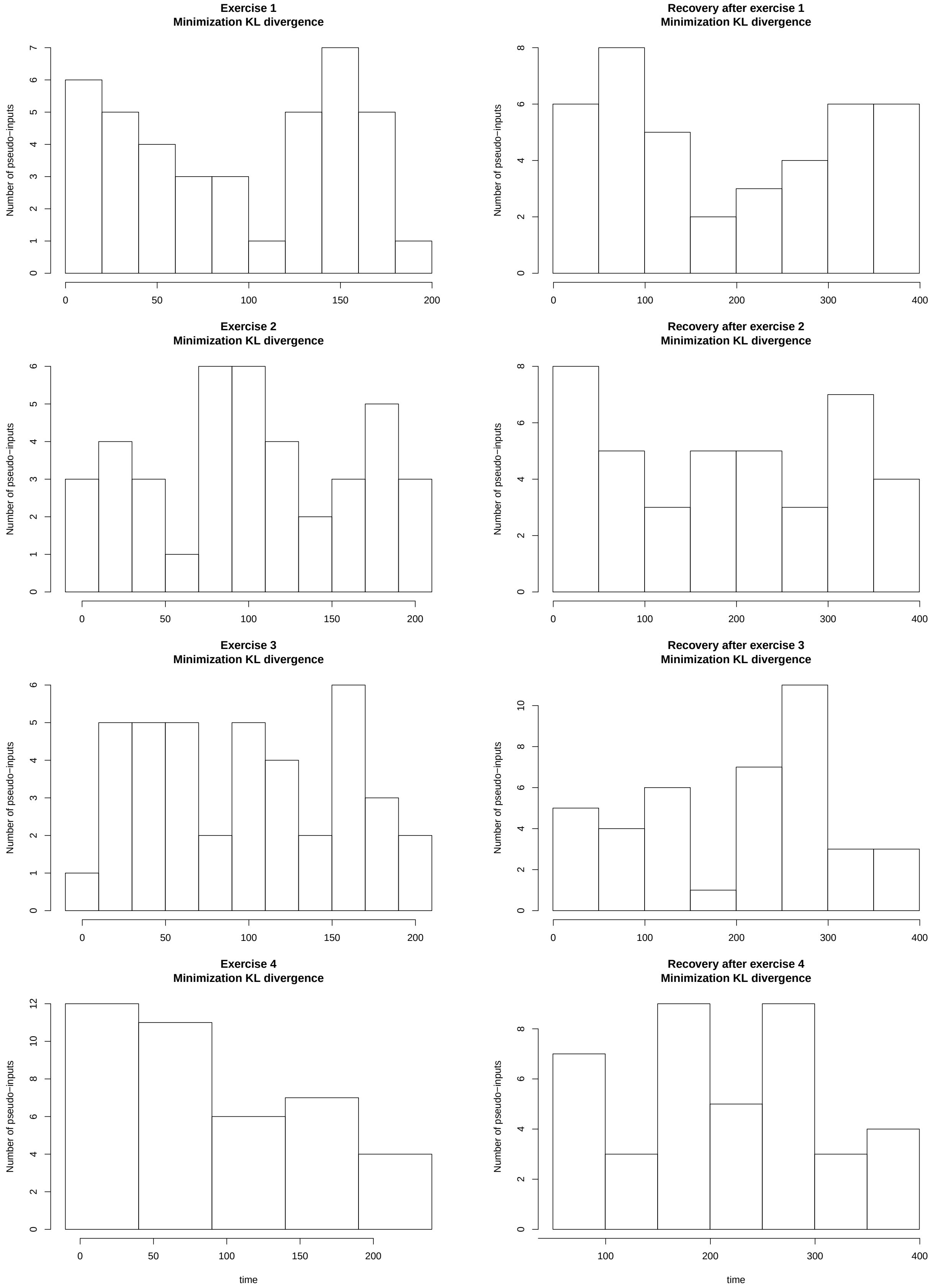


Figure A.5. Distribution of pseudo-input locations for pseudo-input GP fit based on the maximization of the marginal likelihood.



Exercise 1
Minimization KL divergence

Recovery after exercise 1
Minimization KL divergence

Exercise 2
Minimization KL divergence

Recovery after exercise 2
Minimization KL divergence

Exercise 3
Minimization KL divergence

Recovery after exercise 3
Minimization KL divergence

Exercise 4
Minimization KL divergence

Recovery after exercise 4
Minimization KL divergence

FIGURE A.6. Distribution of pseudo-input locations for pseudo-input GP fit based on the minimization of the KL divergence.