Pseudo-Location: A novel predictor for predicting pseudo-temporal gene expression patterns using spatial functional regression

Kyungmin Ahn and Hironobu Fujiwara

RIKEN Center for Biosystems Dynamics Research (BDR), Kobe, Japan

Abstract

Background: It is generally known that analyses of gene expression at single-cell resolution and spatial location of genes will reveal insights into the regulation of gene expression. Although transcriptome-wide single-cell gene expression analyses with genetic spatial information are not yet feasible, many bioinformaticians and biologists are trying to understand and predict gene expression levels using genetic spatial information of genes by implementing statistical regression models.

Methods: We implement a trajectory inference analysis in order to identify the pseudo-temporal gene expression patterns (PTGEPs) for single cell RNA-sequencing (scRNA-seq) data. In here, each PTGEP is obtained as a functional form in which we treat as functional response in spatial functional regression model. For a predictor, we propose a new concept of genetic spatial information, *pseudo-location* by incorporating the chromosome number and molecular starting position of genes. Then we applied and compared several types of functional response regression models to evaluate the prediction performance, including standard function-on-scalar regression model, penalized flexible functional regression model, and Bayes function-on-scalar regression model. We calculate and compared the squared correlation, R_i^2, for each gene under each model to evaluate the prediction performance.

Results: We applied these spatial functional regression model and other function-on-scalar regression models to scRNA-seq data sets including Trapnell2014, Kumar2014, and Shekhar2016. To see the robustness of the prediction performance, we also increase the total number of genes for regression analysis from 1,000 to 3,000, and 10,000. Almost 30% of genes can be predicted for each model and for each data set . By comparing the predictive genes ($R_i^2 > 0.1$) and higher predictive genes ($R_i^2 > 0.3$), we found that spatial functional regression model using pseudo-location is able to provide the better prediction performance compared to other standard functional regression models.

keywords : spatial functional regression, function-on-scalar regression, trajectory inference analysis

1 Introduction

Trajectory inference (TI), also known as a *pseudotime* analysis, in a single-cell RNA-sequencing (scRNA-seq) data analysis has been widely used to explore a cellular dynamic processes such as cell differentiation, cell cycle, and cell activation by computationally ordering cells along the trajectory based on similarities in their expression patterns. Many TI methods have been published and publicly available in computing software such as R packages (e.g. Monocle [31], TSCAN [15], SLICER [34], slingshot [29], SCUBA [17], and many). The trajectory results of these methods can be linear, bifurcating, or a more complex tree-shape or graph structure. In this sense, TI methods allow the objective identification of new subsets of cells, inference of regulatory interaction, and delineation of differentiation tree by giving an unbiased and trascriptome-wide understanding of a stochastic dynamic process. This TI methods can be expanded to differential expression (DE) analysis by comparing pseudo-temporal genes to understand the differentially expressed between groups of cells and assesses the statistical significance of those changes [15, 31, 33]. Since the number of TI methods are published, there are also some articles which compare existing TI methods in several different ways. Cannoodt et al. [2] compared 10 TI techniques for several gene expression datasets highlighting several practical advantages and disadvantages of the individual method. Saelens et al. [26] extended these initial studies to 45 TI algorithms on a total of 110 real and 229 synthetic datasets and developed a set of guidelines to help users select the best method for their dataset.

One of the interesting studies in gene expression data analysis is using genotype data to perform accurate genetic prediction of complex characteristic of genes which can also facilitate genetic network system. This study has been proposed as a critical step for integrating genomic sequencing studies with gene regulatory network (GRN) which can build more powerful and interpretable study in genome-wide association studies (GWAS) [36, 9]. Thus, prediction analysis using gene expression data is a central task in systems biology. Zeng et al. [36] developed a latent Dirichlet process regression which allows flexibility on the a priori effect size distribution and to enable robust phenotype prediction performance across a range of phenotypes. Zhong et al. [38] recently implemented three simple regression models, simple, multiple, and LASSO, concluding that the combination of LASSO regression and all probes (group of atoms or molecules) on the methylation array provides the best prediction for gene expression.

Another interest in prediction analysis is how a topological or spatial location of genes can affect the biological process. It is known that analyses of gene expression at single-cell resolution and genetic spatial information will reveal insights into the regulation of gene expression. Although transcriptome-wide single-cell gene expression analyses with genetic spatial information are not yet feasible, many bioinformaticians and biologists are trying to understand and predict gene expression levels using genetic spatial information of genes based on statistical regression methods. Wilczynski et al. [35] performed a Bayesian approach to model diverse experimental regulatory data which leads to accurate predictions of both spatial and temporal aspects of gene expression. Zeng et al. [37] applied several regression models including linear mixed model (LMM), sparse models, and the hybrid of LMM and showed that gene expression can be predicted with only cis-SNPs using well-developed prediction models. de Luis Balaguer [5] developed a computational pipeline that uses spatial and temporal transcriptomic data to predict interactions among the genes involved in stem cell regulation. All of these articles demonstrated that only a few variables can infer the prediction of gene expression level by incorporating the statistical regression models from *multivariate data analysis* (MDA). Since the expression levels of gene expression data are the vectors of discrete samples, application of statistical methods from MDA would provide an optimal prediction analysis.

In this paper, we are also interested in regression problem using only genetic spatial information of genes for predicting pseudo-temporal gene expression patterns (PTGEPs) rather than discrete gene expression levels. The difference between our interests and previous articles is that each PTGEP is obtained as a functional form after performing trajectory inference methods. Hence, applying *functional data analysis* (FDA) to gene expression data over MDA in predicting PTGEPs is more robust on functional regression problem. When the data observation is a functional from, FDA has many advantages of generating models that can be described by continuous smooth dynamics: 1) capability for the accurate estimates of parameters, 2) smoothing techniques for the effective data noise reduction of the curves, and 3) allow us to analyze irregular time sampling schedules [32]. Moreover, these FDA stochastic methods for forecasting functional data have also major advantages over the standard approaches for better understanding trends. Thus, application of FDA for functional observations is a crucial step for the statistical analysis. Hence, we applied and compared several types of functional regression models to evaluate the prediction performance, including spatial functional regression model (SFR) [6, 10], standard function-on-scalar regression model (FOSR) [24], penalized flexible functional regression model (PFFR) [11], and Bayes function-on-scalar regression model (BFOSR)[12, 13].

This paper is organized as follows. In section 2, we briefly describe the the most recent TI method that we have used for in this paper, *slingshot* [29], and functional response regression models from FDA. We introduce SFR, FOSR, PFFR, and BFOSR. SFR can be implemented using okfd in geofd package from R software. Other function-on-scalar regression models, FOSR, PFFR, and BFOSR are implemented using pffr and bayes_fosr in refund package from R. In section 3, we show the results of application to FDA regression models for several scRNA-seq data sets. We compare the prediction performances by calculating the squared correlation R_i^2 of the predicted responses. In section 4, we discuss about our results with conclusion and give a potential further analysis on this approach. The program implemented in this paper as R package is available as predPTGEP and all the results can be reproducible by using these R codes.

2 Methods

We are interested in prediction performance of FDA regression models when we only consider the genetic spatial information of genes and PTGEPs. In here PTGEPs are treated as functional responses and the genetic spatial information is treated as scalar predictor. For the genetic spatial information, we use chromosome number and molecular starting position of each gene. In here we use starting position of gene since it contains several key regulatory elements of gene transcription, such as TATA-box and transcription factor binding sites. Hence, these two variables can be reparametrized as paired samples for spatial functional regression problem. i.e., the pair of chromosome number (x_1) and molecular starting position (x_2) from each gene is the pair of the form $\mathbf{X} = (x_1, x_2)$. Generally, each chromosome is not connected to each other but we treat them in two dimensional space and we call this genetic spatial information as *pseudo-*

location of gene for scRNA-seq gene expression data (Fig. 1). The genetic spatial information regarding chromosome number and molecular starting position can be obtained from *Genome Browser* and *Ensembl*. In this paper, we use `biomaRt` `R` package to collect the genetic spatial information for the prediction analysis. In fact, human and mouse have 22 pairs and 19 pairs of chromosome (autosomes), respectively. Then each species has one pair of sex chromosomes combining of X and Y (XY for male and XX for female). Hence, we assign 23 for X (20 for mouse) and 24 for Y (21 for mouse) to treat these as independent chromosome numbers for the downstream analysis. Additionally, both species have a Mitochondrial (MT) chromosome, so we assign 25 for MT (22 for Mouse). For example, Fig. 1 shows the pseudo-location for **PadovanMerhar2015** data. The organism of this data is *Homo Sapiens* which contains 22 pairs of autosomes and 1 pair of sex chromosomes. Hence, for statistical analysis, chromosome X, Y, and MT are now denoted as 23, 24, and 25, respectively.

2.1 Trajectory Inference Analysis - Slingshot

Trajectory inference analysis allows us to explore a cellular dynamic process such as the cell differentiation by computationally ordering cells along the trajectory based on the similarities in their gene expression patterns. There are many TI methods which are implemented and publicly available as `R` packages or scripts. In this paper, we have applied one of the most current TI methods, *Slingshot* [29] for identifying the cell lineages and pseudotimes from single-cell gene expression data. Slingshot comprises two main steps: (1) The inference of the global lineage structure using a minimum spanning tree (MST) on the clusters identified by Resampling-based Sequential Ensemble Clustering (RSEC) and (2) the inference of cell pseudotime variables along each lineage using a novel method of simultaneous principal curves. The approach in (1) allows the identification of any number of novel lineages, while also accommodating the use of domain-specific knowledge to supervise parts of the tree; the approach in (2) yields robust pseudotimes for smooth, branching lineages [29]. For fitting problem on relative gene expression level, Gaussian additive model is used.

2.2 Functional Response Regression Model

Ferraty & Vieu [8] define a functional random variable Y in an infinite dimensional space (or functional space $\mathcal{F}$). A functional data set $y_1, \ldots, y_n$ is the observation of n functional variables $Y_1, \ldots, Y_n$ distributed as Y whose realizations (or values) are functions defined on $T = [a, b] \subseteq \mathbb{R}$ which belong to

$$\mathbb{L}^2(T) = \left\{ Y : T \to \mathbb{R}, \text{such that } \int_T Y(t)^2 \, dt < \infty \right\}$$

One assumes here that $\mathbb{L}^2(T)$ is in Hilbert space to allow for this inner product between its elements:

$$\langle Y_1, Y_2 \rangle = \int_T Y_1(t) Y_2(t) \, dt$$

A fast growing subtopic in functional data analysis (FDA) [24] is *regression* involving functional variables, either as predictors or responses or both. Morris [19] categorizes regres-

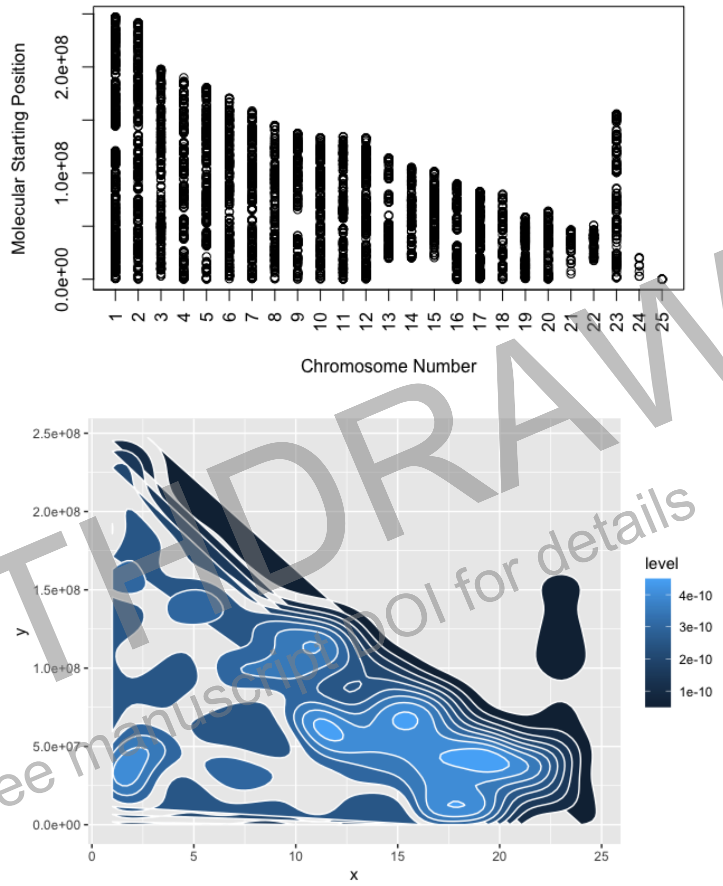


Figure 1: Example of *pseudo-location* for gene expression data (**PadovanMerhar2015**). Genes are located in a chromosome and each chromosome is not connected to each other. Hence, we integrate this genetic spatial information - chromosome number and molecular starting position and project to 2-dimensional space in order to locate each gene. Top panel shows the standard scatter plot between chromosome number and molecular starting position and bottom panel shows the area within a contour plot based on top panel.

sion problems involving functional data into three types: (1) functional predictor regression (scalar-on-function), (2) functional response regression (function-on-scalar) and (3) function-on-function regression. The functional response regression problem (or function-on-scalar) model was first studied by Ramsay [23], Cardot et al. [3], and several other since then. Since the PTEGEPs are functions and the predictors - genetic spatial information are scalar variables, we focus on the functional response regression problem in this paper.

2.2.1 Function-on-Scalar Regression Model

Here the responses are functions over a fixed interval T, call them $\{Y_i \in \mathcal{F}\}$, elements of some pre-specified functional space $\mathcal{F}$, and the predictors are scalar random variables $\{X_i \in \mathbb{R}\}$. A simple and commonly-used model for this problem is the so-called *standard function-on-scalar regression model* given by:

$$Y_i(t) = \sum_{j=1}^{2} X_{ij}\beta_j(t) + \epsilon_i(t), \;\; i = 1, \ldots, n \;, \tag{1}$$

where $Y_i(t)$ is a PTGEP, a sample of functional response, chromosome number and molecular starting position as scalar predictors X_{ij}, $\beta_j(t) \in \mathcal{F}$ is the regression-coefficient function, and $\epsilon_i(t)$ is the residual error assumed to be iid mean zero Gaussians with covariance whose structure describes the within-function covariance. Similar to the linear regression models with Euclidean variables, one can also estimate model parameters here by minimizing the sum of squared errors (SSE):

$$\{\hat{\beta}\} = \underset{\beta \in \mathbb{L}^2}{\operatorname{argmin}} \left[\sum_{i=1}^{n} \int \left(Y_i(t) - \sum_{j=1}^{2} X_{ij}\beta_j(t) \right)^2 dt \right] . \tag{2}$$

To avoid the overfitting and induce smoothness in functional effect of the regression analysis, a regularized term can be added to the Eqn. 2 for a penalization [24].

2.2.2 Penalized Flexible Functional Regression Model

Penalized flexible functional regression or penalized function-on-function regression (PFFR) was first introduced by Goldsmith et al. [11] and the code is publicly available from `R` package `refund`. This regression model can handle both functional and scalar predictor when the responses are functional form. In here we focus on scalar covariates that we only include the scalar terms of predictors in the model

$$Y_i(t) = X_{1i}\beta_1(t) + f(X_{2i}, t) + \epsilon_i(t), i = 1, \ldots, n, \tag{3}$$

where X_{1i} is an effect of chromosome number that are constant over t of $Y(t)$ and $f(X_{2i}, t)$ is nonlinear effect of molecular starting position that is vary smoothly over the index t. This would allow us to investigate the nonlinear relationship between the scalar predictor and functional responses.

2.2.3 Bayesian Functional Response Regression Model

Goldsmith et al. [12, 13] developed a Bayesian framework for penalized spline function which allows the joint modeling of population-level fixed effects, subject-level random effects and residual covariance. This algorithm enables model selection and comparison, which for large datasets is infeasible with competing approaches. The equation of the model is the same as above in Eqn. 3 but Bayes approach is used for the parameter estimation and penalization is used to avoid the overfitting and induce smoothness in functional effects. More details are in [13, 12].

2.2.4 Spatial Functional Regression Model

We define a *spatial* functional process [6] as $\{Y_s(t) : s \in D \subseteq \mathbb{R}^2, t \in T \subseteq \mathbb{R}\}$, such that $Y_s(t)$ is a functional variable for any $s \in D$. Then let $s_1, \ldots, s_n$ be arbitrary pseudo-location in D and assume that we can observe a realization of the functional random process $Y_s(t)$ at these n pseudo-location, $Y_{s_1}(t), \ldots, x_{s_n}(t)$. Then we use ordinary functional kriging predictor (OKFD) [10] for predicting $X_{s_0}(t)$, the functional random process at s_0, where s_0 is a unsampled pseudo-location. The OKFD predictor is defined as follows:

$$\hat{Y}_{s_0}(t) = \sum_{i=1}^{n} \lambda_i Y_{s_i}(t),\ \lambda_1, \ldots, \lambda_n \in \mathbb{R} \tag{4}$$

where λ are such that $\mathbb{E}[\hat{Y}_{s_0} - Y_{s_0}] = 0$. Then the Best Linear Unbiased Predictor (BLUP) of n variables at an unsampled location s_0 can be obtained is obtained by minimizing [10]:

$$\begin{aligned} \sigma^2_{s_0} &= \int_T V(\hat{Y}_{s_0}(t) - Y_{s_0}(t))\, dt \\ &= \sum_{i=1}^{n} \lambda_i \int_T \gamma||s_i - s_0||(t)\, dt - \mu, \text{ s. t. } \sum_{i=1}^{n} \lambda_i = 1 \end{aligned}$$

In here, it is important to assume that the functional random process is second-order stationary and isotropic: the mean and variance functions are constant and the covariance depends only on the distance between sampling points, where we define as follows:

- $E(Y_s(t)) = m(t)$ and $V(Y_s(t)) = \sigma^2(t)$ for all $s \in D$ and all $t \in T$.
- $COV(Y_{s_i}(t) - Y_{s_j}(t)) = C(||s_i - s_j||)(t) = C_{ij}(h,t), s_i, s_j \in D, t \in T$, where $h = ||s_i - s_j||$.
- $\frac{1}{2}V(Y_{s_i}(t) - Y_{s_j}(t)) = \gamma(||s_i - s_j||)(t) = \gamma(h,t), s_i, s_j \in D, t \in T$, where $h = ||s_i - s_j||$.

2.3 Squared Correlation, R_i^2

For comparison of the models, we calculate the squared correlation R_i^2, which is also called as a coefficient of determination. When evaluating the goodness-of-fit in FDA rather than MDA, since we define the function observations in Hilbert space, R_i^2 of each gene can be calculated with inner product in functional space as:

$$R_i^2 = 1 - \frac{\int_T \{Y_i(t) - \hat{Y}_i(t)\}^2\, dt}{\int_T \{Y_i(t) - \bar{Y}(t)\}^2\, dt}$$

where $Y_i(t)$ is the observed PTGEP ($i = 1, \ldots, n$), $\hat{Y}_i(t)$ is the predicted PTGEP, and $\bar{Y}(t)$ is the mean of all PTGEPs. We calculate the squared correlation for each gene and then compare the prediction performances between the models for each data.

3 Results

3.1 Data Description

For the experiment, three scRNA-seq data sets were collected from *conquer* [28] and applied to our prediction evaluations: GSE60749-GPL13112 (here denoted **Kumar2014**) [16], GSE66053-GPL 18573 (**PadovanMerhar2015**) [21], and GSE81903 (**Shekhar2016**) [27]. The brief descriptions of each data set, including the number of cells and genes, are shown in Table 1. For example, **Kumar2014**, **PadovanMerhar2015**, and **Shkhar2016** have 45,686, 65,218, and 45,686 genes, respectively. Detailed information can be found in Table 1.

Data set	Organism	Sequencing Protocol	Cells	Genes	Ref.
Kumar2014	Mus musculus	SMARTerC1	268	45686	[16]
PadovanMerhar2015	Homo Sapiens	Fluidigm C1 Auto Prep	96	65218	[21]
Shekhar2016	Mus Musculus	Smart-Seq2	384	45686	[27]

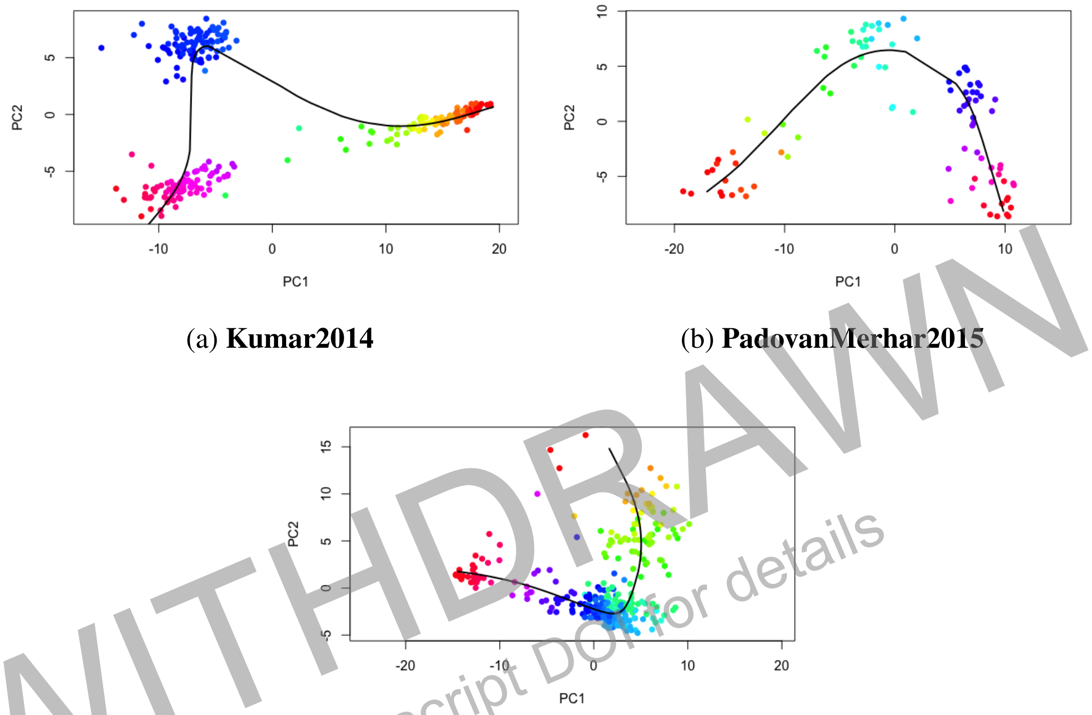
Table 1: Description of scRNA-seq data

For the preprocessing steps, we implemented the fundamental steps for scRNA-seq data analysis such as normalization, scaling, and identification of the highly variable genes using `Seurat R` package [30, 1] to perform the downstream analysis. Then the only predetermined highly variable genes are used for a dimensional reduction [30]. For the methodology of dimensional reduction, we use a principal component analysis (PCA), which is the most popular linear dimensionality reduction algorithm for analyzing gene expression data. Then we visualize the scree plot and perform a jackstraw [4] to determine the optimal number of PCs to reduce the dimensionality of the original data. A scree plot in PCA is a useful tool that visualizes saturation in the relationship between the number of PCs and the percentage of the variance explained. We generally decide the number of principal components that corresponds to the "elbow" part of the curve to have sufficient information of the original data. In this experiment, we chose from PC1 to PC20 for all real data sets for the downstream analysis. After reducing the dimension of the gene expression matrix, we can also cluster the cells using `Seurat R` package and this enables us to simplify the complicated lineages into simple trajectory. In this paper, however, we did not cluster the cell population but focus on all cells in each data set.

3.2 Trajectory Inference

Based on the preprocessed data from `Seurat`, we perform a `Slingshot` [29] to identify the trajectory of cells for each data set. We visualize the distribution or pattern of cell populations by plotting PC1 vs. PC2 and the trajectory for each data set based on TI technique. The plots are shown in Figure 2. Each curve presents the trajectory for each data set, **Kumar2014**, **PadovanMerhar2015**, and **Shekhar2016**, respectively. We also visualize the inferred lineage for the single-trajectory data with points colored by pseudotime.

It is generally known that scRNA-seq gene expression data has may zero values or zero-closed values. Hence, it is important to remove unexpressed genes for downstream analysis.



(a) **Kumar2014**

(b) **PadovanMerhar2015**

(c) **Shekhar2016**

Figure 2: PC1 vs. PC2 plots after performing Principal Component Analysis for each scRNA-seq data set. Different color represents the subpopulation by implementing clustering algorithms from `Seurat` `R` package. The inferred lineage for the single-trajectory data with points are colored by pseudotime from `Slingshot` `R` package.

Thus, we look at the most variable genes by ranking all genes by their variances. Moreover, to see the robustness of the statistical experiment, we arbitrary choose the number of the largest variance of genes for the analysis. We first start with 1,000 the most variable (significant) genes (Notationally, we use $N_s = 1,000$) and then increase to 3,000 and 10,000 genes to see the robustness of the prediction performance.

After choosing the number of significant genes, we identify temporally expressed genes with the most significant time-dependent model fit i.e., those genes whose expression is changing in a continuous manner over pseudotime. To do this, we fit a generalized additive model (GAM) with a locally estimated scatterplot smoothing (LOESS) term for pseudotime. [14]. Then we calculate R_i^2 and p-value for each gene to identify the goodness-of-fit. Since some of regression curves are not fitted well to the genes, total number of significant genes (N_s) is reduced.

3.3 Prediction Performance

After performing TI algorithm, PTGEPs are obtained as function forms. We treat these observations as functional responses in regression model. Moreover, genetic spatial information, which we consider as scalar predictor, can be collected from `R` package, `biomaRt`. We import chromosome number and molecular starting position of each gene to assign the 2-dimensional coordinate of pseudo-location. Since some of genes have no information and some chromosome is not considered as this analysis, the total number of genes for regression analysis is decreased again. We use this final total number of genes (denote as N_g) for prediction analysis. For example, Table 2, 3, and 4 list the total number of genes (N_g for prediction analysis when we choose N_s = 1,000, 3,000, and 10,000, respectively. It is clear to see that number of genes are reduced due to the fitting issue and missing information of genes. In particular, 97, 94, and 14 genes are removed after several steps for **Kumar 2014**, **PadovanMerhar2015**, and **Shekhar2016** in Table 2, Table 3, and Table 4, respectively.

Then we applied several function-on-scalar regression models to compare the prediction performances. The goal in here is how genetic spatial information can predict the PTGEPs well if only the chromosome numbers and molecular starting positions of genes are considered as pseudo-location. In specific, other three function-on-scalar regression models do not assume the predictors as spatial variable. Therefore, we treat chromosome name as a categorical variable and molecular starting position as numerical variable for these regression models. Moreover, to avoid the multicollinearity issue, we normalize the numerical variable for these three regression models. Finally, we randomly split 80% of the total number of genes, N_g, for training set and rest of 20% for test set to perform a regression analysis.

We first compare the performance of predictive genes among four models. In here we define the predictive genes where R_i^2 are greater than 0.1 ($R_i^2 > 0.1$, $i = 1, 2, \cdots, n$). It is generally known that number of predictive genes is very small for prediction analysis using gene expression data especially when we use only few variables [36, 37, 38]. Hence, for the comparison between the models, we focus on the predictive genes for each model. The prediction performances for each total number of genes, $N_s = 1,000$, $N_s = 3,000$, and $N_s = 10,000$ are shown in Table 2, Table 3, and Table 4, respectively.

It is clear to see that the proportion of predictive genes for each model is not powerful. i.e., in Table 2, the highest proportion of predictive genes for all models in each data set (**Kumar2014**, **PadovanMerhar2015**, and **Shekhar2016**) are 28% (51/181), 26% (48/182), and 33% (65/198), respectively. Moreover, increase of total number of genes for prediction analysis also does not increase the proportion of predictive genes. For example, when we increase the number of genes to 3,000, the proportions for each data set become 28% (148/537), 26% (142/548), and 32% (190/591) in Table 3. For $N_s = 10,000$, the proportions for each data set are 29% (513/1744), 33% (469/1407), and 33% (603/1813) in Table 4. Hence, it is unclear to see the different prediction performances or patterns between each sample size.

We are also interested in the quality of predictive genes among four models. First, we compare the higher predictive genes ($R_i^2 > 0.3$). All tables (Table 2, Table 3, and Table 4) show that SFR has the large number of higher predictive genes among four models for each data set. Moreover, we display the box plots and the histograms of predictive genes for each model for $N_s = 10,000$ and they are shown in Fig. 3 and Fig. 4, respectively. Other box plots and histograms for $N_s = 1,000$ and $N_s = 3,000$ are shown in Appendix A. Each panel shows four box plots for four functional regression models of each dataset. In here SFR has the largest median of predictive genes for each data set. The other three regression models show similar prediction performances in which the median of predictive gene is between 0.1 and 0.2. This indicates that SFR using pseudo-location is able to provide the ultimate prediction performances compared to the other functional regression models. On the other hands, the predictions of other

	Threshold	SFR	FOSR	PFFR	BFOSR	N_g (test)
Kumar2014	0.1	51	22	43	44	903 (181)
	0.2	46	5	22	16	
	0.3	41	2	9	9	
PadovanMerhar2015	0.1	48	18	45	29	906 (182)
	0.2	44	3	26	16	
	0.3	33	1	14	10	
Shekhar2016	0.1	49	13	65	44	986 (198)
	0.2	44	2	48	28	
	0.3	40	1	32	15	

Table 2: Number of predictive genes ($R_i^2 > 0.1$) for each threshold when N_s = 1,000 for three scRNA-seq data sets.

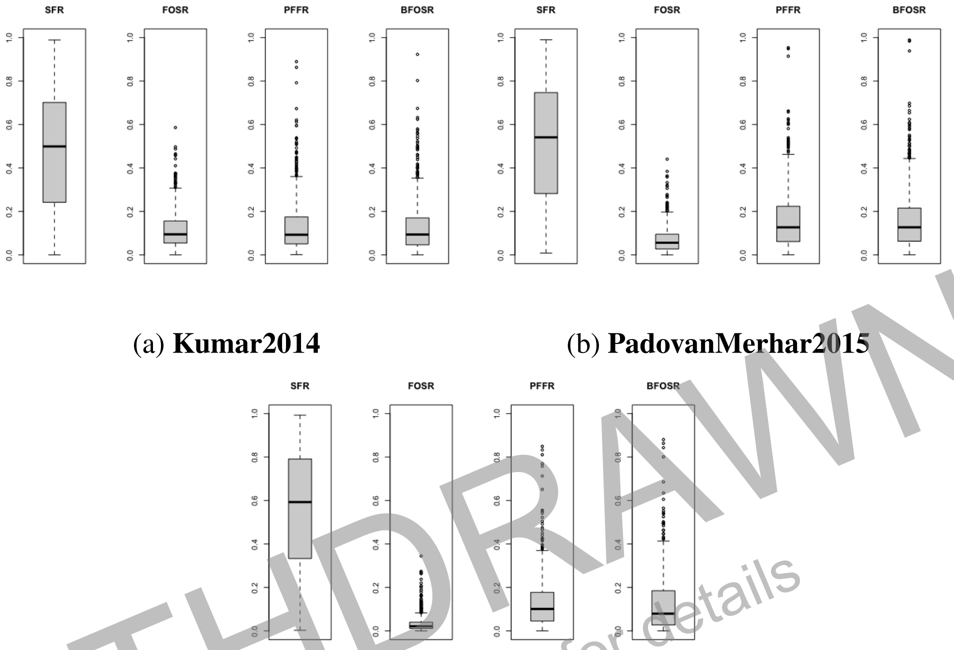
	Threshold	SFR	FOSR	PFFR	BFOSR	N_g (test)
Kumar2014	0.1	135	126	148	125	2681 (537)
	0.2	110	58	80	68	
	0.3	90	25	49	35	
PadovanMerhar2015	0.1	142	34	120	86	2739 (548)
	0.2	124	9	45	42	
	0.3	100	1	15	17	
Shekhar2016	0.1	190	22	135	150	2955 (591)
	0.2	173	5	70	65	
	0.3	150	0	31	31	

Table 3: Number of predictive genes ($R_i^2 > 0.1$) for each threshold when N_s = 3,000 for three scRNA-seq data sets.

	Threshold	SFR	FOSR	PFFR	BFOSR	N_g (test)
Kumar2014	0.1	513	353	373	373	8719 (1744)
	0.2	448	109	165	158	
	0.3	401	29	76	70	
PadovanMerhar2015	0.1	469	156	440	424	7031 (1407)
	0.2	414	25	222	207	
	0.3	370	8	116	109	
Shekhar2016	0.1	603	47	439	398	9063 (1813)
	0.2	544	9	175	212	
	0.3	503	1	70	82	

Table 4: Number of predictive genes ($R_i^2 > 0.1$) for each threshold when N_s = 10,000 for three scRNA-seq data sets.

regression models generally fall around the mean of the curves which implies that the quality of prediction performances is very poor compared to SFR.



(a) **Kumar2014**

(b) **PadovanMerhar2015**

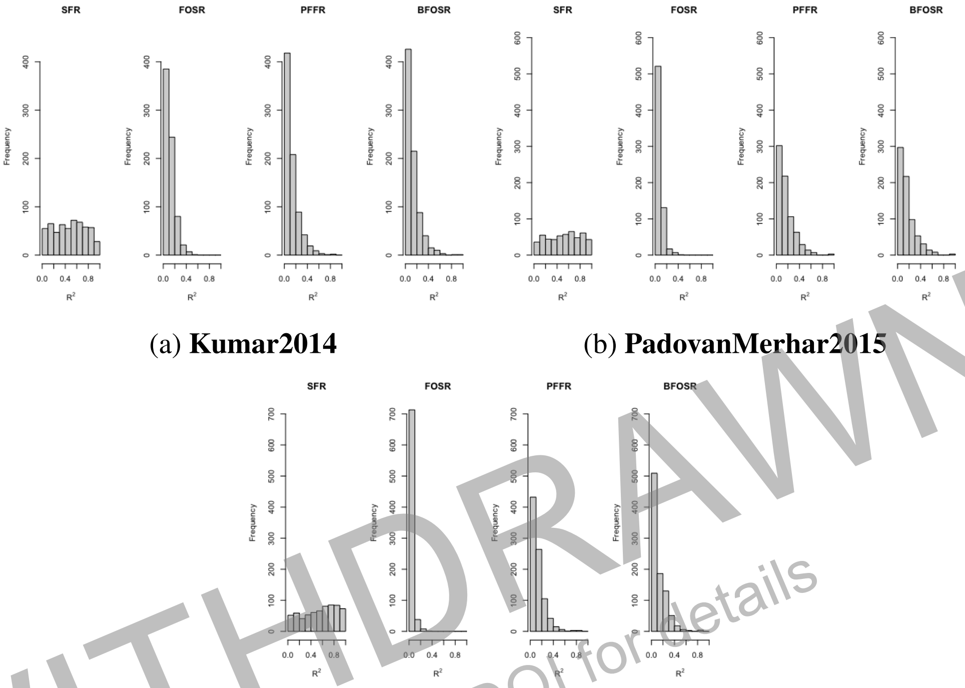
(c) **Shekhar2016**

Figure 3: Box plots for each data set for four functional regression models when $N_s = 10,000$. The prediction performance of SFR using pseudo-location is superior to other functional response regression models.

Gene Ontology analysis of predictive genes We are interested in the potential biological function of the predictive genes ($R_i^2 > 0.1$). To examine the further analysis of these predictive genes, we conducted gene ontology (GO) knowledge base using PANTHER [18] to acquire information of the functions of genes. In this analysis, we use predictive genes from SFR with $N_s = 10,000$ in order to retrieve enough number of genes. At false discovery rate (FDR) of 0.01 or p-value of 0.01, sensory perception of chemical stimulus is among the most significant enriched terms in the **Kumar2014** data, which is also found in [22]. For **Shekhar2016** data, several biological functions are found including neuron development, regulation of gene expression, apoptotic process, and many which are described in [27]. For **PadovanMerhar2015** data, mRNA metabolic process and cell cycle process are found which are also reported in [21].

4 Conclusion and Discussion

We have introduced the new predictor, *pseudo-location*, for defining the location of gene in two-dimensional space using their chromosome number and molecular starting position of genes. Many biologists and bioinformaticians have been developing 3D structure of chromosome, using 3 C-based methods - particularly Hi-C method. Based on these methods, they develop the image statistical analysis by applying the machine learning tools such as regression or classification techniques [25, 7]. Despite the improvement in 3-D structure modeling approaches, the



(a) **Kumar2014**

(b) **PadovanMerhar2015**

(c) **Shekhar2016**

Figure 4: Histograms for each data set for four functional regression models when $N_s = 10,000$. The prediction performance of SFR using pseudo-location is superior to other functional response regression models.

lack of a real structure with which to contrast these models remains a challenge [20]. In this paper, we use the conventionally known quantities from *Ensembl* or *Genome Browser* which does not require the 3D structure of the chromosome and therefore, applying these spatial variables to predict the PTGEPs is simpler and preventing the biased assumption of the chromosome structure. We have demonstrated that the spatial functional regression is able to provide the better prediction performances compared to other functional regression models in which the models that consider the variable as non-spatial values.

In this paper, our analysis is focused on FDA methods rather than MDA since the PTGEPs are the functional forms. Using FDA techniques allow us to measure forecast uncertainty through the estimation of prediction intervals for future data [32] over the standard approach. We have applied three function-on-scalar regression models to compare with spatial function regression model. In FDA, only few function-on-scalar regression methods are published and thus available on R [19]. There are some other functional response regression models but many methods are developing the better estimation of parameters using different techniques rather than different assumption of the models. Hence, we did not include other functional regression models and just present standard one and Bayesian approach for different parameter estimation approach. These two general approaches and algorithms are enough when if the model does not assume the spatial characteristic of the samples.

All of our application to scRNA-seq data is linear case of lineage, but it can be extended to nonlinear case, such as a bifurcating lineage. This can be implemented by applying clustering algorithm to separate the different branches of the trees and fit the function or curve for each gene over pseudotime for each cluster. Clustering algorithm based on `Seurat` `R` package can be used as optional function from our `R` package, `predPTGEP`. Some of the limitations in our approach is that pseudo-location has very limited information since it ignores any topological characteristic of the genes and chromosomes such as Query Hi-C interactions.

We emphasize that the prediction accuracy of these models are still low for most of genes (i.e., proportions of the predictive genes are approximately 30 %), although we discover some gene expression patterns can be effectively predicted by pseudo-location for three real scRNA-seq data. There may be other factors that are also responsible for PTGEPs, such as some other biological factors and interaction between these factors . Nevertheless, we have demonstrated that using only pseudo-location for regression analysis incorporating with spatial functional regression model behave the best prediction performance compared to other existing functional regression models.

References

[1] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411, 2018.

[2] R. Cannoodt, W. Saelens, and Y. Saeys. Computational methods for trajectory inference from single-cell transcriptomics. *European journal of immunology*, 46(11):2496–2506, 2016.

[3] H. Cardot, F. Ferraty, and P. Sarda. Functional linear model. *Statistics & Probability Letters*, 45(1):11–22, 1999.

[4] N. C. Chung and J. D. Storey. Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics*, 31(4):545–554, 2014.

[5] M. A. de Luis Balaguer, A. P. Fisher, N. M. Clark, M. G. Fernandez-Espinosa, B. K. Möller, D. Weijers, J. U. Lohmann, C. Williams, O. Lorenzo, and R. Sozzani. Predicting gene regulatory networks by combining spatial and temporal gene expression data in arabidopsis root stem cells. *Proceedings of the National Academy of Sciences*, 114(36):E7632–E7640, 2017.

[6] P. Delicado, R. Giraldo, C. Comas, and J. Mateu. Statistics for spatial functional data: some recent contributions. *Environmetrics: The official journal of the International Environmetrics Society*, 21(3-4):224–239, 2010.

[7] L. Di Filippo, D. Righelli, M. Gagliardi, M. R. Matarazzo, and C. Angelini. Hiceekr: a novel shiny app for hi-c data analysis. *Frontiers in genetics*, 10:1079, 2019.

[8] F. Ferraty and P. Vieu. *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media, 2006.

[9] E. R. Gamazon, H. E. Wheeler, K. P. Shah, S. V. Mozaffari, K. Aquino-Michaels, R. J. Carroll, A. E. Eyler, J. C. Denny, D. L. Nicolae, N. J. Cox, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, 47(9):1091, 2015.

[10] R. Giraldo, P. Delicado, and J. Mateu. Ordinary kriging for function-valued spatial data. *Environmental and Ecological Statistics*, 18(3):411–426, 2011.

[11] J. Goldsmith, J. Bobb, C. M. Crainiceanu, B. Caffo, and D. Reich. Penalized functional regression. *Journal of computational and graphical statistics*, 20(4):830–851, 2011.

[12] J. Goldsmith and T. Kitago. Assessing systematic effects of stroke on motor control by using hierarchical function-on-scalar regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(2):215–236, 2016.

[13] J. Goldsmith, V. Zipunnikov, and J. Schrack. Generalized multilevel function-on-scalar regression and principal component analysis. *Biometrics*, 71(2):344–353, 2015.

[14] T. Hastie and R. Tibshirani. Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987.

[15] Z. Ji and H. Ji. Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. *Nucleic acids research*, 44(13):e117–e117, 2016.

[16] R. M. Kumar, P. Cahan, A. K. Shalek, R. Satija, A. J. DaleyKeyser, H. Li, J. Zhang, K. Pardee, D. Gennert, J. J. Trombetta, et al. Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature*, 516(7529):56–61, 2014.

[17] E. Marco, R. L. Karp, G. Guo, P. Robson, A. H. Hart, L. Trippa, and G-C Yuan. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proceedings of the National Academy of Sciences*, 111(52):E5643–E5650, 2014.

[18] H. Mi, A. Muruganujan, D. Ebert, X. Huang, and P. D. Thomas. Panther version 14: more genomes, a new panther go-slim and improvements in enrichment analysis tools. *Nucleic acids research*, 47(D1):D419–D426, 2019.

[19] J. S. Morris. Functional regression. *Annual Review of Statistics and Its Application*, 2:321–359, 2015.

[20] O. Oluwadare, M. Highsmith, and J. Cheng. An overview of methods for reconstructing 3-d chromosome and genome structures from hi-c data. *Biological procedures online*, 21(1):7, 2019.

[21] O. Padovan-Merhar, G. P. Nair, A. G. Biaesch, A. Mayer, S. Scarfone, S. W. Foley, A. R. Wu, L. S. Churchman, A. Singh, and A. Raj. Single mammalian cells compensate for differences in cellular volume and dna copy number through independent global transcriptional mechanisms. *Molecular cell*, 58(2):339–352, 2015.

[22] R. N. Perry, M. Moens, and J. T. Jones. *Cyst nematodes*. CABI, 2018.

[23] J. O. Ramsay and C. J. Dalzell. Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):539–572, 1991.

[24] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, 2nd edition, 2005.

[25] M. Rosenthal, D. Bryner, F. Huffer, S. Evans, A. Srivastava, and N. Neretti. Bayesian estimation of three-dimensional chromosomal structure from single-cell hi-c data. *Journal of Computational Biology*, 26(11):1191–1202, 2019.

[26] W. Saelens, R. Cannoodt, H. Todorov, and Y. Saeys. A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547–554, 2019.

[27] K. Shekhar, S. W. Lapan, I. E. Whitney, N. M. Tran, E. Z. Macosko, M. Kowalczyk, X. Adiconis, J. Z. Levin, J. Nemesh, M. Goldman, et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell*, 166(5):1308–1323, 2016.

[28] C. Soneson and M. D. Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nature methods*, 15(4):255, 2018.

[29] K. Street, D. Risso, R. B. Fletcher, D. Das, J. Ngai, N. Yosef, E. Purdom, and S. Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics*, 19(1):477, 2018.

[30] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

[31] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381, 2014.

[32] S. Ullah and C. F. Finch. Applications of functional data analysis: A systematic review. *BMC medical research methodology*, 13(1):43, 2013.

[33] K. Van den Berge, H. R. De Bezieux, K. Street, W. Saelens, R. Cannoodt, Y. Saeys, S. Dudoit, and L. Clement. Trajectory-based differential expression analysis for single-cell sequencing data. *Nature Communications*, 11(1):1–13, 2020.

[34] J. D. Welch, A. J. Hartemink, and J. F. Prins. Slicer: inferring branched, nonlinear cellular trajectories from single cell rna-seq data. *Genome biology*, 17(1):106, 2016.

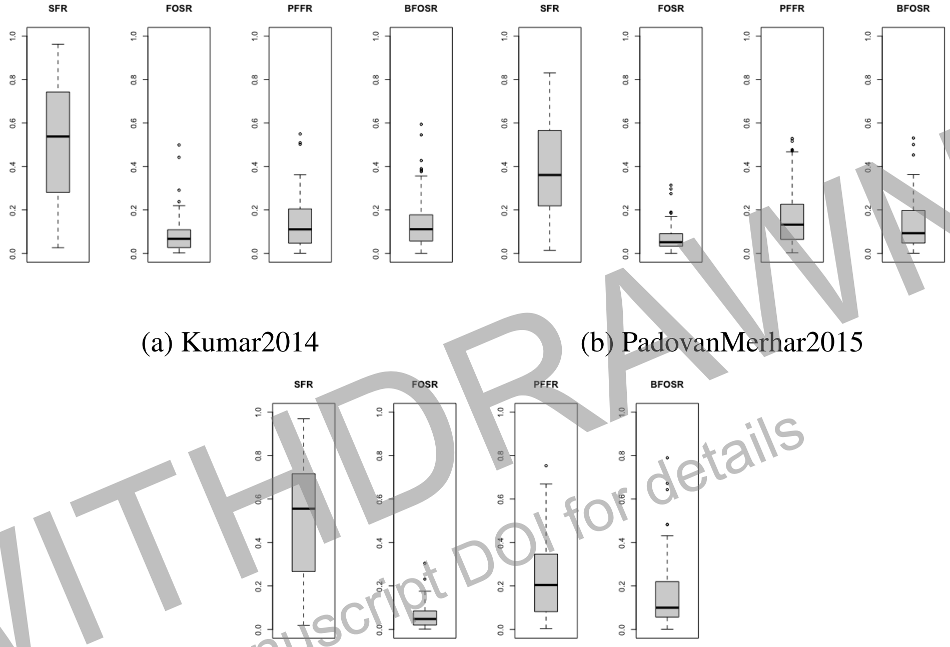
[35] B. Wilczynski, Y-H Liu, Z. X. Yeo, and E. EM Furlong. Predicting spatial and temporal gene expression using an integrative model of transcription factor occupancy and chromatin state. *PLoS computational biology*, 8(12), 2012.

[36] P. Zeng and X. Zhou. Non-parametric genetic prediction of complex traits with latent dirichlet process regression models. *Nature communications*, 8(1):1–11, 2017.

[37] P. Zeng, X. Zhou, and S. Huang. Prediction of gene expression with cis-snps using mixed models and regularization methods. *BMC genomics*, 18(1):368, 2017.

[38] H. Zhong, S. Kim, D. Zhi, and X. Cui. Predicting gene expression using dna methylation in three human populations. *PeerJ*, 7:e6757, 2019.

Appendix A

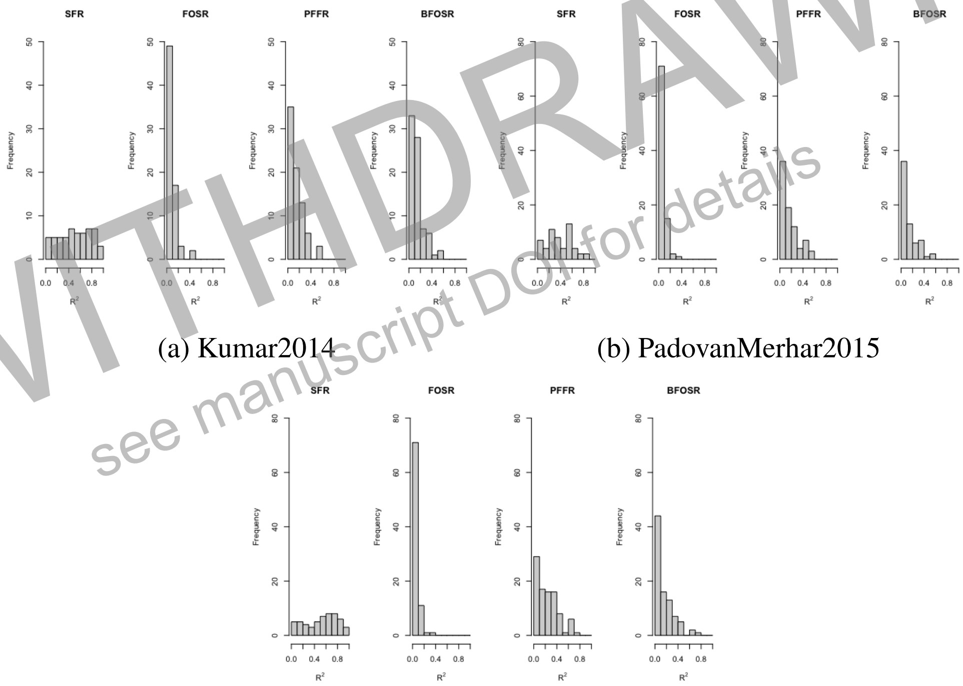


(a) Kumar2014

(b) PadovanMerhar2015

(c) Shekhar2016

Figure A1: Box plots for each data set for four functional regression models when $N_s = 1000$. The prediction performance of SFR using pseudo-location is superior to other functional response regression models.

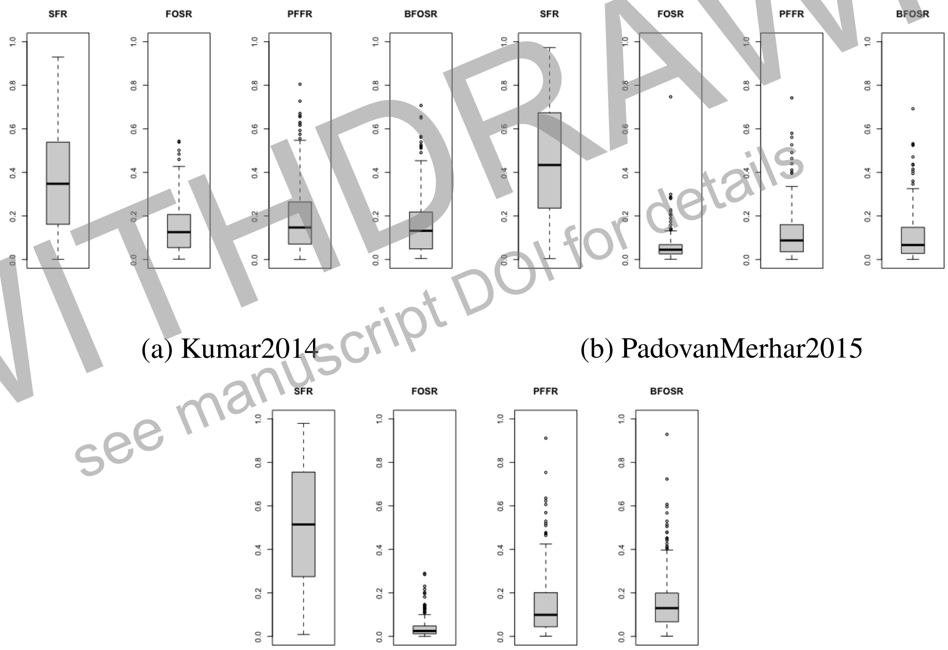


(a) Kumar2014

(b) PadovanMerhar2015

(c) Shekhar2016

Figure A2: Histograms for each data set for four functional regression models when $N_s = 1000$. The prediction performance of SFR using pseudo-location is superior to other functional response regression models.

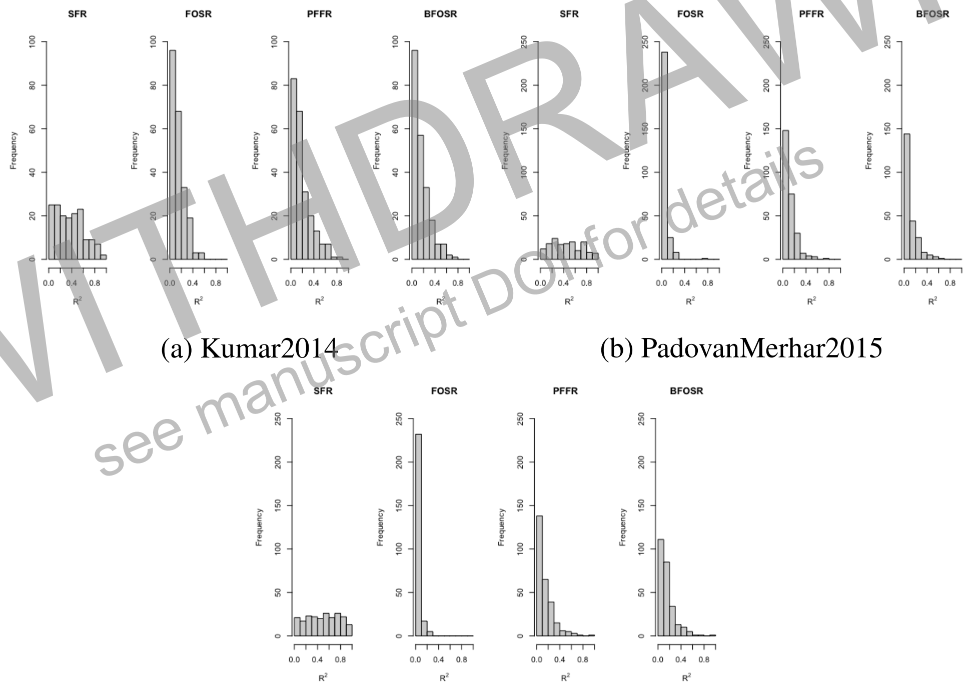


(a) Kumar2014

(b) PadovanMerhar2015

(c) Shekhar2016

Figure A3: Box plots for each data set for four functional regression models when $N_s = 3000$. The prediction performance of SFR using pseudo-location is superior to other functional response regression models.



(a) Kumar2014

(b) PadovanMerhar2015

(c) Shekhar2016

Figure A4: Histograms for each data set for four functional regression models when $N_s = 3000$. The prediction performance of SFR using pseudo-location is superior to other functional response regression models.