Spatially continuous dynamic factor modeling with basis expansion using $L_2$ penalized likelihood

Takamitsu Araki and Shotaro Akaho
Human Informatics Research Institute, National Institute of Advanced Industrial Science and Technology, Central 2, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan
E-mail: tk-araki@aist.go.jp

Abstract. In spatio-temporal data analysis, dimension reduction is necessary to extract intrinsic structures and to avoid over-parametrization problems. The spatial dynamic factor model (SDFM) reduces dimension of the data by decomposing them into spatial and temporal variations. The spatial variation is represented by a few spatially structured vectors, called factor loading vectors. The SDFM cannot be directly applied when the data contain missing values and their observation sites vary with time. We extend the factor loading vector to a smooth continuous function obtained by basis expansion, where we call the extended model the spatially continuous dynamic factor model (SCDFM), and estimate the SCDFM using the maximum $L_2$ penalized likelihood method. We derive model selection criteria to select a regularization parameter and the number of factors. Applications to synthetic and real data show the effectiveness of our modeling strategy in terms of estimation accuracy and stability.

1. Introduction
In many scientific and industrial fields, spatio-temporal data, which depend on time and space, are often observed. The amount of spatio-temporal data are increasing as measurement devices continue to be developed, thereby increasing the importance of statistical modeling of the spatio-temporal data [1]. In spatio-temporal data analysis, when the data at all observation sites are directly modeled by multivariate time series models such as the vector auto-regression moving average model [2], over-parametrization happens and complex models that are difficult to interpret are obtained. In order to avoid over-parametrization, the space-time autoregressive moving average model [3], where many parameters are fixed by topographical information, has been proposed. However, this model is not flexible enough to capture spatio-temporal structures.

For this reason, dimension reduction for the spatio-temporal data is necessary to extract their inherent and essential structures. The spatial dynamic factor model (SDFM; [4, 5, 6]) is a dynamic factor model [7, 8] extended to capture spatial structures. The SDFM reduces dimension of the spatio-temporal data by estimating spatial and temporal variations of the data based on the Bayesian approach. The spatial variation is represented by a few spatially correlated vectors, called factor loading (FL) vectors, which are assumed to follow a Gaussian process. The temporal variation is represented by autoregressive processes of the factors, called factor processes, which are the stochastic processes corresponding to the FL vectors. The estimated FL vectors and factor processes reveal the spatial, temporal and spatio-temporal structures of the data.
Since the SDFM represents spatial variations only on the observation points in discrete form, it cannot be directly applied to a data set that contains missing values or whose observation sites vary with time, such as weather buoy data. The SDFM also has to augment FL vectors and estimate them after the estimation of the other parameters, in order to interpolate unobserved points.

The spatio-temporal random effects model (STREM; [9, 10]) represents spatial variations continuously by basis expansion, which is a linear combination of spatial basis functions. The STREM can be directly applied if the data set contains missing values or if the observation sites change temporally, and can directly interpolate unobserved points using the estimated model. However, the STREM does not compress the dimension of the spatio-temporal data, and the essential spatial, temporal and spatio-temporal structures are not extracted.

We extend the SDFM to a spatially continuous model that has continuous FL functions instead of FL vectors; we call this the spatially continuous dynamic factor model (SCDFM), where the FL functions are constructed by basis expansion. Although the maximum likelihood method has served as a standard method for various statistical models, it often yields unstable and unsmooth FL functions for the SCDFM. Thus, we estimate the parameters in the SCDFM by the maximum penalized likelihood method with an $L_2$ penalty on the coefficients in the FL functions, giving us more stable and smooth FL functions than the maximum likelihood method. We derive an Expectation-Maximization (EM) algorithm [11] to maximize the penalized likelihood parameter and the number of factors. Since the maximum penalized likelihood estimation can also be considered within the Bayesian framework, we derive a model selection criterion from a Bayesian viewpoint [12] for evaluating the SCDFM estimated by the penalized method. In the FL functions estimated by the maximum penalized likelihood method, the number of parameters is not a suitable measure of model complexity, since the complexity also depends on the penalty term. Based on the concept of the effective number of parameters of a regression model [13, 14], we also derive the effective number of parameters of the estimated FL functions from a formula in the EM algorithm. By replacing the number of parameters of the FL functions in consistent AIC (CAIC; [15, 16]) with the derived effective number of parameters, we formally obtain an information criterion for the estimated SCDFM. We use these criteria to choose the regularization parameter and the number of factors.

The maximum likelihood estimation for the SCDFM and the proposed modeling procedures are applied to synthetic data and ozone concentration data, and their effectiveness in terms of estimation accuracy and stability is shown by comparing the SCDFMs estimated by the maximum likelihood method and our modeling procedures.

The rest of the paper is organized as follows. The SDFM is extended to the SCDFM in Section 2. The maximum penalized likelihood method with an $L_2$ penalty for the SCDFM and the EM algorithm are introduced in Section 3. In Section 4, the model selection criterion for evaluating the estimated SCDFM is derived from a Bayesian viewpoint. The effective number of parameters of the estimated FL function is derived, and the model selection criterion with the effective number of parameters is obtained based on the CAIC formula. In Section 5, the performance of our modeling methods is evaluated through applications to synthetic data and ozone concentration data. Finally, we present conclusions in Section 6.

2. Spatially continuous dynamic factor model

We extend the spatial dynamic factor model (SDFM; [4]) in order to allow to analyze directly the spatio-temporal data that have missing values or whose observation sites vary with time. The SDFM consists of an observation model and a system model. The observation model describes a relationship between the observation data and the factors via the factor loading (FL) vectors that represent spatial variations of the data. The system model describes the temporal evolution
of the factors that represent temporal structures of the data. We extend the FL vectors in the observation model to the continuous FL functions expressed by the basis expansion, and we call the model consisting of the system model and the extended observation model the spatially continuous dynamic factor model (SCDFM).

The SDFM assumes that the data \( y_t = (y_{1t}, \ldots, y_{Nt})' \) are observed in locations \( \mathbf{s} = (s_1, \ldots, s_N) \) contained in the region \( \mathbf{S} \), where \( \mathbf{S} \subset \mathbb{R}^p \), for each time \( t = 1, \ldots, T \). The observation model of the SDFM is described as

\[
y_t = \mathbf{Af}_t + \mathbf{e}_t, \quad \mathbf{e}_t \sim N(0, \sigma^2 \mathbf{D}_N),
\]

where \( \mathbf{f}_t = (f_{1t}, \ldots, f_{dt})' \) is a vector of the factors at time \( t \), \( \mathbf{A} = (\mathbf{\lambda}_1, \ldots, \mathbf{\lambda}_d) \) is an FL matrix, where \( \mathbf{\lambda}_j \) is an FL vector, \( \mathbf{D}_N \) is a known \( N \) order positive definite matrix, and \( \sigma^2 \) is a positive scale parameter.

In our study, we consider data whose observation sites or site number may vary with time, that is, the data \( y_t = (y_{1t}, \ldots, y_{N_t})' \) are observed in \( \mathbf{s}_t = (s_{1t}, \ldots, s_{N_t}) \), where \( s_{it} \in \mathbf{S} \). For example, this assumption is satisfied when missing values lie in the data. To cope with the time varying observation sites, we introduce continuous FL functions \( \mathbf{\lambda}_j(\mathbf{s}) \) against the factor loading vectors \( \mathbf{\lambda}_j \) for \( j = 1, \ldots, d \). The FL functions are constructed by the basis expansion as

\[
\mathbf{\lambda}_j(\mathbf{s}) = \sum_{i=1}^{m} b_i(\mathbf{s}) \alpha_{ij} = \mathbf{b}(\mathbf{s})' \mathbf{\alpha}_j, \quad \mathbf{s} \in \mathbf{S}, \quad j = 1, \ldots, d,
\]

where \( \mathbf{b}(\mathbf{s}) = (b_1(\mathbf{s}), \ldots, b_m(\mathbf{s}))' \) is a vector of known spatial basis functions, and \( \mathbf{\alpha}_j = (\alpha_{1j}, \ldots, \alpha_{mj})' \) is a coefficient vector for the \( j \)-th FL function. Then the observation model is expressed as

\[
y_{it} = \mathbf{\lambda}(\mathbf{s}_{it})' \mathbf{f}_t + \epsilon_{it} = \mathbf{b}(\mathbf{s}_{it})' \mathbf{A} \mathbf{f}_t + \epsilon_{it}, \quad i = 1, \ldots, N_t, \tag{1}
\]

where \( \mathbf{\lambda}(\mathbf{s}) = (\mathbf{\lambda}_1(\mathbf{s}), \ldots, \mathbf{\lambda}_d(\mathbf{s}))' \), and \( \mathbf{A} = (\mathbf{\alpha}_1, \ldots, \mathbf{\alpha}_d) \). A vectorized form of the observation model is

\[
y_t = \mathbf{A}(\mathbf{s}_t)' \mathbf{f}_t + \mathbf{e}_t, \quad \mathbf{e}_t \sim N(0, \sigma^2 \mathbf{D}_{N_t}), \tag{2}
\]

where \( \mathbf{A}(\mathbf{s}_t) = (\mathbf{\lambda}(\mathbf{s}_1), \ldots, \mathbf{\lambda}(\mathbf{s}_{N_t}))' \), \( \mathbf{B}_t = (\mathbf{b}(\mathbf{s}_1), \ldots, \mathbf{b}(\mathbf{s}_{N_t}))' \), \( \mathbf{e}_t = (\epsilon_{1t}, \ldots, \epsilon_{N_t})' \), and \( \mathbf{D}_{N_t} \) is a known \( N_t \) order positive definite matrix.

The system model that represents the dynamic evolution of the factors is

\[
\mathbf{f}_t = \mathbf{\Gamma} \mathbf{f}_{t-1} + \mathbf{\omega}_t, \quad \mathbf{\omega}_t \sim N(0, \mathbf{\Sigma}_\omega), \tag{3}
\]

where \( \mathbf{\Gamma} = \text{diag}(\gamma) \), \( \gamma = (\gamma_1, \ldots, \gamma_d)' \), and \( \mathbf{\Sigma}_\omega \) is a positive diagonal matrix.

The SCDFM consists of the extended observation model (2) and the system model (3).

The SCDFM encompasses the SDFM and the STREM without covariates as its special cases. When \( d = m \), \( \mathbf{A} = \mathbf{I}_d \), and \( \mathbf{\Gamma} \) is non-diagonal, the SCDFM is equivalent to the STREM without covariates, and when the observation locations are invariant, that is, \( \mathbf{s}_t = (s_1, \ldots, s_N) \), and \( b_i(\mathbf{s}) = 1(s_i)(\mathbf{s}) \), where \( 1(s_i)(\mathbf{s}) \) is an indicator function, i.e., \( \mathbf{B} = \mathbf{I}_N \), the SCDFM is the same model as the SDFM.
2.1. Scaling indeterminacy

The SCDFM has a scaling indeterminacy in the coefficients of the FL functions (Appendix A), so that a constraint on \( A \) or that on \( \Sigma_w \) is necessary in the estimation of the SCDFM, e.g., \( \| \alpha_j \| = c \) for \( j = 1, \ldots, d \), or \( \Sigma_w = cI_d \), where \( c \) is a constant. In this study, we impose the constraint on \( \Sigma_w, \Sigma_w = cI_d \), because the estimation under the constraint on \( \Sigma_w \) can be conducted by just the maximization without any constraint while the estimation under the constraint on \( A \) requires the maximization with the constraints. After the estimation, the scales of the FL functions are adjusted to the same value for the purpose of model interpretation. Accordingly the system maximization without any constraint while the estimation under the constraint on \( j \) for \( A \) or that on \( \Sigma_w \) are also adjusted to maintain the same likelihood value. Note that the SCDFM doesn’t have a rotational indeterminacy that lies in a traditional factor analysis model, because the factors in the SCDFM independently evolve according to different first AR models (3) whose AR coefficients are all different values. The detail of the indeterminacy is described in Appendix A.

3. Maximum penalized likelihood method with \( L_2 \) penalty

In the estimation of the SCDFM, the maximum likelihood (ML) method often results in overfitting and yields unstable and unsmooth FL functions. We estimate the unknown parameters \( \theta = (A, \gamma, \sigma_z^2) \) using the maximum penalized likelihood (MPL) method with \( L_2 \) penalty for \( A \). The penalty stabilizes the estimation of the FL functions and provides smooth FL functions.

The MPL method estimates the parameters of the SCDFM by maximizing the penalized log-likelihood function of the SCDFM. For the observation data \( y_{1:T} = (y_1, \ldots, y_T) \), the penalized log-likelihood function of the SCDFM is expressed as

\[
l_{p_{\alpha}}(\theta) = \frac{1}{T} \log p(y_{1:T}|\theta) - \frac{\rho_\alpha}{2} \sum_{j=1}^{d} \alpha_j^T K_\alpha^{(j)} \alpha_j, \tag{4}
\]

where \( p \) denotes a probability density function of \( y_{1:T} \) with \( \theta \), \( \rho_\alpha (> 0) \) is a regularization parameter that controls the smoothness of the estimated FL functions, and \( K_\alpha^{(j)} \) are known positive definite matrices. The elements of \( K_\alpha^{(j)} \) depend on the types of the \( L_2 \) penalty for \( A \), which are \( \sum_{j=1}^{d} \| \alpha_j \|^2, \sum_{i=1}^{n-1} \sum_{j=1}^{d} | \alpha_{ij} - \alpha_{i+1,j} |^2 \), where \( b_i \) and \( b_{i+1} \) are adjacent, and so on. We maximize the penalized log-likelihood function using the expectation-maximization (EM) algorithm and the quasi-Newton method.

The EM algorithm is more stable and less sensitive to an initial value than gradient-based optimum algorithms such as a quasi-Newton method, but the parameters updated by the EM algorithm converge very slowly near local optimum solutions. Thus, after the EM algorithm is run a sufficient number of iterations, the quasi-Newton method is applied in order to maximize the penalized log-likelihood function faster.

The EM algorithm maximizes a \( Q \)-function with the \( L_2 \) penalty iteratively to maximize the penalized log-likelihood function. The \( Q \)-function at iteration \( n \), \( Q_{p_{\alpha}}(\theta^{(n-1)}) \), is defined by \( \frac{1}{T} E \left[ \log p(y_{1:T}, f_{0:T}, \theta) | y_{1:T}, \theta^{(n-1)} \right] - \frac{\rho_\alpha}{2} \sum_{j=1}^{d} \alpha_j^T K_\alpha^{(j)} \alpha_j \), where \( E[\cdot | y_{1:T}, \theta^{(n-1)}] \) denotes a conditional expectation with respect to \( f_{0:T} \) given \( y_{1:T} \) with the parameters \( \theta^{(n-1)} \). From
\[ \frac{\partial Q_\rho_0(\theta|\theta^{(n-1)})}{\partial \theta} = 0, \]  
the updates of \( \theta \) at \( n \) iteration are obtained as follows.

\[ \alpha_j^{(n)} = \left\{ \frac{1}{\sum_{t=1}^{T} \langle f_{jt} \rangle^{(n-1)}} B_t D_{N_t}^{-1} y_t - B_t A_{-j}^{(n-1)} \langle f_{jt}(f_{t}e_{t-j})^{-1} \rangle^{(n-1)} \right\}^{-1} \times \sum_{t=1}^{T} B_t D_{N_t}^{-1} \left( \langle f_{jt} \rangle^{(n-1)} y_t - B_t A_{-j}^{(n-1)} \langle f_{jt}(f_{t}e_{t-j})^{-1} \rangle^{(n-1)} \right), \]

\[ \gamma^{(n)} = \text{diag} \left\{ \sum_{t=1}^{T} \langle f_{t-1} \rangle^{(n-1)} \right\}^{-1} \left( \sum_{t=1}^{T} (f_t \circ f_{t-1})^{-1} \right), \]

\[ \sigma_\rho^{2(n)} = \frac{1}{\sum_{t=1}^{T} N_t} \left\{ \sum_{t=1}^{T} y_t D_{N_t}^{-1} y_t - 2 \sum_{t=1}^{T} \langle f_{t} \rangle^{(n-1)} A^{(n)} B_t D_{N_t}^{-1} y_t \right\} + \text{tr} \left( \sum_{t=1}^{T} \langle f_{t} f_{t} \rangle^{(n-1)} \right) A^{(n)} B_t D_{N_t}^{-1} B_t A^{(n)} \right\}, \]

where \( \circ \) denotes an element-wise product, \( A_{-j}^{(n-1)} = (\alpha_1^{(n-1)}, \ldots, \alpha_{j-1}^{(n-1)}, \alpha_{j+1}^{(n-1)}, \ldots, \alpha_d^{(n-1)}) \), \( (f_t)_{-j} \) denotes the vector of the factors at time \( t \) except \( f_{jt} \), and \( \langle \cdot \rangle^{(n-1)} = E[y_{1:T}, \theta^{(n-1)}] \), which is calculated by the Kalman smoothing [17, 18]. After the EM algorithm is completed, the parameters \( A \) and \( \gamma \) are further updated by the quasi-Newton method, setting the EM solutions as its initial values.

To avoid the scaling indeterminacy, the system noise variances \( \Sigma_\omega = \sigma_\omega^2 I_d \) is fixed while the algorithms are running. However, if we update the system noise variance \( \sigma_\omega^2 \) only for the initial iterations in the EM algorithm, the parameters converge faster. Thus, in the EM algorithm we update \( \sigma_\omega^2 \) until \( M \) iterations, e.g., \( M = 50 \), and fix it from the \( M+1 \) iteration. The update of \( \sigma_\omega^2 \) is described in Appendix B.

After optimizing the parameters by the quasi-Newton method, we scale the estimated coefficient vectors \( \hat{A} \) to interpret the estimated model clearly. Accordingly the means and covariances of \( f_0 \) and the system noise variances \( \Sigma_\omega \) are adjusted in order to keep the same likelihood value (Appendix A). That is,

\[ \hat{A}^{\text{scaled}} \leftarrow \hat{A} H, \quad E[f_0]^{\text{scaled}} \leftarrow H^{-1} E[f_0], \]

\[ \text{Cov}[f_0]^{\text{scaled}} \leftarrow H^{-1} \text{Cov}[f_0](H')^{-1}, \quad \Sigma_\omega^{\text{scaled}} \leftarrow H^{-1} \Sigma_\omega (H')^{-1}, \]

where \( H = \text{diag}(c) \), where \( c \) is a \( d \) dimensional vector. For example, \( c = (\epsilon/\|\lambda_1\|_{L_2}, \ldots, \epsilon/\|\lambda_d\|_{L_2}) \) provides the FL functions having the same likelihood value (Appendix A).

4. Model selection criteria

In the MPL estimation with the \( L_2 \) penalty, it is crucial to select appropriately the number of factors \( d \) and the regularization parameter \( \rho_\alpha \), which controls the smoothness of the estimated FL functions. We derive two model selection criteria based on the Bayesian approach and the concept of the effective number of parameters.

4.1. Bayesian approach

From a Bayesian viewpoint, we consider the prior distribution for \( \alpha = (\alpha'_1, \ldots, \alpha'_d)' \) as a multivariate normal distribution given by

\[ \pi(\alpha|\rho_\alpha) = (2\pi)^{-(md)/2}T^{(md)/2}\rho_\alpha^{(md)/2}|K_{\alpha}|^{1/2}\exp\left(-\frac{T\rho_\alpha}{2}\alpha'K_{\alpha}\alpha\right), \]
where $K_{\alpha} = \text{block diag}\{K_{\alpha}^{(1)}, \ldots, K_{\alpha}^{(d)}\}$. We also consider the prior distribution for $\gamma$ and $\sigma^2_t$ as a uniform improper prior distribution $\pi(\gamma, \sigma^2_t) \propto c$. Then, the Bayesian estimation maximizing the posterior probability corresponds to the MPL estimation.

Considering the MPL estimation as the Bayesian estimation, Konishi et al [12] presented a model selection criterion, the generalized Bayesian information criterion (GBIC), for evaluating models estimated by the MPL method. Applying their result, we obtain a model selection criterion for evaluating the SCDFM estimated by the MPL method, as

$$
\text{GBIC}(\rho_0, d) = -2 \log p(y_{1:T}|\hat{\theta}) + T \rho_0 \hat{\alpha}'K_{\alpha}\hat{\alpha} + d \log \left( \frac{T}{2\pi} \right) + \log \left| Q_{\rho_0}(\hat{\alpha}, \gamma) \right| - md \log(\rho_0) - \log |K_{\alpha}|, \tag{10}
$$

where

$$
Q_{\rho_0}(\alpha, \gamma) = -\frac{1}{T} \partial^2 \log p(y_{1:T}|\alpha, \gamma, \hat{\sigma}^2_t) \bigg|_{\alpha=\hat{\alpha}, \gamma=\hat{\gamma}} + \left( \begin{array}{cc} \rho_0K_{\alpha} & 0_{md,d} \\ 0_{d,md} & 0_{d,d} \end{array} \right), \tag{11}
$$

where $0_{m,n}$ denotes an $m \times n$ matrix all of whose elements are zero. The Hessian matrix of the log-likelihood of the SCDFM cannot be calculated analytically, so we calculate it numerically.

To obtain stable numerical calculations of the Hessian matrix, we regard the scale parameter $\hat{\sigma}^2_t$ estimated by the MPL method as a hyper-parameter.

### 4.2. Effective number of parameters

The GBIC for the SCDFM requires the numerical calculation of the Hessian matrix, and the calculation becomes more unstable as the number of parameters increases. For evaluating stably the SCDFM estimated by the MPL method, we introduce AIC-formal criteria [15] including the derived effective number of parameters (ENP; [13, 14]) for $\hat{\Lambda}$ instead of the number of parameters. The criteria don’t require any complicated numerical calculation, and thus they are stable for the SCDFM even with a number of parameters.

The AIC-formal criteria for the SCDFM estimated by the ML method are $-2 \log p(y_{1:T}|\hat{\theta}) + g(T)(md+d)$, where $g(T)$ is a function of $T$, and the number of parameters in the penalty term, $md+d$, represents a measure of model complexity. However, in the FL functions estimated by the MPL method $\hat{\lambda}_j$, the number of parameters $m$ is not a suitable measure of model complexity, since the complexity of the functions depends on the penalty term.

The concept of the number of parameters was extended to the ENP by [13, 14], and others, and the ENP was shown to be a reasonable measure of model complexity especially in regression models. We derive the ENP for $\hat{\lambda}_j$, $df(\hat{\lambda}_j)$, from the EM update equation (5) based on the concept of the ENP for the ridge linear regression [13] as

$$
df(\hat{\lambda}_j) = \text{tr} \left\{ \left( \sum_{t=1}^{T} \langle f_{jt}^2 \rangle B_t^i D_{N_t}^{-1} B_t + \rho_0 T \hat{\sigma}^2_t K_{\alpha}^{(j)} \right)^{-1} \left( \sum_{t=1}^{T} \langle f_{jt}^2 \rangle B_t^i D_{N_t}^{-1} B_t \right) \right\},
$$

where $\langle \cdot \rangle = E[|y_{1:T}, \hat{\theta}|]$. When we assume that the observation sites are temporally invariant, i.e., $N_t = N$ and $B_t = B$, $D_N = I_N$, and that $\text{Var}(f_{jt}) = 0$, for $t = 1, \ldots, T$, then $df(\hat{\lambda}_j) = \text{tr} \left\{ \left( B^i B + (\rho_0 T \hat{\sigma}^2_t) \langle \sum_{t=1}^{T} c_{jt} K_{\alpha}^{(j)} \rangle \right)^{-1} B^i B \right\}$, where $c_{jt} = \langle f_{jt}^2 \rangle = \langle f_{jt}^2 \rangle$. This ENP is the same formula as that for the ridge regression model. Furthermore, when we suppose $\rho_0 = 0$, then $df(\hat{\alpha}_j) = m$. 


By replacing the number of parameters of the FL function $\lambda_j$ in the AIC-formal criteria with the derived ENP for $\lambda_j$, we formally obtain information criteria for the SCDFM estimated by the MPL method as $-2 \log p(\mathbf{y}_1:T|\hat{\Theta}) + g(T)(\sum_{j=1}^d \text{df}(\hat{\lambda}_j) + d)$ . In this study, we use the weight of the penalty in the consistent AIC (CAIC; [16]), i.e., $g(T) = \log(T) + 1$, and we call the criterion modified CAIC (CAICM) hereafter.

We choose the regularization parameter $\rho_0$ and the number of factors $d$ minimizing the GBIC or CAICM. Through all the MPL estimations during selecting $\rho_0$ and $d$, we use the system noise variances of the same value, i.e., $\Sigma_\omega = c I_d$, where $c$ is a constant value for all the estimations.

5. Applications

In this section, we applied the proposed modeling procedure and the ML estimation for the SCDFM to synthetic datasets and ozone concentration data, and compared their results.

5.1. Synthetic data

The synthetic data set was generated from the SCDFM having three factors and corresponding FL functions.

The true FL functions cover the observation region, $S = [0, 10] \times [0, 10]$, and they are constructed by the basis expansion of the $5 \times 5$ Gaussian basis functions, $\exp\left(-\frac{|s-\mu_i|^2}{2\nu_i^2}\right)$, where the centers of the basis functions, $\mu_i$, are allocated on lattice-shaped points (the blue points in Fig. 1(a)), and $\nu_i^2 = 2.5$ for $i = 1, \ldots, 25$. The true FL functions $\lambda_j^*(s)$ are shown in Fig. 1(a), and $L_2$ norms of them are all 10, i.e., $||\lambda_j^*||_{L_2} = 10, j = 1, 2, 3$.

The temporal evolution of the first factor is described as $f_{1t} = T(t) + \zeta_t$, where $\zeta_t = 0.6\zeta_{t-1} + \omega_{1t}$ and $T(t)$ is a linear trend described as $T(t) = t/100 + 4$. The other factors evolve as $f_{jt} = \gamma_j^* f_{j(t-1)} + \omega_{jt}$ for $j = 2, 3$, where $(\gamma_2^*, \gamma_3^*) = (0.55, -0.2)$. The initial values $\zeta_0, f_{2,0}, f_{3,0} \sim N(0, \Sigma_\zeta^*)$, and the system noises $\omega_{1t}, \omega_{2t}, \omega_{3t} \sim N(0, \Sigma_\omega^*)$, where $\Sigma_\omega^* = \text{diag}(0.32^2, 0.17^2, 0.1^2)$. Following these models, the factors $f_j^*$ were simulated for $t = 1, \ldots, T = 50$, as shown in Fig. 2.

Two synthetic datasets were generated on two point sets of 25 observation sites, (i) the centers of the basis functions, and (ii) uniformly random-sampled positions from $S$. Each of the datasets was generated on each of these point sets from the observation model in the SCDFM, Eq. (1), with the simulated factor processes $f_j^*_{1:T}$, the true FL functions $\lambda_j^*(s)$ and the observation noises $\epsilon_{t, i, j} \sim N(0, 0.2^3 I_{25})$. We assumed that 20% of the observations were missing values every time point.

We estimated the SCDFM using the MPL method and ML method from each of the datasets. The model includes the same basis functions as those constructing the true FL functions, and the initial factors are $f_0 \sim N(0, 10 I_d)$. In the MPL estimation, the EM algorithm described in Section 3 was run for 1000 iterations, and then the quasi-Newton method was conducted to maximize the penalized log-likelihood with respect to $\Theta$ and $\gamma_j$, where the initial values were the EM solutions. The regularization parameter and the number of factors were selected by the GBIC and CAICM. The matrices in the penalty term are $K_0^{(d)} = I_m$. The ML estimation was conducted using the same procedure as the MPL estimation, where the EM algorithm was run with $\rho_0 = 10^{-10}$. The number of factors was selected by the AIC and the BIC [19].

The estimated coefficients of the FL functions, $\alpha_{j,a}$, were scaled so that the $L_2$ norms of all the estimated FL functions were equal to 10 that is the $L_2$ norm of the true FL functions. For each of the estimated FL functions, its number and sign were adjusted to the number of the true FL function and the sign of the estimated FL function that minimize the $L_2$ distance between these functions, respectively, i.e., for $a \lambda_i, a = \pm 1, (i,a)^{\text{new}} \leftarrow \arg \min_{j,a} ||\lambda_j^* - a \lambda_i||_{L_2}$.

Through all the estimations in this experiment, the number of factors was selected as the true number 3, by both the GBIC and CAICM in the MPL estimation, and both the AIC and
BIC in the ML estimation. In the dataset observed on (i), the GBIC and CAIC\textsubscript{M} each selected \( \rho_0 = 0.25, 0.05 \) within the three-factor model. In the dataset observed on (ii), both the GBIC and CAIC\textsubscript{M} selected \( \rho_0 = 0.25 \) within the three-factor model. In the SCDFM that has many factors, the determinant of the numerical Hessian matrix in the GBIC often took a negative value, and thus the GBIC was not obtained.

The estimated FL functions, \( \hat{\lambda}_j(\cdot) \), are plotted in Fig. 1(b) and (c), where the FL function estimated by the MPL method with \( \rho_0 \) selected by the CAIC\textsubscript{M} in (i) is not plotted since it is visually indistinguishable from the FL function estimated with the GBIC. The \( L_2 \) errors of the estimated FL functions, \( ||\hat{\lambda}_j - \hat{\lambda}_j||_{L_2} \), are described in Table 1. For the observation point set (i), the MPL method estimated the FL functions slightly more accurately than the ML method (Table 1), but the FL functions estimated by the ML and MPL methods could capture the global structures of the true FL functions (Fig. 1(b)). On the other hand, for the observation point set (ii), the MPL method estimated the FL functions much more accurately than the ML method (Table 1), and the FL functions estimated by the MPL method could capture the global structures of the true FL functions. However, the FL functions by the ML method could not capture the true structures at all (Fig. 1(c)).

| Observation point set | (i) | (ii) |
|----------------------|-----|-----|
| MPL method with GBIC | \( \lambda_1 \) | 0.058 | 0.242 |
| MPL method with GBIC | \( \lambda_2 \) | 3.917 | 3.072 |
| MPL method with GBIC | \( \lambda_3 \) | 2.290 | 5.046 |
| MPL method with CAIC\textsubscript{M} | \( \lambda_1 \) | 0.057 | same as above |
| MPL method with CAIC\textsubscript{M} | \( \lambda_2 \) | 4.416 | |
| MPL method with CAIC\textsubscript{M} | \( \lambda_3 \) | 2.371 | |
| ML method | \( \lambda_1 \) | 0.056 | 12.388 |
| ML method | \( \lambda_2 \) | 4.818 | 13.714 |
| ML method | \( \lambda_3 \) | 2.412 | 14.089 |

Table 1: The \( L_2 \) errors of the FL functions estimated by the MPL method with the GBIC and CAIC\textsubscript{M}, and by the ML method.

We estimated the factors by posterior means of the factors calculated by the Kalman smoothing, i.e., \( \hat{f}_t \equiv E[f_t|y_{1:T}, \hat{\theta}] \). The estimated factor processes \( \hat{f}_{1:T} \) are plotted in Fig. 2, where the estimated factor processes by the MPL method with the CAIC\textsubscript{M} in (i) are not plotted since they are visually indistinguishable from those estimated with the GBIC. Root mean squared errors (RMSEs) of the estimated factor processes, \( \sqrt{\frac{1}{T} \sum_{t=1}^{T} (\hat{f}_{jt} - f^*_jt)^2} \), are described in Table 2. For the observation point set (i), the factor processes estimated by the ML and MPL methods could capture the simulated factor processes accurately (Fig. 2(a)), and the factor processes estimated by the MPL method were slightly more accurate than those by the ML method (Table 2). On the other hand, for the observation point set (ii), the factor processes estimated by the MPL method could capture the temporal variations of the simulated factor processes accurately (Fig. 2(c)), but those by the ML method could not capture them at all (Fig. 2(b)). Hence the factor processes estimated by the MPL method were much more accurate than those by the ML method (Table 2). This is because the MPL method estimated the corresponding FL functions much more accurately.

These results show that the MPL estimation with the GBIC and CAIC\textsubscript{M} of the SCDFM is more stable and accurate than the ML estimation in different observation conditions.

5.2. Ozone concentration data

The ozone (O\textsubscript{3}) is an air pollutant near the ground, and the O\textsubscript{3} concentration data were observed at 26 sites on the ground in New York state (Fig. 3). The data were observed daily for 46 days (Fig. 4), and contain 15 missing values. The dataset is available in the “spTimer”-package of R software [20]. The data were also log-transformed to normalize them. The coordinate of
Figure 1: True and estimated FL functions. Figure (a) shows the true functions $\lambda^*_j$, where blue filled circles denote the centers of the Gaussian basis functions. Figure (b) and (c) show the FL functions estimated from the data observed in (i) and (ii), respectively, where the black open circles denotes the observation points. The FL functions estimated by the MPL and ML methods are plotted in the upper and lower rows, respectively, in Figure (b) and (c). Dotted lines denote contour lines of 0.
Figure 2: The simulated factor processes (black line), the factor processes estimated by the ML method (blue line), and those estimated by the MPL method (red line). Figure (a) shows the estimated factor processes from the data observed in (i). Figure (b) and (c) show the estimated factor processes from the data observed in (ii). In the MPL estimation, Figure (a) shows the estimation result with the GBIC, and Figure (b) and (c) show that with both the GBIC and CAIC_M. In Figure (c), the factor processes estimated by the ML method are omitted, because the temporal variations of the factor processes simulated and estimated by the MPL method cannot be recognized if those by the ML method are included as Figure (b). Dashed lines denote $\hat{f}_{jt} \pm 2 \times$ posterior standard deviations of $f_{jt}$.
Observation point set | (i) $f_1$ | $f_2$ | $f_3$ | (ii) $f_4$ | $f_5$ | $f_6$
---|---|---|---|---|---|---
MPL method with GBIC | 0.0809 | 0.0316 | 0.0201 | 0.0674 | 0.0517 | 0.0355
MPL method with CAIC$_M$ | 0.0902 | 0.0321 | 0.0199 | same as above | |
ML method | 0.0982 | 0.0327 | 0.0199 | 37.179 | 38.6170 | 2.4177

Table 2: The RMSEs of the factor processes estimated by the MPL method with the GBIC and CAIC$_M$, and by the ML method.

the locations was converted from latitude and longitude to orthogonal coordinate, measured in kilometer.

![Map around New York state](image)

Figure 3: The map around New York state, the boundary of which is denoted by black lines. Green points show the observation sites of the $O_3$ concentrations.

![Logarithms of data observed at sites](image)

Figure 4: Figures (a), (b) and (c) show the logarithms of the data observed at the sites “a”, “b” and “c” in Fig. 3, respectively. The blank at the 18th day in the site “c” shows a missing value.

We used the Gaussian basis functions whose number was that of the observation sites, 26. The centers of the basis functions are the observation sites, and each of their spatial extents, $\nu_i$, is the mean of the distances between the center of the basis function, $\mu_i$, and the centers of its three nearest-neighbor basis functions. We estimated the SCDFM using the MPL and ML methods from the dataset through the same setup and procedure conducted in Section 5.1.

The estimated coefficients of the FL functions, $\hat{\alpha}_j$, were scaled so that the average of absolute values of the FL function on the observation sites was 0.5 for each of the estimated FL functions. The order and signs of the FL functions and related parameters were also adjusted with the likelihood value fixed in order to facilitate a comparison between the FL functions estimated by the ML method and those by the MPL method.
The number of factors was selected as five from three to seven by both the CAIC\textsubscript{M} in the MPL estimation and the BIC in the ML estimation. When the number of factors was more than four, the determinant of the numerical Hessian matrix in the GBIC in the MPL estimation often took a negative value, and thus the GBIC was not obtained. The GBIC in the MPL estimation and the AIC in the ML estimation selected a seven-factor model that was obviously redundant since the model had quite similarly shaped FL functions. This shows that the GBIC and AIC could not select an appropriate model. Thus, hereafter we discuss the models selected by the BIC and the CAIC\textsubscript{M}.

The estimated FL functions and factor processes are plotted in Fig. 5 and 6, respectively. In both the ML and MPL estimations, the first FL function takes positive values all over the observation region (Fig. 5(a)(1), (b)(1)), so that the temporal variation of the corresponding first factor represents common temporal variation over the region. The first factor process also represents the nonstationary grand mean component of the temporal variations of the data (Fig. 6 (1)), and thus the time averages of the other factors are almost 0 (Fig. 6 (2)-(5)). The signs of the other FL functions vary by site, and the difference in the signs is considered to represent directions of migrations of the O\textsubscript{3}. For example, the second FL function estimated by the MPL method takes positive values in the west area and negative values in the east (Fig. 5(b)(2)). Since the rise in the second factor increases the second component $f_2\hat{\lambda}_2$ on the west area and decreases that on the east area, the rise in the second factor is considered to represent a migration of the O\textsubscript{3} from the east area to the west. We can find out the spatio-temporal structure of the migration of the O\textsubscript{3} from the estimated factor processes and the structures of the corresponding estimated FL functions.

The FL functions estimated by the MPL method have smooth and clear structures, but those estimated by the ML method have rough and unclear structures, which cannot provide a clear interpretation of the migration of the O\textsubscript{3}. Especially, it is not easy to interpret a flux of the O\textsubscript{3} from the second FL function estimated by the ML method, since the 0-contour line of the FL function is winding, where the line decides the direction of O\textsubscript{3} migration in the second component.

6. Conclusions
We proposed the spatially continuous dynamic factor model (SCDFM) by extending the FL vectors of the spatial dynamic factor model to the continuous FL functions. The SCDFM can be directly applied to the spatio-temporal data whose observation sites and whose number vary with time, where the SDFM cannot be directly applied. To estimate smooth and clear FL functions of the SCDFM, we presented the maximum penalized likelihood method with the $L_2$ penalty for the FL functions. To maximize the penalized likelihood stably, we derived the EM algorithm. For the appropriate choice of the regularization parameter and the number of factors, we presented model selection criteria derived from a Bayesian view point and a concept of the effective number of parameters.

Through the applications to synthetic data and ozone concentration data, we showed that the proposed modeling procedure estimates the SCDFM accurately and stably, and especially provides the smooth and clear FL functions.

Acknowledgments
This study was supported by MEXT KAKENHI Grant Number 25120011.

Appendix A. Indeterminacy
We consider the SCDFM described in (2) and (3). Suppose that the matrix of the coefficients is $\hat{\mathbf{A}} = \mathbf{A}H$, and correspondingly the initial factors and the system noises at all times are $\hat{\mathbf{f}}_0 = H^{-1}\mathbf{f}_0$ and $\hat{\mathbf{\omega}}_t = H^{-1}\mathbf{\omega}_t$, respectively, where $H$ is a diagonal matrix or elementary matrix.
Figure 5: The FL functions estimated by the ML method with the BIC (a) and the MPL method with the CAICM (b). Blue circles denote the observation points, the centers of the Gaussian basis functions, and dotted lines denote contour lines of 0.

that exchanges rows or columns. Then the factors \( \tilde{f}_t \) are independent each other and expressed as \( \tilde{f}_t = H^{-1}f_t \), for \( t = 1, \ldots, T \). The observation \( \tilde{y}_t \) for the model with \( \tilde{A}, \tilde{f}_0 \) and \( \tilde{\omega}_t \) is \( \tilde{y}_t = \tilde{A}\tilde{f}_t + \epsilon_t = Af_t + \epsilon_t = y_t \). Thus, the model converted by \( H \) with \( \tilde{A}, \tilde{f}_0 \) and \( \tilde{\omega}_t \) and the original model with \( A, f_0 \) and \( \omega_t \) generate the same probability density of the observations. That is, the likelihood for any data takes the same value for both the model converted by \( H \) and the original model.

Particularly we calculate the log-likelihood of the SCDFM before converted and converted
by $H$ hereafter. The log-likelihood of the SCDFM before converted is expressed as

$$
\log p(y_{1:T}|A, \Gamma, \sigma^2_\epsilon) = \sum_{t=1}^{T} \log p(y_t|y_{1:t-1}, A, \gamma, \sigma^2_\epsilon)
$$

$$
= -\frac{1}{2} \left\{ \left( \sum_{t=1}^{T} N_t \right) \log 2\pi + \sum_{t=1}^{T} \log |R_{t|t-1}| \right. \\
+ \sum_{t=1}^{T} (y_t - y_{t|t-1})' R_{t|t-1}^{-1} (y_t - y_{t|t-1}) \right\},
$$

where $y_{1:t} = (y_1, \ldots, y_t)$, $p(y_t|y_{1:t-1}, A, \gamma, \sigma^2_\epsilon) = p(y_{t}|A\gamma, \sigma^2_\epsilon)$, $y_{t|t-1} = E[y_t|y_{1:t-1}]$ and $R_{t|t-1} = \text{Cov}[y_t|y_{1:t-1}]$. The conditional means and variances, $y_{t|t-1}$ and $R_{t|t-1}$ are expressed as

$$
y_{t|t-1} = B_t A f_{t|t-1},
$$

$$
R_{t|t-1} = B_t A V_{t|t-1} (B_t A)' + \sigma^2_D N_t,
$$

where $f_{t|t-1} = E[f_t|y_{1:t-1}]$, $V_{t|t-1} = \text{Cov}[f_t|y_{1:t-1}]$.

Suppose that $A$ is replaced by $\hat{A} = A H$, and correspondingly $f_0$ and $\omega_t$ are replaced by $\hat{f}_0 = H^{-1} f_0$ and $\hat{\omega}_t = H^{-1} \omega_t$, respectively. Then $\hat{f}_{t|t-1} = E[\hat{f}_t|\hat{y}_{1:t-1}, H]$ and $\hat{V}_{t|t-1} = \text{Cov}[\hat{f}_t|\hat{y}_{1:t-1}, H]$ are

$$
\hat{f}_{t|t-1} = H^{-1} f_{t|t-1},
$$

$$
\hat{V}_{t|t-1} = H^{-1} \Gamma V_{t|t-1} (H')^{-1} + H^{-1} \Sigma_{\omega}(H')^{-1},
$$

$$
= H^{-1} V_{t|t-1} (H')^{-1},
$$
and $\tilde{R}_{t|t-1} = \text{Cov}[\tilde{y}_t|\tilde{y}_{1:t-1}, H]$ and $\tilde{y}_{t|t-1} = E[\tilde{y}_t|\tilde{y}_{1:t-1}, H]$ 

$$\tilde{y}_{t|t-1} = B_tAHH^{-1}f_{t-1} = y_{t|t-1},$$

$$\tilde{R}_{t|t-1} = B_tAHH^{-1}V_{t|t-1}(H')^{-1}H'(B_tA)' + \sigma^2D_{N_t} = R_{t|t-1}.$$  \hfill(A.2)

From (A.1) and (A.2), the probabilistic densities of $y_{1:T}$ and $\tilde{y}_{1:T}$ are identical, i.e., the likelihoods for the models of $y_{1:T}$ and $\tilde{y}_{1:T}$ are equal.

**Appendix B. EM update of $\Sigma_\omega$**

Suppose $\Sigma_\omega = \sigma^2_\omega I_d$, then the update of $\sigma^2_\omega$ in the EM algorithm (Section 3) is

$$\sigma^2_{2(n)} = \frac{1}{Td} \left\{ \sum_{t=1}^{T} \langle f_t'f_t \rangle^{(n-1)} - 2 \text{tr} \left\{ \left( \sum_{t=1}^{T} \langle f_t'f_{t-1} \rangle^{(n-1)} \right) \Gamma \right\} 
+ \text{tr} \left\{ \left( \sum_{t=1}^{T} \langle f_{t-1}'f_t \rangle^{(n-1)} \right) \Gamma' \Gamma \right\} \right\}.$$  \hfill(B.1)

[1] Cressie N and Wikle C K 2011 *Statistics for spatio-temporal data* (John Wiley & Sons)

[2] Reinsel G C 2003 *Elements of multivariate time series analysis* (Springer Science & Business Media)

[3] Cliff A and Ord J K 1975 *Transactions of the Institute of British Geographers* 119–128

[4] Lopes Hedibert F, Salazar E and Gamerman D 2008 *Bayesian Analysis* 3 759–792

[5] Strickland C, Simpson D, Turner I, Denham R and Mengersen K 2011 *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 60 109–124

[6] Ippoliti L, Valentin P and Gamerman D 2012 *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 61 175–200

[7] Geweke J 1977 *Latent variables in socio-economic models* 1

[8] Stock J H and Watson M 2011 *Oxford Handbook on Economic Forecasting*

[9] Cressie N, Shi T and Kang E L 2010 *Journal of Computational and Graphical Statistics* 19 724–745

[10] Katzfuss M and Cressie N 2011 *Journal of Time Series Analysis* 32 430–446

[11] Dempster A P, Laird N M and Rubin D B 1977 *Journal of the Royal Statistical Society. Series B (Methodological)* 39 1–38

[12] Konishi S, Ando T and Imoto S 2004 *Biometrika* 91 27–43

[13] Hastie T J and Tibshirani R J 1990 *Generalized additive models* vol 43 (CRC press)

[14] Moody J E 1991 The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems *Advances in Neural Information Processing Systems (NIPS)* vol 4 pp 847–854

[15] Akaike H 1973 Information theory and an extension of the maximum likelihood principle *Second International Symposium on Information Theory* (Akademiai Kiado) pp 267–281

[16] Bozdogan H 1987 *Psychometrika* 52 345–370

[17] Ghahramani Z and Hinton G E 1996 Parameter estimation for linear dynamical systems Tech. rep. Technical Report CRG-TR-96-2, University of Toronto, Dept. of Computer Science

[18] Durbin J and Koopman S J 2012 *Time Series Analysis by State Space Methods* (Oxford university press)

[19] Schwarz G 1978 *The annals of statistics* 6 461–464

[20] Bakar K S and Sahu S K 2015 *Journal of Statistical Software* 63 1–32