Supplementary material for '*Regularized Multi-output Gaussian Convolution Process with Domain Adaptation*'

Xinming Wang, Chao Wang, Xuan Song, Levi Kirby, Jianguo Wu

Appendix A
Derivation of covariance function in convolution process

For the convolution process:

$$f_i(\boldsymbol{x}) = g_i(\boldsymbol{x}) * Z(\boldsymbol{x}) = \int_{-\infty}^{\infty} g_i(\boldsymbol{x}-\boldsymbol{u})Z(\boldsymbol{u})d\boldsymbol{u},$$

If $Z(\boldsymbol{x})$ is a commonly used white Gaussian noise process, i.e., $\operatorname{cov}(Z(\boldsymbol{x}), Z(\boldsymbol{x}')) = \delta(\boldsymbol{x}-\boldsymbol{x}')$ and $\mathbb{E}(Z(\boldsymbol{x})) = 0$, then the cross covariance is derived as:

$$\begin{aligned}
&\operatorname{cov}^f_{ij}(\boldsymbol{x},\boldsymbol{x}') = \operatorname{cov}\{g_i(\boldsymbol{x}) * Z(\boldsymbol{x}), g_j(\boldsymbol{x}') * Z(\boldsymbol{x}')\}\\
&= \mathbb{E}\left\{\int_{-\infty}^{\infty} g_i(\boldsymbol{x}-\boldsymbol{u})Z(\boldsymbol{u})d\boldsymbol{u}\int_{-\infty}^{\infty} g_j(\boldsymbol{x}'-\boldsymbol{u}')Z(\boldsymbol{u}')d\boldsymbol{u}'\right\}\\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_i(\boldsymbol{u})g_j(\boldsymbol{u}')\mathbb{E}\left\{Z(\boldsymbol{x}-\boldsymbol{u})Z(\boldsymbol{x}'-\boldsymbol{u}')\right\}d\boldsymbol{u}d\boldsymbol{u}'\\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_i(\boldsymbol{u})g_j(\boldsymbol{u}')\delta(\boldsymbol{x}-\boldsymbol{u}-\boldsymbol{x}'+\boldsymbol{u}')d\boldsymbol{u}d\boldsymbol{u}'\\
&= \int_{-\infty}^{\infty} g_i(\boldsymbol{u})g_j(\boldsymbol{u}-\boldsymbol{v})d\boldsymbol{u}, \qquad (1)
\end{aligned}$$

where $\boldsymbol{v} = \boldsymbol{x}-\boldsymbol{x}'$ and the last equality is based on the property of Dirac function that $\int g(\boldsymbol{u}')\delta(\boldsymbol{u}'-\boldsymbol{x})d\boldsymbol{u}' = g(\boldsymbol{x})$.

For our MGCP structure:

$$\begin{aligned}
y_i(\boldsymbol{x}) &= f_i(\boldsymbol{x}) + \epsilon_i(\boldsymbol{x}) = g_{ii}(\boldsymbol{x}) * Z_i(\boldsymbol{x}) + \epsilon_i(\boldsymbol{x}), i \in \mathcal{I}^S\\
y_t(\boldsymbol{x}) &= f_t(\boldsymbol{x}) + \epsilon_t(\boldsymbol{x}) = \sum_{j\in\mathcal{I}} g_{jt}(\boldsymbol{x}) * Z_j(\boldsymbol{x}) + \epsilon_t(\boldsymbol{x}),
\end{aligned}$$

the source-target covariance function can be calculated as:

$$\begin{aligned}
\operatorname{cov}^f_{it}(\boldsymbol{x},\boldsymbol{x}') &= \operatorname{cov}(f_i(\boldsymbol{x}), f_t(\boldsymbol{x}'))\\
&= \operatorname{cov}\left\{g_{ii}(\boldsymbol{x}) * Z_i(\boldsymbol{x}), \sum_{j\in\mathcal{I}} g_{jt}(\boldsymbol{x}') * Z_j(\boldsymbol{x}')\right\}\\
&= \sum_{j\in\mathcal{I}} \operatorname{cov}\left\{g_{ii}(\boldsymbol{x}) * Z_i(\boldsymbol{x}), g_{jt}(\boldsymbol{x}') * Z_j(\boldsymbol{x}')\right\}\\
&= \int_{-\infty}^{\infty} g_{ii}(\boldsymbol{u})g_{it}(\boldsymbol{u}-\boldsymbol{v})d\boldsymbol{u}, \quad i \in \mathcal{I}^S \qquad (2)
\end{aligned}$$

where $\boldsymbol{v} = \boldsymbol{x}-\boldsymbol{x}'$. In the same way, we can derive the auto-covariance as

$$\begin{aligned}
\operatorname{cov}^f_{ii}(\boldsymbol{x},\boldsymbol{x}') &= \int_{-\infty}^{\infty} g_{ii}(\boldsymbol{u})g_{ii}(\boldsymbol{u}-\boldsymbol{v})d\boldsymbol{u}, i \in \mathcal{I}^S\\
\operatorname{cov}^f_{tt}(\boldsymbol{x},\boldsymbol{x}') &= \sum_{j\in\mathcal{I}}\int_{-\infty}^{\infty} g_{jj}(\boldsymbol{u})g_{jt}(\boldsymbol{u}-\boldsymbol{v})d\boldsymbol{u}.
\end{aligned}$$

Appendix B
Proof of Theorem 1

Suppose that $g_{it}(\boldsymbol{x}) = 0, \forall i \in \mathcal{U} \subseteq \mathcal{I}^S$ for all $\boldsymbol{x} \in \mathcal{X}$. For notational convenience, suppose $\mathcal{U} = \{1, 2, ..., h | h \leq q\}$, then the predictive distribution of the model at any new input $\boldsymbol{x}_*$ is unrelated with $\{f_1, f_2, ..., f_h\}$ and is reduced to:

$$\begin{aligned}
p(y_t(\boldsymbol{x}_*)|\boldsymbol{y}) = \mathcal{N}(&\boldsymbol{k}_+^T\boldsymbol{C}_+^{-1}\boldsymbol{y}_+,\\
&\operatorname{cov}^f_{tt}(\boldsymbol{x}_*,\boldsymbol{x}_*) + \sigma_t^2 - \boldsymbol{k}_+^T\boldsymbol{C}_+^{-1}\boldsymbol{k}_+),
\end{aligned}$$

where $\boldsymbol{k}_+ = (\boldsymbol{K}^T_{h+1,*}, ..., \boldsymbol{K}^T_{q,*}, \boldsymbol{K}^T_{t,*})^T$, $\boldsymbol{y}_+ = (\boldsymbol{y}^T_{h+1}, ..., \boldsymbol{y}^T_q, \boldsymbol{y}^T_t)^T$, and

$$\boldsymbol{C}_+ = \begin{pmatrix} \boldsymbol{C}_{h+1,h+1} & \cdots & \boldsymbol{0} & \boldsymbol{C}_{h+1,t}\\ \vdots & \ddots & \vdots & \vdots\\ \boldsymbol{0} & \cdots & \boldsymbol{C}_{q,q} & \boldsymbol{C}_{q,t}\\ \boldsymbol{C}^T_{h+1,t} & \cdots & \boldsymbol{C}^T_{q,t} & \boldsymbol{C}_{t,t}\end{pmatrix}.$$

Proof. Recall that

$$\begin{aligned}
\operatorname{cov}^y_{jt}(\boldsymbol{x},\boldsymbol{x}') &= \operatorname{cov}^f_{jt}(\boldsymbol{x},\boldsymbol{x}')\\
&= \int_{-\infty}^{\infty} g_{jj}(\boldsymbol{u})g_{jt}(\boldsymbol{u}-\boldsymbol{v})d\boldsymbol{u},\\
\operatorname{cov}^y_{tt}(\boldsymbol{x},\boldsymbol{x}') &= \operatorname{cov}^f_{tt}(\boldsymbol{x},\boldsymbol{x}') + \sigma_t^2\delta(\boldsymbol{x}-\boldsymbol{x}')\\
&= \sum_{h\in\mathcal{I}}\int_{-\infty}^{\infty} g_{hh}(\boldsymbol{u})g_{ht}(\boldsymbol{u}-\boldsymbol{v})d\boldsymbol{u} + \sigma_t^2\delta(\boldsymbol{x}-\boldsymbol{x}'),
\end{aligned}$$

for all $j \in \{1, 2, ..., q\}$, so $g_{it}(\boldsymbol{x}) = 0, i \in \{1, 2, ..., h | h \leq q\}$ implies that $\operatorname{cov}^y_{it}(\boldsymbol{x},\boldsymbol{x}') = 0$ for all $i \in \{1, 2, ..., h\}$ and

$$\operatorname{cov}^y_{tt}(\boldsymbol{x},\boldsymbol{x}') = \sum_{i=h+1}^{t}\int_{-\infty}^{\infty} g_{ii}(\boldsymbol{u})g_{it}(\boldsymbol{u}-\boldsymbol{v})d\boldsymbol{u} + \sigma_t^2\delta(\boldsymbol{x}-\boldsymbol{x}').$$

Therefore, we have that $\boldsymbol{C}_{i,t} = 0, i \in \{1, 2, ..., h\}$ and partition covariance matrix $\boldsymbol{C} = \begin{pmatrix}\boldsymbol{C}_- & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{C}_+\end{pmatrix}$, where $\boldsymbol{C}_- = \begin{pmatrix} \boldsymbol{C}_{1,1} & \boldsymbol{0} & \cdots & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{C}_{2,2} & \cdots & \boldsymbol{0}\\ \vdots & \vdots & \ddots & \vdots\\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{C}_{h,h}\end{pmatrix}$.

The predictive distribution at point $\boldsymbol{x}_*$ is

$$y_t(\boldsymbol{x}_*) \sim N(\boldsymbol{K}_*^T\boldsymbol{C}^{-1}\boldsymbol{y}, \operatorname{cov}^f_{tt}(\boldsymbol{x}_*,\boldsymbol{x}_*) + \sigma_t^2 - \boldsymbol{K}_*^T\boldsymbol{C}^{-1}\boldsymbol{K}_*).$$

Also, based on that $\text{cov}_{it}^{y}(\boldsymbol{x},\boldsymbol{x}')=0$ for all $i\in\{1,2,...,h\}$, we have that $\boldsymbol{K}_*=\left(\boldsymbol{0},\boldsymbol{k}_+^T\right)^T$. Let $\boldsymbol{y}_-=\left(\boldsymbol{y}_1^T,...,\boldsymbol{y}_h^T\right)^T$, then $\boldsymbol{y}=\left(\boldsymbol{y}_-^T,\boldsymbol{y}_+^T\right)^T$. Therefore,

$$\begin{aligned}\boldsymbol{K}_*^T\boldsymbol{C}^{-1}\boldsymbol{y}&=(\boldsymbol{0},\boldsymbol{k}_+^T)\begin{pmatrix}\boldsymbol{C}_- & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{C}_+\end{pmatrix}^{-1}(\boldsymbol{y}_-^T,\boldsymbol{y}_+^T)^T\\ &=(\boldsymbol{0},\boldsymbol{k}_+^T)\begin{pmatrix}\boldsymbol{C}_-^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{C}_+^{-1}\end{pmatrix}(\boldsymbol{y}_-^T,\boldsymbol{y}_+^T)^T\\ &=\boldsymbol{k}_+^T\boldsymbol{C}_+^{-1}\boldsymbol{y}_+,\\ \boldsymbol{K}_*^T\boldsymbol{C}^{-1}\boldsymbol{K}_*&=(\boldsymbol{0},\boldsymbol{k}_+^T)\begin{pmatrix}\boldsymbol{C}_- & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{C}_+\end{pmatrix}^{-1}(\boldsymbol{0},\boldsymbol{k}_+^T)^T\\ &=\boldsymbol{k}_+^T\boldsymbol{C}_+^{-1}\boldsymbol{k}_+.\end{aligned}$$

Note that the auto-covariance matrix of target output f_t, $\boldsymbol{C}_{tt}$, is also unrelated with observed data $\{\boldsymbol{X}_i|i=1,2,...,h\}$ which from source output $\{f_i|i=1,2,...,h\}$. As a result, the predictive distribution is totally independent on these outputs. Proof completes.

Appendix C
Regularity conditions

In this part, we state the regularity conditions for the consistency theorem of the MLE $\hat{\boldsymbol{\theta}}_\#$, which are formulated in [34].

Denote $\boldsymbol{y}$ with total N observations as $\boldsymbol{y}^N$, and let

$$p_k(\boldsymbol{\theta})=\frac{p(\boldsymbol{y}^k|\boldsymbol{\theta})}{p(\boldsymbol{y}^{k-1}|\boldsymbol{\theta})}$$

for each k. Assume $p_k(\boldsymbol{\theta})$ is twice differentiable with respect to $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}^*$. Also assume that the support of $p(\boldsymbol{y}^N|\boldsymbol{\theta})$ is independent of $\boldsymbol{\theta}$ in the neighborhood. Define $\phi_k(\boldsymbol{\theta})=\log p_k(\boldsymbol{\theta})$, and its first derivative $\phi_k'(\boldsymbol{\theta})$, second derivative $\phi_k''(\boldsymbol{\theta})$.

For simplicity and without loss of generality, we only consider the conditions for one-dimensional case. Define $\phi_k^{*\prime}=\phi_k'(\theta^*)$ and $\phi_k^{*\prime\prime}=\phi_k''(\theta^*)$. Let $\mathcal{F}_N$ be the σ-field generated by $y_j, 1\le j\le N$, and $\mathcal{F}_0$ be the trivial σ-field. Define the random variable $i_k^*=var(\phi_k^{*\prime}|\mathcal{F}_{k-1})=\mathbb{E}[(\phi_k^{*\prime})^2|\mathcal{F}_{k-1}]$ and $I_N^*=\sum_{k=1}^N i_k^*$. Define $S_N=\sum_{k=1}^N\phi_k^{*\prime}$ and $S_N^*=\sum_{k=1}^N\phi_k^{*\prime\prime}+I_N^*$. If the following conditions hold:

- (c1) $\phi_k(\theta)$ is thrice differentiable in the neighborhood of θ^*. Let $\phi_k^{*\prime\prime\prime}=\phi_k'''(\theta^*)$ be the third derivative,
- (c2) Twice differentiation of $\int p(\boldsymbol{y}^N|\theta)d\mu^N(\boldsymbol{y}^N)$ with respect to θ exists in the neighborhood of θ^*,
- (c3) $\mathbb{E}|\phi_k^{*\prime\prime}|<\infty$ and $\mathbb{E}|\phi_k^{*\prime\prime}+(\phi_k^{*\prime})^2|<\infty$.
- (c4) There exists a sequence of constants $K(N)\to\infty$ as $N\to\infty$ such that:
  - (i) $K(N)^{-1}S_N\xrightarrow{p}0$,
  - (ii) $K(N)^{-1}S_N^*\xrightarrow{p}0$,
  - (iii) there exists $a(\theta^*)>0$ such that $\forall\epsilon>0$, $P[K(N)^{-1}I_N^*\ge 2a(\theta^*)]\ge 1-\epsilon$ for all $N\ge N(\epsilon)$,
  - (iv) $K(N)^{-1}\sum_{k=1}^N\mathbb{E}|\phi_k^{*\prime\prime\prime}|<M<\infty$ for all N,

then the MLE $\hat{\theta}_\#$ is consistent for θ^*. There exists a sequence r_N such that $r_N\to\infty$ as $N\to\infty$, i.e.,

$$\|\hat{\theta}_\#-\theta^*\|=O_P(r_N^{-1}).$$

Appendix D
Proof of Theorem 2

Suppose that the MLE for $L(\boldsymbol{\theta}|\boldsymbol{y})$, $\hat{\boldsymbol{\theta}}_\#$, is r_N consistent, i.e., satisfying Eq. (19). If $\max\{|\mathbb{P}_\gamma''(\theta_{i0}^*)|:\theta_{i0}^*\neq 0\}\to 0$, then there exists a local maximizer $\hat{\boldsymbol{\theta}}$ of $L_{\mathbb{P}}(\boldsymbol{\theta}|\boldsymbol{y})$ s.t. $\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|=O_P(r_N^{-1}+r_0)$, where $r_0=\max\{|\mathbb{P}_\gamma'(\theta_{i0}^*)|:\theta_{i0}^*\neq 0\}$.

Proof. Recall the assumptions in Section 3.2. For the unpenalized log-likelihood $L(\boldsymbol{\theta})$, the MLE $\hat{\boldsymbol{\theta}}_\#$ is r_N consistent where r_N is a sequence such that $r_N\to\infty$ as $N\to\infty$. And we have that $L'(\boldsymbol{\theta}^*)=O_P(r_N)$ and $\boldsymbol{I}_N(\boldsymbol{\theta}^*)=O_P(r_N^2)$, which are the standard argument based on the consistency of estimator. Based on that, we aim to study the asymptotic properties of the penalized likelihood $L_{\mathbb{P}}(\boldsymbol{\theta})=L(\boldsymbol{\theta})-r_N^2\mathbb{P}_\gamma(\boldsymbol{\theta}_0)$. Here we multiply the penalty function by r_N^2 to avoid that penalty term degenerates as $N\to\infty$. The following proof is similar to that of Fan and Li [33] but based on dependent observations.

To prove theorem 2, we need to show that for any given $\epsilon>0$, there exists a large constant U such that:

$$P\left\{\sup_{\|\boldsymbol{u}\|=U}L_{\mathbb{P}}(\boldsymbol{\theta}^*+r_N^+\boldsymbol{u})<L_{\mathbb{P}}(\boldsymbol{\theta}^*)\right\}\ge 1-\epsilon,\qquad(3)$$

where $r_N^+=r_N^{-1}+r_0$. This implies that with probability at least $1-\epsilon$ there exists a local maximum in the ball $\{\boldsymbol{\theta}^*+r_N^+\boldsymbol{u}:\|\boldsymbol{u}\|\le U\}$. So the local maximizer $\hat{\boldsymbol{\theta}}$ satisfies that $\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|=O_P(r_N^+)$.

By $\mathbb{P}_\gamma(0)=0$, we have

$$\begin{aligned}&L_{\mathbb{P}}(\boldsymbol{\theta}^*+r_N^+\boldsymbol{u})-L_{\mathbb{P}}(\boldsymbol{\theta}^*)\\ &\le L(\boldsymbol{\theta}^*+r_N^+\boldsymbol{u})-L(\boldsymbol{\theta}^*)\\ &\quad-r_N^2\sum_{i=h+1}^q\left[\mathbb{P}_\gamma(|\theta_{i0}^*+r_N^+u_{i0}|)-\mathbb{P}_\gamma(|\theta_{i0}^*|)\right],\end{aligned}$$

where h and q are the number of zero components and all components in θ_{i0}^*, and u_{i0} is the element corresponding to θ_{i0} in $\boldsymbol{u}$. Let $\boldsymbol{I}_N(\boldsymbol{\theta}^*)$ be the finite and positive definite information matrix at $\boldsymbol{\theta}^*$ with N observations. Applying a Taylor expansion on the likelihood function, we have that

$$\begin{aligned}&L_{\mathbb{P}}(\boldsymbol{\theta}^*+r_N^+\boldsymbol{u})-L_{\mathbb{P}}(\boldsymbol{\theta}^*)\\ &\le r_N^+L'(\boldsymbol{\theta}^*)^T\boldsymbol{u}-\frac{1}{2}(r_N^+)^2\boldsymbol{u}^T\boldsymbol{I}_N(\boldsymbol{\theta}^*)\boldsymbol{u}[1+o_P(1)]\\ &\quad-r_N^2\sum_{i=h+1}^q\Big\{r_N^+\mathbb{P}_\gamma'(|\theta_{i0}^*|)\text{sign}(\theta_{i0}^*)u_{i0}\\ &\qquad+\frac{1}{2}(r_N^+)^2\mathbb{P}_\gamma''(|\theta_{i0}^*|)u_{i0}^2[1+o_P(1)]\Big\},\end{aligned}\qquad(4)$$

Note that $\|L'(\boldsymbol{\theta}^*)\|=O_P(r_N)$ and $\boldsymbol{I}_N(\boldsymbol{\theta}^*)=O_P(r_N^2)$. so the first term on the right-hand side of Eq. (4) is on the order $O_P(r_N^+r_N)$, while the second term is $O_P\left((r_N^+r_N)^2\right)$. By choosing a sufficient large U, the first term can be dominated by the second term uniformly in $\|\boldsymbol{u}\|=U$. Besides, the absolute value of the third term is bounded by

$$\sqrt{q-h}\,r_N^2r_N^+r_0\|\boldsymbol{u}\|+(r_Nr_N^+)^2\max\{|\mathbb{P}_\gamma''(\theta_{i0}^*)|:\theta_{i0}^*\neq 0\}\|\boldsymbol{u}\|^2,$$

which is also dominated by second term as it is on the order of $o_P\left((r_Nr_N^+)^2\right)$. Thus, Eq. (3) holds and the proof completes.

APPENDIX E
PROOF OF THEOREM 3

Let $\boldsymbol{\theta}_{10}^*$ and $\boldsymbol{\theta}_{20}^*$ contain the zero and non-zero components in $\boldsymbol{\theta}_0^*$ respectively. Assume the conditions in Theorem 2 also hold, and $\hat{\boldsymbol{\theta}}$ is r_N consistent by choosing proper γ in $\mathbb{P}_\gamma(\boldsymbol{\theta}_0)$. If $\liminf_{N\to\infty} \liminf_{\theta\to 0^+} \gamma^{-1}\mathbb{P}'_\gamma(\theta) > 0$ and $(r_N\gamma)^{-1} \to 0$, then

$$\lim_{N\to\infty} P\left(\hat{\boldsymbol{\theta}}_{10} = \mathbf{0}\right) = 1.$$

Proof. To prove this theorem, we only need to prove that for a small $\epsilon_N = Ur_N$, where U is a given constant and $i = 1, ..., s$,

$$\frac{\partial L_{\mathbb{P}}(\boldsymbol{\theta})}{\partial \theta_{i0}}\theta_{i0} < 0, 0 < |\theta_{i0}| < \epsilon_N. \tag{5}$$

By Taylor's expansion,

$$\begin{aligned}\frac{\partial L_{\mathbb{P}}(\boldsymbol{\theta})}{\partial \theta_{i0}} &= \frac{\partial L(\boldsymbol{\theta})}{\partial \theta_{i0}} - r_N^2\mathbb{P}'_\gamma(|\theta_{i0}|)\mathrm{sign}(\theta_{i0}) \\ &= \frac{\partial L(\boldsymbol{\theta}^*)}{\partial \theta_{i0}} + \left[\partial\left(\frac{\partial L(\boldsymbol{\theta}^*)}{\partial \theta_{i0}}\right)/\partial\boldsymbol{\theta}\right]^T (\boldsymbol{\theta} - \boldsymbol{\theta}^*)[1 + o_P(1)] \\ &\quad - r_N^2\mathbb{P}'_\gamma(|\theta_{i0}|)\mathrm{sign}(\theta_{i0}).\end{aligned}$$

As $\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_{i0}} = O_P(r_N)$, $\partial\left(\frac{\partial L(\boldsymbol{\theta}^*)}{\partial \theta_{i0}}\right)/\partial\theta_j = O_P(r_N^2)$ by the standard argument for r_N consistent estimator, thus

$$\begin{aligned}\frac{\partial L_{\mathbb{P}}(\boldsymbol{\theta})}{\partial \theta_{i0}} &= O_P(r_N) - r_N^2\mathbb{P}'_\gamma(|\theta_{i0}|)\mathrm{sign}(\theta_{i0}) \\ &= r_N^2\gamma\left(O_P(\frac{1}{r_N\gamma}) - \gamma^{-1}\mathbb{P}'_\gamma(|\theta_{i0}|)\mathrm{sign}(\theta_{i0})\right).\end{aligned}$$

Because that $\liminf_{N\to\infty} \liminf_{\theta\to 0^+} \gamma^{-1}\mathbb{P}'_\gamma(\theta) > 0$ and $(r_N\gamma)^{-1} \to 0$, $\frac{\partial L_{\mathbb{P}}(\boldsymbol{\theta})}{\partial \theta_{i0}}$ will be positive while θ_{i0} is negative and vise versa. As a result, Eq. (5) follows. Proof completes.

APPENDIX F
INTERPRETATION OF THE BENCHMARK: MGCP-RF

The illustration of MGCP-RF is shown in Fig. 1.

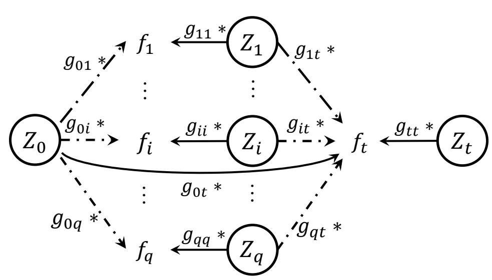

Fig. 1. The structure of MGCP-RF

In this structure, target f_t is generated by three kinds of latent process: $Z_0(\boldsymbol{x})$, $\{Z_i(\boldsymbol{x})\}_{i=1}^q$ and $Z_t(\boldsymbol{x})$. As $Z_0(\boldsymbol{x})$ is the common process shared by sources, the covariance matrix blocks between source f_i and the other outputs are zero only when the scale parameters in $g_{0i}(\boldsymbol{x})$ and $g_{it}(\boldsymbol{x})$ are zero simultaneously. Thus, the marginalized covariance matrix $\boldsymbol{C}_+$ in Theorem 1 will be:

$$\boldsymbol{C}_+ = \begin{pmatrix} \boldsymbol{C}_{h+1,h+1} & \cdots & \boldsymbol{C}_{h+1,q} & \boldsymbol{C}_{h+1,t} \\ \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{C}_{h+1,q}^T & \cdots & \boldsymbol{C}_{q,q} & \boldsymbol{C}_{q,t} \\ \boldsymbol{C}_{h+1,t}^T & \cdots & \boldsymbol{C}_{q,t}^T & \boldsymbol{C}_{t,t} \end{pmatrix}.$$

The difference to MGCP-R is that covariance among the remaining sources $\{f_i\}_{i=h+1}^q$ can be modeled. This structure is indeed more comprehensive but with the cost of a half more parameters than MGCP-R. The cost will increase if we use more latent process to model the correlation among sources.

To realize the effect of shrinking $g_{0i}(\boldsymbol{x})$ and $g_{it}(\boldsymbol{x})$ at the same time, group-L1 penalty is used and the penalized log-likelihood function is:

$$\max_{\boldsymbol{\theta}} L_{\mathbb{P}}(\boldsymbol{\theta}|\boldsymbol{y}) = L(\boldsymbol{\theta}|\boldsymbol{y}) - \gamma\sum_{i=1}^q \sqrt{\alpha_{0i}^2 + \alpha_{it}^2},$$

APPENDIX G
INFLUENCE OF TUNING-PARAMETER

To test the influence of the tuning-parameter γ in our model, we conduct the following experiment. Based on the same dataset in the 1D example of simulation case I, we construct MGCP-R model only with sources f_1 and f_2, and let γ vary from 0 to 10 at a step of 1. Note that MGCP-T is equal to the model with $\gamma = 0$. The boxplot of MAE with respect to different values of γ is shown in Fig. 2. The estimated value of α_{1t}, α_{2t} in one repetition is presented in Fig. 3. It can be seen that as γ increases, source f_2 will be excluded from the prediction of target, leading to an increased prediction error. In practice, cross-validation can be used to select an optimal tuning-parameter.

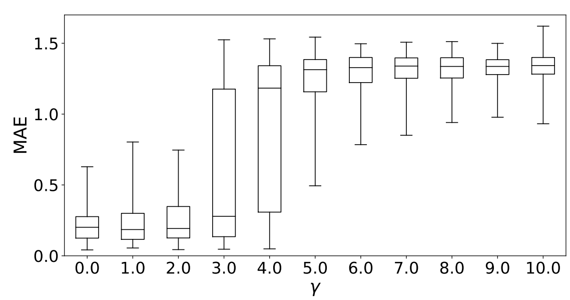

Fig. 2. Prediction error with different γ in 100 repetition.

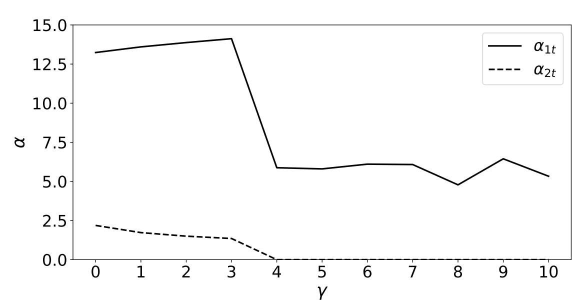

Fig. 3. Estimated values of α_{1t}, α_{2t} in one repetition.