# Learning Localized Spatio-Temporal Models From Streaming Data: Supplementary Material

## 1 Reformulating of fitting criterion

By expanding the objective function in (11), we obtain the equivalent form

$$\widetilde{\mathbf{y}}^{\top}\mathbf{K}_{\theta}^{-1}\widetilde{\mathbf{y}} + \underbrace{n\theta_0 + \sum_{j=1}^{p}\theta_j\|[\mathbf{\Phi}]_j\|_2}_{=\mathrm{tr}\{\mathbf{K}_\theta\}} \tag{16}$$

where $\widetilde{\mathbf{y}} = \mathbf{y} - \mathbf{1}\eta$ when $\mathbf{u}(\mathbf{s},t) \equiv 1$. Next we define an auxiliary variable $\alpha$ that satisfies

$$\alpha \geq \widetilde{\mathbf{y}}^{\top}\mathbf{K}^{-1}\widetilde{\mathbf{y}}$$

or equivalently

$$\begin{bmatrix}\alpha & \widetilde{\mathbf{y}}^{\top} \\ \widetilde{\mathbf{y}} & \mathbf{K}\end{bmatrix} \succeq \mathbf{0} \tag{17}$$

Using the auxiliary variable and the definition of $\mathbf{K}$, we can therefore express the objective function as:

$$\min_{\alpha,\boldsymbol{\theta}} \; \alpha + n\theta_0 + \sum_{j=1}^{p}\theta_j\|[\mathbf{\Phi}]_j\|_2, \tag{18}$$

where $\theta_j$ are nonnegative and $\alpha$ satisfies the constraint. The minimizing $\widehat{\boldsymbol{\theta}}$ is the learned model parameter. This problem is identified as a convex, semidefinite program, cf. [1]. We may also add the following normalization constraint,

$$\mathrm{tr}\{\widetilde{\mathbf{K}} - \mathbf{K}_\theta\} = 0,$$

to match the normalized covariance matrix. This merely adds a linear constraint to problem (18) with a constrained minimizer denoted $\boldsymbol{\theta}^{\star}$. We now prove that $\widehat{\boldsymbol{\theta}} \propto \boldsymbol{\theta}^{\star}$.

Begin by defining a constant $\kappa > 0$, such that $\mathrm{tr}\{\widetilde{\mathbf{K}}\mathbf{K}^{-1}(\boldsymbol{\theta}^{\star})\} = \kappa^2\mathrm{tr}\{\mathbf{K}(\boldsymbol{\theta}^{\star})\}$ at the minimum of (18). We show that $\kappa = 1$ is the only possible value and so both terms in (18) equal each other at the minimum.

Let $\tilde{\boldsymbol{\theta}} = \kappa\boldsymbol{\theta}^{\star}$, and observe that the cost (18) is then bounded by

$$\begin{aligned}(\kappa^2+1)\mathrm{tr}\{\mathbf{K}(\boldsymbol{\theta}^{\star})\} &\leq \mathrm{tr}\{\widetilde{\mathbf{K}}\mathbf{K}^{-1}(\tilde{\boldsymbol{\theta}})\} + \mathrm{tr}\{\mathbf{K}(\widetilde{\boldsymbol{\theta}})\} \\ &= \kappa^{-1}\mathrm{tr}\{\widetilde{\mathbf{K}}\mathbf{K}^{-1}(\widetilde{\boldsymbol{\theta}})\} + \kappa\mathrm{tr}\{\mathbf{K}(\widetilde{\boldsymbol{\theta}})\} \\ &= 2\kappa\mathrm{tr}\{\mathbf{K}(\widetilde{\boldsymbol{\theta}})\}.\end{aligned}$$

Thus $\kappa$ must satisfy $\kappa^2 + 1 \leq 2\kappa$, or $(\kappa-1)^2 \leq 0$. Therefore $\kappa = 1$ is the only solution and both terms must be equal at the minimum. We can thus re-write the minimization of (18) as the following problem

$$\begin{aligned} \min \quad & \alpha \\ \text{subject to} \quad & \operatorname{tr}\{\widetilde{\mathbf{K}}\mathbf{K}_{\theta}^{-1}\} = \alpha, \ \operatorname{tr}\{\mathbf{K}_{\theta}\} = \alpha, \end{aligned} \tag{19}$$

with minimizer $\widehat{\boldsymbol{\theta}}$ and where $\alpha > 0$ is an auxiliary variable.

Next, consider an equivalent problem to (19) obtained by re-defining the variables as $\tilde{\boldsymbol{\theta}} = \rho\alpha^{-1}\boldsymbol{\theta}$. Then $\operatorname{tr}\{\widetilde{\mathbf{K}}\mathbf{K}^{-1}(\boldsymbol{\theta})\} = \rho\alpha^{-1}\operatorname{tr}\{\widetilde{\mathbf{K}}\mathbf{K}^{-1}(\tilde{\boldsymbol{\theta}})\}$ and $\operatorname{tr}\{\mathbf{K}(\boldsymbol{\theta})\} = \alpha\rho^{-1}\operatorname{tr}\{\mathbf{K}(\tilde{\boldsymbol{\theta}})\}$, so that the equivalent problem becomes

$$\begin{aligned} \min \quad & \beta \\ \text{subject to} \quad & \operatorname{tr}\{\widetilde{\mathbf{K}}\mathbf{K}^{-1}\} = \beta, \ \operatorname{tr}\{\mathbf{K}\} = \rho, \end{aligned} \tag{20}$$

where $\beta = \alpha^2\rho^{-1}$. The minimizer of the equivalent problem (20) is therefore $\tilde{\boldsymbol{\theta}} \propto \boldsymbol{\theta}^{\star}$. Problem (20) is however identical to the constrained problem

$$\begin{aligned} \min \quad & \operatorname{tr}\{\widetilde{\mathbf{K}}\mathbf{K}^{-1}\} \\ \text{subject to} \quad & \operatorname{tr}\{\mathbf{K}\} = \rho, \end{aligned} \tag{21}$$

whose minimizer is $\tilde{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}$ when $\rho = \operatorname{tr}\{\widetilde{\mathbf{K}}\}$, which follows from expanding the cost in (11) and the normalization constraint.

Thus we proved that $\widehat{\boldsymbol{\theta}} \propto \boldsymbol{\theta}^{\star}$ and since the predictor is invariant to uniform scaling of $\boldsymbol{\theta}$, that is, $\widehat{y}_{\widehat{\theta}}(\mathbf{s}, t) = \widehat{y}_{\theta^{\star}}(\mathbf{s}, t)$, we see that the normalization constraint is not relevant for the result.

For further details, see [4].

## 2 Equivalent form of the predictor

Consider the following augmented problem

$$\min_{\eta,\, \mathbf{v},\, \boldsymbol{\theta}} \ \theta_0^{-1}\|\mathbf{y} - \mathbf{1}\eta - \boldsymbol{\Phi}\mathbf{v}\|_2^2 + \|\mathbf{v}\|_{\boldsymbol{\Theta}^{-1}}^2 + \operatorname{tr}\{\mathbf{K}_{\theta}\}. \tag{22}$$

Solving for $\eta$ and $\mathbf{v}$ yields the minimizer

$$\mathbf{w}^{\star} = \begin{bmatrix} \eta^{\star} \\ \mathbf{v}^{\star} \end{bmatrix} = \begin{bmatrix} (\mathbf{1}^{\top}\mathbf{K}^{-1}\mathbf{1})^{\dagger}\mathbf{1}^{\top}\mathbf{K}^{-1}\mathbf{y} \\ \boldsymbol{\Theta}\boldsymbol{\Phi}^{\top}\mathbf{K}^{-1}(\mathbf{y} - \mathbf{1}\eta^{\star}) \end{bmatrix}. \tag{23}$$

It can be shown that by inserting the minimizing $\mathbf{v}$ back into (22), we obtain a concentrated cost function which is equal to that in (18). Thus we obtain the sought model parameter $\widehat{\boldsymbol{\theta}}$ from the augmented problem.

Moreover, we can identify $\boldsymbol{\alpha}^{\top}(\mathbf{s}, t)\mathbf{w}^{*} = \widehat{y}_{\theta}(\mathbf{s}, t)$. Thus we obtain both $\boldsymbol{\theta}^{\star}$ and the weights $\mathbf{w}^{\star}$ from the augmented problem. Using these facts, we may alternatively solve for $\boldsymbol{\theta}$ first. The second and third terms in (22) can be written as

$$\|\mathbf{v}\|_{\boldsymbol{\Theta}^{-1}}^2 = \sum_{k=1}^{p} \frac{1}{\theta_k} w_{1+k}^2$$

and

$$\operatorname{tr}\{\mathbf{K}_\theta\} = \sum_{k=1}^{p} \|[\mathbf{\Phi}]_j\|_2^2 \theta_k + n\theta_0,$$

respectively. Then the minimizing hyperparameters $\boldsymbol{\theta}$ in (22) can be expressed in closed-form:

$$\widehat{\theta}_k^\star = \begin{cases} \|\mathbf{y} - [\mathbf{1} \;\; \mathbf{\Phi}]\mathbf{w}\|_2/\sqrt{n}, & k = 0, \\ |w_{1+k}|/\|[\mathbf{\Phi}]_{1+k}\|_2, & k = 1, \ldots, p. \end{cases}$$

Inserting the expression back in to (22) yields a concentrated cost function

$$\sqrt{\|\mathbf{y} - [\mathbf{1} \;\; \mathbf{\Phi}]\mathbf{w}\|_2^2} + \sum_{j=1}^{p} \frac{1}{\sqrt{n}} \|[\mathbf{\Phi}]_j\|_2 |w_{j+1}|$$

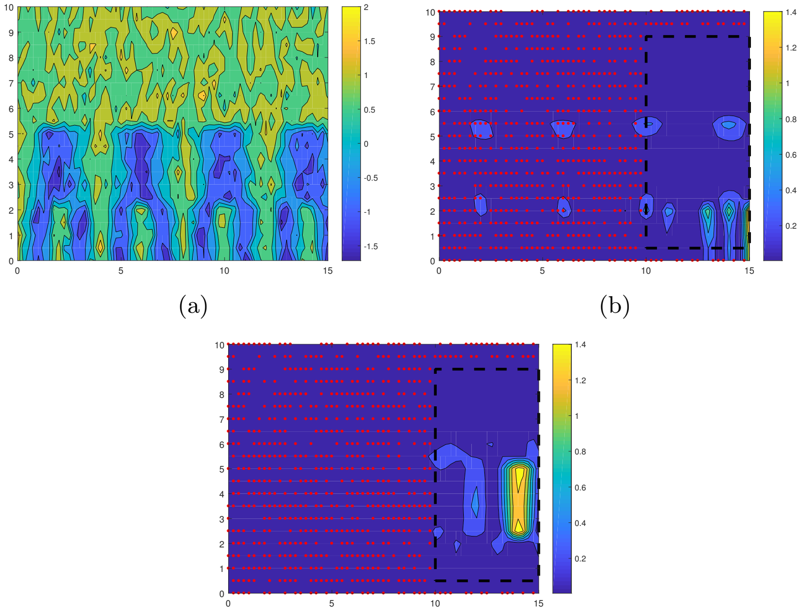
which, after dividing by $n^{-1/2}$, equals that in (14). Thus using minimizing weights $\mathbf{w}^*$, after concentrating the augmented problem with respect to $\boldsymbol{\theta}$, yields $\boldsymbol{\alpha}^\top(\mathbf{s}, t)\mathbf{w}^* = \widehat{y}_{\widehat{\theta}}(\mathbf{s}, t)$.

# 3 Comparison with Gaussian process using spectral mixture kernel

In section 5 (Synthetic data) of the paper, the predictive performance of the proposed method was compared with Gpr using the Matérn ARD covariance function. Here, we show results of comparison with a more expressive covariance function-the spectral mixture kernel used in [3] for the varying seasonalities across space example of section 5.2.

Data generation, proportion of training data and other parameters (number of basis, support) were same as described in section 5.2. The spectral mixture kernel implemented in [2] was used with $Q = 4$ number of spectral components per dimension. As in section 5.2, a contiguous space-time region was selected as a test region to emulate scenarios where data can be missing over large spatial region for some time. The test region is marked by black-dashed box in the figure.

Figure 1b and 1c show the Mean Square Error(MSE) of our proposed method and Gpr respectively. Although compared to Gpr with Matérn ARD (Figure 4*c* in the paper) the MSE in figure 1c is smaller outside the test region but inside the test region the MSE is still higher compared to the MSE of the proposed method.



(a)

(b)

(c)

Figure 1: (a) Realization of the process $y(s, t)$ defined in (15) with varying periods across space. (b) MSE of the proposed method which is able to learn different periodic temporal patterns across space. The red dots denote training points. The black dashed box marks a contiguous test region. (c) MSE of GPR using the spectral mixture kernel with $Q = 4$ spectral components per dimension.

## References

[1] Miguel Sousa Lobo, Lieven Vandenberghe, Stephen Boyd, and Hervé Lebret. Applications of second-order cone programming. *Linear algebra and its applications*, 284(1-3):193–228, 1998.

[2] Carl Edward Rasmussen and Hannes Nickisch. Gaussian processes for machine learning (gpml) toolbox. *Journal of Machine Learning Research*, 11(Nov):3011–3015, 2010.

[3] Andrew Wilson and Ryan Adams. Gaussian process kernels for pattern discovery and extrapolation. In *International Conference on Machine Learning*, pages 1067–1075, 2013.

[4] Dave Zachariah, Petre Stoica, and Thomas B Schön. Online learning for distribution-free prediction. *arXiv preprint arXiv:1703.05060*, 2017.