Product-moment covariance and PM_{10} outlier detection

2021-09-06

Product-moment covariance estimation

Raymaekers, J., & Rousseeuw, P. J. (2021). Fast robust correlation for high-dimensional data. *Technometrics*, 63(2), 184-198.

Estimation procedure

1. Obtain robust location and scale estimate $\hat{\mu}_j$ and $\hat{\sigma}_j$.
2. Transform x_{ij} to

$$x_{ij}^* = g(x_{ij}) = \hat{\mu}_j + \hat{\sigma}_j \psi_{b,c}\left(\frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}\right),$$

where

$$\psi_{b,c}(z) = \begin{cases} z, & 0 \le |z| < b, \\ q_1 \tanh\big(q_2(c - |z|)\big)\text{sign}(z), & b \le |z| < c, \\ 0, & c \le |z|. \end{cases}$$

3. Compute covariance matrix using transformed data x_{ij}^*.

Missing data

- Raymaekers(2021)는 missing이 존재할 경우, $\hat{\mu}_j$으로 대체하여 사용
- 지금처럼 missing이 많은 경우에 모두 $\hat{\mu}_j$으로 대체하는 경우에는 각 curve의 특성이 사라지며, 실제로 시뮬레이션 결과에서도 오히려 `M-est`보다 좋지 않았음
- 따라서 missing을 그대로 두고, 이 값들을 제외하여 covariance를 계산할 경우 `M-est`나 `GK` 보다 좋은 결과를 보여줌
- Imputation으로 L_2-distance가 가장 가까운 일부 curve들의 평균으로 imputation하여 covariance를 계산한 결과가 NA를 제외하고 한 경우보다 약간 더 좋았음
  - 예를 들어, 1st curve의 missing을 제외한 부분과의 distance가 가까운 curve들을 순서대로 나열한 후, 차례대로 missing인 부분에 대해서만 `colMeans`를 하며 NA가 포함되지 않을 때까지의 개수만을 사용하여 평균 계산

Simulation results

- 50 simulations
- 모두 noise variance를 고려하였으며, Yao et al.(2005) 방법으로 계산
- 비교 방법론
  - `Mest` : 기존의 proposed method
  - `GK` : Gnanadesikan and Kettenring (1972) method (박연주 교수님 코드 사용)
  - `PM` : Product-moment method proposed by Raymaekers (2021)
  - `PM-NA` : `PM` 방법에 NA를 그대로 두고 이를 제외하여 covariance를 계산한 방법
  - `PM-Im` : `PM` 방법에서 전체 평균 imputation 대신, distance가 가까운 일부만을 사용한 평균으로 imputation한 방법

Delaigle setting

Method	PVE	Reconstruction	Completion	Eigenfunction
Mest	0.82	0.18 (0.03)	0.40 (0.13)	0.17 (0.11)
Mest-sm	0.97	0.15 (0.03)	0.32 (0.11)	0.14 (0.11)
GK	0.86	0.17 (0.03)	0.38 (0.12)	0.16 (0.11)
GK-sm	0.98	0.15 (0.02)	0.31 (0.10)	0.14 (0.11)
PM	0.86	0.16 (0.03)	0.37 (0.12)	0.16 (0.10)
PM-sm	0.96	0.15 (0.03)	0.33 (0.11)	0.15 (0.10)
PM-NA	0.87	0.15 (0.03)	0.33 (0.11)	0.14 (0.09)
PM-sm-NA	0.97	0.14 (0.02)	0.29 (0.10)	0.12 (0.09)
PM-Im	0.88	0.15 (0.03)	0.31 (0.10)	0.13 (0.10)
PM-sm-Im	0.97	0.14 (0.02)	0.28 (0.10)	0.12 (0.10)

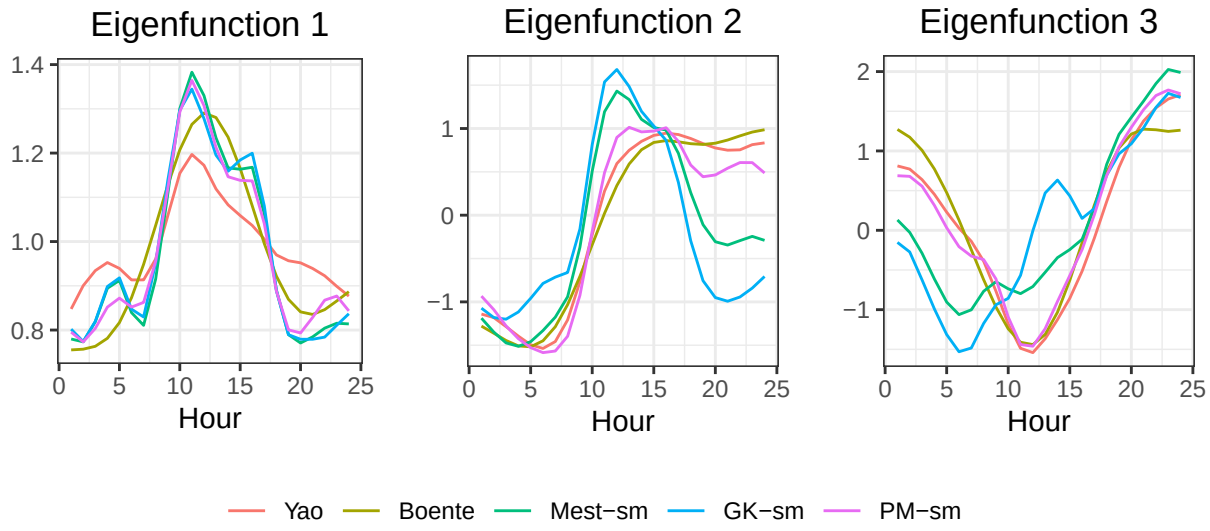
Boente setting

Method	PVE	Reconstruction	Completion	Eigenfunction
Mest	0.84	0.36 (0.07)	0.49 (0.17)	0.94 (0.05)
Mest-sm	0.92	0.32 (0.07)	0.42 (0.16)	0.94 (0.05)
GK	0.88	0.39 (0.09)	0.58 (0.21)	0.89 (0.08)
GK-sm	0.95	0.38 (0.09)	0.54 (0.20)	0.89 (0.08)
PM	0.89	0.34 (0.08)	0.52 (0.21)	0.93 (0.06)
PM-sm	0.91	0.33 (0.08)	0.50 (0.19)	0.93 (0.06)
PM-NA	0.92	0.29 (0.06)	0.35 (0.11)	0.92 (0.06)
PM-sm-NA	0.95	0.28 (0.06)	0.34 (0.10)	0.92 (0.06)
PM-Im	0.94	0.26 (0.05)	0.27 (0.07)	0.94 (0.05)
PM-sm-Im	0.96	0.25 (0.05)	0.27 (0.07)	0.94 (0.05)

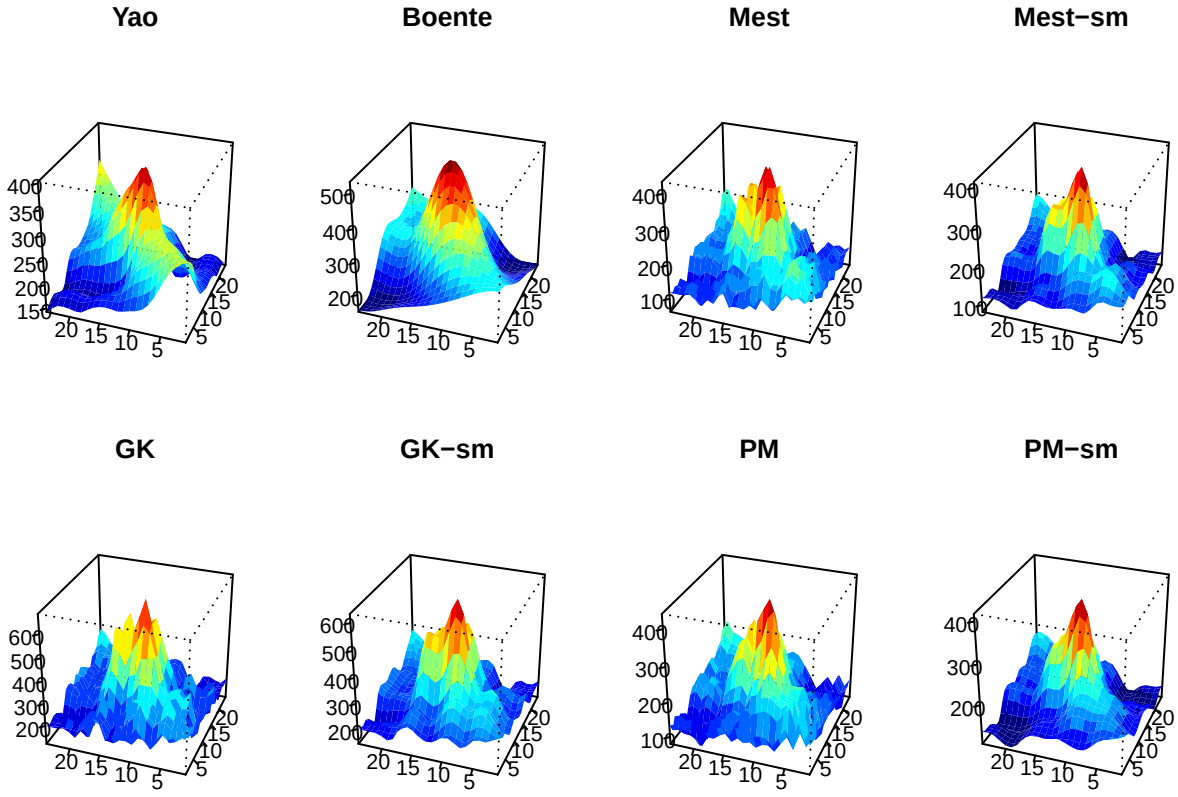
PM_{10} outlier detection

Region 1

Eigenfunctions

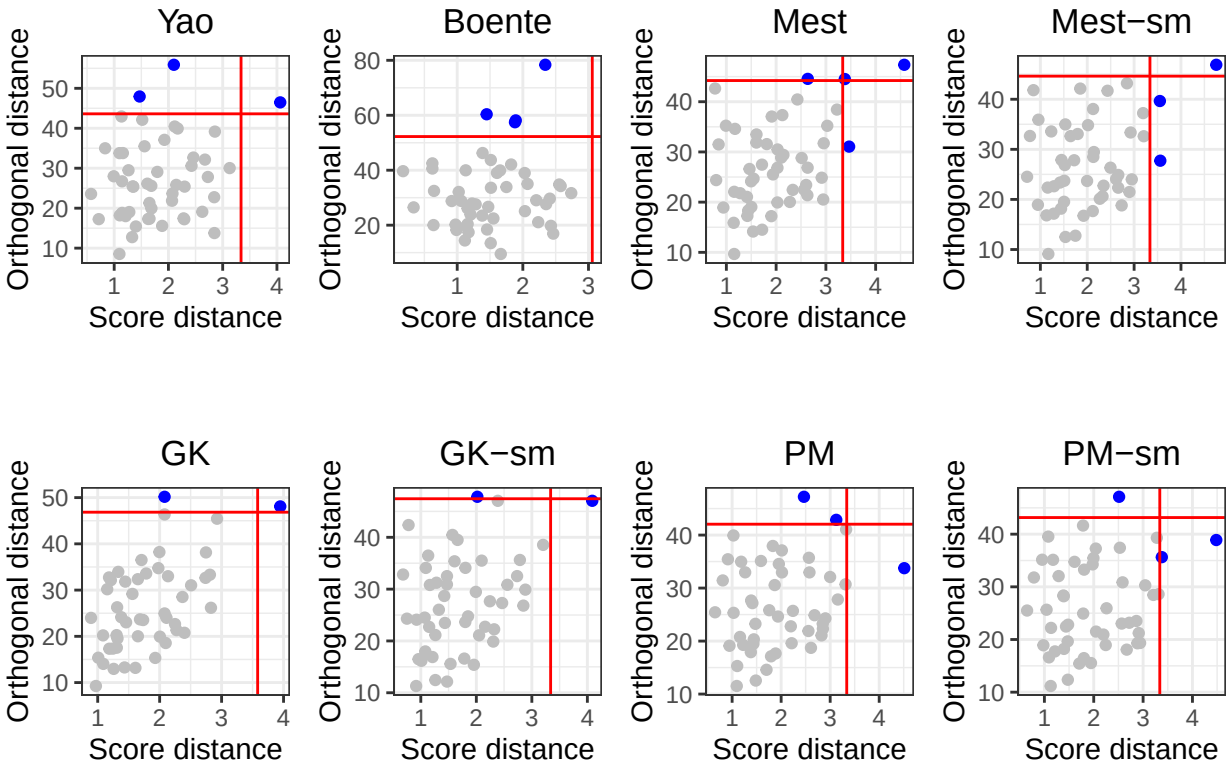


Covariance surfaces



Outlier map

- Score distance vs Orthogonal distance plot



Outlier detection

- 4가지 outlier detection 고려 (1~3의 경우, completion 후에 적용)
  1. robMah : the outlier detection method corresponds to the approach of Rousseeuw and Leroy (1987) using the robust Mahalanobis distance.
  2. LRT : the outlier detection method corresponds to the approach of Febrero et al. (2007) using the likelihood ratio test.
  3. HU : the outlier detection method corresponds to the approach of Hyndman and Ullah (2008) using the integrated square forecast errors.
  4. PCA-dist : Outlie map에서 1사분면에 해당하는 curve를 outlier로 결정

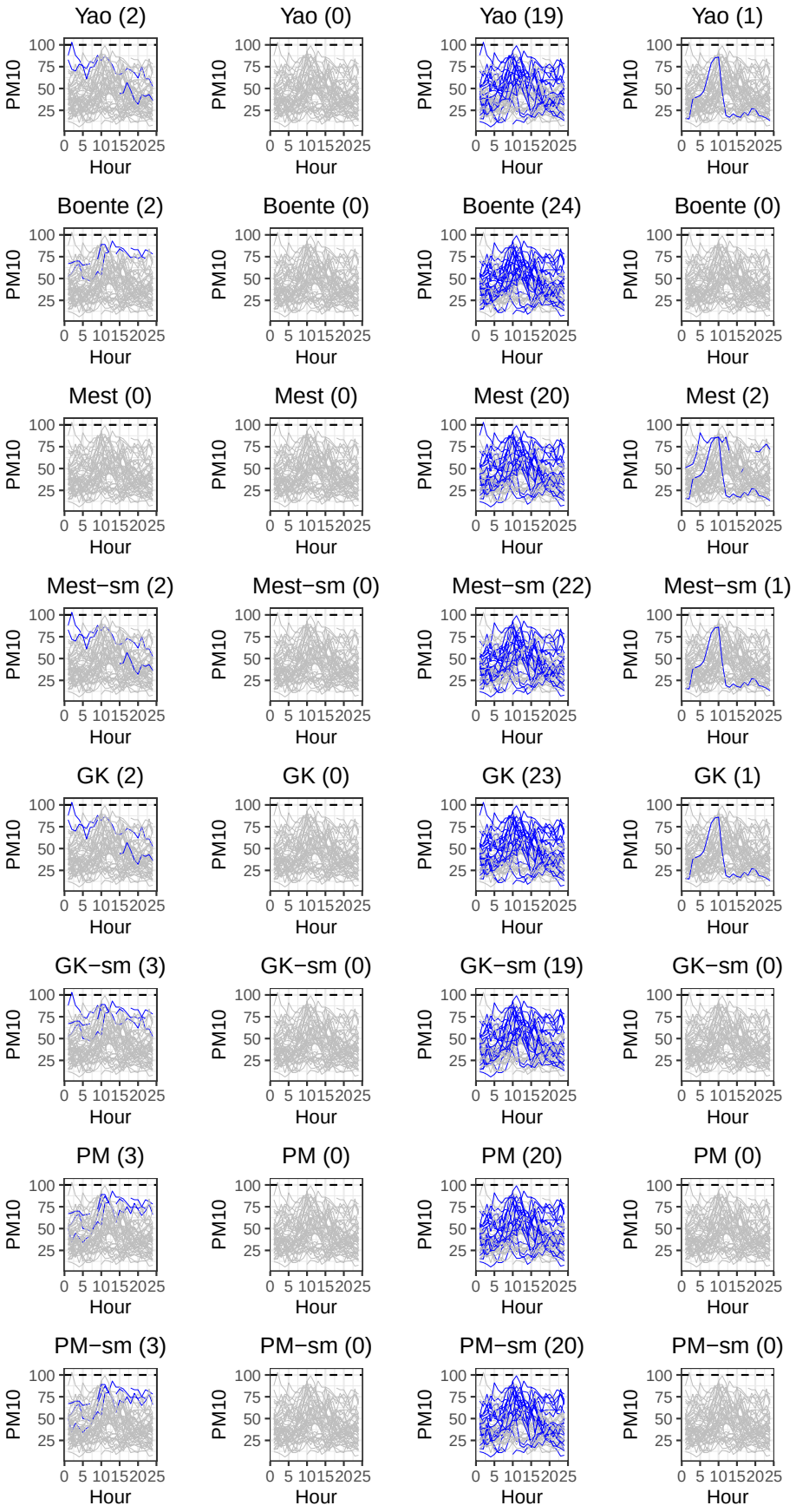
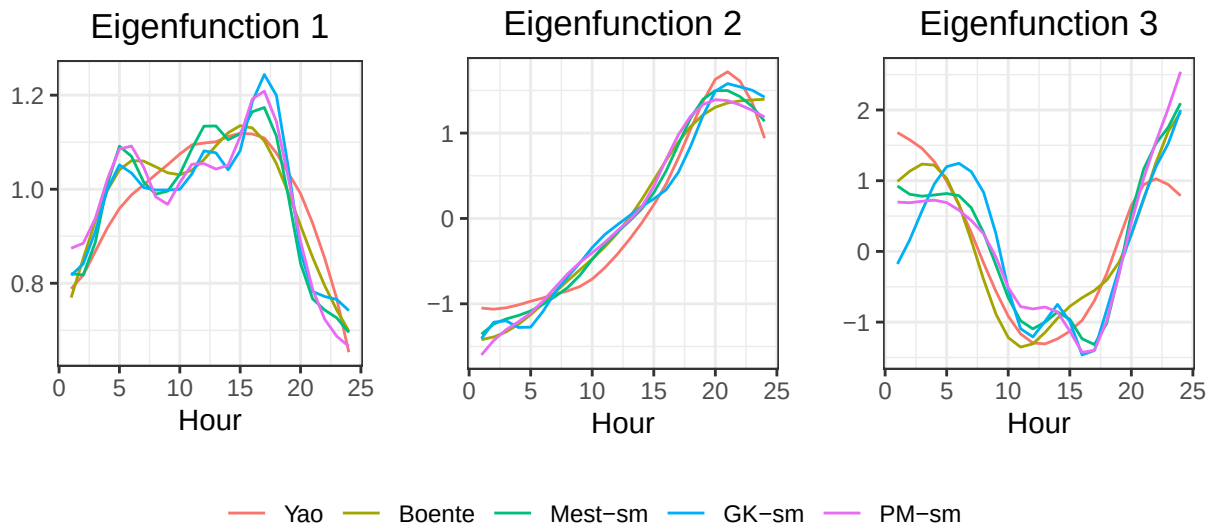


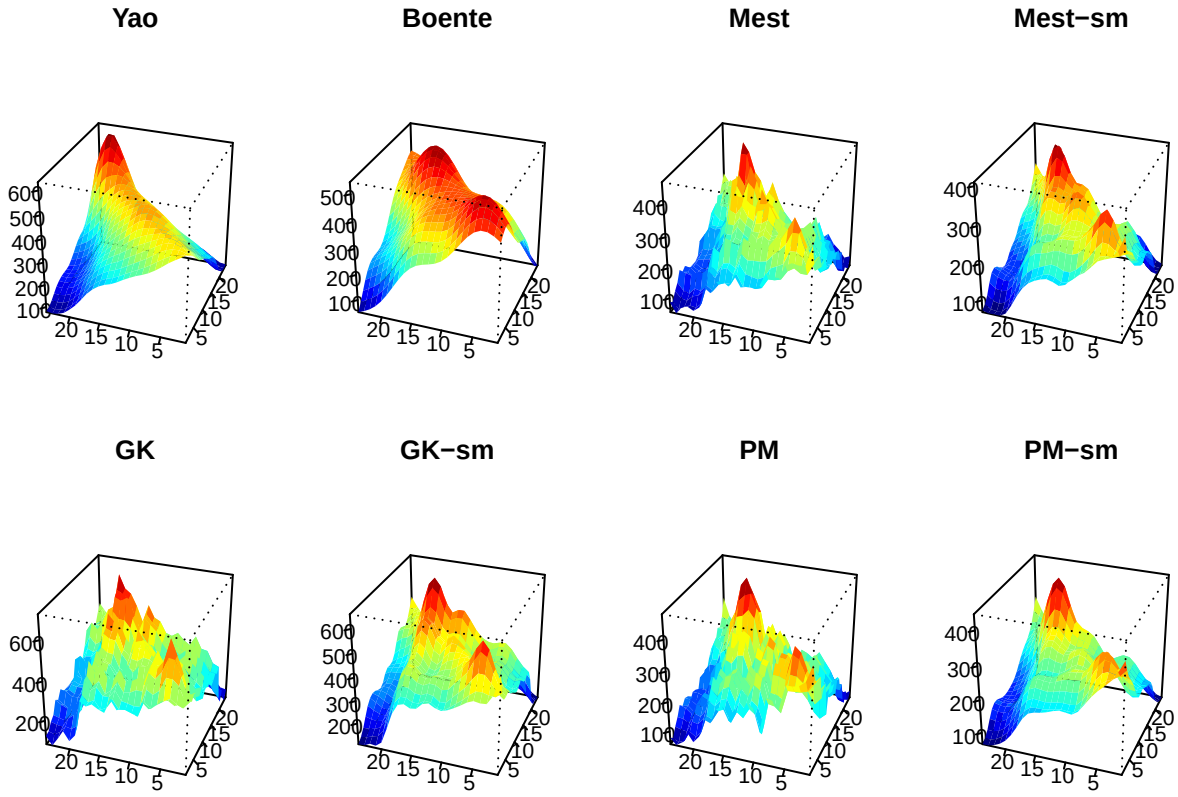
Figure 1: From the left, robMah, LRT, HU and PCA-dist, respectively.

Region 2

Eigenfunctions

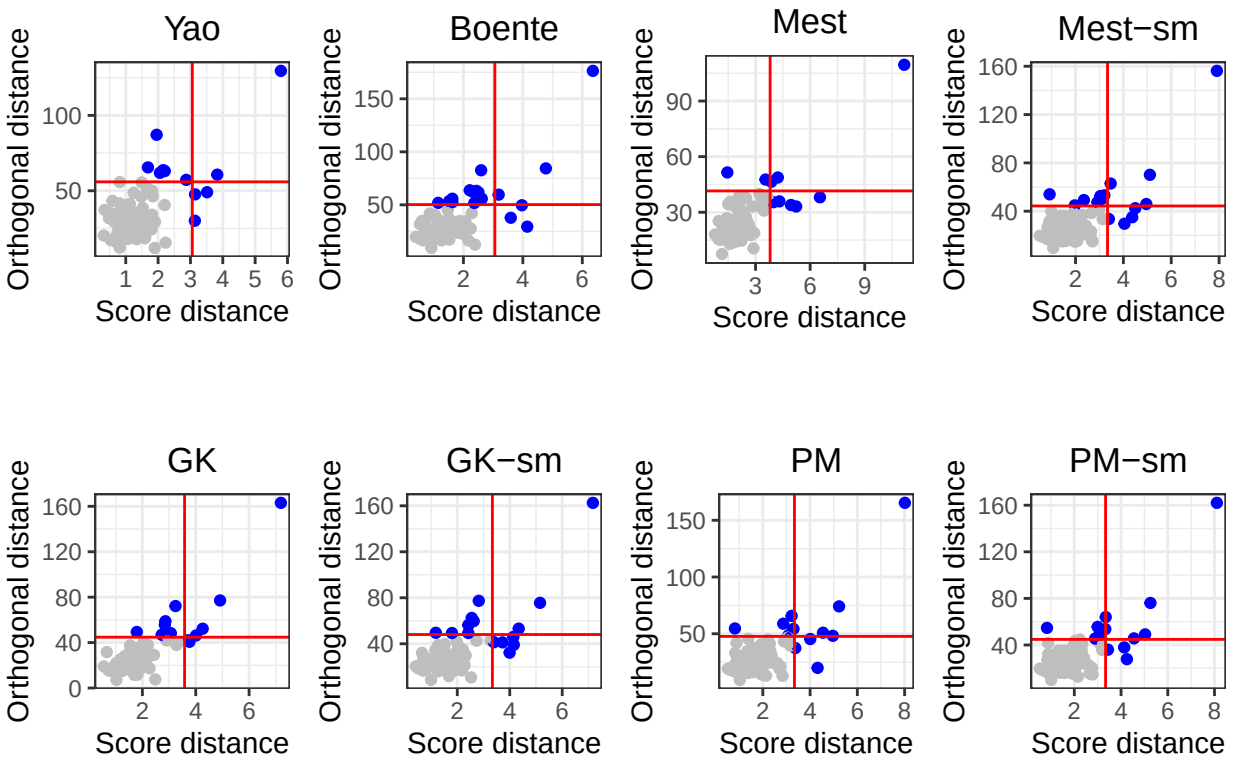


Covariance surfaces



Outlier map

- Score distance vs Orthogonal distance plot



Outlier detection

- 4가지 outlier detection 고려 (1~3의 경우, completion 후에 적용)
  1. robMah : the outlier detection method corresponds to the approach of Rousseeuw and Leroy (1987) using the robust Mahalanobis distance.
  2. LRT : the outlier detection method corresponds to the approach of Febrero et al. (2007) using the likelihood ratio test.
  3. HU : the outlier detection method corresponds to the approach of Hyndman and Ullah (2008) using the integrated square forecast errors.
  4. PCA-dist : Outlie map에서 1사분면에 해당하는 curve를 outlier로 결정

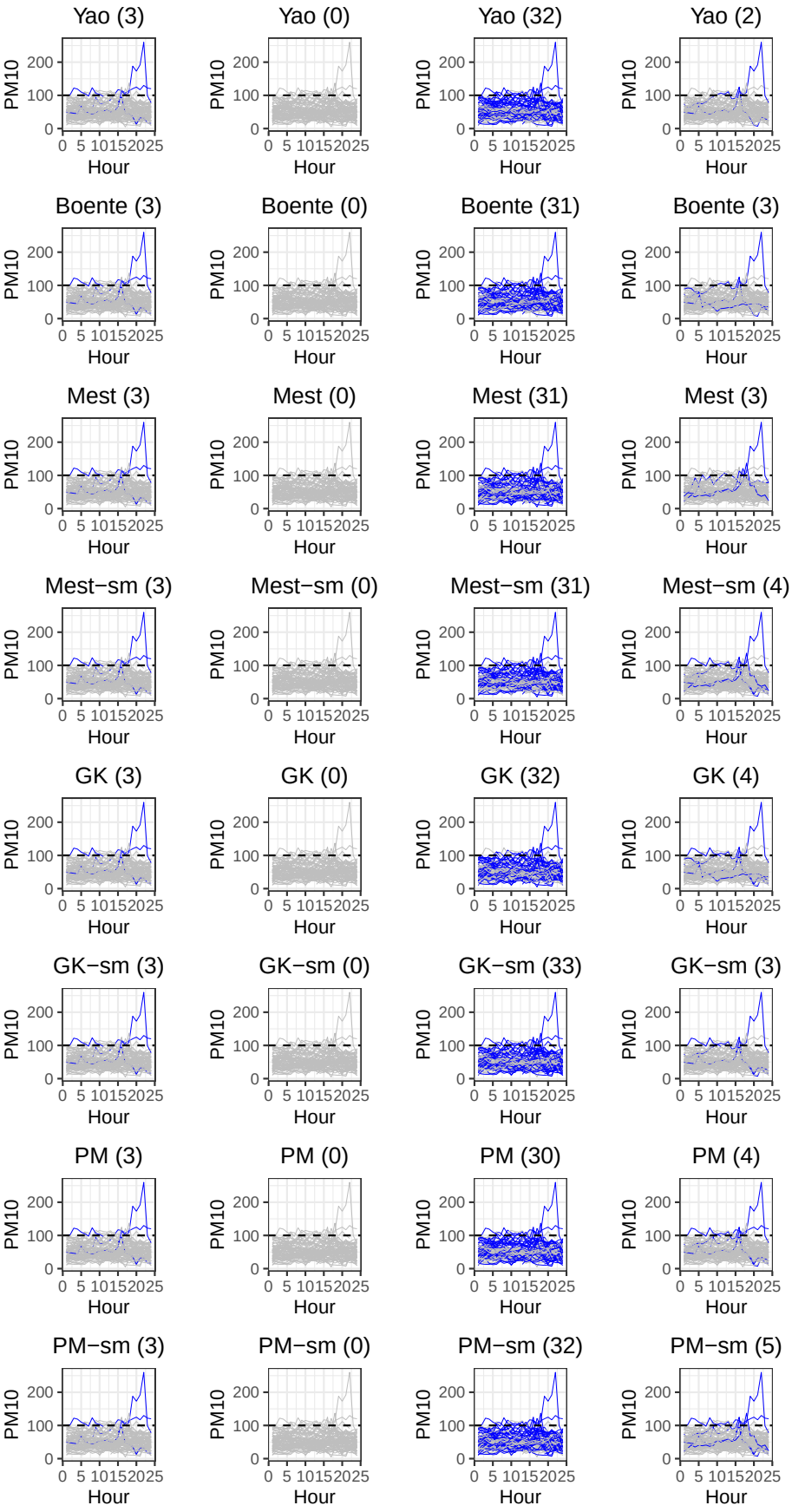
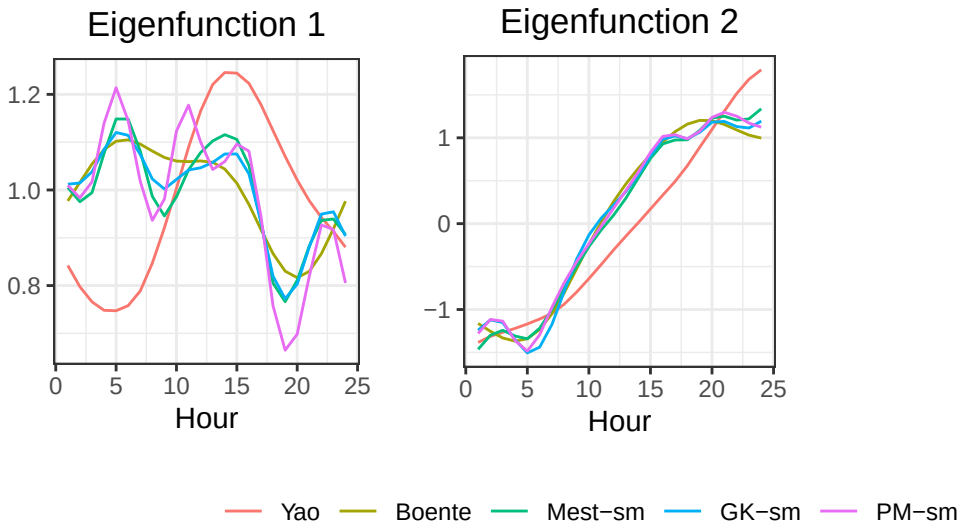


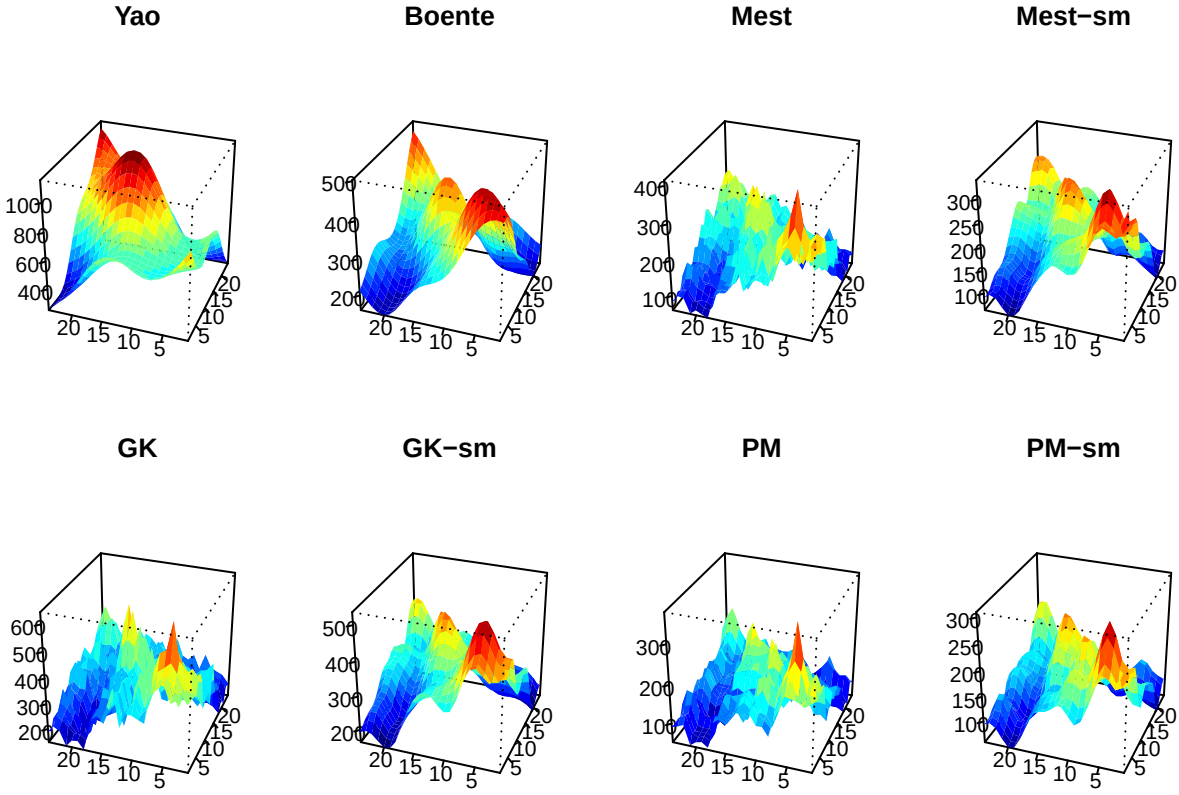
Figure 2: From the left, robMah, LRT, HU and PCA-dist, respectively.

Region 3

Eigenfunctions

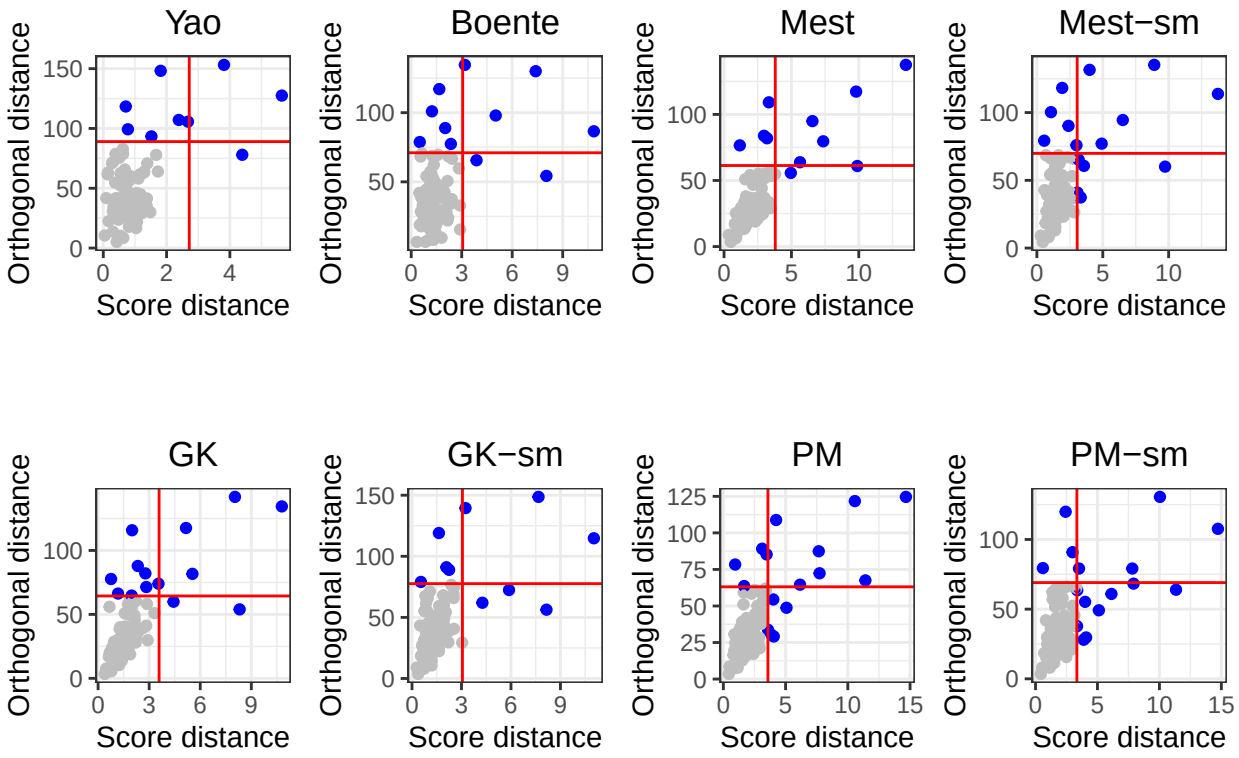


Covariance surfaces



Outlier map

- Score distance vs Orthogonal distance plot



Outlier detection

- 4가지 outlier detection 고려 (1~3의 경우, completion 후에 적용)
  1. robMah : the outlier detection method corresponds to the approach of Rousseeuw and Leroy (1987) using the robust Mahalanobis distance.
  2. LRT : the outlier detection method corresponds to the approach of Febrero et al. (2007) using the likelihood ratio test.
  3. HU : the outlier detection method corresponds to the approach of Hyndman and Ullah (2008) using the integrated square forecast errors.
  4. PCA-dist : Outlie map에서 1사분면에 해당하는 curve를 outlier로 결정

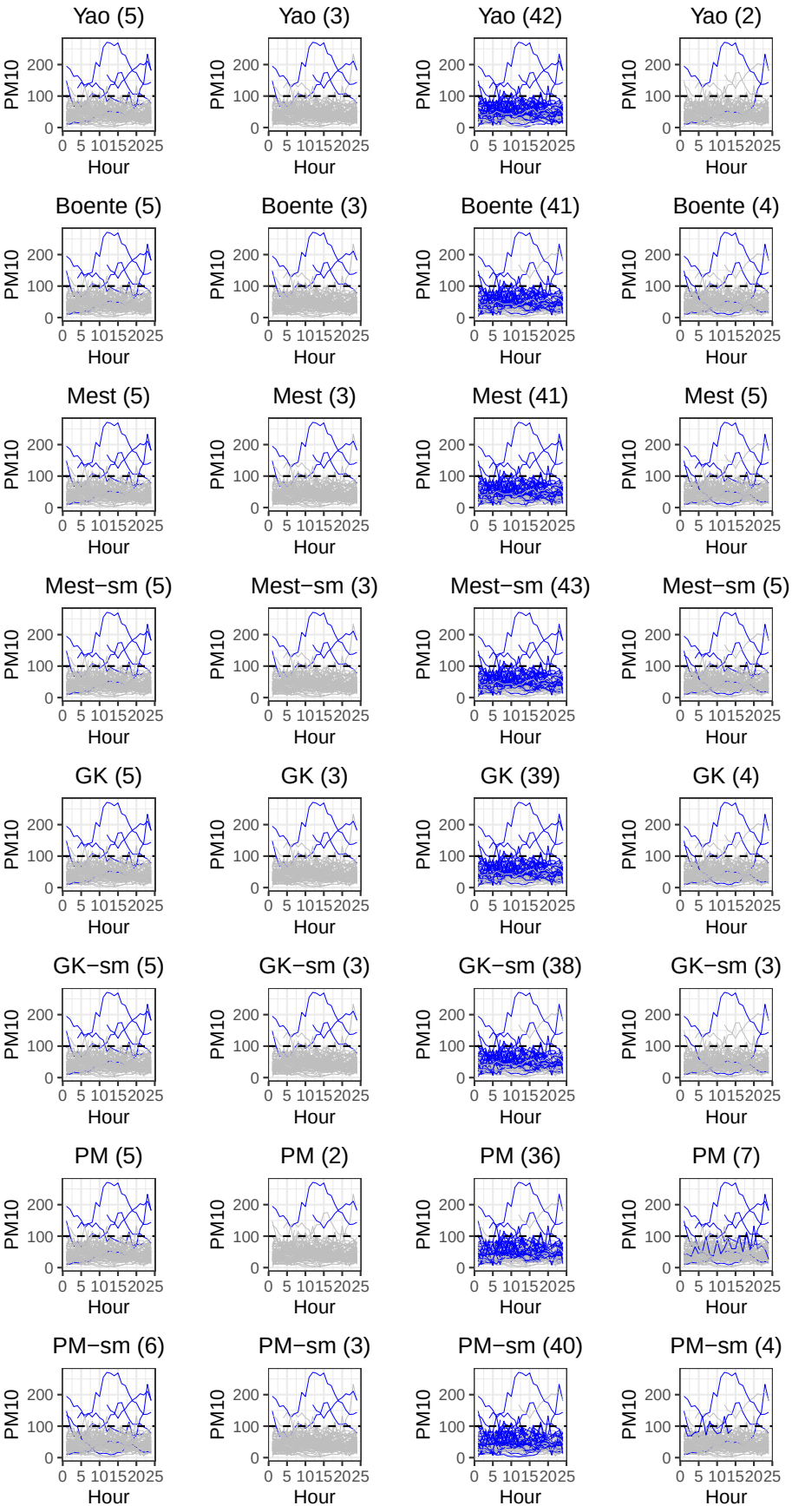
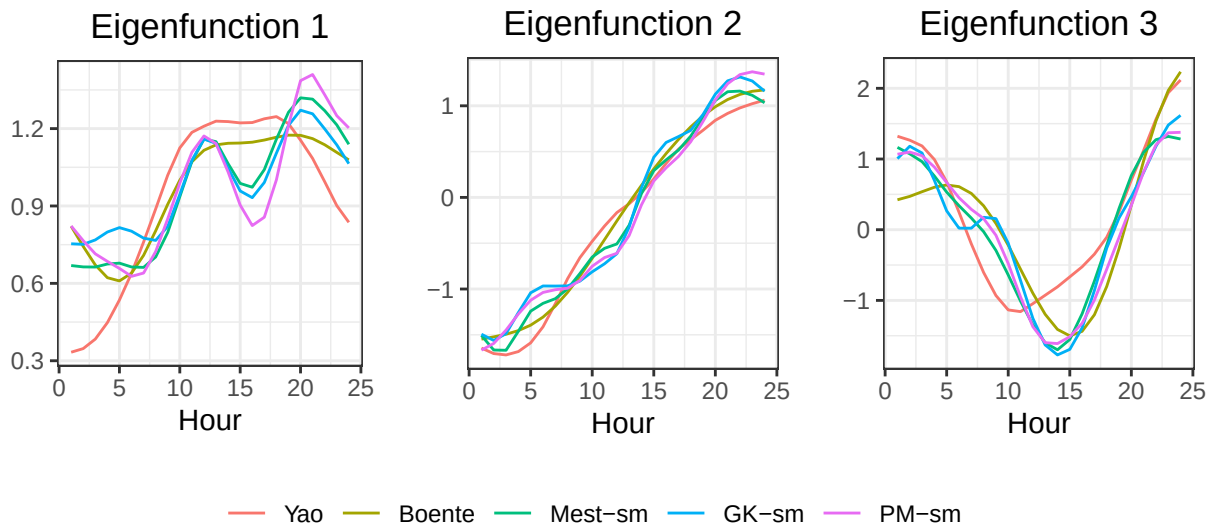


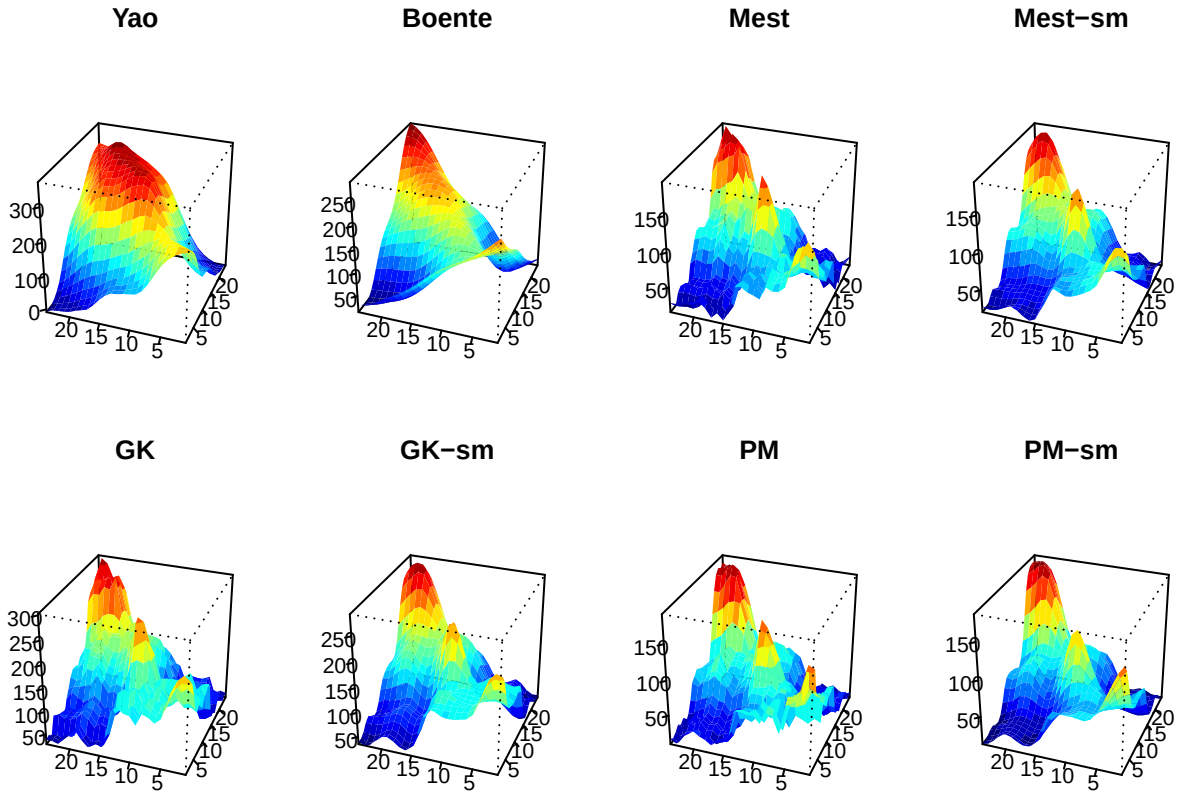
Figure 3: From the left, robMah, LRT, HU and PCA-dist, respectively.

Region 4

Eigenfunctions

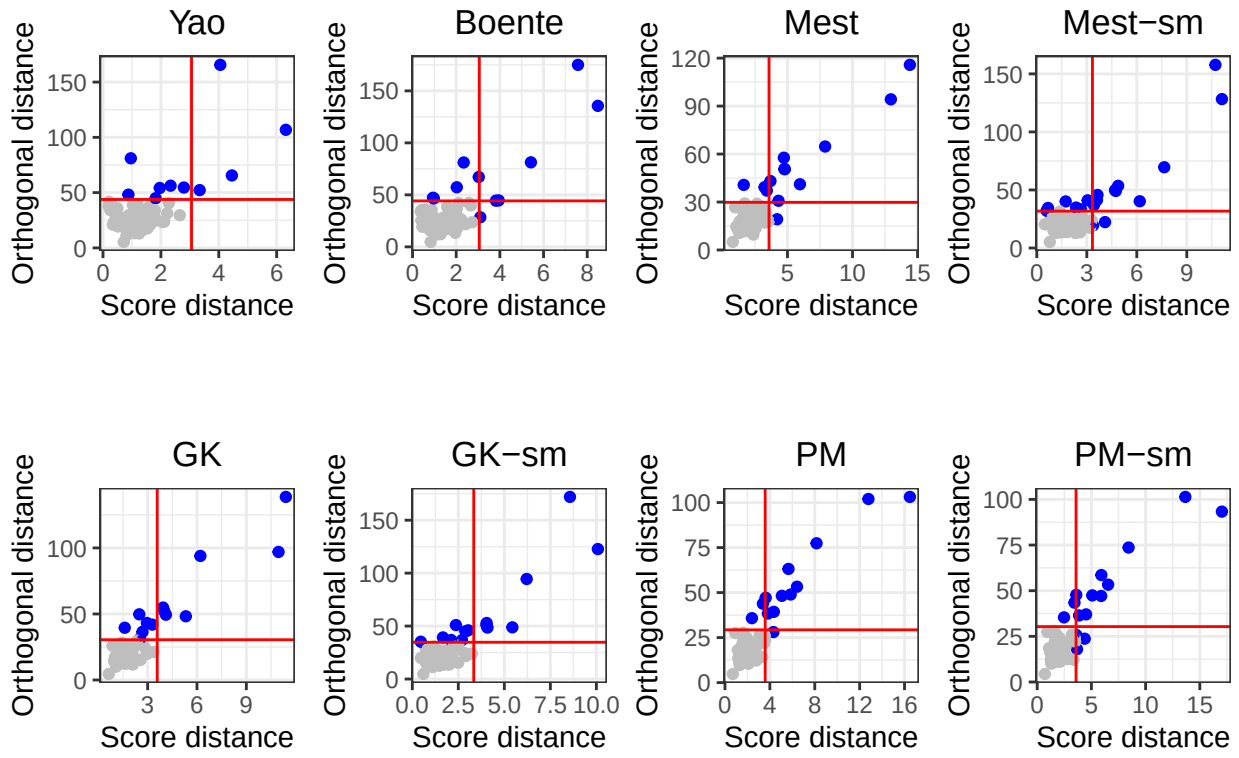


Covariance surfaces



Outlier map

- Score distance vs Orthogonal distance plot



Outlier detection

- 4가지 outlier detection 고려 (1~3의 경우, completion 후에 적용)
  1. robMah : the outlier detection method corresponds to the approach of Rousseeuw and Leroy (1987) using the robust Mahalanobis distance.
  2. LRT : the outlier detection method corresponds to the approach of Febrero et al. (2007) using the likelihood ratio test.
  3. HU : the outlier detection method corresponds to the approach of Hyndman and Ullah (2008) using the integrated square forecast errors.
  4. PCA-dist : Outlie map에서 1사분면에 해당하는 curve를 outlier로 결정

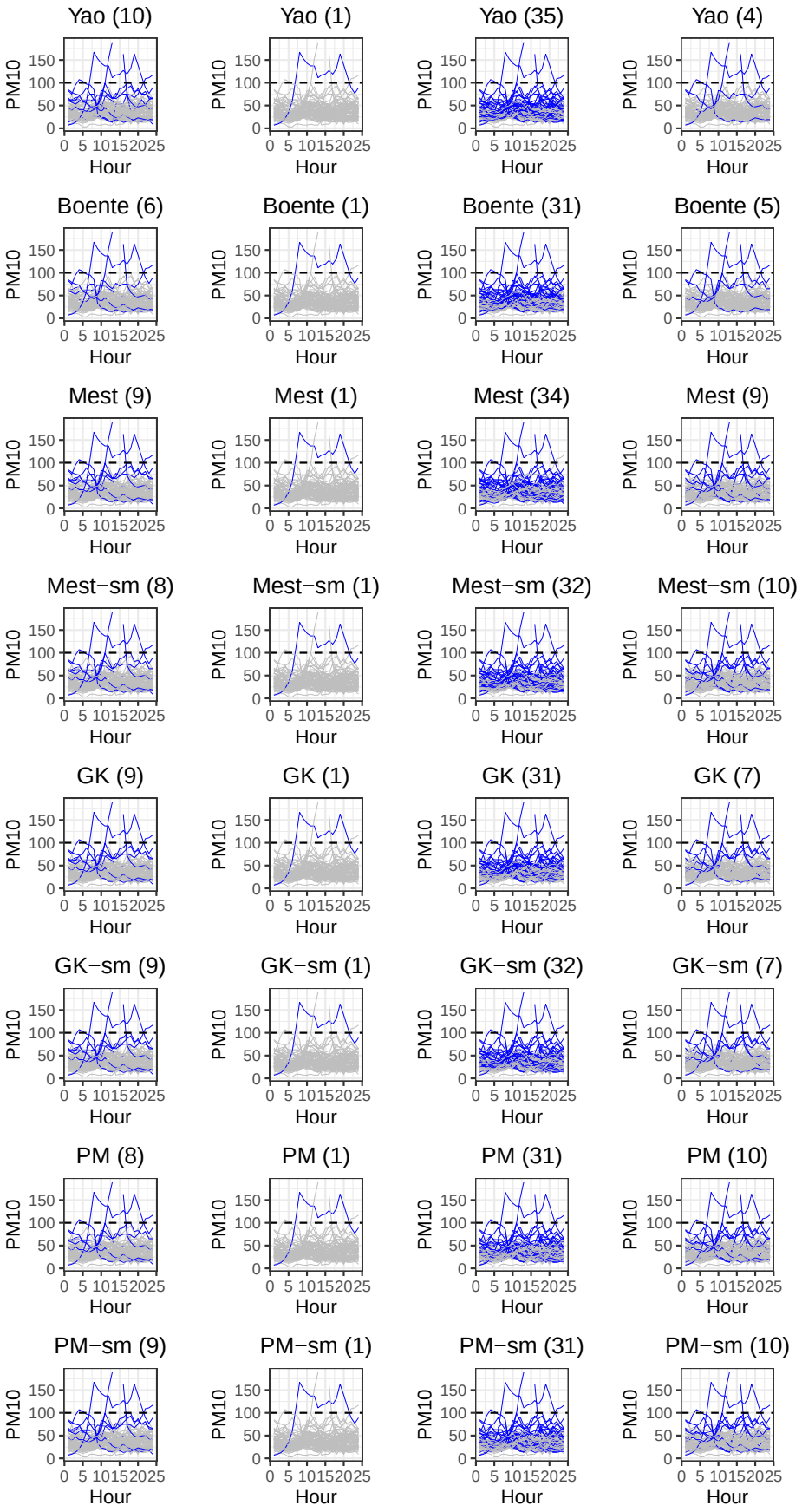


Figure 4: From the left, robMah, LRT, HU and PCA-dist, respectively.