Steven Xu · R&D IoT, Applied Analytical Solutions

About Me

- **Second year Statistics PhD student in NCSU**
- **2019 Summer Intern in Applied Analytical Solutions (ESP testing) group reporting to Yiqing Huang and Arin Chaudhuri.**

Contact Information

Steven.Xu@sas.com

sgxu@ncsu.edu

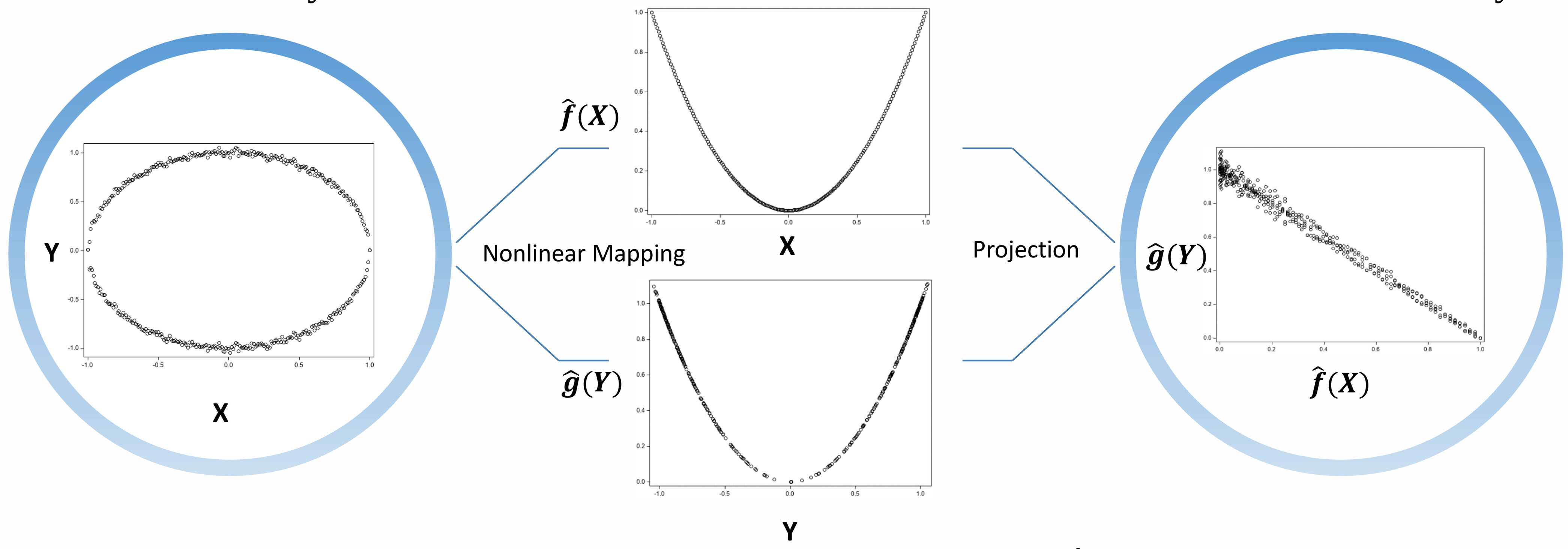
linkedin.com/in/steven-xu-397761115

Achievements

- **Researched on functional canonical correlation analysis, (kernel) canonical variate analysis.**
- **Implemented linear and kernel canonical variate analysis using SAS/IML.**
- **Applied implemented methods on multivariate time series anomaly detection.**
- **SAS Certificate: Certified Base Programmer**
- **Self-Study Event Stream Processing**

Canonical correlation analysis and its variants

Canonical correlation analysis (CCA) describes to what extend two sets of variables are linearly correlated. **Functional CCA** extends this idea to functional data: data that are collected from an underlying smooth process. It enables investigators to analyze the correlation between stochastic series, such as the temperature and precipitation of Canadian cities. CCA can be viewed as a preliminary tool before complex regression study. The limitation of CCA is that it only maximizes the correlation in the vector space. **Kernel CCA** explores maximal correlation in high dimensional feature space, where correlation between nonlinear transformed variables are more likely to be linear. It can reveal latent correlation that is otherwise undetectable by CCA.



Optimization detail for CCA (top) and KCCA

$$\max_{a,b\neq 0} \frac{\widehat{\mathrm{Cov}}[a^T X, b^T Y]}{\widehat{\mathrm{Var}}[a^T X]^{1/2}\widehat{\mathrm{Var}}[b^T Y]^{1/2}}$$

$$\max_{f\in\mathcal{H}_\mathcal{X}, g\in\mathcal{H}_\mathcal{Y}} \frac{\widehat{\mathrm{Cov}}[f(X), g(Y)]}{(\widehat{\mathrm{Var}}[f(X)]+\epsilon_n\|f\|^2_{\mathcal{H}_\mathcal{X}})^{1/2}(\widehat{\mathrm{Var}}[g(Y)]+\epsilon_n\|g\|^2_{\mathcal{H}_\mathcal{Y}})^{1/2}}$$

Figure 1: Example of kernel CCA revealing strong correlation between $\hat{f}(X)$ and $\hat{g}(Y)$, when CCA fails since X and Y are not correlated.

Canonical Variate Analysis

- Originates from state-space modeling
- Maximize correlation between past and future vectors
- Only linear mapping
- Less computation

Kernel method

Kernel Canonical Variate Analysis

- Accounts for nonlinearity in dynamic process
- Requires regularization
- Introduces extra parameters like kernel bandwidth need to be fine tuned

Anomaly Detection

- Abrupt and Incipient faults
- State and residual deviation
- Contributions from one or many processes
- Accuracy and Promptness

Analyze normal operation data $\vec{y}$

Construct past and future vectors
$\vec{y}_{p,k} = [y_{k-1}^T, y_{k-2}^T, \dots, y_{k-p}^T]^T$
$\vec{y}_{f,k} = [y_k^T, y_{k+1}^T, \dots, y_{k+f-1}^T]^T$

Construct variance-covariance matrices
$\boldsymbol{\Sigma}_{ff} = \mathbf{Y}_f\mathbf{Y}_f^T(M-1)^{-1}$
$\boldsymbol{\Sigma}_{pp} = \mathbf{Y}_p\mathbf{Y}_p^T(M-1)^{-1}$
$\boldsymbol{\Sigma}_{fp} = \mathbf{Y}_f\mathbf{Y}_p^T(M-1)^{-1}$

$\boldsymbol{\Sigma}_{ff}^{-1/2}\boldsymbol{\Sigma}_{fp}\boldsymbol{\Sigma}_{pp}^{-1/2} = \mathbf{UDV}^T$

Calculate transformation matrices
$\mathbf{J}_n = \mathbf{V}_n^T\boldsymbol{\Sigma}_{pp}^{-1/2}$, $\mathbf{F} = (\mathbf{I} - \mathbf{V}_n\mathbf{V}_n^T)\boldsymbol{\Sigma}_{pp}^{-1/2}$

Estimate state and residual variable
$\vec{x}_k = \mathbf{J}_n\vec{y}_{p,k}$, $\vec{e}_k = \mathbf{F}\vec{y}_{p,k}$

Calculate monitoring statistics
$\mathrm{T}_k^2 = \vec{x}_k^T\vec{x}_k$, $\mathrm{Q}_k = \vec{e}_k^T\vec{e}_k$

Calculate Upper Control Limit
$\mathrm{T}_{UCL}^2 = \hat{q}_{T^2}(1-\alpha)$, $\mathrm{Q}_{UCL} = \hat{q}_Q(1-\alpha)$

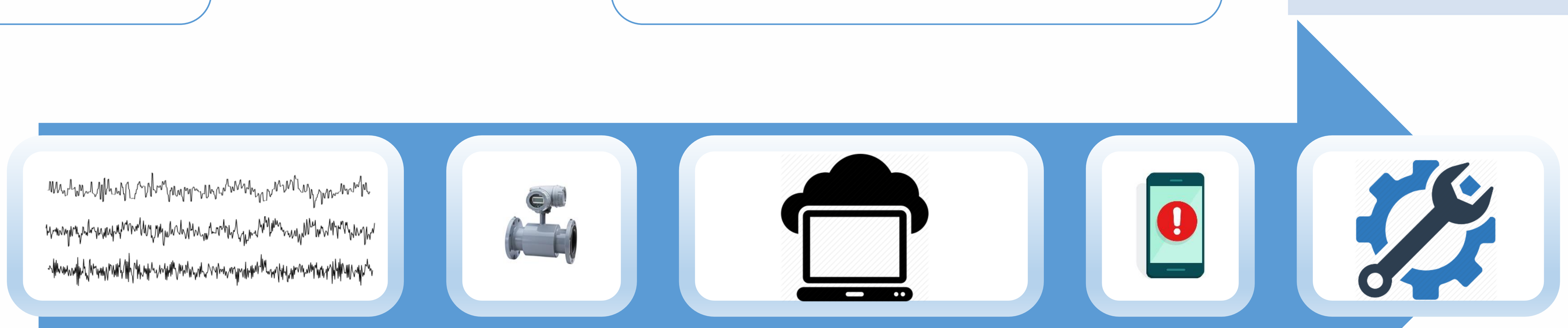


Figure 2: Illustration of online process monitoring and anomaly detection.

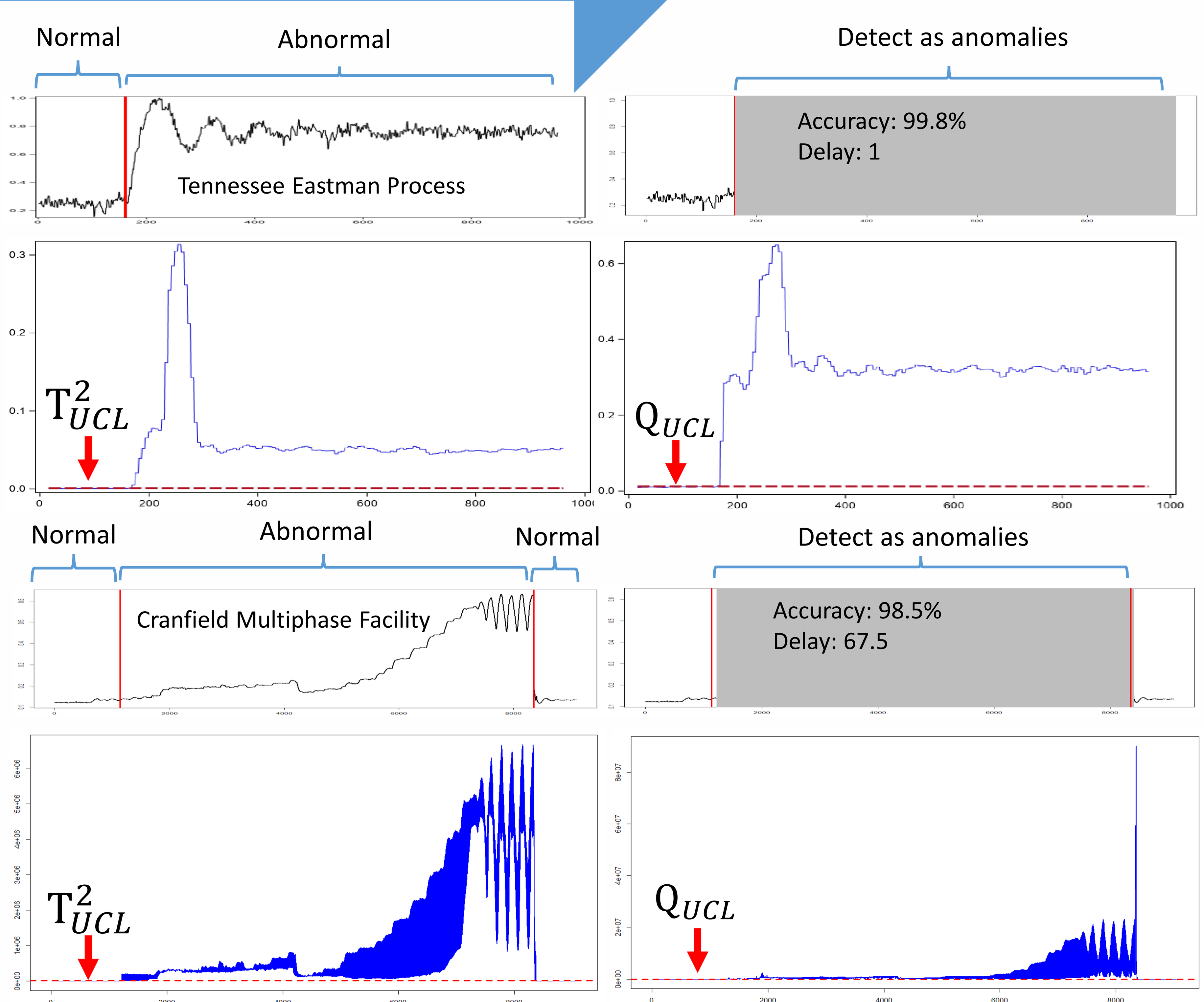
Collect real-time monitoring data $\vec{y}$

Construct past vectors
$\vec{y}_{p,k} = [y_{k-1}^T, y_{k-2}^T, \dots, y_{k-p}^T]^T$

Enter grace period

Potential Fault if
$(\mathrm{T}_k^2 > \mathrm{T}_{UCL}^2) \veebar (Q_k > Q_{UCL})$

Raise alarm if B consecutive potential faults



Tennessee Eastman Process

Fault	1	2	4	5	6	7	8	10	11	12	13	14	15	16	17	18	19	20
Accuracy	99.5%	98.8%	99.9%	99.8%	99.9%	99.9%	98.2%	97.5%	98.7%	99.2%	95.9%	99.7%	98.0%	98.7%	96.3%	94.7%	99.4%	95.9%

THE POWER TO KNOW.