Kernel Learning for SVM-based System Identification.

Matthew Higgs and John Shawe-Taylor

Center for Computational Statistics and Machine Learning
University College London

1. Outline

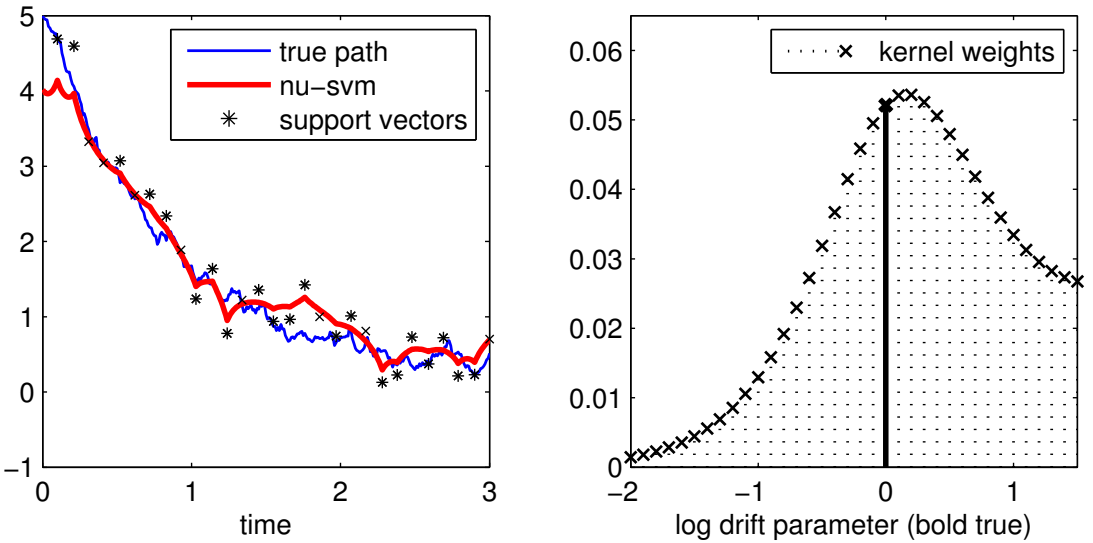
- Application of multiple kernel learning to the problem of system identification for temporal data.
- Aim: To learn the covariance function of a stochastic process.
- Computationally simple approach using out of the box machine learning methods.
- Proposed method for multidimensional, non-linear, non-stationary systems.

2. Simple ν-SVM Based Covariance Learning

- *Primitive* Ornstein-Uhlenbeck (OU) problem,

$$dX = -\gamma X dt + \sigma dW; \quad K(s,t) \propto \sigma^2 \exp(-\gamma|s-t|).$$

- Simple MKL ν-SVM applied to discrete time (noisy) observations of OU path



- Problem: Most dynamical systems are multidimensional, non-linear and/or non-stationary.

3. Vector-valued ν-SVM

- Kernel $\mathbf{K} : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}^d \times \mathbb{R}^d$ and ϵ-insensitive loss $L_{\epsilon,p}(\mathbf{y}, \hat{\mathbf{y}}) = \max(0, ||\mathbf{y} - \hat{\mathbf{y}}||_p - \epsilon)$.
- Vector valued ν-SVM dual form ($\bar{\gamma}_i = \gamma_i^+ - \gamma_i^-$)

$$\max_{\{\gamma_i^+, \gamma_i^-\}} \quad \sum_{i=1}^n \bar{\gamma}_i^\top \mathbf{y}_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \bar{\gamma}_i^\top \mathbf{K}(t_i, t_j) \bar{\gamma}_i$$

$$\text{subject to} \quad \sum_{i=1}^n \bar{\gamma}_i = 0; \quad \sum_{i=1}^n ||\gamma_i^{(\pm)}||_{\frac{p}{p-1}} \leq \nu; \quad ||\gamma_i^{(\pm)}||_{\frac{p}{p-1}} \leq \frac{C}{n}.$$

- Simple MKL still valid for $\mathbf{K}_{\boldsymbol{\eta}} = \sum_{j=1}^p \eta_j \mathbf{K}_j$,

$$\frac{\partial \text{DUAL}(\boldsymbol{\eta}, \{\gamma_i^+, \gamma_i^-\})}{\partial \eta_k} = -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \bar{\gamma}_i^\top \mathbf{K}_k(t_i, t_j) \bar{\gamma}_i.$$

4. GP Approximation of Non-Linear Systems

- Aim: Local GP approximation of a non-linear path.
- Non-stationary kernel, built from RBFs $\psi_i(t_i - t)$ at observation times $\{t_i\}$,

$$\mathbf{K}_{\boldsymbol{\Theta}}(s,t) = \sum_{j=1}^{p} \varphi_j(t|\boldsymbol{\Theta})\varphi_j(s|\boldsymbol{\Theta})\mathbf{K}_j(s,t),$$

$$\varphi_j(t|\boldsymbol{\Theta}) = \sum_{i=1}^{n} \Theta_{i,j}\psi_i(t_i - t); \quad \Theta_{i,j} \in \mathbb{R}_+; \quad \sum_{j=1}^{p} \Theta_{i,j} = 1.$$

- Equivalent to *Localised* MKL; optimisation non-trivial (work in progress), output $\widetilde{\boldsymbol{\Theta}}$, $\{\bar{\gamma}_i^*\}$, ϵ^*,

$$\boldsymbol{\mu}(\cdot) = \sum_{i=1}^{n} \bar{\gamma}_i^* \mathbf{K}_{\widetilde{\boldsymbol{\Theta}}}(t_i, \cdot).$$

5. Statistical Analysis

- GP $\mathbf{X}(\cdot) \sim GP(\boldsymbol{\mu}(\cdot), \mathbf{K}_{\widetilde{\boldsymbol{\Theta}}}(\cdot,\cdot))$ governed by *linear* SDE

$$d\mathbf{X}(t) = \widetilde{\mathbf{f}}_L(\mathbf{X}(t), t)dt + \Sigma d\mathbf{W}(t).$$

- *Any* Q governed by $d\mathbf{Z}(t) = \mathbf{g}(\mathbf{Z}(t), t)dt + \sqrt{\Sigma}d\mathbf{W}(t)$,

$$\mathrm{KL}(GP||Q) = \frac{1}{2}\int_0^T \mathop{\mathbb{E}}_{\mathbf{X}\sim GP(\boldsymbol{\mu},\mathbf{K}_{\widetilde{\boldsymbol{\Theta}}})} \left[||\widetilde{\mathbf{f}}_L(\mathbf{X}(t), t) - \mathbf{g}(\mathbf{X}(t), t)||_\Sigma^2\right] dt.$$

- Assume $\{(\mathbf{y}_i, t_i)\}$ iid; define

$$e_{GP}(\mathbf{y}, t) = \mathop{\mathbb{E}}_{\mathbf{X}\sim GP(\boldsymbol{\mu},\mathbf{K}_{\widetilde{\boldsymbol{\Theta}}})} [\mathbb{1}(L_{\epsilon^*,p}(\mathbf{y}, \mathbf{X}(t)) > 0)];$$

$$e_{GP} = \mathbb{E}_{(\mathbf{y},t)\sim D}[e_{GP}(\mathbf{y}, t)]; \quad \widehat{e}_{GP} = \frac{1}{n}\sum_{i=1}^{n}[e_{GP}(\mathbf{y}_i, t_i)].$$

With probability at least $1-\delta$ over the draw of the sample,

$$\mathrm{KL}_{Bernoulli}(\widehat{e}_{GP}||e_{GP}) \leq \frac{\mathrm{KL}(GP(\boldsymbol{\mu}, \mathbf{K}_{\widetilde{\boldsymbol{\Theta}}})||Q) + \ln\frac{n+1}{\delta}}{n}.$$

References


C. Archambeau, D. Cornford, M. Opper, J. Shawe-Taylor.
Gaussian Process Approximations of SDEs.
JMLR, 2007.


A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet.
Simple MKL.
JMLR, 2008.


M. Göonen, E. Alpaydin.
Localized Multiple Kernel Learning.
ICML, 2008.


M. Seeger.
PAC-Bayesian Generalization Error Bounds for Gaussian Process Classification.
JMLR, 2003.


C. A. Micchelli, M. Pontil.
On Learning Vector Valued Functions.
JMLR, 2003.