Pattern Information Processing
パターン情報処理

Masashi Sugiyama
(Department of Computer Science)
杉山　将（計算工学専攻）

Contact:　W8E-505
sugi@cs.titech.ac.jp
http://sugiyama-www.cs.titech.ac.jp/~sugi/

Learning Methods

- Overfitting
- Subset LS learning
- Sparseness
- Regularization
- Small exercise

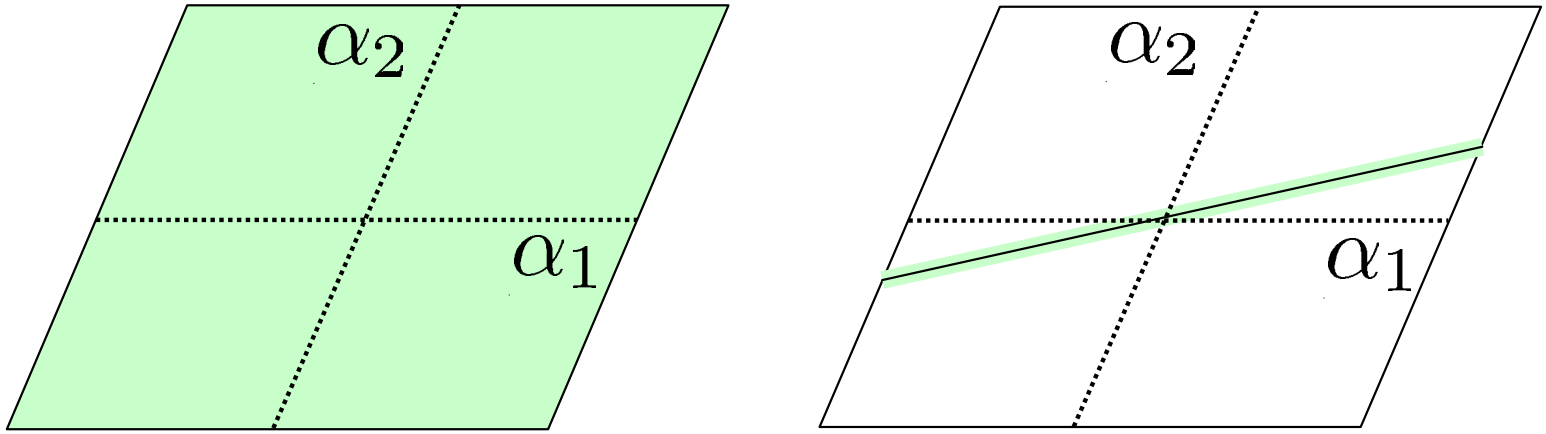
Over-fitting

- When noise is large, the learned function is over-fitted to noisy examples.
- In order to prevent this phenomenon, model should be restricted.

Subspace LS

- Restrict the search space within a subspace

$$\hat{\boldsymbol{\alpha}}_{SLS} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}}\, J_{LS}(\boldsymbol{\alpha})$$
$$\text{subject to } (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\alpha} = \boldsymbol{0}$$



Ordinary LS　　Subspace LS

- Solution: $\boldsymbol{L}_{SLS} = (\boldsymbol{P}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{P})^\dagger \boldsymbol{P}\boldsymbol{X}^\top$

$\dagger$: Moore-Penrose generalized inverse

Moore-Penrose Generalized Inverse

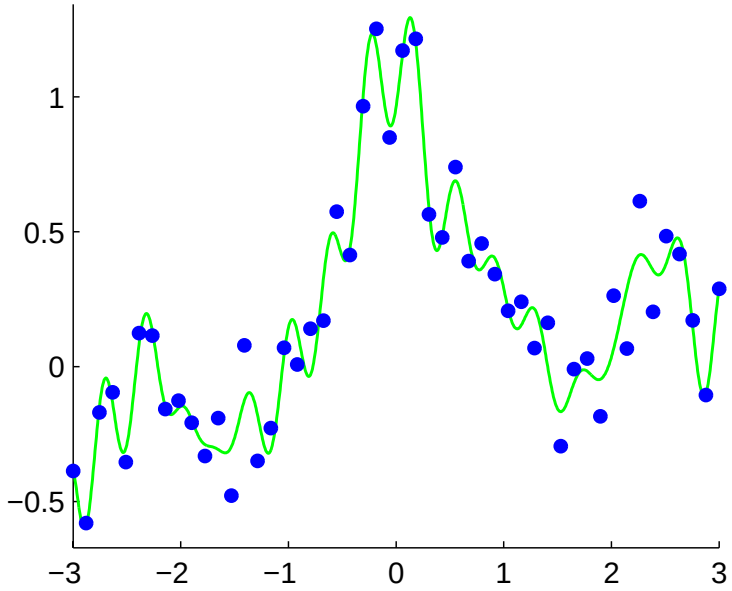
$$\boldsymbol{X} = \boldsymbol{A}^{\dagger} \overset{\text{def}}{\Longleftrightarrow} \begin{cases} \boldsymbol{A}\boldsymbol{X}\boldsymbol{A} &= \boldsymbol{A} \\ \boldsymbol{X}\boldsymbol{A}\boldsymbol{X} &= \boldsymbol{X} \\ (\boldsymbol{A}\boldsymbol{X})^{\top} &= \boldsymbol{A}\boldsymbol{X} \\ (\boldsymbol{X}\boldsymbol{A})^{\top} &= \boldsymbol{X}\boldsymbol{A} \end{cases}$$

$$\exists \boldsymbol{A}^{-1} \Longrightarrow \boldsymbol{A}^{\dagger} = \boldsymbol{A}^{-1}$$

$$\exists (\boldsymbol{A}^{\top}\boldsymbol{A})^{-1} \Longrightarrow \boldsymbol{A}^{\dagger} = (\boldsymbol{A}^{\top}\boldsymbol{A})^{-1}\boldsymbol{A}^{\top}$$

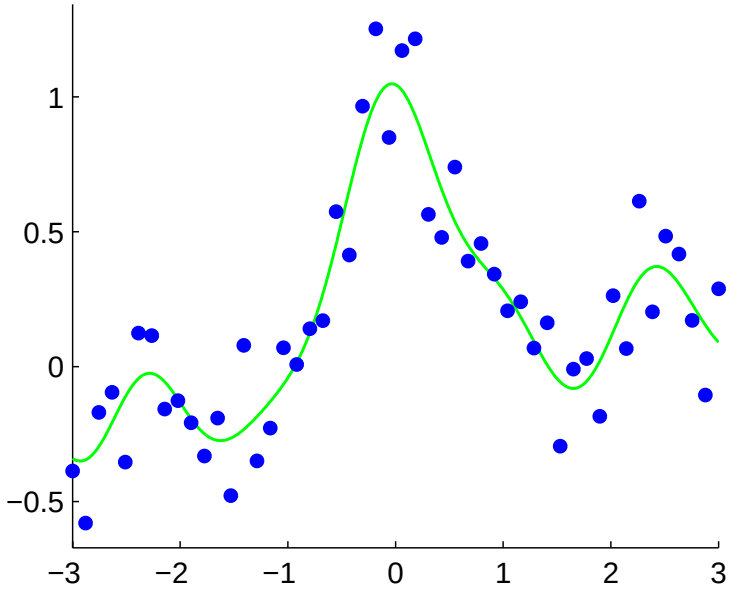
$$\exists (\boldsymbol{A}\boldsymbol{A}^{\top})^{-1} \Longrightarrow \boldsymbol{A}^{\dagger} = \boldsymbol{A}^{\top}(\boldsymbol{A}\boldsymbol{A}^{\top})^{-1}$$

Example of SLS



Full LS

$1, \ldots, \sin 15x, \cos 15x$



Subspace LS

$1, \ldots, \sin 5x, \cos 5x$

- Over-fit can be avoided by properly choosing the subspace.

Sparseness of Solution

- If the subspace is spanned by a subset of basis functions $\{\varphi_i(\boldsymbol{x})\}_{i=1}^{p}$, some of the parameters $\{\alpha_i\}_{i=1}^{p}$ are zero.
- Sparse solution is computationally advantageous in calculating the output values.

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{p} \alpha_i \varphi_i(\boldsymbol{x})$$

- However, the choice of subspaces is discrete.
- Combinatorial explosion! 2^p

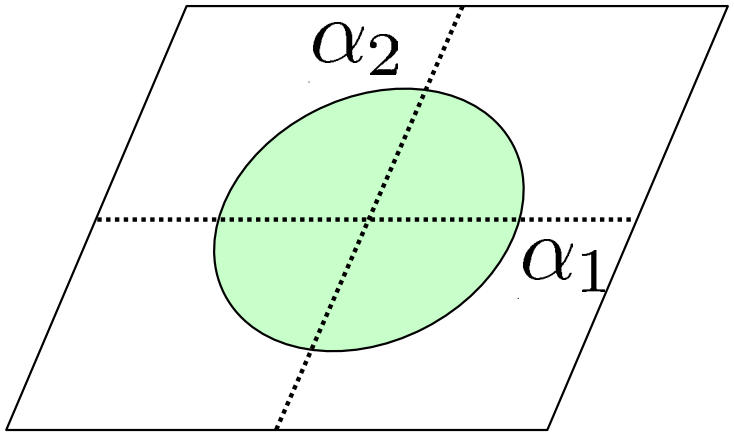
Quadratically Constrained LS

- Restrict the search space within a hyper-sphere.

$$\hat{\boldsymbol{\alpha}}_{QCLS} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} J_{LS}(\boldsymbol{\alpha})$$

$$\text{subject to } \|\boldsymbol{\alpha}\|^2 \leq C$$

$$C \geq 0$$



How to Obtain Solutions

- Lagrangian:

$$J_{QCLS}(\boldsymbol{\alpha}) = J_{LS}(\boldsymbol{\alpha}) + \lambda \|\boldsymbol{\alpha}\|^2$$

- $\lambda\ (\geq 0)$: Lagrange multiplier
- In practice, we start from $\lambda\ (\geq 0)$ and solve

$$\hat{\boldsymbol{\alpha}}_{QCLS} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}}\, J_{QCLS}(\boldsymbol{\alpha})$$

- Solution is given by

$$\boldsymbol{L}_{QCLS} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top$$

- It is often called quadratically regularized LS.

Interpretation of QCLS

- QCLS tries to avoid overfitting by adding penalty to the “goodness-of-fit” term.

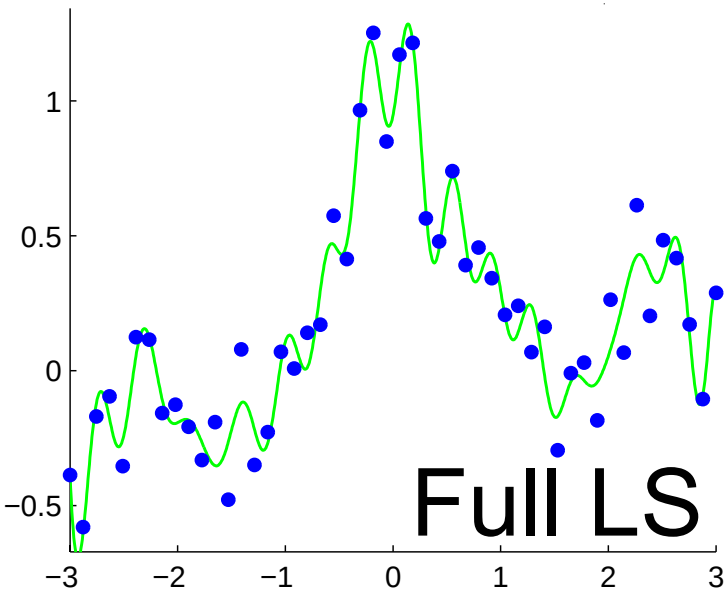
$$J_{QCLS}(\boldsymbol{\alpha}) = \underbrace{J_{LS}(\boldsymbol{\alpha})}_{\text{Good-ness of fit}} + \lambda \underbrace{\|\boldsymbol{\alpha}\|^2}_{\text{Penalty}}$$

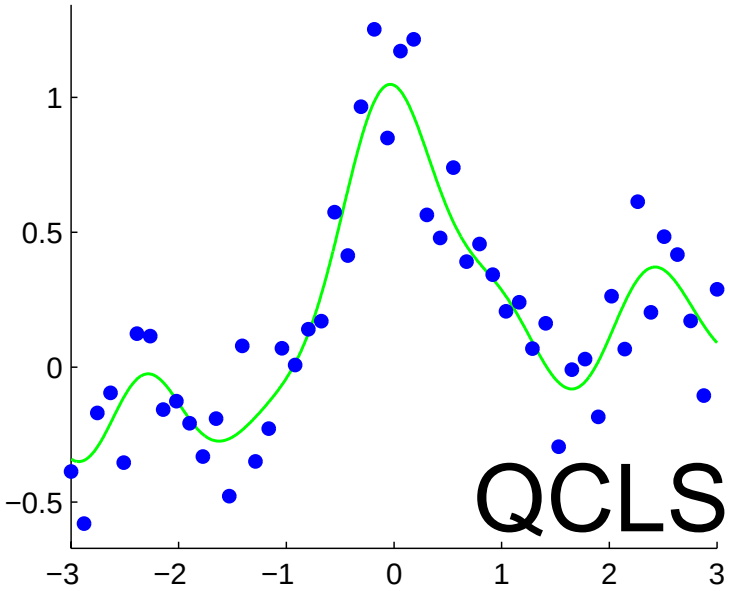
Example of QCLS

- Gaussian kernel model:

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i)$$

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2/2\right)$$





$(\lambda = 1)$

- Over-fit can be avoided by properly choosing the regularization factor.

Property of QCLS

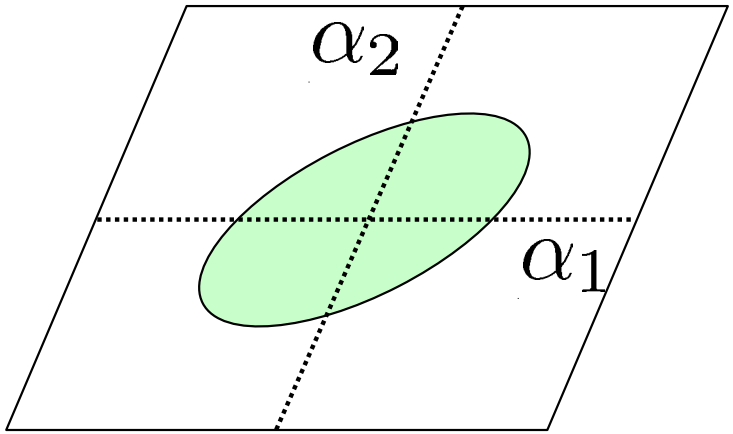
- Choice of models is continuous: λ
- However, solution is not generally sparse.

Generalization

- Restrict the search space within a hyper-ellipsoid.

$$\hat{\boldsymbol{\alpha}}_{QCLS} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}}\, J_{LS}(\boldsymbol{\alpha})$$

$$\text{subject to } \langle \boldsymbol{R}\boldsymbol{\alpha}, \boldsymbol{\alpha} \rangle \leq C$$



$C \geq 0$

$\boldsymbol{R}$:Positive semi-definite matrix

$\forall \boldsymbol{\alpha}, \langle \boldsymbol{R}\boldsymbol{\alpha}, \boldsymbol{\alpha} \rangle \geq 0$

$$\boldsymbol{L}_{QCLS} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{R})^{-1} \boldsymbol{X}^\top$$

Exercise

- Prove that the solution of

$$\hat{\boldsymbol{\alpha}}_{QCLS} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}}\, J_{LS}(\boldsymbol{\alpha}) \quad \text{subject to } \langle \boldsymbol{R}\boldsymbol{\alpha}, \boldsymbol{\alpha} \rangle \leq C$$

is given by

$$\boldsymbol{L}_{QCLS} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{R})^{-1} \boldsymbol{X}^\top$$