Towards Completely-Data-Driven Functional Estimation

Xiaoming Huo

School of Industrial & Systems Engineering
Georgia Institute of Technology, Atlanta

Banff International Research Station
Tuesday, December 13, 2011

Jointly work with Z. Yang & H. Xie & Y. Lu

Outline

Outline

Formulation

- Observe $(X_i, y_i), X_i \in \mathbb{R}^d, p \geq 1, y_i \in \mathbb{R}, i = 1, 2, \ldots, n$
- Relation
$$y_i = f(X_i) + \varepsilon_i,$$
or
$$y_i = f(X_{i1}, X_{i2}, \ldots, X_{id}) + \varepsilon_i,$$
where ε_i's are i.i.d. and satisfy some conditions
- Objective: estimating f
- Desideratum: $f \in \mathscr{F}$

Formulation

- Observe $(X_i, y_i), X_i \in \mathbb{R}^d, p \geq 1, y_i \in \mathbb{R}, i = 1, 2, \ldots, n$
- Relation

$$y_i = f(X_i) + \varepsilon_i,$$

or

$$y_i = f(X_{i1}, X_{i2}, \ldots, X_{id}) + \varepsilon_i,$$

where ε_i's are i.i.d. and satisfy some conditions

- Objective: estimating f
- Desideratum: $f \in \mathscr{F}$

Formulation

- Observe $(X_i, y_i), X_i \in \mathbb{R}^d, p \geq 1, y_i \in \mathbb{R}, i = 1, 2, \ldots, n$
- Relation
$$y_i = f(X_i) + \varepsilon_i,$$
or
$$y_i = f(X_{i1}, X_{i2}, \ldots, X_{id}) + \varepsilon_i,$$
where ε_i's are i.i.d. and satisfy some conditions
- Objective: estimating f
- Desideratum: $f \in \mathscr{F}$

Formulation

- Observe $(X_i, y_i), X_i \in \mathbb{R}^d, p \geq 1, y_i \in \mathbb{R}, i = 1, 2, \ldots, n$
- Relation
$$y_i = f(X_i) + \varepsilon_i,$$
or
$$y_i = f(X_{i1}, X_{i2}, \ldots, X_{id}) + \varepsilon_i,$$
where ε_i's are i.i.d. and satisfy some conditions
- Objective: estimating f
- Desideratum: $f \in \mathscr{F}$

Existing Methods

- Linear regression, $\mathscr{F}$ = a finite-dimensional linear subspace
- Kernel
- Splines
- Wavelets
- ...
- $\mathcal{F}$ is a Sobolev space

$$\mathcal{F} = H^m(\Omega) = \left\{ f : D^\alpha f \in L^2(\Omega), \forall \alpha \in \mathbb{Z}_+^d, |\alpha| \le m \right\}$$

where $\Omega \in \mathbb{R}^d$ is the domain of the function, for $\alpha \in \mathbb{Z}_+^d$, $\alpha = (\alpha_1, \alpha_2, ..., \alpha_d)$ and $|\alpha| = \sum_{i=1}^d \alpha_i$, we define the partial derivative

$$D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_d^{\alpha_d}}$$

Existing Methods

- Linear regression, $\mathscr{F}$ = a finite-dimensional linear subspace
- Kernel
- Splines
- Wavelets
- ...
- $\mathcal{F}$ is a Sobolev space

$$\mathcal{F} = H^m(\Omega) = \left\{ f : D^\alpha f \in L^2(\Omega), \forall \alpha \in \mathbb{Z}_+^d, |\alpha| \leq m \right\}$$

where $\Omega \in \mathbb{R}^d$ is the domain of the function, for $\alpha \in \mathbb{Z}_+^d$, $\alpha = (\alpha_1, \alpha_2, ..., \alpha_d)$ and $|\alpha| = \sum_{i=1}^d \alpha_i$, we define the partial derivative

$$D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_d^{\alpha_d}}$$

Existing Methods

- Linear regression, $\mathscr{F}$ = a finite-dimensional linear subspace
- Kernel
- Splines
- Wavelets
- ...
- $\mathcal{F}$ is a Sobolev space

$$\mathcal{F} = H^m(\Omega) = \left\{ f : D^\alpha f \in L^2(\Omega), \forall \alpha \in \mathbb{Z}_+^d, |\alpha| \leq m \right\}$$

where $\Omega \in \mathbb{R}^d$ is the domain of the function, for $\alpha \in \mathbb{Z}_+^d$, $\alpha = (\alpha_1, \alpha_2, ..., \alpha_d)$ and $|\alpha| = \sum_{i=1}^d \alpha_i$, we define the partial derivative

$$D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_d^{\alpha_d}}$$

Existing Methods

- Linear regression, $\mathscr{F}$ = a finite-dimensional linear subspace
- Kernel
- Splines
- Wavelets
- ...
- $\mathcal{F}$ is a Sobolev space

$$\mathcal{F} = H^m(\Omega) = \left\{ f : D^\alpha f \in L^2(\Omega), \forall \alpha \in \mathbb{Z}_+^d, |\alpha| \leq m \right\}$$

where $\Omega \in \mathbb{R}^d$ is the domain of the function, for $\alpha \in \mathbb{Z}_+^d$, $\alpha = (\alpha_1, \alpha_2, ..., \alpha_d)$ and $|\alpha| = \sum_{i=1}^d \alpha_i$, we define the partial derivative

$$D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_d^{\alpha_d}}$$

Existing Methods

- Linear regression, $\mathscr{F}$ = a finite-dimensional linear subspace
- Kernel
- Splines
- Wavelets
- ...
- $\mathcal{F}$ is a Sobolev space

$$\mathcal{F} = H^m(\Omega) = \left\{ f : D^\alpha f \in L^2(\Omega), \forall \alpha \in \mathbb{Z}_+^d, |\alpha| \leq m \right\}$$

where $\Omega \in \mathbb{R}^d$ is the domain of the function, for $\alpha \in \mathbb{Z}_+^d$, $\alpha = (\alpha_1, \alpha_2, ..., \alpha_d)$ and $|\alpha| = \sum_{i=1}^d \alpha_i$, we define the partial derivative

$$D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_d^{\alpha_d}}$$

Existing Methods

- Linear regression, $\mathscr{F}$ = a finite-dimensional linear subspace
- Kernel
- Splines
- Wavelets
- ...
- $\mathcal{F}$ is a Sobolev space

$$\mathcal{F} = H^m(\Omega) = \left\{ f : D^\alpha f \in L^2(\Omega), \forall \alpha \in \mathbb{Z}_+^d, |\alpha| \leq m \right\}$$

where $\Omega \in \mathbb{R}^d$ is the domain of the function, for $\alpha \in \mathbb{Z}_+^d$, $\alpha = (\alpha_1, \alpha_2, ..., \alpha_d)$ and $|\alpha| = \sum_{i=1}^d \alpha_i$, we define the partial derivative

$$D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_d^{\alpha_d}}$$

Review of Existing Approach

- Penalization estimation approach (equivalently, Lagrange multiplier):

$$\min_{f \in \mathscr{F}} G(f) + \lambda \cdot R(f)$$

where

- $G(f)$: goodness of fit, e.g., $G(f) = \sum_{i=1}^{n} [y_i - f(X_i)]^2$
- $R(f)$: regularity of f, e.g., $R(f) = \int_\Omega \sum_{i,j=1}^{p} \left(\frac{\partial^2 f}{\partial x_i \partial x_j} \right)^2$
- Ω: domain of f, e.g., $\Omega = \mathcal{R}^d$

Review of Existing Approach

- Penalization estimation approach (equivalently, Lagrange multiplier):

$$\min_{f \in \mathscr{F}} G(f) + \lambda \cdot R(f)$$

where

- $G(f)$: goodness of fit, e.g., $G(f) = \sum_{i=1}^{n} [y_i - f(X_i)]^2$
- $R(f)$: regularity of f, e.g., $R(f) = \int_\Omega \sum_{i,j=1}^{p} \left(\frac{\partial^2 f}{\partial x_i \partial x_j} \right)^2$
- Ω: domain of f, e.g., $\Omega = \mathcal{R}^d$

Review of Existing Approach

- Penalization estimation approach (equivalently, Lagrange multiplier):

$$\min_{f \in \mathscr{F}} G(f) + \lambda \cdot R(f)$$

where

- $G(f)$: goodness of fit, e.g., $G(f) = \sum_{i=1}^{n} [y_i - f(X_i)]^2$
- $R(f)$: regularity of f, e.g., $R(f) = \int_\Omega \sum_{i,j=1}^{p} \left(\frac{\partial^2 f}{\partial x_i \partial x_j} \right)^2$
- Ω: domain of f, e.g., $\Omega = \mathcal{R}^d$

Review of Existing Approach

- Penalization estimation approach (equivalently, Lagrange multiplier):

$$\min_{f \in \mathscr{F}} G(f) + \lambda \cdot R(f)$$

where

- $G(f)$: goodness of fit, e.g., $G(f) = \sum_{i=1}^{n} [y_i - f(X_i)]^2$
- $R(f)$: regularity of f, e.g., $R(f) = \int_\Omega \sum_{i,j=1}^{p} \left(\frac{\partial^2 f}{\partial x_i \partial x_j} \right)^2$
- Ω: domain of f, e.g., $\Omega = \mathcal{R}^d$

More on the Existing Approach

- Define a regularity functional. Recall

$$R(f) = \int_\Omega \|\mathbf{H}f\|_F^2$$

- Essence: Need to show that solving problem

$$\min_{f \in \mathscr{F}} \quad R(f)$$
$$\text{subject to} \quad f(X_i) = y_i'$$

ends up with a quadratic form:

$$R(f) = \mathbf{f}^T \mathbf{M} \mathbf{f}$$

where $\mathbf{f} = (f(X_1), f(X_2), \ldots, f(X_n))^T$

- If Ω is irregular, determining analytically the gram matrix $\mathbf{M}$ can be very difficult.

More on the Existing Approach

- Define a regularity functional. Recall

$$R(f) = \int_\Omega \|\mathbf{H} f\|_F^2$$

- Essence: Need to show that solving problem

$$\begin{aligned} \min_{f \in \mathscr{F}} \quad & R(f) \\ \text{subject to} \quad & f(X_i) = y_i' \end{aligned}$$

ends up with a quadratic form:

$$R(f) = \mathbf{f}^T \mathbf{M} \mathbf{f}$$

where $\mathbf{f} = (f(X_1), f(X_2), \ldots, f(X_n))^T$

- If Ω is irregular, determining analytically the gram matrix $\mathbf{M}$ can be very difficult.

More on the Existing Approach

- Define a regularity functional. Recall

$$R(f) = \int_\Omega \|\mathbf{H}f\|_F^2$$

- Essence: Need to show that solving problem

$$\begin{aligned} \min_{f \in \mathscr{F}} \quad & R(f) \\ \text{subject to} \quad & f(X_i) = y_i' \end{aligned}$$

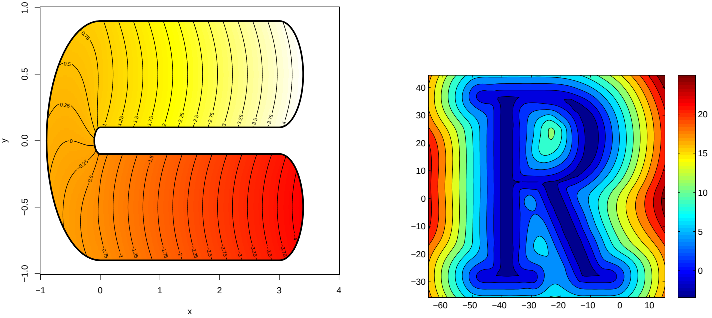
ends up with a quadratic form:

$$R(f) = \mathbf{f}^T \mathbf{M} \mathbf{f}$$

where $\mathbf{f} = (f(X_1), f(X_2), \ldots, f(X_n))^T$

- If Ω is irregular, determining analytically the gram matrix $\mathbf{M}$ can be very difficult.

Example of Irregular Domain



(a) Horseshoe.

(b) Letter R.

Outline

Unbiased Estimation in a Neighborhood

- An unbiased alternative:

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} [y_i - f(X_i)]^2 + \lambda \sum_{i=1}^{n} \|\mathcal{H} f(X_i)\|_F^2.$$

- Let $\mathbf{V}_i, 1 \le i \le k$, denote the k nearest neighbors of $\mathbf{V}_0$. Let $\bar{\mathbf{V}} = \frac{1}{k+1} \sum_{i=0}^{k} \mathbf{V}_i$, i.e., $\bar{\mathbf{V}}$ is the average. Taylor expansion:

$$f(\mathbf{V}_i) \approx f(\bar{\mathbf{V}}) + (\mathbf{V}_i - \bar{\mathbf{V}})^T \mathcal{J} f(\bar{\mathbf{V}}) + \frac{1}{2} (\mathbf{V}_i - \bar{\mathbf{V}})^T \mathcal{H} f(\bar{\mathbf{V}}) (\mathbf{V}_i - \bar{\mathbf{V}}),$$
$$i = 0, 1, \cdots, n,$$

where $f(\bar{\mathbf{V}})$ is the functional value, $\mathcal{J} f(\bar{\mathbf{V}})$ is the Jacobian, and $\mathcal{H} f(\bar{\mathbf{V}})$ is the hessian matrix. Note we have $\mathcal{J} f(\bar{\mathbf{V}}) \in \mathbb{R}^d$ and $\mathcal{H} f(\bar{\mathbf{V}}) \in \mathbb{R}^{d \times d}$.

Unbiased Estimation in a Neighborhood

- An unbiased alternative:

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} [y_i - f(X_i)]^2 + \lambda \sum_{i=1}^{n} \|\mathcal{H} f(X_i)\|_F^2.$$

- Let $\mathbf{V}_i, 1 \le i \le k$, denote the k nearest neighbors of $\mathbf{V}_0$. Let $\bar{\mathbf{V}} = \frac{1}{k+1} \sum_{i=0}^{k} \mathbf{V}_i$, i.e., $\bar{\mathbf{V}}$ is the average. Taylor expansion:

$$f(\mathbf{V}_i) \approx f(\bar{\mathbf{V}}) + (\mathbf{V}_i - \bar{\mathbf{V}})^T \mathcal{J} f(\bar{\mathbf{V}}) + \frac{1}{2} (\mathbf{V}_i - \bar{\mathbf{V}})^T \mathcal{H} f(\bar{\mathbf{V}}) (\mathbf{V}_i - \bar{\mathbf{V}}),$$
$$i = 0, 1, \cdots, n,$$

  where $f(\bar{\mathbf{V}})$ is the functional value, $\mathcal{J} f(\bar{\mathbf{V}})$ is the Jacobian, and $\mathcal{H} f(\bar{\mathbf{V}})$ is the hessian matrix. Note we have $\mathcal{J} f(\bar{\mathbf{V}}) \in \mathbb{R}^d$ and $\mathcal{H} f(\bar{\mathbf{V}}) \in \mathbb{R}^{d \times d}$.

Derivation

- Rewrite as a linear system:

$$\mathbf{f}^* \approx \mathbf{1}_{k+1} \cdot c + \mathbf{V} \cdot \mathbf{J} + \frac{1}{2}\mathbf{C} \cdot \mathbf{H},$$

- A partial implementation of QR-decomposition

$$\begin{bmatrix} \mathbf{1}_{k+1} & \mathbf{V} & \frac{1}{2}\mathbf{C} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{bmatrix} \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{0} & \mathbf{I}_{(d^2+d)/2} \end{bmatrix},$$

where columns of $\mathbf{Q}_1 \in \mathbb{R}^{(k+1)\times(d+1)}$ are orthonormal ($\mathbf{Q}_1^T\mathbf{Q}_1 = \mathbf{I}_{d+1}$), and columns of $\mathbf{Q}_2 \in \mathbb{R}^{(k+1)\times\frac{d^2+d}{2}}$ are orthogonal to the columns of $\mathbf{Q}_1$ (i.e., $\mathbf{Q}_2^T\mathbf{Q}_1 = \mathbf{0}$).

Derivation

- Rewrite as a linear system:

$$\mathbf{f}^* \approx \mathbf{1}_{k+1} \cdot c + \mathbf{V} \cdot \mathbf{J} + \frac{1}{2}\mathbf{C} \cdot \mathbf{H},$$

- A partial implementation of QR-decomposition

$$\left[\begin{array}{ccc} \mathbf{1}_{k+1} & \mathbf{V} & \frac{1}{2}\mathbf{C} \end{array}\right] = \left[\begin{array}{cc} \mathbf{Q}_1 & \mathbf{Q}_2 \end{array}\right] \left[\begin{array}{cc} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{0} & \mathbf{I}_{(d^2+d)/2} \end{array}\right],$$

where columns of $\mathbf{Q}_1 \in \mathbb{R}^{(k+1)\times(d+1)}$ are orthonormal ($\mathbf{Q}_1^T\mathbf{Q}_1 = \mathbf{I}_{d+1}$), and columns of $\mathbf{Q}_2 \in \mathbb{R}^{(k+1)\times\frac{d^2+d}{2}}$ are orthogonal to the columns of $\mathbf{Q}_1$ (i.e., $\mathbf{Q}_2^T\mathbf{Q}_1 = \mathbf{0}$).

More on Derivation

- we have

$$\mathbf{Q}_2^T \mathbf{f}^* = \begin{pmatrix} \mathbf{0} & \mathbf{Q}_2^T \mathbf{Q}_2 \end{pmatrix} \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{0} & \mathbf{I}_{(p^2+p)/2} \end{bmatrix} \begin{bmatrix} c \\ \mathbf{J} \\ \mathbf{H} \end{bmatrix} = \mathbf{Q}_2^T \mathbf{Q}_2 \mathbf{H}.$$

- least-squares estimator of $\mathbf{H}$

$$\hat{\mathbf{H}} = (\mathbf{Q}_2^T \mathbf{Q}_2)^+ \mathbf{Q}_2^T \mathbf{f}^*,$$

where $(\cdot)^+$ denotes a pseudo-inverse of a matrix.

- Frobenius norm of the hessian matrix at a point:

$$\begin{aligned} \|\hat{\mathcal{H}} f(X_i)\|_F^2 &= \|\hat{\mathbf{H}}\|_2^2 = \hat{\mathbf{H}}^T \hat{\mathbf{H}} \\ &= (\mathbf{f}^*)^T \mathbf{Q}_2 (\mathbf{Q}_2^T \mathbf{Q}_2)^+ (\mathbf{Q}_2^T \mathbf{Q}_2)^+ \mathbf{Q}_2^T \mathbf{f}^*. \end{aligned}$$

More on Derivation

- we have

$$\mathbf{Q}_2^T \mathbf{f}^* = \begin{pmatrix} \mathbf{0} & \mathbf{Q}_2^T \mathbf{Q}_2 \end{pmatrix} \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{0} & \mathbf{I}_{(p^2+p)/2} \end{bmatrix} \begin{bmatrix} c \\ \mathbf{J} \\ \mathbf{H} \end{bmatrix} = \mathbf{Q}_2^T \mathbf{Q}_2 \mathbf{H}.$$

- least-squares estimator of $\mathbf{H}$

$$\hat{\mathbf{H}} = (\mathbf{Q}_2^T \mathbf{Q}_2)^+ \mathbf{Q}_2^T \mathbf{f}^*,$$

where $(\cdot)^+$ denotes a pseudo-inverse of a matrix.

- Frobenius norm of the hessian matrix at a point:

$$\begin{aligned} \|\hat{\mathcal{H}} f(X_i)\|_F^2 &= \|\hat{\mathbf{H}}\|_2^2 = \hat{\mathbf{H}}^T \hat{\mathbf{H}} \\ &= (\mathbf{f}^*)^T \mathbf{Q}_2 (\mathbf{Q}_2^T \mathbf{Q}_2)^+ (\mathbf{Q}_2^T \mathbf{Q}_2)^+ \mathbf{Q}_2^T \mathbf{f}^*. \end{aligned}$$

More on Derivation

- we have

$$\mathbf{Q}_2^T \mathbf{f}^* = \begin{pmatrix} \mathbf{0} & \mathbf{Q}_2^T \mathbf{Q}_2 \end{pmatrix} \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{0} & \mathbf{I}_{(p^2+p)/2} \end{bmatrix} \begin{bmatrix} c \\ \mathbf{J} \\ \mathbf{H} \end{bmatrix} = \mathbf{Q}_2^T \mathbf{Q}_2 \mathbf{H}.$$

- least-squares estimator of $\mathbf{H}$

$$\hat{\mathbf{H}} = (\mathbf{Q}_2^T \mathbf{Q}_2)^+ \mathbf{Q}_2^T \mathbf{f}^*,$$

  where $(\cdot)^+$ denotes a pseudo-inverse of a matrix.

- Frobenius norm of the hessian matrix at a point:

$$\begin{aligned} \|\hat{\mathcal{H}} f(X_i)\|_F^2 &= \|\hat{\mathbf{H}}\|_2^2 = \hat{\mathbf{H}}^T \hat{\mathbf{H}} \\ &= (\mathbf{f}^*)^T \mathbf{Q}_2 (\mathbf{Q}_2^T \mathbf{Q}_2)^+ (\mathbf{Q}_2^T \mathbf{Q}_2)^+ \mathbf{Q}_2^T \mathbf{f}^*. \end{aligned}$$

A Quadratic Form

- Denote

$$\mathbf{K}_i = \mathbf{Q}_2(\mathbf{Q}_2^T\mathbf{Q}_2)^+(\mathbf{Q}_2^T\mathbf{Q}_2)^+\mathbf{Q}_2^T.$$

- We have

$$\sum_{i=1}^{n} \|\hat{\mathcal{H}} f(X_i)\|_F^2 = \sum_{i=1}^{n} (\mathbf{f}^T\mathbf{S}_i^T\mathbf{K}_i\mathbf{S}_i\mathbf{f}).$$

- Let $\mathbf{M} = (\mathbf{S}_1^T, \cdots, \mathbf{S}_n^T)\text{diag}\{\mathbf{K}_1, \mathbf{K}_2, \cdots, \mathbf{K}_n\}\begin{pmatrix} \mathbf{S}_1 \\ \vdots \\ \mathbf{S}_n \end{pmatrix}$, we have

$$\sum_{i=1}^{n} \|\hat{\mathcal{H}} f(X_i)\|_F^2 = \mathbf{f}^T\mathbf{M}\mathbf{f}, \tag{1}$$

which is a quadratic function of $\mathbf{f}$.

A Quadratic Form

- Denote
$$\mathbf{K}_i = \mathbf{Q}_2(\mathbf{Q}_2^T\mathbf{Q}_2)^+(\mathbf{Q}_2^T\mathbf{Q}_2)^+\mathbf{Q}_2^T.$$
- We have
$$\sum_{i=1}^n \|\hat{\mathcal{H}}f(X_i)\|_F^2 = \sum_{i=1}^n (\mathbf{f}^T\mathbf{S}_i^T\mathbf{K}_i\mathbf{S}_i\mathbf{f}).$$
- Let $\mathbf{M} = (\mathbf{S}_1^T, \cdots, \mathbf{S}_n^T)\mathrm{diag}\{\mathbf{K}_1, \mathbf{K}_2, \cdots, \mathbf{K}_n\}\begin{pmatrix}\mathbf{S}_1\\ \vdots \\ \mathbf{S}_n\end{pmatrix}$, we have
$$\sum_{i=1}^n \|\hat{\mathcal{H}}f(X_i)\|_F^2 = \mathbf{f}^T\mathbf{M}\mathbf{f}, \tag{1}$$
which is a quadratic function of $\mathbf{f}$.

A Quadratic Form

- Denote
$$\mathbf{K}_i = \mathbf{Q}_2(\mathbf{Q}_2^T\mathbf{Q}_2)^+(\mathbf{Q}_2^T\mathbf{Q}_2)^+\mathbf{Q}_2^T.$$
- We have
$$\sum_{i=1}^n \|\hat{\mathcal{H}}f(X_i)\|_F^2 = \sum_{i=1}^n (\mathbf{f}^T\mathbf{S}_i^T\mathbf{K}_i\mathbf{S}_i\mathbf{f}).$$
- Let $\mathbf{M} = (\mathbf{S}_1^T, \cdots, \mathbf{S}_n^T)\mathrm{diag}\{\mathbf{K}_1, \mathbf{K}_2, \cdots, \mathbf{K}_n\}\begin{pmatrix}\mathbf{S}_1\\ \vdots\\ \mathbf{S}_n\end{pmatrix}$, we have
$$\sum_{i=1}^n \|\hat{\mathcal{H}}f(X_i)\|_F^2 = \mathbf{f}^T\mathbf{M}\mathbf{f}, \tag{1}$$
which is a quadratic function of $\mathbf{f}$.

Close Form Solution

- Problem becomes
$$\min_{\mathbf{f}} \|\mathbf{Y} - \mathbf{f}\|_2^2 + \lambda \mathbf{f}^T \mathbf{M} \mathbf{f},$$
- Close form solution:
$$\hat{\mathbf{f}} = (\mathbf{I}_n + \lambda \cdot \mathbf{M})^{-1} \cdot \mathbf{Y}.$$
- J. Chen and X. Huo (2009). A Hessian regularized nonlinear time series model. *Journal of Computational and Graphical Statistics, 18* (3): 694-716, September.

Close Form Solution

- Problem becomes
$$\min_{\mathbf{f}} \|\mathbf{Y} - \mathbf{f}\|_2^2 + \lambda \mathbf{f}^T \mathbf{M} \mathbf{f},$$
- Close form solution:
$$\hat{\mathbf{f}} = (\mathbf{I}_n + \lambda \cdot \mathbf{M})^{-1} \cdot \mathbf{Y}.$$
- J. Chen and X. Huo (2009). A Hessian regularized nonlinear time series model. *Journal of Computational and Graphical Statistics, 18* (3): 694-716, September.

Close Form Solution

- Problem becomes

$$\min_{\mathbf{f}} \|\mathbf{Y} - \mathbf{f}\|_2^2 + \lambda \mathbf{f}^T \mathbf{M} \mathbf{f},$$

- Close form solution:

$$\hat{\mathbf{f}} = (\mathbf{I}_n + \lambda \cdot \mathbf{M})^{-1} \cdot \mathbf{Y}.$$

- J. Chen and X. Huo (2009). A Hessian regularized nonlinear time series model. *Journal of Computational and Graphical Statistics, 18* (3): 694-716, September.

Outline

Rate of Convergence

- Rate of convergence: How fast does $\frac{1}{n}\|\hat{f}_n - f\|_2^2$ goes to zero?
- Stone (1982) showed that $O(n^{-\frac{2m}{2m+d}})$ is the optimal rate of convergence for nonparametric regression for d-dimensional input, while the up to order m partial derivatives of the underlying function f are in $L^2(\Omega)$.
- Note we have $m = 2$.
- Note: we must have $2m > d$.

Rate of Convergence

- Rate of convergence: How fast does $\frac{1}{n}\|\hat{f}_n - f\|_2^2$ goes to zero?
- Stone (1982) showed that $O(n^{-\frac{2m}{2m+d}})$ is the optimal rate of convergence for nonparametric regression for d-dimensional input, while the up to order m partial derivatives of the underlying function f are in $L^2(\Omega)$.
- Note we have $m = 2$.
- Note: we must have $2m > d$.

Rate of Convergence

- Rate of convergence: How fast does $\frac{1}{n}\|\hat{f}_n - f\|_2^2$ goes to zero?
- Stone (1982) showed that $O(n^{-\frac{2m}{2m+d}})$ is the optimal rate of convergence for nonparametric regression for d-dimensional input, while the up to order m partial derivatives of the underlying function f are in $L^2(\Omega)$.
- Note we have $m = 2$.
- Note: we must have $2m > d$.

Rate of Convergence

- Rate of convergence: How fast does $\frac{1}{n}\|\hat{f}_n - f\|_2^2$ goes to zero?
- Stone (1982) showed that $O(n^{-\frac{2m}{2m+d}})$ is the optimal rate of convergence for nonparametric regression for d-dimensional input, while the up to order m partial derivatives of the underlying function f are in $L^2(\Omega)$.
- Note we have $m = 2$.
- Note: we must have $2m > d$.

Preparation: Inner Product in a RKHS

- For $0 \le \ell \le m$, a semi-inner-product in $W_2^m(\Omega)$ is defined by

$$\langle f, g \rangle_{\Omega,\ell} = \int_\Omega \sum_{|\alpha|=\ell} \frac{\ell!}{\alpha!} (D^\alpha f)(D^\alpha g) \mathrm{d}x, \tag{2}$$

which gives rise to the related semi-norm

$$|f|^2_{\Omega,\ell} = \int_\Omega \sum_{|\alpha|=\ell} \frac{\ell!}{\alpha!} |D^\alpha f|^2 \mathrm{d}x. \tag{3}$$

- With $T = \{X_i\}_{i=1}^n$, we can also give a discrete version of the aforementioned semi-norm as

$$|f|^2_{T,\ell} = \frac{1}{n} \sum_{i=1}^n \sum_{|\alpha|=\ell} \frac{\ell!}{\alpha!} |D^\alpha f(X_i)|^2. \tag{4}$$

- Specially, $|f|^2_{\Omega,0} = \int_\Omega f(x)^2 \mathrm{d}x$ and $|f|^2_{T,0} = \frac{1}{n} \sum_{i=1}^n f(X_i)^2$.

Preparation: Inner Product in a RKHS

- For $0 \le \ell \le m$, a semi-inner-product in $W_2^m(\Omega)$ is defined by

$$\langle f, g\rangle_{\Omega,\ell} = \int_\Omega \sum_{|\alpha|=\ell} \frac{\ell!}{\alpha!}(D^\alpha f)(D^\alpha g)\mathrm{d}x, \tag{2}$$

  which gives rise to the related semi-norm

$$|f|^2_{\Omega,\ell} = \int_\Omega \sum_{|\alpha|=\ell} \frac{\ell!}{\alpha!}|D^\alpha f|^2\mathrm{d}x. \tag{3}$$

- With $T = \{X_i\}_{i=1}^n$, we can also give a discrete version of the aforementioned semi-norm as

$$|f|^2_{T,\ell} = \frac{1}{n}\sum_{i=1}^n \sum_{|\alpha|=\ell} \frac{\ell!}{\alpha!}|D^\alpha f(X_i)|^2. \tag{4}$$

- Specially, $|f|^2_{\Omega,0} = \int_\Omega f(x)^2\mathrm{d}x$ and $|f|^2_{T,0} = \frac{1}{n}\sum_{i=1}^n f(X_i)^2$.

Preparation: Inner Product in a RKHS

- For $0 \le \ell \le m$, a semi-inner-product in $W_2^m(\Omega)$ is defined by

$$\langle f, g \rangle_{\Omega,\ell} = \int_\Omega \sum_{|\alpha|=\ell} \frac{\ell!}{\alpha!} (D^\alpha f)(D^\alpha g) \mathrm{d}x, \tag{2}$$

which gives rise to the related semi-norm

$$|f|^2_{\Omega,\ell} = \int_\Omega \sum_{|\alpha|=\ell} \frac{\ell!}{\alpha!} |D^\alpha f|^2 \mathrm{d}x. \tag{3}$$

- With $T = \{X_i\}_{i=1}^n$, we can also give a discrete version of the aforementioned semi-norm as

$$|f|^2_{T,\ell} = \frac{1}{n} \sum_{i=1}^n \sum_{|\alpha|=\ell} \frac{\ell!}{\alpha!} |D^\alpha f(X_i)|^2. \tag{4}$$

- Specially, $|f|^2_{\Omega,0} = \int_\Omega f(x)^2 \mathrm{d}x$ and $|f|^2_{T,0} = \frac{1}{n} \sum_{i=1}^n f(X_i)^2$.

Preparation: Ideal Quadratic form & Sampling Property

- *Ideal quadratic form.* For $\ell = m$ in (4), we define $\mathbf{E}_{T,m}$ as the matrix representing the quadratic form

$$|f|_{T,m}^2 = \frac{1}{n}\mathbf{f}^T \mathbf{E}_{T,m}\mathbf{f} \tag{5}$$

where $\mathbf{f} = (f(X_1), \cdots, f(X_n))^T$ is the vector of function values at the knots of $T = \{X_i\}_{i=1}^n$.

- *Sampling property.* For the set of sampling points $T = \{X_i\}_{i=1}^n$ in domain Ω, we assume that there exists a constant $B_0 > 0$ such that

$$\frac{\delta_{\max}}{\delta_{\min}} \le B_0, \tag{6}$$

where $\delta_{\max} = \sup_{X \in \Omega} \inf_{X_i \in T} \|X - X_i\|$, and $\delta_{\min} = \min_{j \neq i} \|X_j - X_i\|$.

Preparation: Ideal Quadratic form & Sampling Property

- *Ideal quadratic form.* For $\ell = m$ in (4), we define $\mathbf{E}_{T,m}$ as the matrix representing the quadratic form

$$|f|^2_{T,m} = \frac{1}{n}\mathbf{f}^T \mathbf{E}_{T,m}\mathbf{f} \tag{5}$$

  where $\mathbf{f} = (f(X_1), \cdots, f(X_n))^T$ is the vector of function values at the knots of $T = \{X_i\}_{i=1}^n$.
- *Sampling property.* For the set of sampling points $T = \{X_i\}_{i=1}^n$ in domain Ω, we assume that there exists a constant $B_0 > 0$ such that

$$\frac{\delta_{\max}}{\delta_{\min}} \le B_0, \tag{6}$$

  where $\delta_{\max} = \sup_{X \in \Omega} \inf_{X_i \in T} \|X - X_i\|$, and $\delta_{\min} = \min_{j \neq i} \|X_j - X_i\|$.

Property of the Domain Ω

- A **Lipschitz domain** (or domain with Lipschitz boundary) is a set in Euclidean space whose boundary is sufficiently regular in the sense that it can be thought of as locally being the graph of a Lipschitz continuous function.
- Ω be an open set of $\mathbb{R}^d$ satisfying a **uniform cone condition**: there exist a radius $r > 0$ and an angle $\theta \in (0, \pi/2)$ such that for any $X \in \Omega$ a unit vector $\zeta(X) \in \mathbb{R}^d$ exists such that the cone

$$C(X, \zeta(X), r, \theta) = \{X + t\mathbf{s} : \mathbf{s} \in \mathbb{R}^d, \|\mathbf{s}\| = 1, \zeta(X)^T\mathbf{s} \geq \cos\theta, 0 \leq t \leq r\} \quad (7)$$

  is entirely contained in Ω.
- $U_2^m(\Omega) = \{f \in W_2^m(\Omega) \mid \underline{B}|f|^2_{\Omega,m} \leq |f|^2_{T,m} \leq \overline{B}|f|^2_{\Omega,m}\}$ be a class of functions with bilaterally bounded constraint on their mth-order derivatives, where the constants $\underline{B}, \overline{B} > 0$ do not depend on functions f.

Property of the Domain Ω

- A **Lipschitz domain** (or domain with Lipschitz boundary) is a set in Euclidean space whose boundary is sufficiently regular in the sense that it can be thought of as locally being the graph of a Lipschitz continuous function.
- Ω be an open set of $\mathbb{R}^d$ satisfying a **uniform cone condition**: there exist a radius $r > 0$ and an angle $\theta \in (0, \pi/2)$ such that for any $X \in \Omega$ a unit vector $\zeta(X) \in \mathbb{R}^d$ exists such that the cone

$$C(X, \zeta(X), r, \theta) = \{X + t\mathbf{s} : \mathbf{s} \in \mathbb{R}^d, \|\mathbf{s}\| = 1, \zeta(X)^T \mathbf{s} \geq \cos\theta, 0 \leq t \leq r\} \tag{7}$$

  is entirely contained in Ω.
- $U_2^m(\Omega) = \{f \in W_2^m(\Omega) \mid \underline{B}|f|^2_{\Omega,m} \leq |f|^2_{T,m} \leq \overline{B}|f|^2_{\Omega,m}\}$ be a class of functions with bilaterally bounded constraint on their mth-order derivatives, where the constants $\underline{B}, \overline{B} > 0$ do not depend on functions f.

Property of the Domain Ω

- A **Lipschitz domain** (or domain with Lipschitz boundary) is a set in Euclidean space whose boundary is sufficiently regular in the sense that it can be thought of as locally being the graph of a Lipschitz continuous function.
- Ω be an open set of $\mathbb{R}^d$ satisfying a **uniform cone condition**: there exist a radius $r > 0$ and an angle $\theta \in (0, \pi/2)$ such that for any $X \in \Omega$ a unit vector $\zeta(X) \in \mathbb{R}^d$ exists such that the cone

$$C(X, \zeta(X), r, \theta) = \{X + t\mathbf{s} : \mathbf{s} \in \mathbb{R}^d, \|\mathbf{s}\| = 1, \zeta(X)^T \mathbf{s} \geq \cos\theta, 0 \leq t \leq r\} \tag{7}$$

  is entirely contained in Ω.
- $U_2^m(\Omega) = \{f \in W_2^m(\Omega) \mid \underline{B}|f|^2_{\Omega,m} \leq |f|^2_{T,m} \leq \overline{B}|f|^2_{\Omega,m}\}$ be a class of functions with bilaterally bounded constraint on their mth-order derivatives, where the constants $\underline{B}, \overline{B} > 0$ do not depend on functions f.

A few theorems, Step 1

- *Bound the eigenvalues of the ideal quadratic form.* Let $e_1 \leq \cdots \leq e_n$ be the eigenvalues of $\mathbf{E}_{T,m}$ in ascending order.

Theorem

Let Ω be an open bounded Lipschitz domain satisfying the uniform cone condition, and the sample points $\{X_j\}_{j=1}^n$ fulfill the assumption of (6). Then there exist constants $C_3, C_4 > 0$ such that

$$C_3\rho_j \leq e_j \leq C_4\rho_j,$$

where $\rho_1 \leq \rho_2 \leq \cdots \leq \rho_n$ are the first n eigenvalues of the variational eigenvalue problem

$$\langle \phi, \psi \rangle_{\Omega,m} = \rho \langle \phi, \psi \rangle_{\Omega,0}, \quad \forall\ \psi \in W_2^m(\Omega).$$

A few theorems, Step 1

- *Bound the eigenvalues of the ideal quadratic form.* Let $e_1 \leq \cdots \leq e_n$ be the eigenvalues of $\mathbf{E}_{T,m}$ in ascending order.

Theorem

Let Ω be an open bounded Lipschitz domain satisfying the uniform cone condition, and the sample points $\{X_j\}_{j=1}^n$ fulfill the assumption of (6). Then there exist constants $C_3, C_4 > 0$ such that

$$C_3 \rho_j \leq e_j \leq C_4 \rho_j,$$

where $\rho_1 \leq \rho_2 \leq \cdots \leq \rho_n$ are the first n eigenvalues of the variational eigenvalue problem

$$\langle \phi, \psi \rangle_{\Omega,m} = \rho \langle \phi, \psi \rangle_{\Omega,0}, \quad \forall\ \psi \in W_2^m(\Omega).$$

Step 2: Quote a Known Result

- Using a known rate from functional analysis

Theorem

Let Ω be an open bounded Lipschitz domain satisfying the uniform cone condition, and $\{e_1 \le \cdots \le e_n\}$ the eigenvalues of $\mathbf{E}_{T,m}$ in ascending order. Then there exist constants $C_5, C_6 > 0$ such that for $m(d) = \frac{(d+m-1)!}{d!(m-1)!} < j \le n$ we have

$$C_5 j^{\frac{2m}{d}} \le e_j \le C_6 j^{\frac{2m}{d}}. \tag{8}$$

- Using a known rate from functional analysis

Theorem

Let Ω be an open bounded Lipschitz domain satisfying the uniform cone condition, and $\{e_1 \leq \cdots \leq e_n\}$ the eigenvalues of $\mathbf{E}_{T,m}$ in ascending order. Then there exist constants $C_5, C_6 > 0$ such that for $m(d) = \frac{(d+m-1)!}{d!(m-1)!} < j \leq n$ we have

$$C_5 j^{\frac{2m}{d}} \leq e_j \leq C_6 j^{\frac{2m}{d}}. \tag{8}$$

Step 3: Rates on Eigenvalues

- Property of eigenvalues.

Theorem

Let $\mu_1 \leq \cdots \leq \mu_n$ be the eigenvalues of the matrix $\mathbf{M}$ in (1). There exist constants $C_7, C_8 > 0$ such that for $m^{(d)} < j \leq n$ we have

$$C_7 j^{\frac{2m}{d}} \leq \mu_j \leq C_8 j^{\frac{2m}{d}}.$$

Step 3: Rates on Eigenvalues

- Property of eigenvalues.

Theorem

Let $\mu_1 \leq \cdots \leq \mu_n$ be the eigenvalues of the matrix $\mathbf{M}$ in (1). There exist constants $C_7, C_8 > 0$ such that for $m^{(d)} < j \leq n$ we have

$$C_7 j^{\frac{2m}{d}} \leq \mu_j \leq C_8 j^{\frac{2m}{d}}.$$

Step 4: Rate of Convergence

- Asymptotic rate of convergence

Theorem

Let $\hat{\mathbf{f}}_n(\lambda) = \mathbf{A}_n(\lambda)\mathbf{y} = (\mathbf{I}_n + \lambda \mathbf{M})^{-1}\mathbf{y}$ be the CDS estimator from the multivariate model with the order $m > d/2$ and denote $r_n(\lambda) = n^{-1}\|\hat{\mathbf{f}}_n(\lambda) - \mathbf{f}\|^2$. If $n \to \infty$ and $\lambda \sim n^{-2m/(2m+d)}$ is chosen, then

$$E[r_n(\lambda)] = O(n^{-\frac{2m}{2m+d}}).$$

- The above matches the optimal rate in Stone (1982).

Step 4: Rate of Convergence

- Asymptotic rate of convergence

Theorem

Let $\hat{\mathbf{f}}_n(\lambda) = \mathbf{A}_n(\lambda)\mathbf{y} = (\mathbf{I}_n + \lambda\mathbf{M})^{-1}\mathbf{y}$ *be the CDS estimator from the multivariate model with the order* $m > d/2$ *and denote* $r_n(\lambda) = n^{-1}\|\hat{\mathbf{f}}_n(\lambda) - \mathbf{f}\|^2$*. If* $n \to \infty$ *and* $\lambda \sim n^{-2m/(2m+d)}$ *is chosen, then*

$$E[r_n(\lambda)] = O(n^{-\frac{2m}{2m+d}}).$$

- The above matches the optimal rate in Stone (1982).

Step 4: Rate of Convergence

- Asymptotic rate of convergence

Theorem

Let $\hat{\mathbf{f}}_n(\lambda) = \mathbf{A}_n(\lambda)\mathbf{y} = (\mathbf{I}_n + \lambda\mathbf{M})^{-1}\mathbf{y}$ *be the CDS estimator from the multivariate model with the order* $m > d/2$ *and denote* $r_n(\lambda) = n^{-1}\|\hat{\mathbf{f}}_n(\lambda) - \mathbf{f}\|^2$*. If* $n \to \infty$ *and* $\lambda \sim n^{-2m/(2m+d)}$ *is chosen, then*

$$E[r_n(\lambda)] = O(n^{-\frac{2m}{2m+d}}).$$

- The above matches the optimal rate in Stone (1982).

Outline

Asymptotic Optimality

- Choose the parameter λ via the Generalized Cross Validation
- Let $\hat{\mathbf{f}}_n(\lambda) = \mathbf{A}_n(\lambda)\mathbf{y} = (\mathbf{I}_n + \lambda\mathbf{M})^{-1}\mathbf{y}$ be the estimator of CDS model with the order m and denote $r_n(\lambda) = n^{-1}\|\hat{\mathbf{f}}_n(\lambda) - \mathbf{f}\|^2$. The asymptotic optimality of GCV is defined as

$$\frac{r_n(\hat{\lambda}_G)}{\inf_{\lambda\in\mathbb{R}_+} r_n(\lambda)} \longrightarrow_p 1 \tag{9}$$

where $\longrightarrow_p$ means the convergence in probability.

Asymptotic Optimality

- Choose the parameter λ via the Generalized Cross Validation
- Let $\hat{\mathbf{f}}_n(\lambda) = \mathbf{A}_n(\lambda)\mathbf{y} = (\mathbf{I}_n + \lambda\mathbf{M})^{-1}\mathbf{y}$ be the estimator of CDS model with the order m and denote $r_n(\lambda) = n^{-1}\|\hat{\mathbf{f}}_n(\lambda) - \mathbf{f}\|^2$. The asymptotic optimality of GCV is defined as

$$\frac{r_n(\hat{\lambda}_G)}{\inf_{\lambda \in \mathbb{R}_+} r_n(\lambda)} \longrightarrow_p 1 \tag{9}$$

where $\longrightarrow_p$ means the convergence in probability.

Three Conditions

(A.1) $\inf_{\lambda \in \mathbb{R}_+} nE[r_n(\lambda)] \to \infty$.

(A.2) There exists a sequence $\{\lambda_n\}$ such that $r_n(\lambda_n) \longrightarrow_p 0$ (the convergence in probability).

(A.3) Let $0 \le \kappa_1 \le \cdots \le \kappa_n$ be the eigenvalues of $\mathbf{K}_n(\lambda) = \lambda \mathbf{M}$. For any ℓ such that $\frac{\ell}{n} \to 0$, then $\frac{\left(n^{-1}\sum_{i=\ell+1}^{n} \kappa_i^{-1}\right)^2}{n^{-1}\sum_{i=\ell+1}^{n} \kappa_i^{-2}} \to 0$ as $n \to \infty$.

Three Conditions

(A.1) $\inf_{\lambda \in \mathbb{R}_+} nE[r_n(\lambda)] \to \infty$.

(A.2) There exists a sequence $\{\lambda_n\}$ such that $r_n(\lambda_n) \longrightarrow_p 0$ (the convergence in probability).

(A.3) Let $0 \le \kappa_1 \le \cdots \le \kappa_n$ be the eigenvalues of $\mathbf{K}_n(\lambda) = \lambda \mathbf{M}$. For any ℓ such that $\frac{\ell}{n} \to 0$, then $\frac{\left(n^{-1}\sum_{i=\ell+1}^{n} \kappa_i^{-1}\right)^2}{n^{-1}\sum_{i=\ell+1}^{n} \kappa_i^{-2}} \to 0$ as $n \to \infty$.

Three Conditions

(A.1) $\inf_{\lambda \in \mathbb{R}_+} nE[r_n(\lambda)] \to \infty$.

(A.2) There exists a sequence $\{\lambda_n\}$ such that $r_n(\lambda_n) \longrightarrow_p 0$ (the convergence in probability).

(A.3) Let $0 \le \kappa_1 \le \cdots \le \kappa_n$ be the eigenvalues of $\mathbf{K}_n(\lambda) = \lambda \mathbf{M}$. For any ℓ such that $\frac{\ell}{n} \to 0$, then $\frac{\left(n^{-1}\sum_{i=\ell+1}^{n} \kappa_i^{-1}\right)^2}{n^{-1}\sum_{i=\ell+1}^{n} \kappa_i^{-2}} \to 0$ as $n \to \infty$.

Asymptotic Optimality of GCV

- Formal result

Theorem

Under conditions (A.1), (A.2) and (A.3), $\hat{\mathbf{f}}_n(\hat{\lambda}_G)$ is asymptotically optimal, where $\hat{\lambda}_G$ is the GCV choice.

Asymptotic Optimality of GCV

- Formal result

Theorem

Under conditions (A.1), (A.2) and (A.3), $\hat{\mathbf{f}}_n(\hat{\lambda}_G)$ is asymptotically optimal, where $\hat{\lambda}_G$ is the GCV choice.

Outline

A Permutation Approach

- Sparsity of design matrix $\mathbf{M}$; Recall $\hat{\mathbf{f}} = (\mathbf{I}_n + \lambda \cdot \mathbf{M})^{-1} \cdot \mathbf{Y}$
- The number of nonzeros of $\mathbf{M}$ is strictly less than $(k+1)^2 n$
- Roughly $3kn$ as shown in numerical experiments
- $\mathbf{M}$ can be permutated to a band matrix by the symmetric reverse Cuthill-Mckee ordering (1969) with $O(k \log(k) n)$ complexity
- See an example on the next page...

A Permutation Approach

- Sparsity of design matrix $\mathbf{M}$; Recall $\hat{\mathbf{f}} = (\mathbf{I}_n + \lambda \cdot \mathbf{M})^{-1} \cdot \mathbf{Y}$
- The number of nonzeros of $\mathbf{M}$ is strictly less than $(k+1)^2 n$
- Roughly $3kn$ as shown in numerical experiments
- $\mathbf{M}$ can be permutated to a band matrix by the symmetric reverse Cuthill-Mckee ordering (1969) with $O(k \log(k) n)$ complexity
- See an example on the next page...

A Permutation Approach

- Sparsity of design matrix $\mathbf{M}$; Recall $\hat{\mathbf{f}} = (\mathbf{I}_n + \lambda \cdot \mathbf{M})^{-1} \cdot \mathbf{Y}$
- The number of nonzeros of $\mathbf{M}$ is strictly less than $(k+1)^2 n$
- Roughly $3kn$ as shown in numerical experiments
- $\mathbf{M}$ can be permutated to a band matrix by the symmetric reverse Cuthill-Mckee ordering (1969) with $O(k \log(k) n)$ complexity
- See an example on the next page...

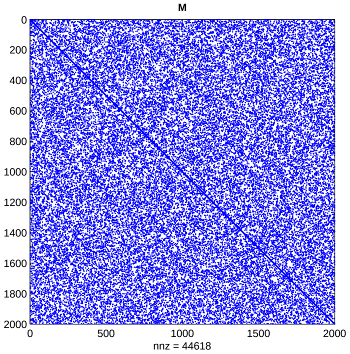
A Permutation Approach

- Sparsity of design matrix $\mathbf{M}$; Recall $\hat{\mathbf{f}} = (\mathbf{I}_n + \lambda \cdot \mathbf{M})^{-1} \cdot \mathbf{Y}$
- The number of nonzeros of $\mathbf{M}$ is strictly less than $(k+1)^2 n$
- Roughly $3kn$ as shown in numerical experiments
- $\mathbf{M}$ can be permutated to a band matrix by the symmetric reverse Cuthill-Mckee ordering (1969) with $O(k \log(k) n)$ complexity
- See an example on the next page...

A Permutation Approach

- Sparsity of design matrix $\mathbf{M}$; Recall $\hat{\mathbf{f}} = (\mathbf{I}_n + \lambda \cdot \mathbf{M})^{-1} \cdot \mathbf{Y}$
- The number of nonzeros of $\mathbf{M}$ is strictly less than $(k+1)^2 n$
- Roughly $3kn$ as shown in numerical experiments
- $\mathbf{M}$ can be permutated to a band matrix by the symmetric reverse Cuthill-Mckee ordering (1969) with $O(k \log(k) n)$ complexity
- See an example on the next page...

Results of Reordering



Complexity After Reordering

- p is the bandwidth of reordered matrix $\mathbf{M}$
- A band LDL^T decomposition procedure, plus other steps, can find all eigenvalues with pn^2 complexity . We observe that $p \approx O((kn)^{0.5})$. So the overall complexity is $O(n^{2.5})$.
- The above so far is an empirical observation

Complexity After Reordering

- p is the bandwidth of reordered matrix $\mathbf{M}$
- A band LDL^T decomposition procedure, plus other steps, can find all eigenvalues with pn^2 complexity . We observe that $p \approx O((kn)^{0.5})$. So the overall complexity is $O(n^{2.5})$.
- The above so far is an empirical observation

Complexity After Reordering

- p is the bandwidth of reordered matrix $\mathbf{M}$
- A band LDL^T decomposition procedure, plus other steps, can find all eigenvalues with pn^2 complexity . We observe that $p \approx O((kn)^{0.5})$. So the overall complexity is $O(n^{2.5})$.
- The above so far is an empirical observation

Outline

Two Major Competitors

- CDS (our method): completely-data-drive smoothing,
- Soap film (Wood *et al.* JRSSB 2008)
  - based on penalty function

$$R(f) = \int_\Omega \left(\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \right)^2 dxdy$$

- TPS: thin-plate splines
  - with $\Omega = \mathbb{R}^d$, you have $\phi(z) = \|z - x_i\|^2 \log \|z - x_i\|$ in 2-D as basis functions

Two Major Competitors

- CDS (our method): completely-data-drive smoothing,
- Soap film (Wood *et al.* JRSSB 2008)
  - based on penalty function

$$R(f) = \int_{\Omega} \left(\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \right)^2 dxdy$$

- TPS: thin-plate splines
  - with $\Omega = \mathbb{R}^d$, you have $\phi(z) = \|z - x_i\|^2 \log \|z - x_i\|$ in 2-D as basis functions

Two Major Competitors

- CDS (our method): completely-data-drive smoothing,
- Soap film (Wood *et al.* JRSSB 2008)
  - based on penalty function

$$R(f) = \int_\Omega \left(\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \right)^2 dxdy$$

- TPS: thin-plate splines
  - with $\Omega = \mathbb{R}^d$, you have $\phi(z) = \|z - x_i\|^2 \log \|z - x_i\|$ in 2-D as basis functions

Two Major Competitors

- CDS (our method): completely-data-drive smoothing,
- Soap film (Wood *et al.* JRSSB 2008)
  - based on penalty function

$$R(f) = \int_\Omega \left(\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \right)^2 dxdy$$

- TPS: thin-plate splines
  - with $\Omega = \mathbb{R}^d$, you have $\phi(z) = \|z - x_i\|^2 \log \|z - x_i\|$ in 2-D as basis functions

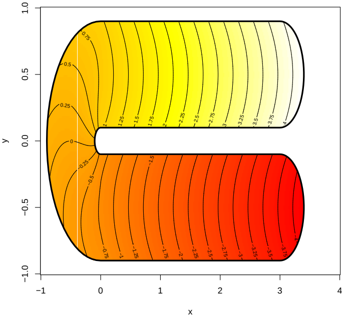
Two Major Competitors

- CDS (our method): completely-data-drive smoothing,
- Soap film (Wood *et al.* JRSSB 2008)
  - based on penalty function

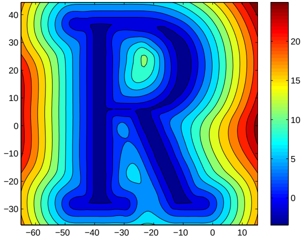
$$R(f) = \int_{\Omega} \left(\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \right)^2 dxdy$$

- TPS: thin-plate splines
  - with $\Omega = \mathbb{R}^d$, you have $\phi(z) = \|z - x_i\|^2 \log \|z - x_i\|$ in 2-D as basis functions

Recall the Irregular Domains



(a) Horseshoe.



(b) Letter R.

Horseshoe Example

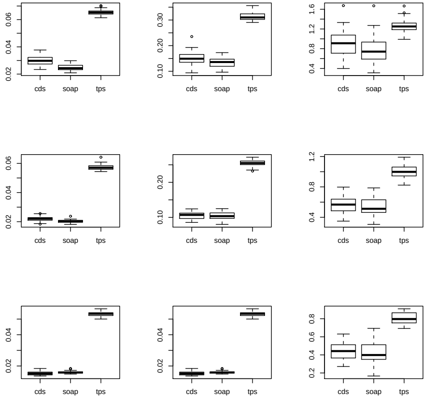
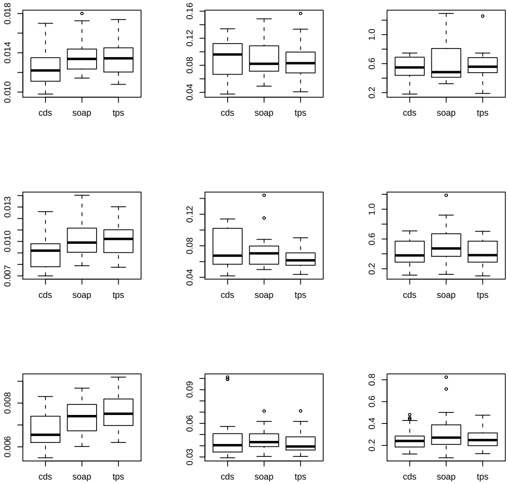


Figure: The first, second, and third rows are for $n = 1000, 2000, 5000$, respectively. From left to right the noise is dominated by $\sigma = 0.1, 1, 10$.

Regular Domain $[0, 1]^2$



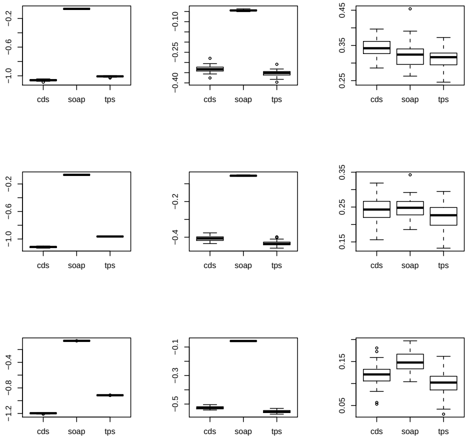


Figure: The RMSE is scaled by $\log_{10}$.

Summary / Future Plans . . .

- The idea of using local estimates to replace an analytical penalty function—*complete data driven approach*—seems appealing.
  1. Numerical experiments demonstrate promise
  2. Theoretical justification is provided
- Acknowledgment: Partially supported by NSF.
- Thank You!

Summary / Future Plans . . .

- The idea of using local estimates to replace an analytical penalty function—*complete data driven approach*—seems appealing.
  1. Numerical experiments demonstrate promise
  2. Theoretical justification is provided
- Acknowledgment: Partially supported by NSF.
- Thank You!

Summary / Future Plans . . .

- The idea of using local estimates to replace an analytical penalty function—*complete data driven approach*—seems appealing.
  1. Numerical experiments demonstrate promise
  2. Theoretical justification is provided
- Acknowledgment: Partially supported by NSF.
- Thank You!

Summary / Future Plans . . .

- The idea of using local estimates to replace an analytical penalty function—*complete data driven approach*—seems appealing.
  1. Numerical experiments demonstrate promise
  2. Theoretical justification is provided
- Acknowledgment: Partially supported by NSF.
- Thank You!

Summary / Future Plans . . .

- The idea of using local estimates to replace an analytical penalty function—*complete data driven approach*—seems appealing.
  1. Numerical experiments demonstrate promise
  2. Theoretical justification is provided
- Acknowledgment: Partially supported by NSF.
- Thank You!