0.5 setgray0 0.5 setgray1

Learning Coordinate Covariances via Gradients

Sayan Mukherjee[1] and Ding-Xuan Zhou[2]

sayan@stat.duke.edu[1], mazhou@cityu.edu.hk[2]

Institute for Genome Sciences and Policy[1]

Institute of Statistics and Decision Sciences[1]

Duke University

Department of Mathematics[2]

City University of Hong Kong[2]

Overview

- Motivation of problem

Overview

- Motivation of problem
- Tikhonov regularization

Overview

- Motivation of problem
- Tikhonov regularization
- Learning the gradient

Overview

- Motivation of problem
- Tikhonov regularization
- Learning the gradient
- A representer theorem

Overview

- Motivation of problem
- Tikhonov regularization
- Learning the gradient
- A representer theorem
- A reduced matrix algorithm

Overview

- Motivation of problem
- Tikhonov regularization
- Learning the gradient
- A representer theorem
- A reduced matrix algorithm
- Convergence of the estimate of the gradient

Overview

- Motivation of problem
- Tikhonov regularization
- Learning the gradient
- A representer theorem
- A reduced matrix algorithm
- Convergence of the estimate of the gradient
- Applications to simulated and real data

Motivation

Classification and regression of high dimensional data given few samples.

The “large p, small n” paradigm.

Tikhonov regularization/shrinkage estimators (for example SVMs) have been successful.

Motivation

Classification and regression of high dimensional data given few samples.
The “large p, small n” paradigm.
Tikhonov regularization/shrinkage estimators (for example SVMs) have been successful.

In a number of problems classical questions from statistical linear modeling have been revived

- variable saliency/significance
- coordinate covariation

However in the “large p, small n” paradigm.

Motivation

Classification and regression of high dimensional data given few samples.
The “large p, small n” paradigm.
Tikhonov regularization/shrinkage estimators (for example SVMs) have been successful.

In a number of problems classical questions from statistical linear modeling have been revived

- variable saliency/significance
- coordinate covariation

However in the “large p, small n” paradigm.

We formulate the problem of learning coordinate covariation and relevance in the framework of Tikhonov regularization or shrinkage estimation.

Tikhonov regularization

$X \subseteq \mathbb{R}^n$ is a compact metric space and $Y \subseteq \mathbb{R}$
a sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in (X \times Y)^m$

Tikhonov regularization

$X \subseteq \mathbb{R}^n$ is a compact metric space and $Y \subseteq \mathbb{R}$
a sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in (X \times Y)^m$

a hypothesis space $\mathcal{H}$ is a set of functions $f : X \to Y \subset \mathbb{R}$

Tikhonov regularization

$X \subseteq \mathbb{R}^n$ is a compact metric space and $Y \subseteq \mathbb{R}$
a sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in (X \times Y)^m$

a hypothesis space $\mathcal{H}$ is a set of functions $f : X \to Y \subset \mathbb{R}$

a loss functional $V(f(x), y) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$

Tikhonov regularization

$X \subseteq \mathbb{R}^n$ is a compact metric space and $Y \subseteq \mathbb{R}$
a sample $\mathbf{z} = \left\{(x_i, y_i)\right\}_{i=1}^m \in (X \times Y)^m$

a hypothesis space $\mathcal{H}$ is a set of functions $f : X \to Y \subset \mathbb{R}$

a loss functional $V(f(x), y) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$

a penalty or smoothness functional $\Omega : \mathcal{H} \to \mathbb{R}_+$ on $\mathcal{H}$ for example $\Omega(f) = \|f\|_K^2$

Tikhonov regularization

$X \subseteq \mathbb{R}^n$ is a compact metric space and $Y \subseteq \mathbb{R}$
a sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in (X \times Y)^m$

a hypothesis space $\mathcal{H}$ is a set of functions $f : X \to Y \subset \mathbb{R}$

a loss functional $V(f(x), y) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$

a penalty or smoothness functional $\Omega : \mathcal{H} \to \mathbb{R}_+$ on $\mathcal{H}$ for example $\Omega(f) = \|f\|_K^2$

$$f_{\mathbf{z}}^V = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{m} \sum_{i=1}^m V(y_i, f(x_i)) + \lambda \Omega(f) \right\}$$

where $\lambda > 0$

Reproducing Kernel Hilbert Spaces

$K : X \times X \to \mathbb{R}$ be continuous, symmetric and positive semidefinite is a Mercer kernel, for example

$$K(w, v) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\|u - v\|^2 / 2\sigma^2)$$

Reproducing Kernel Hilbert Spaces

RKHS is the linear span

$$\mathcal{H}_K = \text{span}\{K_x := K(x, \cdot) : x \in X\}$$

$$\langle K_v, K_u \rangle_K = K(u, v)$$

Reproducing Kernel Hilbert Spaces

RKHS is the linear span

$$\mathcal{H}_K = \text{span}\{K_x := K(x, \cdot) : x \in X\}$$

$$\langle K_v, K_u \rangle_K = K(u, v)$$

reproducing property

$$\langle K_x, f \rangle_K = f(x), \qquad \forall x \in X, f \in \mathcal{H}_K$$

Reproducing Kernel Hilbert Spaces

RKHS is the linear span

$$\mathcal{H}_K = \text{span}\{K_x := K(x, \cdot) : x \in X\}$$

$$\langle K_v, K_u \rangle_K = K(u, v)$$

reproducing property

$$\langle K_x, f \rangle_K = f(x), \qquad \forall x \in X, f \in \mathcal{H}_K$$

$$f_{\mathbf{z}}^V = \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^{m} V(y_i, f(x_i)) + \lambda \|f\|_K^2 \right\}$$

$$f_{\mathbf{z}}^V(x) = \sum_{i=1}^{m} c_i K(x_i, x)$$

optimization over $\{c_i\}_{i=1}^m \in \mathbb{R}^m$

The regression function

the joint

$$\mathbf{z} = \{(x_i, y_i)\}_{i=1}^{m} \sim \rho(x, y)$$

The regression function

the joint

$$\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \sim \rho(x, y)$$

assume $V(y, f(x)) = (y - f(x))^2$
the regression function

$$f_\rho(x) = \int_Y y d\rho(y|x), \qquad x \in X$$

The regression function

the joint

$$\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \sim \rho(x, y)$$

assume $V(y, f(x)) = (y - f(x))^2$
the regression function

$$f_\rho(x) = \int_Y y d\rho(y|x), \qquad x \in X$$

Convergence: as $\lambda = \lambda(m) \to 0$ as $m \to \infty$

$$\|f_{\mathbf{z}}^V - f_\rho\|_\rho \to 0$$

Classification

$Y = \{-1, 1\}$ and $\text{sgn}(f) : X \to Y$
loss function: $V(f(x), y) = \phi(yf(x)) := \log\left(1 + e^{-yf(x)}\right)$

$$f_{\mathbf{z}}^{V} = \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^{m} \log\left(1 + e^{-y_i f(x_i)}\right) + \lambda \|f\|_K^2 \right\}$$

Classification

$Y = \{-1, 1\}$ and sgn$(f) : X \to Y$
loss function: $V(f(x), y) = \phi(yf(x)) := \log\big(1 + e^{-yf(x)}\big)$

$$f_{\mathbf{z}}^{V} = \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^{m} \log\big(1 + e^{-y_i f(x_i)}\big) + \lambda \|f\|_K^2 \right\}$$

classification error

$$\mathcal{R}(\text{sgn}(f)) = \text{Prob}\{\text{sgn}(f(x)) \neq y\}$$

the Bayes (optimal) rule

$$\text{sgn}(f_\rho)$$

Classification

$Y = \{-1, 1\}$ and sgn$(f) : X \to Y$
loss function: $V(f(x), y) = \phi(yf(x)) := \log\left(1 + e^{-yf(x)}\right)$

$$f_{\mathbf{z}}^{V} = \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^{m} \log\left(1 + e^{-y_i f(x_i)}\right) + \lambda \|f\|_K^2 \right\}$$

classification error

$$\mathcal{R}(\text{sgn}(f)) = \text{Prob}\{\text{sgn}(f(x)) \neq y\}$$

the Bayes (optimal) rule

$$\text{sgn}(f_\rho)$$

Convergence: as $\lambda = \lambda(m) \to 0$ as $m \to \infty$

$$\mathcal{R}(\text{sgn}(f_{\mathbf{z}}^{V})) \to \mathcal{R}(\text{sgn}(f_\rho))$$

Learning the gradient

$x = (x^1, x^2, \ldots, x^n)^T \in \mathbb{R}^n$ and the gradient of f_ρ

$$\nabla f_\rho = \left(\frac{\partial f_\rho}{\partial x^1}, \ \ldots, \ \frac{\partial f_\rho}{\partial x^n} \right)^T$$

Learning the gradient

$x = (x^1, x^2, \ldots, x^n)^T \in \mathbb{R}^n$ and the gradient of f_ρ

$$\nabla f_\rho = \left(\frac{\partial f_\rho}{\partial x^1}, \ \ldots, \ \frac{\partial f_\rho}{\partial x^n} \right)^T$$

use of the gradient

- variable selection: $\left\| \frac{\partial f_\rho}{\partial x^i} \right\|$
- coordinate covariation: $\left\langle \frac{\partial f_\rho}{\partial x^i}, \frac{\partial f_\rho}{\partial x^j} \right\rangle$

Formulating the algorithm

Taylor expanding $f(u)$ around x

$$f(u) \approx f(x) + \int_{\Delta x \in \Gamma_x} \langle \nabla f, \Delta x \rangle,$$

where the inner product and neighborhood Γ_x depend on the problem setting

Formulating the algorithm

Taylor expanding $f(u)$ around x

$$f(u) \approx f(x) + \int_{\Delta x \in \Gamma_x} \langle \nabla f, \Delta x \rangle,$$

where the inner product and neighborhood Γ_x depend on the problem setting

manifold setting: ρ_X is concentrated on a manifold $\mathcal{M}$

$$f(u) \approx f(x) + \int_{\Delta x \in \mathcal{M}_x} \langle \nabla_{\mathcal{M}} f, \Delta x \rangle,$$

where $\Delta x \in \mathcal{M}_x$ and the inner product is L_2 over the manifold

Formulating the algorithm

Taylor expanding $f(u)$ around x

$$f(u) \approx f(x) + \int_{\Delta x \in \Gamma_x} \langle \nabla f, \Delta x \rangle,$$

where the inner product and neighborhood Γ_x depend on the problem setting

various algorithms can be realized by a robust mininimization of the error $f(u)$ and its expansion

$$f(x) + \int_{\Delta x \in \Gamma_x} \langle \nabla f, \Delta x \rangle \approx f(x) + \nabla f(x) \cdot (u - x) \text{ for } u \approx x$$

Elements for algorithm

loss function for regression: on sample points $x = x_i, u = x_j$

$$\big(f(u) - f(x) - \nabla f(x) \cdot (u - x)\big)^2 := \big(y_i - y_j + \vec{f}(x_i) \cdot (x_j - x_i)\big)^2 \text{ for } x_i \approx x_j$$

where $x_i \approx x_j$ given by weights: $w_{i,j} > 0$

Elements for algorithm

loss function for classification: on sample points $x = x_i, u = x_j$ and convex function ϕ (logistic)

$$\phi\Big(y_i\big(y_j + \vec{f}(x_i) \cdot (x_i - x_j)\big)\Big) \text{ for } x_i \approx x_j$$

where $x_i \approx x_j$ given by weights: $w_{i,j} > 0$

Elements for algorithm

loss function for regression: on sample points $x = x_i, u = x_j$

$$\big(f(u) - f(x) - \nabla f(x) \cdot (u - x)\big)^2 := \big(y_i - y_j + \vec{f}(x_i) \cdot (x_j - x_i)\big)^2 \text{ for } x_i \approx x_j$$

where $x_i \approx x_j$ given by weights: $w_{i,j} > 0$

weights: a natural choice is a Gaussian

$$w_{i,j} = w_{i,j}^{(s)} = \frac{1}{s^{n+2}} e^{-\frac{|x_i - x_j|^2}{2s^2}} = w(x_i - x_j), \qquad i, j = 1, \ldots, m$$

Elements for algorithm

loss function for regression: on sample points $x = x_i, u = x_j$

$$\big(f(u) - f(x) - \nabla f(x) \cdot (u - x)\big)^2 := \big(y_i - y_j + \vec{f}(x_i) \cdot (x_j - x_i)\big)^2 \text{ for } x_i \approx x_j$$

where $x_i \approx x_j$ given by weights: $w_{i,j} > 0$

weights: a natural choice is a Gaussian

$$w_{i,j} = w_{i,j}^{(s)} = \frac{1}{s^{n+2}} e^{-\frac{|x_i - x_j|^2}{2s^2}} = w(x_i - x_j), \qquad i, j = 1, \ldots, m$$

regularization: $\mathcal{H}_K^n$ is an n-fold of $\mathcal{H}_K$ and $\vec{f} = (f_1, f_2, \ldots, f_n)^T$ with $f_\ell \in \mathcal{H}_K$

$$\langle \vec{f}, \vec{g} \rangle_K = \sum_{\ell=1}^{n} \langle f_\ell, g_\ell \rangle_K \text{ and } \|\vec{f}\|_K^2 = \sum_{\ell=1}^{n} \|f_\ell\|_K^2$$

Gradient algorithms

Definition 1. *The least-square type learning algorithm is defined for the sample* $\mathbf{z} \in Z^m$ *as*

$$\vec{f}_{\mathbf{z},\lambda} := \arg \min_{\vec{f} \in \mathcal{H}_K^n} \left\{ \frac{1}{m^2} \sum_{i,j=1}^{m} w_{i,j}^{(s)} \Big(y_i - y_j + \vec{f}(x_i) \cdot (x_j - x_i) \Big)^2 + \lambda \|\vec{f}\|_K^2 \right\},$$

where λ, s *are two positive constants called the regularization parameters.*

Gradient algorithms

Definition 2. *The least-square type learning algorithm is defined for the sample* $\mathbf{z} \in Z^m$ *as*

$$\vec{f}_{\mathbf{z},\lambda} := \arg \min_{\vec{f} \in \mathcal{H}_K^n} \left\{ \frac{1}{m^2} \sum_{i,j=1}^{m} w_{i,j}^{(s)} \Big(y_i - y_j + \vec{f}(x_i) \cdot (x_j - x_i) \Big)^2 + \lambda \|\vec{f}\|_K^2 \right\},$$

where λ, s *are two positive constants called the regularization parameters.*

Definition 3. *The regularization scheme for classification is defined for the sample* $\mathbf{z} \in Z^m$ *as*

$$\vec{f}_{\mathbf{z},\lambda} = \arg \min_{\vec{f} \in \mathcal{H}_K^n} \left\{ \frac{1}{m^2} \sum_{i,j=1}^{m} w_{i,j}^{(s)} \phi\Big(y_i \big(y_j + \vec{f}(x_i) \cdot (x_i - x_j) \big) \Big) + \lambda \|\vec{f}\|_K^2 \right\}.$$

Remark

Why not estimate f_ρ and then take partial derivatives ?

Remark

Why not estimate f_ρ and then take partial derivatives ?

When we obtain an approximation of f_ρ it is in a particular RKHS.
However, its partial derivatives are not.
Hence, there is no natural ways to find the correlations.
For example for the Gaussian kernel, there are no natural inner products among its partial derivatives, especially when there are no natural coordinates for the underlying manifold.

Representer theorems

For both classification and regression

$$\vec{f}_{\mathbf{z},\lambda}(x) = \sum_{i=1}^{m} c_{i,\mathbf{z}} K(x_i, x),$$

where $c_{i,\mathbf{z}} \in \mathbb{R}^n$.

Representer theorems

For regression

Theorem 1. *For* $i = 1, \ldots, m$, *let* B_i

$$B_i = \sum_{j=1}^{m} w_{i,j}(x_j - x_i)(x_j - x_i)^T \in \mathbb{R}^{n\times n}, \quad Y_i = \sum_{j=1}^{m} w_{i,j}(y_j - y_i)(x_j - x_i) \in \mathbb{R}^n.$$

Then

$$\vec{f}_{\mathbf{z},\lambda}(x) = \sum_{i=1}^{m} c_{i,\mathbf{z}} K(x_i, x)$$

with $c_{\mathbf{z}} = (c_{1,\mathbf{z}}^T, \ldots, c_{m,\mathbf{z}}^T)^T \in \mathbb{R}^{mn}$ *satisfying the linear system*

$$\left\{ m^2 \lambda I_{mn} + \textit{diag}\{B_1, B_2, \cdots, B_m\} \left[K(x_i, x_j) I_n\right]_{i,j=1}^{m} \right\} c = (Y_1^T, Y_2^T, \ldots, Y_m^T)^T.$$

The above is a linear system of size $mn \times mn$ which is prohibitive if $n >> m$.

Reducing the matrix size

Each term in the summation defining B_i is a rank one matrix.
Hence the rank of the $n \times n$ matrix B_i is at most m.
This allows us to reduce the system to $(dm) \times (dm)$ with $d \leq m - 1$.

Denote $V_{\mathbf{x}} = \text{span}\{x_j - x_m\}_{j=1}^{m-1}$, the subspace of $\mathbb{R}^n$ generated by the vectors $\{x_j - x_m\}$.

Reducing the matrix size

Theorem 2. *Let an $n \times d$ matrix $V = (V_1, \dots, V_d)$ have linearly independent column vectors and its column space span$\{V_\ell\}_{\ell=1}^d$ contains $V_{\mathbf{x}}$. Write*

$$x_j - x_m = \sum_{\ell=1}^{d} \widetilde{x}_j^\ell V_\ell = V \widetilde{x}_j$$

with $\widetilde{x}_j \in \mathbb{R}^d$ for each j. Then

$$\vec{f}_{\mathbf{z},\lambda}(x) = \sum_{i=1}^{m} \{\sum_{\ell=1}^{d} \widetilde{c}_{i,\mathbf{z}}^\ell V_\ell\} K(x_i, x)$$

with $\widetilde{c}_{\mathbf{z}} = (\widetilde{c}_{1,\mathbf{z}}^T, \dots, \widetilde{c}_{m,\mathbf{z}}^T)^T \in \mathbb{R}^{md}$ satisfying

$$\left\{ m^2 \lambda I_{md} + \text{diag}\{\widetilde{B}_1, \widetilde{B}_2, \cdots, \widetilde{B}_m\} \left[K(x_i, x_j) I_d\right]_{i,j=1}^m \right\} \widetilde{c} = (\widetilde{Y}_1^T, \widetilde{Y}_2^T, \dots, \widetilde{Y}_m^T)^T.$$

where
$\widetilde{B}_i = \sum_{j=1}^m w_{i,j} (\widetilde{x}_j - \widetilde{x}_i)(x_j - x_i)^T V \in \mathbb{R}^{d \times d}, \quad \widetilde{Y}_i = \sum_{j=1}^m w_{i,j} (y_j - y_i)(\widetilde{x}_j - \widetilde{x}_i) \in \mathbb{R}^d.$

Convergence to the gradient

Proposition 1. *Assume $|y| \leq M$ almost surely. Suppose that for some $0 < \tau \leq 2/3$, $c_\rho > 0$, the marginal distribution ρ_X satisfies*

$$\rho_X\left(\{x \in X : \inf_{u \notin X} |u - x| \leq s\}\right) \leq c_\rho^2 s^{4\tau}, \qquad \forall s > 0,$$

and the density $p(x)$ of $d\rho_X(x)$ exists and satisfies

$$\sup_{x \in X} p(x) \leq c_\rho, \quad |p(x) - p(u)| \leq c_\rho |u - x|^\tau, \qquad \forall u, x \in X.$$

Choose $\lambda = \lambda(m) = m^{-\frac{\tau}{n+2+3\tau}}$ and $s = s(m) = (\kappa c_\rho)^{\frac{2}{\tau}} m^{-\frac{1}{n+2+3\tau}}$. If $\nabla f_\rho \in \mathcal{H}_K^n$ and the kernel K is C^3, then there is a constant $C_{\rho,K}$ such that for any $0 < \delta < 1$ and $m \geq 1$, with confidence $1 - \delta$, we have

$$\|\vec{f}_{\mathbf{z},\lambda} - \nabla f_\rho\|_\rho \leq C_{\rho,K} \log\left(\frac{2}{\delta}\right)\left(\frac{1}{m}\right)^{\frac{\tau}{2(n+2+3\tau)}}.$$

Quantities of interest

Definition 4. *The relative magnitude of the norm for the variables is defined as*

$$s_\ell^\rho = \frac{\|(\vec{f}_{\mathbf{z},\lambda})_\ell\|_K}{\left(\sum_{j=1}^n \|(\vec{f}_{\mathbf{z},\lambda})_j\|_K^2\right)^{1/2}}, \qquad \ell = 1, \dots, n.$$

Quantities of interest

Definition 5. *The relative magnitude of the norm for the variables is defined as*

$$s_\ell^\rho = \frac{\|(\vec{f}_{\mathbf{z},\lambda})_\ell\|_K}{\left(\sum_{j=1}^n \|(\vec{f}_{\mathbf{z},\lambda})_j\|_K^2\right)^{1/2}}, \qquad \ell = 1, \dots, n.$$

Definition 6. *The* **empirical gradient matrix** *(EGM),* $F_{\mathbf{z}}$*, is the* $n \times m$ *matrix whose columns are* $\vec{f}_{\mathbf{z},\lambda}(x_j)$ *with* $j = 1, \dots, m$*. The* **empirical covariance matrix** *(ECM),* $\Xi_{\mathbf{z}}$*, is the* $n \times n$ *matrix of inner products of the gradient between two coordinates*

$$\mathit{Cov}(\vec{f}_{\mathbf{z},\lambda}) := \left[\langle(\vec{f}_{\mathbf{z},\lambda})_p, (\vec{f}_{\mathbf{z},\lambda})_q\rangle_K\right]_{p,q=1}^n = \sum_{i,j=1}^m c_{i,\mathbf{z}} c_{j,\mathbf{z}}^T K(x_i, x_j).$$

Simulated data

Construct a function in an $n = 80$ dimensional space which consists of three linear functions over different partitions of the space. So $\{(x_i, y_i)\}_{i=1}^{30}$ with $y \in \mathbb{R}$ and $x \in \mathbb{R}^{80}$.

Simulated data

$\{x_i\}_{i=1}^{30}$ partition the space

1. For samples $\{x_i\}_{i=1}^{10}$

$$x^j = \mathcal{N}(1, \sigma_x), \text{ for } j = 1, \ldots, 10; \qquad x^j = \mathcal{N}(0, \sigma_x), \text{ for } j = 11, \ldots, 80.$$

2. For samples $\{x_i\}_{i=11}^{20}$

$$x^j = \mathcal{N}(1, \sigma_x), \text{ for } j = 11, \ldots, 20; \qquad x^j = \mathcal{N}(0, \sigma_x), \text{ for } j = 1, \ldots, 10, 21, \ldots, 80.$$

3. For samples $\{x_i\}_{i=21}^{30}$

$$x^j = \mathcal{N}(-1, \sigma_x), \text{ for } j = 41, \ldots, 50; \qquad x^j = \mathcal{N}(0, \sigma_x), \text{ for } j = 1, \ldots, 40, 51, \ldots, 80.$$

Simulated data

Vectors corresponding to different linear functions over partitions

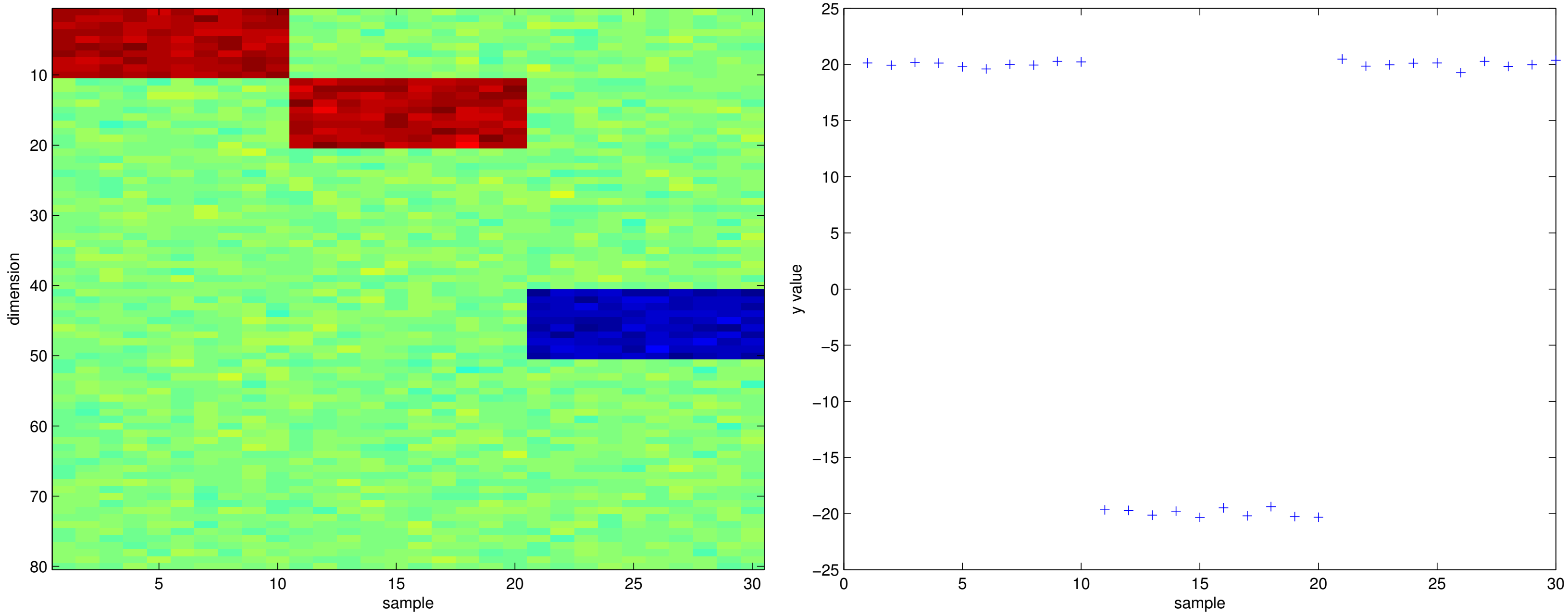
$$
\begin{aligned}
w_1 &= 2 + .5\sin(2\pi i/10) \text{ for } i = 1, ..., 10 \text{ and } 0 \text{ otherwise,}\\
w_2 &= -2 - .5\sin(2\pi i/10) \text{ for } i = 11, ..., 20 \text{ and } 0 \text{ otherwise,}\\
w_3 &= -2 - .5\sin(2\pi i/10) \text{ for } i = 41, ..., 50 \text{ and } 0 \text{ otherwise.}
\end{aligned}
$$

Simulated data

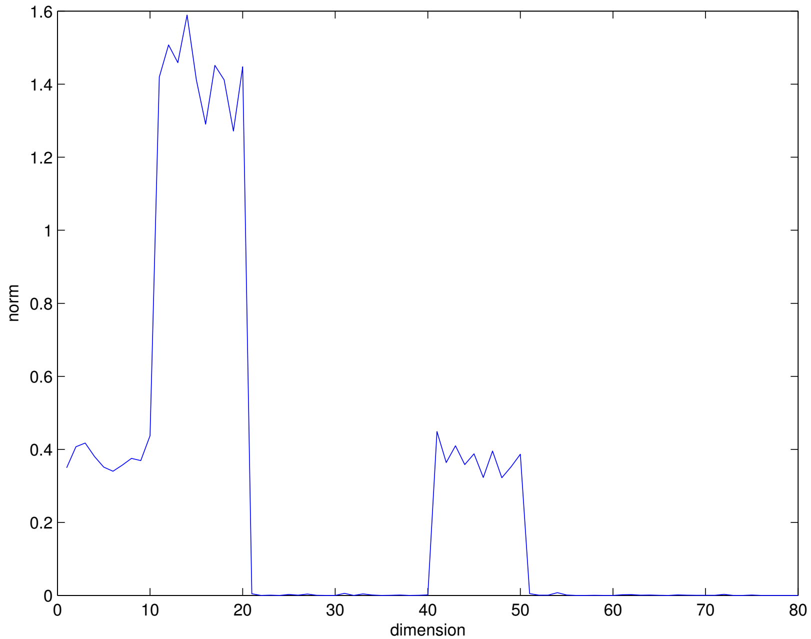
$\{y_i\}_{i=1}^{30}$

1. For samples $\{y_i\}_{i=1}^{10}$
$$y_i = x_i \cdot w_1 + \mathcal{N}(0, \sigma_y),$$
2. For samples $\{y_i\}_{i=11}^{20}$
$$y_i = x_i \cdot w_2 + \mathcal{N}(0, \sigma_y),$$
3. For samples $\{y_i\}_{i=21}^{30}$
$$y_i = x_i \cdot w_3 + \mathcal{N}(0, \sigma_y).$$

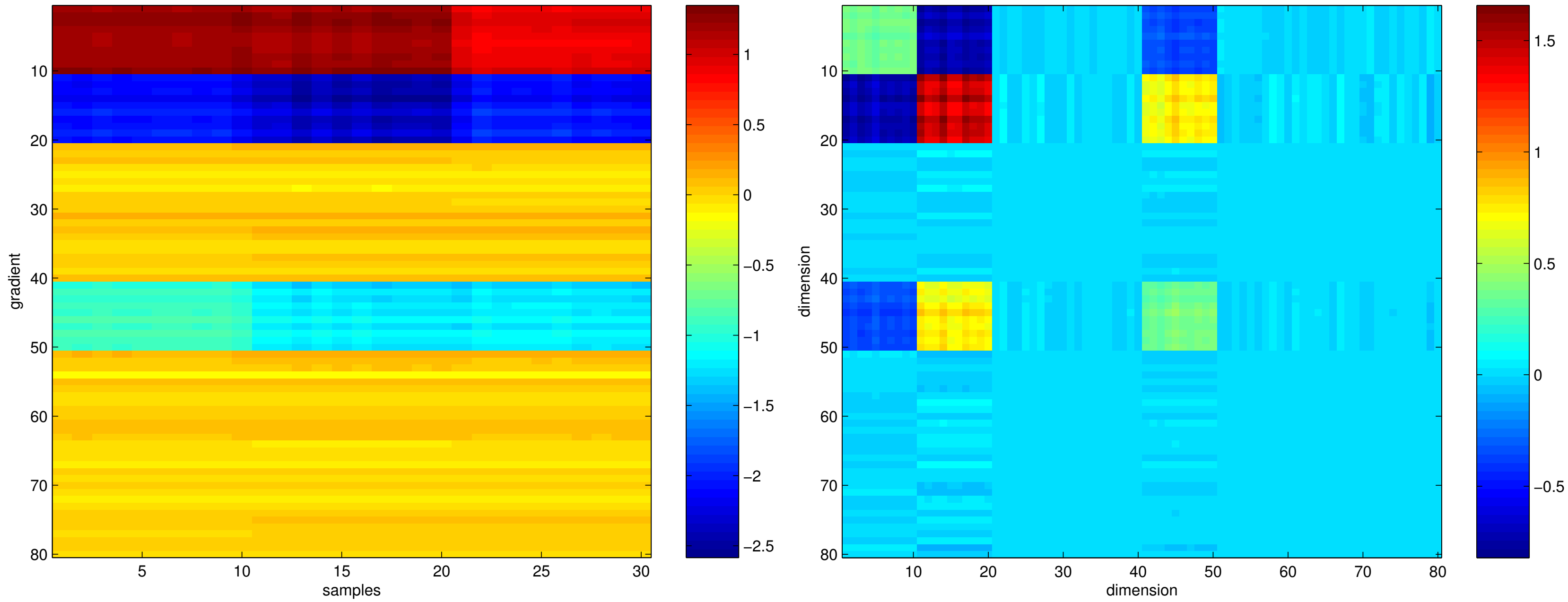
Simulated data



Simulated data



Simulated data



Gene expression data

Expression (number of copies of mRNA) for $7,129$ genes and ESTs were measured over 73 patients with either AML (myeloid leukemia) or ALL (lymphoblastic leukemia)

$\{(x_i, y_i)\}_{i=1}^{73}$ with $x \in \mathbb{R}^{7129}$ and $y \in \{-1, 1\}$

38 samples were used for the training set, 35 for the test set

Gene expression data

Expression (number of copies of mRNA) for $7,129$ genes and ESTs were measured over 73 patients with either AML (myeloid leukemia) or ALL (lymphoblastic leukemia)

$\{(x_i, y_i)\}_{i=1}^{73}$ with $x \in \mathbb{R}^{7129}$ and $y \in \{-1, 1\}$

38 samples were used for the training set, 35 for the test set

genes (S)	5	55	105	155	205	255	305	355	405	455
test errors	1	3	2	1	1	1	1	1	1	1

Gene expression data

Expression (number of copies of mRNA) for $7,129$ genes and ESTs were measured over 73 patients with either AML (myeloid leukemia) or ALL (lymphoblastic leukemia)

$\{(x_i, y_i)\}_{i=1}^{73}$ with $x \in \mathbb{R}^{7129}$ and $y \in \{-1, 1\}$

38 samples were used for the training set, 35 for the test set

genes (S)	5	55	105	155	205	255	305	355	405	455
test errors	1	3	2	1	1	1	1	1	1	1

Decay of norms

The decay of $s^{\rho}_{(\ell)}$ is a measure of how many features are significant

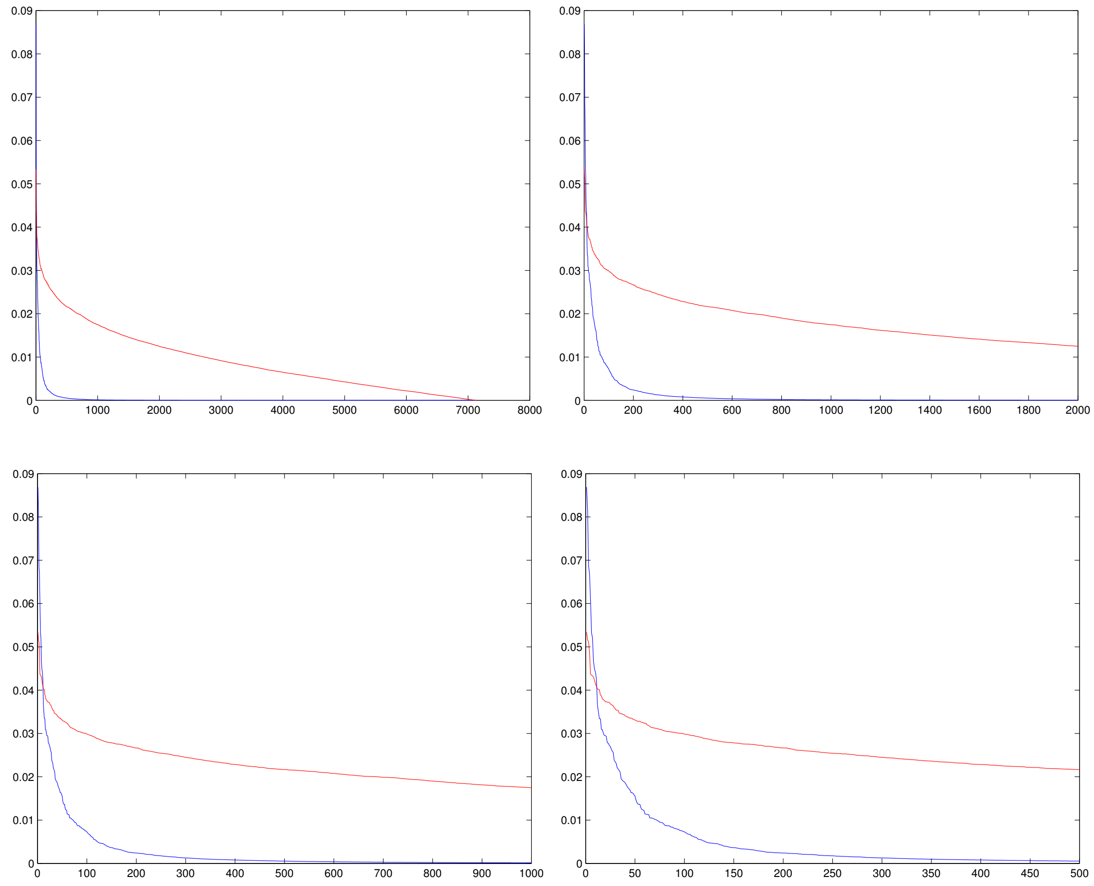
Decay of norms

Fisher score:

$$t_\ell = \frac{|\hat{\mu}_\ell^{\text{AML}} - \hat{\mu}_\ell^{\text{ALL}}|}{\hat{\sigma}_\ell^{\text{AML}} + \hat{\sigma}_\ell^{\text{ALL}}},$$

$$s_\ell^F = \frac{t_\ell}{\left(\sum_{p=1}^n t_p^2\right)^{1/2}}$$

Decay of norms



Discussion

There are many extensions and refinements to this method which we discuss below:

- Logistic regression model: reduced matrix implementation and analysis

Discussion

There are many extensions and refinements to this method which we discuss below:

- Logistic regression model: reduced matrix implementation and analysis
- Fully Bayesian model: compute the full posterior using MCMC

Discussion

There are many extensions and refinements to this method which we discuss below:

- Logistic regression model: reduced matrix implementation and analysis
- Fully Bayesian model: compute the full posterior using MCMC
- Intrinsic dimension: rate of convergence of the gradient has the form of $O(m^{-1/n})$, it would be good to replace n with $n_{\mathcal{M}}$

Discussion

There are many extensions and refinements to this method which we discuss below:

- Logistic regression model: reduced matrix implementation and analysis
- Fully Bayesian model: compute the full posterior using MCMC
- Intrinsic dimension: rate of convergence of the gradient has the form of $O(m^{-1/n})$, it would be good to replace n with $n_{\mathcal{M}}$
- Semi-supervised setting: given $\mathbf{x} = (x_i)_{i=m+1}^{m+u}$ in addition to $\mathbf{z}$

$$\begin{aligned}\vec{f}_{\mathbf{z},\mathbf{x},\lambda,\mu} &= \arg\min_{\vec{f}\in\mathcal{H}_K^n} \Big\{ \tfrac{1}{m^2} \textstyle\sum_{i,j=1}^m w_{i,j}^{(s)} \Big(y_i - y_j + \vec{f}(x_i)\cdot(x_j - x_i) \Big)^2 \\ &\quad + \tfrac{\mu}{(m+u)^2} \textstyle\sum_{i,j=1}^{m+u} W_{i,j} |\vec{f}(x_i) - \vec{f}(x_j)|_{\ell^2(\mathbb{R}^n)}^2 + \lambda \|\vec{f}\|_K^2 \Big\},\end{aligned}$$

where $\{W_{i,j}\}$ are edge weights in the data adjacency graph, μ is another regularization parameter and often satisfies $\lambda = o(\mu)$.

Discussion

There are many extensions and refinements to this method which we discuss below:

- Logistic regression model: reduced matrix implementation and analysis
- Fully Bayesian model: compute the full posterior using MCMC
- Intrinsic dimension: rate of convergence of the gradient has the form of $O(m^{-1/n})$, it would be good to replace n with $n_{\mathcal{M}}$
- Semi-supervised setting: given $\mathbf{x} = (x_i)_{i=m+1}^{m+u}$ in addition to $\mathbf{z}$

$$\begin{aligned}\vec{f}_{\mathbf{z},\mathbf{x},\lambda,\mu} &= \arg\min_{\vec{f}\in\mathcal{H}_K^n} \Big\{ \tfrac{1}{m^2} \textstyle\sum_{i,j=1}^m w_{i,j}^{(s)} \Big(y_i - y_j + \vec{f}(x_i)\cdot(x_j - x_i)\Big)^2 \\ &\quad + \tfrac{\mu}{(m+u)^2} \textstyle\sum_{i,j=1}^{m+u} W_{i,j} |\vec{f}(x_i) - \vec{f}(x_j)|^2_{\ell^2(\mathbb{R}^n)} + \lambda \|\vec{f}\|_K^2 \Big\},\end{aligned}$$

where $\{W_{i,j}\}$ are edge weights in the data adjacency graph, μ is another regularization parameter and often satisfies $\lambda = o(\mu)$.

Acknowledgements

André Elisseeff, Misha Belkin, Aravind Subramanian, Tommy Poggio, and Steve Smale for discussions