Estimation of Gradients and Coordinate Covariation in Classification

Sayan Mukherjee and Qiang Wu

sayan@stat.duke.edu

Institute of Statistics and Decision Sciences

Department of Computer Science

Institute for Genome Sciences & Policy

Duke University

Motivation

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

In a number of problems classical questions from statistical 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 ridge regression or SVMs) have been successful.

In a number of problems classical questions from statistical 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.

Global shrinkage estimators

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

Global shrinkage estimators

$\mathcal{X} \subseteq \mathbb{R}^p$ is a compact metric space, $\mathcal{Y} \in \{-1, 1\}$, and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
a sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$
a hypothesis space $\mathcal{H}$ is a set of functions $f : \mathcal{X} \to \mathbb{R}$

Global shrinkage estimators

$\mathcal{X} \subseteq \mathbb{R}^p$ is a compact metric space, $\mathcal{Y} \in \{-1, 1\}$, and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$

a sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$

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

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

Global shrinkage estimators

$\mathcal{X} \subseteq \mathbb{R}^p$ is a compact metric space, $\mathcal{Y} \in \{-1, 1\}$, and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
a sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$
a hypothesis space $\mathcal{H}$ is a set of functions $f : \mathcal{X} \to \mathbb{R}$
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 $\Omega(f) = \|f\|_K^2$

Global shrinkage estimators

$f^V_{\mathbf{z},\lambda}$ can be interpreted as a MAP estimate

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

where $\lambda > 0$

Reproducing Kernel Hilbert Spaces

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

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

Reproducing Kernel Hilbert Spaces

RKHS is the linear span

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

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

Reproducing Kernel Hilbert Spaces

RKHS is the linear span

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

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

reproducing property

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

Reproducing Kernel Hilbert Spaces

RKHS is the linear span

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

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

reproducing property

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

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

$$f^V_{\mathbf{z},\lambda}(x) = \sum_{i=1}^{n} c_i K(x_i, x)$$

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

Classification

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

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

Classification

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

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

classification error

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

the Bayes optimal classifier

$$\text{sgn}(f_\rho(x)) = 1 \text{ if } \rho(y = 1|x) \geq \rho(y = -1|x) \text{ and } -1 \text{ otherwise.}$$

$$f_\rho(x) = \log\left[\frac{\rho(y = 1|x)}{\rho(y = -1|x)}\right].$$

Classification

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

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

classification error

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

the Bayes optimal classifier

$$\text{sgn}(f_\rho(x)) = 1 \text{ if } \rho(y = 1|x) \geq \rho(y = -1|x) \text{ and } -1 \text{ otherwise.}$$

$$f_\rho(x) = \log\left[\frac{\rho(y = 1|x)}{\rho(y = -1|x)}\right].$$

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

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

Learning the gradient

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

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

Learning the gradient

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

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

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 $f_\rho(u)$

$$f_\rho(x) \approx f_\rho(u) + \nabla f_\rho(x) \cdot (x - u) \quad \text{for} \quad x \approx u.$$

Formulating the algorithm

Taylor expanding $f_\rho(u)$

$$f_\rho(x) \approx f_\rho(u) + \nabla f_\rho(x) \cdot (x - u) \quad \text{for} \quad x \approx u.$$

Estimate f_ρ by g and ∇f_ρ by $\vec{f} = (f_1, f_2, \ldots, f_p)^T : \mathcal{X} \to \mathbb{R}^p$.

$$f_\rho(x) \approx g(u) + \vec{f}(x) \cdot (x - u) \quad \text{for} \quad x \approx u.$$

Elements for algorithm

locality: weight by a Gaussian

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

Elements for algorithm

locality: weight by a Gaussian

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

cost function: $\phi(\eta) = \log\left(1 + e^{-\eta}\right)$

$$\mathcal{E}_{\mathbf{z}}(g, \vec{f}) = \frac{1}{n^2} \sum_{i,j=1}^{n} w_{i,j}^{(s)} \phi\left(y_i(g(x_j) + \vec{f}(x_i) \cdot (x_i - x_j))\right).$$

Elements for algorithm

locality: weight by a Gaussian

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

cost function: $\phi(\eta) = \log\left(1 + e^{-\eta}\right)$

$$\mathcal{E}_{\mathbf{z}}(g,\vec{f}) = \frac{1}{n^2} \sum_{i,j=1}^{n} w_{i,j}^{(s)} \phi\left(y_i(g(x_j) + \vec{f}(x_i) \cdot (x_i - x_j))\right).$$

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

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

Gradient algorithms

Definition 1. *Given a sample* $\mathbf{z}$ *we can estimate the classification function,* $g_{\mathbf{z}}$*, and its gradient,* $\vec{f}_{\mathbf{z}}$*, as follows:*

$$(g_{\mathbf{z}}, \vec{f}_{\mathbf{z}}) = \arg \min_{(g,\vec{f}) \in \mathcal{H}_K^{p+1}} \left[\mathcal{E}_{\mathbf{z}}(g, \vec{f}) + \lambda_1 \|g\|_K^2 + \lambda_2 \|\vec{f}\|_K^2 \right],$$

where $s, \lambda_1, \lambda_2 > 0$ *are the regularization parameters.*

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

Proposition 1. *Given a sample* $\mathbf{z} \in \mathcal{Z}^m$ *the solution takes the form and exists*

$$g_{\mathbf{z}}(x) = \sum_{i=1}^{n} \alpha_{i,\mathbf{z}} K(x, x_i) \quad \textit{and} \quad \vec{f}_{\mathbf{z}}(x) = \sum_{i=1}^{n} c_{i,\mathbf{z}} K(x, x_i)$$

with $c_{\mathbf{z}} = (c_{1,\mathbf{z}}, \ldots, c_{n,\mathbf{z}}) \in \mathbb{R}^{p \times n}$ *and* $\alpha_{\mathbf{z}} = (\alpha_{1,\mathbf{z}}, ..., \alpha_{n,\mathbf{z}})^T \in \mathbb{R}^n$.

Representer theorems

Proposition 2. *Given a sample $\mathbf{z} \in \mathcal{Z}^m$ the solution takes the form and exists*

$$g_{\mathbf{z}}(x) = \sum_{i=1}^{n} \alpha_{i,\mathbf{z}} K(x, x_i) \quad \textit{and} \quad \vec{f}_{\mathbf{z}}(x) = \sum_{i=1}^{n} c_{i,\mathbf{z}} K(x, x_i)$$

with $c_{\mathbf{z}} = (c_{1,\mathbf{z}}, \ldots, c_{n,\mathbf{z}}) \in \mathbb{R}^{p \times n}$ *and* $\alpha_{\mathbf{z}} = (\alpha_{1,\mathbf{z}}, \ldots, \alpha_{n,\mathbf{z}})^T \in \mathbb{R}^n$*.*

The coefficients are computed using Newton's method in what naïvely looks like an optimization problem in $\mathbb{R}^{np \times np}$ which is prohibitive.

Reducing the matrix size

The functional in matrix form

$$\Phi(C,\alpha) = \frac{1}{m^2}\sum_{i,j=1}^{n} w_{i,j}\phi\big(y_i(k_j\alpha + k_iC^T(x_i - x_j))\big) + \frac{\lambda}{2}\big(\alpha^T K\alpha + \mathrm{Tr}(CKC^T)\big),$$

We solve for C, α by setting

$$\nabla\Phi(\alpha, C) = 0$$

using Newton's method.

Reducing the matrix size

The functional in matrix form

$$\Phi(C,\alpha) = \frac{1}{m^2}\sum_{i,j=1}^{n} w_{i,j}\phi\big(y_i(k_j\alpha + k_i C^T(x_i - x_j))\big) + \frac{\lambda}{2}\big(\alpha^T K\alpha + \mathrm{Tr}(CKC^T)\big),$$

We solve for C, α by setting

$$\nabla\Phi(\alpha, C) = 0$$

using Newton's method.

A key quantity in the optimization is the data matrix

$$M_{\mathbf{x}} = (x_1 - x_n, x_2 - x_n, \ldots, x_{n-1} - x_n, x_n - x_n) \in \mathbb{R}^{p\times n}.$$

it has rank $d \leq n-1$ so our optimization is in $\mathbb{R}^{nd\times nd}$ with runtime of $O(nd^2)$ and memory $O(np)$.

Convergence to the gradient

Proposition 3. *If for some constants $c_\rho > 0$ and $0 < \theta \le 1$ the marginal distribution ρ_X satisfies*

$$\rho_X\left(\{x \in \mathcal{X} : d(x, \partial\mathcal{X}) < s\}\right) \le c_\rho s,$$

the density $p(x)$ of ρ_X exists and satisfies

$$\sup_{x \in \mathcal{X}} p(x) \le c_\rho \quad \text{and} \quad |p(x) - p(u)| \le c_\rho |x - u|^\theta, \ \ \forall u, x \in \mathcal{X},$$

then with probability $1 - \delta$

$$\begin{aligned} \|\vec{f}_{\mathbf{z}} - \nabla f_\rho\|_{\rho_X} &\le C \log\left(\frac{2}{\delta}\right) n^{-1/p} \\ \|g_{\mathbf{z}} - f_\rho\|_{\rho_X} &\le C \log\left(\frac{2}{\delta}\right) n^{-1/p}. \end{aligned}$$

Quantities of interest

Definition 2. *The relative magnitude of the norm for the variables is defined as*

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

Quantities of interest

Definition 3. *The relative magnitude of the norm for the variables is defined as*

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

Definition 4. *The* **empirical covariance matrix** *(ECM), $\Xi_{\mathbf{z}}$, is the $p \times p$ matrix of inner products of the gradient between two coordinates*

$$\mathit{Cov}(\vec{f}_{\mathbf{z}}) := \left[\langle (\vec{f}_{\mathbf{z}})_k, (\vec{f}_{\mathbf{z}})_l \rangle_K\right]_{k,l=1}^p = \sum_{i,j=1}^n c_{i,\mathbf{z}} c_{j,\mathbf{z}}^T K(x_i, x_j).$$

Linear example

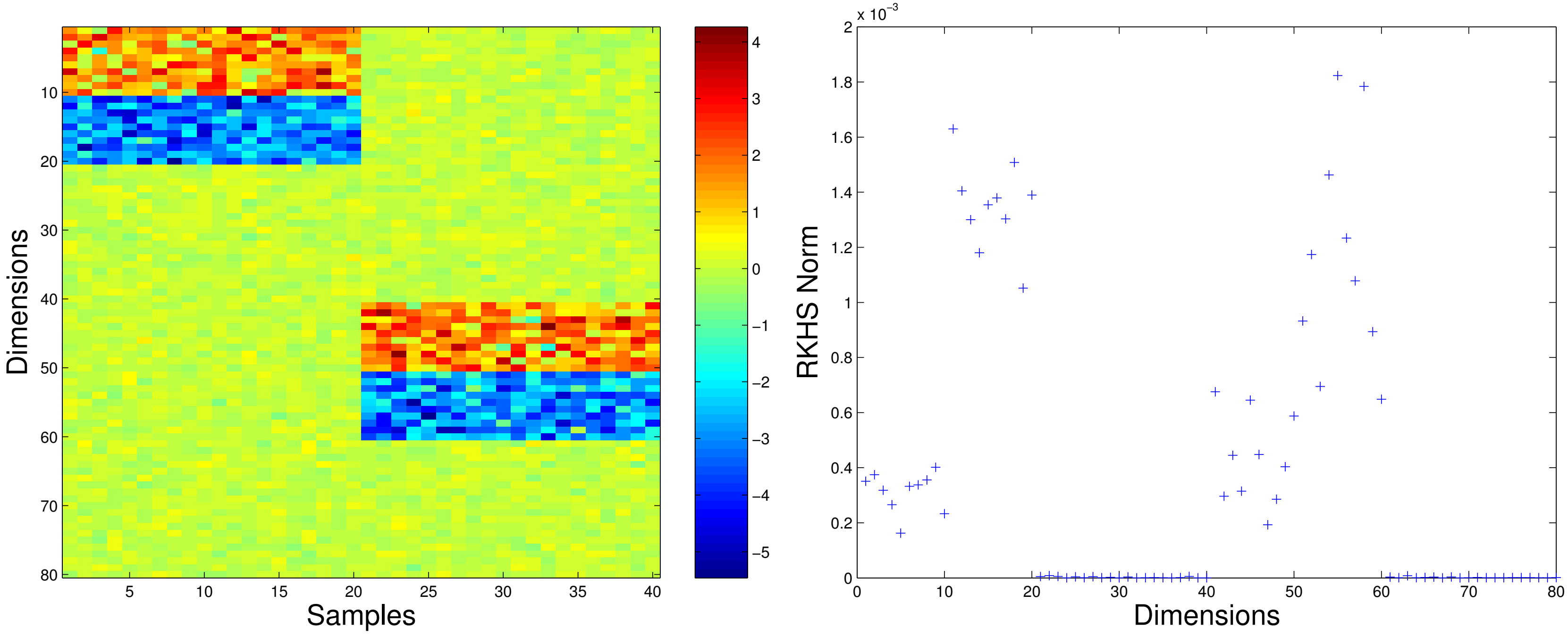
Samples from class $+1$ were drawn from

$$\begin{aligned} x^j &\sim \mathcal{N}(1.5, 1), \text{ for } j = 1, \ldots, 10, \\ x^j &\sim \mathcal{N}(-3, 1), \text{ for } j = 11, \ldots, 20, \\ x^j &\sim \mathcal{N}(0, \sigma_{\text{noise}}), \text{ for } j = 21, \ldots, 80, \end{aligned}$$

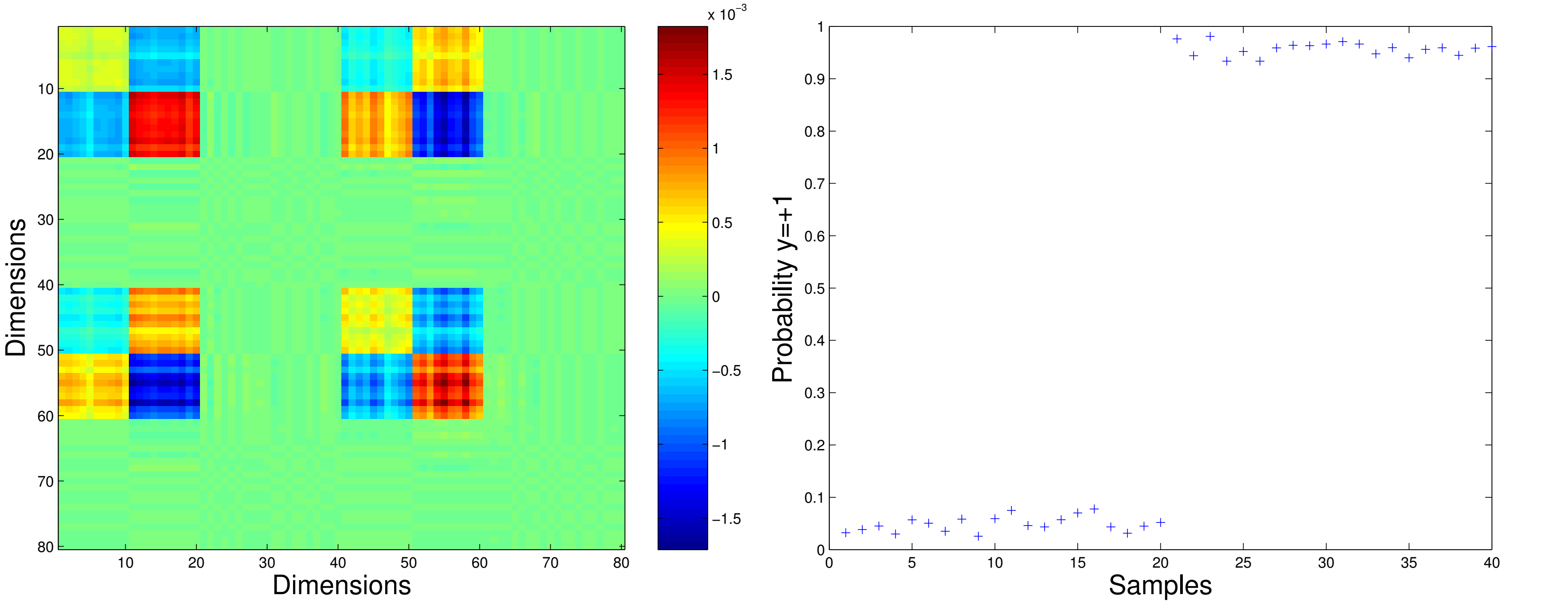
Samples from class -1 were drawn from

$$\begin{aligned} x^j &\sim \mathcal{N}(1.5, 1), \text{ for } j = 41, \ldots, 50, \\ x^j &\sim \mathcal{N}(-3, 1), \text{ for } j = 51, \ldots, 60, \\ x^j &\sim \mathcal{N}(0, \sigma_{\text{noise}})), \text{ for } j = 1, \ldots, 40, 61, \ldots, 80. \end{aligned}$$

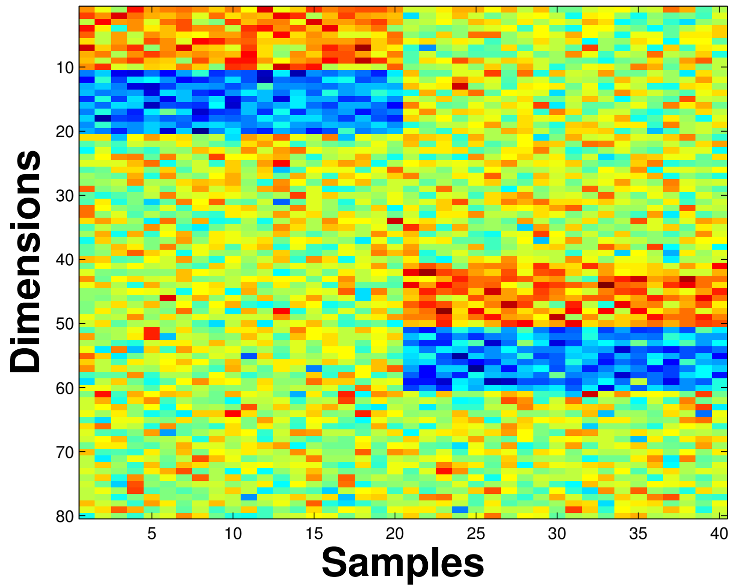
Linear example

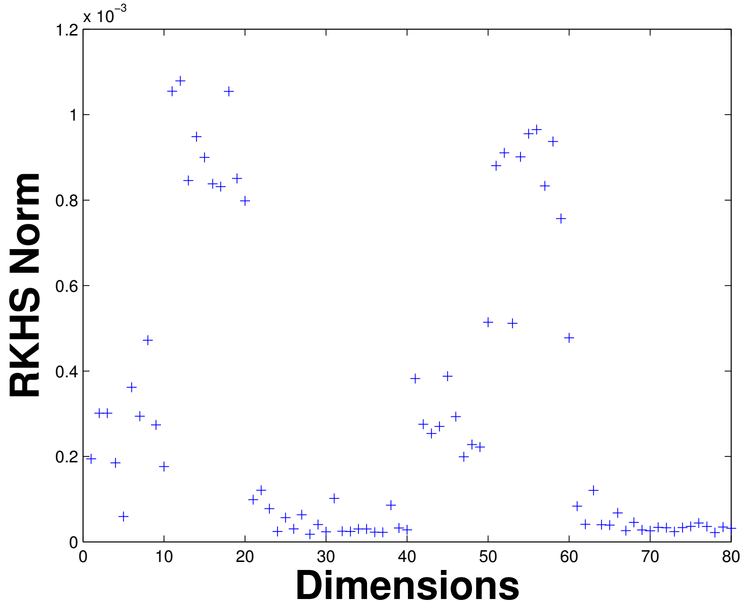


Linear example

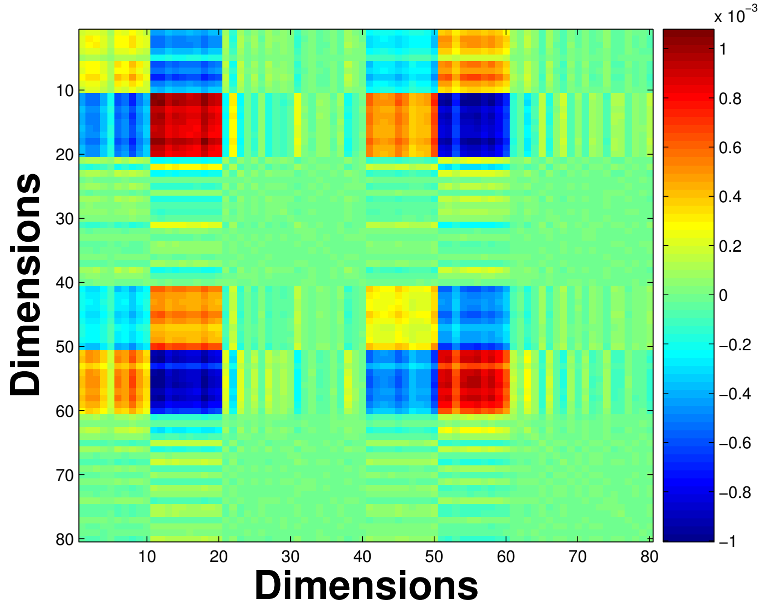


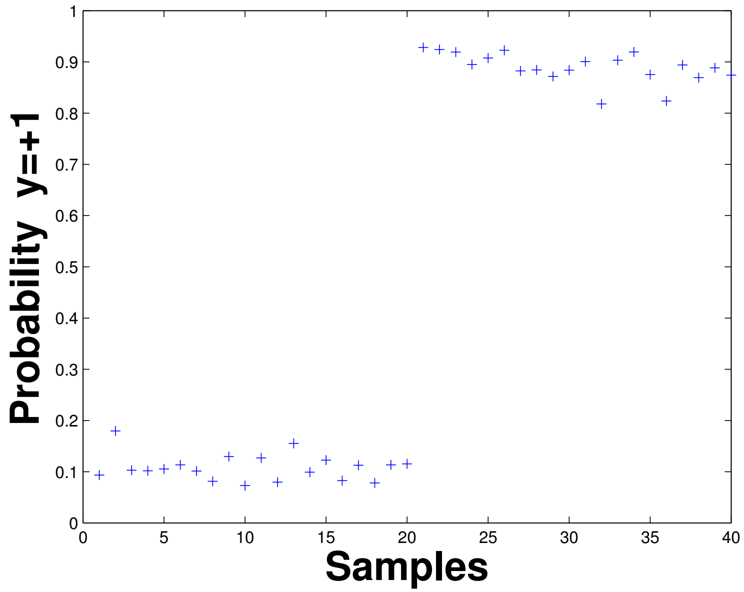
Linear example





Linear example





Nonlinear example

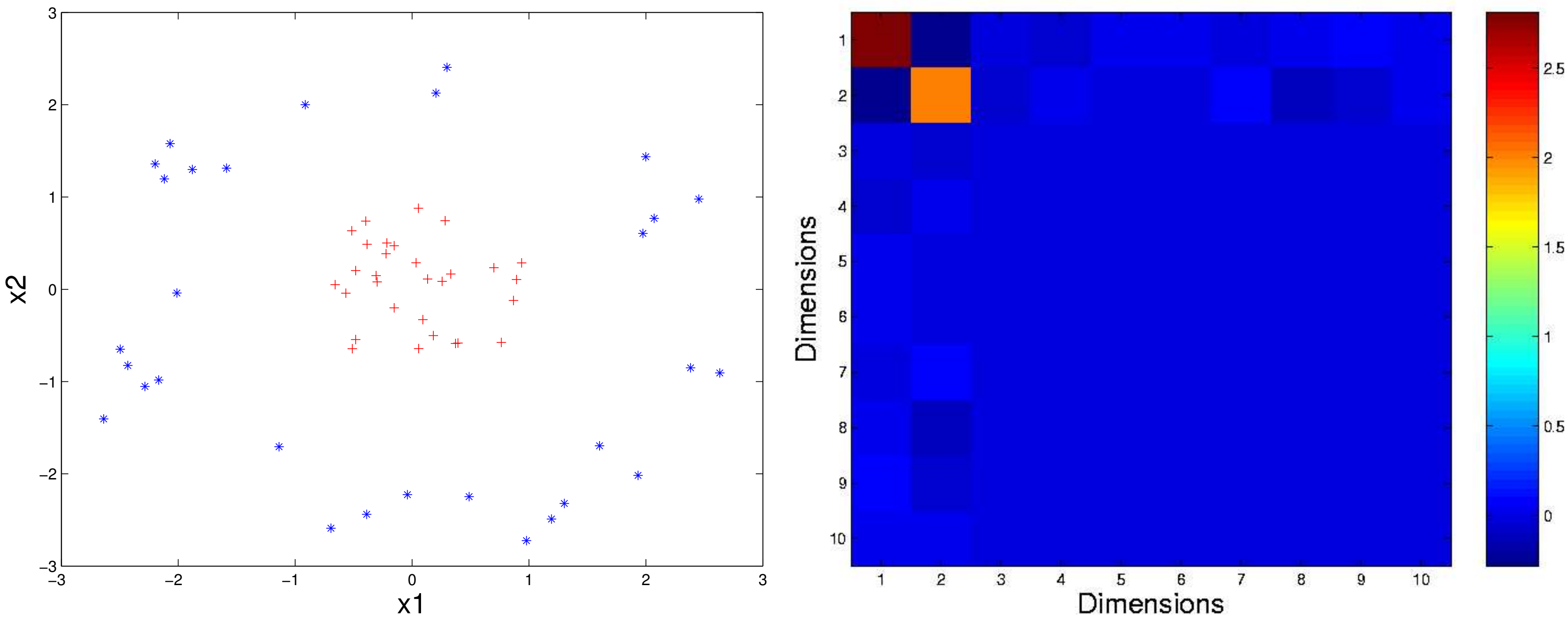
Samples from class $+1$ were drawn from

$$\begin{aligned} (x^1, x^2) &= (r\ \sin(\theta), r\ \cos(\theta)), \text{ where } r \sim U[0,1] \text{ and } \theta \sim U[0, 2\pi], \\ x^j &\sim \mathcal{N}(0.0, .2), \text{ for } j = 3, \ldots, 200, \end{aligned}$$

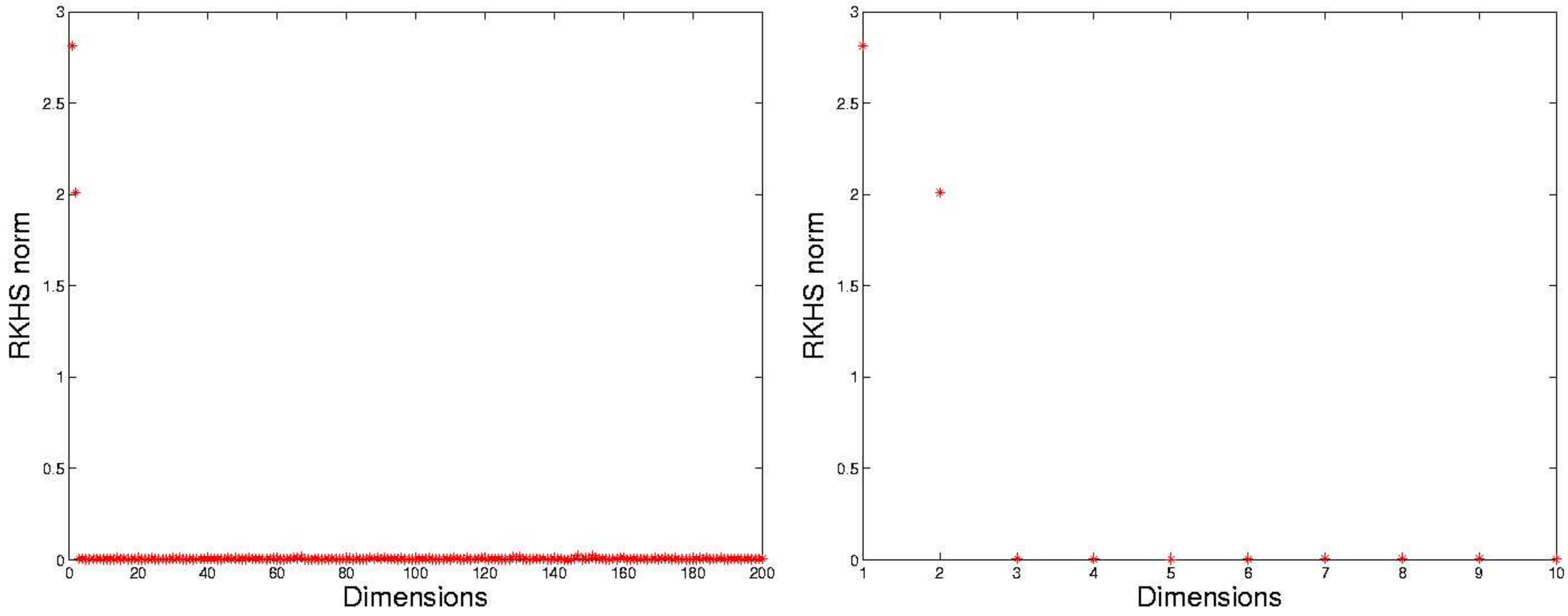
where $U[a, b]$ is the uniform distribution with support on the interval $[a, b]$. Samples from class -1 were drawn from

$$\begin{aligned} (x^1, x^2) &= (r\ \sin(\theta), r\ \cos(\theta)), \text{ where } r \sim U[2,3] \text{ and } \theta \sim U[0, 2\pi], \\ x^j &\sim \mathcal{N}(0.0, .2), \text{ for } j = 3, \ldots, 200. \end{aligned}$$

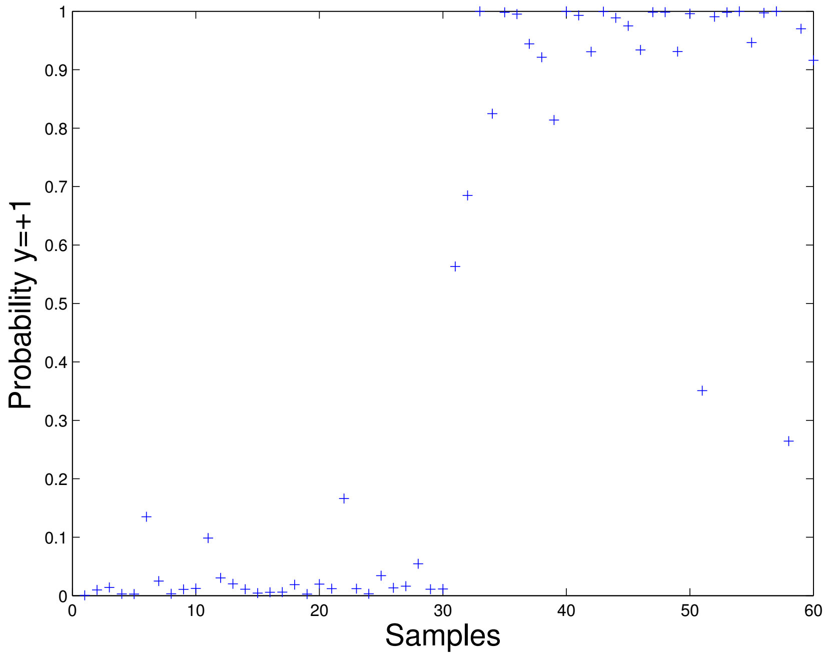
Nonlinear example

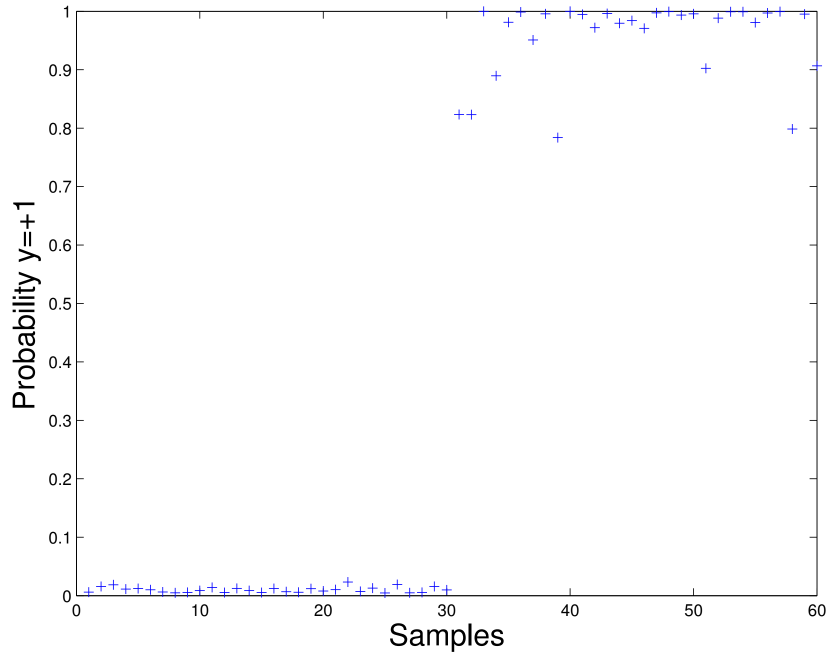


Nonlinear example



Nonlinear example





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)	50	100	200	300	400	500	1,000	3,000	7,129
test errors	2	1	1	1	1	1	1	1	2

Decay of norms

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

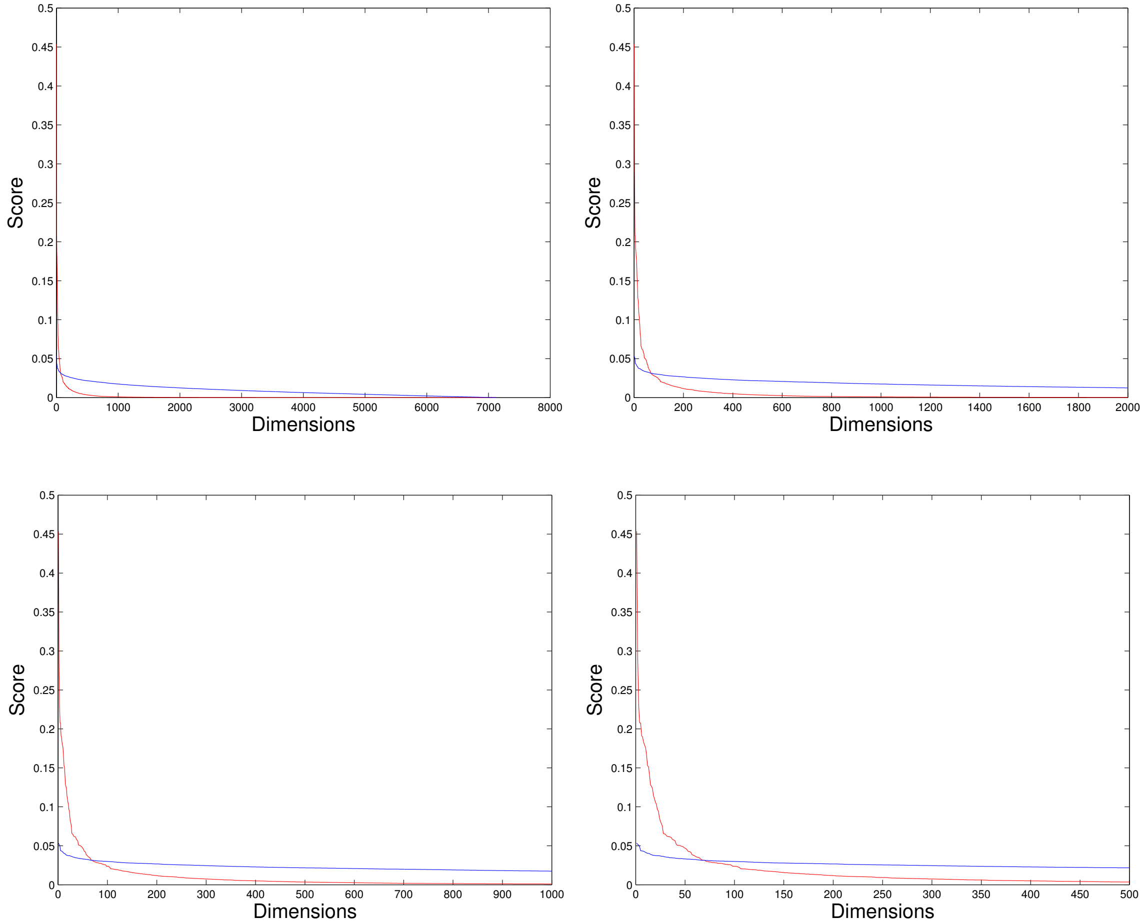
Decay of norms

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



Restriction to a manifold

Assume the data is concentrated on a manifold $\mathcal{M} \subset \mathbb{R}^p$ with $\mathcal{M} \in \mathbb{R}^d$.

Given a smooth orthonormal vector field $\{e_1, \ldots, e_d\}$ we can define the gradient on the manifold $\nabla_{\mathcal{M}} f = (e_1 f, \ldots, e_d f)$.

Restriction to a manifold

Assume the data is concentrated on a manifold $\mathcal{M} \subset \mathbb{R}^p$ with $\mathcal{M} \in \mathbb{R}^d$.

Given a smooth orthonormal vector field $\{e_1, \ldots, e_d\}$ we can define the gradient on the manifold $\nabla_{\mathcal{M}} f = (e_1 f, \ldots, e_d f)$.

For $p \in U \subset \mathcal{M}$ a chart $\mathbf{u} : U \to \mathbb{R}^d$ satisfying $\frac{\partial}{\partial u^i}(p) = e_i(p)$ exists.

The Taylor expansion on the manifold around p

$$f(q) \approx f(p) + \nabla_{\mathcal{M}} f(p) \cdot (\mathbf{u}(q) - \mathbf{u}(p)) \text{ for } q \approx p.$$

Restriction to a manifold

Assume the data is concentrated on a manifold $\mathcal{M} \subset \mathbb{R}^p$ with $\mathcal{M} \in \mathbb{R}^d$.

Given a smooth orthonormal vector field $\{e_1, \ldots, e_d\}$ we can define the gradient on the manifold $\nabla_{\mathcal{M}} f = (e_1 f, \ldots, e_d f)$.

For $p \in U \subset \mathcal{M}$ a chart $\mathbf{u} : U \to \mathbb{R}^d$ satisfying $\frac{\partial}{\partial u^i}(p) = e_i(p)$ exists.

The Taylor expansion on the manifold around p

$$f(q) \approx f(p) + \nabla_{\mathcal{M}} f(p) \cdot (\mathbf{u}(q) - \mathbf{u}(p)) \text{ for } q \approx p.$$

Neither $\mathcal{M}$ nor a local expression of $\mathcal{M}$ are given.

Restriction to a manifold

Assume the data is concentrated on a manifold $\mathcal{M} \subset \mathbb{R}^p$ with $\mathcal{M} \in \mathbb{R}^d$.

Given a smooth orthonormal vector field $\{e_1, \ldots, e_d\}$ we can define the gradient on the manifold $\nabla_{\mathcal{M}} f = (e_1 f, \ldots, e_d f)$.

The Taylor expansion on the manifold around p

$$f(q) \approx f(p) + \nabla_{\mathcal{M}} f(p) \cdot (\mathbf{u}(q) - \mathbf{u}(p)) \text{ for } q \approx p.$$

Assume an an embedding $\varphi : \mathcal{M} \to \mathbb{R}^p$.
$\{(p_i, y_i)\}_{i=1}^n \in \mathcal{M} \times \mathcal{Y}$ are drawn from the manifold
however we are not given a local expression of p_i but its image $x_i = \varphi(p_i) \in \mathbb{R}^p$.

Restriction to a manifold

Assume the data is concentrated on a manifold $\mathcal{M} \subset \mathbb{R}^p$ with $\mathcal{M} \in \mathbb{R}^d$.

Given a smooth orthonormal vector field $\{e_1, \ldots, e_d\}$ we can define the gradient on the manifold $\nabla_{\mathcal{M}} f = (e_1 f, \ldots, e_d f)$.

The Taylor expansion on the manifold around p

$$f(q) \approx f(p) + \nabla_{\mathcal{M}} f(p) \cdot (\mathbf{u}(q) - \mathbf{u}(p)) \text{ for } q \approx p.$$

The Taylor expansion on the manifold around x in terms of $f \circ \varphi^{-1} \in \mathbb{R}^p$

$$(f \circ \varphi^{-1})(u) - (f \circ \varphi^{-1})(x) \approx \nabla(f \circ \varphi^{-1})(x) \cdot (u - x) \text{ for } u \approx x.$$

Restriction to a manifold

Assume the data is concentrated on a manifold $\mathcal{M} \subset \mathbb{R}^p$ with $\mathcal{M} \in \mathbb{R}^d$.

Given a smooth orthonormal vector field $\{e_1, \ldots, e_d\}$ we can define the gradient on the manifold $\nabla_{\mathcal{M}} f = (e_1 f, \ldots, e_d f)$.

Due to this equivalence our gradient algorithm works in the manifold setting without any changes.

We can prove a rate of convergence of

$$\begin{aligned} \|\vec{f}_{\mathbf{z}} - \nabla f_\rho\|_{\rho_X} &\leq C \log\left(\frac{2}{\delta}\right) n^{-1/d_{\mathcal{M}}} \\ \|g_{\mathbf{z}} - f_\rho\|_{\rho_X} &\leq C \log\left(\frac{2}{\delta}\right) n^{-1/d_{\mathcal{M}}}. \end{aligned}$$

Dimensionality reduction

The **empirical covariance matrix** (ECM), $\Xi_{\mathbf{z}}$, is the $p \times p$ matrix of inner products of the gradient between two coordinates

$$\Xi_{\mathbf{z}} := \left[\langle (\vec{f}_{\mathbf{z}})_k , (\vec{f}_{\mathbf{z}})_l \rangle_K \right]_{k,l=1}^{p} .$$

Dimensionality reduction

The **empirical covariance matrix** (ECM), $\Xi_{\mathbf{z}}$, is the $p \times p$ matrix of inner products of the gradient between two coordinates

$$\Xi_{\mathbf{z}} := \left[\langle (\vec{f}_{\mathbf{z}})_k, (\vec{f}_{\mathbf{z}})_l \rangle_K \right]_{k,l=1}^{p}.$$

The covariance matix can be used in the same spirit as the covariance of the data (design) matrix is used in Principle Components Analysis (PCA) to select relevant features.

Dimensionality reduction

The **empirical covariance matrix** (ECM), $\Xi_{\mathbf{z}}$, is the $p \times p$ matrix of inner products of the gradient between two coordinates

$$\Xi_{\mathbf{z}} := \left[\langle \left(\vec{f}_{\mathbf{z}}\right)_k, \left(\vec{f}_{\mathbf{z}}\right)_l \rangle_K \right]_{k,l=1}^{p} .$$

Supervised non-linear dimensionality reduction in the spirit of LLE, ISOMAP, Laplacian Eigenmaps, Hessian Eigenmaps.

Dimensionality reduction

The **empirical covariance matrix** (ECM), $\Xi_{\mathbf{z}}$, is the $p \times p$ matrix of inner products of the gradient between two coordinates

$$\Xi_{\mathbf{z}} := \left[\langle (\vec{f_{\mathbf{z}}})_k, (\vec{f_{\mathbf{z}}})_l \rangle_K\right]_{k,l=1}^{p}.$$

Proposition 4. *Given f on $\mathbb{R}^p$ and assume its gradient exists. A vector $v \in \mathbb{R}^p$ is the k-th important feature if $\|v\| = 1$ and there exist $\{v_i\}_{i=1}^{k-1}$ with $\|v_i\| = 1$ such that*
(1) for all w satisfying $\|w\| = 1$ and $w \perp v_i$, there holds $\|w \cdot \nabla f\|_\infty \leq \|v_i \cdot \nabla f\|_\infty$,
(2) $v = \arg\max \|w \cdot \nabla f\|_\infty$ s.t. $\|w\| = 1$ and $w \perp v_i$,

Replace the L_∞ norm with the RKHS norm, the k-th most important feature is the eigenvector corresponding to the k-th eigenvalue of the covariance matrix Ξ.

Dimensionality reduction

The **empirical covariance matrix** (ECM), $\Xi_{\mathbf{z}}$, is the $p \times p$ matrix of inner products of the gradient between two coordinates

$$\Xi_{\mathbf{z}} := \left[\langle (\vec{f_{\mathbf{z}}})_k, (\vec{f_{\mathbf{z}}})_l \rangle_K \right]_{k,l=1}^{p}.$$

Proposition 5. *Given f on $\mathbb{R}^p$ and assume its gradient exists. A vector $v \in \mathbb{R}^p$ is the k-th important feature if $\|v\| = 1$ and there exist $\{v_i\}_{i=1}^{k-1}$ with $\|v_i\| = 1$ such that*
(1) for all w satisfying $\|w\| = 1$ and $w \perp v_i$, there holds $\|w \cdot \nabla f\|_\infty \leq \|v_i \cdot \nabla f\|_\infty$,
(2) $v = \arg\max \|w \cdot \nabla f\|_\infty$ s.t. $\|w\| = 1$ and $w \perp v_i$,

Replace the L_∞ norm with the RKHS norm, the k-th most important feature is the eigenvector corresponding to the k-th eigenvalue of the covariance matrix Ξ.

This proposition suggests that we project our data matrix onto the top k-eigenvectors. This space should reflect the geometery of the classification or regression function on the manifold.

Dimensionality reduction

The **empirical covariance matrix** (ECM), $\Xi_{\mathbf{z}}$, is the $p \times p$ matrix of inner products of the gradient between two coordinates

$$\Xi_{\mathbf{z}} := \left[\langle (\vec{f}_{\mathbf{z}})_k , (\vec{f}_{\mathbf{z}})_l \rangle_K \right]_{k,l=1}^{p} .$$

Proposition 6. *Given f on $\mathbb{R}^p$ and assume its gradient exists. A vector $v \in \mathbb{R}^p$ is the k-th important feature if $\|v\| = 1$ and there exist $\{v_i\}_{i=1}^{k-1}$ with $\|v_i\| = 1$ such that*
(1) for all w satisfying $\|w\| = 1$ and $w \perp v_i$, there holds $\|w \cdot \nabla f\|_\infty \leq \|v_i \cdot \nabla f\|_\infty$,
(2) $v = \arg\max \|w \cdot \nabla f\|_\infty$ s.t. $\|w\| = 1$ and $w \perp v_i$,

Replace the L_∞ norm with the RKHS norm, the k-th most important feature is the eigenvector corresponding to the k-th eigenvalue of the covariance matrix Ξ.

Since

$$\Xi_{\mathbf{z}} = c_{\mathbf{z}}^T K c_{\mathbf{z}}$$

the n nonzero eigenvalues and corresponding eigenvectors of can be computed without constructing the $p \times p$ matrix, in order $O(n^2 p + n^3)$ time and $O(p \times n)$ memory.

Discussion

Still lots of work left:

- Fully Bayesian model: compute the full posterior using MCMC.

Discussion

Still lots of work left:

- Fully Bayesian model: compute the full posterior using MCMC.
- Semi-supervised version: implement a semi-supervised version.

Discussion

Still lots of work left:

- Fully Bayesian model: compute the full posterior using MCMC.
- Semi-supervised version: implement a semi-supervised version.
- Relation to information geometry: Tthe covariance matrix is a particular case of the non-parametric analog of Fisher's information matrix.