Geometric perspectives for supervised dimension reduction

A Tale of Two Manifolds

S. Mukherjee, K. Mao, F. Liang, Q. Wu, D-X. Zhou, J. Guinney

Department of Statistical Science
Institute for Genome Sciences & Policy
Department of Computer Science
Department of Mathematics
Duke University

December 11, 2009

Information and sufficiency

A fundamental idea in statistical thought is to reduce data to relevant information. This was the paradigm of R.A. Fisher (beloved Bayesian) and goes back to at least Adcock 1878 and Edgeworth 1884.

Information and sufficiency

A fundamental idea in statistical thought is to reduce data to relevant information. This was the paradigm of R.A. Fisher (beloved Bayesian) and goes back to at least Adcock 1878 and Edgeworth 1884.

$X_1, ..., X_n$ drawn iid form a Gaussian can be reduced to μ, σ^2.

Regression

Assume the model

$$Y = f(X) + \varepsilon, \quad \mathbb{E}\varepsilon = 0,$$

with $X \in \mathcal{X} \subset \mathbb{R}^p$ and $Y \in \mathbb{R}$.

Regression

Assume the model

$$Y = f(X) + \varepsilon, \quad \mathbb{E}\varepsilon = 0,$$

with $X \in \mathcal{X} \subset \mathbb{R}^p$ and $Y \in \mathbb{R}$.

Data – $D = \{(x_i, y_i)\}_{i=1}^n \overset{iid}{\sim} \rho(X, Y)$.

Dimension reduction

If the data lives in a p-dimensional space $X \in \mathbb{R}^p$ replace X with $\Theta(X) \in \mathbb{R}^d$, $p \gg d$.

Dimension reduction

If the data lives in a p-dimensional space $X \in \mathbb{R}^p$ replace X with $\Theta(X) \in \mathbb{R}^d$, $p \gg d$.

My belief: physical, biological and social systems are inherently low dimensional and variation of interest in these systems can be captured by a low-dimensional submanifold.

Supervised dimension reduction (SDR)

Given response variables $Y_1, ..., Y_n \in \mathbb{R}$ and explanatory variables or covariates $X_1, ..., X_n \in \mathcal{X} \subset \mathbb{R}^p$

$$Y_i = f(X_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \text{No}(0, \sigma^2).$$

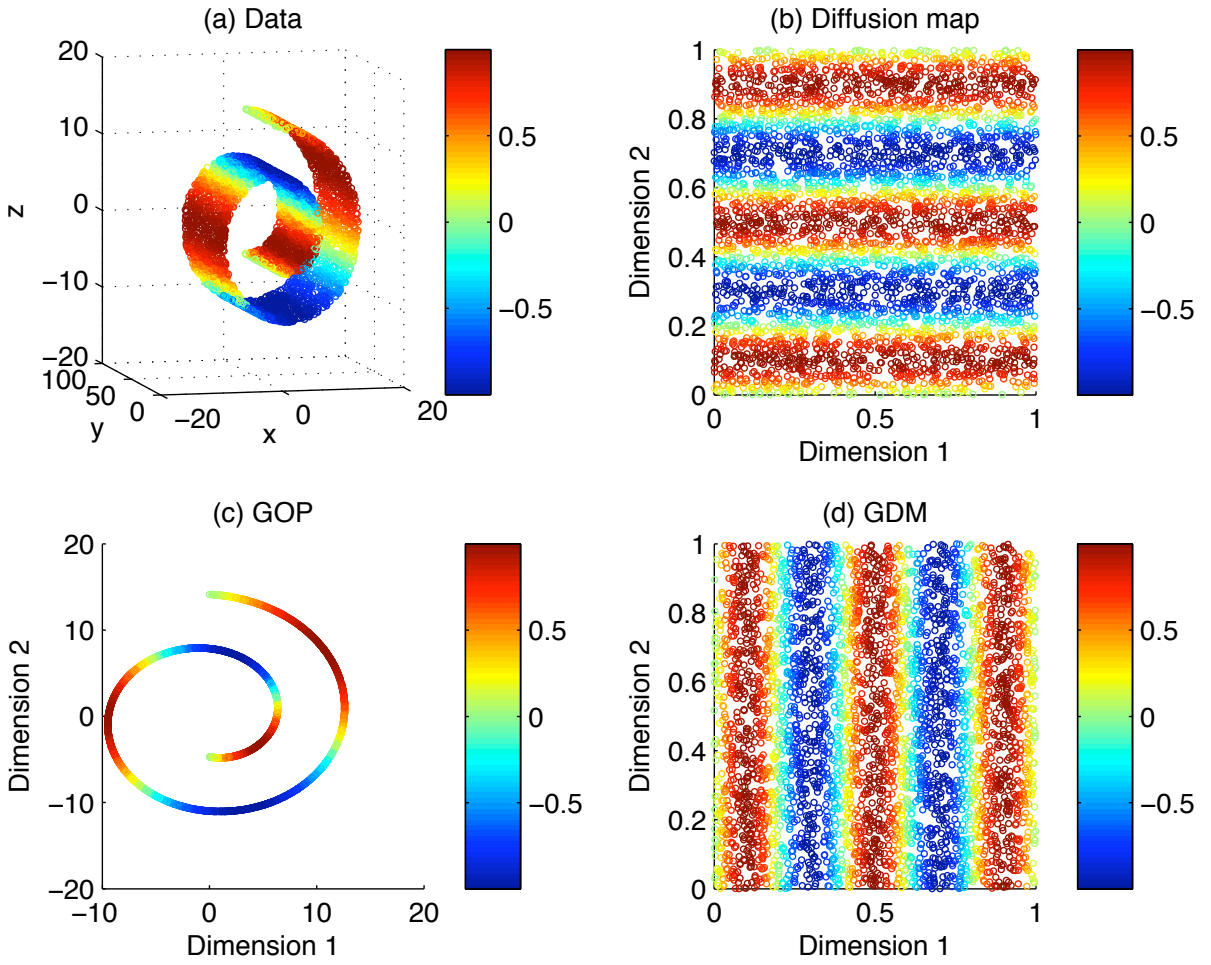
Supervised dimension reduction (SDR)

Given response variables $Y_1, ..., Y_n \in \mathbb{R}$ and explanatory variables or covariates $X_1, ..., X_n \in \mathcal{X} \subset \mathbb{R}^p$

$$Y_i = f(X_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \text{No}(0, \sigma^2).$$

Is there a submanifold $\mathcal{S} \equiv \mathcal{S}_{Y|X}$ such that $Y \perp\!\!\!\perp X \mid P_{\mathcal{S}}(X)$?

Visualization of SDR



Linear projections capture nonlinear manifolds

In this talk $P_{\mathcal{S}}(X) = B^T X$ where $B = (b_1, ..., b_d)$.

Linear projections capture nonlinear manifolds

In this talk $P_{\mathcal{S}}(X) = B^T X$ where $B = (b_1, ..., b_d)$.

Semiparametric model

$$Y_i = f(X_i) + \varepsilon_i = g(b_1^T X_i, \ldots, b_d^T X_i) + \varepsilon_i,$$

span B is the dimension reduction (d.r.) subspace.

SDR model

Semiparametric model

$$Y_i = f(X_i) + \varepsilon_i = g(b_1^T X_i, \ldots, b_d^T X_i) + \varepsilon_i,$$

span B is the dimension reduction (d.r.) subspace.

SDR model

Semiparametric model

$$Y_i = f(X_i) + \varepsilon_i = g(b_1^T X_i, \ldots, b_d^T X_i) + \varepsilon_i,$$

span B is the dimension reduction (d.r.) subspace.

Assume marginal distribution ρ_X is concentrated on a manifold $\mathcal{M} \subset \mathbb{R}^p$ of dimension $d \ll p$.

Gradients and outer products

Given a smooth function f the gradient is

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_p} \right)^T.$$

Gradients and outer products

Given a smooth function f the gradient is
$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_p} \right)^T$.

Define the gradient outer product matrix Γ

$$\begin{aligned} \Gamma_{ij} &= \int_{\mathcal{X}} \frac{\partial f}{\partial x_i}(x) \frac{\partial f}{\partial x_j}(x) d\rho_X(x), \\ \Gamma &= \mathbb{E}[(\nabla f) \otimes (\nabla f)]. \end{aligned}$$

GOP captures the d.r. space

Suppose

$$y = f(X) + \varepsilon = g(b_1^T X, ..., b_d^T X) + \varepsilon.$$

GOP captures the d.r. space

Suppose

$$y = f(X) + \varepsilon = g(b_1^T X, ..., b_d^T X) + \varepsilon.$$

Note that for $B = (b_1, ..., b_d)$

$$\lambda_i b_i = \Gamma b_i.$$

GOP captures the d.r. space

Suppose

$$y = f(X) + \varepsilon = g(b_1^T X, ..., b_d^T X) + \varepsilon.$$

Note that for $B = (b_1, ..., b_d)$

$$\lambda_i b_i = \Gamma b_i.$$

For $i = 1, .., d$

$$\frac{\partial f(x)}{\partial v_i} = v_i^T \left(\nabla f(x)\right) \neq 0 \Rightarrow b_i^T \Gamma b_i \neq 0.$$

If $w \perp b_i$ for all i then $w^T \Gamma w = 0$.

Statistical interpretation

Linear case

$$y = \beta^T x + \varepsilon, \qquad \varepsilon \overset{iid}{\sim} \mathrm{No}(0, \sigma^2).$$

$\Omega = \mathrm{cov}\left(\mathbb{E}[X|Y]\right)$, $\Sigma_X = \mathrm{cov}\left(X\right)$, $\sigma_Y^2 = \mathrm{var}\left(Y\right)$.

Statistical interpretation

Linear case

$$y = \beta^T x + \varepsilon, \qquad \varepsilon \overset{iid}{\sim} \mathrm{No}(0, \sigma^2).$$

$\Omega = \mathrm{cov}\left(\mathbb{E}[X|Y]\right)$, $\Sigma_X = \mathrm{cov}\left(X\right)$, $\sigma_Y^2 = \mathrm{var}\left(Y\right)$.

$$\Gamma = \sigma_Y^2 \left(1 - \frac{\sigma^2}{\sigma_Y^2}\right)^2 \Sigma_X^{-1} \Omega \Sigma_X^{-1} \approx \sigma_Y^2 \, \Sigma_X^{-1} \Omega \Sigma_X^{-1}.$$

Statistical interpretation

For smooth $f(x)$

$$y = f(x) + \varepsilon, \qquad \varepsilon \overset{iid}{\sim} \mathrm{No}(0, \sigma^2).$$

$\Omega = \mathrm{cov}\left(\mathbb{E}[X|Y]\right)$ not so clear.

Nonlinear case

Partition into sections and compute local quantities

$$\mathcal{X} = \bigcup_{i=1}^{\mathcal{I}} \chi_i$$

Nonlinear case

Partition into sections and compute local quantities

$$\begin{aligned} \mathcal{X} &= \bigcup_{i=1}^{\mathcal{I}} \chi_i \\ \Omega_i &= \operatorname{cov}\left(\mathbb{E}[X_{\chi_i} | Y_{\chi_i}]\right) \end{aligned}$$

Nonlinear case

Partition into sections and compute local quantities

$$\begin{aligned} \mathcal{X} &= \bigcup_{i=1}^{\mathcal{I}} \chi_i \\ \Omega_i &= \operatorname{cov}\left(\mathbb{E}[X_{\chi_i} | Y_{\chi_i}]\right) \\ \Sigma_i &= \operatorname{cov}\left(X_{\chi_i}\right) \end{aligned}$$

Nonlinear case

Partition into sections and compute local quantities

$$
\begin{aligned}
\mathcal{X} &= \bigcup_{i=1}^{\mathcal{I}} \chi_i \\
\Omega_i &= \operatorname{cov}\left(\mathbb{E}[X_{\chi_i} | Y_{\chi_i}]\right) \\
\Sigma_i &= \operatorname{cov}\left(X_{\chi_i}\right) \\
\sigma_i^2 &= \operatorname{var}\left(Y_{\chi_i}\right)
\end{aligned}
$$

Nonlinear case

Partition into sections and compute local quantities

$$
\begin{aligned}
\mathcal{X} &= \bigcup_{i=1}^{\mathcal{I}} \chi_i \\
\Omega_i &= \operatorname{cov}\left(\mathbb{E}[X_{\chi_i} | Y_{\chi_i}]\right) \\
\Sigma_i &= \operatorname{cov}\left(X_{\chi_i}\right) \\
\sigma_i^2 &= \operatorname{var}\left(Y_{\chi_i}\right) \\
m_i &= \rho_X(\chi_i).
\end{aligned}
$$

Nonlinear case

Partition into sections and compute local quantities

$$\begin{aligned}
\mathcal{X} &= \bigcup_{i=1}^{\mathcal{I}} \chi_i \\
\Omega_i &= \operatorname{cov}\left(\mathbb{E}[X_{\chi_i} | Y_{\chi_i}]\right) \\
\Sigma_i &= \operatorname{cov}\left(X_{\chi_i}\right) \\
\sigma_i^2 &= \operatorname{var}\left(Y_{\chi_i}\right) \\
m_i &= \rho_X(\chi_i).
\end{aligned}$$

$$\Gamma \approx \sum_{i=1}^{\mathcal{I}} m_i \, \sigma_i^2 \, \Sigma_i^{-1} \, \Omega_i \, \Sigma_i^{-1}.$$

Estimating the gradient

Taylor expansion

$$\begin{aligned} y_i \approx f(x_i) &\approx f(x_j) + \langle \nabla f(x_j), x_j - x_i \rangle \\ &\approx y_j + \langle \nabla f(x_j), x_j - x_i \rangle \quad \text{if } x_i \approx x_j. \end{aligned}$$

Estimating the gradient

Taylor expansion

$$\begin{aligned} y_i \approx f(x_i) &\approx f(x_j) + \langle \nabla f(x_j), x_j - x_i \rangle \\ &\approx y_j + \langle \nabla f(x_j), x_j - x_i \rangle \quad \text{if } x_i \approx x_j. \end{aligned}$$

Let $\vec{f} \approx \nabla f$ the following should be small

$$\sum_{i,j} w_{ij} (y_i - y_j - \langle \vec{f}(x_j), x_j - x_i \rangle)^2,$$

$w_{ij} = \frac{1}{s^{p+2}} \exp(-\|x_i - x_j\|^2 / 2s^2)$ enforces $x_i \approx x_j$.

Estimating the gradient

The gradient estimate

$$\vec{f}_D = \arg\min_{\vec{f} \in \mathcal{H}^p} \left[\frac{1}{n^2} \sum_{i,j=1}^{n} w_{ij} \left(y_i - y_j - (\vec{f}(x_j))^T (x_j - x_i) \right)^2 + \lambda \|\vec{f}\|_K^2 \right]$$

where $\|\vec{f}\|_K$ is a smoothness penalty, reproducing kernel Hilbert space norm.

Estimating the gradient

The gradient estimate

$$\vec{f}_D = \arg\min_{\vec{f} \in \mathcal{H}^p} \left[\frac{1}{n^2} \sum_{i,j=1}^{n} w_{ij} \left(y_i - y_j - (\vec{f}(x_j))^T (x_j - x_i) \right)^2 + \lambda \|\vec{f}\|_K^2 \right]$$

where $\|\vec{f}\|_K$ is a smoothness penalty, reproducing kernel Hilbert space norm.
Goto board.

Computational efficiency

The computation requires fewer than n^2 parameters and is $O(n^6)$ time and $O(pn)$ memory

$$\vec{f}_D(x) = \sum_{i=1}^{n} c_{i,D} K(x_i, x)$$

$c_D = (c_{1,D}, \ldots, c_{n,D})^T \in \mathbb{R}^{np}$.

Computational efficiency

The computation requires fewer than n^2 parameters and is $O(n^6)$ time and $O(pn)$ memory

$$\vec{f}_D(x) = \sum_{i=1}^{n} c_{i,D} K(x_i, x)$$

$c_D = (c_{1,D}, \ldots, c_{n,D})^T \in \mathbb{R}^{np}$.

Define gram matrix K where $K_{ij} = K(x_i, x_j)$

$$\hat{\Gamma} = c_D K c_D^T.$$

Estimates on manifolds

Mrginal distribution ρ_X is concentrated on a compact Riemannian manifold $\mathcal{M} \in \mathbb{R}^d$ with isometric embedding $\varphi : \mathcal{M} \to \mathbb{R}^p$ and metric $d_{\mathcal{M}}$ and $\mathrm{d}\mu$ is the uniform measure on $\mathcal{M}$.
Assume regular distribution

(i) The density $\nu(x) = \frac{\mathrm{d}\rho_X(x)}{\mathrm{d}\mu}$ exists and is Hölder continuous ($c_1 > 0$ and $0 < \theta \leq 1$)

$$|\nu(x) - \nu(u)| \leq c_1 \, d^{\theta}_{\mathcal{M}}(x, u) \quad \forall x, u \in \mathcal{M}.$$

(ii) The measure along the boundary is small: ($c_2 > 0$)

$$\rho_{\mathcal{M}}\left(\{x \in \mathcal{M} : d_{\mathcal{M}}(x, \partial\mathcal{M}) \leq t\} \right) \leq c_2 \, t \qquad \forall t > 0.$$

Convergence to gradient on manifold

Theorem

Under above regularity conditions on ρ_X *and* $f \in C^2(\mathcal{M})$, *with probability* $1 - \delta$

$$\|(\mathrm{d}\varphi)^* \vec{f}_D - \nabla_{\mathcal{M}} f\|^2_{L^2_{\rho_{\mathcal{M}}}} \leq C \log\left(\frac{1}{\delta}\right)\left(n^{-\frac{1}{d}}\right).$$

where $(d\varphi)^*$ *(projection onto tangent space) is the dual of the map* $d\varphi$.

Multi-task learning

Definition

Single Task Notation n_t *samples* (x_i, y_i)
$x_i \in \mathbb{R}^d$
$y_i \in \{-1, 1\}$ *for classification*
Assume to be working in $d \gg n_t$ *paradigm.*

Multi-task learning

Definition

Single Task Notation n_t *samples* (x_i, y_i)
$x_i \in \mathbb{R}^d$
$y_i \in \{-1, 1\}$ *for classification*
Assume to be working in $d \gg n_t$ *paradigm.*

Definition

Multi-task Learning (MTL) Formulation Given T *tasks with* $t \in \{1, \ldots, T\}$

$$F_t(x) = f_0(x) + f_t(x) + \varepsilon, \ \varepsilon \overset{iid}{\backsim} No(0, \sigma^2).$$

Multi-task gradient learning

Estimate not just the functions

$$\{f_0, f_1, ..., f_T\},$$

Multi-task gradient learning

Estimate not just the functions

$$\{f_0, f_1, ..., f_T\},$$

but the gradients as well

$$\{(f_0, \nabla f_0), (f_t, \nabla f_t)_{t=1}^T\}.$$

Multi-task gradient learning

Estimate not just the functions

$$\{f_0, f_1, ..., f_T\},$$

but the gradients as well

$$\{(f_0, \nabla f_0), (f_t, \nabla f_t)_{t=1}^T\}.$$

This provides us with $T + 1$ matrices

1. $\hat{\Gamma}^0$ is the GOP estimate across all the tasks

Multi-task gradient learning

Estimate not just the functions

$$\{f_0, f_1, ..., f_T\},$$

but the gradients as well

$$\{(f_0, \nabla f_0), (f_t, \nabla f_t)_{t=1}^T\}.$$

This provides us with $T + 1$ matrices

1. $\hat{\Gamma}^0$ is the GOP estimate across all the tasks
2. $\hat{\Gamma}^1, \ldots, \hat{\Gamma}^T$ are the task specific GOP estimates.

Principal components analysis (PCA)

Algorithmic view of PCA:

1. Given $X = (X_1, ..., X_n)$ a $p \times n$ matrix construct

$$\hat{\Sigma} = (X - \bar{X})(X - \bar{X})^T$$

Principal components analysis (PCA)

Algorithmic view of PCA:

1. Given $X = (X_1, ..., X_n)$ a $p \times n$ matrix construct

$$\hat{\Sigma} = (X - \bar{X})(X - \bar{X})^T$$

2. Eigen-decomposition of $\hat{\Sigma}$

$$\lambda_i v_i = \hat{\Sigma} v_i.$$

Probabilistic PCA

$X \in \mathbb{R}^p$ is charterized by a multivariate normal

$$X \sim \mathrm{No}(\mu + A\nu, \Delta),$$
$$\nu \sim \mathrm{No}(0, \mathbf{I}_d)$$

$\mu \in \mathbb{R}^p$
$A \in \mathbb{R}^{p \times d}$
$\Delta \in \mathbb{R}^{p \times p}$
$\nu \in \mathbb{R}^d$.

Probabilistic PCA

$X \in \mathbb{R}^p$ is charterized by a multivariate normal

$$X \sim \mathrm{No}(\mu + A\nu, \Delta),$$
$$\nu \sim \mathrm{No}(0, \mathbf{I}_d)$$

$\mu \in \mathbb{R}^p$
$A \in \mathbb{R}^{p \times d}$
$\Delta \in \mathbb{R}^{p \times p}$
$\nu \in \mathbb{R}^d$.

ν is a latent variable

SDR model

Semiparametric model

$$Y_i = f(X_i) + \varepsilon_i = g(b_1^T X_i, \ldots, b_d^T X_i) + \varepsilon_i,$$

span B is the dimension reduction (d.r.) subspace.

Principal fitted components (PFC)

Define $X_y \equiv (X \mid Y = y)$ and specify multivariate normal distribution

$$X_y \sim \mathrm{No}(\mu_y, \Delta),$$
$$\mu_y = \mu + A\nu_y$$

$\mu \in \mathbb{R}^p$
$A \in \mathbb{R}^{p \times d}$
$\nu_y \in \mathbb{R}^d$.

Principal fitted components (PFC)

Define $X_y \equiv (X \mid Y = y)$ and specify multivariate normal distribution

$$X_y \sim \mathrm{No}(\mu_y, \Delta),$$
$$\mu_y = \mu + A\nu_y$$

$\mu \in \mathbb{R}^p$
$A \in \mathbb{R}^{p \times d}$
$\nu_y \in \mathbb{R}^d$.

$B = \Delta^{-1} A$.

Principal fitted components (PFC)

Define $X_y \equiv (X \mid Y = y)$ and specify multivariate normal distribution

$$X_y \sim \mathrm{No}(\mu_y, \Delta),$$
$$\mu_y = \mu + A\nu_y$$

$\mu \in \mathbb{R}^p$
$A \in \mathbb{R}^{p \times d}$
$\nu_y \in \mathbb{R}^d$.

$B = \Delta^{-1} A$.

Captures global linear predictive structure. Does not generalize to manifolds.

Mixture models and localization

A driving idea in manifold learning is that manifolds are locally Euclidean.

Mixture models and localization

A driving idea in manifold learning is that manifolds are locally Euclidean.

A driving idea in probabilistic modeling is that mixture models are flexible and can capture "nonparametric" distributions.

Mixture models and localization

A driving idea in manifold learning is that manifolds are locally Euclidean.

A driving idea in probabilistic modeling is that mixture models are flexible and can capture "nonparametric" distributions.

Mixture models can capture local nonlinear predictive manifold structure.

Model specification

$$X_y \sim \mathrm{No}(\mu_{yx}, \Delta)$$
$$\mu_{yx} = \mu + A\nu_{yx}$$
$$\nu_{yx} \sim G_y$$

G_y: density indexed by y having multiple clusters
$\mu \in \mathbb{R}^p$
$\varepsilon \sim N(0, \Delta)$ with $\Delta \in \mathbb{R}^{p \times p}$
$A \in \mathbb{R}^{p \times d}$
$\nu_{xy} \in \mathbb{R}^d$.

Dimension reduction space

Proposition

For this model the d.r. space is the span of $B = \Delta^{-1}A$

$$Y \mid X \stackrel{d}{=} Y \mid (\Delta^{-1}A)^T X.$$

Sampling distribution

Define $\nu_i \equiv \nu_{y_i x_i}$. Sampling distribution for data

$$x_i \mid (y_i, \mu, \nu_i, A, \Delta) \sim N(\mu + A\nu_i, \Delta)$$
$$\nu_i \sim G_{y_i}.$$

Categorical response: modeling G_y

$Y = \{1, ..., C\}$, so each category has a distribution

$$\nu_i \mid (y_i = k) \sim G_k, \quad c = 1, ..., C.$$

Categorical response: modeling G_y

$Y = \{1, ..., C\}$, so each category has a distribution

$$\nu_i \mid (y_i = k) \sim G_k, \quad c = 1, ..., C.$$

ν_i modeled as a mixture of C distributions $G_1, ..., G_C$ with a Dirichlet process model for ech distribution

$$G_c \sim \mathrm{DP}(\alpha_0, G_0).$$

Categorical response: modeling G_y

$Y = \{1, ..., C\}$, so each category has a distribution

$$\nu_i \mid (y_i = k) \sim G_k, \quad c = 1, ..., C.$$

ν_i modeled as a mixture of C distributions $G_1, ..., G_C$ with a Dirichlet process model for ech distribution

$$G_c \sim \mathrm{DP}(\alpha_0, G_0).$$

Goto board.

Likelihood

$$\text{Lik}(\text{data} \mid \theta) \equiv \text{Lik}(\text{data} \mid A, \Delta, \nu_1, ..., \nu_n, \mu)$$

$$\text{Lik}(\text{data} \mid \theta) \quad \propto \quad \det(\Delta^{-1})^{\frac{n}{2}} \times \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu - A\nu_i)^T \Delta^{-1}(x_i - \mu - A\nu_i)\right].$$

Posterior inference

Given data

$$\mathcal{P}_\theta \equiv \mathsf{Post}(\theta \mid \mathsf{data}) \propto \mathsf{Lik}(\theta \mid \mathsf{data}) \times \pi(\theta).$$

Posterior inference

Given data

$$\mathcal{P}_\theta \equiv \text{Post}(\theta \mid \text{data}) \propto \text{Lik}(\theta \mid \text{data}) \times \pi(\theta).$$

1. $\mathcal{P}_\theta$ provides estimate of (un)certainty on θ

Posterior inference

Given data

$$\mathcal{P}_\theta \equiv \text{Post}(\theta \mid \text{data}) \propto \text{Lik}(\theta \mid \text{data}) \times \pi(\theta).$$

1. $\mathcal{P}_\theta$ provides estimate of (un)certainty on θ
2. Requires prior on θ

Posterior inference

Given data

$$\mathcal{P}_\theta \equiv \text{Post}(\theta \mid \text{data}) \propto \text{Lik}(\theta \mid \text{data}) \times \pi(\theta).$$

1. $\mathcal{P}_\theta$ provides estimate of (un)certainty on θ
2. Requires prior on θ
3. Sample from $\mathcal{P}_\theta$?

Markov chain Monte Carlo

No closed form for $\mathcal{P}_\theta$.

Markov chain Monte Carlo

No closed form for $\mathcal{P}_\theta$.

1. Specify Markov transition kernel

$$K(\theta_t, \theta_{t+1})$$

with stationary distribution $\mathcal{P}_\theta$.

Markov chain Monte Carlo

No closed form for $\mathcal{P}_\theta$.

1. Specify Markov transition kernel

$$K(\theta_t, \theta_{t+1})$$

   with stationary distribution $\mathcal{P}_\theta$.

2. Run the Markov chain to obtain $\theta_1, ..., \theta_T$.

Sampling from the posterior

Inference consists of drawing samples $\theta_{(t)} = (\mu_{(t)}, A_{(t)}, \Delta_{(t)}^{-1}, \nu_{(t)})$ from the posterior.

Sampling from the posterior

Inference consists of drawing samples $\theta_{(t)} = (\mu_{(t)}, A_{(t)}, \Delta^{-1}_{(t)}, \nu_{(t)})$ from the posterior.

Define

$$\begin{aligned}
\theta^{/\mu}_{(t)} &\equiv (A_{(t)}, \Delta^{-1}_{(t)}, \nu_{(t)}) \\
\theta^{/A}_{(t)} &\equiv (\mu_{(t)}, \Delta^{-1}_{(t)}, \nu_{(t)}) \\
\theta^{/\Delta^{-1}}_{(t)} &\equiv (\mu_{(t)}, A_{(t)}, \nu_{(t)}) \\
\theta^{/\nu}_{(t)} &\equiv (\mu_{(t)}, A_{(t)}, \Delta^{-1}_{(t)}).
\end{aligned}$$

Gibbs sampling

Conditional probabilities can be used to sample μ, Δ^{-1}, A

$$\mu_{(t+1)} \mid \left(\text{data}, \theta^{/\mu}_{(t)}\right) \quad \sim \quad \text{No}\left(\text{data}, \theta^{/\mu}_{(t)}\right),$$

Gibbs sampling

Conditional probabilities can be used to sample μ, Δ^{-1}, A

$$\begin{aligned} \mu_{(t+1)} \mid \left(\text{data}, \theta^{/\mu}_{(t)}\right) &\sim \text{No}\left(\text{data}, \theta^{/\mu}_{(t)}\right), \\ {\Delta^{-1}}_{(t+1)} \mid \left(\text{data}, \theta^{/\Delta^{-1}}_{(t)}\right) &\sim \text{InvWishart}\left(\text{data}, \theta^{/\Delta^{-1}}_{(t)}\right) \end{aligned}$$

Gibbs sampling

Conditional probabilities can be used to sample μ, Δ^{-1}, A

$$\begin{aligned}
\mu_{(t+1)} \mid \left(\text{data}, \theta^{/\mu}_{(t)}\right) &\sim \text{No}\left(\text{data}, \theta^{/\mu}_{(t)}\right), \\
{\Delta^{-1}}_{(t+1)} \mid \left(\text{data}, \theta^{/\Delta^{-1}}_{(t)}\right) &\sim \text{InvWishart}\left(\text{data}, \theta^{/\Delta^{-1}}_{(t)}\right) \\
A_{(t+1)} \mid \left(\text{data}, \theta^{/A}_{(t)}\right) &\sim \text{No}\left(\text{data}.\theta^{/A}_{(t)}\right).
\end{aligned}$$

Gibbs sampling

Conditional probabilities can be used to sample μ, Δ^{-1}, A

$$\begin{aligned}
\mu_{(t+1)} \mid \left(\text{data}, \theta^{/\mu}_{(t)}\right) &\sim \text{No}\left(\text{data}, \theta^{/\mu}_{(t)}\right), \\
{\Delta^{-1}}_{(t+1)} \mid \left(\text{data}, \theta^{/\Delta^{-1}}_{(t)}\right) &\sim \text{InvWishart}\left(\text{data}, \theta^{/\Delta^{-1}}_{(t)}\right) \\
A_{(t+1)} \mid \left(\text{data}, \theta^{/A}_{(t)}\right) &\sim \text{No}\left(\text{data}.\theta^{/A}_{(t)}\right).
\end{aligned}$$

Sampling $\nu_{(t)}$ is more involved.

Posterior draws from the Grassmann manifold

Given samples $(\Delta_{(t)}^{-1}, A_{(t)})_{t=1}^{m}$ compute $\mathcal{B}_{(t)} = \Delta_{(t)}^{-1} A_{(t)}$.

Posterior draws from the Grassmann manifold

Given samples $(\Delta_{(t)}^{-1}, A_{(t)})_{t=1}^{m}$ compute $\mathcal{B}_{(t)} = \Delta_{(t)}^{-1} A_{(t)}$.

Each $\mathcal{B}_{(t)}$ is a subspace which is a point in the Grassmann manifold $\mathcal{G}_{(d,p)}$. There is a Riemannian metric on this manifold. This has two implications.

Posterior mean and variance

Given draws $(\mathcal{B}_{(t)})_{t=1}^{m}$ the posterior mean and variance should be computed with respect to the Riemannian metric.

Posterior mean and variance

Given draws $(\mathcal{B}_{(t)})_{t=1}^{m}$ the posterior mean and variance should be computed with respect to the Riemannian metric.

Given two subspaces $\mathcal{W}$ and $\mathcal{U}$ spanned by orthonormal bases W and V the Karcher mean is

$$\begin{aligned} (I - X(X^T X)^{-1} X^T) Y (X^T Y)^{-1} &= U \Sigma V^T \\ \Theta &= \operatorname{atan}(\Sigma) \\ \operatorname{dist}(\mathcal{W}, \mathcal{V}) &= \sqrt{\operatorname{Tr}(\Theta^2)}. \end{aligned}$$

Posterior mean and variance

The posterior mean subspace

$$\mathcal{B}_{\text{Bayes}} = \arg \min_{\mathcal{B} \in \mathcal{G}_{(d,p)}} \sum_{i=1}^{m} \text{dist}(\mathcal{B}_i, \mathcal{B}).$$

Posterior mean and variance

The posterior mean subspace

$$\mathcal{B}_{\text{Bayes}} = \arg \min_{\mathcal{B} \in \mathcal{G}_{(d,p)}} \sum_{i=1}^{m} \text{dist}(\mathcal{B}_i, \mathcal{B}).$$

Uncertainty

$$\text{var}(\{\mathcal{B}_1, \cdots, \mathcal{B}_m\}) = \frac{1}{m} \sum_{i=1}^{m} \text{dist}(\mathcal{B}_i, \mathcal{B}_{\text{Bayes}}).$$

Distribution theory on Grassmann manifolds

If B is a linear space of d central normal vectors in $\mathbb{R}^p$ with covariance matrix Σ the density of Grassmannian distribution $\mathscr{G}_\Sigma$ w.r.t. reference measure $\mathscr{G}_\mathbf{I}$ is

$$\frac{d\mathscr{G}_\Sigma}{d\mathscr{G}_\mathbf{I}}(\langle X\rangle) = \left(\frac{\det(X^T X)}{\det(X^T\Sigma^{-1}X)}\right)^{d/2},$$

where $\langle X\rangle \equiv \text{span}(X)$ where $X = (x_1, \ldots x_d)$.

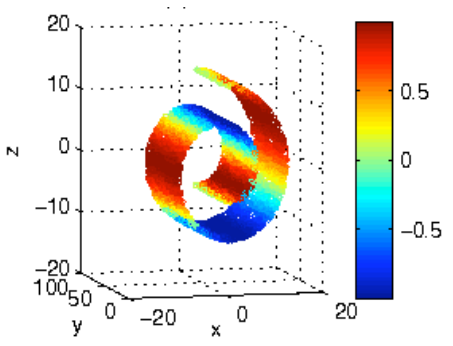
Swiss roll

$$X_1 = t\cos(t), \quad X_2 = h, \quad X_3 = t\sin(t), \quad X_{4,\ldots,10} \overset{iid}{\sim} \text{No}(0,1)$$

where $t = \frac{3\pi}{2}(1+2\theta)$, $\theta \sim \text{Unif}(0,1)$, $h \sim \text{Unif}(0,1)$ and

$$Y = \sin(5\pi\theta) + h^2 + \varepsilon, \quad \varepsilon \sim \text{No}(0, 0.01).$$

Pictures

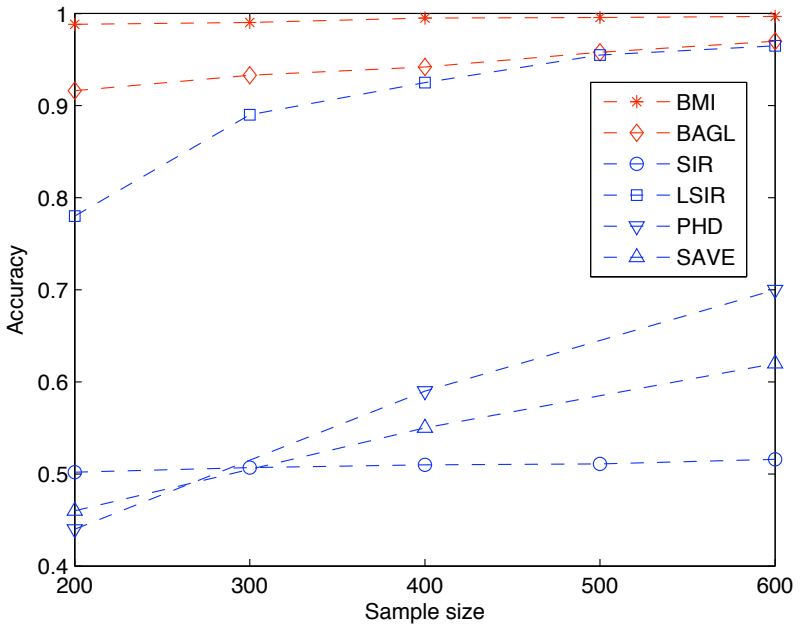


Metric

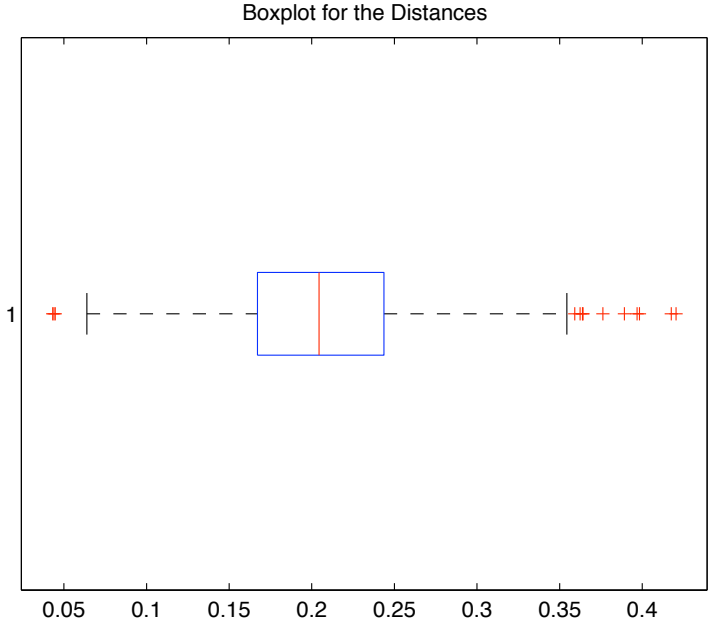
Projection of the estimated d.r. space $\hat{B} = (\hat{b}_1, \cdots, \hat{b}_d)$ onto B

$$\frac{1}{d}\sum_{i=1}^{d} ||P_B \hat{b}_i||^2 = \frac{1}{d}\sum_{i=1}^{d} ||(BB^T)\hat{b}_i||^2$$

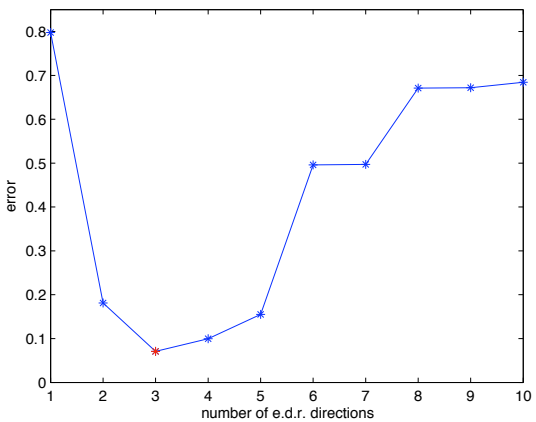
Comparison of algorithms

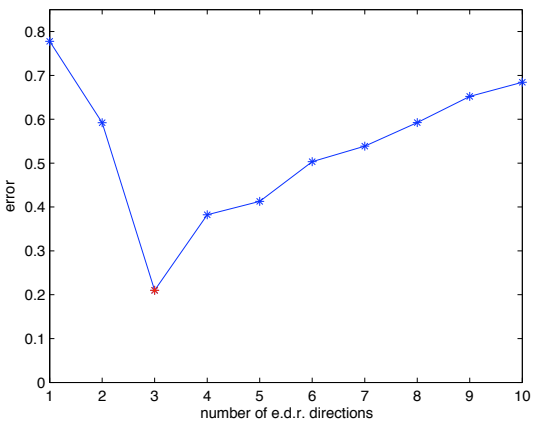


Posterior variance

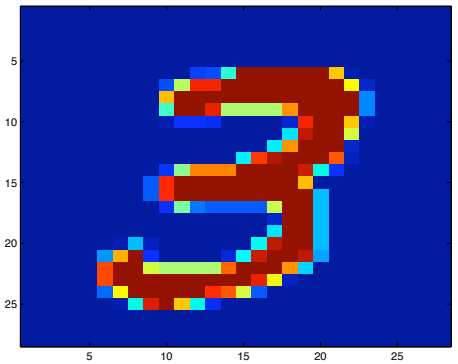


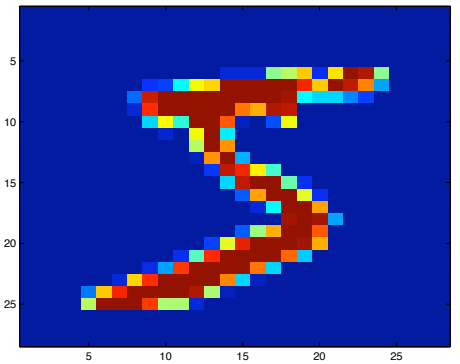
Error as a function of d

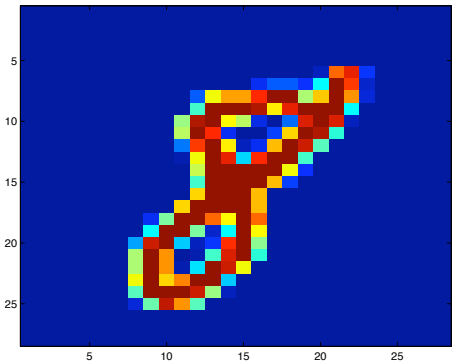




Digits


5
10
15
20
25
5
10
15
20
25


5
10
15
20
25
5
10
15
20
25


5
10
15
20
25
5
10
15
20
25

Two classification problems

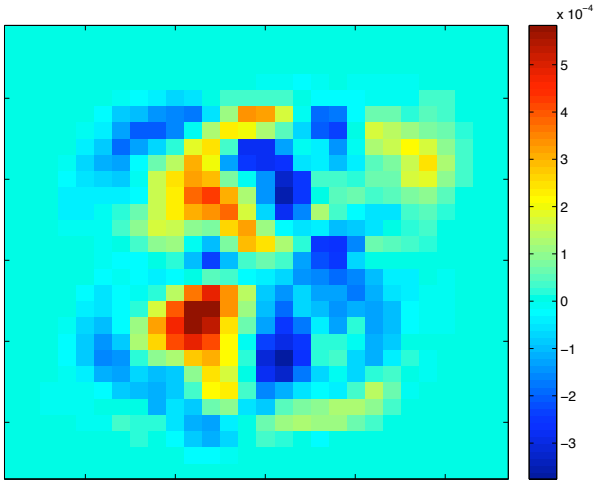
3 vs. 8 and 5 vs. 8.

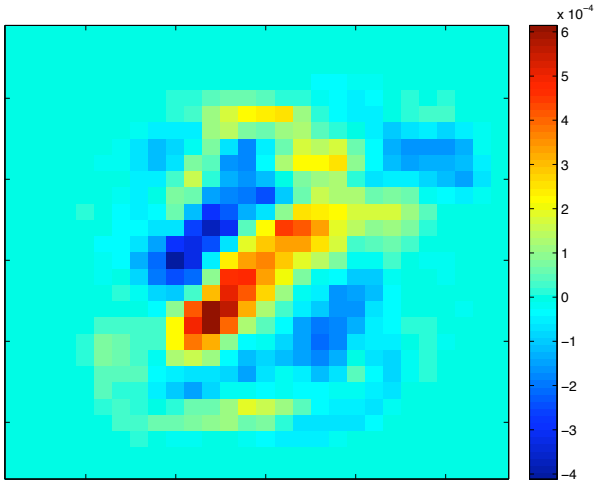
Two classification problems

3 vs. 8 and 5 vs. 8.

100 training samples from each class.

BMI





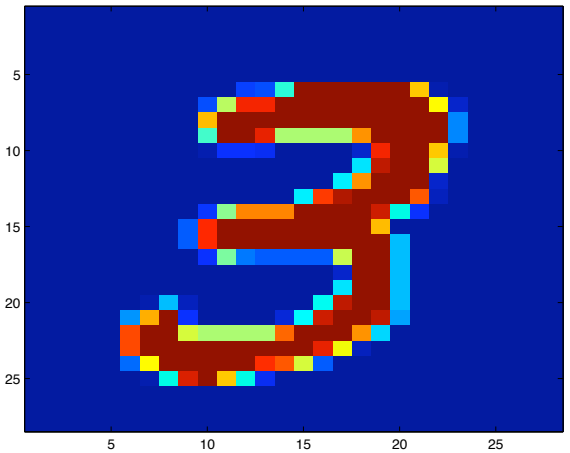
3, 5, 8 Classification Problem

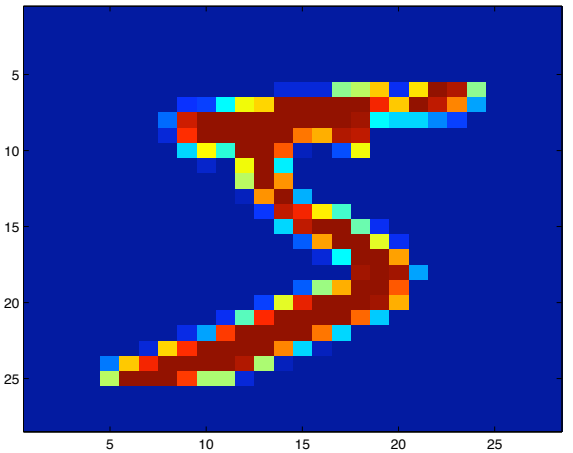
Goal

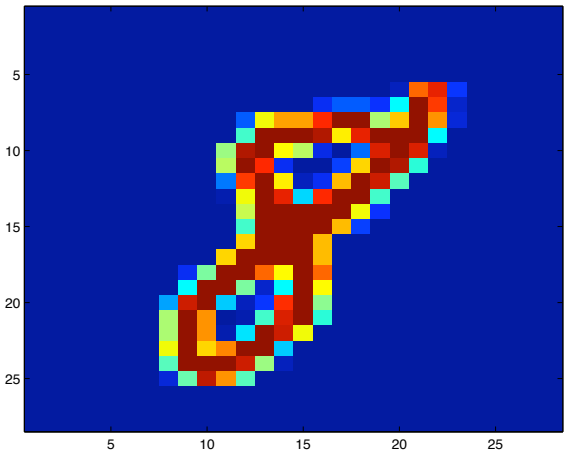
Learn features for predictive model:

- 3 vs 8
- 5 vs 8
- 3 **and** 5 vs 8

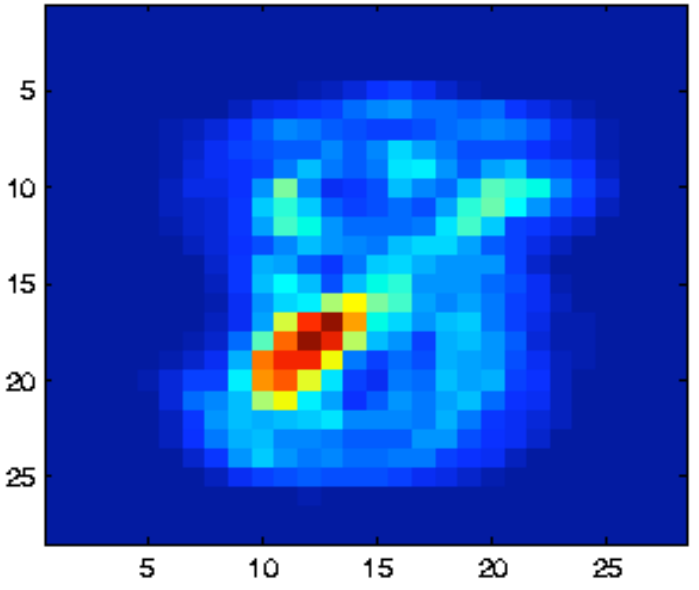
3, 5, 8 Classification problem



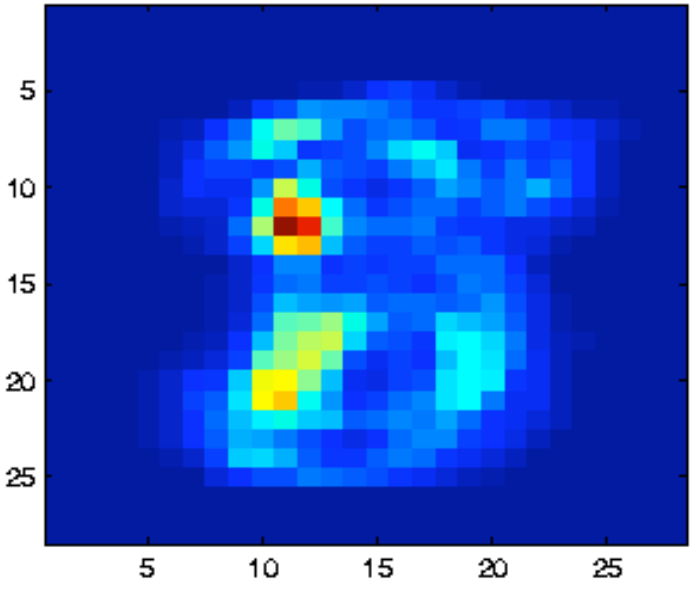




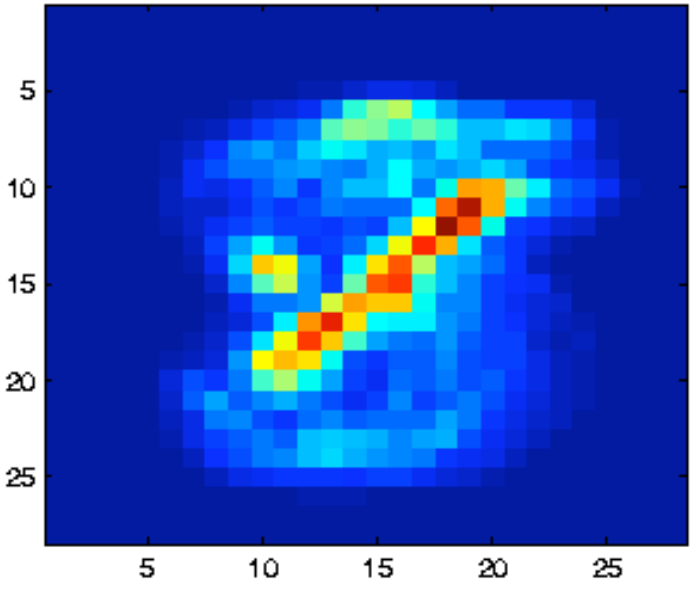
Top features: 3 and 5 vs 8



Top features: 3 vs 8



Top features: 5 vs 8



All ten digits

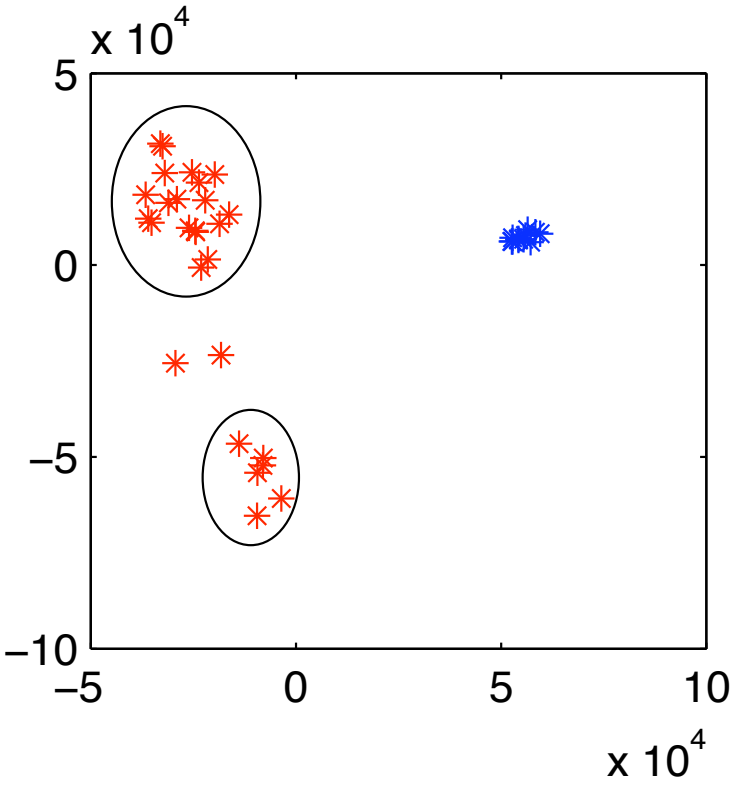
digit	Nonlinear	Linear
0	0.04($\pm$ 0.01)	0.05 ($\pm$ 0.01)
1	0.01($\pm$ 0.003)	0.03 ($\pm$ 0.01)
2	0.14($\pm$ 0.02)	0.19 ($\pm$ 0.02)
3	0.11($\pm$ 0.01)	0.17 ($\pm$ 0.03)
4	0.13($\pm$ 0.02)	0.13 ($\pm$ 0.03)
5	0.12($\pm$ 0.02)	0.21 ($\pm$ 0.03)
6	0.04($\pm$ 0.01)	0.0816 ($\pm$ 0.02)
7	0.11($\pm$ 0.01)	0.14 ($\pm$ 0.02)
8	0.14($\pm$ 0.02)	0.20 ($\pm$ 0.03)
9	0.11($\pm$ 0.02)	0.15 ($\pm$ 0.02)
average	0.09	0.14

Table: Average classification error rate and standard deviation on the digits data.

Cancer classification

$n = 38$ samples with expression levels for $p = 7129$ genes or ests
19 samples are Acute Myeloid Leukemia (AML)
19 are Acute Lymphoblastic Leukemia, these fall into two subclusters – B-cell and T-cell.

Substructure captured



Funding

- IGSP
- Center for Systems Biology at Duke
- NSF DMS-0732260