

# Bayesian simultaneous regression and dimension reduction

## MCMski II

Sayan Mukherjee

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

January 10, 2008

# Table of contents

- 1 Statistical principles
  - Learning gradients
  - Dimension reduction and dependence structure
  - Bayesian formulation

- 2 Illustrations of principles
  - Simulated data
  - Digits

- 3 Modeling tumor progression
  - Pathways and gene sets
  - Progression in prostate cancer

- 4 Thanks

## Motivation and related work

Data generated by measuring thousands of variables lies on or near a low-dimensional manifold or strong dependencies between variables.

## Motivation and related work

Data generated by measuring thousands of variables lies on or near a low-dimensional manifold or strong dependencies between variables.

Manifold learning: LLE, ISOMAP, Laplacian Eigenmaps, Hessian Eigenmaps.

## Motivation and related work

Data generated by measuring thousands of variables lies on or near a low-dimensional manifold or strong dependencies between variables.

Manifold learning: LLE, ISOMAP, Laplacian Eigenmaps, Hessian Eigenmaps.

Simultaneous dimensionality reduction and regression: SIR, MAVE, SAVE.

## Generative vs. predictive modelling

Given data =  $\{Z_i = (x_i, y_i)\}_{i=1}^n$  with  $Z_i \stackrel{iid}{\sim} \rho(X, Y)$ .

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

## Generative vs. predictive modelling

Given data =  $\{Z_i = (x_i, y_i)\}_{i=1}^n$  with  $Z_i \stackrel{iid}{\sim} \rho(X, Y)$ .

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

Two options

- 1 discriminative or regression  $Y|X$
- 2 generative  $X|Y$  (sometimes called inverse regression)

# Regression

Given  $X \in \mathcal{X} \subset \mathbb{R}^p$  and  $Y \subset \mathbb{R}$  and  $p \gg n$  and  $\rho(X, Y)$  we want  $Y|X$ .

A natural idea

$$f_r(x) = \arg \min[\text{var}(f)] = \arg \min \mathbb{E}_Y(Y - f(X))^2,$$

and  $f_r(x) = \mathbb{E}_Y[Y|x]$  provides a summary of  $Y|X$ .

# Inverse regression

Given  $X \in \mathcal{X} \subset \mathbb{R}^p$  and  $Y \subset \mathbb{R}$  and  $p \gg n$  and  $\rho(X, Y)$  we want  $X|Y$ .

$\Omega = \text{cov}(X|Y)$  provides a summary of  $X|Y$ .

# Inverse regression

Given  $X \in \mathcal{X} \subset \mathbb{R}^p$  and  $Y \subset \mathbb{R}$  and  $p \gg n$  and  $\rho(X, Y)$  we want  $X|Y$ .

$\Omega = \text{cov}(X|Y)$  provides a summary of  $X|Y$ .

- 1  $\Omega_{jj}$  – relevance of variable with respect to label
- 2  $\Omega_{ij}$  – covariation with respect to label

# Learning gradients

Model simultaneously  $f_r(x)$  and  $\nabla f_r = \left( \frac{\partial f_r}{\partial x^1}, \dots, \frac{\partial f_r}{\partial x^p} \right)^T$ .

# Learning gradients

Model simultaneously  $f_r(x)$  and  $\nabla f_r = \left( \frac{\partial f_r}{\partial x^1}, \dots, \frac{\partial f_r}{\partial x^p} \right)^T$ .

- 1 regression:  $f_r(x)$
- 2 inverse regression: gradient outer product (GOP)  
 $\Gamma = \mathbb{E}[\nabla f_r \otimes \nabla f_r]$  or

$$\Gamma_{ij} = \left\langle \frac{\partial f_r}{\partial x^i}, \frac{\partial f_r}{\partial x^j} \right\rangle.$$

## Linear case

We start with the linear case

$$y = w \cdot x + \varepsilon, \quad \varepsilon \stackrel{iid}{\sim} \text{No}(0, \sigma^2).$$

$$\Sigma_X = \text{cov}(X), \quad \sigma_Y^2 = \text{var}(Y).$$

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

$\Gamma$  and  $\Omega$  are equivalent modulo rotation and scale.

# Nonlinear case

For smooth  $f(x)$

$$y = f(x) + \varepsilon, \quad \varepsilon \stackrel{iid}{\sim} \text{No}(0, \sigma^2).$$

$\Omega = \text{cov}(X|Y)$  not so clear.

# Nonlinear case

Partition into sections and compute local quantities

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

# Nonlinear case

Partition into sections and compute local quantities

$$\mathcal{X} = \bigcup_{i=1}^{\mathcal{I}} \mathcal{X}_i$$
$$\Omega_i = \text{cov}(\mathbf{X}_{\mathcal{X}_i} | \mathbf{Y}_{\mathcal{X}_i})$$

# Nonlinear case

Partition into sections and compute local quantities

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

$$\Omega_i = \text{cov}(\mathbf{X}_{\mathcal{X}_i} | Y_{\mathcal{X}_i})$$

$$\Sigma_i = \text{cov}(\mathbf{X}_{\mathcal{X}_i})$$

# Nonlinear case

Partition into sections and compute local quantities

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

$$\Omega_i = \text{cov}(\mathbf{X}_{\mathcal{X}_i} | Y_{\mathcal{X}_i})$$

$$\Sigma_i = \text{cov}(\mathbf{X}_{\mathcal{X}_i})$$

$$\sigma_i^2 = \text{var}(Y_{\mathcal{X}_i})$$

# Nonlinear case

Partition into sections and compute local quantities

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

$$\Omega_i = \text{cov}(\mathbf{X}_{\mathcal{X}_i} | Y_{\mathcal{X}_i})$$

$$\Sigma_i = \text{cov}(\mathbf{X}_{\mathcal{X}_i})$$

$$\sigma_i^2 = \text{var}(Y_{\mathcal{X}_i})$$

$$m_i = \rho_{\mathbf{X}}(\mathcal{X}_i).$$

# Nonlinear case

Partition into sections and compute local quantities

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

$$\Omega_i = \text{cov}(\mathbf{X}_{\mathcal{X}_i} | Y_{\mathcal{X}_i})$$

$$\Sigma_i = \text{cov}(\mathbf{X}_{\mathcal{X}_i})$$

$$\sigma_i^2 = \text{var}(Y_{\mathcal{X}_i})$$

$$m_i = \rho_{\mathbf{X}}(\mathcal{X}_i).$$

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

## Gradient estimate – for regression

Taylor expanding  $f(x)$  around data should result in

$$(f(x_j) - f(x_i) + \nabla f(x_i) \cdot (x_j - x_i))^2 \approx 0 \quad \text{for } x_i \approx x_j.$$

## Gradient estimate – for regression

Taylor expanding  $f(x)$  around data should result in

$$(f(x_j) - f(x_i) + \nabla f(x_i) \cdot (x_j - x_i))^2 \approx 0 \quad \text{for } x_i \approx x_j.$$

$$L(f, \vec{f}, \text{data}) = \sum_{ij} w_{ij} (y_j - f(x_i) + \vec{f}(x_i) \cdot (x_j - x_i))^2.$$

## Gradient estimate – for regression

Taylor expanding  $f(x)$  around data should result in

$$(f(x_j) - f(x_i) + \nabla f(x_i) \cdot (x_j - x_i))^2 \approx 0 \quad \text{for } x_i \approx x_j.$$

$$L(f, \vec{f}, \text{data}) = \sum_{ij} w_{ij} (y_j - f(x_i) + \vec{f}(x_i) \cdot (x_j - x_i))^2.$$

Similar idea for classification, link function.

# Gradient estimate

## Optimization Problem

$$(f_D, \vec{f}_D) = \arg \min_{(f, \vec{f}) \in \mathcal{H}_K^{p+1}} \left\{ L(f, \vec{f}, \text{data}) + \lambda_1 \|f\|_K^2 + \lambda_2 \|\vec{f}\|_K^2 \right\}$$

$\vec{f}$  is vector of gradients

$\lambda_1, \lambda_2$  are regularization terms

$L(\cdot)$  is empirical error using convex loss function

## Gradient estimate

### Optimization Problem

$$(f_D, \vec{f}_D) = \arg \min_{(f, \vec{f}) \in \mathcal{H}_K^{p+1}} \left\{ L(f, \vec{f}, \text{data}) + \lambda_1 \|f\|_K^2 + \lambda_2 \|\vec{f}\|_K^2 \right\}$$

$\vec{f}$  is vector of gradients

$\lambda_1, \lambda_2$  are regularization terms

$L(\cdot)$  is empirical error using convex loss function

### Representation form

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

with  $a_D = (a_{1,D}, \dots, a_{n,D}) \in \mathbb{R}^n$  and  $c_D = (c_{1,D}, \dots, c_{n,D})^T \in \mathbb{R}^{np}$ .

## Gradient Outer Product (GOP)

A central quantity in this talk will be the GOP.

### Definition (GOP)

$$\hat{\Gamma} = \vec{f}_D \otimes \vec{f}_D = c_D^T \mathbf{K}_{c_D} \approx \mathbb{E}(\nabla f \otimes \nabla f)$$

# Dimension reduction

## Proposition

*The eigenvectors corresponding to the  $d$  non-zero eigenvalues of  $\Gamma$  span the subspace relevant to prediction.*

Gradients provide information on the predictive directions  $b_i$ ,  
 $i = 1, \dots, d$ .

# Gaussian Markov distributions over graphs

Give a multivariate normal distribution with covariance matrix  $\Sigma$   
the matrix  $P = \Sigma^{-1}$  is the conditional independence matrix

$$P_{ij} = \text{dependence of } i \leftrightarrow j \mid \text{all other variables.}$$

# Gaussian Markov distributions over graphs

Give a multivariate normal distribution with covariance matrix  $\Sigma$   
the matrix  $P = \Sigma^{-1}$  is the conditional independence matrix

$$P_{ij} = \text{dependence of } i \leftrightarrow j \mid \text{all other variables.}$$

Note by construction  $\hat{\Gamma}$  is a covariance matrix of a Gaussian process.

## Gaussian Markov distributions over graphs

Give a multivariate normal distribution with covariance matrix  $\Sigma$   
the matrix  $P = \Sigma^{-1}$  is the conditional independence matrix

$$P_{ij} = \text{dependence of } i \leftrightarrow j \mid \text{all other variables.}$$

Note by construction  $\hat{\Gamma}$  is a covariance matrix of a Gaussian process.

$J = \text{inv}(\hat{\Gamma})$  is the inferred conditional independence matrix.

# Restriction to a manifold

Assume the data is concentrated on a manifold  $\mathcal{M} \subset \mathbb{R}^p$  with  $\mathcal{M} \in \mathbb{R}^d$  and there exists an isometric embedding  $\varphi : \mathcal{M} \rightarrow \mathbb{R}^p$ .

# Restriction to a manifold

Assume the data is concentrated on a manifold  $\mathcal{M} \subset \mathbb{R}^p$  with  $\mathcal{M} \in \mathbb{R}^d$  and there exists an isometric embedding  $\varphi : \mathcal{M} \rightarrow \mathbb{R}^p$ .

## Theorem

*Under mild regularity conditions on the distribution and corresponding density, with probability  $1 - \delta$*

$$\begin{aligned} \|(d\varphi)^* \vec{f}_D - \nabla_{\mathcal{M}} f\|_{\rho_X} &\leq C \log\left(\frac{1}{\delta}\right) n^{-1/d} \\ \|f_D - f\|_{\rho_X} &\leq C \log\left(\frac{1}{\delta}\right) n^{-1/d}, \end{aligned}$$

where  $(d\varphi)^*$  is the dual of the map  $d\varphi$ .

## Bayesian kernel model for regression

$$y_i = f(x_i) + \varepsilon, \quad \varepsilon \stackrel{iid}{\sim} \text{No}(0, \sigma^2).$$

$$f(x) = \int_{\mathcal{X}} K(x, u) Z(du)$$

where  $Z(du) \in \mathcal{M}(\mathcal{X})$  is a signed measure on  $\mathcal{X}$ .

## Bayesian kernel model for regression

$$y_i = f(x_i) + \varepsilon, \quad \varepsilon \stackrel{iid}{\sim} \text{No}(0, \sigma^2).$$

$$f(x) = \int_{\mathcal{X}} K(x, u) Z(du)$$

where  $Z(du) \in \mathcal{M}(\mathcal{X})$  is a signed measure on  $\mathcal{X}$ .

$$\pi(Z|\text{data}) \propto L(\text{data}|Z) \pi(Z),$$

this implies a posterior on  $f$ .

## Priors and integral operators

Integral operator  $\mathcal{L}_K : \Gamma \rightarrow \mathcal{G}$

$$\mathcal{G} = \left\{ f \mid f(x) := \mathcal{L}_K[\gamma](x) = \int_{\mathcal{X}} K(x, u) d\gamma(u), \quad \gamma \in \Gamma \right\},$$

with  $\Gamma \subseteq \mathcal{B}(\mathcal{X})$ .

## Priors and integral operators

Integral operator  $\mathcal{L}_K : \Gamma \rightarrow \mathcal{G}$

$$\mathcal{G} = \left\{ f \mid f(x) := \mathcal{L}_K[\gamma](x) = \int_{\mathcal{X}} K(x, u) d\gamma(u), \quad \gamma \in \Gamma \right\},$$

with  $\Gamma \subseteq \mathcal{B}(\mathcal{X})$ .

A prior on  $\Gamma$  implies a prior on  $\mathcal{G}$ .

## Equivalence with RKHS

For what  $\Gamma$  is  $\mathcal{H}_K = \text{span}(\mathcal{G})$  ?

What is  $\mathcal{L}_K^{-1}(\mathcal{H}_K) = ??$ . This is hard to characterize.

## Equivalence with RKHS

For what  $\Gamma$  is  $\mathcal{H}_K = \text{span}(\mathcal{G})$  ?

What is  $\mathcal{L}_K^{-1}(\mathcal{H}_K) = ??$ . This is hard to characterize.

An appropriate choice for  $\Gamma$  is the union of integrable functions and discrete measures.

## Signed measures are (almost) just right

Nonsingular measures:  $\mathcal{M} = L^1(\mathcal{X}) \cup \mathcal{M}_D$

### Proposition

$\mathcal{L}_K(\mathcal{M})$  is dense in  $\mathcal{H}_K$  with respect to the RKHS norm.

## Signed measures are (almost) just right

Nonsingular measures:  $\mathcal{M} = L^1(\mathcal{X}) \cup \mathcal{M}_D$

### Proposition

$\mathcal{L}_K(\mathcal{M})$  is dense in  $\mathcal{H}_K$  with respect to the RKHS norm.

### Proposition

$\mathcal{B}(\mathcal{X}) \subsetneq \mathcal{L}_K^{-1}(\mathcal{H}_K(\mathcal{X}))$ .

## The implication

Take home message – need priors on signed measures.

A function theoretic foundation for random signed measures such as Gaussian, Dirichlet and Lévy process priors.

## Bayesian kernel model

$$y_i = f(x_i) + \varepsilon, \quad \varepsilon \stackrel{iid}{\sim} \text{No}(0, \sigma^2).$$

$$f(x) = \int_{\mathcal{X}} K(x, u) Z(du)$$

where  $Z(du) \in \mathcal{M}(\mathcal{X})$  is a signed measure on  $\mathcal{X}$ .

## Bayesian kernel model

$$y_i = f(x_i) + \varepsilon, \quad \varepsilon \stackrel{iid}{\sim} \text{No}(0, \sigma^2).$$

$$f(x) = \int_{\mathcal{X}} K(x, u) Z(du)$$

where  $Z(du) \in \mathcal{M}(\mathcal{X})$  is a signed measure on  $\mathcal{X}$ .

$$\pi(Z|\text{data}) \propto L(\text{data}|Z) \pi(Z),$$

this implies a posterior on  $f$ .

## Dirichlet process prior

$$f(x) = \int_{\mathcal{X}} K(x, u)Z(du) = \int_{\mathcal{X}} K(x, u)w(u)F(du)$$

$F(du)$  is a distribution and  $w(u)$  a coefficient function.

## Dirichlet process prior

$$f(x) = \int_{\mathcal{X}} K(x, u)Z(du) = \int_{\mathcal{X}} K(x, u)w(u)F(du)$$

$F(du)$  is a distribution and  $w(u)$  a coefficient function.

Model  $F$  using a Dirichlet process prior:  $DP(\alpha, F_0)$

## Bayesian representer form

Given  $X_n = (x_1, \dots, x_n) \stackrel{iid}{\sim} F$

$$F | X_n \sim \text{DP}(\alpha + n, F_n), \quad F_n = (\alpha F_0 + \sum_{i=1}^n \delta_{x_i}) / (\alpha + n).$$

$$\mathbb{E}[f | X_n] = a_n \int K(x, u) w(u) dF_0(u) + n^{-1} (1 - a_n) \sum_{i=1}^n w(x_i) K(x, x_i),$$

$$a_n = \alpha / (\alpha + n).$$

## Bayesian representer form

Taking  $\lim \alpha \rightarrow 0$  to represent a non-informative prior:

Proposition (Bayesian representer form)

$$\hat{f}_n(x) = \sum_{i=1}^n w_i K(x, x_i),$$

$$w_i = w(x_i)/n.$$

## Bayesian kernel model for gradient estimates

By Taylor expansion

$$y_i = f(x_j) + \vec{f}(x_i) \cdot (x_i - x_j) + \varepsilon_{x_i, x_j}.$$

## Bayesian kernel model for gradient estimates

By Taylor expansion

$$y_i = f(x_j) + \vec{f}(x_j) \cdot (x_i - x_j) + \varepsilon_{x_i, x_j}.$$

By representer form

$$y_i = \alpha_0 + K\alpha + (\iota x_i' - X)CK_i + \varepsilon_i$$

where  $\iota = (1, \dots, 1)'$ ,  $\alpha = (\alpha_1, \dots, \alpha_n)' \in \mathbb{R}^n$ ,  
 $C = (c_1, \dots, c_n) \in \mathbb{R}^{p \times n}$ ,  $X$  is the  $n \times p$  data matrix,  $K_i$  is the  $i$ th column of  $K$ .

## Likelihood: error term and spatial statistics

Intuition: Consider a spatial model (similarity matrix)

$$w_{ij} := \theta \exp\{-\phi \|x_i - x_j\|\},$$

where  $\theta$  and  $\phi$  are parameters of a spatial model.

## Likelihood: error term and spatial statistics

Intuition: Consider a spatial model (similarity matrix)

$$w_{ij} := \theta \exp\{-\phi \|x_i - x_j\|\},$$

where  $\theta$  and  $\phi$  are parameters of a spatial model.  
A natural modeling assumption is

$$\varepsilon_{x_i, x_j} \propto w_{ij}^{-1}.$$

## Likelihood: error term and spatial statistics

Intuition: Consider a spatial model (similarity matrix)

$$w_{ij} := \theta \exp\{-\phi \|x_i - x_j\|\},$$

where  $\theta$  and  $\phi$  are parameters of a spatial model.  
A natural modeling assumption is

$$\varepsilon_{x_i, x_j} \propto w_{ij}^{-1}.$$

Given this spatial structure

$$\varepsilon_i \sim \text{No}_n(0, W_i^{-1})$$

where  $W_i = \text{diag}(w_{x_i, x_1}, \dots, w_{x_i, x_n})$ .

# Likelihood

Given the error model the likelihood is

$$L(\text{data}|f, \vec{f}) \propto \sqrt{\prod_{ij} w_{ij}} \exp \left\{ -\frac{1}{2} \sum_i (e_i' W_i e_i) \right\}$$

# Likelihood

Given the error model the likelihood is

$$L(\text{data} | f, \vec{f}) \propto \sqrt{\prod_{ij} w_{ij}} \exp \left\{ -\frac{1}{2} \sum_i (e_i' W_i e_i) \right\}$$

with

$$K = F \Delta F'$$

$$\Delta := \text{diag}(\lambda_1^2, \dots, \lambda_n^2)$$

$$\alpha = F \Delta^{-1} \beta$$

$$e_i = y_i - \alpha_0 - F \beta - (\iota x_i' - X) C K_i$$

# Prior specification

$$\pi(\alpha_0, \theta) \propto 1/\theta,$$

## Prior specification

$$\begin{aligned}\pi(\alpha_0, \theta) &\propto 1/\theta, \\ \beta &\sim \text{No}(0, T)\end{aligned}$$

## Prior specification

$$\begin{aligned}\pi(\alpha_0, \theta) &\propto 1/\theta, \\ \beta &\sim \text{No}(0, T) \\ T &:= \text{diag}(\tau_1, \dots, \tau_n), \tau_i^{-1} \sim \text{Ga}(a_\tau/2, b_\tau/2),\end{aligned}$$

## Prior specification

$$\begin{aligned}\pi(\alpha_0, \theta) &\propto 1/\theta, \\ \beta &\sim \text{No}(0, T) \\ T &:= \text{diag}(\tau_1, \dots, \tau_n), \tau_i^{-1} \sim \text{Ga}(a_\tau/2, b_\tau/2), \\ C_{kj} &\sim (1 - \pi_k)\delta_0 + \pi_k \text{No}(0, \phi_k^{-1})\end{aligned}$$

## Prior specification

$$\begin{aligned}\pi(\alpha_0, \theta) &\propto 1/\theta, \\ \beta &\sim \text{No}(0, T) \\ T &:= \text{diag}(\tau_1, \dots, \tau_n), \tau_i^{-1} \sim \text{Ga}(a_\tau/2, b_\tau/2), \\ C_{kj} &\sim (1 - \pi_k)\delta_0 + \pi_k \text{No}(0, \phi_k^{-1}) \\ \phi_k &\sim \text{Ga}(\alpha_c/2, \beta_c/2)\end{aligned}$$

## Prior specification

$$\begin{aligned}\pi(\alpha_0, \theta) &\propto 1/\theta, \\ \beta &\sim \text{No}(0, T) \\ T &:= \text{diag}(\tau_1, \dots, \tau_n), \tau_i^{-1} \sim \text{Ga}(a_\tau/2, b_\tau/2), \\ C_{kj} &\sim (1 - \pi_k)\delta_0 + \pi_k \text{No}(0, \phi_k^{-1}) \\ \phi_k &\sim \text{Ga}(\alpha_c/2, \beta_c/2) \\ \pi_k &\sim \text{Beta}(\alpha_\pi, \beta_\pi),\end{aligned}$$

## Prior specification

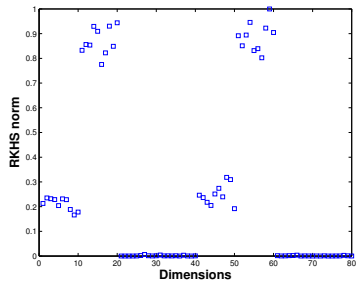
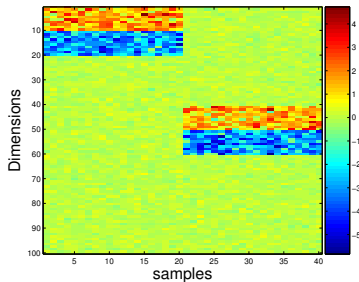
$$\begin{aligned}\pi(\alpha_0, \theta) &\propto 1/\theta, \\ \beta &\sim \text{No}(0, T) \\ T &:= \text{diag}(\tau_1, \dots, \tau_n), \tau_i^{-1} \sim \text{Ga}(a_\tau/2, b_\tau/2), \\ C_{kj} &\sim (1 - \pi_k)\delta_0 + \pi_k \text{No}(0, \phi_k^{-1}) \\ \phi_k &\sim \text{Ga}(\alpha_c/2, \beta_c/2) \\ \pi_k &\sim \text{Beta}(\alpha_\pi, \beta_\pi), \\ \phi &\sim \text{Ga}(a_\phi/2, b_\phi/2)\end{aligned}$$

## Prior specification

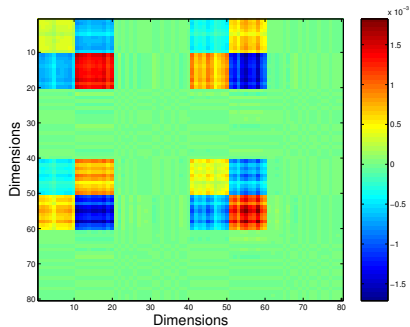
$$\begin{aligned}\pi(\alpha_0, \theta) &\propto 1/\theta, \\ \beta &\sim \text{No}(0, T) \\ T &:= \text{diag}(\tau_1, \dots, \tau_n), \tau_i^{-1} \sim \text{Ga}(a_\tau/2, b_\tau/2), \\ C_{kj} &\sim (1 - \pi_k)\delta_0 + \pi_k \text{No}(0, \phi_k^{-1}) \\ \phi_k &\sim \text{Ga}(\alpha_c/2, \beta_c/2) \\ \pi_k &\sim \text{Beta}(\alpha_\pi, \beta_\pi), \\ \phi &\sim \text{Ga}(a_\phi/2, b_\phi/2)\end{aligned}$$

Standard Gibbs sampler simulates  $p(\alpha, \alpha_0, C, \phi, \theta | \text{data})$ .

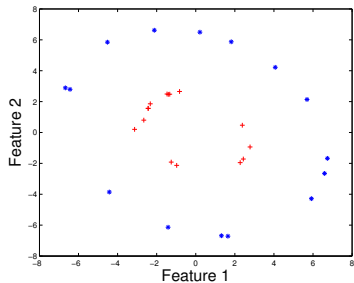
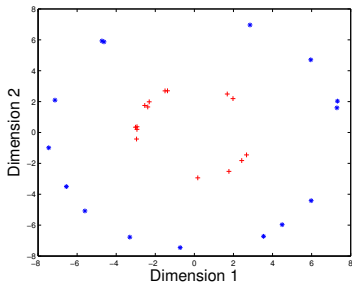
# Linear example



# Linear example



# Nonlinear example



# Digit classification

## Input

MNIST handwritten digits database:  $X_i \in \mathbb{R}^{784}$  : 28 by 28 gray-scale pixel image

# Digit classification

## Input

MNIST handwritten digits database:  $X_i \in \mathbb{R}^{784}$  : 28 by 28 gray-scale pixel image

## Formulation

Problem 1: '3 vs 8' with 50 3's, 50 8's

Problem 2: '5 vs 8' with 50 5's, 50 8's

# Digit classification

## Input

MNIST handwritten digits database:  $X_i \in \mathbb{R}^{784}$  : 28 by 28 gray-scale pixel image

## Formulation

Problem 1: '3 vs 8' with 50 3's, 50 8's

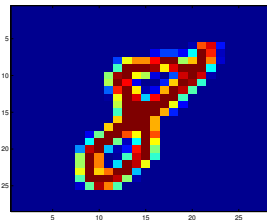
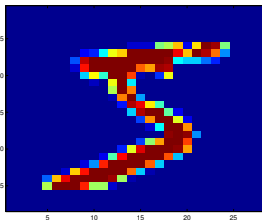
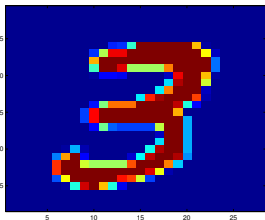
Problem 2: '5 vs 8' with 50 5's, 50 8's

## Goal

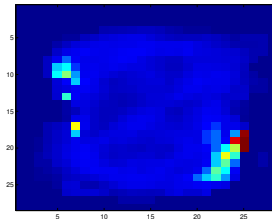
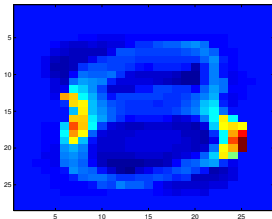
Learn features for predictive model:

- 3 vs 8
- 5 vs 8

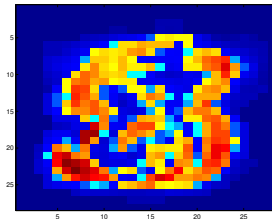
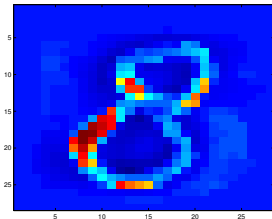
## 3, 5, 8 Classification problem



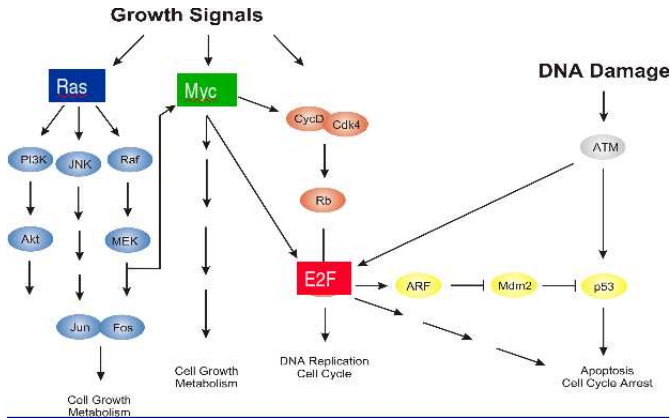
# Top features: 3 vs 8



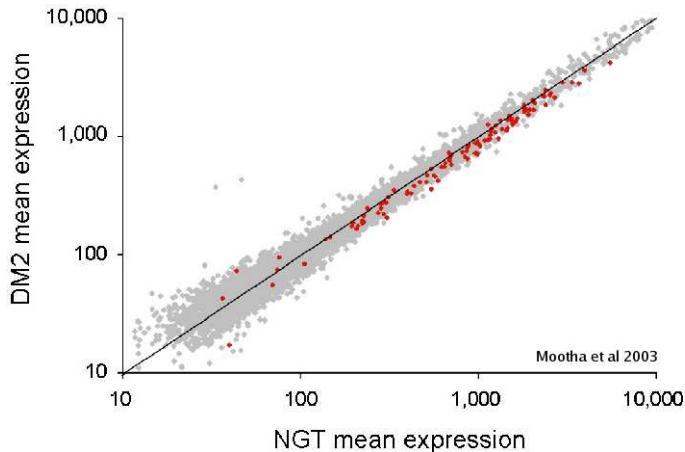
# Top features: 5 vs 8



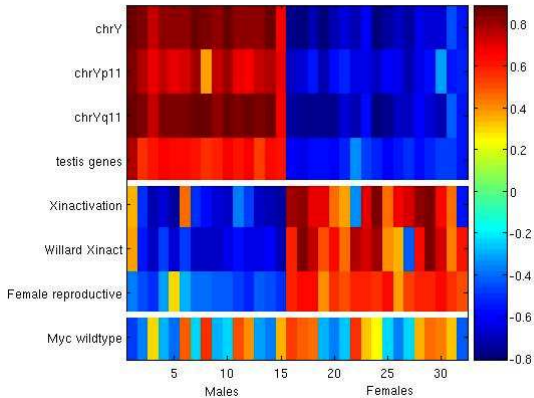
# Genes don't do things



## Diabetes – Oxphos



# Gender



# Gene set database

The gene sets in the database are defined by

- 1 Positional gene sets: cytogenetic bands, 3 megabase windows;

# Gene set database

The gene sets in the database are defined by

- 1 Positional gene sets: cytogenetic bands, 3 megabase windows;
- 2 Motif gene sets: TRANSFAC motifs, Representative motifs;

## Gene set database

The gene sets in the database are defined by

- 1 Positional gene sets: cytogenetic bands, 3 megabase windows;
- 2 Motif gene sets: TRANSFAC motifs, Representative motifs;
- 3 Curated gene sets: Pathways, Literature reviews, Animal models, Clinical phenotypes, Expert curations, Chemical or genetic perturbations.

## Progression of prostate cancer

Gene expression from 22,283 genes. 71 people 22 benign (b) prostate epithelium, 32 primary (p) prostate cancer, 17 metastatic (m) prostate cancer.

## Progression of prostate cancer

Gene expression from 22,283 genes. 71 people 22 benign (b) prostate epithelium, 32 primary (p) prostate cancer, 17 metastatic (m) prostate cancer.

Progression:  $\{b \mapsto p \mapsto m\}$ .

## Progression of prostate cancer

Gene expression from 22,283 genes. 71 people 22 benign (b) prostate epithelium, 32 primary (p) prostate cancer, 17 metastatic (m) prostate cancer.

Progression:  $\{b \mapsto p \mapsto m\}$ .

523 pathway defined gene sets.

## Progression of prostate cancer

Gene expression from 22,283 genes. 71 people 22 benign (b) prostate epithelium, 32 primary (p) prostate cancer, 17 metastatic (m) prostate cancer.

Progression:  $\{b \mapsto p \mapsto m\}$ .

523 pathway defined gene sets.

- 1 Which pathways are involved in all or some stages of progression ?

## Progression of prostate cancer

Gene expression from 22,283 genes. 71 people 22 benign (b) prostate epithelium, 32 primary (p) prostate cancer, 17 metastatic (m) prostate cancer.

Progression:  $\{b \mapsto p \mapsto m\}$ .

523 pathway defined gene sets.

- 1 Which pathways are involved in all or some stages of progression ?
- 2 What are the pathway dependencies (inferring pathway networks) ?

## Progression of prostate cancer

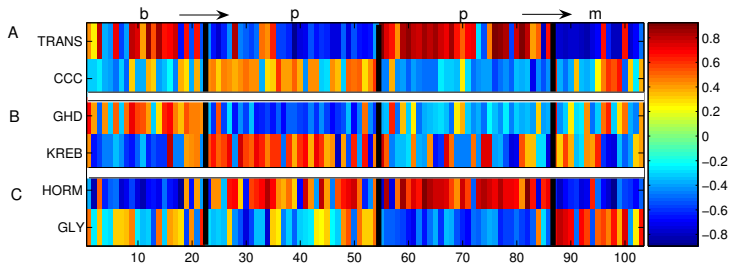
Gene expression from 22,283 genes. 71 people 22 benign (b) prostate epithelium, 32 primary (p) prostate cancer, 17 metastatic (m) prostate cancer.

Progression:  $\{b \mapsto p \mapsto m\}$ .

523 pathway defined gene sets.

- 1 Which pathways are involved in all or some stages of progression ?
- 2 What are the pathway dependencies (inferring pathway networks) ?
- 3 For each relevant pathway infer gene network for pathway.

# Pathways relevant in progression





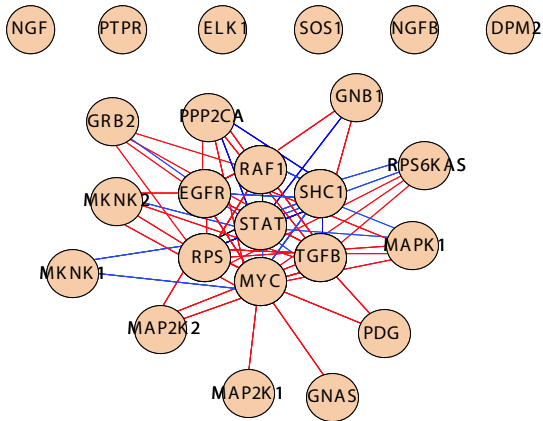
## Refinement of gene sets

- 1 Not all genes in a gene set are relevant in the specific context studied.

## Refinement of gene sets

- 1 Not all genes in a gene set are relevant in the specific context studied.
- 2 Genes not included in the gene set maybe relevant to the specific context studied.

# Gene network for ERK pathway



## Relevant papers

- Learning Coordinate Covariances via Gradients. S. Mukherjee, D-X. Zhou; Journal of Machine Learning Research, 7(Mar):519–549, 2006.
- Estimation of Gradients and Coordinate Covariation in Classification. S. Mukherjee, Q. Wu; Journal of Machine Learning Research, 7(Nov):2481–2514, 2006.
- Characterizing the Function Space for Bayesian Kernel Models. N. Pillai, Q. Wu, F. Liang, S. Mukherjee, R.L. Wolpert; Journal of Machine Learning Research, 8(Aug):1769–1797, 2007.
- Non-parametric Bayesian kernel models. F. Liang, K. Mao, M. Liao, S. Mukherjee, M. West; Biometrika, in submission.
- Learning Gradients: predictive models that infer geometry and dependence. Qiang Wu, Justin Guinney, Mauro Maggioni, Sayan Mukherjee; Journal of Machine Learning Research, submitted.
- Modeling Cancer Progression via Pathway Dependencies. E. Edelman, J. Guinney, J-T. Chi, P.G. Febbo, S. Mukherjee; PLoS Computational Biology, in press.
- Bayesian simultaneous dimension reduction and regression. K. Mao, F. Liang, S. Mukherjee, Q. Wu; in preparation.

# Acknowledgements

People that did the work:

Gradients – Q Wu, D-X Zhou, K Mao, J Guinney

## Acknowledgements

People that did the work:

Gradients – Q Wu, D-X Zhou, K Mao, J Guinney

Computational biology – E Edelman, J Guinney, P Febbo, J-T Chi

# Acknowledgements

People that did the work:

Gradients – Q Wu, D-X Zhou, K Mao, J Guinney

Computational biology – E Edelman, J Guinney, P Febbo, J-T Chi

Bayesian modeling – N Pillai, K Mao, F Liang, M West, R Wolpert

## Acknowledgements

People that did the work:

Gradients – Q Wu, D-X Zhou, K Mao, J Guinney

Computational biology – E Edelman, J Guinney, P Febbo, J-T Chi

Bayesian modeling – N Pillai, K Mao, F Liang, M West, R Wolpert

Funding:

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