Probabilistic models for supervised dimension reduction

Machine learning summer school 2009

Elena Edelman, Justin Guinney, Kai Mao, Sayan Mukherjee, Feng Liang, Qiang Wu

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

June 8, 2009
Formal probabilistic modeling

1. Place manifold learning and geometric ideas in a distributional framework.
Formal probabilistic modeling

1. Place manifold learning and geometric ideas in a distributional framework.

2. Provide an intuition and example of probabilistic (Bayesian) inference.
Formal probabilistic modeling

1. Place manifold learning and geometric ideas in a distributional framework.

2. Provide an intuition and example of probabilistic (Bayesian) inference.

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.

Given $X_1, \ldots, X_n$ drawn form a Gaussian reduce it to $\mu, \sigma^2$. 
Dimension reduction

A modern reprise of this is in machine learning is manifold learning and regression on manifolds.
A modern reprise of this is in machine learning is manifold learning and regression on manifolds.

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

A modern reprise of this is in machine learning is manifold learning and regression on manifolds.

If the data lives in a p-dimensional space $X \in \mathbb{R}^p$ replace $Y \mid X$ with $Y \mid \Theta(X)$ 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 $X_1, ..., X_n \in \mathbb{R}^p$

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

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

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

Is there a submanifold $S \equiv S_{Y|X}$ such that $Y \independent X \mid P_S(X)$?
Visualization of SDR
Linear projections capture nonlinear manifolds

In this talk $P_S(X) = B'X$ where $B = (b_1, \ldots, b_d)$. 
Linear projections capture nonlinear manifolds

In this talk $P_S(X) = B'X$ where $B = (b_1, ..., b_d)$.

Semiparametric model

$$Y_i = f(X_i) + \varepsilon_i = g(b'_1 X_i, \ldots, b'_d X_i) + \varepsilon_i,$$

span $B$ is the dimension reduction (d.r.) subspace.
Three approaches to estimate $B$

1. Forward regression: Directly model $Y \mid P_S \mathbf{X}$. Projection Pursuit Regression (PPR).
Three approaches to estimate $B$

1. Forward regression: Directly model $Y \mid P_S X$. Projection Pursuit Regression (PPR).
2. Inverse regression: Model $X \mid Y$. Sliced inverse regression (SIR).
Three approaches to estimate $B$

1. Forward regression: Directly model $Y \mid P_S X$. Projection Pursuit Regression (PPR).
2. Inverse regression: Model $X \mid Y$. Sliced inverse regression (SIR).
3. Learning gradients: The gradient of the regression function $\nabla f$ spans the d.r. space. principal Hessian directions (pHd).
Two probabilistic models

1. Bayesian mixture of inverses (BMI): Highlights SDR on manifolds as a generative mixture model for $X \mid Y$. 
Two probabilistic models

1. Bayesian mixture of inverses (BMI): Highlights SDR on manifolds as a generative mixture model for $X \mid Y$.

2. Bayesian gradient learning (BAGL): Highlights the differential geometric aspects of SDR.
Principal fitted components (PFC)

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

\[
X_y \sim \text{No}(\mu_y, \Delta), \\
\mu_y = \mu + A\nu_y
\]

\( \mu \in \mathbb{R}^p \)

\( A \in \mathbb{R}^{p \times d} \)

\( \nu_y \in \mathbb{R}^p \).
Principal fitted components (PFC)

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

$$X_y \sim \text{No}(\mu_y, \Delta),$$

$$\mu_y = \mu + A\nu_y$$

$$\mu \in \mathbb{R}^p$$

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

$$\nu_y \in \mathbb{R}^p.$$
Principal fitted components (PFC)

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

$$X_y \sim \text{No}(\mu_y, \Delta),$$

$$\mu_y = \mu + A\nu_y$$

$\mu \in \mathbb{R}^p$

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

$\nu_y \in \mathbb{R}^p$.

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

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.
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

\[ X_{y} \sim \text{No}(\mu_{yx}, \Delta) \]
\[ \mu_{yx} = \mu + A\nu_{yx} \]
\[ \nu_{yx} \sim G_{y} \]

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

Proposition

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

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

Define $\nu_i \equiv \nu_{y_ix_i}$. 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 $G_y$

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

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

$\nu_i$ modeled as a mixture of $C$ distributions $G_1, \ldots, G_C$ with a Dirichlet process model for each distribution

$$G_c \sim \text{DP}(\alpha_0, G_0).$$
Continuous response: modeling $G_y$

$Y \in \mathbb{R}$, so each point has a distributions

$$\nu_i \mid y_i \sim G_{y_i},$$

with $G_{y_j}$ and $G_{y_k}$ dependent if $y_j$ and $y_i$ are close.
Continuous response: modeling $G_y$

$Y \in \mathbb{R}$, so each point has a distribution

$$\nu_i \mid y_i \sim G_{y_i},$$

with $G_{y_j}$ and $G_{y_k}$ dependent if $y_j$ and $y_i$ are close.

$\nu_i$ modeled as a Dependent Dirichlet process model for each distribution $G_{y_i}$. 

---

Probabilistic models for supervised dimension reduction

BMI
Likelihood

\[ \text{Lik}(\text{data} \mid \theta) \equiv \text{Lik}(\text{data} \mid A, \Delta, \nu_1, \ldots, \nu_n, \mu) \]

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

Given data

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

Given data

\[ P_\theta \equiv \text{Post}(\theta \mid \text{data}) \propto \text{Lik}(\theta \mid \text{data}) \times \pi(\theta). \]

1. \( P_\theta \) provides estimate of (un)certainty on \( \theta \)
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 \( \theta \)
2. Requires prior on \( \theta \)
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 \( \theta \)
2. Requires prior on \( \theta \)
3. Sample from \( \mathcal{P}_\theta \)?
Markov chain Monte Carlo

No closed form for $P_\theta$. 
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 $\theta(t) = (\mu(t), A(t), \Delta^{-1}(t), \nu(t))$ 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

\[
\begin{align*}
\theta/\mu(t) & \equiv (A(t), \Delta^{-1}(t), \nu(t)) \\
\theta/A(t) & \equiv (\mu(t), \Delta^{-1}(t), \nu(t)) \\
\theta/\Delta^{-1}(t) & \equiv (\mu(t), A(t), \nu(t)) \\
\theta/\nu(t) & \equiv (\mu(t), A(t), \Delta^{-1}(t)).
\end{align*}
\]
Gibbs sampling

Conditional probabilities can be used to sample $\mu$, $\Delta^{-1}$, $A$

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

Conditional probabilities can be used to sample $\mu, \Delta^{-1}, A$

$$
\mu^{(t+1)} \mid \left( \text{data, } \theta^{/\mu}_{(t)} \right) \sim \text{No} \left( \text{data, } \theta^{/\mu}_{(t)} \right), \\
\Delta^{-1}_{(t+1)} \mid \left( \text{data, } \theta^{/\Delta^{-1}}_{(t)} \right) \sim \text{InvWishart} \left( \text{data, } \theta^{/\Delta^{-1}}_{(t)} \right)
$$
Gibbs sampling

Conditional probabilities can be used to sample $\mu, \Delta^{-1}, A$

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

$$
\Delta^{-1}^{(t+1)} \mid \left( \text{data, } \theta^{/\Delta^{-1}} \right) \sim \text{InvWishart} \left( \text{data, } \theta^{/\Delta^{-1}} \right)
$$

$$
A^{(t+1)} \mid \left( \text{data, } \theta^{/A} \right) \sim \text{No} \left( \text{data, } \theta^{/A} \right).
$$
Gibbs sampling

Conditional probabilities can be used to sample $\mu, \Delta^{-1}, A$

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

$$\Delta^{-1}_{(t+1)} \mid \left( \text{data}, \theta^{/\Delta^{-1}}_{(t)} \right) \sim \text{InvWishart} \left( \text{data}, \theta^{/\Delta^{-1}}_{(t)} \right),$$

$$A_{(t+1)} \mid \left( \text{data}, \theta^{/A}_{(t)} \right) \sim \text{No} \left( \text{data}, \theta^{/A}_{(t)} \right).$$

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

Given samples \((\Delta_{(t)}^{-1}, A(t))^{T}_{t=1}\).
Posterior estimates

Given samples \( (\Delta_{(t)}^{-1}, A_{(t)})_{t=1}^{T} \).

The posterior mean of the d.r. space is

\[
\hat{B} = \frac{1}{T} \sum \Delta_{(t)}^{-1} A_{(t)},
\]
Posterior estimates

Given samples \((\Delta^{-1}_{(t)}, A(t))_{t=1}^T\).

The posterior mean of the d.r. space is

\[
\hat{B} = \frac{1}{T} \sum_{t=1}^T \Delta^{-1}_{(t)} A(t),
\]

the variance of the d.r. space is

\[
\hat{\sigma}^2 \hat{B} = \frac{1}{T} \sum_{t=1}^T \| \Delta^{-1}_{(t)} A(t) - \hat{B} \|^2.
\]
Semiparametric model

\[ Y_i = f(X_i) + \varepsilon_i = g(b_1'X_i, \ldots, b_d'X_i) + \varepsilon_i, \]

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

\[ Y_i = f(X_i) + \varepsilon_i = g(b'_1 X_i, \ldots, b'_d X_i) + \varepsilon_i, \]

\( \text{span } B \) is the dimension reduction (d.r.) subspace.

Assume marginal distribution \( \rho_X \) is concentrated on a manifold \( \mathcal{M} \subset \mathbb{R}^p \) of dimension \( d_M \ll p \).
Gradients and outer products

Given a smooth function $f$ the gradient is

$$\nabla f(X) = \left( \frac{\partial f(X)}{\partial X_1}, \ldots, \frac{\partial f(X)}{\partial X_p} \right)'.$$
Gradients and outer products

Given a smooth function $f$ the gradient is
\[ \nabla f(X) = \left( \frac{\partial f(X)}{\partial X_1}, \ldots, \frac{\partial f(X)}{\partial X_p} \right)'. \]

Define the gradient outer product matrix $\Gamma$

\[
\Gamma_{ij} = \int_X \frac{\partial f(X)}{\partial x_i} \frac{\partial f(X)}{\partial x_j} d\rho_X(X), \\
\Gamma = \mathbb{E}[\nabla f \otimes \nabla f].
\]
GOP captures the d.r. space

Suppose

\[ y = f(X) + \varepsilon = g(b_1' X, \ldots, b_d' X) + \varepsilon. \]
GOP captures the d.r. space

Suppose

\[ y = f(X) + \varepsilon = g(b'_1 X, ..., b'_d X) + \varepsilon. \]

For \( i = 1, \ldots, d \)

\[ \frac{\partial f(x)}{\partial v_i} = v'_i (\nabla f(X)) \neq 0 \Rightarrow b'_i \Gamma b_i \neq 0. \]

If \( w \perp b_i \) for all \( i \) then \( w' \Gamma w = 0. \)
GOP captures the d.r. space

Suppose

\[ y = f(X) + \varepsilon = g(b'_1 X, \ldots, b'_d X) + \varepsilon. \]

For \( i = 1, \ldots, d \)

\[ \frac{\partial f(x)}{\partial v_i} = v'_i (\nabla f(X)) \neq 0 \Rightarrow b'_i \Gamma b_i \neq 0. \]

If \( w \perp b_i \) for all \( i \) then \( w' \Gamma w = 0. \)

\( B = (b_1, \ldots, b_d) \) obtained by spectral decomposition of \( \Gamma. \)
Statistical interpretation

Linear case

\[ y = \beta'x + \varepsilon, \quad \varepsilon \overset{iid}{\sim} \text{No}(0, \sigma^2). \]

\[ \Omega = \text{cov}(\mathbb{E}[X|Y]), \quad \Sigma_X = \text{cov}(X), \quad \sigma^2_Y = \text{var}(Y). \]
Statistical interpretation

Linear case

\[ y = \beta'x + \varepsilon, \quad \varepsilon \text{ iid } \sim \text{No}(0, \sigma^2). \]

\[ \Omega = \text{cov}(\mathbb{E}[X|Y]), \quad \Sigma_X = \text{cov}(X), \quad \sigma^2_Y = \text{var}(Y). \]

\[ \Gamma = \sigma^2_Y \left(1 - \frac{\sigma^2}{\sigma^2_Y}\right)^2 \Sigma_X^{-1} \Omega \Sigma_X^{-1} \approx \sigma^2_Y \Sigma_X^{-1} \Omega \Sigma_X^{-1}. \]
Statistical interpretation

For smooth $f(x)$

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

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

Partition into sections and compute local quantities

\[ X = \bigcup_{i=1}^{I} \chi_i \]
Nonlinear case

Partition into sections and compute local quantities

\[ X = \bigcup_{i=1}^{I} \chi_i \]

\[ \Omega_i = \text{cov} \left( \mathbb{E}[X_{\chi_i} | Y_{\chi_i}] \right) \]
Nonlinear case

Partition into sections and compute local quantities

\[ X = \bigcup_{i=1}^{I} \chi_i \]

\[ \Omega_i = \text{cov} \left( \mathbb{E}[X_{\chi_i} \mid Y_{\chi_i}] \right) \]

\[ \Sigma_i = \text{cov} \left( X_{\chi_i} \right) \]
Nonlinear case

Partition into sections and compute local quantities

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

\[ \Omega_i = \text{cov} \left( \mathbb{E}[X_{\chi_i} | Y_{\chi_i}] \right) \]

\[ \Sigma_i = \text{cov} \left( X_{\chi_i} \right) \]

\[ \sigma^2_i = \text{var} \left( Y_{\chi_i} \right) \]
Nonlinear case

Partition into sections and compute local quantities

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

\[ \Omega_i = \text{cov} \left( \mathbb{E}[X_{\chi_i} | Y_{\chi_i}] \right) \]

\[ \Sigma_i = \text{cov} \left( X_{\chi_i} \right) \]

\[ \sigma_i^2 = \text{var} \left( Y_{\chi_i} \right) \]

\[ m_i = \rho_X \left( \chi_i \right). \]
Nonlinear case

Partition into sections and compute local quantities

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

\[ \Omega_i = \text{cov} (\mathbb{E}[X_{\chi_i} | Y_{\chi_i}]) \]

\[ \Sigma_i = \text{cov} (X_{\chi_i}) \]

\[ \sigma_i^2 = \text{var} (Y_{\chi_i}) \]

\[ m_i = \rho_X (\chi_i) . \]

\[ \Gamma \approx \sum_{i=1}^{I} m_i \sigma_i^2 \Sigma_i^{-1} \Omega_i \Sigma_i^{-1}. \]
Gauss-Markov graphical models

Give a multivariate normal distribution, $x \in \mathbb{R}^p$

$$ p(x) \propto \exp \left( - (x - \mu) C^{-1} (x - \mu)' \right). $$
Gauss-Markov graphical models

Give a multivariate normal distribution, $x \in \mathbb{R}^p$

$$p(x) \propto \exp \left( -(x - \mu)C^{-1}(x - \mu)' \right).$$

The precision matrix $P = C^{-1}$ is also the conditional independence matrix

$$P_{ij} = \text{dependence of } i \leftrightarrow j \mid \text{all other variables}. $$
Gauss-Markov graphical models

By construction $\Gamma$ is the covariance of a Gaussian processes.
Gauss-Markov graphical models

By construction $\Gamma$ is the covariance of a Gaussian processes.

$$\Gamma \approx \sigma_Y^2 \sum_X^{-1} \Omega \sum_X^{-1}.$$
Gauss-Markov graphical models

By construction $\Gamma$ is the covariance of a Gaussian processes.

$$\Gamma \approx \sigma^2_Y \sum^{-1}_X \Omega \sum^{-1}_X.$$ 

Define the (pseudo)-inverse $J = \text{inv}(\Gamma)$ as a conditional independence matrix

$$J_{ij} = \text{dependence of } i \leftrightarrow j \mid \text{all other variables and } y.$$
Gauss-Markov graphical models

By construction $\Gamma$ is the covariance of a Gaussian processes.

$$\Gamma \approx \sigma^2_Y \sum_X^{-1} \Omega \sum_X^{-1}.$$  

Define the (pseudo)-inverse $J = \text{inv}(\Gamma)$ as a conditional independence matrix

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

Define the partial correlation matrix $R$ with entries

$$R_{ij} = -\frac{J_{ij}}{\sqrt{J_{ii}J_{jj}}}.$$
Estimating the gradient

Taylor expansion

\[ y_i \approx f(x_i) \approx f(x_j) + \langle \nabla f(x_j), x_j - x_i \rangle \approx y_j + \langle \nabla f(x_j), x_j - x_i \rangle \text{ if } x_i \approx x_j. \]
Estimating the gradient

Taylor expansion

\[ y_i \approx f(x_i) \approx f(x_j) + \langle \nabla f(x_j), x_j - x_i \rangle \]
\[ \approx y_j + \langle \nabla f(x_j), x_j - x_i \rangle \quad \text{if } x_i \approx x_j. \]

Let \( \bar{f} \approx \nabla f \) the following should be small

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

\[ w_{ij} = \frac{1}{s^{p+2}} \exp \left( -\|x_i - x_j\|^2 / 2s^2 \right) \text{ enforces } x_i \approx x_j. \]
Estimating the gradient

The gradient estimate

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

where $\| \tilde{f} \|_K$ is a smoothness penalty.
Estimating the gradient

The gradient estimate

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

where \( \| \tilde{f} \|_K \) is a smoothness penalty.
Computational efficiency

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

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

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

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

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

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

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

$$\hat{\Gamma} = c_D K c_D'.$$
Bayesian estimates

There is a strong relation between regularization functionals such as

\[
V(\vec{f}) = \frac{1}{n^2} \sum_{i,j=1}^{n} w_{ij} \left( y_i - y_j - (\vec{f}(x_j))' (x_j - x_i) \right)^2 + \lambda \| \vec{f} \|_K^2
\]

and Bayesian inference.
Bayesian estimates

There is a strong relation between regularization functionals such as

\[ V(\vec{f}) = \frac{1}{n^2} \sum_{i,j=1}^{n} w_{ij} \left( y_i - y_j - (\vec{f}(x_j))'(x_j - x_i) \right)^2 + \lambda \| \vec{f} \|^2_K \]

and Bayesian inference.

\[ \text{Lik}(\text{data} \mid \vec{f}) \propto \exp \left( -\frac{1}{n^2} \sum_{i,j=1}^{n} w_{ij} \left( y_i - y_j - (\vec{f}(x_j))'(x_j - x_i) \right)^2 \right), \]
Bayesian estimates

There is a strong relation between regularization functionals such as

$$V(\tilde{f}) = \frac{1}{n^2} \sum_{i,j=1}^{n} w_{ij} \left( y_i - y_j - (\tilde{f}(x_j))'(x_j - x_i) \right)^2 + \lambda \|\tilde{f}\|^2_K$$

and Bayesian inference.

$$\text{Lik}(\text{data} \mid \tilde{f}) \propto \exp \left( -\frac{1}{n^2} \sum_{i,j=1}^{n} w_{ij} \left( y_i - y_j - (\tilde{f}(x_j))'(x_j - x_i) \right)^2 \right),$$

$$\pi(\tilde{f}) \propto \exp(-\lambda \|\tilde{f}\|^2),$$
Bayesian estimates

There is a strong relation between regularization functionals such as

\[ V(\vec{f}) = \frac{1}{n^2} \sum_{i,j=1}^{n} w_{ij} \left( y_i - y_j - (\vec{f}(x_j))'(x_j - x_i) \right)^2 + \lambda \| \vec{f} \|_K^2 \]

and Bayesian inference.

\[ \text{Lik}(\text{data} \mid \vec{f}) \propto \exp \left( -\frac{1}{n^2} \sum_{i,j=1}^{n} w_{ij} \left( y_i - y_j - (\vec{f}(x_j))'(x_j - x_i) \right)^2 \right), \]

\[ \pi(\vec{f}) \propto \exp(-\lambda \| \vec{f} \|_2^2), \]

\[ \text{Post}(\vec{f} \mid \text{data}) \propto \text{Lik}(\text{data} \mid \vec{f}) \times \pi(\vec{f}) = e^{-V(\vec{f})}. \]
Consistency

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

Assume the data is concentrated on a manifold $\mathcal{M} \subset \mathbb{R}^p$ with $\mathcal{M} \in \mathbb{R}^{d_{\mathcal{M}}}$ 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 $1 - \delta$*

$$
\| (d\varphi)^* \bar{f}_D - \nabla_{\mathcal{M}} f \|_{\rho_X} \leq C \log \left( \frac{1}{\delta} \right) n^{-1/d_{\mathcal{M}}}
$$

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

Theorem

Under mild conditions, with probability $1 - \delta$

$$\| \hat{J} - J \|_K \leq C \log \left( \frac{1}{\delta} \right) n^{-1/d_M}.$$
Observations on convergence

Typical theoretical results on convergence of graphical models is in terms of sparsity $s = \| \mathbf{J} \|_0$. 
Observations on convergence

Typical theoretical results on convergence of graphical models is in terms of sparsity $s = \| \hat{J} \|_0$.

Our results are in terms of rank of $\hat{J}$ reflective of intrinsic dimension.
Observations on convergence

Typical theoretical results on convergence of graphical models is in terms of sparsity \( s = \| \mathbf{J} \|_0 \).

Our results are in terms of rank of \( \mathbf{J} \) reflective of intrinsic dimension.

Our belief is the covariance structure is not necessarily sparse but is low rank.
Typical theoretical results on convergence of graphical models is in terms of sparsity $s = \| \hat{J} \|_0$.

Our results are in terms of rank of $\hat{J}$ reflective of intrinsic dimension.

Our belief is the covariance structure is not necessarily sparse but is low rank.
Probabilistic models for supervised dimension reduction

Results on data

Linear classification example

Data
Posterior mean of GOP
Posterior std of GOP
Probabilistic models for supervised dimension reduction

Results on data

Linear classification example

Top d.r.
Mixing

To assess the mixing of the chain we used a trace plot on the EDR space which in this case was the top spectral component

\[ a(t) = v'_(t) v(t+1). \]
Probabilistic models for supervised dimension reduction

Results on data

Linear classification example

Mixing

(f): Trace Plot

Value

Iteration

0.9

0.95

1
Data

100 samples are drawn from the following model

\[ X_1 = \theta_1, \quad X_2 = \theta_1 + \theta_2, \quad X_3 = \theta_3 + \theta_4, \quad X_4 = \theta_4, \quad X_5 = \theta_5 - \theta_4, \quad \theta \sim \text{No}(0,1) \]

and

\[ Y = X_1 + \frac{X_3 + X_5}{2} + \varepsilon, \quad \varepsilon \sim \text{No}(0,0.25). \]
Posterior mean GOP
Posterior std GOP
Posterior mean partial correlation
Posteriors std partial correlation
Inferred graphical model
dig...
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.
Probabilistic models for supervised dimension reduction

Results on data

High-dimensional data

BMI
Probabilistic models for supervised dimension reduction

Results on data

High-dimensional data

BAGL
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

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

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

\[
\frac{1}{d} \sum_{i=1}^{d} \| P_B \hat{b}_i \|^2 = \frac{1}{d} \sum_{i=1}^{d} \| (BB')\hat{b}_i \|^2
\]
Comparison of algorithms
Error as a function of $d$
Funding

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