

# **Non-parametric Bayesian simultaneous dimension reduction and regression on manifolds**

BY KAI MAO

*Department of Statistical Science*

*Duke University, Durham NC 27708-0251, U.S.A.*

km68@stat.duke.edu

QIANG WU

*Department of Statistical Science, Institute for Genome Sciences & Policy, Department of Computer Science*

*Duke University, Durham NC 27708-0251, U.S.A.*

qiang@stat.duke.edu

FENG LIANG

*Department of Statistics*

*University of Illinois at Urbana-Champaign, IL 61820, U.S.A.*

feng@stat.uiuc.edu

and

SAYAN MUKHERJEE

*Department of Statistical Science, Institute for Genome Sciences & Policy, Department of Computer Science*

*Duke University, Durham NC 27708-0251, U.S.A.*

sayan@stat.duke.edu

## **Abstract**

We formulate a Bayesian non-parametric model for simultaneous dimension reduction and regression as well as inference of graphical models. The proposed model holds for both the classical setting of Euclidean subspaces and the Riemannian setting where the marginal distribution is concentrated on a manifold. The method is designed for the high-dimensional setting where the number of variables far exceed the number of observations. A Markov chain Monte Carlo procedure for inference of model parameters is provided. Properties of the method and its utility are elucidated using simulations and real data.

*Some Key Words: Bayesian kernel model, dimension reduction, graphical models*

# 1 Introduction

We formulate a Bayesian non-parametric model for simultaneous dimension reduction with regression. The assumption in our approach is that relevant information in high-dimensional data generated by measuring thousands of variables lies on or near a low-dimensional manifold.

The statistical or mathematical framework we develop is based on ideas in manifold learning (Tenenbaum et al., 2000; Roweis and Saul, 2000; Belkin and Niyogi, 2003; Donoho and Grimes, 2003) as well as simultaneous dimension reduction and regression (Li, 1991; Cook and Weisberg, 1991; Fukumizu et al., 2005; Wu et al., 2007; Xia et al., 2002; Mukherjee and Zhou, 2006; Mukherjee and Wu, 2006; Mukherjee and Zhou, 2006). The Bayesian model and our approach to inference are based on Bayesian non-parametric kernel models (Pillai et al., 2007; Liang et al., 2006, 2007). The inference of a graphical model of the predictive variables is based on the ideas in Wu et al. (2007) that relate dimension reduction and Gauss-Markov graphical models. We will illustrate how our Bayesian model allows for formal inference of uncertainty in dimension reduction as well as inference the uncertainty of conditional dependencies in graphical models.

In Section 2 we state a statistical basis for dimension reduction and state the learning gradients approach developed in Mukherjee and Zhou (2006); Mukherjee and Wu (2006); Mukherjee and Zhou (2006). In Section 3 we develop a fully Bayesian non-parametric model for learning gradients and provide a Markov chain Monte Carlo procedure for inference of model parameters. In Section 4 we illustrate the method and address questions about mixing of the MCMC using simulated data and present analysis on real data. We close with a short discussion.

## 2 Dimension reduction and conditional independence based on gradients

The problem of regression can be summarized as estimating the regression function

$$f(x) = E(Y|X = x)$$

from data  $D = \{L_i = (Y_i, X_i)\}_{i=1}^n$  where  $X_i$  is a vector in a  $p$ -dimensional compact metric space  $X \in \mathcal{X} \subset \mathbb{R}^p$  and  $Y_i \in \mathbb{R}$  is a real valued output. Typically the data are drawn independently and identically from a joint distribution  $\rho(X, Y)$ . Even if  $p$  is large the response variable  $Y$  often depends on a few directions in  $\mathbb{R}^p$ ,

$$Y = f(X) + \varepsilon = g(b'_1 X, \dots, b'_d X) + \varepsilon, \tag{1}$$

where  $\varepsilon$  is noise and  $B = (b'_1, \dots, b'_d)$  is the effective dimension reduction (EDR) space. In this case dimension reduction becomes the central problem in finding an accurate regression model.

### 2.1 Euclidean setting

The central quantity of interest in this paper is the gradient outer product matrix. We first formulate its properties in the Euclidean  $p$ -dimensional ambient space. Assume the regression function  $f(x)$  is smooth, the gradient is given by  $\nabla f = \left( \frac{\partial f}{\partial x^1}, \dots, \frac{\partial f}{\partial x^p} \right)'$  and the the gradient outer product matrix  $\Gamma$  is a  $p \times p$  defined as

$$\Gamma = E_X [(\nabla f) (\nabla f)']. \tag{2}$$

The relation between the gradient outer product matrix and dimension reduction is illustrated by the following observation (Wu et al., 2007). Under the assumptions of the semi-parametric model (1), the gradient outer product matrix  $\Gamma$  is of rank at most  $d$  and if we denote by  $\{v_1, \dots, v_d\}$  the eigenvectors associated to the nonzero eigenvalues of  $\Gamma$  then following holds

$$\text{span}(B) = \text{span}(v_1, \dots, v_d)$$

Another construction of the gradient outer product matrix is in terms of the covariance of the inverse regression matrix  $\Omega_{X|Y} = \text{cov}[E(X|Y)]$  developed in Li (1991). In Wu et al. (2007) it was shown that for a linear regression function

$$\Gamma = \sigma_Y^2 \left(1 - \frac{\sigma_\varepsilon^2}{\sigma_Y^2}\right)^2 \Sigma_X^{-1} \Omega_{X|Y} \Sigma_X^{-1}$$

where  $\sigma_Y^2$  is the variance of the response variable,  $\sigma_\varepsilon^2$  is the variance of the error and  $\Sigma_X = \text{cov}(X)$  is the covariance matrix of the explanatory variables. A similar result holds for nonlinear functions that are smooth (Wu et al., 2007) in this case assuming there exists  $\mathcal{I}$  partitions  $R_i$  of the explanatory variables such that

$$f(x) = \beta'_i x + \varepsilon_i, \quad E\varepsilon_i = 0 \quad \text{for } x \in R_i, \quad (3)$$

then

$$\Gamma = \sum_{i=1}^{\mathcal{I}} \rho(R_i) \sigma_i^2 \left(1 - \frac{\sigma_{\varepsilon_i}^2}{\sigma_i^2}\right)^2 \Sigma_i^{-1} \Omega_i \Sigma_i^{-1}, \quad (4)$$

where  $\Sigma_i = \text{cov}(X \in R_i)$  is the covariance matrix of the explanatory variables in partition  $R_i$ ,  $\sigma_i^2 = \text{var}(Y|X \in R_i)$  is the variance of the response variables in partition  $R_i$ ,  $\Omega_i = \text{cov}[E(X \in R_i|Y)]$  is the the covariance of the inverse regression in partition  $R_i$ , and  $\rho(R_i)$  is the measure of partition  $R_i$  with respect to the marginal distribution.

## 2.2 Manifold setting

The above statements are formulated with respect to Euclidean geometry or linear subspaces. The local nature of the gradient allows for an interpretation of the gradient outer product in the manifold setting (Wu et al., 2007; Mukherjee et al., 2006). In the manifold setting, the support of marginal measure of the explanatory variables is concentrated on a manifold  $\mathcal{M}$  of dimension  $d_{\mathcal{M}} \ll p$ . We assume the existence of an isometric embedding  $\varphi : \mathcal{M} \rightarrow \mathbb{R}^p$  and the observed explanatory variables are the image of points drawn from a distribution concentrated on the manifold,  $x_i = \varphi(q_i)$  where  $(q_i)_{i=1}^n$  are concentrated on the manifold. In this case although global statistics such as  $\Omega_{X|Y}$  are not meaningful the gradient outer product matrix on the manifold can be defined

$$\Gamma_{\mathcal{M}} = E [(\nabla_{\mathcal{M}} f)(\nabla_{\mathcal{M}} f)']$$

in terms of the gradient on the manifold  $\nabla_{\mathcal{M}} f$ . This matrix is a  $d_{\mathcal{M}} \times d_{\mathcal{M}}$  and is the analog of the gradient outer product matrix  $\Gamma$  in the ambient space.

We are neither given the manifold  $\mathcal{M}$  nor the coordinates on the manifold so we cannot compute  $\Gamma_{\mathcal{M}}$ . However, the properties of  $\Gamma_{\mathcal{M}}$  can be understood in terms of  $\Gamma$  and the following relation was derived in Wu et al. (2007)

$$\Gamma = E [(d\varphi(\nabla_{\mathcal{M}}f)) (d\varphi(\nabla_{\mathcal{M}}f))'] .$$

This is due to the following relation from Mukherjee et al. (2006)

$$(d\varphi)^* \vec{f}_D \longrightarrow \nabla_{\mathcal{M}}f \quad \text{as} \quad n \rightarrow \infty,$$

where  $\vec{f}_D$  is a consistent estimator of the gradient in the ambient space and  $(d\varphi)^*$  is the dual of  $d\varphi$ .

The result of these observations is that the EDR directions can be recovered also in the manifold setting and linear projections can be effective for nonlinear manifolds as long as the gradient outer product matrix is low rank.

### 2.3 Conditional independence

The theory of Gauss-Markov graphs (Speed and Kiiveri, 1986; Lauritzen, 1996) was developed for multivariate Gaussian densities

$$p(x) \propto \exp \left( -\frac{1}{2} x^T J X + h^T x \right),$$

where the covariance is  $J^{-1}$  and the mean is  $\mu = J^{-1}h$ . The result of the theory is that the precision matrix  $J$ , given by  $J = \Sigma_X^{-1}$ , provides a measurement of conditional independence. The meaning of this dependence is highlighted by the partial correlation matrix  $R_X$  where each element  $R_{ij}$  is a measure of dependence between variables  $i$  and  $j$  conditioned on all other variables  $S^{/ij}$  and  $i \neq j$

$$R_{ij} = \frac{\text{cov}(x_i, x_j | S^{/ij})}{\sqrt{\text{var}(x_i | S^{/ij})} \sqrt{\text{var}(x_j | S^{/ij})}} = -\frac{J_{ij}}{\sqrt{J_{ii} J_{jj}}}.$$

Under the assumptions implied by equations (2) and (4) the gradient outer product matrix is a covariance matrix. So we can apply the theory of Gauss-Markov graphs to  $\Gamma$  and consider the matrix  $J_{\Gamma} = \Gamma^{-1}$ . The advantage of computing this matrix in the regression and classification framework is that it provides an estimate of the conditional dependence of the explanatory variables with respect to variation of the response variable. The modeling assumption of our construction is that the matrix  $J_{\Gamma}$  is sparse with respect to the factors or directions  $(b'_1, \dots, b'_d)$  rather than the  $p$  explanatory variables. Under this assumption we use pseudo-inverses in order to construct the dependence graph based on the partial correlation  $R_{\Gamma}$ .

### 2.4 Relation to other methods

Dimension reduction based on spectral decomposition of the gradient outer product was related in Mukherjee et al. (2006); Wu et al. (2007) to other simultaneous dimension reduction and regression methods such as Sliced inverse regression (SIR) (Li, 1991), Principal Hessian Directions (PHD) (Li, 1992), and Minimum variance estimation (MAVE) (Xia et al., 2002).

SIR and SAVE use moments of the inverse regression function to retrieve the predictive directions. This can be problematic when moments are degenerate as relevant directions are often lost. PHD estimates the predictive directions using the eigenvectors of the Hessian matrix. This method is constrained due to the requirement of strong distributional assumptions such as normality of the explanatory variables to accurately estimate the Hessian matrix. Another problem with PHD is that directions that are linearly correlated to the output variables are lost. Gradient based methods overcome these limitations.

MAVE uses the gradient outer product matrix implicitly and Outer Product of Gradients (OPG) (Xia et al., 2002) shares the same idea of retrieving predictive directions by eigen-decomposition of the gradient outer product matrix where the gradient is estimated by local polynomial fitting Fan and Gijbels (1996). When  $p > n$ , these two methods cannot be used directly due to overfitting and numerical instability. Kernel gradient learning overcomes by adding a regularization term in the gradient estimate.

### 3 Inference

Many approaches for the inference of gradients exist including various numerical derivative algorithms, local linear smoothing (Fan and Gijbels, 1996), and learning gradients by kernel models (Mukherjee and Zhou, 2006; Mukherjee and Wu, 2006). Our approach will be closely related to the penalized likelihood or regularization models developed in Mukherjee and Zhou (2006); Mukherjee and Wu (2006).

#### 3.1 The model

The starting point for our gradient estimation procedure is the first order Taylor series expansion of the regression function  $f(x)$  around a point  $u$

$$f(x) = f(u) + \nabla f(x)'(x - u) + \varepsilon_d, \quad (5)$$

where the deterministic error term  $\varepsilon_d = O(\|x - u\|^2)$  is a function of the distance between  $x$  and  $u$  and the model

$$y = f(x) + \varepsilon, \quad (6)$$

where  $\varepsilon$  models the stochastic noise. For simplicity we work with a fixed design model with  $(x_i)_{i=1}^n$  given (see (Liang et al., 2006) for the development of the random design setting of which this is a special case). Coupling equations (5) and (6) we can state the following model

$$y_i = \frac{1}{n} \left[ \sum_{j=1}^n f(x_j) + \vec{f}(x_i)'(x_i - x_j) + \varepsilon_{ij} \right], \quad \text{for } i = 1, \dots, n, \quad (7)$$

$$\varepsilon_{ij} = y_i - f(x_j) - \vec{f}(x_i)'(x_i - x_j), \quad \text{for } i, j = 1, \dots, n \quad (8)$$

where  $f$  models the regression function,  $\vec{f}$  models the gradient, and  $\varepsilon_{ij}$  has both stochastic and deterministic components varying monotonically as a function of the distance between two observations  $x_i$  and  $x_j$ . We will model  $\varepsilon_{ij}$  as a random quantity and use a very simple spatial model to specify the covariance structure. Specifically, we first define an association matrix with  $w_{ij} = \exp(-\|x_i - x_j\|^2/2s^2)$  with fixed bandwidth

parameter  $s$ . We then define  $\varepsilon_{ij} \stackrel{iid}{\sim} N(0, (\phi/w_{ij})^{-1})$  where  $\phi$  will be a random scale parameter. Define the vector  $\varepsilon_{i\bullet} = (\varepsilon_{i1}, \dots, \varepsilon_{in})'$ , a joint probability density function on this vector can be used to specify a likelihood function for the data. We specify the following model for  $\varepsilon_{i\bullet}$ .

$$p(\varepsilon_{i\bullet}) \propto \phi^{\frac{n}{2}} \exp \left\{ -\frac{\phi}{2} (\varepsilon'_{i\bullet} W_i \varepsilon_{i\bullet}) \right\}, \quad (9)$$

where the diagonal matrix  $W_i = \text{diag}(w_{i1}, \dots, w_{in})$ .

As in Mukherjee and Zhou (2006); Mukherjee and Wu (2006) we use a non-parametric kernel model which in the fixed design case results in the following representations for the regression function and gradient

$$f(x) = \sum_{i=1}^n \alpha_i K(x, x_i), \quad \vec{f}(x) = \sum_{i=1}^n \mathbf{c}_i K(x, x_i) \quad (10)$$

where  $\alpha = (\alpha_1, \dots, \alpha_n)' \in \mathbb{R}^n$ ,  $C = (\mathbf{c}_1, \dots, \mathbf{c}_n) \in \mathbb{R}^{p \times n}$ . Substituting the above representation in equation (8) results in the following parametrized model

$$y_i = \sum_{k=1}^n \alpha_k K(x_j, x_k) + \sum_{k=1}^n (\mathbf{c}'_k (x_i - x_j)) K(x_i, x_k) + \varepsilon_{ij}, \quad \text{for } i, j = 1, \dots, n. \quad (11)$$

We can rewrite the above in matrix notation where for the  $i$ -th observation

$$y_i \mathbf{1} = K\alpha + D_i C K_i + \varepsilon_{i\bullet},$$

where  $\mathbf{1}$  is the  $n \times 1$  vector of all 1's,  $K_i$  is the  $i$ -th column of the gram matrix  $K$  where  $K_{ij} = k(x_i, x_j)$ ,  $E$  is the  $n \times p$  data matrix and  $D_i = \mathbf{1}x'_i - E$ . This model has a huge number of parameters,  $C$  itself has  $n \times p$  parameters. Many of these parameters are strongly correlated and we can use spectral decompositions to greatly reduce the number of variables. For example the linearization imposed by the first order Taylor series expansion in (7) imposes the constraint that the gradient estimate must be in the span of differences between data points,  $M_X = (x_1 - x_n, x_2 - x_n, \dots, x_{n-1} - x_n) \in \mathbb{R}^{p \times n}$ . The rank of this matrix is  $d \leq \min((n-1), p)$  and the singular value decomposition yields  $M_X = V\Lambda_M U'$  where  $V$  and  $U$  are the left and right eigenvectors and  $\Lambda_M$  is a matrix of the singular values. For a fixed  $d^*$  corresponding to large singular values we select the corresponding left eigenvectors  $\tilde{V} = (v_1, \dots, v_{d^*})$  and define a new set of parameters  $\tilde{C} = \tilde{V}'C$  and define the matrix  $D_i = \tilde{D}_i \tilde{V}'$ . A spectral decomposition can also be applied to the gram matrix  $K$  resulting in  $K = F\Lambda_K F'$ . Note that  $K\alpha = F\beta$  where  $\beta = \Lambda_K F'\alpha$ . We can again select columns of  $F$  corresponding to the largest eigenvalues  $m$ . Given the above re-parametrization we have for the  $i$ -th observation

$$y_i \mathbf{1} = F\beta + \tilde{D}_i \tilde{C} K_i + \varepsilon_{i\bullet}.$$

Given the probability model for the error vector in (9), the likelihood of our model given observations  $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$  is

$$\text{Lik}(D|\phi, \beta, \tilde{C}) \propto \phi^{\frac{n^2}{2}} \exp \left\{ -\frac{\phi}{2} \sum_{i=1}^n \left( y_i \mathbf{1} - F\beta - \tilde{D}_i \tilde{C} K_i \right)' W_i \left( y_i \mathbf{1} - F\beta - \tilde{D}_i \tilde{C} K_i \right) \right\}, \quad (12)$$

where the diagonal matrix  $W_i = \text{diag}(w_{i1}, \dots, w_{in})$ .

The prior specification for the parameters  $(\phi, \beta, \tilde{C})$  are

$$\begin{aligned}\phi &\propto \frac{1}{\omega}, \\ \beta &\sim N(0, \Delta_\psi^{-1}) \text{ where } \Delta_\psi = \text{diag}(\psi_1, \dots, \psi_m) \text{ and } \psi_i \sim \text{Gamma}(a_\psi/2, b_\psi/2), \\ \tilde{C}_j &\sim N(0, \Delta_\varphi^{-1}) \text{ where } \Delta_\varphi = \text{diag}(\varphi_1, \dots, \varphi_{d^*}) \text{ and } \varphi_i \sim \text{Gamma}(a_\varphi/2, b_\varphi/2),\end{aligned}$$

where  $\tilde{C}_j$  is the  $j$ -th column of  $\tilde{C}$  and  $a_\psi, b_\psi, a_\varphi, b_\varphi, \omega$  are pre-specified hyper-parameters, and an improper prior for  $\phi$  is used.

### 3.2 Sampling from the posterior

A standard Gibbs sampler can be used to simulate the posterior density,  $\text{Post}(\phi, \beta, \tilde{C}|D)$ , due to the normal form of the likelihood and conjugacy properties of the prior specifications. The update steps of the Gibbs sampler given data  $D$  and initial values  $(\phi^{(0)}, \beta^{(0)}, \tilde{C}^{(0)})$  follow:

1. Update  $\Delta_\psi$ :  $\Delta_\psi^{(t+1)} = \text{diag}(\psi_1^{(t+1)}, \dots, \psi_m^{(t+1)})$  with

$$\psi_i^{(t+1)} | D, \phi^{(t)}, \beta^{(t)}, \tilde{C}^{(t)} \sim \text{Gamma}\left(\frac{a_\psi + 1}{2}, \frac{b_\psi + (\beta_i^{(t)})^2}{2}\right), \quad i = 1, \dots, m$$

where  $\beta_i^{(t)}$  is the  $i$ -th element of  $\beta^{(t)}$ ;

2. Update  $\Delta_\varphi$ :

$$\begin{aligned}\Delta_\varphi^{(t+1)} &= \text{diag}(\varphi_1^{(t+1)}, \dots, \varphi_{d^*}^{(t+1)}) \\ \varphi_i^{(t+1)} | D, \phi^{(t)}, \beta^{(t)}, \tilde{C}^{(t)} &\sim \text{Gamma}\left(\frac{a_\varphi + 1}{2}, \frac{b_\varphi + \sum_{j=1}^n (\tilde{C}_{ij}^{(t)})^2}{2}\right), \quad i = 1, \dots, d^*,\end{aligned}$$

where  $\tilde{C}_{ij}^{(t)}$  is the  $(i, j)$ -th element of  $\tilde{C}^{(t)}$ ;

3. Update  $\beta$ :

$$\beta^{(t+1)} | D, \tilde{C}^{(t)}, \Delta_\psi^{(t+1)}, \phi^{(t)} \sim N(\mu_\beta, \Sigma_\beta)$$

with

$$\begin{aligned}\Sigma_\beta &= \left( F' \left( \sum_{i=1}^n \phi^{(t)} W_i \right) F + \Delta_\psi^{(t+1)} \right)^{-1}, \\ \mu_\beta &= \phi^{(t)} \Sigma_\beta F' \sum_{i=1}^n W_i (y_i \mathbf{1} - \tilde{D}_i \tilde{C}^{(t)} K_i);\end{aligned}$$

4. Update  $\tilde{C} = (\tilde{C}_1, \dots, \tilde{C}_n)$ :

For  $\tilde{C}_j$  with  $j = 1, \dots, n$

$$\tilde{C}_j^{(t+1)} | D, \tilde{C}_{\setminus j}^{(t)}, \Delta_\psi^{(t+1)}, \phi^{(t)} \sim N(\mu_j, \Sigma_j),$$

where  $\tilde{C}_{\setminus j}^{(t)}$  is the matrix  $\tilde{C}^{(t)}$  with the  $j$ -th column removed.

$$\begin{aligned} b_{ij} &= y_i \mathbf{1} - F \beta^{(t+1)} - \tilde{D}_i \sum_{k \neq j} \tilde{C}_k^{(t)} K_{ik} \\ \Sigma_{j0} &= \left( \phi^{(t)} \sum_{i=1}^n K_{ij}^2 \tilde{D}_i' W_i \tilde{D}_i \right)^{-1}, \\ \mu_{j0} &= \phi^{(t)} \Sigma_j \sum_{i=1}^n K_{ij} \tilde{D}_i' W_i b_{ij} \\ \Sigma_j &= (\Sigma_{j0}^{-1} + \Delta_\phi^{(t+1)})^{-1}, \\ \mu_j &= \Sigma_j (\Sigma_{j0}^{-1} \mu_{j0}). \end{aligned}$$

5. Update  $\phi$ :

$$\phi^{(t+1)} | D, \tilde{C}^{(t+1)}, \beta^{(t+1)} \sim \text{Gamma}(a, b),$$

where

$$\begin{aligned} a &= \frac{n^2}{2}, \\ b &= \frac{1}{2} \left( \sum_{i=1}^n \left[ y_i \mathbf{1} - F \beta^{(t+1)} - \tilde{D}_i \tilde{C}^{(t+1)} K_i \right]' W_i \left[ y_i \mathbf{1} - F \beta^{(t+1)} - \tilde{D}_i \tilde{C}^{(t+1)} K_i \right] \right). \end{aligned}$$

Given draws  $\{\tilde{C}^{(t)}\}_{t=1}^T$  from the posterior we can compute  $\{C^{(t)}\}_{t=1}^T$  from the relation  $\tilde{C} = \tilde{V}' C$ . which allows use to compute a gradient outer product for each draw

$$\Gamma_D^{(t)} = C^{(t)} K K' (C^{(t)})'.$$

Given these instances of the gradient outer product we can compute the posterior mean gradient outer product matrix as well as its variance

$$\hat{\mu}_{\Gamma, D} = \frac{1}{T} \sum_{t=1}^T \Gamma_D^{(t)}, \quad \hat{\sigma}_{\Gamma, D} = \frac{1}{T} \sum_{t=1}^T \|\Gamma_D^{(t)} - \hat{\mu}_{\Gamma, D}\|^2.$$

A spectral decomposition of  $\hat{\mu}_{\Gamma, D}$  provides us with an estimate of the EDR space  $\hat{B}$  and a spectral decomposition of  $\hat{\sigma}_{\Gamma, D}$  provides us with an estimate of the uncertainty of stability in our estimate of the EDR. For inference of conditional independence we first compute the conditional independence and partial correlation matrices

$$J^{(t)} = (\Gamma_D^{(t)})^{-1}, \quad R_{ij}^{(t)} = -\frac{J_{ij}^{(t)}}{\sqrt{J_{ii}^{(t)} J_{jj}^{(t)}}},$$

using using a pseudo-inverse to compute  $(\Gamma_D^{(t)})^{-1}$ . The mean and variance of the posterior estimates of conditional independence as well as partial correlations can be computed as above using  $\{J^{(t)}\}_{t=1}^T$  and  $\{R^{(t)}\}_{t=1}^T$

$$\hat{\mu}_{J, D} = \frac{1}{T} \sum_{t=1}^T J^{(t)}, \quad \hat{\sigma}_{J, D} = \frac{1}{T} \sum_{t=1}^T \|J^{(t)} - \hat{\mu}_{J, D}\|^2,$$

$$\hat{\mu}_{R,D} = \frac{1}{T} \sum_{t=1}^T R^{(t)}, \quad \hat{\sigma}_{R,D} = \frac{1}{T} \sum_{t=1}^T \|R^{(t)} - \hat{\mu}_{R,D}\|^2.$$

### 3.3 Binary regression

The extension to classification problems where responses are  $y_i = 1/0$  using a probit link function is implemented using a set of latent variables  $Z = (z_1, \dots, z_n)'$  modeled as a truncated normal distribution with standard variance. In this setting  $\phi \equiv 1$  and the same Gibbs sampler with a step added to sample the latent variable can be used to sample from the posterior density,  $\text{Post}(\beta, \tilde{C}|D)$ . The update steps of the Gibbs sampler given data  $D$  and initial values  $(Z^{(0)}, \beta^{(0)}, \tilde{C}^{(0)})$  follow:

1. Update  $\Delta_\psi$ :  $\Delta_\psi^{(t+1)} = \text{diag}(\psi_1^{(t+1)}, \dots, \psi_m^{(t+1)})$  with

$$\psi_i^{(t+1)} | D, Z^{(t)}, \beta^{(t)}, \tilde{C}^{(t)} \sim \text{Gamma} \left( \frac{a_\psi + 1}{2}, \frac{b_\psi + (\beta_i^{(t)})^2}{2} \right), \quad i = 1, \dots, m$$

2. Update  $\Delta_\varphi$ :

$$\Delta_\varphi^{(t+1)} = \text{diag}(\varphi_1^{(t+1)}, \dots, \varphi_{d^*}^{(t+1)})$$

$$\varphi_i^{(t+1)} | D, Z^{(t)}, \beta^{(t)}, \tilde{C}^{(t)} \sim \text{Gamma} \left( \frac{a_\varphi + 1}{2}, \frac{b_\varphi + \sum_{j=1}^n (\tilde{C}_{ij}^{(t)})^2}{2} \right), \quad i = 1, \dots, d^*,$$

where  $\tilde{C}_{ij}^{(t)}$  is the  $(i, j)$ -th element of  $\tilde{C}^{(t)}$ ;

3. Update  $\beta$ :

$$\beta^{(t+1)} | D, \tilde{C}^{(t)}, \Delta_\psi^{(t+1)}, Z^{(t)} \sim N(\mu_\beta, \Sigma_\beta)$$

with

$$\Sigma_\beta = \left( F' \left( \sum_{i=1}^n W_i \right) F + \Delta_\psi^{(t+1)} \right)^{-1},$$

$$\mu_\beta = \Sigma_\beta F' \sum_{i=1}^n W_i (z_i^{(t)} \mathbf{1} - \tilde{D}_i \tilde{C}^{(t)} K_i);$$

4. Update  $\tilde{C} = (\tilde{C}_1, \dots, \tilde{C}_n)$ :

For  $\tilde{C}_j$  with  $j = 1, \dots, n$

$$\tilde{C}_j^{(t+1)} | D, \tilde{C}_{\setminus j}^{(t)}, \Delta_\psi^{(t+1)}, Z^{(t)} \sim N(\mu_j, \Sigma_j),$$

where  $\tilde{C}_{\setminus j}^{(t)}$  is the matrix  $\tilde{C}^{(t)}$  with the  $j$ -th column removed.

$$\begin{aligned}
b_{ij} &= z_i^{(t)} \mathbf{1} - F \beta^{(t)} - \tilde{D}_i \sum_{k \neq j} \tilde{C}_k^{(t)} K_{ik} \\
\Sigma_{j0} &= \left( \sum_{i=1}^n K_{ij}^2 \tilde{D}_i' W_i \tilde{D}_i \right)^{-1}, \\
\mu_{j0} &= \Sigma_j \sum_{i=1}^n K_{ij} \tilde{D}_i' W_i b_{ij} \\
\Sigma_j &= (\Sigma_{j0}^{-1} + \Delta_\varphi^{(t+1)})^{-1}, \\
\mu_j &= \Sigma_j (\Sigma_{j0}^{-1} \mu_{j0}).
\end{aligned}$$

5. Update  $Z$ :

For  $i = 1, \dots, n$

$$z_i^{(t+1)} | D, \beta^{(t+1)}, \tilde{C}^{(t+1)} \sim \begin{cases} N^+(\eta_i, 1) & \text{for } y_i = 1 \\ N^-(\eta_i, 1) & \text{for } y_i = 0 \end{cases}$$

where  $N^+$  and  $N^-$  denote the positive and negative truncated normal distributions and  $(\eta_1, \dots, \eta_n)' = F \beta^{(t+1)}$ .

### 3.4 Modeling comments

Many of the modeling decisions made in this paper were for simplicity and efficiency. In this paper we have fixed  $d^*$  and  $m$  rather than allow them to be random quantities. This was done to avoid having to use a reversible jump Markov chain Monte Carlo method. We also did not discuss how to decide how many of the EDR dimensions to keep in our analysis. In theory this can be done from posterior distribution of the eigenvalues of the gradient outer product matrix drawn by simulating from the posterior. We simply use the inflection point of the decay of the posterior mean gradient outer product to select the number of directions.

The greatest simplification with respect to modeling assumptions is the model we used for the covariance structure of the noise,  $\varepsilon_{ij}$ . We currently model the covariance as an independent random variable that is a function of the distance between two points,  $d(x_i, x_j)$ . A more natural approach would be to use a more sophisticated model of the covariance that would respect the fact that  $\varepsilon_{ij}$  and  $v_{ik}$  should covary for  $j \neq k$  again as a function of the distance between  $x_j$  and  $x_k$ . One can either use models on covariance matrices such as Wishart distributions or use ideas from Gauss-Markov graphical models such as given a covariance on all the variables,  $\Sigma_S$ , computing the covariance matrix marginalizing variable  $i$ ,  $\Sigma_{S \setminus i}$ . We implemented this second approach and found that the results were comparable.

Another approach to model the covariation is to use Wishart distributions. Consider  $\Sigma_\varepsilon$  as the following block diagonal matrix  $\Sigma_\varepsilon = \text{diag}(\Sigma_1, \dots, \Sigma_n)$  where  $\Sigma_i$  is an  $n \times n$  matrix. Denote  $W_i = \Sigma_i^{-1}$  as the precision matrix and place a Wishart prior on each  $W_i$  with a suitable scale matrix  $\Lambda_i$  that reflects the positive association between the error scale and the distance. One such scale matrix  $\Lambda_i$  can be constructed by diagonalizing the  $i$ -th row of the similarity matrix  $W$ . In this setting we allow correlations between  $\varepsilon_{ij}$

and  $\varepsilon_{ik}$  – those pairs with the share an explanatory variable. Based on this construction we can propose a full spatial model with

$$\Sigma_\varepsilon = \sigma_s^2 \rho(\phi_s, d_{(ij),(i'j')}) + \text{diag}(\sigma^2/w_{ij}),$$

where the first “spatial” term has a variance parameter  $\sigma_s^2$  and a specified covariogram with some parameter  $\phi_s$  and a suitable distance measure between data pairs, and the second “nugget” effect is the diagonal matrix in the model we currently use in practice.

## 4 Simulated and real data examples

We illustrate the ideas developed and the efficacy of the method on real and simulated data. We first focus on simulated data to ground our argument. We then illustrate the utility of our approach using real data.

### 4.1 Linear regression and dimension reduction

This simple simulation based on binary linear regression model fixes the modeling ideas we have proposed with respect to dimension reduction.

The following data set was used in Mukherjee and Wu (2006). Data was generated by draws from the following two classes of samples:

$$\begin{aligned} X_{j=1,\dots,10}|y=0 &\sim N(1.5, 1), & X_{j=41,\dots,50}|y=1 &\sim N(1.5, 1), \\ X_{j=11,\dots,20}|y=0 &\sim N(-3, 1), & X_{j=51,\dots,60}|y=1 &\sim N(-3, 1), \\ X_{j=21,\dots,80}|y=0 &\sim N(0, 0.1), & X_{j=1,\dots,40,61,\dots,80}|y=1 &\sim N(0, 0.1), \end{aligned}$$

where  $X_j$  is the  $j$ -th coordinate of the 80 dimension random vector  $X$ .

Twenty samples were drawn from each class for the analysis and the data matrix is displayed in Figure 1(a). The posterior mean of the RKHS norm for each of the eighty components is displayed in Figure 1(b) and the expected dimensions (1 to 20 and 41 to 60) have large norms. The posterior mean gradient outer product matrix as well as the uncertainty in the estimates for this are displayed in Figures 1(c) and 1(d). The blocking structure reflects the expected covariance of the predictive variables. In this example there is one effective dimension reduction direction and an estimate of the posterior mean and standard deviation of this is plotted in Figure 1(e).

To illustrate mixing of the Markov chain proposed by our model we examined the mixing of the eigenvector corresponding to the largest eigenvalue of the gradient outer product from each draw from the chain,  $v_{(t)}$  is the eigenvector for the  $t$ -th draw. We examined trace plots of these eigenvectors,  $a_{(t)} = v'_{(t)}v_{(t+1)}$ . Where the scalar value  $a_{(t)}$  is the projection of the previous eigenvector drawn onto the current direction. Figure 1(f) suggests that the chain is mixing and seems to be convergent.

### 4.2 Linear regression and graphical models

A simple linear regression model is used to illustrate inference of conditional dependencies of explanatory variables relevant to prediction.

The explanatory variables are correspond to a five dimension random vector drawn from the following model

$$X_1 = \theta_1, X_2 = \theta_1 + \theta_2, X_3 = \theta_3 + \theta_4, X_4 = \theta_4, X_5 = \theta_5 - \theta_4,$$

where  $\theta \sim N(0, 1)$ . The regression model is

$$Y = X_1 + \frac{X_3 + X_5}{2} + \varepsilon,$$

where  $\varepsilon \sim N(0, 0.25)$ .

One hundred samples were drawn from this model and we estimated to mean and standard deviation of the gradient outer product matrix, see Figure 2(b). The partial correlation matrix and its standard deviation are also displayed in Figure 2(b). The inference consistent with the estimate of the partial correlation structure is that  $X_1, X_3, X_5$  are negatively correlated with respect to variation in the response and  $X_2$  and  $X_4$  are not correlated with respect to variation of the response. This relation is displayed in the graphical model in Figure 2(a) in addition to the graphical model inferred based on the partial correlations corresponding to the covariance of the explanatory variables alone.

### 4.3 Digits analysis

The MNIST digits data (<http://yann.lecun.com/exdb/mnist/>) is commonly used in the machine learning literature to compare algorithms for classification and dimension reduction. The data set consists of 60,000 images of handwritten digits 0, 1,  $\dots$ , 9 where each image is considered as a vector of  $28 \times 28 = 784$  gray-scale pixel intensities. The utility of the digits data is that the effective dimension reduction directions have a visually intuitive interpretation.

For the digits data the following pairs were the most difficult to classify 3 versus 8, 2 versus 7, 4 versus 9. We examined two classification problems 3 versus 8 and 5 versus 8. For both classification problems we found that almost all of the predictive information was contained in the eigenvector corresponding to the top eigenvalue. The vector can be thought of as a  $28 \times 28$  image of what is different between 3 and 8 or 5 and 8 respectively. In Figure 4 we display these images for 3 vs 8 and 5 vs 8, left and right upper panels respectively. These images were computed by applying our model to random draws of 200 samples for the two classification problems and computing posterior estimates of the top eigenvector. We see in the left panel that the upper and lower left regions are the pixels that differentiate a 3 from an 8. Similarly, for the 5 versus 8 example the diagonal from the lower left to the upper right differentiates these digits.

### 4.4 Inference of graphical models for cancer progression

The last example is to illustrate the utility of our model in a practical problem in cancer genetics, modeling tumorigenesis. Genetic models of cancer progression are of great interest to better understand the initiation of cancer as well as the progression of disease into metastatic states. In Edelman et al. (2008) a models of tumor progression in prostate cancer as well as melanoma were developed. One fundamental idea in this paper were that the explanatory variables were summary statistics that assayed the differential enrichment of a priori defined sets of genes in individual samples (Edelman et al., 2006, 2008). These a priori

defined gene sets correspond to genes known to be in signalling pathways or have functional relations. The other fundamental idea is an inference of the interaction between pathways as the disease progresses across stages.

A variation of the analysis in Edelman et al. (2008) is developed in this section. The data consists of 22 benign prostate samples and 32 malignant prostate samples (Tomlins et al., 2007). For each sample we compute the enrichment with respect to 522 candidate gene sets or pathways (Mukherjee and Wu, 2006). Each sample corresponds to a 522 dimensional vector of pathway enrichment. We applied our model to this data set and inferred a mean posterior conditional independence matrix as well as the uncertainty in the estimates of these elements. For visualization purposes we focused on the 16 pathways most relevant with respect to predicting progression, the 15 pathways corresponding to coordinates with the largest RKHS norm. For these pathways we plot the conditional independence matrix and the variance of the elements in the matrix in Figure 5. Red edges correspond to positive partial correlations and blue for negative. The width of the edges correspond to the degree of uncertainty, edges we are more sure of are thicker. This graph offers a great deal of interesting biology to explore some of which is known, see Edelman et al. (2008) for more details. One particularly interesting aspect of the inferred network is the interaction between the ErbB (ERBB) or epidermal growth factor signalling pathway and the the mitogen-activated protein kinase (MAPK) pathway.

## 5 Discussion

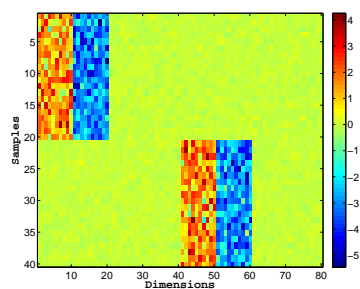
We propose a Bayesian non-parametric model for simultaneous dimension reduction and regression as well as inference of graphical models. This approach applies to the classical Euclidean setting as well as nonlinear manifolds. We illustrate the efficacy and utility of this model on real and simulated data. An implication of our model is that there are fascinating connections between spatial statistics and manifold and nonlinear dimension reduction methods that should be of great interest to the statistics community.

## References

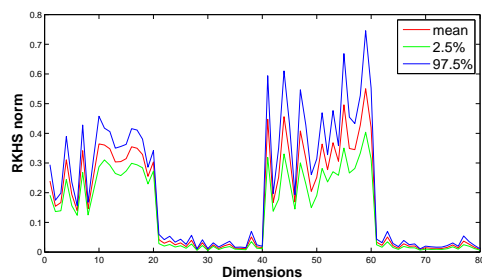
- Belkin, M. and P. Niyogi (2003). Laplacian Eigenmaps for Dimensionality Reduction and Data representation. *Neural Computation* 15(6), 1373–1396.
- Cook, R. and S. Weisberg (1991). Discussion of "sliced inverse regression for dimension reduction". *J. Amer. Statist. Assoc.* 86, 328–332.
- Donoho, D. and C. Grimes (2003). Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences* 100, 5591–5596.
- Edelman, E., J. Guinney, J.-T. Chi, P. Febbo, and S. Mukherjee (2008). Modeling cancer progression via pathway dependencies. *PLoS Comput. Biol.* 4(2), e28.
- Edelman, E., J. Guinney, A. Porello, B. Balakumaran, A. Bild, P. Febbo, and S. Mukherjee (2006). Analysis

- of sample set enrichment scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles. *Bioinformatics*, to appear.
- Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and its Applications*. Chapman and Hall, London.
- Fukumizu, K., F. Bach, and M. Jordan (2005). Dimensionality reduction in supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research* 5, 73–99.
- Hanahan, D. and R. Weinber (2000). The hallmarks of cancer. *Cell* 100, 57–70.
- Lauritzen, S. (1996). *Graphical Models*. Oxford: Clarendo Press.
- Li, K. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* 86, 316–342.
- Li, K. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein’s lemma. *Ann. Statist.* 97, 1025–1039.
- Liang, F., S. Mukherjee, M. Liao, and M. West (2006). Nonparametric Bayesian kernel models. Technical Report 07-10, ISDS Discussion Paper, Duke University.
- Liang, F., S. Mukherjee, and M. West (2007). Understanding the use of unlabelled data in predictive modeling. *Statistical Science* 22(2), 189–205.
- Mukherjee, S. and Q. Wu (2006). Estimation of gradients and coordinate covariation in classification. *J. Mach. Learn. Res.* 7, 2481–2514.
- Mukherjee, S. and D. Zhou (2006). Learning coordinate covariances via gradients. *J. Mach. Learn. Res.* 7, 519–549.
- Mukherjee, S., D.-X. Zhou, and Q. Wu (2006). Learning gradients and feature selection on manifolds. Technical Report 06-20, ISDS, Duke Univ.
- Pillai, N., Q. Wu, F. Liang, S. Mukherjee, and R. Wolpert (2007). Characterizing the function space for Bayesian kernel models. *J. Mach. Learn. Res.* 8, 1769–1797.
- Roweis, S. and L. Saul (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* 290, 2323–2326.
- Speed, T. and H. Kiiveri (1986). Gaussian Markov distributions over finite graphs. *Ann. Statist.* 14, 138–150.
- Tenenbaum, J., V. de Silva, and J. Langford (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* 290, 2319–2323.
- Tomlins, S., R. Mehra, D. Rhodes, X. Cao, L. Wang, S. Dhanasekaran, S. Kalyana-Sundaram, J. Wei, M. Rubin, K. Pienta, R. Shah, and A. Chinnaiyan (2007). Integrative molecular concept modeling of prostate cancer progression. *Nature Genetics* 39(1), 41–51.

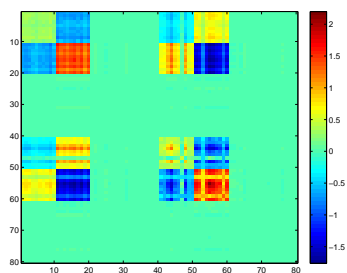
- Wu, Q., J. Guinney, M. Maggioni, and S. Mukherjee (2007). Learning gradients: predictive models that infer geometry and dependence. Technical Report 07-17, ISDS, Duke Univ.
- Wu, Q., F. Liang, and S. Mukherjee (2007). Regularized sliced inverse regression for kernel models. Technical Report 07-25, ISDS, Duke Univ.
- Xia, Y., H. Tong, W. Li, and L.-X. Zhu (2002). An adaptive estimation of dimension reduction space. *J. Roy. Statist. Soc. Ser. B* 64(3), 363–410.
- Yu, C., L. Jia, Y. Meng, J. Zhao, Y. Zhang, X. Qiu, Y. Xu, W. Wen, L. Yao, D. Fan, B. Jin, S. Chen, and A. Yang (2006). Selective proapoptotic activity of a secreted recombinant antibody/aif fusion protein in carcinomas overexpressing her2. *Gene Therapy* 13(4), 313–20.



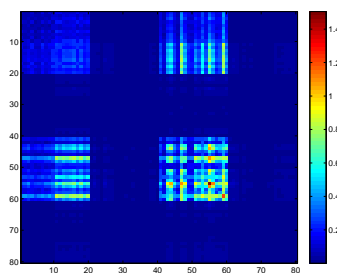
(a) Data matrix



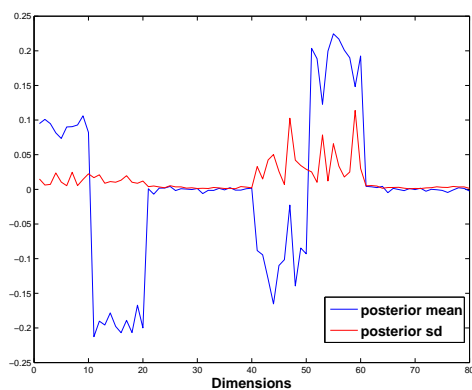
(b) RKHS norm



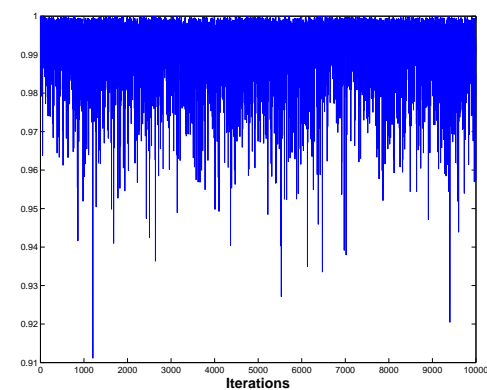
(c) Posterior mean of GOP



(d) Posterior standard deviation of GOP

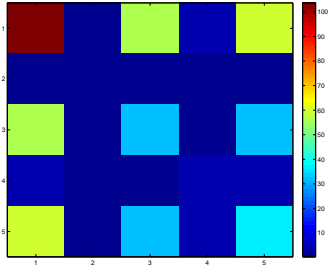


(e) Top edr direction

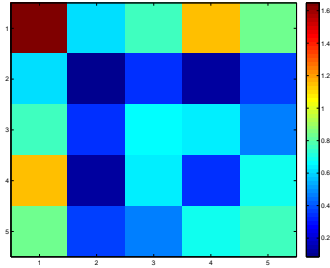


(f) Trace plot

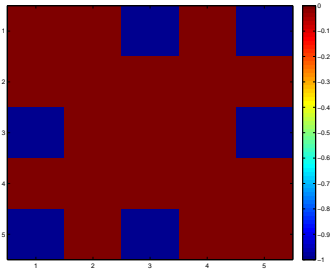
Figure 1: (a) The data matrix with rows corresponding to samples and columns to variables (dimensions); (b) The posterior mean and 2.5% and 97.5% quantiles of the RKHS norm; (c) The posterior mean of the gradient outer product matrix; (d) The posterior standard deviation of the gradient outer product matrix; (e) The posterior mean and posterior standard deviation of the top eigenvector of the gradient outer product matrix – the top edr direction; (f) The trace plot for the inner product of two consecutive draws of the top eigenvector.



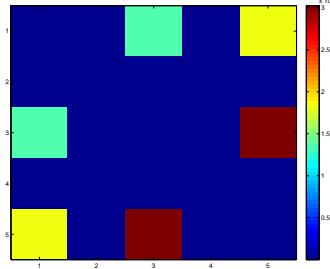
(a) Posterior mean of GOP



(b) Posterior standard deviation of GOP



(c) Posterior mean of partial correlation



(d) Posterior standard deviation

Figure 2: (a) and (b) are the posterior mean and standard deviation for the GOP, respectively; (c) and (d) are the posterior mean and standard deviation for the partial correlation matrix, respectively.

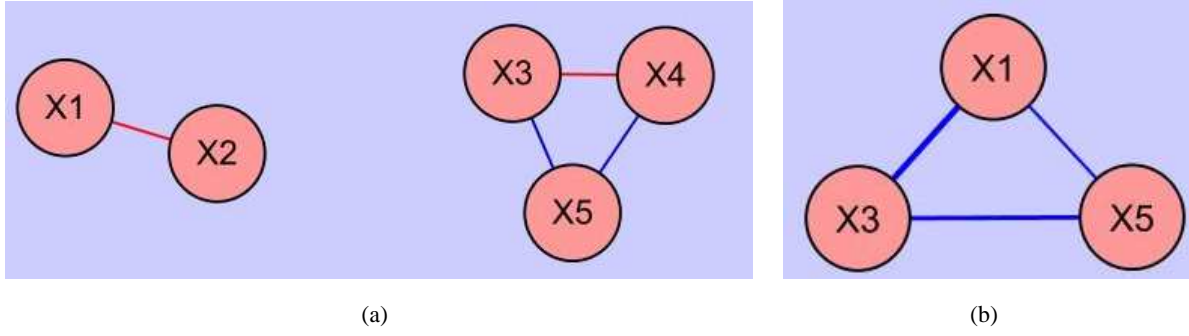
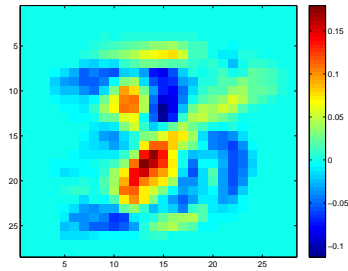
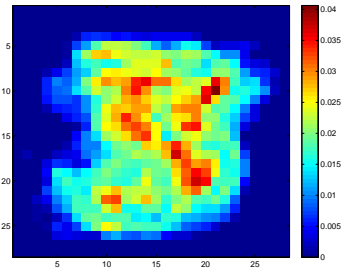


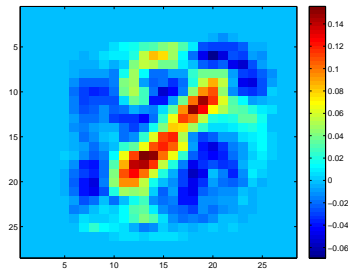
Figure 3: Graphical models inferred from the (a) the gradient outer product matrix and (b) the covariance matrix of the explanatory variables. Each node represents a variable and each edge indicates conditional dependence. The distance of the edge is inversely proportional to the amount of dependence, the thickness of the edge is proportional to the certainty of the inference and blue edges are negative while red edges are positive.



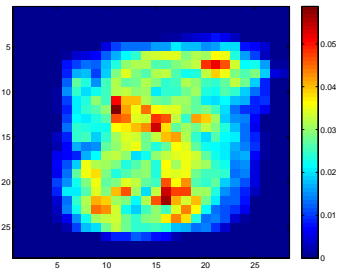
(a) Posterior mean of 3 vs 8



(b) Posterior standard deviation



(c) Posterior mean of 5 vs 8



(d) Posterior standard deviation

Figure 4: (a) The posterior mean of the top feature for 3 versus 8, shown in a  $28 \times 28$  pixel format. (b) The standard deviation of the top top feature. (a) The posterior mean of the top feature for 5 versus 8, shown in a  $28 \times 28$  pixel format. (b) The standard deviation of the top top feature.

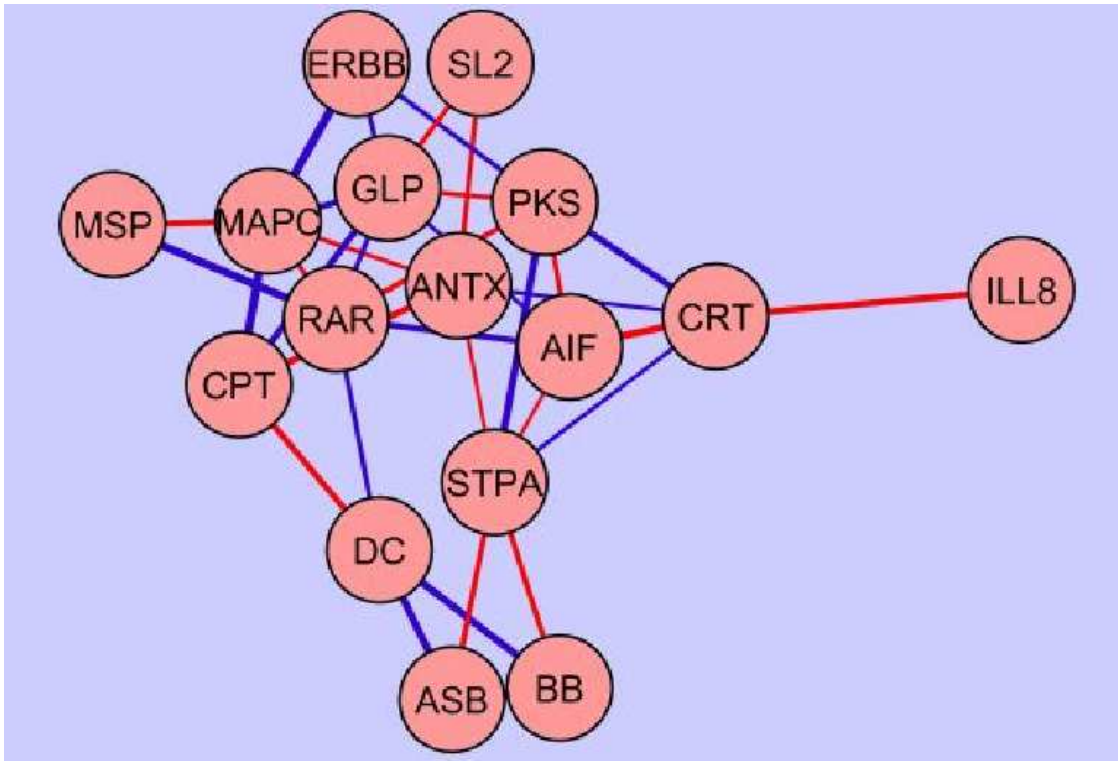


Figure 5: The association graph for the progression of prostate cancer from benign to malignant based on the inferred partial correlation. Red edges correspond to positive partial correlations and blue for negative. The width of the edges correspond to the degree of uncertainty, edges we are more sure of are thicker.