Mach Learn (2011) 83: 265–287
DOI 10.1007/s10994-010-5217-4

Estimating variable structure and dependence in multitask learning via gradients

Justin Guinney · Qiang Wu · Sayan Mukherjee

Received: 6 November 2008 / Revised: 26 March 2010 / Accepted: 30 August 2010 /
Published online: 23 October 2010

Abstract We consider the problem of hierarchical or multitask modeling where we simultaneously learn the regression function and the underlying geometry and dependence between variables. We demonstrate how the gradients of the multiple related regression functions over the tasks allow for dimension reduction and inference of dependencies across tasks jointly and for each task individually. We provide Tikhonov regularization algorithms for both classification and regression that are efficient and robust for high-dimensional data, and a mechanism for incorporating a priori knowledge of task (dis)similarity into this framework. The utility of this method is illustrated on simulated and real data.

Keywords Multitask learning · Dimension reduction · Covariance estimation · Inverse regression · Graphical models

1 Introduction

The problem of dimension reduction in the context of regression models is of fundamental interest in the physical and biological sciences and has a storied history (Fisher 1922; Hotelling 1933; Cook 2007). For much of biological and psychometric data, regression

Editor: Tony Jebara.

J. Guinney
Sage Bionetworks, Seatle, WA 98109, USA
e-mail: justin.guinney@sagebase.org

Q. Wu
Department of Mathematics, Michigan State University, East Lansing, MI 48824, USA
e-mail: wuqiang@math.msu.edu

S. Mukherjee (✉)
Departments of Statistical Science, Computer Science, and Mathematics, Duke University, Durham, NC 27708, USA
e-mail: sayan@stat.duke.edu

modeling needs to be extended to respect dependencies between observations based on temporal, structural, or general subgroup structure due to the way the data is collected. Classic examples of these regression models fall under the purview of Bayesian hierarchical models, hierarchical models with mixed effects, and, in the context of machine learning, multitask models. These models are closely related and can be restated as independent models connected by shared hyper-priors and seek to combine *similar* data for analysis under a single model, rather than each separately. In this paper we develop a method for simultaneous dimension reduction and inference of dependence structure for Bayesian hierarchical or multitask regression models. We first motivate the method with an important applied problem in whole genome analysis or expression analysis in cancer genetics.

Cancer like many complex traits is a heterogeneous disease requiring the accumulation of mutations in order to proceed through tumorigenesis, and an important problem is to predict and infer the mechanism for cancer progression. For any particular cancer the genetic heterogeneity of the disease is caused by two main sources: the stage or phenotypic variation of the disease and variability across individuals. A regression model can be built for each disease state to address the heterogeneity across the disease stage and one can select genes that are strongly correlated with progression. The problem with this stratification approach is the reduction of statistical power from the smaller sample sizes, and the fact that few genes or features individually may be predictive of phenotype. A natural paradigm to address this loss of power is to borrow strength across samples—multitask learning—and borrow strength across genes—simultaneous dimension reduction and regression. We return to this application in Subsect. 5.4.2 where we model the progression of prostate cancer.

This same problem arises in more classical artificial intelligence applications such as digit classification and text categorization since both documents and images of digits have hierarchical structure. In Subsect. 5.2 we illustrate this, demonstrating that inference of the distinction between a "5" and an "8" helps with discriminating a "3" from an "8." In this case we are borrowing strength across the digit images and learning linear combinations of pixels that are predictive subspaces.

The argument behind multitask learning is that pooling related samples (*tasks*) together in a joint analysis can improve predictive accuracy (Evgeniou et al. 2005; Caruana 1997; Ben-David and Schuller 2003; Obozinski et al. 2006; Argyriou et al. 2006; Ando and Zhang 2005; Jebara 2004), especially under conditions where there are few samples. Two interesting examples where this idea is used in therapeutics is to pool across stages in tumor progression (Edelman et al. 2008) or across drug treatments for HIV (Bickel et al. 2008).

Typically in this framework the idea of *data similarity* is traditionally considered in one of two distinct ways: sharing a similar discriminative function, or having variables or features that tend to covary. Our objective is to model these two properties of the data conjointly to uncover shared structure between tasks (dependent task variables) as well as the task specific structure (independent task variables).

We will show that this conjoint analysis across tasks as well as dependence structure results in more accurate predictive models than addressing each task individually. However, a point of emphasis of this paper is that the inference of the predictive geometry and dependencies between variables is of vital importance to interpret the results of our models. This point is stressed in Sect. 5 where we use the dimension reduction and graphical modeling approaches we develop to infer structure in genomic data, scientific documents, and images of digits. The central methodology for learning this structure will be the simultaneous inference of the regression (classification) function and its gradient.

2 Statistical basis for multitask gradient learning

In multitask learning we are given n_t observations for each of $t \in \{1, \ldots, N\}$ tasks where the observations are drawn from a task specific joint distribution function, $(X_{it}, Y_{it})_{i=1}^{n_t} \overset{i.i.d.}{\sim} \rho_t(X, Y)$. The input variables are a subspace of $\mathbb{R}^p$ and the output variable $Y_{it} \in \mathbb{R}$ for regression or $Y_{it} \in \{-1, 1\}$ for classification. The total number of samples is $n = \sum_t n_t$. We will denote the observations from the task t as D_t and D as the set of all the observations: $D = \{D_1, \ldots, D_N\}$. The objective in multitask modeling is to build a regression or classification function, $F_t(x)$, for each task t that has a baseline term $f_0(x)$ over all tasks and a task specific correction $f_t(x)$:

$$Y_t = F_t(X) + \varepsilon = f_0(X_t) + f_t(X_t) + \varepsilon, \quad \varepsilon \overset{iid}{\backsim} \mathrm{No}(0, \sigma^2). \tag{1}$$

The common as well as task specific regression functions are simultaneously learned using all D observations.

The key idea in multitask gradient learning is providing an estimate of the regression functions $\{f_0, (f_t)_{t=1}^N\}$ and their gradients $\{f_0, \nabla f_0, (f_t, \nabla f_t)_{t=1}^N\}$. The gradients provide information both for dimension reduction as well as the inference of a conditional independence graph for the input variables. The central assumption in our model is that each regression function depends on a few dimensions, d, in $\mathbb{R}^p$,

$$Y_t = F_t(X_t) + \varepsilon = g(b_{t1}^T X_t, \ldots, b_{td}^T X_t) + \varepsilon, \tag{2}$$

where ε is noise, X_t is the independent variable for the t-th task, Y_t is the dependent variable, and $B_t = (b_{t1}^T, \ldots, b_{td}^T)$ is the dimension reduction (DR) space for task t.

In a series of papers (Mukherjee and Zhou 2006; Mukherjee and Wu 2006; Mukherjee et al. 2010; Wu et al. 2010) a formal relation between dimension reduction and the conditional independence of predictive variables was developed. The central quantity in this relation is the gradient outer product matrix Γ, a $p \times p$ matrix with elements[1]

$$\Gamma_{ij} = \left\langle \frac{\partial F_t}{\partial x^i}, \frac{\partial F_t}{\partial x^j} \right\rangle_{L^2_{\rho_X}}, \tag{3}$$

where ρ_X is the marginal distribution of the explanatory variables in task t. Using the notation $a \otimes b = ab^T$ for $a, b \in \mathbb{R}^p$, we can write

$$\Gamma_t = \mathbb{E}(\nabla F_t \otimes \nabla F_t).$$

In the single task setting a spectral decomposition of $\Gamma = \Gamma_1$ can be used to compute relevant directions for dimension reduction due to the following observation (Wu et al. 2010, Lemma 1):

Proposition 1 *Under the assumptions of the semi-parametric model* (2) *and* $N = 1$ *task, the gradient outer product matrix* Γ *is of rank at most* d. *Denote by* $\{v_1, \ldots, v_d\}$ *the eigenvectors associated to the nonzero eigenvalues of* Γ *the following holds*

$$\mathrm{span}(B) = \mathrm{span}(v_1, \ldots, v_d).$$

[1]The gradient outer product matrix shares similarity with the Fisher information. The difference is that the Fisher information is an outer product of the gradient of the likelihood with respect to parameters. This characterizes a manifold in parameter space. In our case Γ characterizes a manifold on the data space.

The main argument for this result is the following observation for a vector $v \in \mathbb{R}^p$,

$$\frac{\partial F_1(x)}{\partial v} = v \cdot \nabla F_1$$

is identically zero if F_1 does not depends on v and is not zero if F_1 changes along the direction v.

It was further shown in Wu et al. (2010) that for the single task setting Γ has a natural interpretation as a composition of variances and covariances. For linear functions the following observation was made in Wu et al. (2010, Proposition 1):

Proposition 2 *Given the model*

$$y = \beta^T x + \varepsilon, \qquad \mathbb{E}\varepsilon = 0,$$

and the covariance of the inverse regression, $\Omega_{X|Y} = \mathrm{cov}_Y(\mathbb{E}_X(X \mid Y))$, *the variance of the output variable,* $\sigma_Y^2 = \mathrm{var}(Y)$, *and the covariance of the input variables,* $\Sigma_X = \mathrm{cov}(X)$, *the gradient outer product matrix is*

$$\Gamma = \sigma_Y^2 \left(1 - \frac{\sigma_\varepsilon^2}{\sigma_Y^2}\right)^2 \Sigma_X^{-1} \Omega_{X|Y} \Sigma_X^{-1}, \tag{4}$$

assuming that Σ_X *is full rank.*

This result was extended to any smooth nonlinear function by the following observation based on Wu et al. (2010, Corollary 2):

Proposition 3 *For a smooth function that is locally approximately linear over partitions* R_i *of the input space* $\mathcal{X}$

$$f(x) \approx \beta_i^T x + \varepsilon_i, \qquad \mathbb{E}\varepsilon_i = 0 \quad \textit{for } x \in R_i,$$

and $\mathcal{X} = \bigcup_{i=1}^{\mathcal{I}} R_i$, *we define the following local quantities*: *the covariance of the input variables* $\Sigma_i = \mathrm{cov}(X \in R_i)$, *the covariance of the inverse regression* $\Omega_i = \mathrm{cov}(\mathbb{E}(X \in R_i \mid Y))$, *the variance of the output variable* $\sigma_i^2 = \mathrm{var}(Y \mid X \in R_i)$. *Assuming that matrices* Σ_i *are full rank, the gradient outer product matrix can be computed in terms of these local quantities*

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

where $\rho_X(R_i)$ *is the measure of partition* R_i *with respect to the marginal distribution* ρ_X. *A probabilistic interpretation of* $\rho_X(R_i)$ *is the mixing proportion of partition* R_i.

The above propositions suggest that learning the common gradient outer product as well as the task specific gradient outer products

$$\Gamma^{(f_0)} = \mathbb{E}(\nabla f_0 \otimes \nabla f_0),$$
$$\Gamma^{(f_t)} = \mathbb{E}(\nabla f_t \otimes \nabla f_t),$$
$$\Gamma^{(F_t)} = \mathbb{E}(\nabla F_t \otimes \nabla F_t),$$

can be used to find the common and task specific subspaces $\{B_0, (B_t)_{t=1}^N\}$. We will illustrate the utility of these subspaces in Sect. 5.

For the single task case the above conclusions were shown to extend under weak conditions to the case where the input space is concentrated on a lower dimensional manifold $\mathcal{M}$, $d_{\mathcal{M}} \ll p$ (Mukherjee et al. 2010). The main idea of this paper is that for a variety of algorithms (Xia et al. 2002; Mukherjee and Zhou 2006; Mukherjee and Wu 2006) the gradient estimate in the ambient space converges to the gradient on the manifold. Under mild conditions (see Mukherjee et al. 2010) if the input variables are concentrated on a Riemannian manifold $\mathcal{M}$ of dimension $d_{\mathcal{M}}$ with an unknown isometric embedding $\varphi : \mathcal{M} \to \mathbb{R}^p$ then given a gradient estimate in the ambient space $\hat{\mathbf{f}}$ from n observations with probability $1 - \delta$

$$\|(\mathrm{d}\varphi)^* \hat{\mathbf{f}} - \nabla_{\mathcal{M}} f\|_{L^2_{\rho_{\mathcal{M}}}} \leq C \log\left(\frac{2}{\delta}\right) n^{-1/d_{\mathcal{M}}},$$

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

2.1 Comments

A variety of methods for simultaneous dimension reduction and regression to find directions that are informative with respect to predicting the response variable have been proposed. These methods can be summarized by three categories: (1) methods based on inverse regression (Li 1991; Cook and Weisberg 1991; Fukumizu et al. 2005; Wu et al. 2007), (2) methods based on gradients of the regression function (Xia et al. 2002; Mukherjee and Zhou 2006; Mukherjee and Wu 2006), (3) methods based on combining local classifiers (Hastie and Tibshirani 1996; Sugiyama 2007). In this paper we will build upon the approach outlined in Mukherjee and Zhou (2006), Mukherjee and Wu (2006). Mathematical and statistical relations between some of these approaches are developed in Wu et al. (2010).

3 Learning multitask gradients

3.1 Formulating the optimization problem

Given observations $D = \{D_1, \ldots, D_N\}$ over N tasks, our goal is to estimate the regression or classification functions $\{f_0(x), f_1(x), \ldots, f_T(x)\}$ and gradients $\{\nabla f_0(x), \nabla f_1(x), \ldots, \nabla f_T(x)\}$. These estimates can be used to obtain the gradient outer product matrix specific to each task, $\Gamma^{(f_t)}$, and the baseline gradient outer product for all tasks, $\Gamma^{(f_0)}$. We will formulate the optimization problem to estimate functions and their gradients both for classification and regression, although in this section we limit our discussion to the classification problem as the regression problem is conceptually similar.

For binary classification on a single task, $y_{it} \in \{-1, 1\}$, we first define a convex loss function $\phi(t)$ based on a link function such as the logistic link. Under this model F_t is real-valued and may be smooth. For example, in the case of the logistic function

$$\phi(y F_t(x)) = \log\left(1 + e^{-y F_t(x)}\right)$$

the classification function has a clear statistical interpretation (modeling the conditional probability $\mathbf{P}\mathrm{rob}(y|X)$ as a Bernoulli random variable)

$$\mathbf{P}\mathrm{rob}(y = \pm 1|x) = \frac{1}{1 + e^{-y F_t(x)}}.$$

In this case the classification function is

$$F_t(x) = \ln\left[\frac{\mathbf{P}\text{rob}(y=1|x)}{\mathbf{P}\text{rob}(y=-1|x)}\right]$$

and the gradient of f_t exists under very mild conditions on the underlying marginal distribution. In addition, for a rich enough class of functions $\mathcal{F}_t$ a Bayes optimal classifier exists

$$F_t = \arg\min_{F_t\in\mathcal{F}} \mathbb{E}_{\rho(X_t,Y_t)}[\phi(Y_t F_t(X_t))].$$

Assume that F_t is smooth then the first order Taylor series expansion is written as

$$F_t(x) \approx F_t(u) + \nabla F_t(x)\cdot(x-u), \quad \text{for } x \approx u. \tag{5}$$

If a function f and a vector valued function $\mathbf{f} = (f_1, \ldots, f_p)$ approximates F_t and its gradient well, then given the data $D_t = \{(x_{it}, y_{it})\}_{i=1}^{n_t}$, the expected error

$$\mathbb{E}(Y_t F_t \phi(X_t)) \approx \mathcal{E}^{\phi}_{D_t}(f,\mathbf{f}) = \frac{1}{n_t^2}\sum_{i,j=1}^{n_t} w_{i,j}^{(s)}\phi(y_i(f(x_j)+\mathbf{f}(x_i)\cdot(x_i-x_j)))$$

is small, where $w_{i,j}^{(s)}$ is a weight function with bandwidth s restricting the locality by $w_{i,j}^{s}\to 1$ as $\|x_i - x_j\|\to 0$. Estimates of the classification function and its gradient can be computed by minimizing the above functional with a reproducing kernel Hilbert space penalty added for regularization

$$(f_{D_t}, \mathbf{f}_{D_t}) = \arg\min_{(f,\mathbf{f})\in\mathcal{H}_K^{p+1}} \{\mathcal{E}^{\phi}_{D_t}(f,\mathbf{f}) + \lambda_1\|f\|_K^2 + \lambda_2\|\mathbf{f}\|_K^2\},$$

where f_{D_t} and $\mathbf{f}_{D_t}$ are estimates of F_t and ∇F_t, respectively, $\|\mathbf{f}\|_K^2 = \sum_{i=1}^{p}\|f_i\|_K^2$, and λ_1, λ_2 are regularization parameters. The bandwidth function imposes localization of the samples as required by the Taylor expansion, while the regularization parameters provide numeric stability to the classification and gradient functions estimates.

To extend from a single task to multiple tasks we begin with the hierarchical model in (1)

$$F_t(x) = f_0(x) + f_t(x),$$

and substitute this into (5)

$$F_t(x) \approx f_0(u) + \nabla f_0(x)\cdot(x-u) + f_t(u) + \nabla f_t(x)\cdot(x-u), \quad \text{for } x\approx u. \tag{6}$$

This results in an empirical error functional of the form

$$\mathcal{E}^{\phi}_{D_t}(f_0, f_t, \mathbf{f}_0, \mathbf{f}_t) = \frac{1}{n_t^2}\sum_{i,j=1}^{n_t} w_{i,j;t}\phi(y_{it}((f_0(x_{jt})+f_t(x_{jt}))+(\mathbf{f}_0(x_{jt})+\mathbf{f}_t(x_{jt}))\cdot(x_{it}-x_{jt}))) \tag{7}$$

where $\mathbf{f}_0$ and $\mathbf{f}_t$ are vector valued functions and model the gradient of f_0 and f_t respectively. Since we want to build a model jointly over all tasks and borrow strength across the entire data set $D = \{D_1, \ldots, D_N\}$ we use the average empirical error over the tasks as the error functional for the model

$$\mathcal{E}^{\phi}_{D}(f_0, \{f_t\}_{t=1}^N, \mathbf{f}_0, \{\mathbf{f}\}_{t=1}^N) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{E}^{\phi}_{D_t}(f_0, f_t, \mathbf{f}_0, \mathbf{f}_t).$$

The above functional is regularized by a RKHS penalty resulting in the following penalized error functional which we minimize to obtain our classification function and gradient estimates

$$(f_{D,0}, \{f_{D,t}\}_{t=1}^N, \mathbf{f}_{D,0}, \{\mathbf{f}_{D,t}\}_{t=1}^N)$$
$$= \underset{(f_t,\mathbf{f}_t)_{t=0}^T \in \mathcal{H}_K^{p+1}}{\arg\min} \left\{ \mathcal{E}_D^{\phi}(f_0, \{f_t\}_{t=1}^N, \mathbf{f}_0, \{\mathbf{f}\}_{t=1}^N) \right.$$
$$\left. + \frac{\lambda}{2}(\|f_0\|_K^2 + \|\mathbf{f}_0\|_K^2) + \frac{\mu}{2N}\sum_{t=1}^N (\|f_t\|_K^2 + \|\mathbf{f}_t\|_K^2) \right\}. \tag{8}$$

The regularization parameters μ and λ provide a priori assumptions on task similarity such that when $\frac{\mu}{\lambda}$ becomes small the model puts greater emphasis on the N tasks as independent functions whereas for a large ratio the common model dominates the optimization.

The above optimization problem can be considered as a combination of the gradient estimation ideas in Mukherjee and Zhou (2006), Mukherjee and Wu (2006) with the Tikhonov regularization formulation of multitask learning in Evgeniou and Pontil (2004). The behavior of this optimization problem with respect to the regularization is identical to that of Evgeniou and Pontil (2004). Note that there are identifiability issues with the model stated in (6) unless we assume a priori that $\nabla f_0 \perp \nabla f_t$, i.e. the task corrected gradient is in the null space of the common gradient. This assumption does not effect the model fit but it does effect the interpretation of the model.

3.2 Solving the optimization problem

A key insight in the Tikhonov regularization formulation of multitask learning in Evgeniou and Pontil (2004) was that the multitask problem can be restated as a single task optimization problem over all the data D with a very particular kernel. We will couple this observation with the single task gradient learning results in Mukherjee and Zhou (2006), Mukherjee and Wu (2006) to outline the classification and regression multitask gradient learning algorithms.

3.2.1 Regression

We begin with regression since the resulting optimization problem is simpler. In the regression setting we are given observations from the regression model, $y_{it} \approx F_t(x_{it})$, so we need only estimate the gradients. Assuming a Gaussian noise model and adapting the empirical error derived in (7), this results in the following least square task dependent loss functional

$$\mathcal{E}_{D_t}(\mathbf{f}_0, \mathbf{f}_t) = \frac{1}{n_t^2}\sum_{i,j=1}^{n_t} w_{i,j;t}(y_{it} - y_{jt} - (\mathbf{f}_0(x_{jt}) + \mathbf{f}_t(x_{jt})) \cdot (x_{it} - x_{jt}))^2.$$

Minimizing the regularized version of the above error functional leads to the following optimization problem

$$(\mathbf{f}_{D,0}, \mathbf{f}_{D,t}) = \underset{(\mathbf{f}_t)_{t=0}^T \in \mathcal{H}_K^p}{\arg\min} \left\{ \frac{1}{N}\sum_{t=1}^N \mathcal{E}_{D_t}(\mathbf{f}_0, \mathbf{f}_t) + \frac{\lambda}{2}\|\mathbf{f}_0\|_K^2 + \frac{\mu}{2N}\sum_{t=1}^N \|\mathbf{f}_t\|_K^2 \right\}. \tag{9}$$

The minimizer of this infinite dimensional optimization problem has the following finite dimensional representation

$$\mathbf{f}_0 = \sum_t \sum_i \alpha_{0,t,i} K(x_{it}, \cdot) \qquad \mathbf{f}_t = \sum_i c_{t,i} K(x_{it}, \cdot), \tag{10}$$

with the coefficients $\alpha_{0,t,i},\ c_{t,i} \in \mathbb{R}^p$. This is a result of the representer theorem (Wahba 1990) and was proven in the single task setting in Mukherjee and Zhou (2006, Theorem 5).

Substituting the above representation into (9) and setting the partial derivatives to 0 we obtain the following linear system which we solve to obtain the coefficients

$$\mu c_{t,j} + B_{t,j} \left(\sum_{s=1}^{N} \sum_{l=1}^{n_s} K(x_{ls}, x_{jt}) \alpha_{0,s,l} + \sum_{l=1}^{n_t} K(x_{lt}, x_{jt}) c_{t,l} \right) = Y_{t,j} \tag{11}$$

where

$$\begin{aligned} \alpha_{0,t,i} &= \frac{\mu}{N\lambda} c_{t,i}, \\ B_{t,j} &= \sum_{i=1}^{n_t} \frac{1}{n_t^2} w_{i,j;t} (x_{it} - x_{jt})(x_{it} - x_{jt})^T, \\ Y_{t,j} &= \sum_{i=1}^{n_t} \frac{1}{n_t^2} w_{i,j;t} (y_{it} - y_{jt})(x_{it} - x_{jt}). \end{aligned} \tag{12}$$

The linear system in (11) can be simplified based on ideas developed in Evgeniou and Pontil (2004). Denote the data set $\tilde{D} = \{(\tilde{x}_i, \tilde{y}_i)_{i=1,\ldots,n}\}$ as the samples arranged by task order and t_i as the task of the i-th sample. For example, $\tilde{x}_i = x_{i1}$ when $i \le n_1$. Denote by δ_{st} the Kronecker delta on tasks, $\delta_{st} = 1$ if $s = t$ and $\delta_{st} = 0$ otherwise. Define the kernel

$$\tilde{K}((x, s), (x', t)) = K(x, x') \left(\frac{\mu}{N\lambda} + \delta_{st} \right). \tag{13}$$

Define W_t as the $n_t \times n_t$ matrix with entries $W_t(i, j) = \frac{1}{n_t^2} w_{i,j;t}$ and $W = \operatorname{diag}(W_1, \ldots, W_N)$. Let $\tilde{B}$ be the $np \times np$ matrix composed by $N \times N$ blocks where the (s, t) block is an $n_s p \times n_t p$ sub-matrix with

$$\tilde{B}_{st} = 0 \text{ if } s \neq t \quad \text{and} \quad \tilde{B}_{st} = \operatorname{diag}(B_{t,1}, \ldots, B_{t,n_t}) \text{ if } s = t.$$

Let $\tilde{Y}_t = (Y_{t,1}^T, \ldots, Y_{t,n_t}^T)^T$ and $\tilde{Y} = (\tilde{Y}_1^T, \ldots, \tilde{Y}_N^T)^T$. We can rewrite the linear system (11) as

$$(\mu I_{np} + \tilde{B}(\tilde{K} \otimes I_p))c = \tilde{Y} \tag{14}$$

where I_p is the p-dimensional identity matrix and

$$c = (c_{1,1}^T, \ldots, c_{1,n_1}^T, c_{2,1}^T, \ldots, c_{2,n_2}^T, \ldots, c_{N,1}^T, \ldots, c_{N,n_N}^T)^T.$$

The solution to the linear system (14) results in gradient estimates that minimize the following single-task gradient learning problem

$$\mathbf{f}_{\tilde{D}}(x, t) = \arg\min \sum_{i,j=1}^{n} \tilde{W}_{i,j} (\tilde{y}_i - \tilde{y}_j - \mathbf{f}(\tilde{x}_i, t_i) \cdot (\tilde{x}_i - \tilde{x}_j))^2 + \mu \|\mathbf{f}\|_{\tilde{K}}^2,$$

where

$$\mathbf{f}_{D,0}(x) + \mathbf{f}_{D,t}(x) = \mathbf{f}_{\tilde{D}}(x,t),$$
$$\mathbf{f}_{D,0} = \frac{\mu}{N\lambda}\sum_{t=1}^{N}\mathbf{f}_{D,t}.$$

The linear system (14) is $np \times np$ which when p is large is not practically feasible. However, this linear system can be reduced to an $n^2 \times n^2$ linear system using the matrix reduction argument developed for single-task gradient learning in Mukherjee and Zhou (2006). For the single task setting this formulation is based on the observation that due to the Taylor expansion the gradient estimates will be in the span of the difference between data points

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

This matrix has rank at most $d \leq \min\{p, n-1\}$. A singular value decomposition of $M_{\mathbf{x}}$

$$M_{\mathbf{x}} = V\Sigma U^T = [V_1\ V_2\ \cdots\ V_n]\begin{bmatrix}\operatorname{diag}\{\sigma_1,\sigma_2,\ldots,\sigma_d\} & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}U_1^T\\ U_2^T\\ \vdots\\ U_m^T\end{bmatrix}$$

can be used to reparameterize the gradient estimate in terms of the left singular vectors V

$$\mathbf{f} = \sum_{i=1}^{n}\left\{\sum_{\ell=1}^{d}\tilde{c}_{i,\ell}V_\ell\right\}K(x_i,\cdot).$$

The number of parameters is at most n^2 ($n^2 \leq d \times n$)

$$\tilde{c} = \begin{bmatrix}\tilde{c}_{1,1} & \cdots & \tilde{c}_{1,n}\\ \vdots & \ddots & \vdots\\ \tilde{c}_{1,d} & \cdots & \tilde{c}_{d,n}\end{bmatrix}.$$

This parameterization is used when $p \gg n$ and results in a linear system of equations, see Mukherjee and Zhou (2006, Sect. 3.1).

3.2.2 *Classification*

The minimizer of the infinite dimensional optimization problem in (8) has the following finite dimensional representation

$$\begin{aligned} f_{D,0} &= \sum_{t=1}^{N}\sum_{i=1}^{n_t}\alpha_{0,t,i}K(x_{it},\cdot), \qquad f_{D,t} = \sum_{i=1}^{n_t}\alpha_{t,i}K(x_{it},\cdot),\\ \mathbf{f}_{D,0} &= \sum_{t=1}^{N}\sum_{i=1}^{n_t}c_{0,t,i}K(x_{it},\cdot), \qquad \mathbf{f}_{D,t} = \sum_{i=1}^{n_t}c_{t,i}K(x_{it},\cdot),\end{aligned} \tag{15}$$

with coefficients $\alpha_{0,t,i}, \alpha_{t,i} \in \mathbb{R}$ and $c_{0,t,i}, c_{t,i} \in \mathbb{R}^p$. This is a result of the representer theorem (Wahba 1990) and was proven in the single task setting in Mukherjee and Wu (2006).

Substituting the above representation into (8) and setting the partial derivatives to 0 results in a system of equations equivalent to the following

$$0 = \frac{1}{n_t^2}\sum_{i=1}^{n_t} w_{i,j;t}\phi'(\Upsilon_{i,j,t}) + \mu\alpha_{t,j},$$

$$0 = \frac{1}{n_t^2}\sum_{i=1}^{n_t} w_{i,j;t}\phi'(\Upsilon_{i,j,t})(x_{it} - x_{jt}) + \mu c_{t,j},$$

$$\alpha_{0,t,i} = \frac{\mu}{N\lambda}\alpha_{t,i},$$

$$c_{0,t,i} = \frac{\mu}{N\lambda}c_{t,i}$$

where

$$\Upsilon_{i,j,t} = y_{it}\left[\sum_{s=1}^{T}\sum_{l=1}^{n_s}\alpha_{0,s,l}K(x_{ls}, x_{jt}) + \sum_{l=1}^{n_t}\alpha_{t,l}K(x_{lt}, x_{jt}) + \left(\sum_{s=1}^{T}\sum_{l=1}^{n_s}c_{0,s,l}K(x_{ls}, x_{jt}) + \sum_{l=1}^{n_t}c_{t,l}K(x_{lt}, x_{jt})\right)\cdot(x_{it} - x_{jt})\right].$$

The above system of equations is a $n(p+1)\times n(p+1)$ system and can be solved using Newton's method. Note that when p is very large this is not practical.

To address this computational problem we use the idea of reducing the multitask optimization problem to a single-task optimization problem with a different kernel. This allows us to use the efficient solver developed in Mukherjee and Wu (2006). Denote by $\tilde{D} = \{(\tilde{x}_i, \tilde{y}_i)_{i=1,\ldots,n}\}$ the samples rearranged in task order and t_i is the task associated with sample $\tilde{x}_i$. Define the same kernel $\tilde{K}$ as in the regression setting. Let W_t be the $n_t \times n_t$ matrix with entries $W_t(i,j) = \frac{1}{n_t^2}w_{i,j;t}$ and $\tilde{W} = \operatorname{diag}(\tilde{W}_1, \ldots, \tilde{W}_T)$. Given the kernel $\tilde{K}$, weight matrix $\tilde{W}$, and data $\tilde{D}$ the following single-task optimization problem can be used to compute the coefficients for the multitask problem. The following is a direct result of the computations in Mukherjee and Wu (2006, Sect. 2).

Proposition 4 *Consider the following single-task learning gradient problem*

$$(g_{\tilde{D}}(x,t), \mathbf{g}_{\tilde{D}}(x,t)) = \arg\min_{g,\mathbf{g}\in\mathcal{H}_K^{p+1}}\{\mathcal{E}^{\phi}_{\tilde{D},\tilde{W}}(g,\mathbf{g}) + \mu\|g\|^2_{\tilde{K}} + \mu\nu\|\mathbf{g}\|^2_{\tilde{K}}\}. \tag{16}$$

We have

$$f_{D,0} + f_{D,t} = g_{\tilde{D}}(\cdot,t) \quad \textit{and} \quad \mathbf{f}_0 + \mathbf{f}_t = \mathbf{g}(\cdot,t). \tag{17}$$

A result of this equivalence is that

$$f_{D,0} = \frac{\mu}{N\lambda}\sum_{t=1}^{N} f_{D,t} \quad \text{and} \quad \mathbf{f}_{D,0} = \frac{\mu}{N\lambda}\sum_{t=1}^{N}\mathbf{f}_{D,t}. \tag{18}$$

The system can be reduced to solving a $n^2 \times n^2$ nonlinear system of equations by using the same singular value decomposition used in the classification setting Mukherjee and Wu (2006). This can be solved efficiently using Newton's method when n is small.

4 Dimension reduction, task similarity, and conditional dependencies

The fundamental quantities inferred in the MTGL framework are the $N+1$ gradient outer product matrices $\{\hat{\Gamma}_0, \hat{\Gamma}_1, \ldots, \hat{\Gamma}_N\}$. These matrices and the subspaces spanned by them will be used both for dimension reduction to infer predictive structure as well as learning graphical models to infer the predictive conditional dependencies in the data.

In Sect. 5 we illustrate how we can use these gradient outer product matrices to develop more accurate classifiers as well as better understand the predictive geometrical and dependence structure in the data. This analysis will require three ideas based on the gradient outer product matrices that we now introduce.

4.1 Dimension reduction

A primary purpose in estimating the $N+1$ gradient outer product matrices $\{\hat{\Gamma}_0, \hat{\Gamma}_1, \ldots, \hat{\Gamma}_N\}$ is to estimate the dimension reduction subspace that is common across tasks, $\hat{B}_0$, in addition to the effective dimension reduction subspace for each task $(\hat{B}_t)_{t=1}^N$. The dimension reduction subspace is the span of the gradient outer product $\hat{B}_i = \text{span}(\hat{\Gamma}_i)$ which is computed by spectral decomposition of the gradient outer product matrices. Given $\hat{\Gamma}_i$ the eigenvalues $\{\lambda_1^{(i)} \ldots \lambda_p^{(i)}\}$ and eigenvectors $\{v_1^{(i)}, \ldots, v_p^{(i)}\}$ are computed and the dimension reduction subspace is the span of the eigenvectors corresponding to eigenvalues above a threshold τ, $\hat{B}_i = \text{span}\{v_{k\in K}^{(i)}\}$ where $K = \{i \text{ such that } \lambda_k^i \geq \tau\}$. Alternatively, in the presence of a large "eigengap", τ may be selected by observing the spectral decay.

The immediate application of the dimension reduction subspaces is to project the data onto this space and use this lower dimensional representation for classification or clustering.

4.2 Inference of task similarity

In addition to using the dimension subspaces for better classification accuracy, the overlap between these spaces provides geometric information about the similarity or overlap between tasks. This can be of fundamental interest since a natural question to ask is how related are the tasks and what combinations of variables characterize task similarity. We therefore construct a measure of subspace similarity, or overlap, as a way of measuring the relatedness of linear subspaces. This score serves as a summary statistic of task similarity. We use the following measure:

Definition 1 Let $\{\Gamma_t \ldots \Gamma_T\}$ be the $p \times p$ symmetric (gradient outer product) matrices with entries in $\mathbb{R}$. Without loss of generality, we consider the case of two tasks, where B_1 defines a d-dimensional subspace of Γ_1, and B_2 a f-dimensional subspace of Γ_2. Also, let $\{v_1^{(1)}, \ldots, v_p^{(1)}\}$, $\{\lambda_1^{(1)} \ldots \lambda_p^{(1)}\}$ and $\{v_1^{(2)} \ldots v_p^{(2)}\}$, $\{\lambda_1^{(2)} \ldots \lambda_p^{(2)}\}$ be the eigenvectors, eigenvalues of Γ_1 and Γ_2, respectively. We define the subspace overlap score (SOS) of Γ_1 and Γ_2 as

$$SSscore_{1,2} = \frac{SS_{1\to 2}}{2} + \frac{SS_{2\to 1}}{2} = \frac{\sum_{i=1}^p \lambda_i^{(1)} \| P_\perp^{(2)} v_i^{(1)} \|_{L^2}}{2\sum_{i=1}^p \lambda_i^{(1)}} + \frac{\sum_{i=1}^p \lambda_i^{(2)} \| P_\perp^{(1)} v_i^{(2)} \|_{L^2}}{2\sum_{i=1}^p \lambda_i^{(2)}}. \quad (19)$$

We denote $P_\perp^{(1)}$ as the orthogonal projection matrix onto B_1, and determine the subspace using the top d eigenvectors, for specified $\epsilon \in [0, 1]$, such that

$$\frac{\sum_i^d \lambda_i^{(1)}}{\sum_i^p \lambda_i^{(1)}} < \epsilon.$$

Selection of an appropriate ϵ may be deduced analytically by detection of a significant "eigengap", or reflect instead a preferred threshold for total variance captured. Scores are in the interval [0, 1], and subspaces with complete symmetric overlap will have scores close to 1. In the case where $B_1 \subset B_2$, we would expect $SS_{1\to 2} \approx 1$ and it may therefore be useful to consider the two terms from (19) separately instead of averaged together. We propose this metric due to its intuitiveness—weighted projection of one linear subspace onto another—although we recognize other potential metrics may be suitable, e.g. Kullback-Liebler divergence between Gaussian covariance matrices.

4.3 Inference of graphical models and conditional dependencies

The theory of Gauss-Markov graphs (Speed and Kiiveri 1986; Lauritzen 1996) was developed for multivariate Gaussian densities to model conditional dependencies between variables

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

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. For example, J_{ij} is said to be conditionally independent given all other variables if $J_{ij} \approx 0$. 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 $i \neq j$

$$R_{ij} = \frac{\operatorname{cov}(X_i, X_j | S^{/ij})}{\sqrt{\operatorname{var}(X_i | S^{/ij})}\sqrt{\operatorname{var}(X_j | S^{/ij})}}.$$

The partial correlation matrix is typically computed from the precision matrix J

$$R_{ij} = -J_{ij}/\sqrt{J_{ii}J_{jj}}. \tag{20}$$

In the regression and classification framework inference of the conditional dependence between explanatory variables has limited information. A more useful measure would be the conditional dependence of the explanatory variables conditioned on variation in the response variable. Since the gradient outer product matrices provide estimates of the covariance of the explanatory variables conditioned on variation in the response variable over all tasks and for each task, the inverses of these matrices

$$\{\hat{J}_t = \hat{\Gamma}_t^{-1}\}_{t=1}^N,$$

provide evidence for the conditional dependence between explanatory variables conditioned on the response over all tasks and for each task. See Wu et al. (2010) for more details on the relation between inference of conditional dependencies and dimension reduction.

We will use the inferred conditional dependencies to construct sparse graphs that indicate the dependence structure on simulated and biological data.

5 Experiments

We apply the multitask gradient learning algorithm (MTGL) to simulated and real data for simultaneous classification and inference of the variable dependence structure. We explore

the effect of the regularization parameters in modulating the bias-variance trade-off (Hastie et al. 2001) and its impact on predictive performance. We also compute subspace overlap scores to aid in our interpretation of the structures we infer. We restrict our analysis to the classification setting using only several tasks, although the method generalizes to any number of tasks.

5.1 Simulation

We construct two tasks containing 40 samples each (20 in class **1**, 20 in class $-\mathbf{1}$) in a 120-dimensional space. We generate a data matrix for binary classification that contains features that are common to both tasks as well as features that are specific to each task. The matrix is initialized with background noise drawn from $\mathbf{No}(0, .2)$, defined as normal distribution $\mathbf{No}(\mu, \sigma^2)$. We then generate samples according to the following table, using the notation of x_i as the i-th sample and x^j as the j-th component:

1. task 1, class 1: $\{x_i\}_{i=1}^{20}$

$$x^j \sim \mathbf{No}(2,2),\ \text{for } j = 1,\ldots,10;\quad x^j \sim \mathbf{No}(2,.5),\ \text{for } j = 11,\ldots,20,$$
$$x^j \sim \mathbf{No}(2,2),\ \text{for } j = 61,\ldots,70;\ x^j \sim \mathbf{No}(2,.5),\ \text{for } j = 71,\ldots,80$$

2. task 1, class -1: $\{x_i\}_{i=21}^{40}$

$$x^j \sim \mathbf{No}(2,2),\ \text{for } j = 91,\ldots,100;\ x^j \sim \mathbf{No}(2,.5),\ \text{for } j = 101,\ldots,110,$$

3. task 2, class 1: $\{x_i\}_{i=41}^{60}$

$$x^j \sim \mathbf{No}(2,2),\ \text{for } j = 31,\ldots,40;\ x^j \sim \mathbf{No}(2,.5),\ \text{for } j = 41,\ldots,50,$$
$$x^j \sim \mathbf{No}(2,2),\ \text{for } j = 61,\ldots,70;\ x^j \sim \mathbf{No}(2,.5),\ \text{for } j = 71,\ldots,80$$

4. task 2, class -1: $\{x_i\}_{i=61}^{80}$

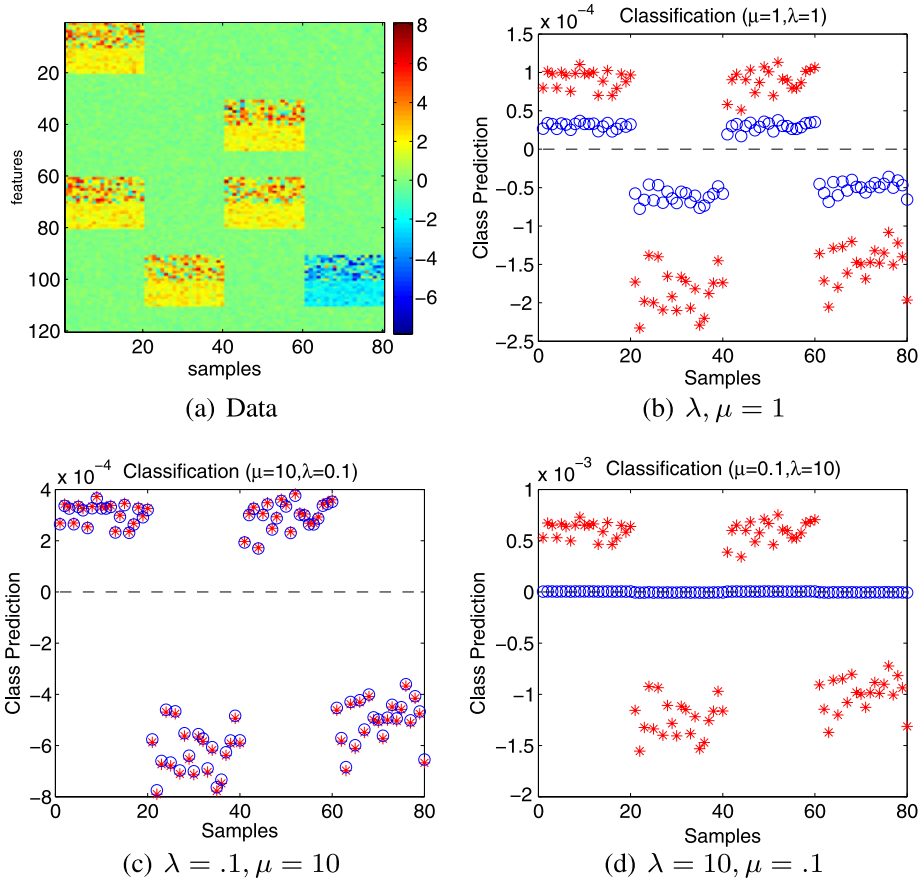
$$x^j \sim \mathbf{No}(-2,2),\ \text{for } j = 91,\ldots,100;\ x^j \sim \mathbf{No}(-2,.5),\ \text{for } j = 101,\ldots,110$$

We run MTGL on the simulated data with variations on the regularization parameters (μ, λ) and observe their effect on predicting class membership for all the samples. Recalling our definition for the multitask function,

$$F_t = f_0 + f_t$$

we can observe the parameters' effects on prediction in Fig. 1b–d for F_t (red) and f_0 (blue). Consistent with our expectations, when $\mu \gg \lambda$, the model behaves as if it is one task and we see $f_t \to 0$, Fig. 1c. Similarly, when $\mu \ll \lambda$, the model behaves as 2 independent tasks and $f_0 \to 0$, Fig. 1d.

We would also like to observe the effect of the regularization parameters on variable selection. We plot the RKHS norm for the common and task specific variables using two sets of regularization parameters (Fig. 2). We observe that the method correctly differentiates task specific components, and that the common components are reflective of the overlapping task variables. The effect of component variance on variable selection is also evident. In general, larger regularization terms will tend to emphasize the mean of the component values across all samples, which we observe with the larger μ and λ parameters, Fig. 2a–c. Conversely, when μ and λ are both small, we observe the opposite effect which is the selection of components with larger variance, Fig. 2d–f.



(a) Data (b) $\lambda, \mu = 1$ (c) $\lambda = .1, \mu = 10$ (d) $\lambda = 10, \mu = .1$

Fig. 1 (**a**) Data matrix **x**, column samples; samples 1 ... 20 correspond to task 1, class +1; samples 21 ... 40, task 1, class −1; samples 41 ... 60 task 2, class +1; samples 61 ... 80 task 2, class −1, (**b**, **c**, **d**) F_t (*red*) and f_0 (*blue*) using varying regularization penalties

We calculate subspace overlap scores (SOS) on this data with $\epsilon = .95$. Between task 1 and task 2 the computed SOS is .18, a low overlap score induced by the negative feature correlations. Taking the absolute values of the gradient outer product matrices reduces this effect and produces a SOS of .63. As we would expect, the subspace of T_1 is primarily contained within the common subspace, where the weighted projection of T_1 onto the common subspace produces a score of .96.

5.2 Dimension reduction on digits

The MNIST digit database (http://yann.lecun.com/exdb/mnist) is an important data set in the machine learning community for benchmarking classification methods. The data set consists of thousands of hand-written numbers (0–9) captured as 784-dimension vectors corresponding to the 28 pixel by 28 pixel image. All images have been centered and normalized. Our experiment uses the **3**, **5**, and **8** digits by considering '3 vs 8' as one task, and '5 vs 8' as a second task. The choice of these digits provides some helpful intuition: the bottom half

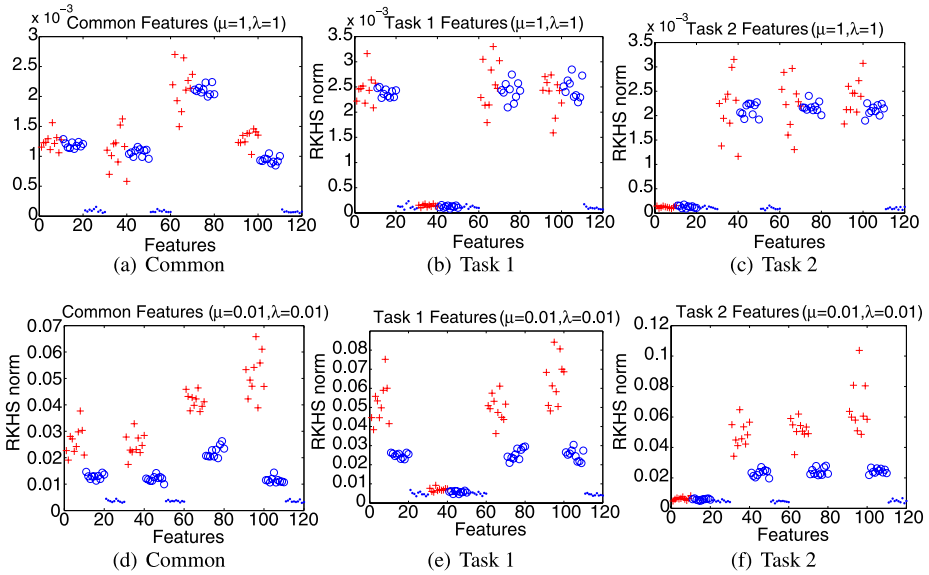


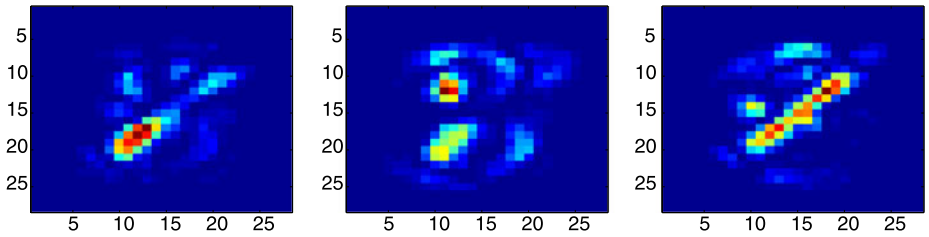
Fig. 2 Variance-bias tradeoff and regularization. (**a**, **b**, **c**) $\mu, \lambda = 1$. (**d**, **e**, **f**) $\mu, \lambda = .01$. High variance features in *red*, low variance features in *blue*

of the 3 and 5 are nearly identical, and the top halves, when taken as a composite, reproduce the top half of an 8. This therefore becomes an interesting classification problem. The goal of our experiment is to locate relevant subspaces within the predictive paradigm, and to compare these subspaces across tasks.

We build our data matrix X with a random selection of 50 3's, 50 5's, and 50 8's, where $X_i \in \mathbb{R}^{784}$ and $i \in \{1, \ldots, 200\}$. We run MTGL on the data and obtain gradient outer product matrices for the common, task 1 (3 vs 8) and task 2 (5 vs 8) models. By a spectral decomposition we can observe the top eigenvector (corresponding to the largest eigenvalue) for each of these matrices and compare the important components. We reshape the top eigenvector back into the 28-by-28 matrix and plot the components, see Fig. 3. The dominant observable features are what we would expect given the canonical forms of the 3, 5, and 8—the significant common features are located in the left lower quadrant of the plot and correspond to the common open loop of the 3 and 5 (Fig. 3a). We observe similar patterns in the task specific plots (Figs. 3b, c).

We would like to demonstrate that the subspaces obtained from the spectral decomposition are relevant for prediction (classification). This is analogous to the well-known PCA regression which constructs a classifier after projection of the data onto a lower dimensional space. We use the top $l = \{1, 2, 3, 4\}$ significant eigenvectors to define a subspace for our data, and predict class membership using the MNIST validation set (3, $n = 1010$; 5, $n = 892$; 8, $n = 974$). Unlike PCA, we have 3 subspaces in which to operate—common, task 1, and task 2—so we utilize the following prediction strategy: we run k-nearest neighbors (kNN) in each of the subspaces separately, and use the consensus of the largest nearest neighbor values to determine the class label. We compare our method's results with PCA regression and support vector machines (SVM, Vapnik 1998), where the SVM is trained within the original component space. All regression models and SVM are trained as '3, 5' vs '8'.

The above experiment is repeated 50 times and summary statistics are generated. We report classification accuracy as well as standard deviations, see Table 1. Here we observe



(a) Common (b) Task 1: 3 vs 8 (c) Task 2: 5 vs 8

Fig. 3 Digit plots of the top eigenvectors after a decomposition of the common, task 1, and task 2 GOP matrices

Table 1 Digit classification after dimension reduction: '3 and 5' vs '8'. Values are percentages (standard deviations) of prediction accuracy

	3	5	8	Total
MTGL	94.3 (2.1)	94.0 (2.6)	82.4 (3.7)	90.2
PCA-R	85.7 (5.2)	74.4 (13.5)	72.2 (6.9)	77.6
SVM	94.0 (2.2)	91.9 (3.4)	80.6 (4.3)	88.8

that MTGL outperforms PCA regression considerably, reflecting the importance of utilizing response variables for dimension reduction. While MTGL outperforms SVM, the difference in accuracy is less significant, although it is important to note that the final regression model for MTGL has many fewer variables than the SVM model.

5.3 Science documents/words

We now consider a data set of 1047 science articles which has been previously shown to have an interpretable hierarchical structure (Maggioni and Coifman 2007). Each article in the document corpora is categorized according to one of the following 8 subjects: Anthropology, Astronomy, Social Sciences, Earth Sciences, Biology, Mathematics, Medicine, or Physics. We restrict our analysis to 2036 words considered most relevant over all the documents. This yields a document-word matrix where entry (i, j) is the frequency of word j in document i. We formulate a multitask learning problem from this data by classifying Earth Sciences and Astronomy as task 1 and 2, respectively, against the remaining subjects. We randomly sample 25 documents from each of these 3 groups as input to MTGL and learn the relevant subspaces, as was done previously. We plot the 2-dimensional embedding of the validation data by projection on the top 2 eigenvectors (Fig. 4). We observe that with just two dimensions, the categories separate well. As we would expect, the Earth Science category is harder to classify since it is more likely to have overlapping terms with subjects such as Biology and Anthropology.

We next use the gradient outer product matrices to extract strongly covarying components (words) by selecting large off-diagonal elements. In general, the covarying terms we observe have a natural interpretation with respect to their corresponding science categories (Table 2). The most significant covarying term for both Astronomy and Earth Sciences is *earth*, a term that we recognize as immediately relevant for both subjects. Within the Astronomy task, *earth* co-varies with the words *star*, *galaxy*, and *universe*; for Earth Science, *earth* strongly co-varies with *water*, *lake*, and *ocean*. (From these results, we are tempted to conclude that the biggest difference between Earth and other planets is the presence or absence of water—an idea not completely devoid of scientific merit.) In the "Common" and "Earth Sciences"

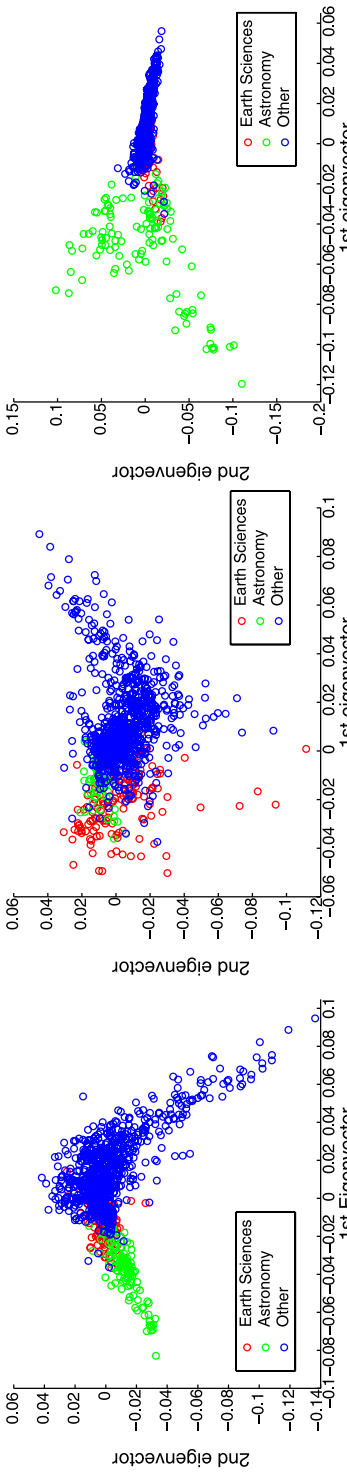


Fig. 4 Earth Sciences & Astronomy vs Other. Plots depict corresponding embedding of data using top 2 eigenvectors

Table 2 Science Documents-Words. Table lists the most highly covarying terms obtained from gradient outer product matrices, with *earth* as one of the most significant terms. Earth Sciences shows greatest covariation with water-related words; Astronomy with planet & star terms

Common	Earth Sciences	Astronomy
star-earth	lake-water	planet-star
planet-earth	earth-water	star-earth
galaxy-earth	gene-cell	galaxy-star
planet-star	water-year	galaxy-earth
earth-water	disease-cell	galaxy-planet
disease-cell	ice-water	astronomer-star
cell-people	ocean-water	universe-star
galaxy-star	human-cell	astronomer-earth
gene-cell	lake-earth	astronomer-planet
earth-year	sea-water	universe-earth

columns in Table 2, we observe biological terms such as *gene*, *cell*, and *disease*. Since Biology is the least distinctive category with terms spanning many other subjects, we would expect to see these uniquely biological terms for better classification. Overall, these results suggest that our method can successfully infer the covariance structure of variables within the predictive setting.

5.4 Graphical models

5.4.1 Simulated data

We begin with a simple, low-dimensional toy example to illustrate the application of the gradient outer product matrices for graphical models, and specifically, how they can be used to infer the full conditional dependencies for the common and task specific variables. We construct the following dependent explanatory variables from the random normal variables $\theta_1, \ldots, \theta_5 \overset{iid}{\backsim} \text{No}(0, 1)$ with

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

and $X_6, \ldots, X_8$ are drawn from independent Gaussians. Response data is modeled as 3 separate tasks

$$\begin{aligned} Y_1 &= X_1 + X_3 + \epsilon, \\ Y_2 &= X_1 + X_5 + \epsilon, \\ Y_3 &= X_3 + X_5 + \epsilon \end{aligned}$$

where $\epsilon \sim \text{No}(0, .5)$. We generate 100 samples for each task and use this data to obtain the estimated covariance matrix $\hat{\Sigma}_X$ and estimated gradient outer product matrices, $\hat{\Gamma}_0, \ldots, \hat{\Gamma}_3$. We compute partial correlations using (20), substituting the pseudo-inverse for the inverse since $\hat{\Sigma}$ and $\hat{\Gamma}$ are rank deficient. In $\hat{R}_X$ we observe significant partial correlations between the X_1 and X_2 variables, and the X_3, X_4, and X_5 variables (see Fig. 5a). Applying this same calculation to the gradient outer product matrices, we recover the response dependent partial correlations (see Figs. 5b, c, d, and e). Here, only the variables X_1, X_3, and X_5 are depicted as relevant, and the task-specific dependencies are correctly recovered.

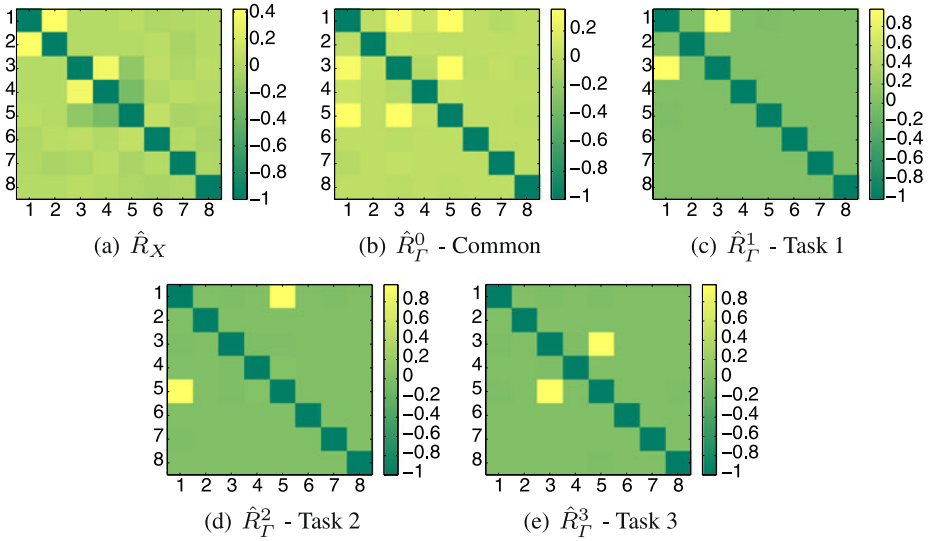


Fig. 5 Conditional Dependence Simulation (**a**) Partial correlation of data matrix X. (**b–e**) Partial correlation matrices using gradient outer products

5.4.2 Prostate cancer

We repeat the previous experiment using microarray expression data obtained from prostate cancer tumors (Tomlins et al. 2007). Each sample is annotated according to one of four stages of tumor progression: benign, low, high, and metastatic. This data set had been previously analyzed using multitask techniques to understand the tumorigenic mechanism common to all stages, as well as each specific stage (Edelman et al. 2008). We repeat the experimental design using a 3-task classification problem: task 1 defined as {benign → low}, task 2 as {low → high}, and task 3 as {high → metastatic}. In Edelman et al. (2008), the goal of the analysis was primarily to identify important genes that characterized these stage transitions. Using MTGL, we can now study the dependency structure across all tasks jointly and potentially identify new sets of co-regulated genes within the context of cancer progression.

The prostate data set is composed of 22 benign, 12 low grade, 20 high grade, and 17 metastatic samples, each sample measuring the expression level of over 12,000 genes. We eliminate those genes with low variance across all samples resulting in 4095 genes or variables. We run the multitask-gradient algorithm on this data to obtain four gradient outer product matrices, one for the common, and one for each of the cancer stage transitions. We compute the subspace overlap scores and report results in Table 3. We infer from these scores that the transition from high grade to metastatic represents the greatest gene expression shift, demonstrated by the largest value (.62) across the common model. The next greatest shift is seen in the transition from benign to low grade. We also observe that the {ben → low} and {low → high} transitions are highly scored (.63) suggesting that the genetic dysregulation between these stage transitions may be one of degree and not kind.

To explore the genetic dependency structure in finer detail, we construct graphical models from the {ben → low} and {high → met} gradient outer products. Since a graphical depiction of all 4095 genes is too complex for visualization purposes here, we select a sub-set of genes using as a threshold the upper-quartile of values along the diagonal of the gradient outer product matrix. We next determine edges in the graphs by taking the pseudo-inverse

Table 3 Prostate cancer: subspace overlap scores

	Common	Ben → Low	Low → High	High → Met
Common	–	.29	.18	.62
Ben → Low	.29	–	.63	.41
Low → High	.18	.63	–	.26
High → Met	.62	.41	.26	–

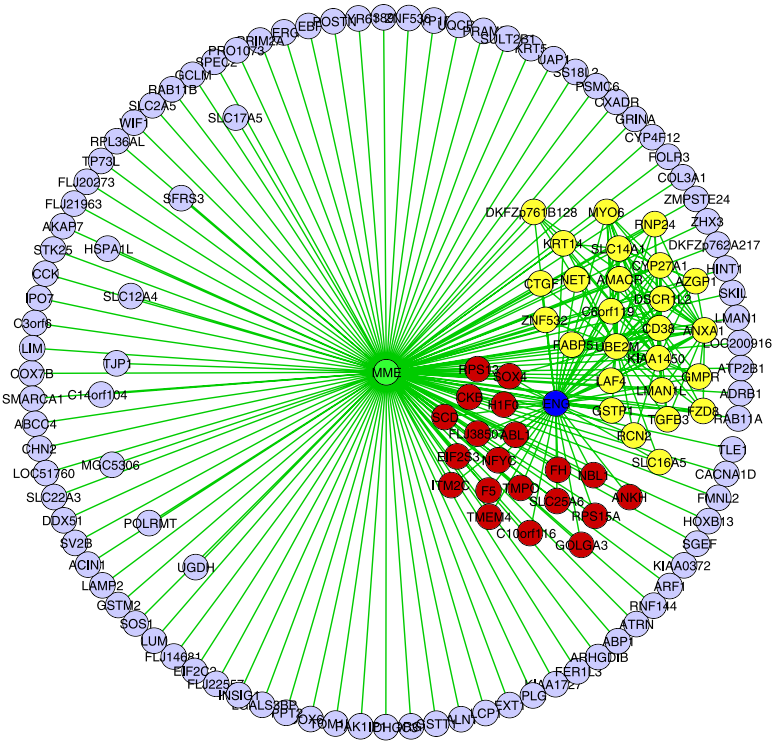


Fig. 6 Graphical model of prostate cancer: benign vs low grade

of the (reduced) GOP matrix—producing a partial correlation matrix—and thresholding this matrix at its upper-quartile. The non-zero elements in this matrix give rise to a sparse gene network.

Figures 6 & 7 recapitulate some of the biological processes and significant genes known in prostate cancer. In the center of the first graph (ben → low), we observe the gene MME (labeled green) connected to all other nodes in the graph, suggesting its strong global dependence. MME has been previously confirmed as strongly differentially expressed in aggressive prostate cancer (Tomlins et al. 2007). Also in this graph, we observe two distinct clusters; we label these C_1 and C_2 and annotate them in the graph with red and yellow, respectively. The genes in cluster C_1 are not connected with each other but do all share an edge with ENG (labled blue) and MME. Cluster C_2, on the other hand, has many interconnections within the cluster in addition to connections with MME and ENG. ENG (Endoglin)

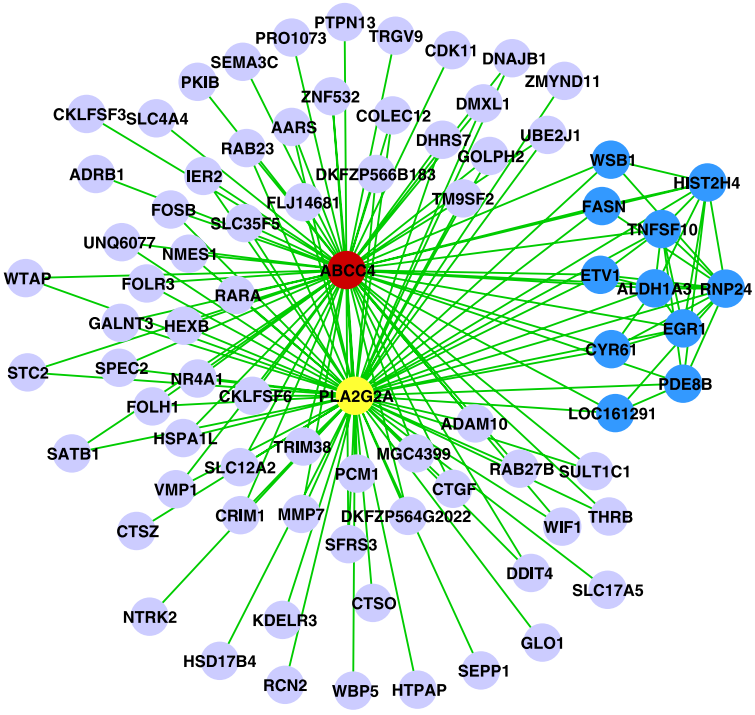


Fig. 7 Graphical model of prostate cancer: high-grade vs metastatic

has been previously implicated in vasculature development (angiogenesis), and is an important hallmark of tumor growth. In cluster C_2 we identify many well known prostate cancer genes including AMACR, ANXA1, CD38 and TFβ3.

In Fig. 7 we depict the gene dependency graph for the {high → met} progression. The labeled is dominated by two genes ABCC4 and PLA2G2A annotated in red and yellow respectively. ABCC4 has the pseudonym MRP (multi-drug resistant protein) and is known to have elevated expression in chemo-insensitive tumors, while PLA2G2A has also been identified in malignant prostate cancer (Jiang et al. 2002). The cluster in blue is strongly interconnected and contains several genes with known roles in prostatic tumor growth.

6 Discussion

We have presented a framework for dimension reduction of multivariate, multitask data in the predictive setting. In addition to finding relevant subspaces, our method is capable of learning the dependency structure of variables, allowing estimation of the full conditional dependency matrix and the construction of graphical models. Our method is based on the simultaneous learning of the regression function and its gradient, formulated as a linear combination of common and task specific components. Assuming smooth functions over all tasks, we can use the Taylor expansion to estimate gradients.

We have shown that dimension reduction can yield subspaces that potentially improve classification accuracy, as was demonstrated with the digits data experiment. However, we do not believe that gradient methods for dimension reduction will always or necessarily

outperform state-of-the-art classification methods such as support vector machines. In some situations classification accuracy is paramount, over and above inference of dependency structure, thereby requiring a parsimonious model with respect to the number of estimated parameters, i.e. Occam's razor. In this paper, the point of classification was to emphasize the relevance of the subspaces obtained for the joint distribution $\rho(X, Y)$, and that explicit modeling of the response variables for dimension reduction can outperform cases where only the marginal distribution $\rho(X)$ is considered.

Moreover, we believe a single, consistent framework is more desirable than multiple disjointed models. While we can imagine methods that consider single tasks separately which then combine results in a post-hoc manner, the efficiency and interpretability gained by a conjoint analysis makes hierarchical and multitask models generally preferable.

The method presented in this paper is based on Tikhonov regularization with an RKHS norm. This allows for the estimates to be effective in high-dimensional problems. However, the use of regularization introduces added parameters that must either be learned or set given some *a priori* knowledge. In the case of classification or regression, the accuracy of the model assessed by cross-validation or generalized approximate cross-validation (GACV) can be used to set the parameters. The MTGL setting introduces additional complexity where decisions concerning emphasis of common or task specific structure must be made. We do not believe prediction accuracy alone is capable of resolving this in many circumstances, and remains an area of open research.

Parametrization choices need not reflect *a priori* knowledge of task similarity; another consideration is the *a posteriori* analysis. This suggests the development of a coherent Bayesian framework for MTGL to allow for a posterior distribution on the regularization parameters and to generalize the types of norms in the regularization terms to a broader class of priors. For MTL a Bayesian model was explored in Xue et al. (2007). Integrating the ideas from Xue et al. (2007) with the non-parametric Bayesian kernel models developed in Liang et al. (2008) should provide a modeling framework for a Bayesian analysis and estimates of uncertainty.

References

Ando, R., & Zhang, T. (2005). A framework for learning predictive structure from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, *6*, 1817–1853.

Argyriou, A., Evgeniou, T., & Pontil, M. (2006). Multi-task feature learning. In *NIPS 20*.

Ben-David, S., & Schuller, R. (2003). Exploiting task relatedness for multiple task learning. In *Proc. of computational learning theory (COLT)*.

Bickel, S., Bogojeska, J., Lengauer, T., & Scheffer, T. T. (2008). Multi-task learning for HIV therapy screening. In *ICML '08: Proceedings of the 25th international conference on machine learning* (pp. 56–63). New York, NY, USA. New York: ACM.

Caruana, R. (1997). Multi-task learning. *Machine Learning*, *28*, 41–75.

Cook, R. (2007). Fisher lecture: Dimension reduction in regression. *Statistical Science*, *22*(1), 1–26.

Cook, R., & Weisberg, S. (1991). Discussion of "sliced inverse regression for dimension reduction". *Journal of the American Statistical Association*, *86*, 328–332.

Edelman, E., Guinney, J., Chi, J., Febbo, P., & Mukherjee, S. (2008). Modeling cancer progression via pathway dependencies. *PLoS Computational Biology*, *4*(2).

Evgeniou, T., & Pontil, M. (2004). Regularized multi-task learning. In *Proc. conference on knowledge discovery and data mining*.

Evgeniou, T., Micchelli, C., & Pontil, M. (2005). Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, *6*, 615–637.

Fisher, R. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Statistical Society A*, *222*, 309–368.

Fukumizu, K., Bach, F., & Jordan, M. (2005). Dimensionality reduction in supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, *5*, 73–99.

Hastie, T., & Tibshirani, R. (1996). Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society. Series B*, *58*(1), 155–176.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. Berlin: Springer.

Hotelling, H. (1933). Analysis of a complex of statistical variables in principal components. *Journal of Educational Psychology*, *24*, 417–441.

Jebara, T. (2004). Multi-task feature and kernel selection for svms. In *Proc. of ICML*.

Jiang, J., Neubauer, B., Graff, J., Chedid, M., Thomas, J., Roehm, N., Zhang, S., Eckert, G., Koch, M., Eble, J., & Cheng, L. (2002). Expression of group iia secretory phospholipase a2 is elevated in prostatic intraepithelial neoplasia and adenocarcinoma. *The American Journal of Pathology*, *160*, 667–671.

Lauritzen, S. (1996). *Graphical models*. Oxford: Clarendon Press.

Li, K. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, *86*, 316–342.

Liang, F., Mukherjee, S., Liao, M., & West, M. (2008). *Nonparametric Bayesian kernel models* (Technical Report 07-25). ISDS, Duke Univ.

Maggioni, M., & Coifman, R. (2007). Multiscale analysis of data sets using diffusion wavelets. In *Proc. data mining for biomedical informatics*.

Mukherjee, S., & Wu, Q. (2006). Estimation of gradients and coordinate covariation in classification. *Journal of Machine Learning Research*, *7*, 2481–2514.

Mukherjee, S., & Zhou, D. (2006). Learning coordinate covariances via gradients. *Journal of Machine Learning Research*, *7*, 519–549.

Mukherjee, S., Wu, Q., & Zhou, D.-X. (2010). Learning gradients on manifolds. *Bernoulli*, *16*(1), 181–207.

Obozinski, G., Taskar, B., & Jordan, M. (2006). *Multi-task feature selection* (Technical report). Dept. of Statistics, University of California, Berkeley.

Speed, T., & Kiiveri, H. (1986). Gaussian Markov distributions over finite graphs. *Annals of Statistics*, *14*, 138–150.

Sugiyama, M. (2007). Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *Journal of Machine Learning Research*, *8*, 1027–1061.

Tomlins, S., Mehra, R., Rhodes, D., Cao, X., Wang, L., Dhanasekaran, S., Kalyana-Sundaram, S., Wei, J., Rubin, M., Pienta, K., Shah, R., & Chinnaiyan, A. (2007). Integrative molecular concept modeling of prostate cancer progression. *Nature Genetics*, *39*(1), 41–51.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

Wahba, G. (1990). *Series in applied mathematics: Vol.* 59. *Splines models for observational data*. Philadelphia: SIAM.

Wu, Q., Guinney, J., Maggioni, M., & Mukherjee, S. (2010). Learning gradients: predictive models that reflect geometry and dependencies. *Journal of Machine Learning Research*, *11*, 2175–2198.

Wu, Q., Liang, F., & Mukherjee, S. (2007). *Regularized sliced inverse regression for kernel models* (Technical report 07-25). ISDS, Duke Univ.

Xia, Y., Tong, H., Li, W., & Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society. Series B*, *64*(3), 363–410.

Xue, Y., Liao, X., Carin, L., & Krishnapuram, B. (2007). Multi-task learning for classification with Dirichlet process priors. *Journal of Machine Learning Research*, *8*, 35–63.