Heterogeneous Multi-task Feature Learning

with mixed $\ell_{2,1}$ regularization

Yuan Zhong[1], Wei Xu[2], Xin Gao[1]

[1]Department of Mathematics and Statistics, York University
[2] Department of Biostatistics, Dalla Lana School of Public Health, University of Toronto

TABLE OF CONTENTS

Introduction

Data integration is the process of extracting information from multiple sources and jointly analyzing different data sets.

- Reduce sample bias;
- Increase prediction accuracy;
- Analyze intrinsic relatedness.

Meanwhile, there are many challenges due to the complexity of the data sets.

- Divergent dimensionality;
- Inconsistent measurements;
- Heterogeneous datasets.

Model Setup

Suppose K different tasks analyze a common set of predictors $M_1, M_2, \cdots, M_{p_n}$:

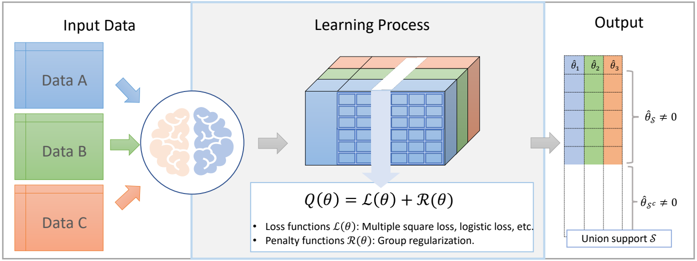
Responses		Linear Predictors					
		M_1	M_2		M_{p_i}		M_{p_n}
Task 1:	Y_1	$\theta_{11} X_{11}$	$+\theta_{12} X_{12}$	...	$+\theta_{1p} X_{1p}$	...	$+\theta_{1p_n} X_{1p_n}$
Task 2:	Y_2	$\theta_{21} X_{21}$	$+\theta_{22} X_{22}$	...	$+\theta_{2p} X_{2p}$	...	$+\theta_{2p_n} X_{2p_n}$
				...			
Task k:	Y_k	$\theta_{k1} X_{k1}$	$+\theta_{k2} X_{k2}$	...	$+\theta_{kp} X_{kp}$	...	$+\theta_{kp_n} X_{kp_n}$
				...			
Task K:	Y_K	$\theta_{K1} X_{K1}$	$+\theta_{K2} X_{K2}$	...	$+\theta_{Kp} X_{Kp}$	...	$+\theta_{Kp_n} X_{Kp_n}$
		$\theta^{(1)}$	$\theta^{(2)}$		$\theta^{(p)}$		$\theta^{(p_n)}$

Each task can be modeled as generalized linear model (GLM) with link function $g_k()$ and linear predictor η_k,

$$g_k(E(Y_k|X_k)) = \eta_k = X_k\theta_k$$

with $X_k = (X_{k1}, X_{k2}, \cdots, X_{kp_n})$ and $\theta_k = (\theta_{k1}, \theta_{k2}, \cdots, \theta_{kp_n})$.

Model Setup



Loss function $\mathcal{L}(\theta)$		Penalty $\mathcal{R}(\theta)$	Literature
Multivariate Square Loss	$+$	Mixed $\ell_{2,1}$ norm Mixed $\ell_{\infty,1}$ norm Ridge penalty	Liu et al. (2009); Lounici et al. (2011) Negahban and Wainwright (2011) Argyriou et al. (2006)
Multivariate Logistic Loss	$+$	Mixed $\ell_{2,1}$ norm	Lapedriza et al. (2007); Zhou et al. (2011);
Hinge loss	$+$	Mixed $\ell_{q,r}$ norm	Rakotomamonjy et al. (2011)
Composite Likelihood Loss	$+$	Group SCAD	Gao and Carroll (2017)

For any doubly indexed vector $v = (v_{11}, v_{12}, \cdots, v_{ij}, \cdots, v_{Kd})^T$, the $\ell_{q,r}$ norm is $\|v\|_{q,r} = \left(\sum_{j=1}^{d} \left(\sum_{i=1}^{K} v_{ij}^q \right)^{\frac{r}{q}} \right)^{\frac{1}{r}}$.

Mixed regularization:

The parameters/coefficients are grouped with the ℓ_2 norm, and for all predictors, the penalization is conducted with parameter λ_n,

$$\mathcal{R}(\theta) = n\lambda_n \|\theta\|_{2,1} = n\lambda_n \sum_{p=1}^{p_n} \|\theta^{(p)}\|_2$$

The sub-differential $\partial\|\theta\|_{2,1}/\partial\theta^{(p)}$ denoted by $z^{(p)}$ satisfies

$$\begin{cases} z^{(p)} = \frac{\theta^{(p)}}{\|\theta^{(p)}\|_2} & \text{if } \theta^{(p)} \neq \mathbf{0} \\ \|z^{(p)}\|_2 < 1 & \text{if } \theta^{(p)} = \mathbf{0} \end{cases}$$

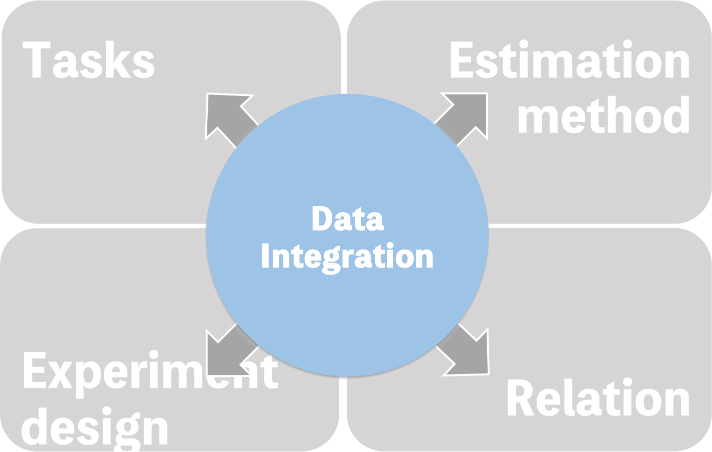
The mixed $\ell_{2,1}$ norm can satisfy following properties:

1. For any subset $\mathcal{E} \in \{1, 2, \cdots, p_n\}$, the mixed norm can be decomposed as $\|\theta\|_{2,1} = \|\theta_{\mathcal{E}}\|_{2,1} + \|\theta_{\mathcal{E}^c}\|_{2,1}$,
2. For any two vector θ_1 and θ_2, $\|\theta_1\|_{2,1} - \|\theta_2\|_{2,1} - z_2^T(\theta_1 - \theta_2) \geq 0$ with z_2 is subdifferential of $\|\theta_2\|_{2,1}$.

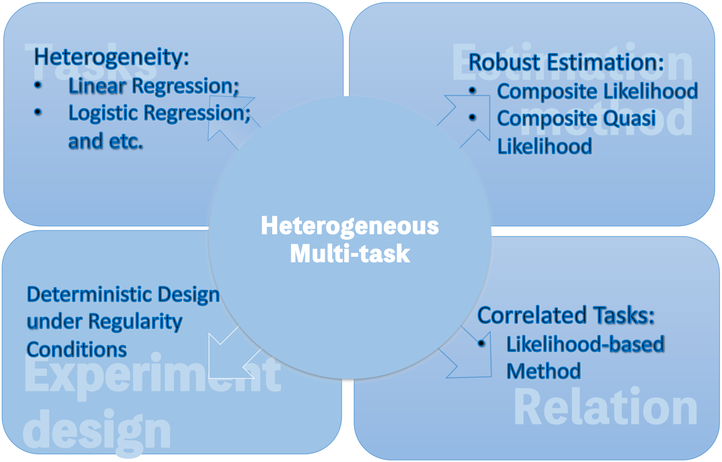
Model Setup

Previous researches:

1. Multiple regression tasks can be modeled by multivariate squared loss:
   - Lounici et al. (2011) showed that the regression coefficients estimated from multi-task learning could satisfy the oracle inequality, and the result can be extended to model multivariate regressions with independent and non-Gaussian errors.
   - Obozinski et al. (2011) and Negahban and Wainwright (2011) verified the union support recovery of the multi-task feature learning with different regularizers over both deterministic and random design.
2. Mixed types of tasks can be modeled by the composite likelihood loss:
   - Gao and Carroll (2017) applied the BIC-type information criterion to show selection consistency with the group smoothly clipped absolute deviations (SCAD) penalty.


Tasks
Estimation method
Data Integration
Experiment design
Relation

Heterogeneous Multi-task Feature Learning



The loss function is negative weighted average of individual losses,

$$\mathcal{L}(\theta) = -\sum_{k=1}^{K} w_k \ell_k(\theta_k; Y_k)$$

with weights w_k assigned based on the relative importance of the data sets.

For response variables from the exponential family, the individual loss can be modeled as

$$\ell_k(\theta_k; Y_k) = \sum_{i=1}^{n_k} \ell_{ki}(\theta_k; y_{ki}) = \sum_{i=1}^{n_k} \frac{y_{ki}\beta_{ki} - b_k(\beta_{ki})}{a(\phi_k)} + c(y_{ki}), \tag{1}$$

with cumulant generating functions $b_k(\beta_{ki})$ and $\partial b_k(\beta_{ki})/\partial\beta_{ki} = E(y_{ki}|x_{k1i}, \ldots, x_{kp_n i}) = \mu_{ki} = g_k^{-1}(\eta_{ki})$ (McCullagh and Nelder, 1989).

To relax distributional assumptions, the quasi log-likelihood function (Wedderburn, 1974) can be used to model the individual loss as follows,

$$\ell_k(\theta_k; Y_k) = \sum_{i=1}^{n_k} \ell_{ki}(\theta_k; y_{ki}) \propto \sum_{i=1}^{n_k} \int_{y_{ki}}^{g_k^{-1}(\eta_{ki})} \frac{y_{ki} - \mu}{V(\mu)\phi_k} d\mu. \tag{2}$$

for linear predictor $\eta_{ki} = \sum_{p=1}^{p_n} x_{kpi}\theta_p$.

Quasi-likelihood function:

- Mean and link function: $E(y_{ki}|x_{k1i},\ldots,x_{kp_ni}) = \mu_{ki} = g_k^{-1}(\eta_{ki})$,
- Variance function: $\mathrm{Var}(y_{ki}|x_{k1i},\ldots,x_{kp_ni}) = \phi_k V(\mu_{ki})$.

Assumption. (Quasi likelihood conditions)

For any θ satisfying $\|\theta-\theta^*\|_1 \le D$, the linear predictor $\|\eta_k\|_\infty \le \infty$ in any kth task. There exist some constants ν, $\sigma_{\max}$, K_1, K_2, and K_3,

- Mean and variance functions are bounded as

$$\max_{k,i}\{|g_k^{-1}(\eta_{ki})|\} \le \nu, \text{ and } V(g_k^{-1}(\eta_{ki})) \le e^3 V(g_k^{-1}(\eta_{ki}^*)) \le \sigma_{\max}^2$$

- The derivatives of mean and variance functions have

$$\left|\max_{k,i}\frac{\partial^2 g_k^{-1}(\eta)}{\partial\eta^2}\Big|_{\eta=\eta_{ki}}\right| \le K_1, \left|\max_{k,i}\frac{V'(g_k^{-1}(\eta))}{V(g_k^{-1}(\eta))}\Big|_{\eta=\eta_{ki}}\right| \le K_2,$$

$$\text{and } \left|\min_{k,i}\frac{\partial g_k^{-1}(\eta)}{\partial\eta}\Big|_{\eta=\eta_{ki}}\right| \ge K_3$$

The sensitivity matrix $H(\theta)$ and variability matrix $J(\theta)$:

$$H(\theta) = E\{n^{-1}\nabla^2\mathcal{L}(\theta)\} \text{ and } J(\theta) = Cov\{n^{-1}\nabla\mathcal{L}(\theta)\}.$$

and $H(\theta) \neq J(\theta)$ for the composite quasi log-likelihood function due to the correlations across different tasks.

The maximum composite quasi-likelihood estimator $\hat{\theta}$ can be used for the inference of correlated platform, which can hold the asymptotic properties based on the information theory

$$\sqrt{n}(\hat{\theta} - \theta^*) \overset{d}{\sim} N_{p_n}(0, G^{-1}(\hat{\theta}))$$

The asymptotic covariance matrix of the maximum composite quasi likelihood estimator can be estimated by the inverse Godambe information matrix $G^{-1}(\theta)$

$$G(\theta) = H(\theta)J^{-1}(\theta)H(\theta),$$

The penalty function is the $\ell_{2,1}$ regularization,

$$\mathcal{R}(\theta) = n\lambda_n \|\theta\|_{2,1} = n\lambda_n \sum_{p=1}^{p_n} \|\theta^{(p)}\|_2$$

The penalized estimate is the solution $\hat{\theta}$ of the estimating equation

$$n^{-1}\nabla Q(\hat{\theta})^T(\hat{\theta}-\theta^*) = n^{-1}\nabla\mathcal{L}(\hat{\theta})^T(\hat{\theta}-\theta^*) + \lambda_n \hat{z}^T(\hat{\theta}-\theta^*) = 0,$$

where $\hat{z}$ is the subdifferential of the mixed $\ell_{2,1}$ norm at the penalized estimate $\hat{\theta}$. If $\hat{\theta}$ correctly recovers the true union support $\mathcal{S}$, then

$$\begin{cases} -\frac{1}{n}\nabla\mathcal{L}(\hat{\theta})^{(p)} = \lambda_n \hat{z}^{(p)}, & \text{for any } p \in \mathcal{S}; \\ \|\frac{1}{n}\nabla\mathcal{L}(\hat{\theta})^{(p)}\|_2 < \lambda_n, & \text{for any } p \in \mathcal{S}^c. \end{cases}$$

Assumption. (Dimensionality)

There exist some constants $0 < 3k_1 + k_2 < 1$, such that $s = O(n^{k_1})$ and $\log(p_n) = O(n^{k_2})$. In addition, the true parameter vector $\|\theta^*\|_1 \leq R$ for some constant $R > 0$.

By Hölder's inequality, the finite sample bound can be obtained by

$$n^{-1}\nabla\mathcal{L}(\theta^*)^T(\hat{\theta}-\theta^*) \leq sup_p\{\|n^{-1}\nabla\mathcal{L}(\theta^*)^{(p)}\|_2\}\|\hat{\theta}-\theta^*\|_{2,1}$$

Assumption. (Design of Study)

For any kth task, let linear predictor be denoted as $\eta_{ki}^* = \sum x_{kpi}\theta_{kp}^*$,

1. The error terms $y_{ki} - g_k^{-1}(\eta_{ki}^*)$ are independent from sub-exponential distributions with ψ_1 norm bounded by some constant $\mathcal{M}$;
2. The covariates in the design matrix satisfy the condition that $\sup_{k,p,i}\{x_{kpi}\} \leq L < \infty$.

The concentration of the score function and Hessian:

$$\sup_p \|\frac{1}{n}\nabla\mathcal{L}(\theta^*)^{(p)}\|_2 = O_p(\sqrt{\frac{K}{n}} + \sqrt{\frac{K\log(p_n)}{n}}),$$

$$\sup_{k,p,p'} \{\frac{1}{n}\nabla^2\mathcal{L}(\theta^*) - H(\theta^*)\}_{[kp,kp']} = O_p(\sqrt{\frac{\log p_n}{n}}),$$

for any $k = 1, 2, \cdots, K$ and $p, p' = 1, 2, \cdots, p_n$.

Assumption. (Restricted Eigenvalues)

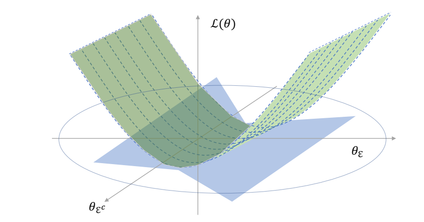
There exist $m = c_0 K s$ for some $c_0 > 0$ and some positive constants $\gamma \geq 2\sqrt{K} + 1$, ρ_- and ρ_+, such that the restricted minimum and maximum eigenvalues of the design matrix

$$\rho_-(m,\gamma) = \inf_k \left\{ u^T \frac{X_k X_k^T}{n} u : u \in \mathcal{C}(m,\gamma) \right\}, \text{ and } \rho_+(m,\gamma) = \sup_k \left\{ u^T \frac{X_k X_k^T}{n} u : u \in \mathcal{C}(m,\gamma) \right\}$$

are bounded by

$$0 < \rho_- \leq \rho_-(m,\gamma) < \rho_+(m,\gamma) \leq \rho_+ < \infty,$$

where $\mathcal{C}(m,\gamma) := \{u : \mathcal{S} \subset J, |J| < m, \|u_{J^c}\|_1 \leq \gamma \|u_J\|_1\}$.



The observed Hessian

$$0 < \kappa_- \leq u^T \frac{\nabla^2 \mathcal{L}(\theta)}{n} u \leq \kappa_+ < \infty$$

for any unit vector $u \in \mathcal{C}(m,\gamma)$.

Assumption. (Mutual Incoherence)

Let the sub-matrices of the expected Hessian matrix be denoted by

$$H^*_{\mathcal{S}\mathcal{S}} = E_{\theta^*}[n^{-1}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}\mathcal{S}}] \text{ and } H^*_{\mathcal{S}^c\mathcal{S}} = E_{\theta^*}[n^{-1}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}^c\mathcal{S}}],$$

where $\mathcal{S}$ is the support of non-zero parameters. For some constant $\xi \in (0,1)$, the inequality holds

$$\sqrt{K}\left|\!\left|\!\left| H^*_{\mathcal{S}^c\mathcal{S}}[H^*_{\mathcal{S}\mathcal{S}}]^{-1} \right|\!\right|\!\right|_\infty \leq 1-\xi.$$

With concentration of Hessian and restricted eigenvalues, the observed Hessian holds the mutual incoherence condition:

$$\sqrt{K}\left|\!\left|\!\left| \frac{1}{n}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}^c\mathcal{S}}\left(\frac{1}{n}\nabla^2\mathcal{L}(\theta^*)_{\mathcal{S}\mathcal{S}}\right)^{-1} \right|\!\right|\!\right|_\infty < 1-\frac{\xi}{2}$$

holds with a probability at least $1-4K\exp\{-C_0\xi^2 n/s^3 + 2\log(p_n)\}$ for some universal constant $C_0 > 0$.

Theorem 1. (Sign Recovery Consistency)

Suppose the penalty parameter chosen as

$$\lambda_n \geq \frac{4\mathcal{M}_*}{\xi}\sqrt{\frac{K}{n}}(1 + \sqrt{2\log(p_n)}), \tag{3}$$

and the minimum non-zero parameter $\min_{k;p\in\mathcal{S}} \theta_{kp} \geq 2\kappa_-^{-1}\sqrt{s}\lambda_n$, the estimator $\hat{\theta}$ satisfies that $\mathrm{sign}(\hat{\theta}) = \mathrm{sign}(\theta^*)$ with probability $1 - 2p_n^{-d} - 2K\exp\{-Cn/s^3 + \log(p_n)\}$ for the universal constants $d > 1$ and $C > 0$.

Theorem 2. (Estimation Error Bound)

Suppose the composite score vector satisfies $\|n^{-1}\nabla\mathcal{L}(\theta^*)\|_\infty \leq \lambda_n/(2\sqrt{K})$, the estimator $\hat{\theta}$ satisfies

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{3\lambda_n\sqrt{s}}{2\kappa_-}; \|\hat{\theta} - \theta^*\|_1 \leq \frac{3\sqrt{K}(\sqrt{K}+1)}{\kappa_-}\lambda_n s;$$

$$(\frac{1}{n}\nabla\mathcal{L}(\hat{\theta}) - \frac{1}{n}\nabla\mathcal{L}(\theta^*))^T(\hat{\theta} - \theta^*) \leq \frac{3(\sqrt{K}+1)(2\sqrt{K}+1)}{2\kappa_-}\lambda_n^2 s$$

with a probability at least $1 - 2\exp\{-C\log(p_n)\}$ for some constant C.

Simulation setups

- Parameters: Non-zero coefficients present different patterns for any $p \in \mathcal{S}$ and $|\mathcal{S}| = \lfloor p_n^{1/2} \rfloor$.

	Coefficient Type	Distribution
Task 1	Large variance	$\theta^*_{1p} \sim N(1,3)$
Task 2	Small variance	$\theta^*_{2p} \sim N(1,1)$
Task 3	Strictly positive	$\theta^*_{3p} \sim \text{Unif}(1,2)$
Task 4	No sign constraint	$\theta^*_{4p} \sim \text{Unif}(-1,1)$

- Responses variable is modeled by the linear function,

$$y_{ki} = \sum_{p=1}^{p_n} x_{kpi}\theta^*_{kp} + \varepsilon_{ki}, \text{ and } x_{kpi} \sim N(0,1)$$

- Error term $(\varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i}, \varepsilon_{4i})^T$ is i.i.d vectors simulated from multivariate Normal distribution MVN$(0, \Sigma)$ or multivariate t distribution $t_{10}(\Sigma)$.

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 & \rho_{14}\sigma_1\sigma_4 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 & \rho_{24}\sigma_2\sigma_4 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 & \rho_{34}\sigma_3\sigma_4 \\ \rho_{14}\sigma_1\sigma_4 & \rho_{24}\sigma_2\sigma_4 & \rho_{34}\sigma_3\sigma_4 & \sigma_4^2 \end{bmatrix}; \begin{pmatrix} \sigma_1^2 \\ \sigma_2^2 \\ \sigma_3^2 \\ \sigma_4^2 \end{pmatrix} = \begin{pmatrix} 9 \\ 4 \\ 4 \\ 1 \end{pmatrix}$$

Combination of two regression tasks and two classification tasks:

Two responses are dichotomized as binary data $y_{ki} = 1$ if $\sum_{p=1}^{p_n} x_{kpi}\theta^*_{kp} + \varepsilon_{ki} \geq 0$.

Simulation I: Moderate correlation $\rho_{kk'} \sim$ Unif(0.4, 0.65).

Model	n= 200 p = 200 PSR	FDR	n= 200 p = 500 PSR	FDR	n= 500 p = 500 PSR	FDR	n= 500 p = 1000 PSR	FDR
Simulation I: Gaussian Error								
MTL	98 (1)	4 (5)	97 (1)	4 (4)	98 (1)	2 (3)	99 (0)	2 (3)
SA 1	81 (7)	8 (8)	86 (4)	11 (7)	87 (4)	7 (5)	89 (3)	7 (4)
SA 2	82 (7)	11 (9)	87 (4)	14 (8)	87 (4)	8 (6)	89 (3)	8 (5)
SA 3	80 (7)	2 (4)	90 (3)	18 (11)	80 (3)	0 (0)	87 (2)	1 (2)
SA 4	83 (6)	22 (11)	91 (3)	35 (9)	83 (4)	15 (7)	88 (2)	21(7)
Simulation II: Heavy-tail Error								
MTL	97 (1)	4 (5)	96 (1)	4 (4)	97 (0)	1 (3)	99 (0)	2 (3)
SA 1	82 (7)	9 (8)	86 (4)	12 (7)	87 (4)	7 (6)	89 (3)	7 (5)
SA 2	82 (7)	12 (9)	87 (4)	15 (8)	87 (4)	8 (6)	90 (3)	9 (5)
SA 3	80 (5)	2 (4)	90 (3)	20 (11)	80 (3)	0 (1)	87 (2)	2 (2)
SA 4	84 (6)	23 (11)	91 (3)	37 (9)	84 (4)	16 (7)	89 (2)	22 (7)

MTL: Multi-task learning; SA: Single-plaform analysis. PSR: Positive selection rates (%); FDR: false discovery rates (%).

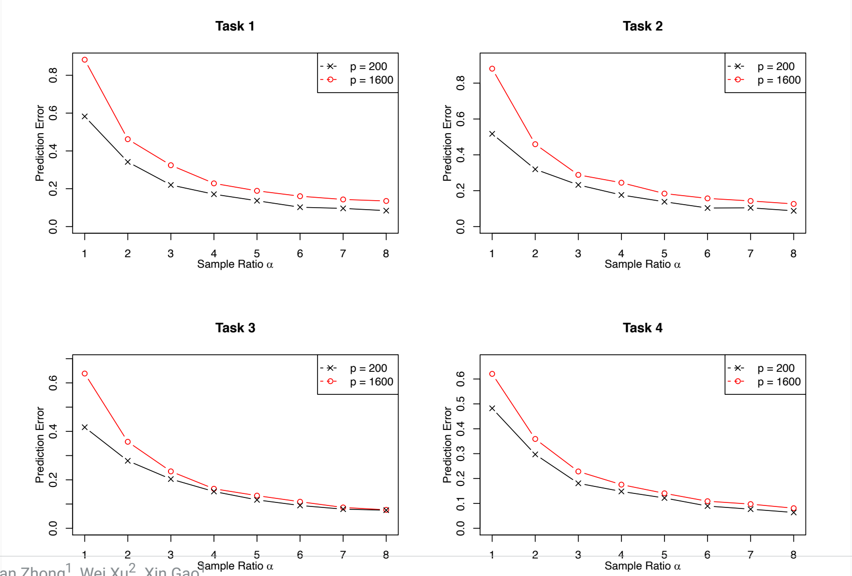
Simulation II: High correlation $\rho_{kk'} = 0.9$.

Model	n= 200 p = 200 PSR	FDR	n= 200 p = 500 PSR	FDR	n= 500 p = 500 PSR	FDR	n= 500 p = 1000 PSR	FDR
Simulation I: Gaussian Error								
MTL	99 (1)	4 (5)	97 (1)	4 (4)	98 (1)	1 (2)	99 (0)	2 (3)
SA 1	81 (8)	8 (8)	87 (4)	10 (7)	87 (4)	6 (5)	89 (3)	6 (4)
SA 2	82 (7)	11 (9)	87 (5)	14 (8)	87 (4)	8 (6)	89 (3)	8 (5)
SA 3	79 (5)	2 (4)	90 (3)	19 (10)	80 (3)	0 (0)	87 (2)	1 (2)
SA 4	83 (6)	23 (12)	91 (3)	36 (10)	83 (4)	15 (8)	88 (2)	21(7)
Simulation II: Heavy-tail Error								
MTL	98 (1)	5 (5)	97 (1)	4 (4)	98 (1)	1 (2)	99 (0)	2 (3)
SA 1	81 (8)	8 (8)	87 (4)	10 (7)	87 (4)	6 (5)	89 (3)	6 (4)
SA 2	82 (7)	10 (8)	87 (5)	14 (8)	87 (4)	8 (6)	89 (3)	8 (5)
SA 3	80 (5)	2 (4)	90 (3)	19 (10)	80 (3)	0 (0)	87 (2)	1 (2)
SA 4	83 (6)	24 (11)	91 (3)	36 (10)	83 (4)	15 (8)	88 (2)	21(7)

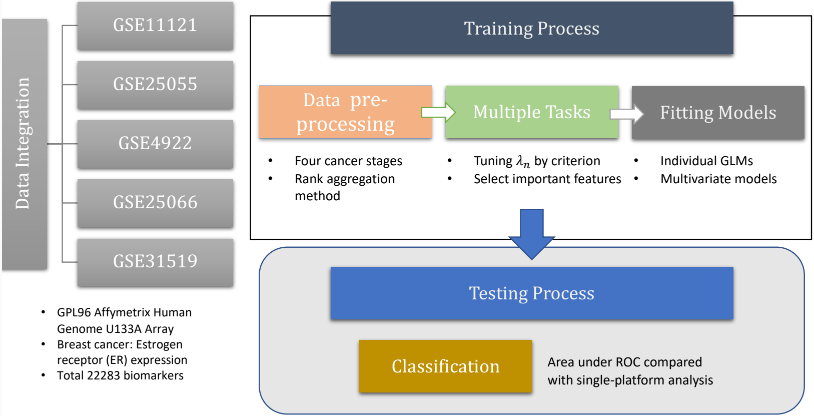
MTL: Multi-task learning; SA: Single-plaform analysis. PSR: Positive selection rates (%); FDR: false discovery rates (%).

Simulation Studies

Prediction Error:$= (n^{-1}\nabla\mathcal{L}(\theta) - n^{-1}\nabla\mathcal{L}(\theta^*))^T(\theta - \theta^*)$ and $\alpha = \frac{n}{s \log p_n}$.



Breast cancer multi-task studies.



- GPL96 Affymetrix Human Genome U133A Array
- Breast cancer: Estrogen receptor (ER) expression
- Total 22283 biomarkers

Table: Performance of the logistic regression models is measured by AUC; performance of the multinomial regression models is measured by the percentage of correct classification.

Tasks	Logistic regression (AUC)			Multinomial regression (% Classification)	
Data	GSE11121	GSE4922	GSE31519	GSE25055	GSE25066
(*n*)	(151)	(188)	(46)	(217)	(358)
MTL	0.81	0.81	0.77	66	66
SA	0.74	0.83	0.65	66	64

Reference

[1] G. Obozinski, M. J. Wainwright, and M. I. Jordan. Support union recovery in high-dimensional multivariate regression. The Annals of Statistics, 39(1):1–47, 02 2011.

[2] K. Lounici, M. Pontil, S. van de Geer, and A. B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. The Annals of Statistics, 39(4):2164–2204, 2011.

[3] X. Gao and R. J. Carroll. Data integration with high dimensionality. Biometrika, 104(2):251–272, 05 2017.

[4] J. Fan, H. Liu, Q. Sun, and T. Zhang. I-lamm for sparse learning: Simultaneous control of algorithmic complexity and statistical error. The Annals of Statistics, 46 2:814–841, 2018.

[5] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. Statist. Sci., 27(4):538–557, 11 2012.

Thank You!

YORK
UNIVERSITÉ
UNIVERSITY