Learning under latent group sparsity via heat flow dynamics on networks

Soumendu Sundar Mukherjee
Indian Statistical Institute, Kolkata

February 7, 2022

Joint work with Subhroshekhar Ghosh (NUS)
arXiv: `https://arxiv.org/abs/2201.08326`

Setting

- $(y_i, X_i)_{i=1}^n$: data
- $X_{n\times p}$: data matrix with rows $X_i^\top$
- Generalised linear model:
$$\mathbb{E}[y_i|X_i] = g^{-1}(X_i^\top \beta).$$
- Regression:
$$y_i = X_i^\top \beta + \epsilon_i, \quad (g(x) = x).$$
- Logistic regression: y_i binary coded.
$$\mathbb{P}[y_i = 1 \mid X_i] = \frac{\exp(X_i^\top \beta)}{1+\exp(X_i^\top \beta)}, \quad \left(g(x) = \log \frac{x}{1-x}\right).$$
- The number of variables p can scale with n.

The problem of grouped variable selection

- β is group-sparse, i.e. there are groups of variables $\mathcal{C}_1, \ldots \mathcal{C}_K$, $[p] = \sqcup_{j \in [K]} \mathcal{C}_j$, such that

$$\text{support}(\beta) = \mathcal{C}_{j_1} \cup \cdots \cup \mathcal{C}_{j_s},$$

  $j_1, \ldots, j_s \in [K]$.
- Want to estimate β from data (y, X).

The lasso penalty [Tibshirani (1996)]

- Penalise via the ℓ_1 norm:

$$\mathrm{L}(\beta) = \|\beta\|_1 = \sum_{j \in [p]} |\beta_j|.$$

- Optimisation problem:

$$\ell(y, X\beta) + \lambda \cdot \mathrm{L}(\beta),$$

where $\ell(y, X\beta)$ is some loss function, e.g., $\frac{1}{2n}\|y - X\beta\|_2^2$, negative log-likelihood, Huber's loss, etc.

The lasso penalty [Tibshirani (1996)]

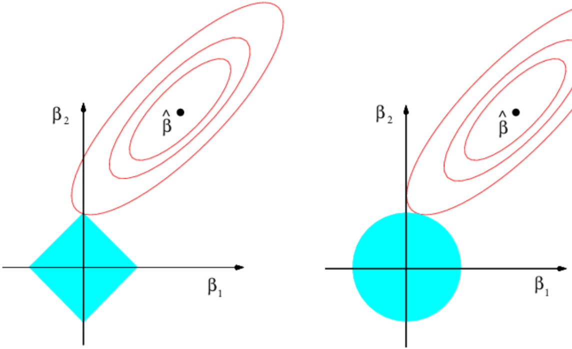


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right)*. The solid blue areas are the constraint regions, $|\beta_1| + |\beta_2| \leq s$ and $\beta_1^2 + \beta_2^2 \leq s$, while the red ellipses are the contours of the RSS.*

Figure taken from *An Introduction to Statistical Learning*, James et al.

The group lasso penalty [Yuan and Lin (2006)]

- Use a (weighted) ℓ_1 penalty on groupwise ℓ_2 norms:

$$\mathrm{GL}(\beta) = \sum_{j \in [K]} \sqrt{|\mathcal{C}_j|} \|\beta_{\mathcal{C}_j}\|_2.$$

- Optimisation problem:

$$\ell(y, X\beta) + \lambda \cdot \mathrm{GL}(\beta),$$

where $\ell(y, X\beta)$ is some loss function, e.g., $\frac{1}{2n}\|y - X\beta\|_2^2$, negative log-likelihood, Huber's loss, etc.

The group lasso penalty [Yuan and Lin (2006)]

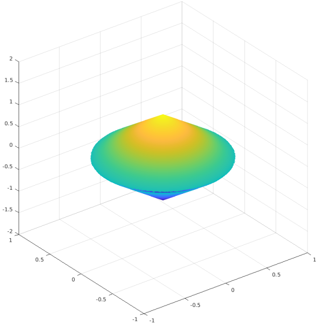


Figure 1: Consider three variables with two groups $\{1, 2\}$ and $\{3\}$. In this display, we plot of the level set $\{\beta \mid \Lambda_t(\beta) \leq 1\}$ for different values of t. The graph G here is the union of an isolated vertex and an edge. The eigengap $\lambda_3 = 2$.

The group lasso penalty [Yuan and Lin (2006)]

- Convex optimisation problem if $\ell(y, X\beta)$ is convex.
- Groups need to be known in advance.

A new penalty based on heat flow

- Assume that the group information comes from a graph G on the variables.
- Let $L = D - A$ denote the (unnormalised) graph Laplacian.
- Recall: G has K connected components $\mathcal{C}_1, \ldots, \mathcal{C}_K$ if and only if L has K zero eigenvalues.
- The eigenspace of 0 is spanned by $\{\frac{\mathbf{1}_{\mathcal{C}_j}}{\sqrt{|\mathcal{C}_j|}}, j \in [K]\}$.

A new penalty based on heat flow

- Consider functions

$$\Psi(\beta) = \beta \odot \beta$$

  and

$$\Psi^{[-1]}(\beta) = (\sqrt{|\beta_1|}, \dots, \sqrt{|\beta_p|})^\top.$$

- We introduce the heat flow penalty

$$\Lambda_t(\beta) := \langle \Psi^{[-1]}(e^{-tL}\Psi(\beta)), \mathbf{1}_p \rangle.$$

But why?

- Let
$$L = \sum_i \lambda_i v_i v_i^\top,$$
$0 = \lambda_1 = \cdots = \lambda_K < \lambda_{K+1} < \cdots < \lambda_p$.

- Note that
$$\begin{aligned} e^{-tL}\Psi(\beta) &= \sum_i e^{-t\lambda_i}\langle v_i, \beta \odot \beta\rangle v_i \\ &= \sum_{i=1}^K \frac{\|\beta_{\mathcal{C}_i}\|_2^2}{|\mathcal{C}_i|}\mathbf{1}_{\mathcal{C}_i} + \sum_{i>K} e^{-t\lambda_i}\langle v_i, \beta \odot \beta\rangle v_i \\ &\approx \sum_{i=1}^K \frac{\|\beta_{\mathcal{C}_i}\|_2^2}{|\mathcal{C}_i|}\mathbf{1}_{\mathcal{C}_i}, \end{aligned}$$
for t large enough (depending on the spectral gap $\lambda_{K+1} > 0$).

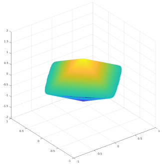
But why?

- Thus

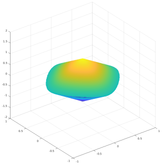
$$\langle \Psi^{[-1]}(e^{-tL}\Psi(\beta)), \mathbf{1}_p \rangle \approx \left\langle \Psi^{[-1]}\left(\sum_{i=1}^{K} \frac{\|\beta_{\mathcal{C}_i}\|_2^2}{|\mathcal{C}_i|} \mathbf{1}_{\mathcal{C}_i} \right), \mathbf{1}_p \right\rangle$$
$$= \mathrm{GL}(\beta).$$

- Learning with heat flow penalty:

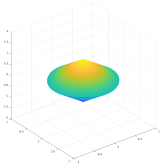
$$\min_{\beta} [\ell(y, X\beta) + \lambda \cdot \Lambda_t(\beta)].$$



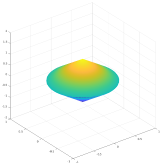
$t = 0.01$



$t = 0.1$



$t = 0.5$



$t = 1$

Figure 2: Consider three variables with two groups $\{1, 2\}$ and $\{3\}$. In this display, we plot of the level set $\{\beta \mid \Lambda_t(\beta) \leq 1\}$ for different values of t. The graph G here is the union of an isolated vertex and an edge. The eigengap $\lambda_3 = 2$.

Properties

- Non-convex, unlike group lasso.
- Subgradient descent of (block) coordinate descent can be performed easily.
- Set $h = e^{-tL}(\beta \odot \beta)$. Then

$$\Lambda_t(\beta) = \sum_{j=1}^{p} \sqrt{|h_j|}.$$

Thus

$$\frac{\partial \Lambda_t(\beta)}{\partial \beta_\ell} = \sum_{j=1}^{p} \partial s(h_j) \frac{\partial h_j}{\partial \beta_\ell} = \sum_{j=1}^{p} \underbrace{\partial s(h_j) 2}_{=:\zeta_j} (e^{-tL})_{j\ell} \beta_\ell,$$

where $s(x) = \sqrt{|x|}$ so that $\partial s(x) = \frac{\operatorname{sign}(x)}{2s(x)}$.

Properties

- Since e^{-tL} is symmetric, we can write

$$\partial \Lambda_t(\beta) = (e^{-tL}\zeta) \odot \beta.$$

- Given $v \in \mathbb{R}^p$,

$$(e^{-tL}v)_i = \mathbb{E}(f_{Z(t)} \mid Z(0) = i),$$

where $(Z(t))_{t \geq 0}$ is the CTRW on G.

- A Monte Carlo estimate can be easily obtained:

$$(\widehat{e^{-tL}v})_i = \frac{1}{B} \sum_{j=1}^{B} f_{Z^{(j)}(t)},$$

where $Z^{(1)}, \ldots, Z^{(B)}$ are B independent random walks started at i.

Advantages

- Learns group structure automatically using only local graph information — privacy friendly.
- Does not need to know the number of groups K.
- Need to compute the end-points of $p \cdot B$ random walks once and for all, can be done in parallel.
- To ensure prediction error $\leq \epsilon$, need $O(\max(\log p, \log(1/\epsilon)))$ many steps in each RW.

Bounds on prediction error

Theorem

Under "some conditions", we have with high probability that

$$\frac{1}{n}\|X(\hat{\beta}_{t,\lambda} - \beta^*)\|_2^2 = O(\|\beta^*\|_{2,1}\lambda|\mathcal{C}_{\max}| + p^{3/2}e^{-t\lambda_g/2}). \qquad (1)$$

If we further assume $\mathrm{RE}(s)$ *holds for* X *with parameter* κ*, then*

$$\frac{1}{n}\|X(\hat{\beta}_{t,\lambda} - \beta^*)\|_2^2 = O\left(\frac{s\lambda^2|\mathcal{C}_{\max}|}{\kappa^2} + p^{3/2}e^{-t\lambda_g/2}\right). \qquad (2)$$

$\mathrm{RE}(s)$ *ensures enough curvature at approximately* *s-group-sparse vectors.*

Simulations - I

- The covariates $\sim \mathcal{N}(0, \Sigma)$, with

$$\Sigma = \begin{pmatrix} \Sigma_{p_1}(\rho_1) & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Sigma_{p_2}(\rho_2) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Sigma_{p_3}(\rho_3) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \Sigma_{p_4}(\rho_4) \end{pmatrix},$$

  where $\Sigma_d(\rho) = (1-\rho)I_d + \rho \mathbf{1}_d \mathbf{1}_d^\top$ is the equi-correlation matrix of order d.

- Take some estimate $\hat{\Sigma}$ of Σ. Let $\hat{R}$ be the corresponding correlation matrix.

- The graph is estimated as follows

$$A_{ij} = \mathbf{1}_{\{|\hat{R}_{ij}| \geq \tau(\hat{R})\}}.$$

Simulations - I

For L we use the Laplacian corresponding to A. Group Lasso is fed the output of spectral clustering on L (with oracle knowledge of K).

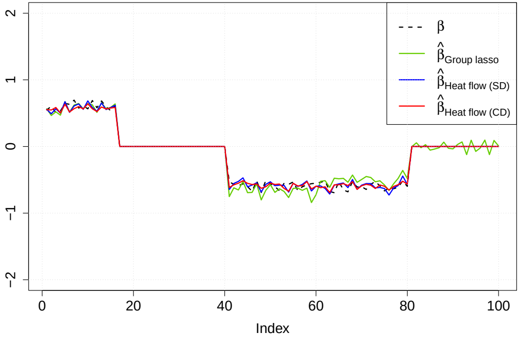


Figure 3: $n = 200, p = 100, (p_1, p_2, p_3, p_4) = (16, 24, 40, 20)$, correlations $(\rho_1, \rho_2, \rho_3, \rho_4) = (0.6, 0.9, 0.7, 0.4)$.

Simulations - I

	Group lasso	Heat flow (SD)	Heat flow (CD)
Prediction error	0.03	0.02	0.03
Estimation error	0.84	0.48	0.46
Sensitivity	1	1	1
Specificity	0.55	1	1

Simulations - II

- Some underlying graph G on p variables with Laplacian L (with a latent block structure).
- The covariates form a *massive* Gaussian Free Field (GFF) on G, i.e. distributed as $\mathcal{N}(0, \Sigma)$, where

$$\Sigma = (L + \epsilon I)^{-1}.$$

- L can be estimated as before, or using graphical lasso.

Simulations - II

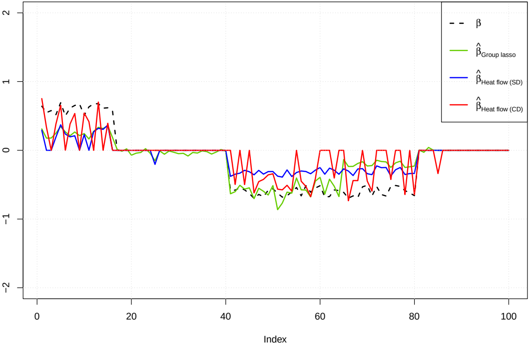


Figure 4: $n = 200, p = 100$, G is generated from a stochastic block model with parameters $K = 4$, $a = 0.5$ and $b = 0.01$.

Simulations - II

	Group lasso	Heat flow (SD)	Heat flow (CD)
Prediction error	0.09	0.12	0.12
Estimation error	2.24	2.59	2.93
Sensitivity	1.00	0.91	0.61
Specificity	0.36	0.98	0.98

Graph estimation from GFF samples

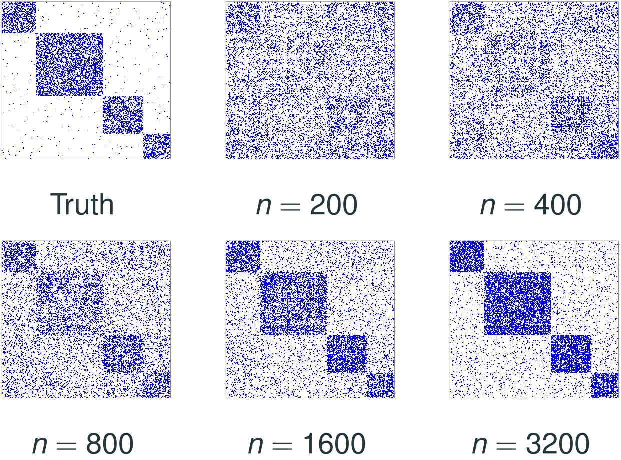


Figure 5: Graph estimated by thresholding an estimate of Σ from a GFF on a graph on $p = 200$ vertices generated from a stochastic block model with parameters $a = 0.5$, $b = 0.01$.

Predicting the monthly temperature at Delhi NCR

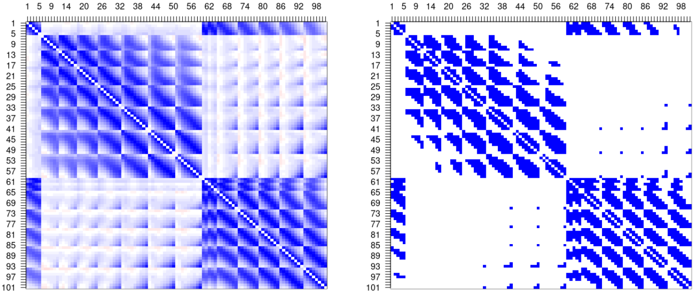
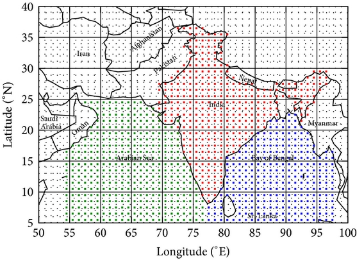
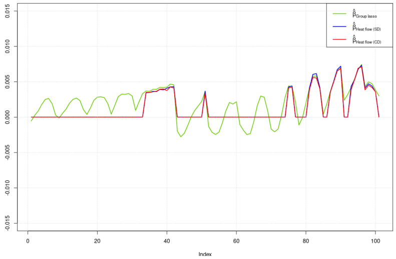


Figure 6: Correlation between monthly precipitation in $2.5^\circ \times 2.5^\circ$ squares on the Arabian Sea and the Bay of Bengal.

Predicting the monthly temperature at Delhi NCR



Random map off the internet



Estimated coefficients