A Projection free method for Generalized Eigenvalue Problem with a nonsmooth Regularizer

Seong Jae Hwang Maxwell D. Collins Sathya N. Ravi Vamsi K. Ithapu Nagesh Adluru Sterling C. Johnson Vikas Singh

University of Wisconsin - Madison

Objective

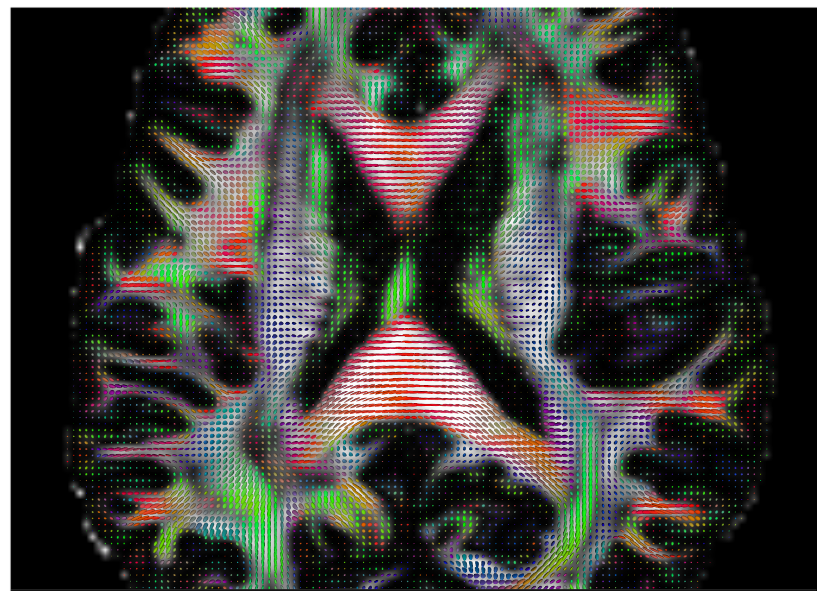
Multi-view statistical analysis of brain imaging data.

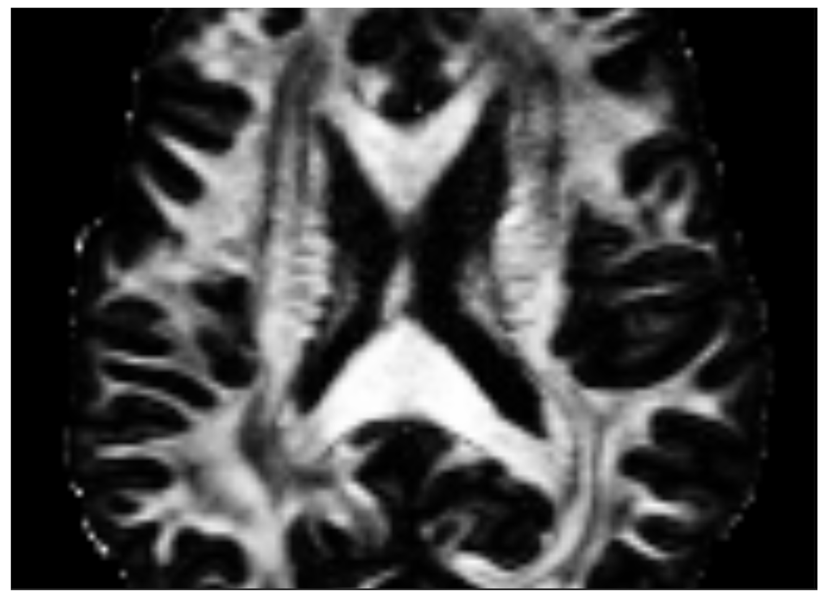
Motivation: Multi-view Data

- Common to see datasets with multiple views (e.g., photos with tags).
- In many cases, there is additional side or even partially available information which may be expensive.
- **Need a general purpose tool for multi-view datasets where some views are partial.**

High Level Setup

- Use **Regularized Generalized Eigenvalue Problem** (R-GEP) as a black-box to incorporate the following information:
  1. **Primary view** (e.g., 3D fractional anisotropy) of the dataset of interest (3D Diffusion Tensor Imaging).
  2. **Secondary view** (e.g., brain connectivity network) of the same dataset of interest.
  3. **Supplementary data** (e.g., cerebrospinal fluid) partially available from more expensive sources such as genotype.





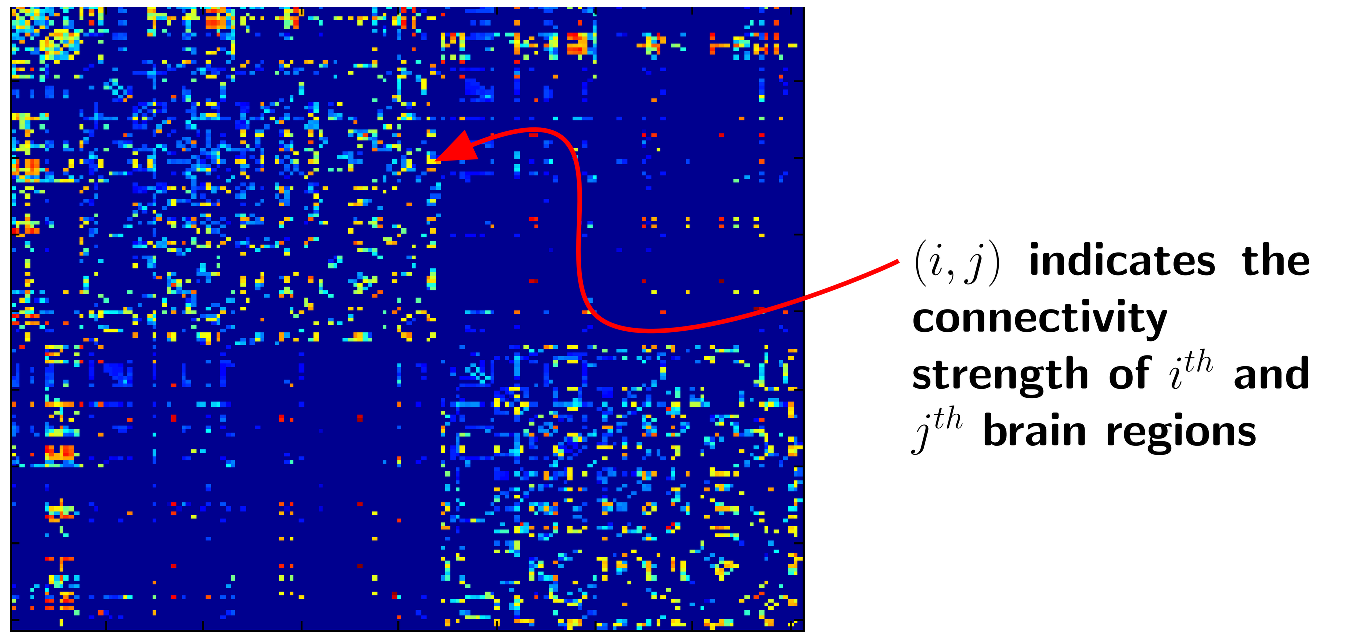


Figure : DTI image (top left) showing tensor directionality, followed by the FA image (top right) and the connectivity matrix (bottom).

Preliminary: Optimization on Manifolds

- Spectral analysis of the similarity/kernel matrix M by via eigendecomposition.
- Ordinary eigenvalue problem formulation where $\mathrm{tr}(\cdot)$ denotes the trace functional:

$$\min_{V \in \mathbb{R}^{N\times p}} \mathrm{tr}(V^TMV) \quad \text{s.t. } V^TV = I \tag{1}$$

- Finds p eigenvectors $V \in \mathbb{R}^{N\times p}$ corresponding to the first p smallest eigenvalues of M.
- $V^TV = I$ imposes an implicit optimization over the Grassmann manifold.

Generalized Eigenvalue Problem (GEP)

- The Generalized Eigenvalue Problem (GEP) involves a matrix pencil $\{M, D\}$ where D is often referred to as a *mass matrix*:

$$\min_{V \in \mathbb{R}^{N\times p}} f(V) := \mathrm{tr}(V^TMV) \quad \text{s.t. } V^TDV = I \tag{2}$$

- Finds p eigenvectors $V \in \mathbb{R}^{N\times p}$ corresponding to the first p smallest eigenvalues of M with respect to D.
- D is derived from a secondary view as a generalized Stiefel manifold constraint.

Regularized Generalized Eigenvalue Problem (R-GEP)

- n': a subset of the original subjects with supplementary information.
- $\mathcal{S}$: similarity matrix of the subjects in n'.
- α: leading eigenvector $\mathcal{S}$, or the "weights" on the subjects in n'.
- $V_{n',1}$: first eigenvector of M restricted to the set n'.
- R-GEP formulation using ℓ_1 norm:

$$\min_{V \in \mathbb{R}^{N\times p}} \mathrm{tr}(V^TMV) + \lambda||V_{n',1} - \alpha||_1 \quad \text{s.t. } V^TDV = I \tag{3}$$

Algorithm: Stochastic Block Coordinate Descent

Algorithm 1 Stochastic block coordinate descent on $\mathrm{GF}_{N,p}$

Require: $f : \mathrm{GF}_{N,p} \to \mathbb{R}$, $D \in \mathbb{R}^{N\times N}$, $V_0 \in \mathrm{GF}_{N,p}(D)$

```
for t = 1, ..., T do
    Select rows I ⊆ {1, ..., N}
    U_0 ← V_{I·} − D_{II}^{-1} D_{Ī I} V_{Ī·}
    [Q QR] ← U_0 for Q nonsingular
    Take G ∈ ∂_{V_{I·}} f(V)
    G' ← G_{IJ} + G_{I J̄} R^T
    Construct a descent curve Y on GF_{i,p}(D_{II}) through U_0 in the direction of −G'
    Pick step size τ_t satisfying Armijo-Wolfe condition
    V_{t+1} ← Y(τ_t)
end for
```

Subproblem Feasibility and Singularity Correction

- $\mathcal{I}, \bar{\mathcal{I}}$: a subset of selected i and non-selected (fixed) row indices of V respectively.
- Given any orthogonal $U \in \mathbb{R}^{i\times r}$, a new iterate satisfying the subproblem constraint is

$$V_{\mathcal{I}\cdot} = D_{\mathcal{I}\mathcal{I}}^{-\frac{1}{2}} U P^{\frac{1}{2}} - D_{\mathcal{I}\mathcal{I}}^{-1} D_{\bar{\mathcal{I}}\mathcal{I}}^T V_{\bar{\mathcal{I}}\cdot}. \tag{4}$$

- Assume w.l.o.g., for a $i \times r$ nonsingular matrix Q and a $r \times (p - r)$ matrix R,

$$V_{\mathcal{I}\cdot} + D_{\mathcal{I}\mathcal{I}}^{-1} D_{\bar{\mathcal{I}}\mathcal{I}}^T V_{\bar{\mathcal{I}}\cdot} = \left[Q \ \ QR\right]. \tag{5}$$

- For $\mathcal{J}$ column indices of Q, and U such that $U^TD_{\mathcal{I}\mathcal{I}}U = P_{\mathcal{J}\mathcal{J}}$, the following are feasible:

$$V_{\mathcal{I}\mathcal{J}} = U - D_{\mathcal{I}\mathcal{I}}^{-1} D_{\bar{\mathcal{I}}\mathcal{I}}^T V_{\bar{\mathcal{I}}\mathcal{J}}, \quad V_{\mathcal{I}\bar{\mathcal{J}}} = UR - D_{\mathcal{I}\mathcal{I}}^{-1} D_{\bar{\mathcal{I}}\mathcal{I}}^T V_{\bar{\mathcal{I}}\bar{\mathcal{J}}}. \tag{6}$$

Descent Curve Construction

- Differentiate $f \circ V(U)$ w.r.t. U, where V is related to U by (4):

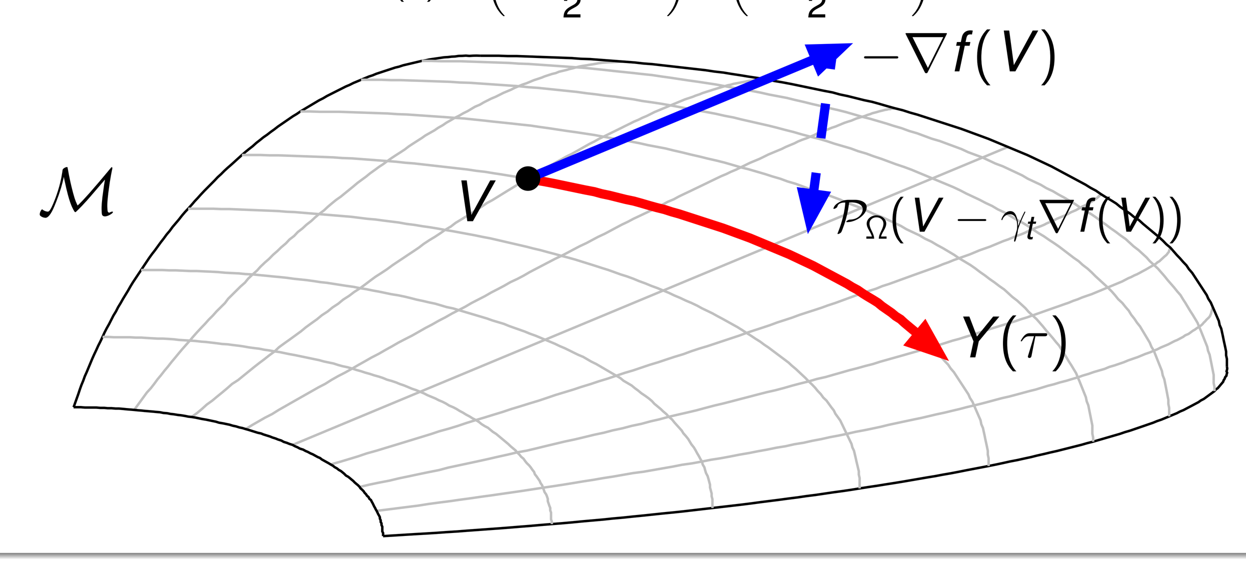
$$\frac{\partial}{\partial U} f \circ V(U) = 2(M_{\mathcal{I}\mathcal{I}} V_{\mathcal{I}\cdot} + M_{\mathcal{I}\bar{\mathcal{I}}} V_{\bar{\mathcal{I}}\cdot}) + \lambda g'(V(U)). \tag{7}$$

- Pick a subgradient $G \in \partial_{V_{\mathcal{I}\cdot}} f \circ V(U)$ and then perform the singularity correction on G:

$$G' = G_{\mathcal{I}\mathcal{J}} + G_{\mathcal{I}\bar{\mathcal{J}}} R^T. \tag{8}$$

- For $A = G'U_0^T - U_0G'^T$, curve Y as a function of τ by the Crank-Nicolson-like design is

$$Y(\tau) = \left(I + \frac{\tau}{2} A D_{\mathcal{I}\mathcal{I}}\right)^{-1} \left(I - \frac{\tau}{2} A D_{\mathcal{I}\mathcal{I}}\right) U_0. \tag{9}$$



Experiments: Brain Imaging Data

- **Wisconsin Alzheimer's Disease Research Center (ADRC) Dataset**
  - 102 middle-aged and older adults (58 healthy, 44 diseased).
  - **Primary view**: 3D volumetric Fractional Anisotropy (FA) imaging data summarizing (possibly losing information) the degree of diffusion of water in each voxel.
  - **Secondary view**: anatomical connectivity of regions derived from 3D Diffusion Tensor Images (DTI) using fiber counting procedures.
  - **Supplementary view**: 7 cerebrospinal fluid (CSF) scores measuring disease related proteins. Informative but partial (60 of 102 subjects) and cannot infer brain regions.

Experiments: Spectral Analysis on Brain Imaging Data

- **R-GEP Setup**
  - M: similarity matrix from FA in subject space.
  - D: CC^T where $C^{N\times r}$ is the subject space factor matrix of rank r of connectivity tensor.
  - α: leading eigenvector of the covariance matrix of CSF in subject space.
- **Comparison of classification/regression powers against three methods**
  1. Baseline model with primary view (FA) only.
  2. Intermediate models with PCA on FA and a GEP (2) without a regularizer.
  3. PCA-avg model where the primary and secondary source kernels are averaged.
- **Experimental Result**
  Introducing additional sources of information always increases the performance (63.4% accuracy for baseline to > 91% for R-GEP). Same is the case with regression results in Table 1 (0.68 correlation coefficient from baseline to > 0.78 for R-GEP).
- **Analysis**
  Based on the Incorporating secondary and incomplete priors increase performance, and our R-GEP model combines these information sources in a meaningful way offering good improvements.

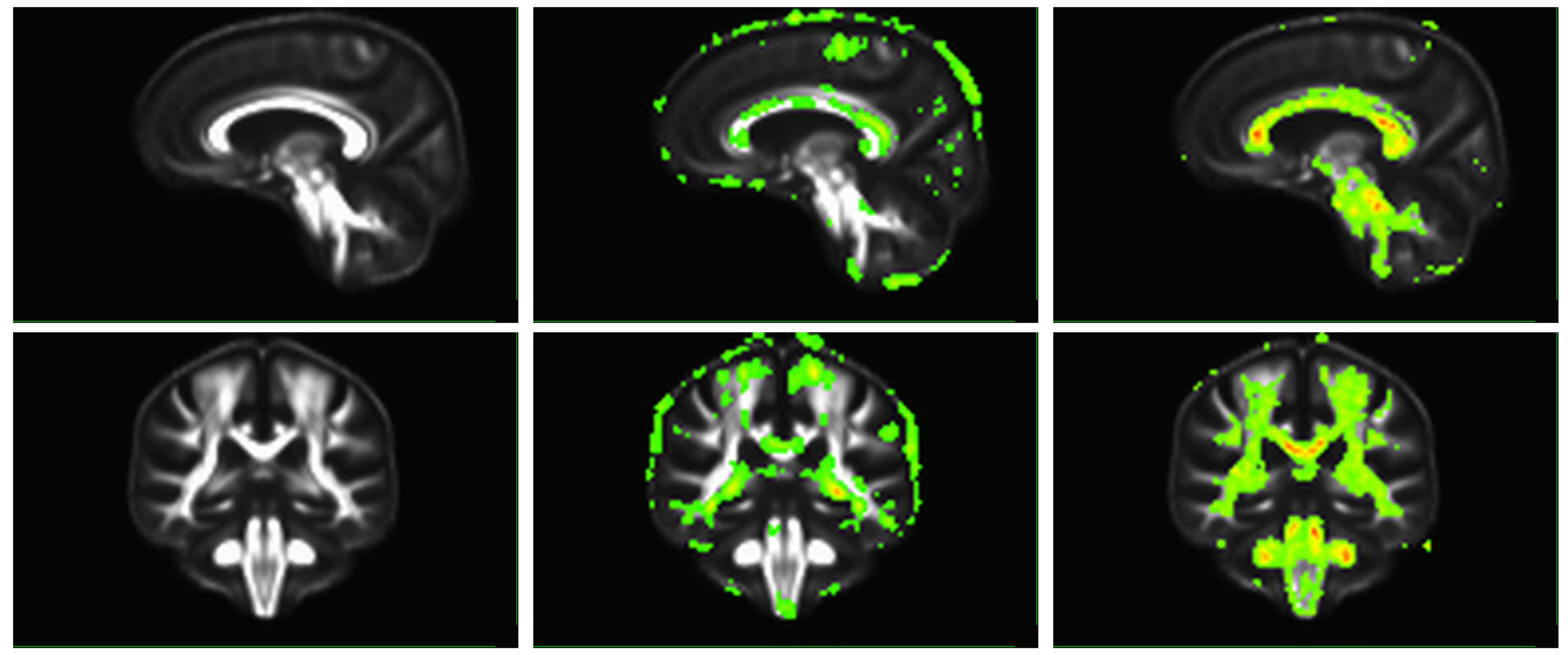


Figure : Feature sensitivity. First column shows the FA image. Second column shows overlays of the weights assigned by baseline linear kernel on this FA image. Last column shows overlays from the base R-GEP case in Table. Green (Red) corresponds to smaller (larger) weights.

p	Baseline	PCA	PCA-avg	GEP (2)			R-GEP (3)		
				$r=1$	5	10	$r=1$	5	10
3	63.41	85.71	85.60	86.62	85.71	85.71	**90.34**	88.53	86.53
5	63.41	82.60	81.69	85.41	85.41	86.32	**89.55**	89.34	87.34
7	63.41	84.30	80.69	84.30	84.30	84.30	86.43	**88.43**	87.53
10	63.41	82.49	83.49	82.39	84.21	86.23	86.53	86.43	**89.32**
13	63.41	83.32	85.62	86.14	84.23	88.14	89.23	**90.05**	88.23

p	Baseline	PCA	PCA-avg	GEP (2)			R-GEP (3)		
				$r=1$	5	10	$r=1$	5	10
3	0.679	0.718	0.647	0.719	0.718	0.718	0.745	**0.771**	0.758
5	0.679	0.719	0.614	0.726	0.737	0.735	**0.769**	0.746	0.749
7	0.679	0.707	0.610	0.707	0.707	0.713	0.763	**0.785**	0.734
10	0.679	0.656	0.622	0.656	0.742	0.719	0.741	**0.762**	0.754
13	0.679	0.717	0.654	0.730	**0.765**	0.745	0.737	0.757	0.754

Table : Healthy vs. diseased classification accuracy (top) and regression correlation coefficient (bottom) (10-fold cross validated) using GEP and R-GEP, compared to the baseline linear classifier and the PCA setup. p is the PCA rank and r is tensor rank.

Acknowledgment

SJH was supported by a University of Wisconsin CIBM fellowship (5T15LM007359-14). We acknowledge support from NIH grants AG040396 and AG021155, NSF RI 1116584 and NSF CAREER award 1252725, as well as UW ADRC (AG033514), UW ICTR (1UL1RR025011), Waisman Core grant (P30 HD003352-45), UW CPCP (AI117924) and NIH grant AG027161.