Speeding up Permutation Testing in Neuroimaging

Chris Hinrichs[†⋆] Vamsi Ithapu[†⋆] Qinyuan Sun[†] Sterling C. Johnson[‡,†] Vikas Singh[†]
[†] University of Wisconsin Madison [‡] William S. Middleton Memorial VA Hospital
http://pages.cs.wisc.edu/~vamsi/pt_fast

Contribution

A novel framework for reducing the computational burden in permutation testing procedure, without sacrificing the fidelity of the estimated Family Wise Error Rate (FWER).

Permutation Testing

Non–parametric random sampling method for estimating test statistic distribution under Global Null hypothesis.

Setup : Given data (v dimensions/features) for n different subjects/instances from two groups/classes.

- Construct the $v \times T$ matrix of univariate test statistics (denoted by $\mathbf{P}$) by *randomly* permuting group labels T times.

Under the Null, distribution of *max* test–statistic is the histogram of the maximum of m entries across each permutation (column of $\mathbf{P}$).

Drawback : For large v, larger T give better estimates (random sampling methods sample often at mode(s) than tails).

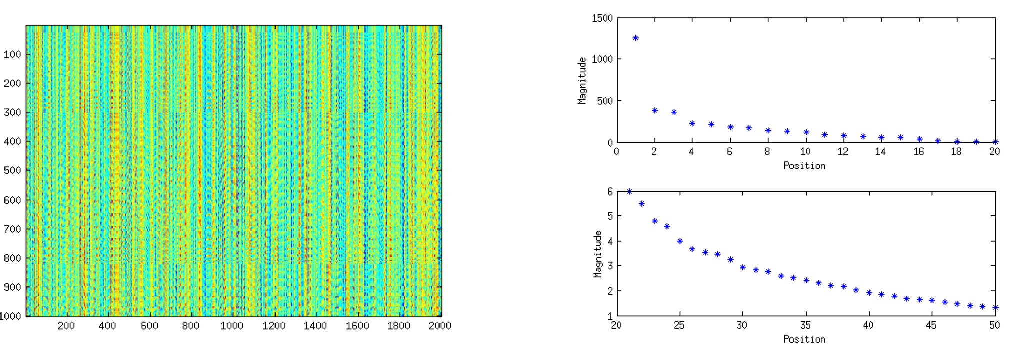
- Neuroimaging : $v \sim 3 \times 10^5$ and $T \sim 10^5$
- Bioinformatics : $v \sim 10^3$ and $T \sim 10^4$

Brute–force computation of $\mathbf{P}$ takes days to weeks in general.

Observation

$\mathbf{P}$ can be decomposed as the sum of a low–rank signal and a high–rank residual.

Example : $v = 1000, T = 2000$ with t–statistic.



The low–rank structure comes from highly correlated covariates (dimensions). The high–rank residual results from the non–linearity in test statistics computation.

Model

$$\mathbf{P} = \mathbf{UW} + \mathbf{S} \qquad \mathbf{P}, \mathbf{UW}, \mathbf{S} \in \mathbb{R}^{v \times T}$$

$\mathbf{U} \in \mathbb{R}^{v \times r}$: orthogonal matrix ($r \ll min(v, T)$)
$\mathbf{W} \in \mathbb{R}^{r \times T}$: coefficient matrix
$\mathbf{S}_{i,j} \sim \mathcal{N}(0, \sigma^2)$ (residual matrix)

Estimating UW :
Low–rank subspace recovery from subsampled data where Ω is the subsampling rate.

$$\min_{\mathbf{U},\mathbf{W}} \|\mathbf{P}_\Omega - \tilde{\mathbf{P}}_\Omega\|_F^2 \quad s.t. \quad \tilde{\mathbf{P}} = \mathbf{UW}$$

Recovering S : *Bias–Variance Dilemma*
A reliable estimate of $\mathbf{S}$ is obtained when the entire matrix $\mathbf{P}$ is used ($\Omega = 1$). For a very sparse subsampling ($\Omega \ll 1$), the variance of $\mathbf{S}$ is grossly underestimated – a sampling artifact that induces a shift/bias in the distribution of sample maximum.

- *Training period* : A training phase is used to estimate this bias by computing *all* entries of $\mathbf{P}$ for $T_b \ll T$ permutations. The recovered sample max distribution is shifted by this estimated bias.

Theoretical Guarantees

Low–rank Perturbation. *Denote that r non-zero eigenvalues of $\mathbf{Q} = \mathbf{UWW}^T\mathbf{U}^T \in \mathbb{R}^{v \times v}$ by $\lambda_1 \geq \lambda_2 \geq, \ldots, \lambda_r > 0$; and let $\mathbf{S}$ be a $v \times T$ random matrix such that $\mathbf{S}_{i,j} \sim \mathcal{N}(0, \sigma^2)$, with unknown σ^2. As $v, T \to \infty$ such that $\frac{v}{T} \ll 1$, the eigenvalues $\tilde{\lambda}_i$ of the perturbed matrix $\mathbf{Q} + \mathbf{SS}^T$ will satisfy*

$$|\tilde{\lambda}_i - \lambda_i| < \delta\lambda_i \qquad i = 1, \ldots, r;$$

$$\tilde{\lambda}_i < \delta\lambda_r \qquad i = r+1, \ldots, v$$

for some $0 < \delta < 1$, whenever $\sigma^2 < \frac{\delta\lambda_r}{T}$

Max Null Recovery. *Let $m_t = \max_i P_{i,t}$ be the maximum observed test statistic at permutation trial t, and similarly let $\hat{m}_t = \max_i \hat{P}_{i,t}$ be the maximum reconstructed test statistic. Further, let the maximum reconstruction error be ϵ, such that $|P_{i,t} - \hat{P}_{i,t}| \leq \epsilon$. Then, for any real number $k > 0$, we have,*

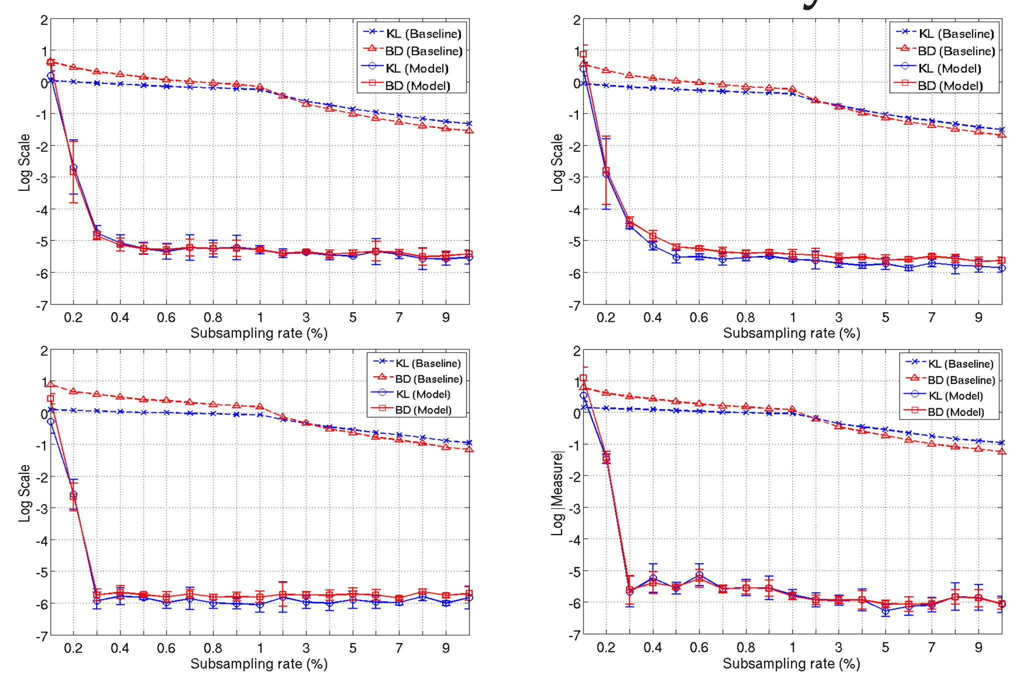
$$Pr\left[m_t - \hat{m}_t - (b - \hat{b}) > k\epsilon\right] < \frac{1}{k^2}$$

where b is the bias due to subsampling, and $\hat{b}$ is its estimate from the training phase.

Evaluations

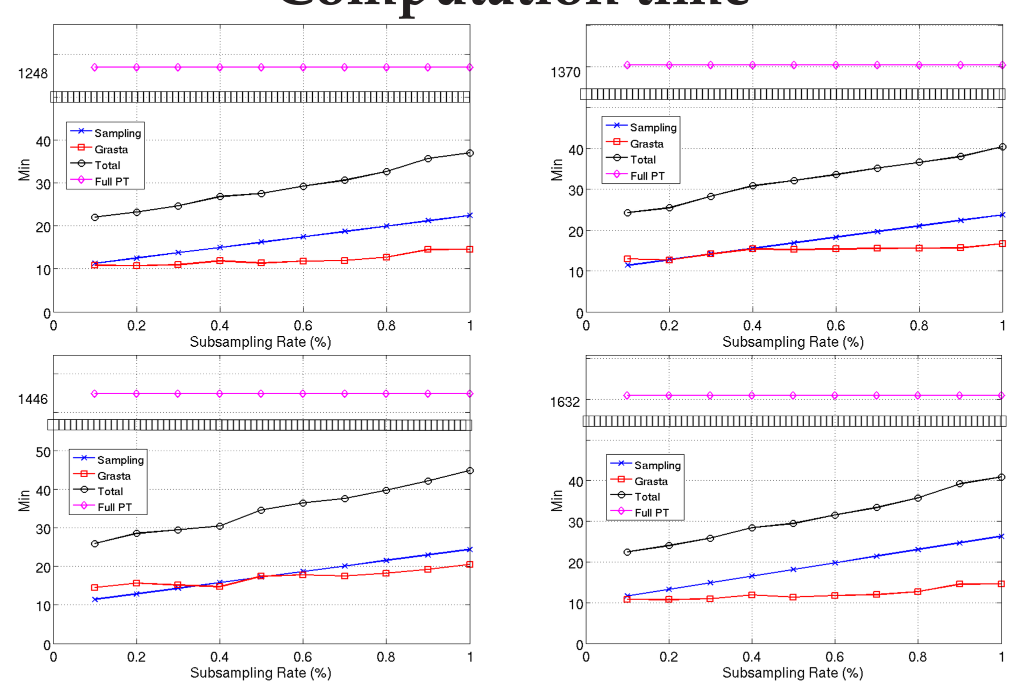
- Healthy vs. demented subjects from 4 datasets ($n = 40, 50, 55, 70$, $v \sim 2.5 \times 10^5$ and $T = 10^4$).
- KL divergence and Bhattacharya Distance (BD) measure reliability of estimated max Null.
- GRASTA algorithm (He et. al. 2011) was used for low–rank subspace recovery. Baseline model estimates max Null directly from subsampled entries.

Maximum Null recovery



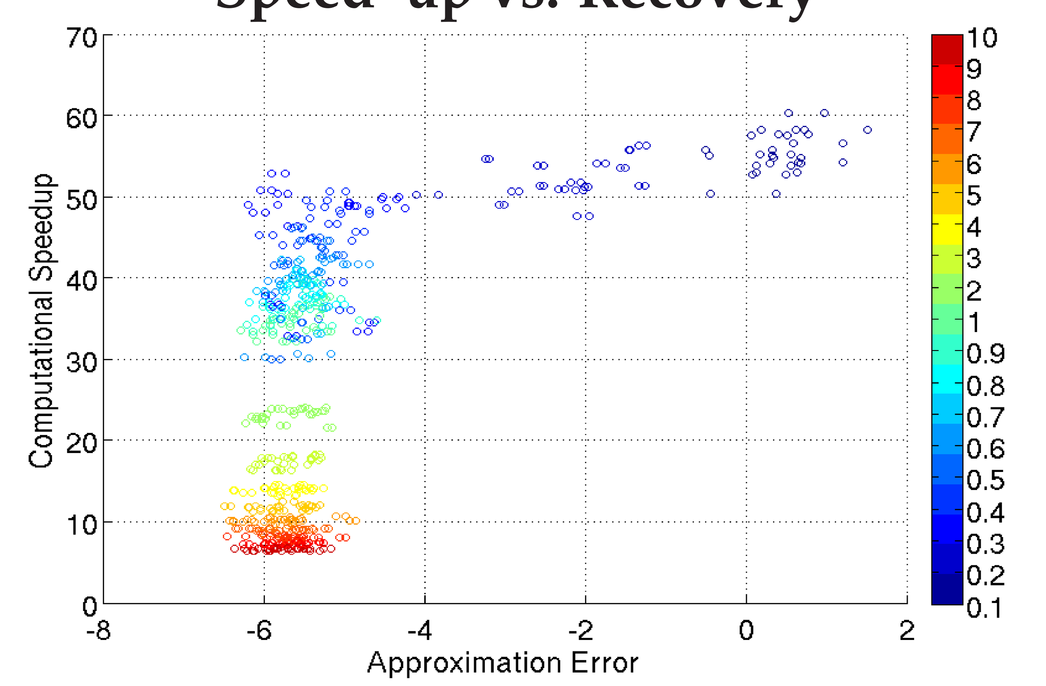
KL and BD of the recovered Null to the true distribution are $< 1e^{-5}$ for sampling rates $> 0.4\%$.

Computation time



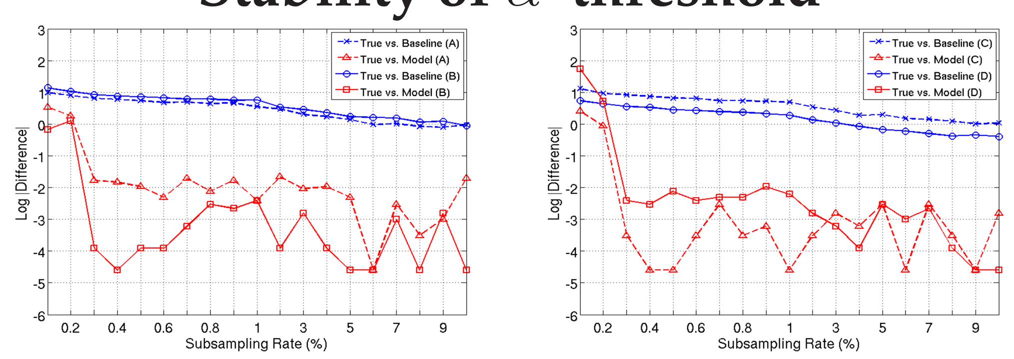
At 0.3% subsampling (where KL and BD are $< 1e^{-4.3}$) the speedup was $> 50\times$.

Speed–up vs. Recovery



At least $30\times$ speedup is achieved in the low sampling regime ($< 1\%$). Around $0.5 - 0.6\%$ subsampling (where KL and BD are $< 1e^{-5}$), the speedup factor averaged over all datasets was $45\times$. Observe the trade–off between speedup factor and approximation error.

Stability of α–threshold



At sampling rates $> 3\%$, mean and maximum difference in $\alpha = 0.95$ threshold over all datasets was 0.04 and 0.18 respectively.

Conclusion

- We approximate the permutation testing matrix by first recovering the major singular vectors followed by estimating the distribution of the residuals.
- The max Null statistic distribution is recovered to a very high degree of accuracy while achieving a computational speed–up of roughly $50\times$.

[⋆] Hinrichs and Ithapu are joint first authors

Advances in Neural Information Processing Systems (NIPS) 2013