Risk Classification with an Adaptive Naive Bayes Kernel Machine Model

Jessica Minnier[1],
Ming Yuan[3], Jun Liu[4], and Tianxi Cai[2]

[1]Department of Public Health & Preventive Medicine, Oregon Health & Science University
[2]Department of Biostatistics, Harvard School of Public Health
[3]Department of Statistics, University of Wisconsin-Madison
[4]Department of Statistics, Harvard University

June 25, 2014

Outline

Background and Motivation

Adaptive Naive Bayes (Blockwise) Kernel Machine Classification

- Goal: genetic data $\rightarrow$ quantify disease risk, predict therapeutic efficacy, determine disease subtypes
- Goal: build an accurate parsimonious prediction model
  - reduce the cost of unnecessary marker measurements
  - improve the prediction precision for future patients
  - improve over modest prediction precision obtained with clinical predictors and/or known risk alleles

Background and Motivation

Adaptive Naive Bayes (Blockwise) Kernel Machine Classification

- Goal: genetic data $\rightarrow$ quantify disease risk, predict therapeutic efficacy, determine disease subtypes
- Goal: build an accurate parsimonious prediction model
  - reduce the cost of unnecessary marker measurements
  - improve the prediction precision for future patients
  - improve over modest prediction precision obtained with clinical predictors and/or known risk alleles
- Complex diseases
  - many alleles contribute to risk
  - many distinct combinations of risk factors lead to disease

Background and Motivation

- Genome wide association studies (GWAS)
  - identifying SNPs associated with disease risk
  - primary goal of testing
  - accurate risk prediction remains difficult
- Common approach:
  - select top ranked SNPs based on large scale testing
  - construct a composite genetic score w/ selected SNPs

Background and Motivation

- Genome wide association studies (GWAS)
  - identifying SNPs associated with disease risk
  - primary goal of testing
  - accurate risk prediction remains difficult
- Common approach:
  - select top ranked SNPs based on large scale testing
  - construct a composite genetic score w/ selected SNPs
  - may not work well due to
    - ⋆ false $+/-$ errors in identifying predictive SNPs
    - ⋆ over-fitting
    - ⋆ using only subset of SNPs available
    - ⋆ additive effects only

Background and Motivation

Recent progress in prediction with high dimensional data

- **Regularized estimation**: LASSO (Tibshirani, 1996); SCAD (Fan and Li, 2001); Adaptive LASSO (Zou, 2006)
- **Machine learning**: Support vector machine (Cristianini, Shawe-Taylor, 2000); Least square Kernel Machine Regression (Liu, Lin, Ghosh, 2007); **Kernel logistic regression** (Zhu and Hastie, 2005; Liu, Ghosh and Lin, 2008)
- **Screening + Regularized estimation**: Sure independence screening (Fan and Lv, 2008; Fan and Song, 2009)

⋆ Global methods: may be unstable for large p, high correlation

Approach

Challenge:

- Prediction models based on univariate testing, additive models, global methods $\rightarrow$ low prediction accuracy, low AUC, missing heritability
- Non-linear effects? testing for interactions $\rightarrow$ low power

Approach

Challenge:

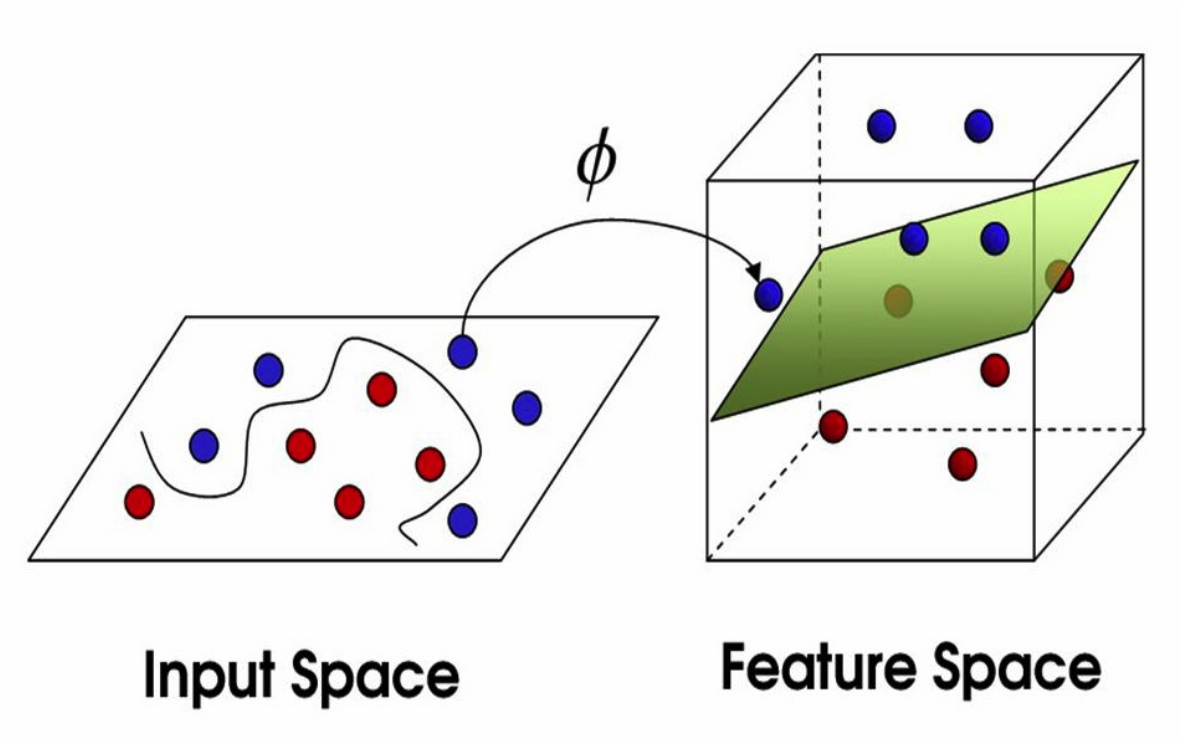
- Prediction models based on univariate testing, additive models, global methods $\rightarrow$ low prediction accuracy, low AUC, missing heritability
- Non-linear effects? testing for interactions $\rightarrow$ low power

Our approach [Minnier et al., 2014]:

- **Blockwise method**:
  - $\star$ leverage biological knowledge to build models at the gene-set level
  - $\star$ genes, gene-pathways, linkage disequilibrium blocks
- **Kernel machine regression:**
  - $\star$ allow for complex and nonlinear effects
  - $\star$ implicitly specify underlying complex functional form of covariate effects via similarity measures (kernels) that define the distance between two sets of covariates

Kernel Methods: similar inputs to similar outputs

- transform data to feature space $\mathcal{H}$ with non-linear map ϕ
- "kernel trick" lets us use $K(,)$ similarity function instead of ϕ
- K induces the feature space



⋆ N. Takahashi's webpage

Previous Methods

Blockwise methods

- Inference: Gene-set testing
  - ⋆ Gene burden tests
  - ⋆ Gene Set Enrichment Analysis (GSEA)
  - ⋆ SNP-set Sequence Kernel Association Test (SKAT, SKAT-O; Wu et al. 2010; Wu, Lee, et al. 2011)

Previous Methods

Blockwise methods

- Inference: Gene-set testing
  - ⋆ Gene burden tests
  - ⋆ Gene Set Enrichment Analysis (GSEA)
  - ⋆ SNP-set Sequence Kernel Association Test (SKAT, SKAT-O; Wu et al. 2010; Wu, Lee, et al. 2011)

Kernel machine methods

- Support Vector Machine (SVM) classification methods
- Inference
  - ⋆ KM SNP-set Testing (Liu et al. 2007, 2008; SKAT methods)
  - ⋆ Gene expression test with kernel Cox model (Li and Luan 2003)

Notations and Model Assumptions

- Data
  - Response: $\mathbf{Y} = (Y_1, ..., Y_n)^{\mathsf{T}}$
  - Predictors: M **blocks** of genomic regions, for $b = 1, ..., M$,

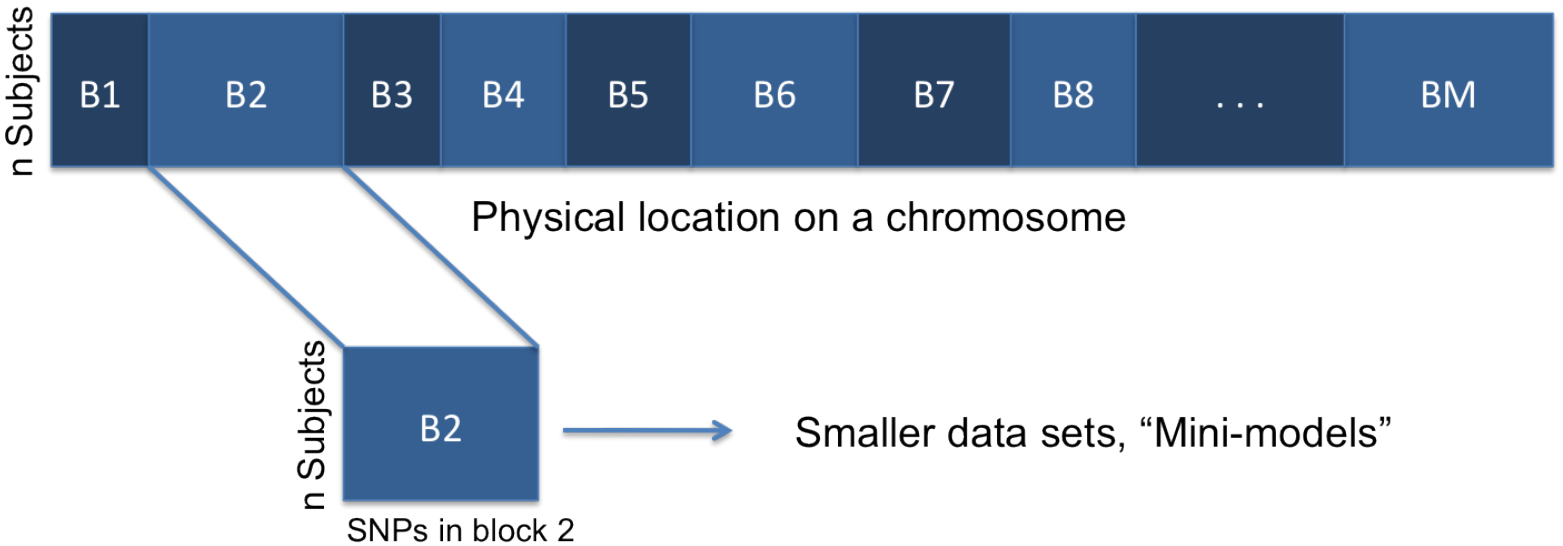
$$\mathbb{X}^{(\mathrm{b})} = (\mathbf{X}_1^{(\mathrm{b})}, ..., \mathbf{X}_n^{(\mathrm{b})})_{n \times p_b}^{\mathsf{T}},$$

Notations and Model Assumptions

- Data
  - Response: $\mathbf{Y} = (Y_1, ..., Y_n)^{\mathsf{T}}$
  - Predictors: M **blocks** of genomic regions, for $b = 1, ..., M$,

$$\mathbb{X}^{(b)} = (\mathbf{X}_1^{(b)}, ..., \mathbf{X}_n^{(b)})_{n \times p_b}^{\mathsf{T}},$$

- **Blockwise**: Partition genome into gene-sets
  - Recombination hotspots, gene-pathways



Notations and Model Assumptions

- Data
  - Response: $\mathbf{Y} = (Y_1, ..., Y_n)^{\mathsf{T}}$
  - Predictors: M blocks of genomic regions, for $b = 1, ..., M$,

$$\mathbb{X}^{(b)} = (\mathbf{X}_1^{(b)}, ..., \mathbf{X}_n^{(b)})_{n \times p_b}^{\mathsf{T}},$$

- Model under *blockwise Naive Bayes (NB)* assumption:

$$\mathbb{X}^{(1)}, ..., \mathbb{X}^{(M)} \mid \mathbf{Y} \quad \text{independent}$$

Notations and Model Assumptions

- Data
  - Response: $\mathbf{Y} = (Y_1, ..., Y_n)^\top$
  - Predictors: M blocks of genomic regions, for $b = 1, ..., M$,

$$\mathbb{X}^{(b)} = (\mathbf{X}_1^{(b)}, ..., \mathbf{X}_n^{(b)})_{n \times p_b}^\top,$$

- Model under *blockwise Naive Bayes (NB)* assumption:

$$\mathbb{X}^{(1)}, ..., \mathbb{X}^{(M)} \mid \mathbf{Y} \quad \text{independent} \qquad \Rightarrow$$

$$\text{logit}\{\text{pr}(Y = 1 \mid \mathbf{X}^{(1)}, ..., \mathbf{X}^{(M)})\} = \text{c} + \sum_{b=1}^{M} \text{logit}\{\text{pr}(Y = 1 \mid \mathbf{X}^{(b)})\}$$

  - NB assumption allows *separate estimation* by block and *reduces overfitting*
  - Performs well for zero-one loss $L(X) = I(\hat{Y}(X) \neq Y)$ [Domingos and Pazzani, 1997]

Notations and Model Assumptions

- Within each region, the effect may be complex and interactive due to
  - multiple causal variants
  - un-typed causal variants in the presence of high LD

Notations and Model Assumptions

- Within each region, the effect may be complex and interactive due to
  - multiple causal variants
  - un-typed causal variants in the presence of high LD

- Blockwise Kernel Machine Regression

$$\text{logit}\{\text{pr}(Y = 1 \mid \mathbf{X}^{(b)})\} = a^{(b)} + h^{(b)}(\mathbf{X}^{(b)})$$

$$h^{(b)}(\mathbf{X}^{(b)}) = \sum_l \beta_l^{(b)} \psi_l^{(b)}(\mathbf{X}^{(b)}) \in \mathcal{H}_{K^{(b)}}$$

  - $\star$ $\{\psi_l^{(b)}\} = \{\sqrt{\lambda_l^{(b)}}\phi_l^{(b)}\}$ implicitly specified via a *symmetric positive definite kernel* $K^{(b)}(\cdot, \cdot)$.

Notations and Model Assumptions

- Within each region, the effect may be complex and interactive due to
  - multiple causal variants
  - un-typed causal variants in the presence of high LD

- Blockwise Kernel Machine Regression

$$\text{logit}\{\text{pr}(Y = 1 \mid \mathbf{X}^{(b)})\} = a^{(b)} + h^{(b)}(\mathbf{X}^{(b)})$$

$$h^{(b)}(\mathbf{X}^{(b)}) = \sum_l \beta_l^{(b)} \psi_l^{(b)}(\mathbf{X}^{(b)}) \in \mathcal{H}_{K^{(b)}}$$

  - $\star$ $\{\psi_l^{(b)}\} = \{\sqrt{\lambda_l^{(b)}}\phi_l^{(b)}\}$ implicitly specified via a *symmetric positive definite kernel* $K^{(b)}(\cdot, \cdot)$.
  - $\star$ $K^{(b)}(\mathbf{X}_i^{(b)}, \mathbf{X}_j^{(b)})$ defines the similarity between $\mathbf{X}_i^{(b)}$ and $\mathbf{X}_j^{(b)}$.

Notations and Model Assumptions

- Within each region, the effect may be complex and interactive due to
  - multiple causal variants
  - un-typed causal variants in the presence of high LD

- Blockwise Kernel Machine Regression

$$\text{logit}\{\text{pr}(Y = 1 \mid \mathbf{X}^{(b)})\} = a^{(b)} + h^{(b)}(\mathbf{X}^{(b)})$$

$$h^{(b)}(\mathbf{X}^{(b)}) = \sum_l \beta_l^{(b)} \psi_l^{(b)}(\mathbf{X}^{(b)}) \in \mathcal{H}_{K^{(b)}}$$

  - $\star$ $\{\psi_l^{(b)}\} = \{\sqrt{\lambda_l^{(b)}}\phi_l^{(b)}\}$ implicitly specified via a *symmetric positive definite kernel* $K^{(b)}(\cdot,\cdot)$.
  - $\star$ $K^{(b)}(\mathbf{X}_i^{(b)}, \mathbf{X}_j^{(b)})$ defines the similarity between $\mathbf{X}_i^{(b)}$ and $\mathbf{X}_j^{(b)}$.
  - $\star$ $\mathcal{H}_{K^{(b)}}$, the functional space spanned by $K^{(b)}(\cdot,\cdot)$, is a reproducible kernel hilbert space (RKHS)

Choices of Kernel Functions

⋆ Linear kernel: $K(\mathbf{X}_i, \mathbf{X}_j) = \rho + \mathbf{X}_i^\top \mathbf{X}_j$,

$$h(\mathbf{X}) = \sum_{k=1}^{p} \widetilde{\beta}_k X_k$$

♯ Fitting logistic regression with linear kernel $\Leftrightarrow$ logistic ridge regression.

Choices of Kernel Functions

⋆ Linear kernel: $K(\mathbf{X}_i, \mathbf{X}_j) = \rho + \mathbf{X}_i^{\mathsf{T}} \mathbf{X}_j$,

$$h(\mathbf{X}) = \sum_{k=1}^{p} \widetilde{\beta}_k X_k$$

♯ Fitting logistic regression with linear kernel $\Leftrightarrow$ logistic ridge regression.

⋆ IBS kernel: $K(\mathbf{X}_i, \mathbf{X}_j) = \sum_{k=1}^{p} (2 - |X_{ik} - X_{jk}|)$,

♯ powerful in detecting non-linear effects with SNP data [Wu et al, 2010]

Estimation of h: Kernel PCA

- primal form: $h = \sum_l \beta_l \psi_l = \sum_l \beta_l \sqrt{\lambda_l} \phi_l$
- Kernel PCA approximation:

$$\mathbb{K} = [K(\mathbf{X}_i, \mathbf{X}_j)]_{1 \le i,j \le n} = \sum_{l=1}^{n} \widehat{\lambda}_l \widehat{\boldsymbol{\phi}}_l \widehat{\boldsymbol{\phi}}_l^{\mathsf{T}}$$

$$\widetilde{\mathbb{K}} = \sum_{l=1}^{\ell_0} \widehat{\lambda}_l \widehat{\boldsymbol{\phi}}_l \widehat{\boldsymbol{\phi}}_l^{\mathsf{T}} = \widehat{\boldsymbol{\Psi}} \widehat{\boldsymbol{\Psi}}^{\mathsf{T}}; \qquad \widehat{\boldsymbol{\Psi}} = [\widehat{\lambda}_1^{\frac{1}{2}} \widehat{\boldsymbol{\phi}}_1, ..., \widehat{\lambda}_{\ell_0}^{\frac{1}{2}} \widehat{\boldsymbol{\phi}}_{\ell_0}]_{n \times \ell_0}$$

⋆ Scholkopf et al. [1999]; Williams and Seeger [2000]; Braun et al. [2008]; Zhang et al. [2010]

- $\widehat{h}^{(\mathrm{b})}(\mathbf{X}^{(\mathrm{b})}) = \widehat{\boldsymbol{\Psi}} \widehat{\boldsymbol{\beta}}$
- obtain $(\widehat{a}, \widehat{\boldsymbol{\beta}})$ as the minimizer of ridge logistic objective function

$$\mathcal{L}(Y, a, \widehat{\boldsymbol{\Psi}} \boldsymbol{\beta}) + \tau \|\boldsymbol{\beta}\|^2$$

Regularized Selection of Informative Regions

- For $b = 1, ..., M$, perform kernel PCA regression and obtain $\widehat{h}^{(b)}$

$$\text{logit}\{\text{pr}(Y = 1 \mid \mathbf{X}^{(b)})\} = a^{(b)} + h^{(b)}(\mathbf{X}^{(b)})$$

- Classify a future subject with $\mathcal{X} = \{\mathbf{X}^{(b)}, b = 1, ..., M\}$ based on

$$\sum_{b=1}^{M} \widehat{h}^{(b)}(\mathbf{X}^{(b)}) \geq c$$

- **Final prediction rule** with **weighted block effects**
  - Some regions may not be predictive of the outcome due to false discovery
  - Inclusion of all regions for prediction may lead to reduced accuracy
  - Regularized estimation of block effects using LASSO:

$$\sum_{b=1}^{M} \widehat{\gamma}_b \widehat{h}^{(b)}(\mathbf{X}^{(b)}) \geq c$$

Regularized Selection of Informative Regions

- For $b = 1, ..., M$, perform kernel PCA regression and obtain $\widehat{h}^{(b)}$

$$\text{logit}\{\text{pr}(Y = 1 \mid \mathbf{X}^{(b)})\} = a^{(b)} + h^{(b)}(\mathbf{X}^{(b)})$$

- Classify a future subject with $\mathcal{X} = \{\mathbf{X}^{(b)}, b = 1, ..., M\}$ based on

$$\sum_{b=1}^{M} \widehat{h}^{(b)}(\mathbf{X}^{(b)}) \geq c$$

- **Final prediction rule** with **weighted block effects**
  - Regularized estimation of block effects using LASSO, pseudo-data $\widehat{\mathbb{H}}$ estimated with cross-validation:

$$\sum_{k=1}^{K} \left[\mathbf{Y}^{\mathsf{T}} \log g(b + \widehat{\mathbb{H}}\boldsymbol{\gamma}) + (1 - \mathbf{Y})^{\mathsf{T}} \log\{1 - g(b + \widehat{\mathbb{H}}\boldsymbol{\gamma})\} \right] - \tau_2 \|\boldsymbol{\gamma}\|_1,$$

$$\sum_{b=1}^{M} \widehat{\gamma}_b \widehat{h}^{(b)}(\mathbf{X}^{(b)}) \geq c$$

Theoretical Results

- Consistency of $\widehat{h}^{(b)}(\mathbf{x})$:
  - $\widehat{h}^{(b)}(\mathbf{x}) \to h^{(b)}(\mathbf{x})$ at $\sqrt{n}$ rate for finite dimensional $\mathcal{H}_K$
  - Relies on convergence of sample eigen-values and -vectors from kernel PCA to the true eigensystem of $\mathcal{H}_K$

$$\widehat{\mathbf{\Psi}} \to \mathbf{\Psi} = \{\psi_1^{(b)}, \ldots, \psi_{\ell_0}^{(b)}\}$$

- Oracle property of $\widehat{\gamma}$:
  - Gene-set selection consistency

$$P(\widehat{\mathcal{A}} = \mathcal{A}) \to 1$$

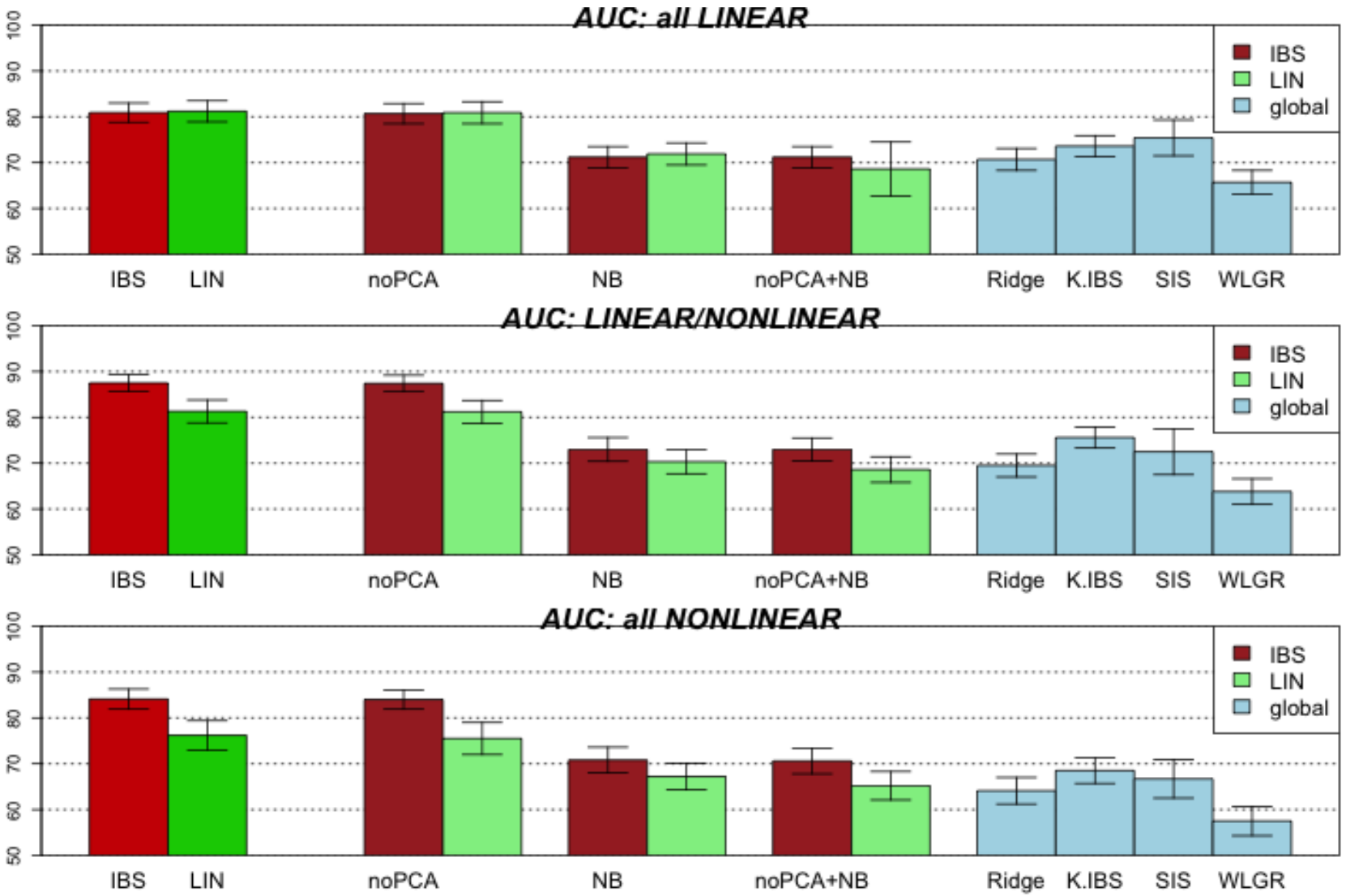
where $\widehat{\mathcal{A}} = \{b | \widehat{h}^{(b)}(\mathbf{x}) \neq 0\}, \mathcal{A} = \{b | h^{(b)}(\mathbf{x}) \neq 0\}$

Simulation Studies for NBKM

- SNP data sampled from gene-sets in a GWAS dataset (from type I diabetes study, Affy 500k)
- 350 regions, 9256 SNPs
- Only the first 4 regions are associated with the outcome
- the joint effects of the SNPs in each of these regions set as
  - linear for the first two regions and non-linear for the other 2 regions
  - linear for all 4 regions
  - nonlinear for all 4 regions

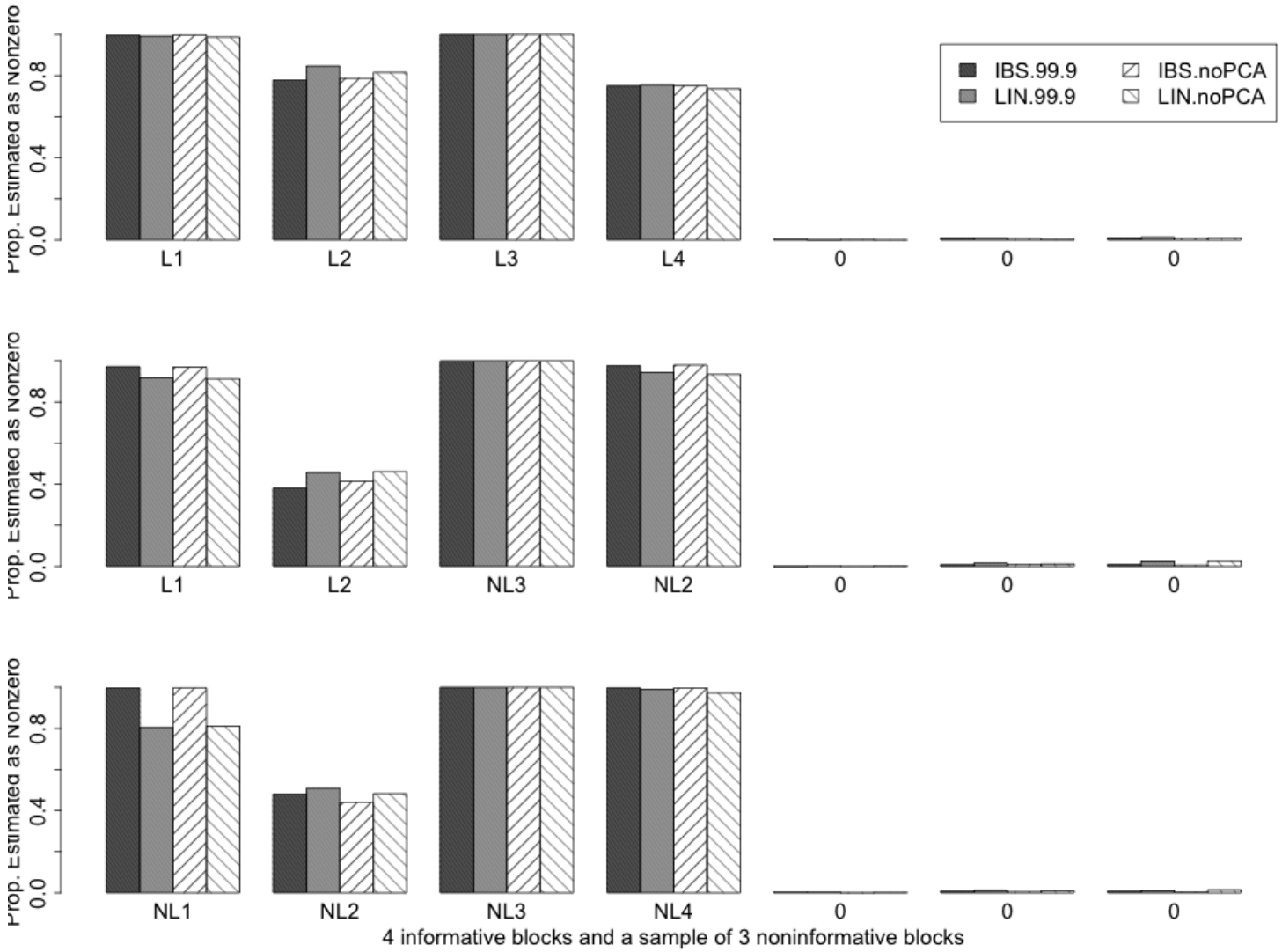
Prediction Accuracy

Simulations: $n_t = 1000, n_v = 500,$ # of genes $= 350$ total # of SNPs $= 9256$



Gene-set selection

Simulations: $n_t = 1000, n_v = 500,$ # of genes $= 350$ total # of SNPs $= 9256$



Genetic Risk of Type I Diabetes

- Autoimmune disease, usually diagnosed in childhood
- T1D
  - 75 SNPs have been identified as T1D risk alleles (National Human Genome Research Institute, Hindorff et al. [2009])
  - 91 genes that either contain these SNPs or flank the SNP on either side on the chromosome

Genetic Risk of Type I Diabetes

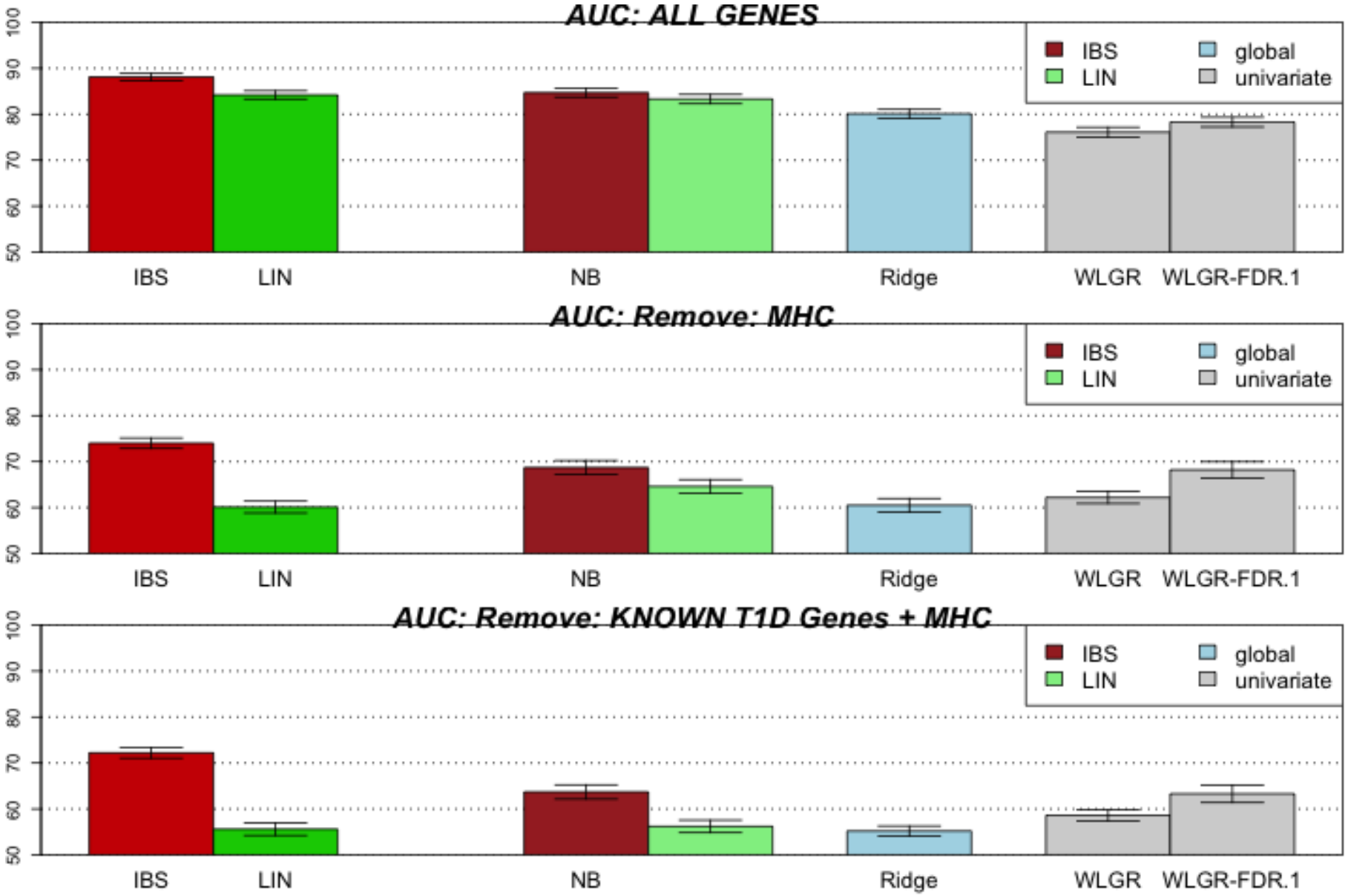
- Autoimmune disease, usually diagnosed in childhood
- T1D
  - 75 SNPs have been identified as T1D risk alleles (National Human Genome Research Institute, Hindorff et al. [2009])
  - 91 genes that either contain these SNPs or flank the SNP on either side on the chromosome
- T1D + Other autoimmune diseases (Rheumatoid arthritis, Celiac disease, Crohns disease, Lupus, Inflammatory bowel disease)
  - 365 SNPs have been identified as T1D+other autoimmune disease risk alleles (NHGRI)
  - 375 genes that either contain these SNPs or flank the SNP on either side on the chromosome

Genetic Risk of Type I Diabetes

GWAS data collected by Welcome Trust Case Control Consortium (WTCCC)

- 2000 cases, 3000 controls of European descent from Great Britain
- segment the genome into gene-sets: gene and a flanking region of 20KB on either side of the gene
- The WTCCC data includes
  - 350 of the gene-sets listed in the NHGRI catalog
  - covering 9,256 SNPs in the WTCCC data

T1D Prediction Results



Conclusions

- Kernel Machine Regression provides a useful tool for incorporating non-linear complex effects
- Blockwise KM regression achieves a nice balance between capturing complex effects and overfitting
- IBS kernel performs well under both linear and non-linear settings

Remarks

- May use SKAT to screen blocks for initial stage
- Can be extended to data with other covariates such as clinical variables
- Possible extensions might incorporate more complex block structure, different types of outcomes, interactions, and beyond!

Thank you!

References I

M. Braun, J. Buhmann, and K. Müller. On relevant dimensions in kernel feature spaces. *The Journal of Machine Learning Research*, 9:1875–1908, 2008.

P. Domingos and M. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine learning*, 29(2): 103–130, 1997.

L. Hindorff, P. Sethupathy, H. Junkins, E. Ramos, J. Mehta, F. Collins, and T. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362, 2009.

D. Liu, D. Ghosh, and X. Lin. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC bioinformatics*, 9(1):292, 2008.

J. Minnier, M. Yuan, J. Liu, and T. Cai. Risk classification with an adaptive naive bayes kernel machine model. *Journal of the American Statistical Association*, page (in press), 2014.

C. Rasmussen and C. Williams. Gaussian processes for machine learning. 2006. *The MIT Press, Cambridge, MA, USA*, 38: 715–719.

B. Scholkopf, S. Mika, C. Burges, P. Knirsch, K. Muller, G. Ratsch, and A. Smola. Input space versus feature space in kernel-based methods. *Neural Networks, IEEE Transactions on*, 10(5):1000–1017, 1999.

C. Williams and M. Seeger. The effect of the input density distribution on kernel-based classifiers. In *Proceedings of the 17th International Conference on Machine Learning*. Citeseer, 2000.

K. Zhang and J. Kwok. Density-weighted nyström method for computing large kernel eigensystems. *Neural computation*, 21(1): 121–146, 2009.

R. Zhang, W. Wang, and Y. Ma. Approximations of the standard principal components analysis and kernel pca. *Expert Systems with Applications*, 37(9):6531–6537, 2010.

Kernel PCA Estimation

Projection

To project the marker effects for a future subject with predictor $\mathbf{X} = \mathbf{x}$, by the Nystrom method [Rasmussen and Williams]

$$\mathbf{x} \mapsto \widehat{\mathbf{V}}_{\mathbf{x}} = \text{diag}\left(\frac{1}{\widehat{\lambda}_1}, ..., \frac{1}{\widehat{\lambda}_{\ell_0}}\right) \widehat{\boldsymbol{\Psi}}^{\mathsf{T}} \left[K(\mathbf{x}, \mathbf{X}_1), \cdots, K(\mathbf{x}, \mathbf{X}_n)\right]^{\mathsf{T}}.$$

Subsequently, we estimate $h(\mathbf{x})$ as

$$\widehat{h}(\mathbf{x}) = \widehat{\boldsymbol{\beta}}^{\mathsf{T}} \widehat{\mathbf{V}}_{\mathbf{x}}$$

Estimation: Dual Form

- By the representer theorem, a properly regularized estimator $h = \sum_l \beta_l \psi_l \in \mathcal{H}_K$ can be represented in the *dual form*

$$h(\cdot) = \sum_{i=1}^{n} \alpha_i K(\mathbf{X}_i, \cdot)$$

$$\Rightarrow \quad \text{logit}\{\text{pr}(Y = 1 \mid \mathbf{X})\} = a + \sum_{i=1}^{n} \alpha_i K(\mathbf{X}_i, \mathbf{X}),$$

- $(a, \boldsymbol{\alpha})$ may be estimated via a kernel ridge regression by minimizing

$$\mathcal{L}(Y, a, \mathbb{K}\boldsymbol{\alpha}) + \lambda \|h\|^2_{\mathcal{H}_K}, \quad \|h\|^2_{\mathcal{H}_K} = \boldsymbol{\alpha}^{\mathsf{T}} \mathbb{K} \boldsymbol{\alpha}$$

  - $\mathbb{K} = [K(\mathbf{X}_i, \mathbf{X}_j)]$
  - See Liu, Ghosh and Lin (2008) for details

Estimation

- By the representer theorem, a properly regularized estimator $h = \sum_l \beta_l \psi_l \in \mathcal{H}_K$ can be represented in the *dual form*

$$h(\cdot) = \sum_{i=1}^{n} \alpha_i K(\mathbf{X}_i, \cdot)$$

and $\boldsymbol{\alpha}$ may be estimated via a kernel ridge regression.

  - ⋆ See Liu et al. [2008] for details

Estimation

- By the representer theorem, a properly regularized estimator $h = \sum_l \beta_l \psi_l \in \mathcal{H}_K$ can be represented in the *dual form*

$$h(\cdot) = \sum_{i=1}^{n} \alpha_i K(\mathbf{X}_i, \cdot)$$

  and $\boldsymbol{\alpha}$ may be estimated via a kernel ridge regression.
  - ⋆ See Liu et al. [2008] for details
- Estimation of h based on $\mathbb{K} = [K(\mathbf{X}_i, \mathbf{X}_j)]$ may not be stable or precise when $\mathbb{K}$ is near singular, especially if λ of K decay quickly
  - primal form: $h = \sum_l \beta_l \psi_l = \sum_l \beta_l \sqrt{\lambda_l} \phi_l$
  - Kernel PCA approximation: $\mathbb{K} = \sum_{l=1}^{n} \widehat{\lambda}_l \widehat{\boldsymbol{\phi}}_l \widehat{\boldsymbol{\phi}}_l^{\mathsf{T}}$

$$\widetilde{\mathbb{K}} = \sum_{l=1}^{\ell_0} \widehat{\lambda}_l \widehat{\boldsymbol{\phi}}_l \widehat{\boldsymbol{\phi}}_l^{\mathsf{T}} = \widehat{\boldsymbol{\Psi}} \widehat{\boldsymbol{\Psi}}^{\mathsf{T}}; \qquad \widehat{\boldsymbol{\Psi}} = [\widehat{\lambda}_1^{\frac{1}{2}} \widehat{\boldsymbol{\phi}}_1, ..., \widehat{\lambda}_{\ell_0}^{\frac{1}{2}} \widehat{\boldsymbol{\phi}}_{\ell_0}]_{n \times \ell_0}$$

  - ⋆ Scholkopf et al. [1999]; Williams and Seeger [2000]; Braun et al. [2008]; Zhang et al. [2010]

Kernel PCA Estimation

Ridge Kernel PCA estimation

- Kernel PCA: $\mathbb{K} \to \widetilde{\mathbb{K}} = \widehat{\boldsymbol{\Psi}}\widehat{\boldsymbol{\Psi}}^{\mathsf{T}}$
- Change of variable: $\widetilde{\mathbb{K}}\boldsymbol{\alpha} = \widehat{\boldsymbol{\Psi}}\boldsymbol{\beta}, \quad \boldsymbol{\beta} = \widehat{\boldsymbol{\Psi}}^{\mathsf{T}}\boldsymbol{\alpha}$
- $\widehat{h}^{(\mathsf{b})}(\mathbf{X}^{(\mathsf{b})}) = \widehat{\boldsymbol{\Psi}}\widehat{\boldsymbol{\beta}}$
- obtain $(\widehat{a}, \widehat{\boldsymbol{\beta}})$ as the minimizer of ridge logistic objective function

$$\mathcal{L}(Y, a, \widehat{\boldsymbol{\Psi}}\boldsymbol{\beta}) + \tau\|\boldsymbol{\beta}\|^2$$

Regularized Selection of Informative Regions

To further improve the predictive accuracy, we estimate $P(Y = 1 \mid \mathcal{X})$ through cross-validated regularization:

- For each k in K folds, divide the data into $(K-1)/K$ of the data $\mathcal{D}_{(-k)} = \{\mathbf{Y}_{(-k)}, \mathbf{X}^{(b)}_{(-k)}\}$ and $1/K$ of the data $\mathcal{D}_{(k)} = \{\mathbf{Y}_{(k)}, \mathbf{X}^{(b)}_{(k)}\}$
  - Estimate $\widehat{\beta}_{(-k)}$ and $\widehat{h}^{(b)}_{(-k)}$ with training set $\mathcal{D}_{(-k)}$
  - Project kernel estimate into $\mathcal{D}_{(k)}$ to obtain $\widehat{h}^{(b)}_{(-k)}(\mathbf{X}^{(b)}_{(k)})$, an vector of length n/K
- Construct synthetic data in

$$\{\mathbf{Y}, \widehat{\mathbb{H}}\}, \quad \text{where} \quad \widehat{\mathbb{H}} = \left[\widehat{h}^{(b)}(\mathbf{X}^{(b)}_i)\right]_{n\times M} = \left[\left\{\widehat{h}^{(b)}_{(-k)}(\mathbf{X}^{(b)}_{(k)})\right\}_{k=1}^{K}\right]_{n\times M}$$

- Estimate $\widehat{\gamma}$ by minimizing the logistic LASSO likelihood using the synthetic data:

$$\widehat{\mathcal{L}}(\gamma) = \ell\left(a_0 + \sum_{b=1}^{M}\sum_{k=1}^{K}\sum_{i=1}^{n/K}\gamma_b \widehat{h}^{(b)}_{(-k)}(\mathbf{X}^{(b)}_{i(k)})\right) + \|\gamma\|_1$$

Numerical Studies: Genetic Risk of Rheumatoid Arthritis

GWAS data collected by:

Training: Brigham Rheumatoid Arthritis Sequential Study (BRASS)

- 483 cases, 1449 controls from Boston area

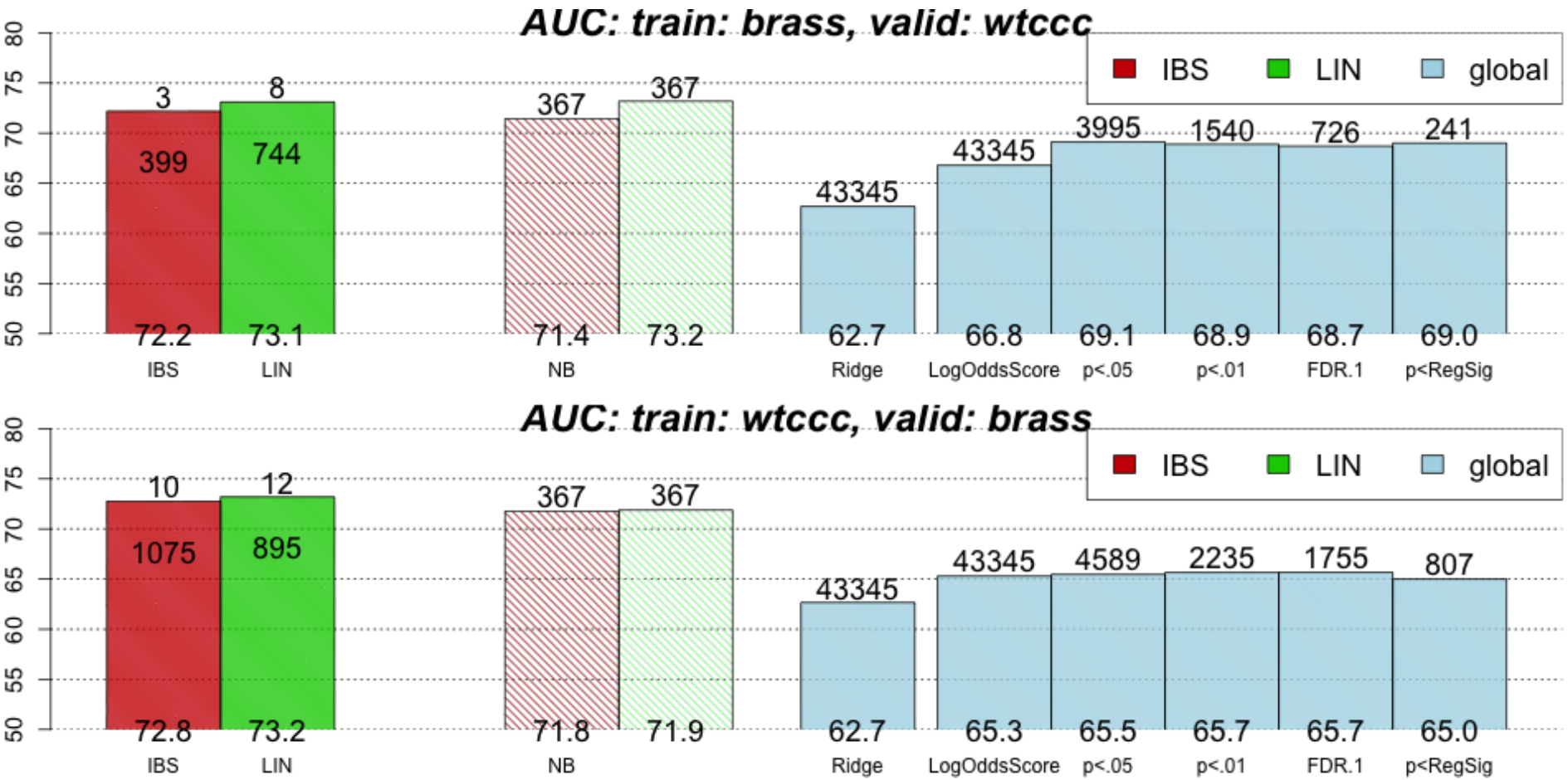
Validation: Welcome Trust Case Control Consortium (WTCCC)

- 1524 cases, 3108 controls of European descent from Great Britain

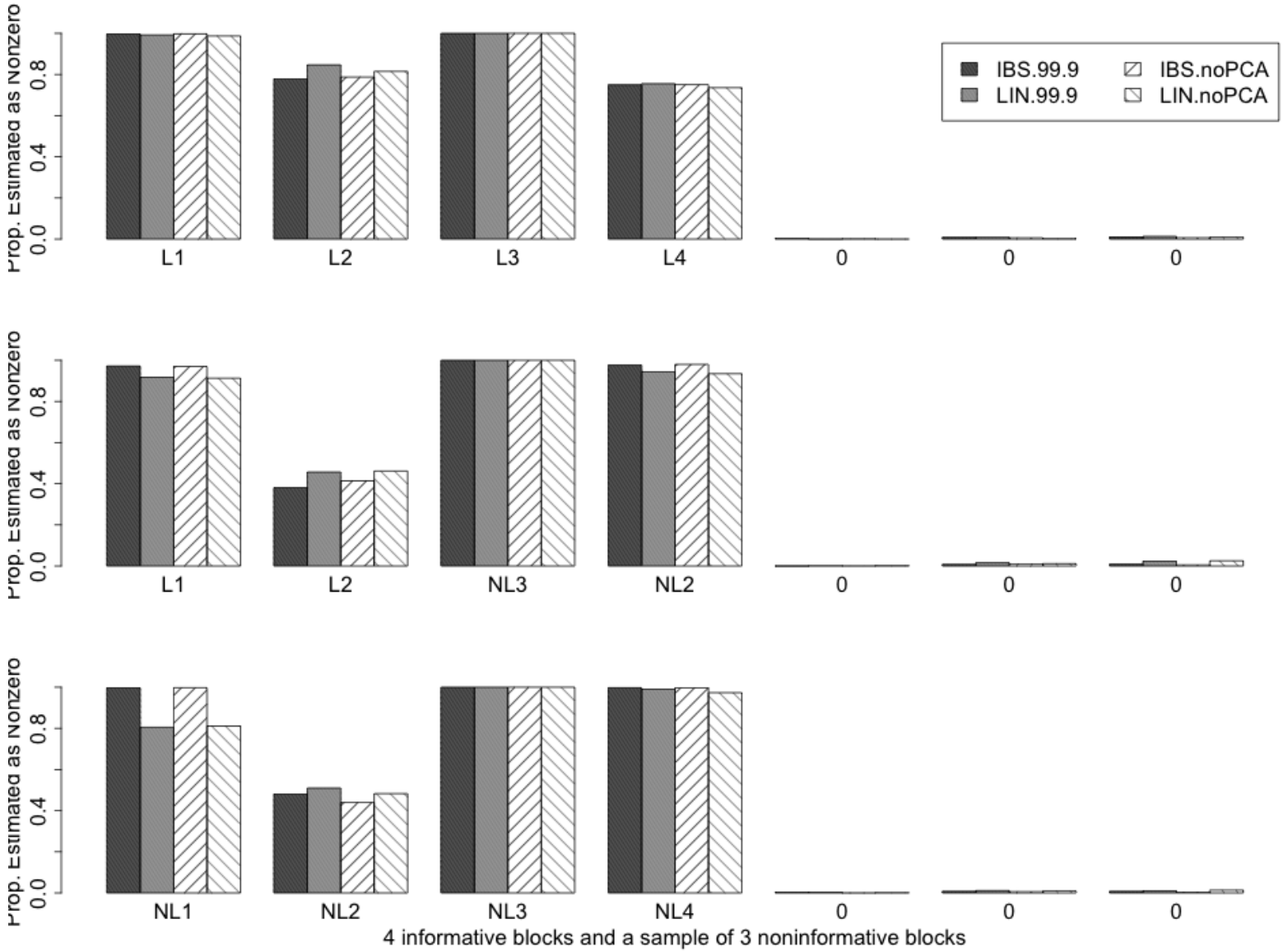
Data:

- segment the genome into gene-sets: gene and a flanking region of 20KB on either side of the gene
- covering 43,345 SNPs imputed to HapMap2 release

RA Validation AUC Results



Simulation $\widehat{\gamma}$ results, proportion nonzero



RA Data Analysis Results, removing regions

BRASS train, WTCCC valid