Mixing Generalized Ridge Regressions

J. Sunil Rao

University of Miami, Division of Biostatistics

Perspectives on High Dimensional Data, Fields Institute, 2011

Joint work with Hemant Ishwaran, Cleveland Clinic

Preamble

- Let $y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ be a response vector and $X = (X_{(1)}, \ldots, X_{(p)})$ an $n \times p$ design matrix.

$$y = X\beta + \varepsilon,$$

  where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is such that $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{E}(\varepsilon_i^2) = \sigma_0^2 > 0$.
- The true value for the coefficient vector $\beta = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ is an unknown value $\beta_0 = (\beta_{0,1}, \ldots, \beta_{0,p})^T$.
- Consider high-dimensional scenarios where $p \gg n$, and consider the problem of estimating $\mathbb{E}(y)$ and β_0.

Many Difficulties

Collinearity Abounds

- This problem poses unique obstacles for prediction and estimation. Sample correlation between variables can be sizeable as an artifact of dimensionality. Groups of variables become highly correlated with other groups sporadically - even if true design matrix is orthogonal (Fan and Lv (2008), Cai and Lv (2007)).
- Multicollinearity is further compounded as variables collected in high-dimensional applications are often naturally correlated (e.g gene expression arrays; SNP arrays, etc).

Back to the future

- Almost 40 years ago, Hoerl and Kennard (1970a,b) proposed generalized ridge regression (GRR), a method specifically designed for correlated and ill-conditioned settings.
- Let $\Lambda = \text{diag}\{\lambda_k\}_{k=1}^{p}$ be a $p \times p$ diagonal matrix with diagonal entries $\lambda_k > 0$. The GRR estimator with ridge matrix Λ is

$$\hat{\beta}_{\text{G}} = (Q + \Lambda)^{-1} X^T y,$$

  where $Q = X^T X$.
- An alternative representation for $\hat{\beta}_{\text{G}}$ is in terms of ℓ_2-penalization:

$$\hat{\beta}_{\text{G}} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ ||y - X\beta||^2 + \sum_{k=1}^{p} \lambda_k \beta_k^2 \right\},$$

Some work already in a similar vein

L1 alternatives

- Much of the recent effort to address high-dimensional problems has focused on ℓ_1-based penalization methods; i.e., lasso-type regularization (Tibshirani, 1996).
- Some of these methods are similar to the GRR in that they allow a unique regularization parameter for each coefficient.
- Examples include adaptive lasso for $p < n$ (Zou, 2006) and for diverging parameters problem (Huang et.al., 2008). Zou and Zhang (2009) also developed adaptive elastic net.
- Given GRR is naturally set up this way, it's natural to wonder how things behave when $p > n$.

A Geometric Approach

- We first recast $\hat{\beta}_G$ as a rescaled ridge estimator. Let $X_* = X\Lambda^{-1/2}$ and $Q_* = X_*^T X_*$.
- $$\begin{aligned} \hat{\beta}_G &= \Lambda^{-1/2}(\Lambda^{-1/2} Q \Lambda^{-1/2} + I_p)^{-1} \Lambda^{-1/2} X^T y \\ &= \Lambda^{-1/2}(Q_* + I_p)^{-1} X_*^T y \\ &= \Lambda^{-1/2} \hat{\beta}_R^*, \end{aligned} \tag{1}$$

  where $\hat{\beta}_R^* = (Q_* + I_p)^{-1} X_*^T y$ is the ridge estimator for the design matrix X_* with ridge parameter $\lambda = 1$.
- Let $X_* = U_* D_* V_*^T$ be the SVD for X_*. Let $d_{1,*} \geq \cdots \geq d_{n,*} \geq 0$ denote the diagonal elements of D_*.
- If $p \geq n$ and $\lambda_k > 0$ for $k = 1, \ldots, p$, then

$$\hat{\beta}_G = \Lambda^{-1/2} V_* S_{*1}^{-1} R_*^T y,$$

where $S_{*1} = \text{diag}\{d_{i,*}^2 + 1\}_{i=1}^n$ and $R_* = U_* D_*$. Requires $O(pn^2)$ operations.

Geometry cont'd

- A celebrated result, due to Penrose (1956), is that the minimum least squares (MLS)estimator exists and is the unique estimator

$$\hat{\beta}_{\text{MLS}} = X^{+}y = \lim_{\lambda \to 0} \hat{\beta}_{\text{R}} = VS_0^{+}R^{T}y,$$

where $S_0^{+} = \text{diag}\{s_{0i}^{+}\}_{i=1}^{n}$ is the Moore-Penrose generalized inverse of S_0 defined by

$$s_{0i}^{+} = \begin{cases} 1/d_i^2 & \text{if } d_i > 0 \\ 0 & \text{otherwise.} \end{cases}$$

- Define the modified MLS estimator as

$$\hat{\beta}_{\text{MLS}}^{*} = \Lambda^{-1/2}X_{*}^{+}y = \Lambda^{-1/2}V_{*}S_{*0}^{+}R_{*}^{T}y.$$

Observe in the special case when $\Lambda = \lambda I_p$ that $\hat{\beta}_{\text{MLS}}^{*} = \hat{\beta}_{\text{MLS}}$.

Theorem

$\hat{\beta}_G$ is the solution to the following optimization problem:

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}}\ \mathbb{Q}(\beta, \hat{\beta}^*_{MLS}) \ \text{ subject to } \ \beta^T \Lambda \beta \leq L,$$

for some $L > 0$, where

$$\mathbb{Q}(\beta, \hat{\beta}^*_{MLS}) = \left(\beta - \hat{\beta}^*_{MLS}\right)^T \left(\Lambda^{1/2} Q_* \Lambda^{1/2}\right) \left(\beta - \hat{\beta}^*_{MLS}\right)$$

is an ellipsoid with contours $\mathbb{S}(c) = \{\beta : \mathbb{Q}(\beta, \hat{\beta}^*_{MLS}) = c^2\}$ centered at $\hat{\beta}^*_{MLS}$.

GRR Shrinkage

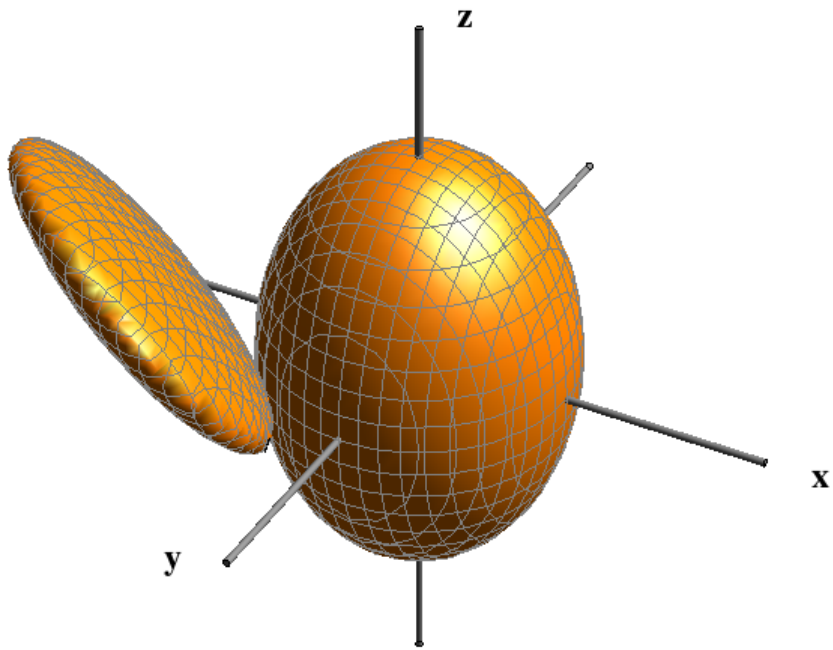


Figure: *Illustration of GRR geometry. Left figure corresponds to a simulation where* $p = 100, n = 25$, *and* $\beta_{0,k} = 0$ *for* $k \geq 3$ *and* $\lambda_k = \infty$ *for* $k \geq 4$. *Only the first 3 coordinates of* $\hat{\beta}_G$ *are nonzero, and these equal the point where the ellipsoid first touches the elliptical constraint region.*

Some consequences

- GRR estimator involves the contours of an ellipsoid centered at the modified MLS. For GRR, the constraint region is also an ellipsoid that depends upon Λ.

Theorem

$\hat{\beta}_G = \Lambda^{-1/2} \sum_{i=1}^{\mathfrak{d}} d_{i,*} \eta_i v_{i,*}$, where $\eta_i = (d_{i,*}^2 + 1)^{-1} u_{i,*}^T y$. That is, $\hat{\beta}_G$ lies in the $\mathfrak{d}$-dimensional subspace $\Lambda(\mathscr{V}_*) = \{\Lambda^{-1/2} v : v \in \mathscr{V}_*\}$, where $\mathfrak{d} = rank(X) \leq n$ and $\mathscr{V}_*$ is the span of $\{v_{1,*}, \ldots, v_{\mathfrak{d},*}\}$; i.e., $\mathscr{V}_*$ is the span of the eigenvectors of Q_*.

- If β_0 non-sparse, there is no guarantee that we can find a Λ to force the distance between GRR estimator and β_0 to zero.

Mixing for Improved Prediction

- Finding a suitable orientation for the GRR may be difficult in non-sparse settings.
- In place of a single GRR estimator, we'll instead use a convex combination of GRR estimators.
- The idea of mixing is widely used in statistics and machine learning (e.g. bagging, random forests, BMA).
- We will construct mixed GRR estimators based on exponential weights and establish that under a dimensionality constraint, that the risk for mixing GRR is at least as small as that of any of its GRR components - mixing GRR can achieve or even surpass the best model without advance knowledge of what that model is.

Mixing GRR

- Let $\mathscr{M}$ denote the space of linear models under consideration and let $M = \#\mathscr{M}$ be its cardinality.
- For each model $m \in \mathscr{M}$, we assume there exists a fixed probability $0 < \pi_m < 1$ indicating the preference for the model, where $\sum_m \pi_m = 1$.
- $\hat{\mu}^m_{\mathrm{G}} = X_m \hat{\beta}^m_{\mathrm{G}} = P_m y$ where

$$P_m = X_m (Q_m + \Lambda_m)^{-1} X_m^T,$$

and $Q_m = X_m^T X_m$. We refer to P_m as the *GRR operator*.

- $$\hat{\mu}_{\mathrm{mix}} = \sum_{m \in \mathscr{M}} \frac{\pi_m w_m}{\sum_{m'} \pi_{m'} w_{m'}} \hat{\mu}^m_{\mathrm{G}},$$

where $w_m = w_m(y)$ are non-negative, data-adaptively selected weights. Normalized weights $\overline{w}_m = \pi_m w_m / \sum_{m'} (\pi_{m'} w_{m'})$. Then, $\hat{\mu}_{\mathrm{mix}}$ can be rewritten as

$$\hat{\mu}_{\mathrm{mix}} = \sum_{m \in \mathscr{M}} \overline{w}_m P_m y.$$

Determining the w_m

- Define

$$\hat{r}_m = ||y - \hat{\mu}_{\mathrm{G}}^m||^2 + \sigma_0^2\big(2\,\mathrm{trace}(P_m) - n\big).$$

This is an unbiased estimator of the risk $r_m = \mathbb{E}||\hat{\mu}_{\mathrm{G}}^m - \mu||^2$.

- $$\nabla\hat{r}_m = \nabla\Big[y^T(I_m - P_m)^2 y + \sigma_0^2\big(2\,\mathrm{trace}(P_m) - n\big)\Big] = 2(I_m - P_m)^2 y.$$

- Theorem 5 of (Leung and Barron, 2006) describes a sharp minimax bound for mixing of the OLS. To obtain a similar result for GRR, it turns out that we need to use weights of the form $w_m = \exp(-\tilde{r}_m/4\sigma_0^2)$, where $\tilde{r}_m$ is chosen to satisfy

$$\nabla\tilde{r}_m = a_1 y - a_2 P_m y,$$

for some constants a_1, a_2 such that $a2/a1 > 0$.

Determining the w_m cont'd

- E.g. this would hold for $\tilde{r}_m = \hat{r}_m$ if P_m were a projection matrix, since then

$$\nabla \tilde{r}_m = 2(I_m - P_m)^2 y = 2(I_m - P_m)y = 2y - 2P_m y.$$

- In order to derive a minimax bound for GRR we must find a $\tilde{r}_m$ that satisfies the gradient equality even when P_m is not a projection matrix.
- Define,

$$\tilde{r}_m = \hat{r}_m - Z_m, \quad \text{where } Z_m = y^T(P_m^2 - P_m)y.$$

- even if P_m is not a projection matrix, we have

$$\nabla \tilde{r}_m = \nabla \hat{r}_m - \nabla Z_m = 2(I_m - P_m)^2 y - 2(P_m^2 - P_m)y = 2y - 2P_m y,$$

- Hence, we define our weights to be:

$$w_m = \exp\left(-f_0\left[\frac{1}{4\sigma_0^2}\hat{r}_m - \frac{1}{4\sigma_0^2}Z_m\right]\right)$$

  where f_0 is any constant such that $0 < f_0 \leq 1$.
- Intuitively, the rationale for these weights is that it down-weights models with large estimated risk $\hat{r}_m$.
- The presence of Z_m in w_m is an adjustment needed to accommodate the GRR operator. Can show that $Z_m \leq 0$. Hence it down-weights GRR operators that are distant from projection operators.
- The constant f_0 acts as a "dampening" parameter that spreads weights over models more evenly for small f_0 values.

Theorem

Assume that $\varepsilon \sim N_n(0, \sigma_0^2 I_n)$ *and that* $\max_m\{\mathbb{E}(\nabla_i|\hat{\mu}^m_{G,i}|)\} < \infty$ *for* $i = 1, \ldots, n$. *If* $\pi_m = 1/M$, *then for our* w_m *:*

$$\underbrace{\mathbb{E}||\hat{\mu}_{mix} - \mu||^2}_{\text{mixing GRR risk}} \leq \underbrace{\min_{m\in\mathscr{M}} \mathbb{E}||\hat{\mu}^m_G - \mu||^2}_{\text{minimax risk}} + \underbrace{4\sigma^2_{0,f} \log M}_{\text{dimensionality effect}}$$

$$+ \underbrace{\mathbb{E}\left(\sum_{m\in\mathscr{M}} \frac{(\hat{r}_m - 4\sigma^2_{0,f} C_m)\exp(-C_m)}{\sum_{m'\in\mathscr{M}} \exp(-C_{m'})} \right)}_{\text{GRR operator effect}} \quad (2)$$

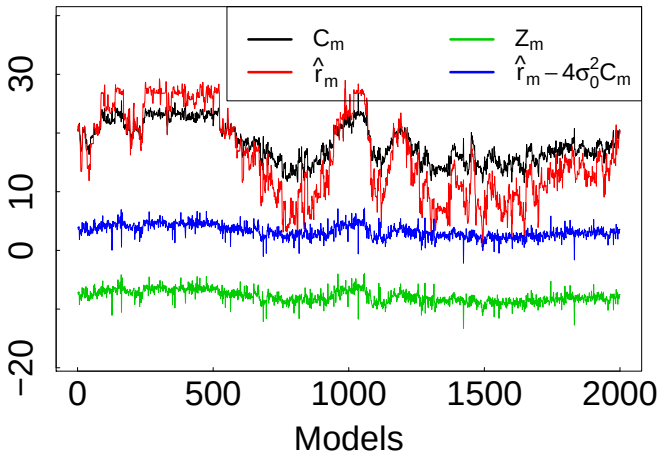
where $\sigma^2_{0,f} = \sigma^2_0 f_0^{-1}$, $C_m = (\hat{r}_m - Z_m + Z_{\hat{m}})/(4\sigma^2_{0,f})$, *and* $\hat{m}$ *is any model achieving the minimum estimated risk* $r_0 = \min_m\{\hat{r}_m\}$.

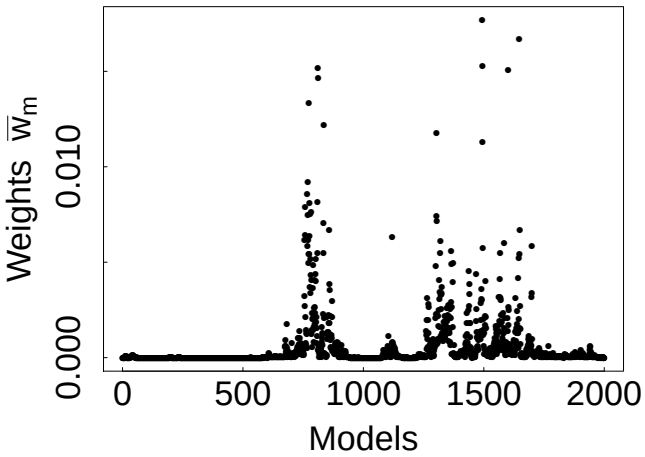
The effect of GRR operator on risk

- Many scenarios yield a negligible value without requiring stringent constraints on Λ_m.
- Consequently in $p > n$ problems, in order for the dimensionality effect to be smaller than the risk, we should choose $4f_0^{-1} \log M$ to be smaller than n. This suggests that $M < \exp(f_0 n/4)$.
- Large values of C_m are exponentially down-weighted and therefore cannot contribute substantially.
- If $Z_m - Z_{\hat{m}}$ is of order $o(\hat{r}_m)$, then C_m is of order $\hat{r}_m$. Thus, a large value of $\hat{r}_m$, corresponding to a large value of C_m, is down-weighted, while a small value of $\hat{r}_m$ yields a small $\hat{r}_m - 4\sigma_{0,f}^2 C_m$.
- Other scenarios exist as well that make this term negligible.

The various components: Diabetes data illustration

Diabetes data with 400 noise variables ($n = 422$, $p = 464$). Models $(\mathscr{C}_m)_{m=1}^{2000}$ and ridge matrices $(\Lambda_m)_{m=1}^{2000}$ used to define the mixing GRR were obtained using a Bayesian computational strategy.





The effect of dimension on risk

- By mixing over different estimators and down-weighting those that are unfavorable in terms of empirical risk, we improve stability, which improves risk behavior.
- At the same time, if we mix over too many estimators, there is a price to pay for dimension.
- Typically, the risk for each model m is order $n\sigma_0^2$. Thus, it is reasonable to anticipate that the minimax risk is also $n\sigma_0^2$.
- Consequently in $p > n$ problems, in order for the dimensionality effect to be smaller than the risk, we should choose $4f_0^{-1} \log M$ to be smaller than n. This suggests that $M < \exp(f_0 n/4)$.
- A strategy for implementing this in practice, is to restrict X to $[nF_0]$ columns for some constant $F_0 \geq 0$, and to define $\mathscr{M}$ to be the set of all possible subsets of the columns of the constrained X matrix

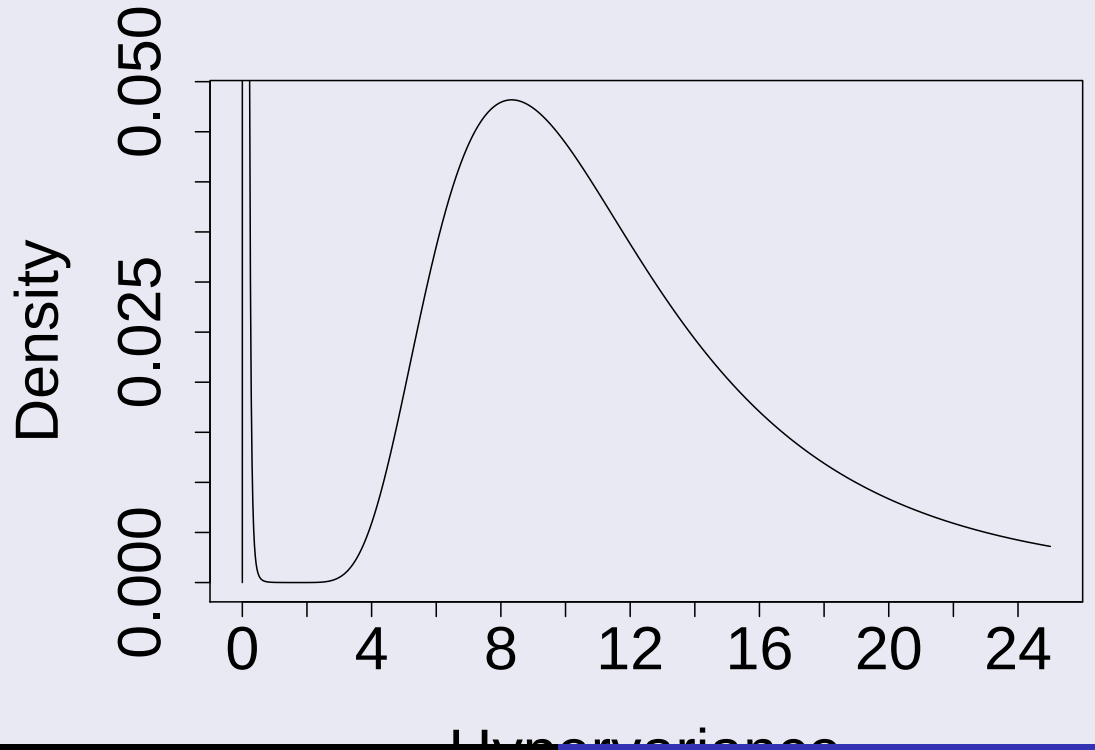
Bayesian Computational Strategy- Spike and Slab Regression

- For our Bayesian model, we rely on the class of rescaled spike and slab models which have been shown to have useful properties in linear regression settings (Ishwaran and Rao, 2005).
-

$$\begin{array}{rcll} (y^\star|X,\beta,\sigma^2) & \sim & \mathsf{N}_n(X\beta, n\sigma^2 I_n), & y^\star = n^{1/2}y \\ (\beta|\gamma) & \sim & \mathsf{N}_p(0,\Gamma), & \Gamma = \mathrm{diag}\{\gamma_k\}_{k=1}^p \\ \gamma & \sim & \Pi(\cdot), & \gamma = (\gamma_1,\ldots,\gamma_p)^T \\ \sigma^2 & \sim & \psi_{a,a}(\cdot), & 0 < a \ll 1, \end{array}$$

where $\psi_{a_1,a_2}(\cdot)$ denotes an inverse-gamma density with shape parameter $a_1 > 0$ and scale parameter $a_2 > 0$.

- A continuous bimodal density is used for γ_k with a spike at a small value ν_0 and a right continuous tail - convenient continuum of small and large values.
- Allows for large values of γ_k which allows the posterior mean for β_k to be large for promising coefficients, and it allows for small γ_k values which shrinks β_k towards zero for non-informative variables.



Calculating the mixing GRR

- Consider the conditional draw for β :
- $$(\beta|\Gamma, \sigma^2, y^\star) \sim \mathsf{N}_p(\mu_\Gamma, \sigma^2\Sigma_\Gamma)$$

  where $\mu_\Gamma = \Sigma_\Gamma X^T y^\star$ and $\Sigma_\Gamma = (Q + n\sigma^2\Gamma^{-1})^{-1}$. This suggests setting $\Lambda_m = n\sigma^2\Gamma^{-1}$.
- Indices k where $\gamma_k = \nu_0$ indicate variables being shrunk towards zero. We shall classify such variables as being excluded from the subset selection.
- Using the mth draw of γ from the Gibbs sampler, we define $\mathscr{C}_m$ to be the set of all indices k where $\gamma_k \neq \nu_0$.

Calculating the mixing GRR cont'd

- Correspondingly, we define Λ_m using the diagonal values of $n\sigma^2\Gamma^{-1}$ with indices in $\mathscr{C}_m$.
- Thus, if $(\Gamma_{(m)}, \sigma^2_{(m)})$ is the mth draw for (Γ, σ^2),

$$\Lambda_m = n\sigma^2_{(m)}\tilde{\Gamma}^{-1}_{(m)},$$

  where $\tilde{\Gamma}_{(m)}$ is the submatrix of $\Gamma_{(m)}$ corresponding to $\mathscr{C}_m$.
- To calculate the mixing GRR we estimate σ^2_0 by $\sum_{m=1}^{M} \sigma^2_{(m)}/M$ and from this calculate $\tilde{r}_m$. The mixing GRR is then defined by setting $\pi_m = 1/M$ and calculating the weights as previously defined.

Other issues

Efficient Gibbs sampling

- We must reduce the computational burden of the draw for β, which is quadratic in p.
- We resolve this by using an SVD similar to what was done in earlier theorem - enabling draw for β to be efficiently calculated in a linear number of operations in p.

Filtering variables

- Use a technique known as grouped complexity filtering where have a separate complexity parameter per group (details in paper).

Hybrid mixing GRR

- Also developed a hybrid mixing GRR where we used the absolute value of the posterior mean of β to generate a sequence of nested models with which to mix. The mth nested model, denoted by $\hat{\mathscr{C}}_m$, is defined to be the first m variables in this ordering.
- Let $\hat{\beta} = n^{-1/2}\mathbb{E}(\beta|y^\star)$ denote the rescaled posterior mean. We solve for $\hat{\Lambda}$ in the following equation

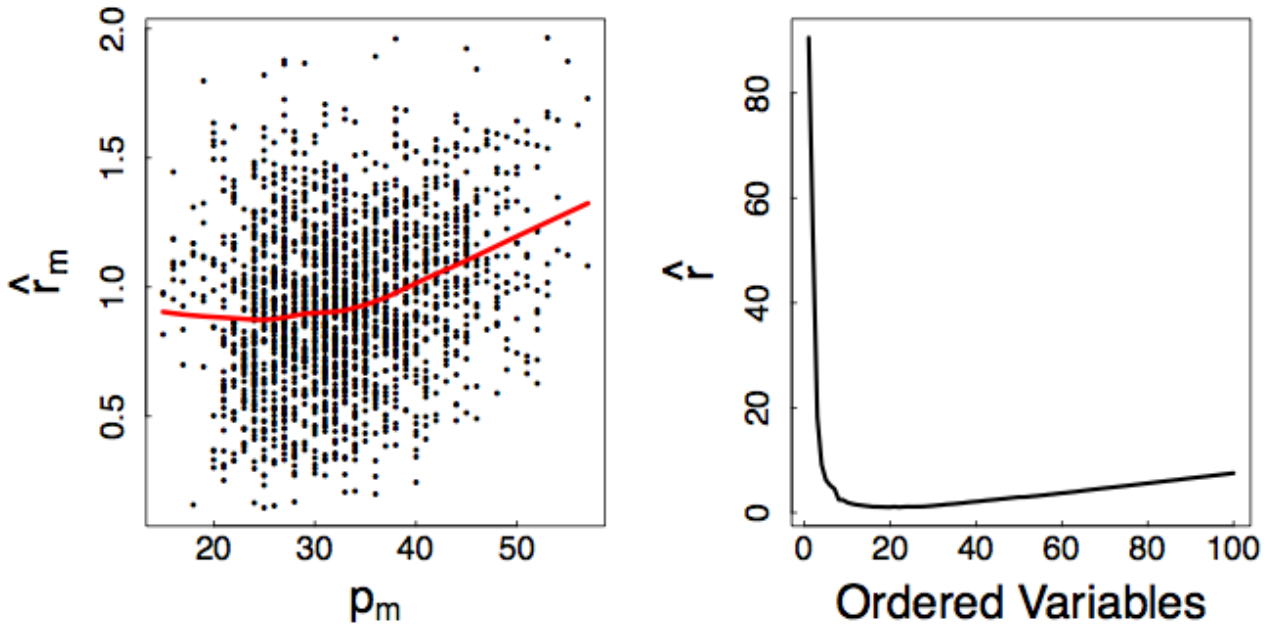
$$\hat{\beta} = (Q + \hat{\Lambda})^{-1} X^T y.$$

  This yields the ridge matrix for the GRR closest to $\hat{\beta}$.
- We define $\hat{\Lambda}_m$ to be the submatrix of $\hat{\Lambda}$ corresponding to those indices in $\mathscr{C}_m$. For each m, let $\overline{w}^m_{\text{hyb}}$ denote our weights calculated using $(\hat{\mathscr{C}}_m, \hat{\Lambda}_m)$.
- Also, let $\hat{\mu}^m_{\text{hyb}}$ be the GRR predictor corresponding to model $(\hat{\mathscr{C}}_m, \hat{\Lambda}_m)$ and let $\hat{r}^m_{\text{hyb}}$ be its estimated risk. The hybrid mixing GRR is defined to be

$$\hat{\mu}_{\text{hyb}} = \sum_{m=1}^{\hat{M}} \overline{w}^m_{\text{hyb}} \hat{\mu}^m_{\text{hyb}}, \quad \text{where } \hat{M} = \underset{1 \le m \le M}{\arg\min} \left\{ \hat{r}^m_{\text{hyb}} \right\}.$$

Correlated simulation illustration - hybrid mixing GRR

Correlated simulation $n = p = 250$. Left side: value of $\hat{r}_m$ for each model m plotted against its model size, p_m (these are used to construct mixing GRR estimator). Right side: value of $\hat{r}$ for hybrid GRR mixing estimator in which GRR estimators were constructed sequentially from variables ordered using the posterior mean of β.



Ultra-high dimensional simulation

- X that was simulated by drawing $n = 200$ rows independently from a p-dimensional multivariate normal with mean zero and covariance matrix, with correlations that decayed according to $\rho_{k,l} = \rho^{|k-l|}$. We sampled ε from a $N_n(0, I_n)$ distribution.
- β_0, nonzero coefficients were clustered into $\mathscr{G}$ distinct clusters, with each cluster comprising five coefficients: one strong coefficient, two moderately strong coefficients, and one weak coefficient. Coefficient values were adjusted by multiplying by a constant so that the theoretical $\mathscr{R}^2$ was equal to 0.9. The theoretical $\mathscr{R}^2$ was defined as

$$\mathscr{R}^2 = \frac{||X\beta_0||^2}{||X\beta_0||^2 + n}.$$

- Looked at $\mathscr{G} = 9$ (i.e. $p_0 = 45$) and $\mathscr{G} = 45$ (i.e. $p_0 = 225$). Looked at $p = 1000, 2000, 4000$.
- Looked at $\rho = 0.25, .90$.

Ultra-high dimensional simulation

Table 1: *Ultra-high dimensional simulations (averaged over 100 runs).*

Sparse signal, low correlation ($n = 200$, $p_0 = 45$, $\rho = 0.25$)

	$p = 1000$				$p = 2000$				$p = 4000$			
	$\hat{p}$	PE	AUC	MSE	$\hat{p}$	PE	AUC	MSE	$\hat{p}$	PE	AUC	MSE
mix	400	1.80	89.14	0.31	400	1.93	84.80	0.28	400	1.97	82.44	0.17
hyb	185	1.74	89.22	0.27	255	1.90	85.13	0.25	285	1.95	82.47	0.16
enet	116	1.74	90.30	0.19	119	1.90	86.92	0.13	117	2.17	82.34	0.08
lasso	116	1.74	90.31	0.19	119	1.90	86.92	0.13	114	2.18	82.31	0.08
anet	93	1.80	89.68	0.21	99	1.87	85.94	0.14	98	1.94	79.98	0.09

Sparse signal, high correlation ($n = 200$, $p_0 = 45$, $\rho = 0.90$)

	$p = 1000$				$p = 2000$				$p = 4000$			
	$\hat{p}$	PE	AUC	MSE	$\hat{p}$	PE	AUC	MSE	$\hat{p}$	PE	AUC	MSE
mix	400	1.25	97.00	0.12	400	1.26	98.41	0.06	400	1.30	99.04	0.03
hyb	47	1.27	96.33	0.15	52	1.26	98.12	0.07	56	1.28	98.87	0.04
enet	70	1.26	86.88	0.07	77	1.27	86.88	0.04	75	1.31	83.59	0.02
lasso	51	1.27	79.44	0.10	57	1.28	79.55	0.05	61	1.31	78.50	0.03
anet	34	1.30	87.72	0.11	45	1.30	87.37	0.05	48	1.39	84.06	0.03

Less sparse signal, low correlation ($n = 200$, $p_0 = 225$, $\rho = 0.25$)

	$p = 1000$				$p = 2000$				$p = 4000$			
	$\hat{p}$	PE	AUC	MSE	$\hat{p}$	PE	AUC	MSE	$\hat{p}$	PE	AUC	MSE
mix	400	1.99	63.44	0.98	400	2.00	62.04	0.51	400	1.99	59.72	0.25
hyb	336	2.01	63.53	0.95	341	1.99	62.03	0.50	335	1.99	59.72	0.25
enet	152	3.30	62.82	0.52	248	4.20	60.83	0.28	426	5.13	58.67	0.15
lasso	78	4.28	60.42	0.56	55	5.70	57.93	0.29	44	6.50	55.85	0.15
anet	109	3.02	60.81	0.56	105	3.79	58.65	0.31	77	4.58	56.32	0.17

Less sparse signal, high correlation ($n = 200$, $p_0 = 225$, $\rho = 0.90$)

	$p = 1000$				$p = 2000$				$p = 4000$			
	$\hat{p}$	PE	AUC	MSE	$\hat{p}$	PE	AUC	MSE	$\hat{p}$	PE	AUC	MSE
mix	400	1.40	95.01	0.11	400	1.44	96.60	0.06	400	1.46	96.32	0.03
hyb	108	1.46	94.61	0.17	115	1.47	96.41	0.09	118	1.48	96.22	0.04
enet	177	1.39	79.89	0.06	204	1.43	80.65	0.03	221	1.48	79.47	0.02
lasso	93	1.44	65.44	0.14	103	1.48	65.27	0.07	111	1.52	64.70	0.04
anet	89	1.52	79.82	0.10	119	1.49	80.66	0.04	159	1.54	79.56	0.02

Ultra-high dimensional simulation - summary

- Overall message for mixing GRR is that it is most adept in high-dimensional correlated settings. This includes both sparse and less sparse scenarios.
- In such settings it has good PE performance and it accurately ranks variables (AUC). The results for the hybrid mixing GRR are interesting. Its PE is often as good as mixing GRR and it is a much sparser estimator.
- Note that AUC numbers are evaluated over the whole solution path for each estimator in order to make results comparable across methods.

Benchmark performance

- 33 datasets of different sample sizes and dimensions
- Some datasets were related to one another. For example, row entries named “x.I” indicate a dataset “x” that was modified to include all pairwise interactions, as well as B-spline basis functions (up to 6 degrees of freedom), for all original variables.
- Data with names “x.noise” indicate a dataset “x” with 1500 noises variables added (sampled independently from a standard normal distribution).
- Data with names “x.I!” were modified similar to x.I, but in addition, all real valued variables were mapped to dummy variables representing a factor with three levels and all pairwise interactions of these dummy variables were added to the design matrix.

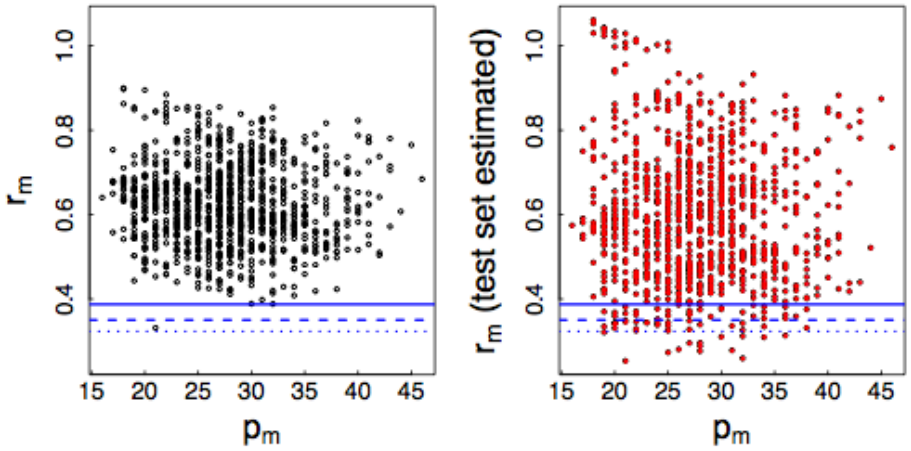
Benchmark performance - results

Table 2: Benchmark results: 10-fold test set prediction error and estimated model size (rounded to the nearest integer).

Dataset	n	p	$\hat{p}_{\text{mix}}$	$\hat{p}_{\text{hyb}}$	$\hat{p}_{\text{enet}}$	PE_{mix}	PE_{hyb}	PE_{enet}
Air	111	5	5	5	4	0.28	0.28	0.29
Air.I!	111	311	111	26	42	0.28	0.32	0.28
Auto	193	65	64	42	42	5.55	5.49	5.28
Auto.spline	193	135	134	72	76	5.08	5.48	5.58
Bodyfat	252	13	13	10	6	20.58	20.67	21.61
CMB	899	4	4	3	3	10.16	10.08	10.22
CMB.noise	899	1504	676	43	15	10.35	10.33	10.41
Correlated	250	1000	249	21	40	1.20	1.28	1.31
Crime	47	15	15	13	7	68974.26	69056.54	93276.54
Diabetes	442	10	10	8	8	3004.97	3010.63	3015.13
Diabetes.I	442	64	64	14	20	2928.47	2929.47	2951.50
Diabetes.I.noise	442	1564	442	47	46	3263.38	3268.62	3314.50
Fitness	31	6	6	3	3	8.32	8.43	9.46
Friedman.1	100	50	50	1	8	1.45	1.38	1.45
Friedman.2	100	10	10	1	1	1.36	1.36	1.37
Friedman.3	100	6	6	1	1	1.36	1.37	1.37
Highway	39	11	11	7	6	0.28	0.31	0.24
Highway.I	39	110	39	15	19	0.18	0.16	0.43
Housing	506	13	13	12	12	23.98	24.17	24.35
Housing.I!	506	658	506	82	40	12.72	13.45	15.40
Housing.I!.noise	506	2158	506	73	48	12.09	12.97	16.10
Iowa	33	9	9	5	4	89.62	84.00	90.90
Iowa.I	33	90	33	9	11	75.31	103.78	85.43
Ozone	203	12	12	7	7	19.89	19.84	19.89
Ozone.I	203	134	134	13	21	14.19	14.63	15.57
Ozone.I.noise	203	1634	203	15	22	15.38	15.16	18.02
Pollute	60	15	15	8	7	1540.32	1563.01	1756.65
Prostate	97	8	8	4	4	0.53	0.53	0.54
Servo	167	19	19	14	13	0.71	0.74	0.79
Servo.I!	167	147	146	16	16	0.32	0.35	0.35
Servo.I!.noise	167	1647	167	12	30	0.40	0.41	0.39
Tecator	215	22	22	22	20	7.11	7.42	7.93
Windmill	1114	12	12	11	10	4.58	4.58	4.59

The effect of data-adaptivity on the minimax rate

Correlated simulation from earlier ($\rho = 0.9$ equicorrelated data with $n = 200$, $p = 250$, and $p_0 = 10$). Blue horizontal lines are test set estimated risk for the mixing GRR (thick, dashed, and dotted lines are $f_0 = 0.1, 0.5, 1.0$, respectively).



Minimax bound does not completely hold, but the results are very close. Only a few models that surpass the mixing GRR.
Moreover, with increasing f_0, the risk for the mixing GRR improves and the bound becomes nearly exact.
A larger f_0 improves the mixing GRR in this example due to the sparsity of the problem because it concentrates the mixing GRR on fewer models.

Conclusions

- Spurious correlations are a real problem in $p >> n$ situations. Hence wanted to study GRR in these problems.
- GRR solution differs from classic $n > p$ setting because it's constrained to lie in a subspace containing the MLS of dimension at most n. This implies that for accurate estimation, the true parameter vector should be sparse. In non-sparse (or less sparse) situations, no guarantee of accurate estimation.
- Introduced mixing GRR ensemble predictor which has a nice finite sample minimax bound which shows that the risk for mixing GRR will never be larger than the risk for any of its constituent components assuming dimensionality is properly constrained.
- Developed Bayesian computational approaches for use in practice.

Acknowledgements

Thanks to NSF and NIH for financial support.