Structure aware error bounds using random projections

Ata Kabán
School of Computer Science
The University of Birmingham
Birmingham B15 2TT, UK
http://www.cs.bham.ac.uk/∼axk

ICMS Edinburgh, 2-6 July 2018.

Two goals

- Suppose we try to learn from random projections of the data – what are generalisation guarantees? & what structural characteristics they depend on?

- From the answers, obtain better guarantees for the original problem & bettter understand what structural characteristics they depend on

Linear Classification

- Given:

· $\mathcal{T}^N = \{(x_n, y_n) : (x_n, y_n) \overset{\text{i.i.d}}{\sim} \mathcal{D}\}_{n=1}^N$, where $\mathcal{D}$ is an unknown distribution over $\mathcal{X} \times \mathcal{Y}, \mathcal{X} \subseteq \mathbb{R}^d, \mathcal{Y} = \{-1, 1\}$

· $\mathcal{H} := \{x \to \text{sign}\,(h^T x) : h \in \mathbb{R}^d, x \in \mathcal{X}\}$

· $\ell : \mathcal{Y} \times \mathcal{Y} \to \{0, 1\}, \ell(\hat{y}, y) = \mathbf{1}(\hat{y} \neq y)$

- Goal: Find $\hat{h} \in \mathcal{H}$ s.t. its risk is as small as possible

$\mathsf{E}[\ell \circ \hat{h}] := \mathsf{E}_{(x,y)\sim\mathcal{D}}[\ell(\hat{h}(x), y)|\mathcal{T}^N]$

- Optimal classifier: $h^* := \underset{h\in\mathcal{H}}{\text{arg inf}}\ \mathsf{E}_{(x,y)\sim\mathcal{D}}[\ell(h(x), y)]$

- Known: $|\mathsf{E}[\ell \circ \hat{h}] - \hat{\mathsf{E}}[\ell \circ \hat{h}]| = \tilde{\Theta}(\sqrt{d/N})$ in general.
- Q: What if $d > N$?

Compressive ERM Classifier

- Let $R \in \mathbb{R}^{k\times d}$, $k \leq d$ be a random matrix with i.i.d. 0-mean (sub-)Gaussian rows.

  · $\mathcal{T}_R^N = \{(Rx_n, y_n)\}_{n=1}^N$ RP of the training set

  · $\mathcal{H}_R := \{Rx \to \text{sign}\,(h_R^T Rx) : h_R \in \mathbb{R}^k \in \mathbb{R}, x \in \mathcal{X}\}$

- Compressive ERM: $\hat{h}_R = \underset{h_R \in \mathcal{H}_R}{\text{arg min}} \frac{1}{N} \sum_{n=1}^N \ell(h_R(Rx_i), y_i)$

- What is the generalization error of $\hat{h}_R$:

$$\mathsf{E}[\ell \circ \hat{h}_R] := \mathsf{E}_{(x,y)\sim\mathcal{D}} \left[\ell(\hat{h}_R(Rx), y) | \mathcal{T}^N, R\right] \leq ?$$

Risk bound for compressive ERM classification

Theorem [K-D, 2017] Let R be a $k \times d$ sub-Gaussian random matrix with i.i.d. entries, $k \leq d$. For any $\delta \in (0,1)$, the following holds for the compressive ERM classifier $\hat{h}_R$ with probability $1-2\delta$:

$$
\begin{aligned}
\mathsf{E}_{x,y}[\mathbf{1}(\hat{h}_R^T Rxy \leq 0)] \quad &\leq \quad \mathsf{E}_{x,y}[\mathbf{1}(h^{*T} xy \leq 0)] + 2c\sqrt{\frac{k+\log(1/\delta)}{N}} \ldots \\
&+ \quad \mathsf{E}_{x,y}[f_k^+(\theta_{xy}^{h^*})] + \min\left\{\frac{1-\delta}{\delta} \cdot \mathsf{E}_{x,y}[f_k^+(\theta_{xy}^{h^*})], \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right\}
\end{aligned}
$$

where $c > 0$ is a constant, θ_u^h is the angle betwen u and h, and $f_k^+(\theta_u^h) := f_k(\theta_u^h) \cdot \mathbf{1}(h^T u > 0)$,
where $f_k(\theta_u^h) = \mathsf{Pr}_R\left\{h^T R^T Ru \leq 0\right\}$.

• On RHRs, first 2 terms match a VC bound for k-dimensional linear classifier – complexity reduced from d to $k < d$.
• Last 2 terms pay the price.
• If k grows to d, we recover classical VC bound.

• Note, no sparse representation was required for the compressive classification to succeed.

• The last 2 terms can be $\leq \epsilon$ despite $k << d$ if we are 'lucky': for $k \geq \frac{8 \log(1/(\epsilon\delta))}{\inf_{(x,y)} \cos^2(\theta_{xy}^{h^*})}$, provided $\inf_{(x,y)} \cos(\theta_{xy}^{h^*}) > 0$

What if $\inf_{(x,y)} \cos(\theta_{xy}^{h^*}) \leq 0$?

Proof (sketch).

- For a fixed instance of R, VC bound in the compressed space gives, $\forall \delta \in (0,1)$ w.p. $1-\delta$ over $\mathcal{T}^N$,

$$\mathsf{E}_{x,y}[\mathbf{1}(\hat{h}_R^T Rxy \leq 0)] \leq \mathsf{E}_{x,y}[\mathbf{1}(h_R^{*T} Rxy \leq 0) + 2c\sqrt{\frac{k+\log(1/\delta)}{N}}$$

RP reduces the complexity term but can increase error of best classifier in the class. By how much?

- Noting that $Rh^* \in \mathcal{H}_R$, and since h_R^* is optimal in $\mathcal{H}_R$,

$$\begin{aligned} & \mathsf{E}[\mathbf{1}(h_R^{*T} Rxy \leq 0)] \leq \mathsf{E}[\mathbf{1}(h^{*T} R^T Rxy \leq 0)] \\ = & \left(\mathsf{E}[\mathbf{1}\left(h^{*T} R^T Rxy \leq 0\right) - \mathbf{1}\left(h^{*T} xy \leq 0\right)]\right) + \mathbf{1}\left(h^{*T} xy \leq 0\right) \\ \leq & \ ... \leq \underbrace{\mathsf{E}[\mathbf{1}\left(h^{*T} R^T Rxy \leq 0\right) \mathbf{1}\left(h^{*T} xy > 0\right)]}_{S} + \mathsf{E}[\mathbf{1}\left(h^{*T} xy \leq 0\right)] \end{aligned}$$

- Finally, bound S from $\mathsf{E}_R[S]$ w.h.p, w.r.t. R. ■

Variant: When $\inf_{x,y}\cos(\theta^h_{xy}) \leq 0$:

Corollary Fix some $\gamma > 0$. Let R be a $k \times d$ sub-Gaussian random matrix with i.i.d. entries, $k \leq d$. For any $\delta \in (0,1)$, w.p. $1-2\delta$ the compressive ERM classifier $\hat{h}_R$ satisfies:

$$\begin{aligned}
\mathsf{E}_{x,y}\{\mathbf{1}(\hat{h}_R^T R x y \leq 0)\} \quad &\leq \quad \mathsf{E}_{x,y}[\mathbf{1}\{\cos(\theta^{h^*}_{xy})] \leq \gamma\} + c\sqrt{\frac{k+\log(1/\delta)}{N}} \ldots \\
&+ \quad \mathsf{E}_{x,y}[f_k^{\gamma}(\theta^{h^*}_{xy})] + \min\left\{\frac{1-\delta}{\delta} \cdot \mathsf{E}[f_k^{\gamma}(\theta^{h^*}_{xy}), \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right\}
\end{aligned}$$

where $c > 0$ is a constant, and $f_k^{\gamma}(\theta^h_u) := f_k(\theta^h_u) \cdot \mathbf{1}(\cos(\theta^h_u) > \gamma)$, where $f_k(\theta^h_u) = \mathsf{Pr}_R\left\{h^T R^T R u \leq 0\right\}$.

Variant: Tighter bound when R is Gaussian

Theorem For all $\delta \in (0,1)$, w.p. $1-2\delta$:

$$\begin{aligned} \mathsf{E}_{x,y}\{\hat{h}_R^T Rxy \leq 0\} &\leq \mathsf{E}_{x,y}[f_k(\theta_{xy}^{h^*})] + 2c\sqrt{\frac{k+\log\frac{1}{\delta}}{N}} \\ &+ \min\left\{\frac{1-\delta}{\delta}\cdot \mathsf{E}_{x,y}[f_k(\theta_{xy}^{h^*})], \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right\} \end{aligned}$$

where c is an absolute constant.

Captures both flipping the prediction from right to wrong, and from wrong to right after RP.

The sign flipping probability

Lemma [Flip probability - Gaussian case] Let R be a 0-mean Gaussian RP matrix. Let $h, x \in \mathbb{R}^d$, and let $\theta = \theta_x^h \in [0, \pi)$ be the angle between them. Assume $h^T x \neq 0$, Then,

1. Exact form:

$$f_k(\theta) := \frac{\Gamma(k)}{(\Gamma(k/2))^2} \int_0^{\frac{1-\cos(\theta)}{1+\cos(\theta)}} \frac{z^{(k-2)/2}}{(1+z)^k} \mathrm{d}z = \Pr\left\{(Rh)^T Rx \leq 0\right\}$$

$$\Pr\left\{\frac{(Rh)^T Rx}{h^T x} \leq 0\right\} = f_k(\theta) \cdot \mathbf{1}(h^T x > 0) + (1 - f_k(\theta)) \cdot \mathbf{1}(h^T x < 0)$$

2. Upper bound: $\Pr\left\{\frac{(Rh)^T Rx}{h^T x} \leq 0\right\} \leq \exp(-k\cos^2(\theta)/2)$

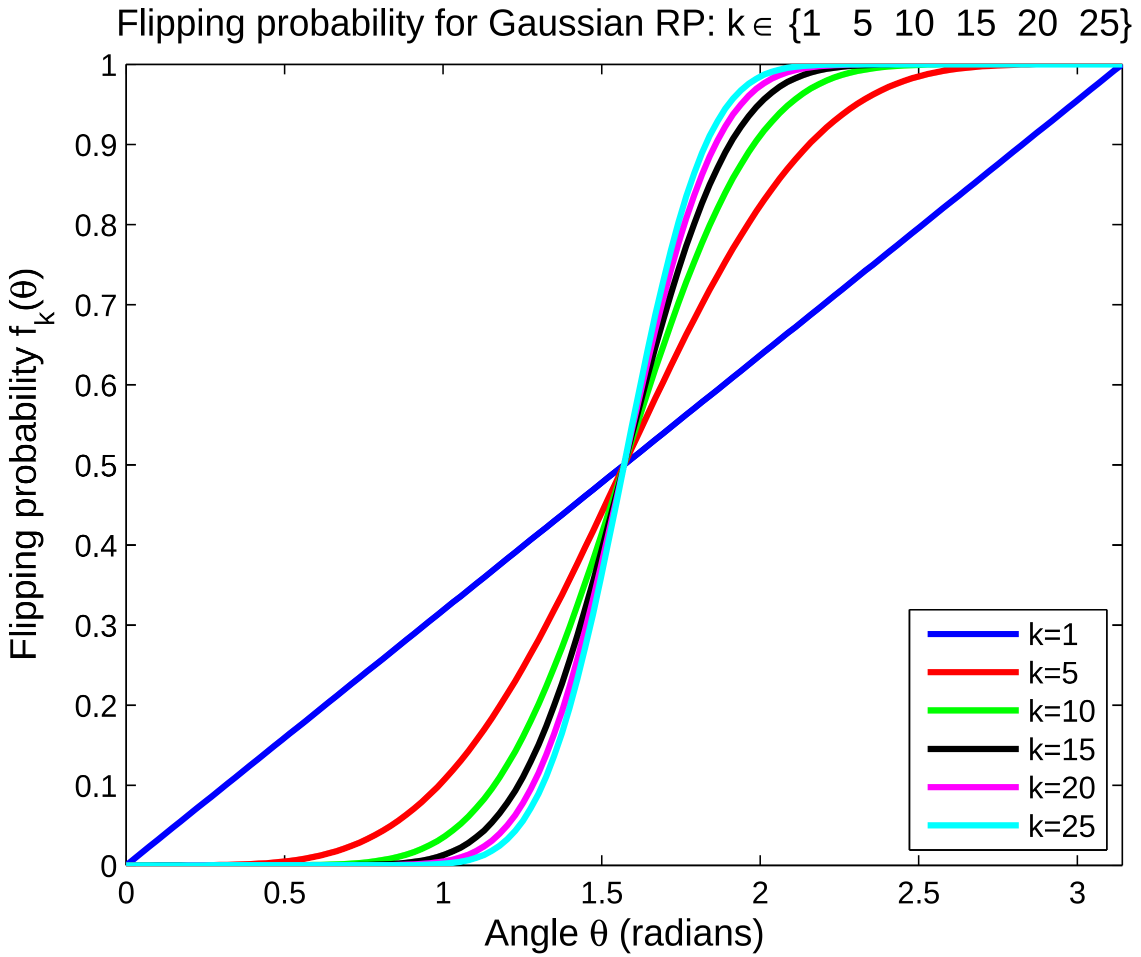


Illustration of the function $f_k(\theta)$ as a function of θ.

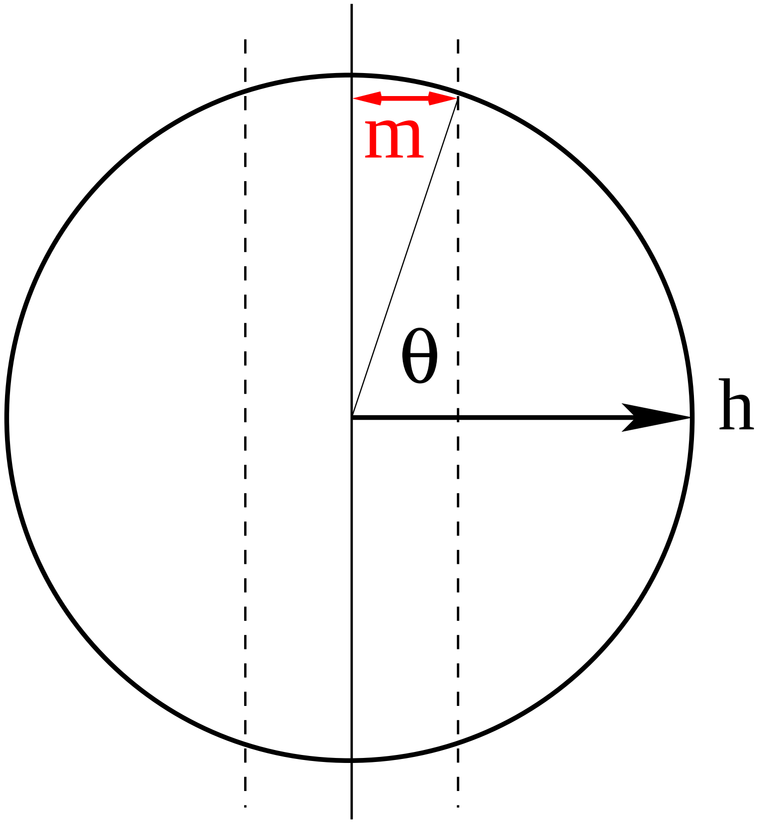
Definition [sub-Gaussian random variable] A zero-mean random variable X is subgaussian with parameter σ^2 if $\exists \sigma^2 > 0$ such that:

$$\mathsf{E}\left\{\exp(\lambda X)\right\} \leq \exp\left\{\sigma^2\lambda^2/2\right\}$$

Lemma [Flip probability - sub-Gaussian case] Let R be a RP matrix with entries r_{ij} drawn i.i.d. from a zero-mean subgaussian distribution, let $h, x \in \mathbb{R}^d$, and let $\theta = \theta_x^h$ be the angle between them. If $h^T x \neq 0$, then:

$$\Pr\left\{\frac{(Rh)^T Rx}{h^T x} \leq 0\right\} \leq \exp(-k\cos^2(\theta)/8) \tag{1}$$

Relation of Sign Flipping Probability to Margin



Flip probability and Margins

$$f_k^+(\theta) \leq \exp(-\tfrac{1}{8}k\cos^2(\theta))$$

$$\cos(\theta) = m$$

Large margin ⇒
small flip probability (no ⇐)

When does RP cost nothing? Explicitly geometry-aware bound

$U := \left\{ \frac{xy}{\|x\|} : x \in \mathcal{X}, y \in \{-1,1\} \right\}$

For $h \in \mathcal{H}$, $\gamma_h := \inf_{u \in U} \cos(\theta_u^h)$

$T_{h,\gamma}^{+} := \left\{ u \in U : \cos(\theta_u^h) \geq \gamma \right\} \subset S^{d-1}$; where $\gamma > 0$

Theorem Let R be a $k \times d, k \leq d$ isotropic subgaussian random matrix with independent rows each having subgaussian norm bounded as $\|R_i\|_{\psi_2} \leq K$. Fix some $\gamma > 0$ s.t. $\gamma \geq \gamma_h$. Then, for any $\delta > 0$ there are absolute constants $C, c > 0$ s.t. w.p. $1 - 2\delta$,

$$\mathsf{E}_{x,y}[\hat{h}_R^T Rxy \leq 0] \leq \mathsf{E}_{x,y}[1\left(\cos(\theta_{xy}^{h^*}) < \gamma\right)] + c\sqrt{\frac{k + \log(1/\delta)}{N}} \qquad (2)$$

provided $k \geq CK^4 \left(w(T_{h,\gamma}^{+}) + \sqrt{\log(1/\delta)} \right)^2 \gamma^{-1}$, where $w(T) \equiv \mathsf{E}_{g \sim \mathcal{N}(0,I)} \sup_{x \in T} g^T x$ denotes Gaussian width of set T.

Back in the original data space

We have seen:

- RP has ability to exploit benign geometric structure – this explains why compressive classification works well on e.g. image data sets, but not so well on noisy medical data.

- Easy problem = has good structure = classifier works well on RP-ed data; difficult problem = the opposite

Can we exploit this insight to better understand the original problem?

- Hypothetical RP of inputs to capture the difficulty of the problem instance.

Back in the original data space

Theorem Fix any positive integer $k(\leq d)$. For any $\delta > 0$, with probability at least $1-\delta$ with respect to the random draws of $\mathcal{T}^N$ of size N, uniformly $\forall h \in \mathcal{H}$ it holds:

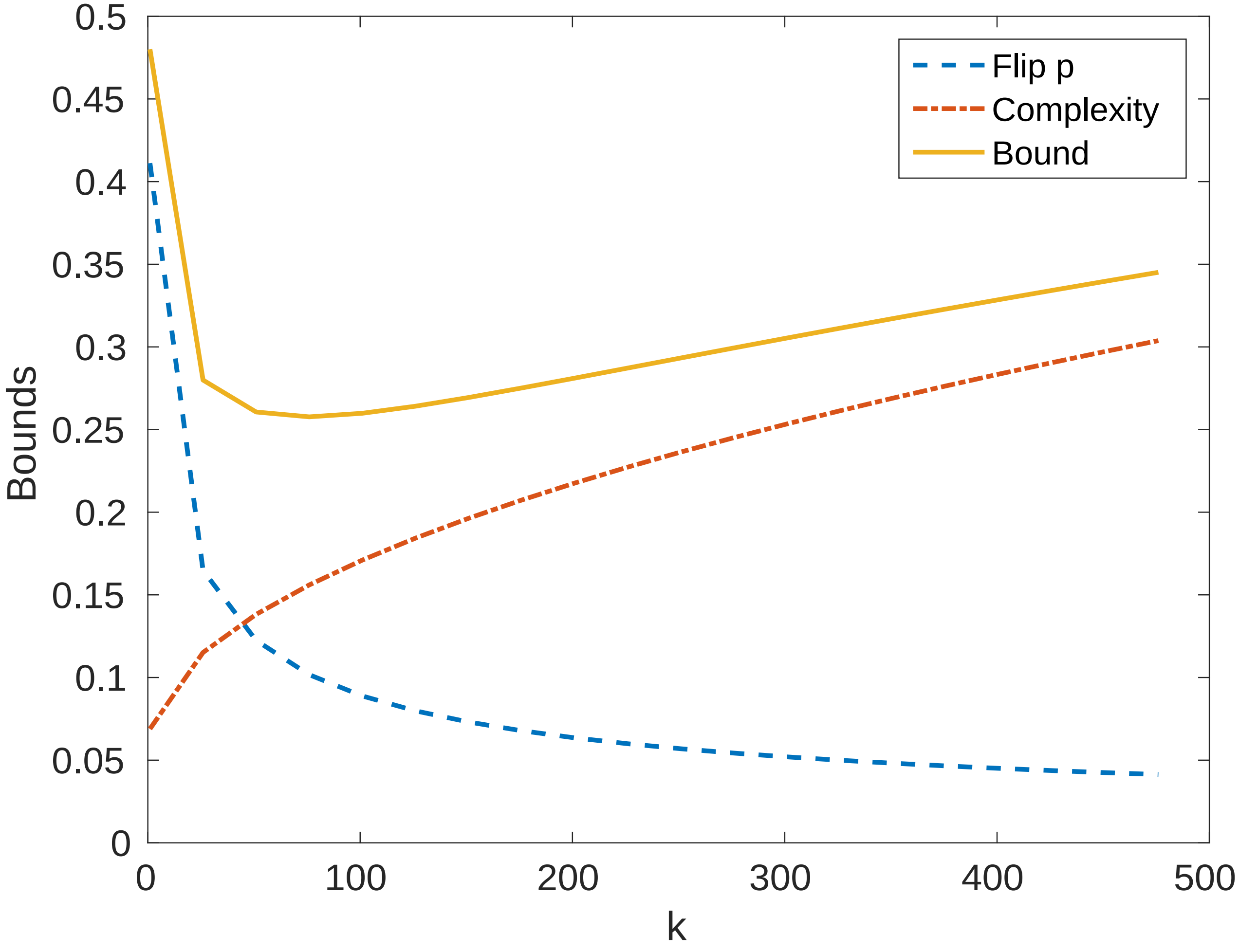
$$\mathsf{E}_{x,y}[\mathbf{1}(h^T xy \leq 0)] \quad \leq \quad \frac{1}{N}\sum_{n=1}^{N} \min(1, 2f_k(\theta^h_{x_n y_n})) + \frac{2\sqrt{2}}{\sqrt{\pi}}\sqrt{\frac{k}{N}} + 3\sqrt{\frac{\log(2/\delta)}{2N}}$$

• The ‘complexity-term’ (k) replaces VC-dimension at the expense of the new empirical error term.

• Empirical error small if ‘benign structure’ present (for instance, margin).

• If $k \to d$ we recover the classical VC bound.

• Informative even if N small: Rather than wishing N was large, choose the matching complexity k, and measure the error from the empirical term.


Flip p
Complexity
Bound
Bounds
k
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0
100
200
300
400
500

Proof.

The following inequality is immediate:

$$\mathsf{E}_{x,y}[\mathbf{1}(h^T xy \leq 0)] \;\leq\; \mathsf{E}_{x,y}[\mathbf{1}(h^T xy \leq 0) + 2f_k(\theta^h_{xy})\mathbf{1}(h^T xy > 0)]$$

By Rademacher complexity bound, for any fixed $k > 0$,

$$\begin{aligned} \mathsf{E}_{x,y}[\mathbf{1}(h^T xy \leq 0)] \;&\leq\; \frac{1}{N}\sum_{n=1}^{N}\left\{\mathbf{1}(h^T x_n y_n \leq 0) + 2f_k(\theta^h_{x_n y_n}) \cdot \mathbf{1}(h^T x_n y_n > 0)\right\} \\ &+\; 2\hat{\mathcal{R}}_N(G_k) + 3\sqrt{\frac{\log(2/\delta)}{2N}} \qquad (3) \end{aligned}$$

where

$$G_k = \{u \to \mathbf{1}(h^T u \leq 0) + 2f_k(\theta^h_u) \cdot \mathbf{1}(h^T u > 0) : h \in \mathbb{R}^d\}$$

To compute the empirical Rademacher complexity, observe:

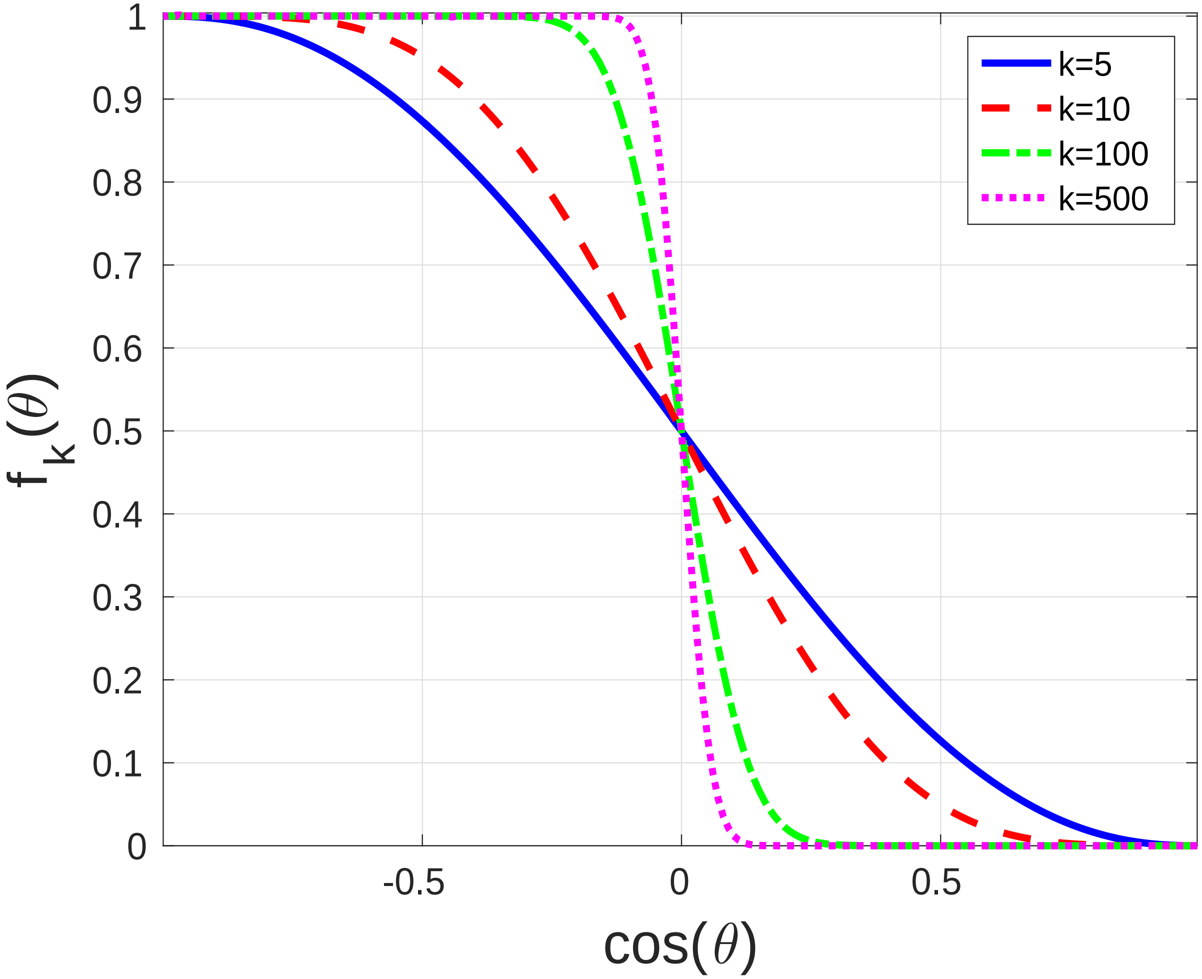
$$G_k = \ell_k \circ \mathcal{F}$$

where $\ell_k : [-1, 1] \to [0, 1]$,

$$\begin{aligned} \ell_k(a) &= 2\frac{\Gamma(k)}{(\Gamma(k/2))^2} \int_0^{\frac{1-a}{1+a}} \frac{z^{(k-2)/2}}{(1+z)^k} dz \cdot \mathbf{1}(a > 0) + \mathbf{1}(a \leq 0) \\ \mathcal{F} &= \{u \to \frac{h^T}{\|h\|} \frac{u}{\|u\|} : h \in \mathbb{R}^d\} \end{aligned}$$

We show that ℓ_k is Lipschitz with constant L: (1) on $a \in [-1, 0]$ it is constant; (2) on $a \in [0, 1]$ by Leibniz integration rule:

$$|\ell_k'(a)| = \left| -2\frac{\Gamma(k)}{2^{k-1}(\Gamma(k/2))^2} (1-a^2)^{\frac{k-2}{2}} \right| \leq 2\frac{\Gamma(k)}{(\Gamma(k/2))^2 \, 2^{k-1}} = L$$



The function $f_k(\theta)$ in the role of a classification loss.

Make L pretty ... (details omitted):

$$L \; = \; 2\frac{\Gamma(k/2+1/2)}{\sqrt{\pi}\ \Gamma(k/2)} \leq \sqrt{\frac{2k}{\pi}}$$

Using this, by Talagrand's contraction lemma,

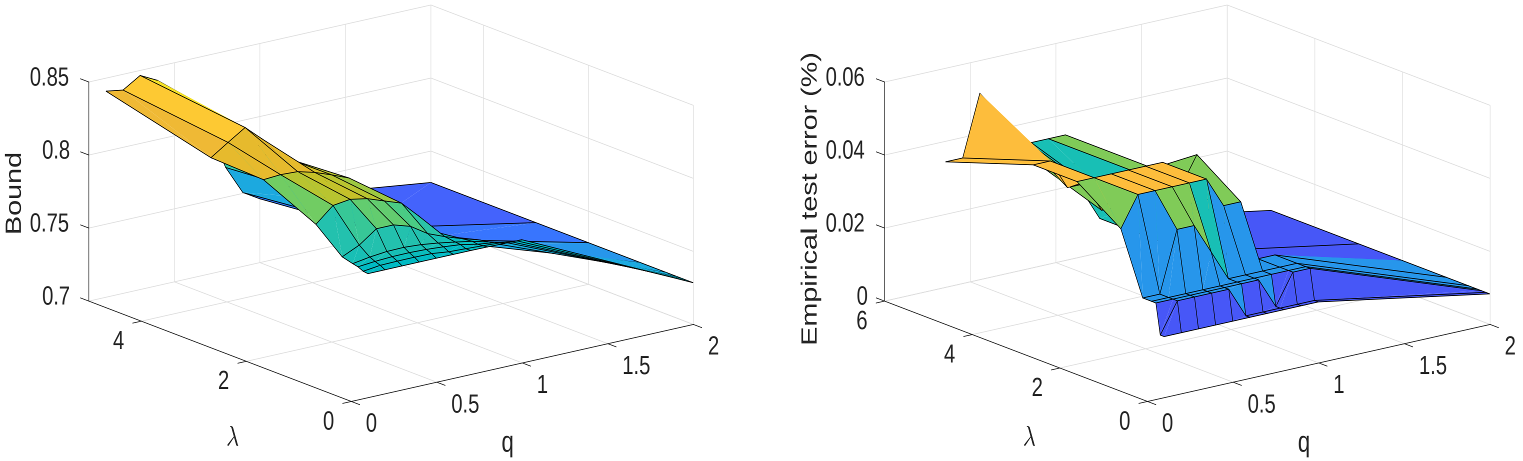
$$\hat{\mathcal{R}}_N(G_k) \; \leq \; \sqrt{\frac{2k}{\pi}} \cdot \hat{\mathcal{R}}_N(\mathcal{F}) \tag{4}$$

Note, $\mathcal{F}$ is a linear function class in $h/\|h\|$. Since and both $h/\|h\|$ and $xy/\|x\|$ have unit norm, so

$$\hat{\mathcal{R}}_N(\mathcal{F}) \leq \frac{1}{\sqrt{N}}$$

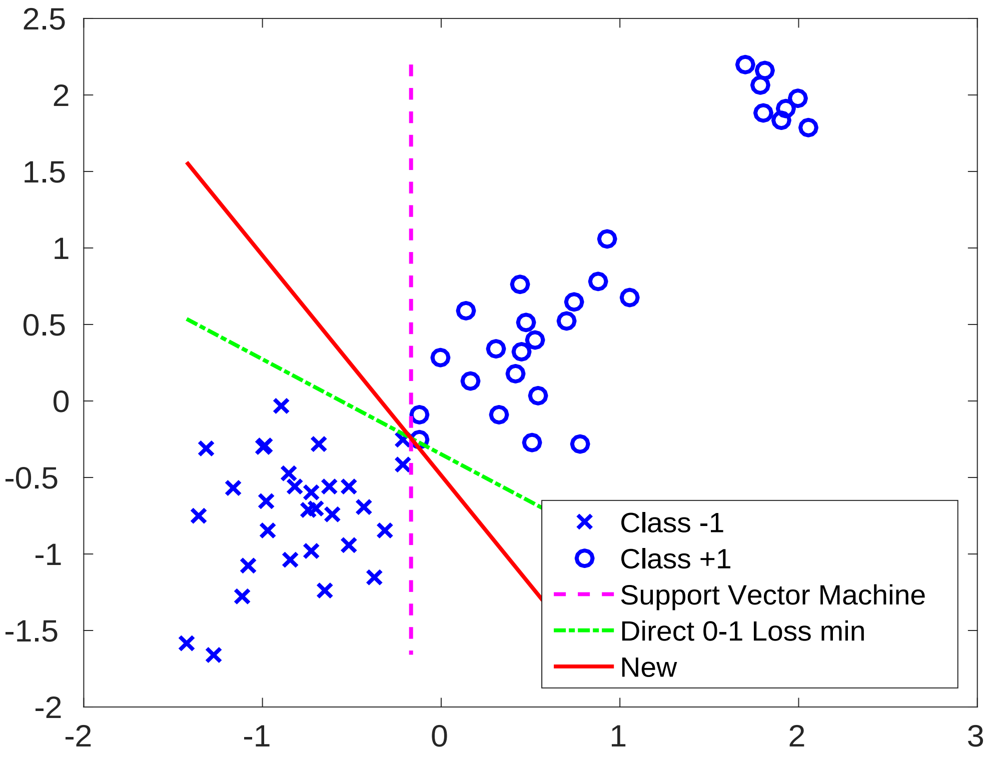
Combining this with eqs (3) and (4) completes the proof. ■

How good is the theory? - Predicting test error from training error



Behaviour of the bound (left) vs. the hold-out test error estimate (right), as the hyper-parameters of an L_q-regularised classifier are varied, on data that deceives sparse classifiers (smaller $q \sim$ sparser parameter vector).

Bound minimising classifier



Bound minimising classifier

Test error rates ± std for the bound optimizer in comparisons. Bold font indicates significant win against SVM at the 0.05 level cf. a paired t-test. Underline in last two columns indicates statistically significant loss of competing methods. No statistically significant loss of our approach has been observed on the data sets tested,

Data set	N	d	New	SVM	0-1 Loss	LDM
Australian	690	14	**0.137**± **0.015**	0.148± 0.013	0.156±0.077	0.149± 0.014
German	1000	24	**0.260**± **0.018**	0.280± 0.016	0.264±0.021	0.315±0.015
Haberman	306	3	**0.265**± **0.025**	0.285± 0.050	0.268±0.024	0.276±0.030
Parkinsons	195	22	**0.141**± **0.032**	0.221± 0.049	0.141± 0.036	0.135±0.034
PlRelax	182	12	**0.285**± **0.029**	0.361± 0.166	0.299±0.035	0.290±0.051
Sonar	208	60	0.256± 0.045	0.271± 0.036	0.245±0.044	0.264±0.044

How good is the theory? - New connections

Theorem Let $k : \mathbb{R}^d \times \mathcal{H} \to \mathbb{N}$ a deterministic function specified independently of $\mathcal{T}^N$. Then $\forall h \in \mathcal{H}$, with probability $1-\delta$ with respect to the random draw of a training set of size N, the generalization error of h is upper bounded as the following:

$$\begin{aligned}\mathsf{E}_{x,y}[h^T xy \leq 0] \quad &\leq \quad \frac{1}{N}\sum_{n=1}^{N} \min(1, 2f_{k(x_n y_n, h)}(\theta^h_{x_n y_n})) + 2\sqrt{\frac{2}{\pi}}\sqrt{\frac{1}{N}\max_{n=1}^{N} k(x_n y_n, h)} \\ &+ \quad 3\sqrt{\frac{\log(2/\delta)}{2N}} + 3\sqrt{\frac{\log(2)}{2}}\sqrt{\frac{1}{N}\max_{n=1}^{N} k(x_n y_n, h)}\end{aligned}$$

Proof.

· SRM allows choosing k after seeing the sample for $3\sqrt{\frac{\log(2)}{2}}\sqrt{\frac{k}{N}}$.
Choose $k := k_{\max} = \max_n k(x_n y_n, h)$.

· Observe that $f_k(\theta^h_{x_n y_n}) \geq f_{k_{\max}}(\theta^h_{x_n y_n})$.

$\Rightarrow$ RHS is upper bound on RHS of previous Thm. ■

Connection with the Large Margin Distribution Machine

Corollary W.p. $1-\delta$ (w.r.t. the training set of size N), $\forall h \in \mathcal{H}$,

$$\begin{aligned} \mathsf{E}_{x,y}[h^T xy \leq 0] \quad &\leq \quad \frac{1}{N}\sum_{n=1}^{N} 2\exp\left(-\frac{h^T x_n y_n}{\|h\|\cdot\|x_n\|}\right) + \frac{4}{\sqrt{\pi}}\frac{1}{\sqrt{N}} \cdot \max_n \sqrt{\frac{\|h\|\cdot\|x_n\|}{|h^T x_n|}} \\ &+ \quad 3\sqrt{\frac{\log(2/\delta)}{2N}} + 3\sqrt{\frac{\log(2)}{N}} \cdot \max_n \sqrt{\frac{\|h\|\cdot\|x_n\|}{|h^T x_n|}} \end{aligned} \tag{5}$$

Proof.

- Bound: $\min(1, 2f_k(\theta)) \leq 2\exp\left\{-\frac{k\cos^2(\theta)\cdot\mathsf{sgn}(\cos(\theta))}{2}\right\}$
- Choose $k(xy,h) := \frac{2}{|\cos(\theta^h_{xy})|}$
- Plug into previous Thm. ■

Lo & behold: Denote $\gamma^h_n = \cos(\theta^h_{x_n y_n})$;

$\frac{1}{N}\sum_{n=1}^{N} \exp\left(-\frac{h^T x_n y_n}{\|h\|\cdot\|x_n\|}\right) = \frac{1}{N}\sum_{n=1}^{N} \exp\left(-\gamma^h_n\right) = 1 - \frac{1}{N}\sum_{n=1}^{N} \gamma^h_n + \frac{1}{N}\sum_{n=1}^{N} (\gamma^h_n)^2 - \ldots$

Linear combination of weighted ensemble: Connecting two views

Consider a linearly weighted ensemble of binary valued base learners from the class $B = \{b : \mathcal{X} \times \{-1,1\}\}$, with weights $\alpha = (\alpha_1, \alpha_2, ..., \alpha_T)$:

$$F_{ens} = \left\{ x \to \sum_{t=1}^{T} \alpha_t b_t(x) : b_t \in B, \sum_{i=1}^{T} |\alpha_i| \leq 1 \right\} \tag{6}$$

Corollary [Margin distribution view] Fix any $k(\leq T)$ positive integer, and $\delta > 0$. With probability $1-\delta$ w.r.t. the training set of size N, uniformly for all $\alpha_t, \sum_{t=1}^{T} |\alpha_t| \leq 1$ and all $b_t \in B, t = 1, ..., T$,

$$\mathsf{E}_{x,y}[\sum_{t=1}^{T} \alpha_t b_t(x) y \leq 0] \leq \frac{1}{N} \sum_{n=1}^{N} \min\left(1, 2 f_k(\theta^{\alpha}_{b(x_n)y_n})\right) + c\sqrt{\frac{k \cdot V(B)}{N}} + 3\sqrt{\frac{\log(2/\delta)}{2N}} \tag{7}$$

where V denotes VC-dimension, and c is an absolute constant.

Linear combination of weighted ensemble: Connecting 2 views

Applying the 'local' Theorem with the choice $k(h, b(x)y) := \frac{2\|b(x)\|_2}{|\cos(\theta^{\alpha}_{b(x)y})|} \cdot \frac{\|\alpha\|_2}{\|\alpha\|_1}$, where b is the vector of binary predictions $(b_t)_{t=1,...,T}$, yields:

Corollary [Exponential loss view] With probability $1-\delta$ w.r.t. the training set of size N, uniformly for all $\alpha_t, \sum_{t=1}^{T} |\alpha_t| \leq 1$ and all $b_t \in B, t = 1, ..., T$,

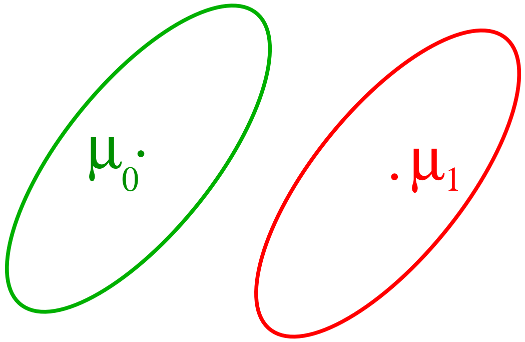
$$\begin{aligned} \mathsf{E}_{x,y}[\alpha^T b(x)y \leq 0] &\leq \frac{1}{N}\sum_{n=1}^{N} 2\exp\left(-\frac{\alpha^T b(x_n)y_n}{\|\alpha\|_1}\right) + 3\sqrt{\frac{\log(2/\delta)}{2N}} \\ &+ \left(c\sqrt{\frac{V(B)}{N}} + 3\sqrt{\frac{\log(2)}{2N}}\right)\sqrt{2T}\max_n \sqrt{\frac{\|\alpha\|_1}{|\alpha^T b(x_n)|}} \end{aligned}$$

Summing up linear discriminative classification

For linear classifiers trained by ERM on i.i.d. traiing sample (and no other assumptions a-priori), the use of RP revealed:

- The task is solvable in a random linear subspace (i.e. with performance guarantees) if the label flipping probabilities under a RP are small. This requirement is more general than large margin.
- The dataspace ERM classifier's error is small under the same conditions.
- We did not require any sparse representation for our bounds to hold, as usually compressed learning approaches do.

Generative classification. Fisher's Linear Discriminant (FLD)



- Simple and popular linear classifier, in widespread application. Classes are modelled as identical multivariate Gaussians.

- Assign class label to any query point according to its Mahalanobis distance from the class means.

- Simple enough to allow a deeper analysis addressing our questions.

Approach

- The FLD model assumes equal class covariances, but the true class covariances may differ.

- We need good bounds on largest and smalest eigenvalue of random-projected projected covariance matrix.

Definition. Let Σ be a trace class covariance matrix in a separable Hilbert space, i.e. $\mathrm{Tr}(\Sigma) < \infty$, and denote by $\lambda_{\mathrm{max}}(\Sigma)$ its largest eigenvalue. The effective rank of Σ is defined as $r(\Sigma) = \frac{\mathrm{Tr}(\Sigma)}{\lambda_{\mathrm{max}}(\Sigma)}$. For the model under consideration we will call this the *effective dimension* of the data.

Dimension-free Bounds on the Extreme Eigenvalues of Weighted Wishart

Lemma Let Σ be a covariance matrix in $\mathbb{R}^d$, and we denote by $\lambda_{\textsf{max}}(\cdot)$ and $\lambda_{\textsf{min}}(\cdot)$ its largest and smallest eigenvalues. Let R be a $k \times d$ random matrix with i.i.d. standard Gaussian entries. For any $\epsilon > 0$ we have w.p. at least $1 - \exp(-\epsilon^2/2)$:

$$\lambda_{\textsf{max}}(R\Sigma R^T) \;\leq\; \left(\sqrt{Tr(\Sigma)} + \sqrt{k \cdot \lambda_{\textsf{max}}(\Sigma)} + \epsilon\right)^2. \tag{8}$$

If $k < \lfloor \frac{\textsf{Tr}(\Sigma)}{\lambda_{\textsf{max}}(\Sigma)} \rfloor$ then for any $\epsilon \in (0,1)$ we have with probability at least $1 - \exp(-\epsilon^2/2)$:

$$\lambda_{\textsf{min}}(R\Sigma R^T) \;\geq\; \left(\sqrt{Tr(\Sigma)} - \sqrt{k \cdot \lambda_{\textsf{max}}(\Sigma)} - \epsilon\right)_+^2. \tag{9}$$

Comparison with (Davidson & Szarek, 2001)

Our proof uses comparison inequalities for the suprema of Gaussian processes, extending work by (Davidson & Szarek, 2001).

Lemma [Davidson & Szarek, 2001] Let R be a $k \times d$ matrix with entries sampled i.i.d from $\mathcal{N}(0,1)$. Then for all $\epsilon > 0$ with probability at least $1 - 2\exp(-\epsilon^2/2)$ we have:

$$(\sqrt{d} - \sqrt{k} - \epsilon)_+^2 \leq \lambda_{\text{min}}(RR^T) \leq \lambda_{\text{max}}(RR^T) \leq (\sqrt{d} + \sqrt{k} + \epsilon)^2. \quad (10)$$

where the lower estimate requires that $k < d$.

Using this, an easy alternative for us would be to write:

$$d \cdot \lambda_{\text{min}}(\Sigma)(1 - \sqrt{k/d} - \epsilon/\sqrt{d})_+^2 \leq \lambda_{\text{min}}(R\Sigma R^T) \leq \lambda_{\text{max}}(R\Sigma R^T) \leq d \cdot \lambda_{\text{max}}(\Sigma)(1 + \sqrt{k/d} + \epsilon/\sqrt{d})^2$$

w.p. $1 - 2\exp(-\epsilon^2/2)$, where $(\cdot)_+ = \max(\cdot, 0)$.

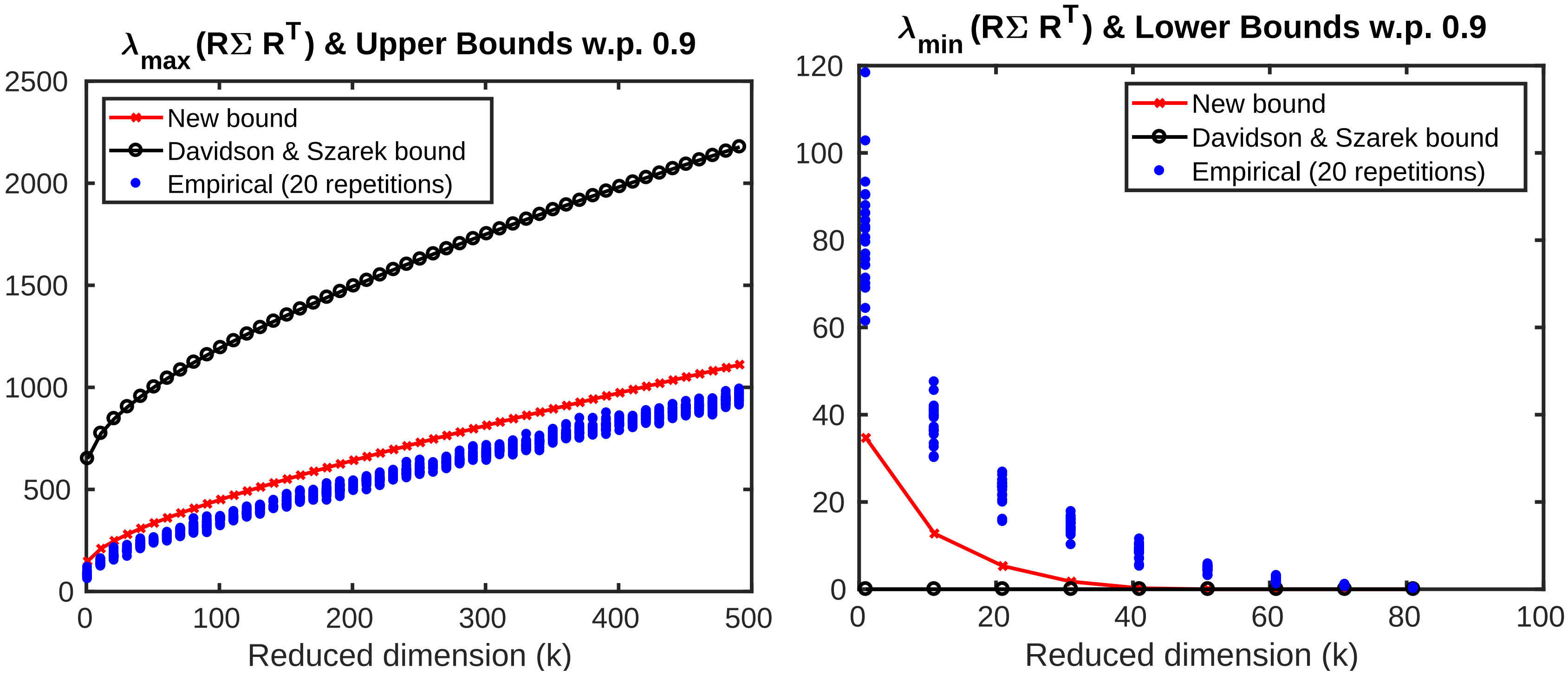


Illustration of the bounds. The $d \times d$ covariance matrix, $d = 500$. Σ has its first 80 eigenvalues equal to 1 and the remaining eigenvalues decay as the sequence $(1/i^2)_{i=1,\dots,d-80}$.

- When$d$ is finite, and $\Sigma = I_d$, we recover exactly the upper and lower estimates of (Davdson & Szarek, 2001), of which the upper bound on the largest eigenvalue is known to be sharp.

- Our λ_{max} bound is tighter than $\lambda_{\text{max}}(R\Sigma R^T) \leq \lambda_{\text{max}}(\Sigma)\lambda_{\text{max}}(RR^T)$ with the bound of (Davdson & Szarek, 2001) on the latter term – indeed, the effective dimension Tr$(\Sigma)/\lambda_{\text{max}}(\Sigma)$ is always no larger than the ambient dimension d.

- There are cases when our λ_{min} bound is also tighter than $\lambda_{\text{min}}(R\Sigma R^T) \geq \lambda_{\text{min}}(\Sigma)\lambda_{\text{min}}(RR^T)$ with the bound of (Davdson & Szarek, 2001) applied on the latter term.

- d can be arbitrary large as our bounds do not depend on d directly.

Application to Compressive FLD

In the true data distribution we assume multivariate Gaussian classes, but the true class covariances need not be shared. Need only that the true class covariances have finite trace.

The FLD model assumes a shared covariance. Covariance misspecification effects are part of our analysis.

Denote by $\hat{\Sigma}$ the ML estimate of the pooled covariance and by $\hat{\mu}_0$ and $\hat{\mu}_1$ the ML estimates of the class means in $\mathbb{R}^d$. Decision function of compressive FLD for an input query point x is:

$$\hat{h}^R(x) = \mathbf{1}\left\{(\hat{\mu}_1 - \hat{\mu}_0)^T R^T (R\hat{\Sigma}R^T)^{-1} R \left(x - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}\right) > 0\right\}$$

Interested in the generalisation error $\Pr_{x,y}[\hat{h}^R(Rx) \neq y | \mathcal{T}_N^R]$.

Theorem. Let $(x, y) \sim \mathcal{D}$ be a query point with unknown label y and Gaussian class conditionals $x|y = i \sim \mathcal{N}(\mu_i, \Sigma_i)$ with trace class covariances, i.e. $\mathsf{Tr}(\Sigma_i) < \infty, \forall i = \{0, 1\}$. Let $\pi_i = \mathsf{Pr}(y = i)$ be bounded away from both 0 and 1. Let R be a $k \times d$ random matrix with i.i.d. standard Gaussian entries. Then, $\forall \epsilon \in (0, 1)$,

$$\mathsf{Pr}_{x,y}[\hat{h}^R(Rx) \neq y|\mathcal{T}_N] \leq ...$$

$$\sum_{i=0}^{1} \pi_i \Phi \left(- \left[[\sqrt{k} - \epsilon]_+ \frac{\left[\sqrt{\|\mu_1 - \mu_0\|^2 + \frac{\mathsf{Tr}(\beta_1 \Sigma_0 + \beta_0 \Sigma_1)}{N \alpha_0 \alpha_1}} - \epsilon \sqrt{\frac{\lambda_{\mathsf{max}}(\alpha_1 \Sigma_0 + \alpha_0 \Sigma_1)}{N \beta_0 \beta_1}} \right]_+}{\sqrt{\mathsf{Tr}(\Sigma_i)} + \sqrt{k \lambda_{\mathsf{max}}(\Sigma_i)} + \epsilon} g(\tilde{\kappa}_i) - \frac{\sqrt{k} + \epsilon}{\sqrt{N \beta_i}} \right]_+ \right)$$

w.p. $1 - 10 \exp(-\epsilon^2/2) - 2 \exp(-\pi_0 N \epsilon^2/3)$, where:

- Φ is the standard Gaussian cumulative distribution function,
- $\alpha_0 = \pi_0(1+\epsilon)$, $\alpha_1 = 1-\pi_0(1-\epsilon)$, $\beta_0 = \pi_0(1-\epsilon), \beta_1 = 1-\pi_0(1+\epsilon)$,
- $g(\tilde{\kappa}_i) = \frac{\sqrt{\tilde{\kappa}_i}}{1+\tilde{\kappa}_i}$,
- $\tilde{\kappa}_i$... (on next slide)

In the case when $\Sigma_0 \neq \Sigma_1$, then

$$\tilde{\kappa}_i = \frac{\left[(\sqrt{N-2}+\sqrt{k}+\epsilon)^2 + (\sqrt{N\alpha_{\neg i}-1}+\sqrt{k}+\epsilon)^2 \tilde{\lambda}_{\mathsf{max}}(M_i) - (\sqrt{N\beta_{\neg i}-1}-\sqrt{k}-\epsilon)_+^2\right]_+}{\left[(\sqrt{N-2}-\sqrt{k}-\epsilon)_+^2 + (\sqrt{N\beta_{\neg i}-1}-\sqrt{k}-\epsilon)_+^2 \tilde{\lambda}_{\mathsf{min}}(M_i) - (\sqrt{N\alpha_{\neg i}-1}+\sqrt{k}+\epsilon)^2\right]_+}, \tag{11}$$

provided that k and N are such that this is finite.

In the above, $M_i := (R\Sigma_i R^T)^{-1/2} R\Sigma_{\neg i} R^T (R\Sigma_i R^T)^{-1/2}$ encodes the mismatch between the true class-covariances after RP, so:

$$\begin{aligned}
\tilde{\lambda}_{\mathsf{max}}(M_i) &= \mathsf{min}\left\{\frac{(\sqrt{\mathsf{Tr}(\Sigma_{\neg i})}+\sqrt{k\cdot\lambda_{\mathsf{max}}(\Sigma_{\neg i})}+\epsilon)^2}{(\sqrt{\mathsf{Tr}(\Sigma_i)}-\sqrt{k\cdot\lambda_{\mathsf{max}}(\Sigma_i)}-\epsilon)_+^2}, \lambda_{\mathsf{max}}(\Sigma_i^+\Sigma_{\neg i})\right\} \quad (12)\\
\tilde{\lambda}_{\mathsf{min}}(M_i) &= \mathsf{max}\left\{\frac{(\sqrt{\mathsf{Tr}(\Sigma_i)}-\sqrt{k\cdot\lambda_{\mathsf{max}}(\Sigma_i)}-\epsilon)_+^2}{(\sqrt{\mathsf{Tr}(\Sigma_{\neg i})}+\sqrt{k\cdot\lambda_{\mathsf{max}}(\Sigma_{\neg i})}+\epsilon)^2}, \lambda_{\mathsf{min}}(\Sigma_i^+\Sigma_{\neg i})\right\} \quad (13)
\end{aligned}$$

where $(\cdot)^+$ stands for any choice of generalised inverse.

Corollary [Asymptotic error bound]. Under the conditions of Theorem, and using the same function $g(\cdot)$, we have:

$$\limsup_{N\to\infty} \mathsf{Pr}_{x,y}[\hat{h}^R(Rx) \neq y|\mathcal{T}_N] \leq \sum_{i=0}^{1} \pi_i \Phi\left(-\frac{(\sqrt{k}-\epsilon)\|\mu_1-\mu_0\|}{\sqrt{\mathsf{Tr}(\Sigma_i)}+\sqrt{k\cdot\lambda_{\mathsf{max}}(\Sigma_i)}+\epsilon}\cdot g\left(\frac{\tilde{\lambda}_{\mathsf{max}}(M_i)}{\tilde{\lambda}_{\mathsf{min}}(M_i)}\right)\right)$$

w.p. $1-4\exp(-\epsilon^2/2)$, where $\tilde{\lambda}_{\mathsf{max}}(M_i)$ and $\tilde{\lambda}_{\mathsf{min}}(M_i)$ are as previously, in eqs. (12)-(13).

Main characteristics:

- upper bound on the Bayes error of FLD
- distance between class means relative to size of covariances $\lambda_{\mathsf{max}}(\Sigma_i)$ plays a crucial role
- $-g(\cdot)$ is the price of covariance misestimation and/or misspecification of a shared covariance: increases with the condition number of $M_i = (R\Sigma_i R^T)^{-1/2} R\Sigma_{\neg i} R^T (R\Sigma_i R^T)^{-1/2}$
- bound is independent of d, scales with the *effective dimension* $\mathsf{Tr}(\Sigma_i)/\lambda_{\mathsf{max}}(\Sigma_i)$

Proof sketch & further interpretation

We start from the error of FLD conditional on training set, applied in the RP space, and decompose it:

$$
\begin{aligned}
&\text{Pr}_{x,y}[\hat{h}^R(Rx) \neq y | \mathcal{T}_N, R] = \ldots \\
&\sum_{i=0}^{1} \pi_i \Phi \left(-\frac{1}{2} \frac{(\hat{\mu}_{\neg i} - \hat{\mu}_i)^T R^T (R\hat{\Sigma}R^T)^{-1} R \, (\hat{\mu}_{\neg i} + \hat{\mu}_i - 2\mu_i)}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T R^T (R\hat{\Sigma}R^T)^{-1} R\Sigma_i R^T (R\hat{\Sigma}R^T)^{-1} R(\hat{\mu}_1 - \hat{\mu}_0)}} \right) = \ldots \\
&\sum_{i=0}^{1} \pi_i \Phi \left(-\frac{1}{2} \frac{(\hat{\mu}_1 - \hat{\mu}_0)^T R^T (R\hat{\Sigma}R^T)^{-1} R \, (\hat{\mu}_1 - \hat{\mu}_0) - 2\,(\mu_i - \hat{\mu}_i)^T R^T (R\hat{\Sigma}R^T)^{-1} R \, (\hat{\mu}_{\neg i} - \hat{\mu}_i)}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T R^T (R\hat{\Sigma}R^T)^{-1} R\Sigma_i R^T (R\hat{\Sigma}R^T)^{-1} R(\hat{\mu}_1 - \hat{\mu}_0)}} \right) \leq \\
&\sum_{i=0}^{1} \pi_i \Phi \left(-[A_i B_i - C_i] \right) \qquad (14)
\end{aligned}
$$

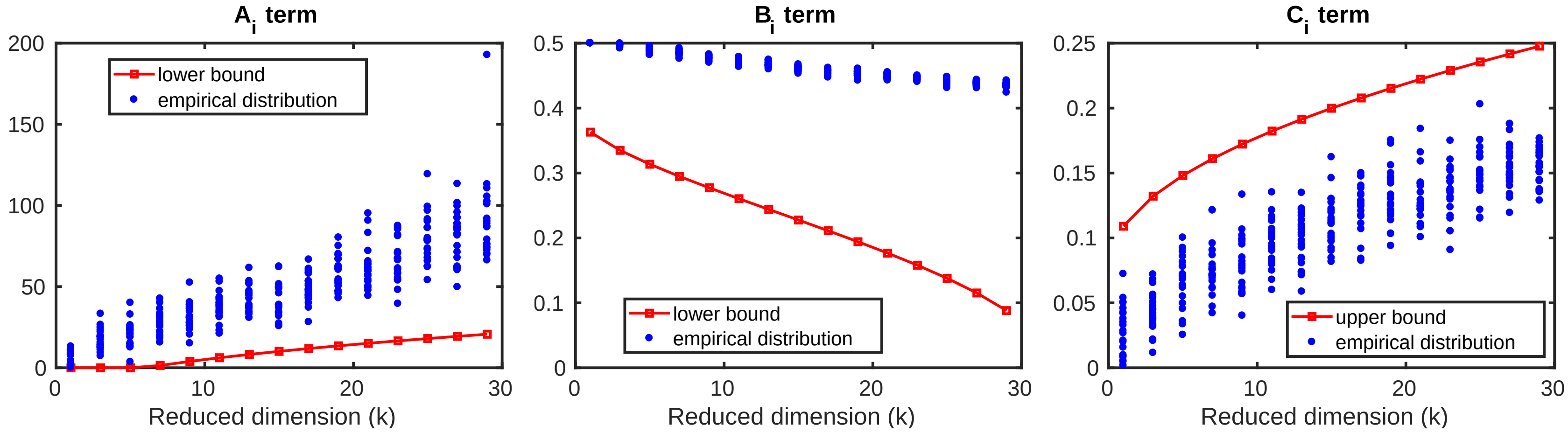
where

$$
\begin{array}{rcll}
A_i & := & \|(R\Sigma_i R^T)^{-\frac{1}{2}} R \, (\hat{\mu}_1 - \hat{\mu}_0) \, \| & \text{we lower bound this} \\
B_i & := & \dfrac{\sqrt{\kappa((R\hat{\Sigma}R^T)^{-\frac{1}{2}} R\Sigma_i R^T (R\hat{\Sigma}R^T)^{-\frac{1}{2}})}}{1 + \kappa((R\hat{\Sigma}R^T)^{-\frac{1}{2}} R\Sigma_i R^T (R\hat{\Sigma}R^T)^{-\frac{1}{2}})} & \text{we lower bound this} \\
C_i & := & \|(R\Sigma_i R^T)^{-\frac{1}{2}} R(\mu_i - \hat{\mu}_i)\| & \text{we upper bound this}
\end{array}
$$

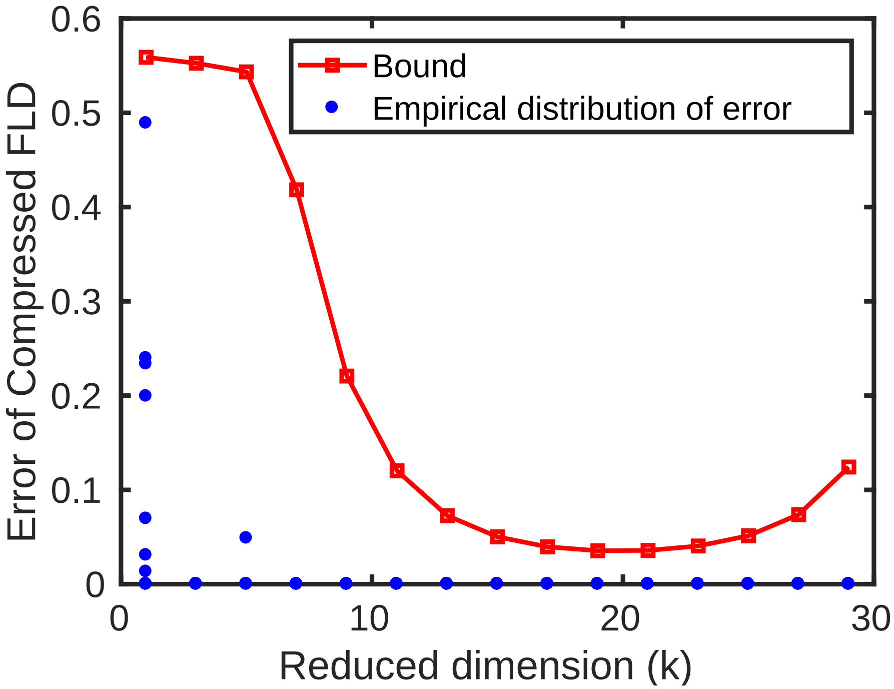
Experiments on synthetic data

- From analysis, low effective dimension of class covariances, and large distance between class means are benign characteristacs.
- Generated such data in $d = 500$:

$\mathrm{Tr}(\Sigma_0)/\lambda_{\max}(\Sigma_0) = 28.09$, $\mathrm{Tr}(\Sigma_1)/\lambda_{\max}(\Sigma_1) = 12.98$; $\|\mu_0 - \mu_1\| = 44.72$. $N_0 = N_1 = 1000$ for training + same amount for testing. Independent repeats 20 times.



Experiments on synthetic data: Bound vs. empirical test error



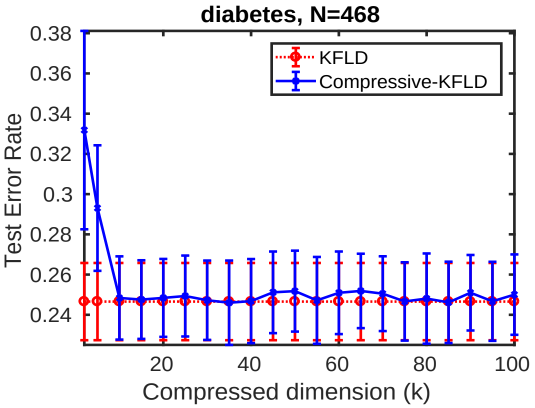
$\mathsf{Tr}(\Sigma_0)/\lambda_{\mathsf{max}}(\Sigma_0) = 28.09$, $\mathsf{Tr}(\Sigma_1)/\lambda_{\mathsf{max}}(\Sigma_1) = 12.98$; $\|\mu_0 - \mu_1\| = 44.72$. $N_0 = N_1 = 1000$ for training + same amount for testing. Independent repeats 20 times.

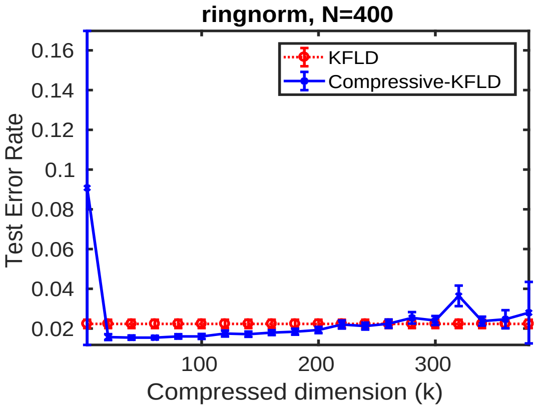
Experiments on real data

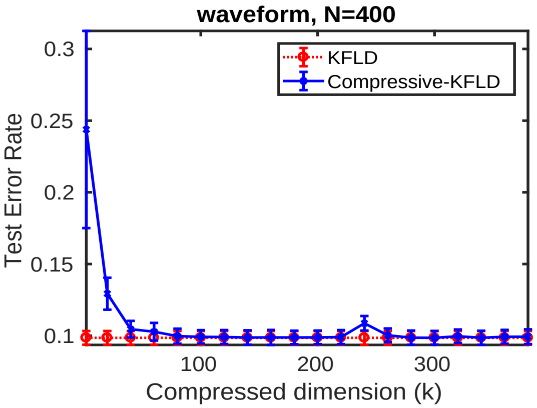
d can be arbitrary large – including functional data.
Kernel methods is a context where functional data appears frequently.

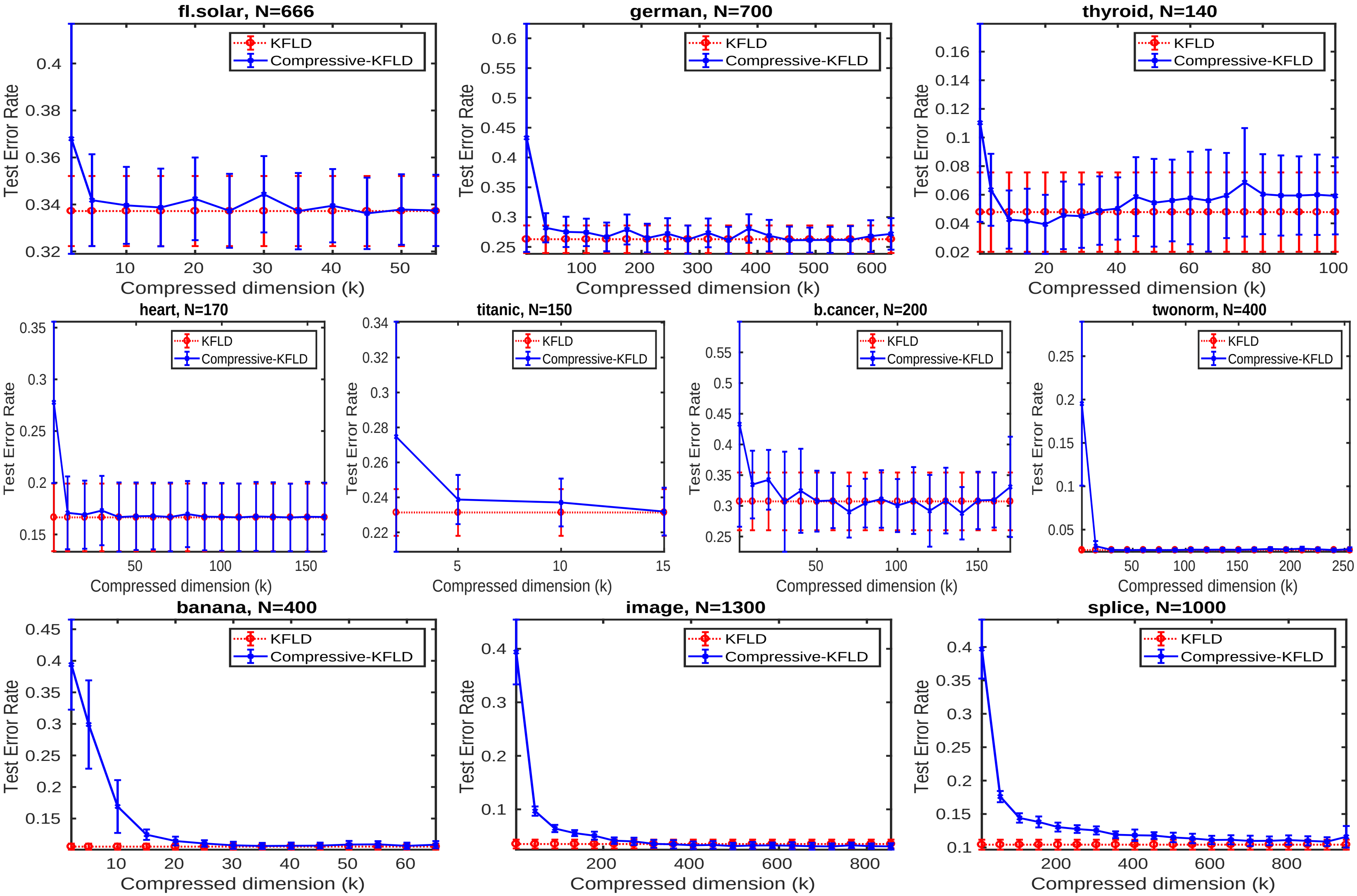
Kernel trick is a smooth transformation – reason to expect low effective dimension & separability of the classes.

13 UCI benchmark data sets previously used in studying kernel-FLD (KFLD).
KFLD with ridge-regularised covariance vs. compressive KFLD no regularisation.








fl.solar, N=666
german, N=700
thyroid, N=140
heart, N=170
titanic, N=150
b.cancer, N=200
twonorm, N=400
banana, N=400
image, N=1300
splice, N=1000
KFLD
Compressive-KFLD
Test Error Rate
Compressed dimension (k)
0.32
0.34
0.36
0.38
0.4
10
20
30
40
50
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6
100
200
300
400
500
600
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
20
40
60
80
100
0.15
0.2
0.25
0.3
0.35
50
100
150
0.22
0.24
0.26
0.28
0.3
0.32
0.34
5
10
15
0.25
0.3
0.35
0.4
0.45
0.5
0.55
50
100
150
0.05
0.1
0.15
0.2
0.25
50
100
150
200
250
0.15
0.2
0.25
0.3
0.35
0.4
0.45
10
20
30
40
50
60
0.1
0.2
0.3
0.4
200
400
600
800
0.1
0.15
0.2
0.25
0.3
0.35
0.4
200
400
600
800

Summing up compressive FLD

- Our analysis disentangles the effects of RP on various components of the error:
  - RP has beneficial effects on both misestimation and misspecification of the class covariances.
  - RP has beneficial effects on misestimation of class centers.
  - RP has detrimental effect by reducing class separability.
- Dimension free bound under mild assumptions.
- The key technical ingredient – new dimension-free bounds on the largest and smallest eigenvalues of the compresed covariance – may be of independent interest.

When a single RP looses too much

Noise can make the data fill more of the ambient space.
Not enough structure for a single RP.

- Can we achieve (or improve on) the classification performance in data space, using a compressive ensemble?
- Can we understand how the ensemble acts to improve performance?
- Can we interpret the pramters of the compressive ensemble in the original data space?

Ensemble of compressive FLDs

Interested in $N \ll d$ setting, which is a common situation e.g. medical imaging, genomics, proteomics, etc.

For a single RP FLD classifier, the decision rule is:

$$\mathbf{1}\left\{(\hat{\mu}_1 - \hat{\mu}_0)^T R^T \left(R\hat{\Sigma}R^T\right)^{-1} R\left(x_q - \frac{\hat{\mu}_1 + \hat{\mu}_0}{2}\right) > 0\right\}$$

For the ensemble we will use:

$$\mathbf{1}\left\{\frac{1}{M}\sum_{i=1}^{M}(\hat{\mu}_1 - \hat{\mu}_0)^T R_i^T \left(R_i\hat{\Sigma}R_i^T\right)^{-1} R_i\left(x_q - \frac{\hat{\mu}_1 + \hat{\mu}_0}{2}\right) > 0\right\}$$

Linear combination rules are a common choice for ensembles.

Observation

We can rewrite decision rule as:

$$\mathbf{1}\left\{(\hat{\mu}_1 - \hat{\mu}_0)^T \ \frac{1}{M}\sum_{i=1}^{M} R_i^T \left(R_i \hat{\Sigma} R_i^T\right)^{-1} R_i \ \left(x_q - \frac{\hat{\mu}_1 + \hat{\mu}_0}{2}\right) > 0\right\}$$

Then, at convergence, enough to consider:

$$\lim_{M\to\infty} \frac{1}{M}\sum_{i=1}^{M} R_i^T \left(R_i \hat{\Sigma} R_i^T\right)^{-1} R_i \ = \ \mathsf{E}\left[R^T \left(R \hat{\Sigma} R^T\right)^{-1} R\right]$$

Ingredients of analysis

• Rows (and columns) of R drawn from a spherical Gaussian, hence for any orthogonal matrix U, $R \sim RU$. Eigendecomposing $\hat{\Sigma} = U\hat{\Lambda}U^T$ and using $UU^T = I$ we find that:

$$\mathsf{E}\left[R^T\left(R\hat{\Sigma}R^T\right)^{-1}R\right] = U\ \mathsf{E}\left[R^T\left(R\hat{\Lambda}R^T\right)^{-1}R\right]U^T \qquad (15)$$

• Furthermore since a matrix A is diagonal if and only if $VAV^T = A$ for all *diagonal* orthogonal matrices $V = \text{diag}\{\pm 1\}$ one can easily show that the expectation on RHS is diagonal.

• Now enough to evaluate the diagonal terms on RHS!

• (Marzetta et al.'11) has an algorithm that calculates this exactly. We are more interested in interpretable estimates, to relate it to $\hat{\Sigma}$.

Ingredients of analysis

Define $\rho := \text{rank}(\hat{\Lambda}) = \text{rank}(\hat{\Sigma})$.
Work with positive semidefinite ordering: $A \succeq B \iff A - B$ is positive semidefinite (p.s.d $\equiv$ symmetric with all eigenvalues $\geqslant 0$).
Upper and lower bound the diagonal matrix expectation (15) in the p.s.d ordering with spherical matrices $\alpha_{\max} \cdot I$, $\alpha_{\min} \cdot I$ to bound its condition number in terms of *data space parameters*:

$$\alpha_{\max} \cdot I \succeq \mathsf{E}\left[R^T \left(R\Lambda R^T\right)^{-1} R\right] \succeq \alpha_{\min} \cdot I$$

Where $\alpha = \alpha(k, \rho, \lambda_{\max}, \lambda_{\min \neq 0})$, k is the projected dimensionality, $\rho = \text{rank}(\hat{\Lambda}) = \text{rank}(\hat{\Sigma})$, $\lambda_{\max}$ and $\lambda_{\min \neq 0}$ are respectively the greatest and least non-zero eigenvalues of $\hat{\Sigma}$.

Results: The regularisation effect

Theorem [D-K, MLJ] Let $\hat{\Sigma} \in \mathcal{M}_{d\times d}$ be a symmetric positive semi-definite matrix with rank $\rho \in \{3, ..., d-1\}$, and denote by $\lambda_{\max}(\hat{\Sigma}), \lambda_{\min\neq 0}(\hat{\Sigma}) > 0$ its greatest and least non-zero eigenvalues. Let $k < \rho - 1$ be a positive integer, and let $R \in \mathcal{M}_{k\times d}$ be a random matrix with i.i.d $\mathcal{N}(0,1)$ entries. Let $\hat{S}^{-1} :=$ E $\left[R^T \left(R\hat{\Sigma}R^T\right)^{-1} R\right]$, and denote by $\kappa(\hat{S}^{-1})$ its condition number, $\kappa(\hat{S}^{-1}) = \lambda_{\max}(\hat{S}^{-1})/\lambda_{\min}(\hat{S}^{-1})$. Then:

$$\kappa(\hat{S}^{-1}) \leqslant \frac{\rho}{\rho - k - 1} \cdot \frac{\lambda_{\max}(\hat{\Sigma})}{\lambda_{\min\neq 0}(\hat{\Sigma})}$$

This theorem implies that for a large enough ensemble the condition number of the sum of random matrices $\frac{1}{M}\sum_{i=1}^{M} R_i^T \left(R_i\hat{\Sigma}R_i^T\right)^{-1} R_i$ is bounded.

Exact Generalisation error of the converged ensemble conditioned on fixed training set

Lemma [D-K, MLJ]. Let $x_q|y_q \sim \mathcal{N}(\mu_y, \Sigma)$, where $\Sigma \in \mathcal{M}_{d\times d}$ is a full rank covariance matrix. Let $R \in \mathcal{M}_{k\times d}$ be a RP matrix with i.i.d. Gaussian entries and denote $S_R^{-1} := \frac{1}{M}\sum_{i=1}^{M} R_i^T \left(R_i\hat{\Sigma}R_i^T\right)^{-1} R_i$. Then the error of the ensemble conditioned on training set equals:

$$\sum_{y=0}^{1} \pi_y \Phi \left(-\frac{1}{2} \frac{(\hat{\mu}_{\neg y} - \hat{\mu}_y)^T S_R^{-1}(\hat{\mu}_0 + \hat{\mu}_1 - 2\mu_y)}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T S_R^{-1} \Sigma S_R^{-1} (\hat{\mu}_1 - \hat{\mu}_0)}} \right)$$

For the converged ensemble, substitute the expectation (15) for S_R^{-1} above.

Generalisation error of the converged ensemble

Theorem [D-K, MLJ]. Let $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^{N}$ be a set of training data of size $N = N_0 + N_1$, subject to $N < d$ and $N_y > 1\ \forall y$. Let x_q be a query point with Gaussian class-conditionals $x_q | y_q \sim \mathcal{N}(\mu_y, \Sigma)$, and let $\Pr\{y_q = y\} = \pi_y$. Let ρ be the rank of the maximum likelihood estimate of the covariance matrix and let $k < \rho - 1$ be a positive integer. Then for any $\delta \in (0, 1)$ we have w.p. $1 - \delta$ w,r,t, random draws of $\mathcal{T}$:

$$\Pr_{x_q,y_q}\left(\hat{h}_{ens}(x_q) \neq y_q\right) \leqslant \sum_{y=0}^{1} \pi_y \Phi\left(-\left[g\left(\bar{\kappa}\left(\sqrt{2\log\frac{5}{\delta}}\right)\right) \times \dots\right.\right. \tag{16}$$

$$\left.\left.\dots \left[\sqrt{\|\Sigma^{-\frac{1}{2}}(\mu_1 - \mu_0)\|^2 + \frac{dN}{N_0 N_1}} - \sqrt{\frac{2N}{N_0 N_1}\log\frac{5}{\delta}}\right]_+ - \sqrt{\frac{d}{N_y}}\left(1 + \sqrt{\frac{2}{d}\log\frac{5}{\delta}}\right)\right]\right)$$

where $\bar{\kappa}(\epsilon)$ is a high probability (w.r.t draws of $\mathcal{T}$) upper bound on the condition number of $\Sigma\hat{S}^{-1}$ (given in the paper) and $g(\cdot)$ is the function $g(a) := \frac{\sqrt{a}}{1+a}$.

Experiments: Datasets

Datasets:

Name	Source	#samples	#features
colon	[Alon et al.]	62	2000
leukemia	[Golub et al.]	72	3571
leukemia large	[Golub et al.]	72	7129
prostate	[Singh et al.]	102	6033
duke	[West et al.]	44	7129

Standardised features to have mean 0 and variance 1 and ran experiments on 100 independent splits. In each split took 12 points for testing, rest for training.

Experiments: Results for $k = \rho/2$

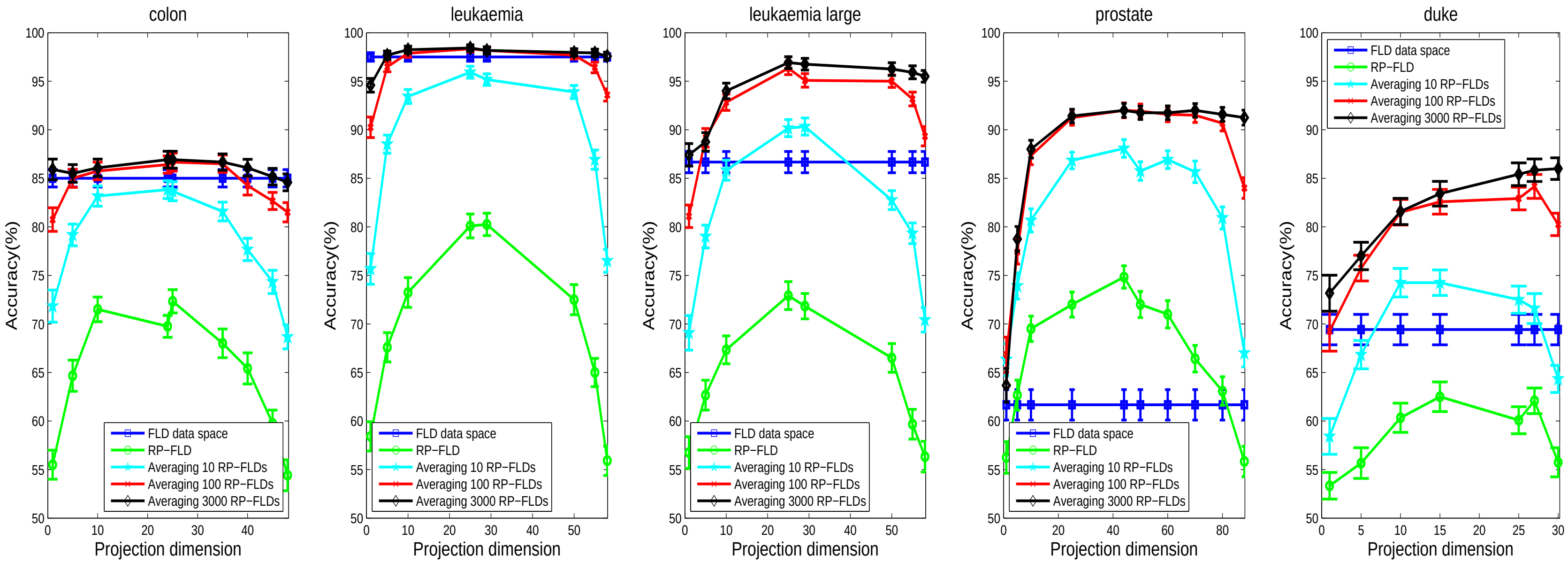
Base learners are compressive FLDs with full covariance and no regularisation when $k \leqslant \rho$ and pseudoinverted FLD when $k > \rho$.

Dataset	$\rho/2$	100 RP-FLD	1000 RP-FLD	SVM
colon	24	13.58 ± 0.89	13.08 ± 0.86	16.58 ± 0.95
leuk.	29	1.83 ± 0.36	1.83 ± 0.37	1.67 ± 0.36
leuk.lg.	29	4.91± 0.70	3.25 ± 0.60	3.50 ± 0.46
prost.	44	8.00 ± 0.76	8.00 ± 0.72	8.00 ± 0.72
duke	15	17.41 ±1.27	16.58 ± 1.27	13.50 ± 1.10

More comparisons: For data space experiments on colon and leukaemia used ridge-regularised FLD for comparison and fitted regularisation parameter using 5-fold CV.

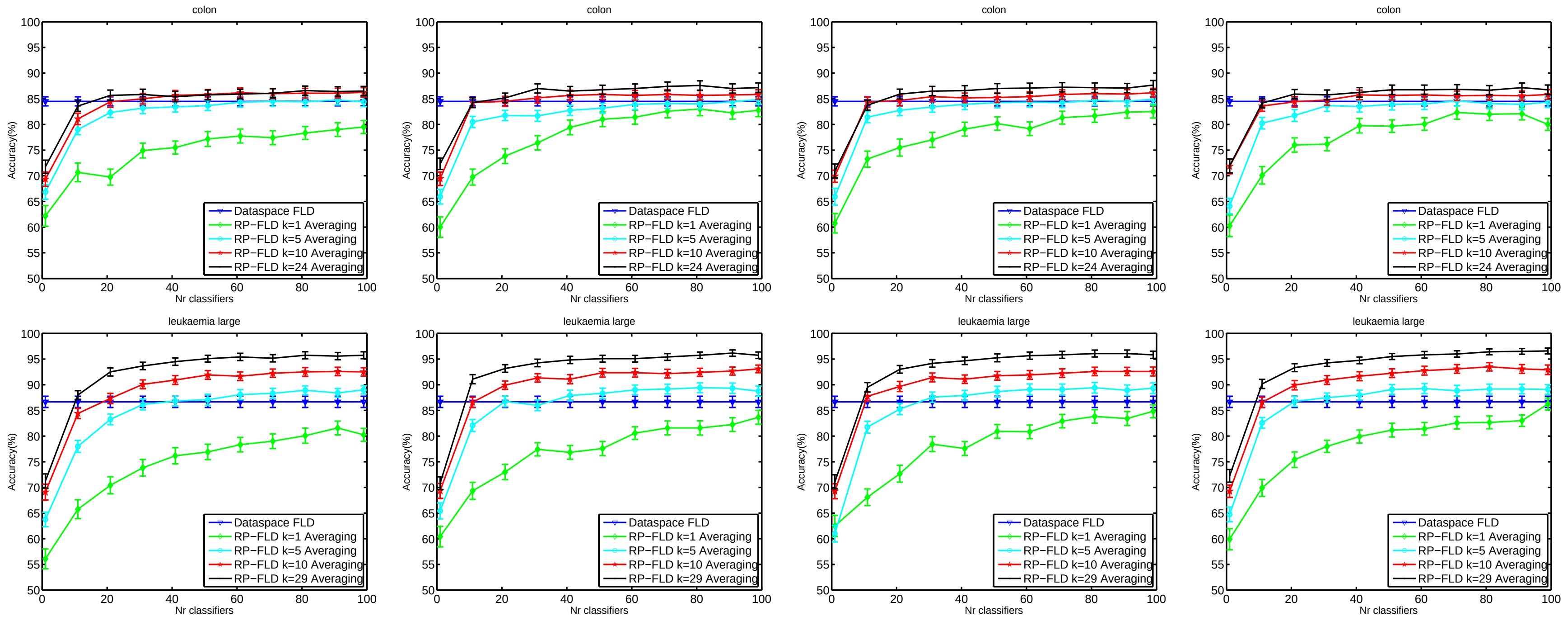
For other datasets we used diagonal FLD in the data space (size, no sig. diff. in error on colon, leuk.).

Experiments – effect of k



Test error rates versus k and error bars mark 1 standard error estimated from 100 runs. In these experiments we used Gaussian random matrices with i.i.d $\mathcal{N}(0, 1)$ entries.

Experiments – different RP matrices



Column 1: Majority Vote using Gaussian random matrices; *Column 2*: Averaging ensemble using Gaussian r.m; *Column 3*: Averaging ensemble using ± 1 random matrices. *Column 4*: Averaging ensemble using the sparse $\{-1, 0, +1\}$ random matrices from [Achlioptas '03].

How far is the finite ensemble from the infinite ensemble?

Approximating the inverse of a singular covariance matrix is a problem of independent interest (e.g. small sample problems with Gaussian mixtures, Gaussian graphical models).

Given M a $d \times d$ psd, rank $\rho < d$ matrix, the precision matrix approximator of our cFLD ensemble:

$$ic_k(\mathcal{M}) = E[R^T(R\mathcal{M}R^T)^{-1}R]$$

was previously proposed as a general purpose method in (Marzetta et al., 2011). A rotation-invarant method, different from sparsity approaches.

In (Marzetta et al., 2011),

- Shown to be always non-singular, regularised inverse of M.

- Empirically demonstrated to outperform the Ledoit-Wolf in terms of Frobenius norm from the unknown true covariance

- Theoretical analyses so far existed only for the infinite ensemble, implying good performance for ‘large-enough’ ensemble.

- How large is ‘large enough’?

Result: How large is large enough?

Theorem [K, ALT'17] Let M be a $d \times d$, rank $3 \leq \rho < d$ psd matrix. Let $R_i, i = 1, ..., m$ be i.i.d. copies of a $k \times d, k < \rho - 1$ random matrix with i.i.d. standard Gaussian entries, k finite. If $\rho - k + 1 \geq \Omega(\log(d - \rho))$, then, $\exists c, \eta > 0$ s.t. $\forall \epsilon \in (0, 1)$, we have:

$$\mathsf{E}\left[\| \frac{1}{m} \sum_{i=1}^{m} R_i^T (R_i M R_i^T)^{-1} R_i - \mathsf{E}\left[R^T (R M R^T)^{-1} R \right] \| \right] \leq \epsilon \cdot \mathsf{E}\left[R^T (R M R^T)^{-1} R \right]$$

provided that the ensemble size is:

$$m \geq C_1(c, \eta) \cdot \frac{d}{\epsilon^{2+2/\eta}}$$

where $\| \cdot \|$ denotes spectral norm, and $C_1(c, \eta)$ is an absolute constants independent of d.

The condition is mild.

- If M was a singular covariance estimate, then ρ is always no larger than the sample size.
- Exponentially many irrelevant features can still fit the bill.

Technical ingredients of proof (1): Upper bound on the spectral norm of a matrix-variate T

Let P and Q be two independent random matrices with i.i.d. standard normal entries, of size $k \times \rho$, and $k \times r$ respectively, and assume that $k < \rho - 1$. So,
$PP^T \sim \mathcal{W}(\rho, I_k)$ is a Wishart matrix independent of Q,
$\mathcal{T} := (PP^T)^{-1/2}Q \sim T_{k\times r}(0, I_k, I_r, \nu)$ has a zero mean matrix-variate T-distribution, with $\nu = \rho - k + 1$.

- $\mathcal{T}^T \sim T_{r\times k}(0, I_r, I_k, \nu)$
- $J_j \sim t_r(0, I_r, \nu) = T_{1\times r}(0, 1, I_r, \nu)$

$$\text{Pr}\left\{\lambda_{\max}\left(Q^T(PP^T)^{-1}Q\right)\cdot\frac{\rho-k-1}{k} \geq t\right\} \leq ?$$

where $\lambda_{\max}$ denotes largest eigenvalue of its argument.

Technical ingredients of proof (1): Upper bound on the spectral norm of a matrix-variate T

Lemma[K, ALT'17] [Chernoff-type bound on square norm of t-vector]
Let $x \sim T_r(0, I_r, \nu)$. Then $\forall t > r$,

$$\Pr\left\{\|x\|^2 > t\right\} \le \left(\frac{r}{t}\right)^{-\frac{r}{2}} \left(\frac{r+\nu}{t+\nu}\right)^{\frac{\nu+r}{2}}$$

- tightens with increasing ν
- recovers χ^2 Chernoff bound as $\nu \to \infty$

Using Lemma,

$$\Pr\left\{\lambda_{\max}\left(Q^T(PP^T)^{-1}Q\right) \cdot \frac{\rho-k-1}{k} \ge t\right\} \le k \cdot \left(\frac{t}{r}\right)^{\frac{r}{2}} \cdot \left(\frac{r+\nu}{t+\nu}\right)^{\frac{\nu+r}{2}}$$

Technical ingredients of proof (2): Tools from random matrix theory

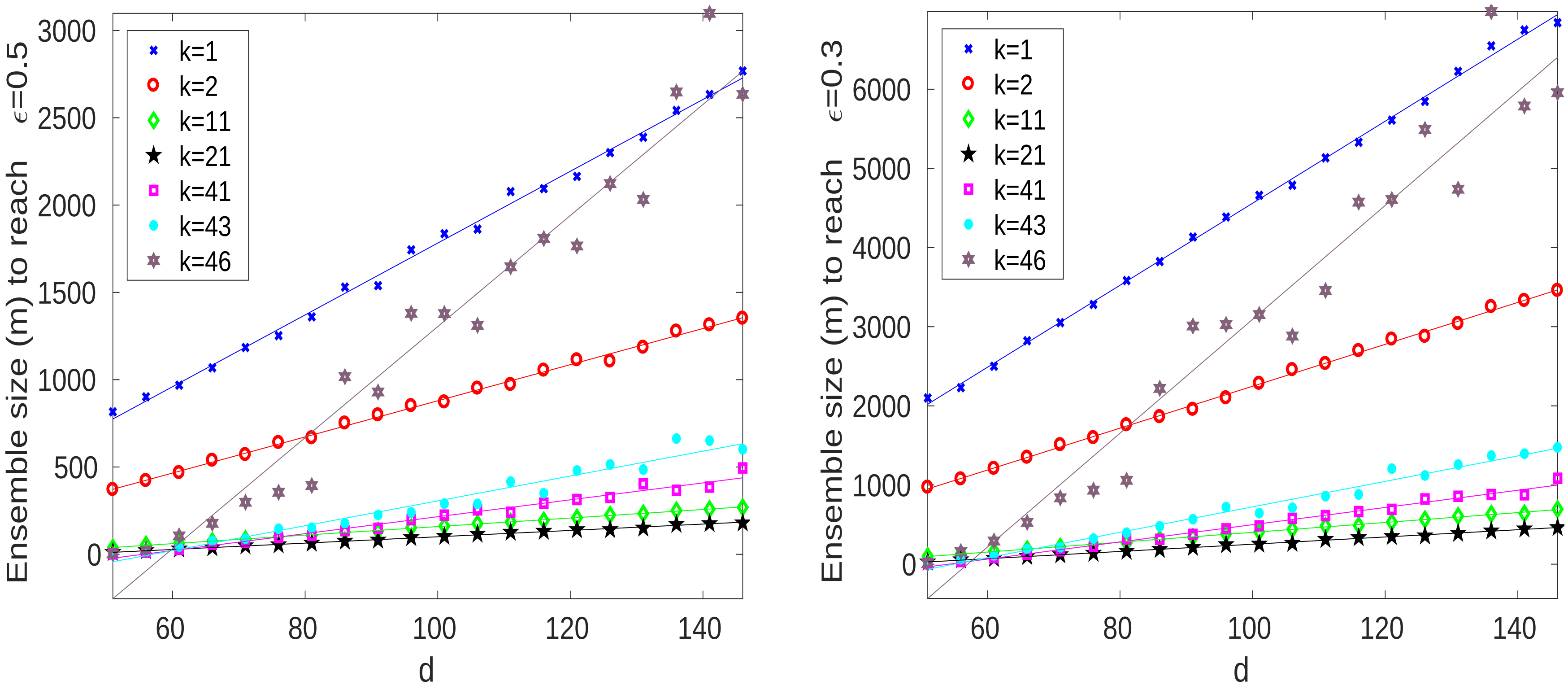
Definition (Youssef, 2013) [Matrix Strong Regularity (MSR) condition] A $d \times d$ psd random matrix U with $\mathsf{E}[U] = I_d$ satisfies MSR if $\exists \eta, c_{MSR} > 0$ constants s.t.

$$\mathsf{Pr}\{\|AUA\| \geq t\} \leq \frac{c_{MSR}}{t^{1+\eta}}, \forall t \geq c_{MSR} \cdot \mathsf{rank}(A), \forall A \text{ orthogonal projection in } \mathbb{R}^d$$

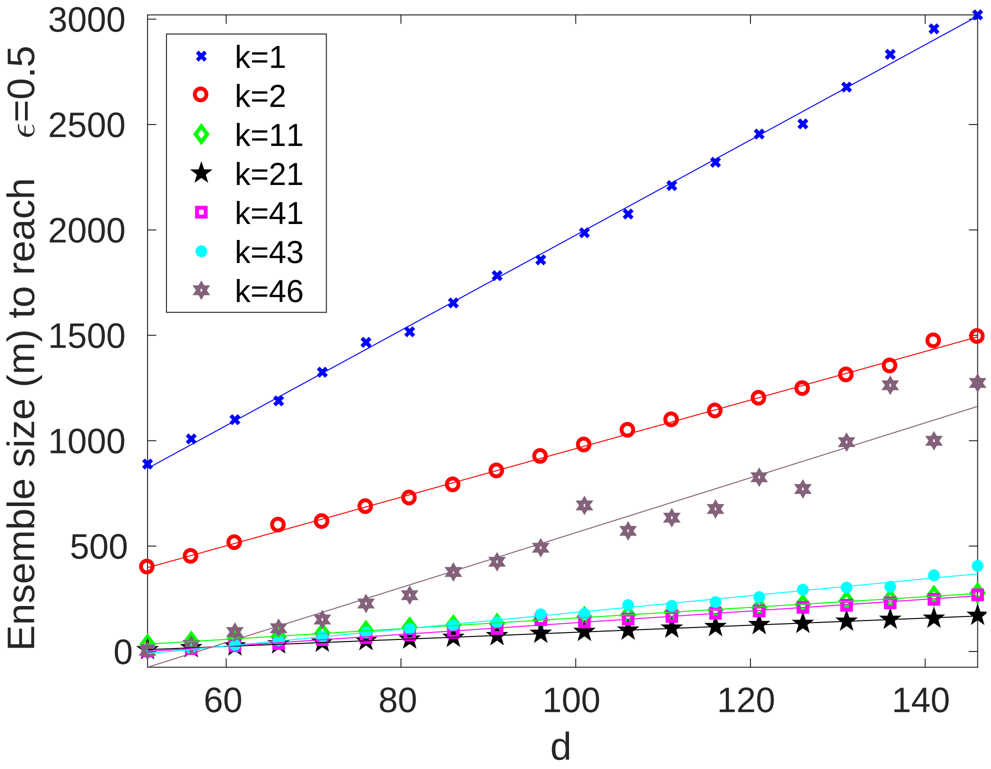
Theorem (Youssef, 2013) [Covariance of random matrices] Let U be a $d \times d$ positive semidefinite matrix having $\mathsf{E}[U] = I_d$ and satisfying the MSR for some $\eta, c_{MSR} > 0$, and let $U_1, U_2, ..., U_m$ be independent copies of U. Then, $\forall \epsilon \in (0,1)$, for $m = C_1 \cdot \frac{d}{\epsilon^{2+2/\eta}}$,

$$\mathsf{E}[\|\frac{1}{m}\sum_{i=1}^{m} U_i - I_d\|] \leq \epsilon$$

where C_1 is a constant that depends only on η and c_{MSR}.



Numerical experiment demonstrating that the required ensemble size grows linearly in d, for $\epsilon = 0.5$, and $\epsilon = 0.3$. Here $\mathcal{M} = \mathcal{M}_0$ with $\rho = 50$.



Singular matrix M of rank $\rho = 50$ with condition number $\kappa(M) = 1.5$ in its range space. Ensemble size required for ϵ convergence of the ensemble scales linearly with d.

Summing up

The ensemble of compressive learners only needs random projections of the data, and can be run in parallel.

Improves performance of ridge regularised dataspace FLD classifier.

Detailed analysis shows it implements a sophisticated regularisation scheme in the original data space.

Results on single compressive learners, as well as on ensembles, suggest that random projections may be used to uncover the structures and problem characteristics that allow effective and efficient learning for high dimensional data.

References & Acknowledgements

- A. Kabán, R.J. Durrant, Structure-aware error bounds for linear classification with the zero-one loss. arXiv:1709.09782, 2017.

- A. Kabán, The Extreme Eigenvalues of Weighted Wishart & the Generalisation of Compressed Discriminant Learning, Submitted, 2018.

- A. Kabán, On Compressive Ensemble Induced Regularization: How Close is the Finite Ensemble Precision Matrix to the Infinite Ensemble? The 28th International Conference on Algorithmic Learning Theory (ALT 2017).

- R.J. Durrant, A. Kabán. Random Projections as Regularizers: Learning a Linear Discriminant from Fewer Observations than Dimensions. Machine Learning 99(2), pp. 257-286, 2015.

This work is funded by the EPSRC Fellowship EP/P004245/1, FORGING: Fortuitous Geometries and Compressive Learning.