Are You a Bayesian or a Frequentist?

Michael I. Jordan

Department of EECS
Department of Statistics
University of California, Berkeley

http://www.cs.berkeley.edu/∼jordan

Statistical Inference

- Bayesian perspective
  - conditional perspective—inferences should be made conditional on the current data
  - natural in the setting of a long-term project with a domain expert
  - the optimist—let's make the best possible use of our sophisticated inferential tool

- Frequentist perspective
  - unconditional perspective—inferential methods should give good answers in repeated use
  - natural in the setting of writing software that will be used by many people with many data sets
  - the pessimist—let's protect ourselves against bad decisions given that our inferential procedure is inevitably based on a simplification of reality

Machine Learning (As Explained to a Statistician)

- A loose confederation of themes in statistical inference (and decision-making)
- A focus on prediction and exploratory data analysis
  - not much worry about "coverage"
- A focus on computational methodology and empirical evaluation, with a dollop of empirical process theory
  - lots of nonparametrics, but not much asymptotics
- Sometimes Bayesian and sometimes frequentist
  - not much interplay

Decision-Theoretic Perspective

- Define a family of probability models for the data X, indexed by a "parameter" θ
- Define a "procedure" $\delta(X)$ that operates on the data to produce a decision
- Define a loss function:

$$l(\delta(X), \theta)$$

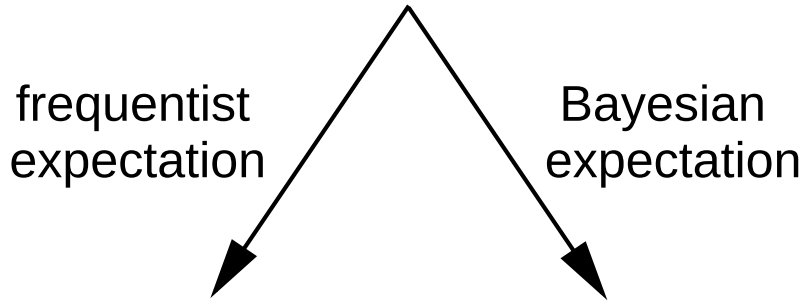
- The goal is to use the loss function to compare procedures, but both of its arguments are unknown

Decision-Theoretic Perspective

- Define a family of probability models for the data X, indexed by a "parameter" θ
- Define a "procedure" $\delta(X)$ that operates on the data to produce a decision
- Define a loss function:

$$l(\delta(X), \theta)$$

- The goal is to use the loss function to compare procedures, but both of its arguments are unknown



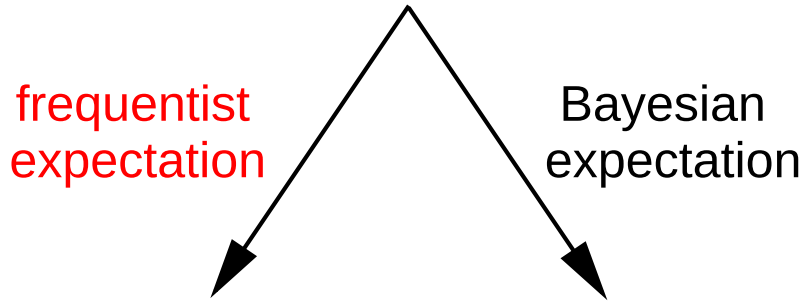
$$R(\theta) = \mathbb{E}_\theta l(\delta(X), \theta) \qquad \rho(X) = \mathbb{E}[l(\delta(X), \theta) \,|\, X]$$

Decision-Theoretic Perspective

- Define a family of probability models for the data X, indexed by a "parameter" θ
- Define a "procedure" $\delta(X)$ that operates on the data to produce a decision
- Define a loss function:

$$l(\delta(X), \theta)$$

- The goal is to use the loss function to compare procedures, but both of its arguments are unknown



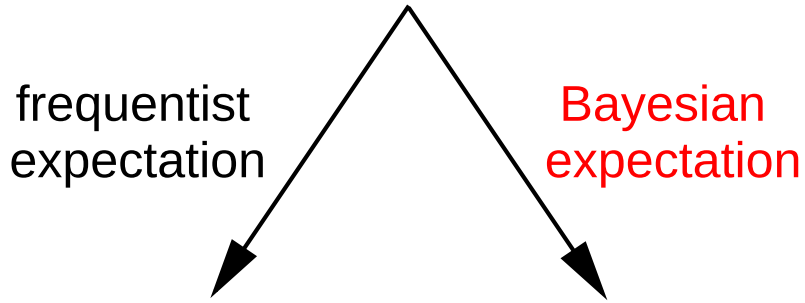
$$R(\theta) = \mathbb{E}_\theta l(\delta(X), \theta) \qquad \rho(X) = \mathbb{E}[l(\delta(X), \theta) \,|\, X]$$

Decision-Theoretic Perspective

- Define a family of probability models for the data X, indexed by a "parameter" θ
- Define a "procedure" $\delta(X)$ that operates on the data to produce a decision
- Define a loss function:

$$l(\delta(X), \theta)$$

- The goal is to use the loss function to compare procedures, but both of its arguments are unknown



$$R(\theta) = \mathbb{E}_\theta l(\delta(X), \theta) \qquad \rho(X) = \mathbb{E}[l(\delta(X), \theta) \,|\, X]$$

Coherence and Calibration

- Coherence and calibration are two important goals for statistical inference
- Bayesian work has tended to focus on coherence while frequentist work hasn't been too worried about coherence
  - the problem with pure coherence is that one can be coherent and completely wrong
- Frequentist work has tended to focus on calibration while Bayesian work hasn't been too worried about calibration
  - the problem with pure calibration is that one can be calibrated and completely useless
- Many statisticians find that they make use of both the Bayesian perspective and the frequentist perspective, because a blend is often a natural way to achieve both coherence and calibration

The Bayesian World

- The Bayesian world is further subdivided into subjective Bayes and objective Bayes
- Subjective Bayes: work hard with the domain expert to come up with the model, the prior and the loss
- Subjective Bayesian research involves (inter alia) developing new kinds of models, new kinds of computational methods for integration, new kinds of subjective assessment techniques
- Not much focus on analysis, because the spirit is that "Bayes is optimal" (given a good model, a good prior and a good loss)

Subjective Bayes

- A fairly unassailable framework in principle, but there are serious problems in practice:
  - for complex models, there can be many, many unknown parameters whose distributions must be assessed
  - independence assumptions often must be imposed to make it possible for humans to develop assessments
  - independence assumptions often must be imposed to obtain a computationally tractable model
  - it is particularly difficult to assess tail behavior, and tail behavior can matter (cf. marginal likelihoods and Bayes factors)
  - Bayesian nonparametrics can be awkward for subjective Bayes

- Also, there are lots of reasonable methods out there that don't look Bayesian; why should we not consider them?

Objective Bayes

- When the subjective Bayesian runs aground in complexity, the objective Bayesian attempts to step in
- The goal is to find principles for setting priors so as to have minimal impact on posterior inference
- E.g., reference priors maximize the divergence between the prior and the posterior
  - which often yields "improper priors"
- Objective Bayesians often make use of frequentist ideas in developing principles for choosing priors
- An appealing framework (and a great area to work in), but can be challenging to work with in complex (multivariate, hierarchical) models

Frequentist Perspective

- From the frequentist perspective, procedures can come from anywhere; they don't have to be derived from a probability model
  - e.g., nonparametric testing
  - e.g., the support vector machine, boosting
  - e.g., methods based on first-order logic
- This opens the door to some possibly silly methods, so it's important to develop principles and techniques of analysis that allow one to rule out methods, and to rank the reasonable methods
- Frequentist statistics tends to focus more on analysis than on methods
- (One general method—the bootstrap)

Frequentist Activities

- There is a hierarchy of analytic activities:
  - consistency
  - rates
  - sampling distributions
- Classical frequentist statistics focused on parametric statistics, then there was a wave of activity in nonparametric testing, and more recently there has been a wave of activity in other kinds of nonparametrics
  - e.g., function estimation
  - e.g., large p, small n problems
- One of the most powerful general tools is empirical process theory, where consistency, rates and sampling distributions are obtained uniformly on various general spaces (this is the general field that encompasses statistical learning theory)

Outline

- Surrogate loss functions, f-divergences and experimental design
- Composite loss functions and multivariate regression
- Sufficient dimension reduction
- Sparse principal component analysis

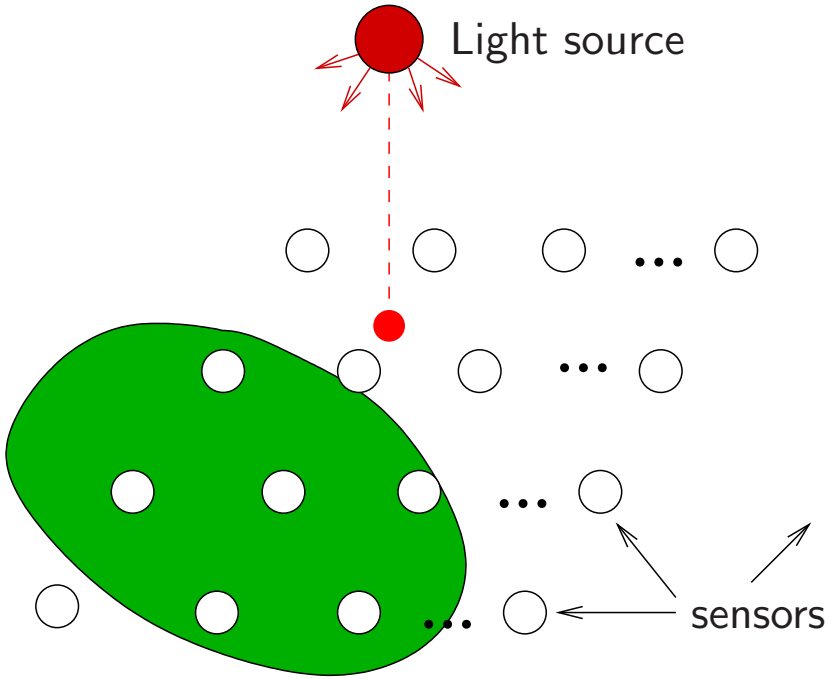
Surrogate Loss Functions, f-Divergences and Experimental Design

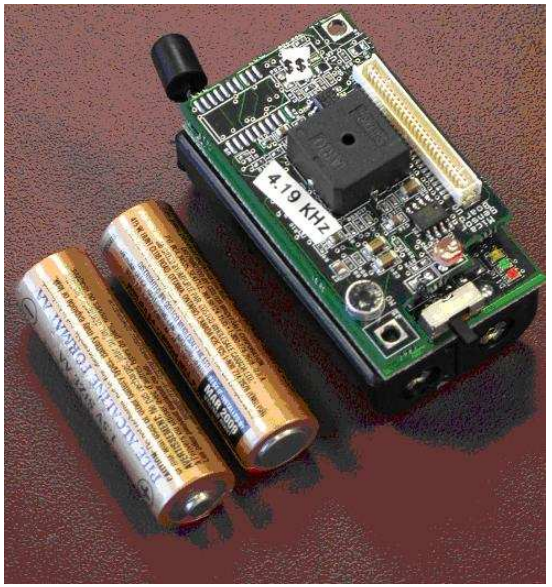
Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2008). On loss functions and f-divergences. *Annals of Statistics, 37*, 876–904.

Bartlett, P., Jordan, M. I., and McAuliffe, J. (2006). Convexity, classification and risk bounds. *Journal of the American Statistical Association, 101*, 138–156.

Nguyen, X., Jordan, M. I., and Sinopoli, B. (2005). *ACM Transactions on Sensor Networks, 1*, 134–152.

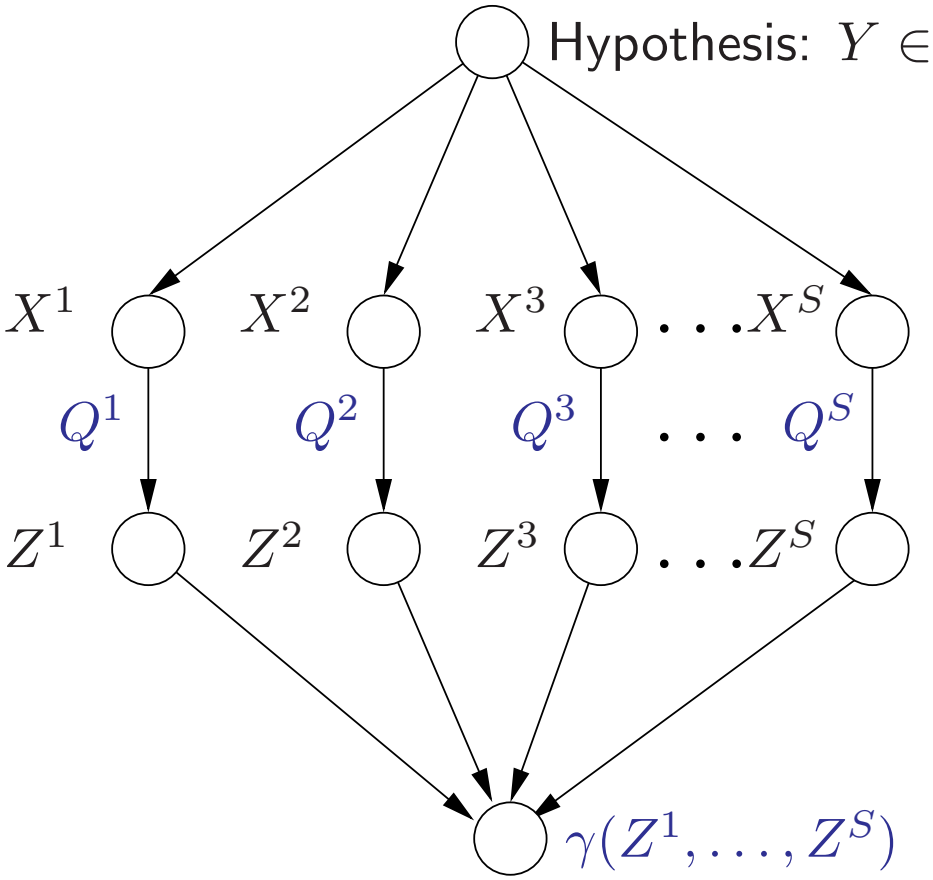
Motivating Example: Decentralized Detection





- Wireless network of motes equipped with sensors (e.g., light, heat, sound)
- Limited battery: can only transmit quantized observations
- Is the light source above the green region?

Decentralized Detection



Observations: $X \in \{1, \ldots, M\}^S$

Quantized versions: $Z \in \{1, \ldots, L\}^S$
$L \ll M$

Decentralized Detection (cont.)

- General set-up:
  - data are (X, Y) pairs, assumed sampled i.i.d. for simplicity, where $Y \in \{0, 1\}$
  - given X, let $Z = Q(X)$ denote the covariate vector, where $Q \in \mathcal{Q}$, where $\mathcal{Q}$ is some set of random mappings (can be viewed as an experimental design)
  - consider a family $\{\gamma(\cdot)\}$, where γ is a discriminant function lying in some (nonparametric) family Γ

- Problem: Find the decision $(Q; \gamma)$ that minimizes the probability of error $P(Y \neq \gamma(Z))$

- Applications include:
  - decentralized compression and detection
  - feature extraction, dimensionality reduction
  - problem of sensor placement

Perspectives

- *Signal processing literature*
  - everything is assumed known except for Q—the problem of "decentralized detection" is to find Q
  - this is done via the maximization of an "f-divergence" (e.g., Hellinger distance, Chernoff distance)
  - basically a heuristic literature from a statistical perspective (plug-in estimation)

- *Statistical machine learning literature*
  - Q is assumed known and the problem is to find γ
  - this is done via the minimization of an "surrogate loss function" (e.g., boosting, logistic regression, support vector machine)
  - decision-theoretic flavor; consistency results

f-divergences (Ali-Silvey Distances)

The f-divergence between measures μ and π is given by

$$I_f(\mu, \pi) := \sum_z \pi(z) f\left(\frac{\mu(z)}{\pi(z)}\right).$$

where $f : [0, +\infty) \to \mathbb{R} \cup \{+\infty\}$ is a continuous convex function

- Kullback-Leibler divergence: $f(u) = u \log u$.

$$I_f(\mu, \pi) = \sum_z \mu(z) \log \frac{\mu(z)}{\pi(z)}.$$

- variational distance: $f(u) = |u - 1|$.

$$I_f(\mu, \pi) := \sum_z |\mu(z) - \pi(z)|.$$

- Hellinger distance: $f(u) = \frac{1}{2}(\sqrt{u} - 1)^2$.

$$I_f(\mu, \pi) := \sum_{z \in \mathcal{Z}} (\sqrt{\mu(z)} - \sqrt{\pi(z)})^2.$$

Why the f-divergence?

- A classical theorem due to Blackwell (1951): *If a procedure A has a smaller f-divergence than a procedure B (for some fixed f), then there exist some set of prior probabilities such that procedure A has a smaller probability of error than procedure B*

- Given that it is intractable to minimize probability of error, this result has motivated (many) authors in signal processing to use f-divergences as surrogates for probability of error

- I.e., choose a quantizer Q by maximizing an f-divergence between $P(Z|Y = 1)$ and $P(Z|Y = -1)$
  - Hellinger distance (Kailath 1967; Longo et al, 1990)
  - Chernoff distance (Chamberland & Veeravalli, 2003)

- Supporting arguments from asymptotics
  - Kullback-Leibler divergence in the Neyman-Pearson setting
  - Chernoff distance in the Bayesian setting

Statistical Machine Learning Perspective

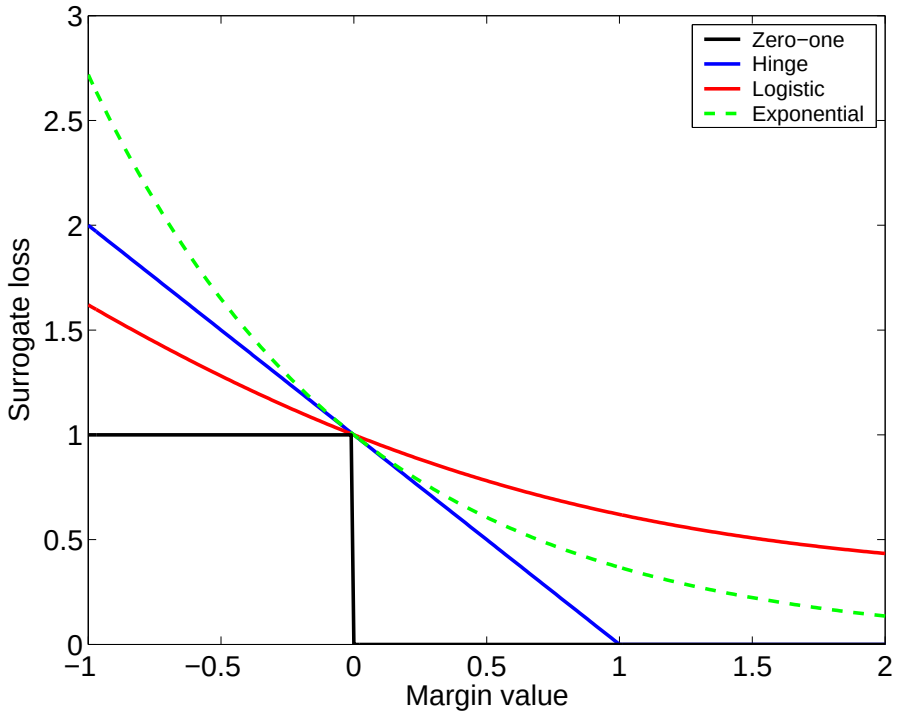
- *Decision-theoretic*: based on a loss function $\phi(Y, \gamma(Z))$
- E.g., 0-1 loss:

$$\phi(Y, \gamma(Z)) = \begin{cases} 1 & \text{if } Y \neq \gamma(Z) \\ 0 & \text{otherwise} \end{cases}$$

  which can be written in the binary case as $\phi(Y, \gamma(Z)) = \mathbb{I}(Y\gamma(Z) < 0)$

- The main focus is on estimating γ; the problem of estimating Q by minimizing the loss function is only occasionally addressed
- It is intractable to minimize 0-1 loss, so consider minimizing a surrogate loss functions that is a convex upper bound on the 0-1 loss

Margin-Based Surrogate Loss Functions



- Define a convex surrogate in terms of the margin $u = y\gamma(z)$
  - hinge loss: $\phi(u) = \max(0, 1 - u)$ — support vector machine
  - exponential loss: $\phi(u) = \exp(-u)$ — boosting
  - logistic loss: $\phi(u) = \log[1 + \exp(-u)]$ — logistic regression

Estimation Based on a Convex Surrogate Loss

- Estimation procedures used in the classification literature are generally M-estimators ("empirical risk minimization")

- Given i.i.d. training data $(x_1, y_1), \ldots, (x_n, y_n)$

- Find a classifier γ that minimizes the empirical expectation of the surrogate loss:

$$\hat{\mathbb{E}}\phi(Y\gamma(X)) := \frac{1}{n}\sum_{i=1}^{n}\phi(y_i\gamma(x_i))$$

  where the convexity of ϕ makes this feasible in practice and in theory

Some Theory for Surrogate Loss Functions

(Bartlett, Jordan, & McAuliffe, JASA 2006)

- ϕ must be classification-calibrated, i.e., for any $a, b \geq 0$ and $a \neq b$,

$$\inf_{\alpha:\alpha(a-b)<0} \phi(\alpha)a + \phi(-\alpha)b > \inf_{\alpha\in\mathbb{R}} \phi(\alpha)a + \phi(-\alpha)b$$

  (essentially a form of Fisher consistency that is appropriate for classification)

- This is necessary and sufficient for Bayes consistency; we take it as the definition of a "surrogate loss function" for classification

- In the convex case, ϕ is classification-calibrated *iff* differentiable at 0 and $\phi'(0) < 0$

Outline

- A precise link between surrogate convex losses and f-divergences
  - we establish a constructive and many-to-one correspondence
- A notion of universal equivalence among convex surrogate loss functions
- An application: Proof of consistency for the choice of a (Q, γ) pair using any convex surrogate for the 0-1 loss

Setup

- We want to find (Q, γ) to minimize the ϕ-*risk*

$$R_\phi(\gamma, Q) = \mathbb{E}\phi(Y\gamma(Z))$$

- Define:

$$\begin{aligned}
\mu(z) &= P(Y = 1, z) = p \int_x Q(z|x) dP(x|Y = 1) \\
\pi(z) &= P(Y = -1, z) = q \int_x Q(z|x) dP(x|Y = -1).
\end{aligned}$$

- ϕ-risk can be represented as:

$$R_\phi(\gamma, Q) = \sum_z \phi(\gamma(z))\mu(z) + \phi(-\gamma(z))\pi(z)$$

Profiling

- Optimize out over γ (for each z) and define:

$$R_\phi(Q) := \inf_{\gamma \in \Gamma} R_\phi(\gamma, Q)$$

- For example, for 0-1 loss, we easily obtain $\gamma(z) = \text{sign}(\mu(z) - \pi(z))$. Thus:

$$\begin{aligned} R_{\text{0-1}}(Q) &= \sum_{z \in \mathcal{Z}} \min\{\mu(z), \pi(z)\} \\ &= \frac{1}{2} - \frac{1}{2} \sum_{z \in \mathcal{Z}} |\mu(z) - \pi(z)| \\ &= \frac{1}{2}(1 - V(\mu, \pi)) \end{aligned}$$

  where $V(\mu, \pi)$ is the variational distance.

- I.e., optimizing out a ϕ-risk yields an f-divergence. Does this hold more generally?

Some Examples

- **hinge loss**:

$$R_{hinge}(Q) = 1 - V(\mu, \pi) \qquad \text{(variational distance)}$$

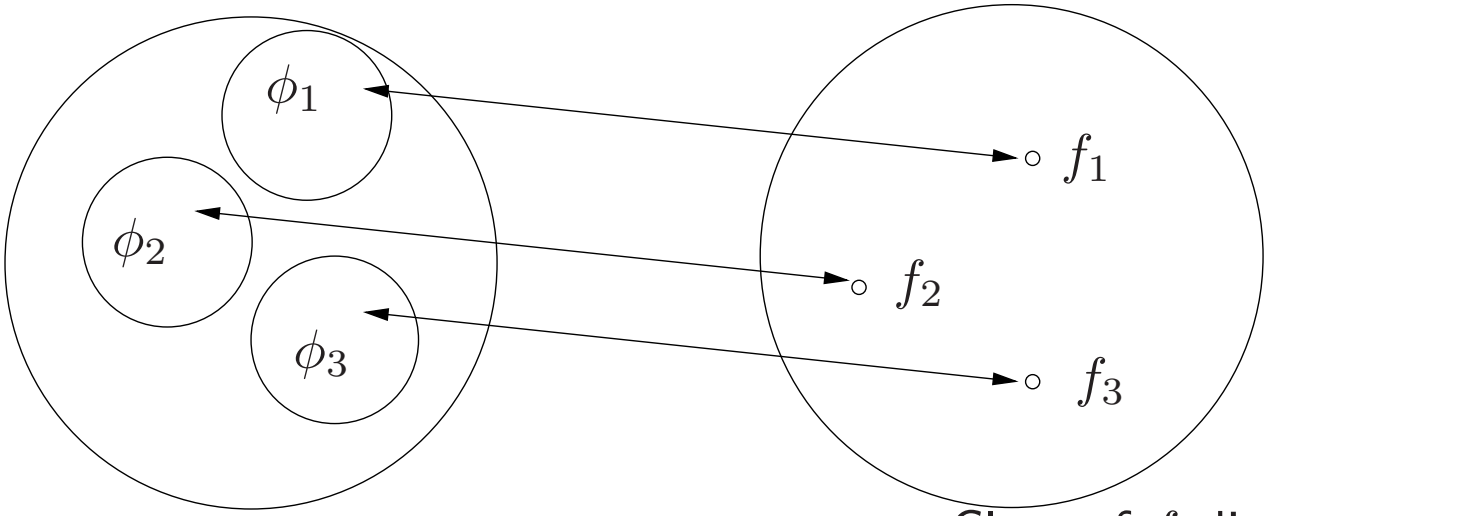
- **exponential loss**:

$$R_{exp}(Q) = 1 - \sum_{z \in \mathcal{Z}} (\sqrt{\mu(z)} - \sqrt{\pi(z)})^2 \qquad \text{(Hellinger distance)}$$

- **logistic loss**:

$$R_{log}(Q) = \log 2 - D(\mu \| \frac{\mu + \pi}{2}) - D(\pi \| \frac{\mu + \pi}{2}) \qquad \text{(capacitory discrimination)}$$

Link between ϕ-losses and f-divergences



Surrogate loss functions

Class of f-divergences

Conjugate Duality

- Recall the notion of *conjugate duality* (Rockafellar): For a lower-semicontinuous convex function $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$, the conjugate dual $f^* : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ is defined as

$$f^*(u) = \sup_{v \in \mathbb{R}} \{uv - f(v)\},$$

  which is necessarily a convex function.

- Define

$$\Psi(\beta) = f^*(-\beta)$$

Link between ϕ-losses and f-divergences

Theorem 1. *(a) For any margin-based surrogate loss function ϕ, there is an f-divergence such that $R_\phi(Q) = -I_f(\mu, \pi)$ for some lower-semicontinuous convex function f.*

In addition, if ϕ is continuous and satisfies a (weak) regularity condition, then the following properties hold:

(i) Ψ *is a decreasing and convex function.*

(ii) $\Psi(\Psi(\beta)) = \beta$ *for all* $\beta \in (\beta_1, \beta_2)$.

(iii) There exists a point u^* *such that* $\Psi(u^*) = u^*$.

(b) Conversely, if f is a lower-semicontinuous convex function satisfying conditions (i–iii), there exists a decreasing convex surrogate loss ϕ that induces the corresponding f-divergence

The Easy Direction: $\phi \to f$

- Recall

$$R_\phi(\gamma, Q) = \sum_{z \in \mathcal{Z}} \phi(\gamma(z))\mu(z) + \phi(-\gamma(z))\pi(z)$$

- Optimizing out $\gamma(z)$ for each z:

$$R_\phi(Q) = \sum_{z \in \mathcal{Z}} \inf_\alpha \phi(\alpha)\mu(z) + \phi(-\alpha)\pi(z) = \sum_z \pi(z) \inf_\alpha \left(\phi(-\alpha) + \phi(\alpha)\frac{\mu(z)}{\pi(z)} \right)$$

- For each z let $u = \frac{\mu(z)}{\pi(z)}$, define:

$$f(u) := -\inf_\alpha (\phi(-\alpha) + \phi(\alpha)u)$$

  - f is a convex function
  - we have

$$R_\phi(Q) = -I_f(\mu, \pi)$$

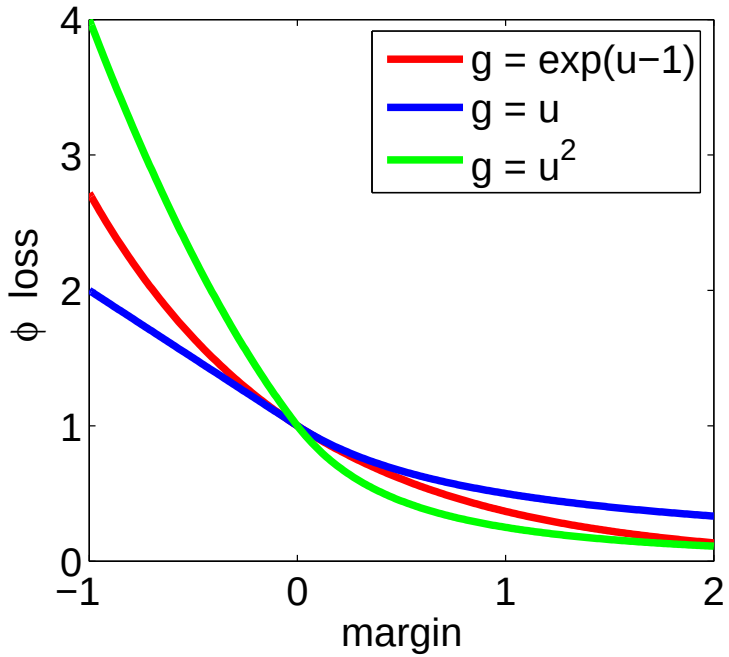
The $f \to \phi$ Direction Has a Constructive Consequence

- Any continuous loss function ϕ that induces an f-divergence must be of the form

$$\phi(\alpha) \quad = \quad \begin{cases} u^* & \text{if } \alpha = 0 \\ \Psi(g(\alpha + u^*)) & \text{if } \alpha > 0 \\ g(-\alpha + u^*) & \text{if } \alpha < 0, \end{cases}$$

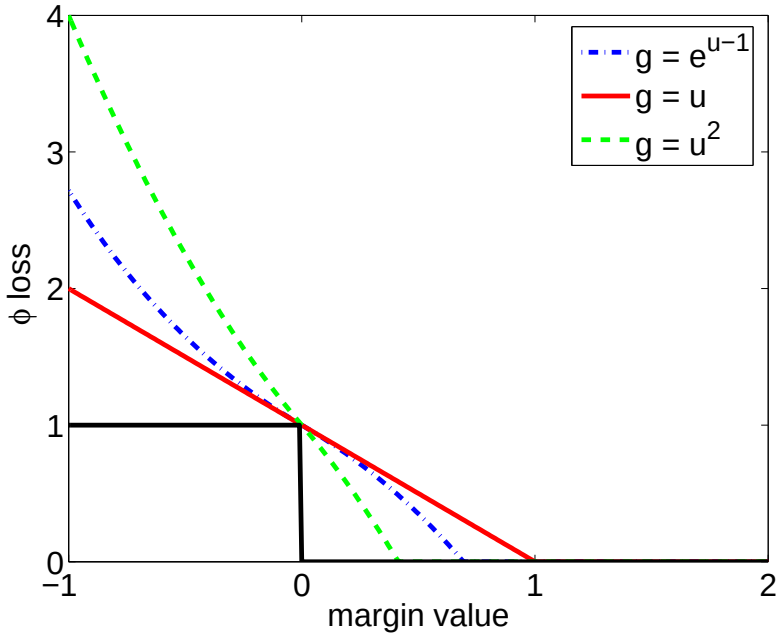
  where $g : [u^*, +\infty) \to \overline{\mathbb{R}}$ is some increasing continuous and convex function such that $g(u^*) = u^*$, and g is right-differentiable at u^* with $g'(u^*) > 0$.

Example – Hellinger distance



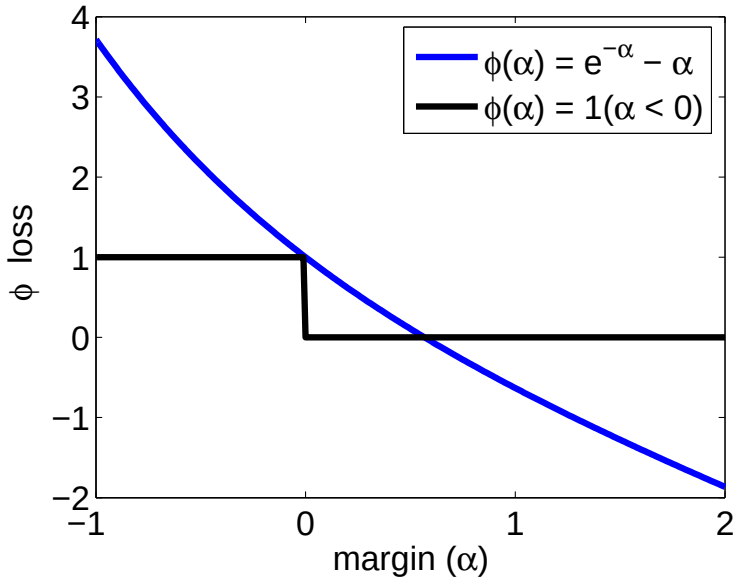
- Hellinger distance corresponds to an f-divergence with $f(u) = -2\sqrt{u}$
- Recover immediate function $\Psi(\beta) = f^*(-\beta) = \begin{cases} 1/\beta & \text{when } \beta > 0 \\ +\infty & \text{otherwise.} \end{cases}$
- Choosing $g(u) = e^{u-1}$ yields $\phi(\alpha) = \exp(-\alpha)$ $\Rightarrow$ exponential loss

Example – Variational distance



- Variational distance corresp. to an f-divergence with $f(u) = -2\min\{u, 1\}$
- Recover immediate function $\Psi(\beta) = f^*(-\beta) = \begin{cases} (2-\beta)_+ & \text{when } \beta > 0 \\ +\infty & \text{otherwise.} \end{cases}$
- Choosing $g(u) = u$ yields $\phi(\alpha) = (1-\alpha)_+$ $\Rightarrow$ hinge loss

Example – Kullback-Leibler divergence



- There is no corresponding ϕ loss for either $D(\mu\|\pi)$ or $D(\pi\|\mu)$

- But the *symmetrized* KL divergence $D(\mu\|\pi) + D(\pi\|\mu)$ is realized by

$$\phi(\alpha) = e^{-\alpha} - \alpha$$

Bayes Consistency for Choice of (Q, λ)

- Recall that from the 0-1 loss, we obtain the variational distance as the corresponding f-divergence, where $f(u) = \min\{u, 1\}$.

- Consider a broader class of f-divergences defined by:

$$f(u) = -c \min\{u, 1\} + au + b$$

- And consider the set of (continuous, convex and classification-calibrated) ϕ-losses that can be obtained (via Theorem 1) from these f-divergences

- We will provide conditions under which such ϕ-losses yield Bayes consistency for procedures that jointly choose (Q, λ)

- (And later we will show that *only* such ϕ-losses yield Bayes consistency)

Setup

- Consider sequences of increasing compact function classes $\mathcal{C}_1 \subseteq \ldots \subseteq \Gamma$ and $\mathcal{D}_1 \subseteq \ldots \subseteq \mathcal{Q}$

- Assume there exists an oracle that outputs an optimal solution to:

$$\min_{(\gamma,Q)\in(\mathcal{C}_n,\mathcal{D}_n)} \hat{R}_\phi(\gamma, Q) = \min_{(\gamma,Q)\in(\mathcal{C}_n,\mathcal{D}_n)} \frac{1}{n}\sum_{i=1}^{n}\sum_{z\in\mathcal{Z}} \phi(Y_i\gamma(z))Q(z|X_i)$$

  and let (γ_n^*, Q_n^*) denote one such solution.

- Let R_{Bayes}^* denote the minimum Bayes risk:

$$R_{Bayes}^* := \inf_{(\gamma,Q)\in(\Gamma,\mathcal{Q})} R_{Bayes}(\gamma, Q).$$

- Excess Bayes risk: $R_{Bayes}(\gamma_n^*, Q_n^*) - R_{Bayes}^*$

Setup

- *Approximation error*:

$$\mathcal{E}_0(\mathcal{C}_n, \mathcal{D}_n) = \inf_{(\gamma, Q) \in (\mathcal{C}_n, \mathcal{D}_n)} \{R_\phi(\gamma, Q)\} - R_\phi^*$$

  where $R_\phi^* := \inf_{(\gamma, Q) \in (\Gamma, \mathcal{Q})} R_\phi(\gamma, Q)$

- *Estimation error*:

$$\mathcal{E}_1(\mathcal{C}_n, \mathcal{D}_n) = \mathbb{E} \sup_{(\gamma, Q) \in (\mathcal{C}_n, \mathcal{D}_n)} \left| \hat{R}_\phi(\gamma, Q) - R_\phi(\gamma, Q) \right|$$

  where the expectation is taken with respect to the measure $\mathbb{P}^n(X, Y)$

Bayes Consistency for Choice of (Q, λ)

Theorem 2.

Under the stated conditions:

$$R_{Bayes}(\gamma_n^*, Q_n^*) - R_{Bayes}^* \quad \leq \quad \frac{2}{c}\left\{2\mathcal{E}_1(\mathcal{C}_n, \mathcal{D}_n) + \mathcal{E}_0(\mathcal{C}_n, \mathcal{D}_n) + 2M_n\sqrt{2\frac{\ln(2/\delta)}{n}}\right\}$$

- Thus, under the usual kinds of conditions that drive approximation and estimation error to zero, and under the additional condition on ϕ:

$$M_n \; := \; \max_{y\in\{-1,+1\}} \; \sup_{(\gamma,Q)\in(\mathcal{C}_n,\mathcal{D}_n)} \; \sup_{z\in\mathcal{Z}} |\phi(y\gamma(z))| \; < \; +\infty,$$

  we obtain Bayes consistency (for the class of ϕ obtained from $f(u) = -c\min\{u, 1\} + au + b$)

Universal Equivalence of Loss Functions

- Consider two loss functions ϕ_1 and ϕ_2, corresponding to f-divergences induced by f_1 and f_2

- ϕ_1 and ϕ_2 are **universally** equivalent, denoted by

$$\phi_1 \overset{u}{\approx} \phi_2$$

if for **any** $P(X,Y)$ and quantization rules Q_A, Q_B, there holds:

$$R_{\phi_1}(Q_A) \leq R_{\phi_1}(Q_B) \Leftrightarrow R_{\phi_2}(Q_A) \leq R_{\phi_2}(Q_B).$$

An Equivalence Theorem

Theorem 3.

$$\phi_1 \overset{u}{\approx} \phi_2$$

if and only if

$$f_1(u) = cf_2(u) + au + b$$

for constants $a, b \in \mathbb{R}$ *and* $c > 0$.

- $\Leftarrow$ is easy; $\Rightarrow$ is not
- In particular, surrogate losses universally equivalent to 0-1 loss are those whose induced f divergence has the form:

$$f(u) = -c \min\{u, 1\} + au + b$$

- Thus we see that *only* such losses yield Bayes consistency for procedures that jointly choose (Q, λ)

Estimation of Divergences

- Given i.i.d. $\{x_1, \ldots, x_n\} \sim \mathbb{Q}$, $\{y_1, \ldots, y_n\} \sim \mathbb{P}$
  - $\mathbb{P}, \mathbb{Q}$ are unknown multivariate distributions with densities p_0, q_0 wrt Lesbegue measure μ on $\mathbb{R}^d$

- Consider the problem of estimating a divergence; e.g., KL divergence:
  - Kullback-Leibler (KL) divergence functional

$$D_K(\mathbb{P}, \mathbb{Q}) = \int p_0 \log \frac{p_0}{q_0} \, d\mu$$

Existing Work

- Relations to entropy estimation
  - large body of work on functional of one density (Bickel & Ritov, 1988; Donoho & Liu 1991; Birgé & Massart, 1993; Laurent, 1996 and so on)
- KL is a functional of two densities
- Very little work on nonparametric divergence estimation, especially for high-dimensional data
- Little existing work on estimating density ratio per se

Main Idea

- Variational representation of f-divergences:

  Lemma 4. *Letting $\mathcal{F}$ be any function class in $\mathcal{X} \to \mathbb{R}$, there holds:*

$$D_\phi(\mathbb{P}, \mathbb{Q}) \geq \sup_{f \in \mathcal{F}} \int f \, d\mathbb{Q} - \phi^*(f) \, d\mathbb{P},$$

  with equality if $\mathcal{F} \cap \partial\phi(q_0/p_0) \neq \emptyset$.

  ϕ^* denotes the conjugate dual of ϕ

- Implications:
  - obtain an M-estimation procedure for divergence functional
  - also obtain the likelihood ratio function $d\mathbb{P}/d\mathbb{Q}$
  - how to choose $\mathcal{F}$
  - how to implement the optimization efficiently
  - convergence rate?

Kullback-Leibler Divergence

- For the Kullback-Leibler divergence:

$$D_K(\mathbb{P}, \mathbb{Q}) = \sup_{g>0} \int \log g \; d\mathbb{P} - \int g d\mathbb{Q} + 1.$$

- Furthermore, the supremum is attained at $g = p_0/q_0$.

M-Estimation Procedure

- Let $\mathcal{G}$ be a function class: $\mathcal{X} \to \mathbb{R}_+$
- $\int d\mathbb{P}_n$ and $\int d\mathbb{Q}_n$ denote the expectation under empirical measures $\mathbb{P}_n$ and $\mathbb{Q}_n$, respectively
- One possible estimator has the following form:

$$\hat{D}_K = \sup_{g \in \mathcal{G}} \int \log g \, d\mathbb{P}_n - \int g d\mathbb{Q}_n + 1.$$

- Supremum is attained at $\hat{g}_n$, which estimates the likelihood ratio p_0/q_0

Convex Empirical Risk with Penalty

- In practice, control the size of the function class $\mathcal{G}$ by using a penalty
- Let $I(g)$ be a measure of complexity for g
- Decompose $\mathcal{G}$ as follows:

$$\mathcal{G} = \cup_{1 \leq M \leq \infty} \mathcal{G}_M,$$

  where $\mathcal{G}_M$ is restricted to g for which $I(g) \leq M$.

- The estimation procedure involves solving:

$$\hat{g}_n = \text{argmin}_{g \in \mathcal{G}} \int g d\mathbb{Q}_n - \int \log g \, d\mathbb{P}_n + \frac{\lambda_n}{2} I^2(g).$$

Convergence Rates

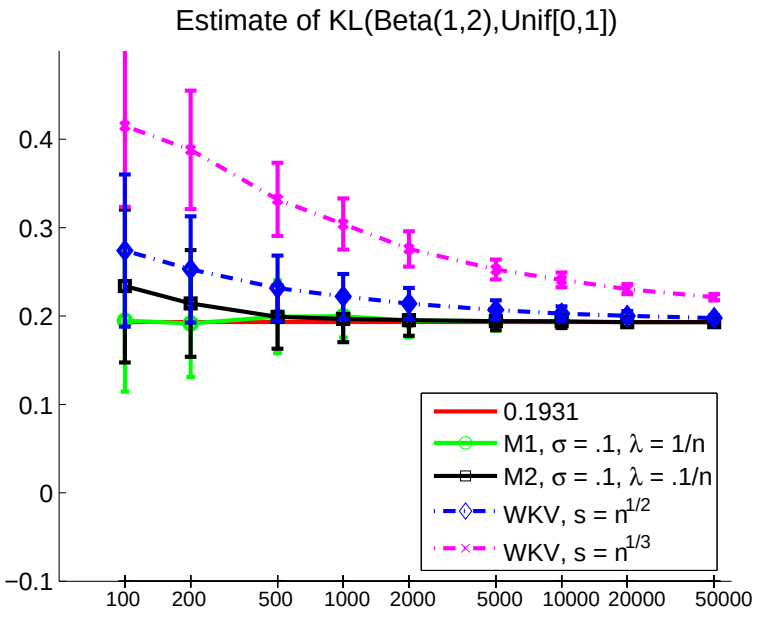
Theorem 5. *When λ_n vanishes sufficiently slowly:*

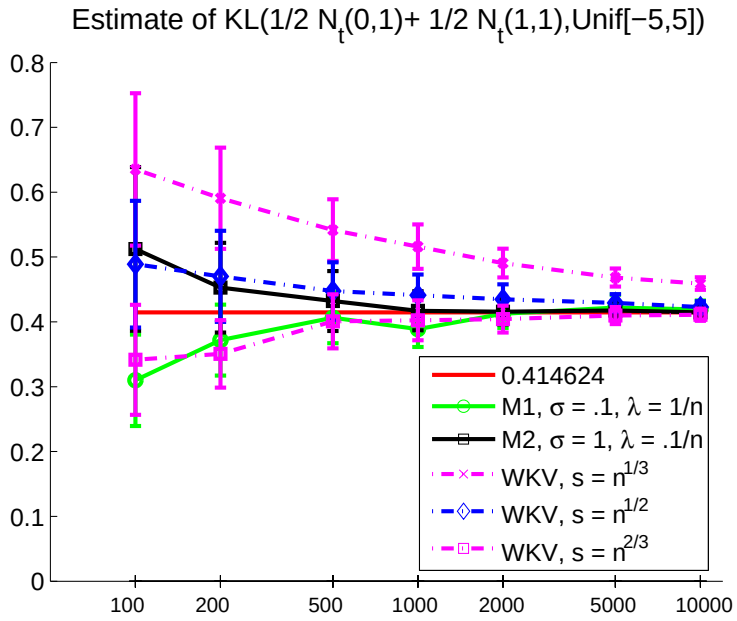
$$\lambda_n^{-1} = O_P(n^{2/(2+\gamma)})(1 + I(g_0)),$$

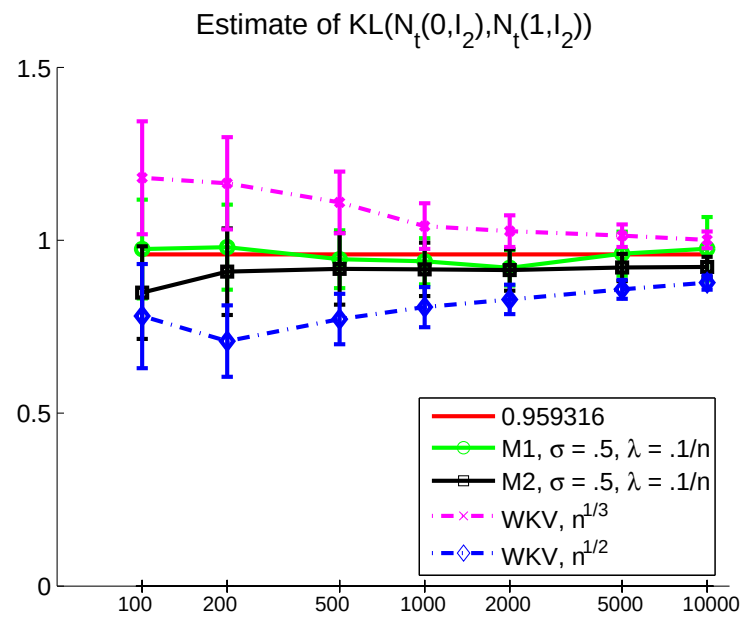
then under $\mathbb{P}$:

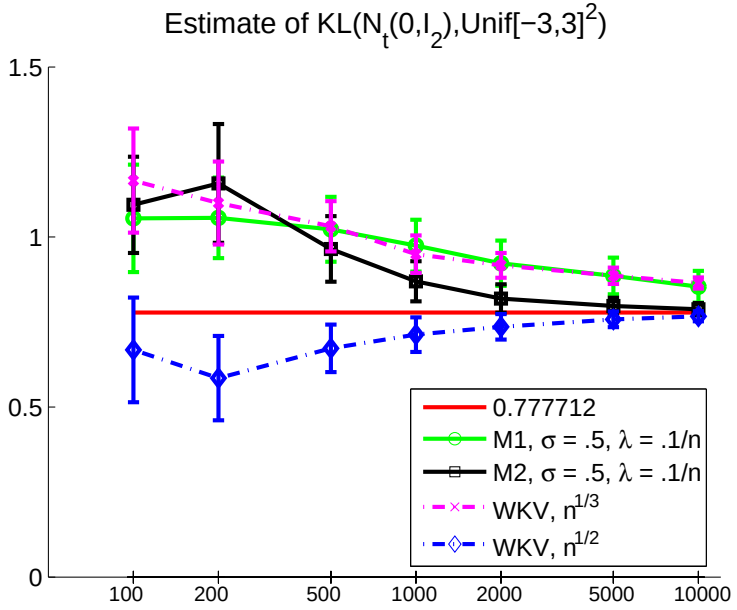
$$h_{\mathbb{Q}}(g_0, \hat{g}_n) = O_P(\lambda_n^{1/2})(1 + I(g_0))$$
$$I(\hat{g}_n) = O_P(1 + I(g_0)).$$

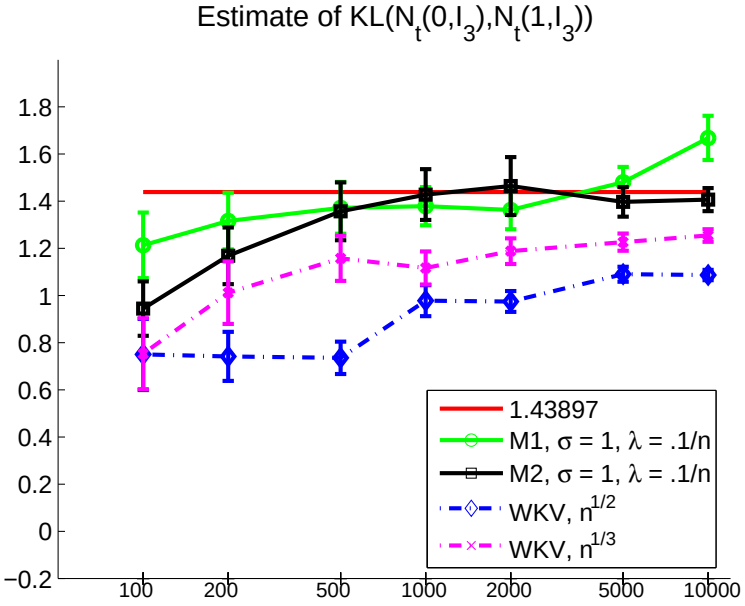
Results

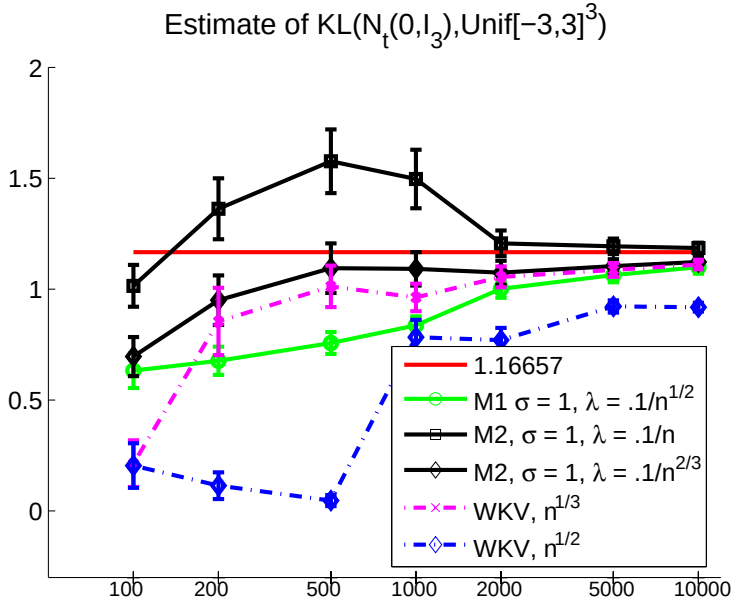

Estimate of KL(Beta(1,2),Unif[0,1])
0.4
0.3
0.2
0.1
0
−0.1
100
200
500
1000
2000
5000
10000
20000
50000
0.1931
M1, σ = .1, λ = 1/n
M2, σ = .1, λ = .1/n
WKV, s = n^1/2
WKV, s = n^1/3


Estimate of KL(1/2 N_t(0,1)+ 1/2 N_t(1,1),Unif[−5,5])
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
100
200
500
1000
2000
5000
10000
0.414624
M1, σ = .1, λ = 1/n
M2, σ = 1, λ = .1/n
WKV, s = n^1/3
WKV, s = n^1/2
WKV, s = n^2/3


Estimate of KL(N_t(0,I_2),N_t(1,I_2))
1.5
1
0.5
0
100
200
500
1000
2000
5000
10000
0.959316
M1, σ = .5, λ = .1/n
M2, σ = .5, λ = .1/n
WKV, n^1/3
WKV, n^1/2


Estimate of KL(N_t(0,I_2),Unif[−3,3]^2)
1.5
1
0.5
0
100
200
500
1000
2000
5000
10000
0.777712
M1, σ = .5, λ = .1/n
M2, σ = .5, λ = .1/n
WKV, n^1/3
WKV, n^1/2


Estimate of KL(N_t(0,I_3),N_t(1,I_3))
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
100
200
500
1000
2000
5000
10000
1.43897
M1, σ = 1, λ = .1/n
M2, σ = 1, λ = .1/n
WKV, n^1/2
WKV, n^1/3


Estimate of KL(N_t(0,I_3),Unif[−3,3]^3)
2
1.5
1
0.5
0
100
200
500
1000
2000
5000
10000
1.16657
M1 σ = 1, λ = .1/n^1/2
M2, σ = 1, λ = .1/n
M2, σ = 1, λ = .1/n^2/3
WKV, n^1/3
WKV, n^1/2

Conclusions

- Formulated a precise link between f-divergences and surrogate loss functions
- Decision-theoretic perspective on f-divergences
- Equivalent classes of loss functions
- Can design new convex surrogate loss functions that are equivalent (in a deep sense) to 0-1 loss
  - Applications to the Bayes consistency of procedures that jointly choose an experimental design and a classifier
  - Applications to the estimation of divergences and entropy

Composite Loss Functions and Multivariate Regression; Sparse PCA

G. Obozinski, B. Taskar, and M. I. Jordan (2009). Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, to appear.

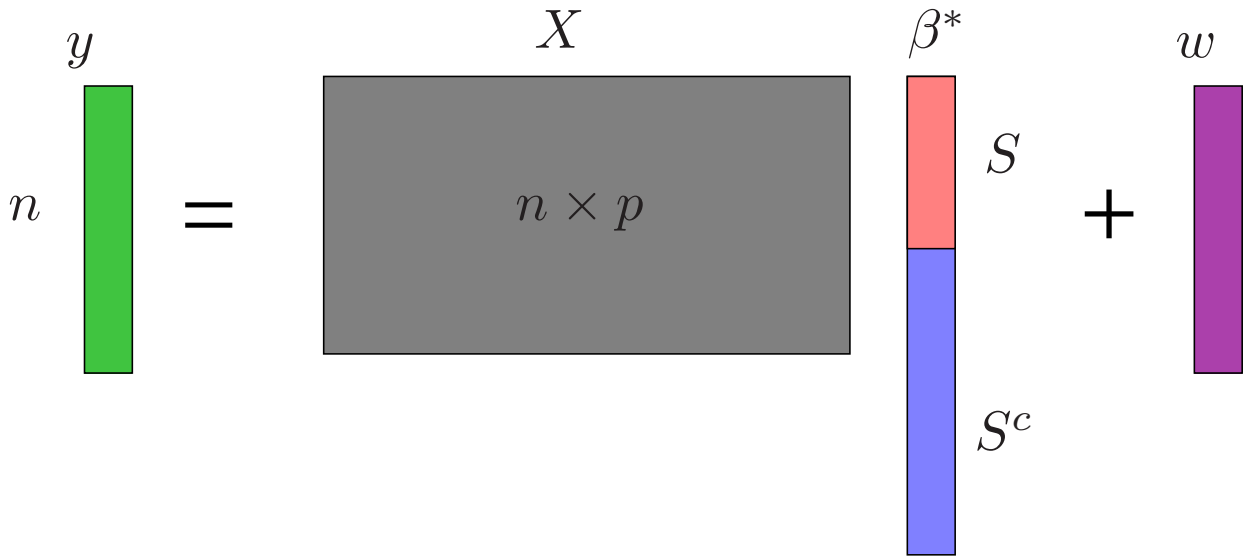
G. Obozinski, M. J. Wainwright, and M. I. Jordan (2009). Union support recovery in multivariate regression. *Annals of Statistics*, under review.

A. Amini and M. J. Wainwright (2009). High-dimensional analysis of semidefinite relaxations for sparse PCA. *Annals of Statistics*, to appear.

Introduction

- classical asymptotic theory of statistical inference:
  - number of observations $n \to +\infty$
  - model dimension p stays fixed
- not suitable for many modern applications:
  - $\{$ images, signals, systems, networks $\}$ frequently large ($p \approx 10^3 - 10^8$)...
  - interesting consequences: might have $p = \Theta(n)$ or even $p \gg n$
- curse of dimensionality: frequently impossible to obtain consistent procedures unless $p/n \to 0$
- can be saved by a lower *effective dimensionality*, due to some form of complexity constraint

Example: Sparse linear regression

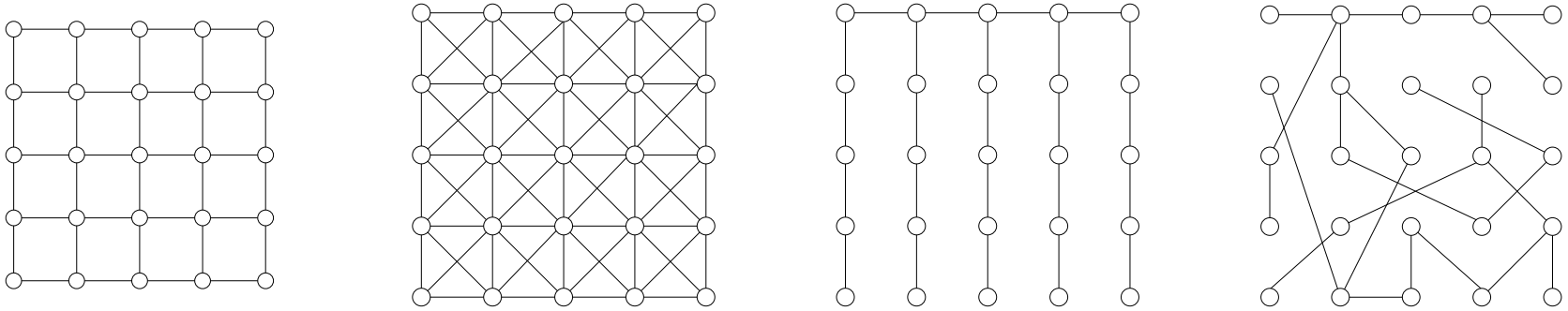


- vector $\beta^* \in \mathbb{R}^p$ with at most $k \ll p$ non-zero entries

- observation model: noisy linear observations $y = X\beta^* + w$
  $X \in \mathbb{R}^{n \times p}$: design matrix
  $w \in \mathbb{R}^{n \times 1}$: noise vector

- various applications (database sketching, imaging, genetic testing...)

Example: Graphical model selection



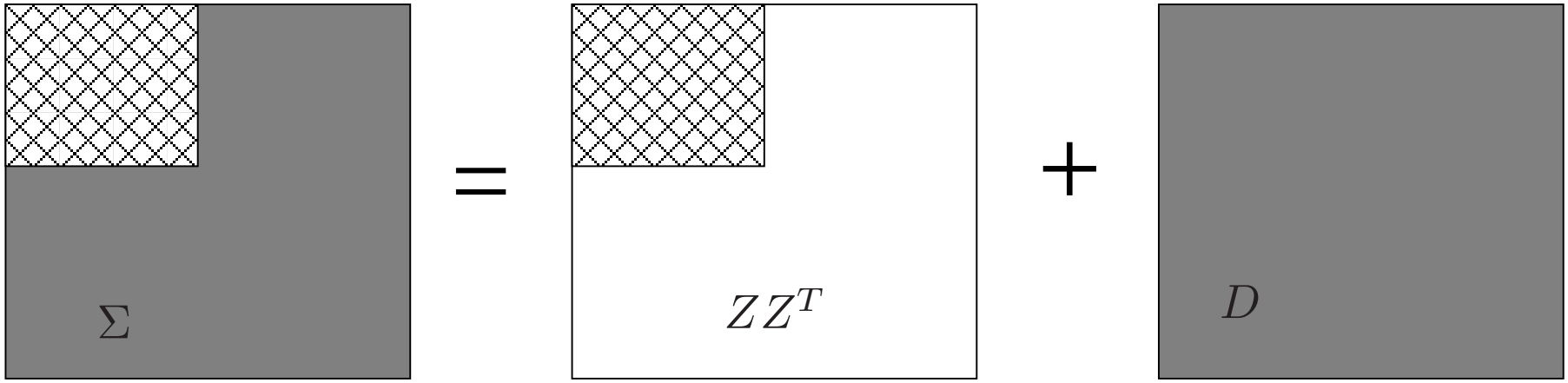
- consider m-dimensional random vector $Z = (Z_1, \ldots, Z_m)$:

$$\mathbb{P}(Z_1, \ldots, Z_m; \beta) \quad \propto \quad \exp\left\{ \sum_{(i,j) \in E} \beta_{ij} Z_i Z_j \right\}.$$

- given n independent and identically distributed (i.i.d.) samples of $\vec{Z}$, identify underlying graph $G = (V, E)$

- lower effective dimensionality: graphs with $k \ll p := \binom{m}{2}$ edges

Example: Sparse principal components analysis



Set-up: Covariance matrix $\Sigma = ZZ^T + D$, where leading eigenspace Z has sparse columns.

Goal: Produce an estimate $\widehat{Z}$ based on samples $X^{(i)}$ with covariance matrix Σ.

Some issues in high-dimensional inference

- Consider some fixed loss function, and a fixed level δ of error.

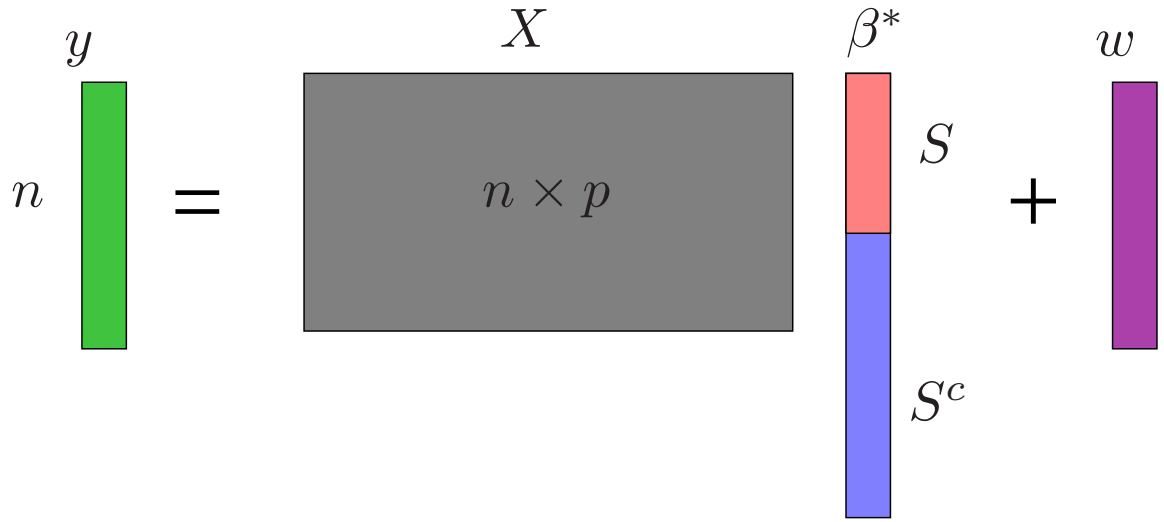
- Given particular (polynomial-time) algorithms
  - for what sample sizes n do they succeed/fail to achieve error δ?
  - when does more computation reduce minimum # samples needed?

Outline

1. Multivariate regression in high dimensions
   (a) Practical limitations: scaling laws for second-order cone programs
   (b) SOCP vs. Lasso: when does more computation reduce statistical error?

2. Sparse principal component analysis in high dimensions
   (a) Thresholding methods
   (b) Semidefinite programming

Optimization-based estimators in (sparse) regression



Regularized QP: $\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \{ \underbrace{\frac{1}{2n}\|y - X\beta\|_2^2}_{\text{Data term}} + \rho_n \underbrace{R(\beta)}_{\text{Regularizer}} \}.$

$R(\beta) = \|\beta\|_2$	Ridge regression	(Tik43, HoeKen70)
$R(\beta) = \|\beta\|_1$	convex ℓ_1-constrained QP	(CheDonSau96; Tibs96)
$R(\beta) = \|\beta\|_0$	Subset selection: combinatorial, NP-hard	(Nat95)
$R(\beta) = \|\beta\|_a$, $a \in (0,1)$	Non-convex ℓ_a regularization	

Different loss functions

Given an estimate $\widehat{\beta}$, how to assess its performance?

1. Predictive loss: compute expected error $\mathbb{E}[\|\widetilde{y} - X\widehat{\beta}\|_2^2]$
   - goal is to construct model with good predictive power
   - β^* itself of secondary interest (need not be uniquely determined)

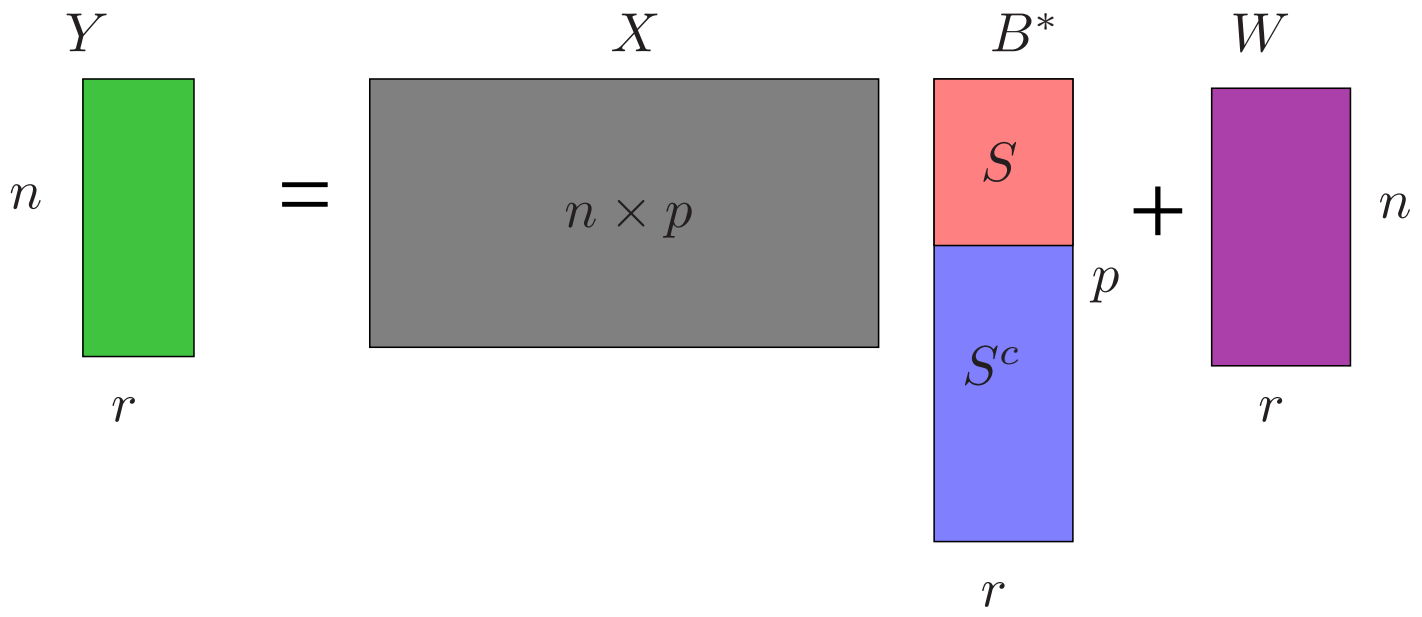
2. ℓ_2-loss $\mathbb{E}[\|\widehat{\beta} - \beta^*\|_2^2]$
   - appropriate when B^* is of primary interest (signal recovery, compressed sensing, denoising etc.)

3. Support recovery criterion: define estimated support

$$S(\widehat{\beta}) \quad = \quad \{i = 1, \ldots, p \mid \widehat{\beta}_i \neq 0\},$$

   and measure probability $\mathbb{P}[S(\widehat{\beta}) \neq S(\beta^*)]$.
   - useful for feature selection, dimensionality reduction, model selection
   - can be used as a pre-processing step for estimation in ℓ_2-norm

§1. Multivariate regression in high dimensions



- signal B^* is a $p \times r$ matrix: partitioned into **non-zero rows** S and **zero rows** S^c
- observe n noisy projections, defined via **design matrix** $X \in \mathbb{R}^{n \times p}$ and **noise matrix** $W \in \mathbb{R}^{n \times r}$
- matrix $\mathbf{Y} \in \mathbb{R}^{\mathbf{n} \times \mathbf{r}}$ **of observations**
- high-dimensional scaling: allow parameters $(n, p, r, |S|)$ to scale

Block regularization and second-order cone programs

(Obozinski, Taskar & Jordan, 2009)

- for fixed parameter $q \in [1, \infty]$, estimate B^* via:

$$\widehat{B} \quad \in \quad \arg \min_{B \in \mathbb{R}^{p \times r}} \left\{ \underbrace{\frac{1}{2n} \|Y - XB\|_F^2}_{\text{Data term } \sum_{j=1}^{n} \sum_{\ell=1}^{r} \left[Y_{j\ell} - (XB)_{j\ell} \right]^2} + \rho_n \underbrace{|||B|||_{1,q}}_{\sum_{i=1}^{p} \|(B_{i1}, \ldots, B_{ir})\|_q} \right\}.$$

- regularization constant $\rho_n > 0$ to be chosen by user
- different cases:
  - $q = 1$: elementwise ℓ_1 norm (constrained QP)
  - $q = 2$: second-order cone program (SOCP)
  - $q = \infty$: block ℓ_1/ℓ_∞ max-norm (constrained QP)

- in all cases, efficiently solvable (e.g., by interior point methods)
- generalization of the Lasso (Tibshirani, 1996; Chen et al., 1998),
- special case of the CAP family (Zhao, Rocha, & Yu, 2006); see also (Turlach et al., 2005; Yuan & Lin, 2006, Nardi & Rinaldo, 2008)

Two strategies

Goal: Model selection consistency: recover union of supports

$$S(B^*) \quad := \quad \{i \in \{1, 2, \ldots, p\} \mid \|B^*_{i1}, \ldots, B^*_{ir}\|_2 \neq 0\}.$$

Different methods:

- *Lasso-based recovery:*
  1. Solve a separate Lasso (ℓ_1-constrained QP) for each column $\ell = 1, \ldots, r$, yielding column vector $\widehat{\beta}_\ell \in \mathbb{R}^p$.
  2. Estimate row support $\widehat{S}_{\text{Lasso}} = \{i \in \{1, 2, \ldots, p\} \mid \widehat{\beta}_{i\ell} \neq 0 \text{ for some } \ell\}$.
- *SOCP-based recovery:*
  1. Solve a single SOCP, obtaining matrix estimate $\widehat{B} \in p \times r$.
  2. Estimate support $\widehat{S}_{\text{SOCP}} = \{i \in \{1, \ldots, p\} \mid \|(\widehat{B}_{i1}, \ldots, \widehat{B}_{ir}\|_2 \neq 0\}$.

Trade-offs:

- Lasso (QP) cheap to solve, but method ignores coupling among columns
- SOCP more expensive, but block-regularizer better tailored to matrix structure

Scaling law for high-dimensional SOCP recovery

(Obozinski, Wainwright & Jordan, 2009)

- SOCP method: $\widehat{B} \in \arg\min_{B \in \mathbb{R}^{p \times r}} \left\{ \frac{1}{2n} \|Y - XB\|_F^2 + \rho_n |||B|||_{1,2} \right\}$.
- Parameters: Problem dimension p; number of non-zero rows k
- Design matrix X: i.i.d. rows from sub-Gaussian distribution, with "suitable" covariance Σ

Theorem: If the *rescaled sample size*

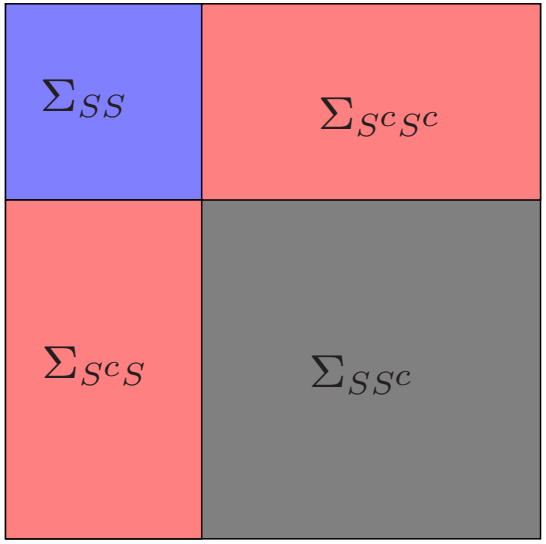
$$\theta_{\mathrm{SOCP}}(n, p, k, \ B^*) \ := \ \frac{n}{\Psi(B_S^*; \Sigma_{SS}) \ \log(p-k)}$$

is greater than a critical threshold $\theta_\ell(\Sigma; \sigma^2)$, then for suitable ρ_n we have with probability greater than $1 - 2\exp(c_2 \log k)$:

(a) the SOCP has a unique solution $\widehat{B}$ s.t. $\widehat{S}(\widehat{B}) \subseteq S(B^*)$, and

(b) It includes all rows i with $\|B_i^*\|_2 \geq c_3 \sqrt{\frac{\max\{k, \log(p-k)\}}{n}}$.

Assumptions on design covariance



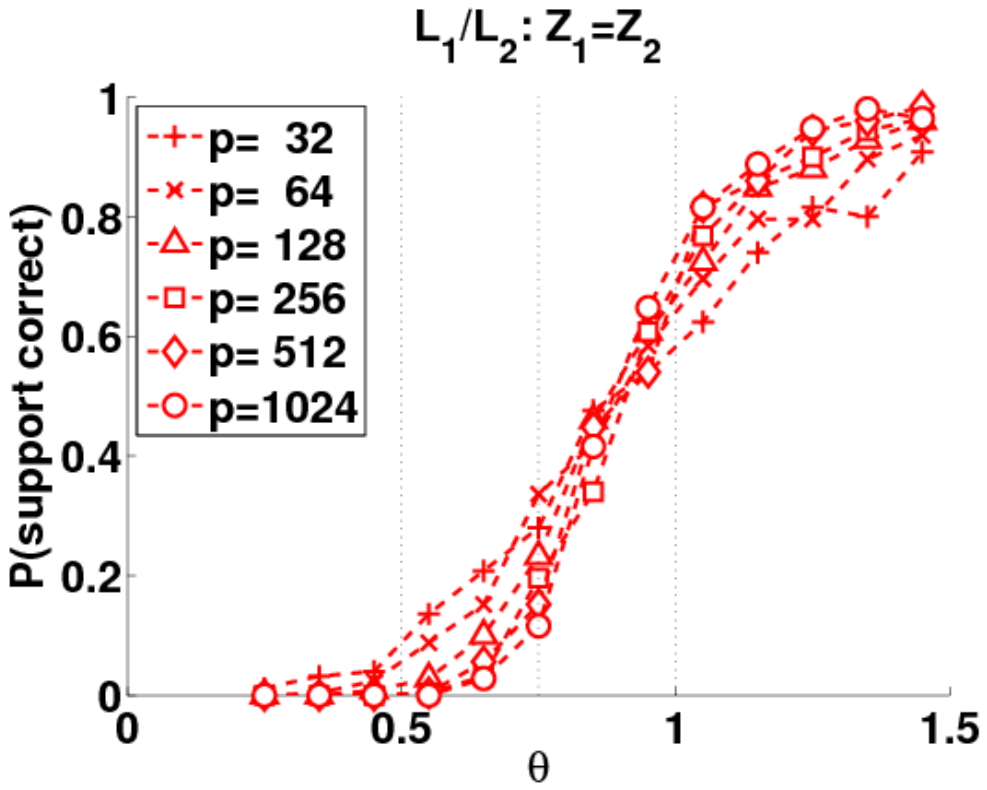
- support set $S = \{i \mid \beta_i^* \neq 0\}$
- complement $S^c := \{1, \ldots p\} \backslash S$.
- random design matrix $X \in \mathbb{R}^{n \times p}$
- rows drawn i.i.d., cov. Σ, sub-Gaussian

1. **Bounded eigenspectrum:** $\lambda(\Sigma_{SS}) \in [C_{min}, C_{max}]$.
2. **Mutual incoherence/irrepresentability:** There exists an $\nu \in (0, 1]$ such that

$$|||\Sigma_{S^cS}(\Sigma_{SS})^{-1}|||_{\infty,\infty} \leq \quad 1 - \nu.$$

Example: if $\Sigma_{SS} = I$, then require $\max_{j \in S^c} \sum_{i \in S} |\Sigma_{ji}| \leq 1 - \nu$.

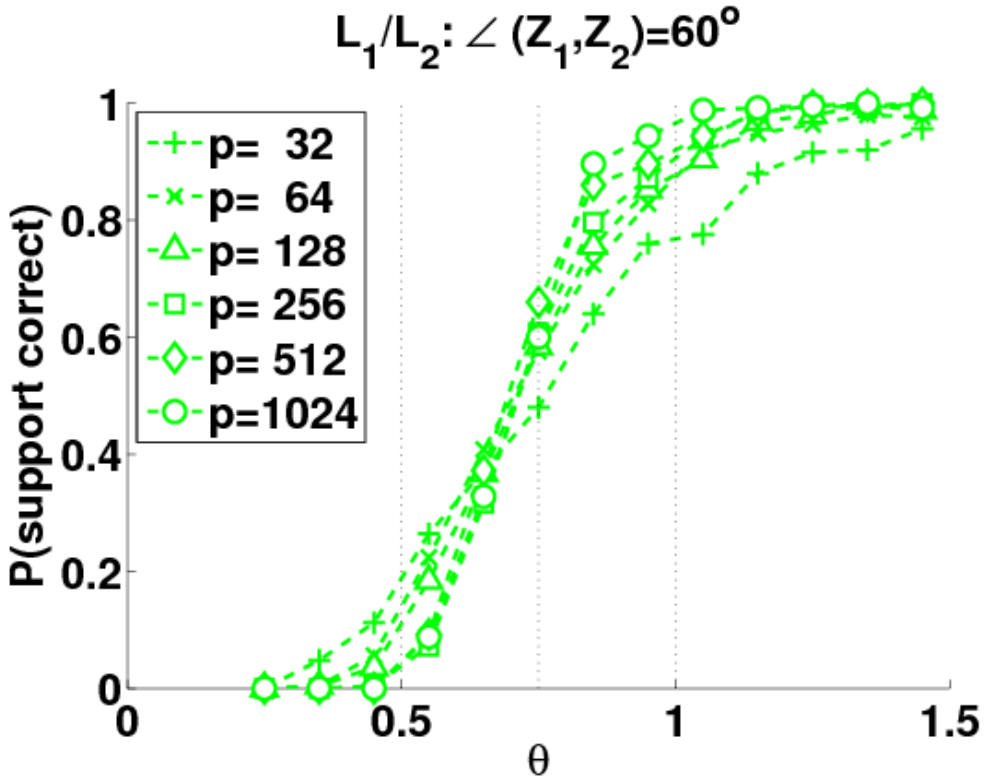
Order parameter captures threshold (Angle 0°)



Prob. success versus rescaled sample size

$$\theta_{\mathrm{SOCP}}(n, p, k, B^*) = \frac{n}{\Psi(B_S^*; \Sigma_{SS}) \log(p-k)}.$$

Order parameter captures threshold (Angle 60°)



Prob. success versus rescaled sample size

$$\theta_{\mathrm{SOCP}}(n, p, k, B^*) = \frac{n}{\Psi(B_S^*; \Sigma_{SS}) \log(p-k)}.$$

Sparsity overlap function Ψ

- form gradient matrix $Z(B_S^*) := \nabla |||B_S|||_{1,2}\Big|_{B_S = B_S^*} \in \mathbb{R}^{k \times r}$
- equivalent to renormalizing B_S^* to have unit ℓ_2-norm rows
- form $r \times r$ Gram matrix:

$$G \quad = \quad Z^T (\Sigma_{SS})^{-1} Z$$

  with $G_{a,b} = \langle\langle Z_a, \ Z_b \rangle\rangle_{(\Sigma_{SS})^{-1}}$
- sparsity overlap function is max. eigenvalue of G:

$$\Psi(B_S^*; \Sigma_{SS}) \quad = \quad |||G|||_2.$$

- measures relative alignments of the renormalized columns of B^*
- **Special case:** Univariate regression ($r = 1$): $Z(\beta_S^*) = k$ for any vector β_S^*

Concrete examples ($k = 4, r = 2$)

Aligned columns

$$B_S^* = \begin{bmatrix} 2 & 2 \\ 10 & 10 \\ 1 & 1 \\ 7 & 7 \end{bmatrix} \qquad Z(B_S^*) = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

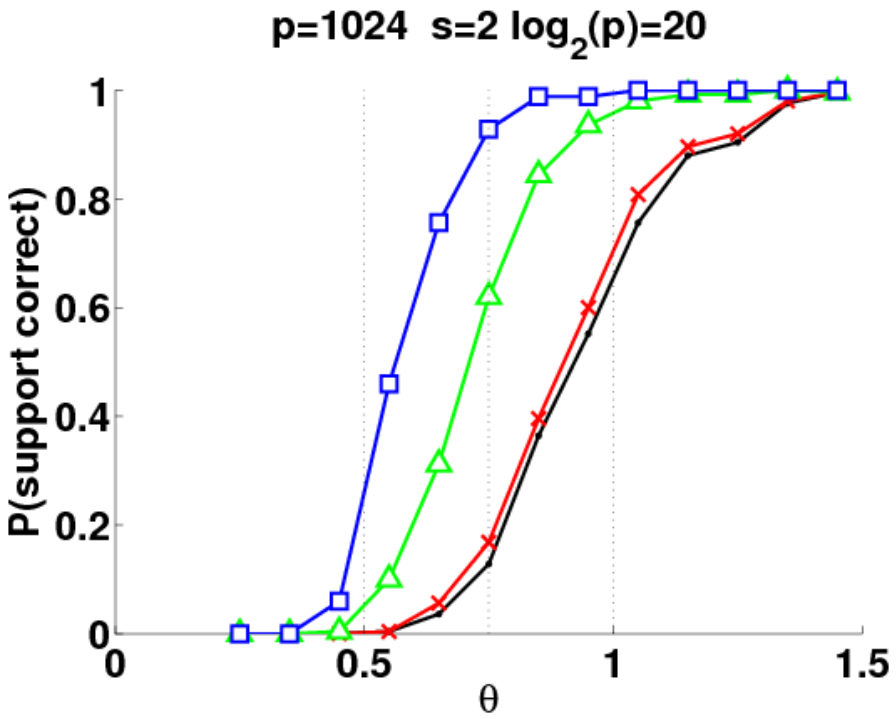
$$G = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \qquad \|\!|G|\!\|_2 = 4$$

Orthogonal columns

$$B_S^* = \begin{bmatrix} 2 & 2 \\ 10 & 10 \\ 1 & -1 \\ 7 & -7 \end{bmatrix} \qquad Z(B_S^*) = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$$

$$G = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \qquad \|\!|G|\!\|_2 = 2$$

Empirical illustration of sparsity-overlap Ψ



- Orthogonal regression: Columns $Z_1 \perp Z_2$
- Intermediate angle: Columns at 60°
- Aligned regression: Columns parallel
- Ordinary Lasso: solve problems separately.

SOCP versus ordinary QP

Corollary: If $\Sigma_{SS} = I_{k \times k}$, SOCP always dominates ordinary QP, with relative statistical efficiency:

$$1 \leq \frac{\max_{\ell=1,\ldots r} k_\ell \log(p - k_\ell)}{\Psi(B_S^*; I) \log(p-k)} \leq r$$

(QP sample size)/(SOCP sample size)

- increased statistical efficiency of SOCP: dependent on orthogonality properties of rescaled columns B_S^*
- up to a factor $1/r$ reduction in number of samples required
- most pessimistic case: no gain for disjoint supports, SOCP can be worse in some cases (if $\Sigma_{SS} \neq I$)

Proof sketch of sufficient conditions

Direct analysis :

Given n observations of $\beta^* \in \mathbb{R}^p$ with $|S(\beta^*)| = k$, oracle decoder performs following two steps:

1. For each subset S of size k, solve the quadratic program:

$$f(S) \quad = \quad \min_{\beta_S \in \mathbb{R}^k} \|Y - X_S \beta_S\|_2^2.$$

2. Output the subset $\widehat{S} = \arg\min_{|S|=k} f(S)$.

- by symmetry of ensemble, may assume that fixed subset S is chosen
- for sets U different from true set S, consider range of *non-overlaps* $t := |U \backslash S| \in \{1, \ldots, k\}$
- number of subsets with non-overlap t given by $N(t) = \binom{k}{t-k} \binom{p-k}{t}$

Error exponents for random projections

- union bound yields upper bound on error probability $\mathbb{P}[\text{error} \mid S \text{ true}]$:

$$\sum_{t=1}^{k} \binom{k}{k-t} \binom{p-k}{t} \mathbb{P}[\text{error on subset with non-overlap } t]$$

- orthogonal projection $\Pi_U^{\perp} := I_{n \times n} - X_U \left[X_U^T X_U\right]^{-1} X_U^T$
- optimal decoder chooses U incorrectly over S if and only if

$$\Delta(U) \quad = \quad \underbrace{\left\| \Pi_U^{\perp} \left(X_{S \setminus U} \beta^*_{S \setminus U} + W \right) \right\|^2}_{\text{effective noise in } U^{\perp}} - \underbrace{\left\| \Pi_S^{\perp} W \right\|^2}_{\text{effective noise in } S^{\perp}} \; < \; 0$$

- use large deviations to establish that

$$\mathbb{P}[\Delta(U) < 0] \quad \leq \quad \exp\left(-n \; F(\|\beta^*_{S \setminus U}\|^2; t) \right).$$

Proof sketch of necessary conditions

- Fano's inequality applied to a restricted ensemble, assuming *fixed choice* of β^*:

$$\beta_i^*[U] \quad = \quad \begin{cases} \beta_{min} & \text{if } i \in U \\ 0 & \text{otherwise.} \end{cases}$$

- by Fano's inequality, probability of success upper bounded as

$$1 - \mathbb{P}[\text{error}] \quad \leq \quad \frac{I(Y; \beta^*)}{\log(M-1)} - o(1),$$

  where
  - $I(Y; \beta^*)$: mutual information between β^* and observation vector Y
  - $M = \binom{p}{k}$: number of competing models

- some work to establish the upper bound holds w.h.p. for X:

$$I(Y, \beta^* \mid X) \quad \leq \quad \frac{n}{2} \log \left[1 + (1 - \frac{k}{p}) k \beta_{min}^2 \right]$$

§2. High-dimensional analysis of sparse PCA

- principal components analysis (PCA): classical method for dimensionality reduction
- high-dimensional version: eigenvectors from sample covariance $\widehat{\Sigma}$ based on n samples in p dimensions
- in general, high-dimensional PCA inconsistent unless $p/n \to 0$ (Joh01, JohLu04)
- natural to investigate more structured ensembles for which consistency still possible even with $p/n \to +\infty$:
  - sparse eigenvector recovery (JolEtal03, JohLu04, ZouEtAl06)
  - sparse covariance matrices (LevBic06,ElKar07)

Spiked covariance ensembles

- sequences $\{\Sigma_p\}$ of spiked population covariance matrices:

$$\Sigma_p = \sum_{i=1}^{M} \alpha_i \beta_i \beta_i^T + \Gamma_p, \qquad \text{with leading eigenvectors } (\beta_i, i = 1, \ldots M).$$

- past work on identity spiked ensembles ($\Gamma_p = I_p$) (Joh01; JohLu04)
- different sparsity models:
  - hard sparsity model: β has exactly k non-zero coefficients
  - weak ℓ_q-sparsity: β belongs to the ℓ_q-"ball":

$$\mathbb{B}_q(R_q) \quad = \quad \left\{ z \in \mathbb{R}^p \mid \sum_{i=1}^{p} |z_i|^q \leq R_q \right\}.$$

- given n i.i.d. samples $\{X_i\}_{i=1}^n$ with $\mathbb{E}[X_i] = 0$ and $\mathrm{cov}(X_i) = \Sigma_p$

SDP relaxation of sparse PCA

(D'Asprémont, El Ghaoui, Jordan & Lanckriet, 2006)

- Courant-Fischer variational principle for maximum eigenvalue/vector (PCA):

$$\lambda_{\max}(Q) \quad = \quad \max_{\|z\|_2=1} z^T Q z.$$

- equivalent/exact semidefinite program (SDP) of max. eigenvector:

$$\lambda_{\max}(Q) \quad = \quad \max_{Z \succeq 0, \mathrm{trace}(Z)=1} \mathrm{trace}\,(Z\,Q).$$

- *SDP relaxation* of sparse PCA:

$$\widehat{Z} \quad = \quad \arg \max_{Z \succeq 0, \mathrm{trace}(Z)=1} \left\{ \mathrm{trace}\,(Z\,Q) - \rho_n \Big(\sum_{i,j} |Z_{ij}| \Big) \right\},$$

with regularization parameter $\rho_n > 0$ chosen by user.

Rates in spectral norm

- given n samples from spiked identity model $\Sigma_p = \alpha z z^T + \sigma^2 I_p$
- eigenvector z in weak ℓ_q-ball $\mathbb{B}_q(R_q)$
- SDP relaxation: $\widehat{Z} \in \arg \min_{Z \succeq 0, \mathrm{trace}(Z)=1} \{ -\mathrm{trace}(Z\widehat{\Sigma}) + \rho_n \sum_{i,j} |Z_{ij}| \}$.

Theorem: (AmiWai08b) Suppose that we apply the SDP to the sample covariance $\widehat{\Sigma}$ with regularization parameter $\rho_n = f(\alpha, \sigma^2)\sqrt{\frac{\log p}{n}}$. Then with probability greater than $1 - c_1 \exp(-c_2 \log p) \to 0$, we have:

$$|||\widehat{Z} - zz^T|||_2 \leq C \; R_q \; (\frac{\log p}{n})^{\frac{1}{2(1+q)}}.$$

Example (Hard sparsity): $q = 0$, and radius $R_q = k$ (# non-zeros)

$$|||\widehat{Z} - zz^T|||_2 \quad \leq \quad C\sqrt{\frac{k^2 \log p}{n}}.$$

Comparison to some known results

- Estimating sparse covariance matrices (BicLev07)
  - Thresholding estimator $T_{\lambda_n}(\widehat{\Sigma})$ achieves rate:

$$|||T_{\lambda_n}(\widehat{\Sigma}) - \Sigma|||_2 \quad \leq \quad C R_q \; \left(\frac{\log p}{n}\right)^{\frac{1-q}{2}}.$$

  - by matrix perturbation results, for "well-separated" eigenvalues, same rate applies to leading eigenvector
  - agrees with SDP result for $q = 0$, but slower rate for $q > 0$
- Minimax rates for $q \in (0, 2)$: (PauJoh08)
  - with $\mathrm{sign}\langle \widehat{z}, \, z \rangle = 1$:

$$\min_{\widehat{z}} \max_{z \in \mathbb{B}_q(R_q)} \mathbb{E}[\|\widehat{z} - z\|_2^2] \quad \geq \quad C \, R_q \; \left(\frac{\log p}{n}\right)^{1-\frac{q}{2}}.$$

  - same rate as normal sequence model (DonJoh94)
  - SDP rate is slower, but approaches minimax rate as $q \to 0$

Model selection consistency for hard sparsity ($q = 0$)

Goal: Given spiked model with k-sparse eigenvector ($z_i = \pm\frac{1}{\sqrt{k}}$), recover support set $S(z) = \{i \in \{1, 2, \ldots, p\} \mid z_i \neq 0\}$ exactly.

Methods:

1. Diagonal thresholding: Complexity $\mathcal{O}(np + p \log p)$ (JohLu04)

   (a) Form sample covariance $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T$.
   (b) Extract top k order statistics $\widehat{\Sigma}_{(11)}, \ldots, \widehat{\Sigma}_{(kk)}$, and estimate support $\widehat{S}(D)$ by rank indices.

2. SDP-based recovery: Complexity $\mathcal{O}(np + p^4 \log p)$ (AspLanGhaJor08)

   (a) Solve SDP with $\rho_n = \alpha/(2\sigma^2 k)$.
   (b) Given solution $\widehat{Z}$, estimate support

$$\widehat{S} \quad := \quad \{i \in \{1, \ldots, p\} \mid \widehat{Z}_{ij} \neq 0 \quad \text{for some } j\}.$$

Sharp threshold for diagonal thresholding

Model: $\Sigma_p = \alpha z z^T + \sigma^2 I_p$

Parameters:
- $p \equiv$ model dimension
- $k \equiv$ number of non-zeroes in spiked eigenvector

Proposition: (AmiWai08a) If $k = \mathcal{O}(p^{1-\delta})$ for any $\delta \in (0,1)$, diagonal thresholding for support recovery controlled by *rescaled sample size*

$$\theta_{\mathrm{thr}}(n,p,k) \quad := \quad \frac{n}{k^2 \log(p-k)}.$$

I.e., there are constants $0 < \tau_\ell^*(\alpha,\sigma^2) \leq \tau_u^*(\alpha,\sigma^2) < \infty$ such that

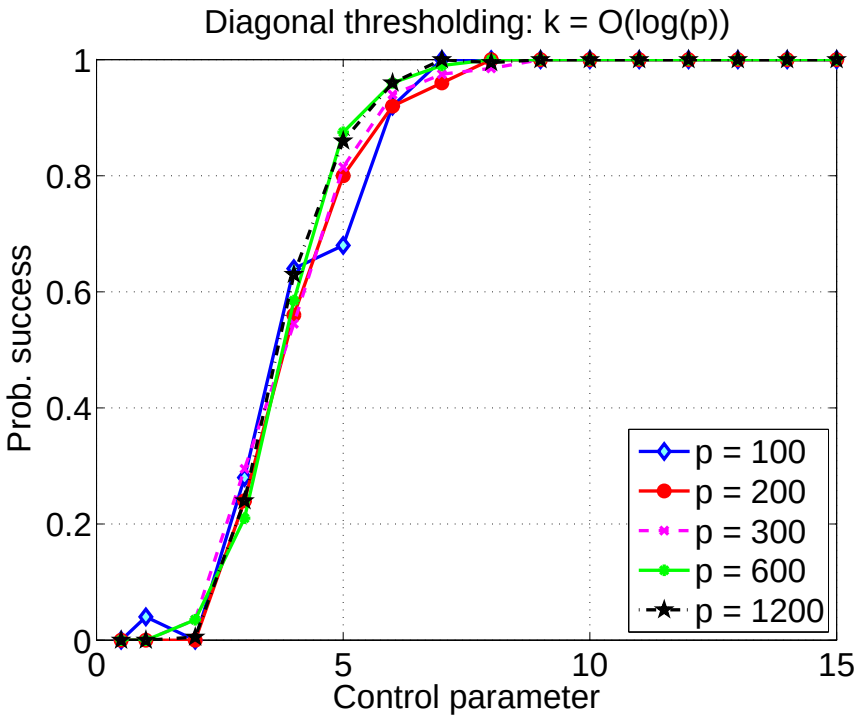
(a) **Success:** If $n > \tau_u^* k^2 \log(p-k)$, then

$$\mathbb{P}[\widehat{S}(D) = S(\beta)] \geq 1 - c_1 \exp\big(-c_2 k^2 \log(p-k))\big) \;\to\; 1.$$

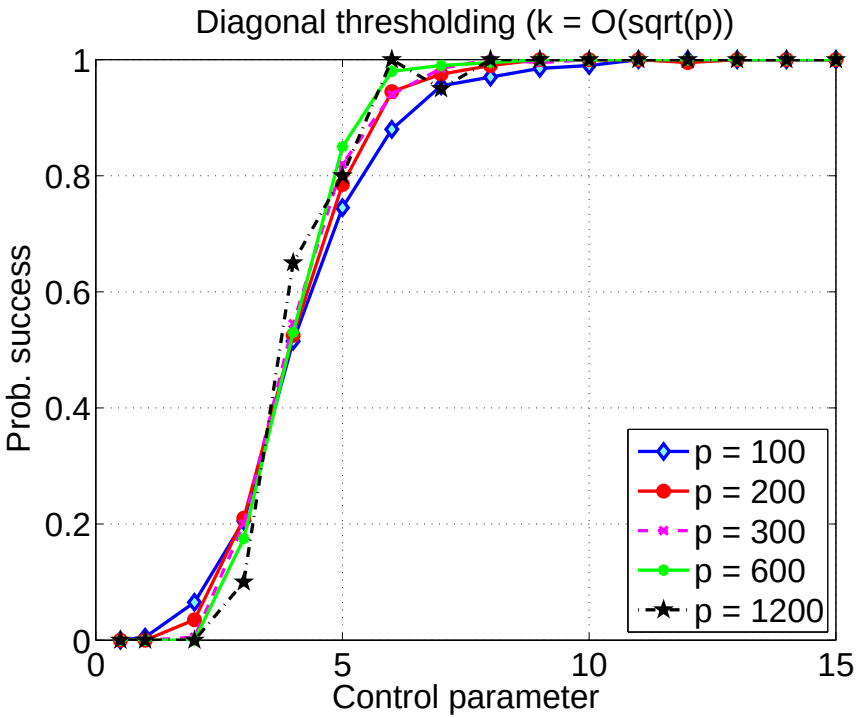
(b) **Failure:** If $n \leq \tau_\ell^* \; k^2 \log(p-k)$, then

$$\mathbb{P}[\widehat{S}(D) = S(\beta)] \leq c_1 \exp\left(-c_2(\log(p-k))\right) \;\to\; 0.$$

Performance of diagonal thresholding



(a) Log. sparsity



(b) Square-root sparsity

Probability of success $\mathbb{P}[S(D) = S(\beta^*)]$ versus rescaled sample size

$$\theta_{\mathrm{thr}}(n, p, k) = \frac{n}{k^2 \log(p - k)}$$

Eigenvector support recovery via SDP relaxation

- spiked identity model $\Sigma_p = \alpha z z^T + \sigma^2 I_p$ with k-sparse eigenvector z
- SDP relaxation: $\widehat{Z} \in \arg \min_{Z \succeq 0, \mathrm{trace}(Z)=1} \left\{ - \mathrm{trace}(Z \widehat{\Sigma}) + \rho_n \sum_{i,j} |Z_{ij}| \right\}$.

Theorem: (AmiWai08a) Suppose that we solve the SDP with $\rho_n = \alpha/(2\sigma^2 k)$. Then there are constants θ_{wr} and θ_{crit} such that

(a) For sample sizes such that $\theta_{\mathrm{thr}}(n, p, k) = \frac{n}{k^2 \log(p-k)} > \theta_{\mathrm{wr}}$, the SDP has a rank one solution w.h.p., and

(b) For problem sequences such that $k = \mathcal{O}(\log p)$, and

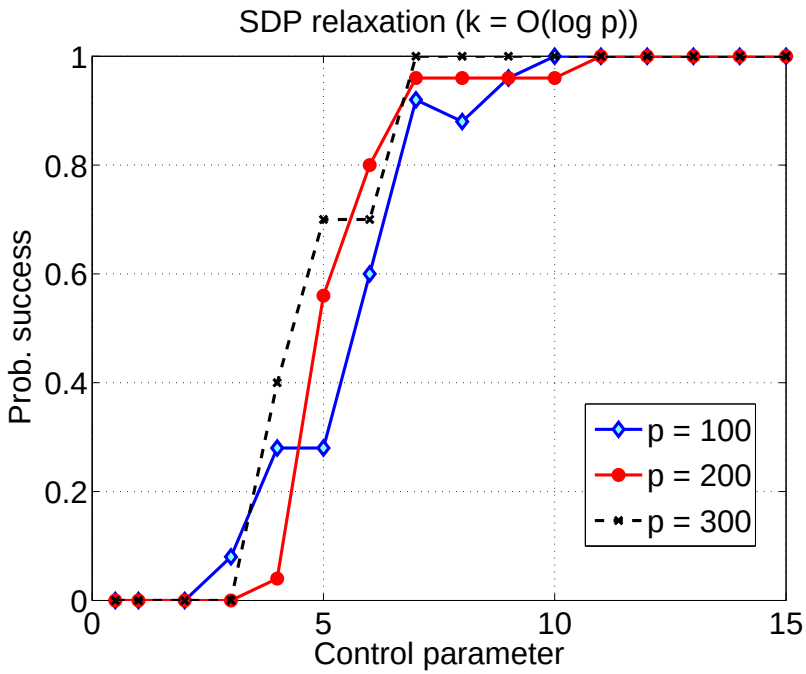
$$\theta_{\mathrm{sdp}}(n, p, k) \quad := \quad \frac{n}{k \log(p-k)} \quad > \quad \theta_{\mathrm{crit}},$$

a rank one solution (when it exists) specifies correct support w.h.p.

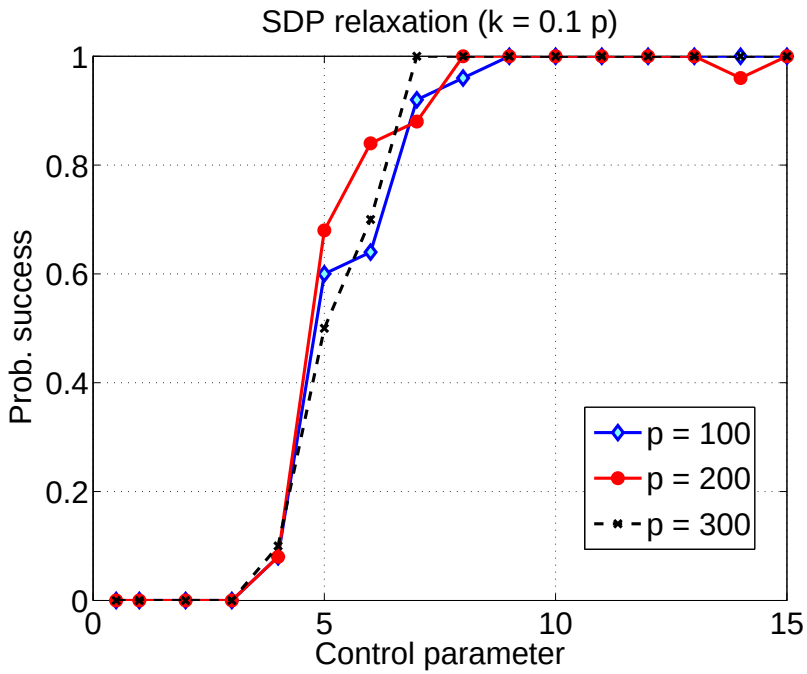
Remarks:

- technical condition $k = \mathcal{O}(\log p)$: likely an artifact

Performance of SDP relaxation



(a) Log. sparsity



(b) Linear sparsity

Probability of success $\mathbb{P}[S(\widehat{\beta}) = S(\beta^*)]$ versus rescaled sample size

$$\theta_{\mathrm{sdp}}(n, p, k) = \frac{n}{k \log(p - k)}.$$

Summary and open directions

1. When does more computation yield greater statistical accuracy?
   - Multivariate regression: second-order cone programming versus quadratic programming (Lasso)
   - Sparse PCA: diagonal thresholding versus SDP relaxation
2. When are polynomial-time algorithms as good as "optimal" algorithms?
   - Multivariate regression: Lasso/SOCP order-optimal for $k = o(p)$
   - Sparse PCA: SDP relaxation order-optimal for $k = \mathcal{O}(\log p)$

Kernel-Based Contrast Functions for Sufficient Dimension Reduction

K. Fukumizu, F. Bach, & M. I. Jordan, (2009).
Annals of Statistics, 37, 1871-1905.

Outline

- Introduction
  – dimension reduction and conditional independence
- Conditional covariance operators on RKHS
- Kernel Dimensionality Reduction for regression
- Manifold KDR
- Summary

Sufficient Dimension Reduction

- Regression setting: observe *(X,Y)* pairs, where the covariate *X* is high-dimensional
- Find a (hopefully small) subspace *S* of the covariate space that retains the information pertinent to the response *Y*
- *Semiparametric formulation*: treat the conditional distribution *p(Y* | *X)* nonparametrically, and estimate the parameter *S*

Perspectives

- Classically the covariate vector X has been treated as ancillary in regression
- The sufficient dimension reduction (SDR) literature has aimed at making use of the randomness in X (in settings where this is reasonable)
- This has generally been achieved via inverse regression
  - at the cost of introducing strong assumptions on the distribution of the covariate X
- We'll make use of the randomness in X without employing inverse regression

Dimension Reduction for Regression

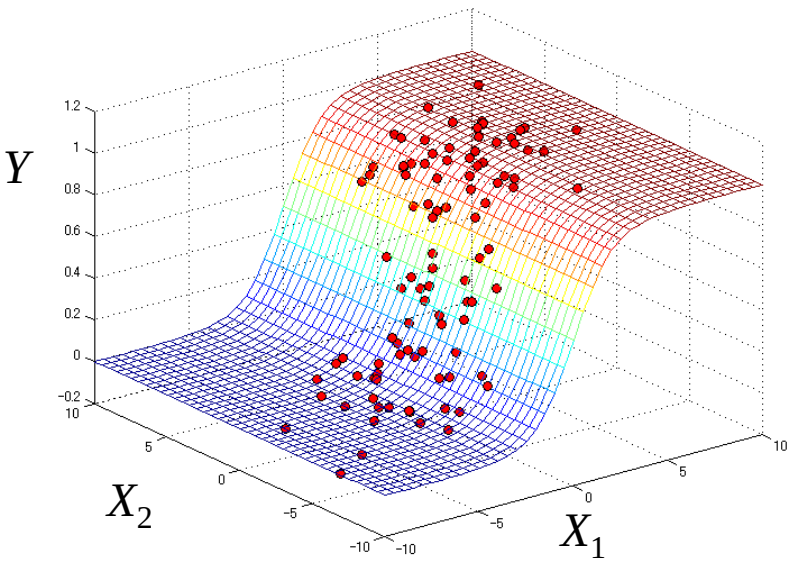
- Regression: $p(Y \mid X)$

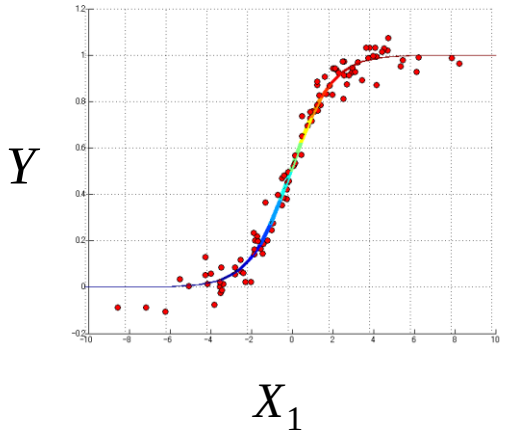
  Y : response variable,

  $X = (X_1,...,X_m)$: m-dimensional covariate

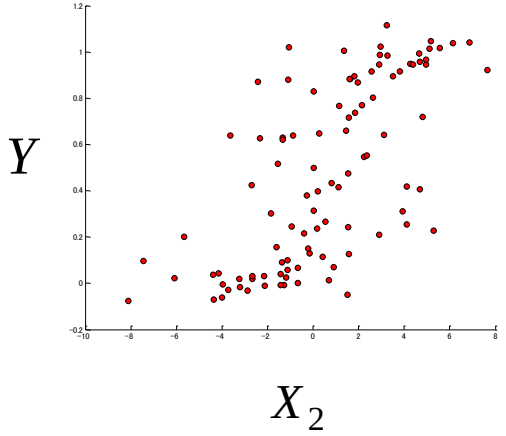
- Goal: Find the central subspace, which is defined via:

$$p(Y \mid X) = \tilde{p}(Y \mid b_1^T X,...,b_d^T X) \quad \left(= \tilde{p}(Y \mid B^T X) \right)$$





$$Y = \frac{1}{1+\exp(-X_1)} + N(0; 0.1^2)$$



central subspace = X_1 axis

Some Existing Methods

- Sliced Inverse Regression (SIR, Li 1991)
  - PCA of $\mathrm{E}[X|Y]$ → use slice of Y
  - Elliptic assumption on the distribution of X
- Principal Hessian Directions (pHd, Li 1992)
  - Average Hessian $\Sigma_{yxx} \equiv E[(Y-\bar{Y})(X-\bar{X})(X-\bar{X})^T]$ is used
  - If X is Gaussian, eigenvectors gives the central subspace
  - Gaussian assumption on X. Y must be one-dimensional
- Projection pursuit approach (e.g., Friedman et al. 1981)
  - Additive model $E[Y|X] = g_1(b_1{}^T X) + ... + g_d(b_\mathrm{d}{}^T X)$ is used
- Canonical Correlation Analysis (CCA) / Partial Least Squares (PLS)
  - Linear assumption on the regression
- Contour Regression (Li, Zha & Chiaromonte, 2004)
  - Elliptic assumption on the distribution of X

Dimension Reduction and Conditional Independence

- $(U, V)=(B^T X, C^T X)$

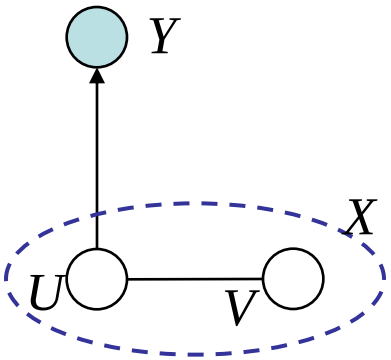
  where C: m x $(m\text{-}d)$ with columns orthogonal to B

- B gives the projector onto the central subspace

  $\Leftrightarrow \quad p_{Y|X}(y \mid x) = p_{Y|U}(y \mid B^T x)$

  $\Leftrightarrow \quad p_{Y|U,V}(y \mid u,v) = p_{Y|U}(y \mid u)$ for all y,u,v

  $\Leftrightarrow$ Conditional independence $Y \perp\!\!\!\perp V \mid U$



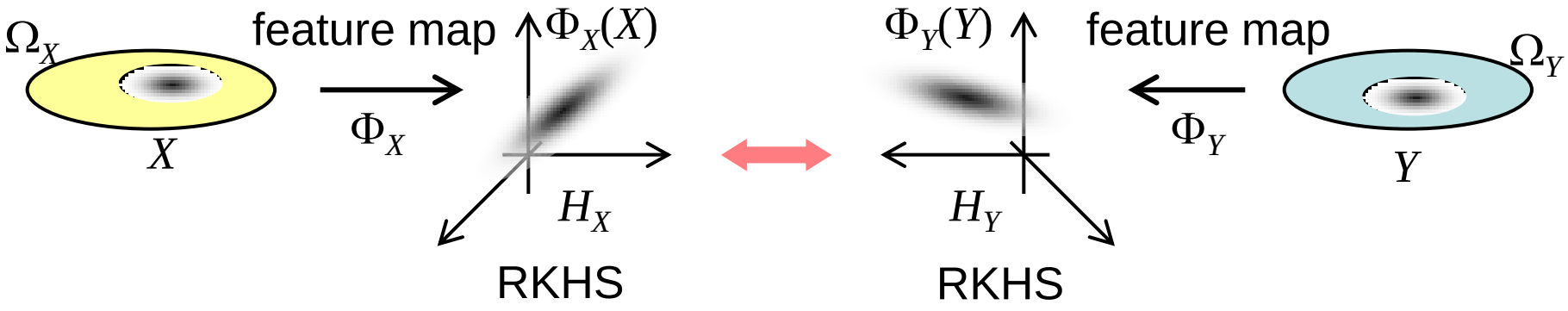
- Our approach: *Characterize conditional independence*

Outline

- Introduction
  – dimension reduction and conditional independence
- Conditional covariance operators on RKHS
- Kernel Dimensionality Reduction for regression
- Manifold KDR
- Summary

Reproducing Kernel Hilbert Spaces

- "Kernel methods"
  - RKHS's have generally been used to provide basis expansions for regression and classification (*e.g.*, support vector machine)
  - *Kernelization*: map data into the RKHS and apply linear or second-order methods in the RKHS
  - But RKHS's can also be used to characterize independence and conditional independence



Positive Definite Kernels and RKHS

- Positive definite kernel (p.d. kernel)

  $k : \Omega \times \Omega \to \mathbf{R}$

  k is positive definite if $k(x,y) = k(y,x)$ and for any $n \in \mathbf{N},\ x_1, \ldots, x_n \in \Omega$ the matrix $\left(k(x_i, x_j)\right)_{i,j}$ (Gram matrix) is positive semidefinite.

  – Example: Gaussian RBF kernel $k(x,y) = \exp\left(-\|x-y\|^2 / \sigma^2\right)$

- Reproducing kernel Hilbert space (RKHS)

  k: p.d. kernel on Ω

  $\Rightarrow$ $\exists H$: reproducing kernel Hilbert space (RKHS)

  1) $k(\cdot, x) \in H$ for all $x \in \Omega$.
  2) $\mathrm{Span}\left\{k(\cdot, x) \mid x \in \Omega\right\}$ is dense in H.
  3) $\left\langle k(\cdot, x), f \right\rangle_H = f(x)$ (reproducing property)

Functional data

$$\Phi : \Omega \to H, \quad x \mapsto k(\cdot, x) \qquad i.e. \quad \Phi(x) = k(\cdot, x)$$

Data: $X_1, \ldots, X_N$ → $\Phi_X(X_1), \ldots, \Phi_X(X_N)$: functional data

Why RKHS?

- By the reproducing property, computing the inner product on RKHS is easy:

$$\langle \Phi(x), \Phi(y) \rangle = k(x, y)$$

$$f = \sum_{i=1}^{N} a_i \Phi(x_i) = \sum_i a_i k(\cdot, x_i), \qquad g = \sum_{j=1}^{N} b_j \Phi(x_j) = \sum_j b_j k(\cdot, x_j)$$

$$\Rightarrow \quad \langle f, g \rangle = \sum_{i,j} a_i b_j k(x_i, x_j)$$

- The computational cost essentially depends on the sample size. Advantageous for high-dimensional data of small sample size.

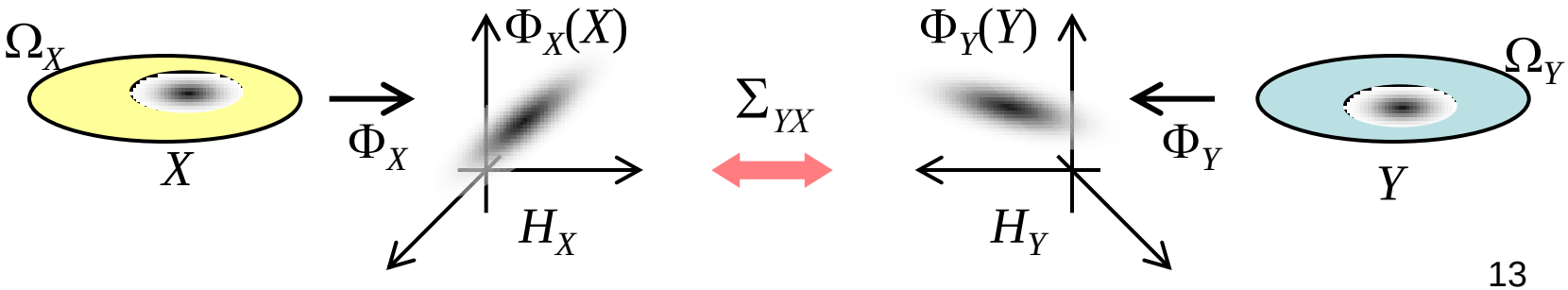
Covariance Operators on RKHS

- X , Y : random variables on Ω_X and Ω_Y , resp.
- Prepare RKHS (H_X, k_X) and (H_Y, k_Y) defined on Ω_X and Ω_Y, resp.
- Define random variables on the RKHS H_X and H_Y by

$$\Phi_X(X) = k_X(\cdot, X) \qquad \Phi_Y(Y) = k_Y(\cdot, Y)$$

- Define the covariance operator Σ_{YX}

$$\Sigma_{YX} = E[\Phi_Y(Y)\langle \Phi_X(X), \cdot \rangle] - E[\Phi_Y(Y)]E[\langle \Phi_X(X), \cdot \rangle]$$



Covariance Operators on RKHS

- Definition

$$\Sigma_{YX} = E[\Phi_Y(Y)\langle \Phi_X(X), \cdot \rangle] - E[\Phi_Y(Y)]E[\langle \Phi_X(X), \cdot \rangle]$$

Σ_{YX} is an operator from H_X to H_Y such that

$$\langle g, \Sigma_{YX} f \rangle = E[g(Y)f(X)] - E[g(Y)]E[f(X)] \ (= \mathrm{Cov}[f(X), g(Y)])$$

for all $f \in H_X, g \in H_Y$

- *cf.* Euclidean case

$V_{YX} = \mathrm{E}[YX^T] - \mathrm{E}[Y]\mathrm{E}[X]^T$: covariance matrix

$$(b, V_{YX} a) = Cov[(b, Y), (a, X)]$$

Characterization of Independence

- Independence and cross-covariance operators

If the RKHS's are "rich enough":

$$X \perp\!\!\!\perp Y \quad \Leftrightarrow \quad \Sigma_{XY} = O$$

$$\Updownarrow$$

$$\mathrm{Cov}[f(X), g(Y)] = 0$$

or

$$E[g(Y)f(X)] = E[g(Y)]E[f(X)]$$

for all $f \in H_X, g \in H_Y$

$\Rightarrow$ is always true

$\Leftarrow$ requires an assumption on the kernel (universality)

e.g., Gaussian RBF kernels are universal

$$k(x, y) = \exp\left(-\|x - y\|^2 / \sigma^2\right)$$

– *cf*. for Gaussian variables,

X and Y are independent $\Leftrightarrow$ $V_{XY} = O$ i.e. uncorrelated

- Independence and characteristic functions

  Random variables X and Y are independent

  $$\Leftrightarrow E_{XY}\left[e^{i\omega^T X}e^{i\eta^T Y}\right]=E_X\left[e^{i\omega^T X}\right]E_Y\left[e^{i\eta^T Y}\right] \qquad \text{for all } \omega \text{ and } \eta$$

  I.e., $e^{i\omega^T x}$ and $e^{i\eta^T y}$ work as test functions

- RKHS characterization

  Random variables $X \in \Omega_X$ and $Y \in \Omega_Y$ are independent

  $$\Leftrightarrow \quad E_{XY}\left[f(X)g(Y)\right]=E_X\left[f(X)\right]E_Y\left[g(Y)\right] \qquad \text{for all } f \in \mathcal{H}_X,\ g \in \mathcal{H}_Y$$

  – RKHS approach is a generalization of the characteristic-function approach

RKHS and Conditional Independence

- Conditional covariance operator

X and Y are random vectors. $\mathcal{H}_X$, $\mathcal{H}_Y$: RKHS with kernel k_X, k_Y, resp.

Def. $\Sigma_{YY|X} \equiv \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$: conditional covariance operator

– Under a universality assumption on the kernel

$$\langle g, \Sigma_{YY|X} g \rangle = E\left[\mathrm{Var}[g(Y) \mid X]\right]$$

cf. For Gaussian $\mathrm{Var}_{Y|X}[a^T Y \mid X = x] = a^T \left(V_{YY} - V_{YX}V_{XX}^{-1}V_{XY}\right)a$

– Monotonicity of conditional covariance operators

$X = (U, V)$: random vectors

$$\Sigma_{YY|U} \geq \Sigma_{YY|X}$$

$\geq$: in the sense of self-adjoint operators

- Conditional independence

Theorem

$X = (U,V)$ and Y are random vectors.

$\mathcal{H}_X$, $\mathcal{H}_U$, $\mathcal{H}_Y$: RKHS with Gaussian kernel k_X, k_U, k_Y, resp.

$$\Rightarrow \qquad Y \perp\!\!\!\perp V \mid U \quad \Leftrightarrow \quad \Sigma_{YY|U} = \Sigma_{YY|X}$$

This theorem provides a new methodology for solving the sufficient dimension reduction problem

Outline

- Introduction
  – dimension reduction and conditional independence
- Conditional covariance operators on RKHS
- Kernel Dimensionality Reduction for regression
- Manifold KDR
- Summary

Kernel Dimension Reduction

- Use a universal kernel for $B^T X$ and Y

$$\Sigma_{YY|B^T X} \geq \Sigma_{YY|X}$$

($\geq$: the partial order of self-adjoint operators)

$$\Sigma_{YY|B^T X} = \Sigma_{YY|X} \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y \mid B^T X$$

- KDR objective function:

$$\min_{B:\, B^T B = I_d} \mathrm{Tr}\left[\Sigma_{YY|B^T X}\right]$$

which is an optimization over the Stiefel manifold

Estimator

- Empirical cross-covariance operator

$$\hat{\Sigma}_{YX}^{(N)} = \frac{1}{N}\sum_{i=1}^{N}\left\{k_Y(\cdot, Y_i) - \hat{m}_Y\right\} \otimes \left\{k_X(\cdot, X_i) - \hat{m}_X\right\}$$

$$\hat{m}_X = \frac{1}{N}\sum_{i=1}^{N} k_X(\cdot, X_i) \qquad \hat{m}_Y = \frac{1}{N}\sum_{i=1}^{N} k_Y(\cdot, Y_i)$$

$\hat{\Sigma}_{YX}^{(N)}$ gives the empirical covariance:

$$\left\langle g, \hat{\Sigma}_{YX}^{(N)} f \right\rangle = \frac{1}{N}\sum_{i=1}^{N} f(X_i) g(Y_i) - \frac{1}{N}\sum_{i=1}^{N} f(X_i) \frac{1}{N}\sum_{i=1}^{N} g(Y_i)$$

- Empirical conditional covariance operator

$$\hat{\Sigma}_{YY|X}^{(N)} = \hat{\Sigma}_{YY}^{(N)} - \hat{\Sigma}_{YX}^{(N)} \left(\hat{\Sigma}_{XX}^{(N)} + \varepsilon_N I\right)^{-1} \hat{\Sigma}_{XY}^{(N)}$$

ε_N: regularization coefficient

- Estimating function for KDR:

$$\mathrm{Tr}\left[\hat{\Sigma}^{(N)}_{YY|U}\right] = \mathrm{Tr}\left[\hat{\Sigma}^{(N)}_{YY} - \hat{\Sigma}^{(N)}_{YU}\left(\hat{\Sigma}^{(N)}_{UU} + \varepsilon_N I\right)^{-1}\hat{\Sigma}^{(N)}_{UY}\right] \qquad U = B^T X$$

$$= \mathrm{Tr}\left[G_Y - G_Y G_U\left(G_U + N\varepsilon_N I_N\right)^{-1}\right]$$

where

$G_U = \left(I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T\right)K_U\left(I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T\right)$: centered Gram matrix

$K_U = k(B^T X_i, B^T X_j)$

- Optimization problem:

$$\min_{B:B^TB=I_d} \mathrm{Tr}\left[G_Y\left(G_{B^TX} + N\varepsilon_N I_N\right)^{-1}\right]$$

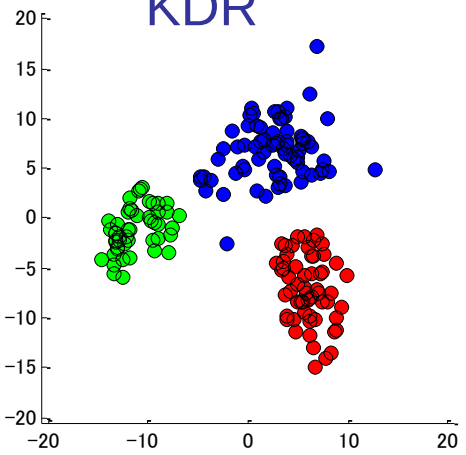
Experiments with KDR

- Wine data
  - Data
    - 13 dim. 178 data
    - 3 classes
    - 2 dim. projection

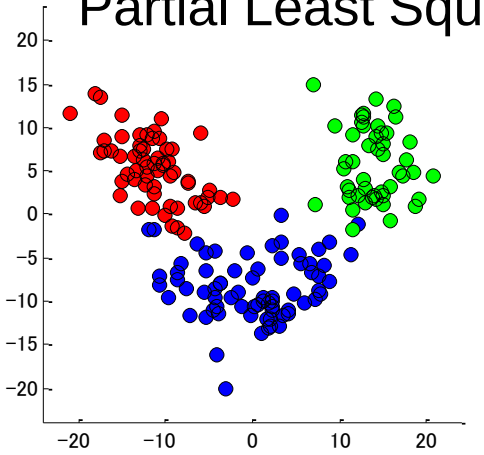
$$k(z_1, z_2) = \exp\left(-\left\|z_1 - z_2\right\|^2 \Big/ \sigma^2\right)$$

$$\sigma = 30$$

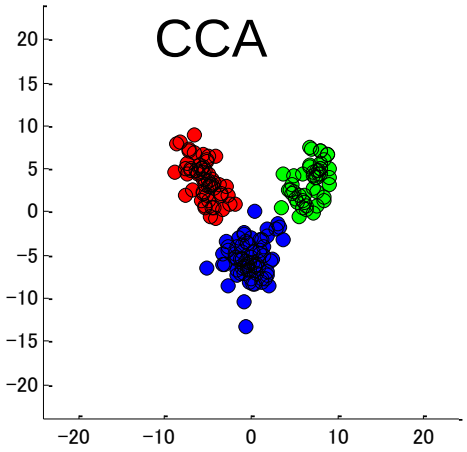
KDR



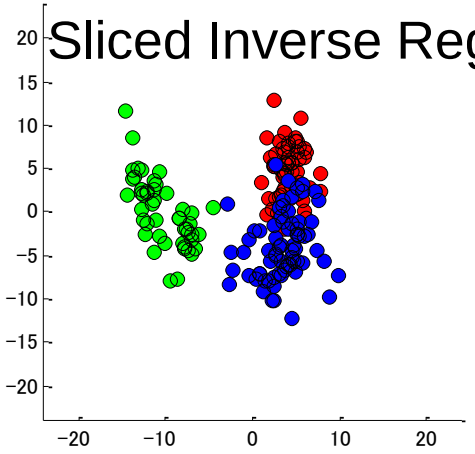
Partial Least Square



CCA



Sliced Inverse Regression



Consistency of KDR

Theorem

Suppose k_d is bounded and continuous, and

$$\varepsilon_N \to 0, \ N^{1/2}\varepsilon_N \to \infty \ \ (N \to \infty).$$

Let S_0 be the set of optimal parameters:

$$S_0 = \left\{ B \mid B^T B = I_d, \mathrm{Tr}\left[\Sigma^B_{YY|X} \right] = \min_{B'} \mathrm{Tr}\left[\Sigma^{B'}_{YY|X} \right] \right\}$$

Then, under some conditions, for any open set $U \supset S_0$

$$\Pr\left(\hat{B}^{(N)} \in U \right) \to 1 \quad (N \to \infty).$$

Lemma

Suppose k_d is bounded and continuous, and

$$\varepsilon_N \to 0,\ N^{1/2}\varepsilon_N \to \infty\ \ (N \to \infty).$$

Then, under some conditions,

$$\sup_{B:B^T B = I_d} \left| \mathrm{Tr}\left[\hat{\Sigma}_{YY|X}^{B(N)} \right] - \mathrm{Tr}\left[\Sigma_{YY|X}^{B} \right] \right| \to 0\ \ (N \to \infty)$$

in probability.

Conclusions

- Are you a Bayesian or a frequentist?
- My own answer is “both,” but there are days where I'm much more clearly one than the other
  - and it is an ongoing intellectual challenge to try to understand the ramifications of this distinction
- I view them as complementary perspectives, but there is a wave/particle uncomfortableness at times
- A main conclusion: machine learning is a part of statistics; don't just read the machine learning literature---read, ponder and contribute to the broad statistical literature