# One-shot learning and big data with $n = 2$

Lee Dicker
Department of Statistics
Rutgers University

Joint work w/Dean Foster

DIMACS, May 16, 2013

# Introduction and overview

# One-shot learning

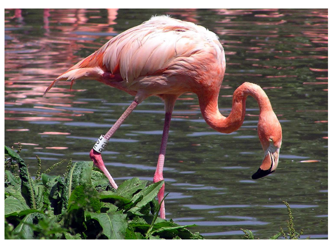
Introduction and overview
Statistical setting
Principal component regression
Weak consistency and big data with $n = 2$
Risk approximations and consistency
Numerical results
Conclusions and future directions

- Humans are able to correctly recognize and understand objects based on very few training examples.
  - e.g. images, words.

Training

Testing

Flamingo ✓



Flamingo ✓

$\longrightarrow$

Flamingo?

Flamingo?

Flamingo?

- Vast literature in cognitive science (Tenenbaum et al., 2006; Kemp et al., 2007), language acquisition (Carey et al., 1978; Xu et al., 2007), and computer vision (Fink, 2005; Fei-Fei et al., 2006)

# One-shot learning



- Successful one-shot learning requires the learner to incorporate strong contextual information into the learning algorithm.
  - Image recognition: Information on object categories.
    - Objects tend to be categorized by shape, color, etc.
  - Word-learning: Common function words are often used in conjunction with a novel word and referent.
    - `This is a KOBA.` Since `this`, `is`, and a are function words that often appear with nouns, `KOBA` is likely the new referent.
- Many recent statistical approaches to one-shot learning are based on hierarchical Bayesian models.
  - Effective in a variety of examples.

- We propose a simple factor model for one-shot learning with continuous outcomes.
  - *Highly* idealized, but amenable to theoretical analysis.
  - Novel risk approximations for:
    - (i) assessing the performance of one-shot learning methods and
    - (ii) gaining insight into the significance of various parameters for one-shot learning.
- The methods considered here are variants of principal component regression (PCR).
  - One-shot asymptotic regime: Fixed $n$, large $d$, strong contextual information.
    - See work by Hall, Jung, Marron, and co-authors on "high dimension, low sample size" data (especially work on PCA and classification).
  - New insights into PCR.
    - Classical PCR estimator is generally inconsistent in the one-shot regime.
    - Bias-correction via expansion.

# Outline



- Statistical setting.
- Principal component regression.
- Weak consistency and big data with $n = 2$.
- Risk approximations and consistency.
- Numerical results.
- Conclusions and future directions.

# Statistical setting

- The observed data consists of $(y_1, \mathbf{x}_1), ..., (y_n, \mathbf{x}_n)$, where $y_i \in \mathbb{R}$ is a scalar outcome and $\mathbf{x}_i \in \mathbb{R}^d$ is an associated $d$-dimensional "context" vector.
- We suppose that $y_i$ and $\mathbf{x}_i$ are related via

$$\begin{aligned} y_i &= h_i\theta + \xi_i, & h_i &\sim N(0,\eta^2),\ \xi_i \sim N(0,\sigma^2), \\ \mathbf{x}_i &= h_i\gamma\sqrt{d}\mathbf{u} + \boldsymbol{\epsilon}_i, & \boldsymbol{\epsilon}_i &\sim N(0,\tau^2 I). \end{aligned}$$

- NB:
  - $h_i, \xi_i \in \mathbb{R}$ and $\boldsymbol{\epsilon}_i \in \mathbb{R}^d$, $1 \leq i \leq n$, are all assumed to be independent.
    - $h_i$ is a latent factor linking $y_i$ and $\mathbf{x}_i$.
    - $\xi_i$ and $\boldsymbol{\epsilon}_i$ are random noise.
  - The unit vector $\mathbf{u} \in \mathbb{R}^d$ and real numbers $\theta, \gamma \in \mathbb{R}$ are non-random.
  - It is implicit in our normalization that the "$\mathbf{x}$-signal" $||h_i\gamma\sqrt{d}\mathbf{u}||^2 \asymp d$ is quite strong.
  - To simplify notation, we let $\mathbf{y} = (y_1, ..., y_n)$ and $X = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$.

- Observe that $(y_i, \mathbf{x}_i) \sim N(0, V)$ are jointly normal with

$$V = \begin{pmatrix} \theta^2\eta^2 + \sigma^2 & \theta\gamma\eta^2\sqrt{d}\mathbf{u}^T \\ \theta\gamma\eta^2\sqrt{d}\mathbf{u} & \tau^2 I + \eta^2\gamma^2 d\mathbf{u}\mathbf{u}^T \end{pmatrix}. \qquad (\dagger)$$

- **Goal:** Given the data $(\mathbf{y}, X)$, devise prediction rules $\hat{y} : \mathbb{R}^d \to \mathbb{R}$ so that the risk

$$R_V(\hat{y}) = E_V\{\hat{y}(\mathbf{x}_{new}) - y_{new}\}^2 = E_V\{\hat{y}(\mathbf{x}_{new}) - h_{new}\theta\}^2 + \sigma^2$$

  is small, where $(y_{new}, \mathbf{x}_{new}) = (h_{new}\theta + \xi_{new}, h_{new}\gamma\sqrt{d}\mathbf{u} + \boldsymbol{\epsilon}_{new})$ has the same distribution as $(y_i, \mathbf{x}_i)$ and is independent of $(\mathbf{y}, X)$.

- $R_V(\hat{y})$ is a measure of *predictive risk*, which is completely determined by $\hat{y}$ and the parameter matrix $V$, given in (†).

## One-shot asymptotic regime



- We are primarily interested in identifying methods $\hat{y}$ that perform well in the *one-shot asymptotic regime*.
- Key features of the one-shot asymptotic regime:

$$\left.\begin{array}{ll}\text{(i)} & n \text{ is fixed} \\ \text{(ii)} & d \to \infty\end{array}\right\}\text{small } n\text{, large } d$$

$$\left.\begin{array}{ll}\text{(iii)} & \sigma^2 \to 0 \\ \text{(iv)} & \inf \eta^2\gamma^2/\tau^2 > 0\end{array}\right\}\text{abundant contextual information}$$

- NB:
  - $\sigma^2$ is the noise-level for the "$y$-data."
  - $\eta^2\gamma^2/\tau^2$ is the signal-to-noise ratio for the "$\mathbf{x}$-data."

# Principal component regression

- By assumption, the data are multivariate normal. Thus,

$$E_V(y_i|\mathbf{x}_i) = \mathbf{x}_i^T \boldsymbol{\beta},$$

where $\boldsymbol{\beta} = \theta\gamma\eta^2\sqrt{d}\mathbf{u}/(\tau^2 + \eta^2\gamma^2 d)$.

- This suggests studying linear prediction rules of the form

$$\hat{y}(\mathbf{x}) = \mathbf{x}^T \hat{\boldsymbol{\beta}}$$

for some estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$.

- Let $l_1 \geq \cdots \geq l_{n \wedge d} \geq 0$ denote the ordered $n$ largest eigenvalues of $X^T X$ and let $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_{n \wedge d}$ denote corresponding eigenvectors with unit length.
  - $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_{n \wedge d}$ are the principal components of $X$.
- Let $U_k = (\hat{\mathbf{u}}_1 \cdots \hat{\mathbf{u}}_k)$ be the $d \times k$ matrix with columns given by $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_k$, for $1 \leq k \leq n \wedge d$. In its most basic form, *principal component regression* involves regressing $\mathbf{y}$ on $XU_k$ for some (typically small) $k$, and taking $\hat{\boldsymbol{\beta}} = U_k(U_k^T X^T X U_k)^{-1} U_k^T X^T \mathbf{y}$.
- In the problem considered here, $\mathrm{Cov}(\mathbf{x}_i) = \tau^2 I + \eta^2 \gamma^2 d \mathbf{u}\mathbf{u}^T$ has a single eigenvector larger than $\tau^2$ and the corresponding eigenvector is parallel to $\boldsymbol{\beta}$. Thus, it is natural to take $k = 1$ and consider the principal component regression (PCR) estimator

$$\hat{\boldsymbol{\beta}}_{pcr} = \frac{\hat{\mathbf{u}}_1^T X^T \mathbf{y}}{\hat{\mathbf{u}}_1^T X^T X \hat{\mathbf{u}}_1} \hat{\mathbf{u}}_1 = \frac{1}{l_1} \hat{\mathbf{u}}_1^T X^T \mathbf{y} \hat{\mathbf{u}}_1.$$

# Weak consistency and big data with $n = 2$

- As a warm-up for the general $n$ setting, we consider the special case where $n = 2$.
- When $n = 2$, the PCR estimator $\hat{\boldsymbol{\beta}}_{pcr}$ has an especially simple form because the largest eigenvalue of $X^T X$ and its corresponding eigenvector are given explicitly by

$$\begin{aligned} l_1 &= \frac{1}{2}\left\{||\mathbf{x}_1||^2 + ||\mathbf{x}_2||^2 + \sqrt{(||\mathbf{x}_1||^2 - ||\mathbf{x}_2||^2)^2 + 4(\mathbf{x}_1^T\mathbf{x}_2)^2}\right\}, \\ \hat{\mathbf{u}}_1 &\propto \frac{l_1 - ||\mathbf{x}_2||^2}{\mathbf{x}_1^T\mathbf{x}_2}\mathbf{x}_1 + \mathbf{x}_2. \end{aligned}$$

- Recall that $\mathbf{x}_i = h_i\gamma\sqrt{d}\mathbf{u}_i + \boldsymbol{\epsilon}_i$. Using the large $d$ approximations

$$\begin{aligned} ||\mathbf{x}_i||^2 &\approx h_i^2\gamma^2 d + \tau^2 d \\ \mathbf{x}_1^T\mathbf{x}_2 &\approx h_1 h_2 \gamma^2 d \end{aligned}$$

leads to...

- **Large $d$ approximation:**

$$\hat{y}_{pcr}(\mathbf{x}_{new}) = \mathbf{x}_{new}^T \hat{\boldsymbol{\beta}}_{pcr} \approx \frac{\gamma^2(h_1^2 + h_2^2)}{\gamma^2(h_1^2 + h_2^2) + \tau^2} h_{new}\theta + e_{pcr},$$

where $e_{pcr} = o_P(1)$, as $d \to \infty$ and $\sigma^2 \to 0$.

- Thus,

$$\begin{aligned} \hat{y}_{pcr}(\mathbf{x}_{new}) - y_{new} &\approx -\frac{\tau^2}{\gamma^2(h_1^2 + h_2^2) + \tau^2} h_{new}\theta + e_{pcr} - \xi_{new} \\ &\to -\frac{\tau^2}{\gamma^2(h_1^2 + h_2^2) + \tau^2} h_{new}\theta \\ &\neq 0, \end{aligned}$$

as $d \to \infty$ and $\sigma^2 \to 0$.

- In other words, $\hat{y}_{pcr}$ is *inconsistent* in the one-shot regime.

- To obtain a consistent method, we multiply the PCR estimator $\hat{\boldsymbol{\beta}}_{pcr}$ by

$$\frac{l_1}{l_1 - l_2} \approx \frac{\gamma^2(h_1^2 + h_2^2) + \tau^2}{\gamma^2(h_1^2 + h_2^2)} > 1.$$

The bias-corrected estimator is

$$\hat{\boldsymbol{\beta}}_{bc} = \frac{l_1}{l_1 - l_2}\hat{\boldsymbol{\beta}}_{pcr} = \frac{1}{l_1 - l_2}\hat{\mathbf{u}}_1^T X^T \mathbf{y}\hat{\mathbf{u}}_1.$$

- When $d$ is large and $\sigma^2$ is small,

$$\hat{y}_{bc}(\mathbf{x}_{new}) - y_{new} \approx \frac{\gamma^2(h_1^2 + h_2^2) + \tau^2}{\gamma^2(h_1^2 + h_2^2)} e_{pcr} + \xi_{new} = o_P(1).$$

- It follows that $|\hat{y}_{bc}(\mathbf{x}_{new}) - y_{new}| \to 0$ in probability; that is, $\hat{y}_{bc}$ is *weakly consistent*.

- On the other hand, $R_V(\hat{y}_{bc}) = \infty$ because $E_V(h_1^2 + h_2^2)^{-1} = \infty$.
  - To obtain finite risk, we must take $n$ a little bit larger.

# Risk approximations and consistency

# Bias-corrected PCR



- When $n = 2$, we found that $\hat{\boldsymbol{\beta}}_{pcr}$ is inconsistent in the one-shot regime; to remedy this, we introduced the bias-corrected PCR estimator.

- A similar phenomenon occurs for arbitrary fixed $n \geq 2$. For $d \geq n \geq 2$, define the bias-corrected PCR estimator

$$\hat{\boldsymbol{\beta}}_{bc} = \frac{l_1}{l_1 - l_n}\hat{\boldsymbol{\beta}}_{pcr} = \frac{1}{l_1 - l_n}\hat{\mathbf{u}}_1^T X^T \mathbf{y}\hat{\mathbf{u}}_1.$$

- Note that

$$||\hat{\boldsymbol{\beta}}_{bc}|| = \frac{l_1}{l_1 - l_n}||\hat{\boldsymbol{\beta}}_{pcr}|| \geq ||\hat{\boldsymbol{\beta}}_{pcr}||.$$

  - $\hat{\boldsymbol{\beta}}_{bc}$ is obtained from $\hat{\boldsymbol{\beta}}_{pca}$ by *expansion*.

- If $n = 2$, then $R_V(\hat{y}_{bc}) = \infty$.
  - Inverse moments of $\chi^2$ random variable.
- When $n$ is larger, there are "enough" degrees of freedom and $R_V(\hat{y}_{bc})$ is finite.

**Theorem:** Suppose that $\eta^2\gamma^2/\tau^2 > c$ for some constant $c > 0$.

(a) If $n \geq 9$ and $d \geq 1$, then

$$R_V(\hat{y}_{pcr}) = \sigma^2 + \theta^2\eta^2 \left(\frac{\eta^2\gamma^2 d}{\eta^2\gamma^2 d + \tau^2}\right)^2 E_V\left\{(\mathbf{u}^T\hat{\mathbf{u}}_1)^2 - 1\right\}^2 + (\text{smaller terms}).$$

(b) If $d \geq n \geq 9$, then

$$R_V(\hat{y}_{bc}) = \sigma^2 + \theta^2\eta^2 \left(\frac{\eta^2\gamma^2 d}{\eta^2\gamma^2 d + \tau^2}\right)^2 E_V\left\{\frac{l_1}{l_1 - l_n}(\mathbf{u}^T\hat{\mathbf{u}}_1)^2 - 1\right\}^2 + (\text{smaller terms}).$$

**Proposition.** Let $W_n \sim \chi_n^2$ be a chi-squared random variable with $n$ degrees of freedom. If $n \geq 9$ is fixed, $d \to \infty$, and $\eta^2\gamma^2/\tau^2 > c$ for some constant $c > 0$, then

$$\begin{aligned} E_V\left\{(\mathbf{u}^T\hat{\mathbf{u}}_1)^2 - 1\right\}^2 &\to E\left\{\frac{\tau^2}{\eta^2\gamma^2 W_n + \tau^2}\right\}^2, \\ E_V\left\{\frac{l_1}{l_1 - l_n}(\mathbf{u}^T\hat{\mathbf{u}}_1)^2 - 1\right\}^2 &\to 0. \end{aligned}$$

**Corollary.** If $n \geq 9$ is fixed, then

$$\begin{aligned} R_V(\hat{y}_{pcr}) &\to \theta^2\eta^2 E\left\{\frac{\tau^2}{\eta^2\gamma^2 W_n + \tau^2}\right\}^2, \\ R_V(\hat{y}_{bc}) &\to 0 \end{aligned}$$

in the one-shot regime, where $d \to \infty$, $\sigma^2 \to 0$, and $\inf \eta^2\gamma^2/\tau^2 > 0$. In particular, $\hat{y}_{pcr}$ is inconsistent, but $\hat{y}_{bc}$ is consistent.

# Numerical results

- We conducted a simulation study to compare the performance of $\hat{y}_{pcr}$ and $\hat{y}_{bc}$.
- We fixed:
  - $\theta = 4$, $\sigma^2 = 1/10$, $\eta^2 = 4$, $\gamma^2 = 1/4$, $\tau^2 = 1$.
    - NB: $\sigma^2 = 1/10$ is fairly small; $\eta^2\gamma^2/\tau^2 = 1$ is reasonably large.
  - $\mathbf{u} = (1, 0, ..., 0) \in \mathbb{R}^d$.
- We simulated 1000 independent datasets with various $d, n$ and computed:
  - Empirical prediction error.
  - Theoretical prediction error (as given by the leading terms in our risk approximations).
  - Relative error,

$$\left| \frac{(\text{Empirical PE}) - (\text{Theoretical PE})}{\text{Empirical PE}} \right| \times 100\%.$$

# Numerical results



$d = 500$

| | | PCR | | Bias-corrected PCR | |
|---|---|---|---|---|---|
| $n = 2$ | Empirical PE | 17.9710 | | 4.6898 | |
| | Theoretical PE (Rel. Err.) | ? | (?) | $\infty$ | ($\infty$) |
| $n = 4$ | Empirical PE | 7.0684 | | 1.0616 | |
| | Theoretical PE (Rel. Err.) | ? | (?) | ? | (?) |
| $n = 9$ | Empirical PE | 1.4555 | | 0.3565 | |
| | Theoretical PE (Rel. Err.) | 1.3959 | (4.10%) | 0.2175 | (38.98%) |
| $n = 20$ | Empirical PE | 0.4485 | | 0.2737 | |
| | Theoretical PE (Rel. Err.) | 0.4330 | (3.45%) | 0.1399 | (48.89%) |

$d = 5000$

| | | PCR | | Bias-corrected PCR | |
|---|---|---|---|---|---|
| $n = 2$ | Empirical PE | 18.1134 | | 1.7101 | |
| | Theoretical PE (Rel. Err.) | ? | (?) | $\infty$ | ($\infty$) |
| $n = 4$ | Empirical PE | 6.0708 | | 0.2378 | |
| | Theoretical PE (Rel. Err.) | ? | (?) | ? | (?) |
| $n = 9$ | Empirical PE | 1.3257 | | 0.1395 | |
| | Theoretical PE (Rel. Err.) | 1.2737 | (3.92%) | 0.1306 | (6.40%) |
| $n = 20$ | Empirical PE | 0.3229 | | 0.1237 | |
| | Theoretical PE (Rel. Err.) | 0.3127 | (3.17%) | 0.1115 | (9.84%) |

# Conclusions and future directions

# Conclusions and future directions

## Conclusions:

- We've proposed a simple factor model and a relevant asymptotic regime for one-shot learning with continuous outcomes.
  - Identified consistent methods.
  - Gained new insights into PCR.
    - Bias-correction via expansion may lead to improved performance.

## Future directions:

- *Classification.*
  - Flexible classification methods based on probit/latent variable models and techniques discussed here.
- *Sparsity.*
  - Sparsity is a major topic in high-dimensional data analysis. How does sparsity fit into one-shot learning?
  - If $\mathbf{u}$ is sparse, then effective one-shot learning may be possible with smaller $\mathbf{x}$-data signal-to-noise ratio.
- *Applications!*