Dimension reduction techniques for efficient subspace approximation

Felix Krahmer

Institute for Numerical and Applied Mathematics, University of Göttingen

12/12/2013

Joint work with Mark Iwen (Michigan State)

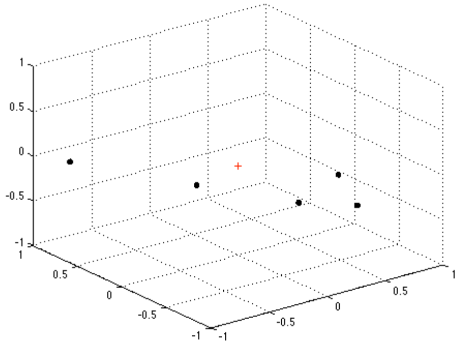
Outline

Outline

Subspace Approximation

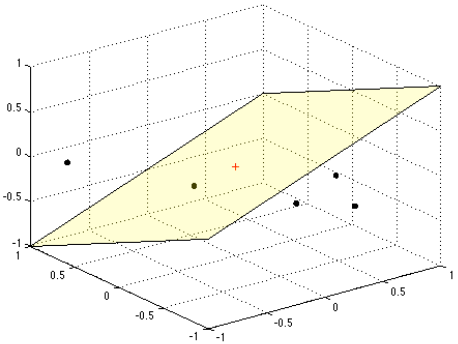
- Set up: We have many data vectors

$$X = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\} \subset \mathbb{R}^N$$



Subspace Approximation

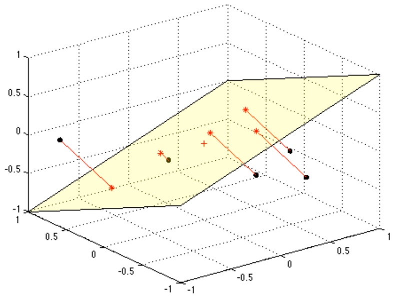
- Set up: We have many data vectors
$$X = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\} \subset \mathbb{R}^N$$
- M and N are both large
- Want to fit X with an n-dimensional hyperplane, $\mathcal{P} \subset \mathbb{R}^N$.



Measuring the Fit

- We can define an error vector $\mathbf{e}_{\mathcal{P}} \in \mathbb{R}^M$ by

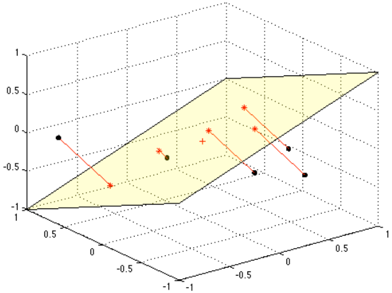
$$(e_{\mathcal{P}})_j := \|\mathbf{x}_j - \Pi_{\mathcal{P}} \mathbf{x}_j\|_2$$



Measuring the Fit

- We can define an error vector $\mathbf{e}_{\mathcal{P}} \in \mathbb{R}^M$ by

$$(e_{\mathcal{P}})_j := \|\mathbf{x}_j - \Pi_{\mathcal{P}} \mathbf{x}_j\|_2$$



Measuring the Fit

The q-distance from X to $\mathcal{P}$ is

$$d^{(q)}(X, \mathcal{P}) := \|\mathbf{e}_{\mathcal{P}}\|_q.$$

The Best Fit

- Let $\Gamma_n(\mathcal{S})$ denote the set of all n-dimensional subspaces of a given higher dimensional subspace $\mathcal{S} \subset \mathbb{R}^N$
- The best fit for $X \subset \mathbb{R}^N$ by an n-dimensional subspace is

$$d_n^{(q)}\left(X, \mathbb{R}^N\right) := \inf_{\mathcal{P} \in \Gamma_n(\mathbb{R}^N)} d^{(q)}(X, \mathcal{P}) = \inf_{\mathcal{P} \in \Gamma_n(\mathbb{R}^N)} \|\mathbf{e}_{\mathcal{P}}\|_q.$$

- We want to find an optimal n-dimensional subspace, $\mathcal{P}_{\mathrm{opt}} \in \Gamma_n\left(\mathbb{R}^N\right)$, such that

$$d^{(q)}(X, \mathcal{P}_{\mathrm{opt}}) = d_n^{(q)}\left(X, \mathbb{R}^N\right).$$

- Note that at least one $\mathcal{P}_{\mathrm{opt}}$ exists (Stiefel manifolds are compact, and $d^{(q)}(X, \cdot)$ is continuous)

The Problem

Our Goal

Compute an n-dimensional subspace, $\mathcal{A} \in \Gamma_n\left(\mathbb{R}^N\right)$, with

$$d^{(q)}(X, \mathcal{A}) \approx d_n^{(q)}\left(X, \mathbb{R}^N\right) := \inf_{\mathcal{P} \in \Gamma_n(\mathcal{S})} \|\mathbf{e}_{\mathcal{P}}\|_q.$$

- q $= 2$: This case seeks to minimize least squares error, and can be solved by computing the top n eigenvectors of $X^{\mathrm{T}}X$.
- Many computational methods solve $q = 2$ case, including randomized methods that succeed with high probability, are stable, and use only $O(MNn)$-flops.
- Other cases have been studied less, although a lot is still known...

The Case $q = \infty$

- The ∞-distance from X to $\mathcal{P}$ is defined by

$$d^{(\infty)}(X, \mathcal{P}) := \max_{\mathbf{x}_j \in X} \|\mathbf{x}_j - \Pi_{\mathcal{P}} \mathbf{x}_j\|_2.$$

- The best fit for $X \subset \mathbb{R}^N$ by an n-dimensional subspace is

$$d_n^{(\infty)}\left(X, \mathbb{R}^N\right) := \inf_{\mathcal{P} \in \Gamma_n(\mathbb{R}^N)} d^{(\infty)}(X, \mathcal{P}) = \inf_{\mathcal{P} \in \Gamma_n(\mathbb{R}^N)} \max_{\mathbf{x}_j \in X} \|\mathbf{x}_j - \Pi_{\mathcal{P}} \mathbf{x}_j\|_2.$$

- Essentially the (Euclidean) Kolmogorov n-width of X

The Case $q = \infty$

- The ∞-distance from X to $\mathcal{P}$ is defined by

$$d^{(\infty)}(X, \mathcal{P}) := \max_{\mathbf{x}_j \in X} \|\mathbf{x}_j - \Pi_{\mathcal{P}} \mathbf{x}_j\|_2.$$

- The best fit for $X \subset \mathbb{R}^N$ by an n-dimensional subspace is

$$d_n^{(\infty)}\left(X, \mathbb{R}^N\right) := \inf_{\mathcal{P} \in \Gamma_n(\mathbb{R}^N)} d^{(\infty)}(X, \mathcal{P}) = \inf_{\mathcal{P} \in \Gamma_n(\mathbb{R}^N)} \max_{\mathbf{x}_j \in X} \|\mathbf{x}_j - \Pi_{\mathcal{P}} \mathbf{x}_j\|_2.$$

- Essentially the (Euclidean) Kolmogorov n-width of X

Our Goal When $q = \infty$

Compute an n-dimensional subspace, $\mathcal{A} \in \Gamma_n\left(\mathbb{R}^N\right)$, with

$$d^{(\infty)}(X, \mathcal{A}) \approx d_n^{(\infty)}\left(X, \mathbb{R}^N\right) := \inf_{\mathcal{P} \in \Gamma_n(\mathcal{S})} \max_{\mathbf{x}_j \in X} \|\mathbf{x}_j - \Pi_{\mathcal{P}} \mathbf{x}_j\|_2.$$

Outline

Previous Work: Slow and accurate

The Problem

Compute an n-dimensional subspace, $\mathcal{A} \in \Gamma_n\left(\mathbb{R}^N\right)$, with

$$d^{(q)}(X, \mathcal{A}) \approx d_n^{(q)}\left(X, \mathbb{R}^N\right) := \inf_{\mathcal{P} \in \Gamma_n(\mathcal{S})} \|\mathbf{e}_{\mathcal{P}}\|_q.$$

- $\underline{2 < q < \infty}$ [Deshpande et al. (2011)]:
  Approximate by relaxing to a convex program, and then "rounding"
  $\longrightarrow O\left(M^{>2}N^{>2}\right)$-flops find $\mathcal{A} \in \Gamma_n\left(\mathbb{R}^N\right)$ with
  $d^{(q)}(X, \mathcal{A}) \leq C \cdot d_n^{(q)}\left(X, \mathbb{R}^N\right)$ w.h.p.
- $\underline{q = \infty}$ [Varadarajan et al. (2007)]
  Relax to a semidefinite program, then "round" its solution
  $\longrightarrow O\left(M^{>2}N^{>2}\right)$-flops find $\mathcal{A} \in \Gamma_n\left(\mathbb{R}^N\right)$ with
  $d^{(\infty)}(X, \mathcal{A}) \leq C\sqrt{\log M} \cdot d_n^{(\infty)}\left(X, \mathbb{R}^N\right)$ w.h.p.

Previous Work: Fast for Small Subspace Dimensions

The Problem

Compute an n-dimensional subspace, $\mathcal{A} \in \Gamma_n\left(\mathbb{R}^N\right)$, with

$$d^{(q)}(X, \mathcal{A}) \approx d_n^{(q)}\left(X, \mathbb{R}^N\right) := \inf_{\mathcal{P} \in \Gamma_n(\mathcal{S})} \|\mathbf{e}_{\mathcal{P}}\|_q.$$

- $\underline{q = \infty}$ [Agarwal et al. (2005)]
  Can compute $\mathcal{A} \in \Gamma_n\left(\mathbb{R}^N\right)$ that has

  $$d^{(\infty)}(X, \mathcal{A}) \leq C\sqrt{\log M} \cdot d_n^{(\infty)}\left(X, \mathbb{R}^N\right)$$

  w.h.p. Uses $MN \cdot 2^{\tilde{O}(n)}$-flops. Based on finding "Core sets".
- $\tilde{O}$ denotes that log-factors are dropped in the O-notation.

New Result: Approximate Solutions for $q = \infty$

Previous Results

- <u>M small:</u>
  $d^{(\infty)}(X, \mathcal{A}) \leq C\sqrt{\log M} \cdot d_n^{(\infty)}\left(X, \mathbb{R}^N\right)$ using $O(M^{>2}N^{>2})$-flops.
- <u>M Large, n Small:</u>
  $d^{(\infty)}(X, \mathcal{A}) \leq C\sqrt{\log M} \cdot d_n^{(\infty)}\left(X, \mathbb{R}^N\right)$ using $O(MN) \cdot 2^{\tilde{O}(n)}$-flops.

Theorem (Iwen, FK (2013))

Let $P = \{\mathbf{p}_1, \ldots, \mathbf{p}_M\} \subset \mathbb{R}^N$ be symmetric, and $n \in \{1, \ldots, N\}$. Then, one can calculate an $\mathcal{A} \in \Gamma_n\left(\mathbb{R}^N\right)$ with

$$d^{(\infty)}\left(P, \mathcal{A}\right) \leq C\sqrt{n \cdot \log M} \cdot d_n^{(\infty)}\left(P, \mathbb{R}^N\right)$$

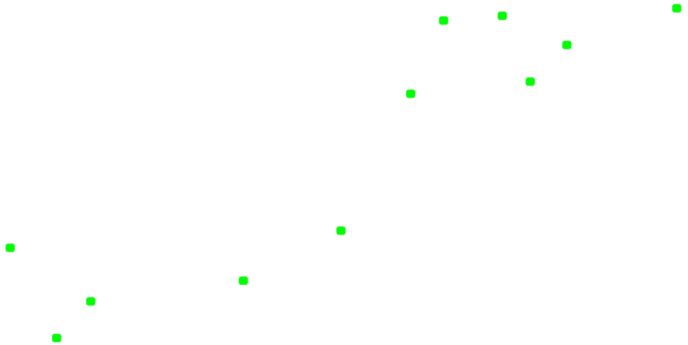
in $O\left(MN^2 + Mn^2 \cdot \log^2 M \cdot \log(n \log M)\right)$-time.

Outline

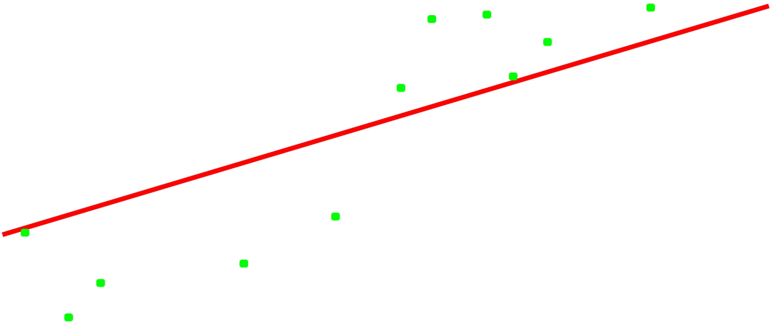
Greedy Least squares

- Consider the n-dimensional least squares approximation $\mathcal{A}_1$ of P. Is it a good approximation in the sense of $d^{(q)}$?
- Compared to $d^{(2)}$, points with large distance from $\mathcal{A}_1$ have a larger impact in $d^{(q)}$, points with a small distance have less of an impact.
- $\mathcal{A}_1$ is a good $d^{(q)}$-approximation to the 90% closest points.
- Remove those 90% points, find the n-dimensional least-squares approximation $\mathcal{A}_2$ of the remaining points. Again this is a good $d^{(q)}$-approximation to 90% of the points.
- Iterate, $O(\log M)$ iterations cover the whole set P.

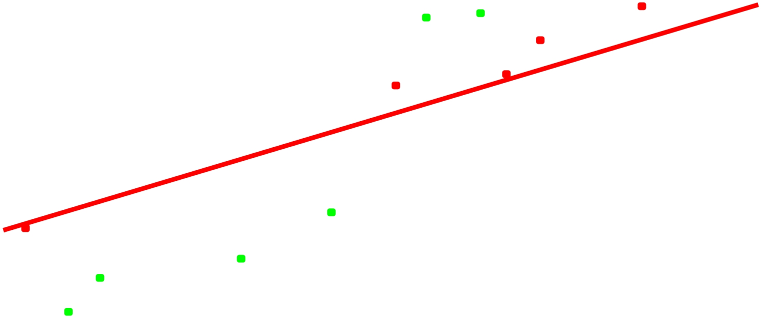
Example



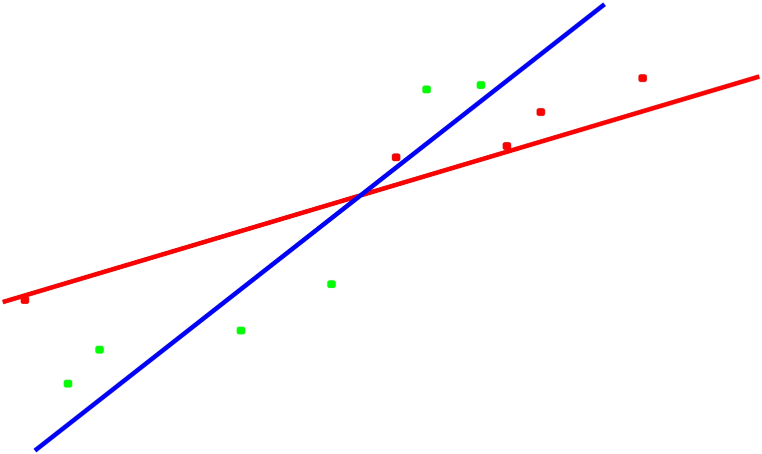
Example



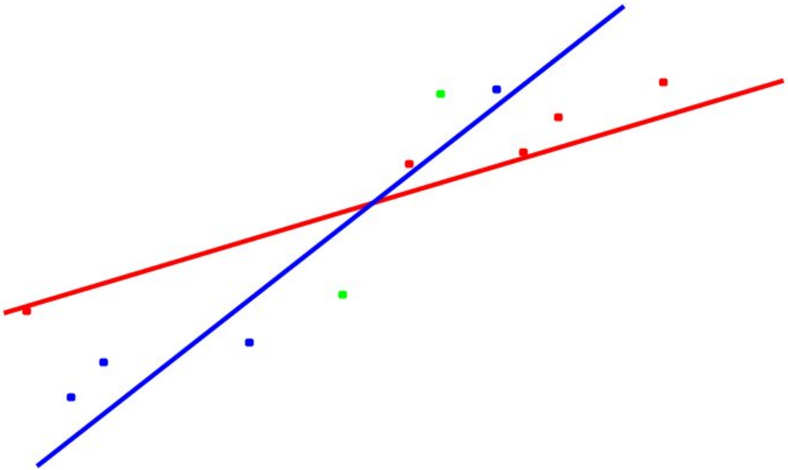
Example



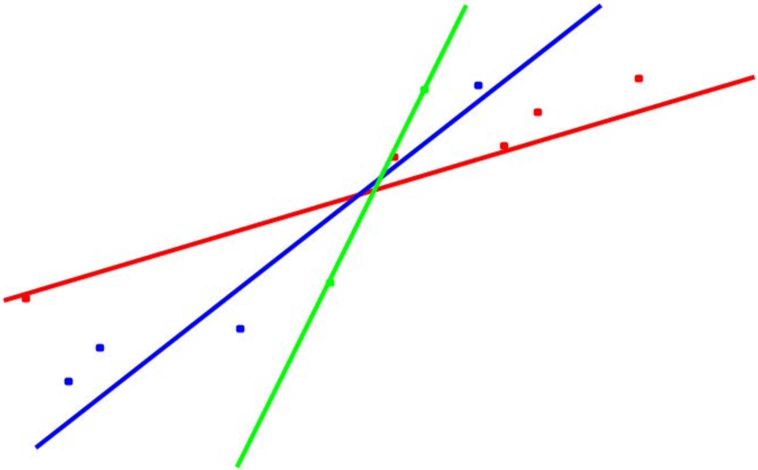
Example



Example



Example



Reducing the problem dimension

- $\mathcal{A} := \bigcup_{\ell=1}^{\log M} \mathcal{A}_\ell$ is an $n \log M$-dimensional subspace that provides a good approximation in the sense of $d^{(q)}$.
- $$d_n^{(p)}(P, \mathcal{S}) \le d^{(p)}(P, \mathcal{S}) + d_n^{(p)}\left(P, \mathbb{R}^N\right). \tag{1}$$

  implies that $\mathcal{A}$ contains a near-optimal n-dimensional subspace approximation.

New Results: Dimensionality Reduction

- Previous Result [Deshpande and Varadarajan (2007)]:
  Finds $O\left((n+1/\epsilon)^{q+3}\right)$-dimensional subspace, $\mathcal{S} \subset \mathbb{R}^N$, with

$$d_n^{(q)}(X, \mathcal{S}) \leq (1+\epsilon) \cdot d_n^{(q)}\left(X, \mathbb{R}^N\right)$$

  w.h.p.. Uses $O(MN(n+1/\epsilon)^{q+3})$-flops for $1 \leq q < \infty$, $\epsilon > 0$.

Theorem (Iwen, FK (2013))

For $P = \{\mathbf{p}_1, \dots, \mathbf{p}_M\} \subset \mathbb{R}^N$ symmetric, one can find a subspace $\mathcal{S} \subset \mathbb{R}^N$ with $\dim \mathcal{S} = O(n \cdot \log M)$ in $O\left(MN^2 + N \cdot n^2 \log^2 M\right)$-time such that

$$d_n^{(p)}(P, \mathcal{S}) \leq \left(1 + C(\log m)^{1/p}\right) \cdot d_n^{(p)}\left(P, \mathbb{R}^N\right)$$

and, for $p = \infty$,

$$d_n^{(\infty)}(P, \mathcal{S}) \leq C \cdot d_n^{(\infty)}\left(P, \mathbb{R}^N\right).$$

Increasing Algorithm Efficiencies

- Apply the dimension reduction technique to find a near-equivalent low-dimensional subproblem
- Solve the low-dimensional problem using existing algorithms
- The problem is smaller size, thus faster to solve.
- For $p = \infty$, use new recovery result based on John's ellipsoid:
  - Accuracy loss of $O(\sqrt{N})$ not competitive in full-dimensional space
  - In reduced dimension, only loss of $O(\sqrt{n \log M})$.

Outline

Summary and Outlook

- Dimension reduction based on greedy least squares
- New suboptimal algorithm using John's ellipsoid
- Together yield fast algorithm for subspace approximation

- log-factor does not seem necessary if $p \to 2$. Can it be removed?

- Preprint available on arxiv 1312.1413