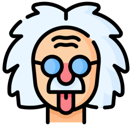
Metric learning from lazy, opinionated crowds

i.e., from limited pairwise preference comparisons

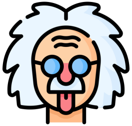
Geelon So (UCSD), agso@ucsd.edu
Signals, Information, and Algorithms Lab @ MIT — March 7, 2024

An opinionated member of society



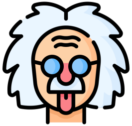
An opinionated member of society

I prefer Blade Runner over Godzilla.



An opinionated member of society

I prefer Blade Runner over Godzilla.



For it is more similar to my favorite movie The Matrix.

Metric learning from preferences

Question:
Suppose a lot of people on the internet tell us these sorts of pairwise movie rankings.

Metric learning from preferences

Question:
Suppose a lot of people on the internet tell us these sorts of pairwise movie rankings.

- Can we learn a metric that captures the similarity of movies in general?

Background

Metric learning: raison d'être

Distance-based algorithms

- nearest neighbor methods
- margin-based classification
- information retrieval
- clustering
- etc.

Metric learning: raison d'être

Distance-based algorithms

- nearest neighbor methods
- margin-based classification
- information retrieval
- clustering
- etc.

- The behavior/performance are often sensitive to the choice of distance.

Metric learning: raison d'être

Distance-based algorithms

- nearest neighbor methods
- margin-based classification
- information retrieval
- clustering
- etc.

- The behavior/performance are often sensitive to the choice of distance.
- Good metrics (e.g. for visual similarity) are hard to construct by hand.

Metric learning: raison d'être

Distance-based algorithms

- nearest neighbor methods
- margin-based classification
- information retrieval
- clustering
- etc.

- The behavior/performance are often sensitive to the choice of distance.
- Good metrics (e.g. for visual similarity) are hard to construct by hand.

Goal of metric learning:
Automatically learn a good metric for these downstream tasks

Metric learning: raison d'être

Distance-based algorithms

- nearest neighbor methods
- margin-based classification
- information retrieval
- clustering
- etc.

- The behavior/performance are often sensitive to the choice of distance.
- Good metrics (e.g. for visual similarity) are hard to construct by hand.

Goal of metric learning:
Automatically learn a good metric for these downstream tasks

- esp. metrics aligning with human values, perception, and preferences.

The alignment problem

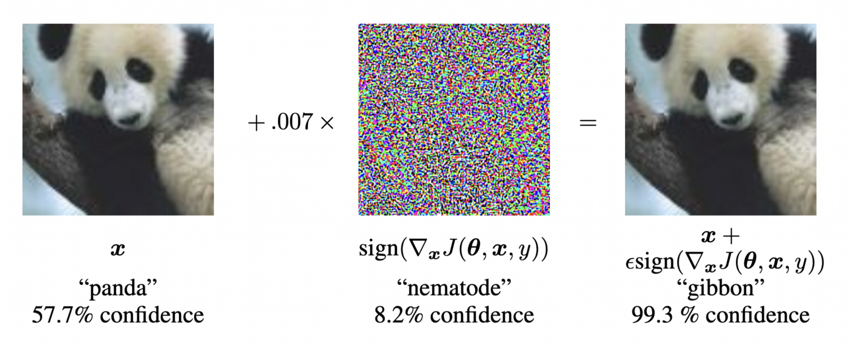


Figure 1: These two images are visually indistinguishable to a human, but very well-separated under the Euclidean distance (Goodfellow et al., 2014).

Related work: metric learning from triplet comparisons

Metric learning from triplet comparisons

Related work: metric learning from triplet comparisons

Metric learning from triplet comparisons

- Goal: given a set of items $\mathcal{X}$, learn a metric ρ over the items.

Related work: metric learning from triplet comparisons

Metric learning from triplet comparisons

- Goal: given a set of items $\mathcal{X}$, learn a metric ρ over the items.
- Feedback: "A is more similar to B than to C."

Related work: metric learning from triplet comparisons

Metric learning from triplet comparisons

- Goal: given a set of items $\mathcal{X}$, learn a metric ρ over the items.
- Feedback: "A is more similar to B than to C."

Schultz and Joachims (2003), Verma and Branson (2015), Mason et al. (2017)

Related work: metric learning from triplet comparisons

Metric learning from triplet comparisons

- Goal: given a set of items $\mathcal{X}$, learn a metric ρ over the items.
- Feedback: "A is more similar to B than to C."

Schultz and Joachims (2003), Verma and Branson (2015), Mason et al. (2017)

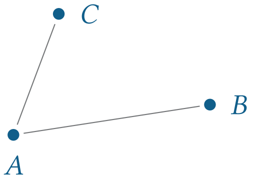


Figure 2: Triplet feedback: "B is closer to A than C is."

Related work: simultaneous metric and preference learning

Problem: triplet comparisons are not a common form of feedback.

Related work: simultaneous metric and preference learning

Problem: triplet comparisons are not a common form of feedback.

- Can be costly to obtain.

Related work: simultaneous metric and preference learning

Problem: triplet comparisons are not a common form of feedback.

- Can be costly to obtain.

Metric learning from pairwise preference comparisons

Related work: simultaneous metric and preference learning

Problem: triplet comparisons are not a common form of feedback.

- Can be costly to obtain.

Metric learning from pairwise preference comparisons

- Assume a user has an *ideal item* A and prefers items more similar to A.

Related work: simultaneous metric and preference learning

Problem: triplet comparisons are not a common form of feedback.

- Can be costly to obtain.

Metric learning from pairwise preference comparisons

- Assume a user has an *ideal item* A and prefers items more similar to A.
- Goal: given a set of items $\mathcal{X}$, learn ideal item A and metric ρ.

Related work: simultaneous metric and preference learning

Problem: triplet comparisons are not a common form of feedback.

- Can be costly to obtain.

Metric learning from pairwise preference comparisons

- Assume a user has an *ideal item* A and prefers items more similar to A.
- Goal: given a set of items $\mathcal{X}$, learn ideal item A and metric ρ.
- Feedback: "I prefer B over C."

Related work: simultaneous metric and preference learning

Problem: triplet comparisons are not a common form of feedback.

- Can be costly to obtain.

Metric learning from pairwise preference comparisons

- Assume a user has an *ideal item* A and prefers items more similar to A.
- Goal: given a set of items $\mathcal{X}$, learn ideal item A and metric ρ.
- Feedback: "I prefer B over C."
    - triplet comparison with a latent comparator (i.e. A is not observed)

Related work: simultaneous metric and preference learning

Problem: triplet comparisons are not a common form of feedback.

- Can be costly to obtain.

Metric learning from pairwise preference comparisons

- Assume a user has an *ideal item* A and prefers items more similar to A.
- Goal: given a set of items $\mathcal{X}$, learn ideal item A and metric ρ.
- Feedback: "I prefer B over C."
    - triplet comparison with a latent comparator (i.e. A is not observed)
    - much more prevalent form of feedback than triplet comparisons

Related work: simultaneous metric and preference learning

Problem: triplet comparisons are not a common form of feedback.

- Can be costly to obtain.

Metric learning from pairwise preference comparisons

- Assume a user has an *ideal item* A and prefers items more similar to A.
- Goal: given a set of items $\mathcal{X}$, learn ideal item A and metric ρ.
- Feedback: "I prefer B over C."
  - triplet comparison with a latent comparator (i.e. A is not observed)
  - much more prevalent form of feedback than triplet comparisons

Xu and Davenport (2020) and Canal et al. (2022)

Preview of results

The hidden cost of weaker feedback

Preview of results

The hidden cost of weaker feedback

- Weaker feedback may make data easier/cheaper to collect.

Preview of results

The hidden cost of weaker feedback

- Weaker feedback may make data easier/cheaper to collect.

$$\text{triplet: } (A; B, C) \qquad \text{vs.} \qquad \text{binary: } (\,?\,; B, C).$$

Preview of results

The hidden cost of weaker feedback

- Weaker feedback may make data easier/cheaper to collect.

$$\text{triplet: } (A; B, C) \qquad \text{vs.} \qquad \text{binary: } (\,?\,; B, C).$$

- However, this can introduce new fundamental regimes where data is unusable.

Preview of results

The hidden cost of weaker feedback

- Weaker feedback may make data easier/cheaper to collect.

$$\text{triplet: } (A; B, C) \qquad \text{vs.} \qquad \text{binary: } (\,?\,; B, C).$$

- However, this can introduce new fundamental regimes where data is unusable.

Paying for weaker feedback

Preview of results

The hidden cost of weaker feedback

- Weaker feedback may make data easier/cheaper to collect.

$$\text{triplet: } (A; B, C) \qquad \text{vs.} \qquad \text{binary: } (\,?\,; B, C).$$

- However, this can introduce new fundamental regimes where data is unusable.

Paying for weaker feedback

- It may be possible to overcome new limits with structural assumptions.

Preview of results

The hidden cost of weaker feedback

- Weaker feedback may make data easier/cheaper to collect.

$$\text{triplet: } (A; B, C) \qquad \text{vs.} \qquad \text{binary: } (\,?\,; B, C).$$

- However, this can introduce new fundamental regimes where data is unusable.

Paying for weaker feedback

- It may be possible to overcome new limits with structural assumptions.
- These assumptions may be realistic (e.g. approximate low rank structures).

Our work in context

Let $\mathbb{R}^d$ be equipped with an unknown Mahalanobis distance ρ_M.

Our work in context

Let $\mathbb{R}^d$ be equipped with an unknown Mahalanobis distance ρ_M.

What is known:

- we can learn ρ_M using $\Theta(d^2)$ measurements from a single user

Our work in context

Let $\mathbb{R}^d$ be equipped with an unknown Mahalanobis distance ρ_M.

What is known:

- we can learn ρ_M using $\Theta(d^2)$ measurements from a single user
- or, using $\Theta(d)$ measurements from $\Omega(d)$ users

Our work in context

Let $\mathbb{R}^d$ be equipped with an unknown Mahalanobis distance ρ_M.

What is known:

- we can learn ρ_M using $\Theta(d^2)$ measurements from a single user
- or, using $\Theta(d)$ measurements from $\Omega(d)$ users
  - $\Theta(d)$ is necessary for simultaneous recovery of metric and ideal points

Our work in context

Let $\mathbb{R}^d$ be equipped with an unknown Mahalanobis distance ρ_M.

What is known:

- we can learn ρ_M using $\Theta(d^2)$ measurements from a single user
- or, using $\Theta(d)$ measurements from $\Omega(d)$ users
  - $\Theta(d)$ is necessary for simultaneous recovery of metric and ideal points

Limitations:

- modern representations of data can be extremely high dimensional

Our work in context

Let $\mathbb{R}^d$ be equipped with an unknown Mahalanobis distance ρ_M.

What is known:

- we can learn ρ_M using $\Theta(d^2)$ measurements from a single user
- or, using $\Theta(d)$ measurements from $\Omega(d)$ users
  - $\Theta(d)$ is necessary for simultaneous recovery of metric and ideal points

Limitations:

- modern representations of data can be extremely high dimensional
- it could be infeasible to obtain $\Theta(d)$ measurements per user

Our work in context

Let $\mathbb{R}^d$ be equipped with an unknown Mahalanobis distance ρ_M.

What is known:

- we can learn ρ_M using $\Theta(d^2)$ measurements from a single user
- or, using $\Theta(d)$ measurements from $\Omega(d)$ users
  - $\Theta(d)$ is necessary for simultaneous recovery of metric and ideal points

Limitations:

- modern representations of data can be extremely high dimensional
- it could be infeasible to obtain $\Theta(d)$ measurements per user

Our work: Let's just give up on trying to learn the ideal points. We ask:

Can we recover the metric using $m \ll d$ measurements per user?

Preliminaries

Formal setting

Representation space

Let $\mathcal{X}$ be a set of items embedded into $\mathbb{R}^d$.

Formal setting

Representation space

Let $\mathcal{X}$ be a set of items embedded into $\mathbb{R}^d$.

- Assume $\mathbb{R}^d$ is equipped with an unknown Mahalanobis distance ρ_M.

Formal setting

Representation space

Let $\mathcal{X}$ be a set of items embedded into $\mathbb{R}^d$.

- Assume $\mathbb{R}^d$ is equipped with an unknown Mahalanobis distance ρ_M.

Feedback model

Users provide preference feedback under the *ideal point model* (Coombs, 1950).

Formal setting

Representation space

Let $\mathcal{X}$ be a set of items embedded into $\mathbb{R}^d$.

- Assume $\mathbb{R}^d$ is equipped with an unknown Mahalanobis distance ρ_M.

Feedback model

Users provide preference feedback under the *ideal point model* (Coombs, 1950).

- Assume each user has a ideal point $u \in \mathbb{R}^d$ (which we cannot observe).

Formal setting

Representation space

Let $\mathcal{X}$ be a set of items embedded into $\mathbb{R}^d$.

- Assume $\mathbb{R}^d$ is equipped with an unknown Mahalanobis distance ρ_M.

Feedback model

Users provide preference feedback under the *ideal point model* (Coombs, 1950).

- Assume each user has a ideal point $u \in \mathbb{R}^d$ (which we cannot observe).
- The user prefers an item x over x' whenever:

$$\rho_M(u, x) < \rho(u, x').$$

Formal setting

Representation space

Let $\mathcal{X}$ be a set of items embedded into $\mathbb{R}^d$.

- Assume $\mathbb{R}^d$ is equipped with an unknown Mahalanobis distance ρ_M.

Feedback model

Users provide preference feedback under the *ideal point model* (Coombs, 1950).

- Assume each user has a ideal point $u \in \mathbb{R}^d$ (which we cannot observe).
- The user prefers an item x over x' whenever:

$$\rho_M(u, x) < \rho(u, x').$$

- We receive measurements from users of the form:

$$(x, x', y) \qquad \text{where} \quad y = \mathbf{1}\{\rho_M(u, x) < \rho_M(u, x')\}.$$

Mahalanobis distances

A *Mahalanobis distance* ρ_M on $\mathbb{R}^d$ is a metric of the form:

$$\rho_M(x, x') = \sqrt{(x - x')^\top M (x - x')} = \|x - x'\|_M,$$

Mahalanobis distances

A *Mahalanobis distance* ρ_M on $\mathbb{R}^d$ is a metric of the form:

$$\rho_M(x, x') = \sqrt{(x - x')^\top M (x - x')} = \|x - x'\|_M,$$

where $M \in \mathrm{Sym}(\mathbb{R}^d)$ is a positive-definite (symmetric) matrix.

Mahalanobis distances

A *Mahalanobis distance* ρ_M on $\mathbb{R}^d$ is a metric of the form:

$$\rho_M(x, x') = \sqrt{(x - x')^\top M (x - x')} = \|x - x'\|_M,$$

where $M \in \mathrm{Sym}(\mathbb{R}^d)$ is a positive-definite (symmetric) matrix.

Geometric interpretation

- $M = A^\top A$ for some $A \in \mathbb{R}^{d \times d}$ since $M \succ 0$.

Mahalanobis distances

A *Mahalanobis distance* ρ_M on $\mathbb{R}^d$ is a metric of the form:

$$\rho_M(x, x') = \sqrt{(x - x')^\top M(x - x')} = \|x - x'\|_M,$$

where $M \in \mathrm{Sym}(\mathbb{R}^d)$ is a positive-definite (symmetric) matrix.

Geometric interpretation

- $M = A^\top A$ for some $A \in \mathbb{R}^{d\times d}$ since $M \succ 0$.
- Let $\Phi(x) = Ax$ be a new (linear) representation. Then:

$$\rho_M(x, x') = \|\Phi(x) - \Phi(x')\|_2.$$

A mathematical simplification

A user with ideal point $u \in \mathbb{R}^d$ can give two types of feedback:

A mathematical simplification

A user with ideal point $u \in \mathbb{R}^d$ can give two types of feedback:

- **Continuous responses:** measurements of the form (x, x', ψ), where:

A mathematical simplification

A user with ideal point $u \in \mathbb{R}^d$ can give two types of feedback:

- **Continuous responses:** measurements of the form (x, x', ψ), where:

$$\psi \equiv \psi_M(x, x'; u) = \|u - x\|_M^2 - \|u - x'\|_M^2.$$

A mathematical simplification

A user with ideal point $u \in \mathbb{R}^d$ can give two types of feedback:

- **Continuous responses:** measurements of the form (x, x', ψ), where:

$$\psi \equiv \psi_M(x, x'; u) = \|u - x\|_M^2 - \|u - x'\|_M^2.$$

  - Not realistic form of feedback, but mathematically easy to work with.

A mathematical simplification

A user with ideal point $u \in \mathbb{R}^d$ can give two types of feedback:

- **Continuous responses:** measurements of the form (x, x', ψ), where:

$$\psi \equiv \psi_M(x, x'; u) = \|u - x\|_M^2 - \|u - x'\|_M^2.$$

  - Not realistic form of feedback, but mathematically easy to work with.

- **Binary responses:** measurements of the form (x, x', y) where:

A mathematical simplification

A user with ideal point $u \in \mathbb{R}^d$ can give two types of feedback:

- **Continuous responses:** measurements of the form (x, x', ψ), where:

$$\psi \equiv \psi_M(x, x'; u) = \|u - x\|_M^2 - \|u - x'\|_M^2.$$

  - Not realistic form of feedback, but mathematically easy to work with.

- **Binary responses:** measurements of the form (x, x', y) where:

$$y = \mathbf{1}\{\psi < 0\}.$$

A mathematical simplification

A user with ideal point $u \in \mathbb{R}^d$ can give two types of feedback:

- **Continuous responses:** measurements of the form (x, x', ψ), where:

$$\psi \equiv \psi_M(x, x'; u) = \|u - x\|_M^2 - \|u - x'\|_M^2.$$

  - Not realistic form of feedback, but mathematically easy to work with.

- **Binary responses:** measurements of the form (x, x', y) where:

$$y = \mathbf{1}\{\psi < 0\}.$$

  - Later, we consider the setting where labels are binary and noisy.

A linear reparametrization (Canal et al., 2022)

Let $x, x' \in \mathbb{R}^d$ be two items. If a user has ideal point $u \in \mathbb{R}^d$, then:

$$\psi_M(x, x'; u) = \underbrace{\langle xx^\top - x'x'^\top, M\rangle}_{(1)} + \underbrace{\langle x - x', v\rangle}_{(2)}, \qquad \text{where } \underbrace{v = -2Mu}_{(3)}.$$

A linear reparametrization (Canal et al., 2022)

Let $x, x' \in \mathbb{R}^d$ be two items. If a user has ideal point $u \in \mathbb{R}^d$, then:

$$\psi_M(x, x'; u) = \underbrace{\left\langle xx^\top - x'x'^\top, M \right\rangle}_{(1)} + \underbrace{\left\langle x - x', v \right\rangle}_{(2)}, \qquad \text{where } \underbrace{v = -2Mu}_{(3)}.$$

1. $\left\langle xx^\top - x'x'^\top, M \right\rangle$ is the trace inner product on $\mathrm{Sym}(\mathbb{R}^d)$, where $\langle A, B\rangle = \mathrm{tr}(AB)$.

A linear reparametrization (Canal et al., 2022)

Let $x, x' \in \mathbb{R}^d$ be two items. If a user has ideal point $u \in \mathbb{R}^d$, then:

$$\psi_M(x, x'; u) = \underbrace{\left\langle xx^\top - x'x'^\top, M \right\rangle}_{(1)} + \underbrace{\left\langle x - x', v \right\rangle}_{(2)}, \qquad \text{where } \underbrace{v = -2Mu}_{(3)}.$$

1. $\left\langle xx^\top - x'x'^\top, M \right\rangle$ is the trace inner product on $\mathrm{Sym}(\mathbb{R}^d)$, where $\langle A, B \rangle = \mathrm{tr}(AB)$.
2. $\left\langle x - x', v \right\rangle$ is the standard inner product on $\mathbb{R}^d$.

A linear reparametrization (Canal et al., 2022)

Let $x, x' \in \mathbb{R}^d$ be two items. If a user has ideal point $u \in \mathbb{R}^d$, then:

$$\psi_M(x, x'; u) = \underbrace{\left\langle xx^\top - x'x'^\top, M \right\rangle}_{(1)} + \underbrace{\left\langle x - x', v \right\rangle}_{(2)}, \qquad \text{where } \underbrace{v = -2Mu}_{(3)}.$$

1. $\left\langle xx^\top - x'x'^\top, M \right\rangle$ is the trace inner product on $\mathrm{Sym}(\mathbb{R}^d)$, where $\langle A, B \rangle = \mathrm{tr}(AB)$.
2. $\left\langle x - x', v \right\rangle$ is the standard inner product on $\mathbb{R}^d$.
3. The reparametrization $v = -2Mu$ is called the user's *pseudo-ideal point*.

A linear reparametrization (Canal et al., 2022)

Let $x, x' \in \mathbb{R}^d$ be two items. If a user has ideal point $u \in \mathbb{R}^d$, then:

$$\psi_M(x, x'; u) = \underbrace{\left\langle xx^\top - x'x'^\top, M \right\rangle}_{(1)} + \underbrace{\left\langle x - x', v \right\rangle}_{(2)}, \qquad \text{where } \underbrace{v = -2Mu}_{(3)}.$$

1. $\left\langle xx^\top - x'x'^\top, M \right\rangle$ is the trace inner product on $\mathrm{Sym}(\mathbb{R}^d)$, where $\langle A, B \rangle = \mathrm{tr}(AB)$.
2. $\left\langle x - x', v \right\rangle$ is the standard inner product on $\mathbb{R}^d$.
3. The reparametrization $v = -2Mu$ is called the user's *pseudo-ideal point*.

Upshot: Reparametrize (M, u) to (M, v). Then, the following map is linear:

$$(M, v) \mapsto \psi_M(x, x'; u).$$

A linear reparametrization (Canal et al., 2022)

Let $x, x' \in \mathbb{R}^d$ be two items. If a user has ideal point $u \in \mathbb{R}^d$, then:

$$\psi_M(x, x'; u) = \underbrace{\left\langle xx^\top - x'x'^\top, M \right\rangle}_{(1)} + \underbrace{\left\langle x - x', v \right\rangle}_{(2)}, \qquad \text{where } \underbrace{v = -2Mu}_{(3)}.$$

1. $\left\langle xx^\top - x'x'^\top, M \right\rangle$ is the trace inner product on $\mathrm{Sym}(\mathbb{R}^d)$, where $\langle A, B \rangle = \mathrm{tr}(AB)$.
2. $\left\langle x - x', v \right\rangle$ is the standard inner product on $\mathbb{R}^d$.
3. The reparametrization $v = -2Mu$ is called the user's *pseudo-ideal point*.

Upshot: Reparametrize (M, u) to (M, v). Then, the following map is linear:

$$(M, v) \mapsto \psi_M(x, x'; u).$$

Thus, there is a reparametrization under which measurements are linear.

Design matrices

Let $\{(x_{i_0}, x_{i_1})\}_{i=1}^{m}$ be a set of item pairs.

Design matrices

Let $\{(x_{i_0}, x_{i_1})\}_{i=1}^{m}$ be a set of item pairs.

- Define the linear map $D : \mathrm{Sym}(\mathbb{R}^d) \oplus \mathbb{R}^d \to \mathbb{R}^m$:

$$D_i(A, w) = \left\langle x_{i_0} x_{i_0}^\top - x'_{i_1} x'^\top_{i_1}, A \right\rangle + \left\langle x_{i_0} - x'_{i_1}, w \right\rangle.$$

Design matrices

Let $\{(x_{i_0}, x_{i_1})\}_{i=1}^{m}$ be a set of item pairs.

- Define the linear map $D : \mathrm{Sym}(\mathbb{R}^d) \oplus \mathbb{R}^d \to \mathbb{R}^m$:

$$D_i(A, w) = \left\langle x_{i_0} x_{i_0}^{\top} - x'_{i_1} x'^{\top}_{i_1}, A \right\rangle + \left\langle x_{i_0} - x'_{i_1}, w \right\rangle.$$

- We call D the design matrix induced by the item pairs.

Metric learning from continuous responses, single user case

Suppose a user provides us with measurements $\left\{(x_{i_0}, x_{i_1}, \psi_i)\right\}_{i=1}^{m}$, where:

$$\psi_i = \psi_M(x_{i_0}, x_{i_1}; u).$$

Metric learning from continuous responses, single user case

Suppose a user provides us with measurements $\left\{(x_{i_0}, x_{i_1}, \psi_i)\right\}_{i=1}^m$, where:

$$\psi_i = \psi_M(x_{i_0}, x_{i_1}; u).$$

- We can recover (M, u) by solving the linear system of equations:

$$D_i(A, w) = \psi_i.$$

Metric learning from continuous responses, single user case

Suppose a user provides us with measurements $\left\{(x_{i_0}, x_{i_1}, \psi_i)\right\}_{i=1}^{m}$, where:

$$\psi_i = \psi_M(x_{i_0}, x_{i_1}; u).$$

- We can recover (M, u) by solving the linear system of equations:

$$D_i(A, w) = \psi_i.$$

- The pair (M, v) of the Mahalanobis matrix and pseudo-ideal point is a solution.

Metric learning from continuous responses, single user case

Suppose a user provides us with measurements $\{(x_{i_0}, x_{i_1}, \psi_i)\}_{i=1}^m$, where:

$$\psi_i = \psi_M(x_{i_0}, x_{i_1}; u).$$

- We can recover (M, u) by solving the linear system of equations:

$$D_i(A, w) = \psi_i.$$

- The pair (M, v) of the Mahalanobis matrix and pseudo-ideal point is a solution.
  - The ideal point can be computed from the pseudo-ideal point since $v = -2Mu$.

Metric learning from continuous responses, single user case

Suppose a user provides us with measurements $\{(x_{i_0}, x_{i_1}, \psi_i)\}_{i=1}^m$, where:

$$\psi_i = \psi_M(x_{i_0}, x_{i_1}; u).$$

- We can recover (M, u) by solving the linear system of equations:

$$D_i(A, w) = \psi_i.$$

- The pair (M, v) of the Mahalanobis matrix and pseudo-ideal point is a solution.
  - The ideal point can be computed from the pseudo-ideal point since $v = -2Mu$.
- To recover the metric and ideal point, $m = \frac{d(d+1)}{2} + d$ measurements is necessary.

Metric learning from continuous responses, multiple user case

Generalization to multiple users

- Suppose K users provide us with measurements (on distinct pairs of items).

Metric learning from continuous responses, multiple user case

Generalization to multiple users

- Suppose K users provide us with measurements (on distinct pairs of items).
- Recover the metric and all ideal points by solving a linear system of equations:

$$\mathbf{D}(A, w_1, \ldots, w_K) = \Psi,$$

where $A \in \mathrm{Sym}(\mathbb{R}^d)$ and each $w_k \in \mathbb{R}^d$.

Metric learning from continuous responses, multiple user case

Generalization to multiple users

- Suppose K users provide us with measurements (on distinct pairs of items).
- Recover the metric and all ideal points by solving a linear system of equations:

$$\mathbf{D}(A, w_1, \ldots, w_K) = \Psi,$$

where $A \in \mathrm{Sym}(\mathbb{R}^d)$ and each $w_k \in \mathbb{R}^d$.

Known results

- At least d measurements from each user is necessary to recover ideal points.

Metric learning from continuous responses, multiple user case

Generalization to multiple users

- Suppose K users provide us with measurements (on distinct pairs of items).
- Recover the metric and all ideal points by solving a linear system of equations:

$$\mathbf{D}(A, w_1, \ldots, w_K) = \Psi,$$

where $A \in \mathrm{Sym}(\mathbb{R}^d)$ and each $w_k \in \mathbb{R}^d$.

Known results

- At least d measurements from each user is necessary to recover ideal points.
- Recovering $(M, u_1, \ldots, u_K)$ is possible:
  - from $2d$ measurements per user if $K = \Omega(d)$

Metric learning from continuous responses, multiple user case

Generalization to multiple users

- Suppose K users provide us with measurements (on distinct pairs of items).
- Recover the metric and all ideal points by solving a linear system of equations:

$$\mathbf{D}(A, w_1, \ldots, w_K) = \Psi,$$

where $A \in \mathrm{Sym}(\mathbb{R}^d)$ and each $w_k \in \mathbb{R}^d$.

Known results

- At least d measurements from each user is necessary to recover ideal points.
- Recovering $(M, u_1, \ldots, u_K)$ is possible:
  - from $2d$ measurements per user if $K = \Omega(d)$
  - from $d + 1$ measurements per user if $K = \Omega(d^2)$.

Basic question: metric learning from lazy crowds

We ask: Suppose we can obtain very few $m \ll d$ measurements per user. Though ideal points can no longer be learned, is metric learning still possible?

Basic question: metric learning from lazy crowds

We ask: Suppose we can obtain very few $m \ll d$ measurements per user. Though ideal points can no longer be learned, is metric learning still possible?

High-level structure:

- Matrix sensing problem: learn the parameters of $M \in \mathrm{Sym}(\mathbb{R}^d)$.

Basic question: metric learning from lazy crowds

We ask: Suppose we can obtain very few $m \ll d$ measurements per user. Though ideal points can no longer be learned, is metric learning still possible?

High-level structure:

- Matrix sensing problem: learn the parameters of $M \in \mathrm{Sym}(\mathbb{R}^d)$.
- We have access to a large pool of sensors (users + items).

Basic question: metric learning from lazy crowds

We ask: Suppose we can obtain very few $m \ll d$ measurements per user. Though ideal points can no longer be learned, is metric learning still possible?

High-level structure:

- Matrix sensing problem: learn the parameters of $M \in \mathrm{Sym}(\mathbb{R}^d)$.
- We have access to a large pool of sensors (users + items).
  - Part of the measurement parameters are latent (unknown ideal points).

Basic question: metric learning from lazy crowds

We ask: Suppose we can obtain very few $m \ll d$ measurements per user. Though ideal points can no longer be learned, is metric learning still possible?

High-level structure:

- Matrix sensing problem: learn the parameters of $M \in \mathrm{Sym}(\mathbb{R}^d)$.
- We have access to a large pool of sensors (users + items).
  - Part of the measurement parameters are latent (unknown ideal points).
  - Previous work: learn latent parameters along with M.

Basic question: metric learning from lazy crowds

We ask: Suppose we can obtain very few $m \ll d$ measurements per user. Though ideal points can no longer be learned, is metric learning still possible?

High-level structure:

- Matrix sensing problem: learn the parameters of $M \in \mathrm{Sym}(\mathbb{R}^d)$.
- We have access to a large pool of sensors (users + items).
  - Part of the measurement parameters are latent (unknown ideal points).
  - Previous work: learn latent parameters along with M.
- Our regime: too few measurements per user to learn latent parameters.

An impossibility result

Setting for impossibility result

Setting.

- Let $\mathcal{X} \subset (\mathbb{R}^d, \rho_M)$ be a countable set of items.

Setting for impossibility result

Setting.

- Let $\mathcal{X} \subset (\mathbb{R}^d, \rho_M)$ be a countable set of items.
- Let user $k \in \mathbb{N}$ have pseudo-ideal point v_k.

Setting for impossibility result

Setting.

- Let $\mathcal{X} \subset (\mathbb{R}^d, \rho_M)$ be a countable set of items.
- Let user $k \in \mathbb{N}$ have pseudo-ideal point v_k.
- We ask $m \leq d$ pairwise comparisons per user over items in $\mathcal{X}$.

Setting for impossibility result

Setting.

- Let $\mathcal{X} \subset (\mathbb{R}^d, \rho_M)$ be a countable set of items.
- Let user $k \in \mathbb{N}$ have pseudo-ideal point v_k.
- We ask $m \leq d$ pairwise comparisons per user over items in $\mathcal{X}$.
  - Let $D^{(k)}$ be the design matrix for user k.

Learning latent parameters is necessary

Theorem (Impossibility result)

For (i) almost all sets $\mathcal{X}$*,*

Learning latent parameters is necessary

Theorem (Impossibility result)

For (i) almost all sets $\mathcal{X}$, (ii) any set of designs $D^{(k)}$,

Learning latent parameters is necessary

Theorem (Impossibility result)

For (i) almost all sets $\mathcal{X}$*, (ii) any set of designs* $D^{(k)}$*, and (iii) any* $M' \in \mathrm{Sym}(\mathbb{R}^d)$*,*

Learning latent parameters is necessary

Theorem (Impossibility result)

For (i) almost all sets $\mathcal{X}$, (ii) any set of designs $D^{(k)}$, and (iii) any $M' \in \mathrm{Sym}(\mathbb{R}^d)$, there exists $v'_k \in \mathbb{R}^d$ such that:

$$D^{(k)}(M, v_k) = D^{(k)}(M', v'_k), \qquad \forall k \in \mathbb{N}.$$

Learning latent parameters is necessary

Theorem (Impossibility result)

For (i) almost all sets $\mathcal{X}$, (ii) any set of designs $D^{(k)}$, and (iii) any $M' \in \mathrm{Sym}(\mathbb{R}^d)$, there exists $v_k' \in \mathbb{R}^d$ such that:

$$D^{(k)}(M, v_k) = D^{(k)}(M', v_k'), \qquad \forall k \in \mathbb{N}.$$

- That is, M' is consistent with observed data.

Learning latent parameters is necessary

Theorem (Impossibility result)

For (i) almost all sets $\mathcal{X}$, (ii) any set of designs $D^{(k)}$, and (iii) any $M' \in \mathrm{Sym}(\mathbb{R}^d)$, there exists $v'_k \in \mathbb{R}^d$ such that:

$$D^{(k)}(M, v_k) = D^{(k)}(M', v'_k), \qquad \forall k \in \mathbb{N}.$$

- That is, M' is consistent with observed data.
- Each user introduces enough degrees of freedom to account for all variation in data.

Learning latent parameters is necessary

Theorem (Impossibility result)

For (i) almost all sets $\mathcal{X}$, (ii) any set of designs $D^{(k)}$, and (iii) any $M' \in \mathrm{Sym}(\mathbb{R}^d)$, there exists $v'_k \in \mathbb{R}^d$ such that:

$$D^{(k)}(M, v_k) = D^{(k)}(M', v'_k), \qquad \forall k \in \mathbb{N}.$$

- That is, M' is consistent with observed data.
- Each user introduces enough degrees of freedom to account for all variation in data.
- Not only is recovery impossible, but we learn nothing at all about M.

Which sets do "almost all" item sets refer to?

Theorem (Impossibility result)

When (i) $\mathcal{X}$ has generic pairwise relations, (ii) . . . the impossibility result holds.

Which sets do “almost all” item sets refer to?

Theorem (Impossibility result)

When (i) $\mathcal{X}$ has generic pairwise relations, (ii) . . . the impossibility result holds.

- We introduce a notion of genericity, slightly stronger than *general linear position*.

Which sets do "almost all" item sets refer to?

Theorem (Impossibility result)

When (i) $\mathcal{X}$ has generic pairwise relations, (ii) . . . the impossibility result holds.

- We introduce a notion of genericity, slightly stronger than *general linear position*.
- Almost all finite sets are generic in this sense (w.r.t. Lebesgue measure on $\mathbb{R}^d$).

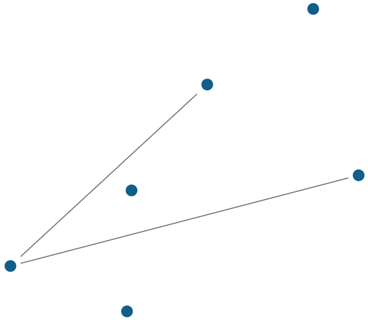
General linear position

Definition

A set $\mathcal{X} \subset \mathbb{R}^d$ is in **general linear position** *if the following is linearly independent:*

$$\{x_i - x_0 : i = 1, \ldots, n\},$$

for any distinct $x_0, x_1, \ldots, x_n \in \mathcal{X}$ and $n \leq d$.



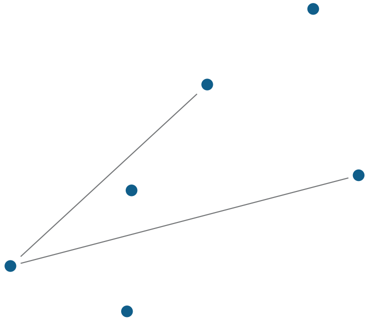
A set of points $\mathcal{X} \subset \mathbb{R}^d$.

General linear position: alternate definition

Definition

A set $\mathcal{X} \subset \mathbb{R}^d$ is in **general linear position** *if for any star graph $G = (V \subset \mathcal{X}, E)$ with $|E| \leq d$, the following is linearly independent:*

$$\{x - x' : (x, x') \in E\}.$$



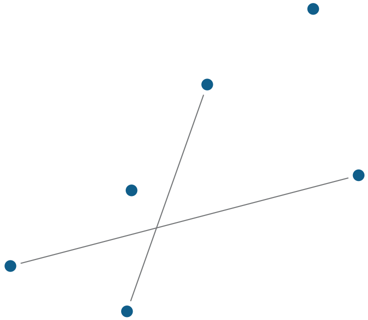
A set of points $\mathcal{X} \subset \mathbb{R}^d$.

Generic pairwise relation

Definition

A set $\mathcal{X} \subset \mathbb{R}^d$ has **generic pairwise relations** *if for any acyclic graph $G = (\mathcal{X}, E)$ with $|E| \leq d$, the following is linearly independent:*

$$\{x - x' : (x, x') \in E\}.$$



A set of points $\mathcal{X} \subset \mathbb{R}^d$.

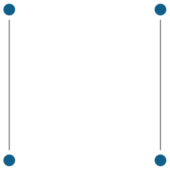
Generic pairwise relations $\Longrightarrow$ general linear position

Proof.

A star graph with at most d edges is an acyclic graph with at most d edges. $\square$

General linear position $\not\Rightarrow$ generic pairwise relations

- ✓ General linear position—no three points are colinear.
- × These points do not have generic pairwise relations.



General takeaway I

(Not) learning from crowd data

- Weaker feedback may make data easier/cheaper to collect
  - e.g. triplet $\rightarrow$ binary feedback (with latent comparator)

General takeaway I

(Not) learning from crowd data

- Weaker feedback may make data easier/cheaper to collect
  - e.g. triplet $\rightarrow$ binary feedback (with latent comparator)
- But we may need to pay for it elsewhere
  - e.g. new fundamental limits/regimes where data carries no information

Metric learning with subspace-cluster structure

Real data often exhibit additional structure

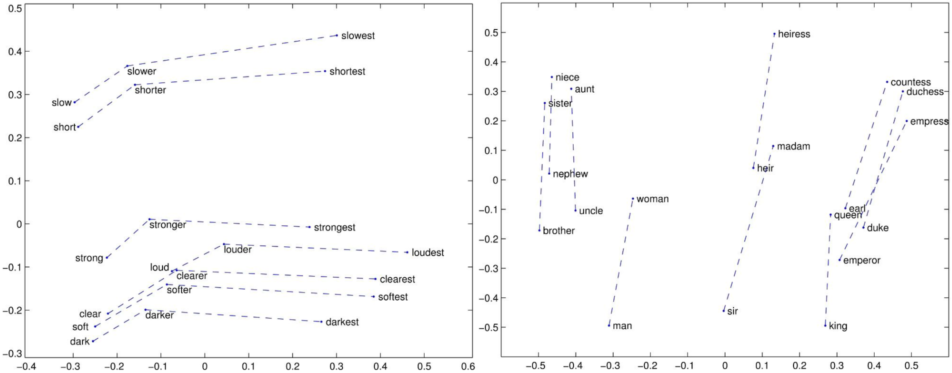


Figure 3: An example of data that approximately *does not have* generic pairwise relations (Pennington et al., 2014).

Subspace-clusterability assumption

Assumption:
There are low-dimensional subspaces of $\mathbb{R}^d$ that are 'rich' with items.

- That is, assume that $\mathcal{X}$ lies on a union of low-rank subspaces.

Subspace-clusterability assumption

Assumption:
There are low-dimensional subspaces of $\mathbb{R}^d$ that are 'rich' with items.

- That is, assume that $\mathcal{X}$ lies on a union of low-rank subspaces.
- e.g. $\mathcal{X}$ is *sparsely encodable*, in the sense of dictionary learning.

Divide-and-conquer approach

A natural approach:

1. Learn the metric restricted to each of the item-rich subspaces.
2. Stitch the subspace metrics together.

Subspace Mahalanobis distances

Definition

Let $V \subset \mathbb{R}^d$ be a subspace. A metric on V is a ***subspace Mahalanobis distance*** *if it is the subspace metric of a Mahalanobis distance ρ on $\mathbb{R}^d$,*

$$\rho\big|_V(x, x') = \rho(x, x'), \qquad \forall x, x' \in V.$$

Why can we divide?

Simple case: both items $\mathcal{X}$ and user ideal point u belong to V.

Why can we divide?

Simple case: both items $\mathcal{X}$ and user ideal point u belong to V.

- Simply reparametrize problem without the extra dimensions $V^\perp$.
- Learn $\rho\big|_V$ like before.

Why can we divide?

Simple case: both items $\mathcal{X}$ and user ideal point u belong to V.

- Simply reparametrize problem without the extra dimensions $V^\perp$.
- Learn $\rho\big|_V$ like before.

General case: we cannot assume the user ideal point u belongs to V.

Why can we divide?

Simple case: both items $\mathcal{X}$ and user ideal point u belong to V.

- Simply reparametrize problem without the extra dimensions $V^\perp$.
- Learn $\rho\big|_V$ like before.

General case: we cannot assume the user ideal point u belongs to V.

- It turns out for any $u \in \mathbb{R}^d$, there exists a phantom ideal point $\tilde{u}$ in V such that:

$$\psi_M(x, x'; u) = \psi_M(x, x'; \tilde{u}), \qquad \forall x, x' \in V.$$

Why can we divide?

Simple case: both items $\mathcal{X}$ and user ideal point u belong to V.

- Simply reparametrize problem without the extra dimensions $V^{\perp}$.
- Learn $\rho\big|_V$ like before.

General case: we cannot assume the user ideal point u belongs to V.

- It turns out for any $u \in \mathbb{R}^d$, there exists a phantom ideal point $\tilde{u}$ in V such that:

$$\psi_M(x, x'; u) = \psi_M(x, x'; \tilde{u}), \qquad \forall x, x' \in V.$$

- We can no longer recover u, but we can learn $\rho\big|_V$.

Why can we recombine?

After dividing, we end up with a collection of subspace metric:

$$\rho\big|_{V_1}, \ldots, \rho\big|_{V_n}.$$

Why can we recombine?

After dividing, we end up with a collection of subspace metric:

$$\rho\big|_{V_1}, \dots, \rho\big|_{V_n}.$$

Result: As long as the subspaces $V_1, \dots, V_n$ *quadratically span* $\mathbb{R}^d$, there is a unique Mahalanobis distance on $\mathbb{R}^d$ generating the joint subspace metrics.

Geometric proof

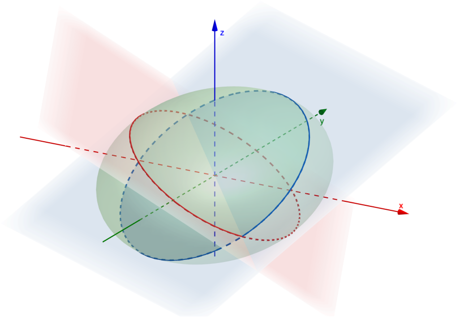


Figure 4: Unit spheres of Mahalanobis distances are ellipsoids in $\mathbb{R}^d$.

Geometric proof

For Mahalanobis distances:

- Metric learning is equivalent to recovering its unit ellipsoid $\mathcal{E}$.

Geometric proof

For Mahalanobis distances:

- Metric learning is equivalent to recovering its unit ellipsoid $\mathcal{E}$.
- Learning the subspace metric on V correspond to recovering the slice $V \cap \mathcal{E}$.

Geometric proof

For Mahalanobis distances:

- Metric learning is equivalent to recovering its unit ellipsoid $\mathcal{E}$.
- Learning the subspace metric on V correspond to recovering the slice $V \cap \mathcal{E}$.

Fact from geometry:
We can reconstruct an ellipsoid given enough low-dimensional slices.

Quadratic spanning

Definition

*The subspaces $V_1, \ldots, V_n \subset \mathbb{R}^d$ **quadratically span** $\mathbb{R}^d$ if the (linear) span satisfies:*

$$\mathrm{Sym}(\mathbb{R}^d) = \mathrm{span}\left(\left\{xx^\top : x \in V_1 \cup \cdots \cup V_n\right\}\right).$$

Metric learning from lazy crowds (simple math setting)

We asked: Suppose we can obtain very few $m \ll d$ measurements per user. Though ideal points can no longer be learned, is metric learning still possible?

Metric learning from lazy crowds (simple math setting)

We asked: Suppose we can obtain very few $m \ll d$ measurements per user. Though ideal points can no longer be learned, is metric learning still possible?

Answer (continuous response model):

- In general, this is not possible.

Metric learning from lazy crowds (simple math setting)

We asked: Suppose we can obtain very few $m \ll d$ measurements per user.
Though ideal points can no longer be learned, is metric learning still possible?

Answer (continuous response model):

- In general, this is not possible.
- If $\mathcal{X}$ is a union of r-dimensional subspaces ($r \ll d$), it is possible with:

number of users	d^2/r
measurements per user	$2r$

General takeaway II

Learning from crowd data

- Fundamental limit overcome using additional structural assumptions
  - e.g. generic pairwise relations $\rightarrow$ subspace-cluster structure

General takeaway II

Learning from crowd data

- Fundamental limit overcome using additional structural assumptions
  - e.g. generic pairwise relations $\rightarrow$ subspace-cluster structure
- These structural assumptions could be (approximately) realistic
  - we could even enforce the structure upsteam
  - e.g. generate representations via dictionary learning

Goals of the rest of the talk

Up to now:

- Fundamental limits of weak and per-user-budgeted crowdsourced data
- Paying for weak feedback if there is additional structure

Goals of the rest of the talk

Up to now:

- Fundamental limits of weak and per-user-budgeted crowdsourced data
- Paying for weak feedback if there is additional structure

Rest of the talk:

- High-level description of statistical/learning-theoretic techniques
- A commonly used model for analyzing preference feedback
- A fundamental open question: crowdsourced sensing with latent parameters

Metric learning from non-idealized data

Divide-and-conquer for idealized data

Divide step:
For each subspace $V_1, \ldots, V_n$, solve a system of linear equations:

$$\mathbf{D}_i(\hat{Q}_i, w_1, \ldots, w_K) = \Psi_i.$$

Divide-and-conquer for idealized data

Divide step:
For each subspace $V_1, \ldots, V_n$, solve a system of linear equations:

$$\mathbf{D}_i(\hat{Q}_i, w_1, \ldots, w_K) = \Psi_i.$$

Recombine step:
Define $\Pi(M) = (Q_1, \ldots, Q_n)$ to be the linear map:

$$\Pi : \text{parameters of Mahalanobis distances} \mapsto \text{parameters of subspace metrics.}$$

Divide-and-conquer for idealized data

Divide step:
For each subspace $V_1, \ldots, V_n$, solve a system of linear equations:

$$\mathbf{D}_i(\hat{Q}_i, w_1, \ldots, w_K) = \Psi_i.$$

Recombine step:
Define $\Pi(M) = (Q_1, \ldots, Q_n)$ to be the linear map:

$$\Pi : \text{parameters of Mahalanobis distances} \mapsto \text{parameters of subspace metrics.}$$

Solve a system of linear equations:

$$\hat{M} = \Pi(\hat{Q}_1, \ldots, \hat{Q}_n).$$

From linear systems to regression

Question: What happens in the non-idealized setting?

From linear systems to regression

Question: What happens in the non-idealized setting?

- Feedback is inexact/binary/noisy.

From linear systems to regression

Question: What happens in the non-idealized setting?

- Feedback is inexact/binary/noisy.
- The set of items is only approximately subspace clusterable.

From linear systems to regression

Question: What happens in the non-idealized setting?

- Feedback is inexact/binary/noisy.
- The set of items is only approximately subspace clusterable.

Divide step:
Prior work shows shows metric learning from non-idealized feedback.

From linear systems to regression

Question: What happens in the non-idealized setting?

- Feedback is inexact/binary/noisy.
- The set of items is only approximately subspace clusterable.

Divide step:
Prior work shows shows metric learning from non-idealized feedback.

- If we get binary responses, solve a binary regression problem instead.

From linear systems to regression

Question: What happens in the non-idealized setting?

- Feedback is inexact/binary/noisy.
- The set of items is only approximately subspace clusterable.

Divide step:
Prior work shows shows metric learning from non-idealized feedback.

- If we get binary responses, solve a binary regression problem instead.

Recombine step:
We need to show that we can recombine estimated subspace metrics.

From linear systems to regression

Question: What happens in the non-idealized setting?

- Feedback is inexact/binary/noisy.
- The set of items is only approximately subspace clusterable.

Divide step:
Prior work shows shows metric learning from non-idealized feedback.

- If we get binary responses, solve a binary regression problem instead.

Recombine step:
We need to show that we can recombine estimated subspace metrics.

- **Algorithm:** perform linear regression instead, and project onto the PSD cone.

Setting for recombination recovery guarantee

Setting:

- Let $\mathcal{V}_n = \{V_1, \ldots, V_n\}$ be a collection of subspaces of $\mathbb{R}^d$.

Setting for recombination recovery guarantee

Setting:

- Let $\mathcal{V}_n = \{V_1, \ldots, V_n\}$ be a collection of subspaces of $\mathbb{R}^d$.
- Let $Q_1, \ldots, Q_n$ be the true parameters of the subspace metrics.

Setting for recombination recovery guarantee

Setting:

- Let $\mathcal{V}_n = \{V_1, \ldots, V_n\}$ be a collection of subspaces of $\mathbb{R}^d$.
- Let $Q_1, \ldots, Q_n$ be the true parameters of the subspace metrics.
- Let $\hat{Q}_1, \ldots, \hat{Q}_n$ be independent estimators of the subspace metrics.

Setting for recombination recovery guarantee

Setting:

- Let $\mathcal{V}_n = \{V_1, \dots, V_n\}$ be a collection of subspaces of $\mathbb{R}^d$.
- Let $Q_1, \dots, Q_n$ be the true parameters of the subspace metrics.
- Let $\hat{Q}_1, \dots, \hat{Q}_n$ be independent estimators of the subspace metrics.
- Let $\hat{M}$ be the projected ordinary least squares solution (on the PSD cone):

$$\hat{M}_{\mathrm{OLS}} = \operatorname*{arg\,min}_{A \in \mathrm{Sym}(\mathbb{R}^d)} \sum_{i=1}^{n} \|\hat{Q}_i - \Pi_{V_i}(A)\|^2$$

Setting for recombination recovery guarantee

Setting:

- Let $\mathcal{V}_n = \{V_1, \ldots, V_n\}$ be a collection of subspaces of $\mathbb{R}^d$.
- Let $Q_1, \ldots, Q_n$ be the true parameters of the subspace metrics.
- Let $\hat{Q}_1, \ldots, \hat{Q}_n$ be independent estimators of the subspace metrics.
- Let $\hat{M}$ be the projected ordinary least squares solution (on the PSD cone):

$$\hat{M}_{\mathrm{OLS}} = \operatorname*{arg\,min}_{A \in \mathrm{Sym}(\mathbb{R}^d)} \sum_{i=1}^{n} \|\hat{Q}_i - \Pi_{V_i}(A)\|^2$$

Setting for recombination recovery guarantee

Setting:

- Let $\mathcal{V}_n = \{V_1, \ldots, V_n\}$ be a collection of subspaces of $\mathbb{R}^d$.
- Let $Q_1, \ldots, Q_n$ be the true parameters of the subspace metrics.
- Let $\hat{Q}_1, \ldots, \hat{Q}_n$ be independent estimators of the subspace metrics.
- Let $\hat{M}$ be the projected ordinary least squares solution (on the PSD cone):

$$\hat{M}_{\mathrm{OLS}} = \underset{A \in \mathrm{Sym}(\mathbb{R}^d)}{\arg\min} \sum_{i=1}^{n} \|\hat{Q}_i - \Pi_{V_i}(A)\|^2$$

$$\hat{M} = \underset{A \succeq 0}{\arg\min} \|\hat{M}_{\mathrm{OLS}} - A\|_F^2.$$

Recombination recovery guarantee

Assumptions:

- The estimators have low-bias: $\| \mathbb{E}[\hat{Q}_i] - Q_i\| \leq \gamma$.

Recombination recovery guarantee

Assumptions:

- The estimators have low-bias: $\| \mathbb{E}[\hat{Q}_i] - Q_i\| \leq \gamma$.
- The estimators have bounded spread: $\|\hat{Q}_i - \mathbb{E}[\hat{Q}_i]\| \leq \varepsilon$.

Recombination recovery guarantee

Assumptions:

- The estimators have low-bias: $\| \mathbb{E}[\hat{Q}_i] - Q_i \| \leq \gamma$.
- The estimators have bounded spread: $\|\hat{Q}_i - \mathbb{E}[\hat{Q}_i]\| \leq \varepsilon$.

Theorem

There is a constant $c > 0$ such that for any $p \in (0, 1]$, with probability at least $1 - p$,

Recombination recovery guarantee

Assumptions:

- The estimators have low-bias: $\| \mathbb{E}[\hat{Q}_i] - Q_i \| \leq \gamma$.
- The estimators have bounded spread: $\| \hat{Q}_i - \mathbb{E}[\hat{Q}_i] \| \leq \varepsilon$.

Theorem

There is a constant $c > 0$ such that for any $p \in (0, 1]$, with probability at least $1 - p$,

$$\|\hat{M} - M\|_F \leq c \cdot \frac{1}{\sigma(\mathcal{V}_n)} \cdot \left(\gamma\sqrt{n} + \varepsilon d \sqrt{\log \frac{2d}{p}} \right),$$

Recombination recovery guarantee

Assumptions:

- The estimators have low-bias: $\| \mathbb{E}[\hat{Q}_i] - Q_i \| \leq \gamma$.
- The estimators have bounded spread: $\| \hat{Q}_i - \mathbb{E}[\hat{Q}_i] \| \leq \varepsilon$.

Theorem

There is a constant $c > 0$ such that for any $p \in (0, 1]$, with probability at least $1 - p$,

$$\|\hat{M} - M\|_F \leq c \cdot \frac{1}{\sigma(\mathcal{V}_n)} \cdot \left(\gamma\sqrt{n} + \varepsilon d \sqrt{\log \frac{2d}{p}} \right),$$

where $\sigma(\mathcal{V})$ quantifies the 'quadratic spread' of subspaces $V_1, \ldots, V_n$ in $\mathrm{Sym}(\mathbb{R}^d)$.

Proof sketch

For simplicity, we just show bound for $\|\hat{M}_{\mathrm{OLS}} - M\|_F$.

Proof sketch

For simplicity, we just show bound for $\|\hat{M}_{\mathrm{OLS}} - M\|_F$.

1. Recall the linear map $\Pi(M) = (Q_1, \ldots, Q_n)$.

Proof sketch

For simplicity, we just show bound for $\|\hat{M}_{\text{OLS}} - M\|_F$.

1. Recall the linear map $\Pi(M) = (Q_1, \ldots, Q_n)$.
2. The OLS solution is computed by the Moore-Penrose pseudoinverse:

$$\hat{M}_{\text{OLS}} = \Pi^+(\hat{Q}_1, \ldots, \hat{Q}_n).$$

Proof sketch

For simplicity, we just show bound for $\|\hat{M}_{\text{OLS}} - M\|_F$.

1. Recall the linear map $\Pi(M) = (Q_1, \ldots, Q_n)$.
2. The OLS solution is computed by the Moore-Penrose pseudoinverse:

$$\hat{M}_{\text{OLS}} = \Pi^+(\hat{Q}_1, \ldots, \hat{Q}_n).$$

3. Since Π^+ is linear, we get the bound:

$$\|\hat{M}_{\text{OLS}} - M\|_F^2 = \|\Pi^+(\hat{Q}_1 - Q, \ldots, \hat{Q}_n - Q_n)\|_F^2 \leq \sigma_{\max}^2(\Pi^+) \sum_{i=1}^{n} \|\hat{Q}_1 - Q_i\|^2.$$

Proof sketch

For simplicity, we just show bound for $\|\hat{M}_{\text{OLS}} - M\|_F$.

1. Recall the linear map $\Pi(M) = (Q_1, \ldots, Q_n)$.
2. The OLS solution is computed by the Moore-Penrose pseudoinverse:

$$\hat{M}_{\text{OLS}} = \Pi^+(\hat{Q}_1, \ldots, \hat{Q}_n).$$

3. Since Π^+ is linear, we get the bound:

$$\|\hat{M}_{\text{OLS}} - M\|_F^2 = \|\Pi^+(\hat{Q}_1 - Q, \ldots, \hat{Q}_n - Q_n)\|_F^2 \leq \sigma_{\max}^2(\Pi^+) \sum_{i=1}^{n} \|\hat{Q}_1 - Q_i\|^2.$$

4. A more fine-grained bound by decomposition: $\hat{Q} - Q = \underbrace{\hat{Q} - \mathbb{E}[\hat{Q}]}_{\text{mean-zero r.v.}} + \underbrace{\mathbb{E}[\hat{Q}] - Q}_{\text{bias}}.$

Proof sketch

For simplicity, we just show bound for $\|\hat{M}_{\mathrm{OLS}} - M\|_F$.

1. Recall the linear map $\Pi(M) = (Q_1, \ldots, Q_n)$.
2. The OLS solution is computed by the Moore-Penrose pseudoinverse:
$$\hat{M}_{\mathrm{OLS}} = \Pi^+(\hat{Q}_1, \ldots, \hat{Q}_n).$$
3. Since Π^+ is linear, we get the bound:
$$\|\hat{M}_{\mathrm{OLS}} - M\|_F^2 = \|\Pi^+(\hat{Q}_1 - Q, \ldots, \hat{Q}_n - Q_n)\|_F^2 \leq \sigma_{\max}^2(\Pi^+) \sum_{i=1}^{n} \|\hat{Q}_1 - Q_i\|^2.$$
4. A more fine-grained bound by decomposition: $\hat{Q} - Q = \underbrace{\hat{Q} - \mathbb{E}[\hat{Q}]}_{\text{mean-zero r.v.}} + \underbrace{\mathbb{E}[\hat{Q}] - Q}_{\text{bias}}$.
   - For independent mean-zero error terms, can apply Chernoff-style concentration.

Interpretation of the bound

Key quantities: $n =$ number of subspaces; $\gamma, \varepsilon =$ subspace recovery bias/accuracy

$$\|\hat{M} - M\|_F \leq c \cdot \frac{1}{\sigma(\mathcal{V})} \cdot \left(\gamma\sqrt{n} + \varepsilon d \sqrt{\log \frac{2d}{p}} \right)$$

Interpretation of the bound

Key quantities: n = number of subspaces; γ, ε = subspace recovery bias/accuracy

$$\|\hat{M} - M\|_F \leq c \cdot \frac{1}{\sigma(\mathcal{V})} \cdot \left(\gamma\sqrt{n} + \varepsilon d \sqrt{\log \frac{2d}{p}} \right)$$

- $\sigma(\mathcal{V}_n)$ grows with the number of subspaces,

$$\sigma(\mathcal{V}_n) = \Omega(\sqrt{n}) \text{ is possible.}$$

Interpretation of the bound

Key quantities: n = number of subspaces; γ, ε = subspace recovery bias/accuracy

$$\|\hat{M} - M\|_F \leq c \cdot \frac{1}{\sigma(\mathcal{V})} \cdot \left(\gamma\sqrt{n} + \varepsilon d\sqrt{\log \frac{2d}{p}}\right)$$

- $\sigma(\mathcal{V}_n)$ grows with the number of subspaces,

$$\sigma(\mathcal{V}_n) = \Omega(\sqrt{n}) \text{ is possible.}$$

As $n \to \infty$, the dominating term is possibly the bias term γ.

- e.g. if the estimators $\hat{Q}$ have a systematic constant biases $\gamma > 0$.

A noisy feedback model with recovery guarantee

Probabilistic model

Generalized linear model:

- Continuous response: (x, x', ψ)

$$\psi \equiv \psi_M(x, x'; u) = D_{x,x'}(M, v).$$

Probabilistic model

Generalized linear model:

- Continuous response: (x, x', ψ)

$$\psi \equiv \psi_M(x, x'; u) = D_{x,x'}(M, v).$$

- Noisy response: (x, x', y)

$$\Pr\left[Y = y \,\middle|\, M, x, x', u\right] = f\big(y \cdot D_{x,x'}(M, v)\big),$$

where f is a (non-linear) *link function.*

Probabilistic model

Generalized linear model:

- Continuous response: (x, x', ψ)

$$\psi \equiv \psi_M(x, x'; u) = D_{x,x'}(M, v).$$

- Noisy response: (x, x', y)

$$\Pr\left[Y = y \,\middle|\, M, x, x', u\right] = f\big(y \cdot D_{x,x'}(M, v)\big),$$

where f is a (non-linear) *link function.*

  - The link function is the first (and only) instance of a non-linearity in this work.

Probabilistic model

Generalized linear model:

- Continuous response: (x, x', ψ)

$$\psi \equiv \psi_M(x, x'; u) = D_{x,x'}(M, v).$$

- Noisy response: (x, x', y)

$$\Pr\left[Y = y \,\middle|\, M, x, x', u\right] = f\big(y \cdot D_{x,x'}(M, v)\big),$$

where f is a (non-linear) *link function.*

  - The link function is the first (and only) instance of a non-linearity in this work.
  - When $f(z) = \frac{1}{1+\exp(-z)}$ is the sigmoid function, this leads to a logistic regression.

Setting for subspace metric recovery

Setting:

- Assume that user provide measurements (x, x', Y) where $Y \in \{-1, +1\}$,

$$\Pr\left[Y = y\right] = f\big(-y \cdot D_{x,x'}(M, v)\big),$$

where f is the sigmoid link function.

Setting for subspace metric recovery

Setting:

- Assume that user provide measurements (x, x', Y) where $Y \in \{-1, +1\}$,

$$\Pr\left[Y = y\right] = f\Big(- y \cdot D_{x,x'}(M, v)\Big),$$

  where f is the sigmoid link function.

- We can perform maximum likelihood estimation:

$$(\hat{M}, \hat{v}_1, \dots, \hat{v}_k) \leftarrow \underset{(A, w_1, \dots, w_K)}{\arg\max} \sum_k \sum_{(x,x',Y)} \log f\Big(- Y \cdot D_{x,x'}(M, v_k)\Big).$$

Setting for subspace metric recovery

Setting:

- Assume that user provide measurements (x, x', Y) where $Y \in \{-1, +1\}$,

$$\Pr\left[Y = y\right] = f\big(-y \cdot D_{x,x'}(M, v)\big),$$

where f is the sigmoid link function.

- We can perform maximum likelihood estimation:

$$(\hat{M}, \hat{v}_1, \ldots, \hat{v}_k) \leftarrow \underset{(A, w_1, \ldots, w_K)}{\arg\max} \sum_k \sum_{(x,x',Y)} \log f\big(-Y \cdot D_{x,x'}(M, v_k)\big).$$

- Assume $\|M\|_\infty \leq 1$ and items and ideal points are contained in unit Euclidean ball.

Analysis via generalization

Theorem (Metric recovery, adapted from Canal et al. (2022))

Let $\mathcal{X}$ quadratically span $\mathbb{R}^d$. There exists designs $D^{(k)}$ asking for m responses from each of K users such that from that data, the maximum likelihood estimator $\hat{M}$ satisfies w.h.p.:

$$\|\hat{M} - M\|_F^2 = \mathcal{O}\left(\sqrt{\frac{d^2 + dK}{mK}}\right).$$

Analysis via generalization

Theorem (Metric recovery, adapted from Canal et al. (2022))

Let $\mathcal{X}$ quadratically span $\mathbb{R}^d$. There exists designs $D^{(k)}$ asking for m responses from each of K users such that from that data, the maximum likelihood estimator $\hat{M}$ satisfies w.h.p.:

$$\|\hat{M} - M\|_F^2 = \mathcal{O}\left(\sqrt{\frac{d^2 + dK}{mK}}\right).$$

- Proof uses standard techniques from generalization theory.

Analysis via generalization

Theorem (Metric recovery, adapted from Canal et al. (2022))

Let $\mathcal{X}$ quadratically span $\mathbb{R}^d$. There exists designs $D^{(k)}$ asking for m responses from each of K users such that from that data, the maximum likelihood estimator $\hat{M}$ satisfies w.h.p.:

$$\|\hat{M} - M\|_F^2 = \mathcal{O}\left(\sqrt{\frac{d^2 + dK}{mK}}\right).$$

- Proof uses standard techniques from generalization theory.
- The $d^2 + dK$ term comes from a metric entropy bound on:

$$\{(A, u_1, \ldots, u_K) : \|A\|_\infty \leq 1 \text{ and } \|u_k\| \leq 1, \forall k\}.$$

Analysis via generalization

Theorem (Metric recovery, adapted from Canal et al. (2022))

Let $\mathcal{X}$ quadratically span $\mathbb{R}^d$. There exists designs $D^{(k)}$ asking for m responses from each of K users such that from that data, the maximum likelihood estimator $\hat{M}$ satisfies w.h.p.:

$$\|\hat{M} - M\|_F^2 = \mathcal{O}\left(\sqrt{\frac{d^2 + dK}{mK}}\right).$$

- Proof uses standard techniques from generalization theory.
- The $d^2 + dK$ term comes from a metric entropy bound on:

$$\{(A, u_1, \ldots, u_K) : \|A\|_\infty \leq 1 \text{ and } \|u_k\| \leq 1, \forall k\}.$$

- When $K \gg d^2$, the dominating term is $\sqrt{d/m}$.

Open question

Weakness of analysis, weakness of naive ERM, or fundamental limit?

Open question

Weakness of analysis, weakness of naive ERM, or fundamental limit?

- The generalization approach actually shows:

$$\|\hat{M} - M\|_F^2 + \sum_{k=1}^{K} \|\hat{v}_k - v_k\|^2 = \mathcal{O}\left(\sqrt{\frac{d^2 + dK}{mK}}\right).$$

Open question

Weakness of analysis, weakness of naive ERM, or fundamental limit?

- The generalization approach actually shows:

$$\|\hat{M} - M\|_F^2 + \sum_{k=1}^{K} \|\hat{v}_k - v_k\|^2 = \mathcal{O}\left(\sqrt{\frac{d^2 + dK}{mK}}\right).$$

- But, we only care about learning the parameters of M.

Open question

Weakness of analysis, weakness of naive ERM, or fundamental limit?

- The generalization approach actually shows:

$$\|\hat{M} - M\|_F^2 + \sum_{k=1}^{K} \|\hat{v}_k - v_k\|^2 = \mathcal{O}\left(\sqrt{\frac{d^2 + dK}{mK}}\right).$$

- But, we only care about learning the parameters of M.
- This analysis does not seem to allow us to decouple estimating $\hat{M}$ and $\hat{v}_k$.

Open question

Weakness of analysis, weakness of naive ERM, or fundamental limit?

- The generalization approach actually shows:

$$\|\hat{M} - M\|_F^2 + \sum_{k=1}^{K} \|\hat{v}_k - v_k\|^2 = \mathcal{O}\left(\sqrt{\frac{d^2 + dK}{mK}}\right).$$

- But, we only care about learning the parameters of M.
- This analysis does not seem to allow us to decouple estimating $\hat{M}$ and $\hat{v}_k$.
- Is the analysis loose? Is there a better algorithm? Is there a fundamental limit?

Implication for metric learning

Suppose K users provide m measurements on rank-r subspaces.

Implication for metric learning

Suppose K users provide m measurements on rank-r subspaces.

Subspace metric error:

$$\gamma + \varepsilon \leq \mathcal{O}\left(\sqrt{\frac{r^2 + rK}{mK}}\right).$$

Implication for metric learning

Suppose K users provide m measurements on rank-r subspaces.

Subspace metric error:

$$\gamma + \varepsilon \leq \mathcal{O}\left(\sqrt{\frac{r^2 + rK}{mK}}\right).$$

Metric error after recombination:

$$\|\hat{M} - M\|_F \leq c \cdot \frac{1}{\sigma(\mathcal{V})} \cdot \left(\gamma\sqrt{n} + \varepsilon d \sqrt{\log \frac{2d}{p}}\right)$$

Implication for metric learning

Suppose K users provide m measurements on rank-r subspaces.

Subspace metric error:

$$\gamma + \varepsilon \leq \mathcal{O}\left(\sqrt{\frac{r^2 + rK}{mK}}\right).$$

Metric error after recombination:

$$\|\hat{M} - M\|_F \leq c \cdot \frac{1}{\sigma(\mathcal{V})} \cdot \left(\gamma\sqrt{n} + \varepsilon d\sqrt{\log\frac{2d}{p}}\right)$$

When $K \gg d$, then there are settings with: $\|\hat{M} - M\|_F = \mathcal{O}\left(\sqrt{\frac{r}{m}}\right)$.

Additional open problems

Further questions

Other structure:

- Low rank metrics; non-linear representations/kernel extension
- Learning with approximate subspace clusters
- Learning with structured user sets

Further questions

Other structure:

- Low rank metrics; non-linear representations/kernel extension
- Learning with approximate subspace clusters
- Learning with structured user sets

Inducing structure:

- What are good representations for human/crowdsourced labeling?

Further questions

Other structure:

- Low rank metrics; non-linear representations/kernel extension
- Learning with approximate subspace clusters
- Learning with structured user sets

Inducing structure:

- What are good representations for human/crowdsourced labeling?

Statistics:

- Other noise/preference models (e.g. Bradley-Terry model)
- Semi-parametric estimation
- Robust recovery

Acknowledgments

Collaborators

Zhi Wang
UC San Diego

Ramya Korlakai Vinayak
UW–Madison

Thank you!

See *https://geelon.github.io/* for preprint.

References

Gregory Canal, Blake Mason, Ramya Korlakai Vinayak, and Robert Nowak. One for all: Simultaneous metric and preference learning over multiple users. *Advances in Neural Information Processing Systems*, 35:4943–4956, 2022.

Clyde H Coombs. Psychological scaling without a unit of measurement. *Psychological review*, 57(3):145, 1950.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Blake Mason, Lalit Jain, and Robert Nowak. Learning low-dimensional metrics. *Advances in neural information processing systems*, 30, 2017.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Matthew Schultz and Thorsten Joachims. Learning a distance metric from relative comparisons. *Advances in neural information processing systems*, 16, 2003.

Nakul Verma and Kristin Branson. Sample complexity of learning mahalanobis distance metrics. *Advances in neural information processing systems*, 28, 2015.

Zhi Wang, Geelon So, and Ramya Korlakai Vinayak. Metric learning from limited pairwise preference comparisons, 2024.

Austin Xu and Mark Davenport. Simultaneous preference and metric learning from paired comparisons. *Advances in Neural Information Processing Systems*, 33:454–465, 2020.