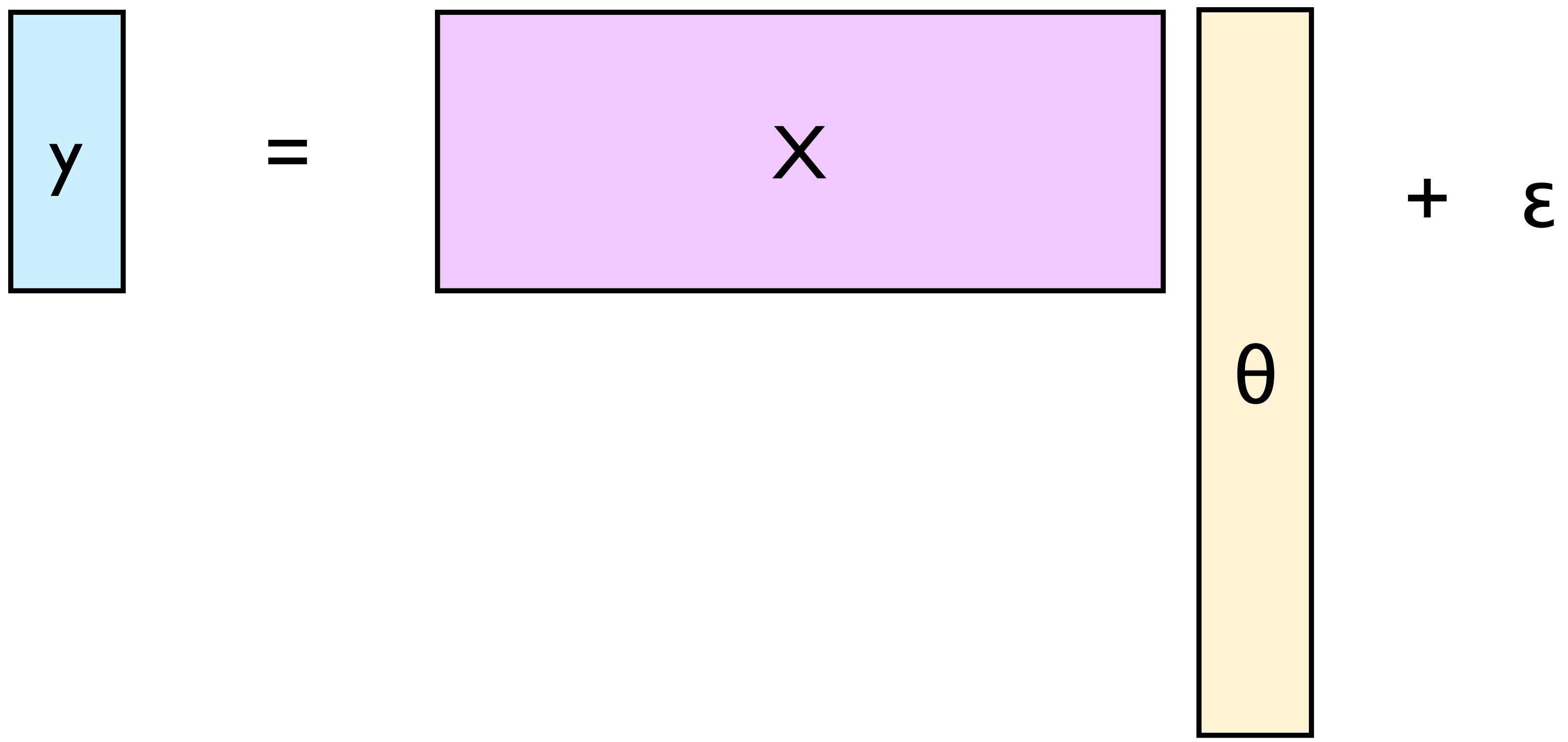
Variable Splitting Methods

Eric Chi

January 15, 2016

Two Typical Problems

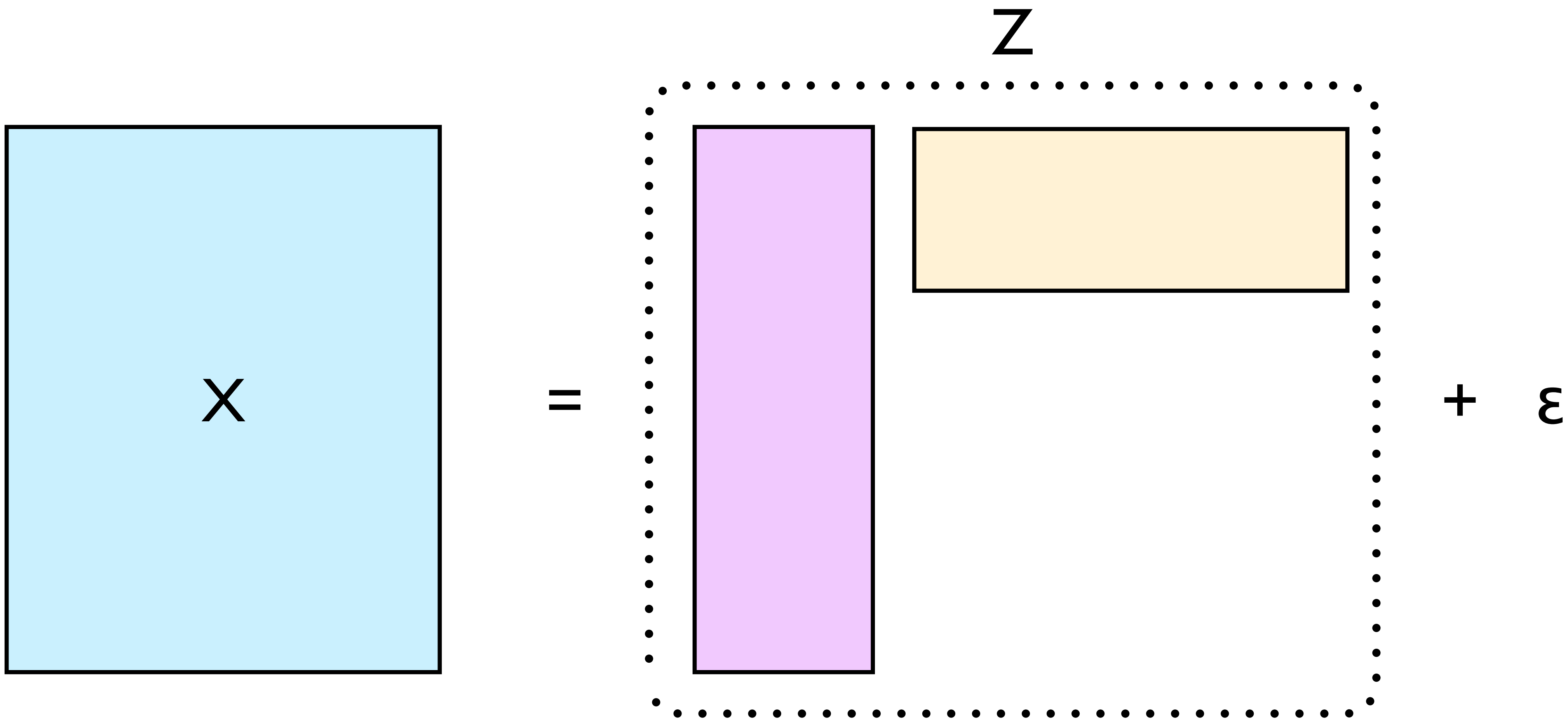


- Regularized estimation to get sparse solutions

$$\hat{\theta} = \arg\min_{\boldsymbol{\theta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1$$

Arises in biomedical problems: genome wide association studies

Two Typical Problems

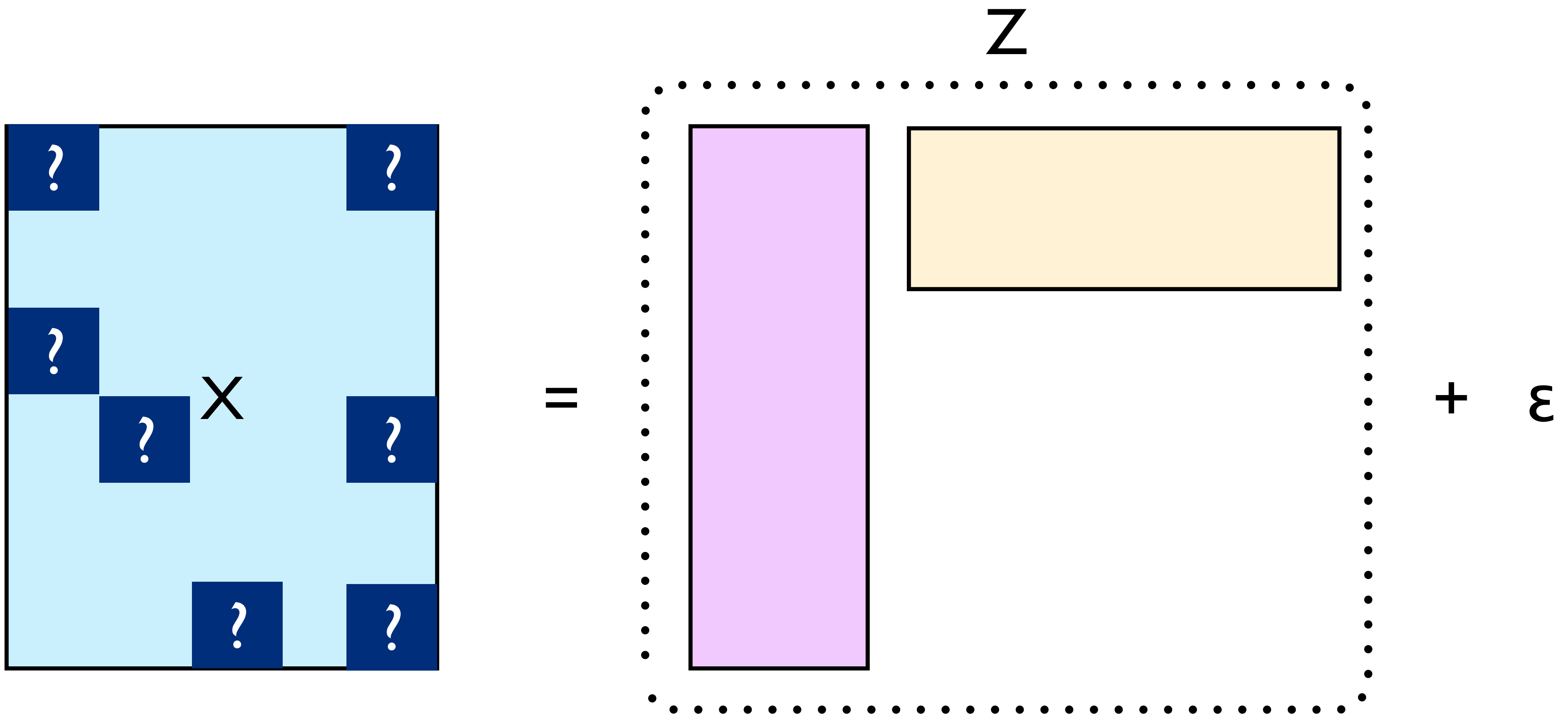


- Regularized estimation to get low-rank solutions

$$\hat{\mathbf{Z}} = \arg\min_{\mathbf{Z}} \frac{1}{2}\|\mathbf{X} - \mathbf{Z}\|_2^2 + \lambda\|\mathbf{Z}\|_*$$

Arises in collaborative filtering: Netflix

Two Typical Problems



- Regularized estimation to get low-rank solutions

$$\hat{\mathbf{Z}} = \arg\min_{\mathbf{Z}} \frac{1}{2}\|\mathbf{X} - \mathbf{Z}\|_2^2 + \lambda\|\mathbf{Z}\|_*$$

Arises in collaborative filtering: Netflix

The Generic Problem

$$\hat{\theta} = \arg\min_{\boldsymbol{\theta}} \underbrace{L(\boldsymbol{\theta})}_{\text{Lack of fit}} + \underbrace{J(\boldsymbol{\theta})}_{\text{Complexity}}$$

Reasons for success:

- Theory: Consistency and convergence rates when $n, p \to \infty$
- Computation: Fast and scalable algorithms for computing $\hat{\theta}$

The Generic Problem

$$\hat{\theta} = \arg\min_{\boldsymbol{\theta}} \underbrace{L(\boldsymbol{\theta})}_{\text{Lack of fit}} + \underbrace{J(\mathbf{D}\boldsymbol{\theta})}_{\text{Complexity}}$$

Reasons for success:

- Theory: Consistency and convergence rates when $n, p \to \infty$
- Computation: Fast and scalable algorithms for computing $\hat{\theta}$

What Variable Splitting Can Do For You

$$\hat{\theta} = \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) + J(\mathbf{D}\boldsymbol{\theta})$$

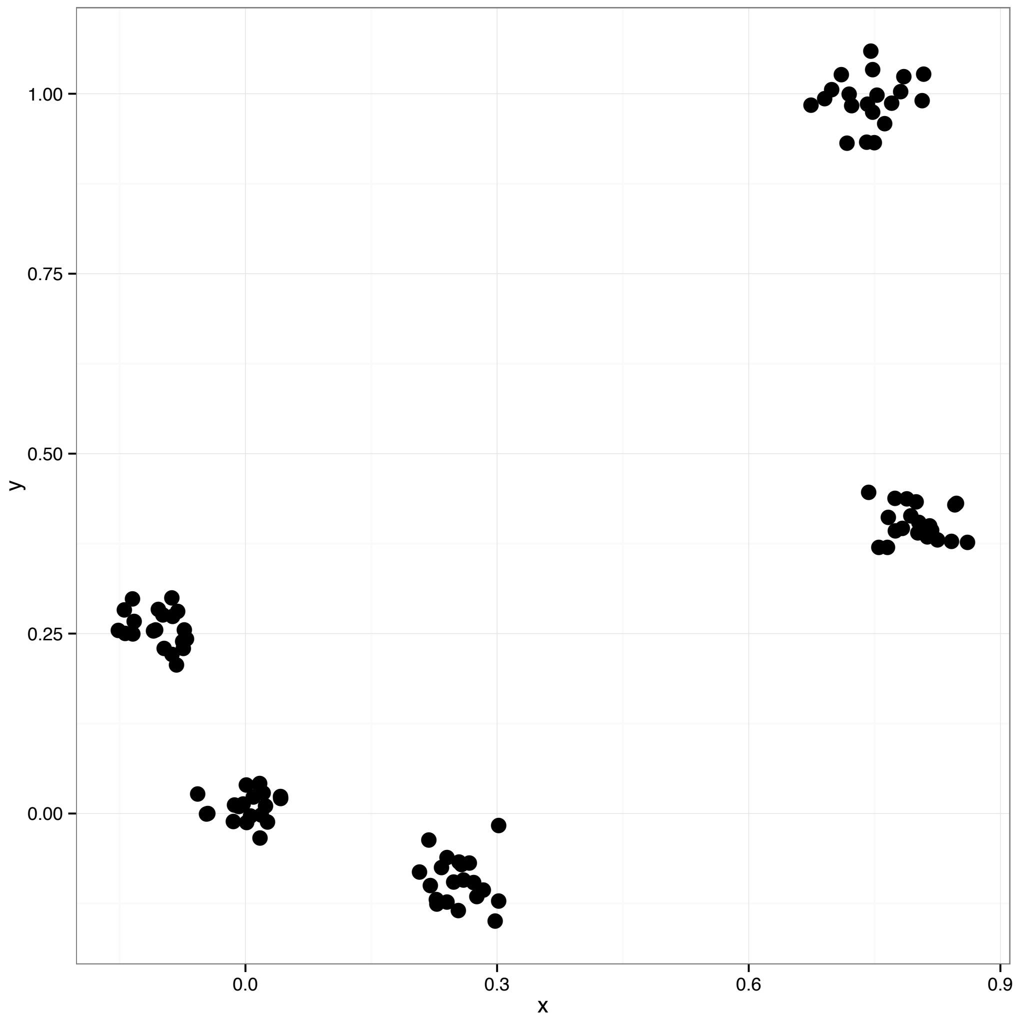
Variable splitting is

- helpful when $J(\boldsymbol{\theta})$ is to work with but $J(\mathbf{D}\boldsymbol{\theta})$ is not.
- typically easy to derive and code
  - e.g. Lasso solver in less than 10 lines of code.
- modestly accurate solutions in 10s to 100s of iterations.

Agenda

- Case Study: Convex Clustering I
- Variable Splitting
  - ADMM
  - AMA
- Case Study: Convex Clustering II
- Case Study: ADMM for Lasso

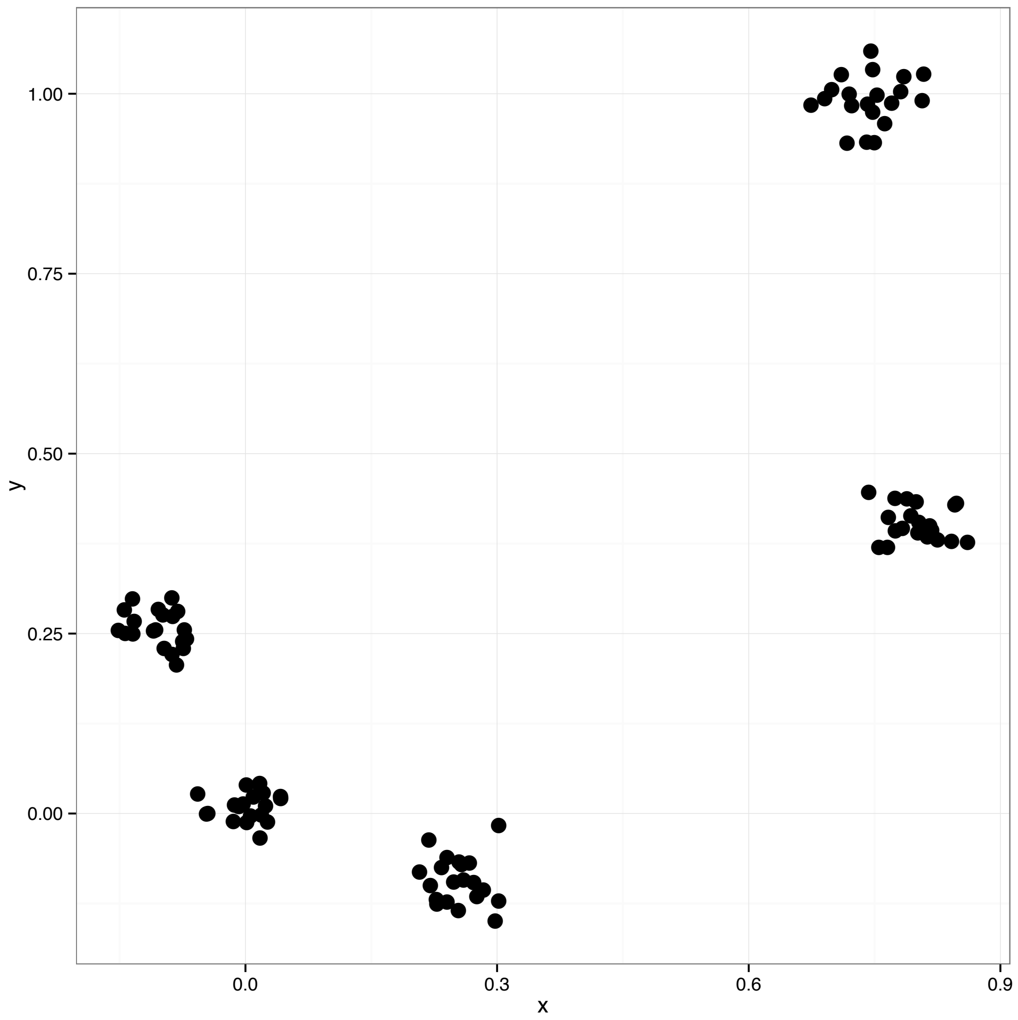
The Clustering Problem



Task:

- Given p points in q dimensions
- $\mathbf{X} \in \mathbb{R}^{q \times p}$
- group similar points together.

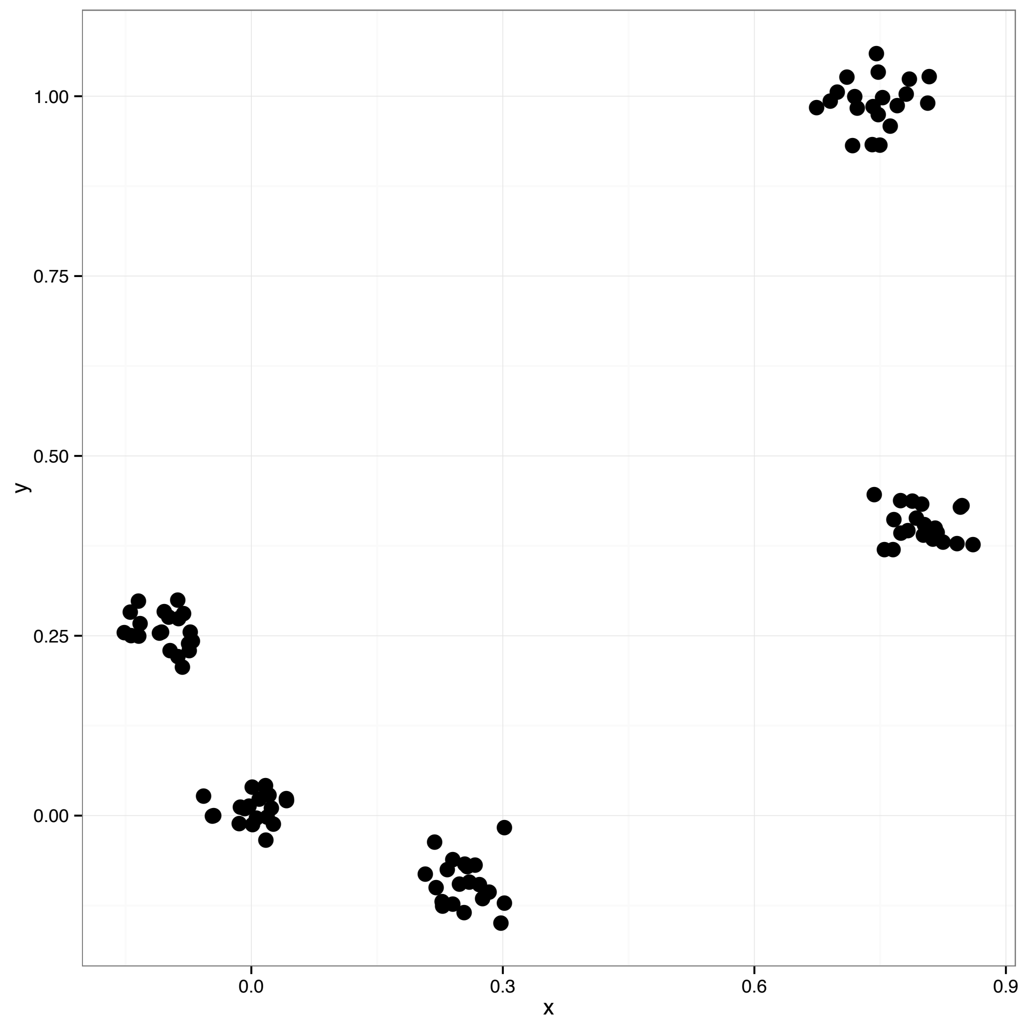
The Clustering Problem



Many approaches:

- *k*-means, mixture models
- Hierarchical clustering
- Spectral clustering, ...

The Clustering Problem



Computational Issues

- Nonconvex formulations
- Local minimizers
- Instability (initializations, tuning parameters, or data)

Convex Clustering

- Pelckmans et al. 2005, Lindsten et al. 2011, Hocking et al. 2011

$$\underset{\mathbf{u}}{\text{minimize}} \; \frac{1}{2}\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{u}_i\|_2^2$$

- Assign a centroid $\mathbf{u}_i$ to each data point $\mathbf{x}_i$.

Convex Clustering

- Pelckmans et al. 2005, Lindsten et al. 2011, Hocking et al. 2011

$$\underset{\mathbf{u}}{\text{minimize}} \ \frac{1}{2}\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{u}_i\|_2^2$$

- Assign a centroid $\mathbf{u}_i$ to each data point $\mathbf{x}_i$.

Too many degrees of freedom!

Convex Clustering

- Pelckmans et al. 2005, Lindsten et al. 2011, Hocking et al. 2011

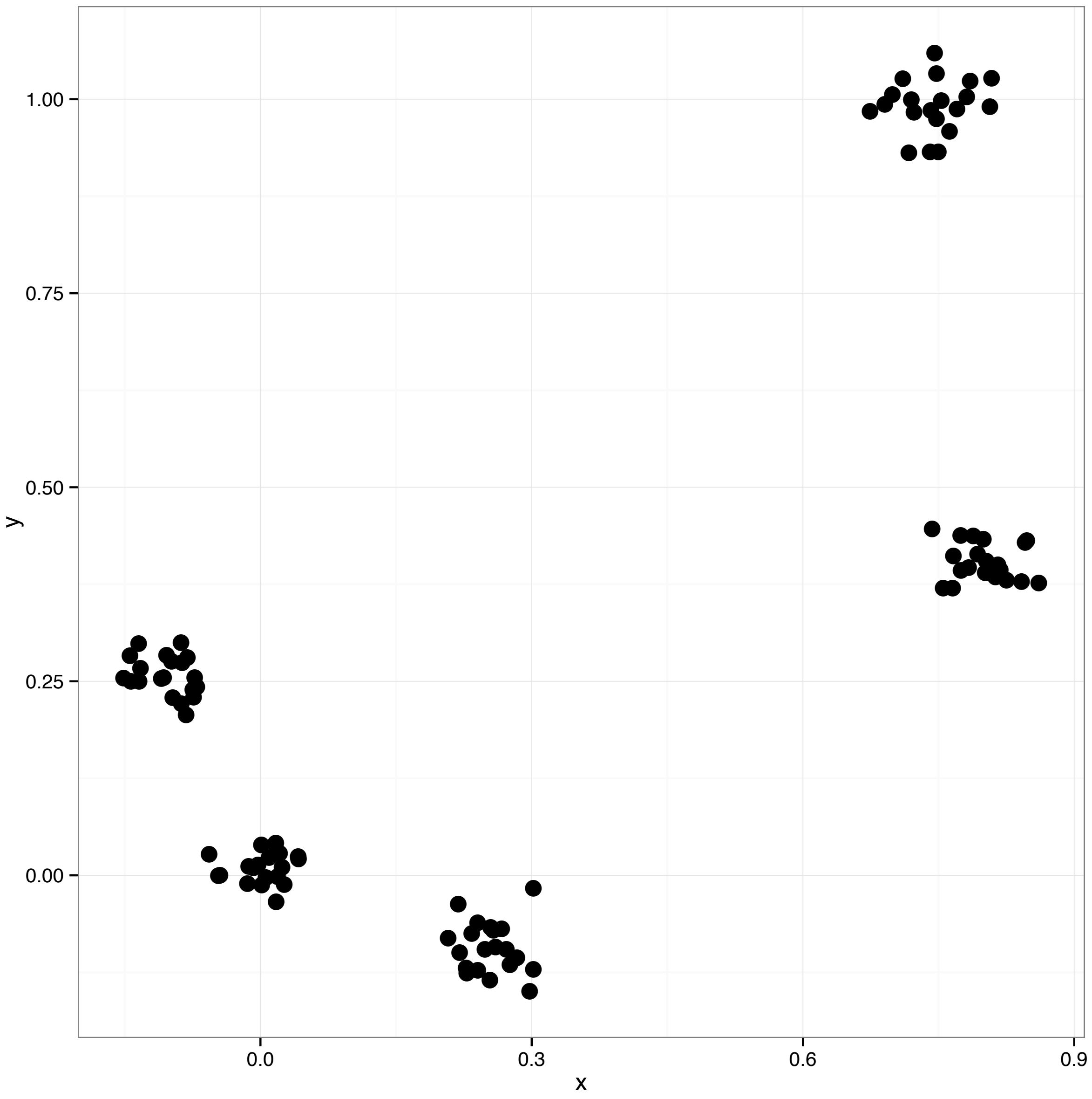
$$\underset{\mathbf{u}}{\text{minimize}}\ \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

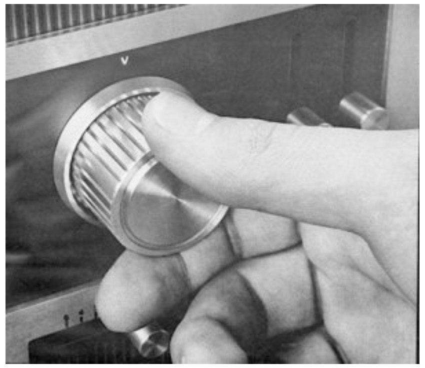
- Assign a centroid $\mathbf{u}_i$ to each data point $\mathbf{x}_i$.
- Convex Fusion Penalty
  - shrinks cluster centroids together
  - sparsity in pairwise differences of centroids

$$\mathbf{u}_i - \mathbf{u}_j = \mathbf{0} \iff \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the same cluster}$$

- γ : tunes overall amount of regularization
- w_{ij} : fine tunes pairwise shrinkage
- Generalizes fused lasso

The Solution Path

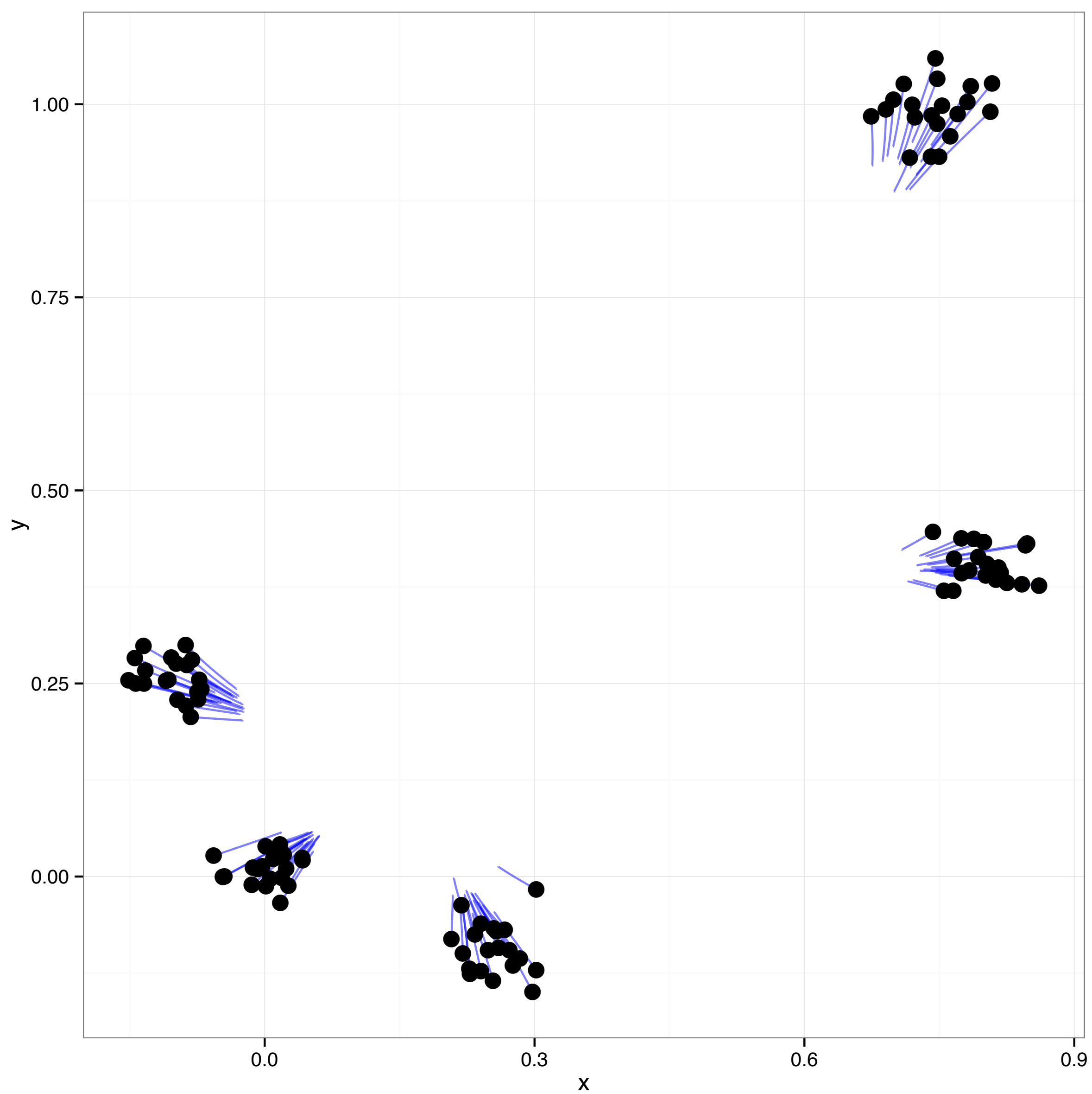


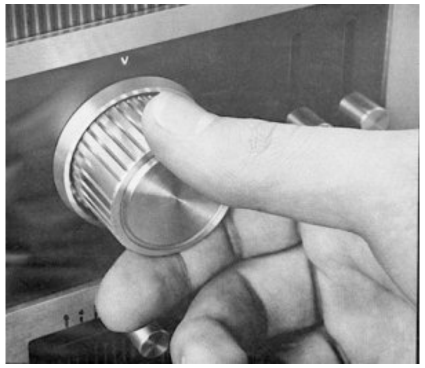


γ

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

The Solution Path

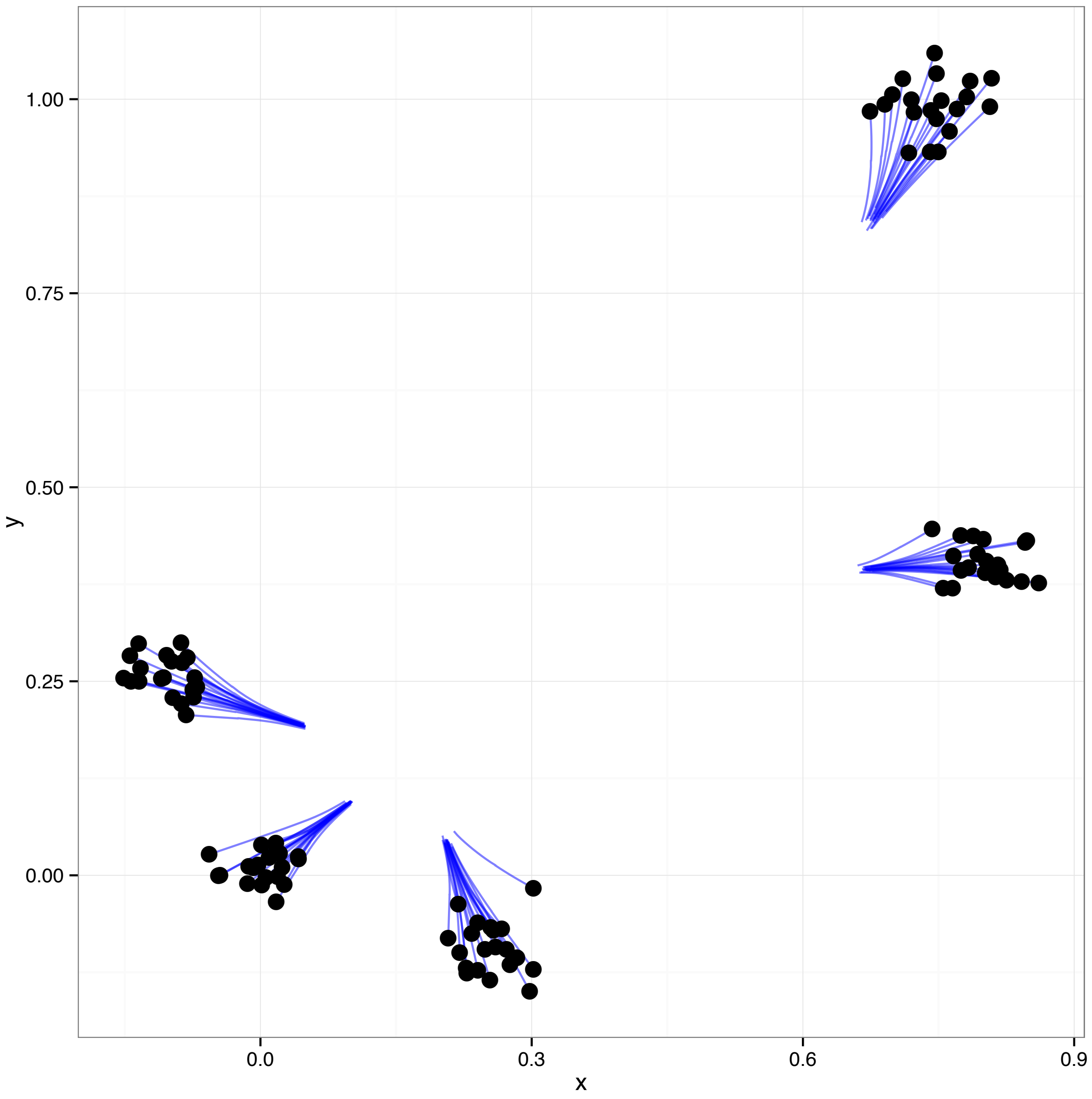


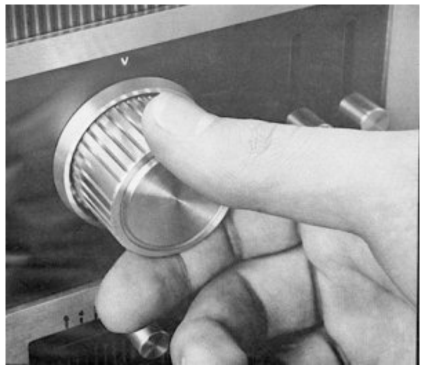


γ

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{i<j} w_{ij} \|\mathbf{u}_i - \mathbf{u}_j\|_2$$

The Solution Path

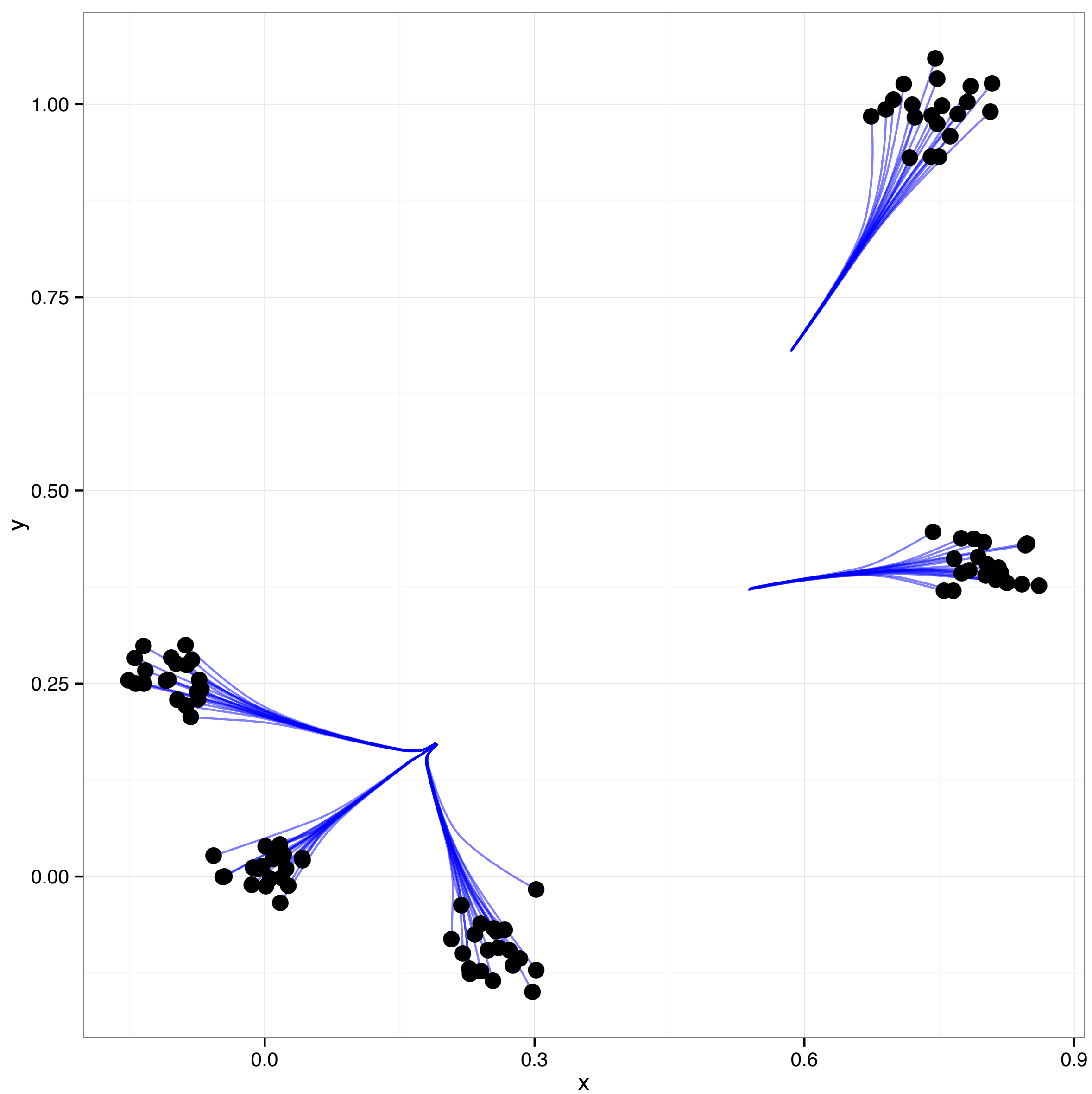


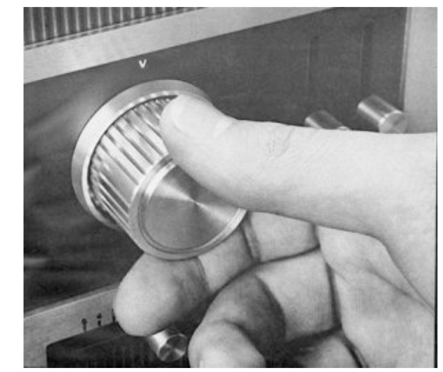


γ

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

The Solution Path

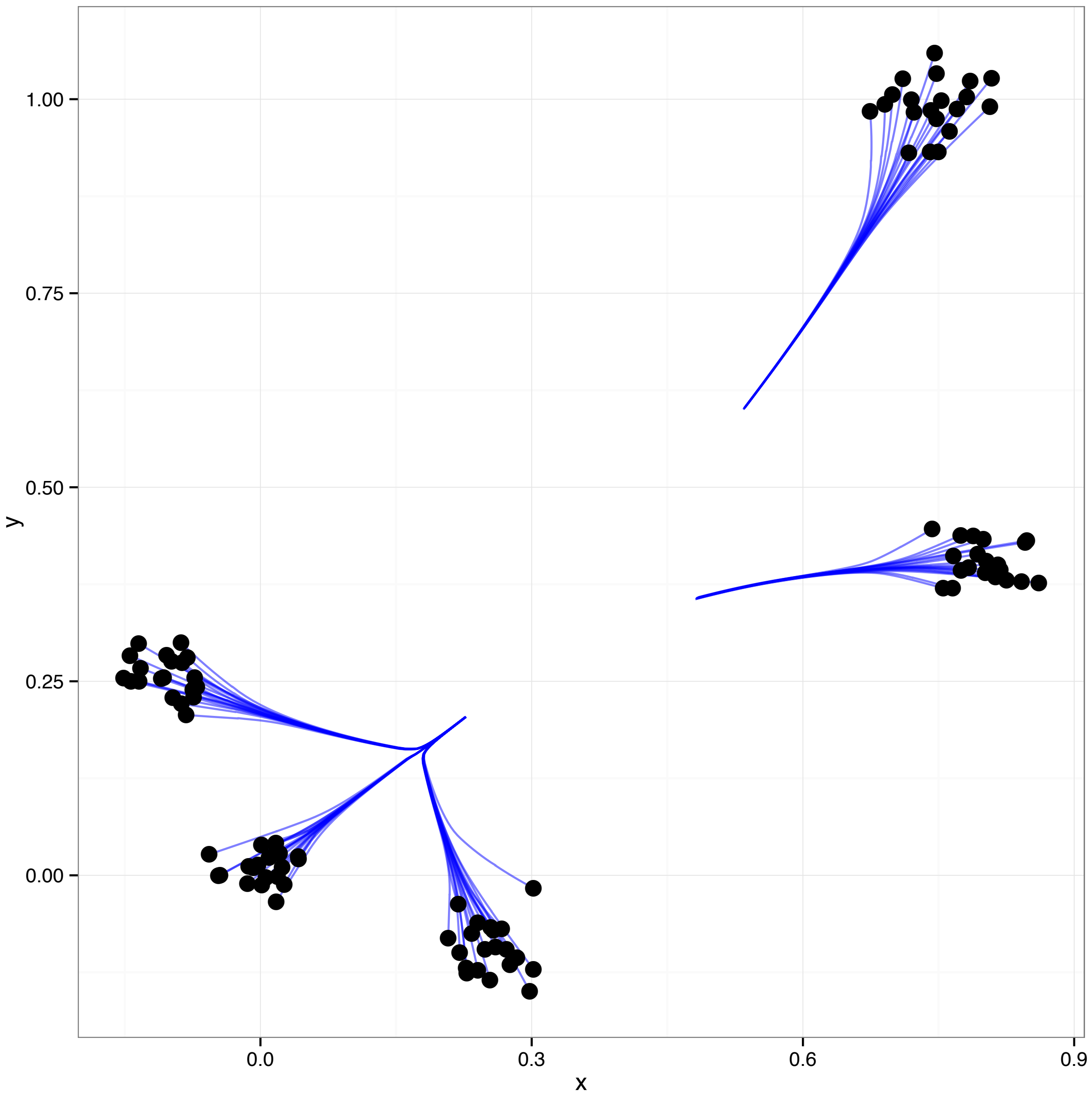


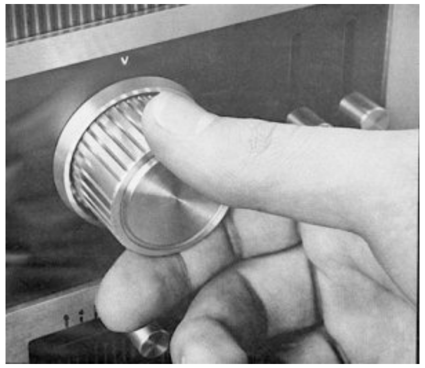


γ

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

The Solution Path

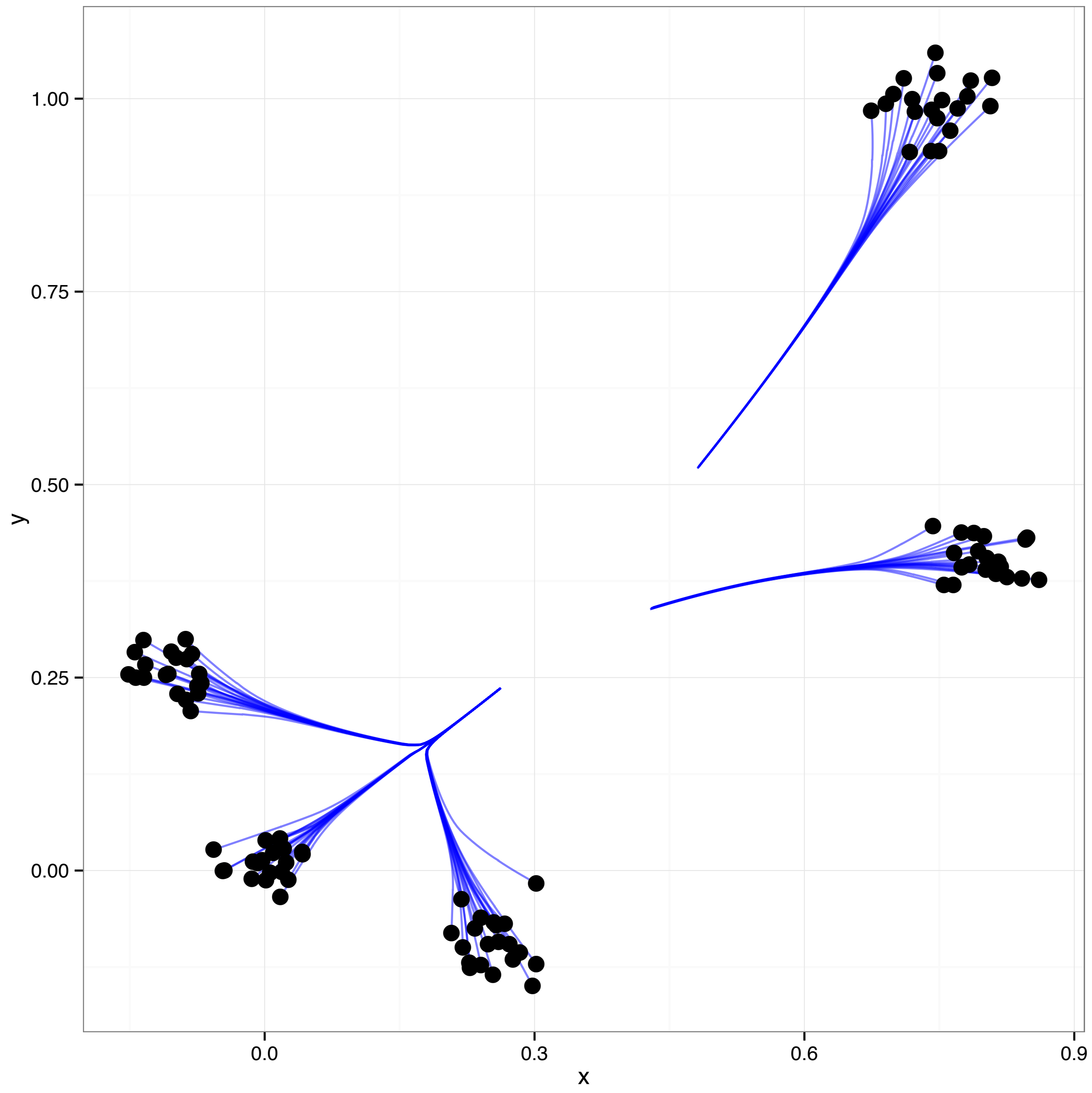


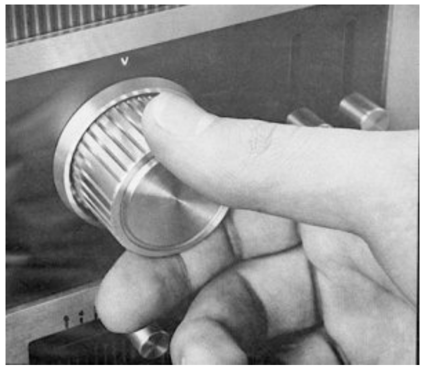


γ

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

The Solution Path

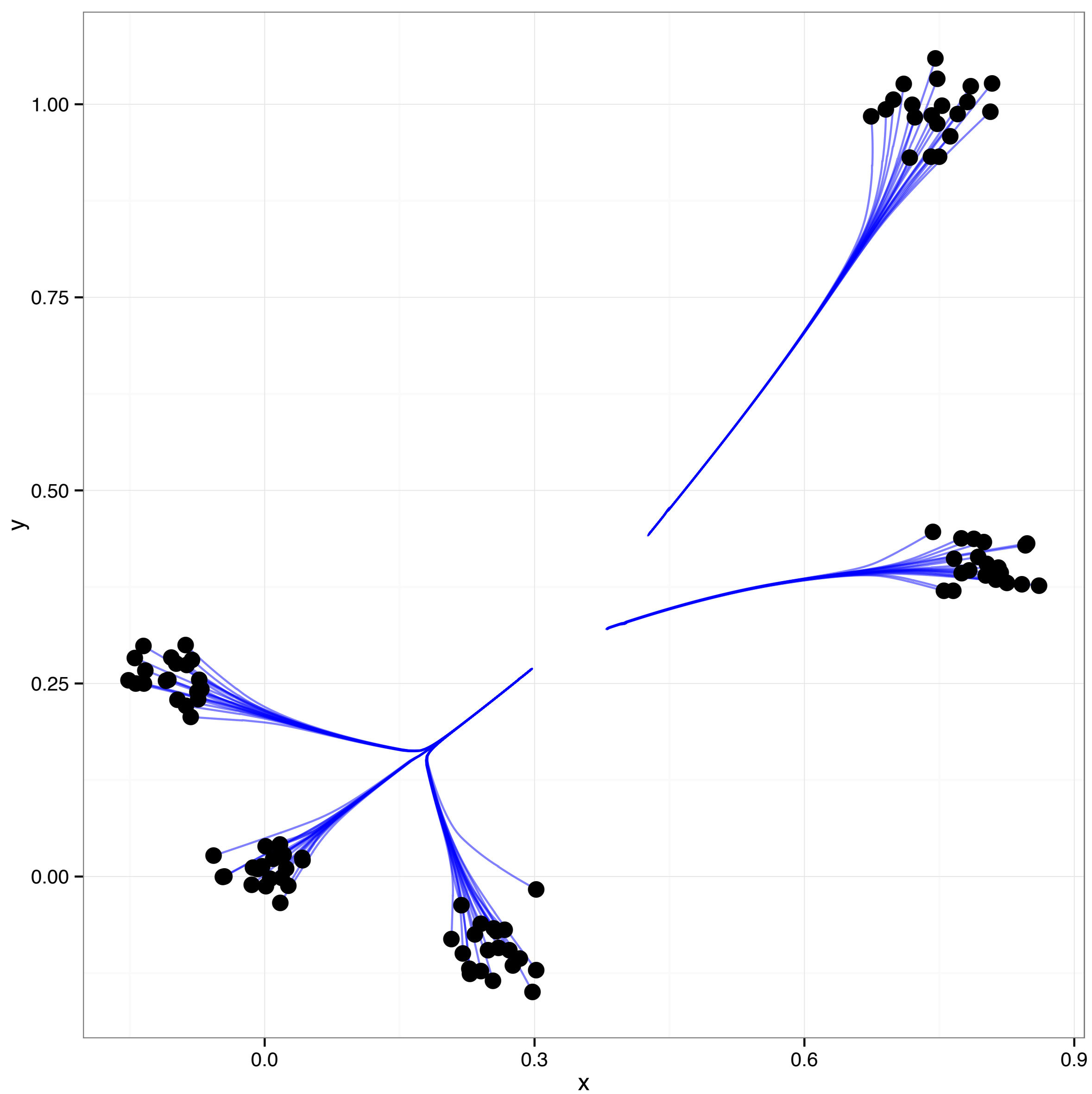


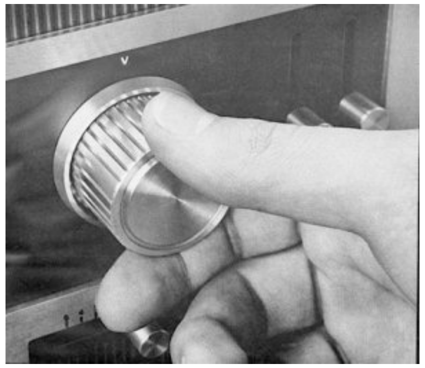


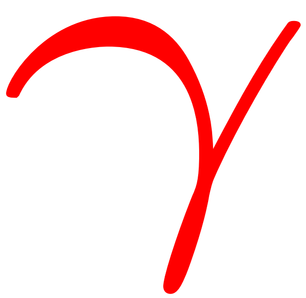
γ

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

The Solution Path

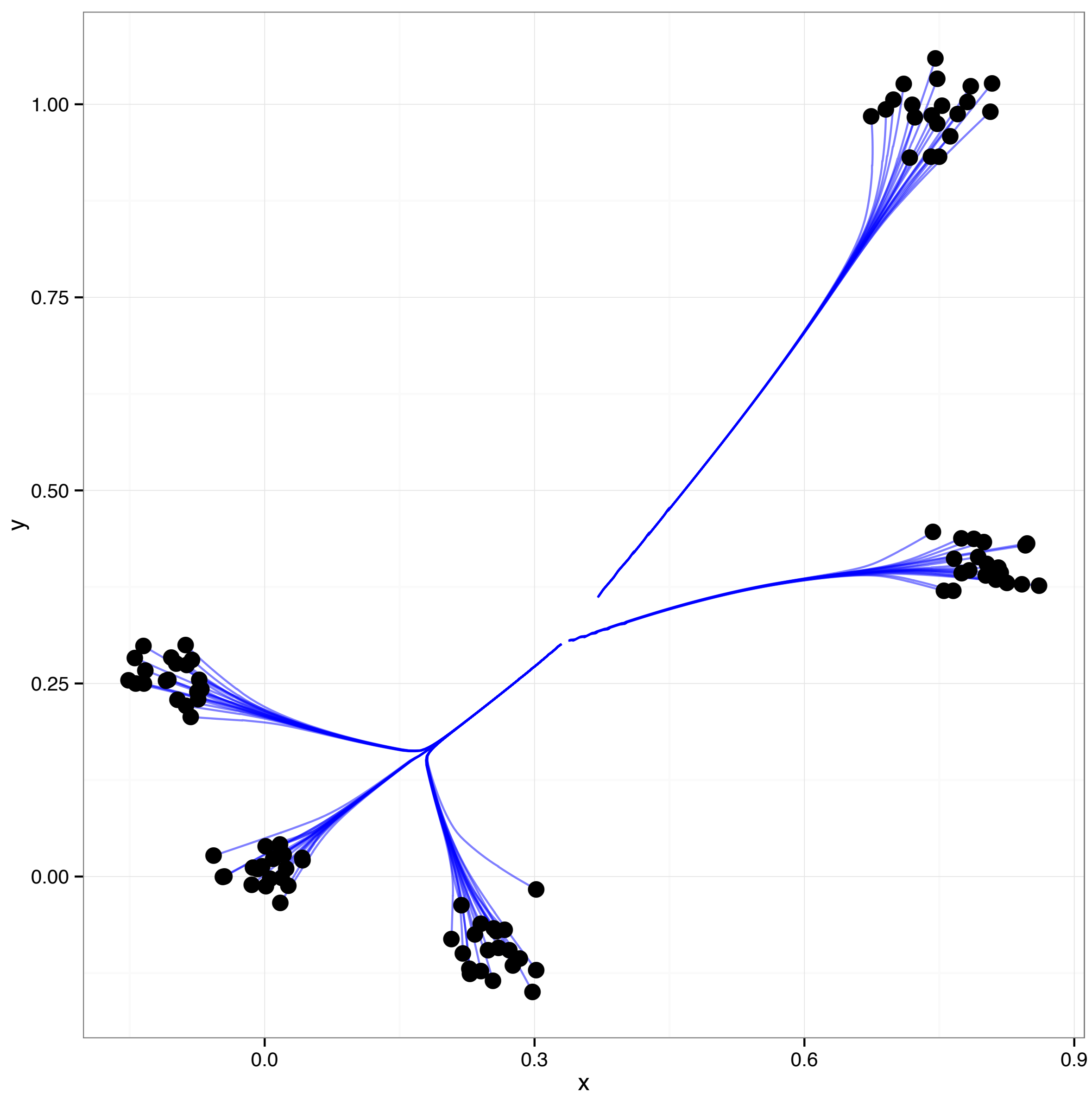


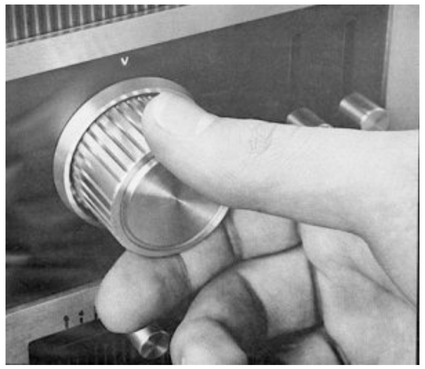




$$\text{minimize } \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

The Solution Path

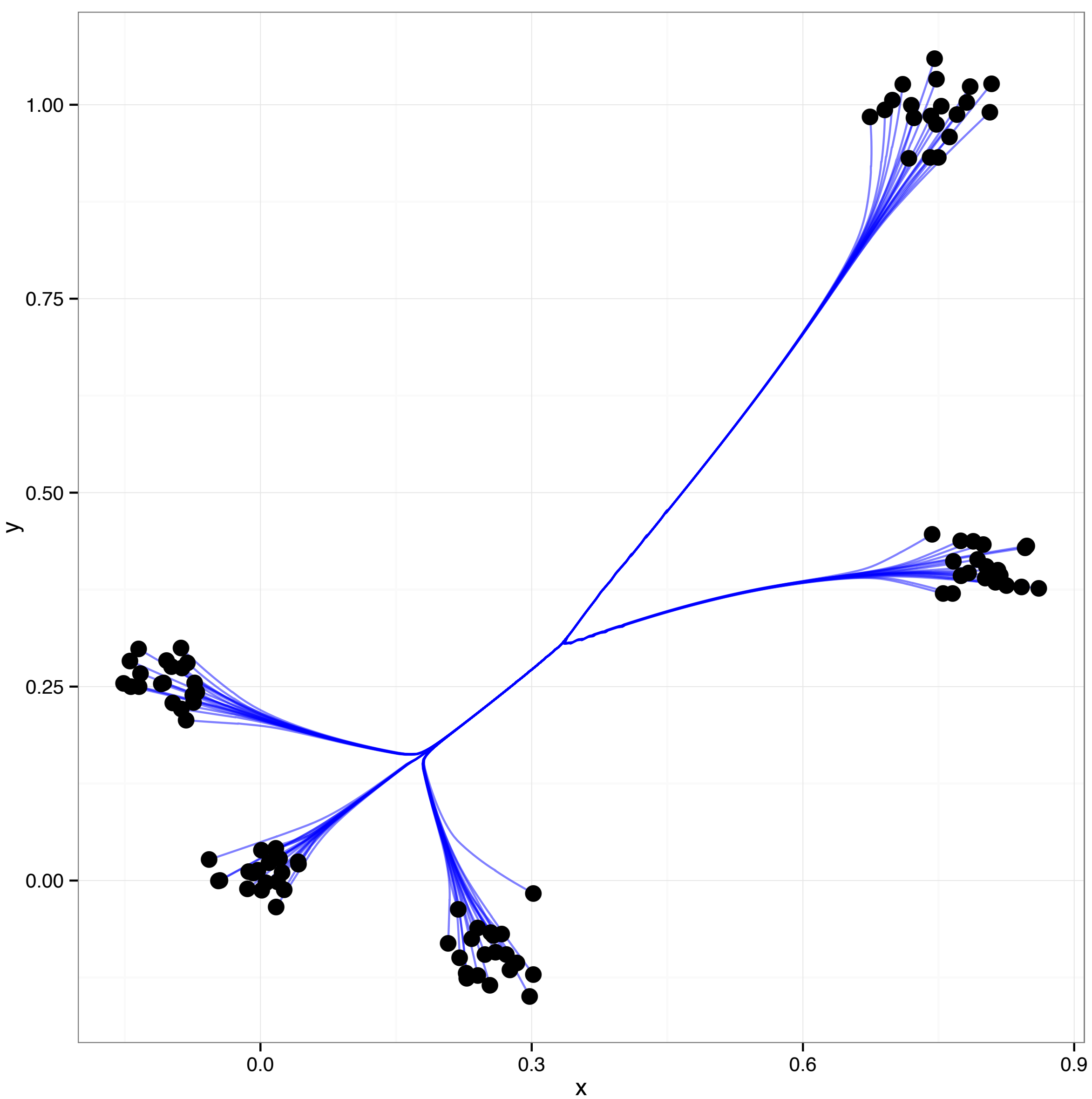


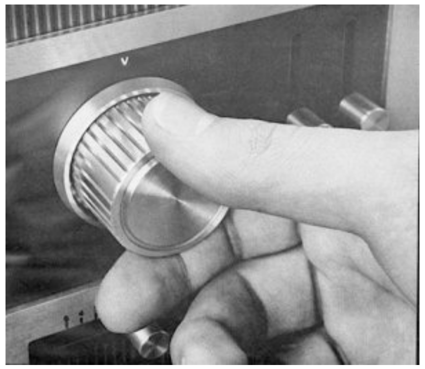


γ

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

The Solution Path

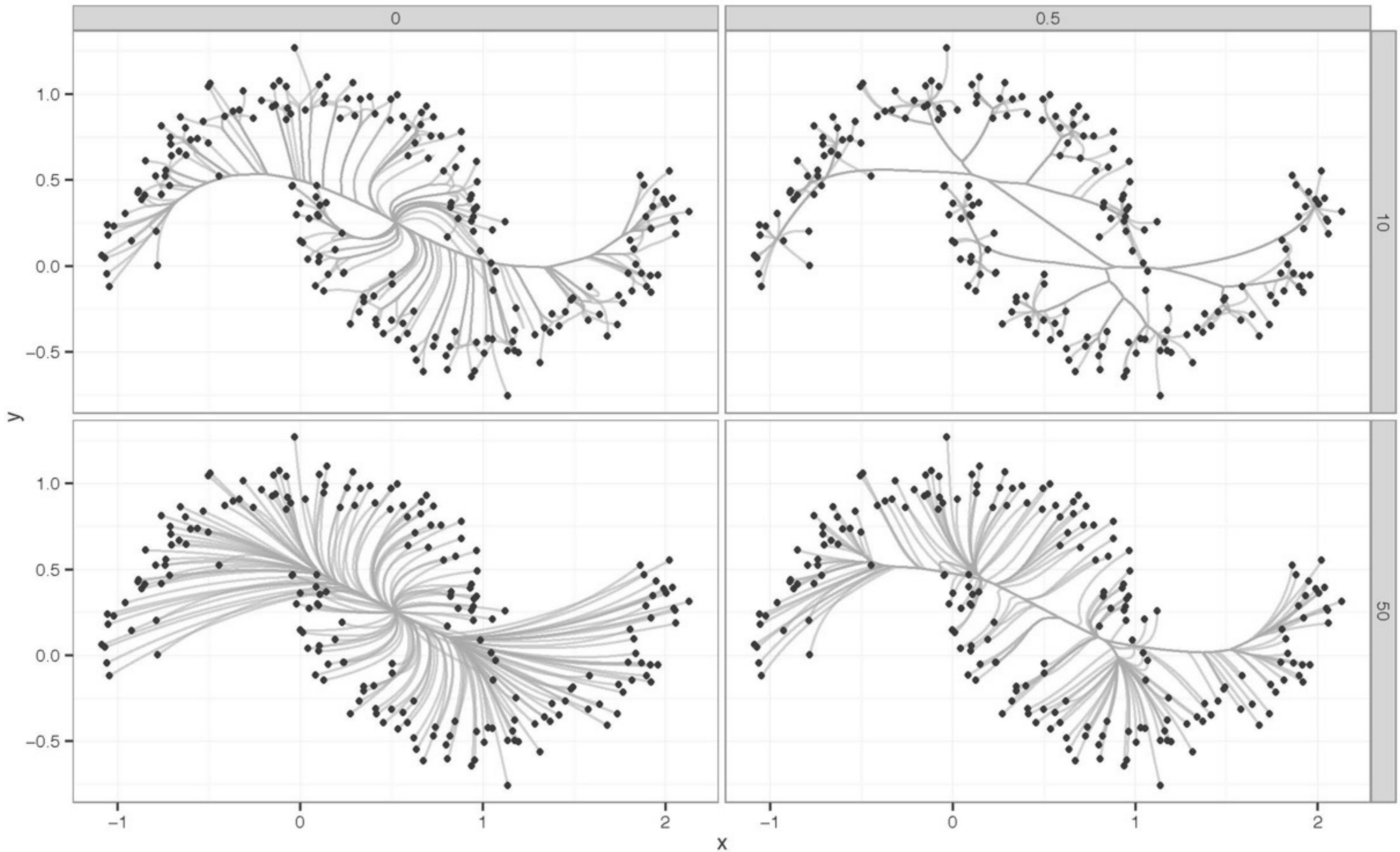




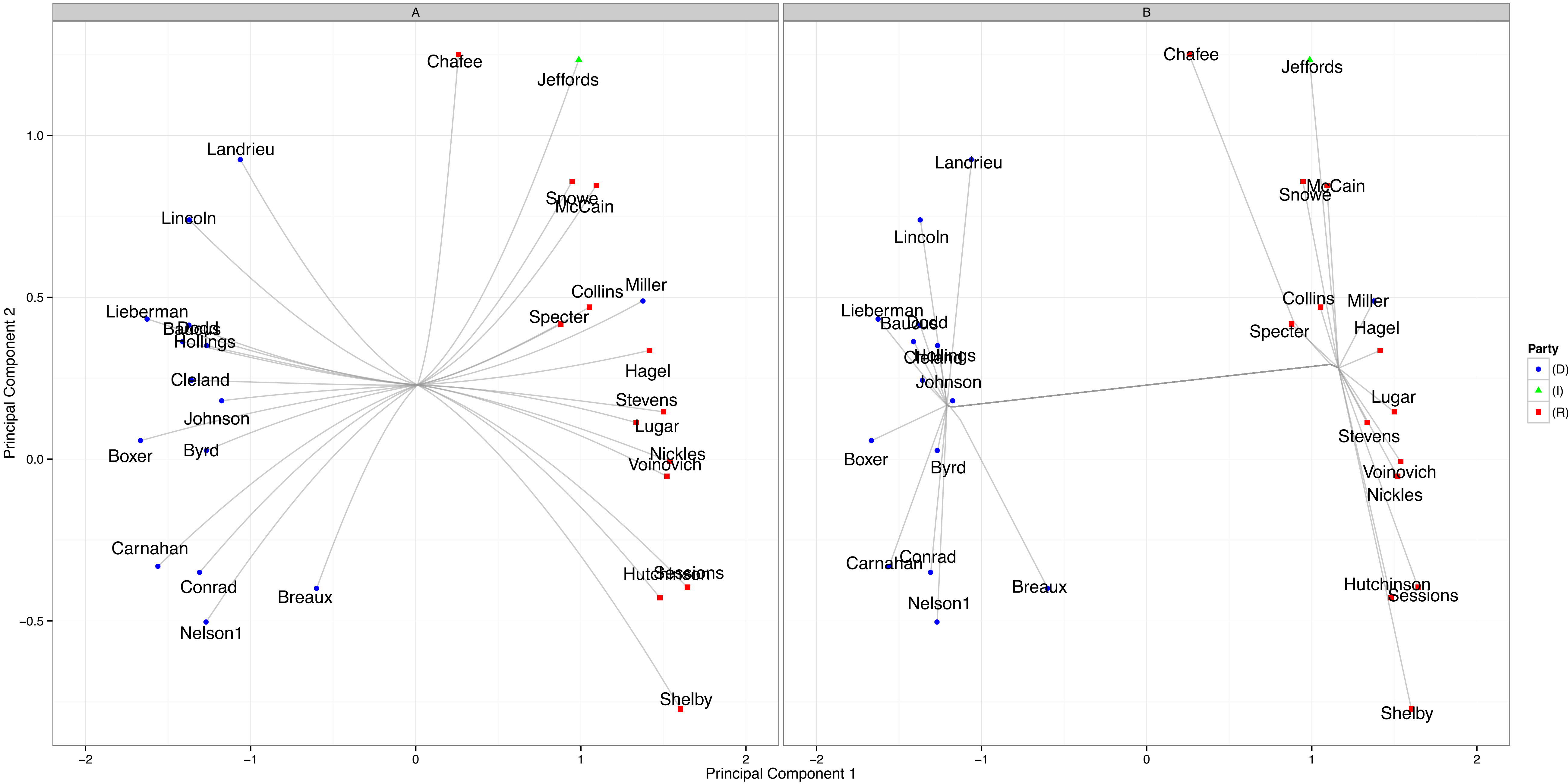
γ

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

Two Interlocking Half-Moons



Senate Voting



Apparently Non-Trivial Optimization Problem

Why is this hard to solve?

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

Apparently Non-Trivial Optimization Problem

Why is this hard to solve?

$$\text{minimize}\ \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

Nonsmooth? Not the issue

Apparently Non-Trivial Optimization Problem

Why is this hard to solve?

$$\text{minimize}\ \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

Affine transformation of u

Apparently Non-Trivial Optimization Problem

Why is this hard to solve?

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{i<j} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2$$

General Recipe:

1. Introduce a dummy variable

$$\text{unconstrained} \rightarrow \text{equality constrained}$$

2. Use iterative method to solve equality constrained version

Convex Clustering: Variable Split Version

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{p} \|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{l} w_l \|\mathbf{v}_l\|$$

$$\text{subject to } \mathbf{u}_{l_1} - \mathbf{u}_{l_2} - \mathbf{v}_l = \mathbf{0}$$

$$l = (l_1, l_2) \text{ with } l_1 < l_2.$$

Equality constrained optimization...

Convex Clustering: Variable Split Version

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{p}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{l} w_l \|\mathbf{v}_l\|$$

subject to $\mathbf{u}_{l_1} - \mathbf{u}_{l_2} - \mathbf{v}_l = \mathbf{0}$

$l = (l_1, l_2)$ with $l_1 < l_2$.

Convex Clustering: Variable Split Version

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{p} \|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \gamma \sum_{l} w_l \|\mathbf{v}_l\|$$

$$\text{subject to } \mathbf{u}_{l_1} - \mathbf{u}_{l_2} - \mathbf{v}_l = \mathbf{0}$$

$l = (l_1, l_2)$ with $l_1 < l_2$.

Lagrange Multipliers

Lagrange Multipliers

$$\begin{aligned}&\text{minimize } f(\mathbf{u}) + g(\mathbf{v})\\&\text{subject to } \mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{v} = \mathbf{c},\end{aligned}$$

$$\mathcal{L}(\mathbf{u}, \mathbf{v}, \boldsymbol{\lambda}) \quad = \quad f(\mathbf{u}) + g(\mathbf{v}) + \langle \boldsymbol{\lambda}, \mathbf{c} - \mathbf{A}\mathbf{u} - \mathbf{B}\mathbf{v} \rangle$$

$$\nabla \mathcal{L}(\mathbf{u}^\star, \mathbf{v}^\star, \boldsymbol{\lambda}^\star) = \mathbf{0}.$$

$$(\mathbf{u}^\star, \mathbf{v}^\star) = \arg\min_{\mathbf{u},\mathbf{v}} \mathcal{L}(\mathbf{u}, \mathbf{v}, \boldsymbol{\lambda}^\star)$$

Typically need to solve this iteratively.

Augmented Lagrangian Method

$$\begin{aligned} &\text{minimize } f(\mathbf{u}) + g(\mathbf{v}) \\ &\text{subject to } \mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{v} = \mathbf{c}, \end{aligned}$$

ALM solves the equivalent problem

$$\begin{aligned} &\text{minimize } f(\mathbf{u}) + g(\mathbf{v}) + \frac{\nu}{2}\|\mathbf{c} - \mathbf{A}\mathbf{u} - \mathbf{B}\mathbf{v}\|_2^2, \\ &\text{subject to } \mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{v} = \mathbf{c} \end{aligned}$$

ALM: Augmented Lagrangian Method

ALM solves the equivalent problem

$$\text{minimize } f(\mathbf{u}) + g(\mathbf{v}) + \frac{\nu}{2}\|\mathbf{c} - \mathbf{A}\mathbf{u} - \mathbf{B}\mathbf{v}\|_2^2,$$

$$\text{subject to } \mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{v} = \mathbf{c}$$

The Augmented Lagrangian

$$\mathcal{L}_{\nu}(\mathbf{u}, \mathbf{v}, \boldsymbol{\lambda}) = f(\mathbf{u}) + g(\mathbf{v}) + \langle \boldsymbol{\lambda}, \mathbf{c} - \mathbf{A}\mathbf{u} - \mathbf{B}\mathbf{v}\rangle + \frac{\nu}{2}\|\mathbf{c} - \mathbf{A}\mathbf{u} - \mathbf{B}\mathbf{v}\|_2^2$$

ALM Updates

$$(\mathbf{u}^{m+1}, \mathbf{v}^{m+1}) = \arg\min_{\mathbf{u},\mathbf{v}} \mathcal{L}_{\nu}(\mathbf{u}, \mathbf{v}, \boldsymbol{\lambda}^m)$$

$$\boldsymbol{\lambda}^{m+1} = \boldsymbol{\lambda}^m + \nu(\mathbf{c} - \mathbf{A}\mathbf{u}^{m+1} - \mathbf{B}\mathbf{v}^{m+1}).$$

ALM: Augmented Lagrangian Method

ALM Updates

$$(\mathbf{u}^{m+1}, \mathbf{v}^{m+1}) = \arg\min_{\mathbf{u},\mathbf{v}} \mathcal{L}_\nu(\mathbf{u}, \mathbf{v}, \boldsymbol{\lambda}^m) \leftarrow \text{Often hard}$$

$$\boldsymbol{\lambda}^{m+1} = \boldsymbol{\lambda}^m + \nu(\mathbf{c} - \mathbf{A}\mathbf{u}^{m+1} - \mathbf{B}\mathbf{v}^{m+1}).$$

ALM: Augmented Lagrangian Method

ALM Updates

$$(\mathbf{u}^{m+1}, \mathbf{v}^{m+1}) = \arg\min_{\mathbf{u},\mathbf{v}} \mathcal{L}_\nu(\mathbf{u}, \mathbf{v}, \boldsymbol{\lambda}^m) \longleftarrow \text{Often hard}$$

$$\boldsymbol{\lambda}^{m+1} = \boldsymbol{\lambda}^m + \nu(\mathbf{c} - \mathbf{A}\mathbf{u}^{m+1} - \mathbf{B}\mathbf{v}^{m+1}).$$

1. Alternating Direction Method of Multipliers (ADMM) (Gabay & Mercier 1976, Glowinski & Marrocco 1975)

2. Alternating Minimization Algorithm (AMA) (Tseng 1991)

ADMM: Alternating Direction Method of Multipliers

ALM Updates

$$(\mathbf{u}^{m+1}, \mathbf{v}^{m+1}) = \arg\min_{\mathbf{u},\mathbf{v}} \mathcal{L}_\nu(\mathbf{u}, \mathbf{v}, \boldsymbol{\lambda}^m) \leftarrow \text{Often hard}$$

$$\boldsymbol{\lambda}^{m+1} = \boldsymbol{\lambda}^m + \nu(\mathbf{c} - \mathbf{A}\mathbf{u}^{m+1} - \mathbf{B}\mathbf{v}^{m+1}).$$

ADMM Updates

$$\mathbf{u}^{m+1} = \arg\min_{\mathbf{u}} \mathcal{L}_\nu(\mathbf{u}, \mathbf{v}^m, \boldsymbol{\lambda}^m)$$

$$\mathbf{v}^{m+1} = \arg\min_{\mathbf{v}} \mathcal{L}_\nu(\mathbf{u}^{m+1}, \mathbf{v}, \boldsymbol{\lambda}^m)$$

$$\boldsymbol{\lambda}^{m+1} = \boldsymbol{\lambda}^m + \nu(\mathbf{c} - \mathbf{A}\mathbf{u}^{m+1} - \mathbf{B}\mathbf{v}^{m+1}).$$

Goal: Simpler algorithms

AMA: Alternating Minimization Algorithm

ALM Updates

$$(\mathbf{u}^{m+1}, \mathbf{v}^{m+1}) = \arg\min_{\mathbf{u},\mathbf{v}} \mathcal{L}_\nu(\mathbf{u}, \mathbf{v}, \boldsymbol{\lambda}^m) \longleftarrow \text{Often hard}$$

$$\boldsymbol{\lambda}^{m+1} = \boldsymbol{\lambda}^m + \nu(\mathbf{c} - \mathbf{A}\mathbf{u}^{m+1} - \mathbf{B}\mathbf{v}^{m+1}).$$

AMA Updates

$$\mathbf{u}^{m+1} = \arg\min_{\mathbf{u}} \mathcal{L}_0(\mathbf{u}, \mathbf{v}^m, \boldsymbol{\lambda}^m)$$

$$\mathbf{v}^{m+1} = \arg\min_{\mathbf{v}} \mathcal{L}_\nu(\mathbf{u}^{m+1}, \mathbf{v}, \boldsymbol{\lambda}^m)$$

$$\boldsymbol{\lambda}^{m+1} = \boldsymbol{\lambda}^m + \nu(\mathbf{c} - \mathbf{A}\mathbf{u}^{m+1} - \mathbf{B}\mathbf{v}^{m+1}).$$

Goal: Simpler algorithms

ADMM Updates

$$\begin{aligned} \mathbf{u}_i &= \frac{1}{1+p\nu}\mathbf{y}_i + \frac{p\nu}{1+p\nu}\bar{\mathbf{x}} \\ \mathbf{y}_i &= \mathbf{x}_i + \sum_{l_1=i}[\boldsymbol{\lambda}_l + \nu\mathbf{v}_l] - \sum_{l_2=i}[\boldsymbol{\lambda}_l + \nu\mathbf{v}_l]. \end{aligned}$$

$$\begin{aligned} \mathbf{v}_l &= \arg\min_{\mathbf{v}} \frac{1}{2}\|\mathbf{v} - (\mathbf{u}_{l_1} - \mathbf{u}_{l_2} - \nu^{-1}\boldsymbol{\lambda}_l)\|_2^2 + \frac{\gamma w_l}{\nu}\|\mathbf{v}\| \\ &= \mathrm{prox}_{\sigma_l\|\cdot\|/\nu}(\mathbf{u}_{l_1} - \mathbf{u}_{l_2} - \nu^{-1}\boldsymbol{\lambda}_l), \end{aligned}$$

where $\sigma_l = \gamma w_l$.

$$\boldsymbol{\lambda}_l = \boldsymbol{\lambda}_l + \nu(\mathbf{v}_l - \mathbf{u}_{l_1} + \mathbf{u}_{l_2}).$$

AMA Updates

$$\mathbf{u}_i = \frac{1}{1+p0}\mathbf{y}_i + \frac{p0}{1+p0}\bar{\mathbf{x}}$$

$$\mathbf{y}_i = \mathbf{x}_i + \sum_{l_1=i}[\boldsymbol{\lambda}_l + 0\mathbf{v}_l] - \sum_{l_2=i}[\boldsymbol{\lambda}_l + 0\mathbf{v}_l].$$

$$\mathbf{v}_l = \arg\min_{\mathbf{v}} \frac{1}{2}\|\mathbf{v} - (\mathbf{u}_{l_1} - \mathbf{u}_{l_2} - \nu^{-1}\boldsymbol{\lambda}_l)\|_2^2 + \frac{\gamma w_l}{\nu}\|\mathbf{v}\|$$
$$= \mathrm{prox}_{\sigma_l\|\cdot\|/\nu}(\mathbf{u}_{l_1} - \mathbf{u}_{l_2} - \nu^{-1}\boldsymbol{\lambda}_l),$$

where $\sigma_l = \gamma w_l$.

$$\boldsymbol{\lambda}_l = \boldsymbol{\lambda}_l + \nu(\mathbf{v}_l - \mathbf{u}_{l_1} + \mathbf{u}_{l_2}).$$

AMA Updates

$$\mathbf{u}_i = \mathbf{x}_i + \sum_{l_1 = i} \boldsymbol{\lambda}_l - \sum_{l_2 = i} \boldsymbol{\lambda}_l$$

$$\begin{aligned}\mathbf{v}_l &= \arg\min_{\mathbf{v}} \frac{1}{2}\|\mathbf{v} - (\mathbf{u}_{l_1} - \mathbf{u}_{l_2} - \nu^{-1}\boldsymbol{\lambda}_l)\|_2^2 + \frac{\gamma w_l}{\nu}\|\mathbf{v}\| \\ &= \mathrm{prox}_{\sigma_l\|\cdot\|/\nu}(\mathbf{u}_{l_1} - \mathbf{u}_{l_2} - \nu^{-1}\boldsymbol{\lambda}_l),\end{aligned}$$

where $\sigma_l = \gamma w_l$.

$$\boldsymbol{\lambda}_l = \boldsymbol{\lambda}_l + \nu(\mathbf{v}_l - \mathbf{u}_{l_1} + \mathbf{u}_{l_2}).$$

Proximal Map

For $\sigma > 0$ the function

$$\mathrm{prox}_{\sigma\Omega}(\mathbf{v}) = \arg\min_{\tilde{\mathbf{v}}} \left[\sigma\Omega(\tilde{\mathbf{v}}) + \frac{1}{2}\|\mathbf{v} - \tilde{\mathbf{v}}\|_2^2\right]$$

is the proximal map of the function $\Omega(\mathbf{v})$.

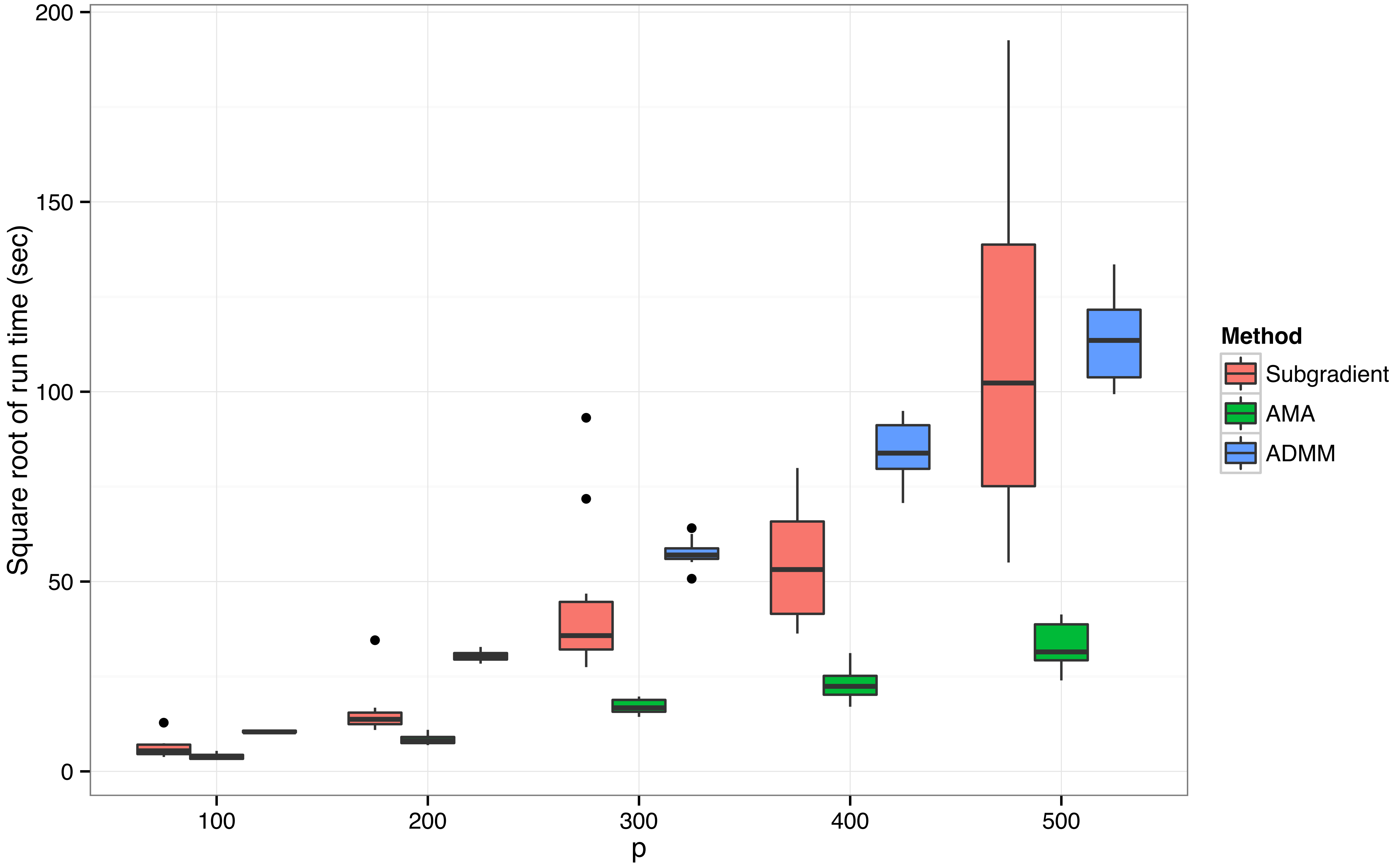
Minimizer always exists and is unique for norms

Proximal maps for common norms

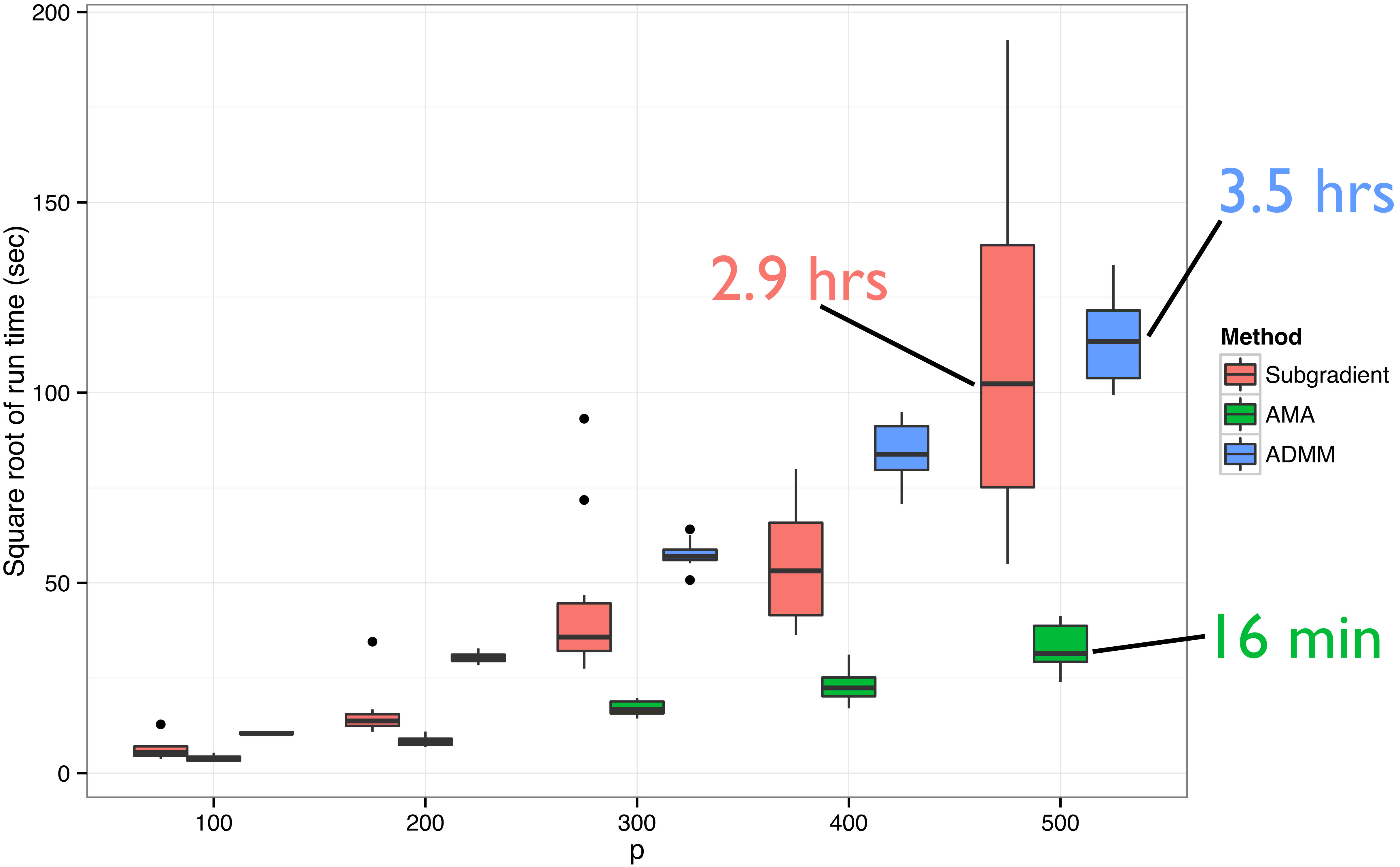
Table: Proximal maps for common norms.

Norm	$\Omega(\mathbf{v})$	$\mathrm{prox}_{\sigma\Omega}(\mathbf{v})$
ℓ_1	$\lVert\mathbf{v}\rVert_1$	$\left[1 - \frac{\sigma}{\lvert v_l\rvert}\right]_+ v_l$
ℓ_2	$\lVert\mathbf{v}\rVert_2$	$\left[1 - \frac{\sigma}{\lVert\mathbf{v}\rVert_2}\right]_+ \mathbf{v}$
ℓ_∞	$\lVert\mathbf{v}\rVert_\infty$	$\mathbf{v} - \mathcal{P}_{\sigma S}(\mathbf{v})$
$\ell_{1,2}$	$\sum_{g\in\mathcal{G}} \lVert\mathbf{v}_g\rVert_2$	$\left[1 - \frac{\sigma}{\lVert\mathbf{v}_g\rVert_2}\right]_+ \mathbf{v}_g$

What's the Difference?



What's the Difference?



Remarks

- Both AMA and ADMM converge
- Both AMA and ADMM can be accelerated
  - Beck and Teboulle (2009)
  - Goldstein, O'Donoghue, and Setzer (2012)
- AMA and ADMM look very similar but...
  - Convergence speed
    - AMA is clearly faster
  - Convergence
    - ADMM converges when $\nu > 0$
    - AMA converges when $\nu \leq 1/p$
  - AMA requires stronger assumptions
    - Smooth part of objective needs to be strongly convex

ADMM solver for Lasso

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \gamma\|\boldsymbol{\theta}\|_1$$

ADMM solver for Lasso

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \gamma\|\mathbf{v}\|_1 \quad \text{subject to} \quad \boldsymbol{\theta} = \mathbf{v},$$

ADMM solver for Lasso

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \gamma\|\mathbf{v}\|_1 \quad \text{subject to} \quad \boldsymbol{\theta} = \mathbf{v},$$

Augmented Lagrangian

$$\mathcal{L}(\boldsymbol{\theta}, \mathbf{v}, \boldsymbol{\lambda}) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \gamma\|\mathbf{v}\|_1 + \frac{\nu}{2}\|\boldsymbol{\theta} - \mathbf{v} + \boldsymbol{\lambda}\|_2^2.$$

ADMM solver for Lasso

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \gamma\|\mathbf{v}\|_1 \quad \text{subject to} \quad \boldsymbol{\theta} = \mathbf{v},$$

Augmented Lagrangian

$$\mathcal{L}(\boldsymbol{\theta}, \mathbf{v}, \boldsymbol{\lambda}) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \gamma\|\mathbf{v}\|_1 + \frac{\nu}{2}\|\boldsymbol{\theta} - \mathbf{v} + \boldsymbol{\lambda}\|_2^2.$$

ADMM Updates

$$\begin{aligned}
\boldsymbol{\theta}^k &= \underset{\boldsymbol{\theta}}{\text{minimize}} \quad \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \frac{\nu}{2}\|\boldsymbol{\theta} - \mathbf{v}^{k-1} + \boldsymbol{\lambda}^{k-1}\|_2^2. \\
\mathbf{v}^k &= \underset{\mathbf{v}}{\text{minimize}} \quad \gamma\|\mathbf{v}\|_1 + \frac{\nu}{2}\|\mathbf{v} - \boldsymbol{\theta}^k - \boldsymbol{\lambda}^{k-1}\|_2^2. \\
\boldsymbol{\lambda}^k &= \boldsymbol{\lambda}^{k-1} + \boldsymbol{\theta}^k - \mathbf{v}^k.
\end{aligned}$$

Getting started

- Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011), "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Found. Trends Mach. Learn.*, 3, 1-122.
- Tseng, P. (1991), "Applications of a Splitting Algorithm to Decomposition in Convex Programming and Variational Inequalities," *SIAM Journal on Control and Optimization*, 29, 119-138.