Conjugate Interior Point Methods for Large-Scale Problems

Aleksandr Y. Aravkin
UW Applied Math

Happy Birthday, Jim!!
WCOM, Seattle
May 14, 2016

Outline

- Modeling with Piecewise linear-quadratic functions
- PLQ class, theory, and calclulus (with J.V. Burke and G. Pillonetto)
- Applications with IPsolve
  - Dynamic systems inference (with J.V. Burke, G. Pillonetto, and B. Bell)
  - Quantile Huber (with A. Lozano, R. Luss, A. Kambadur)
  - Sparse difference graphs (with D. Orban, H. Liu, R. Vanderbei, C. Eisenach, Y. Liu).
  - System identification (with J.V. Burke and G. Pillonetto)
  - Meta-parameters (with P. Zheng, K. Ramamurthy, and JJ Thiagarajan).
- Code: `https://github.com/saravkin/IPsolve`

Optimization in Modeling

$$\begin{aligned} \min_x \quad & V(Ax - b) + J(Gx - w) \\ \text{s.t.} \quad & x \in \mathcal{X}. \end{aligned}$$

- Data misfit V: good results in the face of *large measurement errors*
- Regularization J: *prior information* e.g. sparsity or process model
- Constraints: use information about feasible region

Here, V, J can come from a large class of *piecewise linear quadratic* (PLQ) penalties.

PLQ penalties in practice

Application	Objective	PLQs
Regression	$\|Ax - b\|^2$	ℓ_2^2
Robust regression	$\rho(Ax - b)$	huber
Lasso	$\|Ax - b\|^2 + \lambda\|x\|_1$	$\ell_2^2 + \ell_1$
SVM	$H(\mathbf{1} - Ax) + \frac{1}{2}\|w\|^2$	hinge loss + ℓ_2^2
CVaR	$\frac{1}{N(1-\beta)}H(Ax - \alpha e) + \alpha + \delta_\Delta(x)$	hinge + aff + const
System ID	$\|z - \Phi x\|^2 + \gamma x^T Q^{-1} x$	$\ell_2^2 + \ell_2^2$
Kalman smoother	$\|Hx - z\|_{R^{-1}}^2 + \|Gx - w\|_{Q^{-1}}^2$	$\ell_2^2 + \ell_2^2$

Let's take a closer look at these PLQs, and their conjugates.

Quadratic

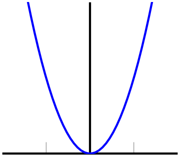


Figure: $\frac{1}{2}x^2 = \sup_{u \in \mathbb{R}} \left\{ ux - \frac{1}{2}u^2 \right\}.$

- linear and nonlinear regression [Fre09, SW03]
- inverse problems [Tik].
- Features: symmetric, smooth, quadratic tail growth.

1- norm

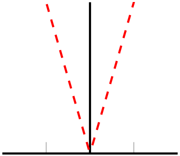


Figure: $|x| = \sup_{u \in [-1,1]} \{ux\}$.

- machine learning and compressed sensing [HT90, EHJT04, Don06],
- image denoising [SED05, MES08, MSNW10],
- seismic image processing [HH08, NKK$^+$10, HFY12, MWLH12].
- Features: symmetric, nonmsooth, linear tail growth

hinge loss

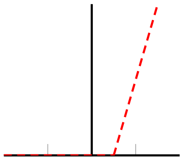


Figure: $h_\epsilon(x) = \sup_{u \in [0,1]} \{u(x - \epsilon)\}$.

- support vector classifiers [EPP00, PV98, SSWB00].
- Conditional Value at Risk and Superquantiles [RU00]
- Features: 1-sided, nonmsmooth, linear tail growth.

quantile loss

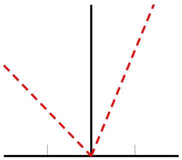


Figure: $q_\tau(x) = \sup_{u \in [-\tau,(1-\tau)]} \{ux\}$.

- heterogeneous datasets [KB78, Buc94]
- computational biology [ZY08]
- survival analysis [KG01]
- economics [KH01, Koe05]
- Features: asymmetric, nonsmooth, linear tail growth

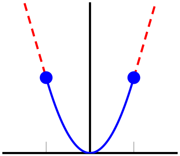


Figure: $h_\kappa(x) = \sup_{u \in [-\kappa, \kappa]} \left\{ ux - \frac{1}{2}u^2 \right\}$.

- robust regression [Hub04, MMY06, BN07, DH81, Cla85, LS98]
- Kalman smoothing [ABP13a]
- System identification [ABP13b]
- robust PCA and robust matrix completion [AKM+14, ABD+16]
- Features: symmetric, smooth, linear tail growth.

quantile huber loss

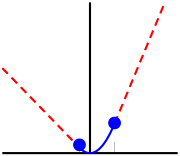


Figure: $h_{\tau,\kappa}(x) = \sup_{u \in [-\kappa\tau, \kappa(1-\tau)]} \left\{ ux - \frac{1}{2}u^2 \right\}.$

- alternative to the quantile penalty in the high-dimensional setting [AKLL14]
- We will come back to this example
- Features: asymmetric, smooth, linear tail growth.

vapnik loss

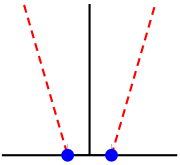


Figure: $\rho_\epsilon(x) = \sup_{u \in [0,1]^2} \left\{ \left\langle \begin{bmatrix} 1 \\ -1 \end{bmatrix} x - \begin{bmatrix} \epsilon \\ \epsilon \end{bmatrix}, u \right\rangle \right\}.$

- support vector regression (SVR) [Vap98, HTF01, SSWB00, SS01]

- Features: ‘deadzone’, symmetric, nonsmooth, linear tail growth.

elastic net

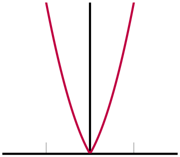


Figure: $\rho(x) = \sup_{u \in [0,1]\times\mathbb{R}} \left\{ \left\langle \begin{bmatrix} 1 \\ 1 \end{bmatrix} x, u \right\rangle - \frac{1}{2} u^T \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} u \right\}.$

- sparse regularization with correlated predictors [ZH05b, ZH05a, LL$^+$10, DMDVR09].

- Features: symmetric, nonsmooth at origin, quadratic tail growth.

PLQ Functions (Rockafellar & Wets)

Explicit encoding of PLQ functions

Define $\rho(C, c, M, b, B; \cdot) : \mathbb{R}^n \to \mathbb{R}$ as

$$\rho(C, c, b, B, M; y) = \sup_{Cu \leq c} \left\{ \langle u, b + By \rangle - \frac{1}{2} \langle u, Mu \rangle \right\}$$

1. $M \in \mathbb{R}^{m \times m}$ is a symmetric positive semidefinite matrix.
2. $b + By$ is an injective affine transformation with $B \in \mathbb{R}^{m \times n}$.
3. $\{u | Cu \leq c\} \subset \mathbb{R}^m$ is a polyhedral containing the origin.

The encoding makes it possible to build general problems from component parts (V, J, affine compositions).

PLQ Representation Calculus

- (Addition)Given two PLQ penalties

$$\rho(c_1, C_1, B_1, b_1, M_1; y) \qquad \text{and} \qquad \rho(c_2, C_2, B_2, b_2, M_2; y)$$

their sum is also a PLQ penalty $\rho(c, C, B, b, M; y)$ with

$$c = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}, \quad C = \begin{bmatrix} C_1 & 0 \\ 0 & C_2 \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad M = \begin{bmatrix} M_1 & 0 \\ 0 & M_2 \end{bmatrix}$$

PLQ Representation Calculus

- (Addition)Given two PLQ penalties

$$\rho(c_1, C_1, B_1, b_1, M_1; y) \qquad \text{and} \qquad \rho(c_2, C_2, B_2, b_2, M_2; y)$$

their sum is also a PLQ penalty $\rho(c, C, B, b, M; y)$ with

$$c = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}, \quad C = \begin{bmatrix} C_1 & 0 \\ 0 & C_2 \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad M = \begin{bmatrix} M_1 & 0 \\ 0 & M_2 \end{bmatrix}$$

- (Affine composition) Given a PLQ penalty $\rho(c, C, b, B, M; y)$, we have

$$\rho(Px - p) = \rho(c, C, b - Bp, BP, M; y) \ .$$

PLQ Representation Calculus

- (Addition)Given two PLQ penalties

$$\rho(c_1, C_1, B_1, b_1, M_1; y) \qquad \text{and} \qquad \rho(c_2, C_2, B_2, b_2, M_2; y)$$

their sum is also a PLQ penalty $\rho(c, C, B, b, M; y)$ with

$$c = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}, \quad C = \begin{bmatrix} C_1 & 0 \\ 0 & C_2 \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad M = \begin{bmatrix} M_1 & 0 \\ 0 & M_2 \end{bmatrix}$$

- (Affine composition) Given a PLQ penalty $\rho(c, C, b, B, M; y)$, we have

$$\rho(Px - p) = \rho(c, C, b - Bp, BP, M; y) \ .$$

- (Smoothing) Smoothing with quadratics preserves PLQ.

$$\rho_\gamma(c, C, M, b, B; y) = \rho(c, C, M + \gamma I, b, B; y)$$

Moreau envelope (Burke & Hoheisel 2013):

$$e_\gamma \rho(c, C, M, b, B; y) = \rho(c, C, M + \gamma BB^T, b, B; y)$$

PLQ Optimization with constraints

Consider now the minimization problem

$$\min_y \rho(c, C, b, B, M; y) \qquad \text{s.t.} \quad Ay \le a.$$

Introduce slack variables s and r:

$$Cu + s = c, \quad Ay + r = a \, .$$

Let q, w be dual variables corresponding to these constraints. The KKT system (optimality conditions) is given by

$$\begin{aligned} 0 &= B^T u + A^T w \\ 0 &= By - Mu - C^T q + b \\ 0 &= Cu + s - c \\ 0 &= Ay + r - a \\ 0 &= q_i s_i \; \forall i \; , \; q, s \ge 0 \\ 0 &= w_i r_i \; \forall i \; , \; w, r \ge 0 \, . \end{aligned}$$

From here, implement an interior point method (see Kojima et al., 1991; Nemirovskii and Nesterov, 1994; Wright, 1997.)

IPsolve: Interior Point via Conjugate Representation

Interior point methods *relax the complementarity conditions*:

$$q_i s_i = \mu \ , \quad w_i r_i = \mu \ ,$$

and μ is aggressively taken to 0, so at the end the true KKT system is (nearly) satisfied. The μ-relaxed KKT system looks like this:

$$F_\mu(\underbrace{s, q, u, r, w, y}_{z}) = \begin{bmatrix} s + Cu - c \\ QS\mathbf{1} - \mu\mathbf{1} \\ By - Mu - C^T q + b \\ r + Ay - a \\ WR\mathbf{1} - \mu\mathbf{1} \\ B^T u + A^T w \end{bmatrix} .$$

To solve, take damped Newton iterations: $\mathbf{F}_\mu^{(1)} \mathbf{\Delta z} = -\mathbf{F}_\mu$, where

$$F_\mu^{(1)} := \begin{bmatrix} I & 0 & C & 0 & 0 & 0 \\ Q & S & 0 & 0 & 0 & 0 \\ 0 & -C^T & -M & 0 & 0 & B \\ 0 & 0 & 0 & I & 0 & A \\ 0 & 0 & 0 & W & R & 0 \\ 0 & 0 & B^T & 0 & A^T & 0 \end{bmatrix} .$$

We can exploit problem structure to make each **Newton** step efficient.

Aravkin, Burke, Pillonetto, JMLR 2013

Exploiting Sparsity of Conjugate Representation:

Instead, can put $c - Cu$ directly into barrier, then take

$$q_i = \mu/(c_i - \langle C_i, u\rangle)\ , \quad w_i r_i = \mu\ ,$$

The new μ-relaxed KKT system looks like this:

$$F_\mu(\underbrace{q, u, r, w, y}_{z}) = \begin{bmatrix} q\text{diag}(c - Cu) - \mu\mathbf{1} \\ By - Mu - C^T q + b \\ WR\mathbf{1} - \mu\mathbf{1} \\ r + Ay - a \\ B^T u + A^T w \end{bmatrix} .$$

To solve, take damped Newton iterations: $\mathbf{F}_\mu^{(\mathbf{1})}\boldsymbol{\Delta}\mathbf{z} = -\mathbf{F}_\mu$, where

$$F_\mu^{(1)} := \begin{bmatrix} \text{diag}(c - Cu) & -QC^T & 0 & 0 & 0 \\ 0 & -C^T & -M & 0 & B \\ 0 & 0 & W & R & 0 \\ 0 & 0 & I & 0 & A \\ 0 & B^T & 0 & A^T & 0 \end{bmatrix} .$$

Smaller system, maintains feasibility of conjugate variables.

Exploiting problem structure

- Each IP iteration requires a matrix solve.
- Reduced the matrix $F^{(1)}$ to upper triangular form:

$$\begin{bmatrix} D & -QC^T & 0 & 0 & 0 \\ 0 & -T & 0 & 0 & B \\ 0 & 0 & W & R & 0 \\ 0 & 0 & 0 & -RW^{-1} & A^T \\ 0 & 0 & 0 & 0 & B^TT^{-1}B + AWR^{-1}A^T \end{bmatrix},$$

$$T = M + CQD^{-1}C^T$$

- D, T, Q, R, W sparse (diagonal).
- B contains the full linear model; A the imposed constraints.
- We are guaranteed $Cu \leq c$ at each iteration.

Applications

Dynamic Systems Inference

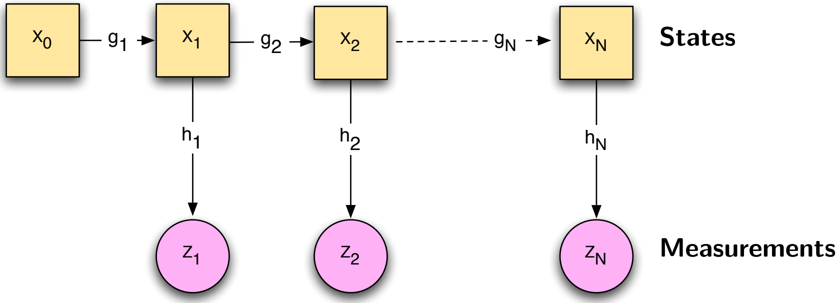
- Kalman filters and smoothers are
  - ubiquitous in navigation systems (planes, space, robots, UAV's)
  - used as a first step for e.g. 3D model reconstruction (NASA)
  - important for climate/weather models (ensemble KF)
  - used in PK/PD modelling to track drug concentrations
  - used in finance (trend filtering)
  - connected to many research areas, including belief propagation/graphical models, functional reconstruction, smoothing splines, dynamic linear models, stochastic differential equations

- Original Kalman filter paper was written in 1960, and the topic still continues to be a hot research area for engineering, statistics, optimization, PDE, SDE, and many applied communities.

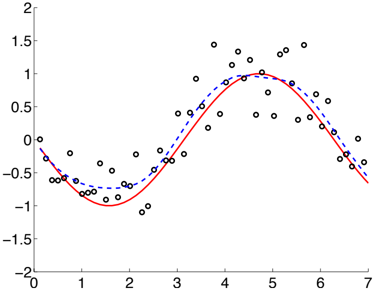
- In 2009, Rudolf Kalman received the National Science Award from President Obama for the Kalman filter.

Graphical Overview of Dynamic Systems

- Goal: to obtain estimates on states $\{x_k\}$ given measurements $\{z_k\}$
- State evolution models $x_k = g_k(x_{k-1}) + w_k$.
- Initialization: $x_1 = x_0 + w_1$.
- Measurement model: $z_k = h_k(x_k) + v_k$



Example : tracking a smooth signal

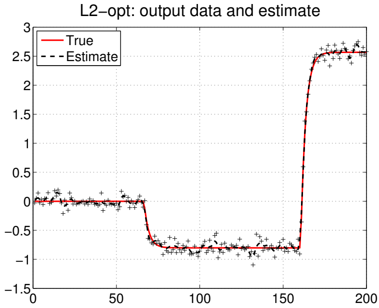


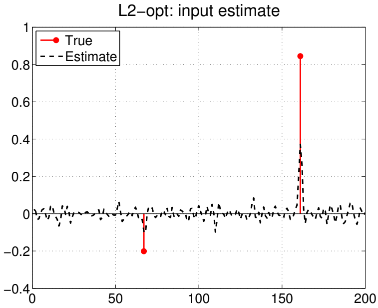
$$\begin{bmatrix}\dot{x}_{k+1}\\ x_{k+1}\end{bmatrix} = \begin{bmatrix} I & 0\\ \Delta t & I\end{bmatrix}\begin{bmatrix}\dot{x}_k\\ x_k\end{bmatrix} + \begin{bmatrix} w_k(1)\\ w_k(2)\end{bmatrix}, \quad Q_k = \sigma_q^2 \begin{bmatrix}\Delta t & \Delta t^2/2\\ \Delta t^2/2 & \Delta t^3/3\end{bmatrix},$$

$$z_k = \begin{bmatrix} 0 & 1\end{bmatrix}\begin{bmatrix}\dot{x}_k\\ x_k\end{bmatrix} + v_k, \quad R_k = \sigma_r^2.$$

Method works well! But what would you do if I told you the signal was always between $[-1, 1]$?

Example 2: impulsive disturbances in DC motor





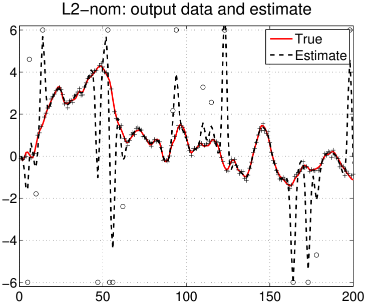
- **State**: angular velocity, angle of motor shaft.
- **Input u:** applied torque (known).
- **disturbances d:** impulsive, unknown.
- **Dynamics:**

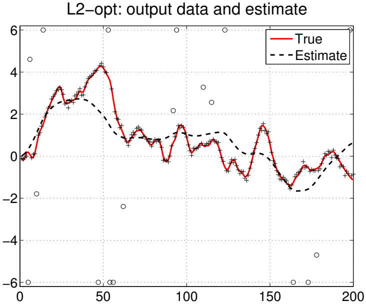
$$x_{k+1} = \begin{pmatrix} 0.71 & 0 \\ 0.08 & 1 \end{pmatrix} x_k + \begin{pmatrix} 11.8 \\ 0.63 \end{pmatrix} (u_t + d_t),$$

$$z_k = \begin{pmatrix} 0 & 1 \end{pmatrix} x_k + e_k$$

- **Right panel:** Best linear estimator of impulsive disturbances d_t is poor. Need a better J to model d_t.

Example 3: measurement outliers





- **Outliers**: 10% of measurements are corrupted by outliers.
- **Left:** Gaussian model with nominal variance (outliers pull estimate)
- **Right:** Best linear estimate (cannot track signal well)
- **Main point:** We need a better V to model measurement errors.

Full-state formulation

$$R = \mathrm{diag}(\{R_k\}) \qquad x = \mathrm{vec}(\{x_k\})$$
$$Q = \mathrm{diag}(\{Q_k\}) \qquad \eta = \mathrm{vec}(\{\overline{x}_0, 0, \ldots, 0\})$$
$$H = \mathrm{diag}(\{H_k\}) \qquad z = \mathrm{vec}(\{z_1, z_2, \ldots, z_N\})$$

$$G = \begin{bmatrix} \mathrm{I} & 0 & & \\ -G_2 & \mathrm{I} & \ddots & \\ & \ddots & \ddots & 0 \\ & & -G_N & \mathrm{I} \end{bmatrix}$$

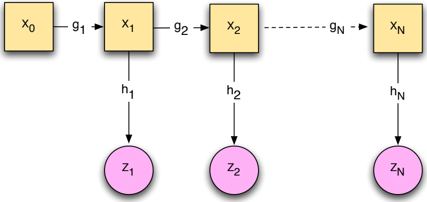
$$\min_x f(x) = \frac{1}{2}\|Hx - z\|^2_{R^{-1}} + \frac{1}{2}\|Gx - \eta\|^2_{Q^{-1}} .$$

Take gradient, set it equal to 0:

$$(H^\top R^{-1} H + G^\top Q^{-1} G)x = H^\top R^{-1} z + G^\top Q^{-1} \eta .$$

Classic algorithms (KF, RTS, MF, M) are easily interpretable as linear algebraic operations on this system.

Aravkin, Bell, Burke, Pillonetto, http://arxiv.org/abs/1303.5237



- We consider the entire class of PLQ smoothers where we use general PLQ penalties for process J and measurement V.
- Solve the optimization problem

$$\min_{x \in \mathcal{X}} J(\mu - Gx) + V(z - Hx) \,.$$

- We also impose constraints $x \in \mathcal{X}$ on the state.

Complexity preserved by interior point

IP iterations require inverting following linear system:

$$\begin{bmatrix} D & -QC^T & 0 & 0 & 0 \\ 0 & -T & 0 & 0 & B \\ 0 & 0 & W & R & 0 \\ 0 & 0 & 0 & -RW^{-1} & A^T \\ 0 & 0 & 0 & 0 & B^TT^{-1}B + AWR^{-1}A^T \end{bmatrix},$$

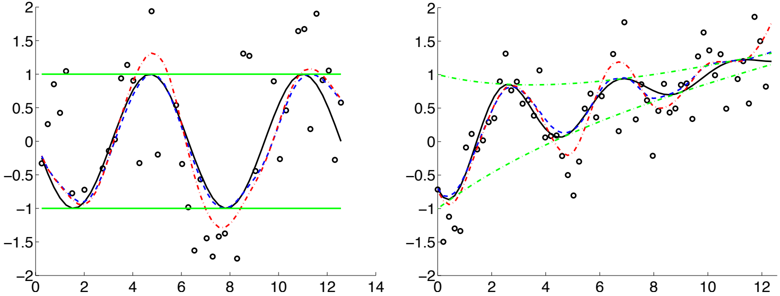
For Kalman smoothing model, we have $B = \begin{bmatrix} Q^{-1/2}G \\ R^{-1/2}H \end{bmatrix}$.

So Schur complement is **block tridiagonal PD**:
$B^TT^{-1}B = \left[G^TQ^{-T/2}T_1^{-1}Q^{-1/2}G + H^TR^{-T/2}T_2^{-1}R^{-1/2}H\right]$

Constraints: $AWR^{-1}A^T$ also block tridiagonal if (...)?

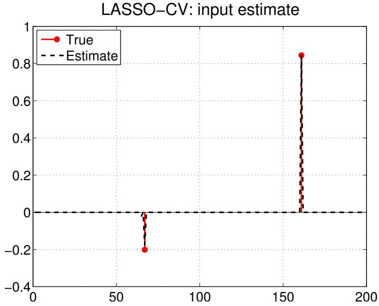
So, we preserve **classical complexity** results, of $O(n^3N)$ per iteration.

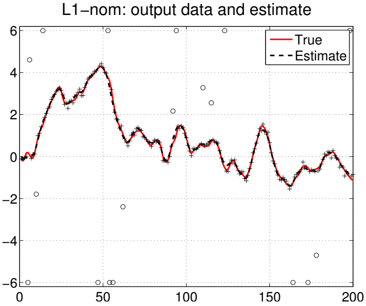
Constrained results



Two examples of linear constraints, using ℓ_2^2 for J and V. Black solid line is true signal, black dash-dot line is unconstrained Kalman smoother, and blue dashed line is the constrained Kalman smoother.

DC rotor results

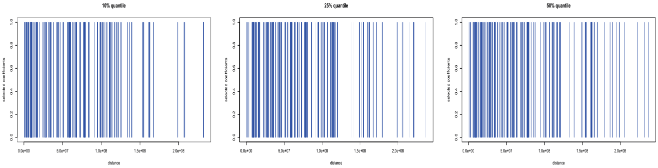




- **Left: impulsive disturbance.** Use ℓ_1 for J, ℓ_2^2 for V.
- **Right: outliers.** Use ℓ_2^2 for J, ℓ_1 or huber for V.

High dimensional sparse regression

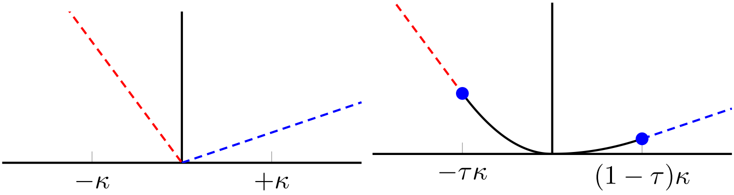
- Genomic data: 206 cases, 18K locations (SNPs).
- Question: which genetic locations (SNPs) are associated with disease?
- Use sparsity (ℓ_1 for J), and *quantile penalty* for V.



(a) 10% Quantile (b) 20% Quantile (c) 50% Quantile

- Figure: top SNPs obtained by sparse QR on Alzheimers data — persistent SNPs include a hotspot associated with APOE gene.
- Background references for QR:
  - Koenker, R. and Bassett, G, Econometrica, pp. 33-50, 1978.
  - Koenker, R. and Geling, O. J. ASA, 96:458-468, 2001.

Quantile vs. Quantile Huber



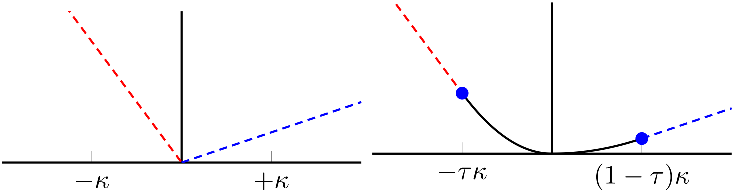
- Convex and nonconvex sparsity formulations:

$$\min_x \rho_q(b - Ax) + \lambda\|x\|_1$$

$$\min_x \rho_q(b - Ax) \quad \text{s.t.} \quad \|x\|_0 \le k.$$

- Run a series of experiments for $q = 0.1, 0.2, \ldots, 0.9$.
- We propose *smooth* (huber) quantile loss (why?).

Quantile vs. Quantile Huber



- Convex and nonconvex sparsity formulations:

$$\min_x \rho_q(b - Ax) + \lambda\|x\|_1$$

$$\min_x \rho_q(b - Ax) \quad \text{s.t.} \quad \|x\|_0 \le k.$$

- Run a series of experiments for $q = 0.1, 0.2, \ldots, 0.9$.
- We propose *smooth* (huber) quantile loss (why?).
- We solve ℓ_0 with a generalized OMP, with ipSolve as subroutine.

Generalized OMP

- **Initialization:** $r = b, S^{(0)} = \emptyset$
- **For**: $j = 1, \ldots$ k:

$$r^{(j)} = b - A_{S^{(j)}} x^{(j)}$$

$$i_{(j)} = \arg\max_i \left| \nabla \rho \left(r^{(j)} \right)^T A_i \right|$$

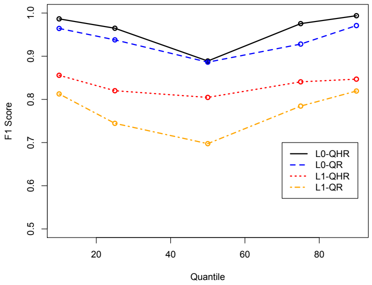
(Maximum projection onto generalized residuals)

$$S^{(j)} = S^{(j-1)} \cup \{ i_{(k)} \}$$

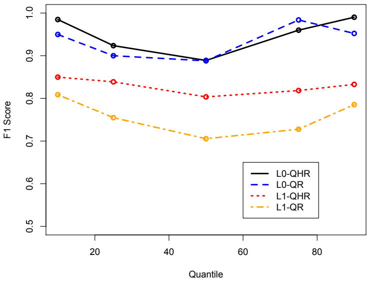
- Refitting Step: ipSolve

$$x^{(j)} = \arg \min_{x : x_i = 0, i \notin S^{(j)}} \rho(b - Ax)$$

Results for simulated experiment



$n = 300, p = 400$



$n = 300, p = 800$

- $F_1 = 2\frac{\text{Pr}\times\text{Rec}}{\text{Pr}+\text{Rec.}}$, $\text{Pr} = \frac{tp}{tp+fp}$, $\text{Rec} = \frac{tp}{tp+fn}$.
- Huber quantile is more accurate than quantile.
- ℓ_0 formulations have superior accuracy to ℓ_1.

Aravkin, Kambadur, Lozano, Luss, ICDM 2014

Large-scale Approach

Inexact methods for linear systems

- We would like inexact methods to handle very large problems, and problems where we can only use matrix-vector products.
- Specialized methods exist for specific applications, mostly for sparse optimization (Koh et al. 2007, Fountoulakis et. al. 2013)
- These methods exploit special structure of 1-norm regularizer to develop efficient preconditioners.
- We want to develop inexact methods for our linear system!

Linear system from conjugate formulation

- Recall the key step is solving $\mathbf{F}_\mu^{(1)} \mathbf{\Delta z} = -\mathbf{F}_\mu$, where

$$F_\mu^{(1)} := \begin{bmatrix} \text{diag}(c - Cu) & -QC^T & 0 & 0 & 0 \\ 0 & -C^T & -M & 0 & B \\ 0 & 0 & W & R & 0 \\ 0 & 0 & I & 0 & A \\ 0 & B^T & 0 & A^T & 0 \end{bmatrix}.$$

- Most structures here are extremely sparse (some diagonal).
- B encodes linear model, A is the constraint matrix.
- System is special, since we have primal, conjugate, and dual variables.

A simple approach

- Start with reduced $F^{(1)}$:

$$\begin{bmatrix} D & -QC^T & 0 & 0 & 0 \\ 0 & -T & 0 & 0 & B \\ 0 & 0 & W & R & 0 \\ 0 & 0 & 0 & -RW^{-1} & A^T \\ 0 & 0 & 0 & 0 & B^T T^{-1} B + AWR^{-1}A^T \end{bmatrix},$$

- There are many ways to solve original system, this is just one of them.
- Strategy: solve lower block approximately, then get exact updates for the other variables.
- For some ML problems, IP iterations require remarkably few linear solver iterations.

Sparse difference graphs

- Data: sample covariance $\Sigma_1,\ \Sigma_2,\ \Delta\Sigma = \Sigma_2 - \Sigma_1$.
- Target: *sparse* estimate of $X = \Theta_1 - \Theta_2 = \Sigma_1^{-1} - \Sigma_2^{-1}$.
- Key identity:

$$\Sigma_1(\Theta_1 - \Theta_2)\Sigma_2 = \Sigma_1(\Sigma_1^{-1} - \Sigma_2^{-1})\Sigma_2 = \Sigma_2 - \Sigma_1 = \Delta\Sigma.$$

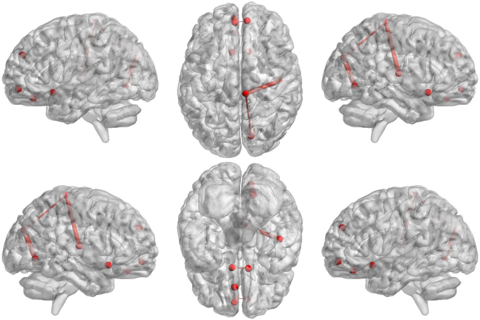
- Formulation:

$$\min_X \|X\|_1 \qquad \text{s.t.} \quad \|\Sigma_1 X \Sigma_2 - \Delta\Sigma\|_\infty \le \lambda$$

  - λ controls fidelity to observed data.
  - Vectorization: $\text{vec}(\Sigma_1 X \Sigma_2) = (\Sigma_2 \otimes \Sigma_1)x$.
  - Special structure: matrix vector products are $O(n^3)$, not $O(n^4)$.

- This is a linear program (!) but using IPsolve and block structure, we have matrix-free implementation.

COBRE Illustration



- 72 healthy and 74 Alzheimers' brains are compared
- lower λ means tighter constraints, less sparsity ($\lambda = 0.5$ shown)
- can potentially be used as a diagnostic tool to detect differences between healthy and unhealthy brains

Joint work with C. Eisenach, H. Liu, Y. Liu, D. Orban, R. Vanderbei

IPsolve vs. CPEX

- COBRE dataset has covariance matrices that are 116×116.
- This gives primal problem dimensions of 13456×13456.

λ	0.9	0.8	0.7	0.6	0.5
Time (s)	23	28	70	113	276

Table: Average runtime for IPSolve in seconds for varying levels of λ.

λ	0.9	0.8	0.7	0.6	0.5
Time (s)	580	659	573	565	575

Table: Runtime for CPLEX in seconds for varying levels of λ.

- Approach scales for larger (synthetic) matrices, and the problem is always more difficult for higher fidelity requirements.

Conclusion

- PLQ penalties appear in a range of data science, high dimensional inference, and machine learning problems.
- We can use conjugate representations of these penalties to design easily customizable interior-point methods.
- Features of PLQ models (smoothness, asymmetry, tail growth) with constraints capture a wide range of useful models for applications.
- Some current work:
  - Semi-parametric inference (e.g. quantile parameter) (Peng, Karthik, JJ)
  - Convex-composite extensions (Jim, Dima)
  - Splitting methods for PLQ (Damek Davis)
  - Robust CVaR (Xin Chen)

Congratulations, Jim, and thank you!!

- Aleksandr Y Aravkin, James V Burke, Dmitriy Drusvyatskiy, Michael P Friedlander, and Scott Roy, **Level-set methods for convex optimization**, arXiv preprint arXiv:1602.01506 (2016).
- A. Aravkin, J.V. Burke, and G. Pillonetto, **Sparse/robust estimation and kalman smoothing with nonsmooth log-concave densities: Modeling, computation, and theory**, Journal of Machine Learning Research **14** (2013), 2689–2728.
- Aleksandr Aravkin, James V. Burke, and Gianluigi Pillonetto, **Linear system identification using stable spline kernels and plq penalties**, Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on, IEEE, 2013, pp. 5168–5173.
- A. Aravkin, P. Kambadur, A.C. Lozano, and R. Luss, **Orthogonal matching pursuit for sparse quantile regression**, Data Mining (ICDM), International Conference on, IEEE, 2014, pp. 11–19.
- A. Aravkin, R. Kumar, H. Mansour, B. Recht, and F.J. Herrmann, **Fast methods for denoising matrix completion formulations,**

with applications to robust seismic data interpolation, SIAM Journal on Scientific Computing **36** (2014), no. 5, S237–S266.

K.P. Bube and T. Nemeth, **Fast line searches for the robust solution of linear systems in the hybrid ℓ_1/ℓ_2 and huber norms**, Geophysics **72** (2007), no. 2, A13–A17.

Moshe Buchinsky, **Changes in the u.s. wage structure 1963-1987: Application of quantile regression**, Econometrica **62** (1994), no. 2, 405–58.

Wei Chu, S Sathiya Keerthi, and Chong Jin Ong, **A unified loss function in bayesian framework for support vector regression**, Epsilon **1** (2001), no. 1.5, 2.

D. Clark, **The mathematical structure of huber'ĂŹs m-estimator**, SIAM Journal on Scientific and Statistical Computing **6:1** (1985), 209–219.

Rudolf Dutter and Peter J Huber, **Numerical Methods for the Nonlinear Robust Regression Problem**, Journal of Statistical Computation and Simulation **13** (1981), 79–113.

- Christine De Mol, Ernesto De Vito, and Lorenzo Rosasco, **Elastic-net regularization in learning theory**, Journal of Complexity **25** (2009), no. 2, 201–230.
- D Donoho, **Compressed sensing**, IEEE Trans. on Information Theory **52** (2006), no. 4, 1289–1306.
- Ofer Dekel, Shai Shalev-Shwartz, and Yoram Singer, **Smooth epsiloon-insensitive regression by loss symmetrization**, Journal of Machine Learning Research, 2005, pp. 711–741.
- B. Efron, T. Hastie, L. Johnstone, and R. Tibshirani, **Least angle regression**, Annals of Statistics **32** (2004), 407–499.
- T. Evgeniou, M. Pontil, and T. Poggio, **Regularization networks and support vector machines**, Advances in Computational Mathematics **13** (2000), 1–150.
- David A Freedman, **Statistical models: theory and practice**, cambridge university press, 2009.

- F.J. Herrmann, M.P. Friedlander, and O. Yilmaz, **Fighting the curse of dimensionality: Compressive sensing in exploration seismology**, Signal Processing Magazine, IEEE **29** (2012), no. 3, 88–100.
- F. J. Herrmann and G. Hennenfent, **Non-parametric seismic data recovery with curvelet frames**, Geophysical Journal International **173** (2008), no. 1, 233–248.
- T.J. Hastie and R.J. Tibshirani, **Generalized additive models**, Monographs on Statistics and Applied Probability, vol. 43, Chapman and Hall, London, UK, 1990.
- T.J. Hastie, R.J. Tibshirani, and J. Friedman, **The elements of statistical learning. data mining, inference and prediction**, Springer, Canada, 2001.
- Peter J Huber, **Robust Statistics**, John Wiley and Sons, 2004.
- R. Koenker and G Bassett, **Regression quantiles**, Econometrica (1978), 33–50.

- R. Koenker and O. Geling, **Reappraising medfly longevity: A quantile regression survival analysis**, Journal of the American Statistical Association **96** (2001), 458âĂŞ468.
- Roger Koenker and Kevin F. Hallock, **Quantile regression**, Journal of Economic Perspectives, American Economic Association (2001), 143–156.
- R. Koenker, **Quantile regression**, Cambridge University Press, 2005.
- Yuh-Jye Lee, Wen-Feng Hsieh, and Chien-Ming Huang, **ϵ-ssvr: a smooth support vector machine for ϵ;-insensitive regression**, Knowledge and Data Engineering, IEEE Transactions on **17** (2005), no. 5, 678–685.
- Qing Li, Nan Lin, et al., **The bayesian elastic net**, Bayesian Analysis **5** (2010), no. 1, 151–170.
- Yuh-Jye Lee and Olvi L Mangasarian, **Ssvm: A smooth support vector machine for classification**, Computational optimization and Applications **20** (2001), no. 1, 5–22.

- W. Li and J. Swetits, **The linear l1 estimator and the huber m-estimator**, SIAM Journal on Optimization **8** (1998), no. 2, 457–475.
- J. Mairal, M. Elad, and G. Sapiro, **Sparse representation for color image restoration**, IEEE Transactions on Image Processing **17** (2008), no. 1, 53–69.
- Ricardo A Maronna, Douglas Martin, and Yohai, **Robust Statistics**, Wiley Series in Probability and Statistics, Wiley, 2006.
- H. Mansour, R. Saab, P. Nasiopoulos, and R. Ward, **Color image desaturation using sparse reconstruction**, Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), March 2010, pp. 778–781.
- Hassan Mansour, Haneet Wason, Tim T Y Lin, and Felix J Herrmann, **Randomized marine acquisition with compressive sampling matrices**, Geophysical Prospecting **60** (2012), no. 4, 648–662.

R. Neelamani, C.E. Krohn, J.R. Krebs, J.K. Romberg, M. Deffenbaugh, and J.E. Anderson, **Efficient seismic forward modeling using simultaneous random sources and sparsity**, Geophysics **75** (2010), no. 6, WB15–WB27.

M. Pontil and A. Verri, **Properties of support vector machines**, Neural Computation **10** (1998), 955–974.

R Tyrrell Rockafellar and Stanislav Uryasev, **Optimization of conditional value-at-risk**, Journal of risk **2** (2000), 21–42.

J.-L Starck, M. Elad, and D. Donoho, **Image decomposition via the combination of sparse representation and a variational approach**, IEEE Transaction on Image Processing **14** (2005), no. 10.

B. Schölkopf and A. J. Smola, **Learning with kernels: Support vector machines, regularization, optimization, and beyond**, (Adaptive Computation and Machine Learning), MIT Press, 2001.

B. Schölkopf, A.J. Smola, R.C. Williamson, and P.L. Bartlett, **New support vector algorithms**, Neural Computation **12** (2000), 1207–1245.

- G.A.F. Seber and C.J. Wild, **Nonlinear regression**, Wiley Series in Probability and Statistics, Wiley, 2003.
- Andreĭ Tikhonov, **Numerical methods for the solution of ill-posed problems**.
- V. Vapnik, **Statistical learning theory**, Wiley, New York, NY, USA, 1998.
- H. Zou and T. Hastie, **Regularization and variable selection via the elastic net**, Journal of the Royal Statistical Society, Series B **67** (2005), 301–320.
- Hui Zou and Trevor Hastie, **Regularization and variable selection via the elastic net**, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67** (2005), no. 2, 301–320.
- Hui Zou and Ming Yuan, **Regularized simultaneous model selection in multiple quantiles regression**, Computational Statistics & Data Analysis **52** (2008), no. 12, 5296–5304.