

# A Review of Some Optimization Techniques in Machine Learning and Statistics

Stephen Wright

Department of Computer Sciences  
University of Wisconsin-Madison

McMaster, June, 2005

# Background

- Many interesting applications in machine learning/statistics give rise to convex quadratic programs with parameters.
- Highly structured, sometimes very large.
- Sometimes the solution is desired/needed for many values of the parameter.
- A clamor for free software!
- Semidefinite programming (SDP) and second-order cone programming (SOCP) also finding applications in these areas.

# Outline

- Describe several parametrized QP applications.
- Discuss properties of solution paths.
- Discuss some of the algorithms used.
- How should we proceed with algorithm design / software implementation?
- Describe a conic optimization application (with SDP and SOCP variables).

# QP Applications in Machine Learning

- Support Vector Machines
  - two-category formulations
  - multicategory variant



## Simplest Formulation

Given data points  $x_1, x_2, \dots, x_m$  in  $\mathbb{R}^n$  with labels  $y_i = \pm 1$ ,  $i = 1, 2, \dots, m$ , find a vector  $w \in \mathbb{R}^n$  and a scalar  $\gamma$  such that the hyperplane defined by  $w^T x + \gamma$  separates the data. That is,

$$y_i = +1 \Rightarrow w^T x_i + \gamma \geq +1,$$

$$y_i = -1 \Rightarrow w^T x_i + \gamma \leq -1.$$

Can show that  $w$  with this property that minimizes  $\|w\|_2$  gives the maximal margin between the data. Hence can formulate the problem as a QP:

$$\min_{w, \gamma} \frac{1}{2} w^T w \quad \text{subject to} \quad y_i(w^T x_i + \gamma) \geq 1, \quad i = 1, 2, \dots, m.$$

## Soft-Margin Classifiers

For nonseparable data sets this problem is infeasible; can modify by penalizing for points on the wrong side of the plane:

$$\min_{w, \gamma, \zeta} \frac{1}{2} w^T w + C e^T \zeta \quad \text{subject to}$$

$$y_i(w^T x_i + \gamma) \geq 1 - \zeta_i, \quad \zeta_i \geq 0, \quad i = 1, 2, \dots, m,$$

where  $e = (1, 1, \dots, 1)^T$  and  $C$  is a (scalar) penalty parameter. This is a soft-margin classifier (C-SVC).

An alternative soft-margin classifier called  $\nu$ -SVC depends on a parameter  $\nu \in [0, 1]$ :

$$\min_{w, \gamma, \zeta, \rho} \frac{1}{2} w^T w - \nu \rho + \frac{1}{m} e^T \zeta \quad \text{subject to}$$

$$y_i(w^T x_i + \gamma) \geq \rho - \zeta_i, \quad \zeta_i \geq 0, \quad i = 1, 2, \dots, m; \quad \rho \geq 0.$$

# Duality and Optimality

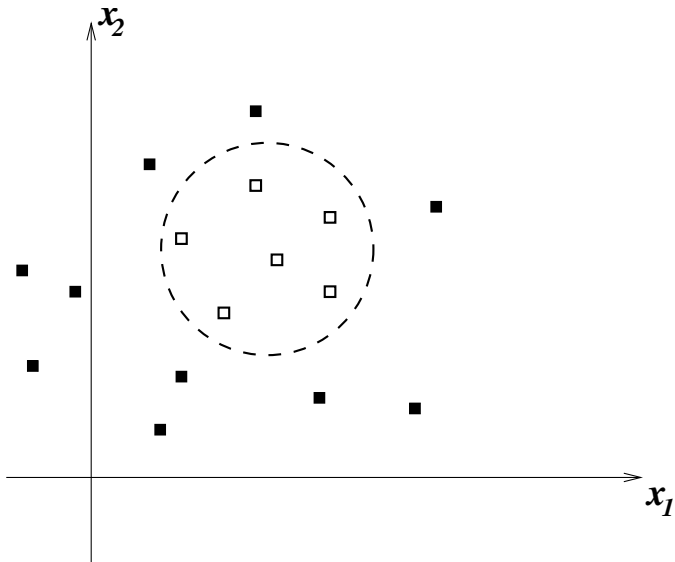
Duality gives an interesting interpretation of the results. Find that the solution  $w$  has the form:

$$w = \sum_{i=1}^m \alpha_i y_i x_i,$$

where  $\alpha_i \geq 0$  are the Lagrange multipliers/dual variables. Hence the classification function  $f$  has the form:

$$f(x) = \sum_{i=1}^m \alpha_i y_i (x_i^T x) + \gamma.$$

We label a new vector  $x$  as  $+1$  if  $f(x) > 0$  and  $-1$  if  $f(x) < 0$ .



# Kernels

Utility of SVMs enhanced greatly by use of kernels. Apply a transformation  $\Phi$  which takes each  $x_i$  into a higher-dimensional *feature space*  $\mathcal{H}$ , then do the separation in  $\mathcal{H}$  space.

Now  $w \in \mathcal{H}$ ; can show that it has the form

$$w = \sum_{i=1}^m \alpha_i y_i \Phi(x_i),$$

where the  $\alpha_i$ 's can be obtained by solving the dual of C-SVC:

$$\max_{\alpha} e^T \alpha - \frac{1}{2} \alpha^T Y^T K Y \alpha \quad \text{subject to} \quad \alpha^T y = 0, \quad 0 \leq \alpha \leq C e$$

where  $Y = \text{diag}(y)$  and  $K$  is the (**dense**) kernel matrix:

$$K_{ij} = \Phi(x_i)^T \Phi(x_j).$$

In fact, can forget about  $\Phi$  and define a kernel function  $k(\cdot, \cdot) : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$  directly, then set

$$K_{ij} = k(x_i, x_j).$$

If  $k$  is positive definite (obvious definition), then there exists a feature space  $\mathcal{H}$  such that  $k(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$ .

Given labelled  $x_1, x_2, \dots, x_m$  and  $k(\cdot, \cdot)$ , can define  $K$  as above, solve the dual on the previous page, define  $\gamma$  as the Lagrange multiplier for  $\alpha^T y = 0$ , and define the classifier to be

$$f(x) = \sum_{i=1}^m \alpha_i y_i k(x_i, x) + \gamma.$$

# Popular Kernels

$$k(x_i, x_j) = \left(x_i^T x_j + 1\right)^d \quad (d \text{ a positive integer})$$

$$k(x_i, x_j) = \exp\left(-\|x_i - x_j\|^2 / (2\sigma^2)\right)$$

where  $\sigma$  is another parameter (enters nonlinearly).



## Multicategory SVM

(*Wahba et al.*) Have  $k \geq 2$  labels. Seek classification functions  $f_1, f_2, \dots, f_k$  such that the a vector  $x$  is classified according to

$$\arg \max_{j=1,2,\dots,k} f_j(x).$$

(When  $k = 2$ , just choose  $f_1 = f$ ,  $f_2 = -f$ , where  $f$  defined above.)

Now everything becomes 2-D:  $\alpha \in \mathbb{R}^{m \times k}$ ,  $\gamma \in \mathbb{R}^k$ , labels  $y \in \mathbb{R}^{m \times k}$ .

If training vector  $x_i$  lies in class  $l$ , set  $y_{il} = +1$  and  $y_{ip} = -1/(k-1)$  for  $p \neq l$ .

# KKT Conditions

$$y_j - \alpha_j - \gamma_j = 0, \quad j = 1, 2, \dots, k,$$

$$n\lambda Kc_j + K\alpha_j + K\gamma = 0, \quad j = 1, 2, \dots, k,$$

$$e^T(\alpha_j + \Delta) = 0, \quad j = 1, 2, \dots, k,$$

$$eb_j + Kc_j - y_j - \xi_j + s_j = 0, \quad j = 1, 2, \dots, k,$$

$$0 \leq \xi_j \perp \gamma_j \geq 0, \quad j = 1, 2, \dots, k,$$

$$\left( \sum_{j=1}^k b_j \right) e + K \left( \sum_{j=1}^k c_j \right) = 0,$$

$$0 \leq s_j \perp \alpha_j \geq 0, \quad j = 1, 2, \dots, k.$$

Dual formulation is

$$\min_{\bar{\alpha}, \alpha_1, \dots, \alpha_k} \frac{1}{2} \sum_{l=1}^k (\alpha_{.l} - \bar{\alpha})^T K (\alpha_{.l} - \bar{\alpha}) + \sum_{i=1}^k \alpha_{.l}^T y_{.l},$$

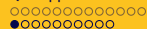
subject to

$$\begin{aligned} 0 &\leq \alpha_{.l} \leq Ce, \quad l = 1, 2, \dots, k, \\ (\alpha_{.l} - \bar{\alpha})^T e &= 0, \quad l = 1, 2, \dots, k. \end{aligned}$$

Again a single scalar parameter  $C$ , but a more complicated QP.

Usually solve SVM and multicategory SVM for wide range of value of the parameter  $C$ , and use some external test to determine the best value to use.

Multicategory SVM (Lee, Lin, Wahba, 2004) use Generalized Approximate Cross Validation (GACV). For each value of  $C$ , requires solving many problems of the form above with slightly different choice of data.



# Applications in Statistics

Regression Problems.

- LASSO
- extended LASSO
- Huber estimation

# LASSO

(Tibshirani, 1996) Linear least squares, with a constraint on  $\ell_1$  norm of the solution  $x$ :

$$\min_x \|Ax - b\|_2^2 \quad \text{subject to} \quad \|x\|_1 \leq t,$$

for some parameter  $t \geq 0$ .

When  $t = 0$ , solution is  $x = 0$ . When  $t \geq \|x_{LS}\|_1$ , where  $x_{LS}$  is the (unconstrained) least-squares solution, we have  $x = x_{LS}$ .

QP formulation:

$$\min_{x,s} \frac{1}{2} x^T A^T A x - b^T A x \quad \text{subject to} \quad -s \leq x \leq s, \quad e^T s \leq t.$$

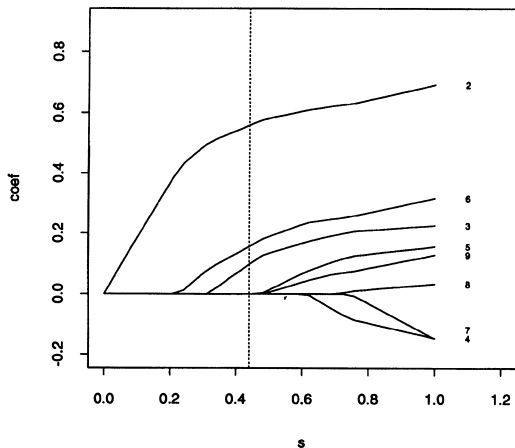
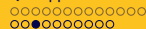


Fig. 5. Lasso shrinkage of coefficients in the prostate cancer example: each curve represents a coefficient (labelled on the right) as a function of the (scaled) lasso parameter  $s = t/|\beta_j^0|$  (the intercept is not plotted); the broken line represents the model for  $\hat{s} = 0.44$ , selected by generalized cross-validation

*Motivation:* Aim to identify the “explanatory” variables: The components of  $x$  that contribute most to minimizing  $\|Ax - b\|_2$ . The  $\ell_1$  constraint tends to produce “sparse” solutions, with few nonzero components—more as  $t$  increases.

Once these variables are identified, solve a reduced least squares problem in which only these variables are allowed to be nonzero.

Variant: Allow other constraints on  $x$ , e.g.  $x \geq 0$ .

## Extended LASSO

(Turlach, Venables, SJW 2005) Have a single coefficient matrix  $A$ , but have  $p$  response variables  $b$ , not just one. Seek a different regressor variable vector  $x_j$  for each response variable  $b_j$ :

$$\min_{x_1, x_2, \dots, x_p} \sum_{j=1}^p \|Ax_j - b_j\|_2^2.$$

Measure the “simultaneous explanatory power” of the  $l$ th component of  $x_1, x_2, \dots, x_p$  by

$$\max(|x_{1l}|, |x_{2l}|, \dots, |x_{pl}|).$$

Motivates the following LASSO-like constraint:

$$\sum_{l=1}^n \max(|x_{1l}|, |x_{2l}|, \dots, |x_{pl}|) \leq t.$$

# Huber Estimation

Like least-squares, but applies a less severe penalty on “outliers”

$$\min_x \sum_{i=1}^m \rho(A_i x - b_i),$$

where  $\rho(\cdot)$  is a loss function defined by

$$\rho(r) = \begin{cases} \frac{1}{2}r^2, & |r| \leq \gamma, \\ \gamma r - \frac{1}{2}\gamma^2, & r > \gamma, \\ -\gamma r - \frac{1}{2}\gamma^2, & r < -\gamma. \end{cases}$$

Same as least squares for  $|r| \leq \gamma$ , outside this range increases linearly with rate  $\gamma$ .

Not immediately obvious how to formulate Huber as a QP! First write optimality conditions

$$\sum_{i=1}^m A_i^T \rho'(A_i x - b_i) = 0,$$

where by differentiating  $\rho$  we get

$$w_i = \rho'(A_i x - b_i) = \begin{cases} A_i x - b_i, & |A_i x - b_i| \leq \gamma, \\ \gamma, & A_i x - b_i \geq \gamma, \\ -\gamma, & A_i x - b_i \leq -\gamma. \end{cases}$$

Hence can express optimality condition succinctly as

$$\begin{aligned} A^T w &= 0, \\ w_i - A_i x + b_i &= 0 \quad \Rightarrow \quad -\gamma \leq w_i \leq \gamma, \\ w_i - A_i x + b_i &< 0 \quad \Rightarrow \quad w_i = \gamma, \\ w_i - A_i x + b_i &> 0 \quad \Rightarrow \quad w_i = -\gamma. \end{aligned}$$

Remarkably, these are the KKT conditions for the following problem:

$$\min_w \frac{1}{2} w^T w + b^T w \quad \text{subject to} \quad A^T w = 0, \quad -\gamma e \leq w \leq \gamma e.$$

A QP with scalar parameter  $\gamma$ .

## Application: Portfolio Optimization (Markowitz)

$x_i$  = proportion to invest in instrument  $i$

$r_i$  = expected rate of return on  $i$

$Q_{ij}$  = covariance between  $i$  and  $j$ .

Objective balances between expected return  $r^T x$  and portfolio variance  $x^T Q x$ . Use a risk tolerance parameter  $\gamma > 0$  to strike the balance:

$$\min_x \frac{1}{2} x^T Q x - \gamma r^T x \quad \text{subject to} \quad e^T x = 1, \quad x \geq 0.$$

Can move the parameter to the constraints by change of variables to  $z = x/\gamma$ .

$$\min_z \frac{1}{2} z^T Q z - r^T z \text{ subject to } e^T z = 1/\gamma, z \geq 0.$$

**An aside:** This model is faulty. The relative allocation shouldn't depend on total amount of funds available. Recently proposed models replace  $x^T Q x$  by  $\sqrt{x^T Q x}$  and use an SOCP formulation.

## General Form of Parametrized QP

All these problems (or their duals) have the form

$$\text{pQP}(t): \min_x \frac{1}{2}x^T Qx + c^T x \text{ subject to } Ax \geq b + tq,$$

where  $Q$  is symmetric positive semidefinite,  $t$  is a scalar parameter, and  $b$  and  $q$  are right-hand side vectors. Wish to solve for all  $t$  in some range.

Note that  $\text{pQP}(t)$  may be infeasible for some  $t$ . The interval of  $t$  values for which it is feasible is connected.

If  $\text{pQP}(t)$  is unbounded for some  $t$  then it is unbounded for all  $t$  (Frank-Wolfe).

Denote by  $V(t)$  the optimal objective for  $\text{pQP}(t)$  for each  $t$ .

Denote by  $\mathcal{X}(t)$  the solution set for  $\text{pQP}(t)$ .

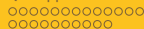
KKT conditions for  $pQP(t)$  are useful in proving results in a elementary fashion. Write these as follows: For some index sets  $\mathcal{A}$  and  $\mathcal{I}$  defined by

$$\mathcal{A} \subset \{1, 2, \dots, m\}, \quad \mathcal{I} = \text{complement of } \mathcal{A},$$

and some vector  $\lambda_{\mathcal{A}}$ , we have

$$\begin{aligned} Qx + c - A_{\mathcal{A}}^T \lambda_{\mathcal{A}} &= 0, \\ A_{\mathcal{A}} x - p_{\mathcal{A}} - tq_{\mathcal{A}} &= 0, \\ A_{\mathcal{I}} x - p_{\mathcal{I}} - tq_{\mathcal{I}} &\geq 0, \\ \lambda_{\mathcal{A}} &\geq 0. \end{aligned}$$

The fact that there are only a finite number of possible choices for  $\mathcal{A}$  ( $2^m$ , to be precise) is useful.



## Elementary Results about $\text{pQP}(t)$

### Theorem

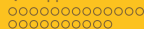
*$\mathcal{X}(t)$  is closed and convex for each  $t$ .*

### Theorem

*The value function  $V(t)$  is convex and continuous in  $t$ .*

### Theorem

*Consider a sequence  $\{t_k\}$  with  $\{t_k\} \rightarrow \bar{t}$ , where  $\mathcal{X}(t_k)$  is nonempty for all  $k$  and  $\text{pQP}(\bar{t})$  is feasible. Then for any sequence  $x(t_k) \in \mathcal{X}(t_k)$ , all limit points of  $\{x(t_k)\}$  lie in  $\mathcal{X}(\bar{t})$ .*



## Piecewise Linearity of Solution Path

Let  $t_L$  and  $t_U$  be the endpoints of the interval for which  $\mathcal{X}(t)$  is nonempty. Then there are a finite number of breakpoints such that the solution set  $\mathcal{X}(t)$  is “linear” in each interval, and “continuous” across the breakpoints.

### Theorem

*There are  $N$  breakpoints  $t_0, t_1, \dots, t_N$  with*

$$t_L = t_0 < t_1 < \dots < t_N = t_U$$

*such that for any  $j = 1, 2, \dots, N$  and any  $t_a$  and  $t_b$  with  $t_{j-1} \leq t_a < t_b \leq t_j$ , and any  $x_a \in \mathcal{X}(t_a)$  and  $x_b \in \mathcal{X}(t_b)$ , we have*

$$(1 - \alpha)x_a + \alpha x_b \in \mathcal{X}((1 - \alpha)t_a + \alpha t_b), \quad \text{for all } \alpha \in [0, 1].$$

# Proof

- Consider the KKT conditions in the form above (based on  $\mathcal{A}$ ).
- For a given  $\mathcal{A}$ , let  $P_{\mathcal{A}}$  be the set of  $t$  values for which the KKT conditions are satisfied by some  $(x, \lambda_{\mathcal{A}})$ , for this  $\mathcal{A}$  and  $t$ .
- Can show that  $P_{\mathcal{A}}$  is empty, or a contiguous interval. Let  $t_L^{\mathcal{A}}$  and  $t_U^{\mathcal{A}}$  be the endpoints.
- Order all the endpoints and discard duplicates to get the breakpoints.
- Within each interval  $(t_j, t_{j+1})$  the collection of possible optimal  $\mathcal{A}$  is constant.

Note possible degeneracy: Active sets nonunique, can change by more than one element across a breakpoint.



## Algorithms: Interior-Point

Can devise interior-point algorithms that respect the structure of each application and solve it fairly efficiently, *but* only for a single value of the parameter  $t$ . Warm starts are not much help.

KKT conditions for pQP( $t$ ):

$$Qx - A^T \lambda = c,$$

$$Ax - s = p + tq,$$

$$s_i \lambda_i = 0, \quad i = 1, 2, \dots, m,$$

$$s \geq 0, \quad \lambda \geq 0$$

Interior-point methods keep  $(s, \lambda) > 0$  and generate Newton-like steps for the three equality conditions.

At each interior-point iteration, solve one or two systems of the form

$$\begin{bmatrix} Q & A^T \\ A & -\Lambda^{-1}S \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta \lambda \end{bmatrix} = \begin{bmatrix} r_x \\ r_s \end{bmatrix},$$

where

$$\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_m), \quad S = \text{diag}(s_1, s_2, \dots, s_m),$$

for some right-hand sides  $r_x, r_s$ .

Can exploit structure in factoring and solving this system efficiently. Best approach is usually to do customized block elimination, ending with a reduced system that can be solved with Cholesky or an LU factorization.



## Experiences with Interior-Point Methods

OOQP (Gertz and SJW, 2003) is an interior-point package for convex QP, written in C++ in an object-oriented fashion to allow customization to special structures.

Solvers for Huber and standard SVM (with C-SVC and  $K_{ij} = x_i^T x_j$ ) are distributed as templates with OOQP, to illustrate how the solver can be customized.

Ferris and Munson (SIOPT, 2003): interior-point SVM.

Codes for LASSO and extended LASSO written from scratch in C and described in (Turlach, Venables, SJW, 2005). Originally written in 1999; predate OOQP.



Code for multiclass SVM written from scratch in C (SJW, 2002-2003). Unusual numerical difficulties: some kernels  $K$  ill conditioned, and could not be as “cavalier” as usual in dealing with small and large elements in  $\Lambda^{-1}S$ .

Other software for SVM can be found on [www.kernel-machines.org](http://www.kernel-machines.org), some based on interior-point. Also customized solvers for very large problems (with many points) that take advantage of the sparsity of the solution (for the standard kernel).



# Algorithms: Pivoting, for Fixed $t$

## LASSO:

- Tibshirani (1996) proposed a “constraint generation” method based on an inefficient formulation with  $2^n$  constraints; also mentioned an efficient formulation (D. Gay) based on a splitting  $x = x^+ - x^-$ .
- Osborne, Presnell, Turlach (2000) propose an active set method customized to special form of the constraint  $\|x\|_1 \leq t$ .



# Homotopy/Pivoting: for all $t$

## LASSO:

- Osborne, Presnell, Turlach (2000) describe a pivoting algorithm that starts with  $t$  near 0 and ramps it up to  $\|x_{LS}\|_1$ , identifying breakpoints. Degeneracy handled in “sledgehammer” fashion: Increase  $t$  slightly past the breakpoint and re-solve from scratch.
- Least-Angle Regression (Efron et al, 2003): Contains a homotopy algorithm for LASSO, closely related to the above.

## Extended LASSO:

- Turlach (2005): unpublished notes and prototype.



## Huber:

- Clark and Osborne (1986!) work with primal Huber formulation, starting with the least-squares solution ( $t$  large) and decreasing it towards zero (the  $\ell_1$  solution).

## SVM:

- Zhu et al. (2003), Hastie et al. (2004).

## Multicategory SVM:

- Lee and Cui (2005)

## General regression setting:

- Rosset and Zhu (2004)



## Is the Unified Viewpoint Useful?

So we can express all these problems (and many others) as parametrized QPs. How does this help?

Better understanding of the solution structure and properties. Leads us to think about algorithms and software. Various possibilities:

- Unified approach to algorithm design—basic approaches that can be reused across different application types, but algorithmic specifics and code redesigned for each application.
- Software tools that can be customized. Reusable components, object-oriented design? (OOQP?) (But code should be easy to work with!)



- A unified package that uses a general sparse solver.
  - Possible with an interior-point approach, but need an excellent symmetric indefinite solver, and still need front-end code to set up data and back-end code to interpret the results.
  - Looks much harder for a pivoting algorithm, if sparse updates are needed.



## Are Other Frameworks Useful?

- Parametrized LCP: More general, pivoting algorithms well studied. Is software available? Is handling of degeneracy etc overkill for the LCPs arising from convex QP?
- Polyhedral convex optimization: Osborne. This viewpoint has been used to develop interesting and efficient homotopy algorithms for various applications. Not as simple as the parametrized QP approach; does not lend itself to a unified approach or a unified code.



## Regularized Kernel Estimation

(Lu, Keles, Wahba, SJW) Given  $N$  objects  $i = 1, 2, \dots, N$  and pairwise dissimilarity measures  $d_{ij}$ ,  $i, j = 1, 2, \dots, N$ . Seek a symmetric, positive semidefinite kernel matrix  $K$  that induces an estimate of dissimilarity  $\hat{d}_{ij}(K)$ .

Do data-fitting between the  $d_{ij}$  and  $\hat{d}_{ij}(K)$ . Data may be redundant or incomplete; may not satisfy a triangle inequality; may encode only a small number of coded levels.

If location of each object  $i$  in Euclidean space is  $x_i$ , natural to define kernel as  $K_{ij} = x_i^T x_j$ . Then

$$\hat{d}_{ij}(K) = K_{ii} + K_{jj} - 2K_{ij} = \|x_i - x_j\|_2^2.$$



## Define

$$B_{ij}(i, i) = B_{ij}(j, j) = 1, \quad B_{ij}(i, j) = B_{ij}(j, i) = -1,$$

then can write data-fitting problem as

$$\min_{K \succeq 0} \sum_{(i,j) \in \Omega} |d_{ij} - B_{ij} \cdot K|$$

where  $\Omega$  is the set of pairs for which distance measures are available.

Introduce a regularization term to reduce rank of  $K$ :

$$\min_{K \succeq 0} \sum_{(i,j) \in \Omega} |d_{ij} - B_{ij} \cdot K| + \lambda \text{trace}(K).$$



Write as an SDP/linear program by setting

$$\min_{K \succeq 0, s \geq 0, t \geq 0} \sum_{(i,j) \in \Omega} s_{ij} + t_{ij} + \lambda I \cdot K, \text{ subject to}$$

$$d_{ij} - B_{ij} \cdot K + s_{ij} - t_{ij} = 0,$$

This problem is challenging for SDP software because of the many constraints ( $|\Omega|$  of them).

Our data set:  $N \approx 280$ , select 55 “buddies” for each object, makes about 14000 constraints. Pushes the limits of standard SDP codes (SDPT3, DSDP).



## An Incremental Method: “Newbie”

Suppose the solution has been obtained for  $N$  objects. Wish to place an additional object, augmenting  $K$  by one row/column, without changing existing entries.

Let  $\Psi$  be the subset of  $\{1, 2, \dots, N\}$  for which estimates of distance  $d_{i,N+1}$  are available.

$$\min_{c \geq 0, b} \sum_{i \in \Psi} \left| d_{i,N+1} - B_{i,N+1} \cdot \begin{bmatrix} K_N & b^T \\ b & c \end{bmatrix} \right|$$

where the matrix is nonnegative definite. In particular we need

$$b \in \text{Range}(K_N), \quad c - b^T K_N^+ b \geq 0.$$

Use eigenvalue decomposition  $K_N = \Gamma \Lambda \Gamma^T$  and define new variables  $\tilde{b}$ ,  $\tilde{c}$  with

$$b = \Gamma \Lambda^{1/2} \tilde{b}.$$

Enforce  $c \geq \tilde{c}^2$  by requiring

$$Z \stackrel{\text{def}}{=} \begin{bmatrix} 1 & \tilde{c} \\ \tilde{c} & c \end{bmatrix} \succeq 0,$$

while  $c - b^T K_N^+ b \geq 0$  is implied by requiring  $x \stackrel{\text{def}}{=} \begin{bmatrix} \tilde{c} \\ \tilde{b} \end{bmatrix}$  to belong to the second-order cone defined by

$$Q_{p+1} = \{x \in \mathbb{R}^{p+1} \mid x_1 \geq \|(x_2, x_3, \dots, x_{p+1})\|_2\}.$$

Enforce the required structure on  $Z$  by setting

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \cdot Z = 1, \quad \begin{bmatrix} 0 & .5 \\ .5 & 0 \end{bmatrix} \cdot Z - \begin{bmatrix} 1 \\ 0 \end{bmatrix}^T x = 0.$$



Defining  $\Sigma = \begin{bmatrix} 0 & 2\Gamma\Lambda^{1/2} \end{bmatrix} \in \mathbb{R}^{N \times p+1}$ , we have

$$\Sigma_{j \cdot x} = 2\Gamma_j\Lambda^{1/2}\tilde{b} = b_j.$$

Hence can capture the discrepancy between  $d_{i,N+1}$  and  $B_{i,N+1} \cdot K_{N+1}$  by the equation

$$d_{i,N+1} - (K_N)_{ii} - \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \cdot Z + \Sigma_{j \cdot x} + u_i - v_i = 0, \quad (u_i, v_i) \geq 0.$$

Objective is  $\sum_{i \in \Psi} (u_i + v_i)$ .

Get a conic program with a  $2 \times 2$  SDP variable, a SOC variable of dimension  $\text{rank}(K_N)$ , and  $2|\Psi|$  linear terms.



## Analyzing the Solution

Having calculated  $K$ , Aim to find a set of low-dimensional  $x_i$  such that  $K_{ij} = x_i^T x_j$ . Do an eigenvalue decomposition

$$K = U\Lambda U^T$$

with  $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_N)$ .

As tuning parameter  $\lambda$  increases, the eigenvalues  $\lambda_i$  drop off more rapidly. Choose a “cutoff” ( $p$ , say) and define the  $x_i \in \mathbb{R}^p$  as follows:

$$\begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix} = \begin{bmatrix} U_{\cdot 1} \sqrt{\lambda_1} & \cdots & U_{\cdot p} \sqrt{\lambda_p} \end{bmatrix}.$$



## Application: Protein Clustering

Try to infer protein function from sequence similarity.

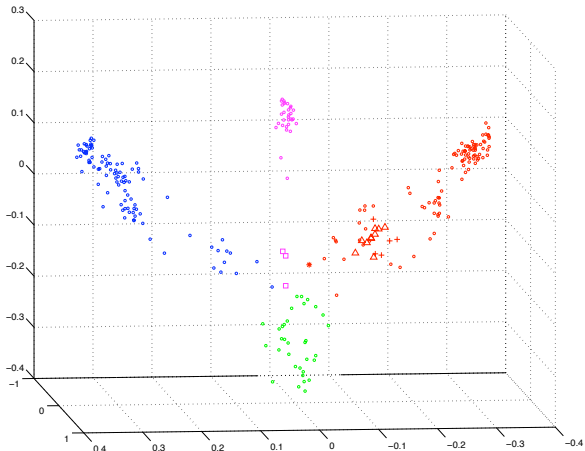
- Cluster proteins in to subfamilies to ease classification (full SDP formulation)
- Assigning new unannotated proteins to the nearest class (“newbie”)

Choose 280 proteins from a database of 630. Four classes: alpha-globins, beta-globins, myoglobins, globins (heterogeneous).

Solve the SDP formulation and reduce to 3 dimensions. The four classes appear as distinct clusters. (Note that the three fish proteins are slightly removed from the other myoglobins.)



## Incremental Formulation





Add three sets of test proteins to the previous set, solve the Newbie problem for each.

- Hemoglobin zeta chain from a goat
- Hemoglobin theta chain from a pig
- 17 Leghemoglobins.

Note that each of the 3 classes fits neatly within one of the existing clusters. Consistent with results of previous studies based on hidden Markov models.



## Incremental Formulation

