A Variety of Regularization Problems

Beginning with a review of some optimization problems in RKHS, and going on to a model selection problem via Likelihood Basis Pursuit (LBP).

LBP based on a paper by Hao Helen Zhang (NC State) with
Grace Wahba, Yi Lin, Meta Voelker, Michael Ferris, Ronald Klein, and Barbara Klein

IPAM "Inverse Problems: Computational Methods and Emerging Applications",
UCLA October 15, 2003

`http://www.stat.wisc.edu/~wahba`
References since 1993 available via the TRLIST link.

We review a selected class of regularization problems which balance distance to the observations with a penalty on the complexity or size of the solution. Considered are a variety of definitions of 'closeness', and several selected penalties, based on RKHS or l_1 norms. A class of tuning methods which generalize the Generalized Cross Validation (GCV) to distance criteria other than least squares are noted. Smoothing Spline ANOVA (SS-ANOVA) models will be described, and their use in a study selecting important important factors affecting the risk of progression of an eye disease, based on data from a demographic study. Based on H. Zhang et al, TR 1059r available via `http://www.stat.wisc.edu/~wahba`, click on "TRLIST".

References

[ZWL02]H. Zhang, G. Wahba, Y. Lin, M. Voelker, M. Ferris, R. Klein, and B. Klein. Variable selection and model building via likelihood basis pursuit. Technical Report 1059, UW-Madison Statistics Department, 2002, under review.[on web]

[CDS98]S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20:33–61, 1998.

[T96]R. J. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, B*, 58:267–288, 1996.

[F98]W. J. Fu. Penalized regression: the bridge versus the LASSO. *Journal of Computational and Graphical Statistics*, 7:397–416, 1998.

References (cont.)

[XW96] Xiang, D. and Wahba, G. A Generalized Approximate Cross Validation for Smoothing Splines with Non-Gaussian Data. Statistica Sinica, 6, 1996, pp.675-692.

[WWGKK95]G. Wahba, Y. Wang, C. Gu, R. Klein, and B. Klein. Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *Ann. Statist.*, 23:1865–1895, 1995.

[W90] G. Wahba. Spline models for observational data, SIAM, 1990.

References (cont.)

[LLW02] Lee, Y, Lin, Y. and Wahba, G. Multicategory Support Vector Machines, Theory, and Application to the Classification of Microarray Data and Satellite Radiance Data ” TR 1064, September 2002. TR1064r, May 2003. The SVM variational problem.

[WLL02] Wahba, G., Lin, Y., and Leng, C. Penalized Log Likelihood Density Estimation via Smoothing-Spline ANOVA and ranGACV - Comments to Hansen and Kooperberg ‘Spline Adaptation in Extended Linear Models’. *Statistical Science* 17:33-37, 2002.

[WLLZ01] Wahba, G., Lin, Y., Lee, Y. and Zhang, H. On the Relation Between the GACV and Joachims’ $\xi\alpha$ Method for Tuning Support Vector Machines, With Extension to the Nonstandard Case ” TR 1039, June 2001.

References (cont.)

[NYGP84] D. Nychka, G. Wahba,S. Goldfarb and T. Pugh. Cross-validated spline methods for the estimation of three dimensional tumor size distributions from observations on two dimensional cross sections, *J. Am. Stat. Assoc.*, 79:832-846, 1984.

OUTLINE

1. Review of positive definite matrices and functions, regularization problems in RKHS

2. Varieties of cost functions and tuning methods

3. SS-ANOVA, or, ANalysis Of VAriance in RKHS

4. The Wisconsin Epidemiological Study of Diabetic Retinopathy (WESDR)

5. What is Likelihood Basis Pursuit (LBP)?

6. Why do l_1 penalties give sparser solutions?

7. Building an overcomplete set of basis functions for LBP, from an SS-ANOVA model.

8.Back to the WESDR results.

9. Closing remarks.

♣♣ 1. Positive definite matrices and functions.

Let $\mathcal{T}$ be an index set. A symmetric function of two variables, $K(s,t), s,t \in \mathcal{T}$ is said to be positive definite (pd) if, for every n and $t_1, \cdots, t_n \in \mathcal{T}$, and every $a_1, \cdots, a_n$,

$$\sum_{i,j=1}^{n} a_i a_j K(t_i, t_j) \geq 0.$$

In the case $\mathcal{T} = \{1, 2, \cdots, N\}$ K reduces to an $N \times N$ matrix. But we will be interested in a (limitless) variety of other index sets-anything on which you can construct a positive definite function:

$\mathcal{T} = (\ldots, -1, 0, 1, \ldots)$
$\mathcal{T} = [0, 1]$
$\mathcal{T} = E^d$ (Euclidean *d*-space)
$\mathcal{T} = \mathcal{S}$ (the unit sphere)
$\mathcal{T} =$ the atmosphere
$\mathcal{T} = \{\diamondsuit, \triangle, \heartsuit\}$ (unordered set)
$\mathcal{T} =$ A Riemannian manifold
$\mathcal{T} =$ A collection of trees

etc, etc.

♣♣ 1. (cont.) Positive definite matrices and functions.

For matrices A and B of appropriate dimensions, the sum $(A \oplus B)$, and the (Kronecker) product,

$$A \otimes B = \begin{pmatrix} a_{11}B & \dots & a_{1n}B \\ a_{21}B & \dots & a_{2n}B \\ \vdots & & \vdots \\ a_{n1}B & \dots & a_{nn}B \end{pmatrix}$$

are pd, and this carries over to positive definite functions on arbitrary domains: Let

$$u, u' \in \mathcal{T}^{(1)}, v, v' \in \mathcal{T}^{(2)}$$

$$s = (u, v), t = (u', v') \in \mathcal{T} = \mathcal{T}^{(1)} \otimes \mathcal{T}^{(2)}$$

$$K_1(u, u'),\ K_2(v, v') \text{ be pd.}$$

Then $K \equiv K_1 \otimes K_2$:

$$K(s, t) = K_1(u, u')K_2(v, v')$$

is pd on $\mathcal{T} \otimes \mathcal{T}$. Thus tensor sums and products of pd functions on arbitrary domains provide an inexhaustible source of models.

♣♣ 1. (cont.) Reproducing kernel Hilbert spaces (RKHS).

Recall: An RKHS (reproducing kernel Hilbert space) is a Hilbert space $\mathcal{H}_K$of functions on a domain $\mathcal{T}$ with all the evaluation functionals $t : f \rightarrow f(t)$ bounded. That is, for each $t \in \mathcal{T}$ there exists a *representer* $\eta_t \in \mathcal{H}_K$ such that $f(t) =< \eta_t, f >_{\mathcal{H}_K}$.

Furthermore, let $K(s,t) =< \eta_s, \eta_t >_{\mathcal{H}_K}$. Thus, K is a uniquely determined pd function, and the famous Moore-Aronszajn theorem says that the converse is true: to each positive definite function on $\mathcal{T} \otimes \mathcal{T}$ there corresponds a unque RKHS $\mathcal{H}_K$, with

$$\eta_t(\cdot) = K(t, \cdot).$$

η_t is the so-called 'representer of evaluation' at t.

♣♣ 1. (cont.) Regularization Problems in RKHS, single smoothing parameter.

The canonical regularization problem in RKHS: Given

$$\{y_i, t_i\}, y_i \in \mathcal{Y}, t_i \in \mathcal{T},$$

and $\{\phi_1, \cdots, \phi_M\}$, M special functions defined on $\mathcal{T}$. Find f of the form

$$f = \sum_{\nu=1}^{M} d_\nu \phi_\nu + h$$

with $h \in \mathcal{H}_K$ to minimize

$$\mathcal{I}_\lambda\{y, f\} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{C}(y_i, f(t_i)) + \lambda \|h\|^2_{\mathcal{H}_K}.$$

$\mathcal{C}$ is a convex function of f for each $y_i \in \mathcal{Y}$ and it is required that the minimizer of $\frac{1}{n}\sum_{i=1}^{n} \mathcal{C}(y_i, f(t_i))$ in the span of the ϕ_ν be unique. $f(t_i)$ may be replaced by $L_i(f)$, where $L_i(f)$ is a bounded linear functional on $\mathcal{H}_K$ and well defined on the ϕ_ν: For example:

$$L_i(f) = \int_{\mathcal{T}} H_i(s) f(s) ds.$$

For some $\mathcal{H}$, observed derivatives can also be used. So a wide variety of observation types can be used.

♣♣ 1. (cont.) Regularization Problems in RKHS, the representer theorem.

$$\mathcal{I}_\lambda\{y, f\} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{C}(y_i, f(t_i)) + \lambda\|h\|^2_{\mathcal{H}_K}.$$

$\mathcal{C}$ measures "fit to data", $\|h\|^2_{\mathcal{H}_K}$ is "complexity" and λ governs their tradeoff. The minimizer of $\mathcal{I}\{f, y\}$ has a representation of the form:

$$f(s) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(s) + \sum_{i=1}^{n} c_i K(t_i, s).$$

$d = (d_1, \cdots d_M)'$ and $c = (c_1, \cdots, c_n)'$ are found using

$$\|\sum_{i=1}^{n} c_i K(t_i, \cdot)\|^2_{\mathcal{H}_K} = c' K_n c$$

where K_n is the $n \times n$ matrix with i, jth entry $K(t_i, t_j)$ [KW71].

♣♣ 1. (cont.) Regularization Problems in RKHS, the representer theorem for indirect observations.

$$\mathcal{I}_\lambda\{y, f\} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{C}(y_i, f(t_i)) + \lambda\|h\|^2_{\mathcal{H}_K}.$$

If $f(t_i)$ is replaced by $L_i(f)$ then $K(t_i, \cdot)$ is replaced by ξ_i obtained by applying L_i to one of the arguments in K, for example if $L_i(f) = \int_{\mathcal{T}} H_i(s)f(s)ds$ then

$$L_i(K(t, \cdot)) = \int_{\mathcal{T}} H_i(s)K(t, s)ds = \xi_i(t),$$

and $(K_n)_{ij}$ is replaced by

$$< \xi_i, \xi_j >= \int_{\mathcal{T}}\int_{\mathcal{T}} H_i(s)K(s, t)H_j(t)dsdt.$$

See [NYGP84]

♣♣ 2. Varieties of Cost Functions and Tuning Methods.

	Cost Function $\mathcal{C}(y, f)$
Regression	
.....	
Gaussian data	$(y-f)^2$
Bernoulli, $f = log[p/(1-p)]$	$-yf + log(1+e^f)$
Other exponential families	other log likelihoods
Data with outliers	robust functionals
Quantile functionals	$\rho_q(y-f)$
.....	
Classification: $y \in \{-1, 1\}$	
.....	
Support vector machines	$(1-yf)_+$
Other "large margin classifiers"	e^{-yf} and other functions of (yf)
.....	
(MV) Density estimation: $y \equiv 1$	$-yf + \int e^f$

(Here $(\tau)_+ = \tau, \tau \geq 0, = 0$ otherwise,
$\rho_q(\tau) = \tau(q - I(\tau \leq 0))$.

♣♣ 2. (cont.) Varieties of cost functions and tuning methods.

Tuning methods for choosing λ from the data:

- Gaussian Data: Generalized Cross Validation (GCV), Generalized Maximum Likelihood (GML)(aka REML), Unbiassed risk (UBR), others (google "methods" "choose" "smoothing parameter" gave 2850 hits)

- Bernoulli Data: Generalized Approximate Cross Validation (GACV) [XW96],other earlier related

- Support Vector Machines: GACV for SVM's [WLZ00] other related, esp. Joachim's $\xi\alpha$ method.

- Multivariate Density Estimation: GACV for density estimation. [WLL02]

- All problems: Leaving-out-one, k-fold cross validation

♣♣♣ 3. Smoothing Spline ANOVA, or, ANalysis Of Variance in RKHS.

Some background:

Let $\mathcal{H}$ be the direct sum of p orthogonal subspaces,

$$\mathcal{H}_K = \sum_{\beta=1}^{p} \oplus \mathcal{H}_\beta$$

In the penalty functional $I_\lambda\{y, f\}$, replace $\lambda\|h\|^2_{\mathcal{H}_K}$ by

$$\sum_{\beta=1}^{p} \lambda_\beta \|P^\beta h\|^2_{\mathcal{H}_K} \equiv \sum_{\beta=1}^{p} \lambda_\beta \|P^\beta h\|^2_{\mathcal{H}_\beta}$$

where P^β is the orthogonal projection of h onto $\mathcal{H}_\beta$. The representer theorem along with some rescaling of components of K can be used to obtain the desired representers with the multiple smoothing parameters $\{\lambda_\beta\}$ explicitly available for tuning [W90][WWGKK95].

♣♣ 3. (cont.) Smoothing Spline ANOVA.

$$t \equiv (t_1, \cdots, t_d) \in \mathcal{T} \equiv \mathcal{T}^{(1)} \otimes \cdots \otimes \mathcal{T}^{(d)}$$

$$f(t) = f(t_1, \cdots, t_d)).$$

Let $d\mu_\alpha$ be a probability measure on $\mathcal{T}^{(\alpha)}$ and define the averaging operator $\mathcal{E}_\alpha$ on $\mathcal{T}$ by

$$(\mathcal{E}_\alpha f)(t) = \int_{\mathcal{T}^{(\alpha)}} f(t_1, \cdots, t_d) d\mu_\alpha(t_\alpha),$$

giving the SS-ANOVA decomposition of f:

$$f(t_1, \cdots, t_d) = \mu + \sum_\alpha f_\alpha(t_\alpha) + \sum_{\alpha\beta} f_{\alpha\beta}(t_\alpha, t_\beta) + \cdots$$

$$\begin{aligned}
\mu &= \prod_\alpha \mathcal{E}_\alpha f \\
f_\alpha &= (I - \mathcal{E}_\alpha) \prod_{\beta \neq \alpha} \mathcal{E}_\beta f \\
f_{\alpha\beta} &= (I - \mathcal{E}_\alpha)(I - \mathcal{E}_\beta) \prod_{\gamma \neq \alpha, \beta} \mathcal{E}_\gamma f \\
\vdots \quad & \quad \vdots \quad \mathcal{E}_\alpha f_\alpha = 0, \;\; \mathcal{E}_\alpha \mathcal{E}_\beta f_{\alpha\beta} = 0, etc.
\end{aligned}$$

♣♣ 3. (cont.) Smoothing spline ANOVA, or, analysis of variance in RKHS.

The idea behind SS-ANOVA is to construct an RKHS $\mathcal{H}$ of functions on $\mathcal{T}$ so that the components of the SS-ANOVA decomposition represent an orthogonal decomposition of f in $\mathcal{H}$. Then RKHS methods can be used to explicitly impose smoothness penalties of the form $\sum_\alpha \lambda_\alpha J_\alpha(f_\alpha) + \sum_{\alpha\beta} \lambda_{\alpha\beta} J_{\alpha\beta}(f_{\alpha\beta}) + \cdots$, (where, however, the series will be truncated at some point.)

♣♣ 3. (cont.) Smoothing Spline ANOVA.

Let $\mathcal{H}^{(\alpha)}$ be an RKHS of functions on $\mathcal{T}^{(\alpha)}$ with $\int_{\mathcal{T}^{(\alpha)}} f_\alpha(t_\alpha) d\mu_\alpha = 0$ for $f_\alpha(t_\alpha) \in \mathcal{H}^{(\alpha)}$, and let $[1^{(\alpha)}]$ be the one dimensional space of constant functions on $\mathcal{T}^{(\alpha)}$.

Construct $\mathcal{H}$ as

$$\mathcal{H} = \otimes_{\alpha=1}^{d} \left[[1^{(\alpha)}] \oplus \mathcal{H}^{(\alpha)} \right]$$

$$= \otimes_{\alpha=1}^{d} [1^{(\alpha)}] \oplus \sum_{j} \mathcal{H}^{(\alpha)} \oplus \sum_{\alpha<\beta} [\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}] \oplus \cdots ,$$

Factors of the form $[1^{(\alpha)}]$ are omitted whenever they multiply a term of a different form. Thus $\mathcal{H}^{(1)}$ is shorthand for $\mathcal{H}^{(1)} \otimes [1^{(2)}] \otimes \cdots \otimes [1^{(d)}]$ (which is a subspace of $\mathcal{H}$).

The components of the ANOVA decomposition will be in mutually orthogonal subspaces of $\mathcal{H}$.

♣♣ 3. (cont.) Smoothing Spline ANOVA.

Consider

$$I = \prod_{\alpha=1}^{d} [\mathcal{E}_\alpha + (I - \mathcal{E}_\alpha)] =$$

$$\prod_{\alpha=1}^{d} \mathcal{E}_\alpha + \sum_{\alpha=1}^{d} (I - \mathcal{E}_\alpha) \prod_{\beta \neq \alpha} \mathcal{E}_\beta$$

$$+ \sum_{\alpha<\beta} (I-\mathcal{E}_\alpha)(I-\mathcal{E}_\beta) \prod_{\gamma \neq \alpha,\beta} \mathcal{E}_\gamma + \cdots + \prod_{\alpha=1}^{d} (I-\mathcal{E}_\alpha).$$

and note that the the terms match up with the expansion

$$\otimes_{\alpha=1}^{d} [1^{(\alpha)}] \oplus \sum_j \mathcal{H}^{(\alpha)} \oplus \sum_{\alpha<\beta} [\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}] \oplus \cdots,$$

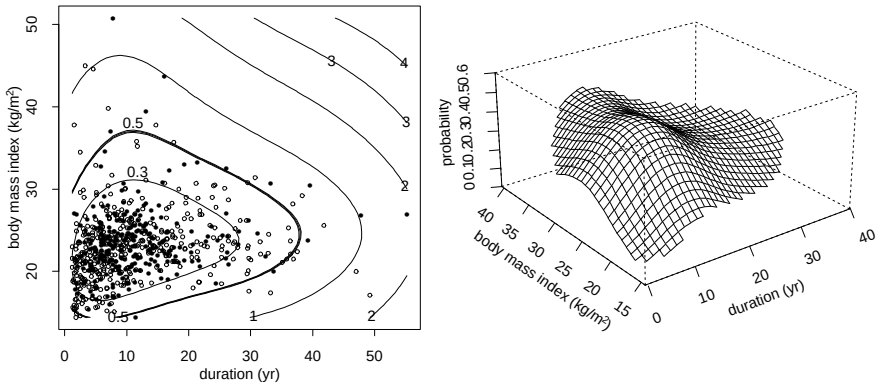
$J_\alpha(f) = \|P^{\mathcal{H}^{(\alpha)}} f\|^2$, and similarly for $J_{\alpha\beta}$. (Details allowing for unpenalized terms omitted here.)

♣♣ 4. The Wisconsin Epidemiological Study of Diabetic Retinopathy (WESDR)

WESDR is an ongoing epidemiological study of a cohort of patients receiving their medical care in an 11-county area in southern Wisconsin. Baseline exam, 1980, with four, ten, fourteen and twenty year followups. (refs in [ZWL02]). [WWGKK95] built a Smoothing Spline ANOVA model for four year risk of progression of diabetic retinopathy from baseline, as a function of three risk factors. We started out with twenty possible risk factors, and narrowed it down to three variables by very laborious means. It would be highly desirable to have a model selection procedure that could simultaneously select important variables/components of a Smoothing Spline ANOVA model. Such a procedure has been obtained in [ZWL02].

♣♣ 4. (cont.). Wisconsin Epidemiological Study of Diabetic Retinopathy (WESDR).

[WWGKK95] looked for the four year risk of progression of diabetic retinopathy from baseline at a cohort of (selected) $n = 669$ younger onset subjects. Let $p(t)$ be the probability of progression for a subject with risk factor vector t and $f = log[p/(1-p)]$. y (coded as 1 or 0) and t is observed for each subject. The model fitted was $f(t) = f(\texttt{dur}, \texttt{gly}, \texttt{bmi}) = \mu + f_1(\texttt{dur}) + a_2 \cdot \texttt{gly} + f_3(\texttt{bmi}) + f_{13}(\texttt{dur}, \texttt{bmi})$ where `dur` = duration of diabetes, `gly` = glycosylated hemoglobin, and `bmi` = body mass index.



(Right: `probability` plotted against `bmi` and `dur`; `gly` at median.) (Note: model is not monotonic in `dur`) From [WWGKK95]

Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR)

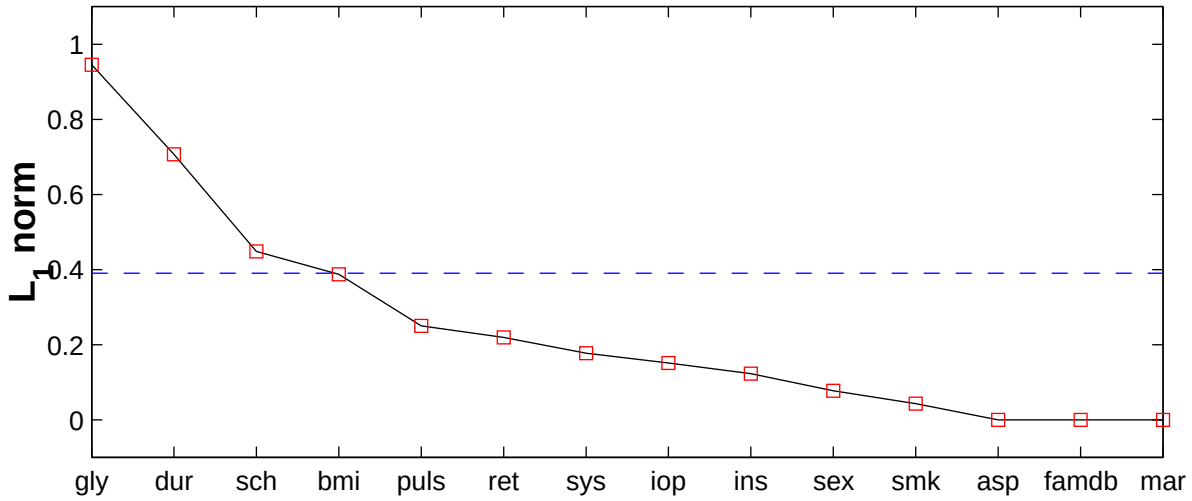
- Continuous covariates:

X_1:	(dur)	duration of diabetes at the time of baseline examination, years
X_2:	(gly)	glycosylated hemoglobin, a measure of hyperglycemia, %
X_3:	(bmi)	body mass index, kg/m^2
X_4:	(sys)	systolic blood pressure, $mmHg$
X_5:	(ret)	retinopathy level
X_6:	$(pulse)$	pulse rate, count for 30 seconds
X_7:	(ins)	insulin dose, kg/day
X_8:	(sch)	years of school completed
X_9:	(iop)	intraocular pressure, $mmHg$

- Categorical covariates:

Z_1:	(smk)	smoking status	(0 = no, 1 = any)
Z_2:	(sex)	gender	(0 = female, 1 = male)
Z_3:	(asp)	use of at least one aspirin for at least three months while diabetic	(0 = no, 1 = yes)
Z_4:	$(famdb)$	family history of diabetes	(0 = none, 1 = yes)
Z_5:	(mar)	marital status	(0 = no, 1 = yes/ever)

♣♣ 4. (cont.) WESDR: The Likelihood Basis Pursuit result for the WESDR data.



L_1 norm scores for the WESDR main effects model, from a Likelihood Basis Pursuit analysis. The method selected `gly`, `dur`, `sch` and `bmi`, in that order, as important variables in a Smoothing Spline ANOVA main effects model.

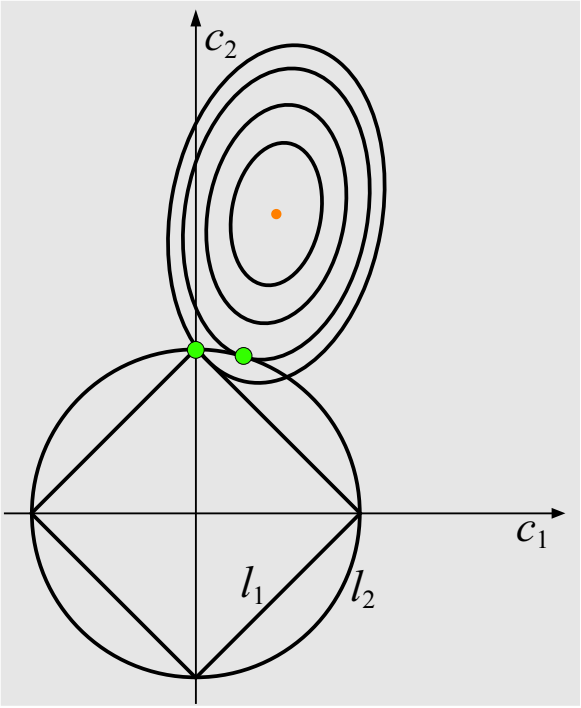
Next: How it was done.

♣♣ 5. What is Likelihood Basis Pursuit (LBP)?

Likelihood Basis Pursuit combines ideas from the LASSO [T96][F98], basis pursuit [CDS98], and Smoothing Spline ANOVA models to generate an overcomplete set of basis functions, which are then used in a penalized likelihood variational problem with l_1 penalties. Basis pursuit uses l_1 penalties, instead of quadratic penalties, to obtain solutions which are relatively sparse in the number of basis functions with non-0 coefficients.

Why do l_1 penalties result in sparser solutions than quatratic penalties?

♣♣ 6. Why do l_1 penalties give sparser solutions?



Circle: $\sum_{j=1}^{N} c_j^2 \leq M$, diamond: $\sum_{j=1}^{N} |c_j| \leq M$.
Ellipses: contours of constant $\sum_{i=1}^{n}(y_i - \sum_{j=1}^{N} x_{ij}c_j)^2$.
Find c to minimize:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{N} x_{ij}c_j)^2$$

subject to $\sum_{j=1}^{N} |c_j|^p \leq M$. Green dots: minimizers for (l. to r.) $p = 1$ and $p = 2$. Note that $c_1 = 0$ for $p = 1$.

♣♣ 6. (cont.) Why do l_1 penalties give sparser solutions ?

The problem: Find c to minimize:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{N} x_{ij}c_j)^2$$

subject to $\sum_{j=1}^{N} |c_j|^p \leq M$ is generally equivalent to the problem: Find c to min

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{N} x_{ij}c_j)^2 + \lambda \sum_{j=1}^{N} |c_j|^p$$

for some $\lambda = \lambda(M)$.

♣♣ 7. Building an overcomplete set of basis functions for Likelihood Basis Pursuit, with a Smoothing Spline ANOVA model.

1. Main effects model, continuous variables.

First: the usual penalized likelihood: Let $l!k_l(u)$ be the lth Bernoulli polynomial, and let $K(u,v) = k_2(u)k_2(v) - k_4(|u-v|), u, v \in [0,1]$ (spline kernel [W90]). Let $x = (x^1, \cdots, x^d)$, and the observations be $\{y_i, x_i\}$ where $x_i = (x_i^1, \cdots, x_i^d), i = 1, \cdots, n$.

The solution to the problem: Find f (in an appropriate space) of the form $f(x) = \mu + \sum_{\alpha=1}^{d} f_\alpha(x^\alpha)$ to min

$$\frac{1}{n}\sum_{i=1}^{n} \mathcal{C}(y_i, f(x_i)) + \sum_{\alpha=1}^{d} \theta_\alpha^{-1} \int (f_\alpha{}'')^2$$

has a representation

$$f(x) = \mu + \sum_{\alpha=1}^{d} b_\alpha k_1(x^\alpha) + \sum_{i=1}^{n} c_i \left(\sum_{\alpha=1}^{d} \theta_\alpha K(x^\alpha, x_i^\alpha) \right)$$

♣♣ 7. (cont.) Building an overcomplete set of basis functions for Likelihood Basis Pursuit, with a Smoothing Spline ANOVA model.

1. Main effects model, continuous variables (cont.).

Since generally an excellent approximation to the solution to the variational problem can be obtained with fewer basis functions, a selected subset, $x_{i_1}, \cdots, x_{i_N}$, of the x_i can be used to generate the solution. Thus

$$\sum_{i=1}^{n} c_i \left(\sum_{\alpha=1}^{d} \theta_\alpha K(x^\alpha, x_i^\alpha) \right)$$

is replaced by

$$\sum_{j*=1}^{N} c_{j*} \left(\sum_{\alpha=1}^{d} \theta_\alpha K(x^\alpha, x_{j*}^\alpha) \right)$$

where $\{x_{j*} = (x_{j*}^1, \cdots, x_{j*}^d),\ j* = 1, \cdots, N\}$ is the selected subset of the x_i.

♣♣ 7. (cont.) Building an overcomplete set of basis functions.

1. Main effects model, continuous variables (cont.). This suggests the overcomplete set of $1 + d + dN$ basis functions:

$$\{1,\ b^{\alpha}(x) \equiv k_1(x^{\alpha}),\ B^{\alpha}_{j*}(x) \equiv K(x^{\alpha}, x^{\alpha}_{j*})\}.$$

for $\alpha = 1, \cdots, d, j* = 1, \cdots, N$

The basis pursuit variational problem is to find $f(x)$,

$$f(x) = \mu + \sum_{\alpha=1}^{d} b_{\alpha} b^{\alpha}(x) + \sum_{\alpha=1}^{d} \sum_{j*=1}^{N} c_{\alpha j*} B^{\alpha}_{j*}(x)$$

to minimize

$$\frac{1}{n} \sum_{i=1}^{n} \mathcal{C}(y_i, f(x_i)) + \lambda_{\pi} \left(\sum_{\alpha=1}^{d} |b_{\alpha}| \right) + \lambda_s \left(\sum_{\alpha=1}^{d} \sum_{j*=1}^{N} |c_{\alpha j*}| \right)$$

2. Two factor interactions basis funtions are built up from tensor products of the main effects basis functions.

♣♣ 7. (cont.) Building an over complete set of basis fuctions .. the importance measure for individual model components.

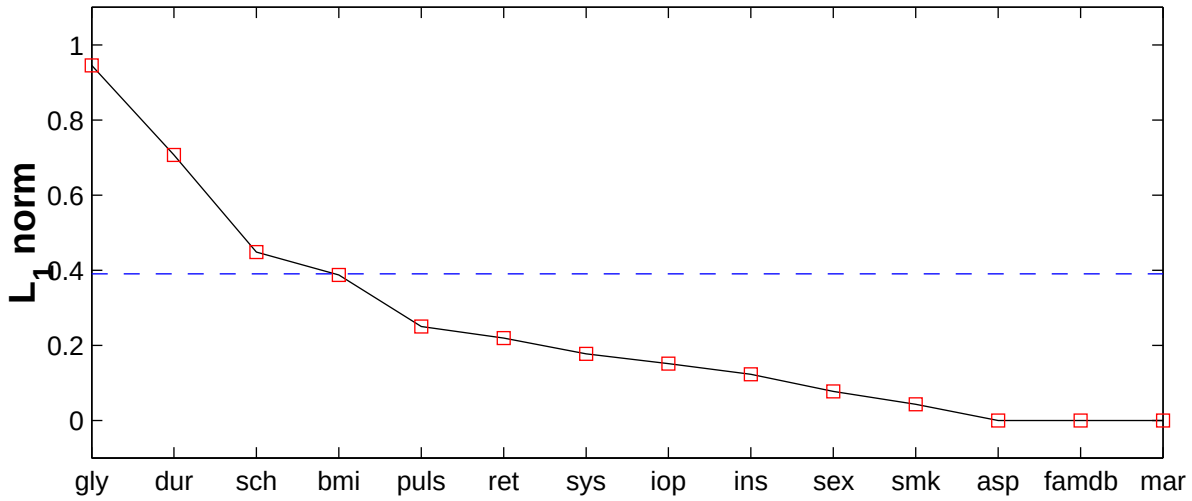
We have adopted the empirical L_1 norm to assess the relative importance of the various terms. (Since the Smoothing Spline ANOVA basis functions all average to 0 this is makes sense). The empirical L_1 norms of the main effects f_α and the two-factor interactions $f_{\alpha\beta}$ are defined as

$$\begin{aligned} L_1(f_\alpha) = & \ \frac{1}{n}\sum_{i=1}^{n} |f_\alpha(x_i^\alpha)| \\ = & \ \frac{1}{n}\sum_{i=1}^{n} |b_\alpha k_1(x_i^\alpha) + \sum_{j=1}^{N} c_{\alpha j*} K_1(x_i^\alpha, x_{j*}^\alpha)| \end{aligned}$$

and

$$\begin{aligned} L_1(f_{\alpha\beta}) = & \ \frac{1}{n}\sum_{i=1}^{n} |f_{\alpha\beta}(x_i^\alpha, x_i^\beta)| \\ = & \ \frac{1}{n}\sum_{i=1}^{n} |b_{\alpha\beta} k_1(x_i^\alpha) k_1(x_i^\beta) \\ + & \ \sum_{j=1}^{N} c_{\alpha\beta j*}^{\pi s} K_1(x_i^\alpha, x_{j*}^\alpha) k_1(x_i^\beta) k_1(x_{j*}^\beta) \\ + & \ \sum_{j=1}^{N} c_{\beta\alpha j*}^{\pi s} K_1(x_i^\beta, x_{j*}^\beta) k_1(x_i^\alpha) k_1(x_{j*}^\alpha) \\ + & \ \sum_{j=1}^{N} c_{\alpha\beta j*}^{ss} K_1(x^\alpha, x_{j*}^\alpha) K_1(x^\beta, x_{j*}^\beta)|. \end{aligned}$$

♣♣ 8. Back to the WESDR results.



L_1 norm scores for the WESDR main effects model.

The importance threshhold is .39 (blue line). The p-values for all four of the selected variables `gly`, `dur`, `sch`, `bmi`, were .02, obtained by a bootstrap procedure [ZWL02]. The solution was very sparse, with about 90% of the coefficients 0.

♣♣ 9. Closing Remarks: Results nice - - the method returned important variables, that had previously been selected by much more tedious methods. Simulation results in [ZWL02] also demonstrated the efficacy of the approach in data sets where 'truth' is known.