Different penalties in the method of regularization: two examples

Yi Lin

University of Wisconsin – Madison

Contents

1. Gaussian reproducing kernel penalty in regression.
2. Sparsity penalty in nonparametric functional ANOVA.

The regression problem

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad i = 1, ..., n,$$

f: unknown regression function to be estimated.

$\mathbf{x}_i = (x_i^{(1)}, ..., x_i^{(d)}) \in R^d$.

ϵ: i.i.d. noise with mean 0 and variance σ^2.

The method of regularization

Find $f \in \mathcal{H}_K$ minimizing

$$\frac{1}{n}\sum_{i=1}^{n}[y_i - f(\mathbf{x}_i)]^2 + \lambda J(f),$$

where $\mathcal{H}_K$ is the RKHS corresponding to the reproducing kernel K, and $J(\cdot)$ is a penalty functional in $\mathcal{H}_K$, typically a squared norm or semi-norm.

Example 1 *Cubic smoothing spline: find $f \in S_2$ minimizing*

$$\frac{1}{n}\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \int_0^1 (f'')^2,$$

where the second order Sobolev Hilbert space S_2 is a RKHS.

Gaussian reproducing kernel

Gaussian reproducing kernel: $G(s,t) \equiv G(s-t)$, where $G(\cdot)$ is the density function of $N(0,\omega^2)$. The corresponding penalty functional is (up to a constant)

$$J_g(f) = \sum_{m=0}^{\infty} \frac{\omega^{2m}}{2^m m!} \int_{-\infty}^{\infty} [f^{(m)}(x)]^2 dx.$$

Periodic Gaussian kernel for estimating periodic functions in $[-\pi,\pi]$: $G^{\infty}(s,t) = G^{\infty}(s-t)$, where $G^{\infty}(r) = \sum_{k=-\infty}^{\infty} G(r-2k\pi)$. The corresponding penalty functional is

$$J_0(f) = \sum_{m=0}^{\infty} \frac{\omega^{2m}}{2^m m!} \int_{-\pi}^{\pi} [f^{(m)}(x)]^2 dx.$$

The corresponding function space (H_{ω}^{∞}) can be seen as Sobolev Hilbert space of infinite order.

The white noise problem

$$Y_n(t) = \int_{-\pi}^{t} f(u)du + n^{-1/2}B(t), \quad t \in [-\pi, \pi],$$

where $B(t)$ is a standard Brownian motion on $[-\pi, \pi]$ and we observe $Y_n = (Y_n(t), -\pi \leq t \leq \pi)$.

Remark 1 *There are results on the equivalence between the white noise problem and*
Gaussian nonparametric regression (Brown and Low, 1996); density estimation (Nussbaum, 1996); spectral density estimation (Golubev and Nussbaum, 1998); nonparametric generalized regression (Grama and Nussbaum, 1997).

Periodic Gaussian regularization

Lin and Brown (2002):

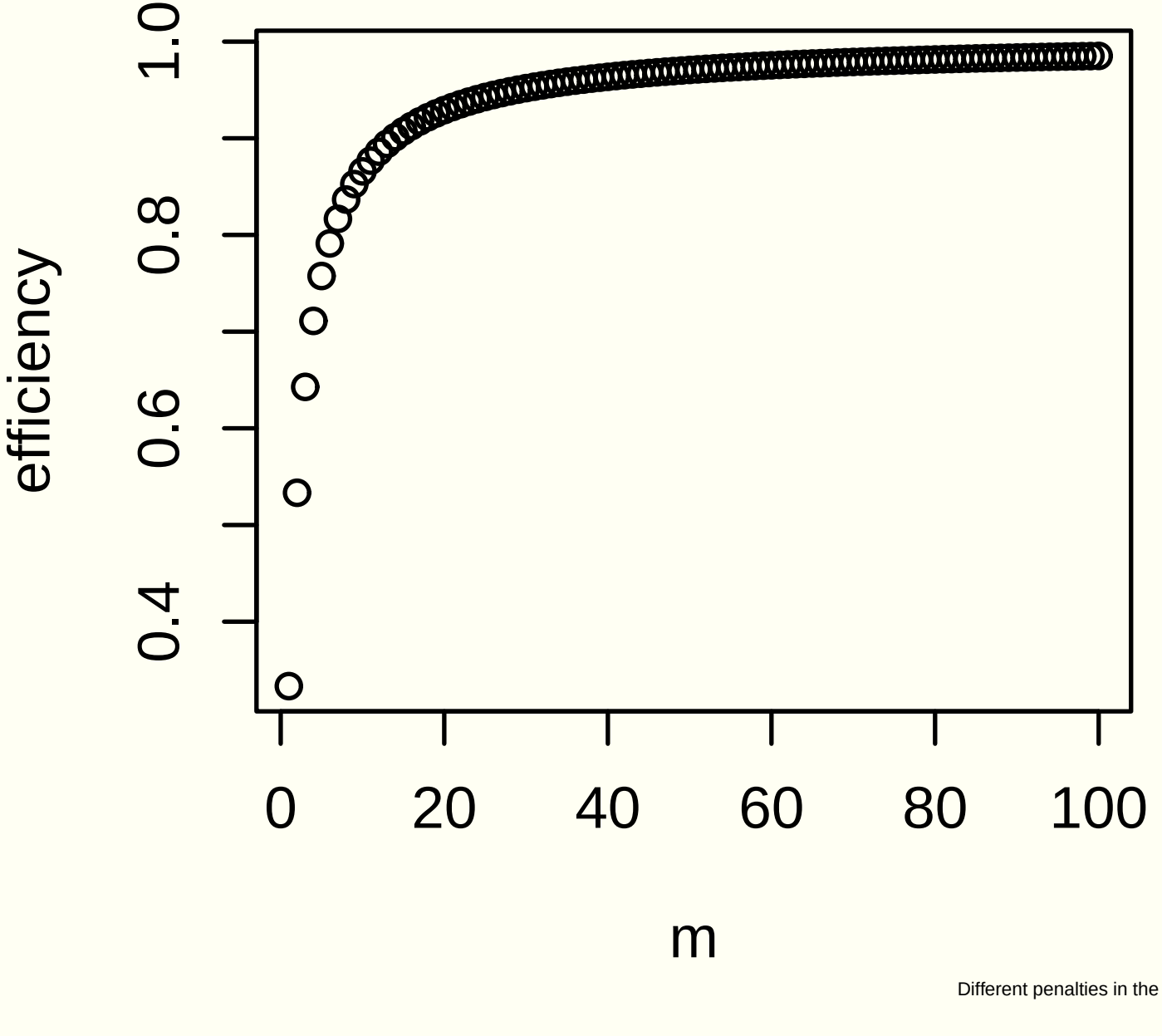
Theorem 1 *The periodic Gaussian regularization in the white noise model is asymptotically minimax in the infinite order Sobolev ball $H_{\omega}^{\infty}(Q)$, if the smoothing parameters are appropriately chosen.*

Theorem 2 *The periodic Gaussian regularization in the white noise model is asymptotically minimax in the analytical function space ball $A_{\alpha}(Q)$, if the smoothing parameters are appropriately chosen.*

Theorem 3 *The periodic Gaussian regularization in the white noise model is rate optimal in the m-th order Sobolev Hilbert ball $H^{m}(Q)$ for $m \geq 1$, if the smoothing parameters are appropriately chosen.*

Theorem 4 *The smoothing parameters can be chosen adaptively without loss of asymptotic efficiency.*

The efficiency in Sobolev balls



Smoothing spline ANOVA

$$f(\mathbf{x}) = b + \sum_{j=1}^{d} f_j(x^{(j)}) + \sum_{j<k} f_{jk}(x^{(j)}, x^{(k)}) + \cdots,$$

where the identifiability of the terms is assured by side conditions through averaging operators. The sequence is usually truncated to enhance interpretability. [Wahba (1990), Wahba et al. (1995), and Gu (2002)]. SS-ANOVA extends the popular additive model.

The function space for SS-ANOVA

Let H^j be the second order Sobolev space of functions of $x^{(j)}$ over $[0,1]$ with inner product $(g_1, g_2) = \int_0^1 g_1 \int_0^1 g_2 + \int_0^1 g_1' \int_0^1 g_2' + \int_0^1 g_1'' g_2''$. We can write $H^j = \{1\} \oplus \bar{H}^j$. Then

$$\otimes_{j=1}^{d} H^j = \{1\} \oplus \sum_{j=1}^{d} \bar{H}^j \oplus \sum_{j<k} [\bar{H}^j \otimes \bar{H}^k] \oplus \cdots .$$

In general, we can write $f = b + \sum_{\alpha=1}^{p} f^\alpha$, with each component function f^α in a different component space $\mathcal{F}^\alpha$ in the above decomposition. Write

$$\mathcal{F} = \{1\} \oplus \left(\bigoplus_{\alpha=1}^{p} \mathcal{F}^\alpha \right).$$

The COSSO

Smoothing spline: find $f \in \mathcal{F}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(\mathbf{x}_i)\right)^2 + \tau^2 \sum_{\alpha=1}^{p} \theta_\alpha^{-1} \|f^\alpha\|^2,$$

where τ and θ_α's are tuning parameters (confounded).
The COSSO (Lin and Zhang, 2002): find $f \in \mathcal{F}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(\mathbf{x}_i)\right)^2 + \tau^2 \sum_{\alpha=1}^{p} \|f^\alpha\|.$$

The COSSO reduces to LASSO (Tibshirani, 1996) in linear models, but with a different interpretation of the penalty.

Theoretical properties of the COSSO

Write $\mathbf{y} = (y_1, ..., y_n)'$, $\mathbf{f} = (f(\mathbf{x}_1), ..., f(\mathbf{x}_n))$. Define $\|\cdot\|_n$ and $\langle \cdot, \cdot \rangle_n$ in R^n as

$$\|\mathbf{f}\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} f^2(\mathbf{x}_i), \quad \langle \mathbf{f}, \mathbf{g} \rangle_n = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i) g(\mathbf{x}_i);$$

Theorem 5 *Consider the additive model with each component function in a second order Sobolev space. Assume ϵ_i's are i.i.d. $N(0, \sigma^2)$ noises. Let $\hat{f}$ be the COSSO estimator. Then we have $\|\hat{\mathbf{f}} - \mathbf{f}\|_n = O_p(n^{-2/5})$ when $\tau \sim n^{-2/5}$.*

Tensor product design case

The design points are

$$\{(x_{i_1,1}, x_{i_2,2}, ..., x_{i_d,d}) : i_k = 1, ..., n_k, k = 1, ..., d\},$$

where $x_{j,k} = j/n_k$, $j = 1, ..., n_k$, $k = 1, ..., d$. We assume $N(0, \sigma^2)$ noises for ϵ's.
It can be shown that the COSSO operates on components in a fashion similar to soft thresholding. If $\lambda \to 0$ and $n\lambda \to \infty$, then with probability tending to one, the COSSO chooses the right model structure.

An equivalent form of the COSSO

Find $\theta = (\theta_1, ..., \theta_p) \geq \mathbf{0}$ and $f \in \mathcal{F}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2 + \lambda_0 \sum_{\alpha=1}^{p} \theta_\alpha^{-1} \|f^\alpha\|^2 + \lambda \sum_{\alpha=1}^{p} \theta_\alpha,$$

where λ_0 is a fixed constant, and λ is a tuning parameter.

Lemma 1 *Set* $\lambda = \tau^4/(4\lambda_0)$.

(i) Let $\hat{f}$ *be a COSSO estimate. Set* $\hat{\theta}_\alpha = \lambda_0^{1/2}\lambda^{-1/2}\|\hat{f}^\alpha\|$, *then the pair* $(\hat{\theta}, \hat{f})$ *minimizes the above.*

(ii) On the other hand, if a pair $(\hat{\theta}, \hat{f})$ *minimizes the above, then* $\hat{f}$ *is a COSSO estimate.*

The algorithm

It can be shown

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i R_\theta(\mathbf{x}_i, \mathbf{x}) + b,$$

where $\mathbf{c} = (c_1, ..., c_n)' \in R^n$, $b \in R$, and $R_\theta = \sum_{\alpha=1}^{p} \theta_\alpha R_\alpha$, with R_α being the reproducing kernel of $\mathcal{F}^\alpha$. Therefore

$$\mathbf{f} = R_\theta \mathbf{c} + b\mathbf{1},$$

and COSSO (the equivalent form) is to minimize

$$\|\mathbf{y} - R_\theta \mathbf{c} - b\mathbf{1}\|_n^2 + \lambda_0 \mathbf{c}' R_\theta \mathbf{c} + \lambda \sum_{\alpha=1}^{p} \theta_\alpha,$$

where $\theta_\alpha \geq 0$, $\alpha = 1, ..., p$.

The algorithm (cont.)

If θ's were known, then the problem is exactly a smoothing spline problem.

If $\mathbf{c}$ and b were known, denote $g_\alpha = R_\alpha \mathbf{c}$, and let G be the $n \times p$ matrix with the α-th column being g_α. To solve for $\theta = (\theta_1, ..., \theta_p)'$,

$$\min(\mathbf{z} - G\theta)'(\mathbf{z} - G\theta),$$

subject to $\theta_\alpha \geq 0$, $\alpha = 1, ..., p$, and $\sum_{\alpha=1}^{p} \theta_\alpha \leq M$, where $\mathbf{z} = \mathbf{y} - (1/2)n\lambda_0 \mathbf{c} - b\mathbf{1}$. This is the nonnegative garrote.

The algorithm iterates between the smoothing spline and the nonnegative garrote. Can be viewed as iterative improvements on the smoothing spline.

The one-step update algorithm

For fixed λ_0 and M,

1. Initialization: fix $\theta_\alpha = 1, \forall \alpha = 1, ..., p$.
2. Solve for $\mathbf{c}$ and b with smoothing spline.
3. For the $\mathbf{c}$ and b obtained in step 2, solve for θ with nonnegative garrote.
4. With the new θ, solve for $\mathbf{c}$ and b with smoothing spline.

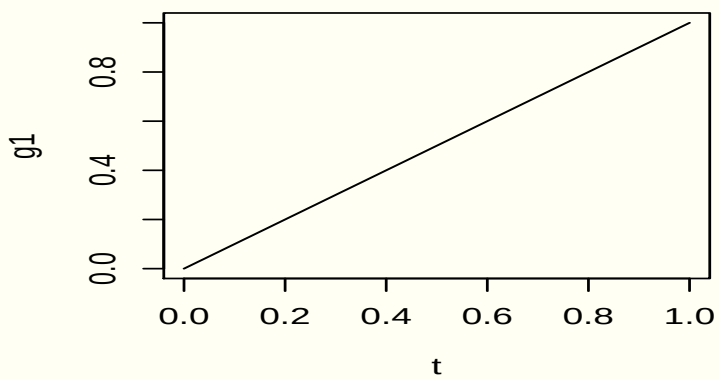
We use 5-fold cross validation to select the tuning parameter. We fix λ_0 at the optimal smoothing spline tuning parameter when θ's are fixed at 1. We tune M between 0 and 35.

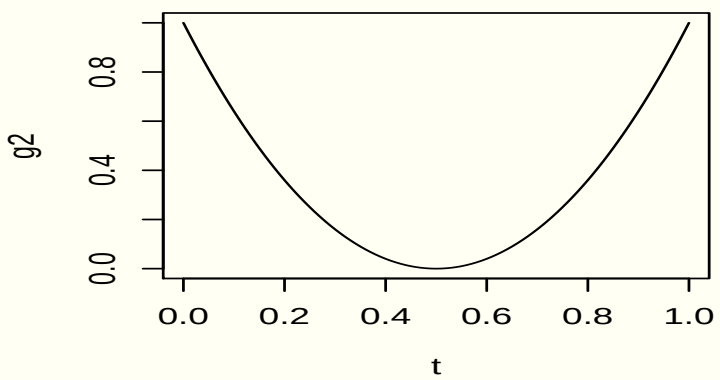
Basic functions used in simulations

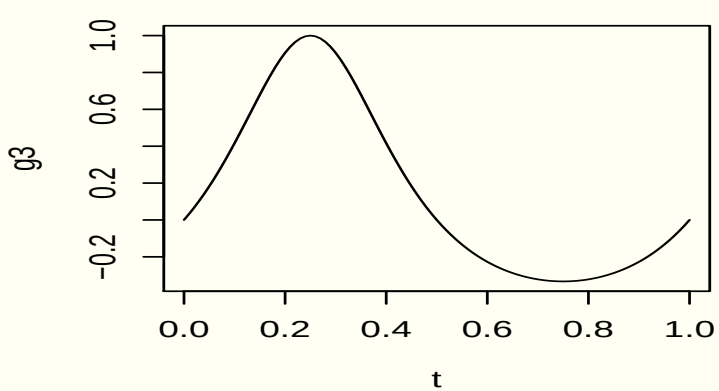
$$g_1(t) = t; \quad g_2(t) = (2t-1)^2; \quad g_3(t) = \frac{\sin(2\pi t)}{2-\sin(2\pi t)};$$

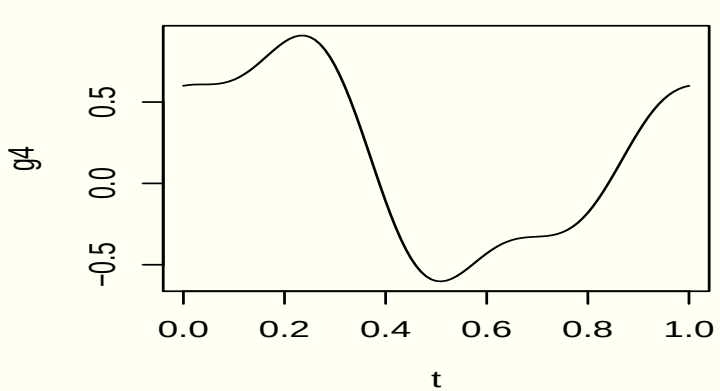
$$\begin{aligned} g_4(t) &= 0.1\sin(2\pi t) + 0.2\cos(2\pi t) + 0.3\sin^2(2\pi t) \\ &\quad +0.4\cos^3(2\pi t) + 0.5\sin^3(2\pi t). \end{aligned}$$

Plots of the basic functions









Distribution of X in the simulations

Compound symmetry: Let W_1, ..., W_d and U be i.i.d from Unif(0,1), and let $X_j = (W_j + tU)/(1+t)$. Therefore $corr(X_j, X_k) = t^2/(1+t^2)$.

(trimmed) AR(1): Let W_1, ..., W_d be i.i.d $N(0,1)$, and let $X_1 = W_1$, $X_j = \rho X_{j-1} + (1-\rho^2)^{1/2} W_j$, $j = 2, ..., d$. Trimmed in [-2.5, 2.5] and scaled to [0, 1].

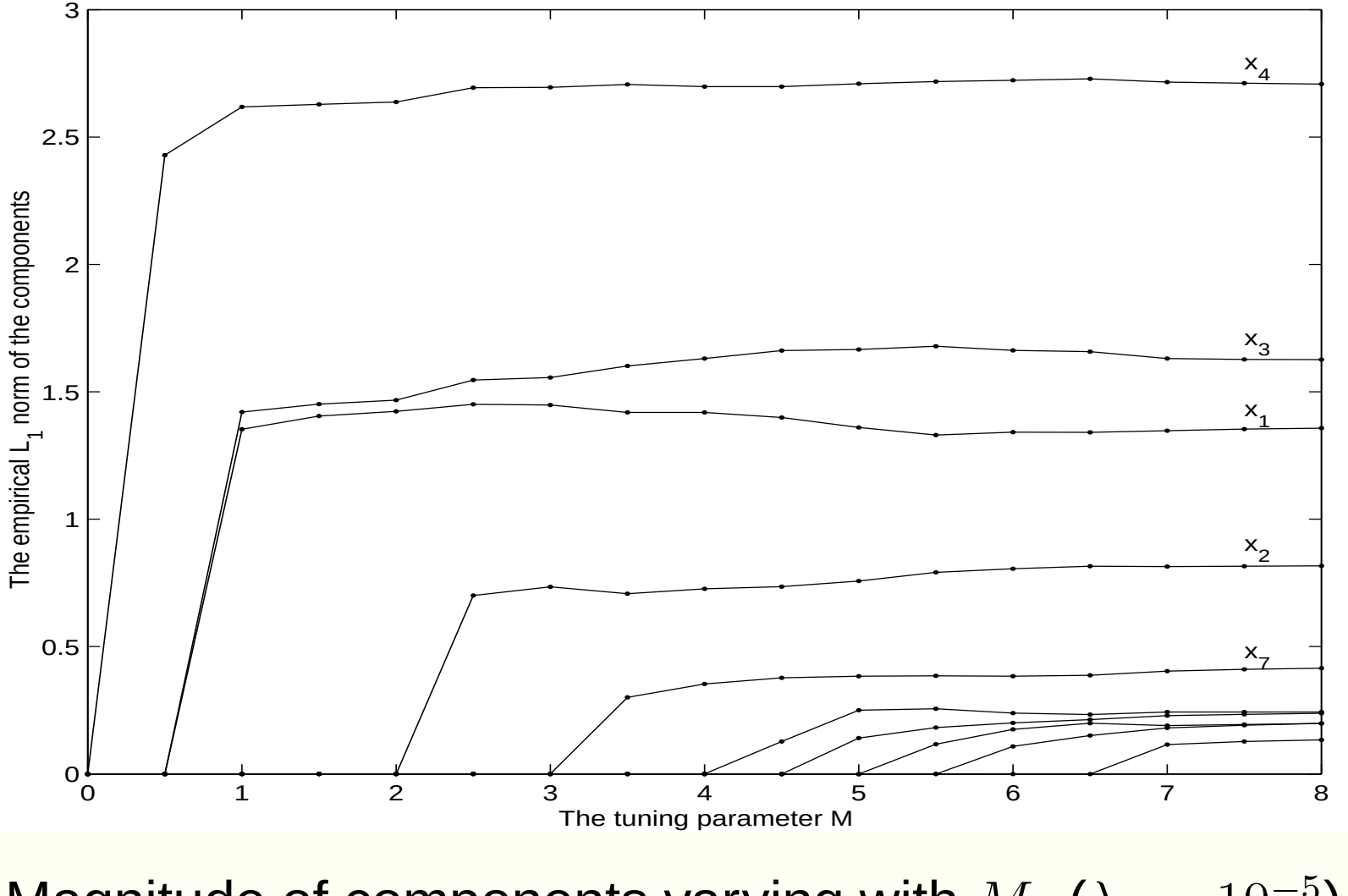
Simulation 1

$n = 100$, $d = 10$. $\epsilon \sim N(0, 1.74)$. Signal to noise ratio is $3 : 1$.

$$f(\mathbf{x}) = 5g_1(x_1) + 3g_2(x_2) + 4g_3(x_3) + 6g_4(x_4).$$

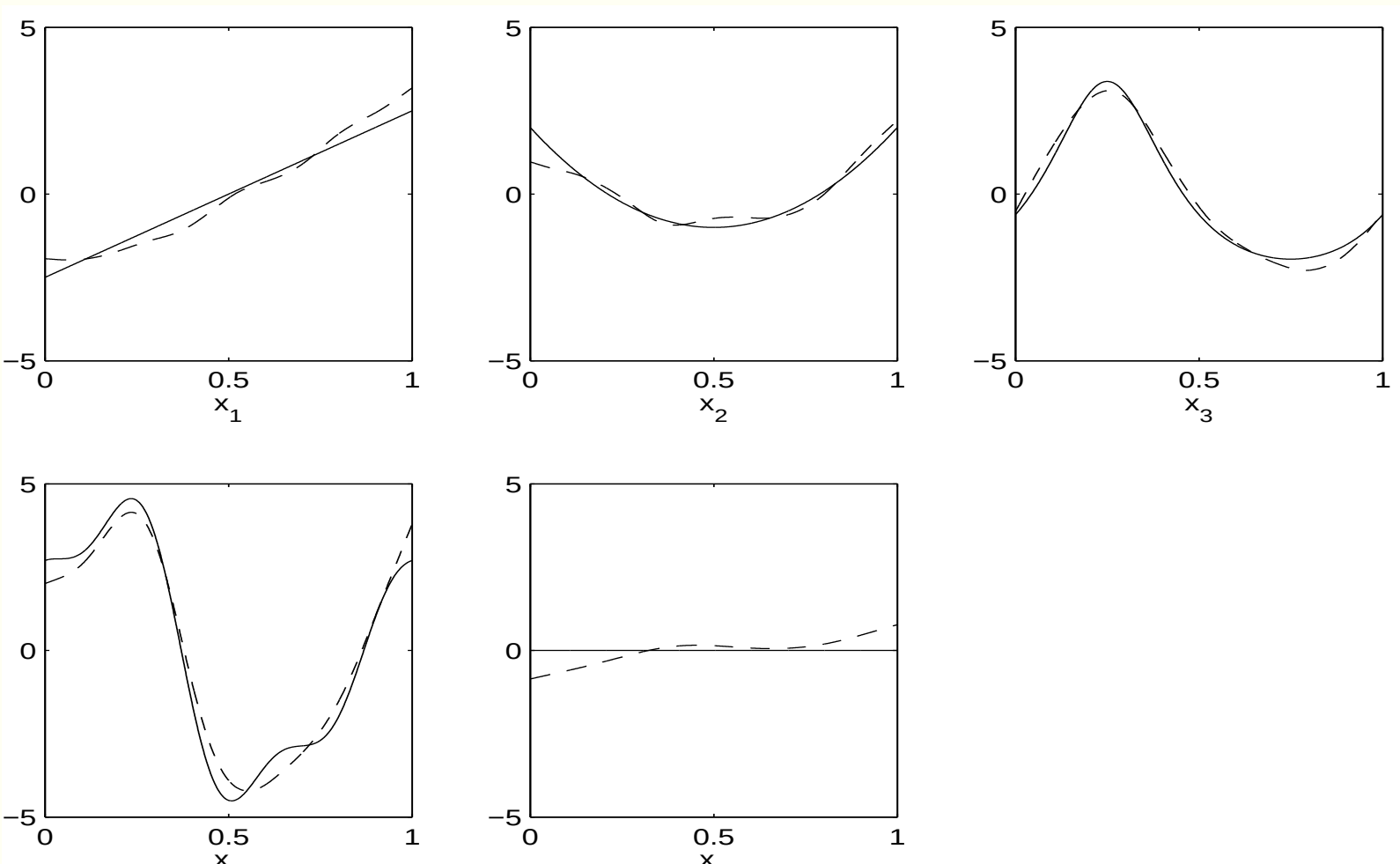
In the uniform setting var$(5g_1(X_1)) = 2.08$, var$(3g_2(X_2)) = 0.80$, var$(4g_3(X_3)) = 3.30$ and var$(6g_4(X_4)) = 9.45$.
Both COSSO and MARS are run with additive models.

Magnitude of the components



Magnitude of components varying with M. ($\lambda_0 = 10^{-5}$)

Estimated function components



The estimated components and the true components. ($M = 3.5$)

Mean integrated squared error (MISE)

Estimated mean integrated squared error (ISE = $E_X[\hat{f}(X) - f(X)]^2$) over 100 runs. The numbers in the parentheses are the standard errors.

Compound Symmetry

	$t = 0$	$t = 1$	$t = 3$
COSSO	0.80 (0.03)	0.97 (0.05)	1.07 (0.06)
MARS	1.57 (0.07)	1.24 (0.06)	1.30 (0.06)

Trimmed AR(1)

	$\rho = -0.5$	$\rho = 0$	$\rho = 0.5$
COSSO	1.03 (0.06)	1.03 (0.06)	0.98 (0.05)
MARS	1.32 (0.07)	1.34 (0.07)	1.36 (0.08)

Term frequencies and model sizes

In the X uniform case, in 100 runs,

Variable	1	2	3	4	5	6	7	8	9	10
COSSO	100	94	100	100	1	1	3	2	4	2
MARS	100	100	100	100	35	35	34	39	28	35

	Model sizes								
	3	4	5	6	7	8	9	10	Mean
COSSO	6	84	7	3	0	0	0	0	4.07
MARS	0	4	24	40	26	6	0	0	6.06

Model sizes in various setting

Mean and standard deviation of the size of the models chosen in 100 runs.

	Comp. symm.		AR(1)		
	$t = 1$	$t = 3$	$\rho = -0.5$	$\rho = 0$	$\rho = 0.5$
COSSO	4.1 (1.2)	4.4 (1.9)	4.1 (1.2)	4.0 (1.0)	3.8 (0.9)
MARS	6.3 (0.9)	6.2 (0.9)	6.1 (1.0)	6.1 (0.8)	5.9 (0.8)

Simulation 2

$n = 500$, $d = 60$, $\epsilon \sim N(0, 0.5184)$.

$$\begin{aligned} f(\mathbf{x}) &= g_1(x_1) + g_2(x_2) + g_3(x_3) + g_4(x_4) \\ &+ 1.5g_1(x_5) + 1.5g_2(x_6) + 1.5g_3(x_7) + 1.5g_4(x_8) \\ &+ 2g_1(x_9) + 2g_2(x_{10}) + 2g_3(x_{11}) + 2g_4(x_{12}). \end{aligned}$$

In the uniform setting var$(g_1(X_1)) = 0.08$, var$(g_2(X_2)) = 0.09$, var$(g_3(X_3)) = 0.21$ and var$(g_4(X_4)) = 0.26$.

Both COSSO and MARS are run with additive models.

100 runs.

MISE and model sizes

Estimated MISE in unit of 10^{-3}

	Comp. symm.		AR(1)	
	$t=0$	$t=1$	$\rho=0.5$	$\rho=-0.5$
COSSO	144 (4)	162 (5)	153 (4)	149 (5)
MARS	353 (7)	302 (7)	286 (6)	280 (5)

Model sizes

	Comp. symm.		AR(1)	
	$t=0$	$t=1$	$\rho=0.5$	$\rho=-0.5$
COSSO	12.0 (0.2)	11.7 (1.4)	12.1 (1.4)	11.9 (1.0)
MARS	35.2 (2.3)	36.1 (2.1)	35.2 (2.5)	35.9 (2.4)

Simulation 3

$d = 10$, uniform setting, $\epsilon \sim N(0, 0.065)$.

$$\begin{aligned} f(\mathbf{x}) &= g_1(x_1) + g_2(x_2) + g_3(x_3) + g_4(x_4) \\ &\quad + g_1(x_3 x_4) + g_2(\frac{x_1 + x_3}{2}) + g_3(x_1 x_2); \end{aligned}$$

	$n = 100$	$n = 200$	$n = 400$
COSSO	0.378 (0.005)	0.094 (0.004)	0.043 (0.001)
MARS	0.239 (0.008)	0.109 (0.003)	0.084 (0.001)

The estimated MISE of COSSO and MARS over 100 runs. Both are run with two-way interaction models.

Circuit examples (Friedman, 1991)

Dependence of the impedance Z and phase shift ϕ on components in the circuit.

$$Z = [R^2 + (\omega L - 1/(\omega C))^2]^{1/2},$$

$$\phi = \tan^{-1}\left[\frac{\omega L - 1/(\omega C)}{R}\right].$$

Input variables uniform in the range $0 \leq R \leq 100$, $40\pi \leq \omega \leq 560\pi$, $0 \leq L \leq 1$, and $1 \leq C \leq 11$.

In the first example, $\epsilon \sim N(0, 15625)$.
In the second example, $\epsilon \sim N(0, 0.01)$.
In both examples, signal to noise ratio is $3:1$.

MISE in circuit examples

Estimating Z (in the unit of 10^3).

	$n = 100$	$n = 200$	$n = 400$
COSSO	1.91 (0.12)	0.85 (0.05)	0.51 (0.03)
MARS	5.57 (0.41)	2.47 (0.16)	1.37 (0.08)

Estimating ϕ (in the unit of 10^{-3}).

	$n = 100$	$n = 200$	$n = 400$
COSSO	12.98 (0.36)	7.96 (0.20)	5.36 (0.10)
MARS	20.59 (0.96)	12.60 (0.71) [a]	8.19 (0.14) [b]

a. Excluded one extreme outlier.

b. Excluded three extreme outliers.

Real examples

Ozone data The daily maximum one-hour-average ozone reading and 8 meteorological variables were recorded in the Los Angeles basin for 330 days of 1976.

Boston housing data Housing values in suburbs of Boston. There are 12 input variables. The sample size is 506.

Tecator data Data recorded on a Tecator Infratec Food and Feed Analyzer working in the wavelength range 850-1050 nm by the Near Infrared Transmission (NIT) principle. Each sample contains finely chopped pure meat with different fat contents. The input vector consists of a 100 channel spectrum of absorbances. As recommended in the document, we use the first 13 principal components to predict the fat content. The total sample size is 215.

Results on real examples

	Ozone	Boston	Tecator
COSSO	16.04 (0.06)	9.89 (0.08)	0.92 (0.02)
MARS	18.24 (0.45)	14.31 (0.34)	4.99 (1.07)

The average prediction error of COSSO and MARS in some real examples as estimated by averaging five 10-fold cross validations. Both COSSO and MARS are run with two-way interaction models.

Summary

Different penalties can be used in the method of regularization for different purposes.

1. The periodic Gaussian kernel regularization adapts to different order of smoothness.
2. The COSSO simultaneously does model selection and estimation. It solves a global minimization problem. This is one reason why it can outperform greedy search type algorithms.

References

Lin, Y. and Brown, L. D. (2002): Statistical properties of the method of regularization with periodic Gaussian reproducing kernel.
Lin, Y. and Zhang, H. (2002): Component selection and smoothing in smoothing spline analysis of variance models.

The papers and software are available at http://www.stat.wisc.edu/˜yilin/papers/papers.html.