Regularization and shrinkage for model selection in sparse GLM models.

Challenging problems in Statistical Learning Workshop

A. Antoniadis
LJK-Université Joseph Fourier.

Grenoble, March 17 & 18, 2011

Introduction

During the 1990s, the nonparametric regression and signal processing literature was dominated by (nonlinear) *wavelet shrinkage* and *wavelet thresholding* estimators.

When sampling points are not equi-spaced, Antoniadis & Fan (2001) address the problem with some new regularization procedures as penalized least squares regression and establish their connexion with model selection in nonparametric regression models.

They suggest using some nonconvex penalties (SCAD) to increase model sparsity and accuracy. This was extended to handle variable selection via penalized ordinary least squares regression in general sparse linear models by Li & Fan (2001).

Summary

Starting from the thresholding rules, we review several thresholding procedures that have been used for wavelet denoising and establish their connexion with penalized ordinary least squares with separable penalties.

When dealing with nonorthogonal designs in high-dimensional linear models sparsity can be achieved via thresholding-based iterative selection procedures for model selection and shrinkage.

Finally, we extend the thresholding iterative procedures to generalized linear models with possibly nonorthogonal designs since one may use them as features selection tools in high-dimensional logistic regression or multinomial regression.

Outline.

Objective: Build a model with a subset of "predictors".

Denoising

– *Wavelet thresholding*

–*Shrinkage and nonlinear diffusion*

Relations to variational methods

– *Convenient penalties*

Extension to nonequispaced designs

– *Connexions with LASSO*

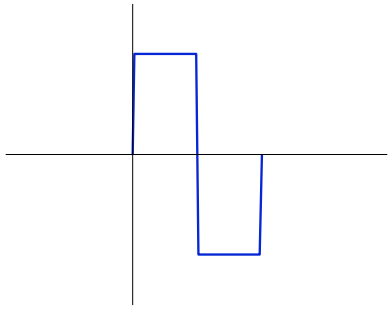
Penalized least squares and iterative thresholding

– *Surrogates and the MM algorithm*

Penalized likelihood and iterative thresholding for GLMs

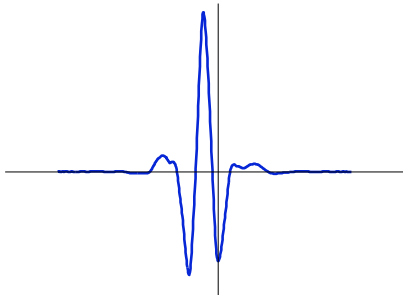
– *Appropriate surrogates*

Wavelet decompositions



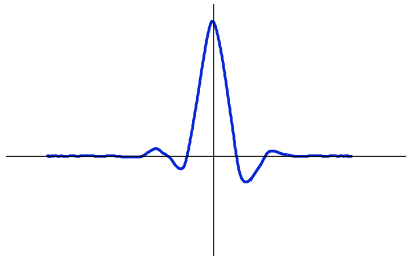
A mother wavelet ψ together with its translations and dilatations

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$$



provide the orthogonal expansion

$$f = \sum_{j,k\in\mathbb{Z}} \langle f, \psi_{j,k}\rangle \psi_{j,k}$$



and with the help of the scaling function ϕ:

$$f = \sum_{k \in \mathbb{Z}} \langle f, \phi_{j_0,k} \rangle \phi_{j_0,k} + \sum_{k \in \mathbb{Z}, j \geq j_0} \langle f, \psi_{j,k} \rangle \psi_{j,k}.$$

The discrete wavelet transform

Given a vector of function values $\mathbf{g} = (g(t_1), ..., g(t_n))'$ at equally spaced points t_i, the discrete wavelet transform of $\mathbf{g}$ is given by $\mathbf{d} = W\mathbf{g}$, where $\mathbf{d}$ is an $n \times 1$ vector comprising both discrete scaling coefficients, $c_{j_0 k}$, and discrete wavelet coefficients, d_{jk}, and W is an orthogonal $n \times n$ matrix associated with the orthonormal wavelet basis chosen.

The $c_{j_0 k}$ and d_{jk} are related to their continuous counterparts $\langle g, \phi_{j_0,k} \rangle$ and $\langle g, \psi_{j,k} \rangle$ (with an approximation error of order n^{-1}) via the relationships

$$c_{j_0 k} \approx \sqrt{n} \, \langle g, \phi_{j_0,k} \rangle \quad \text{and} \quad d_{jk} \approx \sqrt{n} \, \langle g, \psi_{j,k} \rangle.$$

The factor $\sqrt{n}$ arises because of the difference between the continuous and discrete orthonormality conditions.

Denoising by wavelet thresholding

Wavelet series allow a parsimonious and sparse expansion for a wide variety of functions, including inhomogeneous cases.

Due to the orthogonality of the matrix W, the DWT of white noise is also an array of independent $N(0,1)$ random variables, so

$$\begin{aligned} \hat{c}_{j_0 k} &= c_{j_0 k} + \sigma\, \epsilon_{jk}, \quad k = 0, 1, \ldots, 2^{j_0} - 1, \\ \hat{d}_{jk} &= d_{jk} + \sigma\, \epsilon_{jk}, \quad j = j_0, \ldots, J-1, \quad k = 0, \ldots, 2^j - 1, \end{aligned}$$

where $\hat{c}_{j_0 k}$ and $\hat{d}_{jk}$ are respectively the *empirical scaling* and the *empirical wavelet* coefficients of the the noisy data $\mathbf{y}$, and ϵ_{jk} are independent $N(0,1)$ random variables.

Exploiting sparsity

The sparseness of the wavelet expansion makes it reasonable to assume that essentially only a few 'large' d_{jk} contain information about the underlying function g, while 'small' d_{jk} can be attributed to the noise which uniformly contaminates all wavelet coefficients.

Thus, simple denoising algorithms that use the wavelet transform consist of three steps:

1) Calculate the wavelet transform of the noisy signal.

2) Modify the noisy wavelet coefficients according to some rule.

3) Compute the inverse transform using the modified coefficients.

Thresholding rules

Mathematically wavelet coefficients are estimated using either the *hard* or *soft* thresholding rule given respectively by

$$\delta_\lambda^{\mathrm{H}}(\hat{d}_{jk}) = \begin{cases} 0 & \text{if } |\hat{d}_{jk}| \leq \lambda \\ \hat{d}_{jk} & \text{if } |\hat{d}_{jk}| > \lambda \end{cases}$$

and

$$\delta_\lambda^{\mathrm{S}}(\hat{d}_{jk}) = \begin{cases} 0 & \text{if } |\hat{d}_{jk}| \leq \lambda \\ \hat{d}_{jk} - \lambda & \text{if } \hat{d}_{jk} > \lambda \\ \hat{d}_{jk} + \lambda & \text{if } \hat{d}_{jk} < -\lambda. \end{cases}$$

Avantages and disadvantages

Thresholding allows the data itself to decide which wavelet coefficients are significant; hard thresholding (a discontinuous function) is a 'keep' or 'kill' rule, while soft thresholding (a continuous function) is a 'shrink' or 'kill' rule.

Bruce & Gao (1996) and Marron, Adak, Johnstone, Newmann & Patil (1998) have shown that simple threshold values with hard thresholding results in larger variance in the function estimate, while the same threshold values with soft thresholding shift the estimated coefficients by an amount of λ even when $|\hat{d}_{jk}|$ stand way out of noise level, creating unnecessary bias when the true coefficients are large. Also, due to its discontinuity, hard thresholding can be unstable – that is, sensitive to small changes in the data.

Remedies

To remedy the drawbacks of both hard and soft thresholding rules, Gao (1998) considered the *nonnegative garrote* thresholding

$$\delta_\lambda^{\mathrm{G}}(\hat{d}_{jk}) = \begin{cases} 0 & \text{if } |\hat{d}_{jk}| \leq \lambda \\ \hat{d}_{jk} - \frac{\lambda^2}{\hat{d}_{jk}} & \text{if } |\hat{d}_{jk}| > \lambda \end{cases}$$

which also is a "shrink" or "kill" rule (a continuous function). The resulting wavelet thresholding estimators offer, in small samples, advantages over both hard thresholding and soft thresholding.

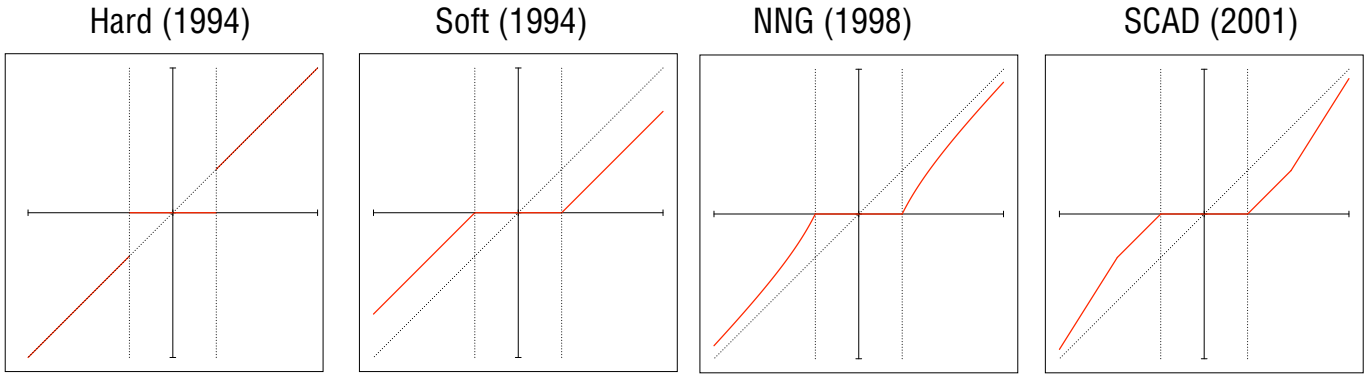
Other rules

In the same spirit to that in Gao (1998), Antoniadis & Fan (2001) (AF for short) suggested the *SCAD* thresholding rule

$$\delta_\lambda^{\text{SCAD}}(\hat{d}_{jk}) = \begin{cases} \text{sign}(\hat{d}_{jk})\max(0, |\hat{d}_{jk}| - \lambda) & \text{if } |\hat{d}_{jk}| \le 2\lambda \\ \frac{(a-1)\hat{d}_{jk} - a\lambda\,\text{sign}(\hat{d}_{jk})}{a-2} & \text{if } 2\lambda < |\hat{d}_{jk}| \le a\lambda \\ \hat{d}_{jk} & \text{if } |\hat{d}_{jk}| > a\lambda \end{cases}$$

which is a "shrink" or "kill" rule (a piecewise linear function). It does not over penalize large values of $|\hat{d}_{jk}|$ and hence does not create excessive bias when the wavelet coefficients are large. AF (2001), based on a Bayesian argument, have recommended to use the value of $\alpha = 3.7$.

Standard thresholding functions δ_λ



Hard : High variance due to discontinuities at $\pm\lambda$

Soft : Oversmoothing (important bias due to constant attenuation)

NNG, SCAD : Compromise between Hard and Soft.

Wavelet shrinkage and nonlinear diffusion

Nonlinear diffusion filtering and wavelet shrinkage are methods that serve the same purpose, namely discontinuity-preserving denoising.

One drawback of the DWT is that the coefficients of the discretized signal are not circularly shift equivariant, so that circularly shifting the observed series by some amount will not circularly shift the discrete wavelet transform coefficients by the same amount, which seriously degrades the quality of the denoising achieved.
The idea of denoising via cycle spinning is to apply denoising not only to $\mathbf{y}$, but also to all possible unique circularly shifted versions of $\mathbf{y}$, and to average the results.

Translation invariant Haar wavelet shrinkage

We can now view a general connection between translation invariant Haar wavelet shrinkage and a discretized version of a nonlinear diffusion. The scaling and wavelet filters h and $\tilde{h}$ corresponding to the Haar transform are

$$h = \frac{1}{\sqrt{2}}(\dots, 0, 1, 1, 0, \dots) \quad \tilde{h} = \frac{1}{\sqrt{2}}(\dots, 0, -1, 1, 0, \dots).$$

Given a discrete signal $f = (f_k)_{k\in\mathbb{Z}}$, we can see that a shift-invariant soft wavelet shrinkage of f on a single level decomposition with the Haar wavelet creates a filtered signal $u = (u_k)_{k\in\mathbb{Z}}$ given by

$$\frac{1}{4}(f_{k-1} + 2f_k + f_{k+1}) + \frac{1}{2\sqrt{2}}\left(-\delta_\lambda^{\mathrm{S}}\left(\frac{f_{k+1}-f_k}{\sqrt{2}}\right) + \delta_\lambda^{\mathrm{S}}\left(\frac{f_k-f_{k-1}}{\sqrt{2}}\right)\right),$$

where $\delta_\lambda^{\mathrm{S}}$ denotes the soft shrinkage operator with threshold λ.

Diffusion

Because the filters of the Haar wavelet are simple difference filters (a finite difference approximation of derivatives) the above rule looks a little like a discretized version of a differential equation.

$$
\begin{aligned}
u_k &= f_k + \frac{f_{k+1} - f_k}{4} - \frac{f_k - f_{k-1}}{4} \\
&\quad + \frac{1}{2\sqrt{2}} \left(-\delta_\lambda^{\mathrm{S}} \left(\frac{f_{k+1} - f_k}{\sqrt{2}} \right) + \delta_\lambda^{\mathrm{S}} \left(\frac{f_k - f_{k-1}}{\sqrt{2}} \right) \right) \\
&= f_k + \left(\frac{(f_{k+1} - f_k)}{4} - \frac{1}{2\sqrt{2}} \delta_\lambda^{\mathrm{S}} \left(\frac{f_{k+1} - f_k}{\sqrt{2}} \right) \right) \\
&\quad - \left(\frac{(f_k - f_{k-1})}{4} - \frac{1}{2\sqrt{2}} \delta_\lambda^{\mathrm{S}} \left(\frac{f_k - f_{k-1}}{\sqrt{2}} \right) \right),
\end{aligned}
$$

we obtain

$$\frac{u_k - f_k}{\Delta t} = (f_{k+1} - f_k)g(|f_{k+1} - f_k|) - (f_k - f_{k-1})g(|f_k - f_{k-1}|),$$

with a function g and a time step size Δt defined by

$$\Delta t\, g(|s|) = \frac{1}{4} - \frac{1}{2\sqrt{2}|s|}\delta_\lambda^{\mathrm{S}}\left(\frac{|s|}{\sqrt{2}}\right).$$

The above appears as a first iteration of an explicit (Euler forward) scheme for a nonlinear diffusion filter with initial state f, time step size Δt and spatial step size 1. Therefore the shrinkage rule corresponds to a discretization of the differential equation $\partial_t u = \partial_x\left((\partial_x u)g(|\partial_x u|)\right)$, with initial condition $u(0) = f$. This equation is a 1-D variant of the Perona-Malik diffusion equation well known in image processing, and the function g is called the *diffusivity*.

Nonlinear diffusion filtering

In the 1-D case the basic idea is to obtain a family $u(x,t)$ of filtered versions of a continuous signal f as the solution of the diffusion process stated in the previous equation with f as initial condition, $u(x,0) = f(x)$ and reflecting boundary conditions.

The diffusivity g controls the speed of diffusion depending on the magnitude of the gradient.

Usually, g is chosen such that it is equal to one for small magnitudes of the gradient and goes down to zero for large gradients. Hence the diffusion stops at positions where the gradient is large. These areas are considered as singularities of the signal.

A connection with shrinkage

A proposition which relates some properties of shrinkage functions and diffusivities which is an easy consequence of the relation between g and δ_λ.

We formulate this relation for the case $\Delta t = 1/4$ which is a common choice and widely used for the Perona-Malik equation.

Let $\Delta t = 1/4$. Then the diffusivity and the shrinkage function are related through

$$g(|x|) = 1 - \frac{\sqrt{2}}{|x|}\delta_\lambda\left(\frac{|x|}{\sqrt{2}}\right).$$

Properties

The following properties hold:

1. If δ_λ performs shrinkage then the diffusion is always forward, i. e.

$$\delta_\lambda(|x|) \leq |x| \Leftrightarrow g(x) \geq 0.$$

2. If δ_λ is differentiable at 0 then, as $x \to 0$,

$$g(x) \to 1 \Leftrightarrow \delta_\lambda(0) = 0 \text{ and } \delta'_\lambda(0) = 0.$$

3. If the diffusion stops for large gradients the shrinkage function has linear growth at infinity, i. e.

$$g(x) \to 0, \text{ as } x \to \infty \Leftrightarrow \frac{\delta_\lambda(x)}{x} \to 1, \text{ as } x \to \infty.$$

Examples

We choose $\Delta t = 1/4$ and derive the corresponding diffusivities by plug in the specific shrinkage function.

Linear shrinkage A linear shrinkage rule, producing linear wavelet denoising is given by $\delta_\lambda(x) = \frac{x}{1+\lambda}$. The corresponding diffusivity is constant $g(|x|) = \frac{\lambda}{(1+\lambda)}$, and the diffusion is linear.

Soft shrinkage The soft shrinkage function $\delta_\lambda(x) = \text{sign}(x)(|x| - \lambda)_+$ gives $g(|x|) = \left(1 - \frac{(|x| - \sqrt{2}\lambda)_+}{|x|}\right)$, which is a stabilized total variation diffusivity.

Hard shrinkage The hard shrinkage function $\delta_\lambda(x) = x(1 - I_{\{|x| \leq \lambda\}}(x))$ leads to $g(|x|) = I_{\{|x| \leq \sqrt{2}\lambda\}}(|x|)$ which is a piecewise linear diffusion that degenerates for

large gradients.

Garrote shrinkage The nonnegative garrote shrinkage $\delta_\lambda(x) = \left(x - \frac{\lambda^2}{x}\right)(1 - I_{\{|x|\leq\lambda\}}(x))$ leads to a stabilized unbounded BFB diffusivity (Keeling and Stollberger (2002)) given by $g(|x|) = I_{\{|x|\leq\sqrt{2}\lambda\}}(|x|) + \frac{2\lambda^2}{x^2} I_{\{|x|>\sqrt{2}\lambda\}}(|x|)$.

Firm shrinkage Firm shrinkage defined yields a diffusivity that degenerates to 0 for sufficiently large gradients:

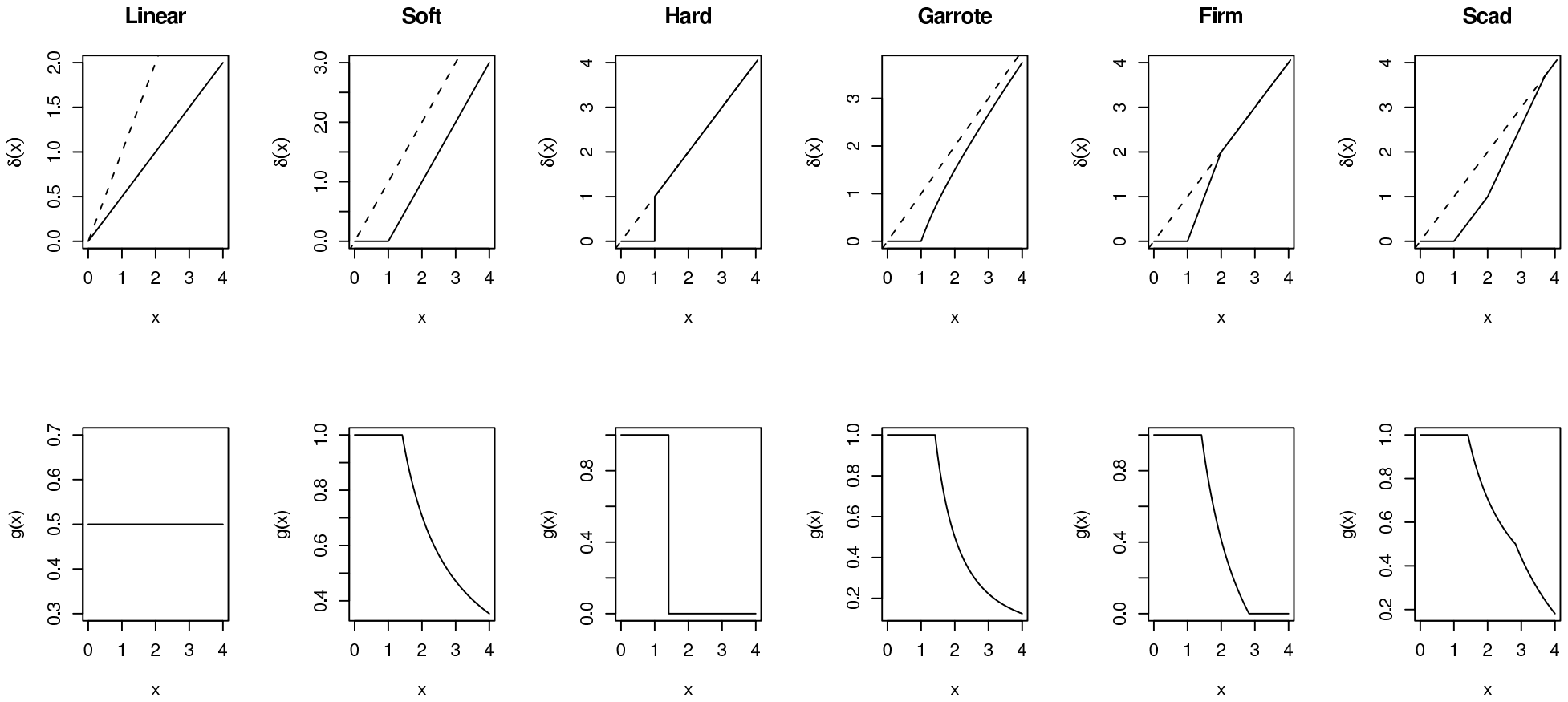
$$g(|x|) = \begin{cases} 1 & \text{if } |x| \leq \sqrt{2}\lambda_1 \\ \frac{\lambda_1}{(\lambda_2 - \lambda_1)}\left(\frac{\sqrt{2}\lambda_2}{|x|} - 1\right) & \text{if } \sqrt{2}\lambda_1 < |x| \leq \sqrt{2}\lambda_2 \\ 0 & \text{if } |x| > \sqrt{2}\lambda_2 \end{cases}.$$

SCAD shrinkage SCAD shrinkage gives also a diffusivity that

degenerates to 0:

$$g(|x|) = \begin{cases} 1 & \text{if } |x| \leq \sqrt{2}\lambda \\ \frac{\sqrt{2}\lambda}{|x|} & \text{if } \sqrt{2}\lambda < |x| \leq 2\sqrt{2}\lambda \\ \frac{a\sqrt{2}\lambda}{(a-2)|x|} - \frac{1}{a-2} & \text{if } 2\sqrt{2}\lambda < |x| \leq a\sqrt{2}\lambda \\ 0 & \text{if } |x| > a\sqrt{2}\lambda \end{cases}.$$

Examples . . .



Shrinkage functions (top) and corresponding diffusivities (bottom). Plotted for $\lambda = 1$, $\lambda_1 = 1$, $\lambda_2 = 2$ (Firm) and $a = 3.7$ (Scad). The dashed line is the diagonal.

From diffusion to skrinkage

The other way round one can ask is how the shrinkage functions for famous diffusivities look like.

The function δ_λ expressed in terms of g looks like

$$\delta_\lambda(|x|) = |x|(1 - g(\sqrt{2}|x|)$$

and the dependence of the shrinkage function on the threshold parameter λ is naturally fulfilled because usually diffusivities involve a parameter too.

Such a remark leads to new shrinkage functions.

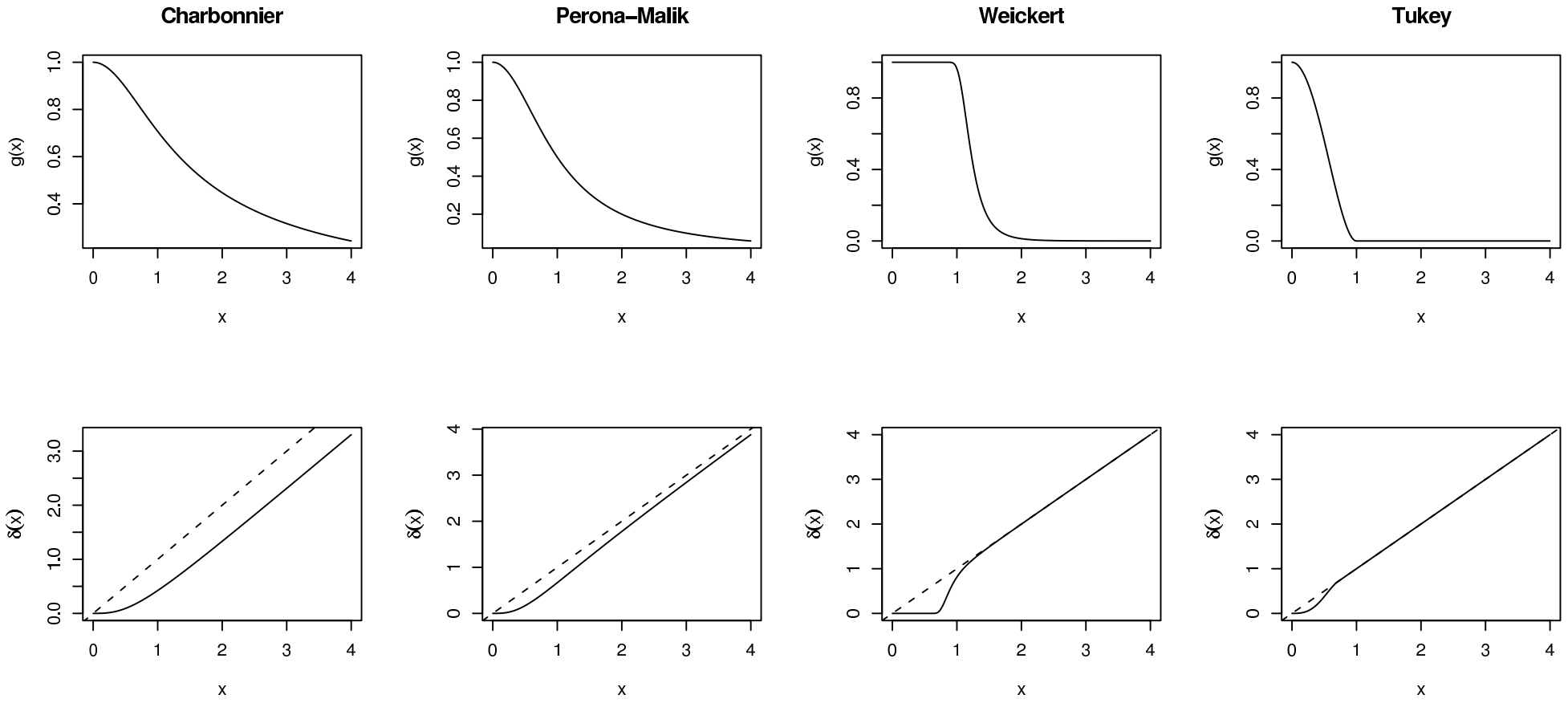
New shrinkage rules

Charbonnier diffusivity The Charbonnier diffusivity (Charbonnier *et al.* (1994)) is given by $g(|x|) = \left(1 + \frac{x^2}{\lambda^2}\right)^{-1/2}$ and corresponds to the shrinkage function $\delta_\lambda(x) = x\left(1 - \sqrt{\frac{\lambda^2}{\lambda^2+2x^2}}\right)$.

Perona-Malik diffusivity The Perona-Malik diffusivity (Perona and Malik (1990)) is defined by $g(|x|) = \left(1 + \frac{x^2}{\lambda^2}\right)^{-1}$ and lead to the shrinkage function $\delta_\lambda(x) = \frac{2x^3}{2x^2+\lambda^2}$.

Weickert diffusivity Weickert (1998) introduced the following diffusivity $g(|x|) = I_{\{|x|>0}(x)\left(1 - \exp\left(-\frac{3.31488}{(|x|/\lambda)^8}\right)\right)$ which leads to the shrinkage function $\delta_\lambda(x) = x\exp\left(-\frac{0.20718\lambda^8}{x^8}\right)$.

Classical diffusivities



"Classical" diffusivites (top) and corresponding shrinkage functions

Motivation for shrinkage

We have developed the *connection* between *diffusivities and shrinkage functions*.

It is well known, the *shrinkage* methods *perform very well* (asymptotic optimality, shown by Johnstone and Donoho.)

But *why* do they work so well? Is there a mathematical motivation for shrinkage methods?

They can all be interpreted as cases of a broad class of *penalized least squares* estimators.

This unified treatment and the general results of AF on penalized wavelet estimators allow a *systematic derivation of oracle inequalities* and minimax properties for a large class *of wavelet estimators*.

Penalized least-squares wavelet estimators

Traditional regularization problem can be formulated in the wavelet domain by finding the minimum in $\boldsymbol{\theta}$ of

$$\ell(\boldsymbol{\theta}) = \|W\mathbf{y} - \boldsymbol{\theta}\|_n^2 + 2\lambda \sum_{i>i_0} p(|\theta_i|),$$

where $\boldsymbol{\theta}$ is the vector of the wavelet coefficients of the unknown regression function g, p is a given penalty function, while i_0 is a given integer corresponding to penalizing wavelet coefficients above certain resolution level j_0.

Here to facilitate the presentation we changed the notation $d_{j,k}$ from a double array sequence into a single array sequence θ_i.

We also use we will use p_λ to denote the penatly function λp in the following.

Separable penalized least-squares

With a choice of an additive penalty $\sum_{i>i_0} p(|\theta_i|)$, the minimization problem becomes separable, i.e. it is equivalent to minimize

$$\ell(\theta_i) = (z_i - \theta_i)^2 + 2\lambda p(|\theta_i|),$$

for each coordinate i larger than i_0. Therefore the estimate of any coordinate θ_i depends solely on the empirical wavelet coefficient z_i.

The performance of the resulting wavelet estimator depends on the penalty and the regularization parameter λ.

Conditions on p

Usually, p is chosen to be symmetric and increasing on $[0, +\infty)$.

AF provide some insights into how to choose a penalty function. A good penalty function should result in

unbiasedness (no over-penalization of large coefficients to avoid unnecessary modeling biases)

sparsity (insignificant coefficients should be set to zero to reduce model complexity)

stability (continuity of the penalty to avoid instability and large variability in model prediction).

We will now show how to derive the penalties corresponding to the thresholding rules defined previously, and check that almost all of them satisfy these conditions.

Shrinkage functions and penalties

Let $\delta_\lambda : \mathbb{R} \to \mathbb{R}$ be a thresholding function that is increasing antisymmetric such that $0 \leq \delta_\lambda(x) \leq x$ for $x \geq 0$ and $\delta_\lambda(x) \to \infty$ as $x \to \infty$.

There exist a continuous positive penalty function p_λ, with $p_\lambda(x) \leq p_\lambda(y)$ whenever $|x| \leq |y|$, such that $\delta_\lambda(z)$ is the unique solution of the minimization problem $\min_\theta (z-\theta)^2 + 2p_\lambda(|\theta|)$ for every z at which δ_λ is continuous.

From the proof of this result one gets an almost analytical expression for p_λ. Denoting by r_λ the generalized inverse function of δ_λ defined by $r_\lambda(x) = \sup\{z | \delta_\lambda(z) \leq x\}$, one gets that, for any $z > 0$, p_λ is defined by

$$p_\lambda(z) = \int_0^z (r_\lambda(u) - u) du.$$

Penalties and thresholding

We find, in particular, the well known ridge regression L_2-penalty

$$p_\lambda(|\theta|) = \lambda|\theta|^2$$

corresponding to the linear shrinkage function, the L_1-penalty

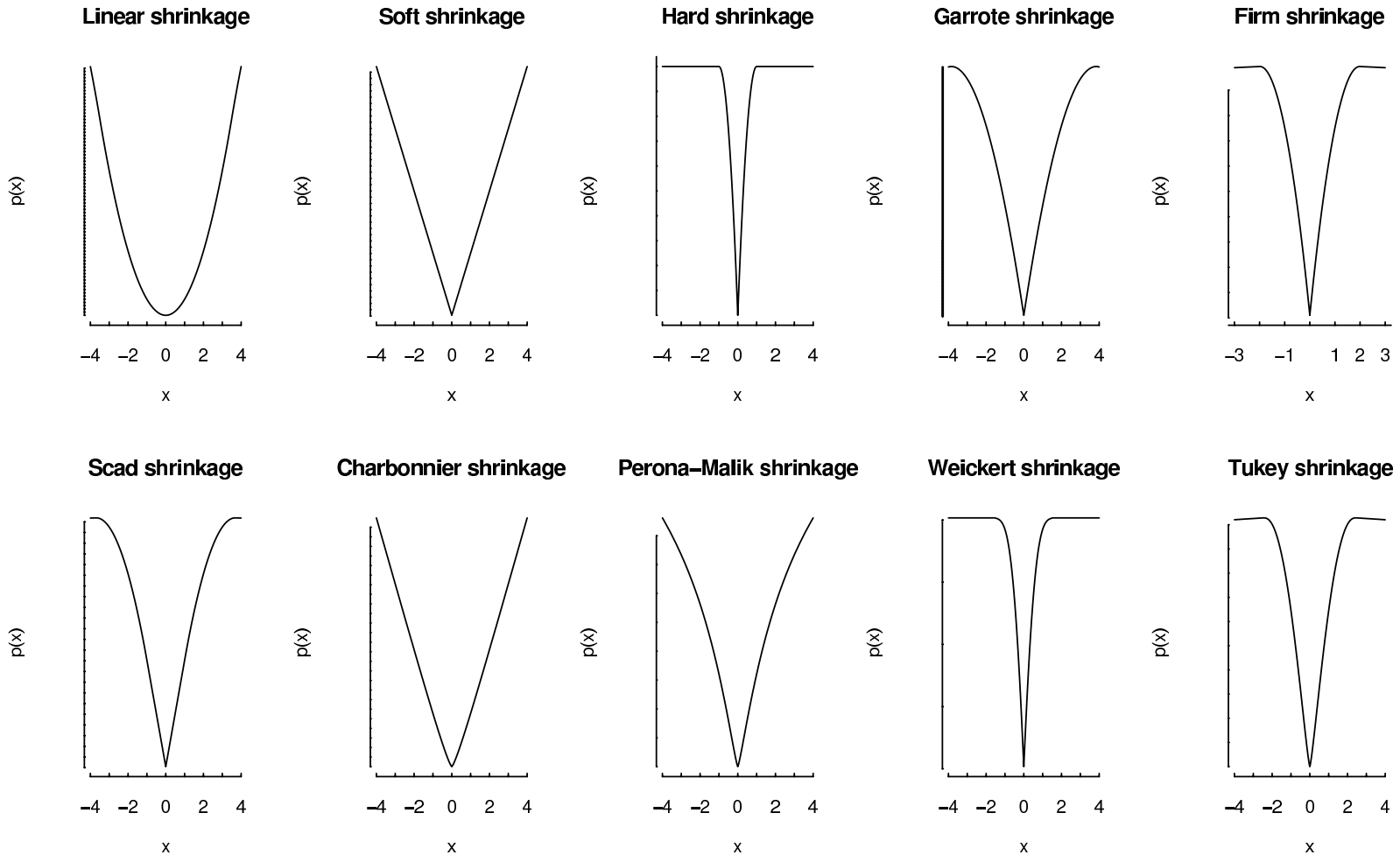
$$p_\lambda(|\theta|) = \lambda|\theta|$$

corresponding to the soft thresholding rule and the hard thresholding penalty function

$$p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 I_{\{|\theta|<\lambda\}}(|\theta|)$$

that results in the hard-thresholding rule.

Penalties



Penalties corresponding to the shrinkage and thresholding functions with the same name

Remarks

The *quadratic penalty*, while continuous is not singular at zero, and the resulting estimator is not thresholded. All other penalties are singular at zero, thus resulting in thresholding rules that enforce sparseness of the solution.

The *hard-thresholding* penalty is not continuous at the threshold, so it may induce the oscillation of the reconstructed signal (lack of stability).

For *soft-thresholding*, the resulting estimator of large coefficients is shifted by an amount of λ (unnecessary bias when the coefficients are large). The same for Charbonnier and Perona-Malick penalties.

All other penalties are singular at zero (encourage sparse solutions), continuous (stable) and do not create excessive bias when the wavelet coefficients are large.

Properties

Most importantly all these other penalties satisfy the conditions of Theorem 1 in AF (2001).

The implication of this fact is that it leads to a systematic derivation of oracle inequalities and minimax properties for the resulting wavelet estimators via Theorem 2 of AF.

In particular, the optimal hard and soft universal threshold $\lambda = \sigma\sqrt{2\log_2 n}$ given by Donoho and Johnstone (1994) leads to a sharp asymptotic risk upper bound and the resulting penalized estimators are adaptively minimax within a factor of logarithmic order over a wide range of Besov spaces.

And the nonequispaced case?

A first possible approach: Assume that $t_i = n_i/2^J$ for some n_i and some resolution J.

Parameter: Let $\mathbf{f}$ be the underlying regression function collected at all dyadic points $\{i/2^J,\ i = 1, \ldots, 2^J\}$.

Apply the Wavelet Transform on $\mathbf{f}$: $\boldsymbol{\theta} = W\mathbf{f}$ and $\mathbf{f} = W^T\boldsymbol{\theta}$, to get an

Overparametrized linear model:

$$\mathbf{Y}_n = A\boldsymbol{\theta} + \varepsilon.$$

The Wavelet basis on which $\mathbf{f}$ is projected is chosen by fixing the resolution J and is truncated by retaining the rows of A. The estimate of $\boldsymbol{\theta}$ and therefore of $\mathbf{f}$ is recovered by penalized least-squares

$$2^{-1}\|\mathbf{Y}_n - A\boldsymbol{\theta}\|^2 + \sum_{i \in I_N} p_\lambda(|\theta_i|)$$

The penalty function p_λ is nonconvex and irregular at point zero.

Computation Challenge:

– Irregular designs: The matrix A is no longer orthonormal. This is a linear regression problem with a number p of unknown parameters much larger than the number n of observations.

A linear regression model

We therefore switch to the problem of obtaining a reasonable estimate for an unknown vector of parameters $\boldsymbol{\beta}$ given a vector $\mathbf{Y}$ of measurements

$$\underbrace{\mathbf{Y}}_{n\times 1} = \underbrace{X}_{n\times p}\underbrace{\boldsymbol{\beta}}_{p\times 1} + \underbrace{\boldsymbol{\epsilon}}_{n\times 1},$$

where X is a known predictor matrix and $\boldsymbol{\epsilon}$ is a (Gaussian) noise error with some variance $\sigma^2\mathbf{I}_n$.

Typically the number p of unknown parameters is much larger than the number n of observations.

Shrinkage, thresholding and complexity

Regularize the solution by minimizing a penalized loss function:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\|\mathbf{Y}-X\boldsymbol{\beta}\|^2+\lambda T(\boldsymbol{\beta})\right\}\Leftrightarrow\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\|\mathbf{Y}-X\boldsymbol{\beta}\|^2\right\}\text{ with } T(\boldsymbol{\beta})\leq t.$$

This is *penalized or constrained least-squares*. The penalty term is usually chosen to encourage sparsity in the optimal $\boldsymbol{\beta}$ while the regularization parameter λ (or t) is connected to the complexity of the model that is fitted. Often need to solve for multiple values of λ e.g. to adjust sparsity to some desired level or perform cross-validation.

Least absolute shrinkage and selection operator

LASSO (Tibshirani,1996 ; Chen, Donoho & Saunders, 1999 (Basis pursuit) ; Donoho *et al.*, 2002 - 2004)

For appropriate values of λ (or t or ϵ) solve the following equivalent optimisation problems:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \|\mathbf{Y} - X\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}$$

$$\Leftrightarrow$$

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \|\mathbf{Y} - X\boldsymbol{\beta}\|^2 \right\} \text{ with } \|\boldsymbol{\beta}\|_1 \leq t$$

$$\Leftrightarrow$$

$$\min \|\boldsymbol{\beta}\|_1 \text{ with } \|\mathbf{Y} - X\boldsymbol{\beta}\|^2 \leq \epsilon.$$

Lasso and thresholding

We already have seen the simple model case with $n = p$ and X is orthonormal: $X^T X = I_p$ (wavelet denoising case). In this case, the LASSO selector is given by the soft thresholding formula

$$\hat{\beta}_j^{soft} = \begin{cases} Z_j - \lambda & \text{si} \quad Z_j \geq \lambda, \\ 0 & \text{if} \quad Z_j < \lambda, \\ Z_j + \lambda & \text{if} \quad Z_j \geq -\lambda, \end{cases} \qquad \text{with } Z_j = (X^T \mathbf{Y})_j.$$

The MSE for this selector is roughly $\lambda^2 + \sum_{i=1}^{p} \min(|Z_j|^2, \lambda^2)$, and this is basically the best possible amongst all selectors in this model.

MM algorithm for optimization

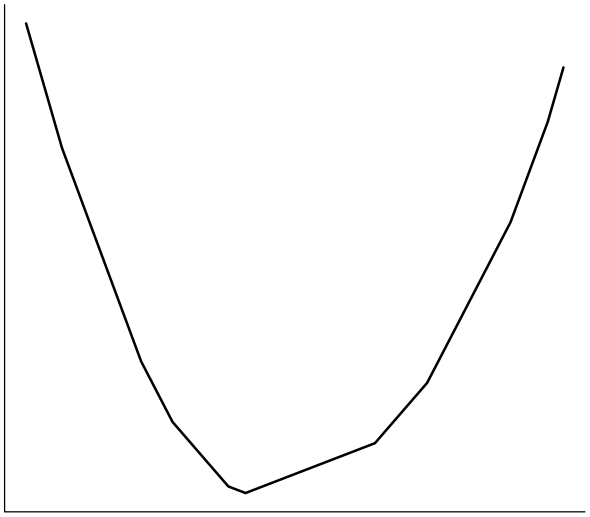
We have seen that optimizing the penalized loss function:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \|\mathbf{Y} - X\boldsymbol{\beta}\|^2 + \lambda T(\boldsymbol{\beta}) \right\}$$

with $T(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$ leads to the LASSO selector which can be easily calculated by soft thresholding when X is orthogonal as in wavelets denoising.

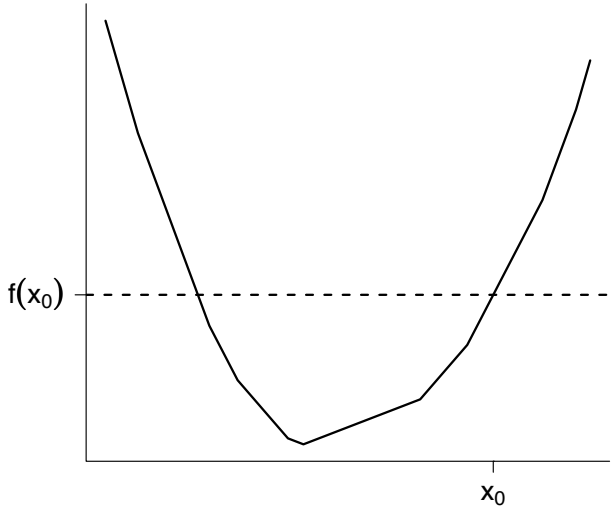
If we concentrate on orthogonal design matrices, ℓ_1 penalty is far from the only choice and as seen before we have several other penalties that lead to good denoising procedures. To retain the separability of the optimization problem when the penalty is separable (univariate minimization) and easy optimization via thresholding and shrinkage in the general case we are going to use an MM algorithm.

A short tutorial on the class of MM algorithms



Goal: Solve difficult minimization problem, like minimizing the function shown here in black.

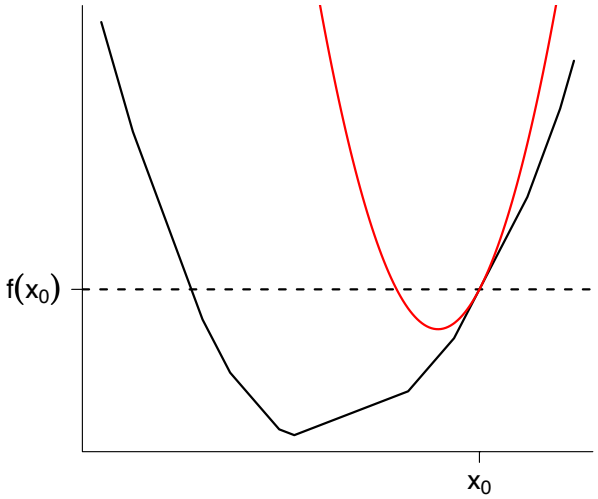
What is an MM algorithm?



- Choose a starting point x_0

Goal: Solve difficult minimization problem, like minimizing the function shown here in black.

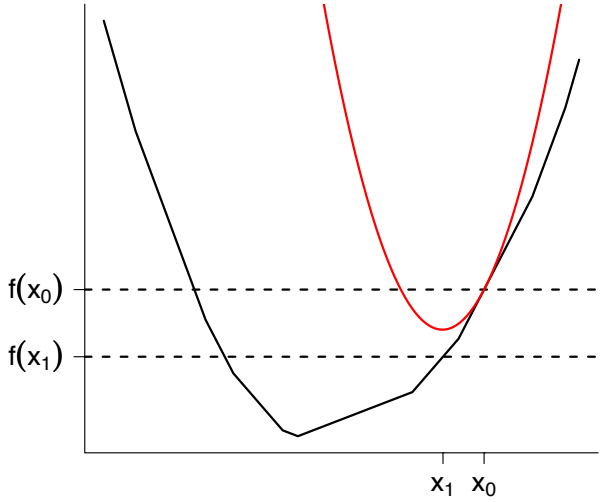
What is an MM algorithm?



- Choose a starting point x_0
- Construct a majorizing function of $f(x)$ at x_0.

Goal: Solve difficult minimization problem, like minimizing the function shown here in black.

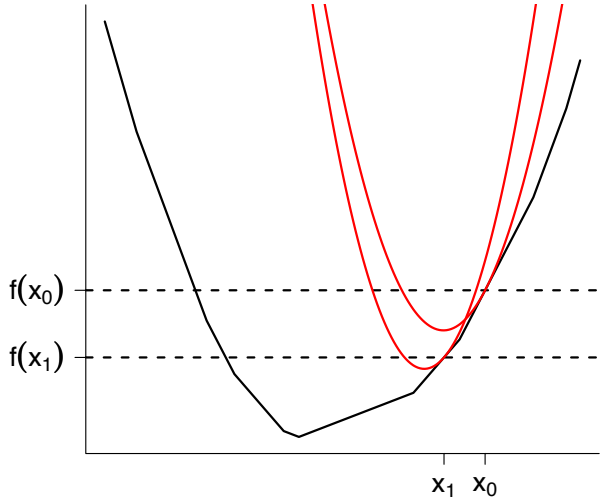
What is an MM algorithm?



- Choose a starting point x_0
- Construct a majorizing function of $f(x)$ at x_0.
- Minimize the majorizer (at x_1).

Goal: Solve difficult minimization problem, like minimizing the function shown here in black.

What is an MM algorithm?



- Choose a starting point x_0
- Construct a majorizing function of $f(x)$ at x_0.
- Minimize the majorizer (at x_1).
- Repeat.

Goal: Solve difficult minimization problem, like minimizing the function shown here in black.

So "MM" stands for "Majorize-Minimize".

Numerical analysis

MM is merely a new name for an old technique. The idea for these algorithms dates back at least as far as Ortega and Rheinboldt (1970). Statisticians have been applying it to various problems for about 30 years.

Multidimensional scaling (de Leeuw and Heiser; Groenen)

Robust regression (Schlossmacher; Huber)

Least squares estimation (Bijleveld and de Leeuw; Kiers and Ten Berge)

Quadratic lower bound principle (Böhning and Lindsay)

Medical imaging (Lange and Fessler; De Pierro)

There are also some surveys of the general method.

Numerical analysis

Kenneth Lange and Draper (2000) have used the term "optimization transfer" for a while but ultimately settled on "MM", which works for both minimization and maximization.

A successful MM algorithm substitutes a simple optimization problem for a difficult optimization problem.

Iteration is the price to pay for simplifying the original problem.

Optimization by an MM algorithm

Let $\boldsymbol{\theta}^{(m)}$ represent a fixed value of the parameter $\boldsymbol{\theta}$, and let $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)})$ denote a real-valued function of $\boldsymbol{\theta}$ whose form depends on $\boldsymbol{\theta}^{(m)}$. The function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)})$ is said to majorize a real-valued function $S(\boldsymbol{\theta})$ at the point $\boldsymbol{\theta}^{(m)}$ provided that

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}) \geq S(\boldsymbol{\theta}), \text{ for all } \boldsymbol{\theta} \quad (1)$$

$$Q(\boldsymbol{\theta}^{(m)}|\boldsymbol{\theta}^{(m)}) \geq S(\boldsymbol{\theta}^{(m)}). \quad (2)$$

The surface $\boldsymbol{\theta} \to Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)})$ lies above the surface $S(\boldsymbol{\theta})$ and is tangent to it at the point $\boldsymbol{\theta} = \boldsymbol{\theta}^{(m)}$.

Ordinarily, $\boldsymbol{\theta}^{(m)}$ represents the current iterate in a search of the minimum of the surface $S(\boldsymbol{\theta})$.

In a majorize-minimize MM algorithm, one minimizes the majorizing function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)})$ rather than the actual function $S(\boldsymbol{\theta})$.

$$\text{MM algorithm} \Leftrightarrow \boldsymbol{\theta}^{(m+1)} = \text{argmin}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)})$$

Monotonicity

If $\boldsymbol{\theta}^{(m+1)})$ is a minimizer of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)})$ then the MM algorithm forces the actual function $S(\boldsymbol{\theta})$ downhill. Indeed, the inequality

$$\begin{aligned} S(\boldsymbol{\theta}^{(m+1)}) &= Q(\boldsymbol{\theta}^{(m+1)}|\boldsymbol{\theta}^{(m)}) + S(\boldsymbol{\theta}^{(m+1)}) - Q(\boldsymbol{\theta}^{(m+1)}|\boldsymbol{\theta}^{(m)}) \\ &\leq Q(\boldsymbol{\theta}^{(m)}|\boldsymbol{\theta}^{(m)}) + S(\boldsymbol{\theta}^{(m)}) - Q(\boldsymbol{\theta}^{(m)}|\boldsymbol{\theta}^{(m)}) \\ &= S(\boldsymbol{\theta}^{(m)}). \end{aligned}$$

follows directly from the fact $Q(\boldsymbol{\theta}^{(m+1)}|\boldsymbol{\theta}^{(m)}) \leq Q(\boldsymbol{\theta}^{(m)}|\boldsymbol{\theta}^{(m)})$ and definitions (1) and (2).

Return to penalized least squares

Recall that we want to minimize the penalized loss function $R_\lambda(\boldsymbol{\beta})$:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\mathbf{Y} - X\boldsymbol{\beta}\|_2^2 + \lambda T(\boldsymbol{\beta}) \right\}$$

with $T(\boldsymbol{\beta})$ one of the separable penalties that are assoiated to "nice" thresholding functions.

Denote by $S_\lambda(\boldsymbol{\beta})$ the above penalized loss function and pick a constant $c > 0$ such that $\lambda_{max}(X^T X) \leq c$. It follows that $cI_p - X^T X$ is strictly positive definite. Since X can be rescaled assume that $c = 1$. Define

$$\Xi(\boldsymbol{\beta}|\boldsymbol{\gamma}) = \frac{1}{2}\|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2^2 - \frac{1}{2}\|X(\boldsymbol{\beta} - \boldsymbol{\gamma})\|_2^2, \tag{3}$$

which depends on an auxiliary p-dimensional vector $\boldsymbol{\gamma}$.

Constructing a majorizing function

Since $I_p - X^T X$ is strictly positive definite, the functional Ξ defined in (3) is strictly convex in $\boldsymbol{\beta}$ for any choice of $\boldsymbol{\gamma}$. Therefore, adding $\Xi(\boldsymbol{\beta}|\boldsymbol{\gamma})$ to $S_\lambda(\boldsymbol{\beta})$ creates a majorizing function for $S_\lambda(\boldsymbol{\beta})$:

$$\begin{aligned} S_\lambda^{sur}(\boldsymbol{\beta}|\boldsymbol{\gamma}) &= \frac{1}{2}\|\mathbf{Y} - X\boldsymbol{\beta}\|_2^2 + \lambda T(\boldsymbol{\beta}) + \Xi(\boldsymbol{\beta}|\boldsymbol{\gamma}) \\ &= \frac{1}{2}\|\mathbf{Y} - X\boldsymbol{\beta}\|_2^2 + \lambda T(\boldsymbol{\beta}) + \frac{1}{2}\langle (I_p - \Sigma)(\boldsymbol{\beta} - \boldsymbol{\gamma}), (\boldsymbol{\beta} - \boldsymbol{\gamma})\rangle \end{aligned}$$

where $\langle \mathbf{x}, \mathbf{w}\rangle = \mathbf{x}^T\mathbf{w}$ and $\Sigma = X^T X$.

Apply the MM methodology

Approach the minimizer of $S_\lambda(\boldsymbol{\beta})$ by the following iterative process:
Starting from an arbitrary chosen $\boldsymbol{\beta}^{(0)}$, determine the minimizer of $S_\lambda^{sur}(\boldsymbol{\beta}|\boldsymbol{\gamma})$ for $\boldsymbol{\gamma} = \boldsymbol{\beta}^{(0)}$; each successive iterate $\boldsymbol{\beta}^{(n)}$ is then the minimizer of the surrogate functional $S_\lambda^{sur}(\boldsymbol{\beta}|\boldsymbol{\gamma})$ anchored at the previous iterate, i.e. $\boldsymbol{\gamma} = \boldsymbol{\beta}^{(n-1)}$.

The iterative algorithm goes as follows:

$$\boldsymbol{\beta}^{(0)} \text{ arbitrary } ; \boldsymbol{\beta}^{(m)} = \text{argmin}_{\boldsymbol{\beta}} S_\lambda^{sur}(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m-1)})$$

Under reasonable conditions on X and for most of the "nice" penalties $T(\boldsymbol{\beta})$ reviewed before the algorithm converges.

Calculus with particular penalties

Suppose first that $\lambda = 0$ in $S_\lambda(\boldsymbol{\beta})$ (no penalization). Then

$$S_0^{sur}(\boldsymbol{\beta}|\boldsymbol{\gamma}) = \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 - \langle \boldsymbol{\beta}, (I-\Sigma)\boldsymbol{\gamma} + X^T\mathbf{y}\rangle + \frac{1}{2}\|\mathbf{y}\|_2^2 + \frac{1}{2}\|\boldsymbol{\gamma}\|_2^2 - \frac{1}{2}\|X\boldsymbol{\gamma}\|_2^2$$

Given the actual anchor $\boldsymbol{\gamma}$, minimizing the above expression with respect to $\boldsymbol{\beta}$ is equivalent in minimizing the following expression $\frac{1}{2}\|\boldsymbol{\beta} - \left[(I-\Sigma)\boldsymbol{\gamma} + X^T\mathbf{y}\right]\|_2^2$, which leads, using $\boldsymbol{\gamma} = \boldsymbol{\beta}^{(n)}$, to the solution

$$\boldsymbol{\beta}^{(n+1)} = \boldsymbol{\beta}^{(n)} + X^T(\mathbf{y} - X\boldsymbol{\beta}^{(n)}),$$

known as the *Landweber* iterative method.

Ridge regression (Tikhonov regularization)

Suppose first that $T(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2$ in $S_\lambda(\boldsymbol{\beta})$. Then, following a calculus similar to that of the previous slide leads to the following iterative procedure for finding the minimum:

$$\boldsymbol{\beta}^{(n+1)} = \frac{1}{\lambda + 1}\left[\boldsymbol{\beta}^{(n)} + X^T(\mathbf{y} - X\boldsymbol{\beta}^{(n)})\right],$$

known as a dumped *Landweber* iterative method.

In both cases, and with reasonable definition on X, the sequence $\boldsymbol{\beta}^{(n)}$ converges to a generalized solution of the function to be minimized.

Iterative shrinkage thresholding

Using same arguments but for a penalty $T(\boldsymbol{\beta})$ associated to a particular thresholding function δ_λ, minimizing the functional $S_\lambda^{sur}(\boldsymbol{\beta}|\boldsymbol{\gamma})$ with anchor $\boldsymbol{\gamma}$ is equivalent in minimizing the expression

$$\frac{1}{2}\|\boldsymbol{\beta} - \left[(I-\Sigma)\boldsymbol{\gamma} + X^T\mathbf{y}\right]\|_2^2 + \lambda T(\boldsymbol{\beta})$$

and leads to

$$\boldsymbol{\beta}^{(n+1)} = \delta_\lambda\left[\boldsymbol{\beta}^{(n)} + X^T(\mathbf{y} - X\boldsymbol{\beta}^{(n)})\right], \tag{4}$$

known to belong to the class of iterative thresholding algorithms (when $T(\boldsymbol{\beta})$ is an ℓ_p penalty, $0 < p < \infty$). These usually converge see e.g. Daubechies, Defrise, De Mol (2004), Combettes & Wajs (2005) and Bredies, Lorenz & Maass (2005).

For particular inverse problems such algorithms have been studied in the recent literature by many authors, especially when considering sparse regularization and compressed sensing. For convex penalties $T(\boldsymbol{\beta})$,

- IST as expectation-maximization (Figuereido and Nowak, 2001, 2003)
- IST as majorization-minimization (De Mol, Defrise, 2002; Daubechies, Defrise, De Mol, 2004; Figuereido, Nowak, Bioucas-Dias, 2005, 2007)

Other authors independently proposed IST-like schemes for signal/image recovery: Starck, Nguyen and Murtagh (2003); Starck, Candès and Donoho (2003); Bect, Blanc-Féraud, Aubert, and Chambolle (2004); Tropp, Donoho and others (2005); Candes (2006); Elad, Matalon and Zibulevsky (2006); Hale, Yin and Zhang (2007), ...

Summary

Consider a thresholding function $\delta_\lambda(\cdot)$ satisfying:
a) $\delta_\lambda(\cdot)$ is an odd function,
b) $\delta_\lambda(\cdot)$ is a shrinkage rule ($0 \leq \delta_\lambda^+(t) \leq t,\ \forall t \geq 0$)
c) δ_λ^+ is not decreasing and coercive.
Most often δ_λ thresholds, i.e. $\delta_\lambda^+(t) = 0$ for $0 \leq t \leq \tau$ for some $\tau \geq 0$.

Then (Antoniadis, 2007) a penalty can be defined with the following 3-step procedure:
1. Define for $u \geq 0$, $\delta_\lambda^{-1}(u) = \sup\{t; \delta_\lambda(t) \leq u\}$ and $\delta_\lambda^{-1}(-u) = -\delta_\lambda^{-1}(u)$.
2. Set $r_\lambda(u) = \delta_\lambda^{-1}(u) - u, \forall u$
3. Put $P_\lambda(\theta) = \int_0^{|\theta|} r_\lambda(u)du$.

Summary

Then (Antoniadis, 2007) the minimization problem

$$\min_{\theta}(t-\theta)^2/2 + P_\lambda(\theta)$$

has a unique optimal solution $\hat{\theta} = \delta_\lambda(t)$ for any t at which $\delta_\lambda(\cdot)$ is continuous.

And one therefore may come back to the original minimization problem using iterative thresholding procedures with thresholding functions such as δ_λ.

For example, when using soft-thresholding one obtains the iterative thresholding algorithm of DDD (2004). If δ_λ is the hard-thresholding then one uses the ℓ_0 penalty and an algorithm by Tropp or Elad, ...

Convergence

If $p < n$ and Σ is not singular, the iterative Landweber mapping is a contraction and the sequence of iterates $\boldsymbol{\beta}^{(n)}$ converges to a stationary point of the function we want to minimize.

But what about the case $p > n$ and Σ singular? DDD (2004) have shown that for soft thresholding the algorithm converges and this is mainly due to the fact that the iterative Landweber iterations operator is *nonexpansive*, i.e. $\|Tx - Ty\| \leq \|x - y\|$. However, most of the thresholding rules that one may consider are usually not nonexpansive and one needs then some appropriate conditions on the design matrix X and the sparsity of $\boldsymbol{\beta}$ (see e.g. Candès and Tao (2007), Foucart (2008), ...).

The bounded curvature condition (BCC)

We will say that a penalty $P_\lambda(\boldsymbol{\beta})$ satisfies the BCC for some positive semi-definite matrix $\mathbf{B}$, if for any $\boldsymbol{\eta} \in \mathbb{R}^p$ one has:

$$P_\lambda(\boldsymbol{\beta} + \boldsymbol{\eta}) \geq P_\lambda(\boldsymbol{\beta}) + \langle \boldsymbol{\eta}, \mathbf{r}_\lambda \rangle - \frac{1}{2}\boldsymbol{\eta}^T \mathbf{B} \boldsymbol{\eta},$$

where $\mathbf{r}_\lambda = r_\lambda(\boldsymbol{\beta})$ is computed component-wize.

Many thresholding rules of practical interest satisfy the BCC with some $\mathbf{B}$. For example soft thresholding with $\mathbf{B} = 0$ because $\|\boldsymbol{\beta}\|_1$ is convex; hard thresholding with $\mathbf{B} = \mathbf{I}_p$; SCAD thresholding with $\mathbf{B} = \mathbf{I}_p/(a-1)$, ...

Convergence with BCC

Given the iterations (4), if $\lambda_{\max}(\Sigma) \leq \max(1, 2 - \lambda_{\max}(\mathbf{B}))$, then

$$R_\lambda(\boldsymbol{\beta}^{(n)}) \geq R_\lambda(\boldsymbol{\beta}^{(n+1)}). \tag{5}$$

Moreover, if $\lambda_{\max}(\Sigma) < \max(1, 2 - \lambda_{\max}(\mathbf{B}))$, there exists a constant $C > 0$ (depending only on X and $\mathbf{B}$) such that

$$R_\lambda(\boldsymbol{\beta}^{(n)}) - R_\lambda(\boldsymbol{\beta}^{(n+1)}) \geq C \cdot \|\boldsymbol{\beta}^{(n)} - \boldsymbol{\beta}^{(n+1)}\|_2^2. \tag{6}$$

We can therefore use the iterative thresholding in the following form

$$\boldsymbol{\beta}^{(n+1)} = \delta_{\lambda/k_0^2}\left[\boldsymbol{\beta}^{(n)} + \frac{1}{k_0^2} X^T(\mathbf{y} - X\boldsymbol{\beta}^{(n)})\right]$$

wher $k_0 = \lambda_{\max}(X) = \|X\|_2$.

Some special cases

Suppose that for the iterations (4), one uses:

- Soft thresholding. If $\lambda_{\max}(X) < \sqrt{2}$ then (6) holds.
- Hard thresholding. If $\lambda_{\max}(X) \leq 1$ then (5) holds and if $\lambda_{\max}(X) < 1$ it is (6) that holds.
- SCAD thresholding. If $\lambda_{\max}(X) < \sqrt{2 - \frac{1}{a-1}}$ then (6) holds.

So given any initial point for $\boldsymbol{\beta}$, if one of these conditions hold then the algorithm converges to a fixed point of (4).

Optimum

Let $\boldsymbol{\beta}^\star$ a fixed point of (4) and suppose that $\lambda_{\max}(\mathbf{B}) \leq 1$.

If

$$\lambda_{\max}(\mathbf{B}) \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq 2 - \lambda_{\max}(\mathbf{B}),$$

then $\boldsymbol{\beta}^\star$ is a global minimizer of $R_\lambda(\boldsymbol{\beta})$.

Although this was known for nonconvex penalties in the orthogonal case, the same conclusion holds as long as X is not too far from orthogonality (characterized in terms of $\mathbf{B}$). This is related to the RIP condition used in sparse learning (see Candès and Tao (2007) and Foucart (2008)) and very closely related and inspired by the Restricted Eigenvalue Property (Donoho, Elad and Temlyakov (2006) and Bickel, Ritov and Tsybakov (2007).

Estimation and risk

Assuming that the errors are Gaussian, that $p_n = O(n^{\xi})$, $n \to \infty$, for some $1 < \xi$ and that the number of $\boldsymbol{\beta}_{0j,n} \neq 0$ is independent of n and finite (S-sparsity). Then, under the assumptions that all entries of the design matrix are uniformly bounded and that the thresholding function used is sandwiched between the soft and the hard thresholding (see AF), then the estimation of $\boldsymbol{\beta}_0$ achieved using (4) is sparsistent and as long as the S-sparsity remains bounded, it leads to an optimal , up to a $\log p$ factor, squared error bound. The proof relies upon similar results by Bunea, Tsybakov and Wegkcamp (2007).

Iterative shrinkage thresholding for Generalized Linear Models

Consider now independent observations $Y_1, \cdots, Y_n$ where Y_i follows a distribution in the natural exponential family $f(y_i; \theta_i) = \exp(y_i\theta_i - b(\theta_i) + c(y_i))$, where θ_i is the natural parameter.

Let $L_i = \log f(y_i, \theta_i)$, $L = \sum L_i$. Clearly, $L_i = y_i\theta_i - b(\theta_i) + c(y_i)$, and thus $\partial L_i/\partial\theta_i = y_i - b'(\theta_i)$, $\partial^2 L_i/\partial\theta_i^2 = -b''(\theta_i)$.

It is well known that $E(\partial L_i/\partial\theta_i) = 0$ and $E(\partial L_i/\partial\theta_i)^2 = -E(\partial^2 L_i/\partial\theta_i^2)$ hold in general for the exponential family.

Therefore, $\mu_i \triangleq E(y_i) = b'(\theta_i)$, $var(y_i) = b''(\theta_i)$.

Generalized Linear Models

Let $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]^T$ be the model matrix.

We will use the *canonical link* function that is, the link function $\boldsymbol{x}_i^T \boldsymbol{\beta} = g(\mu_i)$ determined by $g(\mu_i) = \theta_i$. Obviously $g = (b')^{-1}$.

For instance, when $Y_i \sim \text{Bernoulli}(\pi_i)$, $f(y_i; \theta_i) = \exp\left\{y_i \log \frac{\pi_i}{1-\pi_i} + \log(1-\pi_i)\right\} = \exp\left\{y_i\theta_i - \log(1+e^{\theta_i})\right\}$, for which $\theta_i = \log \frac{\pi_i}{1-\pi_i}$, $\mu_i = \pi_i$, $b(t) = \log(1+e^t)$, and $g(t) = \log \frac{t}{1-t}$ (the logit link).

In the Poisson case where $y_i \sim \text{Poi}(\omega_i)$, $f(y_i; \theta_i) = \frac{1}{y_i!} e^{-\omega_i} \omega_i^{y_i} = \exp(y_i \log \omega_i - \omega_i - \log y_i!) = \exp(y_i\theta_i - e^{\theta_i} + c(y_i))$ with $\theta_i = \log \omega_i$, $\mu_i = \omega_i$, $b(t) = e^t$, and $g(t) = \log t$ (the log link).

A surrogate function

We consider the penalized GLM problem

$$\min_{\boldsymbol{\beta}} -L(\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta}) (\triangleq F(\boldsymbol{\beta})), \tag{7}$$

where

$$L = \sum_{i=1}^{n} L_i,$$

and

$$P_\lambda(\boldsymbol{\beta}) = \sum_{i=1}^{p} P_\lambda(\beta_i)$$

is a (separable) penalty with λ as the regularization parameter.

We assume again that $\boldsymbol{\beta}$ is *sparse*, and use (7) for predictive learning.

Optimization

Directly tackling (7) for a general penalty can be a difficult task. Use instead

$$\begin{aligned} G(\boldsymbol{\beta}, \boldsymbol{\gamma}) &= -\sum_{i=1}^{n} L_i(\boldsymbol{\gamma}) + P(\boldsymbol{\gamma}; \lambda) + \frac{1}{2}\|\boldsymbol{\gamma} - \boldsymbol{\beta}\|_2^2 \\ &\quad - \sum_{i=1}^{n} (b(\boldsymbol{x}_i^T \boldsymbol{\gamma}) - b(\boldsymbol{x}_i^T \boldsymbol{\beta})) + \sum_{i=1}^{n} \mu_i(\boldsymbol{\beta})(\boldsymbol{x}_i^T \boldsymbol{\gamma} - \boldsymbol{x}_i^T \boldsymbol{\beta}), \end{aligned}$$

where $\mu_i = g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta}) = b'(\boldsymbol{x}_i^T \boldsymbol{\beta})$.

Given $\boldsymbol{\beta}$, minimizing G over $\boldsymbol{\gamma}$ is equivalent to

$$\min_{\boldsymbol{\gamma}} \frac{1}{2} \left\| \boldsymbol{\gamma} - \left[\boldsymbol{\beta} + \mathbf{X}^T \boldsymbol{y} - \mathbf{X}^T \boldsymbol{\mu}(\boldsymbol{\beta}) \right] \right\|_2^2 + P(\boldsymbol{\gamma}; \lambda).$$

This problem is an OLS problem with an orthogonal design.

Equivalent optimization

Given γ, minimizing G over $\boldsymbol{\beta}$ is equivalent to

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\gamma - \boldsymbol{\beta}\|_2^2 - \sum_{i=1}^{n}\left[b(\boldsymbol{x}_i^T\gamma) - b(\boldsymbol{x}_i^T\boldsymbol{\beta}) - b'(\boldsymbol{x}_i^T\boldsymbol{\beta})(\boldsymbol{x}_i^T\gamma - \boldsymbol{x}_i^T\boldsymbol{\beta})\right].$$

Taking its derivative with respect to $\boldsymbol{\beta}$ gives

$$(\boldsymbol{I} - \boldsymbol{\mathcal{I}}(\boldsymbol{\beta}))(\boldsymbol{\beta} - \gamma) = \mathbf{0},$$

where $\boldsymbol{\mathcal{I}}(\boldsymbol{\beta}) = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$ with $\boldsymbol{W} = \text{diag}\left\{b''(\boldsymbol{x}_i^T\boldsymbol{\beta})\right\}$. $\boldsymbol{\mathcal{I}}(\boldsymbol{\beta})$ is the observed/expected information matrix $[-\partial^2 L(\boldsymbol{\beta})/\partial\beta_h\beta_l]$ at $\boldsymbol{\beta}$. Intuitively, the optimal value of G is achieved at $\gamma = \boldsymbol{\beta}$ as long as $\boldsymbol{X}$ is scaled down properly. It is easy to verify $\min_{\boldsymbol{\beta}} G(\boldsymbol{\beta}, \boldsymbol{\beta})$ is equivalent to $\min_{\boldsymbol{\beta}} F(\boldsymbol{\beta})$. The advantage of optimizing G instead of F is that given $\boldsymbol{\beta}$, the problem is orthogonal and separable in γ, and we can adopt non convex penalties.

Iterative shrinkage for GLMs

Now given a thresholding function δ corresponding to the penalties already seen, use MM to get the estimates. The iterations simplify to

$$\boldsymbol{\beta}^{(j+1)} = \delta_\lambda(\boldsymbol{\beta}^{(j)} + \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{\mu}(\boldsymbol{\beta}^{(j)}))$$

with $\boldsymbol{X}$ scaled down properly at each iteration with corresponding weights $\text{diag}(\boldsymbol{W}(\boldsymbol{\beta}^{(j)}))$. This provides a generalization of iterative shrinkage for any GLM.

MERCI!