# Penalized likelihood methods for finding important variables in black boxes

A. Antoniadis
Université Joseph Fourier

IFP, March 23, 2007

# Sensitivity Analysis

Consider a model (computer code or black box function) written as a function $f$ defined on a domain of $\mathbb{R}^p$ with values in $\mathbb{R}^m$ :

$$\mathbf{Y} = \mathbf{Y}(\mathbf{X}) = \mathbf{f}(\mathbf{X}),$$

with $\mathbf{X} \in \mathbb{R}^p$ the "inputs" and $\mathbf{Y} \in \mathbb{R}^m$ the outputs.

GSA tries to find the importance and the weight of uncertainties of the components of $\mathbf{X}$ on the variability of the response $\mathbf{Y}(\mathbf{X})$.

As dimension increases many numerical problems become more statistical, because the sample is inevitable sparse and error depends on un-sampled parts of the domain.

# Meta-model (approximation)

Let

$$f(\mathbf{x}) \simeq \sum_{r \in \mathcal{R}} \beta_r \psi_r(\mathbf{x}) + \eta(\mathbf{x}).$$

The functions $\psi_r(\mathbf{x})$ are basis functions of $L^2([0,1]^p)$ such that:

- $\psi_0(\mathbf{x}) = 1, \quad \forall \mathbf{x} \in [0,1]^p$
- $\int \psi_j^2(\mathbf{x}) d\mathbf{x} = 1, \ \forall j \in \mathcal{R}$
- $\int \psi_j(\mathbf{x}) \psi_k(\mathbf{x}) d\mathbf{x} = 0, \quad j \neq k$

$\eta(\mathbf{x})$ is a deterministic truncation error.

Usually, $\{\psi_j\}_{r \in \mathbb{N}^p}$ is a tensor product basis made from univariate basis functions (sinusoids, orthogonal B-splines, orthogonal polynomials, wavelets).

## Interpretation in GSA

- Variance of $f$ is

$$\sigma^2(f) = \sum_{r \in \mathcal{R}, r \neq 0} \beta_r^2 + \int \eta(\mathbf{x})^2 d\mathbf{x}$$

- Importance of a subset $\mathcal{S}$ of coefficients is

$$\sigma_{\mathcal{S}}^2 = \sum_{\mathbf{r} \in \mathcal{S}} \beta_{\mathbf{r}}^2$$

- Normalized version is

$$\sigma_{\mathcal{S}}^2 / \sigma^2(f)$$

## Estimation ($d = \text{card}(\mathcal{R})$)

Let : $Z(\mathbf{x}) = (\psi_0(\mathbf{x}), \ldots, \psi_{d-1}(\mathbf{x}))^T$

The optimal $\boldsymbol{\beta}$ is

$$\begin{aligned} \boldsymbol{\beta}^* &= \text{argmin}_{\boldsymbol{\beta}} \int (f(\mathbf{x}) - Z(\mathbf{x})^T \boldsymbol{\beta})^2 d\mathbf{x} \\ &= \left( \int Z(\mathbf{x}) Z(\mathbf{x})^T d\mathbf{x} \right)^{-1} \int Z(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} \end{aligned}$$

and the ISE is

$$ISE = \int (f(\mathbf{x}) - Z(\mathbf{x})^T \boldsymbol{\beta})^2 d\mathbf{x}.$$

## Orthogonality implies

$$\begin{aligned}\boldsymbol{\beta}^* &= \left(\int Z(\mathbf{x})Z(\mathbf{x})^T d\mathbf{x}\right)^{-1} \int Z(\mathbf{x})f(\mathbf{x})d\mathbf{x} \\ &= \int Z(\mathbf{x})f(\mathbf{x})d\mathbf{x}.\end{aligned}$$

**Observations**

$$\mathbf{x}_i \sim U([0,1]^p),\ 1 \le i \le n,\ i.i.d.$$

**Regression**

$$\hat{\boldsymbol{\beta}} = (\mathcal{Z}^T\mathcal{Z})^{-1}\mathcal{Z}^T\mathcal{Y}, \quad \mathcal{Z}_{n\times d},\ \mathcal{Y}_{n\times 1} \text{ and } \hat{f}(\mathbf{x}) = Z(\mathbf{x})^T\hat{\boldsymbol{\beta}}.$$

## Quasi-regression

For an orthonormal basis of $L^2([0,1]^p)$ the matrix $\left(\int Z(\mathbf{x})Z(\mathbf{x})^T d\mathbf{x}\right) = I$.

For a design derived from a uniform distribution in $[0,1]^p$, the LLN allows to approximate $\mathcal{Z}^T\mathcal{Z}$ by the identity. Therefore

$$\tilde{\boldsymbol{\beta}} = \frac{1}{n}\mathcal{Z}^T\mathcal{Y}.$$

The quality of approximation depends on $d$ and the norm of $\boldsymbol{\beta}$.

Important to penalize $\boldsymbol{\beta}$ for reducing the variability in estimating the sensitivity indices and also to obtain a parcimonious metamodel !

See Efromovich (1992), Owen (1992, 1998), Antoniadis (2005).

# Regularization

A popular method for fitting a *regression* function from data measurements is *regularization*: minimize an objective function which enforces a roughness penalty in addition to coherence with the data.

Penalizing the squared norm of $\boldsymbol{\beta}$ in the Gaussian logl-ikelihood leads to ridge quasi-regression

$$\tilde{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \|\mathbf{Y} - Z\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2.$$

However, when the dimension $d$ is large, it is better to look for sparse models (as in wavelets) by using a “Lasso”-type penalty:

$$\tilde{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \|\mathbf{Y} - Z\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1$$

which leads to soft-thresholded estimators.

## General penalties

Several penalty functions have been used in the literature.

- The $L_2$ penalty $\psi(\beta) = |\beta|^2$ yields a ridge type regression
- The $L_1$ penalty $\psi(\beta) = |\beta|$ results in LASSO (first proposed by Donoho and Johnstone (1994) in the wavelet setting and extended by Tibshirani (1996) for general least squares settings).
- More generally, the $L_q$ $(0 \leq q \leq 1)$ leads to bridge regression (see Frank and Friedman (1993), Ruppert and Carroll (1997), Fu (1998), Knight and Fu (2000), Yu and Ruppert (2001)).

## Conditions on $\psi$

Usually, the penalty function $\psi$ is chosen to be symmetric and increasing on $[0, +\infty)$. Furthermore, $\psi$ can be convex or non-convex, smooth or non-smooth.

In the wavelet setting, Antoniadis and Fan (2001) provide some insights into how to choose a penalty function. A good penalty function should result in

- *unbiasedness*,
- *sparsity*,
- *stability*.

# Examples

| Penalty function | Convexity | Smoothness at 0 | Authors |
|---|---|---|---|
| $\psi(\beta) = \lvert\beta\rvert$ | yes | $\psi'(0^+) = 1$ | (Rudin 1992) |
| $\psi(\beta) = \lvert\beta\rvert^\alpha,\ \ \alpha \in (0,1)$ | no | $\psi'(0^+) = \infty$ | (Saquib 1998) |
| $\psi(\beta) = \alpha\lvert\beta\rvert/(1+\alpha\lvert\beta\rvert)$ | no | $\psi'(0^+) = \alpha$ | (Geman 92, 95) |
| $\psi(0) = 0,\ \psi(\beta) = 1, \forall\beta \neq 0$ | no | discontinuous | Leclerc 1989 |
| $\psi(\beta) = \lvert\beta\rvert^\alpha,\ \alpha > 1$ | yes | yes | Bouman 1993 |
| $\psi(\beta) = \alpha\beta^2/(1+\alpha\beta^2)$ | no | yes | McClure 1987 |
| $\psi(\beta) = \min\{\alpha\beta^2, 1\}$ | no | yes | Geman 1984 |
| $\psi(\beta) = \sqrt{\alpha+\beta^2}$ | yes | yes | Vogel 1987 |
| $\psi(\beta) = \log(\cosh(\alpha\beta))$ | yes | yes | Green 1990 |
| $\psi(\beta) = \begin{cases} \beta^2/2 & \text{if} \quad \lvert\beta\rvert \leq \alpha, \\ \alpha\lvert\beta\rvert - \alpha^2/2 & \text{if} \quad \lvert\beta\rvert > \alpha. \end{cases}$ | yes | yes | Huber 1990 |

Examples of penalty functions

## Discussion

- unbiasedness $\leftrightarrow \dot{\psi}(|\beta|) = 0$
- sparsity $\leftrightarrow |\beta| + \lambda\dot{\psi}(|\beta|) \geq 0$
- stability $\leftrightarrow$ argmin$\{|\beta| + \lambda\dot{\psi}(|\beta|)\} = 0$

From the above, a penalty satisfying the conditions on sparsity and stability must be non-smooth at 0. !

# SCAD

A penalty satisfying all the above is the one associated to the *SCAD* thresholding rule (Antoniadis & Fan (2001))

$$\delta_\lambda^{\mathrm{SCAD}}(\hat{\beta}) = \begin{cases} \mathrm{sign}(\hat{\beta})\max(0, |\hat{\beta}| - \lambda) & \text{if } |\hat{\beta}| \leq 2\lambda \\ \frac{(a-1)\hat{\beta} - a\lambda\,\mathrm{sign}(\hat{\beta})}{a-2} & \text{if } 2\lambda < |\hat{\beta}| \leq a\lambda \\ \hat{\beta} & \text{if } |\hat{\beta}| > a\lambda \end{cases} \tag{1}$$

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

# Penalizing Gaussian Kriging models

Kriging is popular analysis approach for computer experiments when one wants to create a cheap-to-compute meta-model.

The maximum likelihood approach is used to estimate the parameters in the kriging model.

If the likelihood function around the optimum is flat, then the resulting mle estimates of the parameters of the covariance matrix have large variances.

We will use also here a penalization approach to overcome this problem.

# Gaussian Kriging

$\mathbf{x}_i$, $i = 1, \ldots, N$ design points over a $p$-dimensional experimental domain $D$, $y_i = y(\mathbf{x}_i)$ sampled from

$$y(\mathbf{x}_i) = \mu + z(\mathbf{x}_i),$$

where $z(\mathbf{x})$ is a Gaussian process with mean 0 and covariance between $\mathbf{x}_i$ and $\mathbf{x}_j$,

$$r(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \exp\left\{ -\sum_{k=1}^{p} \theta_k |x_{ik} - x_{jk}|^2 \right\},$$

where $\theta_k \geq 0$.

Let $\boldsymbol{\gamma} = (\theta_1, \ldots, \theta_p, \sigma^2)^T$ and define $R(\boldsymbol{\gamma})$ to be the $N \times N$ matrix with the $(ij,)$th element $r(\mathbf{x}_i, \mathbf{x}_j)$.

# Gaussian Kriging (bis)

The density of $\mathbf{y} = (y_1, \ldots, y_N)^T$ is

$$f(\mathbf{y}) = (2\pi)^{-N/2} |R(\boldsymbol{\gamma})|^{-1/2} \exp\left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{1}_N \mu)^T R(\boldsymbol{\gamma})^{-1} (\mathbf{y} - \mathbf{1}_N \mu) \right\}$$

and the log-likelihood is proportional to

$$\ell(\mu, \boldsymbol{\gamma}) = -\frac{1}{2} \log R(\boldsymbol{\gamma})| - \frac{1}{2} (\mathbf{y} - \mathbf{1}_N \mu)^T R(\boldsymbol{\gamma})^{-1} (\mathbf{y} - \mathbf{1}_N \mu).$$

Once $\mu$, $\sigma^2$ and $\gamma$ are estimated, the BLUE can be calculated by

$$\hat{y}(\mathbf{x}) = \hat{\mu} + \mathbf{b}(\mathbf{x}) R(\hat{\boldsymbol{\gamma}})^{-1} (\mathbf{y} - \mathbf{1}_N \hat{\mu}),$$

with variance

$$\hat{\text{var}}[\hat{y}(\mathbf{x})] = \hat{\sigma^2} - \mathbf{b}(\mathbf{x}) R(\hat{\boldsymbol{\gamma}})^{-1} \mathbf{b}(\mathbf{x}),$$

where $\mathbf{b}(\mathbf{x}) = (\hat{r}(\mathbf{x}, \mathbf{x}_1), \ldots, \hat{r}(\mathbf{x}, \mathbf{x}_N))$.

# Why using regularization?

Focus on estimation of $\boldsymbol{\theta}$ (fixing $\mu$ and $\sigma^2$). We have

$$\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}_0) + \ell'(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \ell''(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + O_P(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2).$$

When $\ell(\boldsymbol{\theta})$ is flat around $\boldsymbol{\theta}_0$ then $\ell''(\boldsymbol{\theta}_0)$ is nearly singular and the covariance of $\hat{\boldsymbol{\theta}}$ will be very large. In such situations, a penalized likelihood estimator may be shown to perform better than the MLE. We will therefore use instead of $\ell$ the following criterion

$$Q(\mu, \boldsymbol{\gamma}) = -\frac{1}{2}\log R(\boldsymbol{\gamma})| - \frac{1}{2}(\mathbf{y} - \mathbf{1}_N\mu)^T R(\boldsymbol{\gamma})^{-1}(\mathbf{y} - \mathbf{1}_N\mu) - N\sum_{k=1}^{p}\psi_\lambda(\gamma_k),$$

where $\psi_\lambda(\cdot)$ is any penalty (in particular the SCAD penalty). To compute the solution we use Fisher's scoring algorithm. The regularization parameter $\lambda$ is chosen by $K$-fold cross-validation ( note that GCV cannot be used since Gaussian kriging gives an exact fit at the design points).

## Actual problems

- Better algorithms
- Asymptotic properties
- Selection methods for penalty parameters
- Adequate penalties