On Regularized Estimation and/for Stochastic Differential Equations

Stefano M. Iacus (University of Milan)

Third YUIMA Workshop @ Brixen-Bressanone, 26-06-2019

About regularized estimation

Let us consider the simple regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma_\epsilon^2)$$

$\beta \in \Theta \subseteq \mathbb{R}^{p+1}$. The OLS solution is given by

$$\hat{\beta}^{LS} = \underset{\beta \in \Theta}{\arg\min} \, ||Y - X\beta||^2 = (X'X)^{-1} X'Y$$

unbiased, asymptotically normal and with minimal variance (in the class of linear estimators) provided that $(X'X)^{-1}$ exists or, more specifically, that $X'X$ is full rank.

General motivation

Typical issues are:

- **multi-collinearity**: the regressors are correlated, therefore $X'X$ is not full rank and the variance of estimates may diverge
- there are **far more regressors than observations**, i.e., $n << p$: identifiability problem
- we are not sure about **model specification**: need of model selection

Although the above issues has always been there, they became more and more compelling recently due to the deluge of new data (genomics, finance, social sciences, ecc)

My motivation

Helping the E.U. Commission to setup and early warning and forecasting system for asylum applicants to EU28+ to setup the logistics and necessary HR capabilities before crises explode.

- 28 countries of destination (CoD) from 220 world countries of origin (CoO)
- Frontex data, monthly: irregular border crossings from CoO to about 10 CoD
- EASO data, weekly: from CoD to all CoD
- GDelt data: daily: conflict, social, economics events for (almost) all CoO
- Google Search, weekly: from CoO looking for different searches: the EU28 countries, Visa, Passport, Asylum, etc (around 15)
- adding lagged effects: some populations move from a CoO, then transits to other CoO and then enters EU.
- outcome: applicants in 4 weeks from each CoO to each CoD (and EU in general)

A few hundreds of time series and **weekly frequency** only up to 2016-2017 (about 52*3.5 data). No way to fit, e.g., VAR models. Push factors and triggers are different for each route (model selection problem).

Nevertheless, using regularized estimation seems to produce working results.

Ridge estimation

Ridge regression (Hoerl and Kennard, 1970): solution can be formalized as an **unconstrained but penalized** optimization problem

$$\hat{\beta}_\lambda^R \equiv \underbrace{\text{argmin}_\beta \|Y - X\beta\|^2}_{\text{LS}} + \underbrace{\lambda \sum_{j=1}^{p} \beta_j^2}_{l_2 \text{ - 'penalty'}} = \text{argmin}_\beta RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \quad (1)$$

or as an **unpenalized but constrained** optimization problem

$$\hat{\beta}_\lambda^R = \underset{||\beta||^2 \leq s}{\arg\min} ||Y - X\beta||^2$$

Approximate relationship: $\lambda \sim \frac{1}{s}$

Ridge regression: shrinkage

$$\hat{\beta}_\lambda^R = \text{argmin}_\beta RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

- The term $\lambda \sum_{j=1}^{p} \beta_j^2$ is called a **shrinkage penalty**.
- It depends on the tuning parameter λ:
  - when $\lambda = 0$ the shrinkage penalty term has no effect and $\hat{\beta}_\lambda^R = \hat{\beta}^{LS}$.
  - as λ grows the shrinkage effect increases too and $\beta_1, \ldots, \beta_p$ approach zero.
  - selecting the best value for λ is crucial (data dependent).
  - λ penalizes each β_j differently unless X is standardized

Properties of Ridge estimates

The explicit solution to (1) is: $$\hat{\beta}_{\lambda}^{R} = (X'X + \lambda I)^{-1} X'Y$$

Ridge regression can “solve” the **multi-collinearity** problem.

Properties of Ridge estimates

The explicit solution to (1) is: $\hat{\beta}_\lambda^R = (X'X + \lambda I)^{-1} X'Y$

Ridge regression can "solve" the **multi-collinearity** problem. Let $W = (X'X + \lambda I)^{-1}$, then

$$\text{Bias}(\hat{\beta}_\lambda^R) = -\lambda W \beta, \qquad \text{Var}(\hat{\beta}_\lambda^R) = \sigma_\epsilon^2 W X'X W'$$

Thus, the **bias depends on** λ but and it is possible to prove that

$$\text{Var}\left(\hat{\beta}^{LS}\right) - \text{Var}\left(\hat{\beta}_\lambda^R\right)$$

is a positive definite matrix: **variance shrinkage**.

Properties of Ridge estimates

The explicit solution to (1) is: $\hat{\beta}_\lambda^R = (X'X + \lambda I)^{-1} X'Y$

Ridge regression can "solve" the **multi-collinearity** problem. Let $W = (X'X + \lambda I)^{-1}$, then

$$\text{Bias}(\hat{\beta}_\lambda^R) = -\lambda W \beta, \qquad \text{Var}(\hat{\beta}_\lambda^R) = \sigma_\epsilon^2 W X'X W'$$

Thus, the **bias depends on** λ but and it is possible to prove that

$$\text{Var}\left(\hat{\beta}^{LS}\right) - \text{Var}\left(\hat{\beta}_\lambda^R\right)$$

is a positive definite matrix: **variance shrinkage**. Further

$$MSE\left(\hat{\beta}^{LS}\right) - MSE\left(\hat{\beta}_\lambda^R\right) \lessgtr 0$$

but it is possible to prove (see, Theobald, 1974 and Farebrother, 1976) that **there always exists** a value of $\lambda > 0$ such that the above quantity is **strictly positive**.

Ridge regression: averaging effect

Suppose now that X_1, X_2 are standardized and strongly positively collinear, and their population slopes are β_1 and β_2 respectively, then their Ridge estimates are
Fits of the form:

$$(\beta_1 + \gamma)X_1 + (\beta_2 - \gamma)X_2 = X_1\beta_1 + \gamma X_1 + \beta_2 X_2 - \gamma X_2 = EY + \gamma(X_1 - X_2),$$

have similar MSE values as γ varies, since $X_1 - X_2$ is small when X_1 and X_2 are strongly positively associated.

In other words, OLS can't easily distinguish among these fits.

For example, if $X_1 \approx X_2$, then $3X_1 + 3X_2$, $4X_1 + 2X_2$, $5X_1 + X_2$, etc. all have very similar MSE values.

Ridge regression and collinearity (cont'd)

For large λ, ridge regression favors the fits that minimize:

$$(\beta_1 + \gamma)^2 + (\beta_2 - \gamma)^2.$$

This expression is minimized at $\gamma = (\beta_2 - \beta_1)/2$, giving the fit:

$$\frac{(\beta_1 + \beta_2)X_1}{2} + \frac{(\beta_1 + \beta_2)X_2}{2} = (\beta_1 + \beta_2)\frac{X_1 + X_1}{2}.$$

Therefore, multi-collinear variables **share the same estimated coefficients** or, put it in another way, it averages covariates.

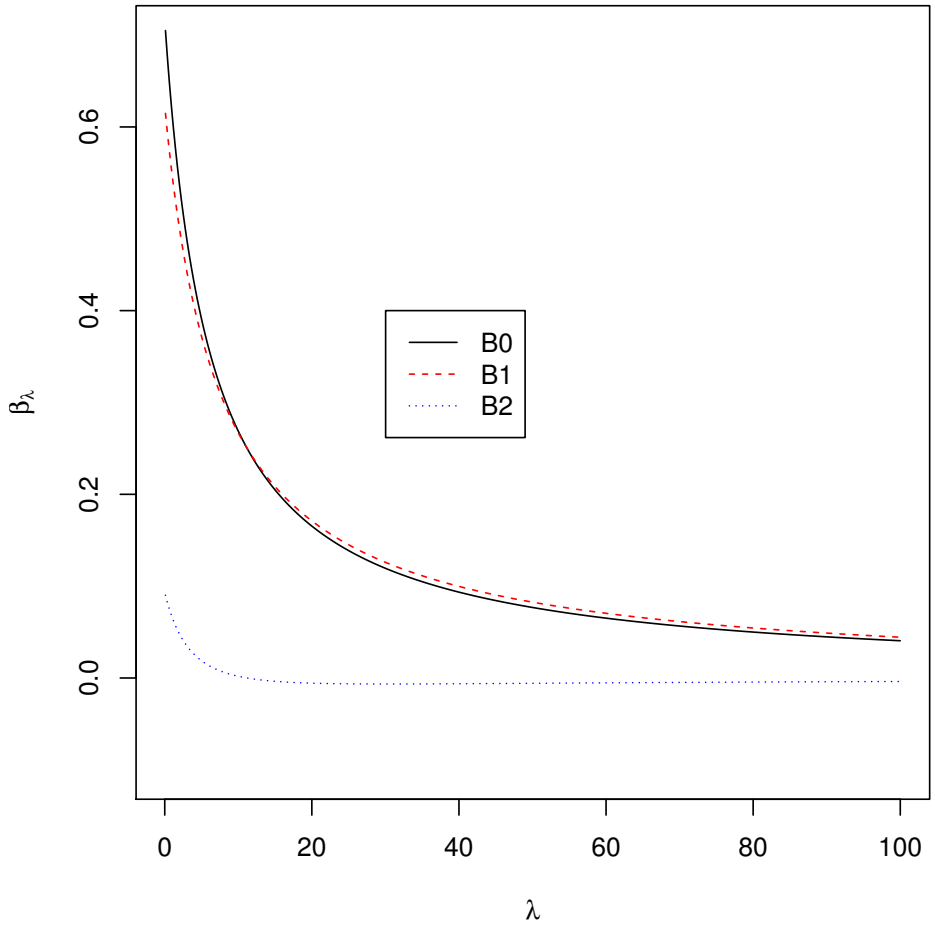
Ridge regression effective degrees of freedom

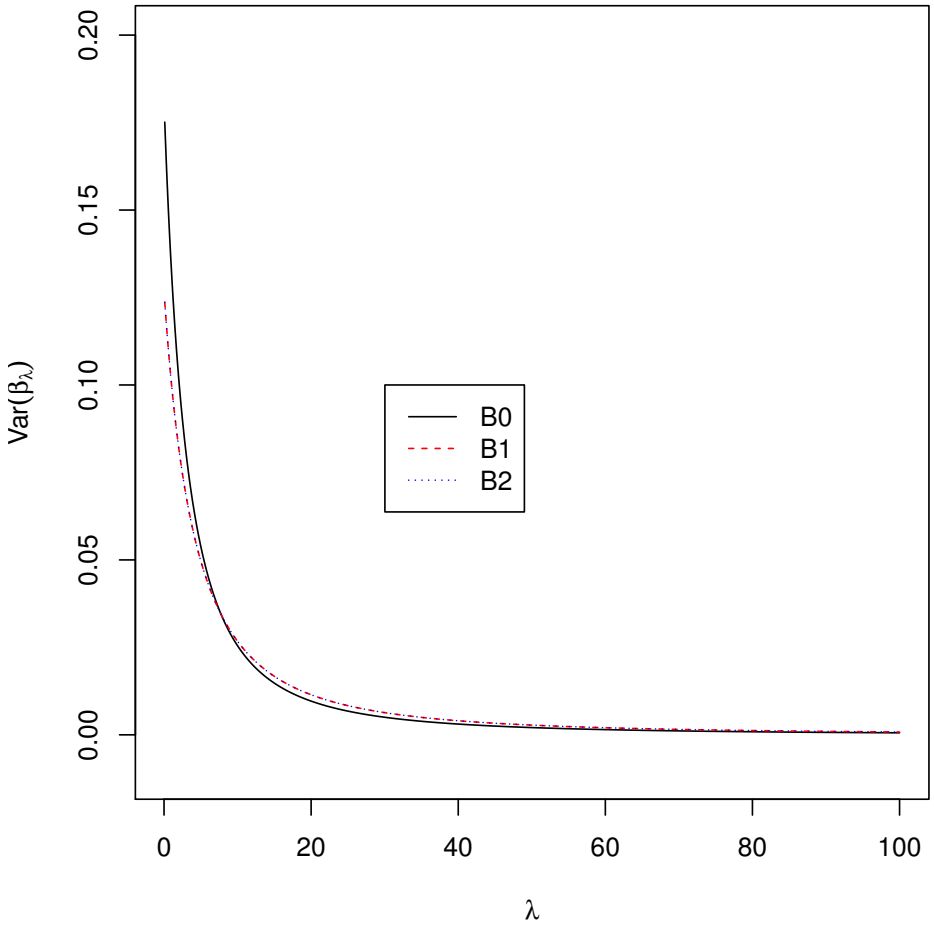
In OLS the trace of the "hat" matrix $H = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'$ is equal to the rank of X, which corresponds to the number of free independent parameters of the linear model = degrees of freedom. In Ridge regression effective degrees of freedom (EDF) are

$$\mathrm{EDF}_\lambda = \mathrm{tr}\left[X(X'X + \lambda I)^{-1}X'\right]$$

When $\lambda = 0$ and there is no multi-collineraity the matrix $X'X$ is full rank, otherwise the EDF converges to 1 as λ grows, i.e. all coefficients other than the intercept are forced to take the value zero.

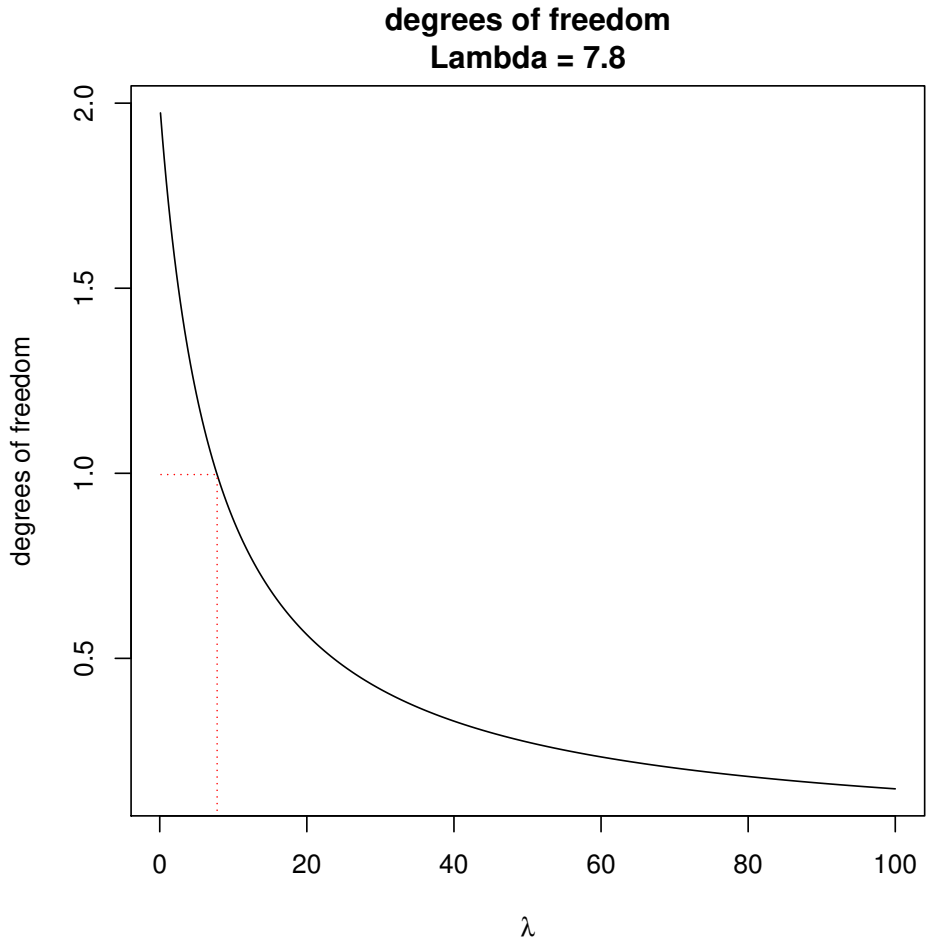
The shrinkage effect of λ

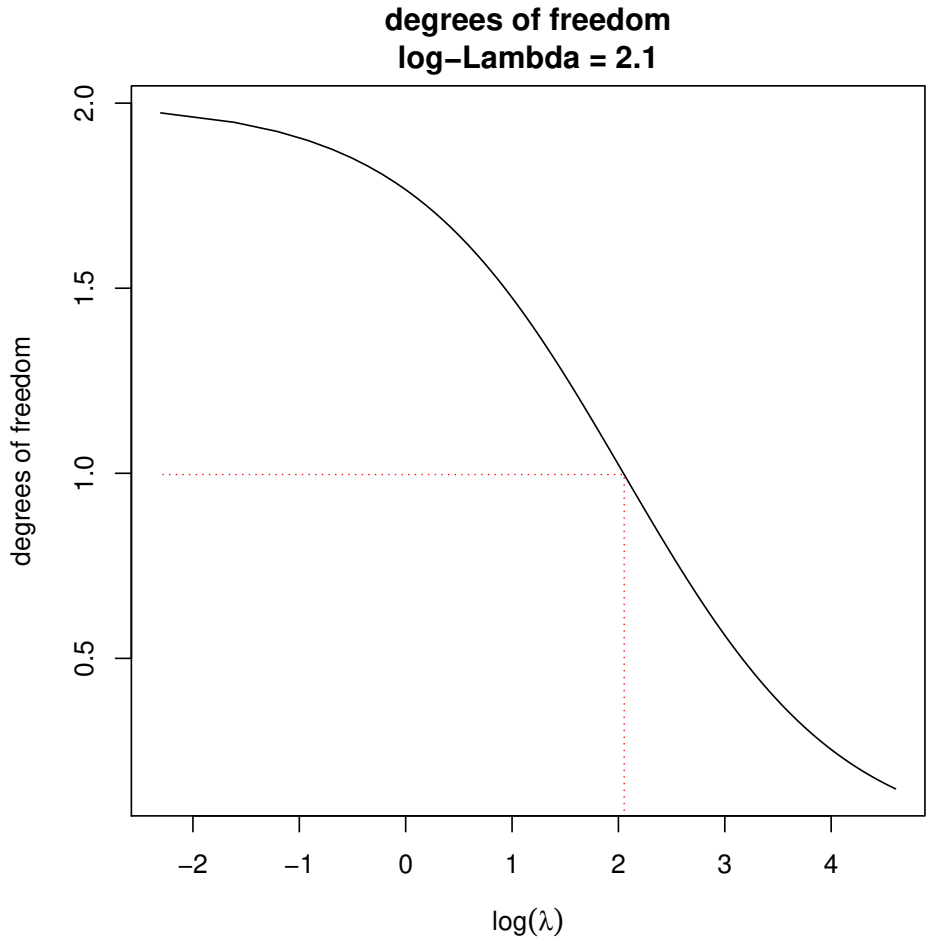




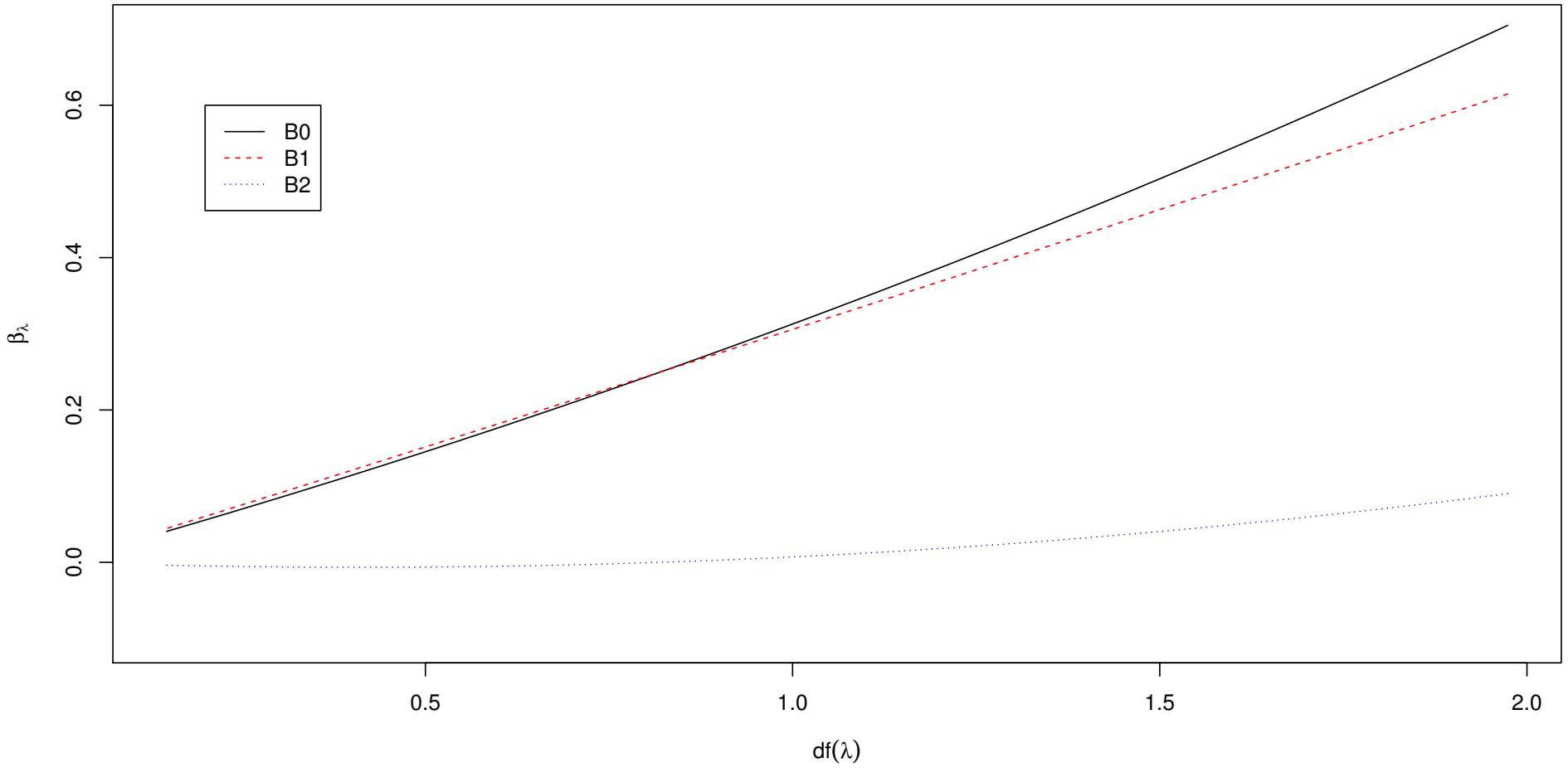
Ridge regression's EDF vs λ

The effective number of degrees of freedom as a function of λ





Ridge regression estimates vs effective degrees of freedom



Ridge regression summary

So, Ridge estimates are

- **biased but with less variance** than OLS
- can address the **multicollinearity** problem (also average effect on the coefficients)
- notice that, incidentally, when $p >> n$, $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist (*ill-posed problem*) and thus Ridge regression can also help.
- but there is **no model selection** effect, i.e., the number of coefficients remains p, unless $\lambda \to \infty$ [$\mathrm{EDF}_\lambda \to 1$]
- Still, another problem remains: **overfitting**.
  - A model with too many predictor variables may be sub-optimal if the true model is sparse (i.e., the response variable Y depends only on a small number of input variables).

Sparse Estimation

Lasso: Least Absolute Selection and Shrinkage Operator

Lasso estimates (see Tibshirani, 1996; Knight and Fu, 2000, Efron *et al.*, 2004) minimize

$$RSS + \lambda \sum_{j=1}^{k} |\beta_j|.$$

The important difference with ridge regression is in the penalty part (l_1 vs l_2). This seemingly tiny difference makes qualitative gaps practically as well as theoretically.

The l_1 penalty causes some coefficients to be shrunken exactly to zero, i.e., the predictive model is **sparse**

Lasso performs both variable selection **and** shrinkage

Bridge estimation

The previous Lasso approach can be generalized further to l_q constraints (**Bridge** estimation), for some $q > 0$, i.e.

$$\hat{\beta} = \arg\min_{\beta} RSS + \lambda \sum_{i=1}^{k} |\beta_i|^q$$

Where Lasso is for $q = 1$, Ridge is for $q = 2$ and the limiting case $q = 0$ is OLS.

Notice that, in the limit as $q \to 0$, this procedure approximates AIC/BIC criteria as

$$\lim_{q \to 0} \sum_{i=1}^{k} |\beta_i|^q = \sum_{i=1}^{k} \mathbf{1}_{\{\beta_i \neq 0\}}$$

as the RHS amounts to the number of non-null parameters.

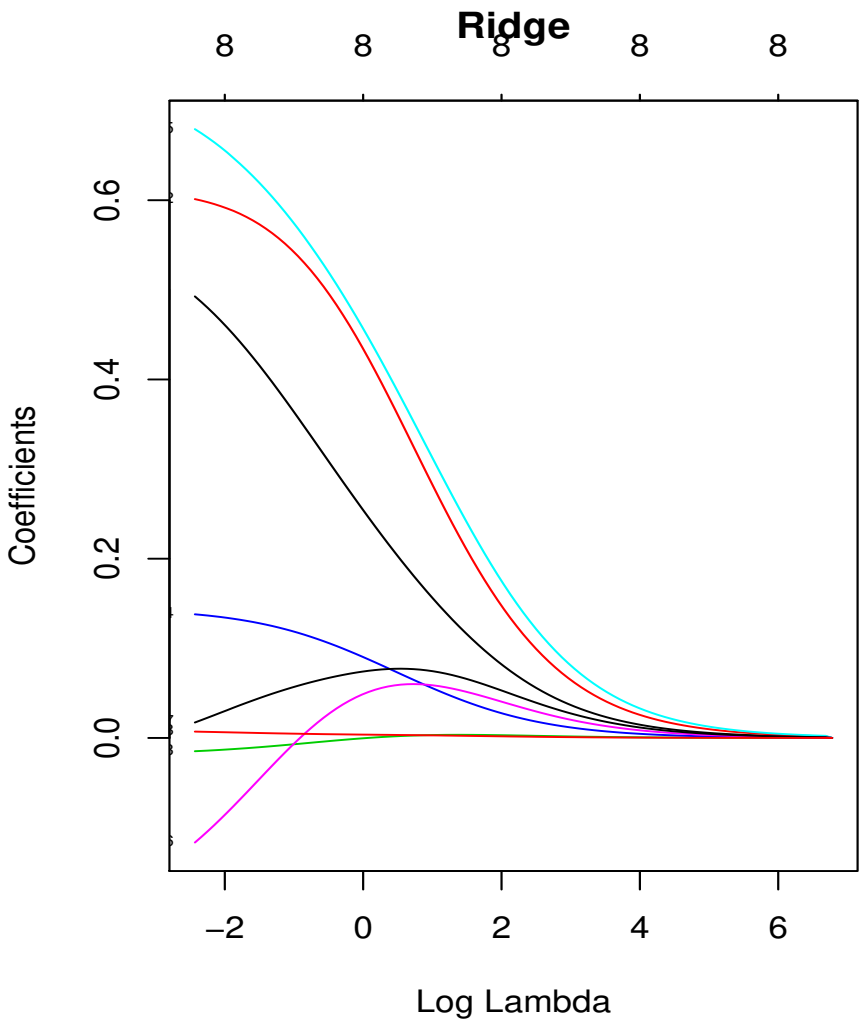
A typical Lasso result

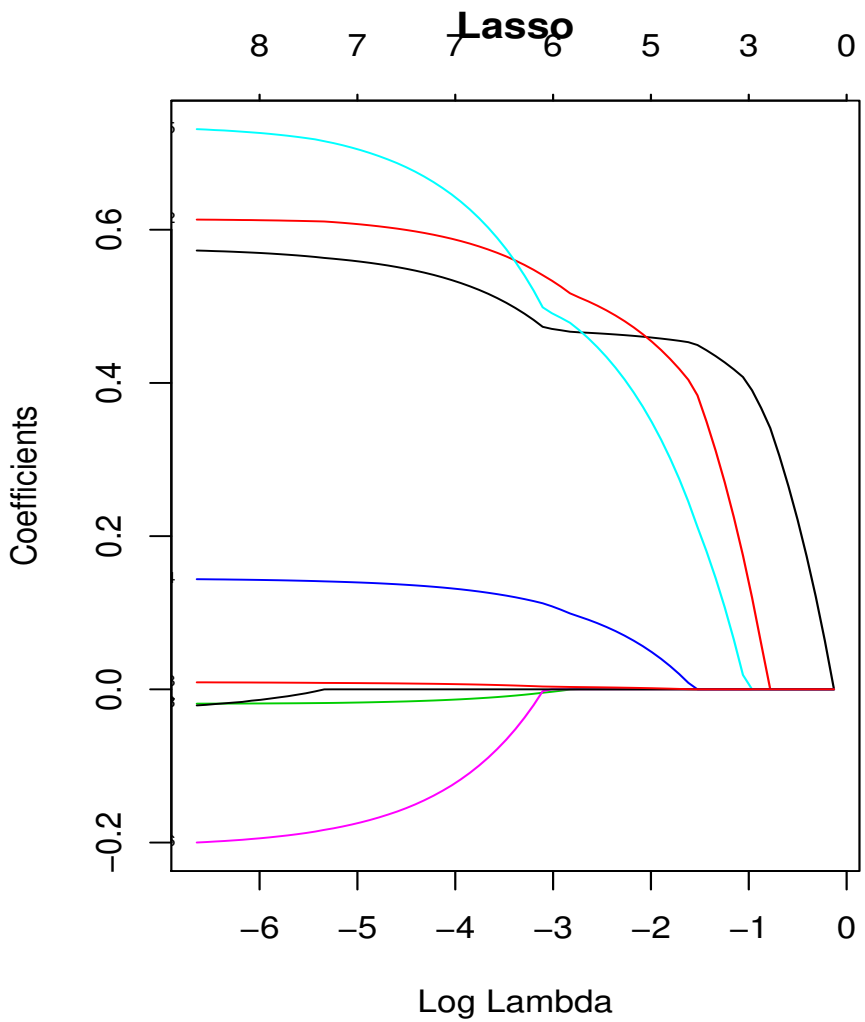
	OLS	OLS-Step*	Ridge	Lasso
(Intercept)	0.43	0.26	-0.15	0.33
lcavol	0.58	0.57	0.27	0.45
lweight	0.61	0.62	0.45	0.40
age	-0.02	-0.02	-0.00	–
lbph	0.14	0.14	0.09	0.01
svi	0.74	0.74	0.47	0.24
lcp	-0.21	-0.21	0.05	–
gleason	-0.03	–	0.07	–
pgg45	0.01	0.01	0.00	0.00
MSE (train)	0.44	0.44	0.56	0.59
R^2 (train)	0.69	0.69	0.61	0.59
MSE (test)	0.52	0.52	0.52	0.47
R^2 (test)	0.50	0.51	0.51	0.55

Lasso solution can achieve both model selection and shrinkage!

$*$: OLS-Step is the OLS model with stepwise regression.

Comparison of shirinkage between the Ridge and the Lasso





Geometric interpretation

Equivalent formulations for Lasso and Ridge

Formulation 1:

$$\hat{\beta}_L = \text{argmin}_\beta \left[RSS + \lambda \sum_{j=1}^{k} |\beta_j| \right], \qquad \hat{\beta}_R = \text{argmin}_\beta \left[RSS + \lambda \sum_{j=1}^{k} \beta_j^2 \right]$$

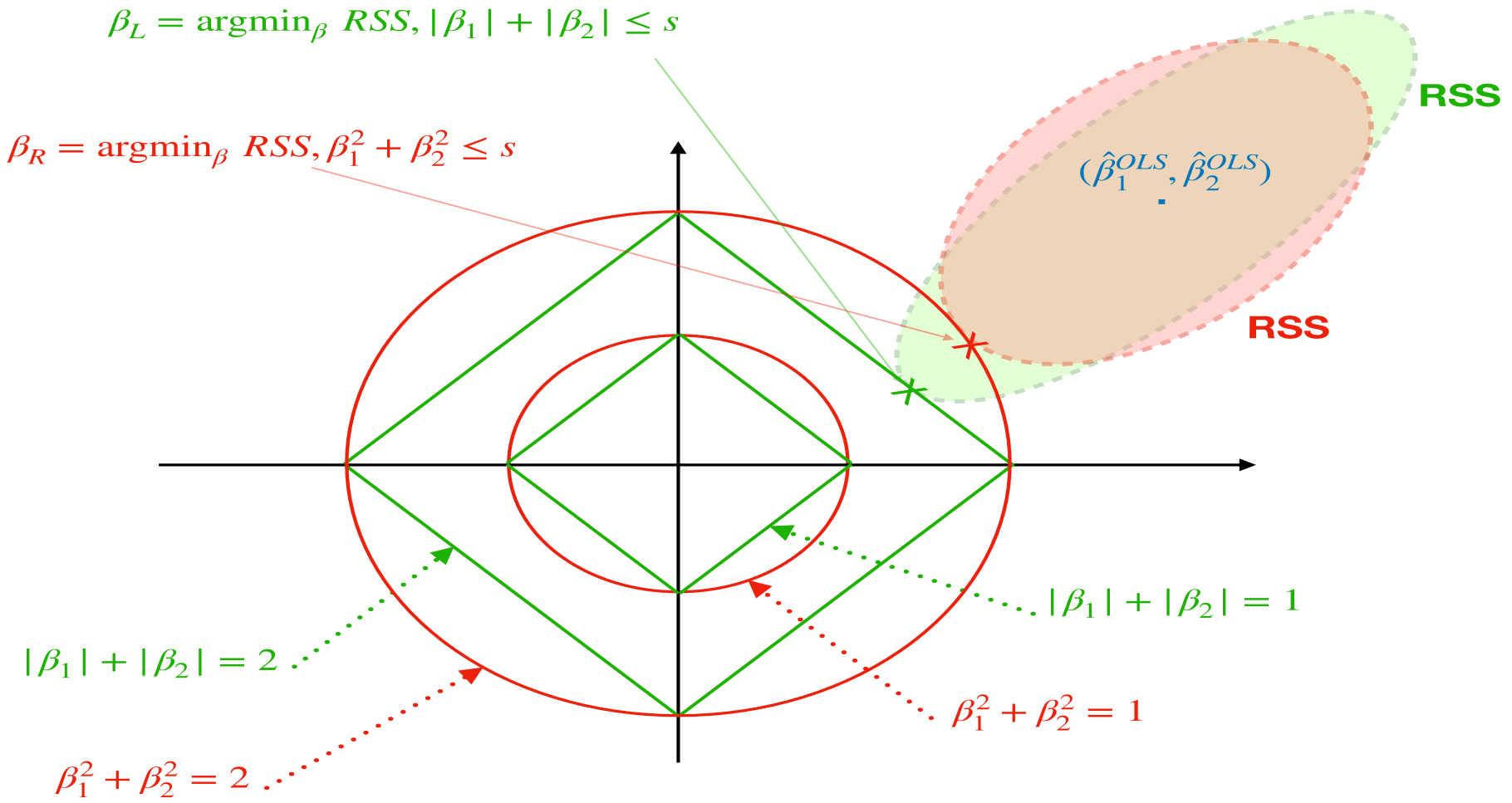
Formulation 2:

$$\hat{\beta}_L = \text{argmin}_\beta \left[\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij})^2 \right]$$

subject to $\sum_{j=1}^{k} |\beta_j| \leq s$ for Lasso and $\sum_{j=1}^{k} |\beta_j|^2 \leq s$ for Ridge.

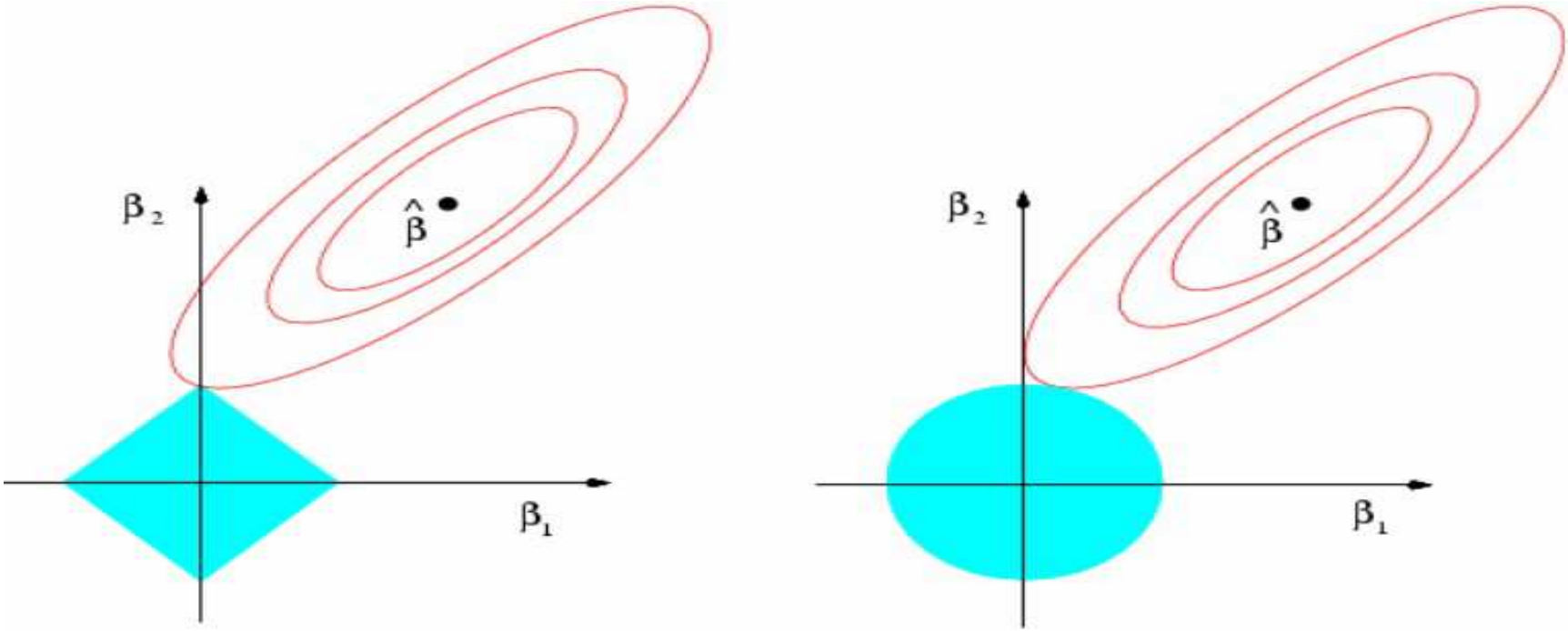
Changing the value for s/λ

$\lambda \uparrow$ or $s \downarrow$, the smaller estimates and viceversa



$\lambda \uparrow \Leftrightarrow s \downarrow \Rightarrow |\beta| \downarrow$

Why Lasso can give sparse solutions, but not Ridge



Left: Lasso, solutions can reach the edge of the diamond for both coefficients while (right) this is not possible for Ridge.

Extensions of the Lasso

Lasso methods for selecting block of predictors at a time (categorical variable represented through dummies). Indeed, individual sparsity does NOT ensure blockwise sparsity: **Group Lasso** (Yuan and Lin, 2006) and **Blockwise Sparse Regression** (Kim et al. 2006).

Beware: sparse methods are OK ONLY IF the true model is sparse.

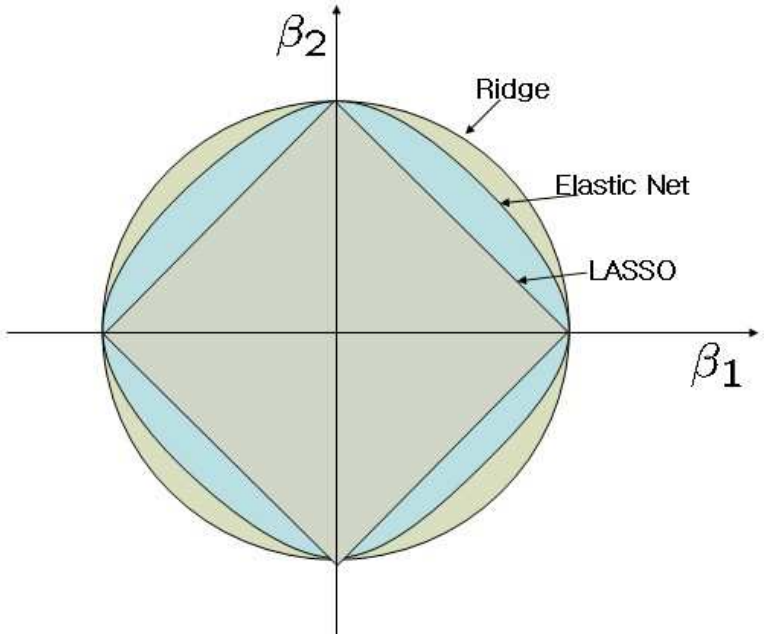
When there is a high correlation between predictors, the average of the correlated predictors (Ridge) might be better than selecting a single predictor (Lasso).

Shrinkage methods with an **oracle property**: asymptotically unbiased and consistent for the non null parameters.

There exists shrinkage methods with less sparse solutions than the Lasso: **Elastic Net** (Zou and Hastie, 2005)

Visual comparison of Elastic Net, Ridge and Lasso

Figure 1: Penalties for Lasso, ridge, and elastic net



Elastic Net: Computation

- The main idea of the **Elastic Net** is to find a compromise between Ridge and Lasso.
- *Naive elastic net* minimizes the following objective function:

$$RSS + \underbrace{\lambda_1 \|\boldsymbol{\beta}\|_1}_{Lasso} + \underbrace{\lambda_2 \|\boldsymbol{\beta}\|_2^2}_{Ridge}.$$

- Penalty for elastic net proposed by Zou and Hastie (2005):

$$\lambda \sum_{j=1}^{p} (\alpha \beta_j^2 + (1-\alpha)|\beta_j|)$$

- The elastic net selects variables (in the way the Lasso does), and shrinks together the coefficients of correlated predictors (in the way the Ridge regression does). For very small, $\alpha > 0$, ENet is almost Lasso but without the unpleasant degeneracies and wild behaviour in the presence of strong correlation.

- Fan and Li (2001) generalized the penalized approaches:

$$\hat{\beta} = \text{argmin}_{\beta} C(\beta) + \sum_{j=1}^{p} J_{\lambda}(|\beta_j|),$$

where J is a penalty function and $C(\beta)$ is a loss function (for example RSS or negative log-likelihood).

  - Hard thresholding (Fan, 1997): $\lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda)$
  - Bridge regression: $J_{\lambda}(\theta) = \lambda|\theta|^q, q > 0$
  - Lasso regression: $J_{\lambda}(\theta) = \lambda|\theta|$
  - Ridge regression: $J_{\lambda}(\theta) = \lambda|\theta|^2$
  - SCAD:

$$J_{\lambda}(\theta) = \begin{cases} \lambda|\theta|, & |\theta| \leq \lambda, \\ -(\theta^2 - 2a\lambda|\theta| + \lambda^2)/[2(a-1)], & \lambda \leq |\theta| \leq a\lambda \\ (a+1)\lambda^2/2, & |\theta| \geq a\lambda. \end{cases}$$

SCAD: Properties

Table 1: Penalty methods, sparsity and unbiasedness

	Bridge($q < 1$)	Lasso	Ridge	SCAD	ENet
sparsity	Yes	Yes	No	Yes	Yes
unbiasedness	Yes	No	No	Yes	No

Adaptive Estimation

Oracle procedures

Let $\mathcal{A} = \{j : \beta_j \neq 0\}$ be the set of true non-zero coefficients in the standard regression model

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

such that $|\mathcal{A}| = p_0 < p$. Denote by $\hat{\beta}(\delta)$ the estimates of an estimation procedure δ. Following Fan and Li (2001), we call δ an **oracle procedure** if $\hat{\beta}(\delta)$ (asymptotically) has the following oracle properties:

- Identifies the right subset model, $\{j : \hat{\beta}_j \neq 0\} = \mathcal{A}$
- Has the optimal estimation rate $\sqrt{n}(\hat{\beta}(\delta)_{\mathcal{A}} - \beta_{\mathcal{A}})$ converges in distribution to $N(0, \Sigma^*)$ where where Σ^* is the covariance matrix of the true subset/reduced model.

Remind that if all coefficients are non-zero, the MLE estimator satisfies

$$\sqrt{n}(\hat{\beta}^{ML} - \beta) \xrightarrow{d} N(0, \mathcal{I}^{-1}(\beta) =: \Sigma^*)$$

In the classic Lasso procedure, the main assumption are that

$$\frac{1}{n}X'X \to C$$

where C is A positive definite matrix. Let us re-order the coefficients β so that the true non-zero coefficients occupy the first positions $1, \ldots, p_0$. Then let

$$C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$$

where C_{11} is $p_0 \times p_0$. Now let $\lambda = \lambda_n$ in the Lasso penalty function

$$\hat{\beta}_n = \arg\min_{\beta} \left(RSS + \lambda_n \sum_{j=1}^{p} |\beta_j| \right)$$

Lasso is not an Oracle procedure!

If λ_n is such that $\lim_{n\to\infty} \lambda_n/n = \lambda_0 \geq 0$, then Lemma 1 (Knight and Fu, 2000):

$$\hat{\beta}_n \xrightarrow{p} \arg\min_{\beta} V_1, \quad \text{with} \quad V_1(u) = (u-\beta)'C(u-\beta) + \lambda_0 \sum_{j=1}^{p} |u_j|$$

and if $\lim_{n\to\infty} \lambda_n/\sqrt{n} = \lambda_0 \geq 0$ then, Lemma 2 (Knight and Fu, 2000):

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} \arg\min_{\beta} V_2$$

with

$$V_2(u) = -2u'W + u'Cu + \lambda_0 \sum_{j=1}^{p} \left(u_j \text{sign}(\beta_j) I_{\{\beta_j \neq 0\}} + |u_j| I_{\{\beta_j = 0\}} \right)$$

with $W = N(0, \sigma^2 C)$.

Lasso is not an Oracle procedure!

Lemma 1 shows that only if $\lambda_0 = 0$ the Lasso estimators are consistent.

Lemma 2 shows that Lasso can be $\sqrt{n}$-consistent under the same conditions. But in general bias remains.

Indeed, it is also possible to prove that

$$\lim_{n \to \infty} P(\mathcal{A}_n = \mathcal{A}) \leq c < 1$$

which means that the true set of non-zero coefficients is not correctly identified even asymptotically.

Adaptive Lasso addresses this problem.

Adaptive Lasso is an Oracle procedure!

Let $\tilde{\beta}$ be a $\sqrt{n}$-consistent estimator of β (e.g. OLS or MLE). Let $\gamma > 0$ and define $\tilde{w}_j = 1/|\tilde{\beta}|^\gamma$, $j = 1, \ldots, p$. The adaptive Lasso estimator is defined as follows

$$\hat{\beta} = \arg\min_{\beta} \left(RSS + \lambda_n \sum_{j=1}^{p} \tilde{w}_j |\beta_j| \right)$$

If $\lambda_n / \sqrt{n} \to 0$ and $\lambda_n n^{\frac{\gamma-1}{2}} \to \infty$, then (Zou, 2006), we have the **oracle** properties:

- consistent variable selection: $\lim_{n\to\infty} P(\mathcal{A}_n = \mathcal{A}) = 1$
- asymptotic normality: $\sqrt{n}(\hat{\beta}_{\mathcal{A}} - \beta_{\mathcal{A}}) \xrightarrow{d} N\left(0, \sigma^2 C_{11}^{-1}\right)$.

Adaptive Elastic Net is also oracle

In its other form, the Elastic Net estimator can be written as follows

$$\hat{\beta}^{ENet} = \left(1 + \frac{\lambda_2}{n}\right) \left\{ \arg\min_{\beta} \left(RSS + \lambda_2 ||\boldsymbol{\beta}||_2^2 + \lambda_1 ||\boldsymbol{\beta}||_1 \right) \right\}$$

then define $\tilde{w}_j = \left(|\hat{\beta}^{ENet}|\right)^{-\gamma}$ and finally

$$\hat{\beta}^{AdaENet} = \left(1 + \frac{\lambda_2}{n}\right) \left\{ \arg\min_{\beta} \left(RSS + \lambda_2 ||\boldsymbol{\beta}||_2^2 + \tilde{\lambda}_1 \sum_{j=1}^{p} \tilde{w}_j |\beta_j| \right) \right\}$$

but because ENet, like Lasso, likes sparse estimation, to avoid division by 0, we can use these weights $\tilde{w}_j = \left(|\hat{\beta}^{ENet}| + 1/n\right)^{-\gamma}$.

Then, again, it is possible to prove that $\hat{\beta}^{AdaENet}$ is **oracle** (see Zou and Zhang, 2009).

Application to Discretely Observed Stochastic Differential Equations

Let X_t be a diffusion process solution to

$$\mathrm{d}X_t = b(\alpha, X_t)\mathrm{d}t + \sigma(\beta, X_t)\mathrm{d}W_t$$

$$\alpha = (\alpha_1, ..., \alpha_p)' \in \Theta_p \subset \mathbb{R}^p, \quad p \geq 1$$

$$\beta = (\beta_1, ..., \beta_q)' \in \Theta_q \subset \mathbb{R}^q, \quad q \geq 1$$

$b : \Theta_p \times \mathbb{R}^d \to \mathbb{R}^d$, $\sigma : \Theta_q \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^m$ and $W_t, t \in [0, T]$, is a standard Brownian motion in $\mathbb{R}^m$.

We assume that the functions b and σ are known up to α and β.

We denote by $\theta = (\alpha, \beta) \in \Theta_p \times \Theta_q = \Theta$ the parametric vector and with $\theta_0 = (\alpha_0, \beta_0)$ its unknown true value.

Sampling scheme

The sample path of X_t is observed only at $n+1$ equidistant discrete times t_i, such that $t_i - t_{i-1} = \Delta_n < \infty$ for $1 \leq i \leq n$ (with $t_0 = 0$ and $t_n = T$). We denote by $\mathbf{X}_n = \{X_{t_i}\}_{0 \leq i \leq n}$ our random sample with values in $\mathbb{R}^{(n+1)\times d}$.

The asymptotic scheme adopted in this talk is the following:

$T = n\Delta_n \to \infty$, $\Delta_n \to 0$ and $n\Delta_n^2 \to 0$ as $n \to \infty$.

This asymptotic framework is called *rapidly increasing design* and the condition $n\Delta_n^2 \to 0$ means that Δ_n shrinks to zero slowly.

Implications: the parameters β are $\sqrt{n}$ – consistent while the parameters α in the drift are only $\sqrt{n\Delta_n}$ – consistent. This requires a non trivial adaptation of the Lasso method.

Regularity conditions

$\mathcal{A}_1$. there exists a constant C such that

$$|b(\alpha_0, x) - b(\alpha_0, y)| + |\sigma(\beta_0, x) - \sigma(\beta_0, y)| \leq C|x - y|;$$

$\mathcal{A}_2$. $\inf_{\beta,x} \det(\Sigma(\beta, x)) > 0$; with $\Sigma(\beta, x) = \sigma(\beta, x)\sigma(\beta, x)'$.

$\mathcal{A}_3$. the process $X_t, t \in [0, T]$, is ergodic for every θ with invariant probability measure μ_θ;

$\mathcal{A}_4$. if the coefficients $b(\alpha, x) = b(\alpha_0, x)$ and $\sigma(\beta, x) = \sigma(\beta_0, x)$ for all x (μ_{θ_0}-almost surely), then $\alpha = \alpha_0$ and $\beta = \beta_0$;

$\mathcal{A}_5$. for all $m \geq 0$ and for all $\theta \in \Theta$, $\sup_t E|X_t|^m < \infty$;

$\mathcal{A}_6$. for every $\theta \in \Theta$, the coefficients $b(\alpha, x)$ and $\sigma(\beta, x)$ are five times differentiable with respect to x and the derivatives are bounded by a polynomial function in x, uniformly in θ;

$\mathcal{A}_7$. the coefficients $b(\alpha, x)$ and $\sigma(\beta, x)$ and all their partial derivatives respect to x up to order 2 are three times differentiable with respect to θ for all x in the state space. All derivatives with respect to θ are bounded by a polynomial function in x, uniformly in θ.

$\mathcal{A}_1$ ensures the existence and uniqueness of a solution to the SDE for the value $\theta_0 = (\alpha_0, \beta_0)$ of $\theta \in \Theta$, while $\mathcal{A}_4$ is the identifiability condition. From now on we assume that the conditions $\mathcal{A}_1 - \mathcal{A}_7$ hold.

We can discretize the SDE

$$X_{t+\mathrm{d}t} - X_t = b(\alpha, X_t)\mathrm{d}t + \sigma(\beta, X_t)(W_{t+\mathrm{d}t} - W_t),$$

and the increments $X_{t+\mathrm{d}t} - X_t$ are then independent Gaussian random variables with mean $b(\alpha, X_t)\mathrm{d}t$ and variance-covariance matrix $\Sigma(\beta, x)\mathrm{d}t$. Therefore the transition density of the process can be written as a simple Gaussian density.

Quasi-likelihood function

$$\mathbb{H}_n(\mathbf{X}_n, \theta) = \frac{1}{2}\sum_{i=1}^{n}\left\{\log \det(\Sigma_{i-1}(\beta)) + \frac{1}{\Delta_n}(\Delta X_i - \Delta_n b_{i-1}(\alpha))'\Sigma_{i-1}^{-1}(\beta)(\Delta X_i - \Delta_n b_{i-1}(\alpha))\right\}$$

where $\Delta X_i = X_{t_i} - X_{t_{i-1}}$, $\Sigma_i(\beta) = \Sigma(\beta, X_{t_i})$ and $b_i(\alpha) = b(\alpha, X_{t_i})$.

This quasi-likelihood has been introduced by, e.g., Yoshida (1992), Genon-Catalot and Jacod (1993) and Kessler (1997) and used to obtain quasi-MLE estimators.

$\mathbb{H}_n$ plays the role of the negative log-likelihood for this model but the results of this part are such that $\mathbb{H}_n$ it can be replaced by any contrast function (see Masuda and Shimizu, 2016) or random field (in the sense of Yoshida, 2011)

The quasi-MLE $\tilde{\theta}_n$ for this model is the solution of the following problem

$$\tilde{\theta}_n = (\tilde{\alpha}_n, \tilde{\beta}_n)' = \arg\min_{\theta} \mathbb{H}_n(\mathbf{X}_n, \theta)$$

Optimality properties of the QMLE estimator

Consider the matrix (of rates of convergence)

$$\varphi(n) = \begin{pmatrix} \frac{1}{n\Delta_n}\mathbf{I}_p & 0 \\ 0 & \frac{1}{n}\mathbf{I}_q \end{pmatrix}$$

where $\mathbf{I}_p$ and $\mathbf{I}_q$ are respectively the identity matrix of order p and q. Let

$$\mathcal{I}(\theta) = \begin{pmatrix} \Gamma_\alpha = [\mathcal{I}_b^{kj}(\alpha)]_{k,j=1,\dots,p} & 0 \\ 0 & \Gamma_\beta = [\mathcal{I}_\sigma^{kj}(\beta)]_{k,j=1,\dots,q} \end{pmatrix}$$

where

$$\mathcal{I}_b^{kj}(\alpha) = \int \frac{1}{\sigma^2(\beta,x)} \frac{\partial b(\alpha,x)}{\partial \alpha_k} \frac{\partial b(\alpha,x)}{\partial \alpha_j} \mu_\theta(dx)\,,$$

$$\mathcal{I}_\sigma^{kj}(\beta) = 2\int \frac{1}{\sigma^2(\beta,x)} \frac{\partial \sigma(\beta,x)}{\partial \beta_k} \frac{\partial \sigma(\beta,x)}{\partial \beta_j} \mu_\theta(dx)\,.$$

Optimality properties of the QMLE estimator

Lemma 1 (see e.g., Kessler, 1997). *Let* $\Lambda_n(\theta) = \varphi(n)^{1/2}\ddot{\mathbb{H}}_n(\mathbf{X}_n, \theta)\varphi(n)^{1/2}$. *Under the conditions* $\mathcal{A}_1 - \mathcal{A}_7$, *and* $n\Delta_n \to \infty$, $n\Delta_n^2 \to 0$, $\Delta_n \to 0$ *as* $n \to \infty$, *the following two properties hold true*

i) *for* $\epsilon_n \to 0$, *as* $n \to \infty$, *then*

$$\Lambda_n(\theta_0) \xrightarrow{p} \mathcal{I}(\theta_0)$$

$$\sup_{||\theta|| \leq \epsilon_n} |\Lambda_n(\theta + \theta_0) - \Lambda_n(\theta_0)| = o_p(1)$$

ii) *for each* $\theta \in \Theta$, $\tilde{\theta}_n$ *is a consistent estimator of* θ *and asymptotically Gaussian, i.e.*

$$\varphi(n)^{-1/2}(\tilde{\theta}_n - \theta) \xrightarrow{d} N(0, \mathcal{I}(\theta)^{-1})$$

The classical adaptive Lasso objective function for the present model is then

$$\min_{\alpha,\beta} \left\{ H_n(\alpha,\beta) + \sum_{j=1}^{p} \lambda_{n,j}|\alpha_j| + \sum_{k=1}^{q} \gamma_{n,k}|\beta_k| \right\}$$

$\lambda_{n,j}$ and $\gamma_{n,k}$ are appropriate sequences representing an adaptive amount of shrinkage for each element of α and β.

Adaptiveness is essential to avoid the situation in which larger parameter are estimated with larger bias (up to missing consistency)

Unfortunately, the above is a non-linear optimization problem under l_1 constraints which might be numerically challenging to solve. Luckily, following Wang and Leng (2007), the minimization problem can be transformed into a quadratic minimization problem (under l_1 constraints) which is asymptotically equivalent to minimizing the original Lasso objective function.

Idea of Quadratic Approximation

By Taylor expansion of the original Lasoo objective function, for θ around $\tilde{\theta}_n$ (the QMLE estimator)

$$\begin{aligned}\mathbb{H}_n(\mathbf{X}_n,\theta) &= \mathbb{H}_n(\mathbf{X}_n,\tilde{\theta}_n) + (\theta-\tilde{\theta}_n)'\dot{\mathbb{H}}_n(\mathbf{X}_n,\tilde{\theta}_n) + \frac{1}{2}(\theta-\tilde{\theta}_n)'\ddot{\mathbb{H}}_n(\mathbf{X}_n,\tilde{\theta}_n)(\theta-\tilde{\theta}_n) \\ &\quad +o_p(1) \\ &= \mathbb{H}_n(\mathbf{X}_n,\tilde{\theta}_n) + \frac{1}{2}(\theta-\tilde{\theta}_n)'\ddot{\mathbb{H}}_n(\mathbf{X}_n,\tilde{\theta}_n)(\theta-\tilde{\theta}_n) + o_p(1)\end{aligned}$$

with $\dot{\mathbb{H}}_n$ and $\ddot{\mathbb{H}}_n$ the gradient and Hessian of $\mathbb{H}_n$ with respect to θ.

We define the adaptive Lasso estimator the solution to the quadratic problem under l_1 constraints

$$\hat{\theta}_n = (\hat{\alpha}_n, \hat{\beta}_n) = \arg\min_{\theta} \mathcal{F}(\theta).$$

with

$$\mathcal{F}(\theta) = (\theta - \tilde{\theta}_n)\ddot{\mathbb{H}}_n(\mathbf{X}_n, \tilde{\theta}_n)(\theta - \tilde{\theta}_n)' + \sum_{j=1}^{p} \lambda_{n,j}|\alpha_j| + \sum_{k=1}^{q} \gamma_{n,k}|\beta_k|$$

We will discuss adaptiveness later

Lasso cautions

- Adaptiveness: without adaptiveness, larger (true) parameters are estimated with more bias because of the penalization
- Speed of convergence: in diffusion models the speed of the parameters in the drift (α) and diffusion (β) are different (big difference w.r.t. i.i.d. models)
- Oracle property: the method should correctly estimate as zero the parameters which are truly zero

Before presenting formally the oracle property of the adaptive Lasso estimator, we will explain in which sense Lasso can be used as a model selector in this framework.

Model selection and causal inference with Lasso

Lasso as non-linear model selectorI

The CKLS model includes a special cases many famous models and is a nice example to apply Lasso to non-linear models. Indeed, fitting Lasso to real data on the CKLS model is a one-step model selection compared to the evalutaion of AIC for all the models below separately

Reference	Model	α	β	γ
Merton (1973)	$\mathrm{d}X_t = \alpha \mathrm{d}t + \sigma \mathrm{d}W_t$		0	0
Vasicek (1977)	$\mathrm{d}X_t = (\alpha + \beta X_t)\mathrm{d}t + \sigma \mathrm{d}W_t$			0
Cox, Ingersoll and Ross (1985)	$\mathrm{d}X_t = (\alpha + \beta X_t)\mathrm{d}t + \sigma \sqrt{X_t} \mathrm{d}W_t$			$1/2$
Dothan (1978)	$\mathrm{d}X_t = \sigma X_t \mathrm{d}W_t$	0	0	1
Geometric Brownian Motion	$\mathrm{d}X_t = \beta X_t \mathrm{d}t + \sigma X_t \mathrm{d}W_t$	0		1
Brennan and Schwartz (1980)	$\mathrm{d}X_t = (\alpha + \beta X_t)\mathrm{d}t + \sigma X_t \mathrm{d}W_t$			1
Cox, Ingersoll and Ross (1980)	$\mathrm{d}X_t = \sigma X_t^{3/2} \mathrm{d}W_t$	0	0	$3/2$
Constant Elasticity Variance	$\mathrm{d}X_t = \beta X_t \mathrm{d}t + \sigma X_t^{\gamma} \mathrm{d}W_t$	0		
CKLS (1992)	$\mathrm{d}X_t = (\alpha + \beta X_t)\mathrm{d}t + \sigma X_t^{\gamma} \mathrm{d}W_t$			

Let X_t be a multidimensional diffusion process solution to

$$\mathrm{d}X_t = \sum_{i=1}^{p} \alpha_i b(X_t)\mathrm{d}t + \sum_{j=1}^{p} \beta_j \sigma(X_t)\mathrm{d}W_t$$

where $b(\cdot)$ and $\sigma(\cdot)$ represent given statistical models. Then, the Lasso estimators of α_i and β_j allows for model selection as well.

Group Lasso idea can also be applied.

A typical usage of Lasso in model selection is the case **causation** (closely related to Granger causation). For example, in a model like this

$$\begin{pmatrix} \mathrm{d}X_t \\ \mathrm{d}Y_t \end{pmatrix} = \begin{pmatrix} \kappa_0 + \mu_{11}X_t + \mu_{12}Y_t \\ \kappa_1 + \mu_{21}X_t + \mu_{22}Y_t \end{pmatrix} \mathrm{d}t + \begin{pmatrix} \sigma_{11}X_t & \sigma_{12}Y_t \\ \sigma_{21}X_t & \sigma_{22}Y_t \end{pmatrix} \begin{pmatrix} \mathrm{d}W_t \\ \mathrm{d}B_t \end{pmatrix}$$

with initial condition $(X_0 = 1, Y_0 = 1)$ and $W_t, t \in [0, T]$, and $B_t, t \in [0, T]$, are two independent Brownian motions.

The case of $\mu_{12} = 0$, $\mu_{21} = 0$, $\sigma_{12} = 0$, $\sigma_{21} = 0$ is of practical interest because the systems becomes

$$\begin{aligned} \mathrm{d}X_t = &\quad \kappa_0 + \mu_{11}X_t + \sigma_{11}X_t\mathrm{d}W_t \\ \mathrm{d}Y_t = &\quad \kappa_1 + \mu_{22}Y_t + \sigma_{22}Y_t\mathrm{d}B_t \end{aligned}$$

Of course this can be generalized to affine diffusion in higher dimension without imposing a specific correlation structure like in the above simple example.

Adaptive Lasso properties

Without loss of generality, we assume that the true model, indicated by $\theta_0 = (\alpha_0, \beta_0)$, has parameters α_{0j} and β_{0k} equal to zero for $p_0 < j \leq p$ and $q_0 < k \leq q$, while $\alpha_{0j} \neq 0$ and $\beta_{0k} \neq 0$ for $1 \leq j \leq p_0$ and $1 \leq k \leq q_0$.

Denote by $\theta^* = (\alpha^*, \beta^*)'$ the vector corresponding to the nonzero parameters, where $\alpha^* = (\alpha_1, ..., \alpha_{p_0})'$ and $\beta^* = (\beta_1, ..., \beta_{q_0})'$, while $\theta^\circ = (\alpha^\circ, \beta^\circ)'$ is the vector corresponding to the zero parameters where $\alpha^\circ = (\alpha_{p_0+1}, ..., \alpha_p)'$ and $\beta^\circ = (\beta_{q_0+1}, ..., \beta_q)'$.

Therefore,

$$\text{TRUE :} \quad \theta_0 = (\alpha_0, \beta_0)' = (\alpha_0^*, \alpha_0^\circ, \beta_0^*, \beta_0^\circ)'$$

$$\text{Lasso :} \quad \hat{\theta}_n = (\hat{\alpha}_n^*, \hat{\alpha}_n^\circ, \hat{\beta}_n^*, \hat{\beta}_n^\circ)'$$

$$\text{MLE :} \quad \tilde{\theta}_n = (\tilde{\alpha}_n^*, \tilde{\alpha}_n^\circ, \tilde{\beta}_n^*, \tilde{\beta}_n^\circ)'$$

Intuition behind adaptiveness

$\mathcal{C}_1$. $\frac{\mu_n}{\sqrt{n\Delta_n}} \to 0$ and $\frac{\nu_n}{\sqrt{n}} \to 0$ where $\mu_n = \max\{\lambda_{n,j}, 1 \leq j \leq p_0\}$ and $\nu_n = \max\{\gamma_{n,k}, 1 \leq k \leq q_0\}$;

$\mathcal{C}_2$. $\frac{\kappa_n}{\sqrt{n\Delta_n}} \to \infty$ and $\frac{\omega_n}{\sqrt{n}} \to \infty$ where $\kappa_n = \min\{\lambda_{n,j}, j > p_0\}$ and $\omega_n = \min\{\gamma_{n,k}, k > q_0\}$.

Intuition behind adaptiveness

$\mathcal{C}_1$. $\frac{\mu_n}{\sqrt{n\Delta_n}} \to 0$ and $\frac{\nu_n}{\sqrt{n}} \to 0$ where $\mu_n = \max\{\lambda_{n,j}, 1 \leq j \leq p_0\}$ and $\nu_n = \max\{\gamma_{n,k}, 1 \leq k \leq q_0\}$;

$\mathcal{C}_2$. $\frac{\kappa_n}{\sqrt{n\Delta_n}} \to \infty$ and $\frac{\omega_n}{\sqrt{n}} \to \infty$ where $\kappa_n = \min\{\lambda_{n,j}, j > p_0\}$ and $\omega_n = \min\{\gamma_{n,k}, k > q_0\}$.

Assumption $\mathcal{C}_1$ implies that the maximal tuning coefficients μ_n and ν_n for the parameters α_j and β_k, with $1 \leq j \leq p_0$ and $1 \leq k \leq q_0$, tends to infinity slower than $\sqrt{n\Delta_n}$ and $\sqrt{n}$ respectively.

Intuition behind adaptiveness

$\mathcal{C}_1$. $\frac{\mu_n}{\sqrt{n\Delta_n}} \to 0$ and $\frac{\nu_n}{\sqrt{n}} \to 0$ where $\mu_n = \max\{\lambda_{n,j}, 1 \leq j \leq p_0\}$ and $\nu_n = \max\{\gamma_{n,k}, 1 \leq k \leq q_0\}$;

$\mathcal{C}_2$. $\frac{\kappa_n}{\sqrt{n\Delta_n}} \to \infty$ and $\frac{\omega_n}{\sqrt{n}} \to \infty$ where $\kappa_n = \min\{\lambda_{n,j}, j > p_0\}$ and $\omega_n = \min\{\gamma_{n,k}, k > q_0\}$.

Assumption $\mathcal{C}_1$ implies that the maximal tuning coefficients μ_n and ν_n for the parameters α_j and β_k, with $1 \leq j \leq p_0$ and $1 \leq k \leq q_0$, tends to infinity slower than $\sqrt{n\Delta_n}$ and $\sqrt{n}$ respectively.

Analogously, we observe that $\mathcal{C}_2$ means that that the minimal tuning coefficient for the parameter α_j and β_k, with $j > p_0$ and $k > q_0$, tends to infinity faster than $\sqrt{n\Delta_n}$ and $\sqrt{n}$ respectively.

Theorem 2. *Under conditions $\mathcal{A}_1 - \mathcal{A}_7$ and $\mathcal{C}_1$, one has that*

$$||\hat{\alpha}_n - \alpha_0|| = O_p\left((n\Delta_n)^{-1/2}\right) \quad \textit{and} \quad ||\hat{\beta}_n - \beta_0|| = O_p\left(n^{-1/2}\right).$$

Theorem 3. *Under conditions $\mathcal{A}_1 - \mathcal{A}_7$ and $\mathcal{C}_2$, we have that*

$$P(\hat{\alpha}_n^{\circ} = 0) \to 1 \qquad \textit{and} \qquad P(\hat{\beta}_n^{\circ} = 0) \to 1. \tag{2}$$

From Theorem 2, we can conclude that the estimator $\hat{\theta}_n$ is consistent.

Theorem 3 says us that all the estimates of the zero parameters are correctly set equal to zero with probability tending to 1

SKIP: Idea of the proof of Theorem 2

One as to prove the existence of a consistent local minimizer; this is implied by that fact that for an arbitrarily small $\varepsilon > 0$, there exists a sufficiently large constant C, such that

$$\lim_{n\to\infty} P\left\{\inf_{z\in\mathbb{R}^{p+q}:||z||=C} \mathcal{F}(\theta_0 + \varphi(n)^{1/2}z) > \mathcal{F}(\theta_0)\right\} > 1-\varepsilon, \tag{3}$$

with $z = (u, v)' = (u_1, ..., u_p, v_1, ..., v_q)'$. After some calculations, we obtain that

$$\mathcal{F}(\theta_0 + \varphi(n)^{1/2}z) - \mathcal{F}(\theta_0)$$

$$\geq z'\varphi(n)^{1/2}\ddot{\mathbb{H}}_n(\mathbf{X}_n, \tilde{\theta}_n)\varphi(n)^{1/2}z + 2z'\varphi(n)^{1/2}\ddot{\mathbb{H}}_n(\mathbf{X}_n, \tilde{\theta}_n)\varphi(n)^{1/2}\varphi(n)^{-1/2}(\theta_0 - \tilde{\theta}_n)$$

$$- \left[p_0\frac{\mu_n}{\sqrt{n\Delta_n}}||u|| + q_0\frac{\nu_n}{\sqrt{n}}||v||\right]$$

$$= \Xi_1 + \Xi_2 - \Xi_3$$

SKIP: Idea of the proof of Theorem 2

Let $\tau_{min}(A)$ is the minimal eigenvalue of A. Then, Lemma 1, being $||z|| = C$, Ξ_1 is uniformly larger than $\tau_{min}(\varphi(n)^{1/2}\ddot{\mathbb{H}}_n(\mathbf{X}_n, \tilde{\theta}_n)\varphi(n)^{1/2})C^2$ and

$$\tau_{min}(\varphi(n)^{1/2}\ddot{\mathbb{H}}_n(\mathbf{X}_n, \tilde{\theta}_n)\varphi(n)^{1/2})C^2 \xrightarrow{p} C^2\tau_{min}(\mathcal{I}(\theta_0)).$$

Moreover, Lemma 1 also implies that

$$||\varphi(n)^{1/2}\ddot{\mathbb{H}}_n(\mathbf{X}_n, \tilde{\theta}_n)\varphi(n)^{1/2}\varphi(n)^{-1/2}(\theta_0 - \tilde{\theta}_n)|| = O_p(1)$$

and then Ξ_2 is bounded and linearly dependent on C.

Therefore, for C sufficiently large, $\mathcal{F}(\theta_0 + \varphi(n)^{1/2}z) - \mathcal{F}(\theta_0)$ dominates $\Xi_1 + \Xi_2$ with arbitrarily large probability. Further, from the condition $\mathcal{C}_1$, one has that $\Xi_3 = o_p(1)$.

Strict convexity of $\mathcal{F}(\theta)$ implies that the consistent local minimum is the consistent global one.

Let $\mathcal{I}_0(\theta_0^*)$ the $(p_0 + q_0) \times (p_0 + q_0)$ submatrix of $\mathcal{I}(\theta)$ at point θ_0^* and introduce the following rate of convergence matrix

$$\varphi_0(n) = \begin{pmatrix} \frac{1}{n\Delta_n}\mathbf{I}_{p_0} & 0 \\ 0 & \frac{1}{n}\mathbf{I}_{q_0} \end{pmatrix}$$

Theorem 4 (Oracle property). *Under conditions $\mathcal{A}_1 - \mathcal{A}_7$ and $\mathcal{C}_1 - \mathcal{C}_2$, we have that*

$$\varphi_0(n)^{-\frac{1}{2}}(\hat{\theta}_n^* - \theta_0^*) \xrightarrow{d} N(0, \mathcal{I}_0^{-1}(\theta_0^*)) \tag{4}$$

where θ_0^* is the subset of non-zero true parameters.

How to choose the adaptive sequences

Clearly, the theoretical and practical implications of our method rely to the specification of the tuning parameter $\lambda_{n,j}$ and $\gamma_{n,k}$.

The tuning parameters should be chosen as is Zou (2006) in the following way

$$\lambda_{n,j} = \lambda_0 |\tilde{\alpha}_{n,j}|^{-\delta_1}, \qquad \gamma_{n,k} = \gamma_0 |\tilde{\beta}_{n,j}|^{-\delta_2} \tag{5}$$

where $\tilde{\alpha}_{n,j}$ and $\tilde{\beta}_{n,k}$ are the unpenalized QML estimator of α_j and β_k respectively, $\delta_1, \delta_2 > 1$. The asymptotic results hold under the additional conditions

$$\frac{\lambda_0}{\sqrt{n\Delta_n}} \to 0, \quad (n\Delta_n)^{\frac{\delta_1 - 1}{2}} \lambda_0 \to \infty$$

and

$$\frac{\gamma_0}{\sqrt{n}} \to 0, \quad n^{\frac{\delta_2 - 1}{2}} \gamma_0 \to \infty$$

as $n \to \infty$.

General regularized estimation

Masuda and Shimizu (2016) generalized the above Lasso approach to the wider class of regularized estimation methods. In their setup, they propose to solve this problem

$$\min_{\Theta} \mathbb{H}_n(\theta)$$

where

$$\mathbb{H}_n = M_n(\alpha, \beta) + R_n^b(\alpha) + R_n^\sigma(\beta)$$

M_n is any contrast function and $R_n^b(\alpha)$ and $R_n^\sigma(\beta)$ are called regularization sequences and generalize the adaptive sequences of previous part of this talk.

In their proof, there is no need of convexity as they adopt the random field approach of Yoshida (2011): **LAQ + PLDI** (locally asymptotic quadratic structure) + (polynomial large deviation inequality).

They can also establish the converge of moment of the estimators (useful in prediction problems), oracle properties, tail probability estimate.

What's next?

The **Elastic Net** method, is a regularization method proposed by Zou and Hastie (2005) in the i.i.d. case $Y = X\theta + \epsilon$, $\epsilon \sim N(0, \sigma^2)$, to solve some problems in the Lasso estimation procedure in the following circumstances

- the number of regressors p is larger than n and **grows** with n
- the regressors are correlated

and later generalized to the case of diverging parameters (Zou and Zhang, 2009).

The ENet estimator is the solution to

$$\left(1 + \frac{\lambda_2}{n}\right) \arg\min_{\theta} \left\{ ||(Y - X\theta)||_2^2 + \lambda_2 \sum_{j=1}^{p} |\theta_j|^2 + \lambda_1 \sum_{j=1}^{p} |\theta_j| \right\}$$

In orthogonal design this asymptotically converges to Lasso, but with correlated regressors X the L_2 penalty increase the accuracy of prediction.

The AdaEnet estimator is a combination of ENet and adaptive Lasso. And works this way

- **Step 1:** let $\hat{\theta}^{ENet}$ be the solution of the previous ENet problem and compute

$$\hat{w}_j = \frac{1}{\left|\hat{\theta}_j^{ENet} + \frac{1}{n}\right|^{\gamma}}, \quad j = 1, \ldots, p, \quad \gamma > 0$$

- **Step 2:**

$$\hat{\theta}^{AdaENet} = \left(1 + \frac{\lambda_2}{n}\right) \arg\min_{\theta} \left\{ ||(Y - X\theta)||_2^2 + \lambda_2 \sum_{j=1}^{p} |\theta_j|^2 + \lambda_1^* \sum_{j=1}^{p} \hat{w}_j |\theta_j| \right\}$$

Under the assumption that $\lim_{n\to\infty} \frac{\log(p)}{\log(n)} = \nu, 0 \leq \nu < 1$ choosing $\gamma > \frac{2\nu}{1-\nu}$ guarantees the oracle property of AdaENet.

Further, if

$$b \leq \lambda_{min}\left(\frac{1}{n}X'X\right) \leq \lambda_{max}\left(\frac{1}{n}X'X\right) \leq B$$

(with $\lambda_{min}(A)$ and $\lambda_{max}(A)$, the smallest and largest eigenvalue of pos. def. A)

$$E\left(||\hat{\theta}(AdaENet) - \theta^*||_2^2\right) \leq 4\frac{\lambda_2^2||\theta^*||_2^2 + Bpn\sigma^2 + \lambda_1^2 E\left(\sum_{j=1}^p \hat{w}_j^2\right)}{(bn+\lambda_2)^2}$$

This **non-asymptotic** bound gives root-(n/p)-consistency of the AdaENet estimator. Asymptotic normality and oracle properties can be attained as well.

Further, if

$$b \leq \lambda_{min}\left(\frac{1}{n}X'X\right) \leq \lambda_{max}\left(\frac{1}{n}X'X\right) \leq B$$

(with $\lambda_{min}(A)$ and $\lambda_{max}(A)$, the smallest and largest eigenvalue of pos. def. A)

$$E\left(||\hat{\theta}(AdaENet) - \theta^*||_2^2\right) \leq 4\frac{\lambda_2^2||\theta^*||_2^2 + Bpn\sigma^2 + \lambda_1^2 E\left(\sum_{j=1}^p \hat{w}_j^2\right)}{(bn+\lambda_2)^2}$$

This **non-asymptotic** bound gives root-(n/p)-consistency of the AdaENet estimator. Asymptotic normality and oracle properties can be attained as well.

Application to dynamical systems with small noise, SDE's with jumps and jump processes: on its way, joint work with A. De Gregorio and N. Yoshida.

THANKS!

References

References

- Azencott, R. (1982) Formule de taylor stochastique et développement asymptotique d'intégrales de feynmann, *Séminaire de Probabilités XVI; Supplément: Géometrie Différentielle Stochastique. Lecture Notes In Math*, **921**, 237–285.
- Dacunha-Castelle, D., Florens-Zmirou, D. (1986) Estimation of the coefficients of a diffusion from discrete observations, *Stochastics*, **19**, 263–284.
- De Gregorio, A., Iacus, S. M. (2012) Adaptive lasso-type estimation for multivariate diffusion processes, *Econometric Theory*, **28**, 838–860.
- Fan, J. (1997). Comments on "Wavelets in Statistics: A Review" by A. Antoniadis". *Journal of the Italian Statistical Association*, **6**, 131-138.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *JASA*, **96**, 1348-1360.
- Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. (2004) Least angle regression, *The Annals of Statistics*, **32**, 407–489.
- Freidlin, M. I., Wentzell, A. D. (1998) *Random perturbations of dynamical systems. 2nd. ed.*, Springer-Verlag, New York.
- Geyer, C.J. (1994) On the asymptotics of constrained M-estimation, *Annals of Statistics*, **22**, 1993–2010.
- Geyer, C.J. (1996) On the asymptotics of convex stochastic optimization, available at `http://www.stat.umn.edu/PAPERS/preprints/convex.ps`.
- Iacus, S. M. (2000) Semiparametric estimation of the state of a dynamical system with small noise, *Stat. Infer. for Stoch. Proc.*, **3**, 277–288.
- Hastie, T., Tibshirani, R. and Friedman, J. (2008), *The Elements of Statistical Learning*, Springer Verlag, New York.
- Iacus, S. M., Kutoyants, Y. (2001) Semiparametric hypotheses testing for dynamical systems with small noise, *Mathematical Methods of Statistics*, **10**(1), 105–120.
- Kato, K. (2009) Asymptotics for argmin processes: Convexity arguments, *Journal of Multivariate Analysis*, **100**(8), 1816–1829.
- Kim, J., Pollard, D. (1990) Cube root asymptotics, *Annals of Statistics*, **18**, 191–219.
- Knight, K., Fu, W. (2000) Asymptotics for lasso-type estimators, *Annals of Statistics*, **28**, 1536–1378.
- Kunitomo, N., Takahashi, A. (2001) The asymptotic expansion approach to the valuation of interest rate contingent claims, *Mathematical Finance*, **11**(1), 117–151.
- Kutoyants, Y. (1984) *Parameter estimation for stochastic processes*, Heldermann, Berlin.
- Kutoyants, Y. (1991) Minimum distance parameter estimation for diffusion type observations, *C.R. Acad. Paris*, **312**, Sér. I, 637–642.
- Kutoyants, Y. (1994) *Identification of Dynamical Systems with Small Noise*, Kluwer Academic Publishers, Dordrecht, The Netherlands.
- Kutoyants, Y., Philibossian, P. (1994) On minimum L_1-norm estimates of the parameter of Ornstein-Uhlenbeck process, *Statistics and Probability Letters*, **20**(2), 117–123.
- Masuda, H., Shimizu, Y. (2016) Moment convergence in regularized estimation under multiple and mixed-rates asymptotics, `https://arxiv.org/abs/1406.6751`
- Takahashi, A., Yoshida, N. (2004) An asymptotic expansion scheme for optimal investment problems, *Stat. Inference Stoch. Process.*, **7**, 153–188.
- Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso, *J. Roy. Statist. Soc. Ser. B*, **58**, 267–288.
- Uchida, M. and Yoshida, N. (2004a) Asymptotic expansion for small diffusions applied to option pricing, *Statist. Infer. Stochast. Process*, **7**, 189–223.
- Uchida, M. and Yoshida, N. (2004b) Information criteria for small diffusions via the theory of Malliavin-Watanabe, *Statist. Infer. Stoch. Process*, **7**, 35–67.

References

- Park, T., and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, **103**(482), 681-686.
- Pollard, D. (1991) Asymptotics for least absolute deviation regression estimators, *Econometric Theory*, **7**, 186–199.
- Yoshida, N. (1992a) Asymptotic expansion for statistics related to small diffusions, *Journal of the Japan Statistical Society*, **22**, 139–159.
- Yoshida, N. (1992b) Asymp. expansion of maximum likelihood estimators for small diffusions via the theory of Malliavin-Watanabe, *P. Theory Rel.. Fields*, **92**, 275–311.
- Yoshida, N. (2003) Conditional expansions and their applications, *Stochastic Process. Appl.*, **107**, 53–81.
- Yoshida, N. (2011) Polynomial type large deviation inequalities and quasi-likelihood analysis for stochastic differential equations, *Ann. Inst. Statist. Math.*, **63**(3), 431–479.
- Yuan, M. and Lin, Y. (2006), Model Selection and Estimation in Regression with Grouped Variables, *JRSS, B*, **68**, 49-67.
- Zou, H. (2006) The adaptive LASSO and its Oracle properties, *J. Amer. Stat. Assoc.*, **101**(476), 1418-1429.
- Zou, H., Hastie, T. (2005) Regularization and variable selection via the elastic net, *J. R. Statist. Soc. B*, **67**(Part 2), 301-302.
- Zou, H., Zhang, E. (2009) On the adaptive elastic-net with a diverging number of parameters, *The Annals of Statistics*, **37**(4), 1733-1751.