Estimation, shrinkage and penalization

Sofia Olhede

November 16, 2020

1. How and why to shrink?

2. Generalised Linear Model

Shrinkage

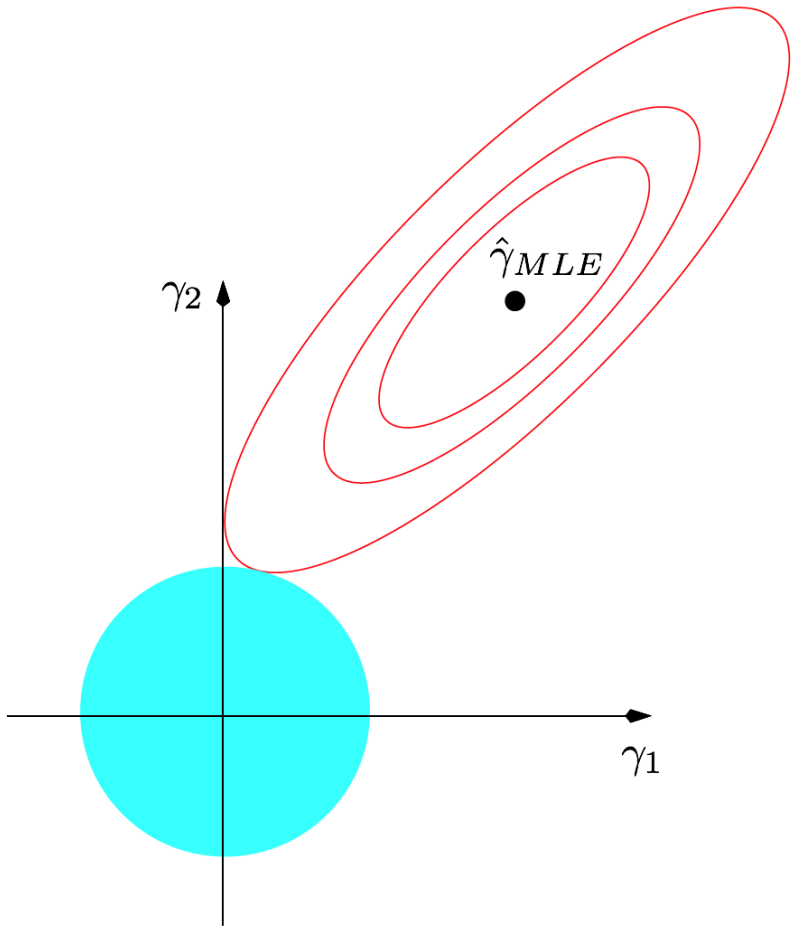


Figure: L^2 Shrinkage [Ridge Regression]

Shrinkage II

Theorem

Let $\boldsymbol{Z}_{n\times q}$ be a matrix of rank $r \leq q$ with centred column vectors of unit norm. Given $\lambda > 0$, the unique minimiser of

$$Q(\alpha, \boldsymbol{\xi}) = \|\boldsymbol{Y} - \alpha\mathbf{1} - \boldsymbol{Z}\boldsymbol{\xi}\|_2^2 + \lambda\|\boldsymbol{\xi}\|_2^2$$

is

$$(\hat{\beta}_0\,, \hat{\boldsymbol{\gamma}}) = (\bar{Y}\,, (\boldsymbol{Z}^\top\boldsymbol{Z} + \lambda\boldsymbol{I})^{-1}\boldsymbol{Z}^\top\boldsymbol{Y}).$$

Proof.

Write

$$\boldsymbol{Y} = \underbrace{(\boldsymbol{Y} - \bar{Y}\mathbf{1})}_{=\boldsymbol{Y}^*\in\mathcal{M}^\perp(\mathbf{1})} + \underbrace{\bar{Y}\mathbf{1}}_{\in\mathcal{M}(\mathbf{1})}$$

Note also that by assumption $\mathbf{1} \in \mathcal{M}^\perp(\boldsymbol{Z})$. Therefore by Pythagoras' theorem

$$\|\boldsymbol{Y} - \hat{\beta}_0\mathbf{1} - \boldsymbol{Z}\hat{\boldsymbol{\gamma}}\|_2^2 = \|\underbrace{(\bar{Y} - \hat{\beta}_0)\mathbf{1}}_{\in\mathcal{M}(\mathbf{1})} + \underbrace{(\boldsymbol{Y}^* - \boldsymbol{Z}\hat{\boldsymbol{\gamma}})}_{\in\mathcal{M}(\boldsymbol{Z})}\|_2^2 = \|(\bar{Y} - \hat{\beta}_0)\mathbf{1}\|_2^2 + \|(\boldsymbol{Y}^* - \boldsymbol{Z}\hat{\boldsymbol{\gamma}})\|_2^2.$$

Shrinkage III

Therefore, $\min_{\alpha,\boldsymbol{\xi}} Q(\alpha,\boldsymbol{\xi}) = \min_{\alpha} ||(\bar{Y}-\alpha)\mathbf{1}||_2^2 + \min_{\boldsymbol{\xi}} \left\{ ||(\boldsymbol{Y}^* - \boldsymbol{Z}\boldsymbol{\xi})||_2^2 + \lambda||\boldsymbol{\xi}||_2^2 \right\}$

Clearly, $\arg\min_\alpha ||(\bar{Y}-\alpha)\mathbf{1}||_2^2 = \bar{Y}$ while the second component can be written

$$\min_{\boldsymbol{\xi}\in\mathbb{R}^q} \left\| \begin{pmatrix} \boldsymbol{Z} \\ \sqrt{\lambda} I_{q\times q} \end{pmatrix} \boldsymbol{\xi} - \begin{pmatrix} \boldsymbol{Y}^* \\ \mathbf{0}_{q\times 1} \end{pmatrix} \right\|_2^2$$

using block notation. This is the usual least squares problem with solution

$$\left[(\boldsymbol{Z}^\top, \sqrt{\lambda}\boldsymbol{I}_{q\times q}) \begin{pmatrix} \boldsymbol{Z} \\ \sqrt{\lambda}\boldsymbol{I}_{q\times q} \end{pmatrix} \right]^{-1} (\boldsymbol{Z}^\top, \sqrt{\lambda}\boldsymbol{I}_{q\times q}) \begin{pmatrix} \boldsymbol{Y}^* \\ \mathbf{0}_{q\times 1} \end{pmatrix} = (\boldsymbol{Z}^\top\boldsymbol{Z} + \lambda\boldsymbol{I})^{-1}\boldsymbol{Z}^\top\boldsymbol{Y}^*$$

Note that $\boldsymbol{Z}^\top\boldsymbol{Z} + \lambda\boldsymbol{I}$ is indeed invertible. Writing $\boldsymbol{Z}^\top\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, we have

$$\boldsymbol{Z}^\top\boldsymbol{Z} + \lambda\boldsymbol{I} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top + \boldsymbol{U}(\lambda\boldsymbol{I}_{q\times q})\boldsymbol{U}^\top = \boldsymbol{U}(\boldsymbol{\Lambda} + \lambda\boldsymbol{I}_{q\times q})\boldsymbol{U}^\top$$

and $\boldsymbol{\Lambda} = \text{diag}\{\underbrace{\lambda_1,\ldots,\lambda_r}_{>0},\underbrace{\lambda_{r+1},\ldots,\lambda_q}_{=0}\}$ ($\boldsymbol{Z}^\top\boldsymbol{Z} \succeq 0$ & $\text{rank}(\boldsymbol{Z}^\top\boldsymbol{Z}) = \text{rank}(\boldsymbol{Z})$).

To complete the proof, observe that $\boldsymbol{Z}^\top\boldsymbol{Y}^* = \boldsymbol{Z}^\top\boldsymbol{Y} - \bar{Y}\boldsymbol{Z}^\top\mathbf{1} = \boldsymbol{Z}^\top\boldsymbol{Y}$. □

Shrinkage IV

Note that if the SVD of $\boldsymbol{Z}$ is $\boldsymbol{Z} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, last steps of previous proof may be used to show that

$$\hat{\boldsymbol{\gamma}} = \sum_{j=1}^{q} \frac{\lambda_j}{\lambda_j^2 + \lambda} (\boldsymbol{V}_j^\top \boldsymbol{Y}) \boldsymbol{U}_j,$$

where the $\boldsymbol{V}_j$s and $\boldsymbol{U}_j$s are the columns of $\boldsymbol{V}$ and $\boldsymbol{U}$, respectively.

Compare this to the ordinary least squares solution, when $\lambda = 0$:

$$\hat{\boldsymbol{\gamma}} = \sum_{j=1}^{q} \frac{1}{\lambda_j} (\boldsymbol{V}_j^\top \boldsymbol{Y}) \boldsymbol{U}_j,$$

which is not even defined if $\boldsymbol{Z}$ is of reduced rank.

Role of λ is to reduce the size of $1/\lambda_j$ when λ_j becomes very small.

Shrinkage V

Proposition

Let $\hat{\gamma}$ be the ridge regression estimator of γ. Then

$$\text{bias}(\hat{\gamma}, \gamma) = -\lambda \left(\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I}_{q \times q} \right)^{-1} \gamma$$

and

$$\text{cov}(\hat{\gamma}) = \sigma^2 (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1} \boldsymbol{Z}^\top \boldsymbol{Z} (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1}.$$

Proof.

Since $\mathbb{E}(\hat{\gamma}) = (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1} \boldsymbol{Z}^\top \mathbb{E}(\boldsymbol{Y}) = (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1} \boldsymbol{Z}^\top \boldsymbol{Z} \gamma$, the bias is

$$\begin{aligned}
\text{bias}(\hat{\gamma}, \gamma) &= \mathbb{E}(\hat{\gamma}) - \gamma = \{(\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1} \boldsymbol{Z}^\top \boldsymbol{Z} - \boldsymbol{I}\}\gamma \\
&= \left\{ \left(\frac{1}{\lambda} \boldsymbol{Z}^\top \boldsymbol{Z} + \boldsymbol{I} \right)^{-1} \left(\frac{1}{\lambda} \boldsymbol{Z}^\top \boldsymbol{Z} + \boldsymbol{I} - \boldsymbol{I} \right) - \boldsymbol{I} \right\} \gamma \\
&= \left\{ \boldsymbol{I} - \left(\frac{1}{\lambda} \boldsymbol{Z}^\top \boldsymbol{Z} + \boldsymbol{I} \right)^{-1} - \boldsymbol{I} \right\} \gamma = - \left(\frac{1}{\lambda} \boldsymbol{Z}^\top \boldsymbol{Z} + \boldsymbol{I} \right)^{-1} \gamma.
\end{aligned}$$

The covariance term is straightforward. □

Shrinkage VI

Corollary (Domination over Least Squares)

Assume that $\mathrm{rank}(\boldsymbol{Z}_{n\times q}) = q$ *and let*

$$\tilde{\gamma} = (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{Y} \qquad \& \qquad \hat{\gamma}_\lambda = (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1}\boldsymbol{Z}^\top \boldsymbol{Y}$$

be the least squares estimator and ridge estimator, respectively. Then,

$$\mathbb{E}\left\{(\tilde{\gamma}-\gamma)(\tilde{\gamma}-\gamma)^\top\right\} - \mathbb{E}\left\{(\hat{\gamma}_\lambda-\gamma)(\hat{\gamma}_\lambda-\gamma)^\top\right\} \succeq 0$$

for all $\lambda \in (0, 2\sigma^2/||\gamma||^2)$.

Ridge estimator uniformly better than least squares! How can this be?
(What happened to Gauss-Markov?)

- Gauss-Markov only covers unbiased estimators – but ridge estimator biased.
- A bit of bias can improve the MSE by reducing variance!
- Also, there is a catch! The "right" range for λ depends on unknowns.
- Choosing a good λ is all about balancing bias and variance.

Shrinkage VII

Proof.

From our bias/variance calculations on the ridge estimator, we have

$$\mathbb{E}\left\{(\tilde{\gamma}-\gamma)(\tilde{\gamma}-\gamma)^\top\right\} - \mathbb{E}\left\{(\hat{\gamma}_\lambda-\gamma)(\hat{\gamma}_\lambda-\gamma)^\top\right\} =$$

$$= \sigma^2(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1} - \sigma^2(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})^{-1}\boldsymbol{Z}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})^{-1} + \lambda\left(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I}_{q\times q}\right)^{-1}\gamma$$

$$= \lambda(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})^{-1}\left(\sigma^2(2\boldsymbol{I}+\lambda(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}) - \lambda\gamma\gamma^\top\right)(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})^{-1}.$$

This last term can be made positive definite if

$$2\sigma^2\boldsymbol{I} + \sigma^2\lambda(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1} - \lambda\gamma\gamma^\top \succeq 0.$$

Noting that we can always write

$$\boldsymbol{I} = \frac{\gamma\gamma^\top}{||\gamma||^2} + \sum_{j=1}^{q-1}\boldsymbol{\theta}_j\boldsymbol{\theta}_j^\top$$

for $\{\gamma/||\gamma||, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{q-1}\}$ an orthonormal basis of $\mathbb{R}^q$ we see that $\lambda \in (0, 2\sigma^2/||\gamma||^2)$ suffices for positive definiteness to hold true.

□

Shrinkage VIII

Role of λ: Regulates Bias–Variance tradeoff

- $\lambda \uparrow$ decreases variance (collinearity) but increases bias
- $\lambda \downarrow$ decreases bias but variance inflated if collinearity exists

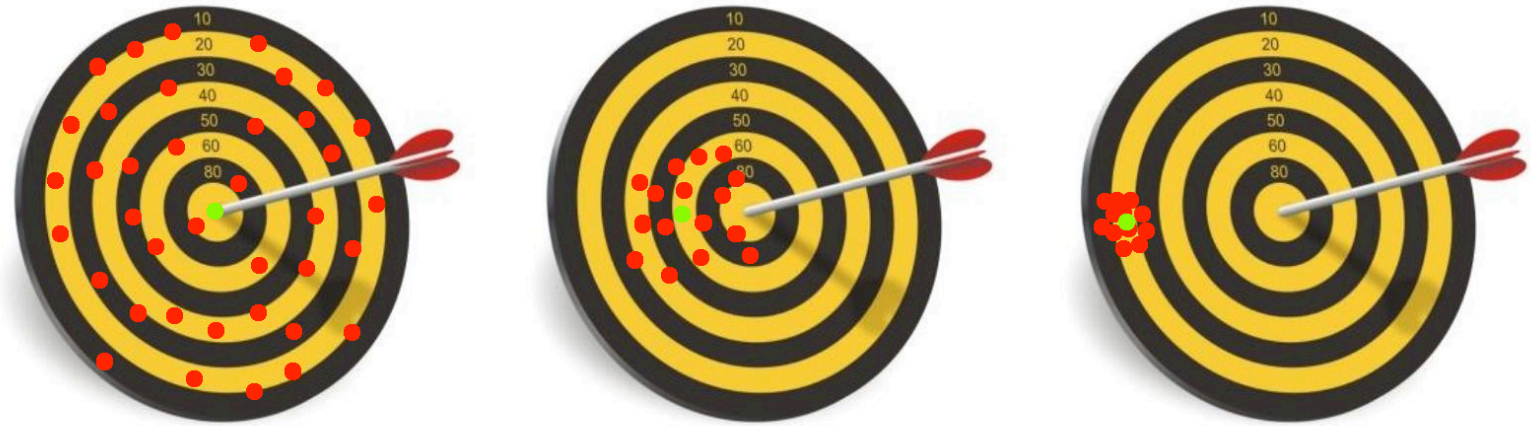
Recall:

$$\mathbb{E}||\hat{\gamma} - \gamma||^2 = \underbrace{||\mathbb{E}\hat{\gamma} - \gamma||^2}_{\text{bias}^2} + \underbrace{\mathbb{E}||\hat{\gamma} - \mathbb{E}\hat{\gamma}||^2}_{\text{variance}=trace[cov(\hat{\gamma})]} + \underbrace{2(\mathbb{E}\hat{\gamma} - \gamma)^\top \mathbb{E}[\hat{\gamma} - \mathbb{E}\hat{\gamma}]}_{=0}$$

Writing $\boldsymbol{Z}^\top \boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ $\text{trace}\{\text{cov}(\hat{\gamma})\} = \sum_{j=1}^{q} \frac{\lambda_i}{\lambda_i^2+\lambda}\sigma^2$

So choose λ so as to optimally to balance bias/variance

Use cross validation!



Shrinkage IX

Motivated from Ridge Regression formulation can consider:

$$\text{min!} \qquad \|\boldsymbol{Y} - \beta_0\mathbf{1} - \boldsymbol{Z}\boldsymbol{\gamma}\|_2^2 \quad \text{subject to} \quad \sum_{j=1}^{p-1}|\gamma_j| = \|\boldsymbol{\gamma}\|_1 \leq r(\lambda)$$

$$\Longleftrightarrow$$

$$\text{min!} \qquad \|\boldsymbol{Y} - \beta_0\mathbf{1} - \boldsymbol{Z}\boldsymbol{\gamma}\|_2^2 + \lambda\|\boldsymbol{\gamma}\|_1.$$

Shrinks coefficient *size* by different version of *magnitude*.

- Resulting estimator non–linear in $\boldsymbol{Y}$
- No explicit form available (unless $\boldsymbol{Z}^\top\boldsymbol{Z} = \boldsymbol{I}$), needs quadratic programming algorithm

Why choose a different type of norm?
L^1 penalty (almost) produces a "continuous" model selection!

Shrinkage X

When the explanatory variables are orthogonal (i.e. $\boldsymbol{Z}^\top \boldsymbol{Z} = \boldsymbol{I}$), then the LASSO[7] exactly performs model selection via thresholding:

Theorem (Orthogonal Design $\leftrightarrow$ Model Selection)

Consider the linear model

$$\underset{n\times 1}{\boldsymbol{Y}} = \beta_0 \underset{1\times 1}{\mathbf{1}}_{n\times 1} + \underset{n\times(p-1)}{\boldsymbol{Z}} \underset{(p-1)\times 1}{\gamma} + \underset{n\times 1}{\varepsilon}$$

where $\boldsymbol{Z}^\top \mathbf{1} = 0$ *and* $\boldsymbol{Z}^\top \boldsymbol{Z} = I$. *Let* $\hat{\gamma}$ *be the least squares estimator of* γ,

$$\hat{\gamma} = (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \boldsymbol{Y} = \boldsymbol{Z}^\top \boldsymbol{Y}.$$

Then, the unique solution to the LASSO problem

$$\min_{\beta_0 \in \mathbb{R}, \gamma \in \mathbb{R}^{p-1}} \left\{ || \boldsymbol{Y} - \beta_0 \mathbf{1} - \boldsymbol{Z}\gamma ||_2^2 + \lambda ||\gamma||_1 \right\}$$

is given by $(\hat{\beta}_0, \check{\gamma}) = (\beta_0, \check{\gamma}_1, \ldots, \check{\gamma}_{p-1})$, *defined as*

$$\hat{\beta}_0 = \bar{Y} \qquad \& \qquad \check{\gamma}_i = \operatorname{sgn}(\hat{\gamma}_i) \left(|\hat{\gamma}_i| - \frac{\lambda}{2} \right)_+, \quad i = 1, ..., p-1.$$

Shrinkage XI

Proof $(*)$.

Note that since $Z^\top \mathbf{1} = 0$ and since β_0 does not appear in the L^1 penalty, we have

$$\hat{\beta}_0 = (\mathbf{1}^\top \mathbf{1})^{-1} \mathbf{1} Y = \bar{Y}.$$

Thus, the LASSO problem reduces to

$$\min_{\beta_0 \in \mathbb{R}, \gamma \in \mathbb{R}^{p-1}} \left\{ || \boldsymbol{Y} - \beta_0 \mathbf{1} - Z\gamma ||_2^2 + \lambda ||\gamma||_1 \right\} = \min_{\gamma \in \mathbb{R}^{p-1}} \left\{ ||\boldsymbol{u} - \boldsymbol{Z}\gamma||_2^2 + \lambda ||\gamma||_1 \right\}.$$

where $\boldsymbol{u} = \boldsymbol{Y} - \bar{Y}\mathbf{1}$ for tidiness. Expanding $||\boldsymbol{u} - \boldsymbol{Z}\gamma||_2^2$ gives

$$\boldsymbol{u}^\top \boldsymbol{u} - 2\boldsymbol{u}^\top \boldsymbol{Z}\gamma + \gamma^\top \underbrace{(\boldsymbol{Z}^\top \boldsymbol{Z})}_{=I} \gamma = \boldsymbol{u}^\top \boldsymbol{u} - 2\underbrace{\boldsymbol{Y}^\top \boldsymbol{Z}}_{=\hat{\gamma}^\top} \gamma + 2\bar{Y} \underbrace{\mathbf{1}^\top \boldsymbol{Z}}_{=0} \gamma + \gamma^\top \gamma$$

Since $\boldsymbol{u}^\top \boldsymbol{u}$ does not depend on γ, we see that the LASSO objective function is

$$-2\hat{\gamma}^\top \gamma + ||\gamma||_2^2 + \lambda ||\gamma||_1.$$

Clearly, this has the same minimizer if multiplied across by $1/2$, i.e.

$$-\hat{\gamma}^\top \gamma + \tfrac{1}{2}||\gamma||_2^2 + \tfrac{1}{2}\lambda ||\gamma||_1 = \textstyle\sum_{i=1}^{p-1} \left(-\hat{\gamma}_i \gamma_i + \tfrac{1}{2}\gamma_i^2 + \tfrac{\lambda}{2}|\gamma_i| \right).$$

Shrinkage XII

Notice that we now have a sum of $p-1$ objective functions, each depending only on one γ_i. We can thus optimise each separately. That is, for any given $i \leq p-1$, we must minimise

$$-\hat{\gamma}_i\gamma_i + \frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}|\gamma_i|.$$

We distinguish 3 cases:

1. Case $\hat{\gamma}_i = 0$. In this case, the objective function becomes $\frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}|\gamma_i|$ and it is clear that it is minimised when $\gamma_i = 0$. So in this case $\check{\gamma}_i = 0$.
2. Case $\hat{\gamma}_i > 0$. In this case, the objective function $-\hat{\gamma}_i\gamma_i + \frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}|\gamma_i|$ is minimised somewhere in the range $\gamma_i \in [0, \infty)$ because the term $-\hat{\gamma}_i\gamma_i$ is negative there (and all other terms are positive). But when $\gamma_i \geq 0$, the objective function becomes

$$-\hat{\gamma}_i\gamma_i + \frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}\gamma_i = \left(\frac{\lambda}{2} - \hat{\gamma}_i\right)\gamma_i + \frac{1}{2}\gamma_i^2.$$

If $\frac{\lambda}{2} - \hat{\gamma}_i \geq 0$, then the minimum over $\gamma_i \in [0, \infty)$ is clearly at $\gamma_i = 0$. Otherwise, when $\frac{\lambda}{2} - \hat{\gamma}_i < 0$, we differentiate and find the minimum at $\gamma_i = \hat{\gamma}_i - \lambda/2 > 0$. In summary, $\check{\gamma}_i = (\hat{\gamma}_i - \lambda/2)_+ = \text{sgn}(\hat{\gamma}_i)(|\hat{\gamma}_i| - \lambda/2)_+$

Shrinkage XIII

3. Case $\hat{\gamma}_i < 0$. In this case, the objective function $-\hat{\gamma}_i\gamma_i + \frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}|\gamma_i|$ is minimised somewhere in the range $\gamma_i \in (-\infty, 0]$ because the term $-\hat{\gamma}_i\gamma_i$ is negative there (and all other terms are positive). But when $\gamma_i \leq 0$, the objective function becomes

$$-\hat{\gamma}_i\gamma_i + \frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}(-\gamma_i) = \left(\frac{\lambda}{2} + \hat{\gamma}_i\right)(-\gamma_i) + \frac{1}{2}\gamma_i^2 = \left(\frac{\lambda}{2} - |\hat{\gamma}_i|\right)(-\gamma_i) + \frac{1}{2}\gamma_i^2.$$

If $\frac{\lambda}{2} - |\hat{\gamma}_i| \geq 0$, then the minimum over $\gamma_i \in (-\infty, 0]$ is clearly at $\gamma_i = 0$, since $-\gamma_i$ ranges over $[0, \infty)$. Otherwise, when $\frac{\lambda}{2} - |\hat{\gamma}_i| < 0$, we differentiate and find the minimum at $\gamma_i = -|\hat{\gamma}_i| + \lambda/2 < 0$, which we may re-write as:

$$-|\hat{\gamma}_i| + \lambda/2 = -\left(|\hat{\gamma}_i| - \lambda/2\right) = \mathrm{sgn}(\hat{\gamma}_i)\left(|\hat{\gamma}_i| - \lambda/2\right).$$

In summary, $\check{\gamma}_i = \mathrm{sgn}(\hat{\gamma}_i)(|\hat{\gamma}_i| - \lambda/2)_+$.

The proof is now complete, as we can see that all three cases yield

$$\check{\gamma}_i = \mathrm{sgn}(\hat{\gamma}_i)\left(|\hat{\gamma}_i| - \frac{\lambda}{2}\right)_+, \quad i = 1, ..., p-1.$$

□

Shrinkage XIV

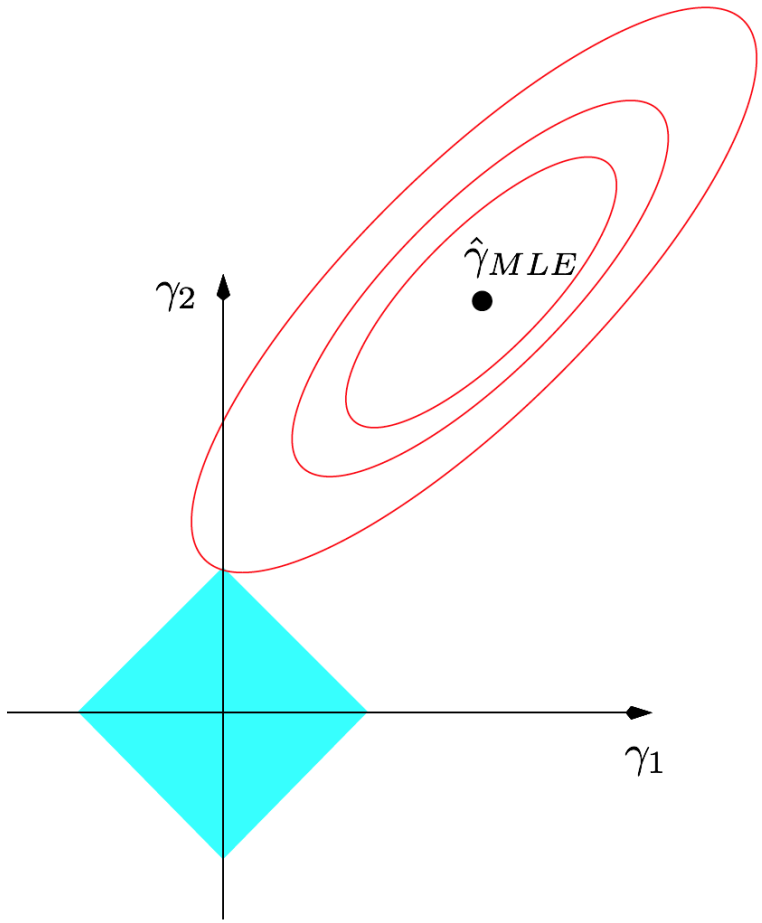


Figure: L^1 Shrinkage (the LASSO)

Shrinkage XV

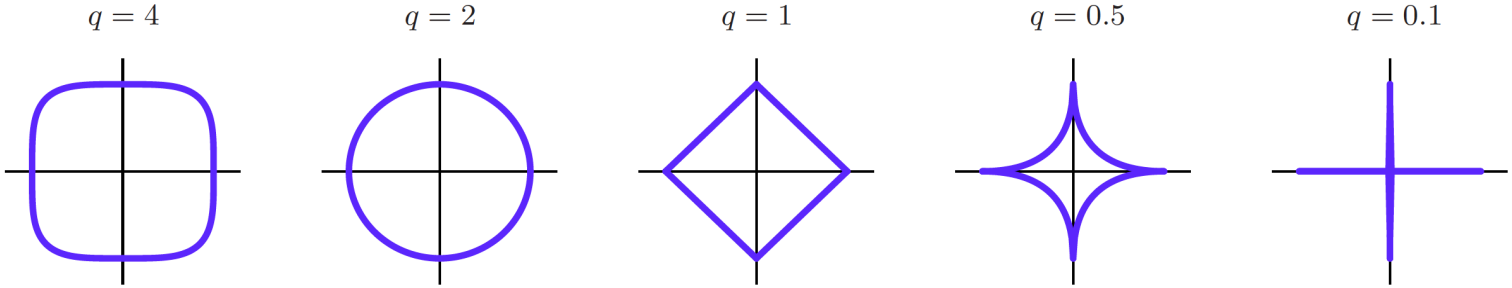
Intuition: L_1 norm induces "sharp" balls!

- Balls more concentrated around the axes
- Induces model selection by regulating the LASSO (through λ)

Extreme case: L^0 "Norm", gives best subset selection!

$$\|\gamma\|_0 = \sum_{j=1}^{p-1} |\gamma_j|^0 = \sum_{j=1}^{p-1} \mathbf{1}_{\{\gamma_j \neq 0\}} = \#\{j : \gamma_j \neq 0\}$$

Generally: $\|\gamma\|_q^q = \sum_{j=1}^{p-1} |\gamma_j|^q$, sharp balls for $0 < q \leq 1$



But L^1 gives sharpest convex ball among these.

Shrinkage XVI

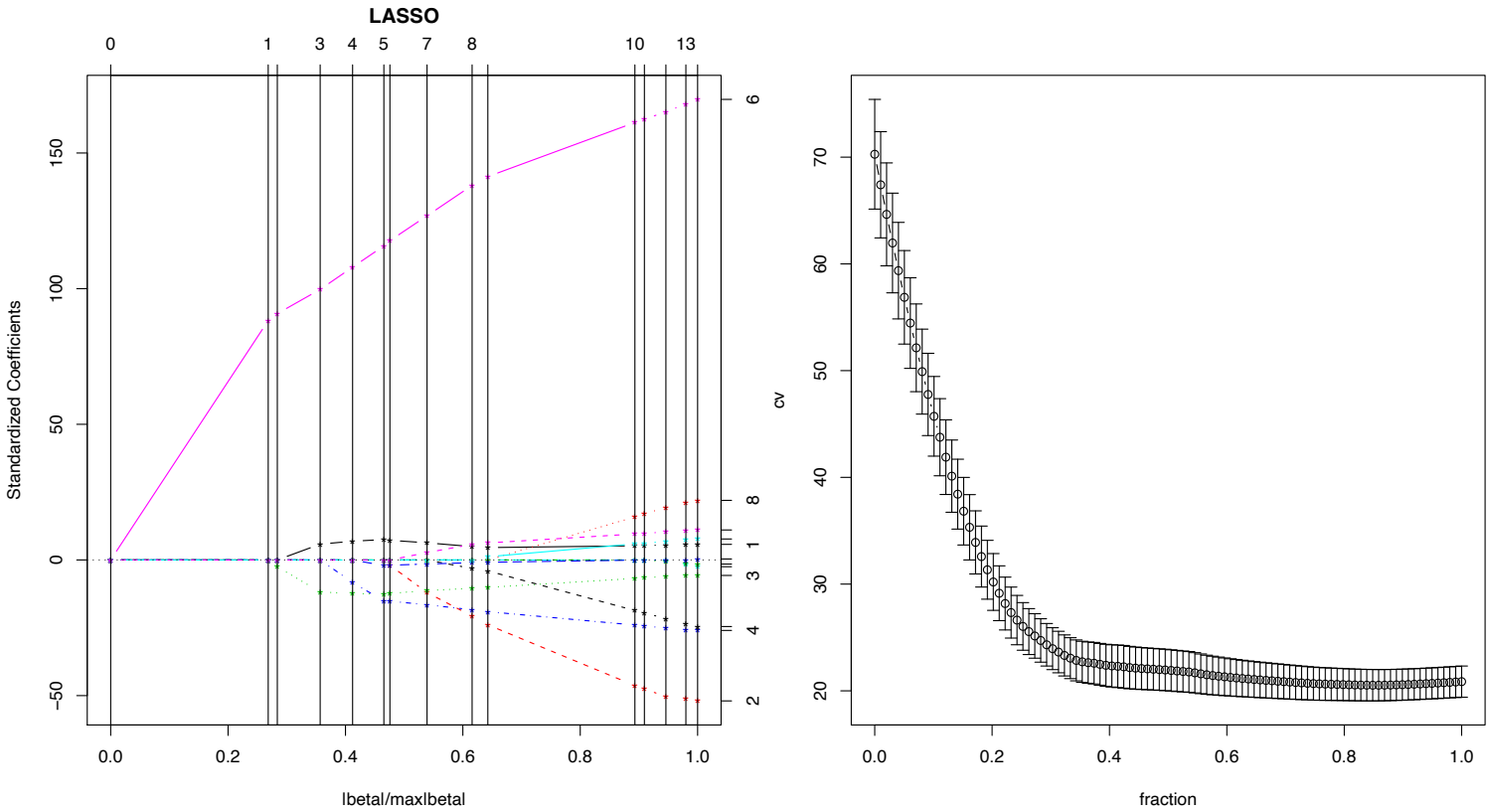


Figure: LASSO and CV for different values of $r(\lambda)/||\hat{\gamma}||_1$ for the bodyfat data (LARS algorithm)

Generalised Linear Model

Back to the big picture:

Y (random output) $\overset{\text{whose law is influenced by}}{\longleftarrow}$ x (non-random input)

General formulation:

$$Y_i \overset{\text{independent}}{\sim} \text{Distribution}\{\underbrace{g(x_i)}_{=\theta_i}\}, \qquad i = 1, ..., n.$$

Distribution / Function g	$g(x_i^\top) = x_i^\top \beta$	g nonparametric
Gaussian	Linear Regression ✓	Smoothing
Exponential Family	GLM ←	GAM

Generalised Linear Model II

Generalised Linear Models: regression with exponential family responses!

GLM for $Y_1, ..., Y_n$

Response vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ has independent entries with joint law

$$f(\boldsymbol{y}; \boldsymbol{\phi}) = \prod_{i=1}^{n} \exp\{\phi_i y_i - \gamma(\phi_i) + S(y_i)\} = \exp\left\{\boldsymbol{\phi}^\top \boldsymbol{y} - \sum_{i=1}^{n} \gamma(\phi_i) + \sum_{i=1}^{n} S(y_i)\right\},$$

where $\boldsymbol{\phi} \in \Phi \subseteq \mathbb{R}^n$ is the natural parameter with Φ open. The parameter varies as a function of the covariates via

$$\boldsymbol{\phi} = \boldsymbol{X}_n \boldsymbol{\beta},$$

for $\boldsymbol{X}_n$ the $n \times p$ covariate matrix of rank p and $\boldsymbol{\beta}$ a p-dimensional parameter.

In our general notation

$$Y_i \overset{\text{independent}}{\sim} f(y_i; \phi_i) = \exp\{\underbrace{(\boldsymbol{x}_i^\top \boldsymbol{\beta})}_{=\phi_i} y_i - \gamma(\underbrace{\boldsymbol{x}_i^\top \boldsymbol{\beta}}_{=\phi_i}) + S(y_i)\}, \qquad i = 1, ..., n.$$

where the row vector $\boldsymbol{x}_i^\top$ is the ith row of $\boldsymbol{X}_n$.

Generalised Linear Model III

Comments:

- Notice that the the sufficient statistic for each marginal distribution $f(y;\phi_i)$ was taken to be the identity $T(Y) = Y$.
- This does not incur any loss of generality for two reasons:
  1. In the three main GLM of interest (Gaussian, Bernoulli, and Poisson) the natural statistic is for $f(y;\phi_i)$ is indeed the identity.
  2. More generally, since we only observe a single observation from each $f(y;\phi_i)$, if the natural statistic were $T(Y_i) \neq Y_i$, we could re-define the response to just be $T_i = T(Y_i)$. The sampling distribution of T_i can be shown to also be a one-parameter exponential family with the same natural parameter.
- Recall from our sampling theory results:
  - $\mu_i = \mathbb{E}[Y_i] = \dfrac{\partial}{\partial \phi_i}\gamma(\phi_i)$
  - $\text{var}[Y_i] = \dfrac{\partial^2}{\partial \phi_i^2}\gamma(\phi_i)$
  - So if γ is invertible (which it is for the three main examples), the variance can be written as a function of the mean:

$$\text{var}[Y_i] = \gamma''([\gamma']^{-1}(\mu_i)) = V(\mu_i).$$

SGeneralised Linear Model IV

Interpretation of $\phi_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$?

- In key cases ϕ_i is directly interpretable. If not, can switch perspective using the mean μ_i as defining parameter, connected to the linear predictor $\boldsymbol{x}_i^\top \boldsymbol{\beta}$ via

$$[\gamma']^{-1}(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} = \phi_i$$

- The function $[\gamma']^{-1}(\cdot)$ is called the natural link function.
- Instead of $[\gamma']^{-1}$ can use other link functions $g(\cdot)$ and postulate

$$g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

  This will also yield a GLM, but now the natural parameter will not be equal to the linear predictor but to some function $u(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ of it.
- In summary, the nomenclature is:
  - $g(\cdot)$ is the link function
  - $h = g^{-1}$ is the inverse link function
  - $g(\cdot) = [\gamma']^{-1}(\cdot)$ is the natural link function
- Will focus on natural links but methods/results generalise quite easily.

Generalised Linear Model V

With natural link, the loglikelihood (up to constants w.r.t. $\boldsymbol{\beta}$) is

$$\ell_n(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \boldsymbol{X}_n^\top \boldsymbol{Y} - \sum_{i=1}^n \gamma(\boldsymbol{x}_i^\top \boldsymbol{\beta})$$

for $\boldsymbol{x}_i^\top$ the ith row of $\boldsymbol{X}_n$. The corresponding $p \times 1$ derivative (*score function*) is

$$\nabla_{\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta}) = \boldsymbol{X}_n^\top \boldsymbol{Y} - \sum_{i=1}^n \boldsymbol{x}_i \gamma'(\boldsymbol{x}_i^\top \boldsymbol{\beta}) = \sum_{i=1}^n \boldsymbol{x}_i \left(Y_i - \mu_i \right) = \boldsymbol{X}_n^\top \left(\boldsymbol{Y} - \boldsymbol{\mu} \right)$$

with $p \times p$ covariance equaling the information matrix and given by

$$\begin{aligned} \operatorname{cov} \left\{ \nabla_{\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta}) \right\} &= \sum_{i=1}^n \boldsymbol{x}_i^\top \operatorname{cov} \left\{ Y_i - \mu_i \right\} \boldsymbol{x}_i = \boldsymbol{X}_n^\top \boldsymbol{V}(\boldsymbol{\beta}) \boldsymbol{X}_n \\ &= -\nabla^2_{\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta}) = \mathcal{I}_n(\boldsymbol{\beta}), \end{aligned}$$

where $\operatorname{cov}\{ \boldsymbol{Y} \} = \boldsymbol{V}(\boldsymbol{\beta}) \succ 0$ is diagonal, with ith diagonal element

$$\operatorname{var}\{ Y_i \} = \gamma''(\phi_i) = \gamma''(\boldsymbol{x}_i^\top \boldsymbol{\beta}).$$

Generalised Linear Model VI

Thus, if the MLE $\hat{\boldsymbol{\beta}}$ exists it is also unique, and must satisfy:

$$\sum_{i=1}^{n} \boldsymbol{x}_i \left(Y_i - \gamma'(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}) \right) = 0$$

By a first order Taylor expansion of γ', we have

$$\gamma'(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}) \approx \gamma'(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) + \boldsymbol{x}_i^\top (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) \gamma''(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}})$$

for some guesstimate $\tilde{\boldsymbol{\beta}}$ near $\hat{\boldsymbol{\beta}}$. Plugging into the score equation yields

$$\sum_{i=1}^{n} \boldsymbol{x}_i^\top \left(Y_i - \gamma'(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) + \boldsymbol{x}_i^\top (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) \gamma''(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) \right) \approx 0$$

$$\implies \sum_{i=1}^{n} \gamma''(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) \boldsymbol{x}_i^\top (Z_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}) = \boldsymbol{X}_n^\top \boldsymbol{V}(\tilde{\boldsymbol{\beta}})(\tilde{\boldsymbol{Z}} - \boldsymbol{X}_n \hat{\boldsymbol{\beta}}) \approx 0$$

where we defined $\tilde{\boldsymbol{Z}} = (\tilde{Z}_1, ..., \tilde{Z}_n)^\top$ to be the adjusted response

$$\tilde{Z}_i = \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}} + \frac{1}{\gamma''(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}})} (Y_i - \gamma'(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}})) \quad \text{so} \quad \tilde{\boldsymbol{Z}} = \boldsymbol{X}_n \tilde{\boldsymbol{\beta}} + \boldsymbol{V}^{-1}(\tilde{\boldsymbol{\beta}})(\boldsymbol{Y} - \boldsymbol{\mu}(\tilde{\boldsymbol{\beta}})).$$

Generalised Linear Model VII

The last expression for the score expression now yields:

$$\hat{\boldsymbol{\beta}} \approx (\boldsymbol{X}_n^\top \boldsymbol{V}(\tilde{\boldsymbol{\beta}})\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^\top \boldsymbol{V}(\tilde{\boldsymbol{\beta}})\tilde{\boldsymbol{Z}}$$

Just a weighted least squares estimate! Where's the catch?

- Weight matrix $\boldsymbol{V}(\tilde{\boldsymbol{\beta}})$ requires specification of an initial guesstimate $\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}$ sufficiently close to $\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}$.
- Luckily we can give such a guesstimate by recalling that $\mu_i = \gamma'(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ so that estimating μ_i by Y_i yields the guesstimate $\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}} \equiv (\gamma')^{-1}(Y_i)$.
- Suggests the following Iteratively Reweighted Least Squares (IRLS)

IRLS

1. Initialize with $\boldsymbol{x}_i^\top \boldsymbol{\beta}^{(0)} \leftarrow \gamma'(Y_i)$ and $Z_i^{(0)} = \boldsymbol{x}_i^\top \boldsymbol{\beta}^{(0)} + \frac{(Y_i - \gamma'(\boldsymbol{x}_i^\top \boldsymbol{\beta}^{(0)}))}{\gamma''(\boldsymbol{x}_i^\top \boldsymbol{\beta}^{(0)})}$
2. Update with $\boldsymbol{\beta}^{(j+1)} \leftarrow (\boldsymbol{X}_n^\top \boldsymbol{V}(\boldsymbol{\beta}^{(j)})\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^\top \boldsymbol{V}(\boldsymbol{\beta}^{(j)})\boldsymbol{Z}^{(j)}$

- Equivalent to Newton-Raphson iteration.
- Not always guaranteed to converge.

Generalised Linear Model VIII

(Approximate) Sampling Distribution of MLE in GLM

Heuristics:

- Suppose we had started iteration at true $\boldsymbol{\beta}$ and stopped at first iterate:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}_n^\top \boldsymbol{V}(\boldsymbol{\beta})\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^\top \boldsymbol{V}(\boldsymbol{\beta})\boldsymbol{Z}$$

  where

$$Z_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \frac{1}{\gamma''(\boldsymbol{x}_i^\top \boldsymbol{\beta})}(Y_i - \gamma'(\boldsymbol{x}_i^\top \boldsymbol{\beta})) \quad \text{so} \quad \boldsymbol{Z} = \boldsymbol{X}_n\boldsymbol{\beta} + \boldsymbol{V}^{-1}(\boldsymbol{\beta})(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}))$$

- This would give us,

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}_n^\top \boldsymbol{V}(\boldsymbol{\beta})\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^\top(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}))$$

- So we would expect $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ and $\text{cov}[\hat{\boldsymbol{\beta}}] = (\boldsymbol{X}_n^\top \boldsymbol{V}^{-1}(\boldsymbol{\beta})\boldsymbol{X}_n)^{-1} = \mathcal{I}_n^{-1}(\boldsymbol{\beta})$.
- And we would conjecture a Gaussian limiting law (paralleling the IID setting)

Under conditions, this is indeed what we obtain.

Result stated in form valid for more general (sufficiently regular) link functions.

Generalised Linear Model IX

Theorem (Asymptotic Normality of MLE in GLM)

In the same context and notation as before, assume that:

(C1) $\boldsymbol{\beta} \in B$ *for* B *an open convex subset of* $\mathbb{R}^p$.

(C2) *The* $p \times p$ *matrix* $\boldsymbol{X}_n^\top \boldsymbol{X}_n$ *is of full rank for all* n.

(C3) *The information diverges, i.e.* $\lambda_{min}\Big(\mathcal{I}_n(\boldsymbol{\beta})\Big) \to \infty$ *as* $n \to \infty$ *for* $\lambda_{\min}(\cdot)$ *the smallest eigenvalue.*

(C4) *Given any parameter* $\boldsymbol{\beta} \in \mathbb{R}^p$ *it holds that*

$$\sup_{\boldsymbol{\alpha} \in N_\delta(\boldsymbol{\beta})} \left\| \mathcal{I}_n^{-1/2}(\boldsymbol{\beta}) \mathcal{I}_n^{1/2}(\boldsymbol{\alpha}) - \boldsymbol{I}_{p\times p} \right\| \to 0$$

$$\forall\, \delta > 0, \text{ where } N_\delta(\boldsymbol{\beta}) = \left\{ \boldsymbol{\alpha} \in \mathbb{R}^p : (\boldsymbol{\alpha} - \boldsymbol{\beta})^\top \mathcal{I}_n(\boldsymbol{\beta})(\boldsymbol{\alpha} - \boldsymbol{\beta}) \leq \delta \right\}.$$

Then, as $n \to \infty$, *provided it exists, the MLE* $\hat{\boldsymbol{\beta}}_n$ *of* $\boldsymbol{\beta}_0$ *is unique & satisfies*

$$\mathcal{I}_n^{1/2}(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, \boldsymbol{I}_{p\times p}).$$

- Recall that under canonical link $\mathcal{I}_n(\boldsymbol{\beta}) = \boldsymbol{X}_n^\top \boldsymbol{V}(\boldsymbol{\beta}) \boldsymbol{X}_n$.