

An introduction to sparsity: modelling, properties and some applications

Benjamin Poignard

Osaka University and RIKEN AIP

Ecodep Seminar

The 24th of March 2021

Agenda

- 1 Big Data and High-dimensionality.
- 2 Sparsity: modelling and properties.
- 3 Some applications.

- 1 Big Data and High-dimensionality
 - New statistical challenges
 - Solutions for the dimensionality problem
- 2 General theory of penalized M-estimators
- 3 Applications

Big Data

New types of data are now available: continuous geolocation of car drivers through the GPS system; facial recognition; price sequences of financial assets; and the like.

These data are massive (gigabytes of data) and large datasets may allow for more flexible relationships than simple linear regressions.

They require new statistical methods to analyze the relationship among these datasets and perform prediction.

High-dimensionality

High-dimensionality is about the large number of parameters we would like to estimate.

It concerns variable selection and common feature extraction.

Example Let $(y_{i,t}, i = 1, \dots, N; t = 1, \dots, T)$ and

$$y_t = \Phi y_{t-1} + u_t,$$

with N : number of variables; T : number of observations;
 $\forall t, u_t \sim \mathcal{N}_{\mathbb{R}^N}(0, \Sigma)$.

Parameter vector: $\beta = (\text{vec}(\Phi)^\top, \text{vec}(\Sigma)^\top)^\top$: $N^2 + N(N+1)/2$
unknown coefficients.

If $N > T$, then $N^2 > NT$: significant model complexity.

Let $Z_t = (y_{1,t}, \dots, y_{N,t})^\top$ and
 $\mathbf{Y} = (Z_2, \dots, Z_T)$, $\mathbf{X} = (Z_1, \dots, Z_{T-1})$, $\mathbf{U} = (u_2, \dots, u_T)$.

Dimensions: $\mathbf{Y} : N \times (T - 1)$, $\mathbf{X} : N \times (T - 1)$, $\Phi : N \times N$.

Regression model: $\mathbf{Y} = \Phi \mathbf{X} + \mathbf{U}$. By OLS

$$\hat{\Phi}^{\text{ols}} = (\mathbf{Y} \mathbf{X}^\top)(\mathbf{X} \mathbf{X}^\top)^{-1}.$$

Problem:

Identifiability: $\mathbf{X} \mathbf{X}^\top$ is non-invertible (rank condition).

Overfitting issue.

- 1 Big Data and High-dimensionality
 - New statistical challenges
 - Solutions for the dimensionality problem
- 2 General theory of penalized M-estimators
- 3 Applications

Solution 1: Factor modelling

Introduction of factor models (Stock and Watson, 1989):

$$y_t = BAy_{t-1} + u_t = BF_{t-1} + u_t,$$

where (F_t) is a much smaller vector than y_t . The B matrix ($N \times q$) contains for example β coefficients (APT).

The vector of interest is $\beta = (\text{vec}(B)^\top, \text{vec}(\Sigma)^\top)^\top$:
 $Nq + N(N + 1)/2$ unknown coefficients.

Example: Application to the portfolio allocation problem (Fan, J and al., 2008).

Solution 2: parameter shrinkage

Let $\mathbf{Y}_v = \text{vec}(\mathbf{Y})$, $\tilde{\mathbf{X}} = \mathbf{Z}^\top \otimes I_N$, $\mathbf{U}_v = \text{vec}(\mathbf{U})$, $\theta = \text{vec}(\Phi)$.

Dimensions: $\mathbf{Y}_v : N(T-1) \times 1$, $\tilde{\mathbf{X}} : N(T-1) \times N^2$, $\theta : N^2 \times 1$.

Regression model: $\mathbf{Y}_v = \tilde{\mathbf{X}}\theta + \mathbf{U}_v$.

Parameter shrinkage (Bayesian approach): constrains the set of parameter values.

The Ridge method (Hoerl and Kennard, 1970) corresponds to

$$\hat{\theta}^{\text{ridge}} = \arg \min_{\theta} (\mathbf{Y}_v - \tilde{\mathbf{X}}\theta)^\top (\mathbf{Y}_v - \tilde{\mathbf{X}}\theta) + \lambda \sum_{k=1}^d \theta_k^2,$$

with $d = N^2$ (dimension of the regression parameters), $\lambda \geq 0$.

The solution is

$$\hat{\theta}^{\text{ridge}} = (\tilde{\mathbf{X}}^{\top} \tilde{\mathbf{X}} + \lambda I_{N^2})^{-1} \tilde{\mathbf{X}}^{\top} \mathbf{Y}.$$

Bayesian interpretation

$$\begin{aligned} \hat{\theta}^{\text{ridge}} &= \arg \max_{\theta} p(\theta | \mathbf{Y}, \mathbf{X}), \\ p(\theta | \mathbf{Y}, \mathbf{X}) &= p(\mathbf{Y}, \mathbf{X} | \theta) p(\theta). \end{aligned}$$

Prior $\text{vec}(\theta) \sim \mathcal{N}(0, \lambda^{-1} I_{N^2})$. Then the maximum a posteriori corresponds to the Ridge regression for a gaussian likelihood.

Solution 3: LASSO

The LASSO (Tibshirani, 1996).

$$\hat{\theta}^{\text{lasso}} = \arg \min_{\theta} (\mathbf{Y}_v - \tilde{\mathbf{X}}\theta)^\top (\mathbf{Y}_v - \tilde{\mathbf{X}}\theta) + \lambda \sum_{k=1}^d |\theta_k|,$$

Bayesian interpretation

Prior $\theta_i \sim \text{Laplace}(0, b)$ with $b = \lambda^{-1}$ the scale parameter such that $\mathbf{p}(\theta) \propto \exp(-\lambda \sum_{k=1}^d |\theta_k|)$. For a Gaussian likelihood, the posterior distribution correspond to the LASSO.

The Laplacian prior assigns more weight to regions near zero than the normal prior.

Parameters of interest

Trade-off: parameter change and parameter weight.

Example VAR for the Nikkei 225: large companies/smaller caps, company sectors, and the like.

Alternative: the Group LASSO (Yuan and Lin, 2006): m groups (known) of parameters with sizes p_1, \dots, p_m . Then

$$\hat{\theta}^{\text{Glasso}} = \arg \min_{\Phi} (\mathbf{Y}_v - \tilde{\mathbf{X}}\theta)^\top (\mathbf{Y}_v - \tilde{\mathbf{X}}\theta) + \lambda \sum_{j=1}^m \eta_j \sqrt{\sum_{k=1}^{p_j} |\theta_k^{(j)}|^2},$$

where $\theta = (\theta_k^{(j)}, j = 1, \dots, m; k = 1, \dots, p_j)$, η_j controls for the group's size.

Sparsity assumption

Data with a large number of variables relative to the sample size are increasingly common. High-dimensional data arise through a combination of:

- (i) the data may be inherently high-dimensional in that many different characteristics per observation are available.
- (ii) even when the number of available variables is relatively small, researchers rarely know the exact functional form with which the small number of variables enters the model of interest.

The key concept underlying the analysis of high-dimensional data: **dimension reduction** or **regularization**.

Producing a useful forecasting model requires regularization; that is, the estimates must be constrained so that overfitting is avoided and useful out-of-sample forecasts can be obtained.

Parameter of interest: $\theta \in \mathbb{R}^d$.

The sparsity assumption can be formulated as:

$$k_0 = \text{card}(\mathcal{A}), \quad \text{with } \mathcal{A} := \left\{ i : \theta_{0,i} \neq 0 \right\},$$

such that $k_0 < d$.

\mathcal{A} : true underlying support (not observed).

Penalisation/Regularization: provides $\hat{\mathcal{A}}$.

Penalty/Regularizer:

- (i) norm with respect to the parameter **non-differentiable** at the origin.
- (ii) the penalty depends on a **tuning/regularization** parameter that enforces a particular type of sparse structure in the solution.

- 1 Big Data and High-dimensionality
- 2 General theory of penalized M-estimators
 - Penalised M-estimation
 - Properties
- 3 Applications

Criterion

n sample $\mathcal{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)$ of n realizations of $\mathbf{X} \in \mathbb{R}^q$.

Loss function $\mathbb{L}_n : \mathbb{R}^{qn} \times \Theta \rightarrow \mathbb{R}$, $\Theta \subseteq \mathbb{R}^d$, defined as

$$\mathbb{L}_n(\theta; \mathcal{X}) = \frac{1}{n} \sum_{i=1}^n \ell(\theta; \mathbf{X}_i).$$

Typically, ℓ : least square error, or minus a log-likelihood function.

Problem of interest:

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \left\{ \mathbb{L}_n(\theta; \mathcal{X}) + \mathbf{p}(\lambda_n, \theta) \right\}.$$

Here $\mathbf{p} : \mathbb{R}^+ \times \Theta \rightarrow \mathbb{R}^+$ is the penalty/regularizer; λ_n is the tuning/regularization parameter.

Motivation for sparsity? How to define sparsity in the model?

Penalty function

LASSO (Tibshirani, 1996), Bridge (Knight and Fu, 2000), SCAD (Fan and Li, 2001) and MCP (Zhang, 2010):

$$\text{Lasso : } \boldsymbol{\rho}(\lambda, \rho) = \lambda|\rho|,$$

$$\text{Bridge : } \boldsymbol{\rho}(\lambda, \rho) = \lambda|\rho|^\gamma, \quad 0 < \gamma < 1,$$

$$\text{MCP : } \boldsymbol{\rho}(\lambda, \rho) = \text{sign}(\rho)\lambda \int_0^{|\rho|} (1 - z/(\lambda b_1))_+ dz,$$

$$\text{SCAD : } \boldsymbol{\rho}(\lambda, \rho) = \begin{cases} \lambda|\rho|, & |\rho| \leq \lambda, \\ -\frac{1}{2(b_2-1)}(\rho^2 - 2b_2\lambda|\rho| + \lambda^2), & \lambda \leq |\rho| \leq b_2\lambda, \\ (b_2 + 1)\lambda^2/2, & |\rho| > b_2\lambda, \end{cases}$$

Here, $b_1 > 0$ and $b_2 > 2$: the larger, the more a LASSO like penalty.

Extension to Group Penalisation, (Group) Fused LASSO, ...

- 1 Big Data and High-dimensionality
- 2 General theory of penalized M-estimators
 - Penalised M-estimation
 - Properties
- 3 Applications

Asymptotic analysis

Focus on the asymptotic behaviour of the sparse M-estimator when $n, d \rightarrow \infty$. Usually, $d = O(n^c)$ with $0 < c < 1$.

Key result: **oracle property** (Fan and Li, 2001) \Rightarrow sparsity-based estimator recovers \mathcal{A} and is asymptotically normally distributed.

Convex penalisation (LASSO, Group LASSO): do not satisfy the oracle property (inherent bias shrinking the large parameters) except under a specific condition (irrepresentable condition); to fix this issue, use **adaptive** version: Zou (2006).

- (i) Knight and Fu (2000): asymptotic properties of LASSO/Bridge when $n \rightarrow \infty$ only within OLS setting.
- (ii) Fan and Li (2001): general penalized likelihood framework with SCAD and oracle property. Fan and Peng (2004): extension to double asymptotic.
- (iii) Zou (2006): adaptive LASSO and oracle property.

Oracle inequalities and support recovery

Derivation of explicit error bounds of the sparse M-estimator and the conditions to establish support recovery.

The curvature of the loss function is a key ingredient:

- (i) restricted eigenvalue conditions: Bickel, Ritov and Tsybakov (2009); van de Geer and Bühlmann (2009).
- (ii) restricted strong convexity (RSC): Negahban, Ravikumar, Wainwright and Yu (2012); Loh and Wainwright (2015, 2017); Poignard and Fermanian (2021).

Loh and Wainwright (2017): support recovery established for non-convex penalty functions when the loss \mathbb{L}_n satisfies the RSC condition.

- 1 Big Data and High-dimensionality
- 2 General theory of penalized M-estimators
- 3 Applications
 - MGARCH
 - Factor modelling

High-dimensional MGARCH

Joint work with J.D. Fermanian (2021).

In finance: need for flexible and realistic joint dynamics for asset returns.

Portfolio size N , which may be large: $N = 50, 100, 1000, \dots$

Quantity of interest: second order conditional moment.

The usual approaches :

- (a) Multivariate GARCH models (MGARCH)
- (b) Multivariate stochastic volatility models (MSV)

Typical problems:

- (i) up to $O(N^4)$, $O(N^2)$ parameters in general. For some restricted "scalar cases", only 3, but questionable.
- (ii) generation of nonnegative definite matrices \Rightarrow some more or less ad-hoc "tricks", "normalizations", etc.
- (iii) inference techniques (two-stage Quasi Maximum Likelihood) without well-founded theoretical foundations
- (iv) modest improvements in terms of forecasting performances

Multivariate GARCH process

A stochastic process $(X_t)_{t=1, \dots, T}$, $X_t \in \mathbb{R}^N$.

Detrended series:

$$X_t = \mu_t + \epsilon_t,$$

$$\mu_t = \mathbb{E}[X_t | \mathcal{F}_{t-1}] = \Phi_0 + \Phi_1 X_{t-1},$$

$$\epsilon_t = H_t^{1/2}(\theta) \eta_t.$$

- (η_t) : strong white noise, $\mathbb{E}[\eta_t] = 0$, $\text{Var}(\eta_t) = I_N$.
- semi-parametric model \Rightarrow specifications of the law of (η_t) and the dynamic of (H_t) .
- $\forall \theta$, $H_t(\theta) \in \mathcal{F}_{t-1}$.

What can we expect from (H_t) ?

- (i) Stability by aggregation.
- (ii) Sufficiently richly parameterized to capture cross-dynamics / parsimony.
- (iii) Easy conditions for positive-definiteness.
- (iv) Avoid excessive inversion of the conditional variance.

A lot of model specification on (H_t) : Vector-GARCH, BEKK, DCC, and the like.

These models typically suffer from the curse of dimensionality:

$$\text{BEKK: } H_t = \Omega + A\epsilon_{t-1}\epsilon_{t-1}^\top A^\top + BH_{t-1}B^\top,$$

with $\Omega \succ 0$, $A, B : N \times N$ matrices.

Solutions to the curse of dimensionality

- (i) Scalar dynamics: scalar BEKK, scalar DCC and the like, which consists in constraining the matrix parameters as scalar parameters.
- (ii) Introduction of factor models: factor GARCH models.
- (iii) Parameter shrinkage:

$$\hat{\theta} = \arg \min_{\theta} \left\{ f^{\text{qMLE}}(\theta; \epsilon_t, t = 1, \dots, T) + \mathbf{p}(\lambda, \theta) \right\}.$$

Problems: smoothness of $f^{\text{qMLE}}(\cdot; \epsilon_t, t = 1, \dots, T)$, numerical estimation.

Multivariate ARCH

Ignoring the autoregressive term, the Vector GARCH becomes

$$H_t = A + \sum_{k=1}^q (I_N \otimes \epsilon_{t-k}^\top) B_k (I_N \otimes \epsilon_{t-k}),$$

where A, B_k are symmetric, non-negative definite. This can be written as a linear model:

$$\epsilon_t \epsilon_t^\top = A + \sum_{k=1}^q (I_N \otimes \epsilon_{t-k}^\top) B_k (I_N \otimes \epsilon_{t-k}) + \zeta_t, \quad \mathbb{E}[\zeta_t | \mathcal{F}_{t-1}] = 0,$$

instead for every couple $(i, j) \in \{1, \dots, N\}^2$ such that $i \leq j$, we have

$$\epsilon_{i,t} \epsilon_{j,t} = a_{i,j} + \sum_{k=1}^q \sum_{r,s=1}^N b_{ijk,rs} \epsilon_{r,t-k} \epsilon_{s,t-k} + \zeta_{ij,t}, \quad \mathbb{E}[\zeta_{ij,t} | \mathcal{F}_{t-1}] = 0,$$

where $B_{ijk} = [b_{ijk,rs}]_{1 \leq r,s \leq N}$. Sparsity assumption on the B_{ijk} coefficients.

- Multivariate ARCH process: estimation by OLS.
- Assume that the above model is the true one, with the true index q_0 . A penalisation procedure with q larger than q_0 would likely set the parameters $b_{ijk,rs}$ to zero when $k > q_0$.
- If the true model is a GARCH type one, then it can be rewritten as the above model with $q = \infty$ (under suitable conditions on the parameters).
⇒ May produce relevant approximations of usual GARCH processes taking q "sufficiently" large.
- Propose several conditional variance specification ensuring the p.d. of H_t : Cholesky-GARCH, projection on space of p.d. matrices.
- Numerical advantage: ability to parallelize the estimation (equation-by-equation).

Consistency and oracle property

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \{ \mathbb{G}_T \ell(\theta) + \mathbf{p}(\lambda_T, \gamma_T \theta) \}, \quad \mathbb{G}_T \ell(\cdot) : \text{OLS type loss,}$$

$$\mathbf{p}(\lambda_T, \gamma_T \theta) = \lambda_T \sum_{k=1}^m \sum_{i=1}^{\mathbf{c}_k} \alpha_{T,i}^{(k)} |\theta_i^{(k)}| + \gamma_T \sum_{l=1}^m \xi_{T,l} \|\theta^{(l)}\|_2 : \text{adaptive SGL.}$$

- (i) Sparsity: Group level (parameters corresponding to a lag) and within each group.
- (ii) $\|\hat{\theta} - \theta_0\| = O_p(T^{-1/2} + \lambda_T T^{-1} a_T + \gamma_T T^{-1} b_T)$, where $a_T := k_0 \cdot (\max_{k \in \mathcal{S}} (\max_{i \in \mathcal{A}_k} \alpha_{T,i}^{(k)}))$, $b_T := k_0 \cdot (\max_{l \in \mathcal{S}} \xi_{T,l})$ with $k_0 = \text{card}(\mathcal{A})$.
- (iii) Oracle property: $\lim_{T \rightarrow \infty} \mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}) = 1$ and

$$\sqrt{T}(\hat{\theta}_{\mathcal{A}} - \theta_{0,\mathcal{A}}) \xrightarrow[T \rightarrow \infty]{d} \mathcal{N}_{\mathbb{R}^{k_0}}(0, \mathbb{H}_{\mathcal{A}\mathcal{A}}^{-1} \mathbb{M}_{\mathcal{A}\mathcal{A}} \mathbb{H}_{\mathcal{A}\mathcal{A}}^{-1}).$$

- 1 Big Data and High-dimensionality
- 2 General theory of penalized M-estimators
- 3 Applications
 - MGARCH
 - Factor modelling

Factor models

Joint work with Y.Terada (2020).

Consider n observations of a p -dimensional i.i.d. random vector (X_i) following the factor structure

$$X_i = \Lambda F_i + \epsilon_i,$$

where (F_i) is the \mathbb{R}^m vector of factor variables and (ϵ_i) the \mathbb{R}^p vector of errors - or idiosyncratic variables (r.h.s. non-observable).

- $\Lambda \in \mathcal{M}_{p \times m}(\mathbb{R})$ is the loading matrix, m is known.
- $\mathbb{E}[F_i] = 0 \in \mathbb{R}^m$, $\mathbb{E}[F_i F_i^\top] = I_m$, $\mathbb{E}[F_i \epsilon_i^\top] = 0 \in \mathcal{M}_{m \times p}(\mathbb{R})$.
- $\mathbb{E}[\epsilon_i \epsilon_i^\top] = \Psi \in \mathcal{M}_{p \times p}(\mathbb{R})$ non-diagonal.

The idiosyncratic components (ϵ_i) are assumed to be correlated: approximate factor models (Chamberlain and Rothschild, 1983).
The quantity of interest is

$$\Sigma(\Lambda, \Psi) := \text{Var}(X_i) = \Lambda \Lambda^\top + \Psi.$$

Inference methods

PCA method

- Provides an easy estimation of Σ together with consistent estimators when p and n are large.
- Implicitly assumes that the idiosyncratic covariance matrix is decomposed as a scalar times an identity matrix: see, e.g., Fan, Liao, Mincheva (2011, 2013) and their **POET** estimator (probability bounds).

QML method

- Eliminates the bias from the cross-sectional heteroscedasticity: see, e.g., Bai and Li (2012, 2016).
- Anderson and Amemiya (1988), Bai and Li (2012, 2016): large sample properties of the likelihood-based factor model estimators, Ψ is diagonal.
- Bai and Li (2016): large sample properties of the QML-based factor model, non-diagonal Ψ and diverging p .

Sparse modelling

A broad range of studies on the sparse estimation of Σ .

Such sparse assumption may not be appropriate: several common factors exist for the underlying structure of the observed variables.

Factor analysis stands as the natural method to appropriately deal with the common factors.

In standard factor analysis, the ϵ_i are assumed uncorrelated $\Rightarrow \Psi$ diagonal (strict factor model).

However, this diagonal assumption is too restrictive in practice: assume the sparsity of the idiosyncratic covariance, which allows for the existence of correlation among the idiosyncratic components (approximate factor model).

Statistical criterion: Gaussian QML

$$\left\{ \begin{array}{l} \hat{\Psi}^g = \arg \min_{\Psi \in \Omega} \left\{ \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi) + \mathbf{p}(\lambda_n, \theta_{\Psi}) \right\}, \text{ where} \\ (\tilde{\Lambda}, \tilde{\Psi}) = \arg \min_{(\Lambda, \Psi) \in \Theta} \left\{ \mathbb{G}_{n,p}(\Lambda; \Psi) \right\}, \text{ with} \\ \mathbb{G}_{n,p}(\Lambda; \Psi) = \frac{1}{2p} \left(\log(|\Sigma(\Lambda, \Psi)|) + \text{tr}(\hat{S}\Sigma(\Lambda, \Psi)^{-1}) \right), \end{array} \right.$$

with $\theta_{\Psi} = \text{vech}(\Psi)$ and

- \hat{S} : sample variance covariance estimator.
- $\mathbf{p}(\lambda_n, \cdot) : \mathbb{R}^{p(p+1)/2} \rightarrow \mathbb{R}$: penalty function with λ_n the regularization parameter.
- $g(\theta_{\Psi}) \leq R$: side condition to manage non-convex problems.

$$\Omega = \left\{ \Psi : \Sigma := \Sigma(\tilde{\Lambda}, \Psi) = \tilde{\Lambda}\tilde{\Lambda}^{\top} + \Psi, \quad \Psi = \Psi^{\top}, \quad \Psi \succ 0, \right. \\ \left. c_1 < \lambda_{\min}(\Psi) < \lambda_{\max}(\Psi) < c_2, \quad a < \lambda_{\min}(2\hat{S} - \Sigma), \quad g(\theta_{\Psi}) \leq R \right\}.$$

Statistical criterion: Least Squares

$$\left\{ \begin{array}{l} \hat{\Psi}^{ls} = \arg \min_{\Psi \in \bar{\Omega}} \left\{ \mathbb{F}_{n,p}(\tilde{\Lambda}; \Psi) + \mathbf{p}(\lambda_n, \theta_{\Psi}) \right\}, \text{ where} \\ \mathbb{F}_{n,p}(\tilde{\Lambda}; \Psi) = \frac{1}{2p} \|\tilde{\Sigma} - \tilde{\Lambda}\tilde{\Lambda}^{\top} - \Psi\|_F^2, \text{ and} \\ (\tilde{\Lambda}, \tilde{\Psi}) = \arg \min_{(\Lambda, \Psi) \in \Theta} \left\{ \mathbb{G}_{n,p}(\Lambda; \Psi) \right\}, \text{ with} \\ \mathbb{G}_{n,p}(\Lambda; \Psi) = \frac{1}{2p} \left(\log(|\Sigma(\Lambda, \Psi)|) + \text{tr}(\hat{S}\Sigma(\Lambda, \Psi)^{-1}) \right), \end{array} \right.$$

with $\theta_{\Psi} = \text{vech}(\Psi)$ and

$$\bar{\Omega} = \left\{ \Psi : \Sigma := \Sigma(\tilde{\Lambda}, \Psi) = \tilde{\Lambda}\tilde{\Lambda}^{\top} + \Psi, \Psi = \Psi^{\top}, \Psi \succ 0, \right. \\ \left. l_1 < \lambda_{\min}(\Psi) < \lambda_{\max}(\Psi) < l_2, g(\theta_{\Psi}) \leq R \right\}.$$

Two-step estimation: motivation

- **Step 1** ($\tilde{\Lambda}, \tilde{\Psi}$): first step estimators (non-penalised) obtained by Gaussian QML function $\mathbb{G}_{n,p}(\cdot; \cdot)$ in the parameter set Θ .
- **Step 2** Solve the penalised Gaussian QML based criterion \Rightarrow
 $\hat{\theta}_{\Psi}^g = \text{vech}(\hat{\Psi}^g)$

Or alternatively

- Step 2** Solve the penalised least squares based criterion \Rightarrow
 $\hat{\theta}_{\Psi}^{ls} = \text{vech}(\hat{\Psi}^{ls})$

Two step method: regularity conditions on the (non-penalised) loss function with respect to Ψ (RSC condition) are satisfied conditionally on the first step estimators.

Bai and Li (2012):

$$\|\tilde{\Lambda} - \Lambda_0\|_F = O_p\left(\sqrt{\frac{p}{n}}\right) + O_p\left(\sqrt{\frac{1}{p}}\right).$$

Error bounds: Gaussian based criterion

Corollary

Assume $\mathbf{p}(\lambda_n, \cdot)$ is μ -amenable, $n \geq CR^2 \alpha_2^{-2} \log(p(p+1)/2)$, with $C > 0$ a sufficiently large constant, with

$\alpha_2 = \{\lambda_{\max}(\tilde{\Lambda}\tilde{\Lambda}^\top) + \lambda_{\max}(\Psi_0) + 1\}^{-3} a/2p$, if

$$4 \max\left\{\frac{\lambda_{\max}(\Psi_0^{-1})^2}{2p} \|\tilde{\Lambda}\tilde{\Lambda}^\top + \Psi_0 - \hat{S}\|_s, \alpha_2 \sqrt{\frac{\log p(p+1)/2}{n}}\right\} \leq \lambda_n \leq \frac{\alpha_2}{6R},$$

where $\Psi_0 \in \Omega$, suppose $\frac{3}{4}\mu < \alpha_1$ with $\alpha_1 = \alpha_2$. Then $\hat{\Psi}^g$ satisfies

$$\begin{aligned} \|\text{vech}(\hat{\Psi}^g) - \text{vech}(\Psi_0)\|_2 &\leq \frac{6\lambda_n \sqrt{k_0}}{4\alpha_1 - 3\mu}, \\ \|\text{vech}(\hat{\Psi}^g) - \text{vech}(\Psi_0)\|_1 &\leq \frac{6(16\alpha_1 - 9\mu)\lambda_n k_0}{(4\alpha_1 - 3\mu)^2}, \end{aligned}$$

with $a \in \Omega$ so that $a > 0$, $k_0 = |\mathcal{A}|$.

Error bounds: Least Squares based criterion

Corollary

Assume $\mathbf{p}(\lambda_n, \cdot)$ is μ -amenable, $n \geq CR^2 \alpha_2^{-2} \log(p(p+1)/2)$, with $C > 0$ a sufficiently large constant, with $\alpha_2 = \frac{1}{p}$, if

$$4 \max \left\{ \frac{1}{p} \|\widehat{\mathbf{S}} - \widetilde{\mathbf{\Lambda}} \widetilde{\mathbf{\Lambda}}^\top - \Psi_0\|_\infty, \frac{1}{p} \sqrt{\frac{\log p(p+1)/2}{n}} \right\} \leq \lambda_n \leq \frac{\alpha_2}{6R},$$





where $\Psi_0 \in \Omega$, suppose $\frac{3}{4}\mu < \alpha_1$ with $\alpha_1 = \alpha_2$. Then $\widehat{\Psi}^{ls}$ satisfies





$$\begin{aligned} \|\text{vech}(\widehat{\Psi}^{ls}) - \text{vech}(\Psi_0)\|_2 &\leq \frac{6\lambda_n \sqrt{k_0}}{4/p - 3\mu}, \\ \|\text{vech}(\widehat{\Psi}^{ls}) - \text{vech}(\Psi_0)\|_1 &\leq \frac{6(16/p - 9\mu)\lambda_n k_0}{(4/p - 3\mu)^2}, \end{aligned}$$





with $k_0 = |\mathcal{A}|$.





Work in progress




- Factor decomposition: $\Sigma = \Lambda\Lambda^\top + \Psi \Rightarrow$ Regularization of Λ .
Major difficulty: quadratic product and identifiability condition (rotational indeterminacy) \Rightarrow Penalised estimation equation framework with explicit management of the rotational indeterminacy.
- Sparse SVAR models via precision matrix: the SVAR coefficients can be interpreted in terms of a directed acyclic graph \Rightarrow sparse precision matrix of a suitable random vector provides sparse SVAR coefficients.
- Feature selection methods: specification of association measure, sure screening properties, management of redundant features.




-  ANDERSON, T.W. AND Y. AMEMIYA (1988): *The asymptotic normal distribution of estimators in factor analysis under general conditions*, The Annals of Statistics, Vol. 16, No. 2, 759-771.
-  BAI, J. AND K. LI (2012): *Statistical analysis of factor models of high dimension*, The Annals of Statistics, Vol. 40, No. 1, 436-465.
-  BAI, J. AND K. LI (2016): *Maximum likelihood estimation and inference for approximate factor models of high dimension*, The Review of Economics and Statistics, Vol. 98, No. 2.
-  BAI, J. AND K. LIAO (2016): *Efficient estimation of approximate factor models via penalized maximum likelihood*, Journal of Econometrics, Vol. 191, 1-18.

-  BICKEL, P.J., Y. RITOV AND A.B. TSYBAKOV (2009): *Simultaneous analysis of LASSO and Dantzig selector*, The Annals of Statistics, Vol. 37, No. 4, 1705–1732.
-  FAN, J. AND R. LI (2001): *Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties*, Journal of the American Statistical Association, 96:456, 1348-1360.
-  FAN, J. AND H. PENG (2004): *Nonconcave Penalized Likelihood with a Diverging Number of Parameters*, The Annals of Statistics, Vol. 32, No. 3, 928-961.
-  FAN, J. Y. FAN AND J. LV (2008): *High dimensional covariance matrix estimation using a factor model*, Journal of Econometrics, Vol. 147, 186-197.

-  FAN, J., Y. LIAO AND M. MINCHEVA (2011): *High-dimensional covariance matrix estimation in approximate factor models*, The Annals of Statistics, Vol. 39, No. 6, 3320-3356.
-  FAN, J., Y. LIAO AND M. MINCHEVA (2013): *Large covariance estimation by thresholding principal orthogonal complements*, Journal of the Royal Statistical Society Series B, Statistical Methodology, Vol. 75, No. 4.
-  HOERL, A.E. AND R.W. KENNARD (1970), *Ridge Regression: Biased Estimation for Nonorthogonal Problems*, Technometrics, Vol. 12, No. 1, 55-67.
-  KNIGHT, K. AND FU, W. (2000): *Asymptotics for LASSO-type Estimators*, The Annals of Statistics, Vol. 28, No. 5, 1356-1378.

-  LOH, P.L. AND M.J. WAINWRIGHT (2017): *Support recovery without incoherence: a case for non-convex regularisation*, The Annals of Statistics, Vol. 45, No. 6, 2455-2482.
-  NEGAHBAN, S.N, P. RAVIKUMAR, M.J. WAINWRIGHT AND B. YU (2012): *A unified framework for high-dimensional analysis of M-estimators with decomposable regularisers*, Statistical Science, Vol. 27, No. 4, 538-557.
-  POIGNARD, B., AND J.D. FERMANIAN (2021): *High-dimensional penalised ARCH processes*, Econometric Reviews, Vol. 40, No. 1, 86–107.
-  POIGNARD, B. AND J.D. FERMANIAN (2021): *Finite sample properties of Sparse M-estimators with Pseudo-Observations*, To appear in Annals of the Institute of Statistical Mathematics.

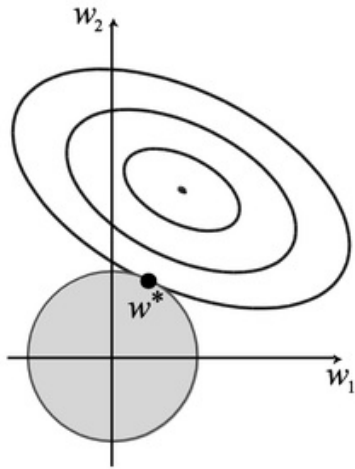
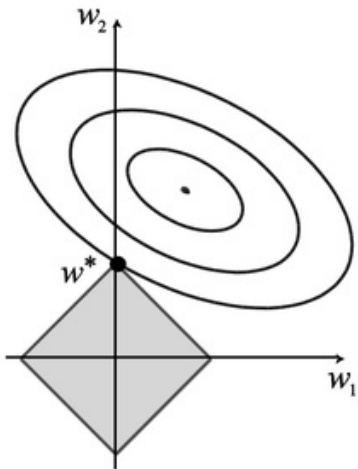
-  POIGNARD, B. AND Y. TERADA (2020): *Statistical Analysis of Sparse Approximate Factor Models*, Electronic Journal of Statistics, Vol. 14, 3315–3365.
-  STOCK, J.H., AND M.W. WATSON (1989), *New indexes of coincident and leading economic indicators*, NBER Macroeconomics Annual, 351-393.
-  TIBSHIRANI, R. (1996): *Regression Shrinkage and Selection via the Lasso*, Journal of the Royal Statistical Society. Series B, Vol. 58, No. 1, 267-288.
-  YUAN, M. AND Y. LIN (2006): *Model Selection and Estimation in Regression with Grouped Variables*, Journal of the Royal Statistical Society. Series B, Vol. 68, No. 1, 49-67.

-  VAN DE GEER, S.A. AND P. BÜHLMANN (2009): *On the conditions used to prove oracle results for the Lasso*, Electronic Journal of Statistics, Vol. 3, 1360-1392.
-  ZHANG, C.-H. (2010): *Nearly unbiased variable selection under minimax concave penalty*, The Annals of Statistics, Vol. 38, 894-942.
-  ZOU, H. (2006): *The adaptive LASSO and its oracle properties*, Journal of the American Statistical Association, Vol. 101, No. 476, 1418-1429.

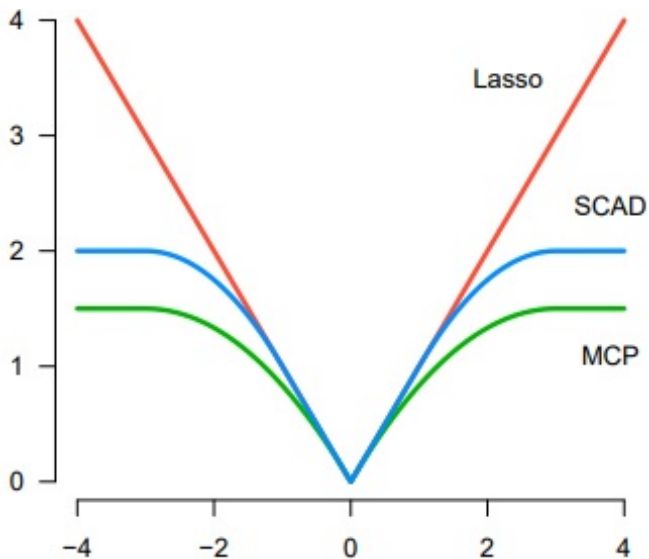
Geometry for Ridge and LASSO

$$\begin{aligned}\hat{\theta}^{\text{ridge}} &= \arg \min_{\theta} (\mathbf{Y}_v - \tilde{\mathbf{X}}\theta)^\top (\mathbf{Y}_v - \tilde{\mathbf{X}}\theta) + \lambda \sum_{k=1}^d \theta_k^2 \\ \Leftrightarrow \hat{\theta}^{\text{ridge}} &= \arg \min_{\theta} (\mathbf{Y}_v - \tilde{\mathbf{X}}\theta)^\top (\mathbf{Y}_v - \tilde{\mathbf{X}}\theta) \text{ s.t. } \sum_{k=1}^d \theta_k^2 \leq t.\end{aligned}$$

$$\begin{aligned}\hat{\theta}^{\text{lasso}} &= \arg \min_{\theta} (\mathbf{Y}_v - \tilde{\mathbf{X}}\theta)^\top (\mathbf{Y}_v - \tilde{\mathbf{X}}\theta) + \lambda \sum_{k=1}^d |\theta_k| \\ \Leftrightarrow \hat{\theta}^{\text{lasso}} &= \arg \min_{\theta} (\mathbf{Y}_v - \tilde{\mathbf{X}}\theta)^\top (\mathbf{Y}_v - \tilde{\mathbf{X}}\theta) \text{ s.t. } \sum_{k=1}^d |\theta_k| \leq t.\end{aligned}$$



[◀ Back to presentation](#)



[◀ Back to presentation](#)

Regularity conditions on the unpenalised loss

$\mathbb{L}_n(\theta)$: often lack of convexity w.r.t. the parameters.

Restricted strong convexity: allows the management of non-convex loss functions (see Negahban et al., 2012).

\mathbb{L}_n satisfies the restricted strong convexity condition (RSC) at θ if there exist two positive functions $\exists \alpha_1, \alpha_2 > 0$ and $\exists \tau_1, \tau_2 \geq 0$ of (θ, n, d) such that, for any $\Delta \in \mathbb{R}^d$,

$$\langle \nabla_{\theta} \mathbb{L}_n(\theta + \Delta) - \nabla_{\theta} \mathbb{L}_n(\theta), \Delta \rangle \geq \alpha_1 \|\Delta\|_2^2 - \tau_1 \frac{\log d}{n} \|\Delta\|_1^2, \text{ if } \|\Delta\|_2 \leq 1,$$

$$\langle \nabla_{\theta} \mathbb{L}_n(\theta + \Delta) - \nabla_{\theta} \mathbb{L}_n(\theta), \Delta \rangle \geq \alpha_2 \|\Delta\|_2 - \tau_2 \sqrt{\frac{\log d}{n}} \|\Delta\|_1, \text{ if } \|\Delta\|_2 \geq 1.$$

Note that the (RSC) property is fundamentally local and that $\alpha_k, \tau_k, k = 1, 2$ depend on the chosen θ .