Exponentially Weighted Aggregate with the Laplace Prior

Joint work with Edwin Grappin and Quentin Paris

3rd UCL Workshop on the Theory of Big Data
Mon 26-Wed 28 June 2017

Arnak S. Dalalyan

ENSAE ParisTech / CREST

Estimation of sparse vectors and matrices

Problem formulation

Regression model and prediction

- **Regression model:** we observe n feature-label pairs $(\boldsymbol{X}_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$ such that

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}^* + \xi_i \quad \Leftrightarrow \quad \boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\xi}$$

  with $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_n] \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

- **Regression vector:** The p-dimensional vector $\boldsymbol{\beta}^*$ is unknown.
- **Deterministic design:** The vectors $\boldsymbol{X}_i$ are deterministic.

Problem formulation

Regression model and prediction

- **Regression model:** we observe n feature-label pairs $(\boldsymbol{X}_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$ such that

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}^* + \xi_i \quad \Leftrightarrow \quad \boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\xi}$$

  with $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_n] \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

- **Regression vector:** The p-dimensional vector $\boldsymbol{\beta}^*$ is unknown.
- **Deterministic design:** The vectors $\boldsymbol{X}_i$ are deterministic.
- **Prediction loss:** The quality of an estimator $\widehat{\beta}_n$ is measured by

$$\ell_n(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}^*) = \frac{1}{n}\|\mathbf{X}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}_n - \boldsymbol{X}_i^\top \boldsymbol{\beta}^*)^2.$$

- **Curse of dimensionality:** OLS behaves poorly if $n \ll p$.

Problem formulation

Trace regression

- **Trace Regression:** we observe n pairs $(\boldsymbol{X}_i, Y_i) \in \mathcal{M}_{m_1,m_2} \times \mathbb{R}$ such that

$$Y_i = \langle \boldsymbol{X}_i, \mathbf{B}^* \rangle + \xi_i$$

  with $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_n] \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ and $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{Tr}(\mathbf{A}^\top \mathbf{B})$.

- The $m_1 \times m_2$-dimensional matrix $\mathbf{B}^*$ is unknown.
- **Deterministic design:** The matrices $\boldsymbol{X}_i$ are deterministic.

Problem formulation

Trace regression

- **Trace Regression:** we observe n pairs $(\boldsymbol{X}_i, Y_i) \in \mathcal{M}_{m_1,m_2} \times \mathbb{R}$ such that

$$Y_i = \langle \boldsymbol{X}_i, \mathbf{B}^* \rangle + \xi_i$$

  with $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_n] \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ and $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{Tr}(\mathbf{A}^\top \mathbf{B})$.

- The $m_1 \times m_2$-dimensional matrix $\mathbf{B}^*$ is unknown.
- **Deterministic design:** The matrices $\boldsymbol{X}_i$ are deterministic.
- **Prediction loss:** The quality of an estimator $\widehat{\mathbf{B}}_n$ is measured by

$$\ell_n(\widehat{\mathbf{B}}_n, \mathbf{B}^*) = \frac{1}{n} \sum_{i=1}^{n} (\langle \boldsymbol{X}_i, \widehat{\mathbf{B}}_n \rangle - \langle \boldsymbol{X}_i, \mathbf{B}^* \rangle)^2.$$

Sparsity, approximate sparsity and mis-specification

- **s-sparsity:** When p is large, to make consistent estimation of $\boldsymbol{\beta}^*$ feasible, we assume that

$$\|\boldsymbol{\beta}^*\|_0 = \#\{j : \beta_j^* \neq 0\} =: s \ll n \wedge p.$$

Sparsity, approximate sparsity and mis-specification

- **s-sparsity:** When p is large, to make consistent estimation of $\boldsymbol{\beta}^*$ feasible, we assume that

$$\|\boldsymbol{\beta}^*\|_0 = \#\{j : \beta_j^* \neq 0\} =: s \ll n \wedge p.$$

- **δ-approximate s-sparsity:** for some $\bar{\boldsymbol{\beta}} \in \mathbb{R}^p$,

$$\|\bar{\boldsymbol{\beta}}\|_0 \leq s \quad \text{and} \quad \|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \delta.$$

Sparsity, approximate sparsity and mis-specification

- **s-sparsity:** When p is large, to make consistent estimation of $\boldsymbol{\beta}^*$ feasible, we assume that

$$\|\boldsymbol{\beta}^*\|_0 = \#\{j : \beta_j^* \neq 0\} =: s \ll n \wedge p.$$

- **δ-approximate s-sparsity:** for some $\bar{\boldsymbol{\beta}} \in \mathbb{R}^p$,

$$\|\bar{\boldsymbol{\beta}}\|_0 \leq s \quad \text{and} \quad \|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \delta.$$

- **Mis-specification:** $\boldsymbol{\beta}^*$ is neither sparse nor approximately sparse, but there is an approximately sparse vector $\bar{\boldsymbol{\beta}}$ with a small prediction error $\ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^*)$.

  This $\bar{\boldsymbol{\beta}}$ is the (δ-approximately s-sparse) oracle.
  Can we mimic the oracle ?

Low-rank, approximate low-rank and mis-specification

- The role of sparsity in matrix estimation problems is most often played by the low rank assumption (sparsity of the spectrum).
- **Matrix sparsity:** $\mathbf{B}^*$ is r-low-rank, if

$$\text{rank}(\mathbf{B}^*) \leq r \ll m_1 \wedge m_2.$$

Low-rank, approximate low-rank and mis-specification

- The role of sparsity in matrix estimation problems is most often played by the low rank assumption (sparsity of the spectrum).
- **Matrix sparsity:** $\mathbf{B}^*$ is r-low-rank, if

$$\text{rank}(\mathbf{B}^*) \leq r \ll m_1 \wedge m_2.$$

- δ**-approximate** r**-low rank:** for some $\bar{\mathbf{B}}$,

$$\text{rank}(\bar{\mathbf{B}}) \leq r \quad \text{and} \quad \|\bar{\mathbf{B}} - \mathbf{B}^*\|_1 \leq \delta.$$

Low-rank, approximate low-rank and mis-specification

- The role of sparsity in matrix estimation problems is most often played by the low rank assumption (sparsity of the spectrum).
- **Matrix sparsity:** $\mathbf{B}^*$ is r-low-rank, if

$$\text{rank}(\mathbf{B}^*) \leq r \ll m_1 \wedge m_2.$$

- δ**-approximate** r**-low rank:** for some $\bar{\mathbf{B}}$,

$$\text{rank}(\bar{\mathbf{B}}) \leq r \quad \text{and} \quad \|\bar{\mathbf{B}} - \mathbf{B}^*\|_1 \leq \delta .$$

nuclear norm = sum of singular values

Low-rank, approximate low-rank and mis-specification

- The role of sparsity in matrix estimation problems is most often played by the low rank assumption (sparsity of the spectrum).
- **Matrix sparsity:** $\mathbf{B}^*$ is r-low-rank, if

$$\text{rank}(\mathbf{B}^*) \leq r \ll m_1 \wedge m_2.$$

- **δ-approximate r-low rank:** for some $\bar{\mathbf{B}}$,

$$\text{rank}(\bar{\mathbf{B}}) \leq r \quad \text{and} \quad \|\bar{\mathbf{B}} - \mathbf{B}^*\|_1 \leq \delta \,.$$

nuclear norm = sum of singular values

- **Mis-specification:** there is an approximately low rank matrix $\bar{\mathbf{B}}$ with a small prediction error $\ell_n(\bar{\mathbf{B}}, \mathbf{B}^*)$.

Lasso, Bayesian Lasso and EWA

Estimating sparse vectors by the lasso

- **Regression model, prediction loss:** $n \times p$ matrix $\mathbf{X}$ and $\boldsymbol{Y} \in \mathbb{R}^p$ such that

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\xi},$$

$$\ell_n(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}^*) = {}^1\!/_n \|\mathbf{X}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)\|_2^2.$$

Estimating sparse vectors by the lasso

- **Regression model, prediction loss:** $n \times p$ matrix $\mathbf{X}$ and $\boldsymbol{Y} \in \mathbb{R}^p$ such that

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\xi},$$

$$\ell_n(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}^*) = {}^1\!/_n \|\mathbf{X}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)\|_2^2.$$

- **Weak sparsity:** Good prediction can be achieved by a nearly sparse vector $\bar{\boldsymbol{\beta}}$.

Estimating sparse vectors by the lasso

- **Regression model, prediction loss:** $n \times p$ matrix $\mathbf{X}$ and $\boldsymbol{Y} \in \mathbb{R}^p$ such that

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\xi},$$

$$\ell_n(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}^*) = {}^1\!/_n \|\mathbf{X}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)\|_2^2.$$

- **Weak sparsity:** Good prediction can be achieved by a nearly sparse vector $\bar{\boldsymbol{\beta}}$.
- **Lasso [Tibshirani 96]:** ℓ_1 penalized least squares

$$\widehat{\boldsymbol{\beta}}^{\mathrm{L}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ {}^1\!/_{2n} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 . \right\}$$

- Fast computation by LARS or convex programming.

Prediction accuracy of the lasso

- **Lasso [Tibshirani 1996]:** ℓ_1 penalized least squares

$$\widehat{\boldsymbol{\beta}}^{\mathrm{L}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ {}^1\!/_{2n} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 . \right\}$$

- **Oracle inequality** [Koltchinskii, Lounici and Tsybakov 2011; D., Hebiri and Lederer 2017]: if $\lambda \geq \sqrt{({}^2\!/_n) \log(p/\alpha)}$ then with probability $\geq 1 - \alpha$

$$\ell_n(\widehat{\boldsymbol{\beta}}^{\mathrm{L}}, \boldsymbol{\beta}^*) \leq \ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) + \min_J \left\{ 4\lambda \|\bar{\boldsymbol{\beta}}_{J^c}\|_1 + ({}^3\!/_\kappa)\lambda^2 |J| \right\}.$$

Prediction accuracy of the lasso

- **Lasso [Tibshirani 1996]:** ℓ_1 penalized least squares

$$\widehat{\boldsymbol{\beta}}^{\mathrm{L}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ {}^1\!/_{2n} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 . \right\}$$

- **Oracle inequality** [Koltchinskii, Lounici and Tsybakov 2011; D., Hebiri and Lederer 2017]: if $\lambda \geq \sqrt{({}^2\!/_n) \log(p/\alpha)}$ then with probability $\geq 1 - \alpha$

$$\ell_n(\widehat{\boldsymbol{\beta}}^{\mathrm{L}}, \boldsymbol{\beta}^*) \leq \ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) + \min_J \left\{ 4\lambda \|\bar{\boldsymbol{\beta}}_{J^c}\|_1 + ({}^3\!/_\kappa)\lambda^2 |J| \right\}.$$

- **Slow rate:** For $J = \varnothing$ the rate is $\sqrt{{}^1\!/_n}$ without κ.

Prediction accuracy of the lasso

- **Lasso [Tibshirani 1996]:** ℓ_1 penalized least squares

$$\widehat{\boldsymbol{\beta}}^{\mathrm{L}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ {}^1\!/_{2n} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 . \right\}$$

- **Oracle inequality** [Koltchinskii, Lounici and Tsybakov 2011; D., Hebiri and Lederer 2017]: if $\lambda \geq \sqrt{({}^2\!/_n) \log(p/\alpha)}$ then with probability $\geq 1 - \alpha$

$$\ell_n(\widehat{\boldsymbol{\beta}}^{\mathrm{L}}, \boldsymbol{\beta}^*) \leq \ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) + \min_J \left\{ 4\lambda \|\bar{\boldsymbol{\beta}}_{J^c}\|_1 + ({}^3\!/_\kappa)\lambda^2 |J| \right\}.$$

- **Slow rate:** For $J = \varnothing$ the rate is $\sqrt{{}^1\!/_n}$ without κ.
- **Fast rate:** For getting the fast rate $({}^s\!/_n) \log(p)$, we do need the restricted eigenvalue (RE) condition $\kappa > 0$ [Bickel, Ritov, Tsybakov (2009), D., Hebiri and Lederer 2017].

Estimating sparse vectors by the Bayesian lasso

- **Regression with Gaussian noise:**
  $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\xi}$, $\boldsymbol{\xi} \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$.

Estimating sparse vectors by the Bayesian lasso

- **Regression with Gaussian noise:**
  $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\xi}$, $\boldsymbol{\xi} \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$.
- **Laplace prior:** sparsity inducing prior

$$\pi_0(\boldsymbol{\beta}) \propto \exp\Big(- (n\lambda/\sigma^2) \|\boldsymbol{\beta}\|_1 \Big).$$

Estimating sparse vectors by the Bayesian lasso

- **Regression with Gaussian noise:**
  $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\xi}$, $\boldsymbol{\xi} \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$.
- **Laplace prior:** sparsity inducing prior

$$\pi_0(\boldsymbol{\beta}) \propto \exp\big(- (n\lambda/\sigma^2) \|\boldsymbol{\beta}\|_1 \big).$$

- **Posterior** density is then

$$\widehat{\pi}_n(\boldsymbol{\beta}) \propto \exp\Big\{ - (1/2\sigma^2) \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 - (n\lambda/\sigma^2) \|\boldsymbol{\beta}\|_1 \Big\}.$$

- **Bayesian lasso** [Park and Casella 2008]: posterior mean

$$\widehat{\boldsymbol{\beta}}^{\mathrm{BL}} = \int \boldsymbol{\beta}\, \widehat{\pi}_n(d\boldsymbol{\beta}).$$

- **Remark:** the MAP coincides with the lasso.

Properties of the Bayesian lasso

- **Pros**
  - Computable by the Gibbs sampling or Langevin-type algorithm.
  - Uncertainty quantification ?
  - Less sensitive to changes in data than the lasso.

Properties of the Bayesian lasso

- **Pros**
  - Computable by the Gibbs sampling or Langevin-type algorithm.
  - Uncertainty quantification ?
  - Less sensitive to changes in data than the lasso.
- **Cons**
  - Computational guarantees weaker than for the lasso.
  - No matter the λ, $\widehat{\boldsymbol{\beta}}^{\mathrm{BL}}$ does not converge at the optimal rate [Castillo, Schmidt-Hieber and van der Vaart 2015], $\forall C > 0$

$$\widehat{\pi}_n\left(\left\{\boldsymbol{\beta} : \ell_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*) \leq \frac{Cs\log p}{n}\right\}\right) \xrightarrow[n\to\infty]{} 0, \qquad \text{a.s.}$$

even when the matrix $\mathbf{X}$ is diagonal.

Properties of the Bayesian lasso

- **Pros**
  - Computable by the Gibbs sampling or Langevin-type algorithm.
  - Uncertainty quantification ?
  - Less sensitive to changes in data than the lasso.
- **Cons**
  - Computational guarantees weaker than for the lasso.
  - No matter the λ, $\widehat{\boldsymbol{\beta}}^{\mathrm{BL}}$ does not converge at the optimal rate [Castillo, Schmidt-Hieber and van der Vaart 2015], $\forall C > 0$

$$\widehat{\pi}_n\left(\left\{\boldsymbol{\beta} : \ell_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*) \leq \frac{Cs\log p}{n}\right\}\right) \xrightarrow[n\to\infty]{} 0, \qquad \text{a.s.}$$

  even when the matrix $\mathbf{X}$ is diagonal.

One solution is to change the prior (spike-and-slab) but then the computation becomes NP-hard.

EWA and PAC-Bayes approach

- **Pseudo-posterior:** for some prior π_0, define

$$\widehat{\pi}_{n,\tau} \in \arg\min_{p(\cdot)} \left\{ \int_{\mathbb{R}^p} \frac{1}{2n} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \, p(d\boldsymbol{\beta}) + \tau \mathcal{K}(p\|\pi_0) \right\},$$

where $\tau > 0$ is a tuning parameter referred to as temperature.

EWA and PAC-Bayes approach

- **Pseudo-posterior:** for some prior π_0, define

$$\widehat{\pi}_{n,\tau} \in \arg\min_{p(\cdot)} \left\{ \int_{\mathbb{R}^p} \frac{1}{2n} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \, p(d\boldsymbol{\beta}) + \tau \mathcal{K}(p\|\pi_0) \right\},$$

  where $\tau > 0$ is a tuning parameter referred to as temperature.

- **Explicit form:**

$$\widehat{\pi}_{n,\tau}(\boldsymbol{\beta}) \propto \exp\left\{ -\frac{1}{2n\tau} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \log \pi_0(\boldsymbol{\beta}) \right\}.$$

EWA and PAC-Bayes approach

- **Pseudo-posterior:** for some prior π_0, define

$$\widehat{\pi}_{n,\tau} \in \arg\min_{p(\cdot)} \left\{ \int_{\mathbb{R}^p} \frac{1}{2n} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \, p(d\boldsymbol{\beta}) + \tau \mathcal{K}(p\|\pi_0) \right\},$$

  where $\tau > 0$ is a tuning parameter referred to as temperature.

- **Explicit form:**

$$\widehat{\pi}_{n,\tau}(\boldsymbol{\beta}) \propto \exp\left\{ -\frac{1}{2n\tau} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \log \pi_0(\boldsymbol{\beta}) \right\}.$$

- **Exponentially weighted aggregate (EWA):**

$$\widehat{\boldsymbol{\beta}}_{\tau}^{\mathrm{EWA}} = \int_{\mathbb{R}^p} \boldsymbol{\beta} \, \widehat{\pi}_{n,\tau}(d\boldsymbol{\beta}).$$

- If $\tau = \sigma^2/n$, then the EWA coincides with the Bayes estimator.

EWA and the Laplace prior

- One can used a scaled Laplace prior in the EWA.
- The resulting (suitably re-parameterized) pseudo-posterior is

$$\widehat{\pi}_{n,\tau}(\boldsymbol{\beta}) \propto \exp\Big\{ -\frac{1}{2n\tau}\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 - \frac{\lambda}{\tau}\|\boldsymbol{\beta}\|_1 \Big\}.$$

- This estimator interpolates the lasso and the Bayesian lasso:
  - For $\tau = 0$, the EWA coincides with the lasso.
  - For $\tau = \sigma^2/n$, the EWA coincides with the Bayes lasso.

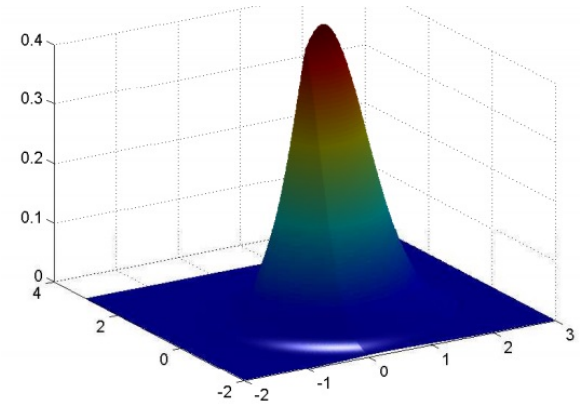
EWA and the Laplace prior

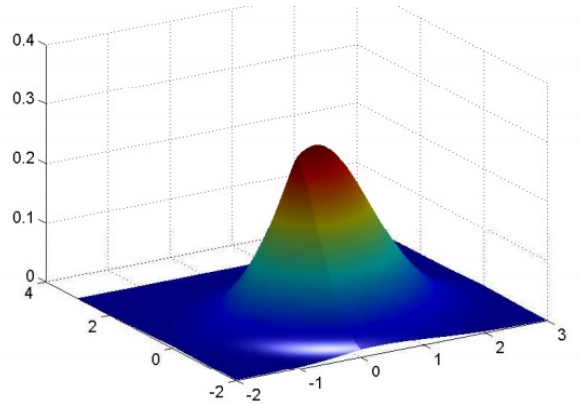
- One can used a scaled Laplace prior in the EWA.
- The resulting (suitably re-parameterized) pseudo-posterior is

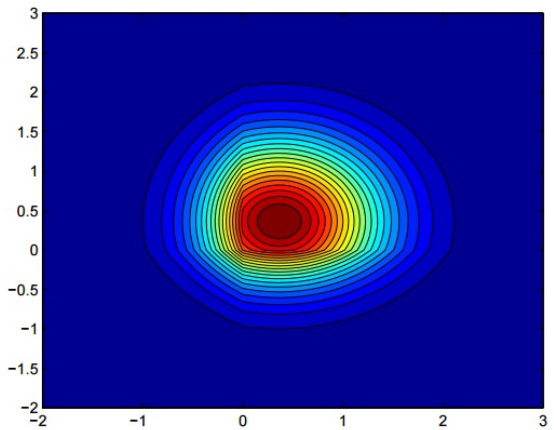
$$\widehat{\pi}_{n,\tau}(\boldsymbol{\beta}) \propto \exp\Big\{ -\frac{1}{2n\tau}\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 - \frac{\lambda}{\tau}\|\boldsymbol{\beta}\|_1 \Big\}.$$

- This estimator interpolates the lasso and the Bayesian lasso:
  - For $\tau = 0$, the EWA coincides with the lasso.
  - For $\tau = \sigma^2/n$, the EWA coincides with the Bayes lasso.
- **Questions:**
  - Is the EWA suitable for prediction under the sparsity scenario? If it is, what is the range of temperature τ providing good prediction accuracy?
  - How do the statistical properties of the EWA compare with those of the lasso?

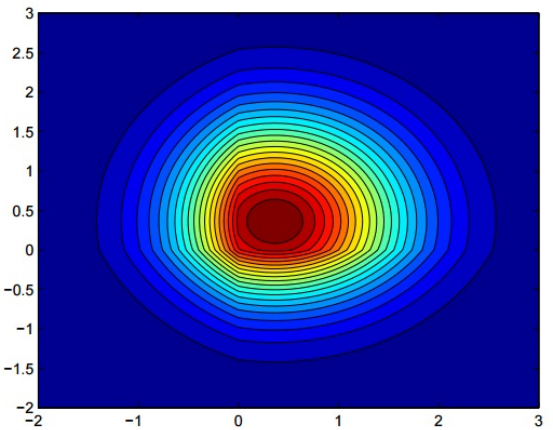
EWA and the Laplace prior







$\tau = 0.5$



$\tau = 0.8$

Main results

Oracle inequality for the EWA

Theorem 1

Assume that $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$ and that $\max_{j \in [p]} {}^1\!/_n \|\boldsymbol{x}^j\|_2^2 \leq 1$. Suppose, in addition, that $\lambda \geq 2\sigma\sqrt{({}^2\!/_n)\log(p/\alpha)}$, for some $\alpha \in (0,1)$. Then, with probability at least $1-\alpha$,

$$\ell_n(\widehat{\boldsymbol{\beta}}_\tau^{\mathrm{EWA}}, \boldsymbol{\beta}^*) \leq \min_{J \subset [p]} \left\{ \ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) + 4\lambda \|\bar{\boldsymbol{\beta}}_{J^c}\|_1 + \frac{9\lambda^2 |J|}{4\kappa} \right\} + 2p\tau. \quad (1)$$

Oracle inequality for the EWA

Theorem 1

Assume that $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$ and that $\max_{j\in[p]} {}^1\!/_n \|\boldsymbol{x}^j\|_2^2 \leq 1$. Suppose, in addition, that $\lambda \geq 2\sigma\sqrt{({}^2\!/_n)\log(p/\alpha)}$, for some $\alpha \in (0,1)$. Then, with probability at least $1-\alpha$,

$$\ell_n(\widehat{\boldsymbol{\beta}}_\tau^{\mathrm{EWA}}, \boldsymbol{\beta}^*) \leq \min_{J\subset[p]} \left\{ \ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) + 4\lambda \|\bar{\boldsymbol{\beta}}_{J^c}\|_1 + \frac{9\lambda^2|J|}{4\kappa} \right\} + 2p\tau. \quad (1)$$

Remarks

1. This is a sharp OI, since the leading constant is 1.
2. This risk bound extends the risk bounds available for the lasso [D., Hebiri and Lederer 2017] to the EWA.
3. The last term in the above risk bound, $2p\tau$, reflects the influence of the temperature τ; if $\tau = \sigma^2/(pn)$ this term is negligible.

Pseudo-posetrior concentration

Theorem 2

Assume that $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$ and that $\max_{j \in [p]} {}^1\!/_n \|\boldsymbol{x}^j\|_2^2 \leq 1$. Assume that $\lambda \geq 2\sigma\sqrt{({}^2\!/_n)\log({}^p\!/_\alpha)}$. Then, with probability at least $1-\alpha$ on data, the pseudo-posterior $\widehat{\pi}_{n,\tau}$ with the Laplace prior puts at least a weight $1 - 2e^{-\sqrt{p}/16}$ on the set

$$\left\{ \boldsymbol{\beta} : \ell_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*) \leq \min_{J \subset [p]} \left(\ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) + 4\lambda \|\bar{\boldsymbol{\beta}}_{J^c}\|_1 + \frac{9\lambda^2 |J|}{2\kappa} \right) + 8p\tau \right\}.$$

Pseudo-posetrior concentration

Theorem 2

Assume that $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$ and that $\max_{j \in [p]} {}^1\!/_n \|\boldsymbol{x}^j\|_2^2 \leq 1$. Assume that $\lambda \geq 2\sigma\sqrt{({}^2\!/_n)\log({}^p\!/_\alpha)}$. Then, with probability at least $1-\alpha$ on data, the pseudo-posterior $\widehat{\pi}_{n,\tau}$ with the Laplace prior puts at least a weight $1 - 2e^{-\sqrt{p}/16}$ on the set

$$\left\{ \boldsymbol{\beta} : \ell_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*) \leq \min_{J \subset [p]} \left(\ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) + 4\lambda \|\bar{\boldsymbol{\beta}}_{J^c}\|_1 + \frac{9\lambda^2 |J|}{2\kappa} \right) + 8p\tau \right\}.$$

Remarks

1. A random sample from $\widehat{\pi}_{n,\tau}$ is almost as good as the pseudo-posterior mean.
2. It might be easier to sample from $\widehat{\pi}_{n,\tau}$ rather than to compute the pseudo-posterior mean.
3. The method is sparsity-adaptive.

Main idea of the proof of Theorem 1

- Let $V_n(\boldsymbol{\beta})$ be the penalized log-likelihood.
- The lasso estimator and the EWA are

$$\widehat{\boldsymbol{\beta}}^{\mathrm{L}} \in \arg\min_{\boldsymbol{\beta}} V_n(\boldsymbol{\beta}), \qquad \widehat{\boldsymbol{\beta}}_{\tau}^{\mathrm{EWA}} = \frac{1}{C} \int_{\mathbb{R}^p} \boldsymbol{\beta} \, \exp\{-V_n(\boldsymbol{\beta})/\tau\} \, d\boldsymbol{\beta}.$$

- **Known:** the proof of the OI for the lasso builds on

$$\nabla V_n(\widehat{\boldsymbol{\beta}}^{\mathrm{L}}) = \mathbf{0} \qquad [\text{rigorously } \mathbf{0} \in \partial V_n(\widehat{\boldsymbol{\beta}}^{\mathrm{L}})].$$

- **New:** the analogue of the above condition is:

$$\int_{\mathbb{R}^p} \nabla V_n(\boldsymbol{\beta}) \, \widehat{\pi}_n(\boldsymbol{\beta}) \, d\boldsymbol{\beta} = C \int_{\mathbb{R}^p} \nabla \big[\widehat{\pi}_n(\boldsymbol{\beta})\big] \, d\boldsymbol{\beta} = \mathbf{0}.$$

- Additionally, for every $j \in [p]$, $\displaystyle\int_{\mathbb{R}^p} \nabla \big[\beta_j \widehat{\pi}_n(\boldsymbol{\beta})\big] \, d\boldsymbol{\beta} = \mathbf{0}$.

Main ingredient of the proof of Theorem 2

Magic theorem [Bobkov and Madiman 2011]

Let $\widehat{\pi}_n(\boldsymbol{u}) \propto \exp(-V_n(\boldsymbol{u})/\tau)$ be a log-concave probability density[a] and let $\boldsymbol{\beta}$ be a random vector drawn from $\widehat{\pi}_n$. Then, for any $t > 0$, the inequality

$$V_n(\boldsymbol{\beta}) \le \int_{\mathbb{R}^p} V_n(\boldsymbol{u})\, \widehat{\pi}_n(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} + \sqrt{p}\tau t$$

holds with probability at least $1 - 2e^{-t/16}$.

[a] This means that V_n is a convex function.

Lasso versus EWA

- Both are computable in polynomial time, but the lasso computation is faster.
- Theoretical guarantees for the lasso and the EWA are equally good, provided that the temperature τ is small.
- The EWA is less sensitive to the variations in the data.
  - The SURE for the lasso [Tibshirani and Taylor, 2012] is given by

$$\widehat{R}^{\mathrm{L}}(\lambda) = \frac{1}{n}\|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{\mathrm{L}}(\lambda)\|_2^2 - \sigma^2 + \frac{2\sigma^2}{n}\mathsf{rank}(\mathbf{X}_{\mathcal{A}(\lambda)}),$$

    where $\mathcal{A}(\lambda) = \{j \in [p] : \beta_j^{\mathrm{L}}(\lambda) \neq 0\}$.
    This function is even not continuous.
  - The SURE for the EWA is

$$\widehat{R}^{\mathrm{EWA}}(\lambda) = \frac{1}{n}\|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\tau}^{\mathrm{EWA}}(\lambda)\|_2^2 - \sigma^2 + \frac{2\sigma^2}{n^2\tau}\mathbf{Var}_{\widehat{\pi}_n(\lambda)}[\mathbf{X}\boldsymbol{\beta}].$$

OI for the EWA in the matrix trace regression

Recall that the ℓ_1-norm $\|\beta\|_1$ is replaced by the nuclear norm $\|\mathbf{B}\|_1$.

Let us define $\upsilon_{\mathbf{X}} = \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top \right\|^{1/2} \bigvee \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i^\top \mathbf{X}_i \right\|^{1/2}$.

Theorem 3

Assume that $\lambda \geq 2\sigma\upsilon_{\mathbf{X}}\{2/n \log((m_1+m_2)/\alpha)\}^{1/2}$, for some $\alpha \in (0,1)$. Then, with probability at least $1-\alpha$, the matrix $\widehat{\mathbf{B}}^{\mathrm{EWA}}$ satisfies

$$\ell_n(\widehat{\mathbf{B}}^{\mathrm{EWA}}, \mathbf{B}^*) \leq \min_J \left\{ \ell_n(\bar{\mathbf{B}}, \mathbf{B}^*) + 4\lambda \|\mathcal{P}_{J^c}(\bar{\mathbf{B}})\|_1 + \frac{9\lambda^2 |J|}{4\kappa} \right\} + 2m_1 m_2 \tau,$$

where the min is over all subsets $J \subset [\mathrm{rank}(\bar{\mathbf{B}})]$.

We have a similar result on pseudo-posterior concentration.

Summary and outlook

- We have obtained nonasymptotic risk bounds for the EWA with Laplace prior, both in vector- and matrix-regression models.
- If the temperature $\tau \leq \sigma^2/(np)$, then the EWA is as good as the lasso.
- The pseudo-posterior has nice concentration properties.
- This can be extended to random design, but more technical and additional terms appear.
- Current/future work:
  - Computational complexity of drawing from $\widehat{\pi}_{n,\tau}$.
  - Assessing the sensitivity of the EWA as compared to the lasso.


thank you
danke
謝謝
ngiyabonga
teşekkür ederim
gracias
merci
obrigado
bedankt
спасибо
dziękuję
sukriya
kop khun krap
go raibh maith agat
mochchakkeram
terima kasih
grazie
arigatô
dakujem
takk
taiku
tapadh leat
мерси
감사합니다
ありがとう
ευχαριστώ
sagolun
hvala
mauruuru
dank je
рахмат
Баярлалаа
enkosi
sobodi
dekuji
mèsi
rahmat
tanemirt
rahmet
diolch
dhanyavadagalu
shukriya
xièxie
najis tuke
kam sah hamnida
trugarez
хвала
asante
manana
obrigada
tenki
mamnun
djiere dieuf
sulpáy
chnorakaloutioun
gratias ago
gràcies
blagodaram
misaotra
matondo
paldies
grazzi
mahalo
mersi
barka
welalin
tack
spas
kia ora
vinaka
спасибі
faafetai lava
kiitos
dankie
dhanyavad
nandri
nanni
bayarlalaa
gracie
köszönöm
akun dankon
aciu
дякую