Adaptive Bayesian estimation and its self-informative limit in an indirect sequence space model

Xavier Loizeau, joint work with Jan Johannes
Ruprecht-Karls-Universität Heidelberg

The Gaussian sequence space model

Consider an indirect Gaussian sequence space model consisting of:

- an unknown parameter of interest $\left(\theta_j^\circ\right)_{j\in\mathbb{N}} = \theta^\circ$,
- a decreasing multiplicative sequence $\left(\lambda_j\right)_{j\in\mathbb{N}} = \lambda$ converging to 0,
- observations $\left(Y_j\right)_{j\in\mathbb{N}} = Y$, contaminated by an additive independent centered Gaussian noise with variance n^{-1},
$$Y = \left(\theta_j^\circ \cdot \lambda_j + \sqrt{n}^{-1} \cdot \xi_j\right)_{j\in\mathbb{N}}, \quad (\xi)_{j\in\mathbb{N}} \sim_{iid} \mathcal{N}(0,1).$$

The goal is to recover θ° and derive an upper bound.

The frequentist model selection

For any index j, an unbiased estimator of θ_j° is Y_j/λ_j. Hence, an intuitive class of estimators are the projection estimators: $\tilde{\theta}^m = \left(Y_j/\lambda_j \mathbb{1}_{\{j\le m\}}\right)_{j\in\mathbb{N}}$ with m in $\mathbb{N}$. The model selection method offers a data driven way to select m in this context:

$$G_n := \max\left\{1 \le j \le n : n^{-1}\lambda_j^{-2} \le \lambda_1^{-2}\right\},$$

$$\widehat{m} := \underset{m\in[\![1,G_n]\!]}{\arg\min}\left\{3m - \sum_{j=1}^m Y_j^2\right\}, \qquad \widehat{\theta} := \left(\tilde{\theta}_j^{\widehat{m}}\right)_{j\in\mathbb{N}}.$$

It is shown in Massart [2003], in the direct case, that this estimator is consistent, converges in probability and $\mathbb{L}^2$-norm, noted $\|\cdot\|$, with minimax optimal rate over some Sobolev ellipsoid:

$$\Theta^\circ := \Theta^\circ(\mathfrak{a}, L^\circ)\left\{\theta : \sum_{j=1}^\infty \frac{1}{\mathfrak{a}_j}\theta_j^2 < L^\circ\right\}.$$

Bayesian paradigm, iterated posterior distribution and self informative limit

We adopt a Bayesian point of view:

- the parameter $\boldsymbol{\theta}$ is a random variable with prior $\mathbb{P}_{\boldsymbol{\theta}}$,
- given $\boldsymbol{\theta}$, the likelihood of Y is $\mathbb{P}_{Y|\boldsymbol{\theta}}^n = \mathcal{N}\left(\boldsymbol{\theta}\lambda, n^{-1}\mathbb{I}\right)$,
- we are interested in the posterior distribution $\mathbb{P}_{\boldsymbol{\theta}^n|Y} \propto \mathbb{P}_{Y|\boldsymbol{\theta}}^n \cdot \mathbb{P}_{\boldsymbol{\theta}}$.

In the spirit of Bunke and Johannes [2005], we then generate a posterior family by introducing an iteration parameter η:

- for $\eta = 1$, the prior distribution is $\mathbb{P}_{\boldsymbol{\theta}^1} = \mathbb{P}_{\boldsymbol{\theta}}$, the likelihood $\mathbb{P}_{Y^1|\boldsymbol{\theta}^1}^n = \mathbb{P}_{Y|\boldsymbol{\theta}}^n$ and the posterior distribution is $\mathbb{P}_{\boldsymbol{\theta}^1|Y^1}^n = \mathbb{P}_{\boldsymbol{\theta}|Y}^n$,
- for $\eta = 2$, we take the posterior for $\eta = 1$ as prior, hence, the prior distribution is $\mathbb{P}_{\boldsymbol{\theta}^2}^n = \mathbb{P}_{\boldsymbol{\theta}^1|Y^1}^n$, the likelihood is kept the same $\mathbb{P}_{Y^2|\boldsymbol{\theta}^2}^n = \mathbb{P}_{Y|\boldsymbol{\theta}}^n$ and we compute the posterior distribution with the same observations Y, which we note $\mathbb{P}_{\boldsymbol{\theta}^2|Y^2}^n$,
- ...
- for any value of $\eta > 1$, the prior is $\mathbb{P}_{\boldsymbol{\theta}^\eta}^n = \mathbb{P}_{\boldsymbol{\theta}^{\eta-1}|Y^{\eta-1}}^n$ and we compute the posterior with the same likelihood $\mathbb{P}_{Y^\eta|\boldsymbol{\theta}^\eta}^n = \mathbb{P}_{Y|\boldsymbol{\theta}}^n$ and same observation Y which gives $\mathbb{P}_{\boldsymbol{\theta}^\eta|Y^\eta}^n$.

This iteration procedure corresponds to giving more and more weight to the observations and make the prior knowledge vanish.

Within this framework we define the family of estimators:

$$\widehat{\theta}^{(\eta)} := \mathbb{E}_{\boldsymbol{\theta}^\eta|Y^\eta}^n[\boldsymbol{\theta}],$$

and call self-informative limit the limit of the estimate with $\eta \to \infty$.

We are interested in the behavior of the family $\left(\mathbb{P}_{\boldsymbol{\theta}^\eta|Y^\eta}^n\right)_{\eta\in\mathbb{N}^\star}$ as n and/or η tend to infinite.

In particular, the question of oracle and minimax concentration (resp. convergence) is answered for any element of the family of posterior distributions (resp. posterior means), including when η tends to infinite.

Hierarchical prior

- Consider a random hyper-parameter M, with values in a subset of $\mathbb{N}$, acting like a threshold:
$$\forall j > m, \quad \mathbb{P}_{\boldsymbol{\theta}_j|M=m} = \delta_0,$$
$$\forall j \le m, \quad \mathbb{P}_{\boldsymbol{\theta}_j|M=m} = \mathcal{N}(0,1).$$
- if we denote $\mathbb{P}_M$ the distribution of M (to be specified later), then
$$\mathbb{P}_{\boldsymbol{\theta}|Y}^n = \sum_{m\in\mathbb{N}} \mathbb{P}_{\boldsymbol{\theta}|M=m,Y}^n \cdot \mathbb{P}_{M=m|Y}^n.$$
- Hence, given M, the posterior is
$$\forall j > m, \quad \boldsymbol{\theta}_j|M=m,Y \sim \delta_0,$$
$$\forall j \le m, \quad \boldsymbol{\theta}_j|M=m,Y \sim \mathcal{N}\left(\frac{Y_j \cdot n \cdot \lambda_j}{1+n\cdot\lambda_j^2}, \frac{1}{1+n\cdot\lambda_j^2}\right).$$

Remark: the family of hierarchical priors with deterministic threshold M is called family of sieve priors.

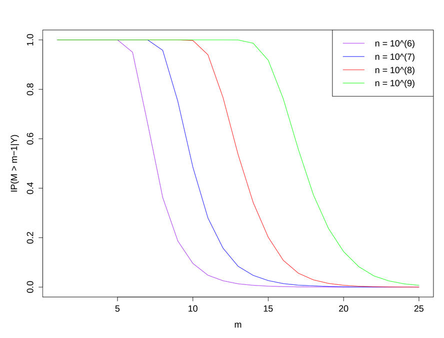


Figure: Survival function of M for different values of n

Existing results

In Johannes et al. [2016], under a pragmatic Bayesian point of view; that is, the existence of a true parameter θ° is accepted; it is shown that, by choosing $\mathbb{P}_M$ suitably:

- the estimator $\widehat{\theta}^{(1)}$ converges with,
  - oracle optimal rate for the quadratic risk which means, $\forall\theta^\circ\in\Theta^\circ, \exists C^\circ\in[1,\infty[: \forall n\in\mathbb{N}, \exists\Phi_n^\circ\in\mathbb{R}$:
$$\inf_{m\in\mathbb{N}} \mathbb{E}_{\theta^\circ}^n\left[\left\|\tilde{\theta}^m - \theta^\circ\right\|^2\right] \ge \Phi_n^\circ,$$
$$\mathbb{E}_{\theta^\circ}^n\left[\left\|\widehat{\theta}^{(1)} - \theta^\circ\right\|^2\right] \le C^\circ\Phi_n^\circ;$$
  - minimax optimal rate for the maximal risk over Θ°, that is to say, $\exists C^\star\in[1,\infty[: \forall n\in\mathbb{N}, \exists\Phi_n^\star\in\mathbb{R}$:
$$\inf_{\tilde{\theta}} \sup_{\theta^\circ\in\Theta^\circ} \mathbb{E}_{\theta^\circ}^n\left[\left\|\tilde{\theta} - \theta^\circ\right\|^2\right] \ge \Phi_n^\star,$$
$$\sup_{\theta^\circ\in\Theta^\circ} \mathbb{E}_{\theta^\circ}^n\left[\left\|\widehat{\theta}^{(1)} - \theta^\circ\right\|^2\right] \le C^\star\Phi_n^\star,$$
  where $\inf_{\tilde{\theta}}$ is taken over all possible estimators of θ°;
- the posterior distribution concentrates with,
  - oracle optimal rate for the quadratic loss which means, $\forall\theta^\circ\in\Theta^\circ, \exists K^\circ\in[1,\infty[$:
$$\lim_{n\to\infty} \mathbb{E}_{\theta^\circ}^n\left[\mathbb{P}_{\boldsymbol{\theta}^1|Y^1}^n\left(\|\boldsymbol{\theta} - \theta^\circ\|^2 \le K^\circ\Phi_n^\circ\right)\right] = 1;$$
  - minimax optimal rate Θ°, that is to say, for any unbounded sequence $K_n\in\mathbb{R}^{\mathbb{N}}$:
$$\lim_{n\to\infty} \sup_{\theta^\circ\in\Theta^\circ} \mathbb{E}_{\theta^\circ}^n\left[\mathbb{P}_{\boldsymbol{\theta}^1|Y^1}^n\left(\|\boldsymbol{\theta} - \theta^\circ\|^2 \le K_n\Phi_n^\star\right)\right] = 1.$$

Iterated posterior distributions

Note that in the framework of our hierarchical prior, we have:

$$\mathbb{P}_{\boldsymbol{\theta}^\eta|Y^\eta}^n = \sum_{m\in\mathbb{N}} \mathbb{P}_{\boldsymbol{\theta}^\eta|M^\eta=m,Y^\eta}^n \cdot \mathbb{P}_{M^\eta=m|Y^\eta}^n,$$
$$\widehat{\theta}^{(\eta)} = \left(\mathbb{E}_{\boldsymbol{\theta}^\eta|M^\eta\ge j,Y^\eta}^n[\boldsymbol{\theta}_j] \cdot \mathbb{P}_{M^\eta|Y^\eta}^n(M^\eta \ge j)\right)_{j\in\mathbb{N}}.$$

Hence, we first compute $\boldsymbol{\theta}_j^\eta|M^\eta, Y^\eta$:

$$\forall j\in\mathbb{N}, \quad \boldsymbol{\theta}_j^\eta|M^\eta \ge j, Y^\eta \sim \mathcal{N}\left(\frac{\eta\cdot Y_j\cdot n\cdot\lambda_j}{1+\eta\cdot n\cdot\lambda_j^2}, \frac{1}{1+n\cdot\eta\cdot\lambda_j^2}\right),$$
$$\boldsymbol{\theta}_j^\eta|M^\eta < j, Y^\eta \sim \delta_0;$$

and then fix the distribution of M^1: $\forall m\in[\![1,G_n]\!]$,

$$\mathbb{P}_{M^1}(M=m) \propto \exp\left(-3\cdot\eta\cdot\frac{m}{2}\right)\cdot\prod_{j=1}^m\left(1+n\cdot\eta\cdot\lambda_j^2\right)^2.$$

Which gives the family of posterior distributions:

$$\mathbb{P}_{M^\eta|Y^\eta}^n(m) \propto \exp\left[-\frac{\eta}{2}\left(3m - \sum_{j=1}^m \frac{\eta\left(Y_j\cdot n\cdot\lambda_j^2\right)^2}{1+\eta\cdot n\cdot\lambda_j^2}\right)\right].$$

Self informative limit and model selection

Consider the limit of the family of posteriors as η tends to infinite:

$$\lim_{\eta\to\infty} \mathbb{P}_{\boldsymbol{\theta}^\eta|M^\eta=m,Y^\eta}^n = \delta_{\tilde{\theta}^m},$$

where $\tilde{\theta}^m$ is the projection estimator on the first m dimensions.

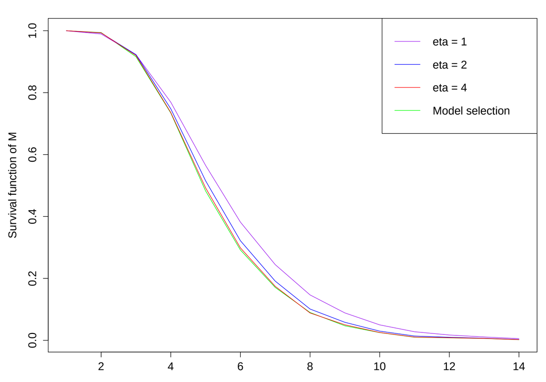
The distribution of M tends to a point mass:

$$\lim_{\eta\to\infty} \mathbb{P}_{M^\eta|Y^\eta}^n = \delta_{\widehat{m}},$$

where $\widehat{m}$ is the choice given by the frequentist model selection presented earlier.

The self-informative limit is equal to the model selection estimator, $\widehat{\theta}$, presented above.

Figure: Survival function of M for different values of η



Notations

Define the following quantities:

$$\mathfrak{b}_m := \sum_{j=m+1}^\infty (\theta^\circ)^2, \quad \Lambda_j := \lambda_j^{-2}, \quad m\cdot\overline{\Lambda}_m := \sum_{j=1}^m \Lambda_j,$$

$$m_n^\circ := \underset{m\in[\![1,G_n]\!]}{\arg\min}\left[\mathfrak{b}_m \vee n^{-1}m\overline{\Lambda}_m\right], \quad \Phi_n^\circ := \left[\mathfrak{b}_{m_n^\circ} \vee n^{-1}m_n^\circ\overline{\Lambda}_{m_n^\circ}\right],$$

$$m_n^\star := \underset{m\in[\![1,G_n]\!]}{\arg\min}\left[\mathfrak{a}_m \vee n^{-1}m\overline{\Lambda}_m\right], \quad \Phi_n^\star := \left[\mathfrak{a}_{m_n^\star} \vee n^{-1}m_n^\star\overline{\Lambda}_{m_n^\star}\right].$$

It is important to note that:

- $\Phi_n^\star$ is the minimax optimal rate over Θ°,
- Φ_n° is the oracle optimal rate over the projection estimators.

Set of assumptions

Define the following assumptions:

$$(\mathbb{H}_\lambda) : \exists a\in\mathbb{R}_+, c\ge 1 : \quad \forall j\in\mathbb{N}, \quad \left(\frac{1}{c}j^{-a} \le \lambda_j \le cj^{-a}\right)$$

$$(\mathbb{H}_1) : 0 < \inf_{n\in\mathbb{N}}\left\{\frac{\left[\mathfrak{b}_{m_n^\circ} \wedge n^{-1}m_n^\circ\overline{\Lambda}_{m_n^\circ}\right]}{\left[\mathfrak{b}_{m_n^\circ} \vee n^{-1}m_n^\circ\overline{\Lambda}_{m_n^\circ}\right]}\right\} \le 1$$

$$(\mathbb{H}_2) : 0 < \inf_{n\in\mathbb{N}}\left\{\frac{\left[\mathfrak{a}_{m_n^\star} \wedge n^{-1}m_n^\star\overline{\Lambda}_{m_n^\star}\right]}{\left[\mathfrak{a}_{m_n^\star} \vee n^{-1}m_n^\star\overline{\Lambda}_{m_n^\star}\right]}\right\} \le 1$$

Note that under $(\mathbb{H}_\lambda)$, there exist a constant L such that,

$$\forall m\in\mathbb{N}, \quad \Lambda_m \le L\overline{\Lambda}_m.$$

Concentration results for the threshold parameter M

For any η in $\overline{\mathbb{N}}$, we have the following results:

1. Under assumptions $(\mathbb{H}_1)$ and $(\mathbb{H}_\lambda)$, define
$$G_n^- := \min\left\{m\in[\![1,m_n^\circ]\!] : \mathfrak{b}_m \le 9L\Phi_n^\circ\right\},$$
$$G_n^+ := \max\left\{m\in[\![m_n^\circ, G_n]\!] : (m - m_n^\circ)\,n^{-1} \le 3\Lambda_{m_n^\circ}^{-1}\Phi_n^\circ\right\},$$
and we then have the following concentration for M,
$$\mathbb{P}_{M^\eta|Y^\eta}^n\left[M > G_n^+\right] \le \exp\left[-\frac{5m_n^\circ}{9L} + \log(G_n)\right],$$
$$\mathbb{P}_{M^\eta|Y^\eta}^n\left[M < G_n^-\right] \le \exp\left[-\frac{7m_n^\circ}{9} + \log(G_n)\right],$$
this means that M^η tends to select an oracle optimal threshold;
2. whereas under $(\mathbb{H}_2)$ and $(\mathbb{H}_\lambda)$, we define
$$G_n^{\star-} := \min\left\{m\in[\![1,m_n^\star]\!] : \quad \mathfrak{b}_m \le 9\left(1\vee L^\circ\right)L\Phi_n^\star\right\},$$
$$G_n^{\star+} := \max\left\{m\in[\![m_n^\star, G_n]\!] : (m - m_n^\star)\,n^{-1} \le 3\Lambda_{m_n^\star}^{-1}\left(1\vee L^\circ\right)\Phi_n^\star\right\},$$
and the following concentration stands,
$$\mathbb{P}_{M^\eta|Y^\eta}^n\left[M > G_n^{\star+}\right] \le \exp\left[-\frac{5\left(1\vee L^\circ\right)m_n^\star}{9L} + \log(G_n)\right],$$
$$\mathbb{P}_{M^\eta|Y^\eta}^n\left[M < G_n^{\star-}\right] \le \exp\left[-\frac{7\left(1\vee L^\circ\right)m_n^\star}{9} + \log(G_n)\right],$$
which means that M^η tends to select a minimax optimal threshold.

Concentration results for θ

For any η in $\mathbb{N}$, we have the following results:

1. under assumptions $(\mathbb{H}_1)$ and $(\mathbb{H}_\lambda)$, for all θ° in Θ°, there exist $K^\circ \ge 1$ and $C^\circ > 1$ such that we have
$$\lim_{n\to\infty} \inf_{\mathbb{Q}_{\boldsymbol{\theta}}} \mathbb{E}_{\theta^\circ}^n\left[\mathbb{Q}_{\boldsymbol{\theta}|Y}^n\left(\|\boldsymbol{\theta} - \theta^\circ\|^2 \ge \Phi_n^\circ\right)\right] = 1,$$
$$\lim_{n\to\infty} \mathbb{E}_{\theta^\circ}^n\left[\mathbb{P}_{\boldsymbol{\theta}^\eta,M^\eta|Y^\eta}^n\left(\left(K^\circ\right)^{-1}\Phi_n^\circ \le \|\boldsymbol{\theta} - \theta^\circ\|^2 \le K^\circ\Phi_n^\circ\right)\right] = 1,$$
$$\mathbb{E}_{\theta^\circ}^n\left[\left\|\widehat{\theta}^{(\eta)} - \theta^\circ\right\|^2\right] \le C^\circ\Phi_n^\circ,$$
where $\inf_{\mathbb{Q}_{\boldsymbol{\theta}}}$ is taken over all possible sieve priors; establishing oracle optimal concentration and convergence of the posterior and Bayes estimate, respectively;
2. whereas under $(\mathbb{H}_2)$ and $(\mathbb{H}_\lambda)$, for a finite constant $C^\star \ge 1$ and any unbounded sequence K_n, we have
$$\lim_{n\to\infty} \inf_{\mathbb{Q}_{\boldsymbol{\theta}}} \sup_{\theta^\circ\in\Theta^\circ} \mathbb{E}_{\theta^\circ}^n\left[\mathbb{Q}_{\boldsymbol{\theta}|Y}^n\left(\|\boldsymbol{\theta} - \theta^\circ\|^2 \ge \Phi_n^\star\right)\right] = 1,$$
$$\lim_{n\to\infty} \sup_{\theta^\circ\in\Theta^\circ} \mathbb{E}_{\theta^\circ}^n\left[\mathbb{P}_{\boldsymbol{\theta}^\eta,M^\eta|Y^\eta}^n\left(\|\boldsymbol{\theta} - \theta^\circ\|^2 \le K_n\Phi_n^\star\right)\right] = 1,$$
$$\sup_{\theta^\circ\in\Theta^\circ} \mathbb{E}_{\theta^\circ}^n\left[\left\|\widehat{\theta}^{(\eta)} - \theta^\circ\right\|^2\right] \le C^\star\Phi_n^\star,$$
where $\inf_{\mathbb{Q}_{\boldsymbol{\theta}}}$ is taken over all possible sieve priors; establishing minimax optimal concentration and convergence of the posterior and Bayes estimate, respectively.

Note that in the case of $\eta\to\infty$, those results are still true and that the concentration corresponds to the convergence in probability as

$$\lim_{\eta\to\infty} \mathbb{E}_{\theta^\circ}^n\left[\mathbb{P}_{\boldsymbol{\theta}^\eta,M^\eta|Y^\eta}^n\left(\|\boldsymbol{\theta} - \theta^\circ\|^2 \le K_n\Phi_n\right)\right] = \mathbb{P}_{\theta^\circ}^n\left[\|\widehat{\theta} - \theta^\circ\|^2 \le K_n\Phi_n\right].$$

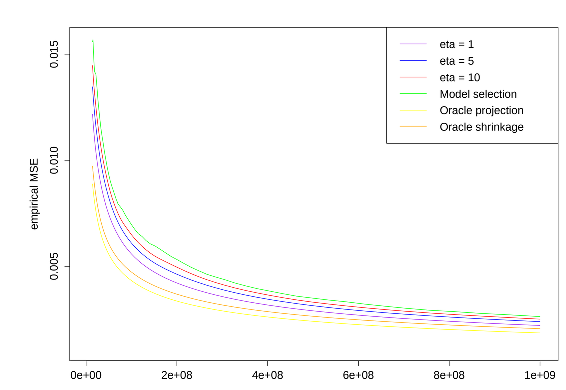


Figure: Estimated mean of the quadratic error of the Bayes estimate for θ° polynomial.

Bibliography

Olaf Bunke and Jan Johannes. Selfinformative limits of bayes estimates and generalized maximum likelihood. *Statistics*, 39(6):483–502, July 2005.

Jan Johannes, Anna Simoni, and Rudolf Schenk. Adaptive bayesian estimation in indirect gaussian sequence space model. 2016.

Pascal Massart. Concentration inequalities and model selection. In Springer, editor, *Ecole d'Été de Probabilités de Saint-Flour XXXIII*, volume 1869, 2003.

Email: loizeau@math.uni-heidelberg.de